UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE CIENCIAS

“ANÁLISIS FUNCIONAL DE TRANSCRIPTOMAS DE *Coffea arabica* L. RELACIONADOS AL ESTRÉS TÉRMICO”

Presentada por:

PIERO ANTONIO PALACIOS BERNUY

Tesis para Optar por el Título Profesional de:

BIÓLOGO

LIMA-PERÚ

2024

ANÁLISIS FUNCIONAL DE TRANSCRIPTOMAS DE Coffea arabica L. RELACIONADOS AL ESTRÉS TÉRMICO

INFORME DE ORIGINALIDAD

8%	8%	2%	1%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.lamolina.edu.pe Fuente de Internet	2%
2	hdl.handle.net Fuente de Internet	1%
3	www.kerwa.ucr.ac.cr Fuente de Internet	<1%
4	www.researchgate.net Fuente de Internet	<1%
5	patents.google.com Fuente de Internet	<1%
6	eprints.ucm.es Fuente de Internet	<1%
7	www.scielo.org.pe Fuente de Internet	<1%
8	repository.javeriana.edu.co Fuente de Internet	<1%
9	repositorio.cinvestav.mx Fuente de Internet	

UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE CIENCIAS

"ANÁLISIS FUNCIONAL DE TRANSCRIPTOMAS DE *Coffea arabica* L. RELACIONADOS AL ESTRÉS TÉRMICO"

Presentada por:

PIERO ANTONIO PALACIOS BERNUY

Tesis para Optar por el Título Profesional de:

BIÓLOGO

Sustentada y aprobada por el siguiente jurado:

Mg. Sc. César Fernando López Bonilla
PRESIDENTE

Mg. Sc. María del Rosario Josefina Castro Muñoz
MIEMBRO

Mg. Sc. Katty Ogata Gutiérrez
MIEMBRO

Ph.D. Roberto Carlos Mansilla Samaniego
ASESOR

Mg. Sc. Cinthia Sheila Quispe Apaza
CO-ASESORA

DEDICATORIA

Dedico esta tesis a mis padres, mis mascotas,

a mi hermana que sin su apoyo no hubiese sido posible realizarla.

En memoria de mi abuelo y parientes que se

encuentran descansando en paz.

AGRADECIMIENTOS

Al Doctor Roberto Mansilla asesor de la tesis en la Universidad Nacional Agraria la Molina, mi inmensa gratitud por la guía, consejos, y el compromiso en la asesoría de la tesis.

A la Universidad Nacional Agraria La Molina y a los profesores de la carrera que me formaron como profesional.

Al Ingeniero Juan Ulloa por su apoyo en el manejo del supercomputador de la Universidad Nacional Agraria La Molina.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El café, uno de los *commodities* más importantes del mundo, está en peligro debido al cambio climático. Específicamente, porque el aumento de la temperatura del aire afecta el rendimiento del cultivo y las cualidades organolépticas del grano. De ahí que es de suma importancia conocer sobre el mecanismo de respuesta de la planta a las elevadas temperaturas y los efectos adversos que estos tendrían sobre el cultivo del café. De acuerdo con esto, la presente tesis tuvo como objetivo analizar los experimentos de transcriptómica (públicos) previamente realizados sobre el efecto de la elevación de la temperatura del aire sobre la planta de café (*Coffea arabica*) para entender mejor la respuesta de esta especie al estrés térmico, por lo que se realizó el análisis funcional de los genes. Para lo cual, primeramente, se mapearon los *reads* de experimentos RNA-Seq frente a los genomas de las especies *Coffea arabica, Coffea canephora y Coffea eugenioies* para comparar las estadísticas de mapeo y determinar la relación genómica con estas especies. Luego, mediante el análisis de la expresión diferencial de genes se determinó que las temperaturas mayores a 37°C activan la respuesta al estrés térmico del café. Seguidamente, mediante el análisis funcional de genes se determinó, que las ontologías relacionadas a la respuesta al calor estaban sobre representadas, por lo tanto, el café presenta una respuesta concertada al estrés térmico. Finalmente, en el análisis dentro de estas ontologías se identificaron a genes que codifican proteínas sHSP (*small heat shock protein*), ribonucleoproteínas para la formación del espliceosoma, proteínas relacionadas a la regulación de auxinas y enzimas que sintetizan terpenoides.

Palabras clave: *reads*, ontología, genes, transcriptómica, elevadas temperaturas, *Coffea arabica*.

ABSTRACT

Coffee, one of the most important commodities in the world, is in danger due to climate change. Specifically, increasing air temperature affects the crop yield and the grain organoleptic quality. Because of this, it is extremely important to know about the molecular mechanism of plant responses to high temperatures and their adverse effects on the coffee crop. According to this, the goal of the present thesis was analyze previously performed transcriptomic experiments related to the effect of high temperature on the coffee plant (*Coffea arabica*) (public) ; in order to have better understanding about the response of this species to thermal stress the gene functional analysis was performed. Firstly, the RNA-Seq experiments reads were mapped against the *Coffea arabica* genomes, *Coffea canephora* and *Coffea eugenioies* to compare mapping statistics and determine the genomic relationship with these species. Then, by analyzing the differential expression of genes it was determined that temperatures greater than 37°C activate the response to thermal stress of coffee. Then, through the functional analysis of genes, it was determined, that the ontologies related to heat response are over-represented, and the coffee plant has a concerted response to thermal stress. Finally, within these ontologies were identified genes that encode sHSP proteins (small heat shock protein), ribonucleoproteins related to the spliceosome formation, proteins related to the regulation of auxins and enzymes that synthesize terpenoids.

Keywords: reads, ontology, genes, transcriptomics, high temperatures, *Coffea arabica*.

I. INTRODUCCIÓN

El cultivo del café es una de las principales fuentes de ingresos económicos en muchos países en vías de desarrollo, debido a que provee una de las bebidas más consumidas en el mundo. Estos ingresos se han incrementado considerablemente desde 1960 hasta la fecha debido a la creciente demanda de países consumidores (importadores) como Francia, Estados Unidos de Norteamérica y Canadá (FAOSTAT, 2021; Observatory of Economic Complexity, 2020). Lo que ha motivado que la producción del grano de café (verde) haya pasado de 7 millones a 9 millones de toneladas a nivel mundial en el periodo 2000-2021 (FAOSTAT, 2022), Actualmente, la demanda del grano de café sigue siendo alta ya que, sólo para el mes de junio del año 2022, se exportaron 11.11 millones de sacos de café (comparados con 10.97 millones de sacos de grano de café de junio del 2021) a un precio de 4.5 dólares por kilogramo en promedio; además, para junio del 2022 ya se han exportado 98.77 millones de sacos lo que equivale al 62% de todo lo exportado en el 2018 (International Coffee Organization, 2022). Por otro lado, cabe mencionar que, de los 195 países reconocidos por la ONU, sólo 70 son exportadores de café; y el 50 % de la producción mundial se concentra en solo tres países (Brasil, Vietnam y Colombia) de los 70.

Al mismo tiempo, la mayoría de los productores son pequeños agricultores (~25 millones de productores), y estos producen la mayoría del café en el mundo (~80 %). Sin embargo, los beneficios de la industria del café no se han reflejado en sus canastas familiares debido a la falta de asociación entre ellos, lo que a su vez ocasiona entre otras dificultades como por ejemplo el acceso a servicios bancarios para la agricultura. Adicionalmente, estos pequeños productores enfrentan nuevos desafíos

como el cambio climático y condiciones naturales cada vez más difíciles para el cultivo de esta planta (Amrouk, 2018; FAO, 2015; MINAGRI, 2018).

En el Perú, el grano de café es el principal producto agrícola de exportación y su producción se distribuye en espacios territorialmente diversos, específicamente en

doce regiones del país. Una de las características de Perú, como país productor de café, es que el 85% de los caficultores son pequeños productores, y el rendimiento promedio de producción a nivel nacional es de 72 kg/ha (MINAGRI, 2018).

Respecto al cambio climático, se ha reportado que desde 1850 hasta 2018 el incremento de la temperatura media superficial del planeta fue de ~0.87°C y el de la temperatura promedio del aire superficial de ~1.53°C. Estos cambios de temperatura, que también ocasionan cambios en la precipitación mundial, han alterado el inicio y fin de las estaciones (alterando la fenología de las plantas); además, han ocasionado la disminución de los rendimientos de cultivos, la reducción de la disponibilidad de agua fresca, han puesto a la biodiversidad bajo mayor estrés (descenso de la biodiversidad 11-14%) y han incrementado la mortalidad de los árboles debido a que estos no pueden adaptarse rápido a los cambios climáticos (Shukla et al., 2019).

Entre las consecuencias del cambio climático descritas en el párrafo anterior, el incremento de la temperatura del aire superficial es el que más afecta el rendimiento del café y a la calidad de su grano. Esto se debe a que el crecimiento y fructificación del café presentan un rango óptimo de temperatura que se encuentra entre 18 y 25°C, por lo que, al incrementarse la temperatura del aire superficial, sumado al estrés térmico del aire afecta las cualidades organolépticas del grano. Por lo tanto, ocasionan que se perjudique la exportación debido a que no cumplen con los estándares mínimos de calidad que exigen los paises importadores (Camargo 2010).

Debido a la posición que el café ocupa en el mundo como uno de los principales commodities agrícolas y, el proceso actual de cambio climático que venimos enfrentando; es necesario realizar investigaciones sobre los efectos del aumento de la temperatura del aire sobre el cultivo de café. Específicamente, se debe ahondar sobre el efecto del incremento de la temperatura del aire sobre la expresión génica y

cómo esta expresión modula mecanismos moleculares que le permiten a esta especie responder a este tipo de estrés.

En el presente estudio se evaluaron datos de transcriptómica de la base de datos *Sequence Read Archive* (SRA) del *National Center for Biotechnology Information* (NCBI) producidos a partir de experimentos en donde plantas de *Coffea arabica* han sido sometidas a estrés térmico (elevación de la temperatura del aire).

Este estudio tuvo como objetivo la identificación de grupos funcionales de genes que se activaron en respuesta a este tipo de estrés. Esto permitirá identificar procesos moleculares que se activan cuando *Coffea arabica* es sometida a elevadas temperaturas del aire y, ayudará a entender la respuesta concertada de genes a este tipo de estrés.

1.1. Objetivos

1.1.1. Objetivos Generales

Analizar el efecto de la elevación de la temperatura del aire en plantas de café (*Coffea arabica*) utilizando datos producidos por RNA-Seq obtenidos de la base de datos del NCBI.

1.1.2. Objetivos Específicos

- Revisar la base de datos de secuencias de RNA del NCBI (*National Center of Biotechnology Information*) para buscar datos crudos de experimentos sobre estrés térmico realizados a partir de las hojas de genotipos de *Coffea arabica*.
- Realizar el control de calidad y preprocesamiento de archivos FASTQ.
- Alinear, mapear y cuantificar *reads* de RNA-Seq sobre los genomas y anotaciones de *Coffea arabica*, *Coffea eugenioides* y *Coffea canephora* para comparar el porcentaje de mapeo.
- Realizar el análisis de expresión diferencial de genes (DEG) con los datos de mapeo de *Coffea arabica.*
- Realizar el análisis funcional de genes.
- Analizar la información de cuantificación, expresión diferencial y enriquecimiento funcional.

II. REVISIÓN DE LITERATURA

2.1. El Café

El café es uno de los *commodities* más importantes y demandados en el mundo cuya exportación refleja una gran suma de ingresos para varios países (Aga, 2005). En lo que se refiere a su exportación, según la Organización Internacional del Café (*International Coffee Organization*) y la Organización de las naciones Unidas para la alimentación y agricultura (*Food and Agriculture Organization of the United Nations*: FAO) para junio del año 2022 se exportaron 98.77 millones de costales de café a nivel mundial a un precio de 4.5 dólares por kilogramo en promedio. Además, en el año 2020, el Observatorio de la Complejidad Económica (*Observatory of Economic Complexity*) menciona que se exportó un total de 30.8 mil millones de dólares en café; siendo los principales países exportadores Brasil, Suiza, Alemania, Colombia y Vietnam con valores de exportación (en mil millones) de 5.08, 2.71, 2.59, 2.54 y 2.24 respectivamente. Estas características del café (gran demanda y exportación) hacen que su valor en el mercado internacional sea muy susceptible a factores de la economía mundial, como el incremento (o disminución) de la globalización, el movimiento de divisas y los precios internacionales. Estos posibles cambios o variaciones en la economía mundial afectan la cadena de valor de los *commodities* en el sector agrícola, por ejemplo, la producción del café (FAO, 2015).

En los últimos 50 años, se ha incrementado la producción de café y su consumo debido a la gran variedad de productos derivados del mismo. Desde el año 2008 el valor de la producción de café en el mundo ha crecido alrededor de 3.5 por ciento por año, llegando a 158.9 millones de sacos en la campaña de 2017-2018 (FAO,

2015; MINAGRI, 2018; FAO, 2021). En la actualidad, de los 195 países reconocidos por la ONU, sólo 70 son productores y exportadores. Sin embargo, la capacidad de producción de estos países varía de tal manera que el 50% de la producción mundial se concentra principalmente en tres países: Brasil, Vietnam y Colombia. Es así que, para enero del año 2022, Brasil exportó 3.2 millones de bolsas de 60 kg de café entre

las variedades arábicas y robusta, Vietnam exportó 2.8 millones de bolsas de 60 kg y Colombia exportó 1.04 millones de bolsas de 60 kg (*International Coffee Organization*, 2022). En contraste a esto, la mayoría de los productores en el mundo son pequeños productores (~25 millones de productores); y, contrariamente al sentido común, muchos de los pequeños productores (los cuales producen alrededor del 80% del café en el planeta) no reciben los beneficios de las grandes cantidades de exportación. Además, actualmente se enfrentan desafíos como el cambio climático y condiciones naturales más difíciles para el crecimiento de la planta de café (FAO, 2015; Amrouk, 2018).

Por otro lado, la producción comercial se basa principalmente en dos especies: *Coffea arabica* y *Coffea canephora*. Alrededor del 60% del café en el mundo es de las variedades arábicas (*Coffea arabica*), mientras que el resto (40%) pertenece a la variedad robusta (*Coffea canephora*). A las variedades arábicas se las considera de mejor calidad y por ende tiene mayor precio; mientras que, robusta es más dura y de sabor más amargo y es de menor precio en el mercado, por lo que es preferida en países como China e Indonesia (*International Coffee Organization*, 2022).

En Perú, donde se cultiva casi exclusivamente *C. arabica*, el café ha sido por mucho tiempo el principal producto agrícola de exportación. El 62% de las exportaciones están concentradas hacia Estados Unidos (27%), Alemania (25%) y Bélgica (10%) llegando a un valor de 667 millones de dólares en exportación en el 2020 (MINAGRI, 2018; *International Coffee Organization, 2020*). Respecto a la producción, una de las características principales del país es que el 85% de los caficultores son pequeños productores y producen entre una a cinco hectáreas con un rendimiento promedio de 72 kg/ha (MINAGRI, 2018). En el 2021 según el INEI (Instituto Nacional de Estadística e Informática), Perú produjo 69 mil 139 toneladas de grano de café. De toda esta producción, las principales regiones productoras fueron Ucayali, San Martín, Cusco, Cajamarca y Amazonas que en su conjunto aportaron el 71.1% a nivel nacional (INEI, 2021).

A favor de los productores, el café en Perú se desarrolla de manera óptima en la vertiente oriental de la Cordillera de los Andes, desde los 600 hasta los 2000 m.s.n.m; lo que ha permitido ser cultivado en 12 de las 25 regiones del país, siendo propio de la cordillera de los andes y de la selva peruana, permitiendo presentar distintos

sabores, aromas y acidez. Esto hace que se puedan cultivar en el Perú las variedades comerciales de *Coffea arabica* como Típica, Caturra y Bourbon (Mansilla, 2018).

A pesar de los beneficios climáticos que ofrece el país, los productores se enfrentan al gran problema de no estar asociados a una institución gestora (~70% de productores pequeños), lo que ocasiona la poca capacidad coordinada de respuesta ante desafíos como el cambio climático, pestes como la roya amarilla del café y a competidores internacionales (MINAGRI, 2018).

2.2. Clasificación Taxonómica

El género *Coffea* L. comprende 103 especies que se encuentran naturalmente en África tropical, Madagascar, La Unión de las Comoras y en las islas Mascareñas (Davis y Rakotonasolo, 2008). *Coffea* pertenece junto a otros 10 géneros a la tribu Coffeae. *Coffea* se divide en dos subgéneros, el primer subgénero *Coffea* contiene a 95 especies dentro de los cuales se puede encontrar a las especies *Coffea arabica* L., *Coffea canephora* A.Froehner y *Coffea liberica* Hiern. El segundo subgénero *Baracoffea* contiene a nueve especies (Davis y Rakotonasolo, 2008).

La clasificación taxonómica según Leroy (1980) y Davis et al. (2007) citados por Mansilla (2021) es:

División: Angiospermae

Clase: Eudycotiledoneae

Subclase: Asteridae

Orden: Gentianales

Familia: Rubiaceae citado por Davis et al. (2007)

Subfamilia: Ixoroidae

Tribu: Coffeeae

Género: *Coffea* citado por Leroy (1980)

Especie: *Coffea arabica* L.

2.3. Morfología de *Coffea arabica*

Coffea arabica es un árbol pequeño perenne cuyo ciclo de vida dura entre 20-25 años, y alcanza su máxima producción entre los 6 y 8 años (Arcila et al., 2007). Puede llegar a medir entre 4 a 5 metros de altura. Posee crecimiento de las ramas de tipo dimórfico, las ramas verticales (ortotróficas) forman ramas horizontales (plagiotróficas) "paralelas" al suelo (Aga, 2005).

Las hojas tienen forma elíptica con disposición opuesta. Crecen desde la parte terminal de las ramas verticales y desde los nudos de las ramas horizontales. Las inflorescencias son de tipo cima con eje corto y se desarrollan de yemas ubicadas en ramas horizontales. Cada inflorescencia lleva de una a cinco flores. Las flores tienen un pedicelo corto y un cáliz rudimentario. Los pétalos están fusionados y forman una corola pentaloba (Aga, 2005).

El androceo, consta de cinco estambres insertados en el tubo de la corola y sobrepasan la altura de los pétalos. El gineceo está compuesto de un ovario súpero bilocular que contiene dos óvulos anátropos, un pistilo que consiste en un ovario inferior y un estilo largo con dos lóbulos estigmáticos y un estilo fino y largo. Los granos de polen son pesados y pegajosos (Aga, 2005; Mansilla, 2021). El sistema de reproducción de *Coffea arabica* se basa en la auto-fertilización; sin embargo, a diferencia de otros cultivos tropicales, *Coffea arabica* no se propaga clonalmente, sino por semilla (Scalabrin et al., 2020).

El fruto de café es una drupa ovalada ligeramente aplanada que usualmente contiene dos semillas de entre 10 a 17 mm. El fruto maduro tiene un exocarpo que puede tener coloraciones amarillo pálido, rojo, rojo intenso o violeta. El mesocarpo es carnoso rico en azúcares y el endocarpo es duro. Cada semilla está envuelta de una capa plateada (testa) la cual es el remanente del perispermo. El grano de café consiste en un endospermo y un embrión en la parte basal de la semilla. Esta semilla es de tipo recalcitrante con una viabilidad entre 3 a 6 meses (Aga, 2005; Mansilla, 2021).

La fase de desarrollo vegetativo del café ocurre desde la germinación hasta la primera floración y puede durar 19 meses. En adelante las fases de crecimiento vegetativo y reproductivo ocurren al mismo tiempo por el resto de la vida de la planta. Desde el

momento de la floración hasta la maduración del fruto pasan alrededor de 8 meses (Arcila et al., 2007).

2.4. Origen del Café

El centro de origen de *Coffea arabica* L. se encuentra en las tierras altas (1500-2800 m.s.n.m.) del suroeste de Etiopía y sureste de Sudán entre las latitudes de 4°N y 9°N, con un patrón de distribución típico de los poliploides: expansión periférica rodeando la distribución de las especies diploides (Aga, 2005; Lashermes et al., 1996, Anthony et al., 2002 y Tesfaye et al., 2007, como se citó en Labouisse et al., 2008; Camargo, 2010; Scalabrin et al., 2020). Entre las características climáticas de esa región africana están: promedio de temperatura anual que varía en un rango de 18-22°C y una media de precipitación anual que varía entre 1600-2000 mm, con un periodo de sequía que dura cuatro meses durante los meses más fríos (Camargo, 2010).

Coffea arabica es la única especie tetraploide del género *Coffea*. Se cree que la hibridación natural de dos especies de café ancestrales, *Coffea eugenioides* y *Coffea canephora* dieron origen a *Coffea arabica* hace aproximadamente 665000 años (Scalabrin, 2020). Esta hipótesis es apoyada por experimentos realizados en meiosis, donde se observó que los cromosomas de especies actuales de *C. arabica* y *C. canephora* tienen un mejor emparejamiento de lo que se esperaba (Aga, 2005; Scalabrin et al., 2020).

Respecto a la historia de *Coffea arabica*, se indica que antiguamente Yemén fue el país que servía como la única fuente de germoplasma de café; por lo que Lineo le proporcionó el nombre de arábica a esta especie debido a que en su época lo que poco que se sabía de la planta era que la infusión provenía de la península Arábica. Sin embargo, esto no sugiere que Yemén sea su centro de origen (Aga, 2005). Posteriormente, se descubrió que cerca al siglo XIV, las semillas de café fueron llevadas desde el suroeste de Etiopía hacia Yemén, donde el cultivo de café se extendió en dirección a Moccha y El Cairo al final del siglo XV (Scalabrin et al., 2020).

En 1670, algunas semillas fueron llevadas desde Yemén hacia la India por contrabando y en 1715, desde Yemén hacia la Isla Borbón. Luego, desde la India los holandeses llevaron semillas hacia Indonesia entre los años 1696 y 1699, a partir de

las que se originaron la variedad Típica de *Coffea arabica*. Esta variedad llegó al continente sudamericano en 1723. Por otro lado, las semillas que llegaron a Borbón desde Yemén originaron a la variedad *Bourbon*, la cual llegó al este de África desde América a mediados del siglo XIX. A principios del siglo XX, en el Este de África comenzó el cultivo de café con las variedades introducidas Típica, *Bourbon* y de la India, así como, con algunas variedades locales de Etiopía (Scalabrin et al., 2020).

2.5. Diversidad Genética del Café

La alta susceptibilidad de las variedades comerciales del café a pestes, enfermedades y temperaturas altas por el calentamiento global puede ser atribuida a una base genética angosta. Tanto café como plátano, soya, maíz y cacao tienen un potencial limitado de adaptación natural a los cambios ambientales debido a su baja diversidad genética (Aerts et al., 2017).

Coffea arabica tiene uno de los niveles más bajos de diversidad entre las especies comerciales, y ese nivel de diversidad (diversidad nucleotídica: 2.3×10^{-4}) sólo es comparable con la del trigo harinero (diversidad nucleotídica: 5.7×10^{-4}) (Scalabrin et al., 2020). Esta baja diversidad genética de *Coffea arabica* se debe principalmente a cuellos de botella originados durante la diseminación global de la especie y a una predominante autogamia que conlleva a altos niveles de endogamia (Aerts et al., 2017; Scalabrin et al., 2020). Además, *Coffea arabica* proviene de un evento reciente de alopoliploidización que explica también la presencia de pocos rearreglos cromosómicos, por ejemplo: en la punta del cromosoma 7 (correspondiente al subgenoma de *Coffea canephora*) se observó que hay una región correspondiente al subgenoma de *Coffea eugenioides* lo que se traduce en que *Coffea arabica* lleva 4 copias de esta región en su genoma (Scalabrin et al., 2020).

Entre las variedades de *Coffea arabica* se observaron dos grupos bien marcados, uno que consiste en accesiones silvestres de Etiopía y otro proveniente del café cultivado proveniente de Yemén (Lashermes et al., 1996; Silvestrini et al., 2008). La diversidad genética conservada en las variedades autóctonas del sudoeste de Etiopía muestra fenotipos deseables para el mejoramiento genético como bajo contenido de cafeína, mayor grado de calidad o resistencia a nemátodos y enfermedades que afectan al fruto (Aerts et al., 2017). Este reservorio de diversidad genética en Etiopía puede ser

explotado mediante introgresiones con *Coffea arabica* como se hizo para obtener al híbrido Timor (Scalabrin et al., 2020).

2.6. Origen y Diversidad Genética del Café en Perú

La Junta Nacional del Café menciona que las primeras plantas de café arribaron a Perú a mediados del siglo XVIII, y fueron sembradas en la provincia de Chinchao, en la Región de Huánuco. Además, menciona que otras plantas de café llegaron desde Guayaquil (Ecuador) y fueron llevadas a la selva central del Perú. Sin embargo, no se tiene información de si llegaron plantas de otros países vecinos productores de café como Brasil y Colombia. Esto ocasiona que no se tenga una ruta histórica y geográfica clara del origen genético de los cultivares que crecen en el país (Mansilla, 2021). Si bien el origen del café peruano es un problema pendiente, existen genotipos (ecotipos) que producen café de alta calidad organoléptica en las condiciones climáticas de la selva peruana; esto hace que sean tomadas en cuenta para posibles programas de mejoramiento genético (Mansilla, 2021).

Sin embargo, para empezar un programa de mejoramiento genético es necesario conocer la diversidad genética de la población inicial. Esto motivó que Mansilla (2021) hiciera un estudio en la región de Pasco-Perú sobre la diversidad genética y estructura poblacional de 17 accesiones de *Coffea arabica* en el fundo de Santa Teresa. En este estudio al evaluar marcadores del tipo SRAP y SSR, se observó un alto porcentaje de bandas polimórficas (51.59%); lo que indica una elevada variabilidad molecular en el germoplasma de Pasco. Por otro lado, se encontraron bajos índices de diversidad genética (Índice de Shannon: I=0.1614) en la misma región (Pasco), lo que evidencia que las accesiones no serían muy diferentes (escaso flujo génico) de los cultivares a partir de los cuales se generaron cuando fueron introducidas al Perú. Esto indicaría que los genotipos peruanos de *Coffea arabica* mantienen el pool genético de su introducción a Perú y podrían ser utilizados para el mejoramiento genético.

2.7. El Cambio Climático y su Efecto en el Cultivo del Café

El cambio climático tiene diversas consecuencias a nivel mundial sobre los ecosistemas. Estas consecuencias se ven acentuadas por las necesidades crecientes del ser humano y el poco manejo sostenible de los recursos ecosistémicos. Como resultado de este mal manejo, para el año 2015 cerca de ¾ de tierras con superficie de hielo fueron afectadas a nivel mundial. También, el uso humano de las tierras a intensidades variables afecta entre el 60 y 85% de bosques y entre 70 y 90% de otros ecosistemas naturales. Además, el mal uso de la tierra ha causado un descenso de la biodiversidad global de alrededor de 11 y 14% (Shukla et al., 2019).

Adicionalmente a las consecuencias sobre las tierras, desde 1850 hasta el 2018 el cambio de la temperatura media superficial fue de ~0.87°C y el de la temperatura del aire superficial de ~1.53°C. Estos cambios de temperatura, que también ocasionan cambios en la precipitación mundial, han alterado el inicio y fin de las estaciones, han contribuido a la disminución de los rendimientos de cultivos, redujeron la disponibilidad de agua fresca, han puesto a la biodiversidad bajo mayor estrés e incrementan la mortalidad de los árboles debido a que estas especies no están adaptadas a cambios climáticos tan rápidos (Shukla et al., 2019).

Agregando a lo anterior, los choques de calor están siendo más frecuentes e intensos debido a la emisión de gases de efecto invernadero antropogénicos; incrementando sequías en la Amazonía, noreste de Brasil, Mediterráneo, Patagonia, la mayoría de África y el noreste de China. Debido a lo descrito, se proyecta que el incremento de sequías va a aumentar en zonas que de por sí ya sufren de sequías como el sur de la Amazonía y el sur de África (Shukla et al., 2019).

En el cultivo del café, los efectos del cambio climático sobre la producción pueden generar grandes riesgos sociales en la seguridad alimentaria de los países productores, ya que puede interferir en la intrincada y frágil relación existente entre los productores, la seguridad regional y su economía, así como, en los aspectos sociales y políticos de los países (Fetzek, 2017). Además, específicamente el estrés por calor reduce la formación de la fruta y acelera el desarrollo resultando en pérdidas de rendimiento, deterioro en la calidad del producto e incrementándose las pérdidas (Shukla et al., 2019).

2.8. Efecto de la Temperatura en el Cultivo de Café

El café arábico crece y fructifica bien en zonas tropicales y es afectado por cambios en sus condiciones ambientales tales como la variación de las horas de luz, frecuencia de lluvias y la temperatura del aire superficial, los que afectan la fenología del cultivo ocasionando cambios en el rendimiento y la calidad del grano (Camargo, 2010).

La temperatura óptima del café tiene un rango de 18 a 23°C; aunque, existen algunos cultivares que crecen óptimamente entre 24 y 25°C. A 30°C el crecimiento puede ser reprimido en cierta medida e incluso si se combina con periodos de sequía puede ocasionar el aborto de las flores. La alta irradiancia puede causar el sobrecalentamiento de las hojas, e incluso en casos extremos, al estar los estomas cerrados se puede llegar a temperaturas de 40°C en las hojas (Camargo, 2010).

Las altas temperaturas, sobre todo en época de verano, puede ocasionar maduración excesiva del fruto. Los árboles de café pueden sufrir cambios fisiológicos por estrés que conlleva a la reducción en la eficiencia fotosintética. Respecto a las flores, se puede acelerar la ruptura de la dormancia de las yemas florales y causar cambios en la fase reproductiva que conlleva a cambios en la producción (desfase de producción) y calidad del grano de café (Camargo, 2010).

2.9. Estrés a Elevadas Temperaturas del Aire en la Planta de Café

Las altas temperaturas pueden causar grandes efectos a nivel morfológico, fisiológico y bioquímico limitando el crecimiento de las plantas y su productividad (Hassan et al., 2021; Marques et al., 2021). Por otro lado, se ha observado que variedades de *Coffea arabica* como Icatú sometidas a 25°C y 37°C muestran respuestas similares a nivel transcriptómico, y que a temperaturas de 40°C se aumenta la expresión de genes (Marques et al., 2021).

El estrés térmico causa la expresión de proteínas de unión al RNA, expresión de proteínas productoras de metabolitos secundarios, expresión de rutas relacionadas al tilacoide, fotosistema, y a la fotosíntesis que tienen relación con una habilidad intrínseca para mantener la actividad fotosintética a altas temperaturas (de Oliveira et al. 2020; Marques et al., 2021). Además, se ha sugerido que los microtúbulos actúan como sensores del estrés abiótico, cambiando el transporte celular, la división,

y la formación de la pared celular (cambios en la permeabilidad la pared celular y membrana celular) en respuesta a estrés de tipo temperatura y sequía (Hassan et al., 2021; Marques et al., 2021).

En relación con los fotosistemas, a pesar de que son sensibles a cambios de temperatura ya que se afecta el transporte de electrones, se ha observado que en café tienen tolerancia al estrés a 37°C y son afectados moderadamente a temperaturas mayores a 40°C gracias a moléculas que disipan la energía y el flujo electrónico alrededor del fotosistema I (Marques et al., 2021).

Los antioxidantes y lípidos también tienen una función esencial en la respuesta al estrés térmico. A nivel cloroplastídico, enzimas con actividad antioxidante y una dinámica intensa de lípidos en la membrana cloroplastídica, junto a desaturaciones de lípidos contribuyen como apoyo a la actividad fotosintética a altas temperaturas (Hassan et al. 2021; Marques et al., 2021).

2.10. Mecanismos Moleculares de Tolerancia a Elevadas Temperaturas en Plantas

Las células en respuesta a elevadas temperaturas activan un mecanismo universal de respuesta a este tipo de estrés, esta respuesta se conoce como *heat shock response* (HSR). En eucariotas, la HSR es mediada por factores de transcripción de estructura conservada llamados *heat shock factors* (HSF), por *heat shock proteins* (HSP) y por moléculas sensoras como especies reactivas de oxígeno (ROS), fosfolípidos, rutas de señalización de calcio y redes de interacción con hormonas (Liu y Charng, 2012; Liu et al., 2015).

La tolerancia al calor o también llamada termotolerancia consiste en dos, la basal y la adquirida. La termotolerancia basal es la habilidad innata de las plantas de sobrevivir a temperaturas por encima de la temperatura óptima de crecimiento. Por otro lado, la termotolerancia adquirida se refiere a la habilidad de lidiar con temperaturas letales después de la adecuación a temperaturas medias, con una analogía a la capacidad humana de adecuarse a mayores altitudes progresivamente (Liu et al., 2015).

Dentro de la termotolerancia basal al estrés por calor, se ha reportado un requerimiento del funcionamiento de las rutas de señalización correspondientes al ácido salicílico, ácido jasmónico y etileno, junto a recolectores de ROS (Liu et al., 2015). Mientra que la termotolerancia adquirida, conlleva a una interacción de HSFs y HSPs.

Los HSFs tienen una conformación trimérica y forman una familia multigénica, la cual está clasificada en tres clases mayores (Clase A, Clase B y Clase C) y varias subclases (Liu y Charng, 2012; Liu et al., 2015). Los HSF de la subclase HsfA1 son considerados como reguladores principales de la HSR. Además, los HSF tienen un rol central en la expresión de los genes de las HSPs al unirse a los elementos cis.

Las HSPs están categorizadas en cinco clases basadas en el peso molecular; Hsp100, Hsp90, Hsp70, Hsp60 y Hsps pequeñas (sHsps). Las HSPs funcionan como chaperonas moleculares y tienen un rol importante al estabilizar proteínas y membranas, y ayudar al replegamiento de proteínas bajo estrés térmico (Liu y Charng, 2012; Liu et al., 2015).

2.11. Genoma de *Coffea arabica*

Coffea arabica es la única especie alopoliploide tetraploide (alotetraploide) en el género *Coffea* (2n = 4x = 44), con una predominante autogamia (auto compatible) mientras que las otras especies del género son diploides (2n = 2x = 22) y auto estériles (Lashermes et al., 1999, Scalabrin et al., 2020, Mekbib et al., 2022). El peso del genoma de especies del género *Coffea* estimado mediante citometría de flujo varía entre 1.03 y 1.76 pg de ADN por núcleo (De Kochko et al., 2010). El genoma de *Coffea arabica* está conformado por la asociación de dos genomas C^a (genoma similar al de *Coffea canephora*) y E^a (genoma similar al de *Coffea eugenioides*) (Lashermes et al., 1999). Debido a que las especies parentales pertenecen al mismo género, *Coffea arabica* tiene dos genomas parcialmente homólogos (debido a su reciente formación evolutiva) a los que se les denomina homeólogos (Mansilla, 2021).

Coffea canephora está ampliamente distribuído en la zona oeste-central de África, mientras que *Coffea eugenioides* está distribuida en zonas frías y secas como las regiones montañosas altas. Entre esas zonas altas del oeste central de África pudo

haberse dado el evento de hibridación interespecífica hace aproximadamente 665000 años (Lashermes et al., 1999, Romero et al., 2014). Según Scalabrin et al. (2020), en los primeros estadíos de *Coffea arabica* luego del evento de hibridación, sucedió un "shock genómico" (pérdida de genes y/o recombinación homóloga) leve que causó el reemplazo homólogo de la punta del cromosoma 7 de 1.2 Mbp. En la actualidad, los cromosomas de especies diploides del género *Coffea* tienen como características morfológicas: pequeños, metacéntricos y submetacéntricos (Romero et al., 2014).

2.12. RNA y Transcriptoma

El ácido ribonucleico (RNA), es un ácido nucleico que tiene estructura similar al DNA. Es un polímero organizado en una cadena larga de ribonucleótidos monofosfatados. Tiene dos diferencias fundamentales con el DNA; presenta dentro de su estructura a uracilo (U) en lugar de timina (T) y el grupo 2´-OH en la ribosa. Es menos estable que la molécula de DNA y susceptible a rupturas nucleofílicas por iones hidróxido. Además, puede presentarse como una estructura geométrica referida como hélice A (el DNA presenta hélice B), que es una estructura de doble hélice conformada por dos RNA a la cual se pueden unir proteínas (Tanaka et al., 1999; Murray et al. 2003, como se citó en Martins et al., 2014). Además, el RNA tiene la capacidad de formar estructuras secundarias y terciarias (Martins et al., 2014).

Las células usan el RNA de diferentes maneras ya que el RNA tiene diferentes formas como RNA mensajero (mRNA), RNA ribosomal (rRNA), RNA de transferencia (tRNA), RNA pequeño (sRNA), microRNA (miRNA), picoRNA (piRNA), RNA circulares (circRNA). Dentro de las funciones o procesos donde participa el RNA se puede mencionar a la traducción en el ribosoma, catálisis, splicing, interferencia (iRNA), estructura del ribosoma, regulación génica (lncRNA y sncRNA), etc (Martins et al., 2014, Hrdlickova et al., 2017).

2.13. Transcriptoma

El transcriptoma es el conjunto de RNAs transcriptos de una célula, porción de tejido u órgano y, la cantidad y variedad de éstos depende de la interacción del organismo biológico con su entorno. El objetivo del estudio del transcriptoma es identificar y catalogar los diferentes transcriptos incluyendo tanto RNAs codificantes como RNA no codificantes. Además, es de interés determinar la estructura génica con relación a patrones de *splicing* y modificaciones post transcripcionales, y cuantificar los niveles de expresión génica de cada transcripto bajo una determinada condición (Wang et al., 2009; Mansilla, 2021).

2.14. Estudio del Transcriptoma

En los años 90, emergieron enfoques de alto rendimiento para estudiar el transcriptoma ya que en años previos todo lo que se conocía provenía de estudios bioquímicos con limitaciones en la cantidad de información que proporcionaban (Hrdlickova et al., 2017).

Las metodologías han ido madurando y mejorando con el tiempo. En los 90´s se desarrollaron metodologías como el etiquetado de secuencias expresadas (EST), el análisis en serie de expresión de genes (SAGE) y los microarreglos (*Microarrays*). EST examinaba la expresión de genes al secuenciar parcialmente el DNA complementario (cDNA) de los RNA, SAGE secuenciaba una región etiquetada pequeña de 15 bp o 21 bp y la tecnología de *Microarrays* se basa en la hibridación de objetivos marcados fluorescentemente que derivan de transcriptos hacia las sondas fijadas en superficie sólida (las cuales se fabricaron a través de la impresión o síntesis in situ); estas técnicas tenían inconvenientes debido a que, o eran muy costosas, o tenían poca sensibilidad (Hrdlickova et al., 2017). Posteriormente, a principios de los 2000 emergió la tecnología de secuenciación de siguiente generación (NGS), la cual permite una secuenciación profunda-paralela que brinda una cantidad de datos gigantesca en poco tiempo. Cabe resaltar que la tecnología NGS se desarrolló principalmente para secuenciar el DNA; sin embargo, es posible la secuenciación del RNA (RNA-Seq) al secuenciar el cDNA proveniente del transcriptoma (Hrdlickova et al., 2017).

2.14.1. Secuenciación de Siguiente Generación

La tecnología de secuenciación de siguiente generación (NGS), aplica la tecnología de secuenciación por síntesis. Esta tecnología está basada en la inmovilización de una cadena simple desnaturalizada de DNA sobre una superficie ya sea una porción de vidrio o nano perlas. Esta inmovilización permite varios ciclos de interacción repetitivos e iterativos con reactivos como nucleótidos A, C, G, T y cebadores. La detección luminiscente de la incorporación de cada nucleótido a la síntesis de la cadena complementaria del DNA inmovilizado se realiza mediante sensores de imágenes de alta resolución. Finalmente, estos datos de imagen son tratados mediante algoritmos para deducir la secuencia del DNA (Weber, 2015).

La primera metodología exitosa fue la secuenciación 454, la cual se basa en la inmovilización en nano perlas y la detección del pirofosfato inorgánico liberado durante la incorporación de un nuevo nucleótido. Luego, el pirofosfato es convertido a ATP por acción de la enzima sulfurilasa seguido de la generación de luminiscencia por la acción de la enzima luciferasa. Inicialmente se obtenían fragmentos secuenciados de 100 bp y aproximadamente 250000 fragmentos por cada proceso de secuenciación; sin embargo, posteriores mejoras en la técnica permitieron obtener fragmentos de 400 bp y hasta 1 millón de fragmentos por proceso (Weber, 2015).

Posteriormente, la empresa Illumina optó por realizar inmovilización en una superficie de vidrio e incorporar la amplificación por puente, seguido de la detección de cuatro colores fluorescentes al insertar un nuevo nucleótido a la cadena complementaria de la secuencia inmovilizada. La técnica de Illumina empezó obteniendose fragmentos de 25 bp de longitud; pero en la actualidad, ya se pueden obtener fragmentos hasta de 300 bp en equipos MiSeq y de 150 bp en equipos HiSeq; además, en un solo ciclo de secuenciación en un equipo HiSeq se pueden obtener 5 billones de fragmentos, lo cual es suficiente para 500 reacciones de RNA-Seq. Finalmente, es importante detallar que la amplificación por puente usada por Illumina permite generar dos tipos de fragmentos de DNA, *single-end* y *paired-end*. Para el segundo, *paired-end* se refiere a la generación fragmentos de ambos lados de una hebra de cDNA inmovilizado (ambos lados del puente); es decir, después de la secuenciación, cada hebra de cDNA ha generado dos fragmentos de DNA. Por el contrario, los fragmentos *single-end* solo generan un fragmento de DNA por cada

hebra de cDNA; es decir, solo se genera información de un lado del puente (Weber, 2015).

2.14.2. Secuenciación de RNA

La técnica de secuenciación de RNA (RNA-Seq) emergió poco después de que se inventara la tecnología NGS. Esta técnica está basada en la transcripción inversa del RNA para generar cDNA y en la amplificación por la reacción en cadena de la polimerasa (PCR). RNA-Seq actualmente es el método principal para estudiar el transcriptoma de las células, es decir, para estudiar el análisis de expresión génica de transcriptos ya sea previamente anotados o nuevos por los siguientes motivos: no depende del conocimiento *a priori* de la secuencia, permite el descubrimiento y cuantificación de secuencias, facilita la detección de variantes génicas, hace posible la cuantificación de la expresión génica, y puede detectar transcriptos de baja expresión bajo condiciones de cambio sutiles (Weber, 2015; Zhao et al., 2015; Shaarschmidt et al., 2020).

2.14.3 Aspectos Específicos de RNA-Seq

El RNA-Seq se puede separar en tres procesos importantes antes de entrar al secuenciador. Esos procesos son: capturar el grupo de RNAs de interés, convertir el RNA a cDNA con tamaños específicos de fragmentos y unir secuencias adaptadoras al cDNA conservando la direccionalidad para poder amplificarlas y secuenciarlas (Wang et al., 2009; Hrdlickova et al., 2017).

Al mencionar grupos de RNAs, específicamente Hrdlickova et al. (2017) se refieren a dos grupos de interés que los investigadores pueden enriquecer para sus experimentos en transcriptómica: exclusivamente RNAs poliadenilados y RNAs poliadenilados junto con RNAs no poliadenilados. Respecto al primer grupo, cuando se estudian eucariotas, la purificación (captura) de RNA poliadenilados (Poly(A)-RNA) es la aplicación más común en los estudios de transcriptómica. La razón se basa en que, en eucariotas, la mayoría de RNAs codificantes de proteínas y algunos lncRNAs (>200nt) contienen una cola poli-A. Esta captura de los Poly(A)-RNAs principalmente se hace mediante perlas magnéticas o perlas de celulosa recubiertas con moléculas oligo-dT debido a que es una técnica que no produce sesgos.

Por otro lado, respecto al segundo grupo, cuando también es de interés las moléculas de RNA no poliadeniladas o cuando se trata de RNAs de procariotas, es necesario degradar el rRNA (es el más abundante, pero de poco interés en estudios). Entre los métodos para eliminar el rRNA, están las técnicas que se basan en la hibridación a sondas específicas con DNA biotinilado, seguido de la degradación con perlas de estreptavidina; las técnicas que se basan en la unión del rRNA con DNA antisentido seguido de la degradación con la enzima RNase H; y finalmente, las técnicas que se basan en la unión del rRNA a cebadores no tan aleatorios (Hrdlickova et al., 2017).

Después de la selección del RNA de interés para el estudio, este tiene que ser fragmentado en un rango de tamaño definido para poder hacer la transcripción inversa. Puede ser fragmentado con cationes divalentes como Mg^{+2}, Zn^{+2} (no es completamente aleatorio) o con enzimas como RNase III (puede causar sesgos) (Hrdlickova et al., 2017).

En los protocolos de RNA-Seq, para la preparación de las bibliotecas de cDNA generalmente se realiza la transcripción inversa usando cebadores hexaméricos aleatorios después de la etapa de purificación de RNA y antes de la etapa de amplificación por PCR. Aunque esta etapa tiene la ventaja de ser simple, usualmente se pierde la información de la direccionalidad del RNA que ha sido convertido en cDNA; es decir, la dirección de la hebra del mRNA proveniente de la transcripción en el núcleo o nucleolo. Esta falta de la información de la hebra dificulta la identificación de RNA antisentido (importante mediador de la regulación génica) y de nuevos transcriptos; además de causar medidas de expresión imprecisas debido a la sobreposición de genes (Zhao et al., 2015; Hrdlickova et al., 2017).

En consecuencia, para conservar la direccionalidad de la secuencia de cDNA, se usan protocolos que retienen la información de la hebra original (*stranded* RNA-seq). Entre estos protocolos se puede mencionar a los que unen los adaptadores directamente a los fragmentos de RNA, los que incorporan dUTP cuando se sintetiza la segunda hebra del cDNA y los que incorporan etiquetas mientras que se sintetiza el cDNA. Entre los mencionados, el protocolo de dUTP es el que mejor resultados ha tenido. y se caracteriza porque utiliza moléculas de dUTP para sintetizar la segunda hebra del cDNA. Esta hebra marcada con uracilos (U) es degradada antes de la etapa de amplificación por PCR mediante la enzima DNA uracil glicosilasa

(UDG), la cual corta al uracilo de la segunda hebra del cDNA. En consecuencia, solo la primera hebra cDNA (la que se formó por transcripción inversa) con sus secuencias adaptadoras será amplificada, brindando información acerca de la direccionalidad de los fragmentos RNA (Zhao et al., 2015; Hrdlickova et al., 2017).

Debido a los límites de detección que presentan los secuenciadores de DNA, las bibliotecas de cDNA deben ser amplificadas por PCR (8-12 ciclos) antes de ser secuenciadas. Sin embargo, la amplificación por PCR genera sesgos debido a la cantidad relativa de fragmentos presentes y a la longitud de estos. Por lo tanto, para eliminar estos sesgos se han generado métodos como etiquetas moleculares conocidas como identificadores moleculares únicos (UMIs) para distinguir los fragmentos resultantes de la PCR. Estos UMIs son colocados dentro de la región de los adaptadores antes de la amplificación; además, tienen diferentes tamaños y complejidad, y pueden poseer secuencias definidas o aleatorias (Hrdlickova et al., 2017).

2.15. Análisis Bioinformático de RNA-Seq

Hasta la fecha el único consenso acerca del mejor flujograma de trabajo para analizar datos de RNA-Seq resultó del primer workshop Grupo de Trabajo de Análisis (AWG) organizado por la NASA en el año 2018, donde se reunieron investigadores de varias instituciones con el objetivo de llegar a un punto en común sobre un flujograma para el análisis de datos de transcriptómica provenientes de experimentos espaciales. En este flujograma consenso se definieron cuatro etapas esenciales para el análisis de datos de RNA-Seq. Estas etapas son: (1) preprocesamiento de datos de RNA-Seq, (2) procesamiento de datos de RNA-Seq, (3) análisis de expresión diferencial de genes y (4) análisis de categorías funcionales de genes (Overbey et al., 2021).

2.15.1. Preprocesamiento de Datos de RNA-Seq

El control de calidad (QC) y el preprocesamiento son críticos para obtener datos de alta calidad y de alta confianza para el posterior análisis. Estos procesos se realizan sobre los archivos FASTQ (archivos provenientes del proceso de secuenciación que contienen las secuencias e identificadores de calidad de los fragmentos secuenciados), los cuales pueden presentar contaminación de adaptadores, sesgos en el contenido de bases nitrogenadas y secuencias sobrerrepresentadas (Chen et al., 2018).

Existen varias herramientas bioinformáticas que se encargan del control de calidad y/o del preprocesamiento de datos. Entre estas herramientas se pueden mencionar algunas como FASTQC, Cutadapt, Trimmomatic y Fast; a continuación, se mencionarán las características más resaltantes de estos programas bioinformáticos. El programa FASTQC es un programa escrito en lenguaje Java enfocado en el control de calidad de los archivos FASTQ. Este control de calidad es, en otras palabras, un análisis exploratorio de datos que se enfoca en visualizar características de los fragmentos secuenciados (*reads*) como, por ejemplo: la calidad de cada nucleótido que conforman los *reads*, la calidad de cada *read*, el porcentaje de GC de cada *read*, detectar la presencia de adaptadores, etc (Andrews, 2010; Overbey et al., 2021). Siguiendo con las herramientas, el programa Cutadapt es una herramienta escrita en lenguaje Python enfocada en el corte de secuencias adaptadoras, usando como metodología el emparejamiento de los *reads* con secuencias adaptadoras predefinidas por el investigador y también tiene opciones de filtrado de *reads* que presentan baja calidad (Martin, 2011). Similar a Cutadapt, el programa Trimmomatic escrito en lenguaje Java, tiene la particularidad de facilitar el procesamiento de *paired-end reads* (véase las secciones “Secuenciación de siguiente generación” y “Aspectos específicos de RNA-Seq”); además, se encarga del corte de secuencias adaptadoras con la misma metodología de Cutadapt y puede desechar nucleótidos de los *reads* que presentan baja calidad, proceso conocido en inglés como *trimming* (Bolger et al. 2014). Por último, el programa Fastp (escrito en lenguaje C++) cuenta con funciones para el control de calidad (antes y después del preprocesamiento), corte de secuencias adaptadoras, corrección de nucleótidos para *paired-end reads*, corte de colas poli-G,

preprocesamiento de identificadores UMI y corte de nucleótidos de baja calidad por *read* (Chen et al., 2018).

Adicional a las herramientas descritas arriba, el programa MultiQC es una herramienta que permite crear resúmenes de reportes estadísticos de varias muestras de forma simultánea; evitando el tedioso trabajo de analizar la calidad varias muestras una por una (Ewels et al., 2016; Chen et al., 2018).

2.15.2. Procesamiento de Datos de RNA-Seq

El procesamiento de datos consta básicamente de dos etapas: la primera es el uso de herramientas bioinformáticas para mapear y alinear los *reads* a una secuencia nucleotídica y la segunda consta de la cuantificación de esos *reads* alineados.

2.15. 2.a Alineamiento y Mapeo de *Reads*

Los archivos FASTQ preprocesados deben ser alineados de forma precisa para los posteriores análisis. Si bien este proceso es simple en procariotas, cuando se habla sobre eucariotas surge el problema de ¿cómo manejar los *reads* que provienen de sitios de *splicing?* Para entender este problema es necesario imaginar a estos *reads* como bloques de LEGO bicolores, dónde un color corresponde a un exón o gen y el otro color corresponde a otro exón o gen. De modo que, para solucionar este problema, se necesitan hacer dos tareas: primero, alinear de manera precisa los *reads* de los archivos FASTQ (estos *reads* pueden contener *mismatches*, inserciones y delecciones causados por variaciones genómicas o errores de secuenciación) y segundo, mapear las secuencias de los *reads* (secuencias que han sido unidas por el proceso de *splicing*) que provienen de regiones no continuas del genoma (Dobin et al., 2013).

Por otro lado, dos tipos de enfoques se pueden utilizar en RNA-Seq para alinear los *reads*: ensamblar fragmentos *de novo* en secuencias contiguas o alinear los *reads* a una referencia genómica. En el caso del alineamiento de los *reads* a una secuencia de referencia, es preferible que esta sea el genoma o transcriptoma de la especie en estudio. Algunas dificultades se pueden presentar si se usa como referencia el transcriptoma, ya que este puede estar incompleto y, si se usa el genoma de referencia de una especie relacionada, este puede no presentar o presentar genes que posea la

especie de estudio (Weber, 2015). El alineamiento de *reads* a una referencia permite obtener datos como conteos de los *reads* sobre genes y/o exones, identificación de isoformas inducidas por *splicing*, cuantificar la cobertura de los *reads* sobre genes, verificar modelos de genes o identificar variantes génicas (de tipo SNPs y/o InDels) (Weber, 2015, Schaarschmidt et al., 2020; Overbey et al., 2021).

Entre los programas con diferentes algoritmos que se usan para el mapeo de *reads* se puede mencionar a BWA, HISAT2 y STAR. El programa BWA (*Burrows-Wheeler-Alignment*), es un programa específicamente desarrollado para mapear secuencias de DNA a un genoma de referencia y su uso se extendió para RNA-Seq, donde se usa la transformación de Burrows-Wheeler (BWT) y realiza el algoritmo de búsqueda hacia atrás (*backward search*). El programa STAR (*Spliced Transcripts Alignment to a Reference*) es un programa desarrollado específicamente para RNA-Seq basado en la búsqueda por extensión de "semillas" (*seed-extension*) y puede detectar diferentes empalmes de *splicing* (Dobin et al., 2013). El programa HISAT2 (*Hierarchical Indexing for Spliced Alignment of Transcripts 2*) es también un programa que detecta sitios de *splicing* al usar un algoritmo basado en grafos que puede alinear tanto secuencias de DNA como RNA (Schaarschmidt et al., 2020).

Los programas HISAT2 y STAR, aparte de tener mejores desempeños frente a BWA y otros programas no descritos (Bowtie 2 y Tophat2), tienen similar desempeño al evaluar el porcentaje de *reads* mapeados y el mapeo correcto a la secuencia de referencia (Corchete et al., 2020; Schaarschmidt et al., 2020).

2.15.2.b Cuantificación de *Reads*

La etapa de alineamiento brinda como resultados archivos en formato SAM (*Sequence Alignment Map*) y/o BAM (*Binary Alignment Map*). Específicamente, los archivos en formato BAM sirven para la cuantificación del número de *reads* mapeados a genes y/o transcriptos (Overbey et al., 2021).

Entre los *softwares* más usados para el conteo de *reads* se pueden mencionar a *Htseq-count* y a RSEM *(RNA-Seq by Expectation Maximization*). El programa Htseq-count, como preferencia solo cuenta los *reads* que se mapean a una sola posición genómica y descarta a los *reads* que son mapeado a múltiples loci. Entre las opciones que ofrece *Htseq-count* para manejar los *reads* que se mapean a varios loci están: contar

fraccionalmente y asignar de manera aleatoria con una distribución uniforme. Para el caso del conteo fraccional, si un *read* se mapea a 3 lugares, cada lugar recibe un conteo de $1/3$ y; para el caso del conteo aleatorio, cada posición genómica tiene la misma probabilidad de recibir ese *read* como parte de su conteo, pero solo uno de ellos recibe ese conteo. Además, htseq-count tiene tres opciones de conteo de *reads*: *union* (unión), *intersection-strict* (intersección estricta) y *intersection-nonempty* (intersección no vacía). En la opción *union*, el *read* para ser contado debe pertenecer solo a un gen, la opción *intersection-strict*, no cuenta a los *reads* que tienen secciones que sobresalen del gen y, la opción *intersection-nonempty* es similar a la opción *union*; sin embargo, la diferencia radica en que *intersection-nonempty* si contabiliza a los *reads* que pueden alinearse parcialmente a dos genes que se solapan en hebras opuestas. Por ejemplo, si el gen A ubicado en la hebra (+) del genoma de referencia se intercepta parcialmente con el gen B ubicado en la hebra (-), el *read* es contado por *intersection-nonempty* si y solo si puede ser distribuido entre los dos genes (Anders et al., 2015).

Reanudando con la explicación de los *softwares*, el programa RSEM está diseñado para contar los *reads* que se mapean a múltiples regiones genómicas e identificar isoformas de RNA, algo que no puede realizar *htseq-count*. RSEM logra este mapeo múltiple mediante máxima verosimilitud (*Maximum Likelihood*: ML) usando el algoritmo de Maximización del Valor Esperado (*Expectation-Maximization*: EM) para estimar los parámetros del modelo estadístico. Adicionalmente, si se decide computar estimaciones *a posteriori* de la media (PME) y obtener intervalos creíbles al 95%, RSEM cambia a su versión bayesiana con la distribución de Dirichlet como modelo *a priori* (con parámetros 1 lo que ocasiona que sea parecida a la distribución uniforme) para generar estimaciones. Entre las estimaciones que genera RSEM ya sea por ML o por el teorema de Bayes, están la longitud efectiva del gen, la abundancia o conteo esperado (un valor no entero) de fragmentos que derivan de un gen o isoforma y la fracción estimada de transcriptos de cada gen o isoforma, la cual si se le multiplica por 10^6 se forma una medida llamada transcriptos por millón (TPM) que indica la cantidad de transcriptos en un millón de transcriptos. La longitud efectiva del gen proviene de un promedio ponderado (*weighted mean*) de la longitud efectiva de las isoformas o transcriptos que puede tener ese gen. La longitud efectiva de un transcripto o isoforma es de manera simplificada la longitud del transcripto (L)

menos la longitud del fragmento secuenciado (F) más uno: L-F+1. Sin embargo, debido a que los fragmentos secuenciados poseen diferentes tamaños, la longitud efectiva de un transcripto es: L-promedio(F)+1. El conteo esperado es el conteo de reads que son mapeados a un solo locus más una fracción (asignada probabilísticamente) de los reads que son mapeados a múltiples loci. TPM es una medida normalizada que toma en cuenta la profundidad de secuenciamiento y la longitud del gen. Primero se divide entre la longitud efectiva de cada gen y luego se

divide entre la suma de los reads escalada por un factor de 10^{-6} . FPKM (o RPKM para experimentos single-end) es similar a TPM con la diferencia que invierte el orden de las operaciones; es decir, primero divide el total de reads escalados por un factor de 10^{-6} y luego se divide la longitud efectiva del gen (Li y Dewey, 2011, Starmer J., 2015).

Además, RSEM también puede modelar el sesgo producido en el extremo 3´ por protocolos de RNA-seq donde se usan primers dT para filtrar a los mRNA y finalmente, también permite especificar si el protocolo que se usó para preparar la biblioteca de RNA es del tipo *stranded* lo que permite hacer estimaciones del conteo de *reads* más precisas (Li & Dewey, 2011).

RSEM es ejecutado en dos pasos, el primero tiene por objetivo generar transcriptos de referencia usando los archivos GTF y el genoma de referencia; el segundo, utilizando los transcriptos de referencia generados en el paso uno, usa el archivo de alineación BAM y estima las abundancias mediante el algoritmo EM (Li & Dewey, 2011; Overbey et al., 2021).

2.15.3. Análisis de Expresión Diferencial de Genes

El análisis de expresión diferencial de genes es una tarea importante en el proceso de RNA-Seq. Tiene como objetivo encontrar genes (también se puede aplicar a exones e isoformas) que se expresan diferencialmente entre grupos de muestras dentro de un experimento (Love et al., 2014).

Para realizar el análisis de expresión diferencial de genes se utilizan los datos de conteo de genes o isoformas provenientes de la etapa de procesamiento de datos. Sin embargo, estos datos (al ser conteos) presentan dificultades tales como no cumplir

con los supuestos de normalidad y homocedasticidad necesarios para realizar metodologías como análisis de la varianza (ANOVA). Además, los experimentos de RNA-Seq tienen un número pequeño de muestras y pocas repeticiones biológicas por condición experimental, esto dificulta el proceso de inferencia ya que, al tener pocas repeticiones no se pueden usar aproximaciones como el teorema del límite central (CTL) y no se puede modelar cada gen separadamente por la incertidumbre que generan pocas repeticiones. Estos problemas hacen que sea necesario optar por otras metodologías como el modelamiento estadístico de los datos de conteo junto con enfoques de máxima verosimilitud y/o enfoques bayesianos para estimar el cambio logarítmico doble entre grupos experimentales y sus respectivos estadísticos (Love et al., 2014).

2.15.3.a Análisis de Expresión Diferencial para Datos de Conteo de Secuencias

El programa Análisis de expresión diferencial para datos de conteo de secuencias (*Differential expression analysis for sequence count data 2*: DESeq2) fue diseñado para analizar a nivel de genes los conteos en bruto de RNA-seq modelándolos con la distribución binomial negativa. Utiliza estimadores de contracción (*Shrinkage estimators*) provenientes del método empírico de Bayes, que permiten mayor estabilidad y reproducibilidad de los resultados en comparación con métodos basados solamente en máxima verosimilitud. Para generar los estimadores de contracción, DESeq2 usa modelos jerárquicos, también conocidos como métodos empíricos de Bayes, en los que genera modelos *a priori* mediante máxima verosimilitud para la estimación de parámetros como la dispersión y la estimación logarítmica del cambio doble (*log 2 fold-change*: LFC) (Love et al., 2014). Estos modelos jerárquicos (aplicados ampliamente en el análisis de datos de alto rendimiento) se basan en dos niveles de modelamiento estadístico: el primer nivel describe la variabilidad a través de las muestras/unidades por cada gen, y el segundo describe la variabilidad entre genes por cada muestra (Irizarry y Love, 2016).

2.15.3.b Análisis de Variables Sustitutas

Adicionalmente a las variables de interés del estudio de RNA-Seq, existen fuente de ruido o factores no medidos o no modelados o muy difíciles de capturar que pueden influenciar la expresión de cualquier gen. En particular, estos factores, además de generar dependencia entre genes, crean fuentes adicionales de expresión diferencial que no son fluctuaciones específicas de los genes; sino, fuentes comunes de variación que pueden observarse en múltiples genes. Estos factores pueden ser técnicos, ambientales, demográficos, o genéticos (Leek y Storey, 2007).

El programa *Surrogate Variable Analysis* (SVA) se encarga de identificar y estimar estos factores en las matrices de conteo de genes para poder introducirlos en la expresión diferencial y obtener mayor poder en la comparación de tratamientos. SVA estima estos factores en cuatro pasos:

1. Remueve la señal que se debe a la(s) variable(s) de interés de la matriz de expresión normalizada. Luego aplica la descomposición del valor singular (*Singular Value Decomposition:* SVD) a la matriz residual para obtener los vectores singulares de la matriz V^T que reproducen estos factores desconocidos. A estos vectores se les aplica una prueba estadística para determinar a los vectores singulares que representan mayor variación de la que se esperaría por aleatoriedad (Leek y Storey, 2007).
2. Se identifica un subconjunto de genes que generan el patron de los vectores singulares significativos a través de un análisis de significancia entre los genes y la matriz residual (Leek y Storey, 2007).
3. Para cada subconjunto de genes, se contruye una variable sustituta basada en la matriz V^T y la matriz de conteo normalizada.
4. Se incluye todas las variables sustitutas como covariables en la especificación del modelo de DESeq2 (Leek y Storey, 2007).

2.15.3.c Estimación *a posteriori* Aproximada para el Modelo Lineal Generalizado

Es un reto estimar con precisión los LFC de genes que tienen baja expresión, o genes con alto coeficiente de variación. En consecuencia, cuando no se toma en cuenta estos dos factores importantes, se obtienen LFC con alta varianza que se traducen en LFC´s grandes. Para reducir este problema, DESeq2 usa como modelo *a priori* una distribución normal estándar adaptativa ($N \sim (0, \sigma_i^2)$) para producir un ajuste de estos LFC. Sin embargo, al usar esta distribución y al filtrar los genes con poca información, resulta en la pérdida de genes con suficiente "señal", o en un sobreajuste agresivo del verdadero valor del LFC (Zhu et al., 2018).

Como consecuencia de este reto, Zhu et al. (2018) crearon el programa *apeglm* disponible en el paquete *apeglm* de Bioconductor. La solución radica en cambiar el modelo normal estándar por una distribución de Cauchy adaptativa como modelo *a priori* de los LFC. *Apeglm* ajusta los LFC de la siguiente manera:

- Ajusta cada coeficiente (β) a la vez. Si los β no están disponibles para un gen, el modelo toma como escala (S) 1, lo que hace que la distribución de Cauchy se asemeje a una distribución de t de Student con 1 grado de libertad (Zhu et al., 2018).
- Ajusta la escala (S) del modelo a la distribución observada de los LFC obtenidos por máxima verosimilitud en la etapa de expresión diferencial (Zhu et al., 2018).
- El estimado ajustado (LFC ajustado) es el valor que tiene la máxima probabilidad *a posteriori* (*máximum a posteriori probability:* MAP) de la distribución de probabilidad *a posteriori* (Zhu et al., 2018).

Este método ayuda a obviar la necesidad de filtrar genes debido a que toma en cuenta la información estadística de los datos para estimar el tamaño de efecto. Finalmente, para los genes que tienen bajos conteos, este método ajusta hacia el cero los respectivos LFC, con el objetivo de aliviar el problema de las estimaciones poco confiables que éstos conllevan a tener pocos conteos (Zhu et al., 2018).

2.15.4. Análisis de Categorías Funcionales de Genes

El análisis de expresión diferencial de genes da como resultado una lista de genes con sus asociados tamaño de efecto (LFC) y estadísticos (estadístico de t y p-valores ajustados) (Marini et al., 2021). Esta lista es extremadamente útil para identificar a los genes que pueden tener roles importantes (genes candidatos) en el tratamiento estudiado (Irizarry et al., 2011). Sin embargo, extraer el significado biológico (funcional) de una lista con miles de genes es un reto, debido a que: la interpretación de los genes significantes depende de la experiencia del investigador, los genes estadísticamente significativos pueden no tener un tema en común (ruta metabólica, proceso biólogico) y la replicación del experimento puede llevar a una lista con poco solapamiento de genes significativos debido a la estocasticidad de las células; es decir baja reproducibilidad del experimento (Subramanian et al., 2005; Khatri et al., 2012).

Estos retos han generado un enfoque para simplificar el análisis y extraer información de los patrones funcionales mediante la agrupación de una larga lista de genes en grupos pequeños de genes relacionados entre sí (*gene sets*) basados en un conocimiento previo (Irizarry et al., 2011; Marini et al., 2021). Este conocimiento previo se refiere a que, a lo largo de los años, las investigaciones en el ámbito genético han generado información de genes de interés y los han asociado a funciones biológicas, procesos moleculares, funciones estructurales y en consecuencia de esto han asociado genes y sus productos a rutas metabólicas. Todo este conocimiento se ha transformado en bases de datos de ontología como *Gene Ontology* (GO), *Kyoto Encyclopedia of Genes and Genomes* (KEGG), *QuickGO*, *Reactome*, y *Plant Reactome*, que describen y asocian procesos biológicos, componentes o estructuras moleculares, en las cuales se sabe la relación existente entre genes (Khatri et al., 2012). Por lo tanto, el usar estas bases de datos reduce la complejidad de la lista de genes procedentes de la etapa de expresión diferencial e incrementa el poder explicativo de los mecanismos biológicos observados en el experimento (Marini et al., 2021).

Siguiendo este razonamiento, se han generado tres enfoques de análisis en el tiempo para usar estas bases de datos de ontología: Análisis de sobrerrepresentación (ORA),

Puntuación de clases funcionales (FCS) y Enfoques basados en topología de rutas (PT) (Khatri et al., 2012).

El enfoque ORA se diseñó para extraer información de estudios de *microarrays* y luego fue adaptado para RNA-Seq. ORA necesita dos cosas, una lista de genes candidatos expresados diferencialmente que hayan pasado un límite o criterio como 5% de *false discovery rate* (FDR) y una base de datos de ontología que agrupe a los genes en *gene sets*. La estrategia es, para cada *gene set*, contar cuantos genes de la lista aparecen en el *gene set*; por ejemplo: un *gene set* puede tener 5 genes de una lista de 300 genes; este proceso se realiza para cada *gene set* de interés (recordar que los *gene sets* pueden ser rutas metabólicas, procesos biológicos, etc. como se menciona en el primer párrafo de la sección). Posteriormente, cuando ya se tiene los conteos de los genes que pertenecen a cada *gene set*, se realizan pruebas de hipótesis (pueden ser basadas en distribuciones estadísticas como hipergeométrica, binomial o chi-cuadrado) comparando los genes pertenecientes a un *gene set* versus los que no pertenecen (Khatri et al., 2012).

Sin embargo, dentro de las limitaciones del enfoque ORA están: usar sólo a genes más significantes que hayan pasado un límite arbitrario y descartando a los que no lo pasaron, usar sólo el conteo de genes dentro de cada clase funcional, no usar ningún estadístico proveniente de la expresión diferencial y, asumir que cada gen es independiente de otro gen ignorando la complejidad de la interacción entre genes (Irizarry et al., 2011; Khatri et al., 2012).

El segundo enfoque, llamado puntuación de clases funcionales (FCS), se basa en la premisa de que, si bien los genes individuales pueden tener un impacto grande en la célula, los cambios débiles pero coordinados de grupos de genes relacionados pueden también tener cambios significativos. En el análisis, primero se necesita un estadístico a nivel de genes como el LFC. Luego, se agrupan los genes con sus estadísticos en *gene sets* y, para cada *gene set* se calculan estadísticos de enriquecimiento; los cuales, dependen de la proporción de genes diferencialmente expresados dentro del grupo de genes, del tamaño del grupo de genes y de la correlación de los genes dentro de ese grupo. Finalmente, se calcula el nivel de significancia del grupo funcional mediante permutaciones y escogiendo entre dos tipos de hipótesis nula: competitiva (se hace la permutación a los genes para cada

grupo de genes y se compara los genes dentro del *gene set* contra los genes fuera del *gene set*) y autocontenida (se hace la permutación a los nombres de los fenotipos para cada muestra y se compara el grupo de genes dentro de un *gene set* consigo mismo ignorando los genes fuera del grupo de genes) (Khatri et al., 2012; Geistlinger et al., 2021).

El enfoque FCS no requiere un límite o criterio (como 5% de FDR) ya que se usan todos los genes de la expresión diferencial de genes, es decir, los que pasan un límite (FDR) y los que no. Además, FCS usa la información de estadísticos como LFC, el valor de t o el *z-score* para detectar cambios coordinados en el mismo gene set lo que permite adicionar la correlación de genes dentro del *gene set* (Khatri et al., 2012).

El tercer enfoque, se basa en la topología de rutas (PT). Además de utilizar información de los genes pertenecientes a un *gene set*, también usa información de la interacción de los productos de los genes, cómo la forma de la interacción (activación o inhibición) y el lugar donde se realiza (citoplasma, núcleo, etc.). En general, sigue los mismos pasos que el enfoque FCS; sin embargo, se diferencia en que usa la información topológica de los genes para calcular los estadísticos a nivel de gen. Este enfoque toma en cuenta que la red de genes cambia en diferentes condiciones experimentales. Por otro lado, dentro de las limitaciones actuales (las cuales son la razón de su uso escaso en especies no modelo) están: la ruta metabólica depende del tipo de célula y de la condición experimental, la falta de información específica que relaciona SNPs, isoformas de transcriptos, rutas metabólicas y la interacción de éstas para estudiar un fenotipo dinámico en el tiempo (Khatri et al., 2012).

2.15.4.a. Análisis Rápido de Enriquecimiento de *Gene Sets*

El Análisis rápido de enriquecimiento de *gene set* (*Fast gene set enrichment analysis*: FGSEA) es un algoritmo perteneciente a los enfoques FCS que extrae información esencial de la expresión diferencial de genes.

El objetivo de GSEA es determinar si los genes de un *gene set* se posicionan al inicio o al final de una lista de genes que ha sido ordenada (*ranked*) usando un estadístico de interés como el LFC. Si sucede que los genes del *gene set* se agrupan al inicio o al final de la lista ordenada, se puede decir que el *gene set* está correlacionado con la clase fenotípica (tratamiento) en estudio (Subramanian et al., 2005).

El algoritmo GSEA primero calcula estadísticos de enriquecimiento (ES) para cada *gene set* con el algoritmo de camino aleatorio (*random walk*) basados en un estadístico de interés proveniente de la expresión diferencial de genes (los ES corresponden a un estadístico ponderado similar a Kolgomorov-Smirnov) (Subramanian et al., 2005).

Después, GSEA realiza permutaciones para generar una distribución nula de los estadísticos ES y luego calcula los *p-values* de estos estadísticos. Para el ajuste de pruebas de hipótesis múltiples, los ES se normalizan de acuerdo con el tamaño de cada *gene set* obteniendo estadísticos de enriquecimiento normalizados (NES) y así, obtener *p-values* ajustados (Subramanian et al., 2005).

GSEA presenta problemas al estimar los *p-values* cuando en un experimento se tienen pocas muestras ya que los estima como cero. La solución práctica es incrementar el número de permutaciones dependiendo del experimento, pero, por la implementación de GSEA le es imposible superar niveles permutaciones mayores a 10^4. FGSEA soluciona el problema de GSEA al tener un mejor algoritmo de implementación basado en programación dinámica. FGSEA es capaz de estimar p-valores muy pequeños (10^{-100}) con gran precisión en minutos o incluso segundos (Korotkevich et al., 2021).

III. METODOLOGÍA

3.1. Materiales

- Set de datos de RNA-Seq. Los datos de RNA-Seq fueron obtenidos de la base de datos del *Sequence read archive* (SRA) del *National Center of Biotechnology Information* (NCBI) que provienen de experimentos realizados para determinar el efecto de las elevadas temperaturas en hojas de plantas de *Coffea arabica*.
- Servidor HPC Bioinformática UNALM.
- Laptop TOSHIBA.

3.2. Metodología

3.2.1. Búsqueda, Selección y Descarga de Datos del SRA del NCBI

La búsqueda y selección de los datos de secuenciación de RNA-Seq es el primer paso y el más importante debido a que es la etapa en la que se seleccionó el organismo, se escogió el tipo de secuenciación y tratamiento de interés; también, se definieron los límites acerca de los protocolos experimentales de interés para el estudio.

3.2.1. a Búsqueda y Selección de Bioproyectos

- La búsqueda de los archivos FASTQ se realizó a través de la página web del NCBI. Primero, se buscó el término "coffea arabica" en la base de datos del SRA. Luego, se enviaron los resultados de la búsqueda al selector de ejecución (Run Selector) (Leinonen et al., 2010).

- La tabla de metadatos que mostró el Run selector, fué filtrada para que los datos sólo correspondan a la especie *Coffea arabica*. La columna de

- *Bioproject* fue usada para buscar en la base de datos de Bioproyectos del NCBI a los que tengan asociados un artículo científico (Barret et al., 2012).

- Los bioproyectos seleccionados fueron buscados de nuevo en el *Run Selector* del NCBI y se descargó la tabla de metadatos.

3.2.1. b. Selección de Archivos FASTQ

- Se aplicaron filtros a los metadatos según los siguientes criterios:
    - El tejido a estudiar fue la hoja de *Coffea arabica*.
    - El efecto temperatura fue el factor de estudio y sus niveles se determinaron de acuerdo con los tratamientos de temperatura disponible en los metadatos.
    - Bioproyectos con menos de dos réplicas biológicas fueron descartados.
- Cada bioproyecto tiene asociado corridas (*Runs*) en la metadata que corresponden a códigos de identificación de cada muestra secuenciada en la base de datos del SRA. Los códigos de identificación fueron buscados en el *Run Selector* y se presionó el botón Galaxy para enviar la lista de accesiones (muestras) a la plataforma Galaxy.

3.2.1. c. Descarga de Archivos FASTQ

- Los archivos FASTQ fueron descargados en la plataforma web Galaxy (www.usegalaxy.org), para lo cual, luego de abrir la plataforma, en el panel izquierdo se seleccionó la opción "*Get Data*" (Afgan et al., 2018).
- Luego se seleccionó la herramienta "*Faster Download and Extract Reads in FASTQ*" y se procedió de la siguiente manera:
    - En la opción "*select input type*" se seleccionó la opción "*List of SRA accession, one per line*" de la barra desplegable.
    - De manera automática apareció "SRA" debajo de la opción "*sra accession list*".
    - Finalmente. se presionó el botón "E*xecute*".

3.2.2. Control de Calidad y Preprocesamiento de Archivos FASTQ

Para el control de calidad y preprocesamiento de los *reads* se seleccionó el *software* Fastp (Chen et al., 2018), ya que es una herramienta que realiza el control de calidad y preprocesamiento de datos (filtrado de *reads*, depuración de *reads*, corte de secuencias adaptadoras, corte de colas *polyG/polyX*) en una sola lectura de un archivo FASTQ.

El programa Fastp fue ejecutado dentro de la plataforma Galaxy, en la que ya se tenían descargados los archivos FASTQ, procediendo de la siguiente manera:

- En la sección "*Single-end or paired reads*" se seleccionó "*Single-end*" o "*Paired Collection*" dependiendo de si el bioproyecto generó *reads single-end* o *paired-end*.

- El programa usó los parámetros en *default* excepto para datos *paired-end*, para los cuales, se activó la función de corrección de bases por sobreposición (*Base correction by overlap analysis*). Además, en la opción de "*output*", se seleccionó las opciones *HTML report* y *JSON report*.

- Los parámetros por *default* son los siguientes:

    - Opciones de recorte de adaptadores: Fastp indentifica automáticamente las secuencias adaptadoras de los *reads.* Para *reads single-end* identifica automáticamente las secuencias adaptadoras mediante un algoritmo de árbol y para los *reads paired-end* solapa a los pares de *reads* y las secuencias que sobresalen en los extremos las considera como adaptadoras.

    - Opciones de filtrado: Fastp filtra a los *reads* que tengan secuencias menores a Q15, *reads* que posean bases N (no identificadas) mayores a 5, *reads* con menos de 15 pares de bases y *reads* que posean una complejidad de secuencia menor a 30% (porcentaje de bases consecutivas).

- Opciones de modificación de *reads*: Fastp recorta el final de los *reads* si estos poseen mas de 10 guaninas consecutivas (*PolyG tail trimming*).

De manera similar, se ejecutó el software MultiQC (Ewels et al., 2016) dentro de la plataforma Galaxy para resumir los reportes estadísticos de la calidad y filtrado de *reads* de todas las muestras de cada bioproyecto.

3.2.2.a. Análisis de componentes principales sobre las estadísticas de alineamiento

Con el objetivo de analizar los patrones de mapeo entre las especies *Coffea arabica, Coffea eugenioides* y *Coffea canephora*, se hizo el análisis de componentes principales (PCA) sobre las estadísticas de STAR. Este método consta de realizar el método de descomposición matricial SVD para obtener tres matrices: U , $\sum$ y V^T. U es la matriz que contiene a los autovectores normalizados ordenados en forma decreciente dependiendo de la variabilidad explicada, $\sum$ es una matriz diagonal que contiene las desviaciones estándar de los autovectores (raíz cuadrada de los autovalores) ordenados decrecientemente dependiendo de la variabilidad explicada y V^T es la matriz traspuesta de los autovectores ordenados en forma descreciente dependiendo de la variabilidad explicada. Al graficar a los componentes principales, se usa a los vectores normalizados de la matriz U (por ejemplo, los dos primeros vectores; es decir, las primeras dos componentes) junto a los primeros valores de $\sum$ (siguiendo el ejemplo a los dos primeros valores elevados al cuadrado) los nos dan información de la variabilidad explicada (Bagheri, 2020).

Adicionalmente, se añadieron en el gráfico de PCA a las variables que conforman las estadísticas de mapeo. Esta técnica se conoce como PCA Biplot el cual consiste en graficar a las variables dependiendo de la correlación que tengan con los autovectores de la matriz U y la correlación entre las mismas variables. Por ejemplo, las variables correlacionadas positivamente apuntan a una dirección similar, las variables correlacionadas negativamente apuntan a lados opuestos y la distancia de las variables del origen mide la calidad de la representación de la variable en el gráfico de PCA (Kassambara, 2017).

3.2.3. Procesamiento de Archivos FASTQ

La etapa de procesamiento contó con dos etapas: el alineamiento-mapeo de *reads* y la cuantificación de los *reads*.

Para el alineamiento-mapeo de *reads,* se utilizó el software STAR (Dobin et al., 2013) debido a que está diseñado específicamente para procesar datos de RNA-seq; además, STAR es rápido, sensible para detectar isoformas por *splicing*; permite alinear *reads single-end* y *paired-end*, y, manejar protocolos *stranded*; además, es aplicable a organismos que tienen regiones genómicas con baja y alta complejidad; asimismo, permite identificar *reads* quiméricos (*reads* formados por más de 1 gen) y la salida de alineamientos tanto al genoma como al transcriptoma (Overbey et al., 2021).

El mapeo se realizó frente a genomas y anotaciones de *Coffea arabica* (GCA_003713225.1 Cara_1.0), *Coffea eugenioides* (GCA_003713205.1 Ceug_1.0) y *Coffea canephora* (AUK_PRJEB4211_v1), los dos primero disponibles en la web de *Genome* del NCBI y el tercero en la web de *Ensmbl plants*.

En la ejecución de STAR se realizaron dos pasos:

1. Se crearon archivos genómicos indexados con el genoma en formato FASTA y la anotación en formato GTF.
2. Par el mapeo sobre los archivos genómicos indexados se usaron los *reads* preprocesados con el modo de dos pasadas (*Two-pass mode*). El modo *two-pass* permitió detectar sitios de splicing en la primera pasada de mapeo y permitió generar una nueva referencia donde se incluyeron los sitios de splicing que no estaban contenidos el archivo genómico indexado. Luego los *reads* fueron re-mapeados a esta nueva referencia para mejorar la cuantificación de isoformas (Overbey et al., 2021).

Para la cuantificación de *reads*, se usó el software RSEM (Li y Dewey, 2011), utilizando los archivos BAM resultantes de la ejecución de STAR que solo pertenezcan a la especie *Coffea arabica*. RSEM tiene la capacidad de contar *reads*

que son mapeados a múltiples regiones del genoma e identificar isoformas mediante máxima verosimilitud.

Para el uso de RSEM también fueron necesarios pasos distintos:

1. El primer paso usó el genoma de referencia en formato FASTA y la anotación en formato GTF para crear archivos genómicos indexados.
2. EL segundo pasó usó los archivos indexados y los *reads* mapeados por STAR en formato BAM para asignar conteos a nivel de genes e isoformas por cada muestra (Overbey et al., 2021).

3.2.4. Análisis de Expresión Diferencial en Genes con los Datos de Mapeo de *Coffea arabica*

La expresión diferencial fue analizada mediante el paquete DESeq2 (Love et al., 2014) del lenguaje de programación R para lo cual se utilizaron los conteos asignados mediante el software RSEM. La elección de DESeq2 se basa en que el uso de estimadores de contracción (*shrinkage estimators*) mejoran la reproducibilidad del análisis en comparación con métodos basados en máxima verosimilitud. Además, DESeq2 al estar basado en el método empírico de Bayes, le da ventajas frente a programas como EdgeR y lima-voom debido a que aumenta la probabilidad de evitar los errores de tipo I (falsos negativos) (Love et al., 2017). Adicionalmente, debido a que usa estimadores de contracción le permite reportar LFC con un fundamento estadístico sólido que lo hace útil para comparar los LFC entre experimentos de RNA-Seq. Finalmente, ofrece la transformación rlog la cual sirve para facilitar el análisis exploratorio de datos (Love et al., 2017).

Durante la ejecución del flujograma de DESeq2, primero se realizó la importación de los archivos de conteo con la función tximport de Bioconductor ya que tienen implementada una forma fácil de agrupar los resultados de RSEM (Soneson et al., 2015). Luego, se colapsaron las réplicas técnicas con la función *collapseReplicates* la cual simplemente suma los conteos de las réplicas técnicas para considerar la suma como una réplica biológica (Love et al., 2017). Seguido a esto, se realizó la transformación de la matriz de conteos mediante la función rlog (logaritmo regularizado). Esta transformación es parecida a la transformación log2 para genes

con muchos conteos; aunque, con rlog, los genes con pocos conteos son ajustados para reducir su heterocedasticidad y por ende ajustar la varianza.

Luego de la transformación rlog, se realizó el análisis exploratorio de datos con visualizaciones multivariadas como el análisis de componentes principales (PCA) para buscar efectos de tipo *batch* o alguna estructura indeseada entre los datos debido a que los datos proceden de diferentes experimentos.

Debido a que se observaron ciertos patrones en los datos que correspondían a variables no biológicas, se usó la función *svaseq* del paquete SVA de Bioconductor para estimar las variables "ocultas". En este punto, se incluyeron las variables sustitutas (*surrogate variables*) al flujograma de DESeq2 para obtener el análisis de expresión diferencial de genes. Finalmente, se usó el modelo *apeglm* con la función *lfcshrink* de DESeq2 para ajustar los LFC que tienen bajos conteos.

3.2.5. Análisis Funcional de Genes

El análisis funcional tuvo tres pasos: La obtención de ontologías de los transcriptos, la construcción del objeto de anotación para *Coffea arabica* y el análisis de enriquecimiento funcional de *gene sets*.

3.2.5.a Obtención de Ontologías de los Transcriptos

El primer paso fue realizar la búsqueda de secuencias similares usando el software BLAST (*Basic Local Alignment Search Tool*) sobre el transcriptoma de *Coffea arabica* (Camacho et al., 2009), para ellos se tuvo en cuenta lo siguiente:

- Se utilizó el transcriptoma de *Coffea arabica* (GCA_003713225.1 Cara_1.0) disponible en la web de *Genome* del NCBI.
- Se utilizó la herramienta *Blastx* de la página web Galaxy implementada por el NCBI usando el algoritmo *blastx-fast* y la base de datos de secuencia de referencia de proteínas no redundantes del 2018 (RefSeq non-redundant proteins: nr).

- Con los resultados de *blastx-fast* se ejecutó el software Blast2GO para obtener las ontologías de las secuencias de referencia encontradas (Conesa y Götz, 2008), mediante el siguiente procedimiento:

    1. Se importaron los resultados de *blastx-fast*.
    2. Mapeo de las ontologías.
    3. Anotación de las ontologías.
    4. Exportación de las anotaciones en formato tabla.

3.2.5.b Construcción del objeto de anotación para *Coffea arabica*

La construcción del objeto de anotación de *Coffea arabica* se realizó con el paquete AnnotationDbi de Bioconductor.

Los pasos realizados fueron los siguientes:

- Se descargó de la página web *Datasets* del NCBI (https://www.ncbi.nlm.nih.gov/datasets/) la información de los genes de *Coffea arabica.*
- Se construyó el objeto de anotación SQLite utilizando la función *makeOrgPackage*.

3.2.5.c. Análisis de Enriquecimiento Funcional de Genes

El análisis de enriquecimiento funcional se realizó en RStudio con el paquete fgsea (Korotkevich et al., 2021) para lo cual se utilizaron los resultados del análisis de expresión diferencial obtenidos mediante DESeq2 y los resultados de la construcción del objeto de anotación.

3.2.6. Interpretación de Datos de RNA-Seq

Para facilitar la interpretación de los datos de RNA-seq, se ejecutó la función *Genetonic* del paquete GeneTonic de Bioconductor (Marini et al., 2021), el cual integra los resultados del análisis de expresión diferencial y el análisis de enriquecimiento funcional, para ello utilizó los siguientescuatro componentes generados previamente:

- La matriz de expresión de genes que se encuentra dentro del objeto generado por DESeq2.
- Los resultados del análisis de expresión diferencial de genes.
- Los resultados del análisis de enriquecimiento funcional de genes.
- Un objeto de anotación que cuenta con dos columnas (los ID de los genes y sus nombres) que se construyó con el paquete de AnnotationDbi de Bioconductor.

IV. RESULTADOS Y DISCUSIÓN

4.1. Resultados

4.1.1. Búsqueda, Selección y Descarga de Datos del SRA del NCBI

De acuerdo a los pasos seguidos en la sección 6.2.1 de la metodología, el resultado del filtrado de parámetros como tejido, número de réplicas biológicas y tratamiento del estudio fueron los siguientes bioproyectos: PRJNA630692, PRJEB5543, PRJNA606444 y PRJNA609253. A continuación, en la tabla 1 se presentan las principales características de estos experimentos de transcriptómica sobre hojas de *Coffea arabica.*

De los cuatro bioproyectos sólo se consideraron dos debido a las siguientes razones:

- El bioproyecto PRJEB5543 es un experimento realizado sobre genotipos silvestres de *Coffea arabica* con el objetivo de analizar la estructura genómica. En una comunicación por correo con el autor del trabajo (el doctor Philipe Lashermes), mencionó que los archivos fastq del secuenciamiento de RNA no eran ideales para estudios fisiológicos debido a que las plantas estuvieron en un invernadero (en Montpellier, Francia) con temperatura variable (18°C-30°C) que dependía de la temperatura externa, por esta razón no se usó este experimento en el análisis.

- El bioproyecto PRJNA606444, es un experimento cuyas réplicas biológicas están incluídas en el bioproyecto PRJNA630692 debido a que son trabajos secuenciales realizados por los mismos autores. Debido a esto, no es necesario tomarlo en cuenta ya que al analizar el bioproyecto PRJNA630692 ya se están analizando las muestras del bioproyecto PRJNA606444.

Finalmente, los bioproyectos selecionados fueron PRJNA630692 y PRJNA609253 debido a que contaban con las características como: condiciones controladas de temperatura, tenían un artículo científico asociado lo que permite revisar la metodología de la obtención de los datos de transcriptómica y tenían más de dos réplicas biológicas. Los metadatos de estos experimentos se encuentran en el anexo 1.

Tabla 1: Características de los bioproyectos

País del experimento	Número de muestras	Temperaturas analizadas	Especificación de la hebra	Tipo de secuen-ciamiento	Bioproyectos
Brasil	20	23°C y 30°C	*Unstranded*	Paired-end	PRJNA609253
Portugal	23	25°C	*Stranded*	Single-end	PRJNA606444
Francia	8	No especifica	*Unstranded*	Single-end	PRJEB5543
Portugal	72	25°C, 37°C y 42°C	*Stranded*	Single-end	PRJNA630692

4.1.2. Control de Calidad y Preprocesamiento de Archivos FASTQ

En el análisis de la calidad de los *reads* del bioprojecto PRJNA630692 se obtuvo en promedio 94.46% de reads mayores a Q30 con una longitud de 50 bp (Figura 1) y el porcentaje de GC en promedio fue de 45.68% (Tabla 2). Además, las estadísticas del filtrado de *reads* (Tabla 3, Figura 2) mostraron que los *reads* de baja calidad son un orden de 10 veces menor a los *reads* que pasaron el filtro.

Tabla 2: Estadísticas resumen de la calidad de *reads* del bioproyecto PRJNA630692

	Porcentaje de duplicación	Porcentaje > Q30*	Porcentaje de GC*	Porcentaje de *reads* que pasaron el filtro
Promedio	38.62%	94.46%	45.68%	96.64%
Mediana	39.25%	94.50%	45.30%	96.60%

*Q30: Porcentaje de bases con un valor de calidad de 30 a más.

*GC: Contenido de guanina-citosina de un *read.*

Tabla 3: Estadísticas resumen del filtrado de *reads* del bioproyecto PRJNA630692

	Pasaron filtro	Baja calidad	Muchos N*	Muy cortos*
Promedio	3.62×10^6	1.16×10^5	4.94×10^2	6.75×10^3
Mediana	3.86×10^6	1.16×10^5	3.87×10^2	5.95×10^3

*Muchos N: *Reads* que poseen más de cinco bases no identificadas.

*Muy cortos: *Reads* con una longitud menor a 15 pares de bases.

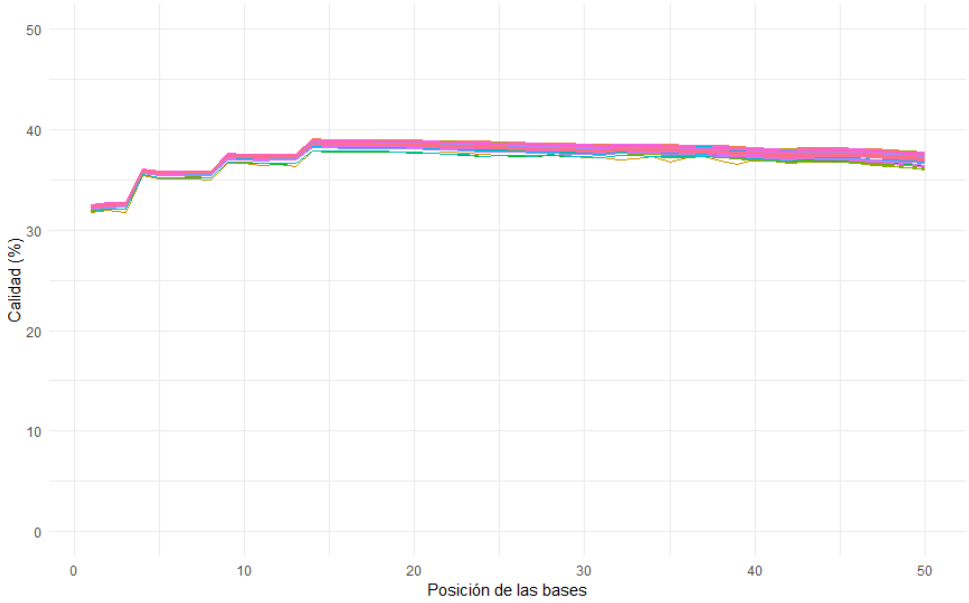

Figura 1: Gráfica del control de calidad de los *reads* del Bioproyecto PRJNA630692

Los colores de la figura representan a las 72 muestras que posee el bioproyecto.

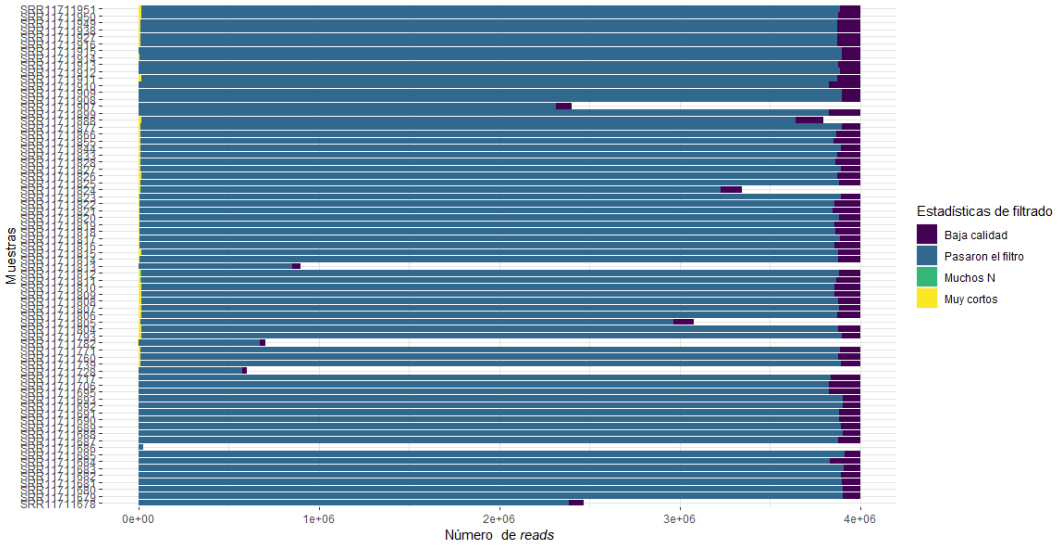

Figura 2: Gráfica donde se muestra la proporción de *reads* que pasaron el filtrado del bioproyecto PRJNA630692

Las figuras 1 y 2 y la tabla 2 mostraron que es un experimento de buena calidad y que es útil para seguir con el desarrollo de la metodología. Para más información véase el anexo 2 que incluye los datos de calidad, los datos de filtrado y los archivos de calidad (en formato HTML) de cada muestra por separado respectivamente.

En el análisis de la calidad de los *reads* del bioprojecto PRJNA609253, las muestras tuvieron en promedio 95.94% de reads mayores a Q30 con una longitud de 125 bp (Figura 3 y 4) y el porcentaje de GC en promedio fue de 45.23% (Tabla 4). Además, las estadísticas del filtrado de *reads* (Tabla 5, Figura 5) mostraron que de baja calidad son un orden de 10^3 veces menor a los *reads* que pasaron el filtro.

Tabla 4: Estadísticas resumen de la calidad de *reads* del bioproyecto PRJNA609253

	Porcentaje de duplicación	Porcentaje > Q30*	Porcentaje de GC*	Porcentaje de reads que pasaron el filtro
Promedio	11.94%	95.94%	45.23%	99.78%
Mediana	10.35%	95.95%	45.10%	99.80%

*Q30: Porcentaje de bases con un valor de calidad de 30 a más.

*GC: Contenido de guanina-citosina de un *read.*

Tabla 5: Estadísticas resumen de *reads* del bioproyecto PRJNA609253

	Pasaron filtro	Baja calidad	Muchos N*	Muy cortos*
Promedio	1.83×10^7	3.02×10^4	4.00×10^3	4.80×10^3
Mediana	1.87×10^7	2.88×10^4	3.99×10^3	0.00

*Muchos N: *Reads* que poseen más de cinco bases no identificadas.

*Muy cortos: *Reads* con una longitud menor a 15 pares de bases.

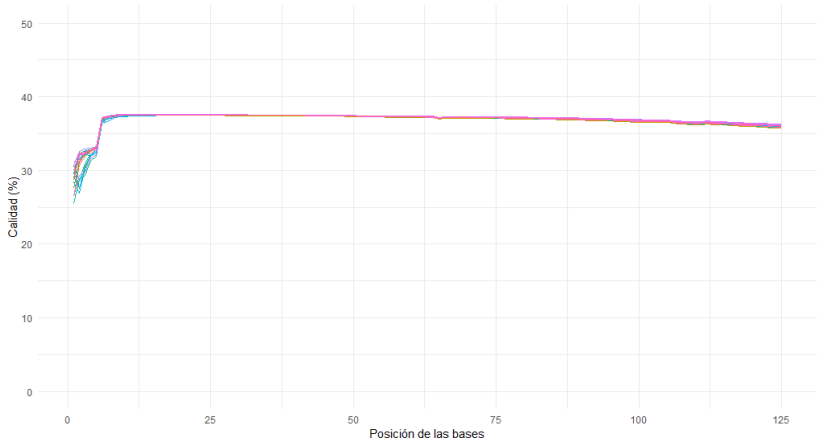

Figura 3: Gráfica del control de calidad de los *reads forward* del bioproyecto PRJNA609253

Los colores de la figura representan a las 20 muestras que posee el bioproyecto.

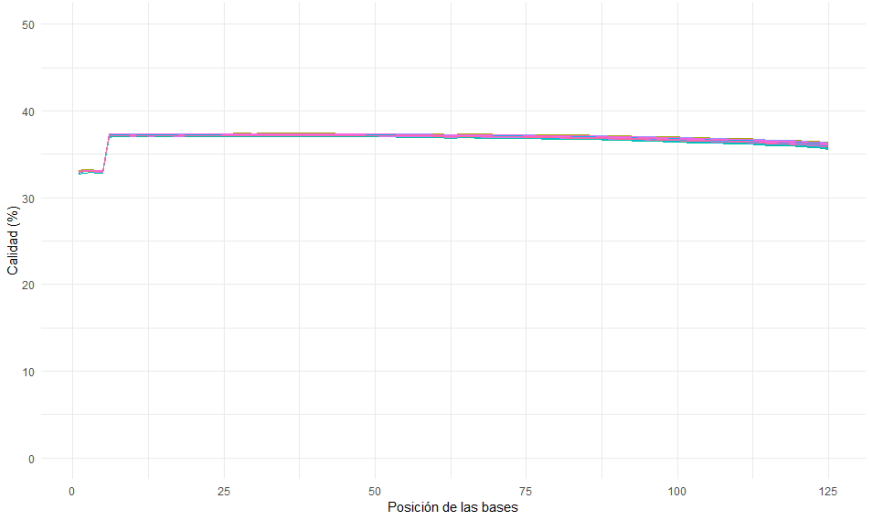

Figura 4: Gráfica del control de calidad de los *reads reverse* del bioproyecto PRJNA609253

Los colores de la figura representan a las 72 muestras que posee el bioproyecto.

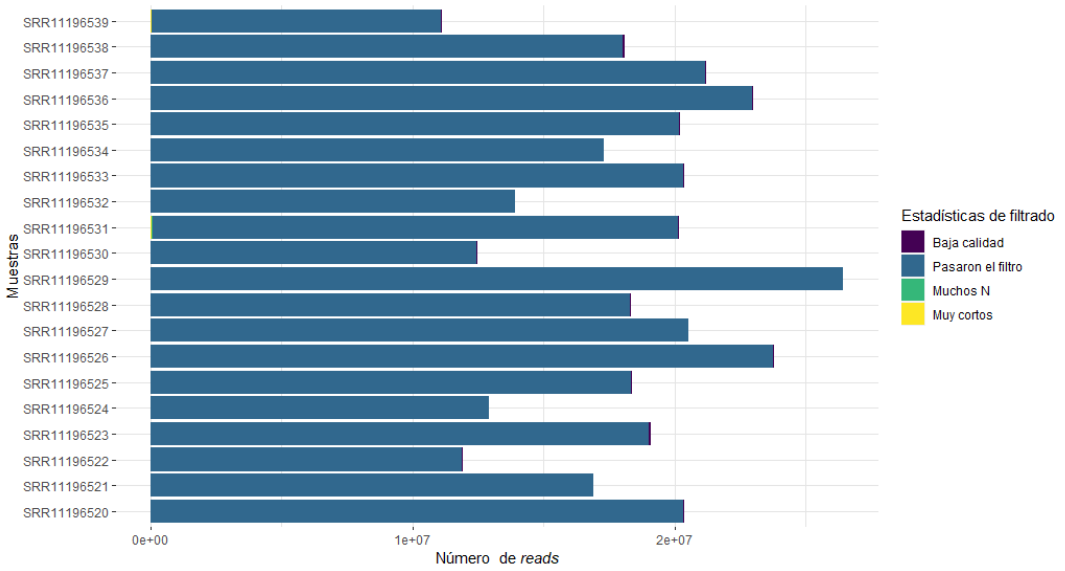

Figura 5: Gráfica donde se muestra la proporción de *reads* que pasaron el filtrado del bioproyecto PRJNA609253

Las figuras 3, 4 y 5 y la tabla 5 mostraron que es un experimento de buena calidad y que es útil para seguir con el desarrollo de la metodología. Para más información véase el anexo 3 que incluyen que incluyen los datos de calidad de los *reads forward* y *reverse*, los datos de filtrado y los archivos de calidad (en formato HTML) de cada muestra por separado respectivamente.

4.1.3. Procesamiento de Archivos FASTQ

4.1.3.a. Alineamiento

El alineamiento con el *software* STAR permitió comparar el porcentaje de *reads* mapeados a sitios únicos, el número total de sitios de splicing, el porcentaje de missmatch por base nucleotídica y el porcentaje de *reads* mapeados a múltiples loci (ver anexo 4).

Para cada una de estas variables se hizo un gráfico de Gardner-Altman (Ho et al., 2019) que permite observar la distribución de los datos, el tamaño de efecto (es la resta del promedio de dos variables) acompañado de intervalos de confianza y una distribución empírica realizada en base a 5000 permutaciones de los datos.

4.1.3.a.1. Porcentaje de mapeos únicos

El porcentaje de mapeos únicos es el porcentaje de reads que han sido mapeados a un único loci.

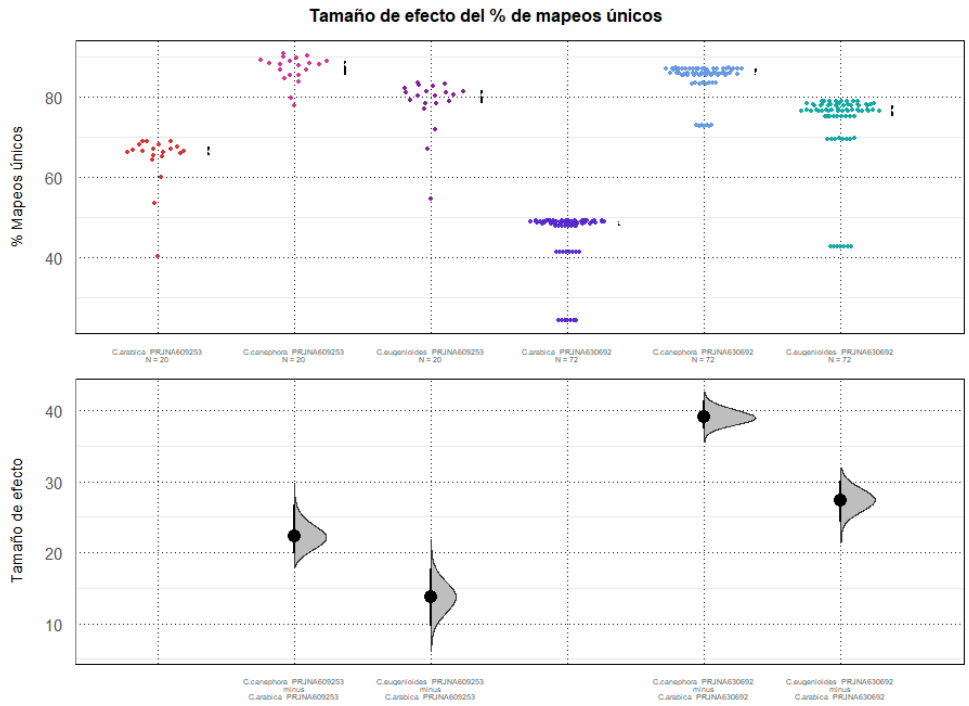

Figura 6: Gráfico Gardner-Altman que muestra del porcentaje de mapeos únicos sobre tres genomas de *Coffea* y los tamaños de efecto que muestran los bioproyectos PRJNA630692 y PRJNA609253

En la tabla 6 y en la figura 6 se puede ver que las especies *Coffea canephora* y *Coffea eugenioides* tienen un mayor porcentaje de mapeos únicos que *Coffea arabica* en ambos bioproyectos.

En ambos experimentos el mayor valor de diferencia lo presenta *Coffea* canephora utilizando a *Coffea arabica* como control (Tabla 6).

Tabla 6: Tamaño de efecto del porcentaje de mapeos únicos

Grupo control	Grupo de prueba	Diferencia	Intervalo de confianza inferior	Intervalo de confianza superior
C.arabica_PRJNA609253	C.canephora_PRJNA609253	22.35750	1.99×10^1	2.67×10^1
C.arabica_PRJNA609253	C.eugenioides_PRJNA609253	13.71350	9.58	1.78×10^1
C.arabica_PRJNA630692	C.canephora_PRJNA630692	39.07958	3.74×10^1	4.13×10^1
C.arabica_PRJNA630692	C.eugenioides_PRJNA630692	27.38583	2.43×10^1	3.01×10^1

4.1.3.a.2. Número total de sitios de *splicing*

El número total de *reads* mapeados a sitios de *splicing* es un parámetro de mapeo de STAR que cuenta el número de *reads* que son mapeados únicamente a un solo sitio de *splicing*. El conteo de los *reads* se hace sobre los sitios de *splicing* canónicos y no canónicos (Sibley et al., 2016) que presenta el genoma de una especie.

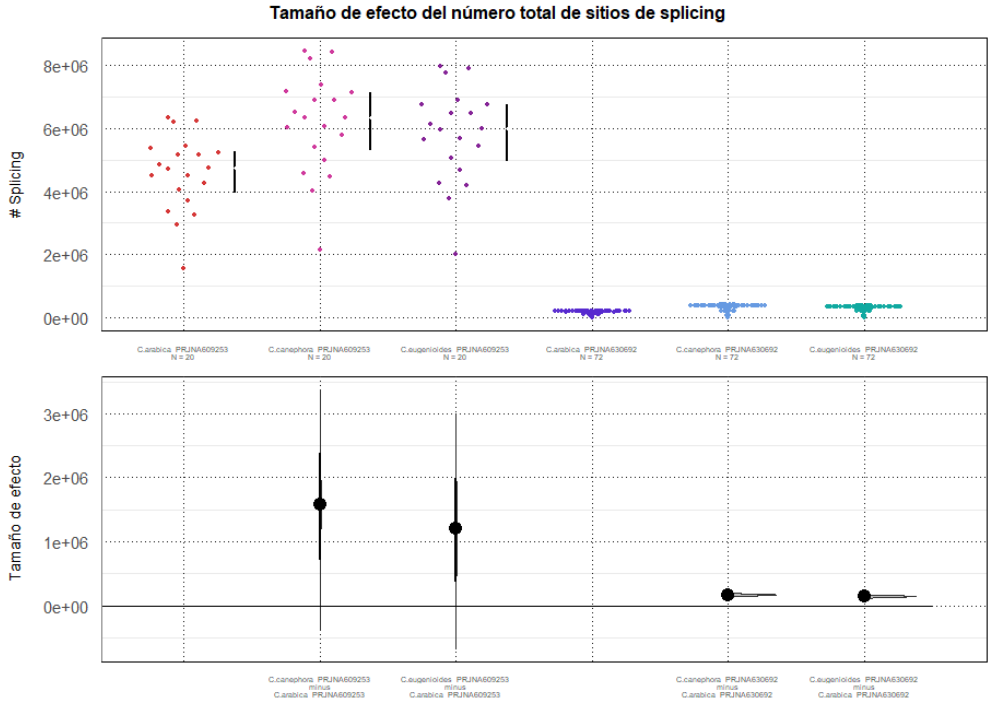

Figura 7: Gráfico Gardner-Altman del número total de reads mapeados a sitios de *splicing* sobre tres genomas de *Coffea* y los tamaños de efecto que muestran los proyectos PRJNA630692 y PRJNA609253

Se puede ver en la figura 7 que las especies *Coffea canephora* y *Coffea eugenioides* tienen un mayor número aparente de *reads* mapeados únicos a sitios de *splicing* que *Coffea arabica* en ambos bioproyectos. En la tabla 7 se puede observar los tamaños de efecto y sus intervalos de confianza.

Tabla 7: Tamaño de efecto del número total de sitios de *splicing*

Grupo control	Grupo de prueba	Diferencia	Intervalo de confianza inferior	Intervalo de confianza superior
C.arabica_PRJNA609253	C.canephora_PRJNA609253	1576218.4	7.09×10^5	2.38×10^6
C.arabica_PRJNA609253	C.eugenioides_PRJNA609253	1208992.7	3.76×10^5	2.00×10^6
C.arabica_PRJNA630692	C.canephora_PRJNA630692	167962.9	1.40×10^5	1.89×10^5
C.arabica_PRJNA630692	C.eugenioides_PRJNA630692	145509.7	1.20×10^5	1.65×10^5

4.1.3.a.3. Porcentaje del *ratio* de *missmatch* por base nucleotídica

El porcentaje del ratio de missmatch por base nucleotídica es un parámetro de mapeo de STAR que divide el número de missmatch en la posición de una base nucleotídica entre el número total de nucleótidos alineados a dicha posición.Se puede ver en la figura 8 que las especies *Coffea canephora* y *Coffea eugenioides* tienen un mayor porcentaje de *missmatch* por base nucleotídica en promedio que *Coffea arabica* para ambos bioproyectos y, que el porcentaje de *missmatch* por base nucleotídica es mayor en *Coffea canphora* que en *Coffea eugenioides*. Además, se puede observar que algunas muestras mapeadas al genoma de *Coffea canephora* presentan *outliers*. En la tabla 8 se puede observar los tamaños de efecto y los intervalos de confianza.

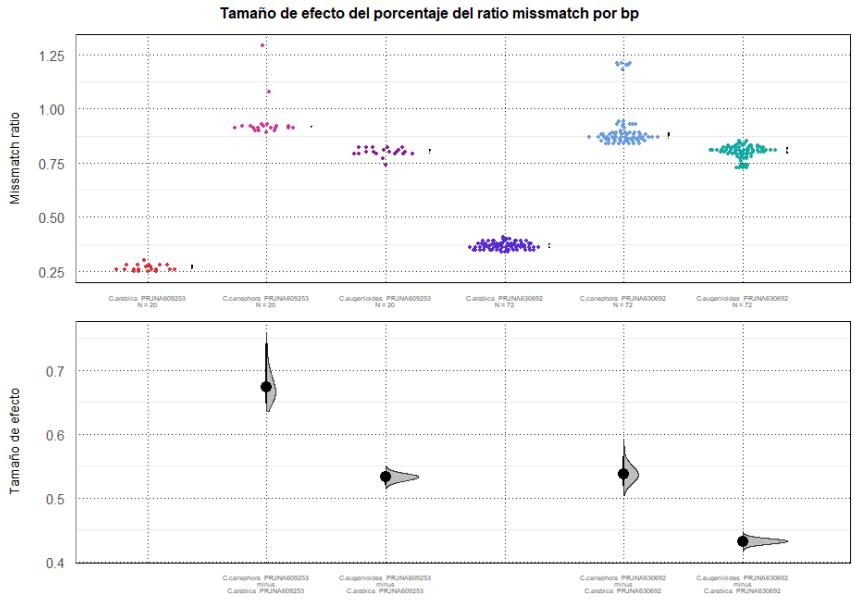

Figura 8: Gráfico Gardner-Altman del porcentaje del *ratio* de missmatch por base nucleotídica sobre tres genomas de *Coffea* y los tamaños de efecto que muestran los proyectos PRJNA630692 y PRJNA609253

Tabla 8: Tamaño de efecto del porcentaje del *ratio* de missmatch por base nucleotídica

Grupo control	Grupo de prueba	Diferencia	Intervalo de confianza inferior	Intervalo de confianza superior
C.arabica_PRJNA609253	C.canephora_PRJNA609253	0.6735000	6.48×10^{-1}	7.42×10^{-1}
C.arabica_PRJNA609253	C.eugenioides_PRJNA609253	0.5325000	5.20×10^{-1}	5.41×10^{-1}
C.arabica_PRJNA630692	C.canephora_PRJNA630692	0.5375000	5.18×10^{-1}	5.66×10^{-1}
C.arabica_PRJNA630692	C.eugenioides_PRJNA630692	0.4323611	4.24×10^{-1}	4.40×10^{-1}

4.1.3.a.4. Porcentaje de mapeos múltiples

El porcentaje de mapeos múltiples es el porcentaje de *reads* que han sido mapeados a más de un loci. En la figura 9, se puede ver que los mapeos realizados sobre los genomas *Coffea canephora* y *Coffea eugenioides* tienen un menor porcentaje de mapeos múltiples que sobre *Coffea arabica* en ambos bioproyectos. En la tabla 9 se puede observar los tamaños de efecto y los intervalos de confianza.

4.1.3.a.5 Análisis de componentes principales Biplot

Se hizo el análisis de componentes principales biplot (PCA-Biplot: *Biplot principal component analysis*) debido a que al tener tantas variables (multivariado) era necesario un método lineal de reducción de dimensiones para poder hacer un análisis exploratorio. Las variables resultantes del alineamiento de *reads*, mediante STAR, con las que se realizó el PCA-Biplot fueron las siguientes: porcentaje de mapeo único de *reads*, número total de sitios de *splicing*, número total de sitios de *splicing* no canónicos, *ratio* de *missmatch* por base nucleotídica, porcentaje de mapeo a múltiples loci, número de sitios de *splicing* GT/AG, número de sitios de *splicing* GC/AG, número de sitios de *splicing* AT/AC.

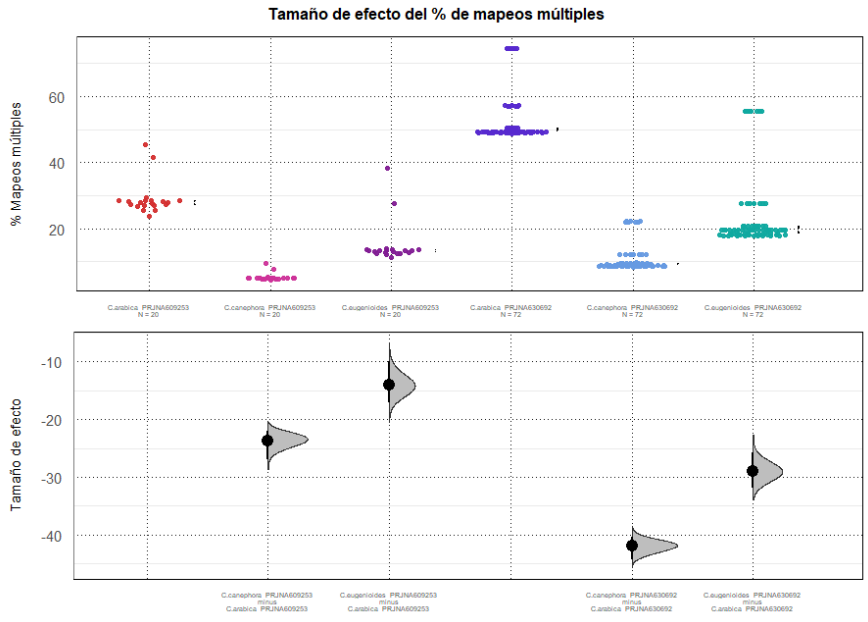

Figura 9: Gráfico Gardner-Altman del porcentaje de mapeos múltiples sobre tres genomas de *Coffea* y los tamaños de efecto que muestran los proyectos PRJNA630692 y PRJNA609253

Tabla 9: Tamaño de efecto del porcentaje de mapeos múltiples

Grupo control	Grupo de prueba	Diferencia	Intervalo de confianza inferior	Intervalo de confianza superior
C.arabica_PRJNA609253	C.canephora_PRJNA609253	-23.69150	-2.70×10^{1}	-2.21×10^{1}
C.arabica_PRJNA609253	C.eugenioides_PRJNA609253	-14.02750	-1.71×10^{1}	-9.91
C.arabica_PRJNA630692	C.canephora_PRJNA630692	-42.01653	-4.43×10^{1}	-4.03×10^{1}
C.arabica_PRJNA630692	C.eugenioides_PRJNA630692	-29.06042	-3.19×10^{1}	-2.58×10^{1}

La figura 10 muestra el agrupamiento de las muestras de acuerdo con las especies *Coffea arabica, Coffea canephora* y *Coffea eugenioides* en la dimensión dos y muestra agrupamiento según el bioproyecto en la dimensión uno. Además, también muestra la influencia que tuvieron las variables para formar cada componente principal. Las flechas opuestas indican existencia de una correlación negativa, mientras que las flechas con menor ángulo entre ellas indican una correlación positiva y la longitud de las flechas indica las calidades de representación de la variable en el gráfico. De acuerdo a esto, se observó que las variables porcentaje de mapeo único de reads, ratio de missmatch por base nucleotídica y porcentaje de mapeo a múltiples loci, influyen en el agrupamiento de las muestras en la dimensión dos; mientras que, las variables número total de sitios de splicing, número total de sitios de splicing no canónicos, número de sitios de splicing GT/AG, número de sitios de splicing GC/AG y número de sitios de splicing AT/AC, influyen en el agrupamiento de las muestras en la dimensión uno.

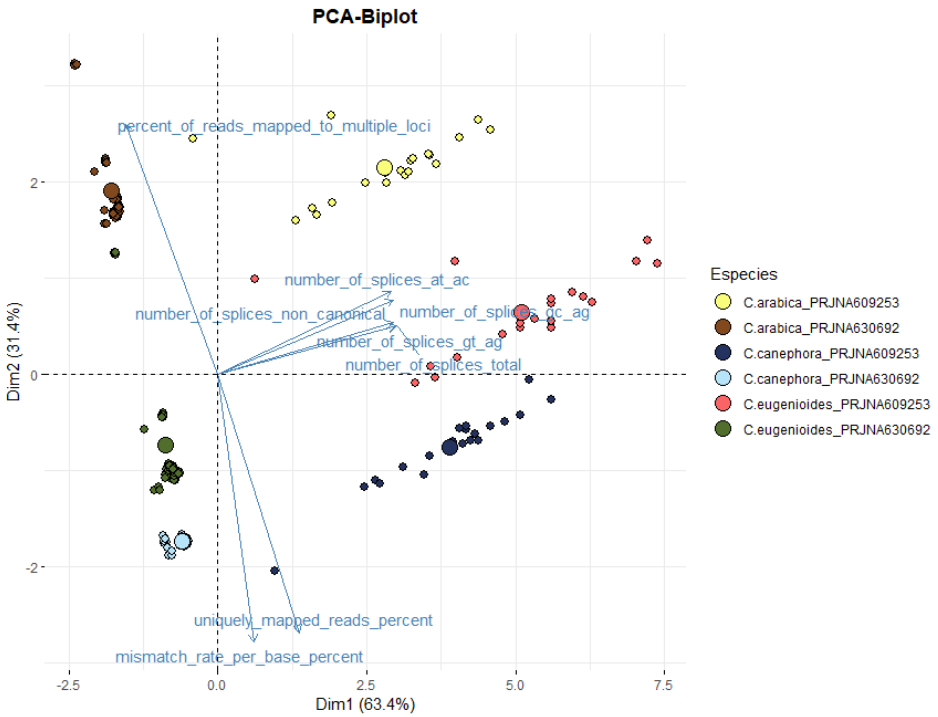

Figura 10: PCA-Biplot del mapeo sobre tres genomas de *Coffea* que muestran los bioproyectos PRJNA630692 y PRJNA609253

4.1.3.b. Conteo de *reads* con RSEM

El *software* RSEM generó para cada muestra dos archivos en formato tabular, uno contiene la tabla del conteo de *reads* por gen y el otro del conteo de *reads* por isoforma. La tabla de conteo de *reads* por gen (que es la de interés) cuenta con las columnas: longitud del gen, longitud efectiva del gen, conteo esperado, transcriptos por millón (TPM) y fragmentos por kilo base por millón (FPKM). (Li y Dewey, 2011, Starmer J., 2015). A continuación, en la tabla 10 se muestra las primeras 10 filas de la tabla de conteos de *reads* por gen de la muesta SRR11196520 (ver anexos 2 y 3).

4.1.4. Análisis de Expresión Diferencial en Genes

El análisis de expresión diferencial se hizo con las tablas de conteo de genes de *Coffea arabica* elaborados por el *software* RSEM. Primero se realizó un análisis exploratorio de los datos normalizados con la función *rlog* del paquete DESeq2 de Bioconductor con la finalidad de detectar alguna estructuración indeseada de los datos. Este análisis exploratorio se reportó por medio de mapas de calor y de análisis de componentes principales.

4.1.4.a. Análisis exploratorio de la matriz de conteos normalizada

Las figuras 11 y 12 muestran las gráficas de mapas de calor generados con las 29 muestras que incluyen a ambos experimentos seleccionados, después de haberse colapsado las réplicas técnicas, ya que sólo son de interés las réplicas biológicas (ver código en R en el anexo 5).

Se evidenció un agrupamiento (o estructura) entre los lugares de donde provienen las muestras. Además, se observaron agrupamientos entre las temperaturas 37°C - 42°C y 23°C - 30°C y, entre los cultivares Icatú y Acaua-Catuaí IAC 144 (estos cultivares pertenecen a los bioproyectos pero no son motivo de análisis en la tesis).

Tabla 10: Conteo de *reads* por gen de la muestra SRR11196520

ID* de genes	ID* de transcriptos	Longitud del transcripto	Longitud efectiva del transcripto	Conteo esperado	TPM*	FPKM*
CoarCr001	CoarCr001	1491	1291.92	6016.5	726.65	579.22
CoarCr002	CoarCr002	2810	2610.92	89.39	5.35	4.26
CoarCr003	CoarCr003	103	0	0	0	0
CoarCr004	CoarCr004	121	0	0	0	0
CoarCr005	CoarCr005	121	0	0	0	0
CoarCr006	CoarCr006	103	0	0	0	0
CoarCr007	CoarCr007	2810	2610.92	32033.61	1914.38	1525.98
CoarCr008	CoarCr008	1491	1291.92	6016.5	726.65	579.22

*ID: Se refiere al código de identificación en la base de datos del NCBI

*TPM: Transcriptos por millón

*FPKM: Fragmentos por kilobase por millón

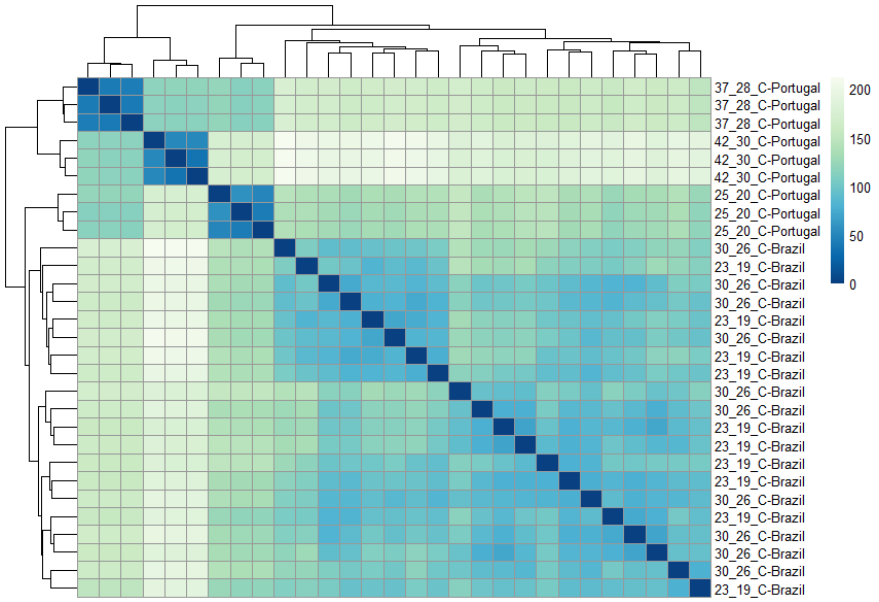

Figura 11: Mapa de calor de las muestras de los bioproyectos PRJNA630692 y PRJNA609253 (Temperatura-Lugar)

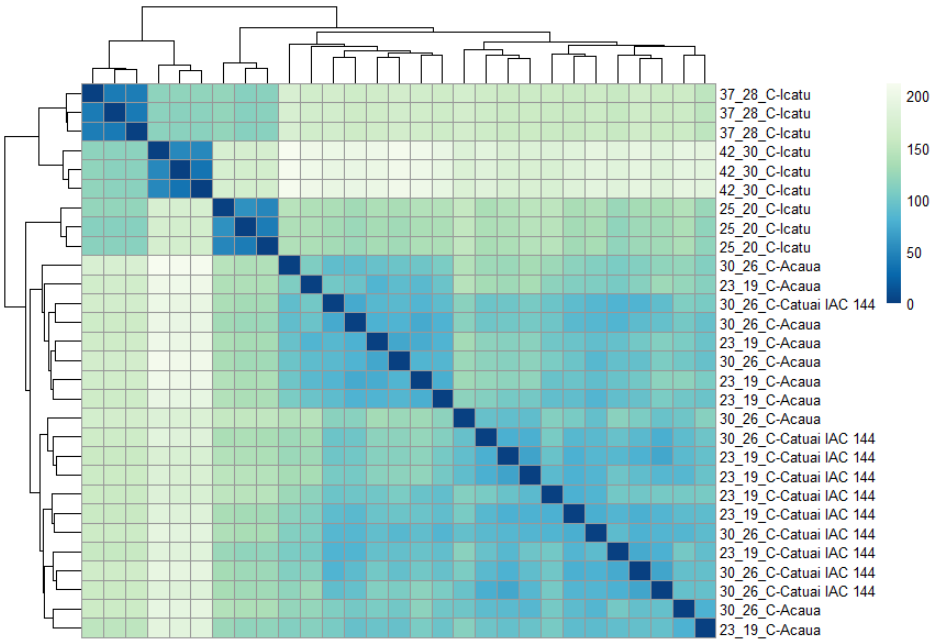

Figura 12: Mapa de calor de las muestras de los bioproyectos PRJNA630692 y PRJNA609253 (Temperatura-Cultivar)

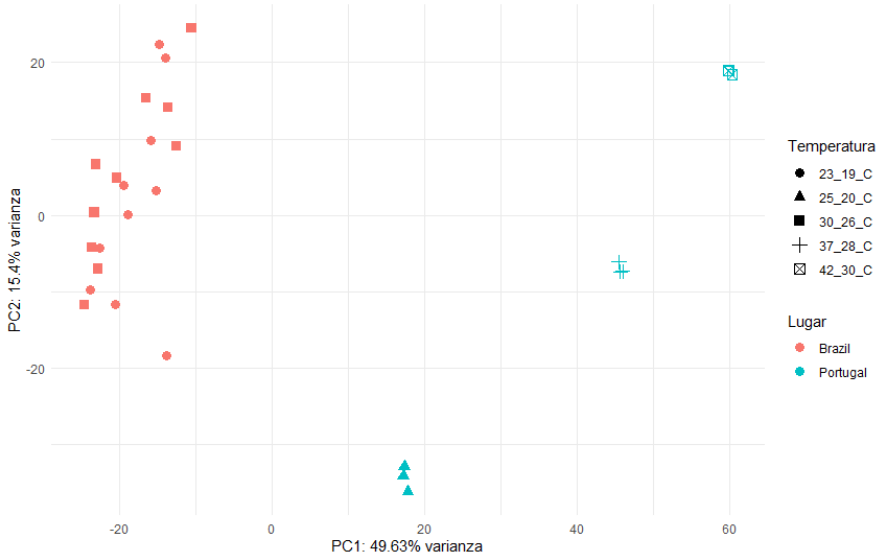

Figura 13: Análisis de componentes principales de las muestras de los bioproyectos PRJNA630692 y PRJNA609253 (Temperatura-Lugar)

En la figura 12, la componente uno explica el 49.63% de la varianza de las 29 muestras y, separa a las muestras que provienen de los dos lugares diferentes. Además, esta dimensión agrupa de manera similar al observado en el mapa de calor (Figura 10).

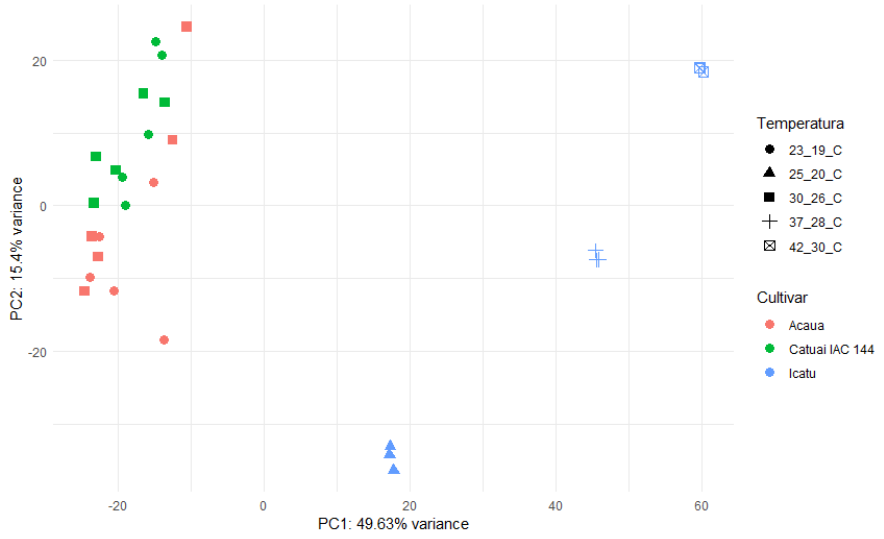

Figura 14: Análisis de componentes principales de las muestras de los bioproyectos PRJNA630692 y PRJNA609253 (Temperatura-Cultivar)

La figura 13 agrupa a los cultivares de acuerdo a su lugar de procedencia. Además, también se puede observar que los cultivares se agrupan de manera similar en el mapa de calor de la figura 11. En la componente dos (15.4% de la varianza explicada) se puede ver una ligera separación de los cultivares Acaua y Catuaí IAC 144.

Sin embargo, debido a que en el presente trabajo sólo es de interés la variable temperatura y no cultivares ni el lugar de realización del experimento, se realizó el análisis de variables sustitutas para estimar la estructura de las variables "Cultivar" y "Lugar" e incluirlas como efecto de tipo *batch* en el diseño experimental a ser considerado en el análisis con DESeq2.

4.1.4.b. Análisis de expresión diferencial de genes

El análisis de expresión diferencial de genes se hizo con el siguiente modelo:

$$\sum_{Batch} + Temperatura$$

Los efectos de tipo *batch* se calcularon con la función *svaseq* del paquete SVA de Bioconductor. Se estimaron dos variables sustitutas con el fin de captar las dos primeras dimesiones que "encapsulan" la mayor variabilidad correspondiente a los patrones que no pertenecen a la variable "Temperatura". Luego, se procedió a calcular la expresión diferencial con la función *DESeq*.

La tabla de resultados se estimó con la función *lfcShrink* especificando el modelo "apeglm" como modelo *a priori* para mejorar la estimación de los LFC. Esta tabla de resultados fue elaborada para los contrastes de la tabla 11 (ver anexo 6 para revisar las muestras por temperatura):

Tabla 11: Especificación de contrastes para la expresión diferencial de genes con los bioproyectos PRJNA630692 y PRJNA609253

Contrastes
25/20 vs 23/20*
30/26 vs 23/20*
37/28 vs 23/19*
42/30 vs 23/19*

*Los valores numéricos antes y después del símbolo "/" se refieren a las temperaturas de día y de noche respectivamente.

Para cada uno de los contrastes se elaboró el gráfico Bland-Altman (MA *plot*) que sitúan en el eje *x* al promedio de los conteos normalizados de los genes en la escala log10 y en el eje *y*, al respectivo LFC de ese gen. En los gráficos de las figuras 15, 16, 17 y 18 los puntos rojos representan a los genes diferencialmente expresados y los azules oscuros el caso contrario.

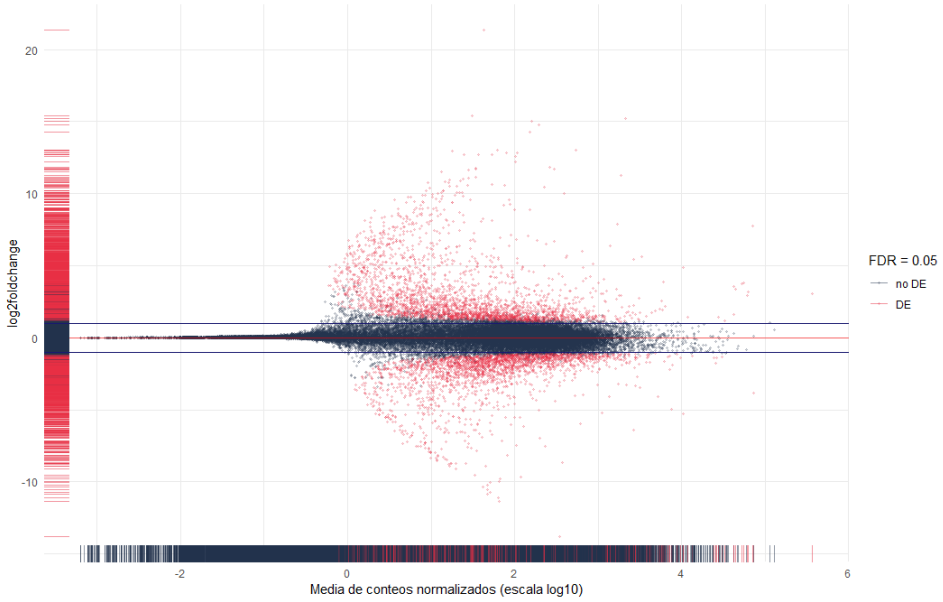

Figura 15: Gráfico Bland-Altman del contraste 25°C vs 23°C
FDR de 0.05 y límite de LFC ±1

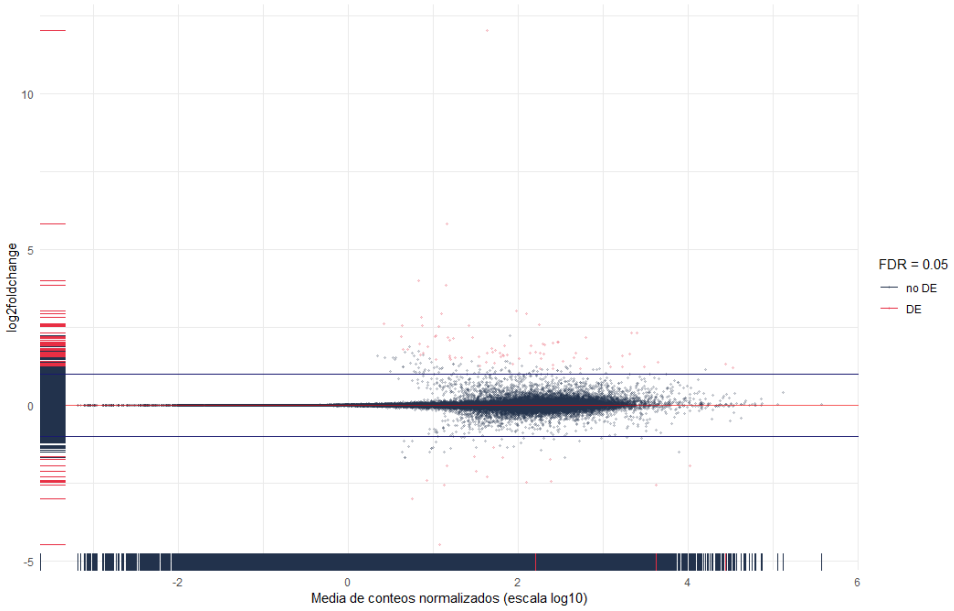

Figura 16: Gráfico Bland-Altman del contraste 30°C vs 23°C
FDR de 0.05 y límite de LFC ± 1

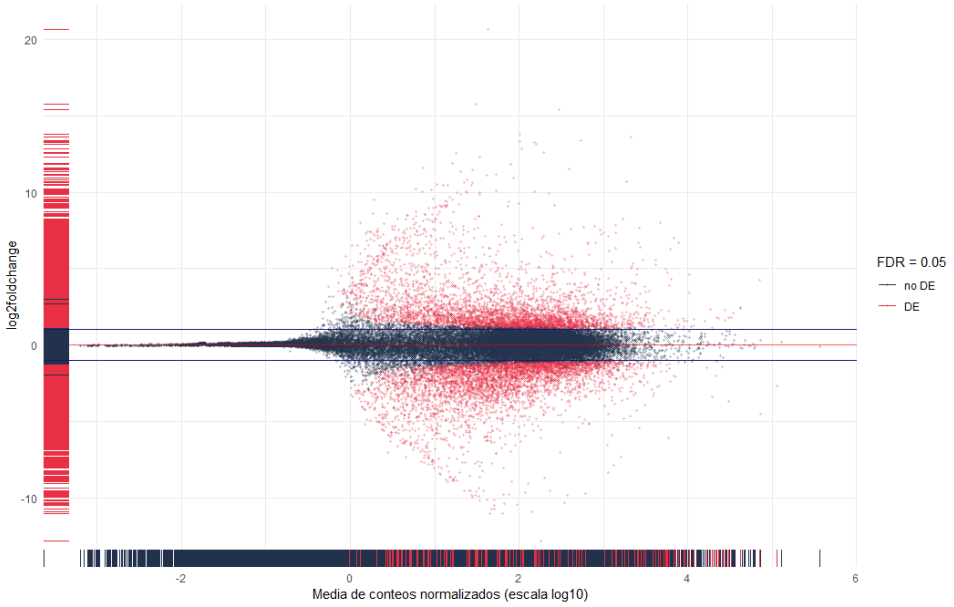

Figura 17: Gráfico Bland-Altman del contraste 37°C vs 23°C
FDR de 0.05 y límite de LFC ± 1

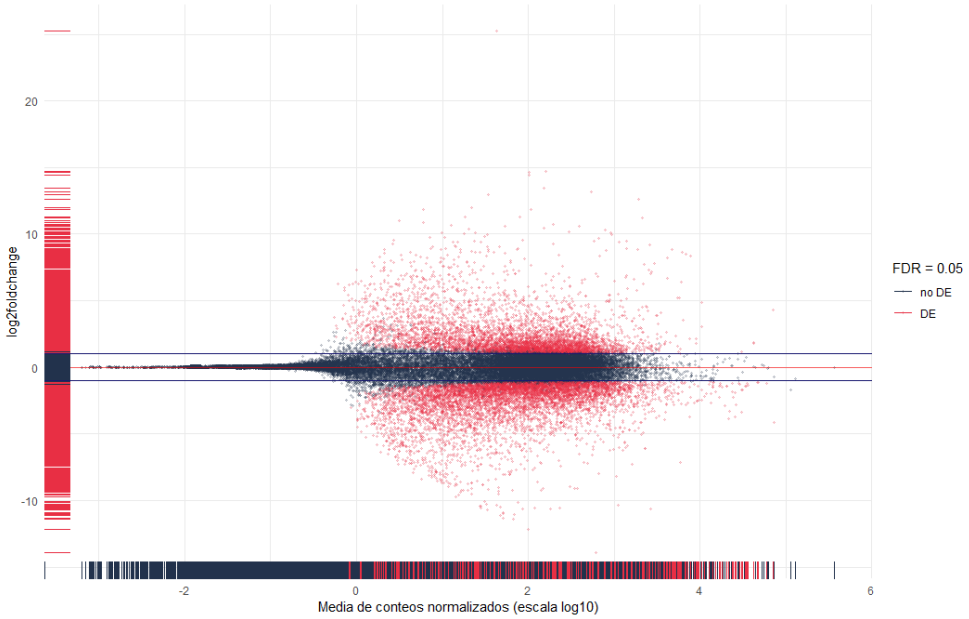

Figura 18: Gráfico Bland-Altman del contraste 42°C vs 23°C
FDR de 0.05 y límite de LFC ± 1

4.1.5. Análisis Funcional de Genes

El análisis de funcional de genes constó de tres componentes. El primero, Obtención de Ontologías de los Transcriptos, que generó una tabla proveniente del *software* Blast2GO que relaciona ontologías, transcriptos y probables funciones de los transcriptos. El segundo, construcción del objeto de anotación para *Coffea arabica*, que relaciona genes, transcriptos, ontologías, el código ENTREZ de los genes y a qué cromosoma pertenecen; construido a partir de la tabla de anotación proveniente de Blast2GO y de la tabla con información sobre los genes de *Coffea arabica* descargada de la basa de datos *Datasets* del NCBI. El tercero, análisis de enriquecimiento funcional de genes, obtenido con la función *fgsea* del paquete fgsea de Bioconductor; el cual usó los LFC de la expresión diferencial y el objeto de anotación construido para *Coffea arabica*.

4.1.5.a. Obtención de Ontologías de los Transcriptos

A continuación, en la tabla 12 se muestra las 10 primeras filas de una tabla de 80667 filas que es resultado del uso de BLAST2GO. La tabla muestra los códigos de identificación de los transcriptos, la descripción asociada resultante del uso de *blastx-fast*, la longitud del transcripto en pares de bases, el código de identificación de la ontología del gen (GO: *gene ontology*) en caso tuviese, los nombres de las ontologías identificados en tres grupos (P: proceso celular, F: función molecular y C: componente celular) y la media de similitud (obtenida en la etapa de "blasteo") con proteínas de la base de datos de proteínas no redundantes del 2018 (por ejem: *Coffea canephora, Arabidopsis thaliana, Oryza sativa, etc*).

4.1.5.b. Construcción del objeto de anotación para *Coffea arabica*

La tabla 13 muestra el resultado de la unión de la tabla 12 resultante del uso de BLAST2GO y la tabla de genes descargada de *Datasets* del NCBI. La tabla 13 muestra las cuatro primeras filas de una tabla de 2190684 filas.

Esta tabla muestra el código de los genes (ENTREZ ID o GID), el cromosoma al que pertenecen, la forma en la que han sido descubiertos (EVIDENCE), el nombre del gen, el tipo de gen, el símbolo del gen y las ontologías a las que pertenecen (ver anexo 7 correspondiente al código R para construir el objeto de anotación).

Tabla 12: Anotación de transcriptos con BLAST2GO

Media de similitud (%)	Nombres de GO*	GO* ID*	Longitud	Descripción	ID* Transcripto
94.85	NA	NA	1198	, AMMECR1	XM_027255494.1
94.85	NA	NA	1396	, AMMECR1	XM_027248934.1
89.09	P: tRNA modification; F: tRNA-uridine aminocarboxypropyltransferase activity	P: GO:0006400; F: GO:0016432	1362	, contains DTW domain	XM_027206459.1
63.87	P: tRNA modification; F: tRNA-uridine aminocarboxypropyltransferase activity	P: GO:0006400; F:GO:0016432	1500	, contains DTW domain	XM_027208424.1

ID: Código de identificación.

GO: Código de identificación de la ontología.

|Tabla 13: Anotación de genes *Coffea arabica*

Símbolo	Todas las ontologías	Ontología	Todos los GO*	GO*	Tipo de gen	Nombre del gen NCBI	Evidencia	Cromosomas	GID*
LOC113688587	BP*	MF*	GO:0006139	GO:0016432	PROTEIN_CODING	uncharacterized LOC113688587	IEA*	5e	113688587
LOC113688587	BP *	MF *	GO:0006400	GO:0016432	PROTEIN_CODING	uncharacterized LOC113688587	IEA *	5e	113688587
LOC113725648	NA	NA	NA	NA	PROTEIN_CODING	uncharacterized protein At2g38710-like	NA	2c	113725648
LOC113730656	NA	NA	NA	NA	PROTEIN_CODING	uncharacterized protein At2g38710-like	NA	2e	113730656

*Evidencia: Forma en la que se descubrió el gen.
*GO: Código de identificación de la ontología.
*GID: Código de identificación del gen en la base de datos del NCBI.
*IEA: Inferido desde anotación electrónica.
*BP: Proceso Biológico.
*MF: Función molecular.

4.1.5.c. Análisis de Enriquecimiento Funcional de Genes

El análisis funcional de genes se realizó utilizando los contrastes especificados en la tabla 11, los cuales son los mismos de la etapa de expresión diferencial de genes (ver sección 7.4.2.). Las figuras 19, 20, 21 y 22 muestran los diez primeros *genesets* más significativos con un puntaje normalizado de enriquecimiento (*normalized enrichment score*: NES) mayor a cero y los diez primeros *genesets* más significativos con un NES menor a cero (ver anexo 5).

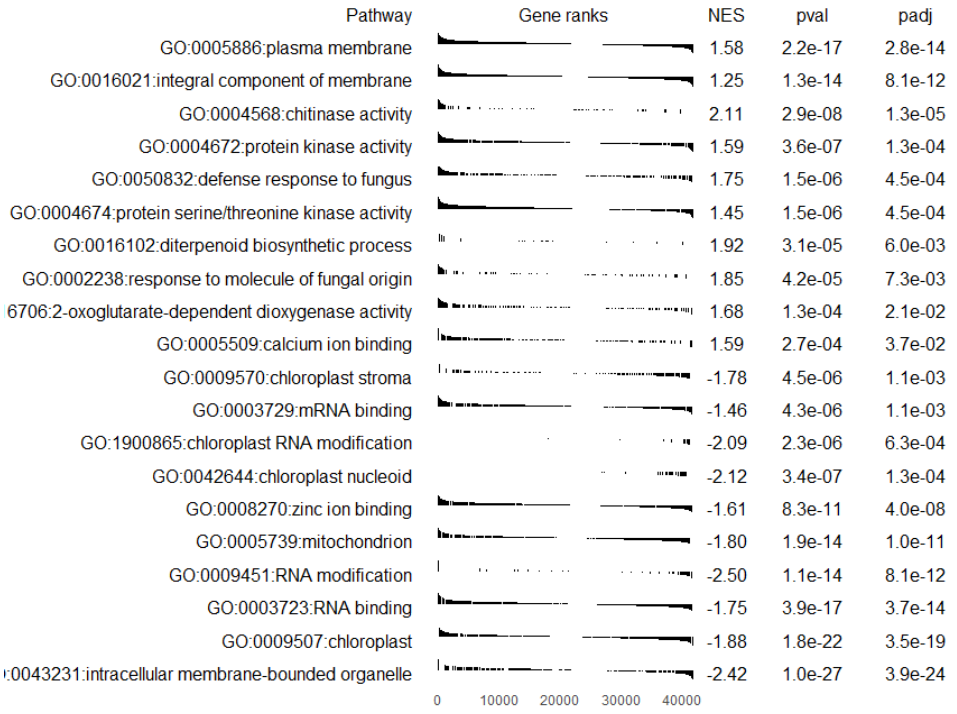

Figura 19: Genesets más significativos hallados en el análisis funcional del contraste 25°C vs 23°C

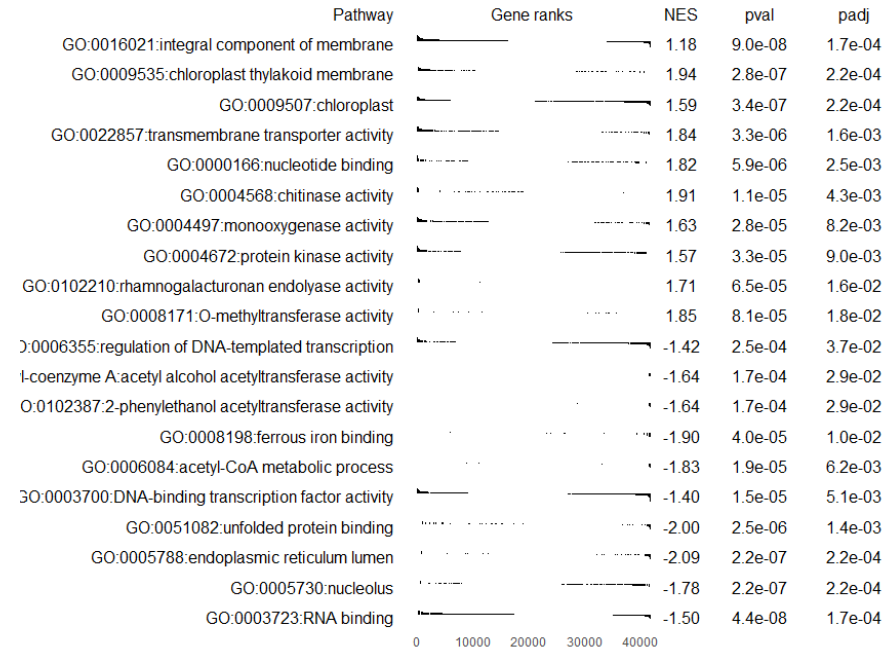

Figura 20: Genesets más significativos hallados en el análisis funcional del contraste 30°C vs 23°C

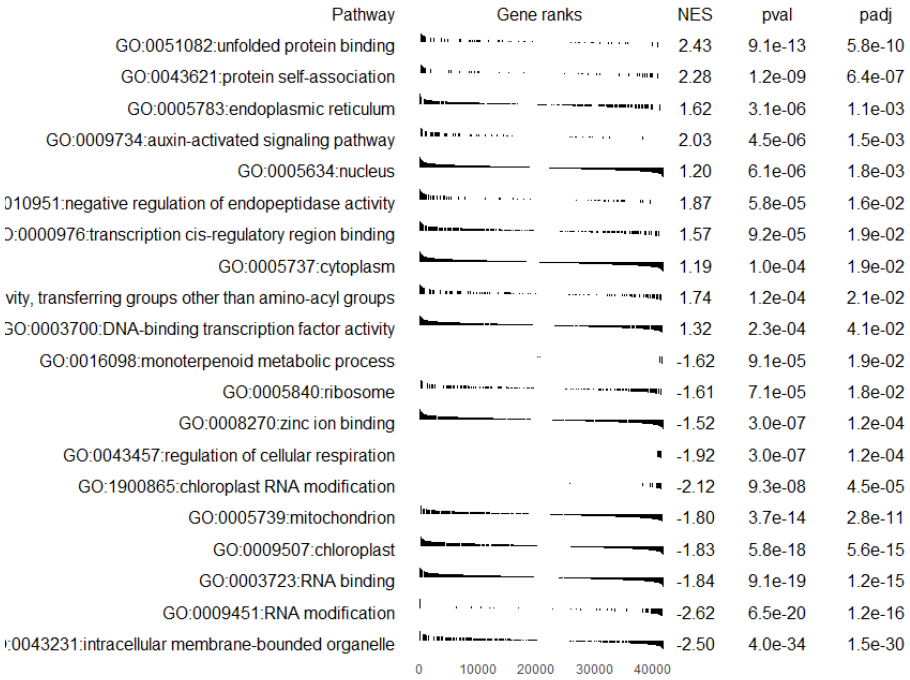

Figura 21: Genesets más significativos hallados en el análisis funcional del contraste 37°C vs 23°C

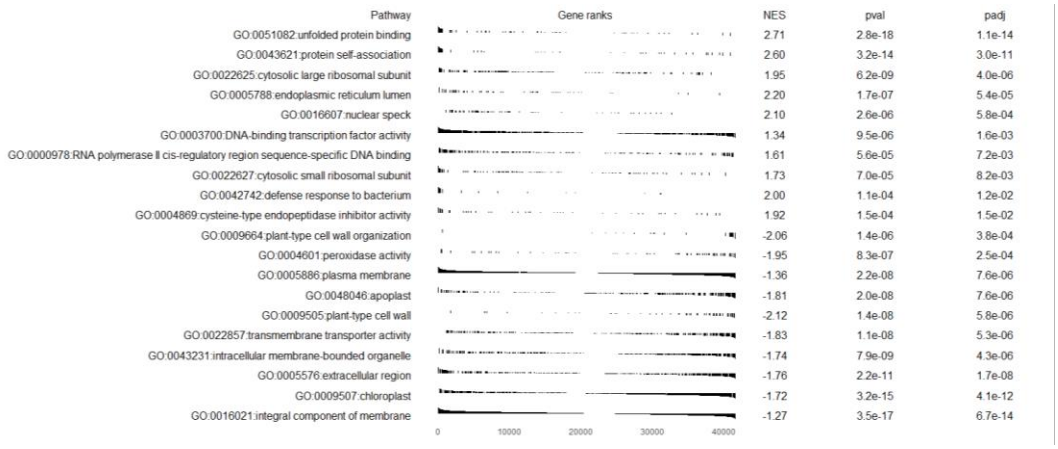

Figura 22: Genesets más significativos hallados en el análisis funcional del contraste 42°C vs 23°C

4.1.5.d. Interpretación de Datos de RNA-Seq

La función *Genetonic* del paquete Genetonic de Bioconductor, tuvo la finalidad de permitir comprender la inmensa cantidad de información de la expresión diferencial de genes; el enriquecimiento funcional de genes; y la anotación de los genes y sus ontologías (ver código R en el anexo 8). Para ello se generó mapas de calor de genes y *genesets* coloreados con los LFC de los genes (figuras 23, 24, 25 y 26) para cada contraste, así como, un gráfico de tela de araña con los valores NES de cada *geneset* provenientes del enriquecimiento funcional de los contrastes. Se hizo el gráfico de tela de araña solo para los contrastes de altas temperaturas (37°C vs 23°C y 42°C vs 37°C) (figura 27), debido a que las ontologías mostradas para los contrastes 30°C vs 23°C y 25°C vs 23°C (figuras 19, 20, 23 y 24) no muestran que esté ocurriendo una respuesta concertada al aumento de temperatura; en lugar a ello, mostraron procesos relacionados a la defensa contra hongos y del normal funcionamiento molecular al interior de la hoja.

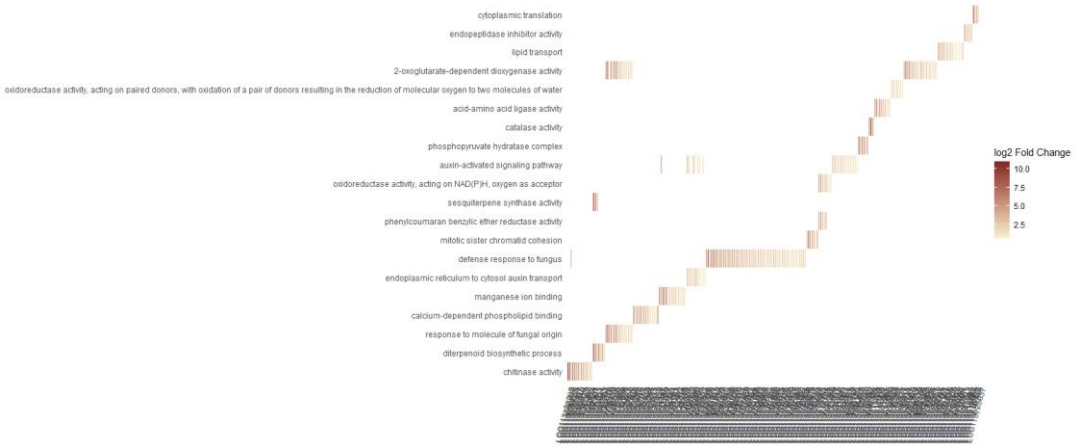

Figura 23: Mapa de calor genes-*genesets* del contraste 25°C vs 23°C

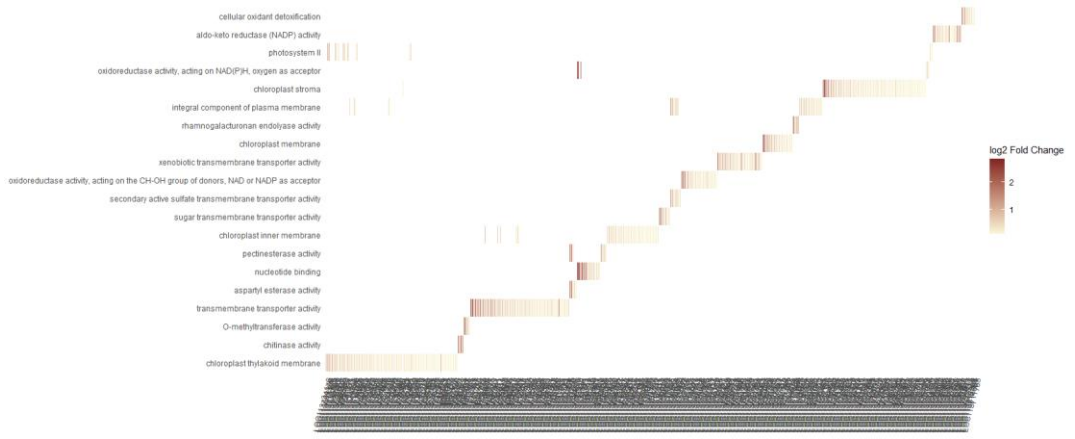

Figura 24: Mapa de calor genes-*genesets* del contraste 30°C vs 23°C

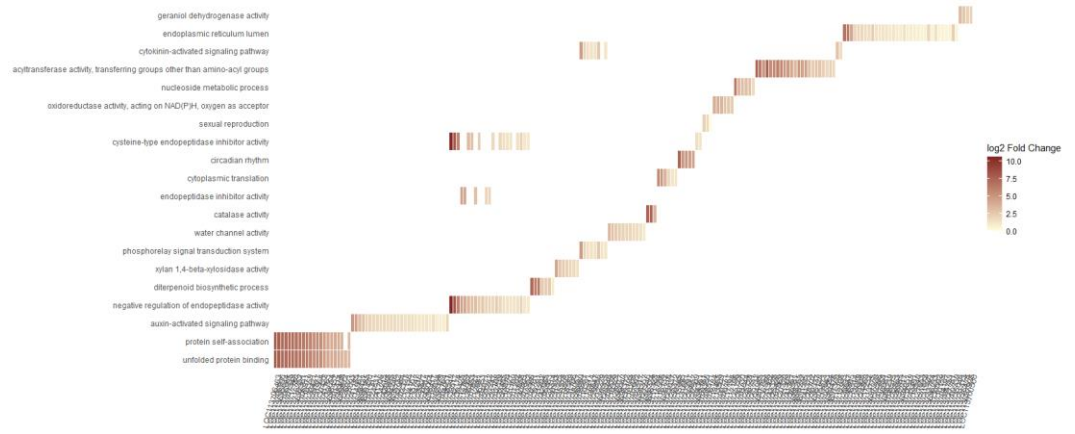

Figura 25: Mapa de calor genes-*genesets* del contraste 37°C vs 23°C

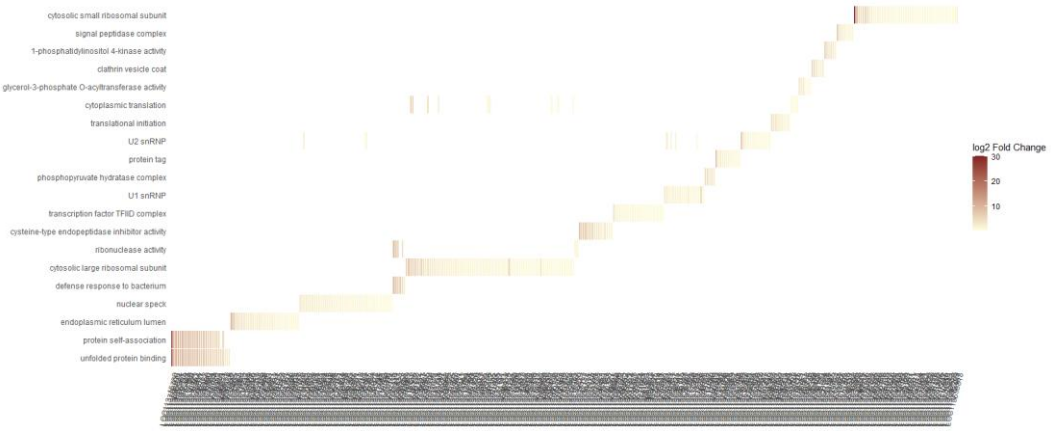

Figura 26: Mapa de calor genes-*genesets* del contraste 42°C vs 23°C

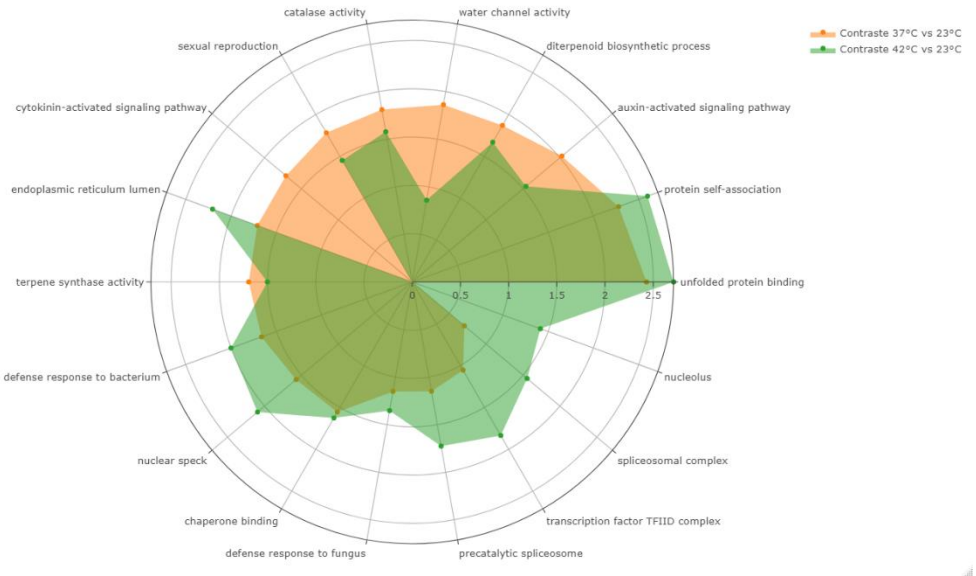

Figura 27: Gráfico de radar que relacionas los *genesets* de los contrastes 37°C vs 23°C y 42°C vs 23°C

4.2. DISCUSIÓN

4.2.1. Control de Calidad

Los *reads* de los bioprojectos PRJNA630692 y PRJNA609253 generaron en promedio 94.46% y 95.95% (ver sección: 7.2. Control de Calidad y Preprocesamiento de Archivos FASTQ) de *reads* respectivamente con calidad mayor a Q30. Estos valores son similares a los obtenidos por Mansilla (2021), quien obtuvo en promedio 94.22% de *reads* con calidad mayor a Q30, en secuencias obtenidas tambien de muestras de hojas de *C. arabica*. Sin embargo, en el trabajo de Marques et al. (2021) reportan 92.85% de *reads* con calidad mayor a Q30 para el bioproyecto PRJNA630692, que es menor a lo estimado. Esto se debería a que ellos usaron el programa *Trimmomatic* para el control de calidad y especificaron parámetros más conservadores. Por otro lado, Budzinski et al. (2021), quienes secuenciaron hojas de *Coffea arabica* para analizar el efecto del ácido hexanoico en la respuesta al estrés, obtuvieron 77.12% de *reads* con calidad mayor a Q30; lo cual posiblemente se debe al uso del programa llamado *Cutadapt* y un script personalizado en PERL (lenguaje de programación) que tendría parámetros más conservadores incluso a los usados por Marques et al. (2021); otra posible explicación es la ocurrencia de algún problema técnico durante el proceso de obtención de las secuencias.

Respecto a los contenidos de guanina y citosina (GC) de los *reads*, los bioproyectos PRJNA630692 y PRJNA609253 contienen 45.68% y 45.23%, respectivamente. Ivamoto et al. (2017), analizaron la diferencia en el transcriptoma de hojas, flores y el perisperma del fruto de *Coffea arabica* obteniendo en promedio para esos órganos 41.23% de contenido de GC un valor menor a los obtenidos posiblemente porque este valor proviene de diferentes órganos de la planta. Valores cercanos a los obtenidos son los de Mansilla (2021), quien reporta un rango entre 45.1% y 49.3%.

Al parecer, valores cercanos a 45% del contenido de GC serían caracterísiticos para *Coffea arabica*.

4.2.2 Alineamiento y Mapeo con STAR

El alineamiento de los *reads* con STAR produjo 45.57% y 64.57% de mapeos únicos frente al genoma de referencia de *Coffea arabica* para los bioproyectos PRJNA630692 y PRJNA609253, respectivamente (Figura 6). Respecto a los mapeos únicos sobre el genoma de *Coffea canephora,* STAR produjo 84.65% y 86.93% para los bioproyectos PRJNA630692 y PRJNA609253, respectivamente (Figura 6). Y, respecto a los mapeos únicos sobre el genoma de *Coffea eugenioides*, STAR produjo 72.96% y 78.29% para los bioproyectos PRJNA630692 y PRJNA609253 respectivamente (Figura 6).

Así mismo, Budzinski et al. (2021) y Marques et al. (2021) obtuvieron resultados similares analizando el transcriptoma de *Coffea arabica*. Marques et al. (2021) realizaron el mapeo de los *reads* sobre genoma de *Coffea canephora* y obtuvieron 81.49% de mapeos únicos, valor que es cercano al que se obtuvo para el bioproyecto PRJNA630692 (bioproyecto de su autoria). Considerando que también ellos utilizaron el programa STAR para el mapeo, la pequeña diferencia en los resultados se debe a que ellos usaron los párametros por defecto, a diferencia del presente trabajo, en el que se utilizaron los parámetros recomendados por Overbey et al. (2021) en el modo *two-pass mode* (dos pasadas de mapeo), que permite mejorar la cantidad de *reads* mapeados e identificar mejor los sitios de *splicing* de los *reads* (Dobin, 2013). En el mapeo de *reads* sobre el genoma de *Coffea arabica* se obtuvo un rango similar (55.18% - 63.35%) al de Budzinski et al. (2021), quienes utilizaron HISAT2 sobre cuatro variedades de café. De acuerdo con estos resultados, es claro que los porcentajes de mapeos únicos sobre los genomas de referencia de las especies diploides (*Coffea canephora* y *Coffea eugenioides*) son mayores a los de *Coffea arabica* debido a que este último es una especie alotetraploide cuyo genoma está conformado por subgenomas provenientes de las dos primeras especies diploides y, en consecuencia, los *reads* de *Coffea arabica* son mapeados a sus dos subgenomas diploides que presentan similitudes entre ellas. Estos resultados tienen consonancia con el trabajo de Scalabrin et al. (2021), que secuenció la variedad Borbón de *C. arabica,* y encontró que los dos subgenomas de *Coffea arabica* tienen un gran parecido, los niveles de diversidad entre subgenomas es alto y *Coffea eugenioides* contribuye ligeramente con más genes (22888 genes) al genoma de *Coffea arabica*

que *Coffea canephora* (21254 genes). De acuerdo con lo reportado por Scalabrin et al. (2021), y a los resultados obtenidos de mapeos únicos (figura 6) se diría que a pesar de que *C. eugenioides* aporta más al genoma del tetraploide, *C. canephora* tiene mayor aporte al genoma de *Coffea arabica* para la respuesta a la alta temperatura del aire.

Teniendo en cuenta los porcentajes de mapeos, se observa que los mapeos múltiples son contrarios a los mapeos únicos; es decir, sobre las especies que se tuvieron mayor porcentaje de mapeos únicos se obtuvieron menor porcentaje de mapeos múltiples y viceversa (figuras 6 y 9). Esto es de esperar debido a que el genoma de *Coffea arabica* es alopoliploide y se cree que se formó por la hibridización de *Coffea canephora* y *Coffea eugenioides* (Scalabrin et al., 2020).

Respecto al número total de sitios de splicing (figura 7), se observó que el mapeo sobre *Coffea arabica* presentó menor número de *reads* mapeados únicos a los sitios de splicing que las otras dos especies. Esto se explica de la misma forma que el porcentaje de mapeos únicos de *reads*, ya que *C. arabica* al tener un genoma constituído por dos subgenomas de especies cercanas, los *reads* tienden a ser mapeados a múltiples sitios de *splicing* y no a un solo locus (Scalabrin et al., 2021). Por lo tanto, al ser un alotetraploide, tendría aproximadamente el doble de sitios de *splicing* que las especies diploides que constituyen su genoma.

La importancia de cuantificar a los *reads* mapeados a los sitios de *splicing* tiene que ver directamente con la cuantificación de las isoformas de un determinado gen, ya que a mayor cantidad de *reads* mapeados a sitios de *splicing*, mayor será el número de las isoformas que se pueden detectar en la condición evaluada (por ejemplo: respuesta al estrés); Además, es indicio de diversidad de proteínas y/o regulación molecular de una especie y, en consecuencia, de las herramientas biológicas para responder a ciertos estímulos (Sibley et al., 2016).

Los sitios de *splicing* se pueden dividir en canónicos y no canónicos. El *splicing* canónico es el proceso de reconocimiento y extracción de las secuencias intrónicas contenidas en un mRNA inmaduro. Se le dice canónico debido a que el espliceosoma reconoce secuencias motivo conservadas (GT/AG, GC/AG y AT/AC) en los extremos 5´ y ´3 de los intrones para realizar el proceso de *splicing* (Frey y Pucker,

2020). El *splicing no canónico* no sigue las reglas del proceso normal de *splicing* (retiro de intrones y unión de exones); por el contrario, se basa en procesos poco convencionales que se pueden dividir en varios tipos: sitios de *splicing* crípticos, exones crípticos, microexones, sitios de *splicing* recursivos, retención de intrones, intrones exónicos (*exitrons*), RNA circulares, RNA quiméricos y sitios atípicos de *splicing*, Desde el punto de vista evolutivo, el *splicing* no canónico incrementa la probabilidad de aparición de nuevas isoformas que son "probadas" por la evolución, y, en caso sean beneficiosas, se añaden a las rutas génicas de la célula (Sibley et al., 2016). Para el caso de las tres especies mencionadas de café, y en base a *splicing* no canónico, *Coffea arabica* al contener los dos subgenomas tendría el doble de posibilidades de probar tipos de *splicing* no canónicos y por ende de generar nuevas isoformas y reiniciar o modificar rutas metabólicas que le beneficien en su evolución.

El porcentaje del *ratio* de *missmatch* por base nucleotídica (Figura 8) es una medida de calidad del mapeo, también indica si el experimento tiene una buena calidad o si el genoma de referencia usado pasa el mapeo no es el indicado. En una buena biblioteca secuenciada con la plataforma ILLUMINA (es el caso de los experimentos analizados) este valor se encuentra por debajo de 0.5%. Un porcentaje mayor a 0.5% puede indicar baja calidad de secuenciamiento o puede ser efecto propio del tipo de genoma de referencia usado en el alineamiento con STAR (Dobin y Gingeras, 2015).

El porcentaje del *ratio* de *missmatch* por base nucleotídica para el bioproyecto PRJNA630692 fue de 0.36%, 0.91% y 0.8% para las especies *Coffea arabica*, *Coffea canephora* y *Coffea eugenioides* respectivamente; y, para el bioproyecto PRJNA609253 fue de 0.27%, 0.94% y 0.79% para las especies *Coffea arabica*, *Coffea canephora* y *Coffea eugenioides* respectivamente.

Estos resultados indican que el genoma de *Coffea arabica*, como era de esperarse, es un buen genoma para mapear los *reads* ya que los datos provienen de hojas de *Coffea arabica*. Por otro lado, también muestran que *Coffea arabica* tendría más pares de bases similares a *Coffea eugenioides* que a *Coffea canephora*. Sin embargo, esto podría deberse a una parte del proceso de mapeo llamado *Clipping*. *Clipping* se define como el alineamiento parcial de un *read* a la referencia genómica, donde los extremos de los *reads* no se alinean con la referencia (Tang, 2018; The SAM/BAM Format Specification Working Group, 2022). Aquí un ejemplo:

Referencia genómica: AGTCGCCCGTCTAGCATACGCA

Read: gggGCGGGCA-ATCGTATgggg

En este ejemplo podemos ver que, si bien el *read* ha sido alineado a la referencia, los extremos no son compatibles y, por ende, el porcentaje de *missmatch* aumentaría. Teniendo en cuenta este ejemplo, el mapeo de los *reads* a *Coffea canephora* genera mayor cantidad de *reads* únicos mapeados, pero estos son mapeados con el llamado *clipping* lo que se conlleva a que se incremente el *ratio* de *missmatch*. Además, si bien el mapeo al genoma de *Coffea eugenioides* presenta menos cantidad de *reads* únicos mapedos, estos son mapeados con menos *missmatches* (posiblemente debido a la contribución de *Coffea eugenioides* al genoma de *Coffea arabica* como mencionan Scalabrin et al. (2020) en su investigación sobre el genoma de *Coffea arabica*). Esto indicaría que el genoma de *Coffea eugenioides* presenta mayor similitud de secuencias genómicas con *Coffea arabica* pero, menor aporte a la respuesta al estrés por calor de acuerdo al resultado de *reads* únicos mapeados. Contrariamente, también indicaría que *Coffea canephora* posee menor similitud de secuencias genómicas con *Coffea arabi*ca (debido al *clipping* y al genoma propio de *Coffea canphora*); pero, mayor contribución a la respuesta al estrés por calor (debido al mayor porcentaje de mapeos únicos).

Finalmente, el análisis de componente principales biplot de la figura 10 permite ver en modo de resumen las estadísticas de mapeo discutidas sobre *Coffea arabica*, *Coffea canephora* y *Coffea eugenioides*. Estos PCA muestran que el mapeo sobre las especies diploides es similar en la segunda componente (31.4% de la varianza explicada), lo que no hace más que reforzar lo discutido hasta el momento.

4.2.3. Análisis de Expresión Diferencial de Genes

El análisis exploratorio de la matriz de conteos normalizada (ver sección 7.4.1.) mostró que las réplicas biológicas del bioprojecto PRJNA609253 no tenían una clara agrupación respecto a las temperaturas 23°C y 30°C (figuras 11, 12, 13 y 14). Por el contrario, las muestras del bioproyecto PRJNA630692 si mostraron una clara separación entre los tratamientos de temperatura. Esta falta de agrupamiento de las muestras del bioproyecto PRJNA609253 podría deberse a que las variedades Acaua y Catuaí (pertenecientes al bioproyecto) tienen una respuesta similar tanto a la

temperatura de 23°C como a 30°C. De Oliveira et al. (2020) mencionan que, ambos cultivares muestran una respuesta transcripcional similar al alza de la temperatura debido a que podrían tener respuestas similares a la termorregulación y a la relocación de recursos energéticos. Además, también mencionan que *Coffea arabica*, como especie, presenta una gran resiliencia en los procesos relacionados a la fotosíntesis como respuesta a temperaturas altas (termotolerancia). Por otro lado, las figuras 11 y 13 muestran un claro agrupamiento de las muestras respecto al lugar de procedencia de los datos de secuenciamiento. Este agrupamiento mostrado sería un efecto de tipo *batch*, originado posiblemente por factores humanos, de instrumentos usados, reactivos usados, geográficos, etc; y no sería ocasionado directamente por la biología del café. Es probable que factores geográficos en invernaderos tengan influencia en la expresión génica; sin embargo, al no ser medidos, tomados en cuenta o no ser de interés para el experimento, es mejor considerarlos como efectos de tipo *batch* y estimarlos indirectamente.

Leek y Storey (2007) (ver sección 4.15.3.b.) explican que los factores técnicos, ambientales, demográficos (poblaciones de especies), o incluso los genéticos pueden perjudicar el análisis de expresión diferencial si no son tomados en cuenta. De acuerdo a esto y a lo expuesto del análisis exploratorio de la matriz de conteo, se hizo el análisis de variables sustitutas o "ocultas" con la función svaseq de Bioconductor; el cual estima a los patrones ocultos de la matriz de conteos normalizada de la siguiente manera: genera una matriz residual que no tiene el efecto de interés (en este caso el de la temperatura del aire), luego aplica la descomposición SVA a esta matriz residual, de la cual se identifican a los genes que generan la mayor variabilidad y, para estos genes, se generan variables sustitutas que se incluyen en el modelo de expresión diferencial (Leek y Storey, 2007).

Después de realizado el análisis de expresión diferencial, tomando en cuenta a las variables "ocultas" y ajustando los LFC con el modelo apeglm, se obtuvo los *MAplot* (figuras 15, 16, 17 y 18). El contraste 30°C vs 23°C (Figura 16) muestra pocos genes expresados diferencialmente en comparación a los contrastes 25°C vs 23°C (Figura 15), 37°C vs 23°C (Figura 17) y 42°C vs 23°C (Figura 18). Por otro lado, el contraste 25°C vs 23°C (Figura 15) debería tener un *MAplot* similar al contraste 30°C vs 23°C (Figura 16) de acuerdo al fundamento de resiliencia al calor de *Coffea arabica*, lo

cual no sucede; esto se analizará a mayor profundidad cuando se discutan los resultados del análisis funcional de genes. Respecto a los contrastes 37°C vs 23°C (Figura 17) y 42°C vs 23°C (Figura 18), estos muestran una gran expresión diferencial de genes. Esto concuerda con lo analizado por Ding et al. (2020), quienes obtuvieron patrones similares de expresión evaluando las hojas de tomate en contrastes de 42°C versus 23°C por cuatro horas. Igualmente, Wang et al. (2019) con hojas de la planta Chieh-Qua (una variedad de calabaza) y contraste de 45°C versus 30°C obtuvieron patrones de expresión similares a los del presente trabajo.

4.2.4. Análisis Funcional de Genes

Los resultados del análisis de enriquecimiento funcional de genes (figuras 19, 20, 21 y 22) dan mayor información sobre los contrastes evaluados en comparación a los gráficos de *MAplot* del análisis expresión diferencial de genes. Esto se debe a que, para este análisis, se agruparon a los genes en base a procesos en común (en este caso ontologías), independientemente de si son o no diferencialmente expresados (ver sección 4.15.4, párrafo 6).

El contraste 25°C vs 23°C (Figura 19) muestra que los diez primeros *genesets* (con NES positivos) son procesos que involucran la defensa contra el ataque de hongos, el mantenimiento de la membrana celular, fosforilación de proteínas y síntesis de terpenoides. Por el contrario, los diez últimos *genesets* (con NES negativos) son procesos que involucran la unión y modificación del RNA a nivel citoplasmático y a nivel de organelos como el cloroplasto, la mitocondria, así mismo, en procesos que corresponden a las membranas intracelulares y de organelos. Estos resultados tienen similitudes a los obtenidos por Li et al. (2018) y Sjokvist et al. (2018) quienes estudiaron la interacción hongo-planta mediante la ténica de RNA-Seq y usando el método ORA (ver sección 4.15.4) para analizar las ontologías. Li et al. (2018) analizaron el transcriptoma de *Triticum aestivum* (trigo) infectado con el hongo *Rhizophagus irregularis* encontrando *genesets* de procesos que involucran a la membrana tanto celular como intracelular, procesos de modificaciones de proteínas y actividad de tipo quinasa. Por otro lado, Sjokvist et al. (2018) analizaron el transcriptoma de *Hordeum vulgare* (cebada) cuando es infectado por *Ramularia collo-cygni*, encontrando *genesets* de procesos que involucran modificación de proteínas, transporte, metabolismo del DNA/RNA, actividad quinasa, actividades

catalíticas, así como, relacionados con las membranas celular e intracelular. Por lo tanto, de acuerdo a lo encontrado por Li et al. (2018) y Sjokvist et al. (2018), las muestras sometidas a 25°C estarían presentando una expresión que corresponde a infección por algún hongo. Esto toma mayor peso cuando se observa que las ontologías GO:0004568 (actividad quitinasa), GO:0058032 (respuesta de defensa a hongos) y GO:000238 (respuesta a la molécula de origen fúngico) son significativas con un NES positivo. Además, la figura 22 nos muestra que algunos genes correspondientes a las ontologías actividad quitinasa y a la defensa contra hongos tienen un LFC mayor a 7.5 (que sería un ratio del orden de 2 elevado a la 7.5 lo cual es bastante alto) lo que posiblemente estaría mostrando que las plantas usadas (25°C) para el experimento han sido infectadas y no serían útiles para mayor discusión.

La figura 20 correspondiente al contraste 30°C vs 23°C muestra que los diez primeros *genesets* (con NES positivos) son de procesos que involucran al cloroplasto, membrana celular, transporte transmembrana, actividad quitinasa y fosforilación de proteínas. Por el contrario, los diez últimos *genesets* (con NES negativos) son procesos que involucran regulación de la transcripción, unión de factores de transcripción al DNA, procesos metabólicos que involucran al acetil-CoA (como ciclo de krebs, síntesis de lípidos, síntesis de terpenos, etc.) y actividad de chaperonas (*unfolded protein binding*). De Oliveira et al. (2020) al analizar el mismo contraste (con el método ORA para ontologías), encontraron procesos ontológicos relacionados a la defensa de la planta y a la respuesta a estímulos bióticos; lo que tendría relación con la presencia de la actividad quitinasa. Además, seleccionaron 10 genes para analizar la expresión mediante PCR en tiempo real y, observaron que uno de ellos (el gen *sHSP-like*) estaba regulado hacia abajo; lo que tiene concordancia con la actividad chaperona encontrada con un NES negativo. Estos resultados tienen concordancia con lo discutido sobre el análisis exploratorio de la matriz de conteo, indicando que una temperatura de 30°C no es lo suficientemente alta para observar sobreexpresión de los procesos relacionados al estrés por temperatura; y por lo contrario, de acuerdo a lo observado en la figura 23, los procesos ontológicos están más relacionados al transporte transmembrana y al metabolismo de la célula y no a la respuesta al alza de la temperatura, lo cual indica que *Coffea arabica* a 30°C tiene una buena termotolerancia (Liu et al., 2015).

La figura 21 correspondiente al contraste 37°C vs 23°C muestra que los diez primeros *genesets* (NES positivos) son de procesos que involucran a la actividad de chaperonas, plegamiento de proteínas, activación de rutas dependientes de auxinas y unión de factores de transcripción al DNA. Por el contrario, los diez últimos *genesets* (con NES negativos) son procesos que involucran modificación y unión al RNA y procesos relacionados al cloroplasto y mitocondria. Marques et al. (2021) analizaron el efecto de la temperatura (37°C y 42°C) y el efecto de niveles de CO_2 sobre el transcriptoma de *Coffea arabica* mediante el método ORA; encontrando procesos relacionados a la actividad de chaperonas (*unfolded protein binding*), actividad sintasa de terpenos, plegamiento de proteínas, unión de proteínas relacionadas al estrés por calor y unión del RNA. Lo encontrado por Marques et al. (2021) tiene procesos que se solapan con los resultados del contraste 37°C vs 23°C (figura 21) y; además, el mapa de calor de dicho contraste (figura 24) muestra que los genes que están incluídos en los procesos de respuesta al calor tienen un LFC mayor a 7.5 (valor alto) indicando que a 37°C ya se estaría activando la termotolerancia adquirida descrita por Liu et al. (2015).

Por último, la figura 22 correspondiente al contraste 42°C vs 23°C muestra que los diez primeros *genesets* (NES positivo) corresponden a procesos que involucran a la actividad de chaperonas, plegamiento de proteínas, estímulo de los factores de *splicing* y unión de factores de transcripción al DNA. Por otro lado, los diez últimos *genesets* (con NES negativos) son de procesos que involucran organización de la pared celular, transporte transmembrana y procesos relacionados a la membrana celular. Respecto al trabajo de Marques et al. (2021) analizando el efecto de 42°C sobre *Coffea arabica*, se encontraron procesos similares a la temperatura de 37°C. Estos procesos son actividad de chaperonas (*unfolded protein binding*), unión de proteínas relacionadas al estrés por calor y plegamiento de proteínas. De manera similar, Mansilla (2021) analizó el transcriptoma de café sometido a un estrés por alta temperatura del aire, 45°C, encontrando genes relacionados al plegamiento de proteínas, activación de chaperonas y a transcriptos que regulan la respuesta al calor. Por lo tanto, temperaturas de 37°C o mayores, estarían activando el proceso *heat shock response* (HSR) (ver sección 4.10.); esto se corrobora en la figura 26, donde se muestra que los genes involucrados en el proceso de HSR tienen un LFC mayor 20, indicando que a 42°C ya se tiene activada la respuesta al estrés por altas

temperaturas,; así como, de la termotolerancia adquirida ya que, procesos como *unfolded protein binding* (GO:0051082) y DNA-*binding transcription factor activity* (GO:0003700) (Figuras 22 y 26) , tienen relación con las *heat shock proteins* (HSP) y los *heat shock factors* (HSF), respectivamente (Liu et al., 2015).

4.2.5. Análisis de la información de cuantificación, expresión diferencial y enriquecimiento funcional

El gráfico de tela de araña (figura 27), permite comparar los patrones de activación de ontologías de los contrastes 37°C vs 23°C y 42°C vs 23°C (23°C es el control). En el gráfico se pudo observar que las ontologías (*unfolded protein binding, protein self-association, chaperone binding, precatalytic spliceosome, spliceosomal complex, transcription factor TFIID complex, nuclear speck, etc*.) tienen directa relación con el proceso celular de HSR, lo que concuerda con el resultado del análisis de enriquecimiento funcional. Entre ellas, podemos destacar a cuatro grupos, que se formaron en base a la diferencia de los contrastes 37°C vs 23°C y 42°C vs 23°C de la figura 27, para enmarcar mejor la discusión: ontologías relacionadas a la HSR, ontologías relacionadas a la transcripción del mRNA, ontologías relacionadas a auxinas y ontologías relacionadas a los metabolitos secundarios.

El primer grupo incluye a las ontologías *unfolded protein binding*, *protein self-association* y *chaperone binding*. Los genes que se encontraron dentro de este primer grupo, corresponden a la familia de proteínas sHSP (*small heat shock protein*); específicamente, para las temperatursa de 37°C y 42°C, se identificaron a las clases I, II, III, IV (citosólicas), MT (mitocondrial), PX (peroxisomal) y al gen que codifica la proteína dnaJ (co-chaperona). Según Waters y Vierling (2020) y Al-Whaibi (2011), las sHSP´s son proteínas ATP-independientes que previenen la desnaturalización y agregación irreversible de las proteínas. Además, al unirse a estas proteínas parcialmente desnaturalizadas, permiten que las HSP (70 y 100), dependedientes de ATP, se unan a las proteínas desnaturalizadas y se puedan re-plegar a su forma funcional. Por otro lado, sólo en plantas se ve que existen sHSP que son compartimentalizadas; es decir, que la respuesta al calor por las sHSP es específica para el cloroplasto, mitocondria, peroxisoma, retículo endoplasmático, núcleo y citosol, aunque, existen sHSP que pueden trasladarse de mitocondria a cloroplasto y viceversa (Waters y Vierling, 2020). Además, en base a su función

chaperona, Al-Whaibi (2011) menciona que las sHSP tienen un rol fundamental en el control de calidad de la membrana plasmática y contribuyen a su mantenimiento especialmente bajo condiciones de estrés. De manera similar, según Waters y Vierling (2020) las sHSP ayudan al plegamiento de factores de transcripción y de traducción, asi como, al mantenimiento de los fotosistemas del tilacoide durante condiciones de estrés. En resumen, la presencia de estas ribonucleoproteínas indica que *Coffea arabica* estaría protegiendo el funcionamiento de la membrana plasmática, los fotosistemas del tilacoide y estarían previniendo la desnaturalización de las proteínas bajo condiciones de elevadas temperaturas (37°C y 42°C).

Dentro del grupo de ontologías relacionadas a la transcripción podemos mencionar a *precatalytic spliceosome, spliceosomal complex, nuclear speck y transcription factor TFIID complex*. Los genes que conforman estos *genesets* corresponden a los relacionados a la formación y activación del complejo espliceosoma. El espliceosoma, que es la asociación RNA-proteína más compleja que presentan las células eucariotas, se encarga del proceso de *splicing* alternativo de las móleculas de pre-mRNA dentro del núcleo. Este complejo está conformado por cinco diferentes subunidades pequeñas de ribonucleoproteínas (snRNP) y de varias proteínas que actúan como cofactores (Matera y Wang, 2014). Estas cinco snRNP (U1, U2, U4, U5 y U6) actúan de manera secuencial para completar el *splicing* del pre-mRNA. En este proceso secuencial, primero la U1 reconoce secuencias motivo en el extremo 5´ de un intrón del pre-mRNA y se une a estas secuencias; cabe resaltar que esta unión es estabilizada por las proteínas SR (proteínas ricas en serina/arginina). Luego, la U2 se une a una adenina del extremo 3´del intrón con ayuda de las proteínas SF1 y U2AF formando el complejo E, que se conoce como *Exon definition* (reconocimiento de exones). Subsecuentemente, helicasas de tipo DExD/H ayudan a que los extremos de los intrones se acerquen (*Intron definition*) formando el complejo A conocido como pre-espliceosoma. Una vez acercados los intrones mediante las proteínas U1 y U2, las proteínas U4, U5 y U6 se unen al complejo A mediante la proteína helicasa de tipo DExD/H formando el complejo B. El cual, es objeto de rearreglos espaciales formando el complejo B activado, (complejo B*) que cataliza la primera reacción de *splicing* (formando un exón libre y un exón unido al intrón) formando el complejo C (espliceosoma catalítico). Este complejo C, luego de rearreglos proteícos forma el complejo post-espliceosoma, el cual realiza la segunda reacción catalítica uniendo a

los exones y separando el intrón con ayuda de la helicasa DExD/H (Matera y Wang, 2014). Entre los genes que conforman la ontología *nuclear speck*, están HSFs y proteínas relacionadas a la formación y activación del spliceosoma. La zona *nuclear speck* (mota nuclear) es una zona nucleoplásmica de reserva ubicada en los bordes del nucleolo y es rica en factores envueltos en el *splicing* del pre-mRNA como proteínas snRNP y SR (Matera y Wang, 2014). Es de destacar esta zona, porque la mayoría de la actividad *splicing* se ubica entre los bordes de los *speckles* (motas) y la cromatina del nucleolo (Matera y Wang, 2014). Finalmente, los genes correspondientes a la ontología *transcription factor TFIID complex* codifican a proteínas de la familia de factores de inicio de la transcripción (TFIIA, TFIIB, TFIIE y TFIIF) que permiten la unión de la RNA polimerasa II (Matera y Wang, 2014; Patel et al., 2020).

El tercer grupo que corresponde a los genes relacionados a auxinas incluye factores de transcripción Aux/IAA represores de auxina, PIN relacionados a la acumulación y transporte de auxinas, BIG GRAIN relacionados al transporte de auxinas, y tetraspaninas relacionados a receptores de superficie en la membrana plasmática (Uniprot, 2021). Sharif et al. (2022) en una revisión de literatura sobre el rol de las auxinas en el crecimiento y estrés de *cucumis sativus* L. (pepino), menciona que los genes PIN y Aux/LAX codifican proteínas que transportan y regulan la homeostasis de las auxinas; además, que los genes Aux/IAA, TIR1/AFB y factores receptores de auxinas (ARF) son responsables de una regulación estricta en diferentes procesos que incluyen el desarrollo embrionario, aborto de semillas y establecimiento del fruto. Asimismo, Sharif et al. (2022) mencionan que las auxinas tienen un rol fundamental en la temomorfogénesis de la planta que incluye la elongación del hipocótilo y la elongación de las hojas. Por otro lado, Chen et al. (2022) en un análisis del metaboloma y transcriptoma de *Eriobotrya japónica* (níspero) expuesto a altas temperaturas (40°C), encontraron que los niveles de las homonas auxina, ácido abcísico (ABA), ácido salicílico (SA), ácido giberélico (GA), ácido jasmónico (JA) y etileno estaban elevados, mientras que la hormona citoquinina estaba en bajas concentraciones. Además, en el análisis ontológico realizado con el método ORA de los genes diferencialmente expresados, encontraron que la ruta metabólica "transducción de señal mediada por auxina" estaba sobreexpresada y, que concuerda con los análisis previamente realizados en arabidopsis y soya, en las que también se

encontraron niveles incrementados de auxinas y de expresión de genes biosínteticos de auxina bajo condiciones de estrés térmico. Chen et al. (2022) proponen una ruta reguladora de genes en respuesta al estrés térmico en la cual las proteínas HSP promueven la protección y mantenimiento de la fotosíntesis y regulan la señalización de auxinas lo cual mejora la respuesta al calor.

En el último grupo que corresponde a ontologías relacionadas a metabolitos secundarios tenemos a genes relacionados a la síntesis de terpenoides como: (-)-germacreno sintetasa D, cis-abienol sintetasa, viridifloreno sintetasa, S-linalol sintetasa, 1-deoxi-D-xilulosa-5-fosfato sintetasa y 4-difosfocitidil-2-C-methil-D-eritritol quinasa. En la base de datos de Uniprot (Uniprot, 2021), encontramos que la enzima (-)-germacreno sintetasa D es parte de la ruta de síntesis del sesquiterpeno (-)-germacreno a partir de la molécula farnesil pirofosfato; la enzima cis-abienol sintetasa es parte de la ruta de síntesis del diterpeno cis-abienol (parte de la síntesis de giberelinas) a partir de geranilgeranil pirofosfato; la enzima viridifloreno sintetasa es parte de la ruta de síntesis del sesquiterpeno viridifloreno a partir de farnesil pirofosfato; la enzima S-linalol sintetasa es parte de la ruta de síntesis del monoterpeno S-linalol (molécula antibacteriana) a partir de geranil pirofosfato y por último; la 1-deoxi-D-xilulosa-5-fosfato sintetasa (DXS, primera enzima de la ruta de metil eritrol fosfato:MEP) y la 4-difosfacitidil-2-C-methil-D-eritritol quinasa (CMK) son parte de la ruta MEP para la síntesis de terpenoides (Yan et al., 2022).

La presencia de terpenoides en la respuesta al estrés por temperatura ha sido estudiada por Akhi et al. (2021) y Yan et al. (2022), en cuyos trabajos se menciona que, la naturaleza anfipática de los terpenoides mejora las interacciones hidrofóbicas entre las proteínas de membrana y los lípidos de membrana (mejora la estabilidad de membrana) y, remedian el daño causado por las especies reactivas de oxígeno frente a condiciones de estrés. Respecto a los genes que sintetizan terpenos (Terpeno sintetasa), Yan et al. (2022) menciona que son clasificados en siete subfamilias (a,b,c,d,e/f,g y h) de las cuales, la subfamilia g sólo se encuentra en gimnospermas y la familia h sólo se encuentra en *Selaginella moellendorffii* y; la cantidad de estos genes varía dependiendo de la especie.

Yan et al. (2022) en *Rosa chinensis* (rosa) evaluaron la expresión de genes para la síntesis de terpenoides, mediante PCR en tiempo real, en estrés de temperatura

(35°C); encontrando que el 78% de los genes que sintetizan terpenoides en esa especie, estuvieron regulados hacia arriba. Por otro lado, Akhi et al. (2021) mencionan en su revisión de literatura que una variedad transgénica de tabaco (*Nicotiana attenuata*), la cual se modificó para sobreexpresar la síntesis de una enzima terpeno sintetasa, poseía resistencia al estrés térmico. En base a esto, la presencia de los genes que codifican a las enzimas cis-abienol sintetasa, DXS y CMK estaría indicando que *Coffea arabica* sintetiza terpenoides en el cloroplasto usando la ruta MEP para la protección de la fotosíntesis de los efectos del estrés por alta temperatura. Y, la presencia de los genes que codifican a las enzimas la enzima (-)-germacreno sintetasa D, viridifloreno sintetasa y S-linalol sintetasa estaría indicando que *Coffea arabica* estaría usa la ruta del mevalonato (MVA) para atenuar los efectos del estrés térmico en el citosol (Yan et al., 2022).

V. CONCLUSIONES

- La búsqueda de experimentos en la base de datos de secuencias SRA del NCBI resultó en la selección de los bioproyectos PRJNA630692 y PRJNA609253.
- La comparación de las estadísticas de mapeo permitió determinar que el transcriptoma de *Coffea arabica* en condiciones de estrés a altas temperaturas se asemeja más al genoma de *Coffea eugenioides* que al genoma de *Coffea canpehora*.
- Mediante el análisis de expresión diferencial se ha podido evidenciar diferencias de expresión génica en *Coffea arabica* cuando las temperaturas son mayores a 37°C.
- El análisis funcional de genes permitió identificar ontologías relacionadas a la termotolerancia adquirida de *Coffea arabica* en temperaturas de 37°C y 42°C. Lo cual indicaría que esta especie presenta resiliencia a elevadas temperaturas del aire.
- El aumento de la temperatura del aire sobre las hojas de *Coffea arabica* permitió identificar la sobreexpresión de genes relacionados a la formación del complejo espliceosoma.
- El aumento de la temperatura del aire sobre las hojas de *Coffea arabica* permitió identificar la sobreexpresión de genes relacionados al transporte, acumulación, transducción y represión de auxinas.
- El aumento de la temperatura del aire sobre las hojas de *Coffea arabica* permitió identificar la sobreexpresión de genes relacionados a la síntesis de terpenoides.

VI. RECOMENDACIONES

- Realizar el análisis transcriptómico del RNA total de *Coffea arabica* en condiciones de elevadas temperaturas (mayores a 37°C) para evaluar las isoformas codificantes y no codificantes del mRNA y, ahondar en el proceso de respuesta al estrés de temperatura.

- Realizar el secuenciamiento de regiones motivo de los factores de transcipción de *Coffea arabica* en condiciones de elevadas temperaturas (mayores a 37°C) con la técnica de secuenciamiento de inmuno precipitación de la cromatina (chromatine inmuno precipitation sequencing: ChIP-Seq) para evaluar a los factores de trnascripción presentes en el estrés por temperatura.

- En caso de realizar el análisis transcriptómico del mRNA de *Coffea arabica*, tomar en cuenta en la etapa preparación de la biblioteca genómica la opción de usar *kits* que permitan obtener la dirección de la que proviene el mRNA (metodología *stranted*).

- Es recomendable usar el *software* RSEM para la etapa de conteo de *reads* ya que, permite contar a los *reads* que son mapeados a múltiples loci y no deja que pierda esa información importante.

- Identificar el número de genes que codifican a las proteínas sHSP en el genoma de *Coffea arabica* para evaluar la posible resiliencia de esta especie frente a elevadas temperaturas.

- Identificar el número de genes que codifican a las enzimas terpeno sintasa en el genoma de *Coffea arabica* para evaluar la diversidad de estás enzimas presentes en la especie y su accionar frente a elevadas temperatura del aire.

- Realizar el análisis mediante PCR en tiempo real a los genes que codifican a las proteínas sHSP y a enzima terpeno sintasa para entender mejor lo que ocurre en *Coffea arabica* bajo elevadas temperaturas del aire.

VII. BIBLIOGRAFÍA

1. Aerts, R., Geeraert, L., Berecha, G., Hundera, K., Muys, B., De Kort, H. y Honnay, O. (2017). Conserving wild Arabica coffee: Emerging threats and opportunities. *Agriculture, Ecosystems & Environment*, *237*, 75-79.

2. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Cech, M., Chilton J., Clements D., Coraor N., Grüning B., Guerler A., Hillman-Jackson J., Hiltemann S., Jalili V., Rasche H., Soranzo N., Goecks J., Taylor J., Nekrutenko A., Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, *46*(W1): W537-W544.

3. Aga, E. (2005). *Molecular genetic diversity study of forest coffee tree (Coffea arabica L.) populations in Ethiopia: Implications for conservation and breeding* [Doctoral dissertation, Swedish University of Agricultural Sciences]. https://pub.epsilon.slu.se/918/1/agalatest_version.pdf

4. Akhi, M., Haque, M., y Biswas, M. (2021). Role of secondary metabolites to attenuate stress damages in plants. *In Antioxidants-Benefits, Sources, Mechanisms of Action*. IntechOpen.

5. Al-Whaibi, M. H. (2011). Plant heat-shock proteins: a mini review. *Journal of King Saud University-Science*, 23(2), 139-150.

6. Amrouk, E. M. (2018). Depressed international coffee prices: insights into the nature of the price decline. *FAO Food Outlook*, 25-28.

7. Anders, S., Pyl, P. T., y Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.

8. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

9. Anthony, F., Bertrand, B., Quiros, O., Wilches, A., Lashermes, P., Berthaud, J. y Charrier, A. (2001). Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. *Euphytica*, 118:53–65

10. Arcila, J., Farfán, F., Moreno A., Salazar L., e Hincapié, E. (2007). Sistemas de producción de café en Colombia. Cernicafé, 309p.

11. Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K., Rsenchuk, S., Tatusova, T., Yaschenko, E. y Ostell, J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic acids research*, *40*(D1), D57-D63.

12. Bagheri, R. (2020). Understanding Singular Value Decomposition and its Application in Data Science. *Towards Data Science*. https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d.

13. Bolger, A., Lohse, M., y Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

14. Budzinski, I., Camargo, P., Rosa R., Calzado N., Ivamoto-Suzuki S. y Domingues D. (2021). Transcriptome Analyses of Leaves Reveal That Hexanouc Acid Priming Differentially Regulate Gene Expression in Contrasting *Coffea arabica* Cultivars. *Frontiers in Sustainable Food systems*.

15. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. y Madden, T. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1):421.

16. Camargo, M. (2010). The impact of climatic variability and climate change on Arabic coffee crop in Brazil. *Bragantia*, *69*(1), 239-247.

17. Chen, Y., Deng, C., Xu, Q., Chen, X., Jiang, F., Zhang, Y., Hu, W., Zheng, S., Su, W., y Jiang, J. (2022). Integrated analysis of the metabolome, transcriptome and miRNome reveals crucial roles of auxin and heat shock proteins in the heat stress response of loquat fruit. *Scientia Horticulturae*, 294, 110764.

18. Chen, S., Zhou, Y., Chen, Y. y Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.

19. Conesa, A. y Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*.

20. Corchete, L. , Rojas, E. , Alonso-López, D., De Las Rivas, J., Gutiérrez, N. y Burguillo, F. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific reports*, *10*(1):19737.

21. Davis, A., Chester, M., Maurin, O. y Fay, M. (2007). Searching for the relatives of *Coffea* (Rubiaceae, Ixoroideae): the circumscription and phylogeny of Coffeeae based on plastid sequence data and morphology. *American Journal of Botany*, *94*(3), 313-329.

22. Davis, A. y Rakotonasolo, F. (2008). A taxonomic revision of the baracoffea alliance: nine remarkable Coffea species from western Madagascar. *Botanical Journal of the Linnean Society*, *158*(3), 355-390.

23. De Kochko, A., Akaffou, S., Andrade, A., Campa, C., Crouzillat, D., Guyot, R., Hamon, P., Ming, R., Mueller, L., Poncet, V., Tranchant-Dubreuil, C. y Hamon, S. (2010). Advances in Coffea genomics. In *Advances in botanical research* (Vol. 53, pp. 23-63). Academic Press.

24. Ding, H., Mo. S., Qian, Y., Yuan, G., Wu, X. y Ge, C. (2020). Integrated proteome and transcriptome analyses revealed key factors involved in tomato (*Solanum lycopersicum*) under high temperature stress. *Food and Energy Security*, 9(4), e239.

25. Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut P., Chaisson M. y Gingeras, T. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

26. Dobin, A., y Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*, 51(1), 11-14.

27. Ewels, P., Magnusson, M., Lundin, S. y Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048.

28. FAO. (2015). Statistical pocketbook of the Food and Agricultural Organization for the United Nations.

29. FAO. (2021). Food Outlook: Biannual Report on Global Food Markets. Rome. https://doi.org/10.4060/cb4479en.

30. FAOSTAT. (2022). Cultivos y productos de ganadería. https://www.fao.org/faostat/es/.

31. Fetzek, S. (2017). Climate, coffee, and security. *Epicenters of climate and security: the new geostrategic landscape of the Anthropocene*. *The Center for Climate and Security*.

32. Frey, K. y Pucker B. (2020). Animal, fungi, and plant genome sequences harbor different non-canonical splice sites. *Cells*, 9(2), 458.

33. Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., Zimmer, R. y Waldron, L. (2021). Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics*, *22*(1), 545-556.

34. Hassan, M. , Chattha, M. U., Khan, I., Chattha, M. B., Barbanti, L., Aamer, M., Iqbal M., Nawaz M., Mahmood A., Ali A. y Aslam, M. (2021). Heat stress in cultivated plants: Nature, impact, mechanisms, and mitigation strategies—A review. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 155(2), 211-234.

35. Hrdlickova, R., Toloue, M. y Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews RNA*: *8*(1), e1364.

36. INEI. (2021). Producción de café se incrementó 17,0% en julio de 2021. https://www.inei.gob.pe/prensa/noticias/produccion-de-cafe-se-incremento-170-en-julio-de-2021-13123/.

37. *International Coffee Organization*. (2022). *Trade statistics – June 2022*. Revisado el 26 de agosto del 2022, de https://www.ico.org/.

38. Irizarry, R., Wang, C., Zhou, Y. y Speed, T. (2009). Gene set enrichment analysis made simple. *Statistical methods in medical research*, *18*(6), 565-575.

39. Irizarry, R. y Love, M. (2016). Data analysis for the life sciences with r. CRC Press.

40. Ivamoto, S., Reis, O., Domingues, D., Dos Santos, T., De Oliveira, F., Pot, D., Leroy T., Esteves L., Falsarella M., Guimaraes G. y Pereira, L. (2017). Transcriptome analysis of leaves, flowers and fruits perisperm of *Coffea arabica* L. reveals the differential expression of genes involved in raffinose biosynthesis. *PloS one*, 12(1), e0169595.

41. Kassambara, A. (2017). Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra (Vol. 2). Sthda.

42. Khatri, P., Sirota, M. y Butte, A. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, *8*(2), e1002375.

43. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. y Sergushichev, A. (2021). Fast gene set enrichment analysis. *BioRxiv*, 060012.

44. Labouisse, J. P., Bellachew, B., Kotecha, S. y Bertrand, B. (2008). Current status of coffee (*Coffea arabica* L.) genetic resources in Ethiopia: implications for conservation. *Genetic Resources and Crop Evolution*, *55*(7), 1079-1093.

45. Lashermes, P., Trouslot, P., Anthony, F., Combes, M. C. y Charrier, A. (1996). Genetic diversity for RAPD markers between cultivated and wild accessions of Coffea arabica. *Euphytica*, *87*(1), 59-64.

46. Lashermes, P., Combes, M., Robert, J., Trouslot, P., D'Hont, A., Anthony, F. y Charrier, A. (1999). Molecular characterisation and origin of the *Coffea arabica* L. genome. *Molecular and General Genetics*, *261*(2), 259-266.

47. Leinonen, R., Sugawara, H., Shumway, M. y *International Nucleotide Sequence Database Collaboration*. (2010). The sequence read archive. *Nucleic acids research*, *39*(suppl_1), D19-D21.

48. Leek, J. y Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics, 3(9), e161.

49. Li, B. y Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12:323*.

50. Li, M., Wang, R., Tian, H., y Gao, Y. (2018). Transcriptome responses in wheat roots to colonization by the arbuscular mycorrhizal fungus *Rhizophagus irregularis*. *Mycorrhiza*, 28(8), 747-759.

51. Liu, H. C. y Charng, Y. Y. (2012). Acquired thermotolerance independent of heat shock factor A1 (HsfA1), the master regulator of the heat stress response. *Plant signaling & behavior*, *7*(5), 547-550.

52. Liu, J., Feng, L., Li, J. y He, Z. (2015). Genetic and epigenetic control of plant heat responses. *Frontiers in plant science*, *6*, 267.

53. Love, M., Huber, W. y Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*, 550.

54. Love, M., Anders, S. y Huber, W. (2017). Analyzing RNA-seq data with DESeq2. *R package reference manual*.

55. Mansilla Samaniego, R. C. (2021). *Transcriptoma de Cafetos (Coffea arabica L.) cultivados en el Perú como respuesta a la elevación de la temperatura del aire*. [Doctoral dissertation, Universidad Nacional Agraria la Molina]. *Repositorio Institucional Universidad Nacional Agraria la Molina.*

56. Marini, F., Ludt, A., Linke, J. y Strauch, K. (2021). GeneTonic: an R/Bioconductor package for streamlining the interpretation of RNA-seq data. *BMC bioinformatics*, *22*, 610.

57. Marques, I., Fernandes, I., Paulo, O., Lidon, F., DaMatta, F., Ramalho, J. y Ribeiro-Barros, A. (2021). A transcriptomic approach to understanding the combined impacts of supra-optimal temperatures and CO2 revealed different responses in the polyploid Coffea arabica and its diploid progenitor *C. canephora*. *International journal of molecular sciences*, *22*(6), 3125.

58. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing *reads. EMBnet. journal*, *17*(1), 10-12.

59. Martins, R., Queiroz, J. A. y Sousa, F. (2014). Ribonucleic acid purification. *Journal of Chromatography A*, *1355*, 1-14.

60. Matera, A., y Wang, Z. (2014). A day in the life of the spliceosome. *Nature reviews Molecular cell biology*, 15(2), 108-121.

61. Mekbib, Y., Tesfaye, K., Dong, X., Saina, J., Hu, G. W., & Wang, Q. F. (2022). Whole-genome resequencing of *Coffea arabica* L. (Rubiaceae) genotypes identify SNP and unravels distinct groups showing a strong geographical pattern. *BMC plant biology*, *22*, 69.

62. Ministerio de Agricultura y Riego. 2018. Plan nacional de acción del café peruano 2018- 2030. Lima – Perú. 101 pp.

63. Montagnon, C., Mahyoub, A., Solano, W. y Sheibani, F. (2021). Unveiling a unique genetic diversity of cultivated *Coffea arabica* L. in its main domestication center: Yemén. *Genetic Resources and Crop Evolution*, *68*(6), 2411-2422.

64. Murray, R., Granner, D., Mayes, P. y Rodwell, V. (2003). Illustrated Biochemistry. *Mc Graw Hill*.

65. *Observatory of Economic Complexity*. (2020). *Coffee*. Revisado el 29 de agosto del 2022, de https://oec.world/en/profile/hs/coffee.

66. Overbey, E., Saravia-Butler, A., Zhang, Z., Rathi, K., Fogle, H., da Silveira, W., Barker, R., Bass, J., Beheshti, A., Berrios, D., Blaber, E., Cekanaviciute, E., Costa, H., Davin, L., Fish, K., Gebre, S., Geniza, M., Gilbet, R., Gilroy, S., ... y Galazka, J. (2021). NASA GeneLab RNA-seq consensus pipeline: Standardized processing of short-read RNA-seq data. *Iscience*, *24*(4), 102361.

67. Patel, A., Greber, B. J., y Nogales, E. (2020). Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. *Current opinion in structural biology*, 61, 17-24.

68. Romero, G., Vásquez, L., Lashermes, P. y Herrera, J. (2014). Identification of a major QTL for adult plant resistance to coffee leaf rust (*Hemileia vastatrix*) in the natural Timor hybrid (*Coffea arabica* x *C. canephora*). *Plant breeding*, *133*(1), 121-129.

69. Scalabrin, S., Toniutti, L., Di Gaspero, G., Scaglione, D., Magris, G., Vidotto, M., Pinosio, S., Cattonaro, F., Magni, F., Jurman, I., Cerutti, M., Liverani, F., Navarini, L., Del Terra, L., Pellegrino, G., Ruosi, M., Vitulo, N., Valle, G., Pallavicini, A., ... y Bertrand, B. (2020). A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific reports*, *10*, 4642.

70. Schaarschmidt, S., Fischer, A., Zuther, E. e Hincha, D. (2020). Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *International journal of molecular sciences*, *21*(5), 1720.

71. Sharif, R., Su, L., Chen, X., y Qi, X. (2022). Involvement of auxin in growth and stress response of cucumber. *Vegetable Research*, 2(1), 1-9.

72. Shukla, P., Skea, J., Slade, R., Van Diemen, R., Haughey, E., Malley, J., Pathak, M. y Portugal, J. (2019). Technical summary. Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems Geneva, Switzerland: The Intergovernmental Panel on Climate Change (IPCC).

73. Sibley, C., Blazquez, L. y Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews Genetics*. 17, 407–421. https://doi.org/10.1038/nrg.2016.46.

74. Sjokvist, E., Lemcke, R., Kamble, M., Turner, F., Blaxter, M., Havis, N. H., ... & Radutoiu, S. (2019). Dissection of Ramularia leaf spot disease by integrated analysis of barley and *Ramularia collo-cygni* transcriptome responses. *Molecular plant-microbe interactions*, 32(2), 176-193.

75. Soneson, C., Love, M., y Robinson, M. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4.

76. Starmer, J. (15 de julio de 2015). RPKM, FPKM and TPM clearly explained. "STATQUEST!!! An epic journey through statistics and machine learning". https://statquest.org/rpkm-fpkm-and-tpm-clearly-explained/

77. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. y Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550.

78. Tanaka, Y., Fujii, S., Hiroaki, H., Sakata, T., Tanaka, T., Uesugi, S., Tomita, K. y Kyogoku, Y. (1999). A′-form RNA double helix in the single crystal structure of r (UGAGCUUCGGCUC). *Nucleic acids research*, 27(4), 949-955.

79. Tang, Dave. (2018). SAM. https://davetang.org/wiki/tiki-index.php?page=SAM.

80. Tesfaye, G., Govers, K., Oljira, T., Bekele, E. y Borsch, T. (2007). Genetic diversity of wild *Coffea arabica* in Ethiopia: analyses based on plastid, ISSR and microsatellite markers. In: 21st International Coffee Science Conference, Montpellier, 11–15 September 2006, [CD-ROM], pp 802–810.

81. The SAM/BAM Format Specification Working Group. (2022). Sequence Alignment/Map Formart Specification. https://samtools.github.io/hts-specs/SAMv1.pdf.

82. UniProt: the universal protein knowledgebase in 2021. (2021). *Nucleic acids research*, vol. 49, no D1, p. D480-D489.

83. Wang, Z., Gerstein, M. y Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

84. Wang, M., Jiang, B., Liu, W., Lin, Y. E., Liang, Z., He, X., y Peng, Q. (2019). Transcriptome analyses provide novel Insights into heat stress responses in Chieh-Qua (*Benincasa hispida* Cogn. var. Chieh-Qua How). *International journal of molecular sciences*, 20(4), 883.

85. Waters, E., y Vierling, E. (2020). Plant small heat shock proteins–evolutionary and functional diversity. *New Phytologist*, 227(1), 24-37.

86. Weber, A. P. (2015). Discovering new biology through sequencing of RNA. *Plant physiology*, *169*(3), 1524-1531.

87. Yan, Y., Li, M., Zhang, X., Kong, W., Bendahmane, M., Bao, M., y Fu, X. (2022). Tissue-Specific Expression of the Terpene Synthase Family Genes in Rosa chinensis and Effect of Abiotic Stress Conditions. *Genes*, 13(3), 547.

88. Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., Von Schack, D. y Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC genomics*, *16*, 675.

89. Zhu, A., Ibrahim, J. y Love, M. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, *35*(12), 2084-2092.

VIII. ANEXOS

- Anexo 1: Metadatos de los bioproyectos- https://github.com/pipaber/Thesis-coffea-arabica/blob/main/Metadata_bioprojects_coffeaarabica.xlsx
- Anexo 2: Datos del bioproyecto PRJNA630692- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/PRJNA630692
- Anexo 3: Datos del bioproyecto PRJNA609253- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/PRJNA609253
- Anexo 4: Tabla de estadísticas de STAR- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/STAR%20aligment%20stats
- Anexo 5: Código R para análisis de expresión diferencial de genes, análisis funcional de genes y Genetonic- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/R%20code%20DEseq2%20and%20fgsea
- Anexo 6: Muestras por temperatura- https://github.com/pipaber/Thesis-coffea-arabica/blob/main/R%20code%20DEseq2%20and%20fgsea/muestras_portemperatura.rtf
- Anexo 7: Código R para la contrucción del objeto de anotación- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/Annotation%20of%20coffea%20arabica
- Anexo 8: Código R para el análisis con Genetonic (tomar en cuenta que en el anexo 5 se generan los archivos necesarios para correr el código de este anexo)- https://github.com/pipaber/Thesis-coffea-arabica/tree/main/Genetonic_code