PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN BIOLOGÍA COMPUTACIONAL

Análisis de expresión génica en genes del maní (*Arachis hypogaea* L.), relacionados a reacciones alergénicas en el ser humano.

Trabajo previo a la obtención del título de Magíster en Biología Computacional

IVONNE DE LOS ANGELES VACA SUQUILLO

Tutor: MSc. Laura González PhDc.

Quito, 2023-2024

AGRADECIMIENTOS

Agradezco infinitamente a Dios, a mis padres y a mi compañero de vida, por su amor y apoyo incondicional, en cada día. Gracias por ser mi fortaleza.

Agradezco a todos los integrantes del grupo de investigación BIOARN, de la Universidad Politécnica Salesiana, por su enorme apoyo, sin este no habría sido posible alcanzar esta meta.

Un enorme agradecimiento a mi querida tutora Laura, por su gran apoyo en el desarrollo de la presente investigación. Mis deseos de éxitos en su vida.

Agradezco a mis maestros y compañeros de maestría, por las lecciones aprendidas y por el increíble trabajo en equipo.

Gracias a todos mis amigos, hermanos y hermanas de vida, que siempre confiaron en mí y me alentaron con sus palabras, gracias por su cariño sincero.

"Todos somos maestros y aprendices, cada día"

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

INDICE DE ANEXOS

1. Introducción

1.1. Problema

El maní (*Arachis hypogaea* L.) es una planta originaria de Sudamérica, ocupa el cuarto lugar de importancia en el mundo para la producción de aceite vegetal comestible; su semilla contiene antioxidantes, grasas, carbohidratos, fibras crudas, vitaminas, minerales y proteínas (Montero, 2020). Se conoce que las semillas del maní contienen más de 32 proteínas diferentes, 18 tienen propiedades alergénicas, de las cuales se han identificado 11 (Iqbal et al., 2016), estas pertenecen a dos familias principales de globulinas: araquinas y conaraquinas (Viquez et al., 2003). Según Faustinelli (2012), las proteínas más alergénicas en la semilla del maní son *Ara h 1-3*; estas representan el 12, 16 y 10% del total proteínas de maní, respectivamente (Pi et al., 2019).

En el maní estas proteínas son más importantes que otros alérgenos alimentarios porque han demostrado ser extremadamente resistentes a la digestión, la desnaturalización por el calor, la acidez, los álcalis y las actividades proteolíticas (Iqbal et al., 2016). Estas proteínas alergénicas plantean dificultades en la salud, para una población importante de personas sensibles. En Norteamérica el maní está entre los principales alimentos alergénicos en niños, con un 1.8%; mientras que en Europa su prevalencia es del 0.2% (Peralta et al., 2016), y afecta entre el 0.6% y el 1.5% de los niños en los países desarrollados (Iqbal et al., 2016). Las alergias alimentarias (AA) son un problema emergente de salud pública en las áreas industrializadas del mundo, en el caso del maní hasta una pequeña cantidad (desde 100 µg de proteína) puede desencadenar reacciones alérgicas graves (Zhang et al., 2015), los síntomas de las alergias al maní varían desde una urticaria leve hasta un shock anafiláctico potencialmente mortal (Zhang et al., 2015).

El estudio de regiones genéticas y variantes asociadas con la producción de proteínas alergénicas, permite identificar estadios de semilla con baja expresión de estos genes o usar tecnologías de edición genética para silenciar la expresión de las proteínas con propiedades alergénicas, para el desarrollo de estrategias de mejoramiento de la especie. También podría facilitar la mejora de la comprensión de la alergenicidad del maní.

1.2. Justificación

El maní tiene una concentración significativa de proteínas y es variable en el proceso de formación de la semilla (Faustinelli, 2012); según Basha (1988), el contenido de proteína

incrementa conforme la semilla madura, por esta razón es importante cuantificar su expresión desde semillas en desarrollo hasta semillas maduras; ya que, las cantidades de proteína que originan reacciones alérgicas varían desde solo 100 µg hasta 1 g, resultando ser responsables de más del 50 % de las muertes por AA (Iqbal et al., 2016), y a pesar de ello a la fecha no existe una terapia eficaz para las alergias al maní; además, resulta necesario establecer al menos un método confiable para su detección y cuantificación en el maní (Zhang et al., 2015).

Los resultados del análisis de expresión de los genes resultan una poderosa herramienta para analizar la función de los genes y su cuantificación (Bi et al., 2010); sin embargo, se sabe poco sobre la expresión de las proteínas el maní durante el desarrollo de la semilla y germinación (Kang et al., 2007). El presente estudio permitirá definir la cantidad de moléculas de ARN y proteínas alergénicas producidas en la semilla de maní en sus diferentes estadios de desarrollo, desde semilla tierna, en desarrollo o también llamada inicial (R5, Tabla 1), una semilla completa o llena (R6), hasta una semilla madura (R7-R8) (Boote, 1982). Adicionalmente, mediante el análisis de expresión génica se identifican los genes más expresados, que pueden ser usados en conjunto con las pruebas previamente mencionadas, para reducir la obtención de resultados falsos negativos y aumentar la sensibilidad de las mismas. A futuro los genes identificados, podrían ser validados mediante ingeniería genética (Natukunda et al., 2022), en busca de una mejor identificación o determinar el estadio adecuado con menor contenido de alérgenos.

1.3. Objetivos

1.3.1. Objetivo general

Analizar los cambios en la expresión de genes asociados a ciertos alérgenos, en tres estadios de desarrollo de la semilla del maní

1.3.2. Objetivos

- Colectar datos de expresión génica desde NCBI, de genes del maní en diferentes estadios de la semilla.
- Mapear los reads de ARN a un genoma de referencia.
- Determinar la cantidad de expresión mediante el conteo de reads mapeados, en semillas de maní en desarrollo, completas y maduras.

2. Revisión de literatura

2.1.Maní

2.1.1. Origen

Arachis hypogaea L. (Linnaeus, 1753), es conocido como maní o cacahuate, cuyo centro de origen se consideran Bolivia y regiones de América del Sur. Su clasificación botánica (Gantait et al., 2019), es:

Reino: Plantae

Division: Tracheophyta

Class: Magnoliophyta

Order: Fabales

Family: Fabaceae

Genus: *Arachis*

Especie: *Arachis hypogaea*

La especie se dividide en dos subespecies (Gantait et al., 2019; Montero, 2020; Singh et al., 2021):

Subespecie *hypogaea*: posee hojas color verde oscuro, presenta una inflorescencia simple, sus flores se encuentra en disposición alternada, el tallo principal no florece, sus semillas revelan latencia, su fruto es de maduración tardía.

Subespecie *fastigiata*: posee hojas color verde claro, presenta inflorescencia simple o compuesta, su disposición no es secuencial, el tallo principal florece, sus semillas carecen de latencia, su fruto es de maduración temprana.

2.1.2. Etapas de crecimiento del maní

Se describen las etapas de crecimiento del maní, conforme a las condiciones vegetativas (V) y reproductivas (R); las etapas V corresponden a la formación de tallo y hojas (VE, V0, V1 y VN); mientras que las etapas R (R1 a R9) hacen referencia a la floración y fructificación (Boote, 1982).

2.1.2.1. Etapas de crecimiento de la semilla

A continuación, se profundiza sobre la formación y maduración del fruto (semillas), siendo R5, R6 y R7 las etapas correspondientes al proceso de interés para la presente investigación.

La etapa R5 de las semillas corresponde a la mitad de su maduración, en esta etapa alcanza un peso fresco entre 20 a 60 mg por semilla (Pattee et al., 1974), ya ha pasado su fase de endospermo líquido (Boote, 1982).

En la etapa R6 se reconoce a la semilla como "completa"; sin embargo, no llena por completo el fruto (cápsula), su peso seco es menos de la mitad de la semilla madura (43 %), por ello se recalca que R6 no representa la fase final de crecimiento (Boote, 1982).

La etapa R7 es una fase activa de llenado de la semilla y se presentan cambios en la coloración del pericarpio interno; a pesar de ser conocida como "madurez inicial", puede superar el 90% del peso seco de un fruto maduro, acercándose al final del llenado del fruto (Boote, 1982). Según Pattee et al. (1974), los pesos máximos se alcanzan en este estadio.

2.1.3. Proteínas en la semilla del maní

El maní es una planta leguminosa, su grano es oleaginoso y de gran importancia en el mundo, debido a su valor nutricional, contiene proteínas (16 a 36 %), aceites (36 a 54 %) y carbohidratos (10 a 20 %) (Singh et al., 2021). La semilla se consume en fresco y procesado, como fuente de proteínas de alta calidad, vitaminas (E, K, B) y fibra (Gantait et al., 2019).

Las proteínas de la semilla se consideran una reserva de aminoácidos, que son usados para la germinación y su crecimiento (Singh et al., 2021).

2.1.3.1. Alérgenos en la semilla del maní

Los principales alérgenos del maní son las proteínas de almacenamiento de las semillas (SSP siglas en inglés) y están compuestos por tres fracciones según sus propiedades de sedimentación (Singh et al., 2021):

- 2S, pertenecen a la familia de las albuminas.
- 7S pertenecen a la familia de las globulinas, del tipo vicilinas (conaraquina).
- 11S, las más diversas, pertenecen a la familia de las globulinas, del tipo leguminas (araquina).

El contenido proteico total del maní está compuesto por más de 32 proteínas diferentes, 18 de estas son alergénicas de las cuales se han identificado alrededor de 11, mediante SDS PAGE (Pele, 2010; Singh et al., 2021), estos alérgenos han sido asignados con el prefijo *Ara* (desde Ara h 1 hasta Ara h 18) (Tabla 1) (Singh et al., 2021).

Los alérgenos más abundantes en el maní son Ara h 1, 2, 3 y 6, siendo Ara h 2 y 6 los más importantes respecto a la alergia alimentaria (Iqbal et al., 2016; Marsh et al., 2022).

Tabla 1. *Alérgenos del maní*

Alérgeno	Superfamilia	Familia	Función Biológica
Ara h 1	Cupina	Globulina 7S, Vicilina	
Ara h 2	Prolamina	Albúmina 2S, Conglutina	Inhibidor de tripsina
Ara h 3	Prolamina	Globulina 11S, Legumina, araquina	Inhibidor de tripsina
Ara h 4	Prolamina	Isoforma de Ara h 3, se renombrada como Ara h 3.02	
Ara h 5	Profilina	Profilina	Regula polimerización de la actina
Ara h 6	Prolamina	Albúmina 2S, Conglutina	Desgranulación de los basófilos
Ara h 7	Prolamina	Albúmina 2S, Conglutina	Inhibidor de amilasa/tripsina
Ara h 8	Bet v 1 Family (homólogo)	Proteína relacionada con la patogénesis (PR-10)	Proteger a la planta de patógenos
Ara h 9	Prolamina	Proteína de transferencia de lípidos no específica, categoria tipo 1	Transferir lípidos entre membranas
Ara h 10	Glicosil transferasa GT-C	Oleosina	Estabilidad estructural en cuerpos oleosos vegetales, durante la maduración
Ara h 11	Glicosil transferasa GT-C	Oleosina	Estabilidad estructural en cuerpos oleosos vegetales, durante la maduración
Ara h 12		Defensinas	
Ara h 13		Defensinas	
Ara h 14	Glicosil transferasa GT-C	Oleosina	proteínas estructurales anfifílicas
Ara h 15	Glicosil transferasa GT-C	Oleosina	proteínas estructurales anfifílicas
Ara h 16		Proteína de transferencia de lípidos no específica, categoria tipo 2	
Ara h 17		Proteína de transferencia de lípidos no específica, categoria tipo 1	
Ara h 18		Ciclofilina	

Nota. Singh et al. (2021). Toomer (2018). Iqbal et al. (2016). Pele (2010).

2.1.4. Genoma del maní

Arachis hypogaea es una especie alotetraploide natural, ya que contiene los juegos completos de cromosomas de sus ancestros (Bertioli et al., 2016, 2019; Kunta et al., 2021). Se considera que el maní cultivado evolucionó desde una hibridación interespecífica simple entre dos especies diploides (Gantait et al., 2019), cuyos progenitores son *Arachis duranensis* (AA, 2n = 2x = 20) y *Arachis ipaensis* (BB, 2n = 2x = 20). El maní tiene una conformación del genoma AABB (2n = 4x = 40 cromosomas), su tamaño del genoma es de ~2.7 Gb y tiene un contenido repetitivo estimado del 64 % (Gantait et al., 2019; Montero, 2020; Singh et al., 2021).

2.1.4.1. Genoma de referencia del maní

El genoma de referencia de maní, inició con el mapeo del cultivar tetraploide "Tifrunner", con 11 poblaciones (10 líneas endogámicas recombinantes y 1 de retrocruzamiento). Para la generación del mapa genético en consenso se emplearon varios sistemas de marcadores, y se adicionaron 5 poblaciones más, fue construido usando 3693 loci marcadores, cubriendo una distancia de 2651 cM (Vishwakarma et al., 2017).

El genoma de referencia se encuentra disponible en la base de datos de National Center for Biotechnology Information (NCBI), identificado como "*Ensamblaje del genoma arahy.Tifrunner.gnm1.KYV3*", con su ID: GCF_003086295.2 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003086295.2/).

El ensamblaje de este genoma es de tipo haploide, a nivel de cromosomas, entre sus características constan (NCBI, 2018):

- Tamaño del genoma 2.6Gb
- Número de cromosomas 20
- Número de scaffolds 384
- Scaffold N50 135.2Mb
- Scaffold L50 9
- Número de contigs 4039
- GC (%) 36
- Cobertura del genoma 48.0x

2.2. RNA-seq y Expresión génica diferencial

2.2.1. RNA-seq

El análisis de secuencias de ARN (RNA-seq) consiste en una diversidad de métodos, tanto experimentales como computacionales, para determinar la identidad y abundancia de ARN en muestras bajo condiciones de interés (Korpelainen et al., 2015).

Este análisis inicia con el aislamiento del ARN a partir de una muestra biológica, la secuenciación y termina con el análisis bioinformático de los datos (Korpelainen et al., 2015). A continuación, se describe el aislamiento y secuenciación de ARN.

2.2.2. Aislamiento de ARN

Todo parte del dogma central de la biología molecular, este se describe en dos procesos: el primer paso es la transcripción, es el paso de ADN a ARN mediado por la enzima ARN polimerasa; y el segundo paso es la traducción, es la conversión del ARN mensajero (ARNm)

en proteína. El conjunto completo de ARN de una célula o tejido es conocido como transcriptoma (Blumenberg, 2019), que es el objeto de estudio en la presente investigación. El ARN es aislado típicamente de células o tejido, fresco o congelado, usando kits comerciales, que permite obtener gran cantidad de ARN total. Casi todas las muestras están contaminadas con ADN, por ello es muy común tratarlo con ADNasas. Después se evalúa la calidad para posteriormente preparar la librería, para ello se requiere de 0.1 a 10 µg de ARN total. La preparación de la librería tiene dos propósitos: representar fielmente la molécula de ARN y transformar el ARN en ADN complementario (ADNc) de doble cadena (ds cDNA en inglés) (Korpelainen et al., 2015).

2.2.3. Secuenciación de ARN

El estudio del transcriptoma inició con los microarreglos (microarrays en inglés), sin embargo, actualmente han sido reemplazados por diferentes técnicas de secuenciación de ARN (RNA-Seq), que requieren mucho menos materiales y permiten el análisis del transcriptoma de una sola célula (Blumenberg, 2019).

Korpelainen et al. (2015), proponen las siguientes plataformas de ARN-Seq (Tabla 2), como las principales:

- Illumina: es un método de secuenciación por síntesis muy popular(Križanović et al., 2018). Una vez lista la librería, el ADNc de doble cadena se liga a adaptadores que se unen por complementariedad a los extremos de las secuencias, estos pasan por una celda de flujo (*flow cell*), y los adaptadores se unen a la placa. Las cadenas son amplificadas (usando polimerasas) a manera de puente obteniendo varias copias, el proceso se llama amplificación de clústeres. Posteriormente se elimina una cadena, para realizar la secuenciación por síntesis, en cada ronda se adiciona un nucleótido con una señal fluorescente distintiva, lo que permite visualizar el nucleótido y su localización, esto proporciona una secuencia de nucleótidos precisa de la pieza original de ADNc. La secuenciación puede ser *single read* (SR, lectura única), si se realiza desde un solo extremo de la doble cadena de ADNc; mientras que, puede ser *paired-end read* (PE, lectura emparejada) al realizarse desde ambos extremos de la cadena. Illumina permite la secuenciación masiva, por ejemplo: el equipo Hi-Seq 2500 puede producir

cerca de 6Gb de datos en una sola lectura, y el sistema MySeq puede producir 8.5Gb de datos en dos días de ejecución (Korpelainen et al., 2015).

- Ion Torrent: utiliza la librería ligada a un adaptador, las secuencias serán hibridas a perlas, seguido se hace una amplificación de emulsión (polimerasas), estas se colocarán en un chip de silicio, para dar paso a la secuenciación por síntesis; cuando se agrega un nucleótido, se libera un protón, el equipo tiene detectores altamente sensibles que identifica los cambios en el pH. En caso de nucleótidos repetidos, esta plataforma puede detectar cambios mayores en el pH y usa esta medición para leer el polímero (Korpelainen et al., 2015). Ion Torrent produce menos lecturas comparada con las demás técnicas, pero los hace en menos tiempo de ejecución (Vlasova-St. Louis, 2021).

- Pacific Biosciences (PacBio): es una plataforma de la tercera generación, su metodología es denominada *Secuenciación en tiempo real de molécula sencilla* (SMRT, siglas en inglés) (Korpelainen et al., 2015; Križanović et al., 2018). Inicia ligando adaptadores a los extremos de la secuencia, creando una sola molécula circular. Su tecnología de celdas SMRT, contienen cámaras denominadas *guía de onda de "modo cero"* (ZMWs, siglas en inglés), son cámaras de espacio restringido que contienen, una molécula de ADN polimerasa e inmovilizan la molécula de ADNc, esta última es secuenciada en tiempo real, usando nucleótidos trifosfatos fluorescentes distintivos. Su principal ventaja es la velocidad, puede ejecutarse en un par de horas, y producir hasta 250Mb de datos en una sola ejecución (Korpelainen et al., 2015).

- Nanopore Technologies: presentada en 2014 por Oxford Nanopore Technologies (ONT MinION), es una plataforma de tercera generación (Križanović et al., 2018), que usa una sola enzima para separar la cadena de ADN y guiarlo a través de una membrana porosa que está unida a una proteína, por estos poros pasan iones que generan una corriente eléctrica continua, cada nucleótido al atravesar impide el flujo de corriente de manera específica, lo que genera un cambio que es medido, permitiendo de esta manera identificar la secuencia de la cadena (Korpelainen et al., 2015).

Tabla 2. *Comparación de metodologías de RNA-seq*

Método	Longitud de lectura	Exactitud	Tiempo por corrida	Ventajas	Desventajas
Pacific biosciences (PacBio)	> 100 000 bases	87.00%	30 min a 20 horas	• Rápido	• Costoso
Ion Torrent	> 600 pb	99.60%	2 horas	• Rápido • Equipo menos costoso	• Errores de homopolímero
Illumina	50 - 600 pb	99.90%	1 a 11 días	• Alto rendimiento de secuencia	• Requiere alta concentración de ADN • Equipo costoso
Nanopore (ONT MinION)	> 500 kb	92-97%	1 min a 48 horas	• Portátil	• Menor rendimiento • Menor precisión

Nota. Blumenberg (2019).

2.2.4. Aplicaciones de RNA-seq

El desarrollo de la secuenciación de nueva generación y los avances bioinformáticos han abierto nuevos campos para explorar, RNA-seq ha permitido capturar de manera precisa todas las moléculas de ARN (codificante y no codificante), por ejemplo, se ha empleado para la investigación y diagnóstico de enfermedades (Vlasova-St. Louis, 2021). Mediante la secuenciación de ARN se han alcanzado varios logros: primero, el definir la secuencia de ARN, que ha permitido identificar genes o regiones no codificantes; segundo, determinar la estructura de la molécula, lo que permite señalar su tipo (ARNt o miARN); tercero, comprobar la abundancia de las lecturas de ARN, lo que facilita comparar entre muestras específicas, por ejemplo individuos sanos vs enfermos, organismos en diferentes etapas de crecimiento, confrontar diferentes tejidos, entre otros.

RNA-seq es una poderosa tecnología con varias aplicaciones, que van desde el descubrimiento de genes y variantes de empalme hasta el análisis de expresión diferencial, la detección de genes y edición de ARN (Korpelainen et al., 2015). Adicionalmente, varios ensayos de expresión génica por RNA-seq, han generado información sobre la respuesta del hospedero a los patógenos, alérgenos, entre otros factores diversos (Vlasova-St. Louis, 2021).

2.2.5. Expresión diferencial

2.2.5.1. Expresión génica

Los organismos eucariotas, están formados por diversas células que se originan por diferenciación (Anello et al., 2021). Las células se diferencian dando lugar a varios tipos y funciones, debido a diferencias cualitativas y cuantitativas en su expresión génica (Gibney & Nolan, 2010).

La actividad de un gen inicia con la transcripción en el núcleo, es el paso de ADN en ARNm maduro, posteriormente el ARNm se transporta hacia el citoplasma, para unirse a los ribosomas, en donde continúa con la traducción, que consiste en la síntesis de polipéptidos (proteínas) (Gibney & Nolan, 2010). La expresión génica abarca desde la activación del gen (transcripción), hasta que la proteína madura se ubica en el tejido adecuado y cumple su función, por tanto, la proteína contribuye a la expresión del fenotipo celular (Hernández et al., 1995).

2.2.5.2. Análisis de expresión diferencial (ED)

Los organismos poseen muchos genes, de los cuales solo una parte se expresa, es decir, que no se requieren ni todas las proteínas ni los mismos niveles de manera simultánea, para esto existe un sistema de regulación que controla la transcripción y la traducción de los genes, en determinadas condiciones (Anello et al., 2021).

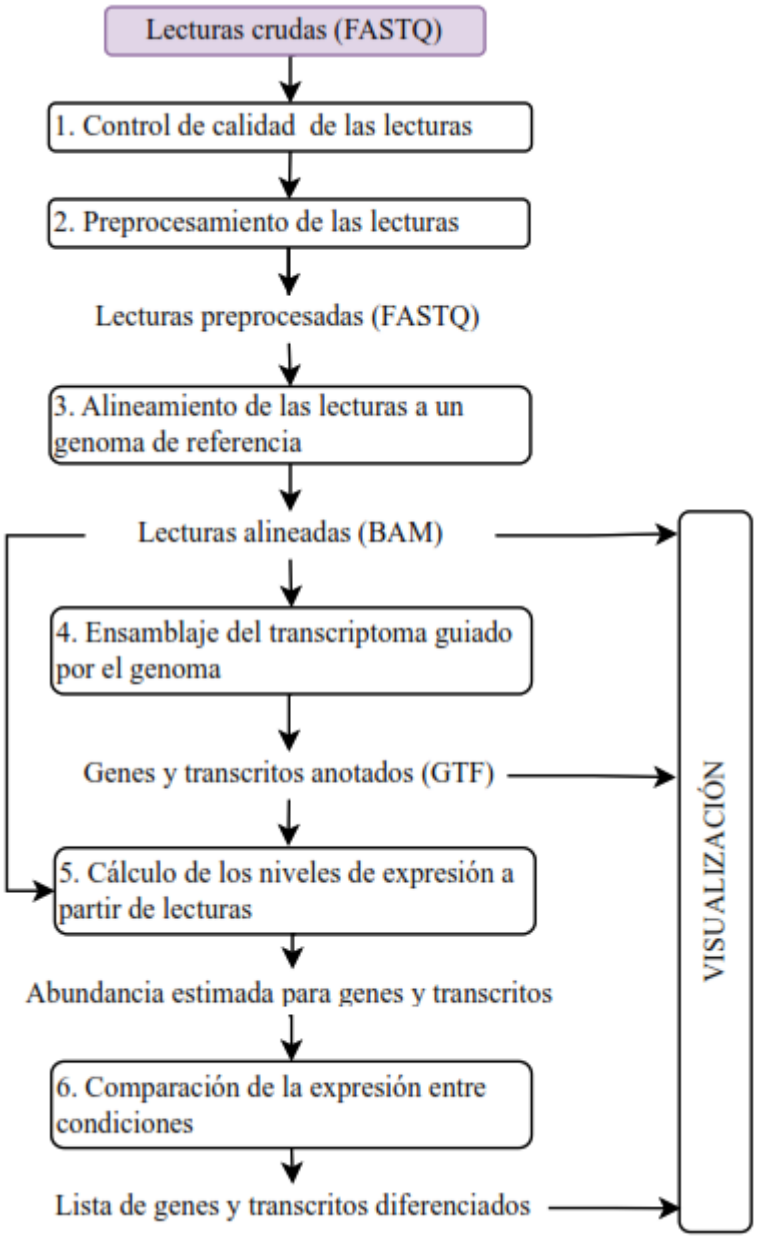
El análisis de expresión diferencial (ED) se basa en la identificación de estos genes o transcritos, que se expresan en cantidades significativamente diferentes, en distintos grupos de muestras, condiciones biológicas, tejidos diferentes, fases de desarrollo distintas u otra condición (tratamientos) (Korpelainen et al., 2015).

2.3. RNA-seq: Análisis bioinformático de la data

Después del aislamiento de ARN a partir de una muestra biológica y obtener las lecturas de RNA-seq, sigue el análisis bioinformático de los datos, para la investigación sobre los niveles de expresión de cada gen, existen muchos caminos, para ello Korpelainen et al. (2015), sugieren definir un flujo de trabajo (Figura 1).

Para el análisis de datos de RNA-seq existen varios enfoques y herramientas para cada paso del análisis, por ello es importante seleccionar el programa más adecuado (Korpelainen et al., 2015); a continuación, se exponen los programas más utilizados para cada paso.

Figura 1. *Flujo de trabajo general para el análisis de expresión diferencial, consta de varios pasos interrelacionados. Los formatos típicos de los archivos de salida se indican entre paréntesis.*



Nota. Figura traducida desde Korpelainen et al. (2015)

2.3.1. Control de calidad

Las lecturas crudas pueden presentar problemas, por diferentes razones que afectan la calidad de la secuencia, lo que puede afectar el alineamiento al genoma de referencia. Estos inconvenientes se pueden corregir usando prgramas para revisar la calidad de las lecturas (Babraham Institute, 2020; Korpelainen et al., 2015).

El control de calidad (QC) es el primer paso para empezar con el análisis de datos de secuenciación, este mide las estadísticas en un archivo de lecturas secuenciadas para evaluar si los datos son adecuados para el análisis posterior (de Sena Brandine & Smith, 2021), la herramienta más popular para este proceso es FastQC (Babraham Institute, 2020), actualmente existe otra herramienta similar llamada Falco (de Sena Brandine & Smith, 2021).

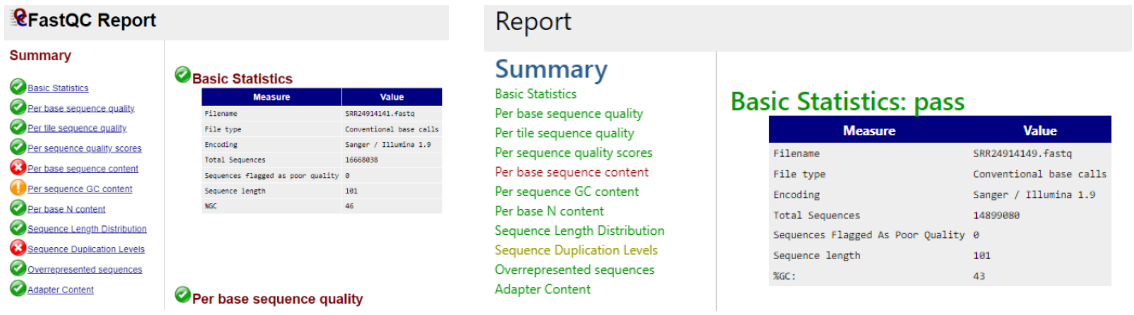
2.3.1.1. *FastQC*

FatsQC es un programa de control de calidad (QC), que se puede usar mediante línea de comando o interfaz gráfico, usa archivos de entrada con formato FASTQ, SAM (*Sequence Alignment Map*) o BAM (*Binary Alignment Map*). Genera un reporte con varias métricas de calidad, el resultado final presenta las estadísticas básicas (Figura 2), calidad y contenido de las bases, longitud de la lectura, presencia de bases ambiguas, secuencias sobrerrepresentadas y duplicadas, entre otros (Korpelainen et al., 2015). FastQC es la herramienta más popular utilizada para el análisis de lecturas de Illumina (Conesa et al., 2016).

2.3.1.2. *Falco (FastQC Alternative Code)*

Este software es una emulación de la popular herramienta FastQC, genera resultados equivalentes (Figura 2), se ejecuta más rápido ya que requiere menos memoria y su visualización en HTML es más flexible. Fue creada para sistemas operativos basados en UNIX, con código abierto en C++. El código fuente de Falco está disponible en: https://github.com/smithlabcode/falco (de Sena Brandine & Smith, 2021).

Figura 2. *Reporte de resultados del análisis de calidad de una secuencia (Resumen), FastQC (izquierda) y Falco (derecha).*



Tanto FastQC como Falco, ofrecen los resultados similares (Figura 2), entre las principales métricas que presentan están (Babraham Institute, 2020; de Sena Brandine & Smith, 2021):

- *Estadísticas Básicas*: presenta el nombre del archivo, tipo, codificación, secuencias totales, secuencias filtradas, longitud de secuencia (un solo valor si todas son iguales), porcentaje de guanina-citosina (%GC). No presentan advertencias o error.
- *Calidad de secuencia por base*: muestra un gráfico de cajas, provee una descripción del rango de valores de calidad (eje y) en todas las bases (eje x). El informe emite una advertencia si el cuartil inferior es menor a 10, o si la mediana es inferior a 25 en cualquiera de las posiciones de las bases; mientras que, emite error cuando el nivel inferior es menor a 5 o la mediana menor a 20.
- *Contenido de secuencia por base:* informa sobre la proporción de lectura que tiene cada nucleótido por posición. Este módulo emite una advertencia si la diferencia entre A y T, o G y C es superior al 10% en cualquier posición. Muestra un error si la diferencia es superior al 20% en cualquier posición.
- *Contenido de guanina-citosina (GC) por secuencia:* permite determinar el contenido GC en toda la secuencia y se compara contra una distribución normal modelada. Este módulo emite una advertencia si el contenido GC de cualquier base se desvía más de un 5% del contenido GC medio; mientras que, es un error si se desvía más de un 10%.

2.3.1.3. Preprocesamiento de las lecturas (limpieza)

El preprocesamiento de las secuencias se enfoca en eliminar adaptadores y descartar lecturas de baja calidad (Conesa et al., 2016). Conforme los resultados generados en el control de calidad, en caso de presentar advertencias o errores, se sugiere realizar la limpieza de las secuencias, para ello existen diversas herramientas, la más popular es Trimmomatic (Babraham Institute, 2020; Korpelainen et al., 2015).

Trimmomatic es una herramienta versátil basada en Java, desarrollada para el preprocesamiento de lecturas, permite eliminar adaptadores y recortar lecturas de diferentes maneras en función de la calidad, además filtrar lecturas en función de su calidad y longitud (Korpelainen et al., 2015). Funciona con archivos FASTQ, usando las puntuaciones de calidad en escala de Phred (Bolger et al., 2014; Korpelainen et al., 2015).

2.3.2. Alineamiento o mapeo a un genoma de referencia

El alineamiento permite verificar el grado de similitud de una secuencia, en este caso, a un genoma de referencia, el mapeo proporciona información sobre la localización genómica. Esta tarea es complicada porque existen millones de lecturas o *reads* cortos y hay genomas muy grandes que pueden tener secuencias repetidas, lo que dificulta el alineamiento (Korpelainen et al., 2015).

Para estudios de ARN-seq, en donde las lecturas se asignan a genomas que contienen intrones, se debe usar un alineador empalmado, como puede ser STAR o Tophat (Korpelainen et al., 2015). Otros autores sugieren usar Hisat2, que está basado en implementaciones de HISAT y Bowtie 2 (Kim & Park, 2020).

2.3.2.1. STAR (Spliced Transcripts Alignment to a Reference)

Su nombre se refiere a "Alineamiento de transcritos que han sufrido splicing a una referencia"(Guillén, 2019). Este algoritmo fue diseñado para el mapeo de datos de RNA-seq, para alinear las secuencias no contiguas directamente con el genoma de referencia y utiliza una estrategia novedosa para alineaciones empalmadas (Dobin et al., 2013; Guillén, 2019). El código fuente y los binarios de STAR pueden descargarse de GitHub, desde https://github.com/alexdobin/STAR (Dobin, 2019).

El flujo de trabajo de STAR consta de dos pasos (Dobin, 2019):

- *Generación del índice genómico*, para ello STAR requiere las secuencias genómicas de referencia (archivo FASTA) y las anotaciones (archivo GTF o GFF).
- *Mapeo de lecturas al genoma*, en este paso STAR requiere el índice genómico y las lecturas de RNA-seq (FASTA o FASTQ). Como resultado escribe varios archivos de salida, como alineaciones (SAM/BAM), estadísticas de resumen de mapeo, lecturas no mapeadas, entre otros.

2.3.2.2. HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2)

HISAT2 es un programa de alineación rápido y sensible para mapear lecturas de secuenciación de próxima generación (ADN o ARN), frente a un genoma de referencia, utilizando un índice gráfico de Ferragina Manzini (GFM). HISAT2 utiliza un gran conjunto

de pequeños índices (56 Kbp), llamados índices locales, que combinados con varias estrategias de alineación cubren colectivamente todo el genoma, generando alineaciones en formato SAM. El software se desarrolló basado en implementaciones HISAT y Bowtie2 (Kim & Park, 2020).

Es frecuente que los archivos de salida formato SAM sean procesados con otras herramientas como Samtools, para ordenar las lecturas, eliminar lecturas no alineadas o transformar el archivo a una versión binaria (BAM) (Corchete, 2019).

2.3.3. Cuantificación de lecturas

Se procederá a la asignación de las lecturas a un gen o transcrito determinado, este proceso se conoce como contaje o cuantificación (Corchete, 2019). Al disponer del genoma de referencia anotado, las lecturas mapeadas pueden se contadas por características genómicas basadas en la información de localización. El número de lecturas generadas depende de varios factores, como profundidad de secuenciación y longitud del transcrito; y otros factores que pueden ser más difíciles de precisar, como: composición del transcriptoma, sesgo GC, entre otros. La forma más sencilla de estimar la expresión es contar las lecturas por genes, se pueden usar varias herramientas como HTSeq, BEDTools y Qualimap (Korpelainen et al., 2015).

2.3.3.1. HTSeq

HTSeq contabiliza el número de lecturas alineadas para cada gen, que se ajustan a cada uno de sus exones (Corchete, 2019). Htseq-count forma parte del paquete HTSeq, requiere de las lecturas alineadas en formato SAM/BAM y la anotación del genoma como archivo GFF/GTF. Htseq-count encuentra los exones con los que se solapan las lecturas y luego agrupa los recuentos a nivel de exón basándose en el ID del gen de los exones en el archivo GTF (Korpelainen et al., 2015).

2.3.4. Análisis de expresión génica diferencial

2.3.4.1. Expresión génica diferencial

El análisis de datos de secuenciación de ARN (RNA-seq) busca identificar genes que se expresan en cantidades significativamente diferentes en distintos grupos de muestras (Love et al., 2014). Este análisis se refiere a la identificación de genes que se expresan diferente, al

comparar grupos de muestras, condiciones biológicas, diferentes tejidos, entre otros (Korpelainen et al., 2015).

La selección del software de análisis de expresión diferencial, debe vincularse al número de réplicas biológicas. En caso de contar con 5 o más replicas, puede ser beneficioso utilizar un método de análisis no paramétrico; mientras que, si se cuenta con pocas réplicas se recomiendan los métodos paramétricos, que asumen una distribución basada en datos empíricos, como los paquetes paquetes DESeq y edgeR. Para trabajar con DESeq o edgeR, se requiere de los objetos matriz de diseño y matriz de contraste (Korpelainen et al., 2015):

Matriz de diseño: es un objeto R que describe el diseño del experimento.

Matriz de contraste: para comprobar la expresión diferencial, se necesita una matriz de contraste que describa las comparaciones que desea hacer (no necesario en DESeq2).

2.3.4.2. DESeq2

Es un paquete de R/Bioconductor. DESeq2 estima la dependencia entre la media y la varianza en los datos de contaje y lleva a cabo un análisis de expresión diferencial basado en la distribución binomial negativa (distribución gamma-Poisson) (Corchete, 2019), proporciona métodos para comprobar la expresión diferencial mediante el uso de modelos lineales generalizados binomiales negativos, las estimaciones de la dispersión y los cambios logarítmicos de pliegues incorporan distribuciones previas basadas en datos (Korpelainen et al., 2015; Love et al., 2016).

El modelo DESeq2 corrige internamente el tamaño de la biblioteca, por lo que no deben utilizarse como entrada los valores transformados o normalizados (Love et al., 2016), este paquete detecta y corrige las estimaciones de dispersión demasiado bajas (Love et al., 2014).

3. Metodología

3.1.Muestra

Las secuencias fueron obtenidas de la base de datos SRA de NCBI, en archivos formato FASTQ. A continuación, se describen las características de la biblioteca (Tabla 3):

- *Especie: Arachis hypogaea,* cultivar Hanoch
- *Accesión del proyecto:* PRJNA982443
- *Fecha de registro:* 11 de junio de 2023
- *Tipo de dato:* Lecturas de secuencias sin procesar
- *Tipo de ensayo:* RNA-seq
- *Kit de extracción:* TruSeq RNA Library Prep Kit v2
- *Instrumento:* Illumina HiSeq 2000
- *Diseño de biblioteca:* single (LibraryLayout)
- *Tejido:* semilla en diferentes estadios de desarrollo
- *Repeticiones:* tres repeticiones por estadio
- *Presentación:* ARO, adi faigenboim; 2023-06-11
- *Publicado:* 2023-06-14
- *Tipo de acceso:* Público

Tabla 3. *Codificación de las secuencias muestras y los estadios de desarrollo de la semilla*

Estadio de la semilla	Código y repetición	BioSample (Muestra)	SRA	Run
Desarrollo	R5_r1	SAMN35707368	SRS17970673	SRR24914149
Desarrollo	R5_r2	SAMN35707368	SRS17970673	SRR24914148
Desarrollo	R5_r3	SAMN35707368	SRS17970673	SRR24914147
Completa	R6_r1	SAMN35707369	SRS17970674	SRR24914146
Completa	R6_r2	SAMN35707369	SRS17970674	SRR24914145
Completa	R6_r3	SAMN35707369	SRS17970674	SRR24914144
Madura	R7_r1	SAMN35707370	SRS17970675	SRR24914143
Madura	R7_r2	SAMN35707370	SRS17970675	SRR24914142
Madura	R7_r3	SAMN35707370	SRS17970675	SRR24914141

Nota. Link al proyecto en NCBI https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA982443

3.2. Genes alérgenos de la semilla del maní (Ara h)

Varios genes codifican para las diferentes proteínas alérgenas del maní (Ara h), algunos de estos se encuentran identificados en las diferentes bases de datos, para contar con la

información de los mismos se realizó una recopilación de los genes en la base de datos de NCBI (NCBI, 2023; Palladino & Breiteneder, 2018).

3.3. Control de calidad de las lecturas

Para realizar el control de calidad las lecturas fueron evaluadas mediante los programas FastQC, versión 0.11.9 (Babraham Institute, 2020) y Falco versión 0.3.0 (de Sena Brandine & Smith, 2021), usando los parámetros predeterminados de las herramientas.

3.4. Preprocesamiento de las lecturas

Debido a los resultados obtenidos en la métrica *contenido de secuencia por base*, durante el control de calidad de todas las secuencias (Anexo 1), se decidió hacer un preprocesamiento de las nueve lecturas. Las secuencias fueron filtradas, mediante la herramienta Trimmomatic versión 0.39 (Bolger et al., 2014), los parámetros usados para la limpieza de las secuencias, fueron:

- -jar, para la ejecución de Trimmomatic
- -phred33, especifica la codificación de calidad base
- SE (single end), para datos de un solo extremo
- HEADCROP:14, indica número de bases que se eliminarán desde el inicio de la lectura.

Nuevamente se evaluó la calidad de los archivos resultantes, con Falco versión 0.3.0, usando los parámetros predeterminados.

3.5. Alineamiento a un genoma de referencia y cuantificación

Para el alineamiento se utilizó el genoma de referencia del maní, obtenido de NCBI, codificado como GCF_003086295.2, y las secuencias recortadas en Trimmomatic. El alineamiento se realizó con dos herramientas, STAR y HISAT2.

3.5.1. Alineamiento – STAR

Se utilizó el software STAR versión 2.7.11a (Dobin, 2019; Dobin et al., 2013). Este proceso consta de dos pasos: generación del índice genómico y el mapeo de lecturas al genoma.

Generación del índice genómico

Para la generación del índice genómico, se utilizó dos archivos:

- Genoma de referencia del maní en formato FASTA
- Anotaciones del genoma en formato GTF

Y se utilizaron las siguientes opciones básicas (Dobin, 2019):

- --runMode *genomeGenerate*, para la generación de índices genómicos.
- --genomeDir especifica la ruta al directorio donde se almacenará el índice.
- --genomeFastaFiles refiere al archivo FASTA, al genoma de referencia.
- --sjdbGTFfile especifica la ruta al archivo de anotación (GTF).
- --runThreadN 14, corresponde al número de núcleos disponibles en el servidor.

Mapeo de lecturas al genoma

Para el mapeo de las nueve secuencias (formato FASTQ) al índice previamente generado, se utilizó los siguientes parámetros básicos (Dobin, 2019):

- --runThreadN 14
- --genomeDir
- --readFilesIn, especifica la ruta a las secuencias que serán alineadas.

Además, los siguientes parámetros avanzados (Dobin, 2019):

- --outFileNamePrefix, produce varios archivos de salida.
- --outSAMtype, genera alineaciones directamente en formato binario BAM
- Unsorted/SortedByCoodinate, *bamsort* lee un archivo BAM, SAM o CRAM, lo ordena por coordenadas lexicográficas por id de secuencia de referencia y posición en la secuencia de referencia (Tischler, 2016)
- --outFilterMultimapNmax 5, número máximo de loci a los que puede asignarse la lectura (alineaciones múltiples); si se supera, la lectura se considera no mapeada
- --outFilterMismatchNmax 1, número máximo de desajustes por par, la alineación sólo se mostrará si no tiene más desajustes que este valor.

3.5.2. Alineamiento – HISAT2

Se utilizó HISAT2 versión 2.2.1, desarrollado por Kim & Park (2020). Inicialmente se generó un índice anotado a partir del genoma del maní, este proceso genera archivos con extensión ht2 que se utilizaron para el alineamiento.

Generación del índice genómico

Los pasos para la generación del índice fueron (Kim & Park, 2020):

a. Extracción de los sitios de empalme, para ello de utilizó el comando *hisat2_extract_splice_sites.py*, que es una utilidad proporcionada por HISAT2, para extraer los sitios de empalme a partir de un archivo GFF o GTF, hacia un archivo tabular (.tab).
b. Se usó el comando principal de la herramienta *hisat2_extract_exons.py*, para extraer información sobre los exones a partir de un archivo GFF o GTF, misma que se guarda en un archivo tabular (.tab).
c. Para la creación del índice, en este proceso se requirió tres archivos, el genoma de referencia en formato FASTA, y los dos archivos tabulares generados en los pasos previos. Los parámetros de la corrida fueron:
    - hisat2-build que construye el índice.
    - -p 12 corresponde al número de núcleos disponibles en el servidor.
    - --ss especifica el archivo tabular que contiene información sobre los sitios de empalme.
    - --exon especifica el archivo tabular que contiene información sobre los exones.

Mapeo de lecturas al genoma

Se continuó con el alineamiento de las nueve lecturas (formato FASTQ), utilizando el índice originado a partir del genoma de referencia. Los parámetros de la corrida fueron (Kim & Park, 2020):

- hisat2, comando para alinear secuencias.
- -p 18, corresponde al número de núcleos disponibles en el servidor.
- --dta (downstream-transcriptome-assembly), con esta opción HISAT2 requiere longitudes de anclaje más largas para el descubrimiento *de novo* de sitios de empalme, conduce a menos alineaciones con anclajes cortos.

- -x, especifica la ruta al índice del genoma.
- -U indica que se está utilizando un archivo de secuencias de ARN-seq de una sola lectura.
- -S especifica el nombre del archivo de salida, en este caso en formato SAM.

En todos los casos, los alineamientos contenidos en los archivos BAM resultantes se ordenaron por posición y por nombre utilizando samtools para adaptarlos a los requisitos de los algoritmos de contaje que se utilizaron posteriormente.

3.5.3. Cuantificación de las lecturas con HTSeq

Se usó la herramienta HTSeq versión 2.35, para realizar el conteo y posterior estimación de la expresión. Se utilizó el archivo con lecturas de secuenciación alineadas (archivos de salida de STAR, formato BAM), y el genoma anotado (formato GTF). Se empleó la herramienta Htseq-count, y los parámetros utilizados fueron (Anders, 2018):

- htseq-count, comando usado para el conteo de lecturas.
- -f bam, es el formato de los archivos de entrada.
- -r pos, especifica cómo se asignan las lecturas cuando caen en regiones superpuestas, en este caso según la posición de inicio de la lectura
- -s no, significa que no se realizará ninguna corrección de la orientación
- -t exon, es el tipo de característica, para RNA-seq con un archivo GTF.
- -i gene_id, el atributo que se usa para identificar los recuentos en la tabla de salida.
- Se utilizó el modo predeterminado (*union*) para manejar lecturas que se superponen a más de una característica.

3.6. Análisis de Expresión Diferencial

Para el análisis de expresión diferencial, se utilizó el lenguaje de programación de R, en el entorno de RStudio versión 4.2.3, usando el paquete de Bioconductor, denominado DESeq2. A partir de los archivos de salida de HTSeq (formato txt), se realizaron combinaciones de todos los conteos de los diferentes estadios de la semilla (en bash), para posteriormente compararlos en pares, entre las tres condiciones (Tabla 3), de la siguiente manera:

- DC: Desarrollo versus Completa
- DM: Desarrollo versus Madura
- CM: Completa versus Madura

3.6.1. DESeq2

En R los datos en formato txt, fueron importados, generando un dataframe de los conteos. Las funciones utilizadas para correr el paquete Deseq2 fueron (Corchete, 2019; Korpelainen et al., 2015; Love et al., 2016):

- *read.table*, para leer datos desde un archivo TSV. Los parámetros usados fueron header=T, sep="\t", row.names=1.
- *rowSums,* para el prefiltrado de los conteos, con la condición de que sea superior a 3 en más de dos de las muestras (filtrado independiente), que elimina genes con baja expresión.
- *data.frame* para construir la matriz de diseño del experimento (Figura 3), que contiene los datos para las respectivas comparaciones.
- *DESeqDataSetFromMatrix* construye un objeto *DESeqDataSet* a partir de los archivos de conteos y la matriz de diseño (Korpelainen et al., 2015), para realizar análisis de expresión génica diferencial utilizando DESeq2. Además, se específica como variable de diseño la "condición", partiendo del ejemplo de la Figura 3, la condición es la última columna (desarrollo, completa y madura).
- *factor*, convierte los valores únicos de una columna en un factor (variable categórica), ya que para el análisis con *DESeq* se requiere establecer el orden de los factores (Corchete, 2019), para el análisis en pares de condiciones.
- *DESeq* es una función en el paquete DESeq2, que realiza el análisis de expresión génica diferencial. Los resultados pueden ser indagados mediante los comandos *results*, *resultsNames*() y *summary*().

Figura 3. *Matriz de diseño del experimento para posteriormente hacer las comparaciones (captura desde RStudio).*

	Gene_ID		condicion
1	R5_r1	R5	desarrollo
2	R5_r2	R5	desarrollo
3	R5_r3	R5	desarrollo
4	R6_r1	R6	completa
5	R6_r2	R6	completa
6	R6_r3	R6	completa

3.6.1.1. Gráficos

Para realizar los gráficos se ordenaron los resultados por el p-value ajustado, conforme Corchete (2019), recomienda el uso de la función *order* (). Adicionalmente se contabilizaron los genes significativos en el p-value ajustado 0.01; usando el siguiente comando:

sum(condición, na.rm = TRUE)

***Análisis de Componentes Principales (PCA)*:** permite visualizar la variabilidad entre las muestras. Requiere como entrada el objeto ("vsd") generado a partir de la función *vst*(), que realiza la transformación de estabilización de varianza, lo que significa que la varianza de la expresión génica se estabiliza en relación con el nivel medio de expresión (Love et al., 2016). Para generar este gráfico se usó el comando:

plotPCA(vsd, intgroup = c("condicion"))

Conteo de genes: Para la valoración gráfica del conteo de genes, se usó *plotCounts* sobre el resultado del análisis de expresión génica diferencial (*DESeq*) (Love et al., 2016).

Mapas de calor: Con los datos de varianza estabilizada se usa el comando *pheatmap(),* para generar un mapa de calor, sobre una matriz de distancias (*sampleDistMatrix*). Se específica *sampleDists()*, para agrupar las muestras en términos de distancia; tanto en filas como en columnas.

***Gráfico de dispersión tipo MA (Media-Adjusted)*:** para este gráfico se usó la función *DESeq2::plotMA*(), utilizado en análisis de expresión génica diferencial, permite visualizar las diferencias entre las medias de expresión de dos condiciones, por ejemplo: "Desarrollo versus Completa".

Gráfico de volcanes: es un gráficos de dispersión, para generarlo se utilizó la función *EnhancedVolcano*() que es parte del paquete EnhancedVolcano. Sirve para visualizar resultados de análisis de expresión génica diferencial, para el diagrama básico se usó el siguiente comando (Blighe et al., 2023):

EnhancedVolcano(res, lab = rownames(res), x = 'log2FoldChange', y = 'pvalue')

Adicionalmente, en el mismo gráfico, se seleccionaron las etiquetas de los genes que codifican para los alérgenos: Ara h 1, 2 y 3; para visualizar su comportamiento en cada comparación de estadios realizada.

4. Resultados y discusión

4.1. Genes alérgenos de la semilla del maní (Ara h)

A continuación, se presenta la recopilación de información de los genes alérgenos del maní, descritos en la base de datos NCBI (NCBI, 2023; Palladino & Breiteneder, 2018). Mediante la revisión de los alérgenos, se condensó información sobre la descripción del gen, código del gen, cromosoma y posición en el cromosoma de los alérgenos Ara h 1, 2, 3, 5, 6, 8, 10, 11, 14 y 15 (Tabla 4).

Tabla 4. *Genes que codifican para los alérgenos Ara h, descritos en NCBI*

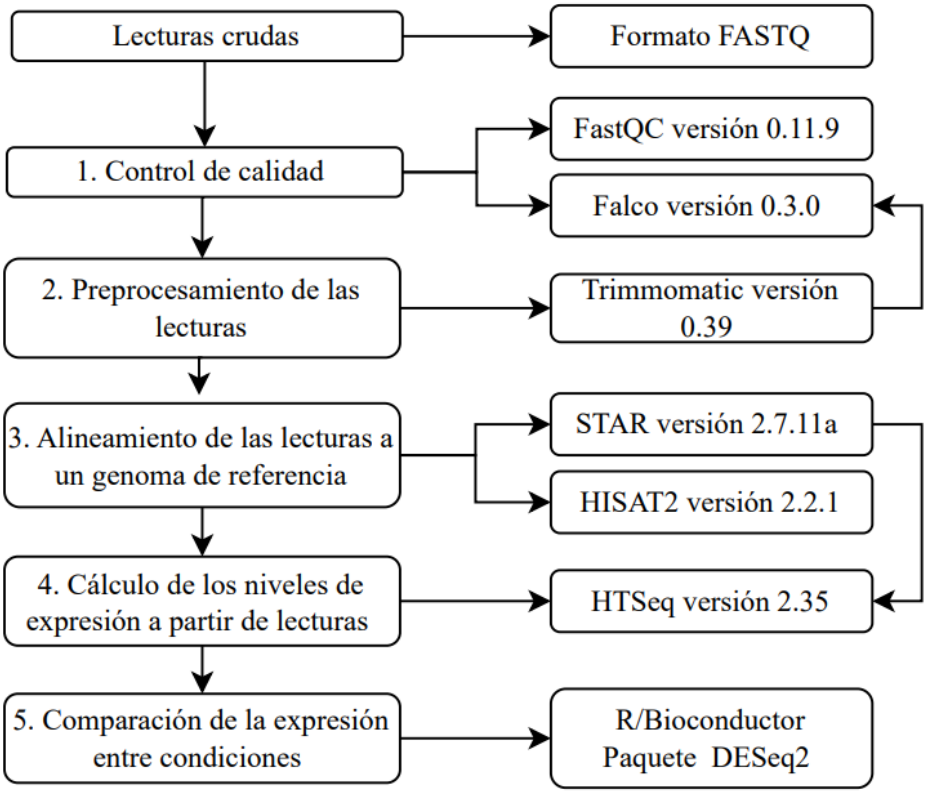
ALÉRGENO	DESCRIPCIÓN DEL GEN	CÓDIGO DEL GEN	CROMOSOMA	POSICIÓN
Ara h 1	Allergen Arah1	LOC112776552	Arahy.19	NC_037636.1:157323835-157326288
Ara h 1	allergen Ara h 1	LOC112711772	Arahy.09	NC_037626.1:110643169-110645412
Ara h 2	Conglutin-7-like	LOC112707245	Arahy.08	NC_037625.1:35990900-35991697
Ara h 2	conglutin-7-like	LOC112771110	Arahy.18	NC_037635.1 (15347052..15347887)
Ara h 3	Arachin Ahy-3-like	LOC112695262	Arahy.06	NC_037623.1 (1856158..1858326, complement)
Ara h 3	Arachin Ahy-3-like	LOC112758686	Arahy.16	NC_037633.1 (26550549..26553059, complement)
Ara h 5	Profilin-1	LOC112802061	Arahy.05	NC_037622.1:95252501-95254198
Ara h 6	Conglutin	LOC112771111	Arahy.18	NC_037635.1:15355187-15355937
Ara h 6	Conglutin-like	LOC112771582	Arahy.18	NC_037635.1 (15404224..15404974)
Ara h 8	Pathogenesis-related protein 2	LOC112723125	Arahy.11	NC_037628.1 (142222599..142224056, complement)
Ara h 10	P24 oleosin	LOC112755203	Arahy.16	NC_037633.1 (137133853..137134849)
Ara h 10	P24 oleosin	LOC112697300	Arahy.06	NC_037623.1 (104906153..104907145)
Ara h 11	Oleosin 1	LOC112784540	Arahy.20	NC_037637.1 (137183341..137184034, complement)
Ara h 11	Oleosin 1	LOC112716547	Arahy.10	NC_037627.1 (110744736..110745426, complement)
Ara h 14	Oleosin 5	LOC112801943	Arahy.05	NC_037622.1 (92473265..92474181)
Ara h 14	Oleosin 5	LOC112751413	Arahy.15	NC_037632.1 (152600196..152601038, complement)
Ara h 15	Oleosin Zm-I	LOC112706321	Arahy.08	NC_037625.1 (16295765..16296774, complement)

Nota. NCBI (2023).

4.2. Flujo de trabajo

Se realizó un flujo de trabajo para el análisis de datos procedente de RNA-seq (Figura 4), conforme se desarrolló en la presente investigación. Este flujo de trabajo puede ser empleado para análisis de expresión diferencial de especies diversas.

Figura 4. Diagrama de flujo para el análisis de expresión génica en genes de *Arachis hypogaea* L.



4.3. Control de calidad de las lecturas

Mediante las herramientas FastQC y FALCO, se generaron las principales métricas para el control de calidad de las nueve secuencias, entre estas encontramos: estadísticas básicas, calidad de secuencia por base, contenido de secuencia por base y contenido de guanina-citosina (GC) por secuencia (Anexo 1).

A partir de las estadísticas básicas generadas con FASTQC y Falco, se realizó un resumen de los resultados obtenidos (Tabla 5), en este se puede observar el total de lecturas en cada corrida, encontrándose en un rango entre 14,769,537 y 16,668,038 lecturas. La longitud de las secuencias reporta un único valor, por lo que se concluye que las secuencias obtenidas

tienen una longitud de 101 pb. Korpelainen et al. (2015), indican que las lecturas de Illumina originalmente son de longitud uniforme.

Tabla 5. *Estadísticas básicas generadas por las herramientas FastQC y Falco, para las lecturas crudas.*

Estadio de la semilla	Código y repetición	Total de secuencias	Longitud de la secuencia	%GC	Secuencias de mala calidad	Calidad media de la secuencia (puntuación de Phred)	Contenido de secuencia por base
Desarrollo	R5_r1	14,899,080	101	43	0	30	Falla
Desarrollo	R5_r2	16,013,160	101	43	0	30	Falla
Desarrollo	R5_r3	16,060,756	101	44	0	30	Falla
Completa	R6_r1	16,412,045	101	46	0	30	Falla
Completa	R6_r2	16,005,552	101	46	0	30	Falla
Completa	R6_r3	16,219,449	101	46	0	30	Advertencia
Madura	R7_r1	16,784,061	101	46	0	30	Falla
Madura	R7_r2	14,769,537	101	46	0	30	Falla
Madura	R7_r3	16,668,038	101	46	0	30	Falla

El contenido de GC fue de 43 y 44 en las semillas en desarrollo (R5), y fue de 46 en las semillas completas (R6) y maduras (R7), presentando una distribución normal. La calidad media para todas las secuencias en los dos programas (FastQC y Falco), fue igual a 30, indicando una buena calidad de las lecturas (Tabla 5, Anexo 1). Los valores de calidad normalmente oscilan entre 0 y 40, idealmente la mayoría de las lecturas deberían tener una calidad base media de 25 o mayor (Korpelainen et al., 2015).

De las nueve corridas, ocho reportaron *error* y una reportó *advertencia* (Completa, R6_r3) para el *contenido de secuencia por base* (Tabla 5). Esta métrica emite un *error* por la diferencia entre A y T, o G y C superior al 20%, en cualquier posición; y emite una *advertencia* al ser superior al 10% (Babraham Institute, 2020).

Con base en los resultados del reporte (Tabla 5), todas las lecturas fueron preprocesadas con un recorte de las primeras 14 bases. El control de calidad después del recorte se realizó con la herramienta Falco (Figura 4, Anexo 2), las métricas: total de secuencias, porcentaje de GC, secuencias de mala calidad y calidad media de la secuencia, se mantuvieron iguales. La longitud de todas las secuencias se redujo a 87, debido a que se eliminaron 14 bases al principio de las mismas. Y el contenido de secuencia por base, no presentó error ni advertencia (Anexo 3).

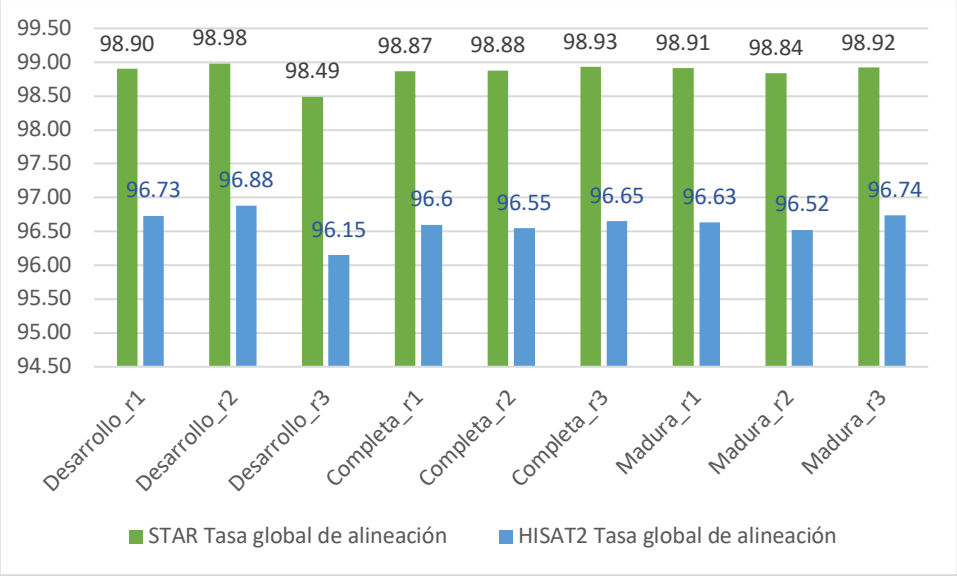
4.4. Alineamiento

El alineamiento se realizó mediante dos herramientas, STAR y HISAT2. Al realizar el análisis de la distribución de los datos generados, mediante Shapiro-Wilks, se encontró una distribución normal de los mismos (p-value = 0.034). Se realizó el análisis de varianza (ANOVA), STAR versus HISAT2, seguida de la prueba post-hoc de Tukey. Se encontró que no presentan diferencias significativas en los resultados de conteo (valor de p = 0.6612), es decir que los datos del conteo para las dos herramientas son similares al 95% de confianza.

4.4.1. STAR versus HISAT2

La tasa de alineamiento global en ambas herramientas supera el 96% (Figura 5). Este porcentaje indica: el número de lecturas asociadas al genoma de referencia respecto al total de la muestra. A mayor porcentaje, mayor calidad del resultado (Campoy-García, 2019).

Figura 5. *Tasa de alineamiento global de HISAT2 versus STAR*

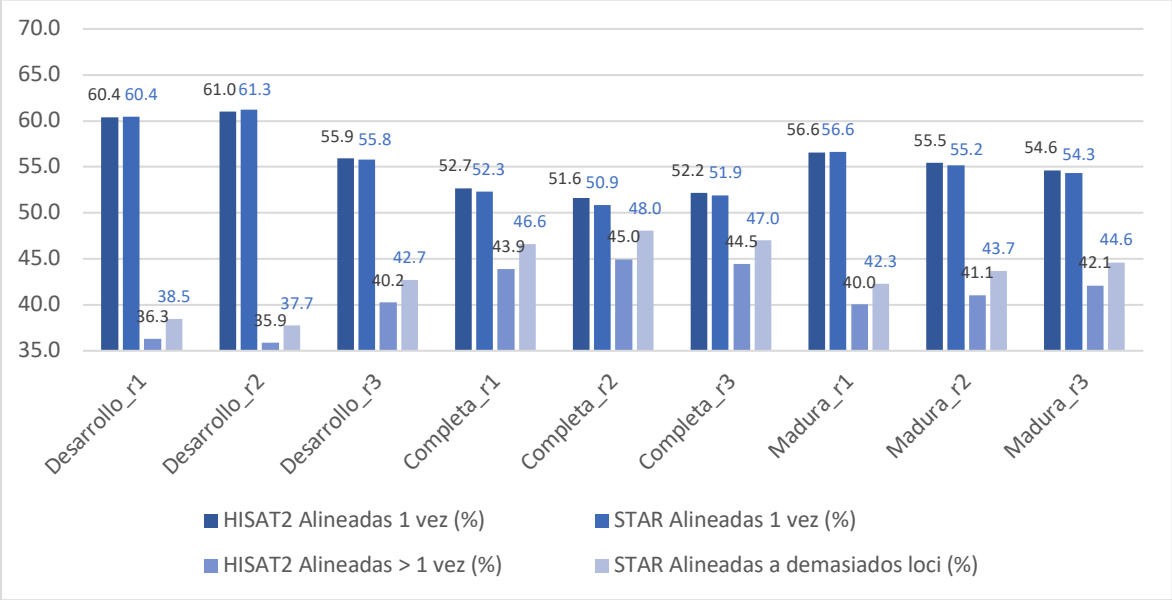


Al comparar los resultados de los conteos obtenidos se observa que son muy similares, tanto para lecturas alineadas una vez, como para lecturas alineadas más de una vez (Figura 6), respecto a lecturas no alineadas STAR presenta un menor porcentaje (Anexo 4). Corchete (2019), reporta resultados similares en su investigación, ya que STAR fue el algoritmo con mayor porcentaje de lecturas únicas alineadas de forma concordante, y presentó el menor porcentaje de fallas en el mapeo.

Debido a los resultados presentados se podrían usar ambas herramientas indistintamente, en la presente investigación para continuar con análisis diferencial, se seleccionaron los

resultados de STAR para continuar con las siguientes fases (Figura 4). STAR presenta algunas ventajas además de su velocidad, realiza una búsqueda no sesgada de empalmes porque no necesita información previa sobre su localización, divide una lectura en trozos (por defecto 50 bases) y encuentra la mejor porción que se puede mapear para cada trozo, busca coincidencias exactas y utiliza el genoma en forma de sufijo sin comprimir. STAR se adapta mejor a los desajustes, funciona más rápido, produce más alineaciones, y puede detectar empalmes de manera imparcial (Korpelainen et al., 2015).

Figura 6. *Tasa de alineamiento de HISAT2 versus STAR, para lecturas alineadas una vez y lecturas alineadas más de una vez*



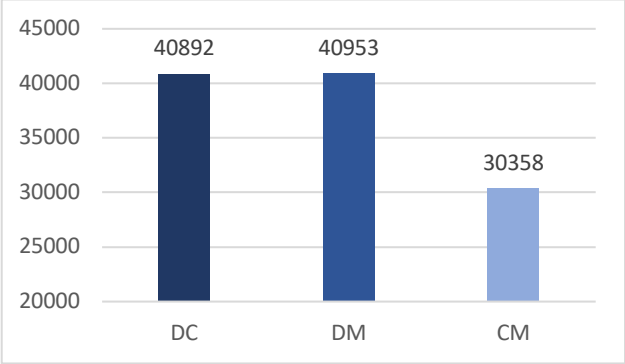
4.5. Análisis de Expresión Diferencial

A partir de los archivos de conteo (salida de HTSeq), se encontró que para Ara h 1 (LOC112776552) el rango de conteo fue de 8266 (R5) y 689,155 (R7), para Ara h 1 (LOC112711772) el rango de conteo estuvo entre 10,129 (R5) y 556734 (R7). Mientras que para Ara h 2 (LOC112707245) el menor conteo fue de 5566 (R5) y el mayor fue 359,125 (R7), y para Ara h 2 (LOC112771110) 3064 fuel el menor conteo (R5), y 293,476 el mayor valor de conteo (R7). Para Ara h 3 (LOC112695262) el rango de conteo se encontró entre 20,315 (R5) y 417,746 (R6); mientras que, para Ara h 3 (LOC112758686) el rango de conteo fue de 3 (R5) y 88 (R6) (Tabla 4). A partir de estos archivos se realizó el análisis de expresión diferencial con DESeq, en comparaciones por pares.

4.5.1. Filtrado independiente

El archivo de conteos original contiene 104,401 genes, estos se reducen después del prefiltrado de los conteos (Figura 7, Anexo 5). La condición planteada fue que sea superior a 3 en más de dos de las muestras. Es recomendable filtrar los transcritos de baja expresión, antes del análisis diferencial basado en el recuento (Korpelainen et al., 2015).

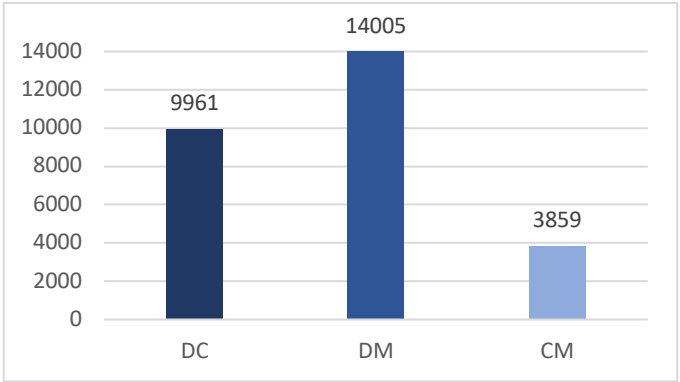
Figura 7. *Conteo final de genes para cada comparación de estadios, después del filtrado independiente.*



4.5.2. Resultados Deseq

A continuación, en la figura 8 se observa el total de genes con valores de p ajustados, menores o iguales a 0.05, indicando el total de genes expresando diferencialmente por par de comparaciones (Anexo 5). Se observa que la comparación de las semillas completa versus madura, es la que menos genes muestran diferencias significativas en la expresión entre las condiciones.

Figura 8. *Conteo de genes expresados diferencialmente (padj=0.05) para cada comparación de estadios, después del análisis de expresión diferencial (DESeq)*

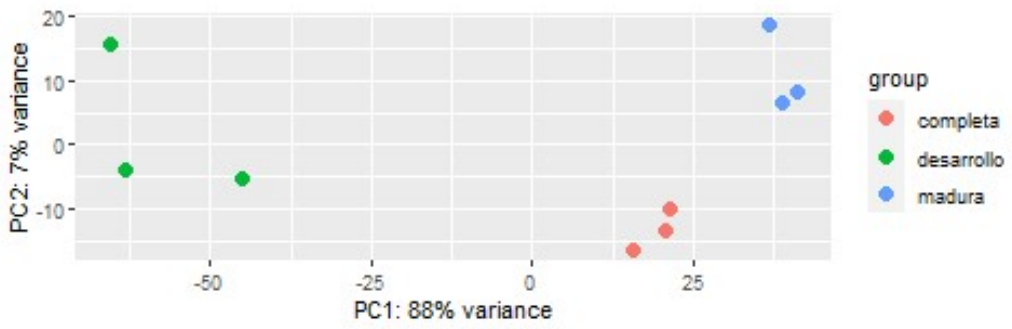


4.5.2.1. Análisis de componentes principales (PCA)

Inicialmente, se realizó una corrida general de todas las muestras, y se obtuvo un gráfico de componentes principales (PCA), para visualizar la variabilidad entre todas las muestras; este gráfico permite reducir la dimensionalidad de un conjunto de datos multivariado (Korpelainen et al., 2015).

Se observa que hay tres grupos claros, las repeticiones correspondientes a cada estadio de la semilla tienden a agruparse, es decir se agrupan los perfiles de expresión similares (Figura 9). Se observa que las muestras presentan una mayor variabilidad entre los diferentes estadios de las semillas, que entre repeticiones del mismo estadio. Aparentemente, el grupo de semillas completas muestran mayor similitud con las semillas maduras.

Figura 9. *Análisis de componentes principales (PCA), para los tres estadios de la semilla.*



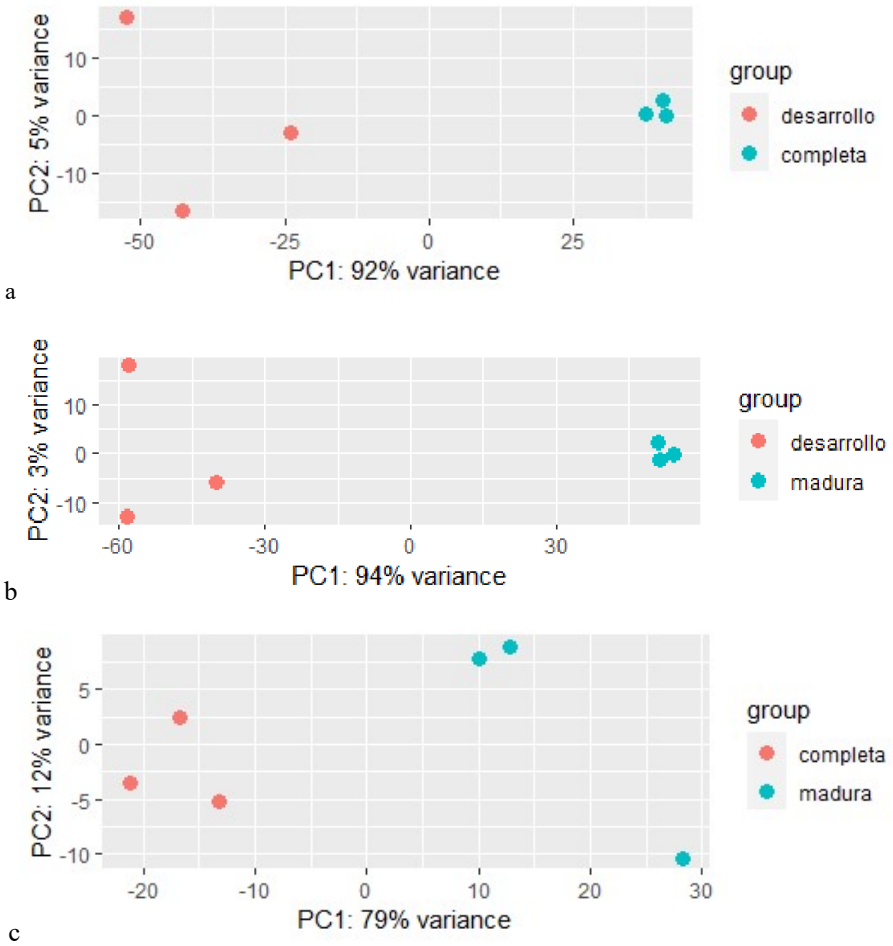
El PCA confirma que los componentes, reproducen los tres grupos en estudio, se observa en color verde las semillas en Desarrollo, en rojo las semillas Completas y en azul las semillas Maduras (Figura 9). Mediante estos gráficos de PCA se comprobó la similitud entre repeticiones, previo al análisis de expresión diferencial.

En la figura 10, se muestran los resultados de las diferentes comparaciones en pares. En todas las figuras (10a, 10b, 10c), se observa que el PC1 presenta altos porcentajes de la variación total, en los datos de expresión, en un rango entre el 79 y 94%; mientras que, el porcentaje de la variación total de PC2 varía entre 3 y12 %.

Por la suma de componentes principales (PC1 y PC2) de cada comparación, se obtiene que DC (Desarrollo versus Completa) presenta un 97% de variación, DM (Desarrollo versus Madura) presenta un 97% de variación y CM (Completa versus Madura) presenta un 91% de variación. Esto indica que ambos componentes explican la mayoría de la varianza que existe entre las comparaciones. Se puede decir entonces, que los diferentes estadios de la

semilla sí están generando diferencia significativa, pues se observa el agrupamiento de muestras por estadio.

Figura 10. *Análisis de componentes principales (PCA), para las comparaciones en pares de los tres estadios de la semilla.*



Nota: La figura 10a, corresponde a la comparación DC (Desarrollo versus Completa). La figura 10b, corresponde a la comparación DM (Desarrollo versus Madura). La figura 10c, corresponde a la comparación CM (Completa versus Madura).

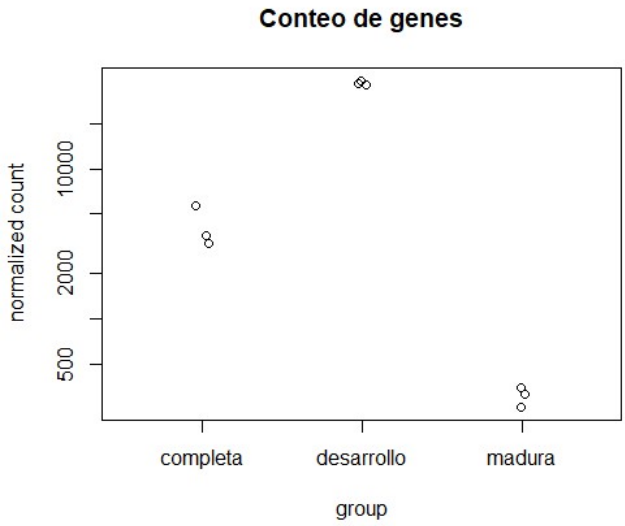
4.5.3. Conteo de genes

La gráfica de conteo de genes, se generó mediante la función *plotCounts,* para una visualización rápida de los genes con mayor expresión diferencial, es decir aquellos con el valor de p ajustado más bajo (menor a 0.05).

Se observa la diferencia de conteo entre los diferentes estadios y sus comparaciones en pares, el conteo normalizado de genes se encuentra en un rango alrededor de 500 para semilla

madura, cerca de 2000 para semilla completa, y es mayor a 10,000 para semilla en desarrollo (Figura 11, Anexo 6). Esto nos indica que hay diferencias de conteo de genes con alta expresión diferencial entre los estadios; las semillas en desarrollo presentan el mayor conteo, seguidas por semillas completas y finalmente el menor conteo se observa en semillas maduras.

Figura 11. *Conteo de genes con mayor expresión diferencial (valor de p ajustado 0.05), para las comparaciones de los tres estadios de la semilla.*

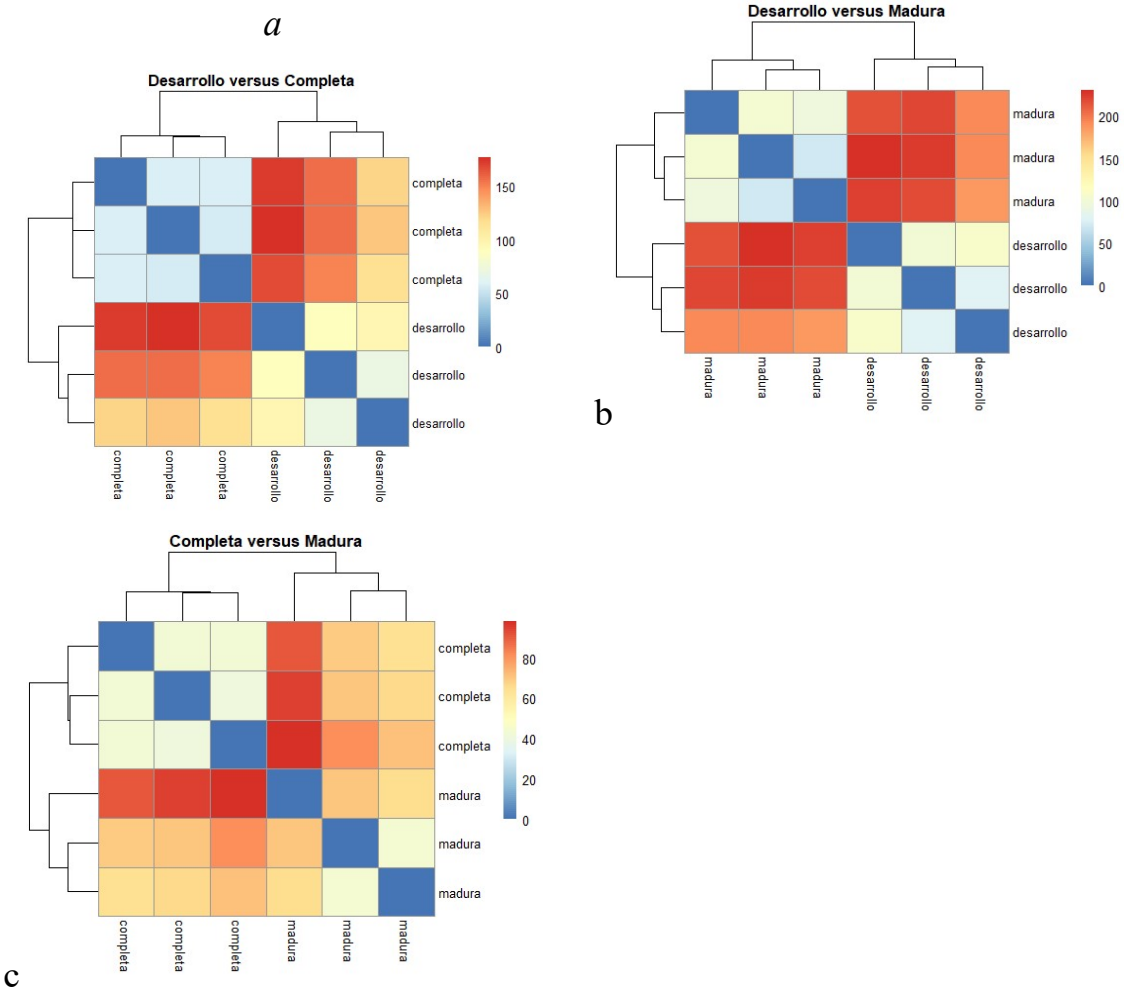


4.5.4. Mapas de calor

En los mapas de calor se observa que todas las repeticiones correspondientes a un mismo estadio de la semilla forman un solo grupo (*cluster*), indicando que son similares (Figura 12). Los mapas de calor dan una descripción general de las similitudes y diferencias entre muestras (Love et al., 2015, 2016). Conforme se visualiza en la escala, el color azul significa una distancia más corta (0) entre las muestras (Figura 12), lo que significa que hay mayor similitud entre ellas; mientras que, conforme los cuadrantes se tornan hacia el color rojo, hay mayor diferencia entre las muestras.

Al observar las líneas de distancia de agrupamiento (clusters), en todas las comparaciones las repeticiones de un estadio de las semillas forman un mismo grupo, diferente a las semillas en el otro estadio (Figura 12), por ejemplo, semillas en desarrollo versus completa.

Figura 12. *Mapas de calor para las comparaciones en pares de los tres estadios de la semilla, de los valores normalizados.*



Nota: La figura 12a, corresponde a la comparación DC (Desarrollo versus Completa). La figura 12b, corresponde a la comparación DM (Desarrollo versus Madura). La figura 12c, corresponde a la comparación CM (Completa versus Madura).

Al contrastar entre los cuadrantes desarrollo versus completa (superior derecho o inferior izquierdo), se observa diferencia en la expresión de genes (Figura 12a), pues se identifica que la coloración varía entre amarillo y rojo (100 a 150). Los cuadrantes de semillas en desarrollo versus madura, en la figura 12b, indican la mayor diferencia en la expresión de genes, puesto que la coloración varía entre anaranjado y rojo (cercano a 200). Mientras que, al contrastar repeticiones de un mismo estadio para las comparaciones DC y DM, se observa colores celestes y cremas, indicando una distancia corta entre ellos, por lo tanto, similitud.

Finalmente, los cuadrantes de semillas completa versus madura, en la figura 12c, identifica diferencia en la expresión de genes, ya que la coloración varía entre amarillo y rojo (60 a

80). En todos los casos al contrastar entre repeticiones de un mismo estadio, se observa coloraciones anaranjadas, que implica que entre repeticiones hay diferencias, siendo completa el estadío con menor diferenciación entre réplicas.

4.5.5. Gráfico de dispersión tipo MA (Media-Adjusted)

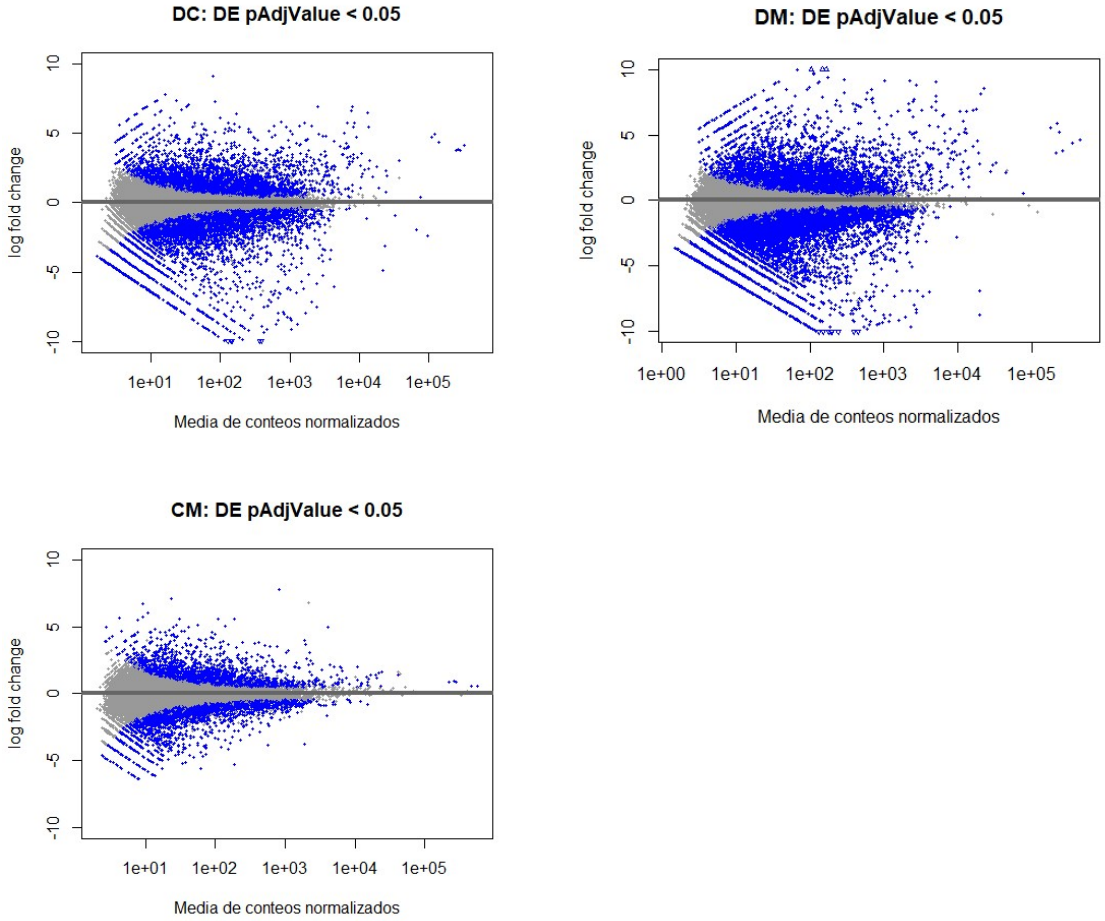
El término expresión génica diferencial hace referencia a la cantidad relativa de mRNA que es observado en una condición respecto a una segunda condición (Corchete, 2019). Al comparar semillas en diferentes estadios de crecimiento (Figura 13), se puede visualizar en gris los genes que se expresan en una proporción similar, mientras que en azul se observan los genes con mayor expresión diferencial ($\alpha = 0.05$). Los gráficos MA facilitan visualizar la diferencia en la expresión génica entre dos condiciones, los puntos se colorearán de azul si el valor p ajustado es menor a 0.05 (diferencia significativa), caso contrario se pintan de color gris. El *eje Y* expresa el cambio proporcional en la expresión de un gen entre las condiciones (como Log_2 *fold change*, LFC); y el *eje X* muestra el nivel de expresión promedio normalizado (Love et al., 2015; McDermaid et al., 2019).

En los gráficos cada punto representa un solo gen, los puntos azules son aquellos que caen por encima o debajo del umbral del eje Y, son los genes que tienen niveles de expresión altamente diferenciales. En la comparación entre semillas en desarrollo versus semillas completas (DC), se visualiza que existen genes regulados positivamente y negativamente (Figura 13DC). Además, se identifica que existen genes subexpresados (LFC más bajo), aquellos fuera de la ventana (triángulos en el borde inferior), es decir, las semillas completas de maní producen menos copias de ciertas proteínas con relación a las semillas en desarrollo.

En la figura 13DM, se observa la comparación entre semillas en desarrollo versus semillas maduras (DM), ésta es la que mayor cantidad de genes regulados positivamente y negativamente, presenta. También se observan los triángulos en la parte superior e inferior de la ventana, indicando que hay genes cuya regulación de expresión es aún mayor (LFC más alto) o menor (LFC más bajo), dejando claro que las semillas en desarrollo presentan patrones de expresión muy diferentes en comparación con las semillas maduras. Según McDermaid et al. (2019), el gráfico MA con una gran cantidad de puntos de datos que caen por encima de un umbral en el eje Y, indicaría una cantidad más significativa de genes que

están regulados positivamente, mientras que por debajo de -1 indicaría altos niveles de regulación negativa en los genes, tal como se observa en la figura 13DM.

Figura 13. *Gráficos de dispersión tipo MA para las comparaciones en pares de los tres estadios de la semilla (p-value adjusted <0.01).*



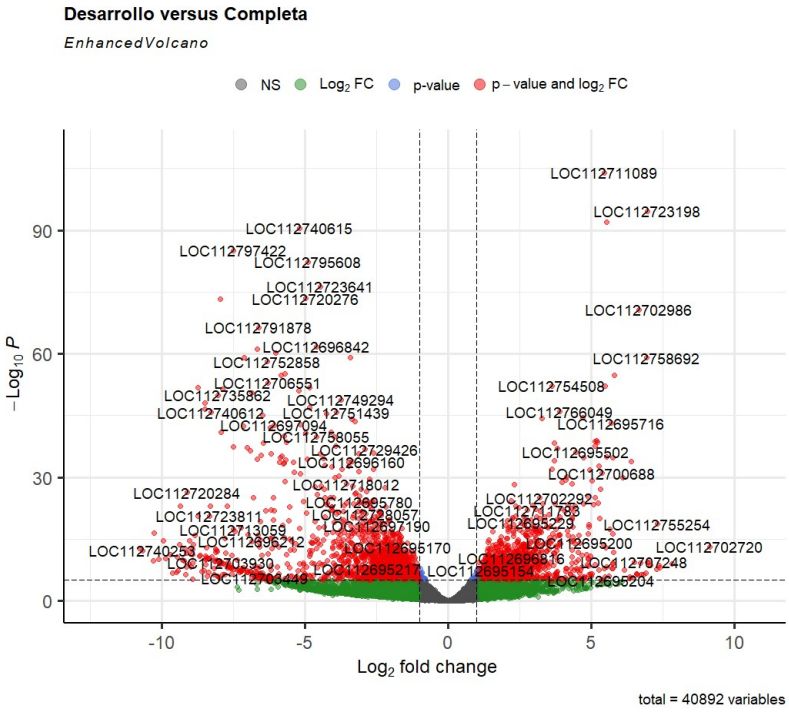
Nota: DC, corresponde a la comparación Desarrollo versus Completa. DM, corresponde a la comparación Desarrollo versus Madura. CM, corresponde a la comparación Completa versus Madura.

Para la comparación de semillas completas versus semillas maduras (Figura 13CM), se observa que es la que menos cantidad de genes (puntos azules) dispersos presenta, sin embargo, indica que hay una regulación de expresión diferente y por tanto sus patrones de expresión no son similares. Los puntos azules que se encuentran en los cuadrantes superior e inferior derecho (figuras 13Dc y 13DM), son los genes con una mayor media de recuentos normalizados y LFC altos, siendo los más interesantes para ser analizados a futuro (McDermaid et al., 2019).

4.5.6. Gráfico de volcanes

El gráfico de volcán permite realizar una comparación entre dos condiciones, presenta en el *eje X* la magnitud de la diferencia en la expresión génica (LFC), y el *eje Y* representa la significación estadística de la diferencia (-log10 del valor de p ajustado).

Figura 14. *Gráfico de volcán de la comparación de semillas en desarrollo versus completas (DC).*

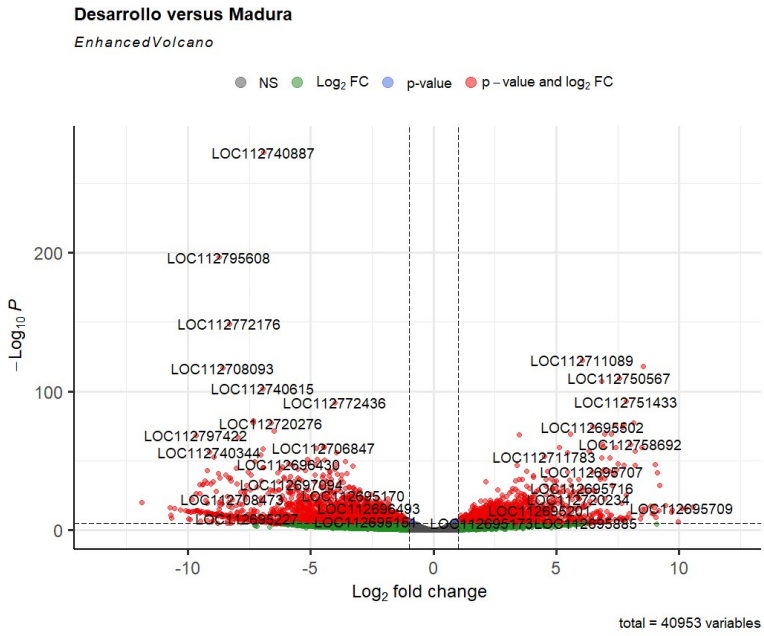


Los genes que no tienen una diferencia estadísticamente significativa, son los puntos que ubican cerca del origen del gráfico; mientras que, los genes que tienen una diferencia estadísticamente significativa y una magnitud de cambio mayor se representan como puntos que están más alejados del origen del gráfico (McDermaid et al., 2019).

La figura 14 muestra la comparación de semillas de maní en desarrollo versus semillas completas, observándose claramente los genes que presentan patrones de expresión diferentes, están con los puntos rojos. Los genes con menor valor de p ajustado (*-log₁₀ P* alrededor de 90), se ubican en la parte superior, son los que presentan mayor diferencia estadísticamente significativa. A la izquierda se encuentran los genes de las semillas completas que se subexpresan en comparación a las semillas en desarrollo, entre ellos se

pueden mencionar LOC112740615 y LOC112797422, ambos no caracterizados en NCBI, y LOC112795608 conocido como proteína específica de las semillas similar a la subtilisina (NCBI, 2023). A la derecha se encuentran los genes de las semillas completas que se sobreexpresan en comparación a las semillas en desarrollo, entre ellos están LOC112711089 (Inhibidor de la proteinasa tipo Bowman-Birk A-II) y LOC112750567 (proteína SLE2) (NCBI, 2023).

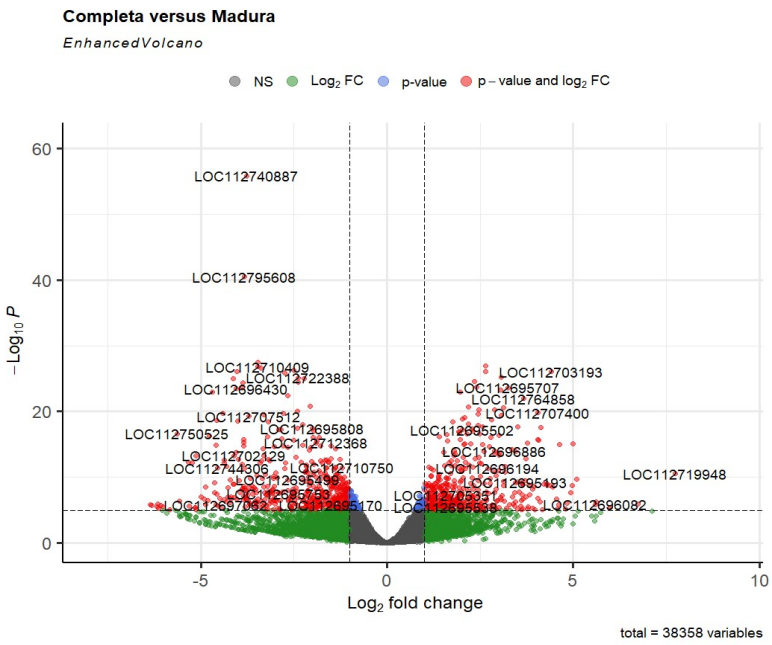
Figura 15. *Gráfico de volcán de la comparación de semillas en desarrollo versus maduras (DM).*



La figura 15 muestra la comparación de semillas de maní en desarrollo versus semillas maduras, los puntos rojos muestran los genes que presentan patrones de expresión diferentes. Hay más genes con menor valor de p ajustado (*-log$_{10}$ P* alrededor de 200), ubicados en la parte superior, en el cuadrante izquierdo, indicando que la proporción de expresión de las medias fue menor para las semillas maduras en comparación a las semillas en desarrollo.

Entre los genes identificados en este cuadrante están LOC112740887 y LOC112795608 identificados como "proteasa similar a la subtilisina SBT5.3", LOC112772176 y LOC112708093, conocidos como "proteína de la familia de la glicosil hidrolasa 5". En el cuadrante derecho el gen con menor valor de p ajustado fue LOC112711089, nombrado como "Inhibidor de la proteinasa tipo Bowman-Birk A-II" (NCBI, 2023).

Figura 16. *Gráfico de volcán de la comparación de semillas completas versus maduras (CM)*



La figura 16 permite visualizar los genes que presentan patrones de expresión diferentes (puntos rojos), en la comparación de semillas de maní en desarrollo versus semillas completas. Los genes con menor valor de p ajustado (*-log₁₀ P* entre 40 y 60), se ubican en la parte superior izquierda, indicando mayor diferencia estadísticamente significativa. Entre los genes que se subexpresan en las semillas maduras en comparación con las semillas completas, se pueden mencionar LOC112740887 y LOC112795608, identificados como "proteasa similar a la subtilisina SBT5.3" (NCBI, 2023).

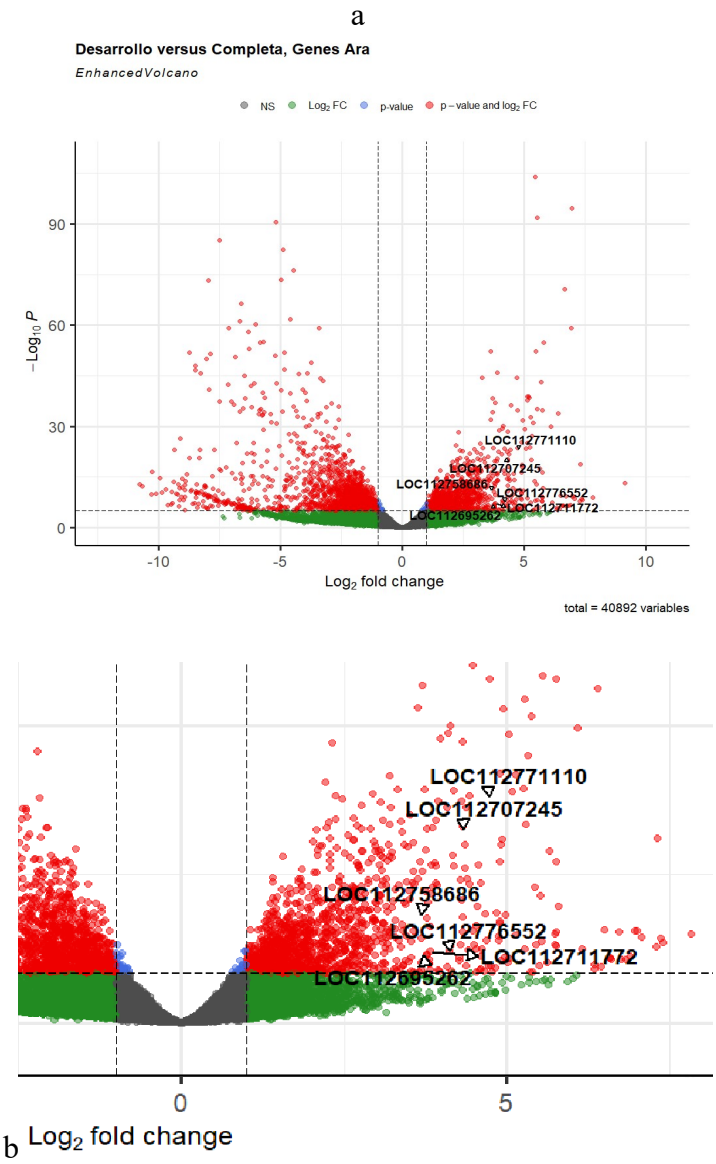
4.6. Expresión diferencial de los genes Ara h 1, 2 y 3

A continuación, se muestra la expresión diferencial de los genes alérgenos Ara h 1, 2 y 3, en tres condiciones de desarrollo de la semilla del maní (D: en desarrollo, C: completa y M: madura).

4.6.1. Semillas en desarrollo versus semillas completas (DC)

En la comparación de expresión de las semillas de maní en desarrollo versus completas, se encontró que los genes Ara h 1, 2 y 3, se expresan diferencialmente, ya que reportan un valor de p ajustado menor a 0.01 (Figura 17, anexo 7).

Figura 17. *Gráfico de volcán de la comparación de semillas en desarrollo versus completa (DC), identificando los genes Ara h1 (*LOC112776552 y LOC112711772*), Ara h2 (LOC112707245 y LOC112771110) y Ara h3 (LOC112695262 y LOC112758686).*



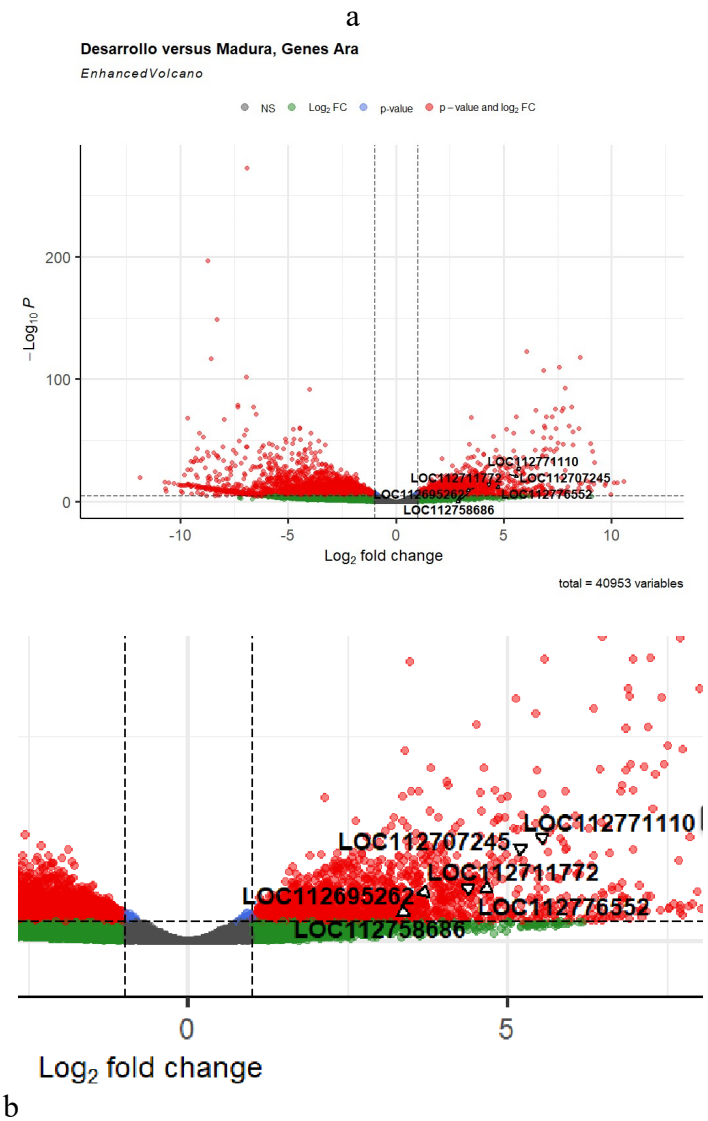
Nota: a. Gráfico de volcán completo. b. Ampliación del gráfico de volcán.

Los genes Ara h2, LOC112707245 y LOC112771110, incrementan su expresión en 4.33 y 4.71 LFC, respectivamente, en las semillas completas en relación con la semilla en desarrollo (Figura 17b). El cambio de expresión de los genes Ara h1, LOC112776552 y LOC112711772 fue de 4.11 y 3.83 LFC, respectivamente. Mientras que los genes Ara h3, LOC112695262 y LOC112758686, presentan un cambio de expresión en 3.79 y 3.69 LFC, respectivamente (Anexo 7a, tabla 4). Lo que significa que las semillas completas expresan entre 3.69 y 4.71 veces más genes codificantes para alérgenos, que las semillas en desarrollo.

4.6.2. Semillas en desarrollo versus semillas maduras (DM)

En la comparación de expresión de las semillas de maní en desarrollo versus maduras, se encontró que los genes Ara h1, Ara h2 y Ara h3, se expresan diferencialmente, ya que presentan un valor de p ajustado menor a 0.01 (Figura 18, anexo 7).

Figura 18. *Gráfico de volcán de la comparación de semillas en desarrollo versus madura (DM identificando los genes Ara h1 (*LOC112776552 y LOC112711772*), Ara h2 (LOC112707245 y LOC112771110) y Ara h3 (LOC112695262 y LOC112758686).*



Nota: a. Gráfico de volcán completo. B. Ampliación del gráfico de volcán.

Los genes Ara h2, LOC112707245 y LOC112771110, incrementan su expresión en 5.19 y 5.54 LFC, respectivamente, en las semillas completas en relación con la semilla en desarrollo

(Figura 18b). El cambio de expresión de los genes Ara h1, LOC112776552 y LOC112711772 fue de 4.66 y 4.38 LFC, respectivamente. Mientras que los genes Ara h3, LOC112695262 y LOC112758686, presentan un cambio de expresión en 3.66 y 3.37 LFC, respectivamente (Anexo 7b, tabla 4). Lo que significa que las semillas maduras expresan entre 3.66 y 5.54 veces más genes codificantes para alérgenos, que las semillas en desarrollo.

4.6.3. Semillas completas versus semillas maduras

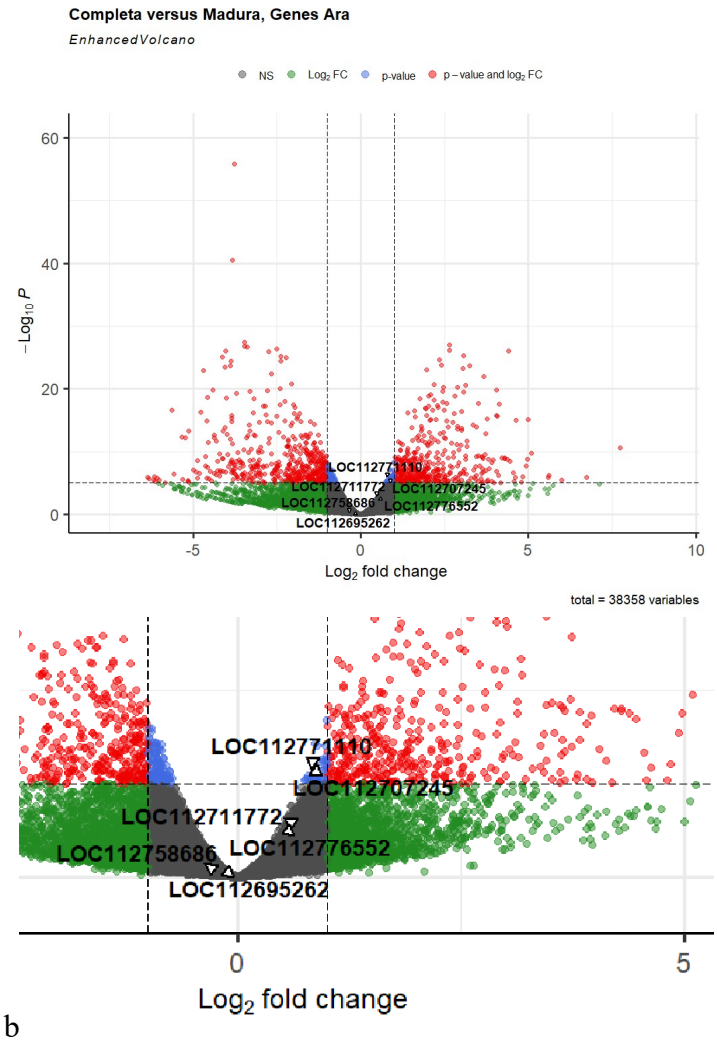
En la comparación de expresión de las semillas de maní en desarrollo versus completa, se encontró que los genes Ara h2, presentan un valor de p ajustado menor a 0.01; los genes Ara h1, de 0.036 y 0.014; mientras que los genes Ara h3, presentan valores de p ajustado de 0.833 y 0.670 (Anexo 7c).

Respecto a LFC los genes Ara h2, LOC112707245 y LOC112771110, presentan valores de 0.87 y 0.84 LFC, respectivamente, los genes Ara h1, LOC112776552 y LOC112711772 fue de 0.56 y 0.57 LFC, respectivamente; mientras que, para los genes Ara h3, LOC112695262 y LOC112758686, fue de -0.104 y -0.292 LFC, para las semillas completas en relación con la semilla madura (Anexo 7c, tabla 4).

En el gráfico de volcán (Figura 19), al evaluar simultáneamente LFC (eje x) con el valor de p ajustado (eje y), se observa que los genes de alérgenos estudiados en esta investigación se encuentran muy cercanos al origen (zona cero), mientras más cerca de cero indica menos cambio. En este caso los genes Ara h1 y Ara h3, que se encuentran en la zona gris y Ara h2 en la zona azul, indicando que no presentan cambios de expresión significativos entre semillas maduras con relación a semillas completas.

Figura 19. *Gráfico de volcán de la comparación de semillas en completa versus madura (CM), identificando los genes Ara h1 (*LOC112776552 y LOC112711772*), Ara h2 (LOC112707245 y LOC112771110) y Ara h3 (LOC112695262 y LOC112758686).*

a



b

Nota: a. Gráfico de volcán completo. b. Ampliación del gráfico de volcán.

4.6.4. Discusión

Arachis hypogaea es una fuente de los alérgenos alimentarios más graves, entre los principales se encuentran Ara h 1, 2 y 3 (I. H. Kang, 2004; I. H. Kang et al., 2007; Palladino & Breiteneder, 2018), estos son los principales desencadenantes de reacciones alérgicas en Estados Unidos (Palladino & Breiteneder, 2018). Los alérgenos en la semilla de maní se sintetizan y degradan durante sus diferentes etapas de formación (I. H. Kang, 2004), es por ello que se encontró diferentes niveles de expresión de alérgenos Ara h 1, 2 y 3 (Anexo 7), en cada etapa de formación de la semilla de maní.

En los resultados del presente estudio, se encontró que la presencia de alérgenos es mayor en las semillas completas y maduras en relación con las semillas en desarrollo (Figuras 17 y 18, anexo 7), lo que puede estar relacionado con la expresión de los alérgenos según las partes de la semilla; además con la formación y crecimiento del cotiledón y el embrión.

Respecto a la expresión de alérgenos, en la formación de la semilla de maní, los ejes embrionarios y los cotiledones los expresan; mientras que, desaparecen durante la germinación (Jiang et al., 2011; Kang, 2004; Kang et al., 2007), pues no hay transcripción para los alérgenos en raíces, hojas y flores (I. H. Kang et al., 2007), lo que indica que las proteínas Ara h, se expresan únicamente en las semillas.

Sustentado lo expuesto, los resultados de Jiang et al. (2011), reportaron que el nivel de transcripción de Ara h 3 fue 5.13 mayor en los cotiledones que en el embrión, Ara h 1 presentaba en niveles similares (cotiledones y embrión), y Ara h 2 los cotiledones presentaban 0.33 veces lo que el embrión. Además, Kang et al. (2007), concluye que la acumulación de los alérgenos está regulada a nivel transcripcional, en su investigación reportaron que el incremento de Ara h 1 y 2 durante el desarrollo de la semilla, fue similar a la acumulación de transcripción (ARNm) (Jiang et al., 2011; Kang et al., 2007).

Con relación a la formación del cotiledón y el embrión, en este estudio se evaluaron tres estadios de formación de las semillas de maní: en desarrollo (D: R5), son aquellas que han pasado la fase de endospermo líquido, se hace visible la forma del cotiledón (Boote, 1982), pero aún es muy pequeño razón por la cual la expresión de alérgenos es menor. Las semillas completas (C: R6) presentan cotiledones bien formados, representa cerca del 50% de peso seco; y las semillas maduras (M: R7), alcanzan o hasta superan el 90% del peso seco (Boote, 1982; Pattee et al., 1974), alcanzando los mayores niveles de expresión de alérgenos.

También fueron reportados resultados similares por Kang (2004), que encontró que los niveles de transcripción de alérgenos en las semillas maduras fueron mayores en comparación con las semillas inmaduras. Esto se debe a que las semillas maduras acumulan las proteínas de almacenamiento (alérgenos), que actúan como una fuente de aminoácidos; por lo tanto, son utilizados para proporcionar energía durante el crecimiento de las semillas, para el proceso de germinación y para sintetizar nuevas proteínas (Palladino & Breiteneder, 2018). Ara h 1 y 3 son proteínas de almacenamiento de semillas de la familia globulina, y

Ara h2 es una proteína de almacenamiento de semillas de la familia albúmina (Palladino & Breiteneder, 2018; Zhuang & Dreskin, 2013).

La presente investigación permitió identificar la expresión diferencial de los alérgenos Ara h 1, 2 y 3 en tres estadios de la semilla de maní, siendo las semillas completas y maduras las que mayor cantidad de alérgenos presentan. Estos hallazgos proporcionan una base sólida para futuras investigaciones que podrían incluir la evaluación de la expresión diferencial de otros alérgenos tipo Ara h, la comparación de diferentes variedades de maní y la evaluación de la semilla madura en diferentes condiciones de almacenamiento y procesamiento. Estos estudios adicionales permitirían identificar variedades y procesos que generen maní con una menor cantidad de alérgenos, lo que contribuiría a mejorar la seguridad alimentaria y reducir el riesgo de alergias al maní.

5. Conclusiones

- Existe expresión genética diferencial de los genes Ara h 1,2 y 3 entre los diferentes estadios de la semilla de maní. Siendo la semilla en desarrollo la que presenta menor expresión de los alérgenos.
- Los genes Ara h 1 y 2, son los que mayor expresión genética diferencial presentan, en las semillas completas y maduras, en comparación con las semillas en desarrollo.
- Las semillas en estadios completo y maduro, son las que menor expresión diferencial presentan para los genes Ara h 1 y 2; sin embargo, no presentan diferencias para Ara h 3.

6. Bibliografía

Anders, S. (2018). *HTSeq Documentation* (Release 0.10.0). https://htseq.readthedocs.io/_/downloads/en/release_0.10.0/pdf/

Anello, M., Arnal, N., & Barbisan, G. (2021). *Elementos de genética para estudiantes de Ciencias Biológicas* (C. Catanesi & E. Villegas Castagnasso, Eds.). Universidad Nacional de La Plata (UNLP). http://sedici.unlp.edu.ar/bitstream/handle/10915/129625/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Babraham Institute. (2020). *FastQC* (0.11.9). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Basha, S. M. (1988). Resolution of peanut seed proteins by high-performance liquid chromatography. *Journal of Agricultural and Food Chemistry*, *36*(4), 778–781. https://doi.org/10.1021/jf00082a027

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., … Ozias-Akins, P. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics*, *48*(4), 438–446. https://doi.org/10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., Leal-Bertioli, S. C. M., Ren, L., Farmer, A. D., Pandey, M. K., Samoluk, S. S., Abernathy, B., Agarwal, G., Ballén-Taborda, C., Cameron, C., Campbell, J., Chavarro, C., Chitikineni, A., Chu, Y., … Schmutz, J. (2019). The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nature Genetics*, *51*(5), 877–884. https://doi.org/10.1038/s41588-019-0405-z

Bi, Y.-P., Liu, W., Xia, H., Su, L., Zhao, C.-Z., Wan, S.-B., & Wang, X.-J. (2010). EST sequencing and gene expression profiling of cultivated peanut (*Arachis hypogaea* L.). *Genome*, *53*(10), 832–839. https://doi.org/10.1139/G10-074

Blighe, K., Rana, S., & Lewis, M. (2023, December 16). *EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling*. Bioconductor.Org. https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html

Blumenberg, M. (2019). *Transcriptome Analysis* (M. Blumenberg, Ed.). IntechOpen. https://doi.org/10.5772/intechopen.77860

Bolger, A. M., Lohse, M., & Usadel, B. (2014). *Trimmomatic: A flexible read trimming tool for Illumina NGS data*. Bioinformatics. http://www.usadellab.org/cms/index.php?page=trimmomatic

Boote, K. J. (1982). Growth Stages of Peanut (*Arachis hypogaea* L.)1. *Peanut Science*, *9*(1), 35–40. https://doi.org/10.3146/i0095-3679-9-1-11

Campoy-García, M. E. (2019). *Comparación de métodos de cálculo de expresión génica basados en RNA-seq* [Universitat Oberta de Catalunya (UOC)]. https://openaccess.uoc.edu/bitstream/10609/90627/6/elenacampoygTFM0119memoria.pdf

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Corchete, L. (2019). *Expresión génica en mieloma múltiple: análisis de datos de RNA-seq y microarrays en combinación con estudios de metaanálisis y predicción de respuesta al tratamiento* [Universidad de Salamanca]. https://gredos.usal.es/handle/10366/140388

de Sena Brandine, G., & Smith, A. D. (2021). Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Research*, *8*, 1874. https://doi.org/10.12688/f1000research.21142.2

Dobin, A. (2019). *STAR manual 2.7.0a*. https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Faustinelli, P. C. (2012). *Obtención de maní transgénico mediante el silenciamiento de un gen alergénico*. Universidad Nacional de Córdoba.

Gantait, S., Panigrahi, J., Patel, I. C., Labrooy, C., Rathnakumar, A. L., & Yasin, J. K. (2019). *Advances in Plant Breeding Strategies: Nut and Beverage Crops* (J. M. Al-Khayri, S. M. Jain, & D. V. Johnson, Eds.; Vol. 4). Springer International Publishing. https://doi.org/10.1007/978-3-030-23112-5

Gibney, E. R., & Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity*, *105*(1), 4–13. https://doi.org/10.1038/hdy.2010.54

Guillén, C. (2019). *COMPARATIVA DE PIPELINES DE ANÁLISIS COMPUTACIONAL PARA LA DETECCIÓN DE TRANSCRITOS CON USO DIFERENCIAL* . Universidad Autónoma de Madrid .

Hernández, A., Martín Vasallo, P. A., Torres, A., & Salido, E. (1995). Análisis del RNA: Estudio de la expresióngénica. *BIOLOGIA MOLECULAR Y NEFROLOGIA*, *XV*(2), 67–84. https://www.revistanefrologia.com/es-pdf-X0211699595022846

Iqbal, A., Shah, F., Hamayun, M., Ahmad, A., Hussain, A., Waqas, M., Kang, S.-M., & Lee, I.-J. (2016). Allergens of *Arachis hypogaea* and the effect of processing on their detection by ELISA. *Food & Nutrition Research*, *60*(1), 28945. https://doi.org/10.3402/fnr.v60.28945

Jiang, S., Wang, S., Sun, Y., Zhou, Z., & Wang, G. (2011). Molecular characterization of major allergens Ara h 1, 2, 3 in peanut seed. *Plant Cell Reports*, *30*(6), 1135–1143. https://doi.org/10.1007/s00299-011-1022-1

Kang, I. H. (2004). *STUDIES OF THREE MAJOR PEANUT ALLERGENS* [DEGREE OF DOCTOR OF PHILOSOPHY, UNIVERSITY OF FLORIDA]. https://ufdcimages.uflib.ufl.edu/UF/E0/00/43/81/00001/UFE0004381.pdf

Kang, I. H., Srivastava, P., Ozias-Akins, P., & Gallo, M. (2007). Temporal and Spatial Expression of the Major Allergens in Developing and Germinating Peanut Seed. *Plant Physiology*, *144*(2), 836–845. https://doi.org/10.1104/pp.107.096933

Kim, D., & Park, C. (2020). *hisat2* (2.2.1). https://github.com/DaehwanKimLab/hisat2

Korpelainen, E., Tuimala, J., Somervuo, P., Huss, M., & Wong, G. (2015). *RNA-seq Data Analysis* (N. F. Britton, X. Lin, H. M. Safer, M. V. Schneider, M. Singh, & A. Tramontano, Eds.). Chapman and Hall/CRC. https://doi.org/10.1201/b17457

Križanović, K., Echchiki, A., Roux, J., & Šikić, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*, *34*(5), 748–754. https://doi.org/10.1093/bioinformatics/btx668

Kunta, S., Agmon, S., Chedvat, I., Levy, Y., Chu, Y., Ozias-Akins, P., & Hovav, R. (2021). Identification of consistent QTL for time to maturation in Virginia-type Peanut (Arachis hypogaea L.). *BMC Plant Biology*, *21*(1), 186. https://doi.org/10.1186/s12870-021-02951-5

Linnaeus, C. (1753). *Species plantarum.: Vol. II*. Holmiae . https://www.biodiversitylibrary.org/item/13830#page/1/mode/1up

Love, M. I., Anders, S., & Huber, W. (2016). *Differential analysis of count data – the DESeq2 package*. https://cdimage.debian.org/mirror/bioconductor.org/packages/3.3/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

Love, M. I., Anders, S., Kim, V., & Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, *4*, 1070. https://doi.org/10.12688/f1000research.7035.1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Marsh, J. T., Palmer, L. K., Koppelman, S. J., & Johnson, P. E. (2022). Determination of Allergen Levels, Isoforms, and Their Hydroxyproline Modifications Among Peanut Genotypes by Mass Spectrometry. *Frontiers in Allergy*, *3*. https://doi.org/10.3389/falgy.2022.872714

McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2019). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics*, *20*(6), 2044–2054. https://doi.org/10.1093/bib/bby067

Montero, J. (2020). Importancia nutricional y económica del maní (Arachis hypogaea L.). *Revista de Investigación e Innovación Agropecuaria y de Recursos Naturales*, *7*(2).

Natukunda, M. I., Mantilla-Perez, M. B., Graham, M. A., Liu, P., & Salas-Fernandez, M. G. (2022). Dissection of canopy layer-specific genetic control of leaf angle in Sorghum bicolor by RNA sequencing. *BMC Genomics*, *23*(1), 95. https://doi.org/10.1186/s12864-021-08251-4

NCBI. (2018). *Genome assembly arahy.Tifrunner.gnm1.KYV3*. National Library of Medicine . https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003086295.2/

NCBI. (2023, December 9). *Results "Ara h arachis hypogaea."* https://www.ncbi.nlm.nih.gov/gene/?term=Ara+h+arachis++hypogaea

Palladino, C., & Breiteneder, H. (2018). Peanut allergens. *Molecular Immunology*, *100*, 58–70. https://doi.org/10.1016/j.molimm.2018.04.005

Pattee, H. E., Johns, E. B., Singleton, J. A., & Sanders, T. H. (1974). Composition Changes of Peanut Fruit Parts During Maturation1. *Peanut Science*, *1*(2), 57–62. https://doi.org/10.3146/i0095-3679-1-2-6

Pele, M. (2010). Peanut allergens. *Romanian Biotechnological Letters* , *15*(2), 5204–5212. https://web.archive.org/web/20180411213142id_/http://www.rombio.eu/rbl2vol15/19%20Pele.pdf

Peralta, C. T., Aguilera, I. R., Tordecilla, F. R., Guzmán, M. M. A., & Ferrer, P. (2016). Alergia alimentaria a maní: conceptos clínicos, diagnósticos y terapéuticos. *Revista Del Hospital Clínico de La Universidad de Chile*, *26*, 285–292.

Pi, X., Wan, Y., Yang, Y., Li, R., Wu, X., Xie, M., Li, X., & Fu, G. (2019). Research progress in peanut allergens and their allergenicity reduction. *Trends in Food Science & Technology*, *93*, 212–220. https://doi.org/10.1016/j.tifs.2019.09.014

Singh, A., Raina, S. N., Sharma, M., Chaudhary, M., Sharma, S., & Rajpal, V. R. (2021). *Functional uses of peanut (Arachis hypogaea L.) seed storage proteins. Grain and seed proteins functionality* (José Carlos Jiménez López, Ed.; First). IntechOpen.

Tischler, G. (2016). *bamsort - sort BAM files by coordinate or query name*. Ubuntu Manuals .

Toomer, O. T. (2018). Nutritional chemistry of the peanut (*Arachis hypogaea)*. *Critical Reviews in Food Science and Nutrition*, *58*(17), 3042–3053. https://doi.org/10.1080/10408398.2017.1339015

Viquez, O. M., Konan, K. N., & Dodo, H. W. (2003). Structure and organization of the genomic clone of a major peanut allergen gene, Ara h 1. *Molecular Immunology*, *40*(9), 565–571. https://doi.org/10.1016/j.molimm.2003.09.002

Vishwakarma, M. K., Nayak, S. N., Guo, B., Wan, L., Liao, B., Varshney, R. K., & Pandey, M. K. (2017). *The Peanut Genome* (R. K. Varshney, M. K. Pandey, & N. Puppala, Eds.). Springer International Publishing. https://doi.org/10.1007/978-3-319-63935-2

Vlasova-St. Louis, I. (2021). *Applications of RNA-Seq in Biology and Medicine* (I. Vlasova-St. Louis, Ed.). IntechOpen. https://doi.org/10.5772/intechopen.91555

Zhang, W.-J., Cai, Q., Guan, X., & Chen, Q. (2015). Detection of peanut (Arachis hypogaea) allergen by Real-time PCR method with internal amplification control. *Food Chemistry*, *174*, 547–552. https://doi.org/10.1016/j.foodchem.2014.11.091

Zhuang, Y., & Dreskin, S. C. (2013). Redefining the major peanut allergens. *Immunologic Research*, *55*(1–3), 125–134. https://doi.org/10.1007/s12026-012-8355-x

7. Anexos

Anexo 1. *Gráficas de control de calidad de la secuencia SRR24914149, estadísticas básicas, calidad por base. Columna izquierda, FastQC. Columna derecha, Falco.*

<table>
<tr><th>FastQC</th><th>FALCO</th></tr>
<tr><td>

Basic Statistics

Measure	Value
Filename	SRR24914149.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	14899080
Sequences flagged as poor quality	0
Sequence length	101
%GC	43

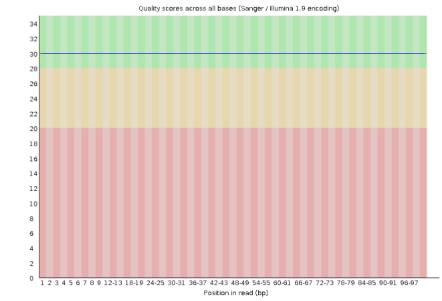
</td><td>

Basic Statistics: pass

Measure	Value
Filename	SRR24914141.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16668038
Sequences Flagged As Poor Quality	0
Sequence length	101
%GC:	46

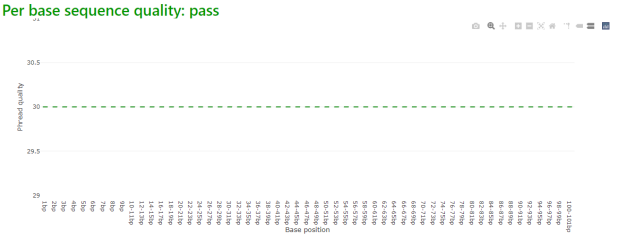
</td></tr>
<tr><td>

Per base sequence quality



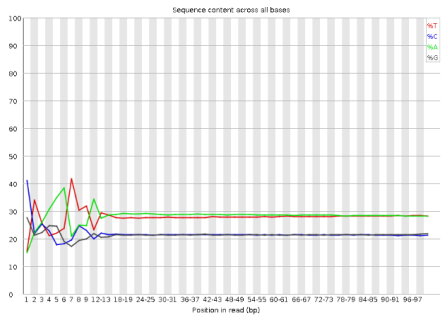
</td><td>

Per base sequence quality: pass



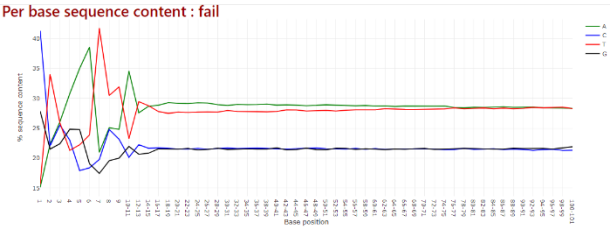
</td></tr>
<tr><td>

Per base sequence content



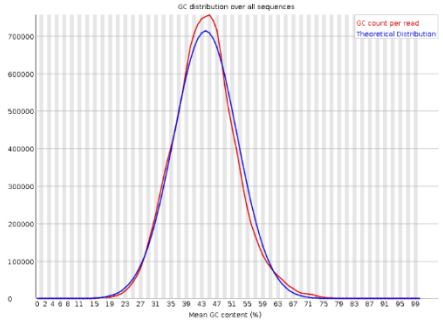
</td><td>

Per base sequence content : fail



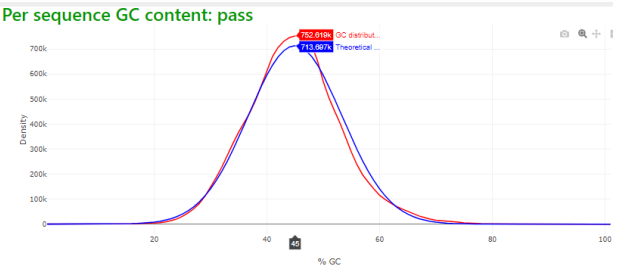
</td></tr>
<tr><td>

Per sequence GC content



</td><td>

Per sequence GC content: pass



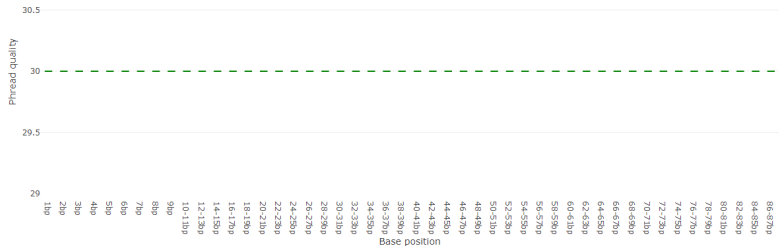
</td></tr>
</table>

Anexo 2. *Gráficas de control de calidad de la herramienta Falco, de la secuencia SRR24914149 después del preprocesamiento, estadísticas básicas, calidad por base.*

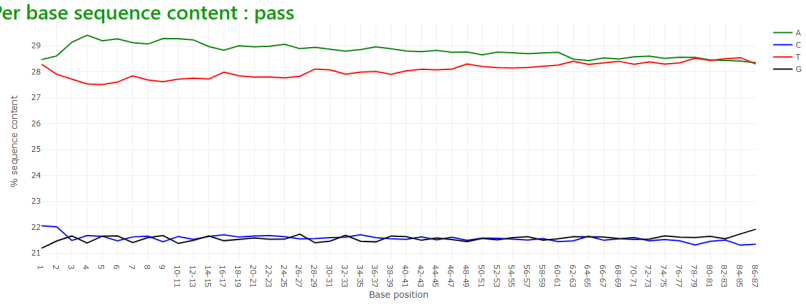
Basic Statistics: pass

Measure	Value
Filename	SRR24914149_trim.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	14899080
Sequences Flagged As Poor Quality	0
Sequence length	87
%GC:	43

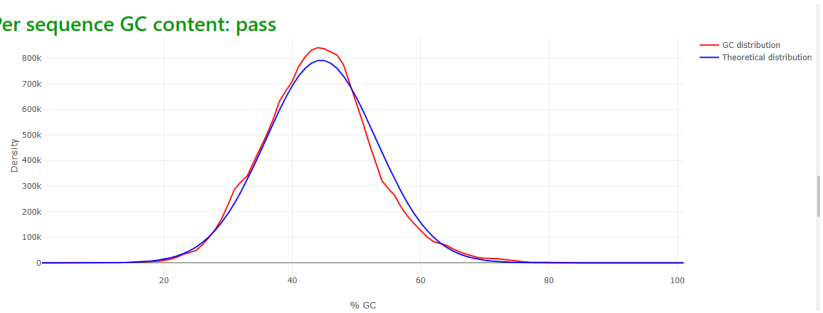
Per base sequence quality: pass



Per base sequence content : pass

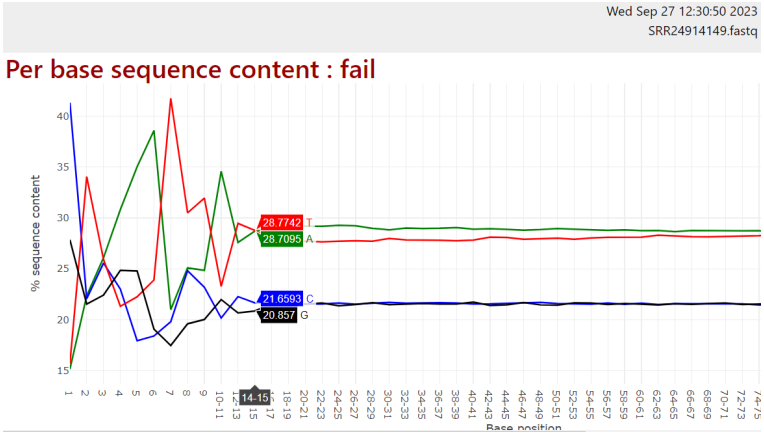


Per sequence GC content: pass

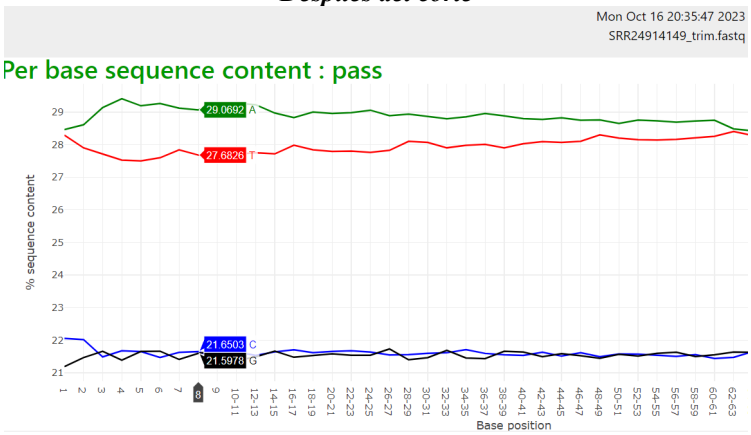


Anexo 3. *Reporte de Falco para el Contenido de secuencia por base, de la lectura de una semilla en Desarrollo-r1 (SRR24914149). Arriba: Antes del preprocesamiento. Abajo: después del corte de las 14 primeras bases.*

Lectura cruda



Después del corte



Anexo 4. *Reporte de MultiQC (v1.12), visualización de los archivos de salida de STAR vs HISAT2, de la lectura de una semilla Madura-r1 (SRR24914143).*

STAR

HISAT2

Anexo 5. *Resumen de los resultados entre las diferentes comparaciones*

Desarrollo vs Completa

```
> resultsNames(dds1)
[1] "Intercept"                          "condicion_completa_vs_desarrollo"
> summary(res05)

out of 40892 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 4558, 11%
LFC < 0 (down)     : 5403, 13%
outliers [1]       : 32, 0.078%
low counts [2]     : 0, 0%
(mean count < 2)

> res05
log2 fold change (MLE): condicion completa vs desarrollo
Wald test p-value: condicion completa vs desarrollo
DataFrame with 40892 rows and 6 columns
                baseMean log2FoldChange     lfcSE      stat     pvalue      padj
               <numeric>      <numeric> <numeric> <numeric>  <numeric> <numeric>
DDI71_pgp066     4.29387      1.8604929  0.957932  1.942198 0.05211319 0.1492810
LOC112695120   491.09466      0.5495468  0.196997  2.789624 0.00527693 0.0261828
LOC112695121   154.93265      0.6336154  0.240901  2.630188 0.00853377 0.0380791
LOC112695122    71.00662     -0.0659883  0.293132 -0.225114 0.82189037 0.9059065
LOC112695123    10.91711     -0.1304045  0.588176 -0.221710 0.82453973 0.9074701
...                   ...            ...       ...       ...        ...       ...
LOC114927817 5.22755e+00      0.1120815  0.838741  0.133631 0.89369462 0.9450492
LOC114927830 3.15388e+01      0.9448759  0.449788  2.100714 0.03566609 0.1128565
TRNAF-GAA_33 2.66159e+00     -4.5332438  1.575884 -2.876636 0.00401939 0.0211503
__no_feature 5.02695e+05      0.0144273  0.166578  0.086610 0.93098155 0.9649655
__ambiguous  7.52583e+04      0.4084952  0.163274  2.501905 0.01235271 0.0504782
```

Desarrollo vs Madura

```
> resultsNames(dds1)
[1] "Intercept"                          "condicion_madura_vs_desarrollo"
> summary(res05)

out of 40953 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 6142, 15%
LFC < 0 (down)     : 7863, 19%
outliers [1]       : 31, 0.076%
low counts [2]     : 0, 0%
(mean count < 1)

> res05
log2 fold change (MLE): condicion madura vs desarrollo
Wald test p-value: condicion madura vs desarrollo
DataFrame with 40953 rows and 6 columns
                baseMean log2FoldChange     lfcSE      stat     pvalue      padj
               <numeric>      <numeric> <numeric> <numeric>  <numeric> <numeric>
LOC112695120   442.1819      0.4044001  0.181581  2.227105  0.0259403 0.0701558
LOC112695121   146.5302      0.6111190  0.252221  2.422946  0.0153952 0.0459085
LOC112695122    72.8496      0.1423863  0.281549  0.505725  0.6130496 0.7351134
LOC112695123    14.3717      0.7214200  0.512510  1.407622  0.1592431 0.2847643
LOC112695125   179.8594     -0.0307891  0.224685 -0.137032  0.8910054 0.9350358
...                  ...            ...       ...       ...        ...       ...
LOC114927817 4.48073e+00     -0.240425  1.008477 -0.238404 0.81156797 0.8817934
LOC114927830 2.64061e+01      0.639672  0.530563  1.205648 0.22795323 0.3696866
TRNAF-GAA_33 2.73856e+00     -3.446966  1.522477 -2.264052 0.02357091 0.0648275
__no_feature 5.04386e+05      0.155423  0.162318  0.957524 0.33830300 0.4894999
__ambiguous  7.65411e+04      0.566181  0.187437  3.020650 0.00252233 0.0103612
```

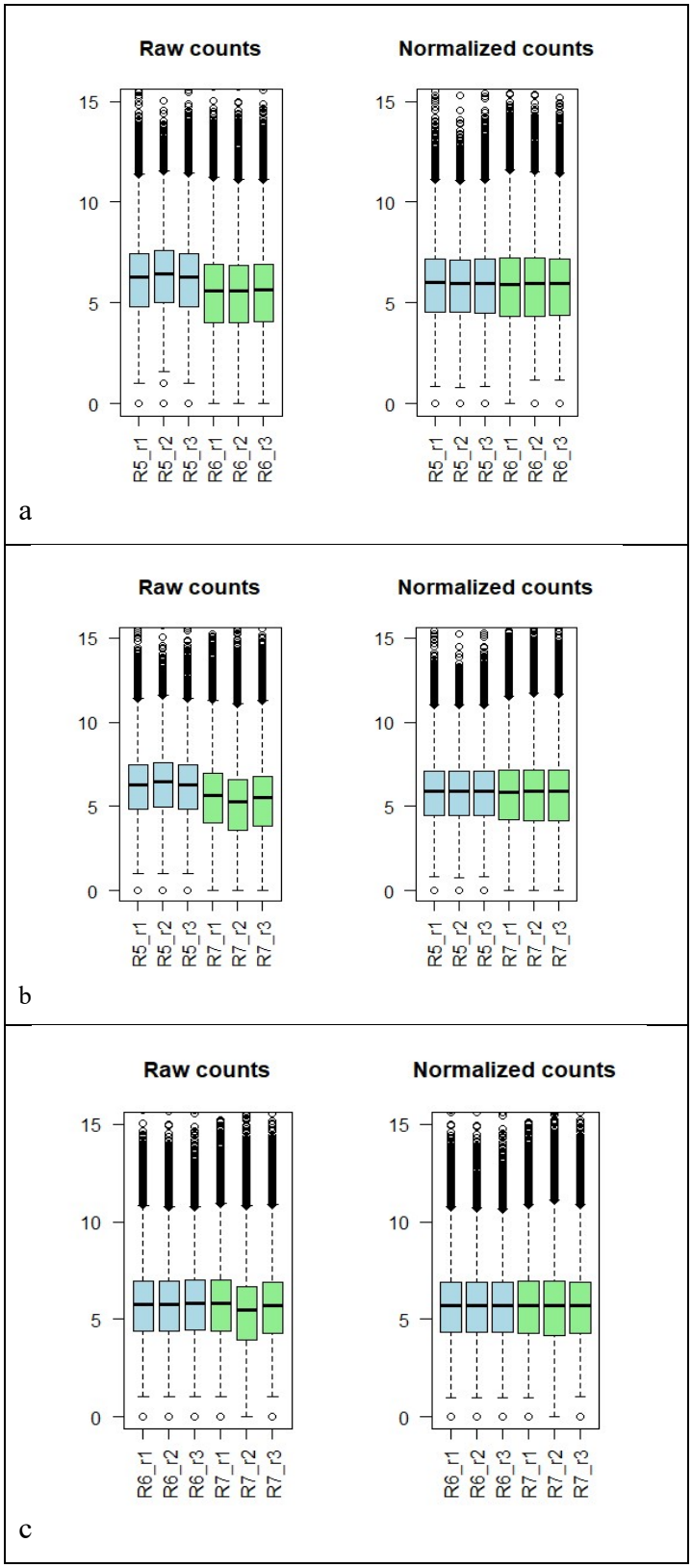
Completa vs Madura

```
> resultsNames(dds1)
[1] "Intercept"                    "condicion_madura_vs_completa"
> summary(res05)

out of 38358 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1717, 4.5%
LFC < 0 (down)     : 2144, 5.6%
outliers [1]       : 3, 0.0078%
low counts [2]     : 744, 1.9%
(mean count < 4)
```

```
> res05
log2 fold change (MLE): condicion madura vs completa
Wald test p-value: condicion madura vs completa
DataFrame with 38358 rows and 6 columns
                baseMean log2FoldChange     lfcSE        stat    pvalue      padj
               <numeric>      <numeric> <numeric>   <numeric> <numeric> <numeric>
DDI71_pgp066     2.84284    -2.87795539  1.374843 -2.09329762 0.0363226        NA
LOC112695120   420.30591    -0.12034507  0.184799 -0.65122043 0.5149042  0.804440
LOC112695121   141.47244     0.00126972  0.228739  0.00555094 0.9955710  0.998421
LOC112695122    56.55202     0.23327478  0.290487  0.80304750 0.4219473  0.744943
LOC112695123    11.06472     0.87748245  0.592516  1.48094270 0.1386218  0.440235
...                   ...            ...       ...         ...       ...       ...
LOC114927817 3.68015e+00    -0.33051259  1.051506 -0.31432317  0.753276        NA
LOC114927830 2.84344e+01    -0.28734581  0.430413 -0.66760476  0.504386  0.799167
TRNAN-GUU_23 2.80459e+00    -0.00304325  1.197599 -0.00254113  0.997972        NA
__no_feature 4.02002e+05     0.16567589  0.144168  1.14918407  0.250480  0.594936
__ambiguous  6.87162e+04     0.18181090  0.156898  1.15878078  0.246546  0.589987
```

Anexo 6. *Gráfico de cajas comparando el conteo de lecturas en crudo versus el conteo normalizado, para cada comparación de estadios de semilla.*



Nota: a es la comparación DC, b es la comparación DM y c es la comparación CM

Anexo 7. *Tablas de resumen de DESeq2*

7a. Tabla de resumen de DESeq2 para la comparación DC

,"baseMean","log2FoldChange","lfcSE","stat","pvalue","padj"
LOC112707245,139769.01842623,4.33382296756471,0.461127218539525,9.39832391870237,5.54391244833415e-21,1.09432010936683e-18
LOC112771110,109808.775583533,4.71851658805541,0.463973589237634,10.1697956467921,2.70475524658736e-24,7.36775329170397e-22
LOC112776552,325334.220938645,4.11631101980907,0.721548265816504,5.70483114549663,1.16458478003743e-08,3.37617091600579e-07
LOC112711772,263754.640247734,3.82557349872568,0.706421253143552,5.41542809152755,6.11422411318895e-08,1.51778370148785e-06
LOC112695262,243246.890758229,3.79412360862387,0.747558692146447,5.07535214088662,3.8677902395091e-07,7.8821899843562e-06
LOC112758686,53.3581460825665,3.6896716847682,0.525190568692003,7.02539593191362,2.13460063898985e-12,1.32843092831858e-10

7b. Tabla de resumen de DESeq2 para la comparación DM

,"baseMean","log2FoldChange","lfcSE","stat","pvalue","padj"
LOC112707245,235944.394281599,5.19103423355808,0.519267889544812,9.99683272945785,1.57349237830588e-23,2.1901515341848e-21
LOC112771110,181817.052667955,5.54017681269818,0.524170713882287,10.5694131052547,4.1307151402243e-26,7.13236814212063e-24
LOC112776552,445487.090639984,4.66653630622082,0.627364873523307,7.43831301872762,1.01979188847734e-13,4.31560741057596e-12
LOC112711772,361170.813137489,4.38117225849753,0.588345313508769,7.4466000797544,9.57763874008791e-14,4.06572751578711e-12
LOC112695262,213700.024235411,3.66873097144204,0.515658342228399,7.11465455128246,1.12193485311951e-12,3.98540087320805e-11
LOC112758686,41.8548435449479,3.37875925529391,0.613251878344126,5.50957832272293,3.59694321668836e-08,5.60312563050327e-07

7c. Tabla de resumen de DESeq2 para la comparación CM

,"baseMean","log2FoldChange","lfcSE","stat","pvalue","padj"
LOC112707245,283395.138215587,0.878095812128318,0.185450436110962,4.734935277277 65, 2.19125014526668e-06,0.000101873211299035
LOC112771110,221593.65775668,0.842397897334152,0.169321526341582,4.97513763037273,6.52012978083804e-07,3.66685597865166e-05
LOC112776552,573046.251044099,0.569160801137482,0.193596774090557,2.9399291584851 1,0.003282872948509,0.0362657234850411
LOC112711772,460151.684232992,0.575041755719037,0.174902139639532,3.2877914295627 3,0.00100976617693643,0.0146315004595379
LOC112695262,328602.684641453,-0.104578787775258,0.180484742603481,-0.579432844387373,0.562297147685034,0.833117282785161
LOC112758686,67.6251961684893,-0.292871791212343,0.296519653660352,-0.987697738065663,0.323300702620133,0.670728842323306