Análisis transcriptómico asociado a la producción de β-caroteno en Yuca

Tatiana Melissa Ovalle Rivera
Máster en Bioinformática y Bioestadística
MO.137 – FM – Estadística y Bioinformática

Lorena Pantano Rubiño
Profesora

Junio 13, 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis transcriptómico asociado a la producción de β-caroteno en Yuca*
Nombre del autor:	*Tatiana Melissa Ovalle Rivera*
Nombre del consultor/a:	*Lorena Pantano Rubiño*
Nombre del PRA:	*Maria Jesús Marco Galindo*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Máster en bioinformática y bioestadística*
Área del Trabajo Final:	M0.137 – TFM – Estadística y Bioinformática 38 Aula 1
Idioma del trabajo:	*Español*
Palabras clave	*RNA-seq, DESeq2, β-caroteno*

Resumen:

La desnutrición afecta a millones de personas, principalmente debido a que los cultivos críticos para la seguridad alimentaria contienen bajos niveles de micronutrientes. La yuca (*Manihot esculenta* Crantz) es el alimento básico de más de 500 millones de personas y los programas de mejoramiento se han enfocado en aumentar los niveles de β-caroteno en las raíces. Con el fin de apoyar los procesos de mejoramiento, el objetivo de este trabajo consistió en identificar la expresión y comprender que proporción del genoma se activa en la producción elevada de este compuesto. Para ello, seis genotipos de una familia de segregación, y los dos progenitores se escogieron para llevar a cabo la secuenciación del RNA. La plataforma bioinformática para este estudio incluyó: Mapeo con HISAT2, ensamblaje del transcriptoma con StringTie, cuantificación de la expresión con Kallisto, expresión diferencial con DESeq2 y DEGreport para mostrar los resultados. Finalmente, la anotación, función y regulación se obtuvo a partir del portal PlantRegMap. Los resultados obtenidos mostraron que las regiones del genoma que se sobre-expresan durante la acumulación de β-caroteno están relacionadas con la respuesta a estrés por calor, por temperatura e intensidad lumínica y en caso de no acumularse β-caroteno se presentan una diversidad de rutas de síntesis activas en la que se destaca la síntesis de esteroles. Adicionalmente 3 genes reguladores fueron identificados en el transcriptoma de raíz y 42 genes para el transcriptoma de hoja. El uso de la información generada en este estudio presenta múltiples implicaciones para el programa de mejoramiento y mejora la comprensión de la acumulación del β-caroteno en la raíces de yuca.

Abstract:

Malnutrition affects thousands of millions of people in the world, because of critical crops to food security contain low levels of micronutrients. Cassava (*Manihot esculenta* Crantz) is the staple food of more than 500 million people in the world and the breeding programs have been focused in increase the β-carotene level in roots. To support the breeding processes, this research aimed to identify differential expression in genotypes with different levels of β-carotene in roots and understand what genome proportion is activated during high production of this compound. To accomplish this goal, six genotypes from a segregating family and two parental were chosen to conduct RNA sequencing. The bioinformatics pipeline for this research included: HISAT for mapping, StringTie for transcriptome assembly, Kallisto for expression quantification, DESeq2 for differential expression analysis and DEGreport to obtain a graphical view of the results. The annotation, function and regulation was obtained from PlantRegMap. The results showed that over expressed genome regions during β-carotene accumulation are associated to heat stress, temperature and luminous intensity, when there is no β-carotene accumulation, there are other activated synthesis pathways, mainly the sterole synthesis. Additionally, three regulator genes were identify in the root transcriptome and 42 genes in leaf transcriptome. The impact of the information from this research has several implications for the Cassava Breeding Program and improve the comprehension of the β-carotene accumulation in cassava roots.

Índice

Lista de Figuras

Lista de Tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

La vitamina A está presente en productos de origen animal y vegetal, encontrándose en estos últimos como el precursor caroteno, el cual se convierte en vitamina A en el cuerpo humano. Existen varios carotenoides en las plantas, pero el más importante para la nutrición humana es el β-caroteno, que se convierte en vitamina A por acción enzimática en la pared intestinal.

La carencia de Vitamina A afecta el sistema inmunológico de aproximadamente el 40% de los niños menores de cinco años en los países en desarrollo, anualmente supone la muerte de un millón de niños, produce enfermedades oculares como la xeroftalmía, queratomalacia y manifestaciones oftálmicas graves que producen la destrucción de la córnea y ceguera, principalmente en niños de corta edad (UNICEF).

A pesar de que alimentos como aceites de hígado de pescado, yema de huevo y productos lácteos presentan altos niveles de vitamina A, en países en desarrollo el 80% de la ingesta de este micronutriente depende del consumo del caroteno de alimentos de origen vegetal. Se han planteado varias estrategias para aumentar el consumo de la vitamina A, principalmente, el enriquecimiento de los alimentos de consumo regular (Latham, 2002).

La yuca (*Manihot esculenta* Crantz) es el alimento básico de más de 500 millones de personas en el trópico. Es una especie de la familia Euphorbiaceae que se caracteriza por la capacidad de almacenar almidón en sus raíces y que le da su valor económico y estatus de planta cultivada. Aunque el sistema radicular presenta una baja densidad de raíces, la capacidad de penetración profunda hace que la planta pueda soportar periodos prolongados de sequía, lo que ha favorecido su cultivo en zonas tropicales, en suelos ácidos e infértiles (Ospina y Ceballos, 2002).

Debido a que la yuca presenta un alto consumo en regiones donde la vitamina A esta ausente en la dieta, ha entrado a formar parte de la estrategia para mejorar la calidad nutricional de los cultivos de importancia mundial para la seguridad alimentaria. La biofortificación es un proceso llevado a cabo por los programas de mejoramiento con el objetivo de incrementar el valor nutricional (minerales, proteínas y vitaminas). El incremento en los niveles de la provitamina A o β-caroteno mediante el mejoramiento convencional ha sido alentador (subió de 2µg/g a 22µg/g), sin embargo, los requerimientos nutricionales para este micronutriente son más elevados. Es por esta razón que identificar los genes que se asocian con su producción, y más aún con su acumulación, permitirá optimizar los procesos de mejoramiento y dirigir la selección de materiales mejorados en un menor tiempo y con mejores resultados.

Con este propósito, dirigimos nuestra investigación en el área de la bioinformática como ciencia integradora del campo de la biología molecular moderna con el análisis de gran cantidad de datos mediante las tecnologías y herramientas computacionales, que ha permitido el gran avance en el

conocimiento de las ciencias "ómicas" incluyendo la genómica, transcriptómica, proteómica y las múltiples interacciones presentes en estas áreas.

En particular, la transcriptómica que analiza aquellas secuencias de ARN que han sido transcritas a partir de ciertas regiones de ADN debido a condiciones biológicas y ambientales particulares. Por lo anterior, se seleccionó la tecnología del RNA-seq para obtener el perfil de expresión de genotipos contrastantes en la acumulación de β-caroteno, a partir de tejido foliar y radicular, que permitan entender los procesos biológicos y moleculares que regulan esta característica.

1.2. Objetivos del Trabajo

Objetivos Generales

- ✓ Identificar las regiones codificantes asociadas a la producción de β-caroteno en yuca *(Manihot esculenta)*.
- ✓ Caracterizar las rutas metabólicas activas en la producción de β-caroteno.

Objetivos Específicos

- ✓ Evaluar la expresión génica en genotipos de yuca con diferentes niveles de β-caroteno.
- ✓ Seleccionar los genes que presenten expresión diferencial significativa entre los grupos de individuos contrastantes fenotípicamente.
- ✓ Anotar función y ruta metabólica asociada a los genes diferencialmente expresados.

1.3. Enfoque y método seguido

La tecnología de secuenciación de nueva generación del transcriptoma (RNA-seq) permite cuantificar y perfilar los niveles de expresión génica. Para lo cual se ha desarrollado una gran cantidad de herramientas y técnicas bioinformáticas para el control de calidad, pre-procesamiento, alineamiento, análisis cuantitativo y expresión diferencial (Tabla 1).

Tabla 1: Soluciones bioinformáticas para el análisis transcriptómico

Pasos del Análisis	Programa	Descripción
Control de calidad	Fastqc (Van Verk *et al*., 2013)	Calidad de las lecturas, distribución por nucleótido, contenido GC, secuencias sobre representadas y frecuencia de k-meros.
Mapeo	TOPHAT (Trapnell *et al*., 2009)	Utiliza Bowtie2 para alinear lecturas a un genoma de referencia. Permite ensayos hebra-específico. Identificación de genes novedosos.
	HISAT2 (Kim D. *et al*., 2015)	Alinea las lecturas de RNA-seq a un genoma y descubre los sitios de empalme de la transcripción. Utiliza menos memoria y es mucho más rápido.
Ensamblaje de transcritos	Cufflinks (Trapnell *et al*. 2010)	Utiliza el mapa contra el genoma para ensamblar las lecturas en transcritos. Abundancias normalizadas por FPKMs
	StringTie (Pertea et al., 2016)	Ensambla los alineamientos en transcritos completos y parciales, creando múltiples isoformas según sea necesario, y estima los niveles de expresión de todos los genes y transcritos.
Cuantificación	Kallisto (Bray *et al*., 2016)	Cuantifica la abundancia de expresión a nivel de isoforma de manera rápida y precisa. Utiliza un pseudoalineamiento.
Detección DEGs	Cuffdiff (Trapnell *et al*., 2013)	Identifica los transcritos diferencialmente expresados y revela el empalme diferencial y los cambios de preferencia del promotor.
	DESeq2 (Love *et al*., 2014)	Distribución binomial negativa para controlar la sobre dispersión. Utiliza un Fold Change ajustado y estimaciones del error estándar, mejorando la detección de los DEGs.
	SAMseq (Li and Tibshirani 2013)	Método no paramétrico Baja tasa de falsos positivos. Necesita 4 réplicas biológicas
Reporte	CummeRbund (Goff *et al*. 2013)	Permite el análisis, manipulación y visualización de los archivos de salida de Cuffdiff .
	DEGreport (Pantano, 2017)	Procesa QCs, figuras y análisis después de la expresión diferencial. Recibe diferentes formatos.

DEG (Differential expression Genes)
FPKMs (Fragments Per Kilobase of Transcript per Million Mapped Reads)

Los programas y herramientas seleccionadas para este estudio fueron FastQC para el control de calidad de las lecturas, HISAT para mapear las lecturas al genoma de referencia de *Manihot esculenta* v.6.0 disponible en Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). El ensamblaje de los transcriptomas (hoja, raíz, hoja-raíz) se llevó a cabo con StringTie y la cuantificación de la expresión con Kallisto. La selección de estos programas está basada en la

rapidez de ejecución, el poco requerimiento de memoria, así como la sensibilidad y precisión de sus algoritmos.

El análisis de expresión diferencial se realizó con DESeq2 principalmente porque realiza un análisis sobre los genes y no sobre los transcritos, dado que la anotación del genoma de yuca no está completa. DEGreport se ha escogido para mostrar los resultados de los genes diferencialmente expresados basados en la media del "*fold changes*" y la variabilidad de cada gen seleccionado. Finalmente se relacionaron los genes sobre expresados con bases de datos mediante el portal PlanRegMap para evaluar función y asociar a rutas metabólicas.

1.4. Planificación del Trabajo

Para la ejecución de los programas desarrollados para la plataforma Linux se tuvo acceso al servidor CASSFE, que pertenece a la unidad de bioinformática del Centro Internacional de Agricultura Tropical (CIAT). Los paquetes estadísticos disponibles en Bioconductor fueron ejecutados a través de la interfaz de R-Studio, y el análisis de anotación y enriquecimiento se llevó a cabo mediante el portal PlantRegMap (Jin *et al*., 2016).

Estos recursos computacionales se utilizaron para llevar a cabo las siguientes tareas:

- ✓ Limpieza de lecturas
- ✓ Mapeo
- ✓ Ensamblaje del transcriptoma
- ✓ Cuantificación de la expresión génica
- ✓ Análisis de expresión diferencial
- ✓ Función y asociación con rutas metabólicas

Las actividades asociadas a las tareas anteriormente mencionadas se desarrollaron de acuerdo con el siguiente cronograma:

ACTIVIDADES	Inicio programado	Duración semanas	Inicio Real	Duración real	% Procesado
Limpieza de las lecturas	1	1	1	1	100%
Mapeo	2	1	2	1	100%
Ensamblaje del transcriptoma	3	1	3	1	100%
Cuantificación de la expresión	4	1	4	1	100%
Análisis de expresión diferencial	5	4	5	4	100%
Gráficas y fíguras y listados de DEGs	9	2	9	2	100%
Función de genes	11	3	11	3	100%
Asociación a rutas metabólicas	14	2	14	2	100%
Revisión bibliográfica	1	15	1	15	100%

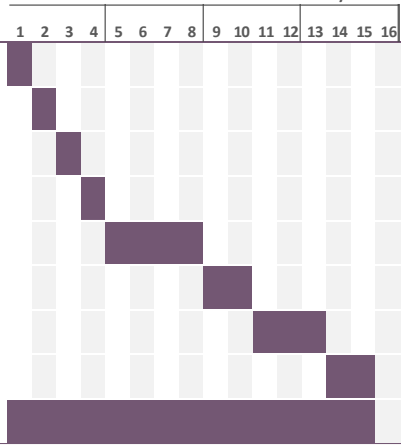

Figura 1: Cronograma de actividades

1.5. Resultados obtenidos

- Con el fin de desarrollar un análisis holístico, se ensamblaron tres transcriptomas: para raíz, para hoja, y para raíz y hoja.

- Para los transcriptomas obtenidos se identificaron los genes diferencialmente expresados en cada uno de los grupos fenotípicos.

- A partir de los genes diferencialmente expresados se identificaron los procesos biológicos activos, presentes en la acumulación de β-caroteno en las raíces de yuca.

1.6. Resumen de los capítulos de la memoria

Capítulo 2: revisa la teoría y la fundamentación de la especie de estudio, los programas de análisis utilizados, los métodos estadísticos relevantes, y finalmente se exponen las referencias bibliográficas que utilizaron estas metodologías en el desarrollo de trabajos similares.

Capítulo 3: expone la siguiente metodología: i.) diseño experimental, ii.) selección de material vegetal, iii.) extracción y secuenciación del RNA, iv.) ensamblaje de transcriptomas, v.) análisis de expresión diferencial, y vi.) análisis de enriquecimiento y regulación.

Capítulo 4: presenta los resultados obtenidos para cada uno de los ítems de la metodología.

Capítulo 5: discute los resultados obtenidos y plantea las perspectivas y el impacto de este trabajo.

Capítulo 6: expone las conclusiones y recomendaciones.

Capítulo 7: lista las referencias bibliográficas utilizadas.

2. Teoría y Fundamentación

2.1. La yuca

2.1.1. Taxonomía de la yuca

Reino: Plantae
División: Magnoliophyta
Clase: Magnoliopsida
Orden: Malpighiales
Familia: Euphorbiacea
Género: *Manihot*
Especie: *Manihot esculenta*

2.1.2. Características de la yuca

Manihot esculenta Crantz es un arbusto leñoso perenne, monoico, con 36 cromosomas (2n=36) y un tamaño de genoma de 740 Mb. Las hojas de la yuca son simples, compuestas por la lámina foliar y el peciolo, la lámina foliar es palmeada y profundamente lobulada (Figura 2A). Los tallos presentan ramificación simpodial y son esenciales en el cultivo debido a que la yuca se propaga vegetativamente por medio de porciones de tallo las cuales se conocen como estacas (Figura 2D). Está estrategia de propagación clonal es la más utilizada por los agricultores, ya que permite conservar las características de interés de una variedad. La yuca también se puede propagar a través de semilla sexual que es ampliamente utilizada en los programas de mejoramiento con el fin de obtener nuevas variedades (Figura 2E), (Ospina y Ceballos, 2002).

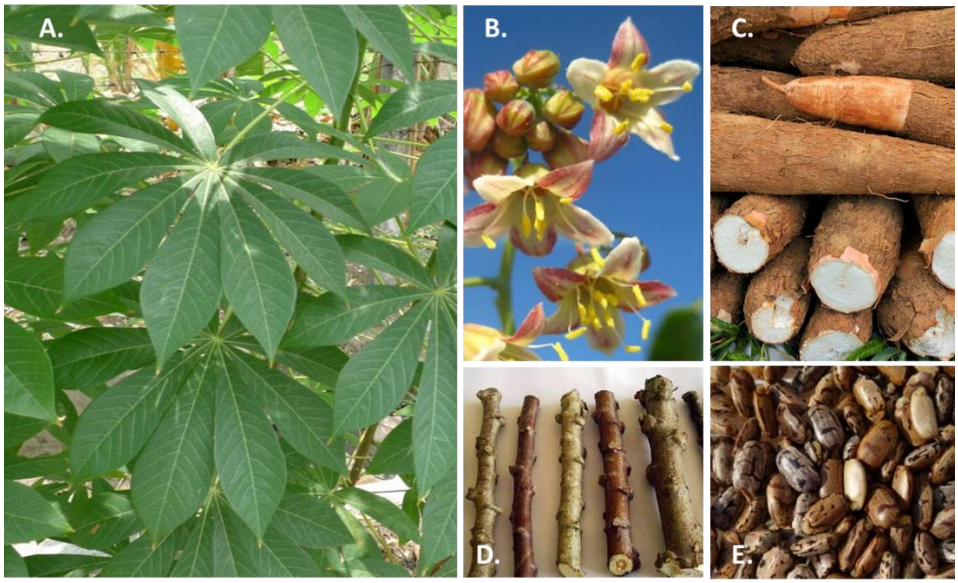

Figura 2: Secciones de la planta de yuca. A. Tejido foliar. B. Frutos y florescencia. C. Raíces de almacenamiento. D. Cortes longitudinales de secciones de tallo para ser empleadas como estacas para propagación clonal. E. Semilla botánica.

El sistema radicular presenta dos tipos de raíces, fibrosas y tuberosas, las cuales presentan baja densidad y tienen la capacidad de profundizar el subsuelo, lo que explica su adaptación a periodos prolongados de sequía (El-Sharkawy 2004). La raíz sigue su proceso de acumulación de almidón hasta la época de cosecha, alrededor de los 12 meses, dependiendo de la variedad (Figura 2C), (Ospina y Ceballos, 2002).

Durante los tres primeros meses después de la siembra, la formación de tejido foliar tiene prioridad sobre el desarrollo de las raíces tuberosas. Una vez pasa este tiempo, la formación de hojas y raíces tuberosas se realiza simultáneamente. La producción de hojas dura aproximadamente seis meses más, tiempo en el cual la planta ajusta su eficiencia fotosintética a una variedad de factores como la temperatura, la intensidad de luz, el estado fisiológico, la apertura estomática y a factores genéticos (Ospina y Ceballos, 2002).

Las raíces de yuca se caracterizan por almacenar grandes cantidades de almidón y niveles muy bajos de proteínas, grasas y vitaminas. Sin embargo, algunas accesiones provenientes principalmente del Amazonas (Brasil-Colombia) presentan altos niveles de provitamina A, que han sido utilizados por el programa de mejoramiento del CIAT durante la última década para desarrollar variedades de yuca con altos niveles de β-caroteno, hierro y zinc, aprovechando la variación genética característica de este cultivo (Figura 3) (Ceballos *et al*., 2013). El reto del mejoramiento ha radicado en incrementar la poca materia seca y pobres propiedades culinarias característica de los clones que presentan alto contenido de β-caroteno en las raíces (Moorthy *et al*., 1990).

Figura 3: Harinas de raíz de yuca provenientes de diferentes variedades , exhibiendo niveles diferenciales en su contenido de β-caroteno. La imagen en la esquina superior izquierda corresponde a la variedad con el nivel más alto de β-caroteno. La imagen en la esquina inferior derecha corresponde a la variedad con el nivel más bajo de contenido de β-caroteno.

2.1.3. Importancia del cultivo de la yuca

El cultivo de la yuca está ampliamente distribuido por toda la región tropical y es de gran importancia socioeconómica para agricultores y consumidores de pocos recursos económicos (Figura 4). Se cultiva en 103 países y representa la cuarta fuente más importante de carbohidratos después del arroz, la caña de azúcar y el maíz, siendo un aporte a la alimentación de más de 1000 millones de personas en Asia, África y América Latina. Para el año 2016 se reportó que el área mundial cultivada fue de 23.482.052 de hectáreas con una producción de 277.102.564 toneladas (FAOSTAT, 2016).

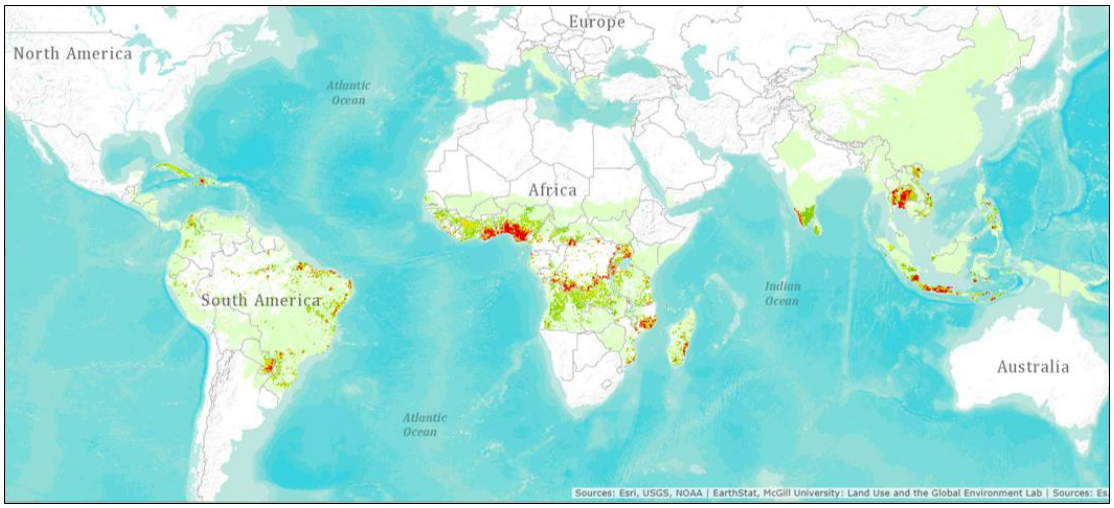

Figura 4: Area cultivada de yuca disponible en el portal del RTB Maps v1. El color rojo representa la mayor cantidad de hectáreas cultivadas.

2.2. Carotenoides

Los carotenoides son una familia de compuestos químicos que pertenecen a la familia de los terpenos, en general formados por 8 unidades de isopreno, cuya síntesis se da a partir del isonpentenil pirofosfato. Se caracterizan por sus cadenas cortas hidrocarbonadas y por su coloración roja, naranja y amarilla, la cual se debe concretamente a la oscilación de los electrones a lo largo de la cadena hidrocarbonada insaturada. Los carotenoides están presentes en plantas, animales, hongos, algas y bacterias donde cumplen diferentes funciones: son responsables del color de las flores y los frutos, hacen parte de plumas y picos en algunas aves, del exoesqueleto de algunos crustáceos y de la piel de algunos peces. Sin embargo, los carotenos en las plantas cumplen funciones fisiológicas esenciales en la fotosíntesis, principalmente en los procesos de captación de luz, por la presencia de 7 o más enlaces dobles conjugados, de fotoprotección, disipación de excesos de energía y la desactivación del oxígeno singlete, entre otras (Meléndez *et al.,* 2007).

2.2.1. β-caroteno

Entre los carotenoides, el β-caroteno es el compuesto más común en las plantas. Se distingue por presentar dos anillos beta en ambos extremos de la molécula (Figura 5). Su biosíntesis tiene lugar en la membrana que rodea al cloroplasto siguiendo la ruta metabólica del isoprenoide, la cual también está implicada en la producción de moléculas responsables del aroma y compuestos como los esteroides y el caucho (Figura 6) (DellaPenna y Pogson, 2006).

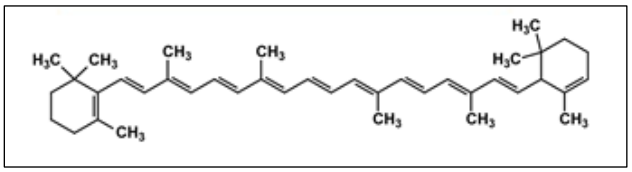

Figura 5: Estructura química de la molécula de β-caroteno.

La formación del β-caroteno se da a partir del compuesto geranil geranil pirofosfato que en las plantas puede formarse a partir de dos rutas metabólicas, en el citoplasma vía mevalonato (MEV) o en los plastidios vía 2C-metil-D-eritritol 4-fosfato (MEP). Los procesos bioquímicos de esta ruta incluyen la adición de unidades de cinco carbonos o múltiplos, seguido de reordenamientos moleculares, ciclizaciones y la adición de grupos funcionales (DellaPenna y Pogson, 2006).

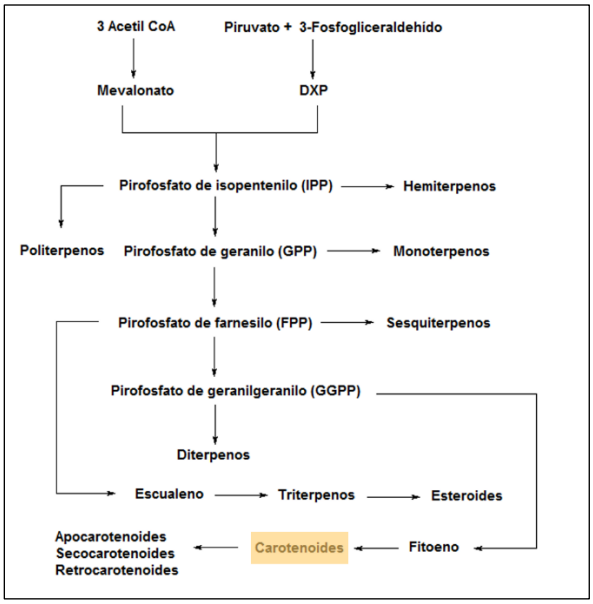

Figura 6: Esquema general de la ruta metabólica de síntesis de carotenos vía ruta del isoprenoide. tomado de (DellaPenna y Pogson, 2006)

El primer paso específico de la síntesis de β-caroteno es la condensación de dos moléculas de geranil geranil pirofosfato, mediado por la enzima fitoeno sintasa para formar la molécula de fitoeno, la cual al presentar cuatro desaturaciones sucesivas catalizadas por las enzimas fitoeno desaturasa y caroteno desaturasa, forma la molécula de licopeno. Este al ser modificado por la enzima β-licopeno ciclasa forma la molécula de 40 carbonos que corresponde al β-caroteno (Figura 7) (DellaPenna y Pogson, 2006).

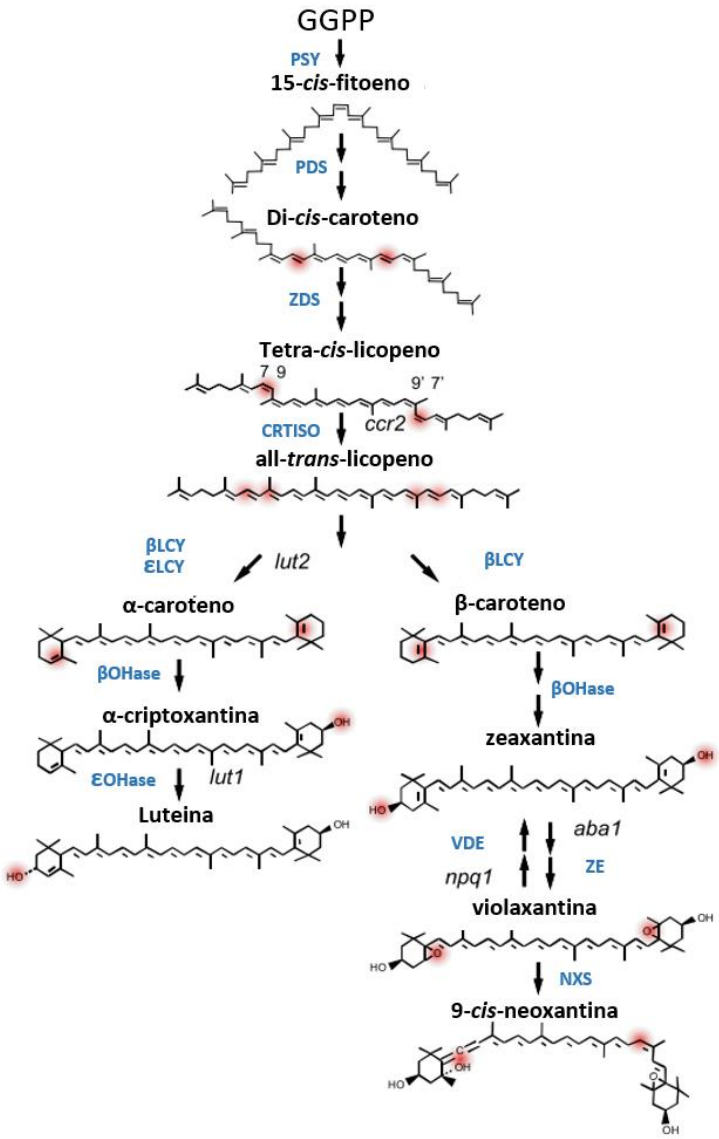

Figura 7: Esquema de la ruta metabólica de la biosíntesis del β-caroteno y otros carotenoides a partir del geranil, geranil pirofostato (GGDP). Adaptado de DellaPenna y Pogson, 2006. En azul se indican las enzimas que catalizan cada reacción. PSY (Fitoeno sintasa), PDS (Fitoeno desaturasa), ZDS (Caroteno desaturasa), CRTISO (Caroteno isomerasa), βLCY (β-licopeno ciclasa), εLCY (ε-licopeno ciclasa), βOHase (β-caroteno hidroxilasa), εOHase (ε-carotneo hidroxilasa), VDE (violaxantina de epoxidasa), ZE (zeaxantina hipoxidasa), NXS (neoxantina sintasa).

El β-caroteno en las plantas presenta varias funciones, i) tiene un papel en la fotoprotección de los fotosistemas presentes en el cloroplasto debido a sus cualidades antioxidantes para evitar la formación de oxígeno reactivo durante las reacciones químicas, dado que la formación de este tipo de oxígeno puede limitar la conversión exitosa de la luz en energía química o alterar el equilibrio químico de una célula vegetal. ii) En la fotosíntesis al absorber la luz necesaria para el proceso, debido a que presenta el espectro de absorción más amplio de luz, permitiendo la máxima conversión de energía. iii) La pigmentación con matices amarillos y anaranjados de diferentes partes de la planta como flores, frutos, tallos y raíces (Nissar *et al*, 2015). En el caso de la pigmentación de flores y frutos atraen insectos y favorecen la polinización y dispersión de semillas (Yuan *et al., 2015*).

También se ha establecido la importancia de este compuesto en la salud humana, debido a que el β-caroteno es un precursor de la vitamina A. La tasa de absorción de este carotenoide se estima entre el 9 y el 22%, está restringida al duodeno del intestino delgado y depende de la presencia del receptor de la proteína de membrana *scavenger* clase B (SR-B1). Una molécula de β-caroteno puede ser escindida por la enzima intestinal β, β-caroteno 15,15`-monooxigenasa en dos moléculas de vitamina A (Biesalski *et al*.,2007).

2.3. RNA-seq

En un principio, la tecnología de los *microarrays* de expresión permitió obtener el nivel de expresión de los genes, transcritos y exones de los organismos de estudio, que componen el transcriptoma. Sin embargo, la tecnología de la secuenciación de RNA (RNA-seq), permite la detección de los genes expresados estén anotados o no y dado que se basa en los diferentes métodos de secuenciación de alto rendimiento y de nueva generación no depende de oligosNT previamente desarrollados, lo que permite extraer el perfil de expresión del transcriptoma en corto tiempo y con mayor cobertura (Zhao *et al*., 2014).

Un transcriptoma es el conjunto completo de transcritos y el número de copias de cada uno de ellos, para una etapa del desarrollo específico, un tipo de tejido o un estado fisiológico particular. La comprensión del transcriptoma permite entender y develar los elementos regulatorios génicos, componentes moleculares de células y tejidos, funciones biológicas, descifrar señales ambientales y su efecto en la expresión del genoma, así como también, evidenciar la coexpresión, relacionar redes biológicas, la biología de los sistemas y percibir eventos epigenéticos. (Jazayeri *et al*.,2015)

Se ha establecido que la tecnología del RNA-seq se compone de 3 partes para su ejecución (Figura 8). La primera parte se lleva a cabo en el laboratorio y se ha denominado "*wet-lab*" en la cual se realiza principalmente el diseño experimental, la extracción del RNA y la preparación de las librerias para secuenciar. La segunda parte determina la plataforma de secuenciación y la obtención de las lecturas y este paso se conoce como "*in equipo*". Finalmente se tiene el componente *"in silico"*, en el cual se ejecuta el análisis de los datos, que incluye el control de calidad de lecturas, mapeo, ensamblaje del transcriptoma con o sin genoma de referencia y los análisis de expresión diferencial (Jazayeri *et al*., 2015).

Figura 8: Esquema general de los pasos a seguir en la técnica de RNA-seq.

2.4. Programas y herramientas bioinformáticas

Los siguientes programas han sido desarrollados con el fin de realizar los diferentes componentes del análisis transcriptómico:

2.4.1. FastQC

Dada la gran cantidad de secuencias que se pueden obtener con las nuevas metodologías de ultrasecuenciación, es necesario evaluar la calidad de las mismas antes de continuar con los análisis y generar asunciones biológicas. FastQC es una herramienta que permite realizar un control de calidad sobre las secuencias para detectar problemas que se originan en el secuenciador o en las librerías (Van Verk *et al*., 2013).

FastQC admite archivos tipo FastQ, Casava FastQ, colorspace FastQ, FastQ comprimido, SAM, BAM y realiza el control de calidad mediante una serie de análisis por módulos, incluyendo:

- Estadística básica: número total de secuencias, longitud de las secuencias y contenido de GC.
- Calidad de la secuencia por base: permite evidenciar la calidad promedio de cada base por posición.
- Calidad de secuencia: relaciona el número total de las secuencias con la puntuación de la calidad media de la secuencia.
- Contenido de secuencia por base: Esta basado en el porcentaje de bases llamadas de cada uno de los cuatro nucleótidos para cada posición de la secuencia.
- Contenido de GC por secuencia: compara la distribución observada del contenido de GC con una distribución teórica.

- Contenido de N por base: determina los puntos de la secuencia en la que no puede ser llamado una base.
- Nivel de duplicación en la secuencia: detecta aquellas lecturas que están varias veces en el conjunto de secuencias.
- Secuencias sobre-representadas: detecta las secuencias que son iguales o mayores al 0.1% del total de las lecturas.
- Contenido de adaptadores: busca la secuencia de los adaptadores con los que fue realizada la librería.
- Contenido Kmer: determina el número de un motivo a lo largo de la secuencia dado un valor k.

FastQC produce un archivo html donde presenta los resultados de cada uno de los análisis en forma gráfica y adicionalmente cuenta con un sistema de etiqueta para valorar si un análisis pasó, falló o necesita ser revisado. Sin embargo, es necesario ser cauteloso con este sistema de etiqueta, debido a que los parámetros de corte han sido desarrollados con base en asunciones que aplican sólo para datos de secuencia de ADN de genoma completo.

2.4.2. HISAT2

De su acrónimo en inglés "*Hierarchical indexing for spliced alignment of transcripts*", es un programa de alineamiento altamente eficiente para mapear lecturas provenientes de secuenciación de nueva generación principalmente de secuenciación de RNA. El mapeo puede realizarse contra un sólo genoma de referencia o contra un conjunto de genomas de referencia con requerimientos bajos de memoria.

Utiliza Bowtie 2 para implementar un esquema de indexación el cual está basado en la transformadas de Borrows-Wheeler y el índice de Farragina-Manzini (FM). HISAT utiliza 2 tipos de índices, uno general que representa el genoma completo y una serie de índices locales que en conjunto cubren todo el genoma y aumentan la velocidad en las extensiones de los alineamientos, lo que permite procesar genomas de cualquier tamaño, incluyendo aquellos genomas que superan los 4 billones de bases en un tiempo mucho menor que el soportado por TOPHAT y otros programas de alineamientos conocidos.

Los algoritmos desarrollados para HISAT aumentan la sensibilidad en la alineación debido a que son específicos para manejar diferentes lecturas que abarcan intrones. Combinado con la indexación jerárquica aumentan la velocidad e igualan o exceden en muchos casos la precisión de los mejores alineadores disponibles.

HISAT procesa cada lectura por separado, identifica las posibles regiones de origen en el genoma objetivo mediante el uso del índice FM global que restringe las regiones a unos pocas y selecciona uno de los índices locales que determina el lugar del alineamiento. En caso de que la librería se haya construido utilizando pares de secuencias, la alineación se hace por separado y si alguna de las lecturas no se alinea mediante el uso de los índices, la otra lectura sirve de punto de referencia para el alineamiento.

La ventaja de tener dos tipos de índices se basa en el hecho, de que la memoria central de una computadora normal presenta dos tipos de memoria de acceso, aleatorio (RAM) y memoria caché. El índice FM global normalmente se ejecuta utilizando la memoria RAM lo cual hace lento el proceso; sin embargo, los índices locales al ser más pequeños se ejecutan desde la memoria caché aumentado la velocidad en el alineamiento (Kim D. *et al*. 2015).

2.4.3. StringTie

Es un método computacional para el ensamblaje de transcriptomas y estimación de niveles de expresión en simultáneo. Identifica entre el 36 al 60% más transcritos que el siguiente mejor ensamblador (Cufflinks). Para la reconstrucción precisa de genes y mejores estimaciones de los niveles de expresión, StringTie primero agrupa las lecturas en clústeres, luego identifica los transcritos presentes mediante un gráfico de empalme para cada clúster y finalmente, para cada transcrito crea una red de flujo separada para estimar su nivel de expresión utilizando un algoritmo de flujo máximo. En la Figura 9 se observa la comparación entre los tres programas de ensamblaje más conocidos y su flujo de trabajo. El transcriptoma puede ensamblarse con o sin genoma de referencia, sin perder precisión en la reconstrucción de genes (Pertea *et al*., 2016).

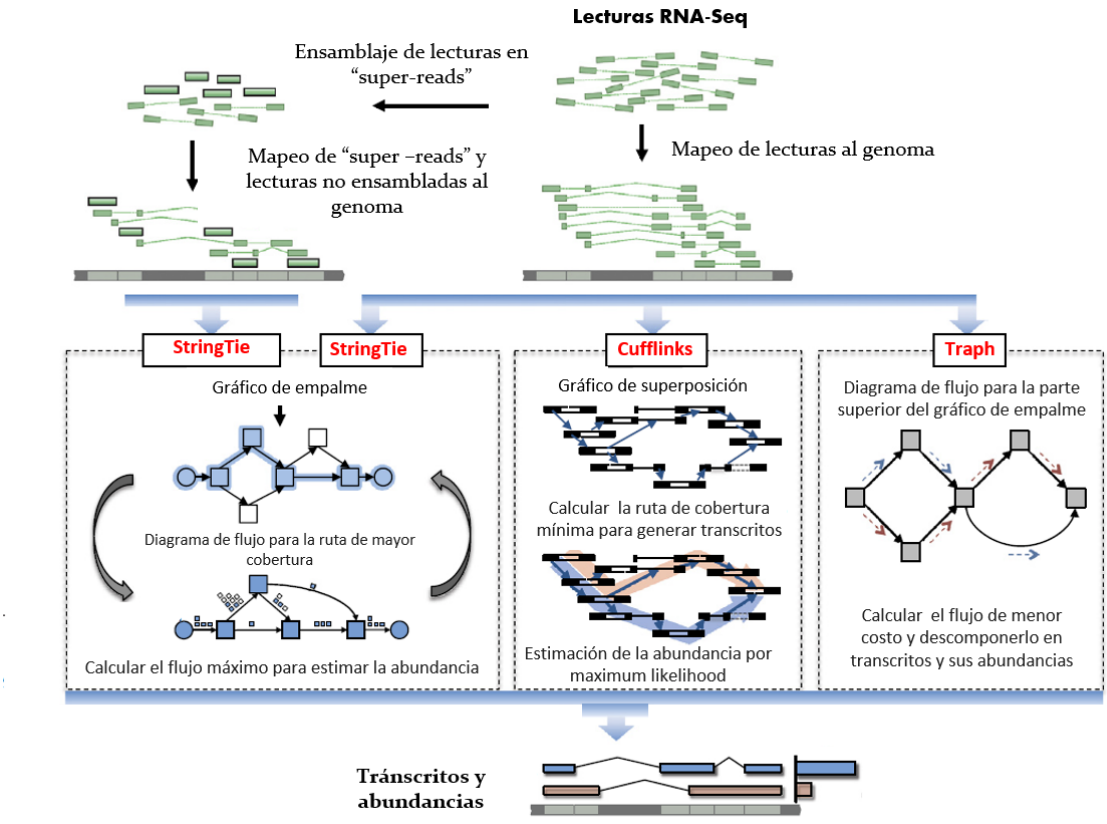

Figura 9: Flujo de trabajo para el ensamblaje de transcriptomas con los programas StringTie, Cufflinks y Traph. Esta figura ha sido adaptada del artículo de Pertea *et. al.,* 2015.

2.4.4. Kallisto

Es un programa para cuantificar la abundancia de los transcritos a partir de los datos de RNA-seq. Está basado en un pseudoalineamiento para una rápida detección de compatibilidad entre las lecturas y los sitios diana del genoma, evitando los alineamientos de bases individuales. El pseudoalineamiento de las lecturas preserva la información clave necesaria para la cuantificación, la cual se lleva a cabo de manera rápida y confiable.

Dentro de las características que sobresalen de Kallisto frente a otros programas de cuantificación está el hecho que permite un *bootstrapping* eficiente que se obtiene mediante repeticiones del algoritmo EM, con lo cual se puede calcular con precisión la incertidumbre en las estimaciones de abundancia. Además de que la cuantificación se hace en corto tiempo, la construcción del índice e incluso la compilación del programa es fácil y rápida, lo que permite un análisis interactivo donde los datos no duran largo tiempo procesándose y no se requiere computación en la nube, lo que significa un análisis portátil, económico y seguro. (Bray *et al.,* 2016).

2.4.5. DESeq2

Es un método de análisis diferencial que realiza conteo de lecturas por gen, para evidenciar cambios sistemáticos a través de condiciones experimentales. DESeq2 es un paquete disponible en R/Bioconductor que permite un análisis más cuantitativo de datos comparativos de RNA-seq mediante el uso de estimadores reducidos para la dispersión y el LFC (*Logarithmic fold change*). También realiza estimaciones del error estándar, clasificación y visualización de genes, pruebas de hipótesis y la transformación logarítmica regularizada para la evaluación de la calidad y la agrupación de datos sobre-dispersados, logrando alta sensibilidad y precisión mientras controla la tasa de falsos positivos.

Los archivos de entrada en DESesq2 puede ser los archivos de abundancia obtenidos con programas de cuantificación de la expresión como Kallisto, Salmon o Sailfish, para luego crear un matriz de conteo a nivel de gen mediante el paquete tximport. De forma alternativa, también se puede utilizar una matriz de conteo de lecturas proveniente de otra fuente. La matriz (K) con los genes (i) en cada fila y las muestras (j) en cada columna. La entrada K_{ij} de la matriz indica el número de lecturas secuenciadas que han sido mapeados inequívocamente a un gen en una muestra. El análisis se realiza mediante un modelo lineal generalizado para cada gen. Para la modelización se utiliza una distribución binomial negativa con media μ_{ij} y la dispersión α_i. La media es tomada como una cantidad q_{ij} la cual es proporcional a la concentración de cDNA proveniente de genes en la muestra, escalada por un factor de normalización $s_{ij,}$ entonces ($\mu_{ij} = s_{ij}q_{ij}$). El uso de modelos lineales provee una gran flexibilidad, así como también, el análisis con diseños complejos (Love *et al*., 2014).

2.4.6. DEGreport

Es un paquete disponible en R/Bioconductor que genera un reporte HTML con QCs, figuras, análisis y listas de genes de acuerdo a la media del *fold change* y a la variabilidad de cada gen

seleccionado, una vez se tengan los resultados del análisis diferencial utilizando DESeq2 o alguna herramienta similar. DEGreport cuenta con 31 funciones documentadas que permiten realizar un análisis minucioso de los resultados de los análisis de expresión diferencial (Pantano, 2017). Entre las funciones que sobresalen están:

- ✓ CreateReport: Genera la matriz de conteo, con p-valores y el *fold change* de un análisis de expresión diferencial y crea un reporte para ayudar a detectar posibles problemas con los datos.
- ✓ DEGpatterns: Hace agrupación de genes usando el perfil de expresión. Esta función no calcula la diferencia significativa entre los grupos, por lo tanto, es necesario partir de una matriz filtrada que contenga sólo los genes que son significativamente diferentes.
- ✓ degResults: Realiza un reporte completo con gráficos generados utilizando la función ggplot2, a partir del análisis hecho en DESeq2.

2.4.7. PlantRegMap

Es un novedoso portal que permite acceder a los recursos de regulación y herramientas de análisis enfocado a plantas, consolidados en la base de datos PlantTFDB 4.0 que proporciona información comprensiva y de alta calidad de factores de transcripción y sus interacciones regulatorias específicas en 165 especies de plantas. La base de datos tiene tres métodos para generar las anotaciones. i) Un conjunto de motivos relacionados con factores de transcripción derivado de experimentos; ii) múltiples tipos de elementos regulatorios identificados a partir de las plataformas de secuenciación de nueva generación; iii) interacción regulatoria curada de la literatura e inferida a partir de motivos relacionados con factores de transcripción y elementos regulatorios. El portal cuenta con un conjunto de herramientas que permiten predecir la regulación y realizar el análisis de enriquecimiento, este último presenta en las opciones avanzadas la categoría de anotación GO que se quiere evaluar (proceso biológico, función molecular y componente celular), permite escoger el p-valor con el cuál se va hacer la selección de enriquecimiento y presenta una serie de opciones de formato para exportar los resultados (Jin *et al*., 2016).

3. Materiales y Métodos

3.1. Lugar de ejecución del proyecto e instituciones participantes

Este proyecto se desarrolló por completo en las instalaciones del Centro Internacional de Agricultura Tropical (CIAT) Cali, Colombia. Con la participación de los programas de Genética, Mejoramiento y el laboratorio de Calidad Nutricional de Yuca, bajo la dirección del líder del programa Luis Augusto Becerra Lopez-Lavalle y la tutoría de Lorena Pantano, profesor adjunto de la *Universitat Oberta de Catalunya* (UOC). Con la financiación del programa de *Harvest Pluss* y el programa *Root, Tubers and Bananas* (RTB).

3.2. Proceso RNA seq

3.2.1. Componente "*wet-lab*"

3.2.1.1. Material vegetal

El programa de mejoramiento de yuca del CIAT generó una familia de segregación a partir de parentales que presentan niveles contrastantes de β-caroteno, CR87 (niveles altos) y PER297 (niveles bajos). La F1 ha sido nombrada como GM905; los cruces entre dos hermanos (GM905-57 y GM905-60) dieron origen a las muestras utilizadas en este estudio. La progenie de este cruce fue evaluada mediante HPLC y NIRS para determinar el contenido de β-caroteno y otros metabolitos presentes en las raíces durante tres años. Las muestras seleccionadas para este trabajo se caracterizaron por presentar los niveles más bajos, intermedio y más altos de β-caroteno durante las tres evaluaciones (Figura 10).

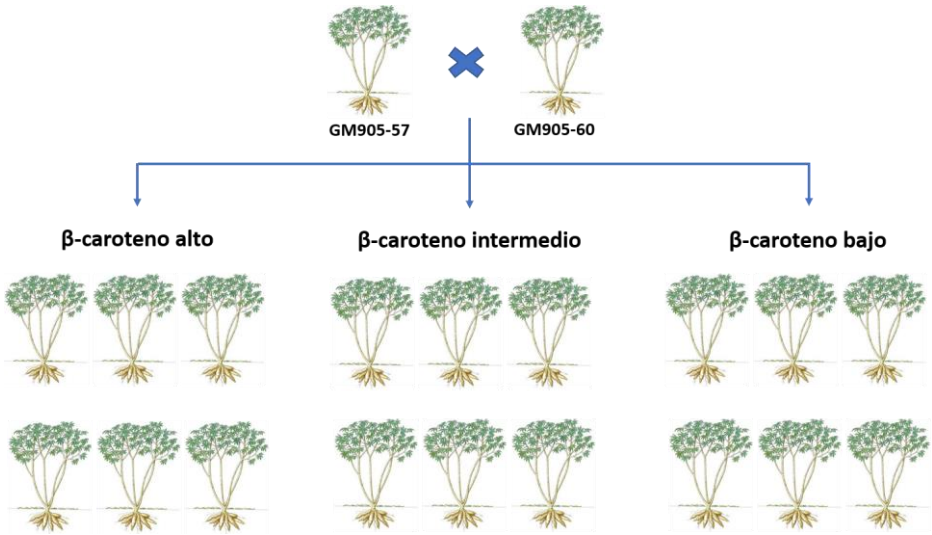

Figura 10: Diseño experimental para la selección de muestras utilizadas en este estudio.

3.2.1.2. Extracción de RNA y preparación de librerías

Muestras de raíz y hoja se colectaron en tres plantas diferentes de cada genotipo a los 7 meses después de la siembra, utilizando nitrógeno líquido para preservar la integridad del RNA. La extracción de RNA se llevó a cabo utilizando el protocolo descrito por Chang *et al*. 2013, y la verificación de la integridad del RNA se evaluó usando el kit de "RNA 6000 Nano" de Agilent con un bioanalizador Agilent 2100 (Agilent Technologies). Las muestras se secaron en un *speed vac* utilizando *RNA-stable* para preservarlas y a continuación se enviaron al Instituto de Genómica de Beijing (BGI) donde se confirmó su calidad y se prepararon las librerías usando el "*kit sample preparation TruSeq RNA*" de Illumina (Figura 11).

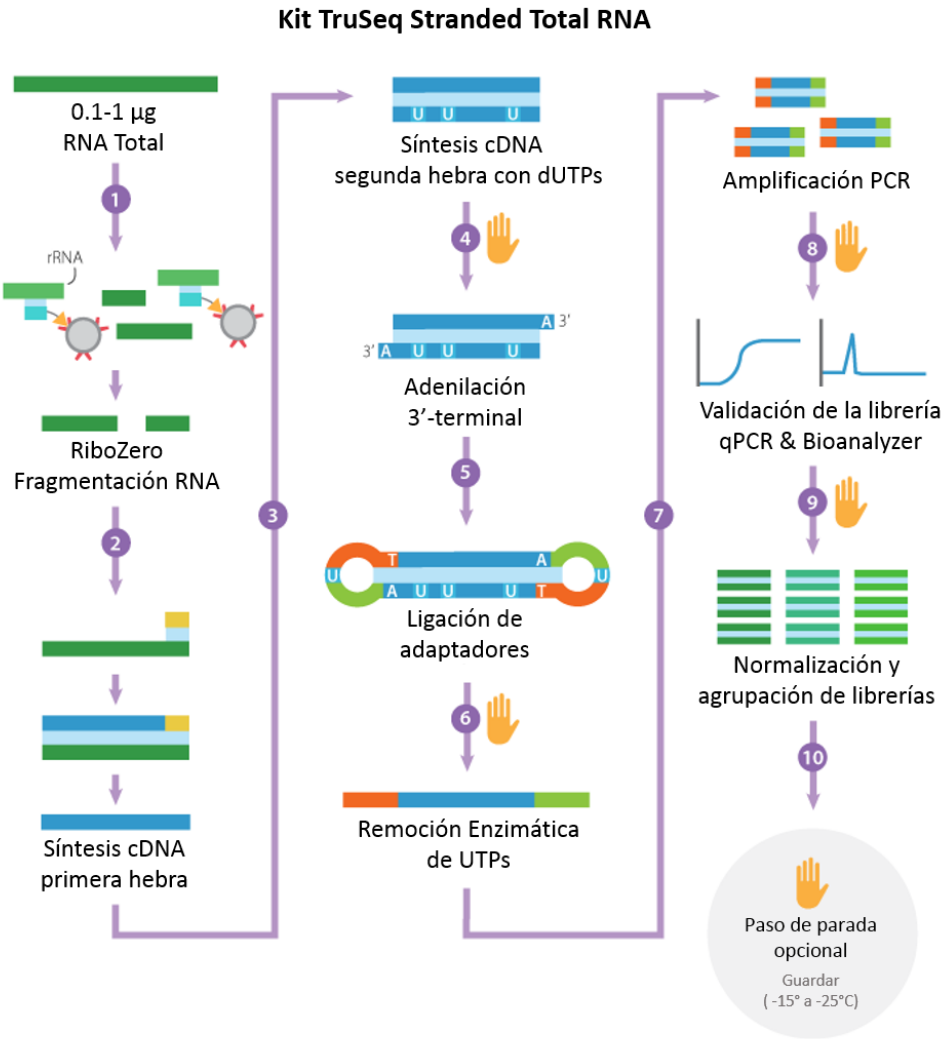

Figura 11: Esquema general del protocolo de preparación de librerías mediante el *kit sample preparation TruSeq RNA* de Illumina (adaptado del protocolo TruSeq Stranded Total RNA Kit).

3.2.2. Componente "*in equipo*"

3.2.2.1. Diseño y secuenciación

El tipo de librería fue *paired-end* para las muestras provenientes del tejido de hoja y raíz. Adicionalmente las muestras de hoja se corrieron en dos *lanes* de secuenciación denominadas L1 y L2. La secuenciación se llevó a cabo mediante la plataforma HiSeq 2000 de Illumina, con una longitud de secuenciación de 100pb.

3.2.3. Componente "*in sillico*"

Debido a que hay muestras de hoja y de raíz se decidió realizar tres transcriptomas. El primero a partir de los datos provenientes de raíz, el segundo a partir de los datos de hoja y finalmente un tercer transcriptoma que presenta los datos en conjunto de hoja y raíz. La Figura 12presenta el esquema general del análisis bioinformático de las secuencias provenientes de RNA-seq de este estudio. A continuación, se detalla el procedimiento y los programas utilizados en cada uno de los componentes de esta estrategia. Los *scripts* que contienen todos los parámetros de ejecución están disponibles en https://github.com/tmovalle/TFM_UOC.

3.2.3.1. Control de calidad de las lecturas

El primer paso involucra evaluar la calidad de la secuencia, el contenido de guanina (G) y citosina (C), la presencia de adaptadores, entre otros componentes. Este análisis se realizó mediante la herramienta FastQC (Van Verk *et al*., 2013) debido a que las lecturas provienen de la plataforma de secuenciación Illumina.

3.2.3.2. Mapeo

Una vez las lecturas han pasado el control de calidad es posible la identificación de transcritos mediante el mapeo de las lecturas del RNA-seq al genoma de referencia o al transcriptoma. La versión de HISAT2-2.1.0 se seleccionó para mapear las lecturas al genoma de referencia de la yuca Mesculenta_305_v.6.fa disponible en la base de datos de Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html).

El formato de los archivos de entradas fue FASTQ. y el *script* runMapping.sh presenta los detalles para su ejecución. Con el fin de cumplir los argumentos para correr HISAT2, fue necesario calcular el tamaño de los intrones para cada uno de los genes del genoma de la yuca utilizando el *script* en *Perl* llamado IntronSizes.pl disponible públicamente y un *script* en Phyton denominado IntronSizes.py. El primero busca los tamaños de los diferentes intrones en los genes y el segundo evalúa cual es el intrón más grande por gen y si encuentra un valor mayor lo guarda como el tamaño del intrón más grande.

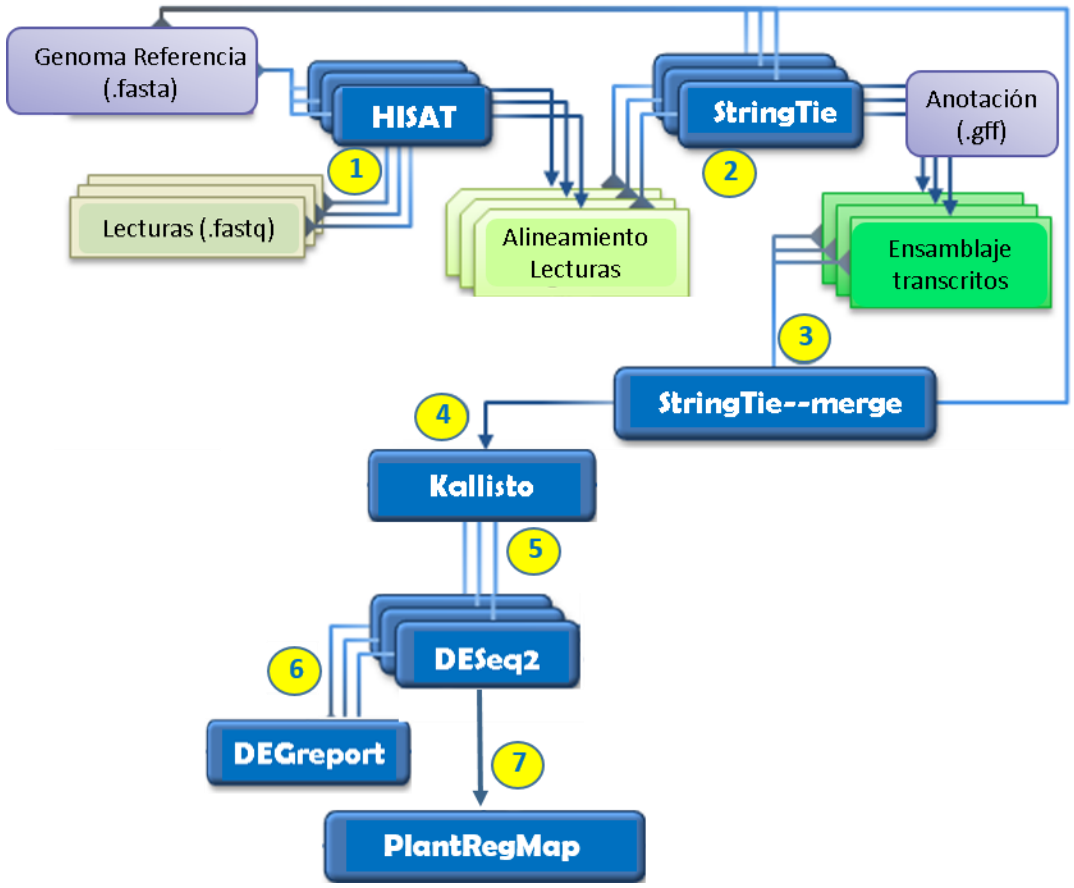

Figura 12: Estrategia de análisis "*in silico*" de las secuencias provenientes de RNA-seq.

3.2.3.3. Ensamblaje y cuantificación del transcriptoma

La versión de StringTie 1.3.3b se utilizó para el ensamblaje del transcriptoma, con los archivos tipo BAM provenientes del mapeo utilizando el *script* runStringtie.sh, el cual cuenta con una función *merge* para unir todos los archivos *.gtf de las diferentes muestras en un solo transcriptoma. Adicionalmente el *script* incorpora una pequeña función de Cufflinks (Trapnell *et al*. 2010) llamada "gffread" y sirve para transformar el transcriptoma de formato GTF, el cual solo presenta coordenadas de los transcritos, a formato FASTA, donde se obtiene las secuencias a partir del genoma de referencia.

La cuantificación de la expresión se realizó mediante el programa Kallisto v0.43.1 y el transcriptoma ensamblado con StringTie. La cuantificación se realizó para cada muestra utilizando el *script* runKallisto.sh.

3.2.3.4. Análisis de expresión diferencial

El análisis de expresión diferencial se realizó con el paquete DESeq2 a partir de los archivos de cuantificación de transcritos llamados abundance.tsv obtenidos con Kallisto. Los archivos *.tsv se

importaron mediante el paquete de R *tximport* y DESeq2 generó la matriz de conteo a nivel de gen. La matriz presenta conteos no normalizados ni escalados debido a que el modelo de DESeq2 corrige internamente el tamaño de la librería.

Las variables tipo factor a tener en cuenta en el modelo son el fenotipo que presenta tres niveles (alto, intermedio, bajo) y el tipo de tejido que presenta dos niveles (hoja, raíz). El uso de estas variables depende del transcriptoma analizado, la variable denominada tejido está presente únicamente en el análisis de expresión diferencial del transcriptoma que contiene los datos en conjunto de hoja y raíz. De las dos pruebas de hipótesis que presenta DESeq2 se utilizó la prueba de hipótesis LRT "*Likelihood Ratio Test*" debido a que la variable fenotipo presenta tres niveles y en uno de los transcriptomas se evalúan dos variables. LRT examina dos modelos, un modelo completo que presenta todos los términos y un modelo reducido donde uno de los términos del modelo completo se eliminó. El LRT utiliza un análisis de desviación (ANODEV), donde la varianza captura la diferencia del "*likelihood" entre* el modelo completo y el reducido.

La visualización de los resultados del análisis de expresión diferencial se llevó a cabo en el paquete de R DEGreport y la lista de los DEGs presentes en los diferentes niveles del fenotipo se obtuvo mediante la función degPatterns.

3.2.3.5. Anotación, función y regulación

El análisis de enriquecimiento de los DEGs obtenidos con DESeq2 se realizó en el portal PlantRegMap. Primero se hizo el análisis con todos los DEGs para cada transcriptoma y luego se evaluó de acuerdo con el fenotipo (alto, intermedio y bajo). El análisis de enriquecimiento se basó en las anotaciones GO disponibles para yuca y se evaluaron los términos asociados a i) componente celular, ii) función molecular y iii) proceso biológico que tuvo un p-valor menor a 0.01. Para el ruido de fondo (*background*) se contó con la referencia de 160 especies vegetales. Este mismo portal se utilizó para generar la red de regulación operante en la producción de β-caroteno.

4. Resultados

4.1. Selección material vegetal

En la familia de segregación GM3736 se evaluaron diferentes metabolitos de la ruta metabólica del β-caroteno durante tres años consecutivos, utilizando HPLC y NIRS. Adicionalmente, el contenido de materia seca, también se midió, dada su importancia a nivel de mejoramiento. Las evaluaciones mostraron la alta correlación de los metabolitos durante los tres años de evaluación, mostrando la importancia del componente genético en estás características. Igualmente, se observó la relación directa entre el fitoeno y el fitoflueno con el β-caroteno y la relación indirecta entre materia seca y β-caroteno (Figura 13).

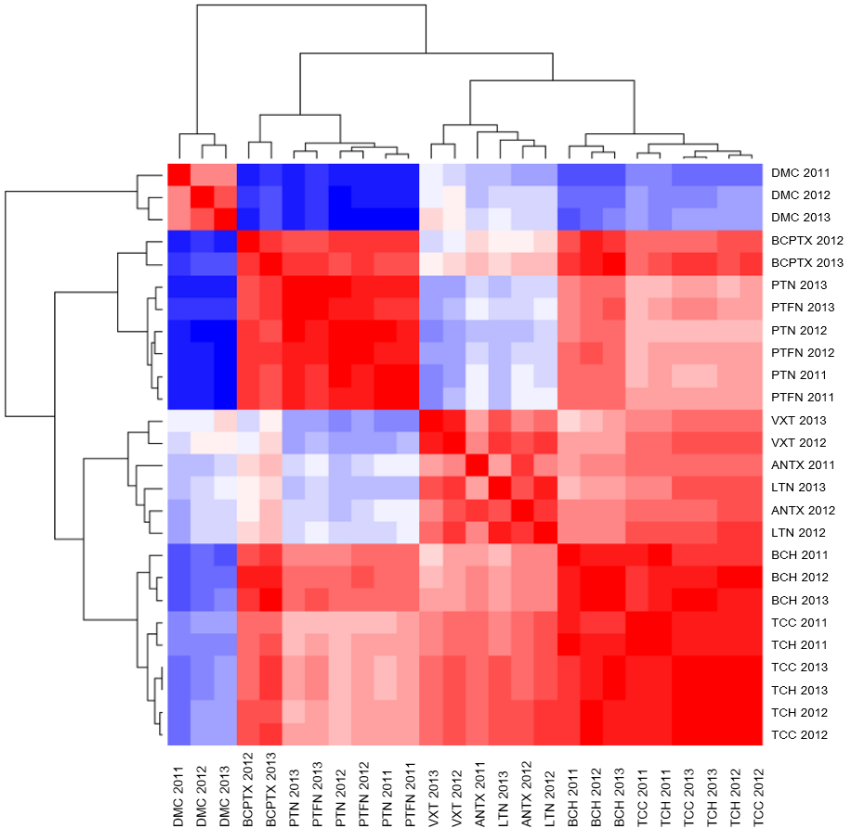

Figura 13: Correlación de los diferentes metabolitos implicados en la ruta metabólica del β-caroteno y la característica de contenido de materia seca (DMC), evaluados en los años 2011, 2012 y 2013. BCH: Betacaroteno-HPLC, TCC: Carotenos totales-Colorimetría, TCH: Carotenos totales-HPLC, PNT: Fitoeno, PTFN: Fitoflueno, VXT: Violaxantina, ANTX: Antoxianina, LTN: Luteina, BCPTX:Betacriptoxantina.

Se determinó que el contenido de β-caroteno en la familia de estudio presentó un comportamiento similar durante los tres años de evaluación lo que se evidencia en el diagrama de caja de la Figura 14.

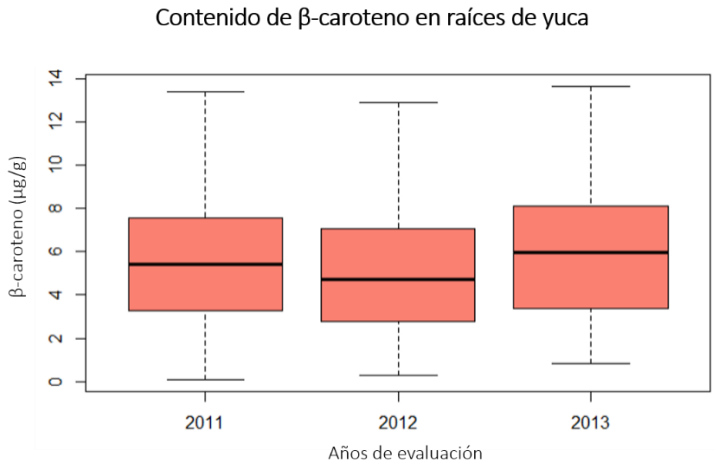

Figura 14: Representación gráfica del contenido de β-caroteno obtenidos en los tres años de evaluación mediante un diagrama de caja.

La selección del material vegetal para este estudio se basó en escoger los materiales que presentaron los valores más bajos, intermedios y altos de β-caroteno durante los tres años de evaluación (Figura 15).

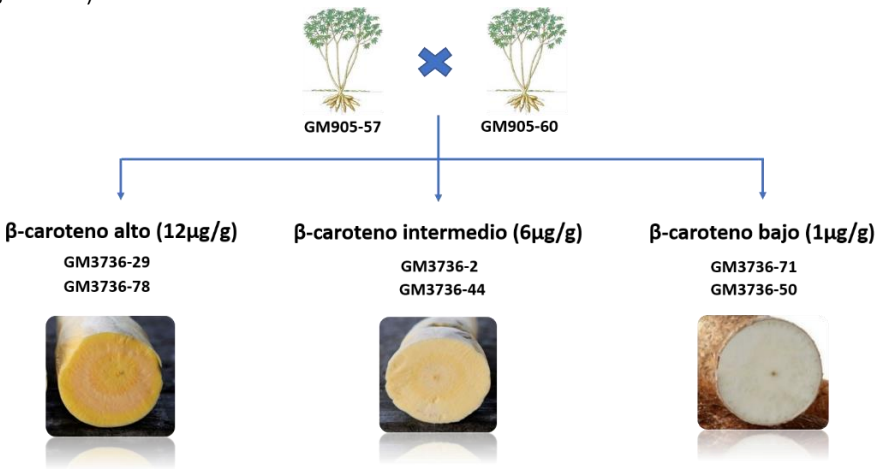

Figura 15: Esquema general de selección de materiales para RNA-seq.

4.2. Integridad del RNA

De los ocho genotipos de estudio seleccionados se extrajo RNA para 24 muestras provenientes de tejido de raíz y 24 muestras de tejido foliar. La integridad del RNA medida en el bioanalizador Agilent 2100 presentó un valor del RIN (RNA *Integrity Number*) mayor a 6 en todas las muestras, lo que indica la alta calidad del RNA extraído (Figura 16).

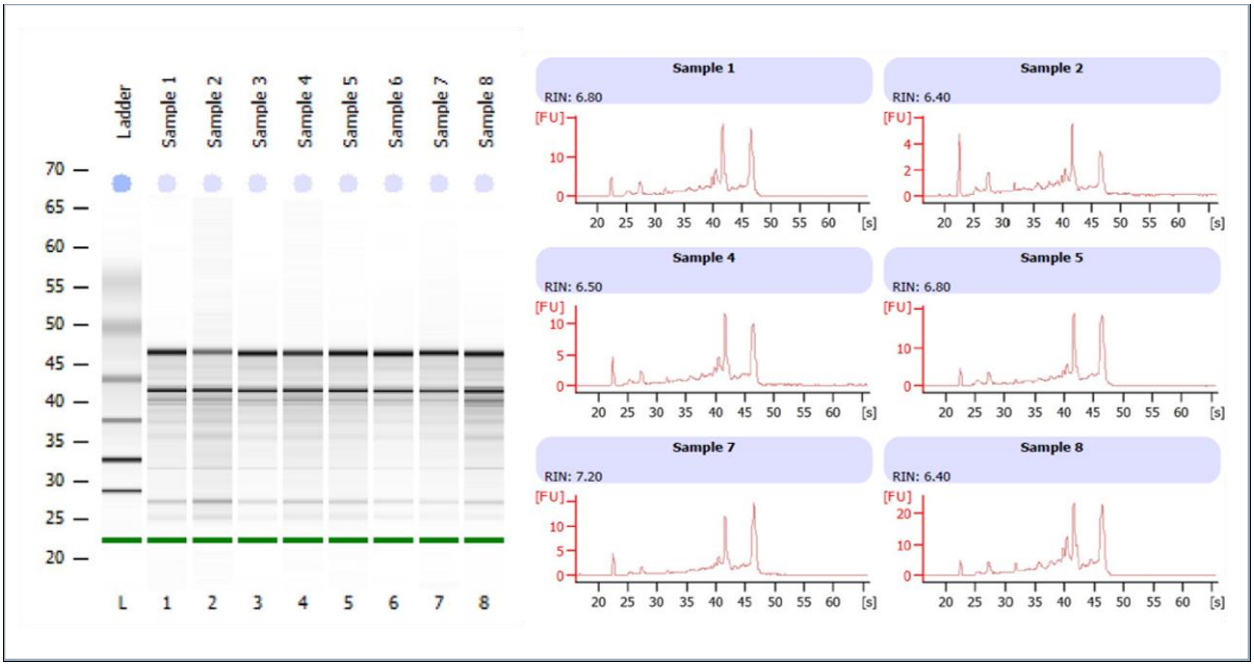

Figura 16: Evaluación de la integridad del RNA extraído en el bioanalizador de Agilent 2100. El gel de calidad ubicado en la parte izquierda de la figura presenta el RNA de las 8 primeras muestras del estudio y el marcador utilizado (L). El valor del RIN y los picos característicos del RNA se observan en la parte derecha de la figura.

La secuenciación del RNA y limpieza de las lecturas se llevó a cabo en el instituto BGI y los archivos con las secuencias limpias fueron recibidos para el procesamiento. Sin embargo, se realizó un nuevo chequeo de la calidad de estas lecturas.

4.3. Calidad de las secuencias

La evaluación de la calidad de las secuencias mediante el programa FastQC mostró un total de secuencias en promedio de 22.000.000 con una longitud de 100 pb y un promedio de 44% de contenido de GC. Todas las muestras presentaron niveles de buena calidad de la secuencia por base, y por secuencia la cual se evidencia en la distribución de la calidad promedio de la secuencia y el valor del *Phred* que para estas secuencias fue de 40.

La evaluación del contenido de GC mostró que es muy similar al de la distribución teórica, así mismo, se comprobó la ausencia de adaptadores en la secuencia. Finalmente, se observó que el 15% de las secuencias presentan 10 veces el mismo fragmento, lo cual se debe a que en las librerías de RNA-seq, se espera que algunas secuencias se produzcan con mucha frecuencia, y otras sean muy poco frecuentes (transcripciones con bajo número de copias), por lo que un cierto nivel de duplicación en la parte de la librería es inevitable. En términos generales, los módulos importantes para determinar la calidad de la secuencia presentaron el símbolo de aprobación (Figura 17).

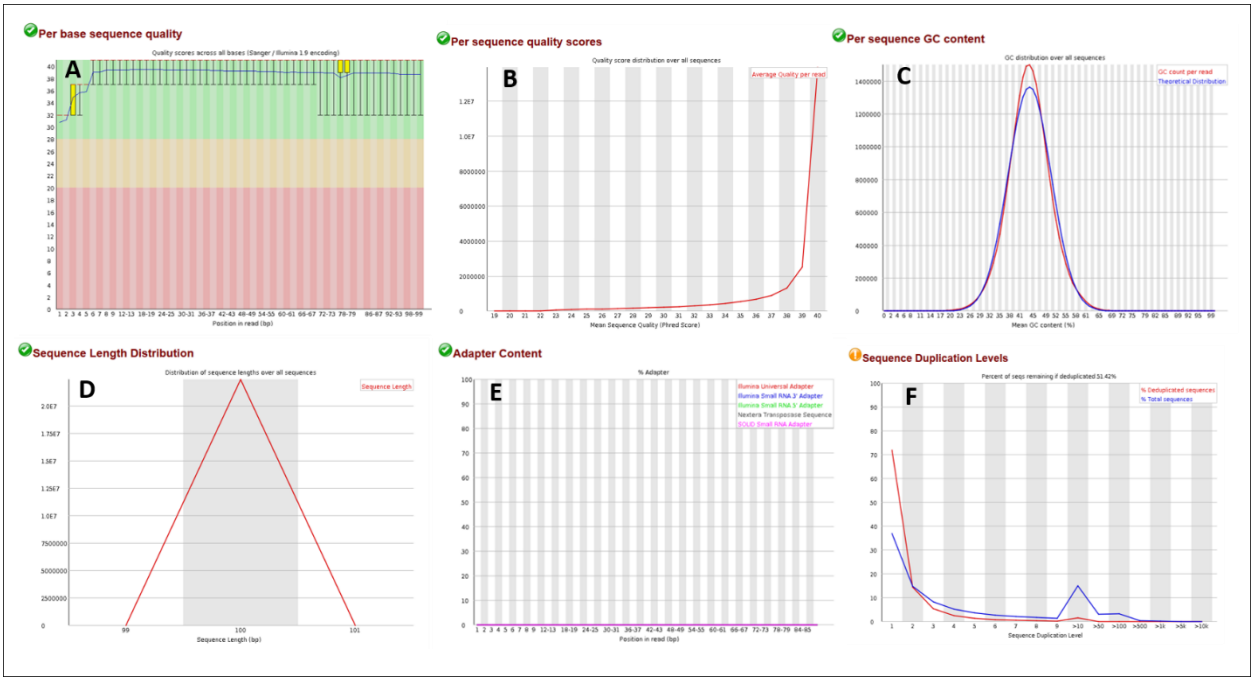

Figura 17: Evaluación de la calidad de las secuencias de RNA utilizando el programa FastQC. A. calidad de secuencia por base. B puntaje de la calidad por secuencia. C. Contenido de GC por secuencia. D. Distribución de la longitud de la secuencia. F. Niveles de duplicación en la secuencia.

4.4. Mapeo

El mapeo de las lecturas provenientes de RNA-seq al genoma de referencia de *Manihot esculenta* versión 6.1 utilizando el programa HISAT2 presentó una tasa de alineamiento del 90% para las 22305694 lecturas provenientes de raíz y las 15052368 lecturas del tejido foliar. El porcentaje de lecturas mapeadas en regiones únicas del genoma, como aquellas que mapearon más de una vez se puede observar en la Figura 18.

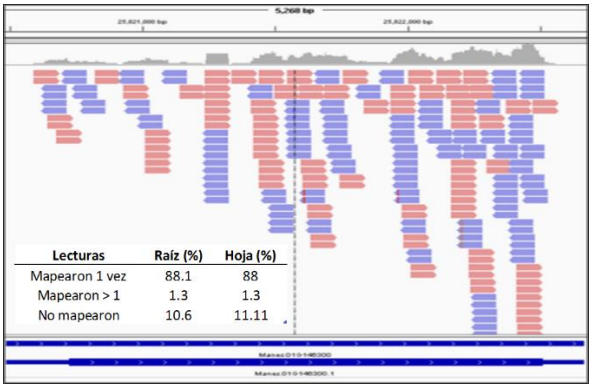

Figura 18: Estadísticas del mapeo y visualización del alineamiento de las lecturas RNA-seq al genoma de referencia utilizando el programa Integrative Genomics Viewer versión 8.1 (IGV). Las barras de color azul y rojo muestran el *strand* positivo (+) y el *strand* negativo (-).

4.5. Ensamblaje y cuantificación del transcriptoma

Todos los transcritos fueron ensamblados mediante el programa StringTie y se produjeron tres transcriptomas, i) raíz, ii) hoja y iii) raíz y hoja. Las estadísticas del ensamblaje muestran que en los tres transcriptomas se capturaron en promedio 10.5% de nuevos exones, 8.2% de nuevos intrones para un total de 20% loci nuevos que corresponden en promedio a 8042 loci (Tabla 2). La cuantificación se llevó a cabo mediante el programa Kallisto para cada uno de los transcriptomas donde se obtuvieron los archivos de abundancia para cada muestra del estudio que presentan el nombre del transcrito, la longitud, el conteo estimado y el valor tpm.

Tabla 2: Estadísticas del ensamblaje de los tres transcriptomas de yuca.

Transcriptomas	Raíz (%)	Hoja(%)	Combiando Raíz -Hoja (%)
Exones faltantes	0	0	0
Nuevos exones	8.4	10.9	12.4
Intrones faltantes	0	0	0
Nuevos intrones	6.4	8.6	9.8
Loci faltantes	0	0	0
Nuevos loci	15.3	20.3	24.3

4.6. Análisis de expresión diferencial

La búsqueda de genes diferencialmente expresados en cada transcriptoma se realizó con el programa DESeq2 utilizando los archivos de abundancia anteriormente obtenidos para cada una de las muestras. La selección de estos genes se basó en que presentaran una tasa de descubrimiento falsa menor al 5% para una comparación de fenotipo en particular y un conteo de lecturas mayor a 10.

La variabilidad dentro de los grupos fue modelada mediante el parámetro de dispersión, con el propósito de evaluar la fiabilidad del modelo. Dado que la exactitud de la estimación de la dispersión es crítica para la inferencia estadística de la expresión diferencial, si un gen tiene una expresión diferencial significativa depende no solo de su LFC sino también de su variabilidad dentro del grupo, como se muestra en la Figura 19.A, donde los puntos negros son las estimaciones de dispersión para cada gen por separado, la línea de tendencia roja, que muestra la dependencia de las dispersiones en la media y los puntos azules representa las estimaciones finales luego de reducir la estimación de cada gene hacia la línea de tendencia, los cuales se usan en la prueba de hipótesis. Los círculos azules sobre la nube principal de puntos son genes que tienen altas estimaciones de dispersión genética y están etiquetadas como valores atípicos de dispersión. Estas estimaciones, por lo tanto, no se redujeron hacia la línea de tendencia ajustada. En la figura 19.B se observan los genes que presentaron un LFC positivo (LFC>0) o negativo (LFG<0), los cuales se clasificaron como *up-regulated* y *down-regulated* respectivamente.

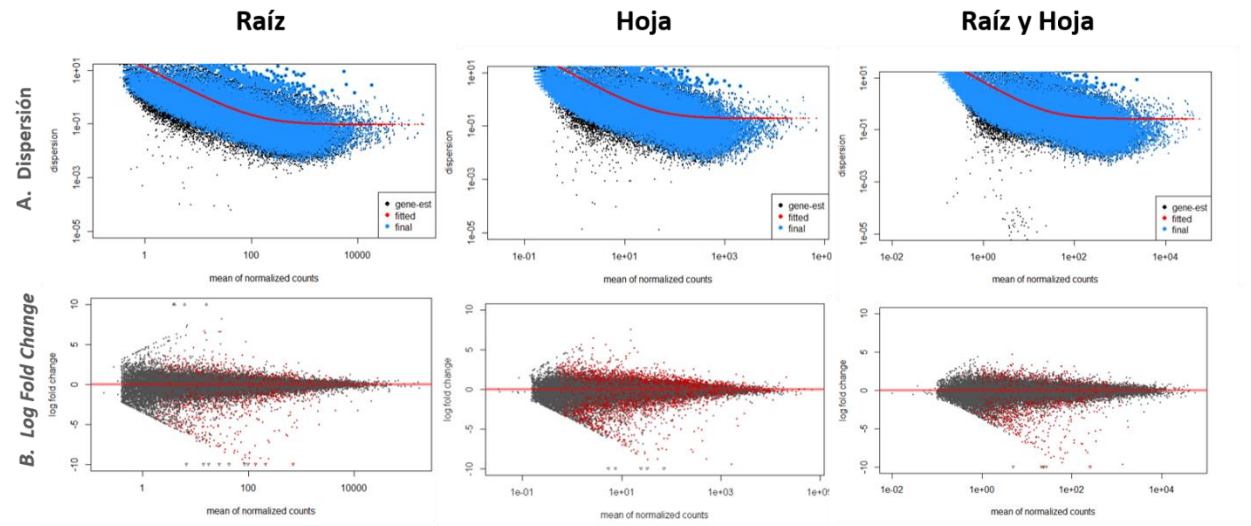

Figura 19: Dispersión y *Log fold change* para cada uno de los genes evaluados en DESeq2.

Con el fin de minimizar las diferencias entre las muestras con pocos conteos y normalizar con respecto al tamaño de la librería se utilizó la función *rlog* que transforma el conteo de datos a una escala de log2 (Figura 20) y de esta manera obtener una mejor representación del agrupamiento de las muestras mediante un análisis de componentes principales (PCA).

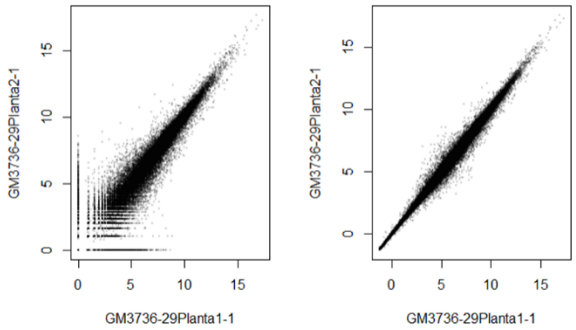

Figura 20: Transformación de la data utilizando la función *rlog* disponible en DESeq2.

El análisis de componentes principales mostró que la variabilidad en la acumulación de β-caroteno en raíces esta explicada principalmente en un 32.22% por el componente uno (Figura 21).

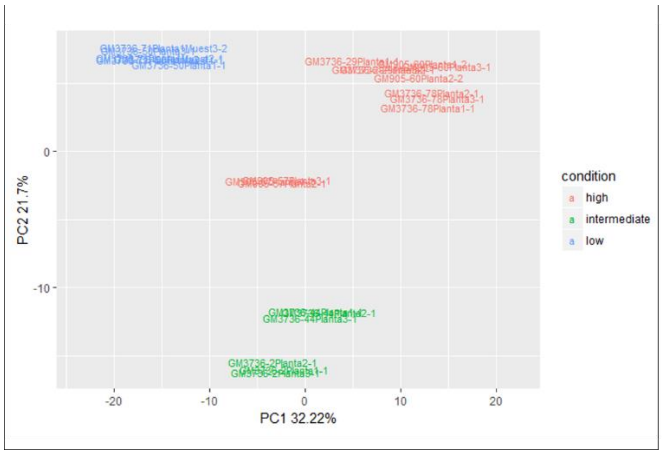

Figura 21: Análisis de componentes principales basado en los datos de expresión. Los colores muestran a que fenotipo pertenece cada muestra. Azul (Bajo β-caroteno), verde (intermedio β-caroteno) y salmón (Alto β-caroteno).

Adicionalmente el PCA nos permitió evidenciar para los datos del transcriptoma de raíz y hoja, la variabilidad sesgada al tipo de tejido, debido a que el componente uno está explicando el 86% de la varianza encontrada (Figura 22). De igual manera un mapa de calor nos permitió observar los datos de expresión de un conjunto de genes en todas las muestras provenientes de raíz y hoja y un análisis jerárquico mostró la agrupación de estas muestras con base en el tipo de tejido (Figura 23).

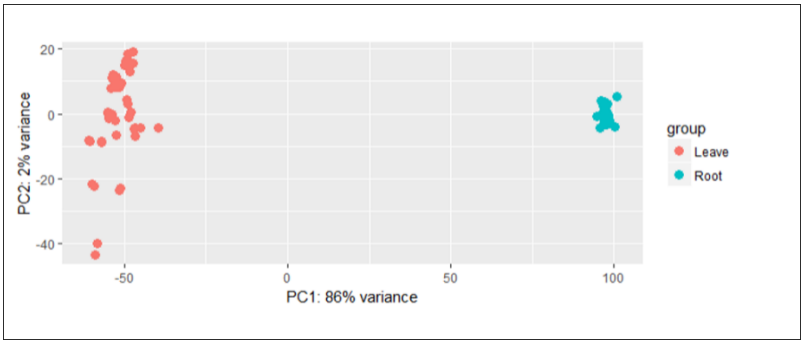

Figura 22: Análisis de componentes principales basado en los datos de expresión. Los colores muestran el tipo de tejido de donde proceden los datos. Color salmón (hoja), color verde (raíz).

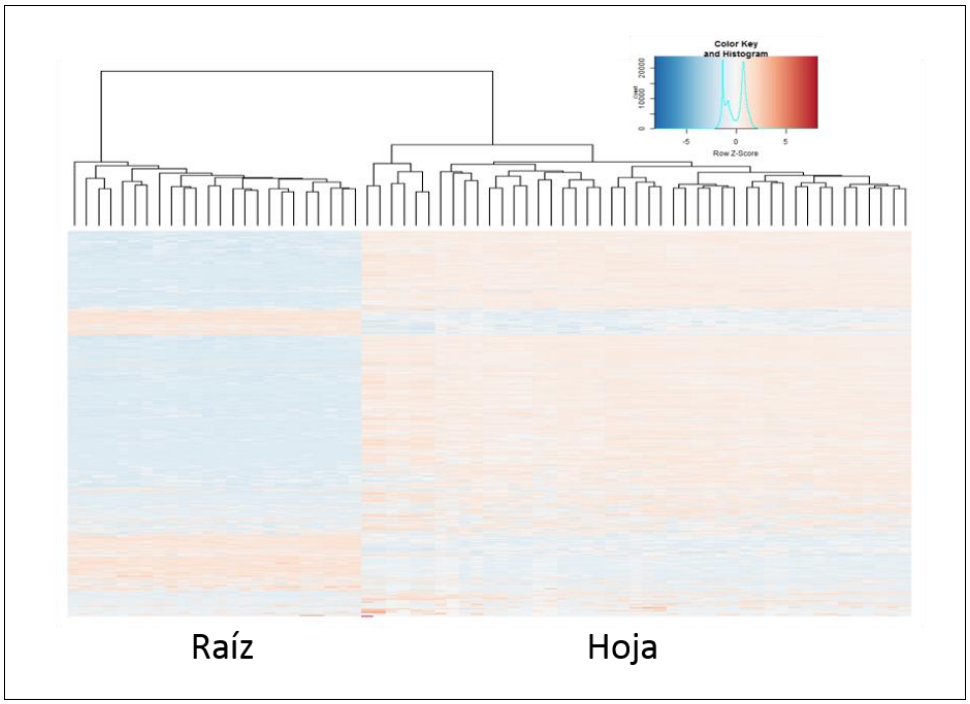

Figura 23: Visualización de los valores de expresión para un subconjunto de genes mediante un mapa de calor (*heapmap*) y el agrupamiento jerárquico de las muestras provenientes de tejido de raíz y hoja.

De los 42472 transcritos evaluados en raíz solo el 1.5% (627 genes) se clasificaron como *up-regulated* y el 1.9 % (810 genes) como *down-regulated*. Los valores atípicos encontrados fueron del 0.21% (91 genes) y los genes con bajos conteos llegaron al 3.9% (1657 genes). Para el caso de hoja, se obtuvieron 51,796 transcritos de los cuales el 7.9% (4096 genes) están *up-regulated* y 8.4% (4343 genes) están *down-regulated.* No se encontraron valores atípicos y 2031 genes presentaron conteos bajos, lo que equivale al 3.9%. Adicionalmente se obtuvieron 55542 transcritos para el transcriptoma de hoja y raíz, de los cuales 5.4% (2975 genes) están *up-regulated* y 5.6% (3091 genes) están *down-regulated.* Los valores atípicos encontrados fueron del 0.088% (49 genes) y 3232 genes presentaron bajos conteos lo que equivale al 5.8%. Los tres análisis incluyeron un valor de p ajustado menor a 0.05. El programa DEGreport permitió la visualización de los genes *up-regulated* y *down-regulated* (Figura 24).

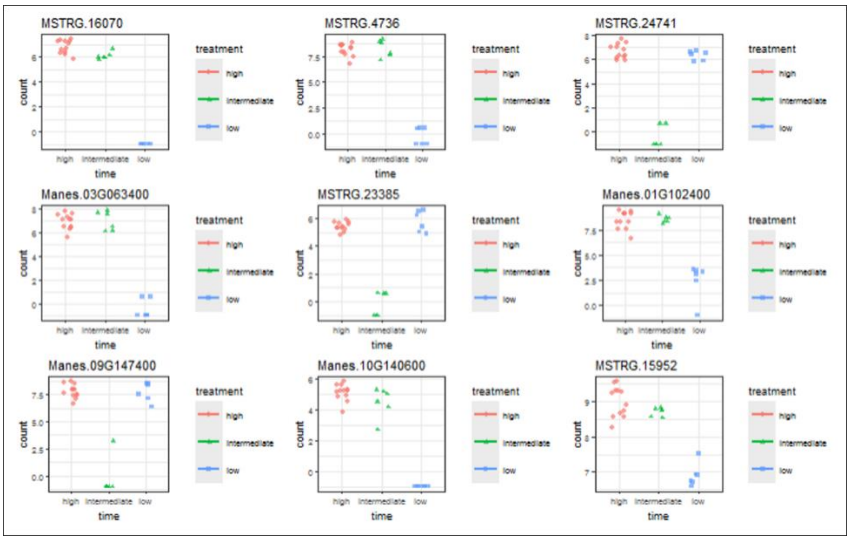

Figura 24: Visualización de un conjunto de genes diferencialmente expresados en el transcriptoma de raíz.

La agrupación de los genes diferencialmente expresados para cada fenotipo se realizó mediante la función degPatterns del paquete DEGreport. De los 1437 genes del transcriptoma de raíz diferencialmente expresados, 565 genes están sobre-expresados en los genotipos que acumulan alto β-caroteno, 363 genes en los genotipos que acumulan niveles intermedios de β-caroteno y 509 genes en los genotipos que acumulan bajas concentraciones de β-caroteno (Figura 25).

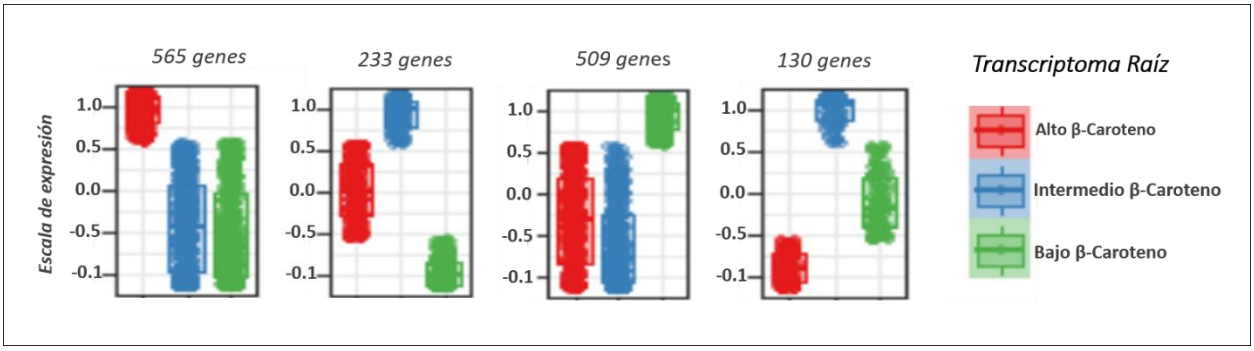

Figura 25: Visualización de los de genes diferencialmente expresados en el transcriptoma de raíz.

En contraste, el transcriptoma de hoja presentó un número más elevado de genes diferencialmente expresados. En total 8439 genes, de los cuales 749 están sobre-expresados en los genotipos que acumulan alto β-caroteno, 3500 genes en los que acumulan niveles intermedios de β-caroteno, 998 genes en genotipos que acumulan niveles bajos de β-caroteno. Adicionalmente, 3192 genes están sobre-expresados en genotipos que acumulan tanto altos como bajos niveles de β-caroteno, los cuales no se asocian con la acumulación del β-caroteno en la yuca (Figura 26).

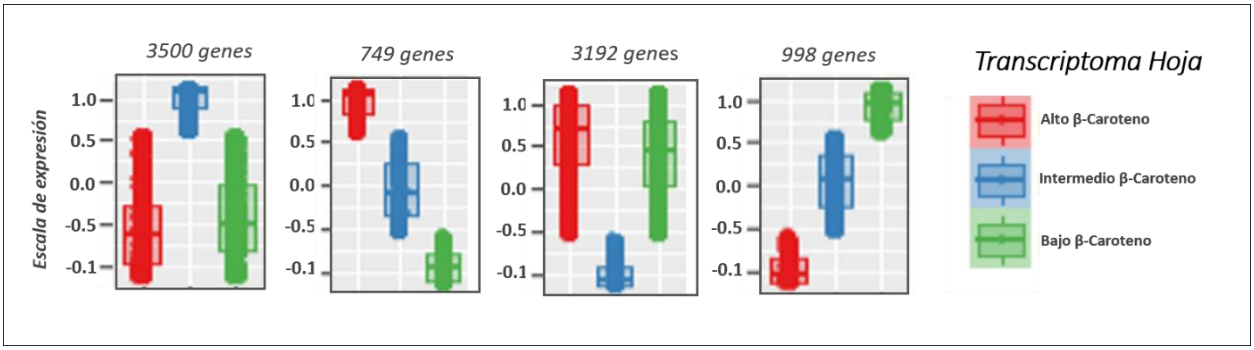

Figura 26: Visualización de los genes diferencialmente expresados en el transcriptoma de hoja.

Finalmente, el transcriptoma de raíz y hoja presentó 6066 genes diferencialmente expresados, de los cuales, 830 genes estuvieron sobre-expresados en los genotipos de acumulan altos niveles de β-caroteno, 2449 genes en los genotipos con niveles intermedios de β-caroteno y 669 genes que acumulan niveles bajos de β-caroteno. 2118 genes estuvieron sobre-expresados en genotipos que acumulan alto y bajo contenido de β-caroteno, que no se pueden asociar con la acumulación de β-caroteno en yuca.

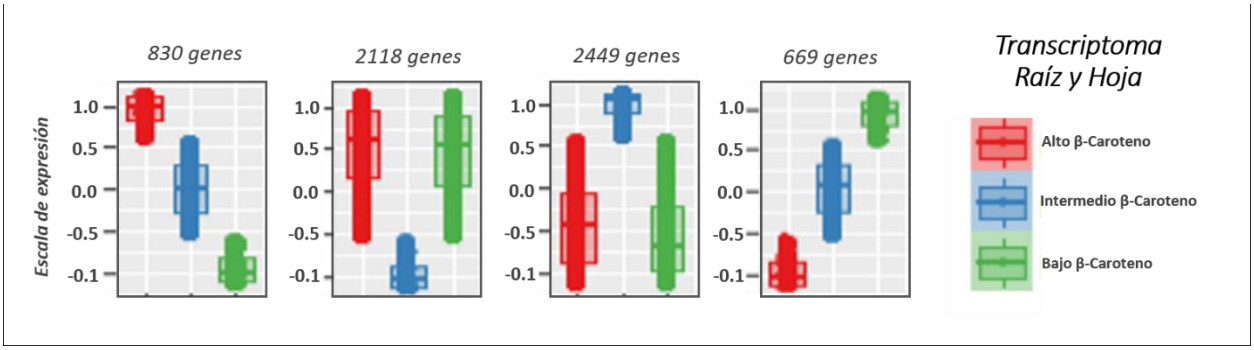

Figura 27: Visualización de los genes diferencialmente expresados en el transcriptoma de raíz y hoja.

4.7. Anotación, función y regulación

4.7.1. Análisis de enriquecimiento basado en el transcriptoma de raíz

De los 1437 genes diferencialmente expresados, 757 genes presentaron gen ID y de estos, 548 presentan anotación GO. El análisis de anotación y enriquecimiento realizado en el portal PlantRegMap, mostró para el transcriptoma de raíz un total de 22 términos enriquecidos (p-valor<0.01), 12 están en procesos biológicos, entre los cuales se destaca la respuesta al calor y la regulación de la muerte celular programada (Figura 28); dos términos asociados a parte intracelular en la categoría de componente celular y 8 términos enriquecidos asociados a las funciones moleculares donde resaltan las funciones de unión a ADP y co-represor de la transcripción (Figura 29).

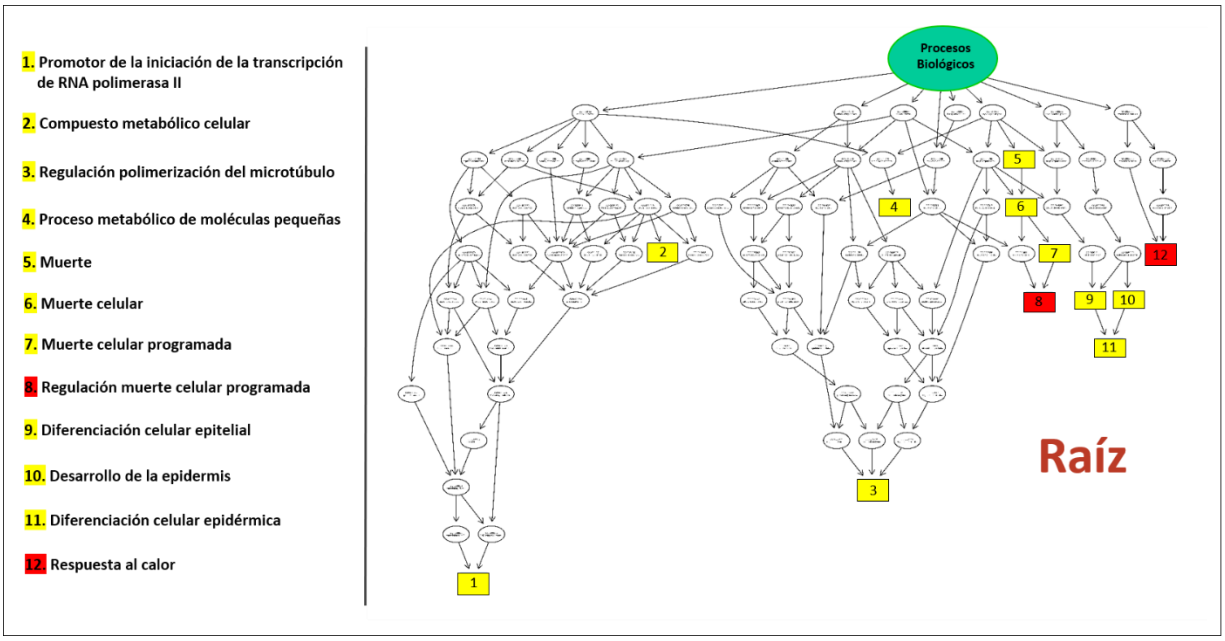

Figura 28: Términos GO enriquecidos en el transcriptoma de raíz asociados a procesos biológicos. Los términos resaltados en color se seleccionaron por presentar un valor de p < 0.01, los términos señalados en rojo presentan el p-valor más bajo del análisis.

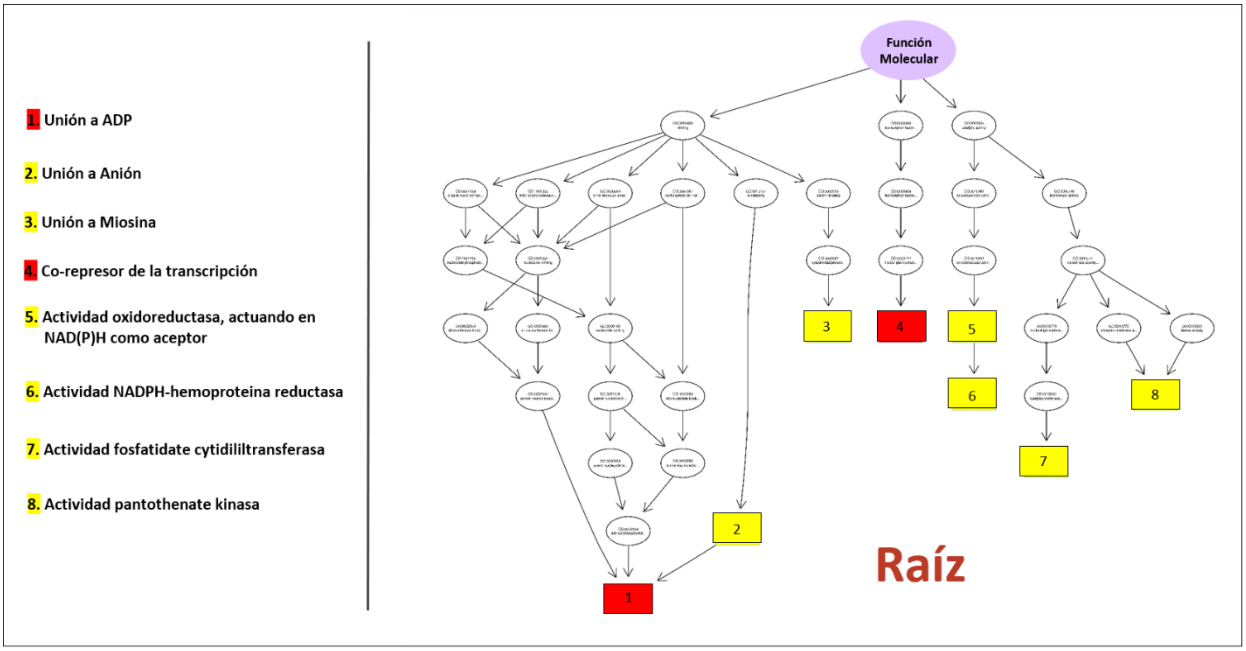

Figura 29: Términos GO enriquecidos en el transcriptoma de raíz asociados a funciones biológicas. Los términos resaltados en color se seleccionaron por presentar un valor de p < 0.01, los términos señalados en rojo presentan el p-valor más bajo del análisis.

Debido a que el objetivo de este trabajo fue buscar e identificar las funciones y rutas metabólicas asociadas a la acumulación de β-caroteno en las raíces de yuca, se realizó nuevamente el análisis de enriquecimiento con los genes diferencialmente expresados para cada uno de los fenotipos.

En los genotipos que acumulan altos niveles de β-caroteno 14 términos GO se presentaron enriquecidos, seis términos involucrados en funciones biológicas, destacándose la respuesta al calor; cuatro términos en la categoría de componente celular, principalmente como parte intracelular y como complejo de la RNA polimerasa. Adicionalmente, en la categoría de función biológica sobresale la actividad óxido-reductasa y la actividad NADPH-hemoproteína reductasa (Figura 30).

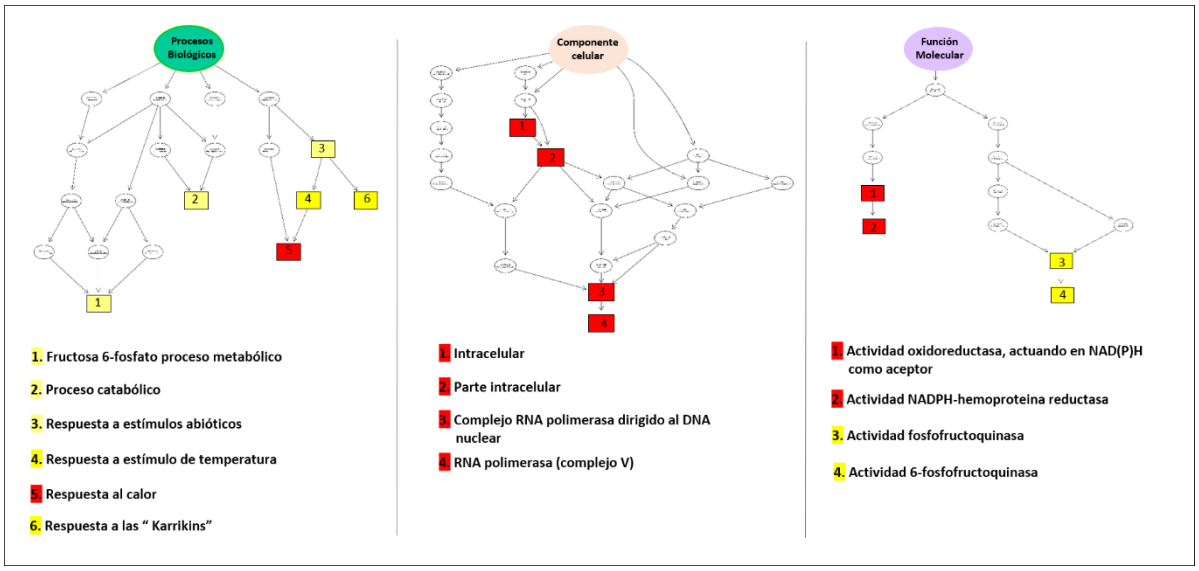

Figura 30: Términos GO enriquecidos en los genotipos que acumulan altos niveles de β-caroteno. Los términos resaltados en color se seleccionaron por presentar un valor de $p < 0.01$, los términos señalados en rojo presentan el p-valor más bajo del análisis.

Una característica relevante del análisis de enriquecimiento en los genotipos que producen bajos niveles de β-caroteno que presentó 22 términos enriquecidos, fue la gran cantidad de términos para la categoría de procesos biológicos, 14 en total comparado con los seis términos de la misma categoría en los genotipos que producen alto β-caroteno. Entre los 14 términos sobresale el proceso biosintético de compuestos organonitrogenados. Así mismo, el componente celular está representado por el complejo macromolecular y las funciones biológicas principales son la actividad catalítica y como constituyente estructural del ribosoma (Figura 31).

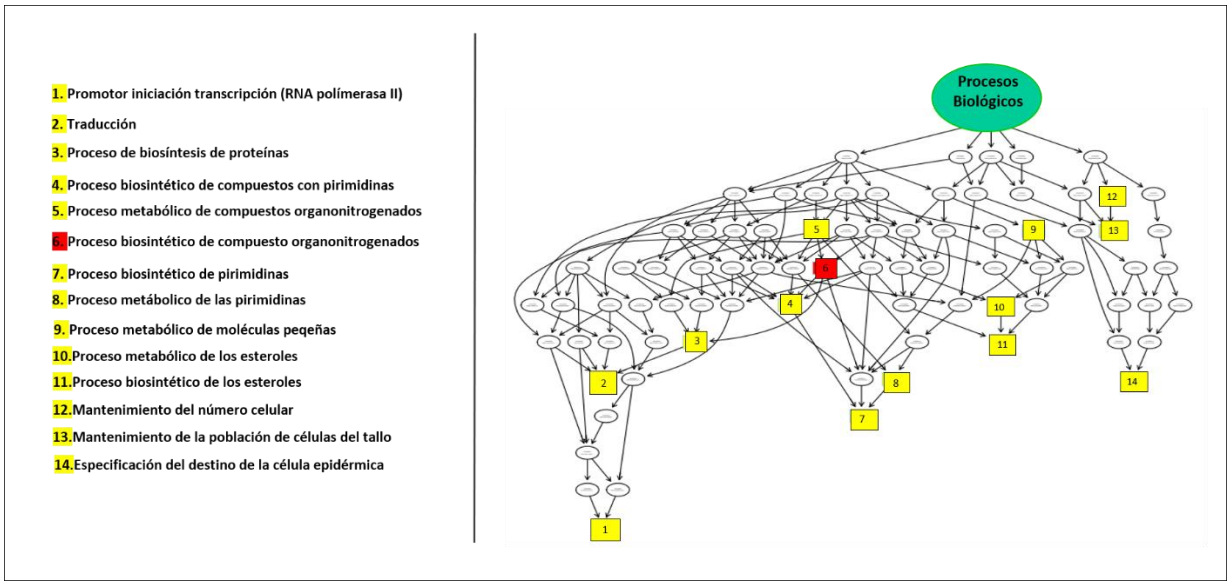

Figura 31: Términos GO enriquecidos en los genotipos que acumulan bajos niveles de β-caroteno. Los términos resaltados en color se seleccionaron por presentar un valor de p < 0.01, los términos señalados en rojo presentan el p-valor más bajo del análisis.

El grupo de genotipos que acumulan concentraciones intermedias de β-caroteno en sus raíces, presentan una cualidad interesante, debido a que un grupo de sus genes que están sobre expresados también lo están en genotipos con altos niveles de β-caroteno aunque con una menor escala en la expresión. Los 9 término enriquecidos están presentes en procesos biológicos y de componente celular relacionado con los microtúbulos, la función molecular relevante está dada por la proteína Acyl transportadora con actividad desaturasa (Figura 32).

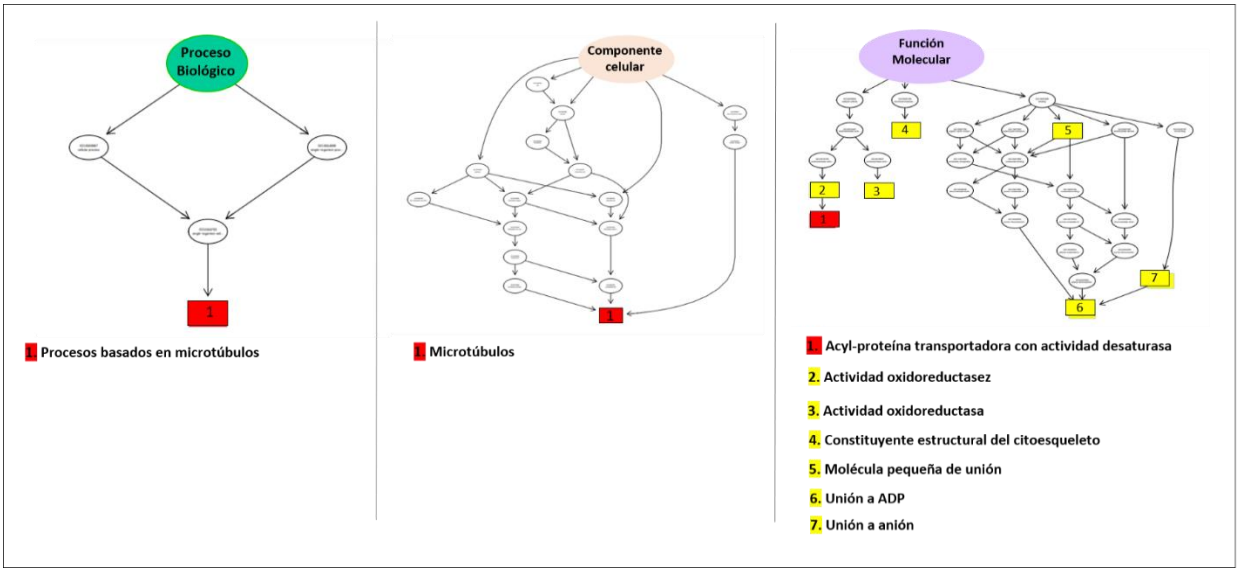

Figura 32: Términos GO enriquecidos en los genotipos que acumulan niveles intermedios de β-caroteno pero que los genes diferencialmente expresados están en intermedios seguidos en expresión por los genotipos altos. Los términos resaltados en color se seleccionaron por presentar un valor de p < 0.01, los términos señalados en rojo presentan el p-valor más bajo del análisis.

Así mismo, pasa con un grupo específico de genes sobre expresados en este fenotipo intermedio que están solo en los que producen bajo β-caroteno y donde las funciones moleculares predominantes se dan en los receptores de señalización transmembranales y los canales de iones.

4.7.2. Análisis de enriquecimiento basado en el transcriptoma de hoja

Los resultados del análisis de enriquecimiento basados en el transcriptoma de hoja muestran 52 términos enriquecidos en procesos biológicos (p-valor<0.01). Principalmente en dos procesos generales: i) biosíntesis de compuestos organonitrogenados, amidas, péptidos, sustancias orgánicas y vitaminas y ii) respuesta al calor, temperatura, alta intensidad lumínica y en general a estreses abióticos (Figura 33). En la categoría de componente celular 26 términos fueron relevante y se asocian con organelos, parte intracelular, citoplasma, ribosoma y en la categoría de función molecular es relevante la actividad de molécula estructural, unión a: ADP, moléculas pequeñas y nucleótidos.

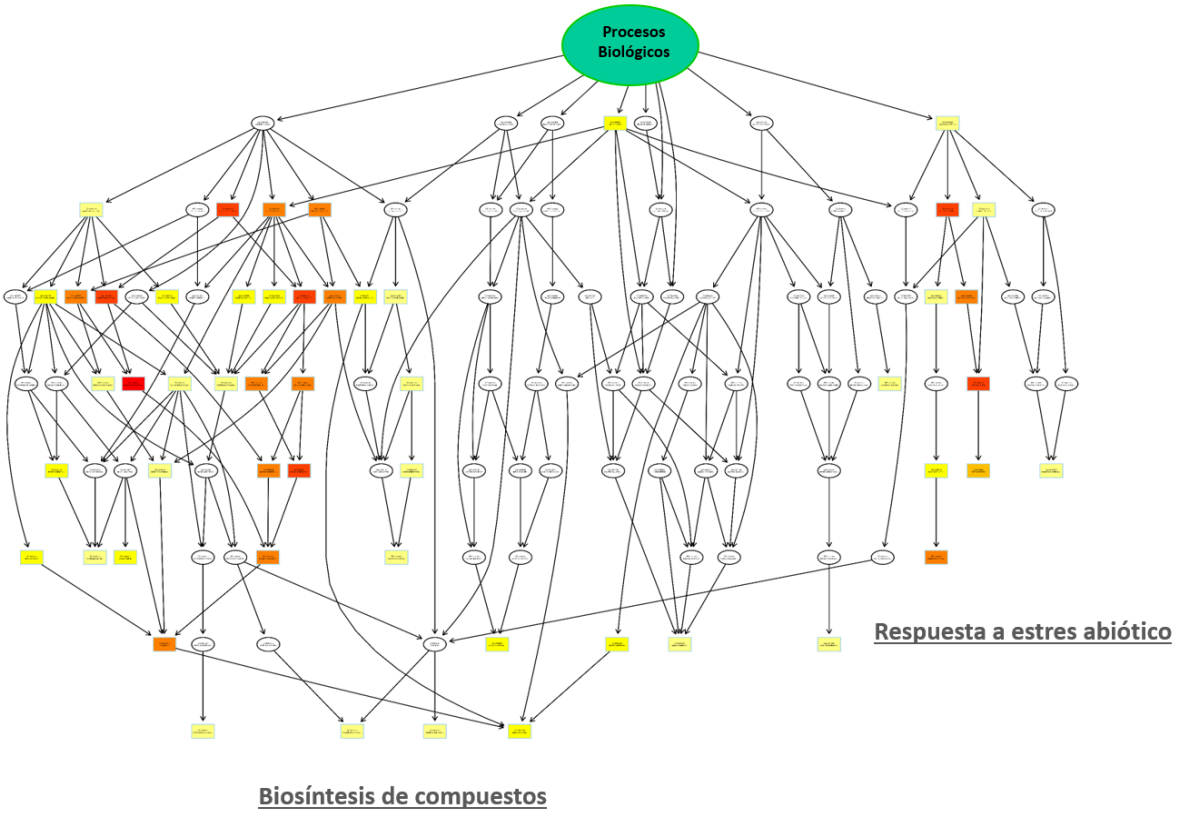

Figura 33: Términos GO enriquecidos asociados a procesos biológicos en el transcriptoma de hoja. Los términos resaltados en color se seleccionaron por presentar un valor de p < 0.01, los términos señalados en rojo presentan el p-valor más bajo del análisis.

En los genotipos que acumulan alto β-caroteno encontramos que los procesos biológicos estadísticamente significativos (p-valor<0.01 & q-valor<0.05) están asociados con la respuesta al estrés por calor, por temperatura y estrés abiótico en general. Así como también, con los procesos

metabólicos de ácidos nucleicos y adicionalmente con la respuesta al peróxido de hidrógeno (Tabla 3).

Tabla 3: Lista de procesos biológicos enriquecidos en los genotipos que acumulan β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:0009408	Respuesta al calor	24	4.29	5.80E-12	2.76E-08
GO:0009266	Respuesta a estímulos de temperatura	29	11.21	2.90E-06	6.90E-03
GO:0009628	Respuesta a estrés abiótico	65	37.49	7.50E-06	1.19E-02
GO:0044260	Proceso metabólico macromolecular	200	159.27	2.70E-05	2.67E-02
GO:0010467	Expresión de genes	106	73.19	2.80E-05	2.67E-02
GO:0006950	Respuesta al estrés	86	56.85	4.30E-05	2.68E-02
GO:0016070	Proceso metabólico del RNA	87	57.72	4.40E-05	2.68E-02
GO:0090304	Proceso metabólico ácido nucleico	99	67.97	4.50E-05	2.68E-02
GO:0042542	Respuesta al peróxido de hidrógeno	8	1.4	6.40E-05	3.39E-02

En los genotipos que acumulan bajos niveles de β-caroteno encontramos que el proceso biológico con mayor significancia estadística (p-valor<0.01 & q-valor<0.05) es el metabolismo de compuestos organonitrogenados. Seguido por el metabolismo de moléculas pequeñas, la biosíntesis de vitaminas y el ciclo del ácido tricarboxílico con una tasa mayor de falsos descubrimientos (Tabla 4). En la categoría de función molecular, la actividad catalítica presenta la mayor significancia (p-valor=1.50E-07 & q-valor=3.49E-04) y en la categoría de componente celular el citoplasma es el más significativo con valores de q < 0.0001.

Tabla 4: Lista de procesos biológicos enriquecidos en los genotipos que acumulan bajos niveles de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:1901564	Metabolismo comp. organonitrogenados	62	35	5.00E-06	2.38E-02
GO:1901566	Biosíntesis de comp. organonitrogenados	47	24.56	1.20E-05	2.86E-02
GO:0044281	Proceso metabólico de moléculas pequeñas	51	31.23	0.00031	3.33E-01
GO:0009110	Biosíntesis de vitaminas	7	1.33	0.00034	3.33E-01
GO:0006099	Ciclo del ácido tricarboxílico	5	0.65	0.00042	3.33E-01

En los genotipos con niveles intermedios de β-caroteno, los procesos biológicos enriquecidos presentan valores de p < 0.01. y los valores q son mayores a 0.05. Sin embargo, los procesos biológicos coinciden con los obtenidos para genotipos que acumulan altos niveles de β-caroteno (Tabla 5). La función molecular unión a cofactores está significativamente enriquecida en estos materiales (valor q < 0.02).

Tabla 5: Lista de procesos biológicos en los genotipos con niveles intermedios de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:0010286	Aclimatación al calor	7	1.29	0.00021	3.20E-01
GO:0009266	Respuesta a la temperatura	33	17.29	0.00028	3.20E-01
GO:0016192	Transporte vesicular	26	12.41	3.00E-04	3.20E-01
GO:0009408	Respuesta al calor	17	6.61	0.00032	3.20E-01
GO:0098771	Homeóstasis iónico	12	3.83	4.00E-04	3.20E-01

4.7.3. Análisis de enriquecimiento basado en el transcriptoma de hoja y raíz

De los 6066 genes diferencialmente expresados, 3232 presentaron gen ID y 2286 genes tuvieron anotación GO. El análisis de enriquecimiento presentó 43 términos enriquecidos. Los genotipos con altos niveles de β-caroteno basado en el transcriptoma de hoja y raíz presentaron los procesos biológicos relacionados con el catabolismo glucosídico, el metabolismo, procesos de óxido reducción y respuesta al calor, entre otros (Tabla 6).

Tabla 6: Lista de procesos biológicos enriquecidos en los genotipos con niveles altos de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:1901658	Proc. catabólico glucosílico	6	0.88	4.00E-05	1.90E-01
GO:0008152	Proceso metabólico	1133	1073.52	0.00017	2.26E-01
GO:0055114	Proceso de óxido-reducción	216	172.74	0.00018	2.26E-01
GO:0009408	Respuesta al calor	30	15.17	0.00019	2.26E-01
GO:0006301	Reparación postreplicación	5	0.77	0.00028	2.67E-01
GO:0006631	Proceso de óxido-reducción	36	20.34	0.00045	3.57E-01

Las funciones moleculares estadísticamente significativas se asociaron con la actividad oxido reductasa, unión al hierro, el grupo hemo y al tetrapirrol. Sin embargo, la Tabla 7 muestra adicionalmente otras funciones moleculares con menor soporte estadístico, pero con un gran soporte biológico como lo es la actividad de la terpeno sintetasa.

Tabla 7: Lista de funciones moleculares enriquecidas en los genotipos con niveles altos de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:0016705	Actividad óxido-reductasa	80	46.83	1.10E-06	2.56E-03
GO:0005506	Unión al hierro	78	48.44	1.40E-05	1.55E-02
GO:0020037	Unión hemo	85	54.45	2.00E-05	1.55E-02
GO:0046906	Unión tetrapirrol	85	55.42	3.80E-05	2.21E-02
GO:0004618	Act. fosfoglicerato quinasa	4	0.64	0.0017	5.65E-01
GO:0003678	Act. DNA helicasa	11	4.19	0.0021	6.11E-01
GO:0004816	Act. asparagina-tRNA ligasa	3	0.43	0.0046	1.00E+00
GO:0010333	Actividad terpeno sintasa	12	5.37	0.0057	1.00E+00

En los genotipos con bajo β-caroteno 26 términos resultaron enriquecidos, siendo el término de mayor significancia la actividad óxido-reductasa del componente función molecular, seguido de la unión al ácido ascórbico, unión a vitaminas y la actividad Acyl coA oxidasa. De la misma forma, en el proceso biológico resaltan los procesos de óxido-reducción.

En los genotipos con niveles intermedios de β-caroteno 60 términos están enriquecidos, en el componente de proceso biológico, tres términos presentan la mayor significancia estadística, procesos de óxido-reducción, metabolismo de ácidos grasos y respuesta al ácido jasmónico (Tabla 8). En el componente función molecular presenta 7 términos enriquecidos relacionados con la actividad óxido-reductasa, unión a tetrapirroles y actividad catalítica (Tabla 9).

Tabla 8: Lista de procesos biológicos enriquecidos en los genotipos con niveles intermedios de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:0055114	Proceso de óxido-reducción	113	72.97	1.10E-06	5.24E-03
GO:0006631	Metabolismo ácidos grasos	23	8.59	1.60E-05	3.81E-02
GO:0009753	Respuesta al ácido jasmónico	17	5.53	3.40E-05	5.40E-02
GO:0006629	Metabolismo lipídico	54	31.35	6.40E-05	7.62E-02

Tabla 9: Lista de funciones moleculares enriquecidas en los genotipos con niveles intermedios de β-caroteno

GO.ID	Término	Genes	Valor esp.	p-valor	q-valor
GO:0016705	Actividad óxido-reductasa	56	19.67	1.50E-12	3.49E-09
GO:0005506	Unión al hierro	55	20.35	1.80E-11	2.10E-08
GO:0020037	Unión hemo	55	22.88	1.50E-09	1.16E-06
GO:0046906	Unión tetrapirrol	55	23.28	2.90E-09	1.69E-06
GO:0016491	Actividad óxido-reductasa	123	75.76	3.70E-08	1.72E-05
GO:0016717	Actividad óxido-reductasa	6	0.77	6.70E-05	2.60E-02
GO:0003824	Actividad catalítica	444	391.32	9.10E-05	3.03E-02

4.7.4. Análisis de regulación

Todos los genes diferencialmente expresados fueron evaluados para determinar su regulación. Con base en el transcriptoma de raíz el análisis mostró 3 genes con capacidad de regulación génica, Manes.14G079000, Manes.09G032800 y Manes.09G147400. La Figura 34 muestra la red de regulación de estos genes. Las anotaciones GO para estos genes están relacionados con el proceso que modula la frecuencia, velocidad o extensión de la transcripción celular con plantilla de ADN, la función molecular mediante la cual un producto génico interactúa de forma selectiva y no covalente con el ADN y un factor de transcripción relacionado con la familia MYB respectivamente.

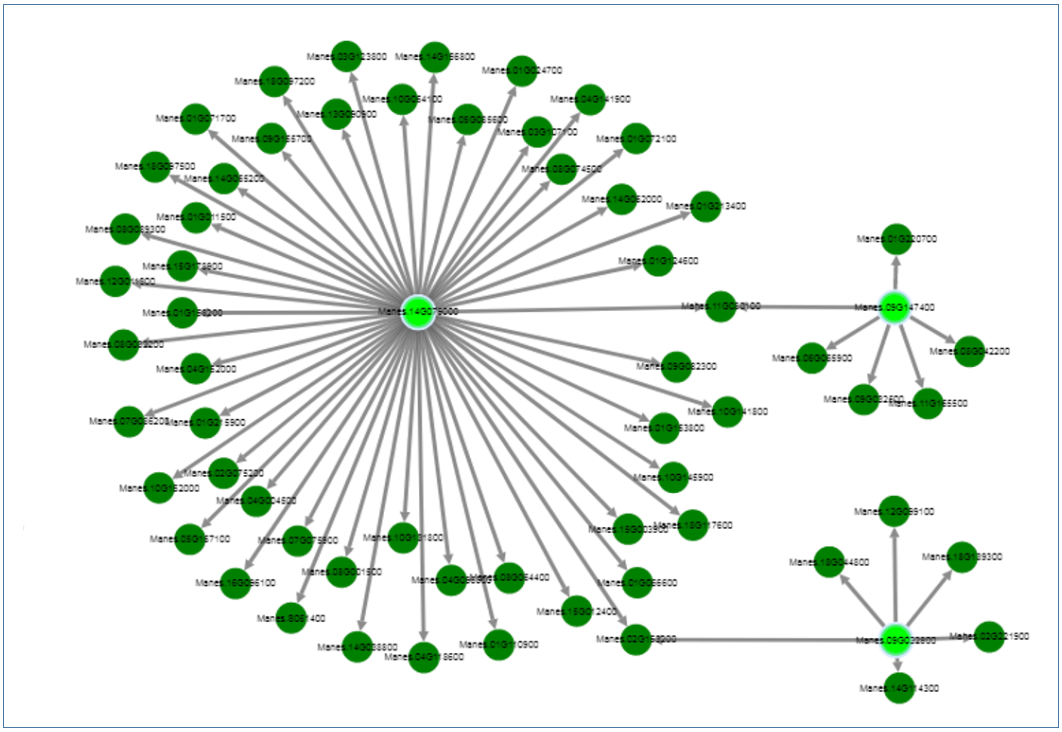

Figura 34: Red de regulación génica presente en raíz. El color verde claro indica los genes reguladores y el color verde oscuro indica los genes regulados.

El análisis de regulación para los genes diferencialmente expresados en genotipos con altos niveles de β-caroteno mostró dos de los tres genes reguladores: Manes.14G079000 y Manes.09G032800 (Figura 35).

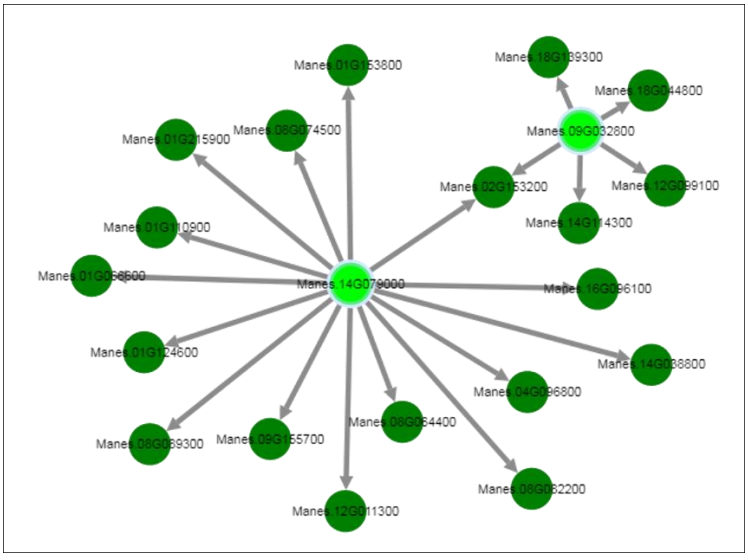

Figura 35: Red de regulación génica presente en genotipos que acumulan alto contenido de β-caroteno.

El análisis de regulación para hoja mostró 42 genes reguladores, de los cuales, dos están en genotipos con bajo β-caroteno, nueve están presentes en los genotipos que presentan alto β-caroteno y 15 genes están en los genotipos con niveles intermedio de β-caroteno (Tabla10).

Tabla 10: Lista de genes reguladores en la hoja de yuca para los diferentes fenotipos

Bajo β-caroteno	Alto β-caroteno	Intermedio β-caroteno
Manes.17G009200	Manes.01G007600	Manes.01G141500
Manes.17G055700	Manes.01G199900	Manes.01G242300
	Manes.05G015600	Manes.02G066400
	Manes.07G135300	Manes.05G037100
	Manes.09G031700	Manes.09G034900
	Manes.09G032800	Manes.09G127300
	Manes.14G079000	Manes.09G185300
	Manes.17G016000	Manes.10G052500
	Manes.18G079600	Manes.12G009000
		Manes.13G107400
		Manes.15G039700
		Manes.15G130800
		Manes.18G001000
		Manes.18G015400
		Manes.18G029600

5. Discusión de Resultados

La secuenciación de nueva generación ha permitido develar el componente génico operante en los organismos a una escala antes impensable. Especialmente, la técnica de RNA-seq ha permitido entender el alcance y la complejidad de los transcriptomas eucariotas y abarcar el estudio de características de interés, basado en miles de genes que soportan los procesos biológicos, permitiendo un mayor entendimiento de una respuesta o característica dada unas condiciones particulares (Wang *et al*.,2009).

Sin embargo, el éxito de la técnica radica en disminuir los potenciales sesgos inherentes a las metodologías experimentales y análisis de datos. El diseño experimental es crucial para que los datos generados tengan el potencial de responder las preguntas biológicas de interés (Conesa *et al*., 2016). Razón por la cual nuestro estudio presentó tres replicas biológicas por genotipo y dos genotipos por condición. Adicionalmente, con el objetivo de obtener la expresión diferencial con base en la acumulación de β-caroteno, las muestras seleccionadas en este estudio presentaron un alto grado de parentesco, debido a que pertenecen a un cruce F2. Aunque esta condición no está documentada para los análisis de RNA-seq, dada la variación típica de la especie *Manihot esculenta*, en la que se ha encontrado variaciones del tamaño de genoma superiores a 400 Mb (Novoa, 2017) el comparar genotipos emparentados disminuye el ruido de fondo inherente a la variación genética *per se*.

El análisis de datos permitió procesar cerca de 22 000 000 de lecturas en promedio por muestra, con un puntaje Phred mayor a 30 obtenido con el programa FastQC, el 88% de las lecturas mapearon al genoma de referencia. El mapeo de las lecturas con HISAT2 se hizo contra el genoma de referencia y no contra el transcriptoma de referencia con el objetivo de obtener nuevos transcritos. Identificándose un 20% de loci nuevos, los cuales representan una fuente de información aún no estudiada.

El transcriptoma de hoja presentó el mayor número de genes diferencialmente expresados, 8439 genes en total en comparación con los 1437 genes presentes en el transcriptoma de raíz, lo cual se debe principalmente a la mayor actividad biológica del tejido foliar, donde resaltan los procesos fotosintéticos y respiratorios de la planta. Este resultado concuerda con lo reportado por Qing *et al*. en el 2009, en el estudio comparativo de perfiles de expresión utilizando la técnica de *microarrays,* en hojas y raíces de maíz bajo condiciones de estrés salino.

Por su parte, el análisis de anotación y función permitió establecer que la producción y acumulación del β-caroteno en las raíces de yuca está asociada principalmente con la respuesta al estrés por calor, temperatura e intensidad lumínica. Este proceso biológico está muy relacionado con la función que cumple el β-caroteno en las plantas, debido a que son moléculas fotoprotectoras que ayudan a estabilizar las moléculas reactivas del oxígeno, dispersan el calor y absorben la luz (Nissar *et al*, 2015). Un resultado que soporta el papel del β-caroteno como agente estabilizante de las moléculas reactivas del oxígeno, es la actividad óxido-reductasa que en este estudio está enriquecida. Otra función molecular relevante en la acumulación de β-caroteno es la

actividad de la fosfofructoquinasa la cual cataliza la reacción esencialmente irreversible de fosforilación de la fructosa 6 fosfato a fructosa-1,6-difosfato en la glucolisis, proceso catabólico que produce dos moléculas de piruvato a partir de una molécula de glucosa (Melo y Cuamatzi, 2007). El piruvato y el 3 fosfogliceraldehido son los precursores de la ruta del isoprenoide vía MEP, por la cual se forman los β-carotenos en los plastidios.

En la acumulación baja de β-caroteno se presenta una mayor actividad biológica enriquecida, en la cual múltiples procesos biosintéticos están operantes. Sin embargo, resalta el hecho de que la ruta biosintética del esterol este sobre-expresada, dado que presenta el precursor fosfato de farnesilo (FPP) compartido con la ruta del β-caroteno (DellaPenna y Pogson, 2006). Entonces la acumulación baja de β-caroteno está dada por la poca disponibilidad de precursores debido a que múltiples rutas están operantes.

Un aspecto importante en la producción de β-caroteno esta relacionado con la regulación génica. En raíz se establecieron dos genes reguladores: Manes.14G079000 en el cromosoma 14 y Manes.09G032800 en el cromosoma 9 que regulan 19 genes en total y se caracterizan por ser factores de transcripción putativos. En contraste, con los 9 genes reguladores presentes en hoja, que regulan cerca de 470 genes. Destacándose el gen Manes.07G135300 que modula la expresión de 253 genes asociados con la acumulación de β-caroteno.

Es de resaltar que los dos genes reguladores en raíz también están presentes en hoja regulando más de 100 genes. Todos los genes reguladores tienen características de factores de transcripción, y en el caso del gen Manes17G016000 se identificó como el factor de transcripción MYC3 que está involucrado en la respuesta al ácido jasmónico (Fernández *et al*., 2011) y con la respuesta de defensa de las plantas a la herbivoría (Schweizer *et al*., 2013) . Así mismo, el gen Manes.01G007600 se relacionó con el factor de transcripción WRKY2 el cual incluye la respuesta a estrés biótico y abiótico (Niu *et al*., 2012), senescencia, germinación de semillas y procesos de embriogénesis (Ueda *et al*., 2011).

Dada la importancia de la regulación génica en la expresión de características de interés, los genes reguladores encontrados en este estudio se utilizarán en PCRs cuantitativos con el fin de comprobar su expresión diferencial y se utilizarán en el desarrollo de marcadores de selección. Adicionalmente el principal gen de la regulación en raíz Manes.14G079000 será candidato para investigaciones enfocadas en la manipulación genética para la obtención de variedades con altos niveles de β-caroteno.

Dado que la familia GM3736 a la cual pertenecen los genotipos de este estudio, tiene datos obtenidos del mapeo de QTLs desarrollado a partir de secuenciación tipo RAD, datos de genoma completo y datos de metabolómica, los resultados de este trabajo serán integrados con la información genómica y metabolómica para dilucidar la acumulación de β-caroteno en raíces de yuca.

6. Conclusiones

1. El uso de muestras emparentadas aumenta el potencial de la técnica de RNA seq para dilucidar los procesos biológicos que regulan una característica en particular.

2. El uso de transcriptomas para cada tipo de tejido permite realizar un análisis más focalizado y mejora el potencial de análisis de los mecanismos operantes compartimentalizados.

3. La producción y acumulación de β-caroteno se ve favorecida por la disponibilidad del piruvato y por la actividad terpeno sintasa, favoreciendo la síntesis de precursores. Así como también se ve favorecida por la respuesta de la planta frente a estreses abióticos como el calor y la luz.

4. La acumulación baja de β-caroteno se asocia con una actividad metabólica diversa, en especial con la síntesis de esteroles que utiliza los mismos precursores del β-caroteno.

5. Dos genes reguladores tienen el control sobre la acumulación en raíz y 9 genes regulan en la hoja. Estos genes son de vital importancia en los procesos de mejoramiento y presentan la base para futuras investigaciones.

7. Referencias

Biesalski, H.K., Chichili, G.R., Frank, J., von Lintig, J. & Nohr, D. Conversion of β-carotene to retinal pigment. Vitamins and hormones. pp 117-130, 75, 2007.

Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near optimal probabilistic RNA seq quantification. Nature Biotechnology. pp 525-527, 34, 2016.

Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A.L., Pizarro, M., Calle, F. & Dufour, D. Rapid Cycling Recurrent Selection for Increased Carotenoids Content in Cassava Roots. Crop Science, vol. 53. 2013.

Chang, S. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter. pp 113-116, 11(2), 1993.

Conesa, A., Madrigal, P., Tarazona, S. A survey of best practices for RNA-seq data analysis. Genome Biology. pp 13-32, 17, 2016.

DellaPenna, D. & Pogson, B.J. Vitamin synthesis in plants: tocopherols and carotenoids. Annu Rev Plant Biol. pp 711-738, 57, 2006.

El-Sharkawy, M.A. Cassava biology and physiology. *Plant Molecular Biology.* pp 481-501, 56(4), 2004

Fernández-Calvo, P., Chini, A., Fernández-Barbero, G., Chico, J.M., Gimenez-Ibañez, S., Geerinck, J., Eeckhout, D., Schweizer, F., Godoy, M., Franco-Zorrilla, M., Pauwels, L., Witters, E., Puga, M.I., Paz-Ares, J., Goossens, A., Reymond, P., de Jaeger, G. & Solanoa, R. The Arabidopsis bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. The Plant Cell. Pp 701-715, 23, 2011.

FAO. 2016. Disponibl en http://www.fao.org/faostat/en/#data/QC. Fecha de revisión: 29 de Mayo de 2018.

Goff, L., Trapnell, C. & Kelley, D. cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.20.0. 2013.

Jazayeri, S.M., Melgarejo, L.M. & Romero, H.M. RNA-Seq: a glance at technologies and methodologies. Acta biol Colomb. pp 23-35, 20(2), 2015.

Jin J.P., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J.C. & Gao, G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research. pp D1040-D1045, 45(D1), 2016.

Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. pp 357-360, 12, 2015.

La carencia de vitaminas y minerales afecta al desarrollo de un tercio de la población mundial. UNICEF, s.f. Disponible en https://www.unicef.es/noticia/la-carencia-de-vitaminas-y-minerales-afecta-al-desarrollo-de-un-tercio-de-la-poblacion. Fecha de revisión: 9 de abril de 2018.

Latham, M. Nutrición humana en el mundo en desarrollo. Colección FAO: Alimentación y nutrición. Roma, 2002. Disponible en: http://www.fao.org/docrep/006/w0073s/w0073s0j.htm

Li, J. & Tibshirani, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. pp 519-536, 22(5), 2013.

Love, M., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. pp 550, 15, 2014.

Meléndez-Martínez, A.J., Vicario, I.M. & Heredia, F.J. Pigmentos carotenoides: consideraciones estructurales y fisicoquímicas. ALAN. 57(2), 2007.

Melo, V. & Cuamatzi, O. Bioquímica de los procesos metábolicos. 2 ed., Barcelona, 2007.

Moorthy, S.N., Jos, J.S., Nair, R.B. & Sreekumari, M.T. Variability of bcarotene content in cassava germplasm. Food Chem. pp 223-236, 36,1990.

Nisar, N., Li, L., Lu, S., Khin, N.C. & Pogson, B.J. Carotenoid Metabolism in Plants. Molecular Plant. pp 68-82, 8(1), 2015.

Niu, C-F., Wei, W., Zhou, Q-Y., Tian, A-G., Hao, Y-J., Zhang, W-K., Ma, B., Lin, Q., Zhang, Z-B., Zhang, J-S. & Chen, S-Y. Wheat WRKY genes TaWRKY2 and TaWRKY19 regulate abiotic stress tolerance in transgenic Arabidopsis plants. Plant, Cell and Environment. pp 1156-1170, 35, 2012.

Novoa, N. Evaluación del tamaño del genoma en 7 subpoblaciones de yuca (Manihot esculenta Crantz) representativas del banco de germoplasma de CIAT por citometría de flujo. Tesis de Pregrado, Departamento de Ciencias Básicas, Universidad de la Salle, 2017.

Ospina, B. & Ceballos, H. La yuca en el tercer milenio: sistemas modernos de producción, procesamiento, utilización y comercialización. Publicación CIAT No. 327, Cali, 2002. ISBN 958-694-043-8.

Phytozome. Disponible en https://phytozome.jgi.doe.gov/pz/portal.html. Fecha de revisión: 1 de febrero de 2018.

Pantano, L. QCs, figures and analyses after differential expression with DESeq 2 or other similar tool. DEGreport: Report of DEG analysis. R package version 1.13.8. 2017.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. pp 1650-1667, 11(9), 2016.

Quing, D., Lu, H., Li, N., Dong, H., Dong, F & Li, Y. Comparative profiles of gene expression in leaves and roots of maize seedlings under conditions of salt stress and the removal of salt stress. Plant Cell Physiol. pp 889-903, 50(4), 2009.

Schweizer, F., Fernández-Calvo, P., Zander, M., Diez-Diaz, M., Fonseca, S., Glauser, G., Lewsey, M.G., Ecker, J.R., Solano, R. & Reymonda, P. Arabidopsis basic Helix-Loop-Helix transcription factors MYC2, MYC3, and MYC4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. The Plant Cell. pp 3117-3132, 25, 2013.

Trapnell, C., Pachter, L. & Salzberg, S. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. pp 1105-1111, 25 (9), 2009.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. pp 511-515, 28 (5), 2010.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. & Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. pp 46-53, 31, 2013.

Ueda, M., Zhang, Z. & Laux, T. Transcriptional activation of Arabidopsis axis patterning genes WOX8/9 links zygote polarity to embryo development. Developmental Cell. pp 264-270, 20, 2011.

Van Verk, M.C., Hickman, R., Pieterse, C.M. & Van Wees, S.C. RNA-Seq: revelation of the messengers. Trends Plant Sci. pp 175-179, 18 (4), 2013.

Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. pp 57-63, 10(1), 2009.

Wilson, M.C., Mutka, A.M., Hummel, A.W., Berry, J., Chauhan, R.D., Vijayaraghavan, A., Taylor, N.J., Voytas, D.F., Chitwood, D.H. & Bart, R.S. Gene expression atlas for the food security crop cassava. New Phytologist. pp 1632-1641, 213, 2017.

Yuan, H., Zhang, J. Nageswaran, D. & Li, L. Carotenoid metabolism and regulation in horticultural crops. Horticulture research. 2, 2015.

Zhao, S., Fung-Leung W-P., Bittner, A. Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Plos One. 9(1), 2014.