UNIVERSIDAD DE ALMERÍA

Análisis bioinformático de expresión génica diferencial en tomate

Bioinformatic gene expression analysis in tomato

Máster en Biotecnología Industrial y Agroalimentaria

Curso 2020/2021

Convocatoria de julio 2021

Autor: Juan José Iglesias Bueno

Directores: Fernando Juan Yuste Lisbona

Juan Capel Salinas

Resumen

En los últimos años el cultivo de tomate, *Solanum lycopersicum* (Linneo, 1753), ha pasado a ser el segundo de mayor importancia a nivel mundial. Resulta interesante buscar formas de aumentar la producción que a la vez mejoren las propiedades organolépticas o aporten diferencias en el producto que atraigan al consumidor. Entre los estados miembros de la Unión Europea, España es el segundo mayor productor, por detrás de Italia. En España la producción se concentra en Almería y Extremadura.

Se ha estudiado una línea mutante de tomate, *eno* (*excessive number of floral organs)*, estos mutantes desarrollan un mayor número de órganos florales y frutos multiloculares de mayor tamaño. Se ha analizado el ARNm expresado en plantas silvestres y plantas mutantes con el fin de analizar la expresión diferencial durante tres etapas del desarrollo de la flor. Se espera así encontrar qué genes producen los diferentes cambios fenotípicos.

Se han encontrado 4958 genes que se han expresado de forma diferencial. Se ha analizado la función de estos y se ha confirmado que la mutación *eno* produce cambios en los genes expresados que participan en el desarrollo meristemático, de la flor y del fruto.

Abstract

In recent years tomato, *Solanum lycopersicum* (Linnaeus, 1753), has become the second most important vegetable crop worldwide. It is essential to find ways to increase the yield that also improve the organoleptic properties or that provide differences in the final product that are appealing to the consumer. Among the European Union member countries, Spain is the second largest producer, only surpassed by Italy. Inside Spain, the production is concentrated in Almeria and Extremadura.

In this study we have analyzed a tomato mutant called *eno (excessive number of floral organs*), these mutants develop an increased number of floral organs and bigger, multilocular fruits. The mRNA expressed in the wild type and the mutant plants during three different stages of flower development has been analyzed to study its differential expression. That way we expect to find the genes that produced the phenotype changes.

4958 differentially expressed genes have been found. Their function has been analyzed and the changes that the *eno* mutation produces in the expressed genes involved in the meristem development, flower development and fruit development have been confirmed.

Índice

1. Introducción

1.1 El tomate: morfología y taxonomía

El tomate (*Solanum lycopersicum L.,* también conocido como *Lycopersicon esculentum)* es una planta herbácea. Es anual, bianual o perenne. Inicialmente erecta, aunque pasa a ser procumbente, sus ramas llegan a medir 4 metros desde el centro (Peralta et al*.,* 2019*).*

Esta planta posee un sistema radicular secundario muy ramificado, su raíz principal es corta y débil. La base del tallo puede emitir raíces al ponerse en contacto con la tierra, una característica que es aprovechada en varias técnicas de cultivo, como el repicado y el rehundido.

En el tallo principal aparecen hojas, inflorescencias y tallos secundarios, que a su vez repiten el desarrollo de hojas, flores y tallos terciarios. El grosor del tallo en la base, una vez desarrollado oscila entre 2 y 4 centímetros. Sus hojas se sitúan en el tallo de forma alternada, son hojas compuestas y de forma imparipinnada, posee de cinco a nueve foliolos, los cuales son peciolados, de borde dentado y están recubiertos de pelos glandulares.

Las flores son inflorescencias en racimo, cada inflorescencia está formada por entre seis y quince flores, según la variedad. La fecundación de las flores es autógama. El fruto es una baya formada por varios lóculos con placenta en los cuales se asientan las semillas (Serrano, 2009).

Su clasificación taxonómica es:

Reino:	Plantae
Subreino:	Viridiplantae
Infrarreino:	Streptophyta
Superdivisión:	Embryophyta
División:	Tracheophyta
Subdivisión:	Spermatophytina
Clase:	Magnoliopsida
Superorden:	Asteranae
Orden:	Solanales
Familia:	Solanaceae
Género:	*Solanum L*.
Especie:	*Solanum lycopersicum L.*

(ITIS Standard Report Page: *Solanum lycopersicum*, 2019).

1.2 Interés socioeconómico y producción de tomate

El tomate es el segundo cultivo vegetal de mayor importancia comercial, justo después de la patata. La producción mundial es de 100 millones de toneladas de fruto, usando 3,7 millones de hectáreas (Figuras 1 y 2).

El tomate es un cultivo de crecimiento rápido, con un periodo de crecimiento de 90 a 150 días. Su temperatura óptima de crecimiento es de 18 a 25 ºC durante el día y del 10 a 20 ºC durante la noche. Es muy sensible a las temperaturas bajas. Una elevada humedad reduce la producción ya que aumenta la frecuencia de plagas y enfermedades. Puede crecer en una gran variedad de suelos, pero se recomienda que sea seco y tenga un pH de entre 5 y 7. Es relativamente sensible a la salinidad del suelo (Tomato, Food and Agriculture Organization of the United Nations, 2019).

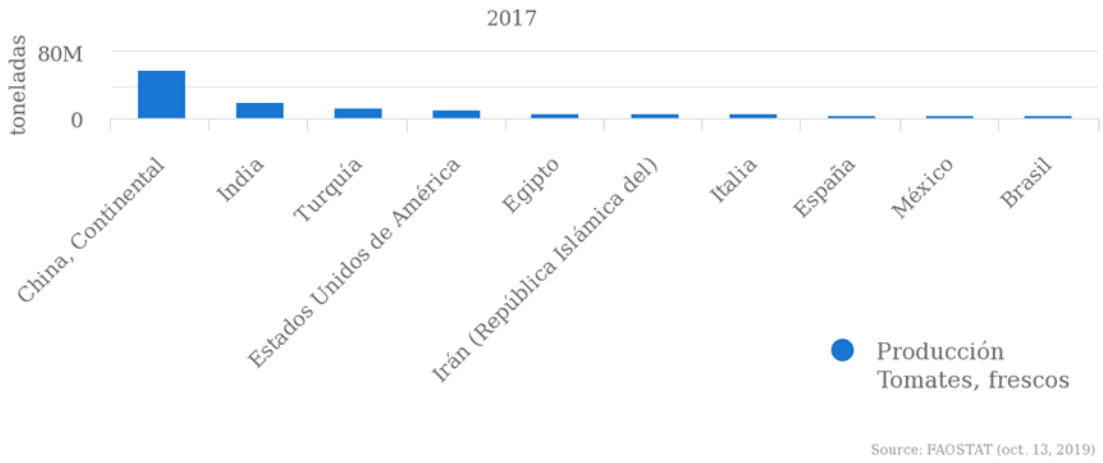

Figura 1. Principales países productores de tomate (FAOSTAT, 2017).

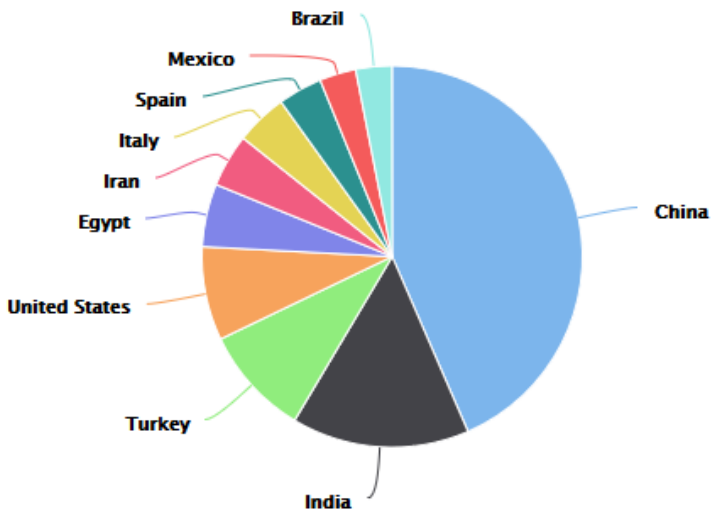

Figura 2. Principales países productores de tomate (FAOSTAT, 2017).

Andalucía exporta más de la mitad de su producción y es la principal suministradora de frutas y hortalizas a la Unión Europea, destino del 96 % de las ventas. Los cinco principales mercados de destino son Alemania, Francia, Países Bajos, Reino Unido e Italia. Más de la mitad de las ventas al exterior de productos hortofrutícolas se concentra en la provincia de Almería (Figuras 3 y 4).

Los tomates son uno de los productos de mayor demanda, en el año 2015 se exportaron 641170 toneladas por un valor total de 665 millones de euros. Almería concentra un 54 % de las exportaciones de tomate españolas. El 80 % del tomate de invernadero se produce mediante control biológico de plagas (Informe sobre "el sector hortofrutícola en Andalucía para su internacionalización", 2017).

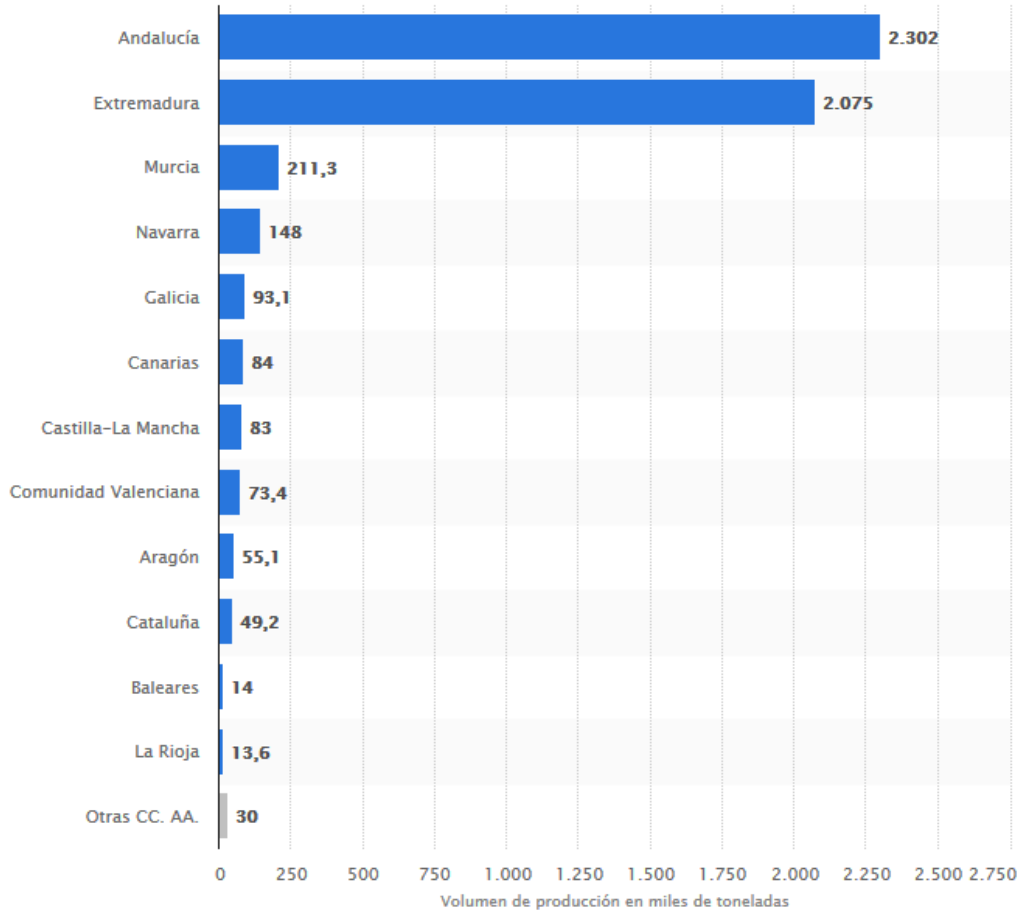

Figura 3. Producción de tomates en España en 2017, por comunidades autónomas, en miles de toneladas (Pérez, 2018).

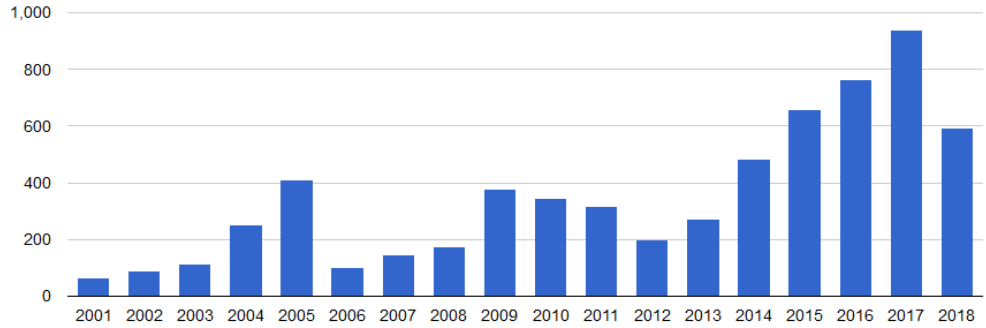

Figura 4. Producción de tomate en Andalucía, en millones de kg (El sector del tomate – Observatorio del tomate para industria, 2018).

1.3 RNA-Seq como herramienta para el análisis de la expresión génica

La secuenciación masiva de ARN (RNA-Seq) es una herramienta indispensable para estudiar los transcriptomas, ya sea para el análisis de expresión diferencial de genes o de los diferentes *splicing* de los ARNm. Gracias al desarrollo de sistemas de alto rendimiento o tecnologías NGS *(Next-Generation Sequencing*), también puede usarse el RNA-Seq para estudiar la traducción o la estructura del ARN.

Las fases esenciales del análisis diferencial de la expresión génica son la extracción de ARN, el enriquecimiento del ARN mensajero o la eliminación del ARN ribosomal, la síntesis de ADN complementario y la preparación de un adaptador apropiado para la genoteca a secuenciar. Luego, la genoteca se secuencia con una cobertura de 10 a 30 millones de *reads* por muestra utilizando una plataforma de secuenciación como por ejemplo *Illumina.*

Los pasos finales son computacionales, se alinean o ensamblan los *reads* secuenciados a un transcriptoma, se cuantifican los *reads* que coinciden con determinados transcritos, se filtra y normaliza entre muestras y se realiza un análisis estadístico de los cambios significativos en la expresión de los genes o tránscritos entre los grupos de muestras (Stark et al., 2019).

El análisis de la expresión génica es ampliamente usado para identificar los mecanismos que controlan procesos celulares o fisiológicos en plantas. Así, por ejemplo, Lemmon et al. (2016) llevaron a cabo una estrategia basada en RNA-Seq para identificar los genes clave implicados en la variación de la arquitectura de la inflorescencia de las solanáceas, centrándose en cinco especies que desarrollan diferentes tipos de inflorescencias, desde una sola flor hasta muy ramificadas. La expresión diferencial de genes ortólogos permitió establecer correlaciones entre la tasa de maduración de los meristemos y la complejidad de la inflorescencia. Así, la inflorescencia más ramificada de *S. peruvianum* mostró un marcado retraso en la tasa de maduración, mientras que *Nicotiana benthamiana* y *Capsicum annuum*, de una sola flor, presentaron una maduración acelerada justo antes de la iniciación floral. Además, el estudio de Lemmon et al. (2016) destaca que los cambios heterocrónicos en dos factores de transcripción

COMPOUND INFLORESCENCE (*S*) y *ANANTHA* (*AN*), previamente caracterizados como genes de identidad de los meristemos de inflorescencia y floral, respectivamente (Lippman et al., 2008), definen la complejidad de la inflorescencia, lo que denota que los cambios transcripcionales en reguladores clave del desarrollo tienen efectos significativos en los rasgos agronómicos.

1.4 Importancia de los meristemos en el desarrollo de inflorescencias y flores

La actividad meristemática ejerce un papel clave en la producción de flores y el rendimiento de los cultivos. A diferencia de *Arabidopsis*, donde la transición floral es un suceso único debido a su naturaleza monopodial, el patrón de desarrollo del tomate es de naturaleza simpodial, de forma que el desarrollo del meristemo apical del tallo es determinado, y su crecimiento finaliza con la aparición de la primera inflorescencia. A continuación, se genera un nuevo brote vegetativo a partir del meristemo axilar de la hoja más joven situada antes de la inflorescencia denominado meristemo simpodial. A partir de este meristemo simpodial, se desarrolla un primer segmento vegetativo simpodial constituido por tres hojas y que se determina con la formación de una nueva inflorescencia. La reiteración de estos segmentos simpodiales de forma sucesiva confiere a la planta de tomate un patrón de crecimiento indeterminado (Lozano et al., 2009).

Una vez ocurre la transición floral, el meristemo apical del tallo adquiere identidad reproductiva, convirtiéndose en meristemo de inflorescencia, cuya actividad permite la diferenciación de meristemos florales donde se establecen los órganos de cada verticilo floral. El locus *S* codifica un factor de transcripción de tipo WOX que actúa regulando la maduración y terminación del meristemo de inflorescencia y, por tanto, el establecimiento de los meristemos florales. Es por ello por lo que los mutantes *s* presentan inflorescencias muy ramificadas y un elevado número de flores (Park et al., 2012).

Durante las etapas iniciales del desarrollo floral, el conjunto de células meristemáticas permanece constante gracias al mecanismo de retroalimentación CLAVATA (CLV)-WOX (WUS), una vía de señalización muy conservada en diversas especies. *WUS* determina como proliferan las células madre y hace que se exprese *CLV3*, que codifica un péptido ligando que se une a varios receptores de membrana que desencadenan una cascada de señalización que acaba disminuyendo la actividad de *WUS*. En tomate, estudios recientes demuestran que dos mutaciones en la ruta meristemática CLV-WUS, *locule number* (*lc*) y *fasciated* (*fas*), son las principales responsables del aumento del tamaño del fruto que tuvo lugar durante el proceso de domesticación y mejora de esta especie. La mutación *lc* afecta a una región reguladora del gen ortólogo a *WUS*, mientras que *fas* se localiza en el promotor del ortólogo a *CLV3*. Ambas mutaciones actúan de forma sinérgica aumentando el tamaño del meristemo floral, lo que da lugar a un incremento en el número de carpelos en la flor y con ello, al desarrollo de frutos pluriloculares de gran tamaño (Xu et al., 2015).

1.5 *ENO* determina el tamaño del fruto regulando la actividad del meristemo floral

El gen *EXCESSIVE NUMBER OF FLORAL ORGANS* (*ENO*, Solyc03g117230) sintetiza un factor de transcripción de la superfamilia APETALA2/Ethylene Responsive Factor (AP2/ERF) en concreto un ERF del grupo VIII. Este factor de transcripción regula la actividad en los meristemos florales (Yuste-Lisbona et al., 2020).

ENO es capaz de unirse a una región específica del promotor de *SlWUS*, lo cual sugiere que ENO regula los dominios de expresión de *SlWUS* de manera directa, restringiendo la proliferación de células madre en las flores. La mutación *eno* se debe a un polimorfismo de nucleótido único (SNP) en el codón de inicio del gen Solyc03g117230 así como otro SNP y un InDel en la región 5'UTR del mismo locus (Figura 5). De este modo, el cambio de adenina por timina en el codón de inicio de la traducción da lugar a un cambio en el marco abierto de lectura que genera una proteína no funcional. Dicha mutación causa un incremento del tamaño del meristemo floral que está asociado a una expansión de los dominios de expresión de *SlWUS*, lo que a su vez conduce al desarrollo de flores con un mayor número de órganos en los verticilos más internos (pétalos, estambres y carpelos) y, en consecuencia, frutos multiloculares de gran tamaño.

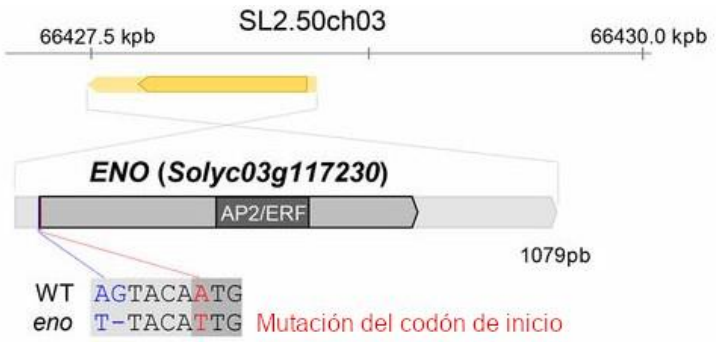

Figura 5. Caracterización molecular de la mutación *eno*. El gen *ENO* se localiza en el cromosoma 3, la región codificante aparece de color gris oscuro y la región no traducida (UTR) de color gris claro. La mutación en el codón de inicio de la traducción aparece marcada en color rojo. Las mutaciones en la región 5'-UTR aparecen en color azul. Adaptado de Yuste-Lisbona et al. (2020).

Además, a través de la caracterización de mutantes dobles (*eno:lc* y *eno:fas*) y triple (*eno:lc;fas*), se ha demostrado que la mutación *eno* tiene un efecto sinérgico con las mutaciones *lc* y *fas* que permite obtener frutos aún mayores (Figura 6), lo cual indica que las mutaciones *eno*, *fas* y *lc* afectan a genes diferentes, pero funcionalmente relacionados, que son necesarios para regular el tamaño del meristemo floral (Yuste-Lisbona et al., 2020).

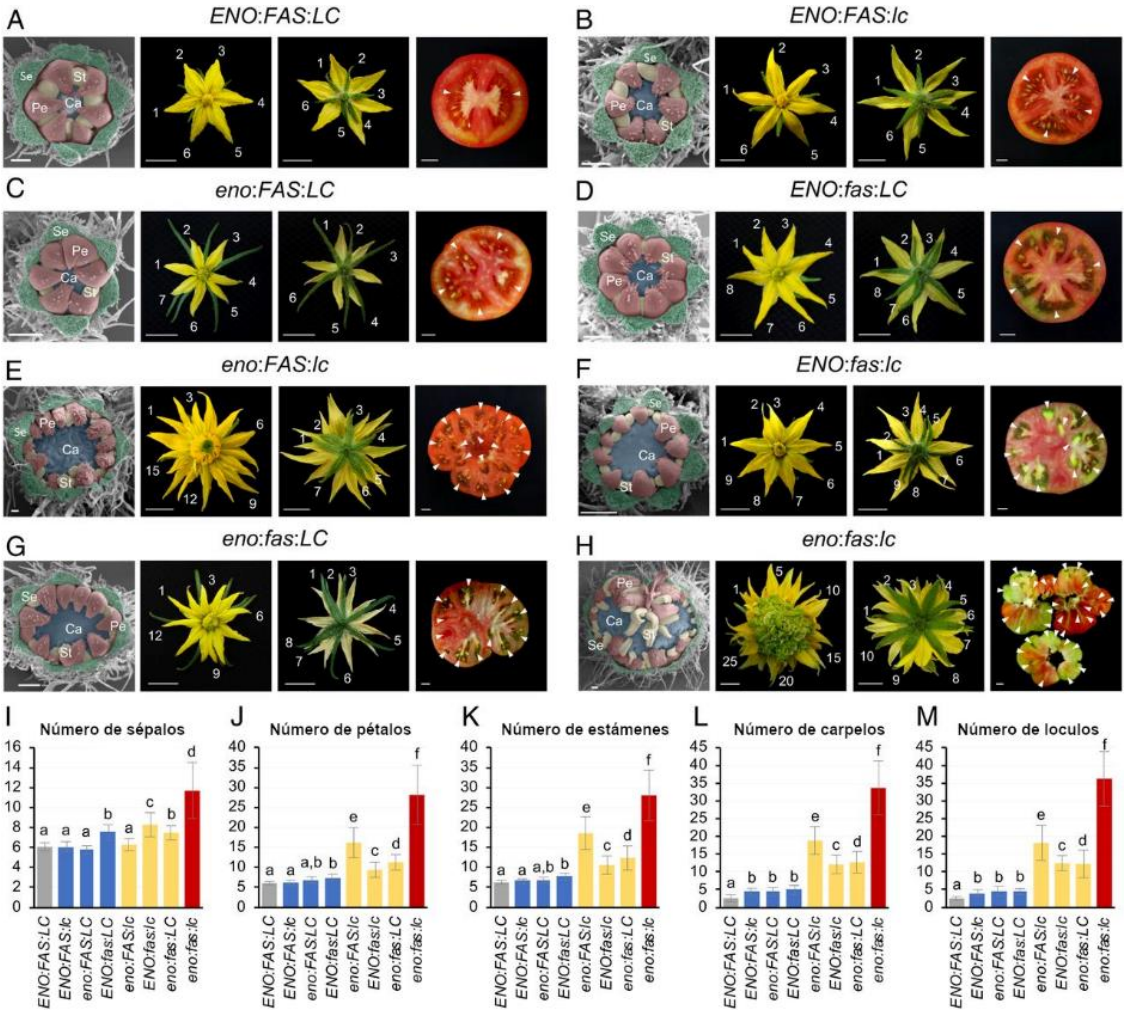

Figura 6. Interacciones genéticas de los loci *ENO*, *FAS* y *LC*. Fotografías de los meristemos florales, flores y frutos de las diferentes combinaciones alélicas de los loci *ENO*, *FAS* y *LC* (A-H). Se, sépalos; Pe, pétalos, Sta, estambres y Ca, carpelos. Se eliminaron los sépalos para hacer las fotografías de los meristemos florales. Se numera el número de pétalos y sépalos. Los lóculos aparecen indicados por flechas blancas. La escala de los meristemos florales corresponde a 200 µm, mientras que la de flores y frutos equivale a 1 cm. Número de sépalos (I), número de pétalos (J), número de estambres (K), número de carpelos (L), número de lóculos (M) en plantas silvestres (gris), mutantes simples (azul), mutantes dobles (amarillo) y el mutante triple (rojo) para los alelos *eno*, *fas* y *lc*. Adaptado de Yuste-Lisbona et al. (2020).

Asimismo, el estudio de la historia evolutiva de los *loci ENO*, *FAS* y *LC* revela que en *S. pimpinellifolium*, ancestro silvestre inmediato del tomate cultivado, existe un haplotipo del gen *ENO* asociado al incremento del número de lóculos en el fruto, y que es portador de una deleción en su promotor. Dicho haplotipo fue seleccionado durante el proceso de domesticación, estableciendo así el fondo genético propicio para aumentar el tamaño de fruto en los cultivares actuales a través de las mutaciones *lc* y *fas*, estas últimas seleccionadas durante el proceso de mejora genética del tomate (Yuste-Lisbona et al., 2020).

2 Justificación y objetivos

Debido a la elevada demanda a nivel mundial del tomate, resulta de especial interés desarrollar nuevas variedades que permitan obtener una mayor producción o bien que consigan un nicho propio en el mercado debido a sus diferencias con las variedades existentes; consiguiendo así aprovechar el interés económico existente de esta especie. En este contexto, este trabajo de investigación se propone analizar las interacciones génicas que el gen *ENO* puede mantener con otros genes ya descritos e implicados en el control del desarrollo reproductivo de tomate. Con tal propósito, dado que el incremento en el tamaño de los frutos del mutante *eno* es consecuencia de alteraciones ocasionadas durante el desarrollo floral, se analizarán mediante RNA-Seq réplicas biológicas de flores en diferentes etapas de desarrollo de plantas de fenotipo silvestre y mutante. Así, los objetivos específicos de este Trabajo Fin de Máster se exponen a continuación:

1. Analizar la expresión génica diferencial entre muestras silvestres y mutantes.
2. Evaluar la calidad de las muestras RNA-Seq y del análisis de expresión diferencial realizado.
3. Definir cuáles son los genes y mecanismos moleculares alterados por la mutación *eno* durante el desarrollo floral de tomate.

3 Material y métodos

3.1 Bases de datos y navegadores genómicos

Las bases de datos genómicos son repositorios online de información específica relacionada con la biología molecular, descritas para uno o más genes o bien de una población específica de la especie. Estas bases de datos forman una parte integral de la bioinformática, pues alojan gran cantidad de información que resulta indispensable para el funcionamiento correcto de los distintos programas usados durante el análisis de las muestras. Debido a la gran cantidad de datos disponibles, es frecuente encontrar bases de datos especializadas en determinadas especies o familias.

Estas bases de datos pueden contar además con su propio navegador genómico, que puede ser un programa o bien una aplicación web, que permite analizar el genoma de la especie junto con las anotaciones génicas de interés. Esto facilita enormemente la visualización y extracción de la información, debido tanto a la facilidad propia que aporta usar una interfaz gráfica como a que estos navegadores contextualizan las anotaciones y muestras analizadas, al permitir observar secciones específicas del genoma estudiado (Kent et al., 2002).

Gene Ontology (GO) es la mayor base de datos especializada en el funcionamiento de los genes y de los productos génicos. Toda la información que contiene está disponible tanto en formatos destinados a ser interpretados por humanos como en formatos destinados a máquinas, lo que la hace fundamental para el análisis computacional de los resultados de experimentos de biología molecular y de genética. Las ontologías de esta base de datos permiten tener una descripción consistente de los productos génicos entre las diferentes bases de datos existentes (Ashburner et al., 2000; The Gene Ontology Consortium, 2018). Gracias a Gene Ontology ha sido posible relacionar la información obtenida en las otras bases de datos.

Sol Genomics Network (SGN) es una base de datos genómicos y fenotípicos específica de la familia Solanaceae. Contiene datos del genoma completo de diversas especies, entre ellas el tomate. Además, Sol Genomics Network posee un navegador genómico basado en JBrowse especializado en el trabajo con genes de tomate (Fernandez-Pozo et al., 2014; Buels et al., 2016). Se ha utilizado para obtener versiones actualizadas del genoma de *S. lycopersicum* y sus anotaciones.

Uniprot es una base de datos que reúne secuencias y anotaciones de proteínas. Su objetivo es ofrecer una base de datos de secuencias de proteínas y su función completa y de alta calidad. Aproximadamente medio millón de estas anotaciones son extraídas de la literatura por expertos, lo que asegura su calidad (The UniProt Consortium, 2018). Su uso ha permitido conocer las funciones de las proteínas correspondientes a los *reads* analizados.

3.2 Formatos utilizados en el análisis

La información contenida en estas bases de datos suele usar formatos específicos para facilitar así la compatibilidad entre múltiples programas. Es frecuente trabajar con

archivos de formato FASTA (.fasta o .fa), GFF (.gff, .gtf o .gff3), BED (.bed) o BAM (.bam). Los formatos FASTA, GFF y BED son archivos de texto con una estructura propia que permite usarlos con facilidad cuando se trabaja con múltiples programas, mientras que los archivos de formato BAM son binarios por lo tanto es necesario usar un programa específico para poder leerlos.

Los archivos de formato FASTA se usan para guardar secuencias de nucleótidos o aminoácidos. Están formados por dos secciones, la primera se denomina encabezado y comienza con el carácter ">" seguido del nombre o identificador de la secuencia estudiada. En la siguiente línea comienza la secuencia y se ignoran todos los caracteres que no sean letras. Un archivo puede contener varias secuencias siempre y cuando cada una tenga su propio encabezado. Los resultados de la secuenciación se dan en este formato y son los que se han alineado con *TopHat*.

Los archivos GFF (*General Feature Format*) se usan para almacenar información de un conjunto de entradas, cada entrada ocupa una línea, cada línea está dividida en columnas. Estas columnas están delimitadas mediante tabulaciones y contienen datos como el nombre de la secuencia, el nombre del programa o base de datos que generó la entrada, la clase de entrada (*Gene*, *Variation*, *Similarity*), la posición de inicio y final de la entrada, la calidad de la entrada, la hebra + (sentido) o – (antisentido) y el marco de lectura. Gracias a estos archivos es posible acceder a una gran cantidad de información con solo conocer el nombre de la secuencia. Su uso ha sido necesario para realizar el conteo de los *reads* alineados con *HTSeq.*

Los archivos BED (*Browser Extensible Data)* permiten definir los datos que se van a mostrar de una anotación. Están formados por tres columnas obligatorias y hasta nueve columnas adicionales opcionales. Los campos obligatorios son el cromosoma, la posición de inicio de la entrada en el cromosoma y la posición final. Las columnas opcionales permiten añadir información como el nombre de la anotación, la hebra, el número de exones contenidos o que partes resaltar de la anotación. Estos archivos son especialmente útiles a la hora de trabajar con navegadores genómicos.

El formato BAM (*Binary Alignment Map)* es la versión binaria y comprimida de los archivos SAM (*Sequence Alignment Map)*, que se usa para almacenar grandes cantidades de alineamientos de secuencias. Estos archivos se dividen en dos secciones, una de encabezado y otra de alineamientos. Las líneas de encabezado empiezan con el carácter "@". Todas las líneas están delimitadas mediante tabulaciones. Este formato tiene 11 campos obligatorios y múltiples campos opcionales. Entre ellos destacan el nombre del *read*, el nombre de la secuencia de referencia, la calidad de mapeo, la calidad de la entrada o la posición de inicio del alineamiento (Li, H. et al., 2009). Este tipo de archivos se ha usado para almacenar la información obtenida tras realizar el alineamiento de los *reads* al genoma de referencia.

3.3 Alineamiento de los *reads*

En primer lugar, se ha usado el programa *TopHat* para alinear los *reads* de RNA-Seq al genoma de *Solanum lycopersicum* con el fin de identificar las distintas uniones de exones que se han producido durante el *splicing* del ARN. *TopHat* se diseñó para trabajar con *reads* obtenidos mediante *Illumina Genome Analyzer*.

TopHat primero mapea los *reads* de RNA-Seq al genoma e identifica los posibles exones ya que muchos *reads* se alinearán de forma contigua sobre el genoma. Con esa información inicial*, TopHat* crea una base de datos con las posibles uniones de exones producidas tras el *splicing* y luego mapea los *reads* frente a esas hipotéticas uniones de exones para confirmarlas (Figura 7).

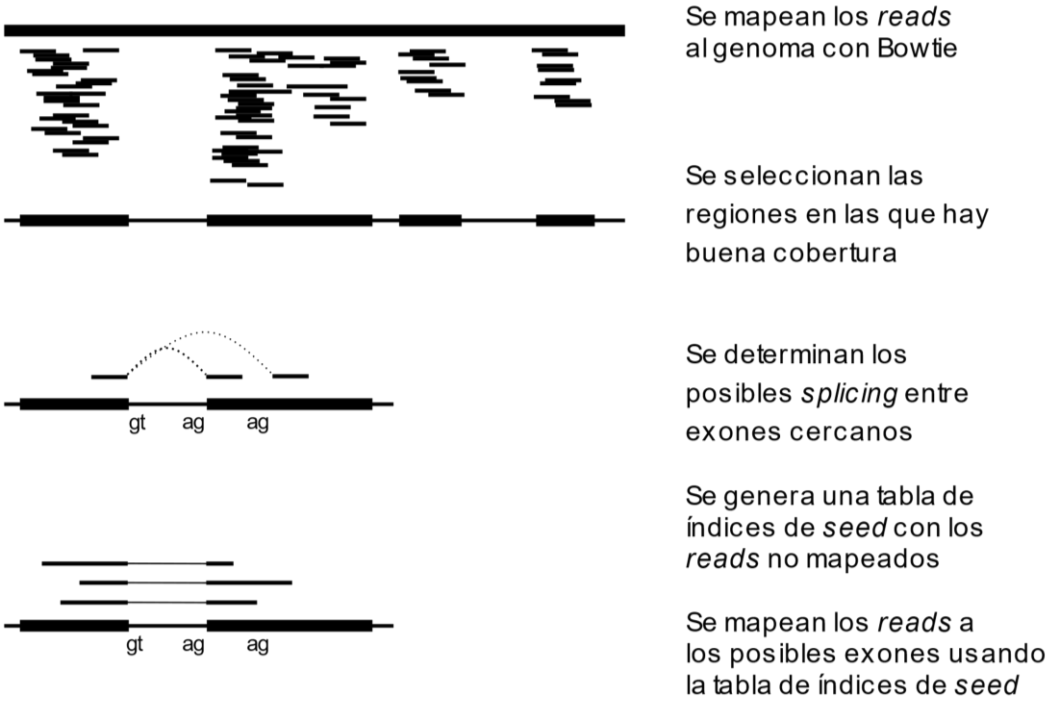

Figura 7. Funcionamiento de *TopHat,* adaptada de Trapnell et al. (2009).

Con el fin de evitar problemas con exones de menor tamaño, *TopHat* fragmenta todos los *reads* en fragmentos más pequeños que se mapean independientemente. Para generar la base de datos de posibles uniones de exones, *TopHat* analiza dos parámetros:

i) El primero y más importante es ver cuando dos fragmentos del mismo *read* son alineados a una determinada distancia en la misma secuencia genómica, o bien cuando un fragmento interno no puede alinearse, lo que sugiere que el *read* abarca múltiples exones. Para confirmar un *splicing*, se genera una tabla con los *seed*, que están formados por parte de la secuencia que está antes y después de los exones. Luego este *seed* se usa para realizar comparaciones adicionales con los *reads* no mapeados inicialmente (Figura 8).

ii) El segundo parámetro son las "*coverage islands*", formadas por las regiones del mapeo inicial en las que se acumulan gran cantidad de *reads*. Dos "*coverage islands*" vecinas suelen ser empalmadas juntas durante el *splicing*, apareciendo ambas en el transcriptoma; por lo que *TopHat* busca formas de unirlas junto con un intrón.

Tras ejecutar *TopHat* se han obtenido una serie de archivos, de entre estos archivos destaca el archivo "accepted_hits.bam", que contiene una lista de los alineamientos en formato BAM. El archivo "junctions.bed" contiene una lista de las uniones de exones encontradas por *TopHa*t. Cada unión está formada por dos bloques BED conectados y se les asigna una puntuación (Trapnell et al., 2009).

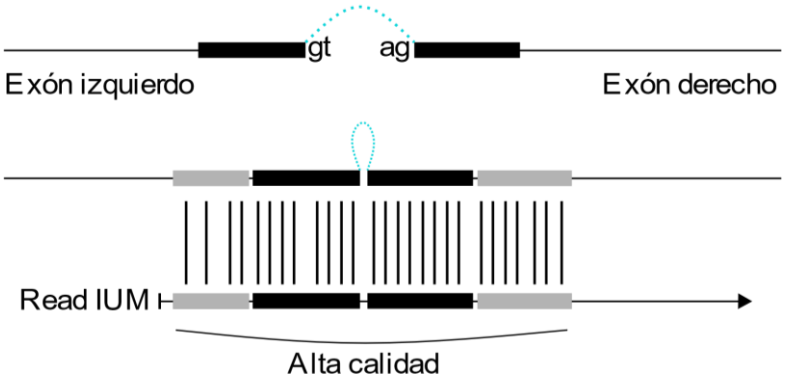

Figura 8. Uso de los *Initially UnMapped* (IUM) o *reads* no mapeados inicialmente, y de los *seed,* que aparecen de color gris en la figura, para confirmar posibles sitios en los que ha habido *splicing.* Adaptada de Trapnell et al. (2009).

3.4 Cuantificación de los *reads*

Para llevar a cabo el conteo de los *reads* se ha utilizado *HTSeq. HTSeq* es un paquete de Python que realiza múltiples análisis cuyo fin es preparar los resultados obtenidos de un programa para ser usados como *input* del siguiente. *HTSeq* permite realizar análisis como obtener resúmenes estadísticos para estudiar la calidad de los datos, calcular vectores de cobertura que luego pueden visualizarse en navegadores genómicos, leer anotaciones en formato GFF o contar el número de *reads* que hay en un archivo.

Para contar los *reads* se ha utilizado el *script htseq-count*, el cual permite contar cuantos *reads* han sido mapeados sobre cada uno de los exones o unión de exones. Si alguno de los *reads* mapeados no se corresponde con un exón, este aparecerá como "*no_feature*" en los resultados. Si un *read* se corresponde con más de un exón, este aparecerá como "*ambiguous*" (Figura 9).

Al evaluar la expresión diferencial se han tenido en cuenta solo aquellos *reads* que se mapean de forma no ambigua a un único gen, descartando aquellos que se asignan a múltiples posiciones o que se solapan con más de un gen. Esto es debido a que si hubiese dos genes que tuviesen alguna similitud en su secuencia, de los cuales uno se expresa diferencialmente y el otro no, aquellos *reads* que se corresponden al gen

expresado diferencialmente harían que el otro gen también se considerase diferencialmente expresado, generando un falso positivo.

Para ejecutar el script es necesario aportar además de los archivos de alineamiento un fichero GFF de la especie estudiada. Estos archivos contienen información de cada "*Genomic Feature*" de la especie de estudio. En este caso el archivo usado considera los exones como *feature*.

El resultado obtenido es un fichero que contiene una lista de los exones junto con el número de *reads* que se le han alineado y el número de *reads* ambiguos o que no tienen un exón que le corresponda (Anders et al., 2015).

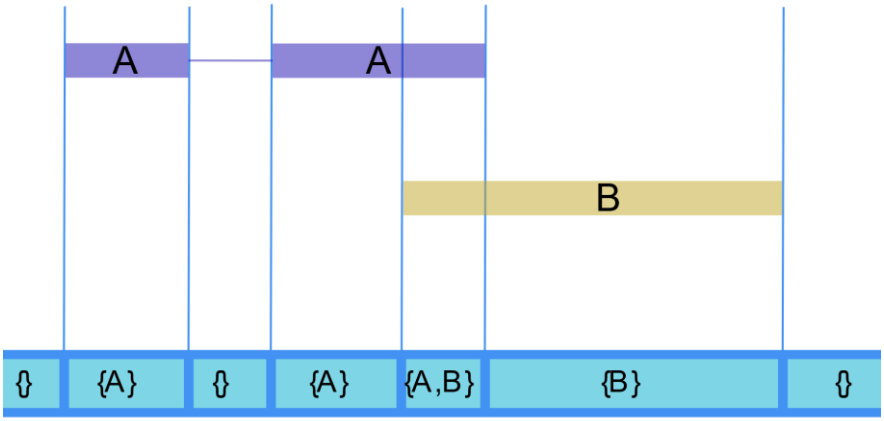

Figura 9. Ejemplo de cómo *HTSeq* representa como se solapan los metadatos de la anotación. Cada exón se alinea sobre el conjunto formado por los *reads*, y los divide de forma interna en fragmentos, a los cuales les asigna un valor que corresponde al conjunto formado por aquellos elementos que se solapan en ese fragmento determinado. Adaptada de Anders et al. (2015).

3.5 Análisis de expresión diferencial

Para analizar la expresión diferencial se ha utilizado el paquete *DESeq2* de R. *DESeq2* usa modelos lineales generalizados de la distribución binomial negativa para analizar la expresión diferencial (Love et al., 2014).

Se utilizó *HTSeq* para realizar el conteo debido a que *DESeq2* contiene funciones específicas que le permiten exportar los resultados de *HTSeq* fácilmente. Debe tenerse en cuenta que *DESeq2* trabaja con datos que no están normalizados, realiza sus propias normalizaciones y correcciones de forma automática.

Al realizar los análisis estadísticos con *DESeq2* se usa un objeto de clase *DESeqDataSet*, la utilidad de usar este tipo de objetos es su compatibilidad con otros paquetes y a que las filas pueden asociarse con rangos genómicos, lo que facilita el trabajar con estos resultados. Se debe tener especial cuidado al generar el parámetro

design, este parámetro expresa las variables que va a usar el modelo e influye en las dispersiones y *log_2 fold changes* obtenidos (Love et al., 2014).

Antes de ejecutar el análisis se realizó un prefiltrado para eliminar aquellas filas que tenían un número bajo de *reads*. Con respecto a los niveles, se asignó como nivel de referencia a los mutantes usando el argumento *contrast.*

El análisis se ha realizado usando la función *DESeq*. Las tablas de resultados generadas contienen los *log_2 fold change*, el error estándar de este y el resultado de la prueba de Wald junto con su *p-value* y *p-value* ajustado.

El *log_2 fold change* se puede interpretar de la siguiente forma:

$$fold\ change = \frac{conteo\ normalizado\ en\ plantas\ mutantes}{conteo\ normalizado\ en\ plantas\ silvestres}$$

Los *p-value* ajustados se han calculado a partir de los *p-value* utilizando el procedimiento de Benjamini-Hochberg el cual permite reducir la tasa falsos positivos (*false discovery rate, FDR*), este ajuste de la tasa permite reducir el número de *p-value* que dan valores menores del límite establecido (por ejemplo; 0,05) debidos al azar (Benjamini y Hochberg, 1995).

Tras obtener estos resultados se han ordenado las tablas en función del *p value*, y se ha asignado un valor absoluto mínimo para *log_2 fold change* para así ir filtrando los resultados. Cuando se obtienen los resúmenes de los resultados es importante indicar el valor apropiado del parámetro *alpha*, en este caso se ha usado 0,05.

Al interpretar los resultados se debe tener en cuenta que hay parámetros que se han tenido que estimar durante el análisis. Por ejemplo, la variabilidad dentro de cada grupo (la variabilidad entre réplicas) debe de estimarse de forma precisa ya que su valor es crítico para poder inferir si existe expresión diferencial. Para este tipo de experimento en el que se ha trabajado con una muestra y dos réplicas para cada condición se genera una gran dispersión entre los estimadores de la variabilidad para cada gen, por lo que estos estimadores no pueden usarse directamente ya que reducirían la precisión del análisis de la expresión diferencial.

Para evitar ese problema se ha compartido la información de distintos genes, es decir, DESeq2 asume que los distintos genes que tienen una expresión promedio similar tienen también una dispersión similar.

En primer lugar, se trata cada gen por separado y se calculan estimadores de máxima verosimilitud para la dispersión de cada gen. Luego, se ajusta una curva a esos valores, lo que permite obtener un estimador preciso de la dispersión esperada para los genes de un nivel de expresión dado, aunque no representa la desviación de cada gen individual. Finalmente, se contraen los estimadores de la dispersión de cada uno de los genes hacia los valores predichos por la curva para obtener así los valores de dispersión finales.

La intensidad de la contracción se determina mediante métodos bayesianos, en los que influyen la estimación de cómo de cerca están los valores reales de dispersión de los valores ajustados y de los grados de libertad. Ese enfoque tiene en consideración la variación específica de cada gen hasta el punto en que lo permitan los datos iniciales, mientras que el uso de la curva ajustada permite ayudar en la estimación cuando se trabaja con muestras con menos cantidad de información de cada gen individual.

El uso de esta contracción ha permitido evitar posibles falsos positivos que pueden derivar de subestimaciones de la dispersión. Para aquellos genes que tenían un estimador de la dispersión individual que se encuentre a una distancia mayor que dos desviaciones estándar por encima de la curva se ha utilizado ese estimador de la estimación en lugar del obtenido tras la contracción (Love et al., 2014).

3.5.1 El modelo matemático de DESeq2

El análisis de la expresión diferencial de DESeq2 usa el siguiente modelo lineal generalizado:

$$K_{ij\sim} BN\left(\mu_{ij}, \alpha_i\right)$$

$$\mu_{ij} = s_j q_{ij}$$

$$log_2\left(q_{ij}\right) = x_{j.}\beta_i$$

Donde K_{ij} representa el número de *reads* para cada gen i, muestra j y se modeliza usando la distribución binomial negativa con una media ajustada μ_{ij} y un parámetro de dispersión específico de cada gen α_i.

La media ajustada μ_{ij} está formada por un factor de tamaño específico de cada muestra $s_{j.}$ y de un parámetro q_{ij} que es proporcional a la concentración de fragmentos esperada para cada muestra j. Los coeficientes β_i dan los *log₂ fold changes* de cada gen i para cada columna j de la matriz del modelo X.

El parámetro de dispersión α_i define la relación entre la varianza de los conteos observador y su media, que dependerá tanto del factor de tamaño s_j como del parámetro dependiente de la covarianza q_{ij} (Love et al., 2014).

$$Var\left(K_{ij}\right) = E\left[\left(K_{ij} - \mu_{ij}\right)^2\right] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Los pasos llevados a cabo por DESeq2 para realizar el análisis son:

1) Estimar los parámetros de tamaño s_j.

2) Estimar los parámetros de dispersión α_i.

3) Ajustar la binomial negativa para β_i y calcular los estadísticos de Wald.

3.6 Material vegetal

Dado que el incremento en el tamaño de los frutos del mutante *eno* es consecuencia de alteraciones ocasionadas durante el desarrollo floral, se ha analizado mediante RNA-Seq tres réplicas biológicas de flores en diferentes etapas de desarrollo de plantas de fenotipo silvestre y mutante, a saber: botones florales de 3 a 6 mm (BF0) y de 6 a 12 mm (BF1) de longitud, así como flores en antesis (AD). Para ello, el ARN total celular de las 18 muestras recolectadas (9 del control y 9 del mutante *eno*) se extrajo con Trizol (Invitrogen), siguiendo el protocolo proporcionado por el fabricante. Dicho ARN fue enviado al centro de genómica del Instituto Max Planck de Mejora Vegetal, donde se prepararon las librerías de cDNA siguiendo el protocolo proporcionado por Illumina Inc. Finalmente, estas librerías fueron secuenciadas con el equipo HiSeq 2000 de Illumina Inc.

4 Resultados y discusión

4.1 Rendimiento de los alineamientos y normalización de los *reads*

Tras realizar el alineamiento con *TopHat* se han analizado los resultados obtenidos. Para poder visualizar el archivo "accepted_hits.bam" se ha usado el programa *Samtools*, en la Tabla 1 pueden verse parte de los primeros cinco resultados de la muestra "mut_AD.1" como ejemplo.

Como puede verse en la Tabla 1, los resultados obtenidos tienen un formato pensado para transmitir información entre programas, lo que los hace algo complejos de interpretar. La columna FLAG contiene valores enteros en base 2, que se corresponden a un valor en binario, lo que permite contener mucha información usando solo unos pocos caracteres. En este caso, solo aparece el valor "16", que significa que la secuencia del *read* analizado es complementaria de forma reversa. En la columna RNAME se puede ver que los 5 *reads* se han alineado sobre la misma secuencia génica, la columna POS indica la coordenada en la que empieza a alinearse la primera base del segmento.

Con la información de RNAME y POS puede verse que los segmentos de los *reads* pueden solaparse, como se mostraba en la Figura 7. En este caso se observa que existe mejor cobertura en la región que cubren los tres últimos segmentos.

La columna CIGAR (*Concise Idiosyncratic Gapped Alignment Report)* representa las diferencias encontradas en el alineamiento. Como puede verse aparecen las letras M, D e I, que son la inicial de *Match* (coinciden)*, Deletion* (deleción) e *Insertion* (inserción). Así pues, M indica que en las posiciones siguientes coinciden el *read* y el genoma de referencia, D indica que no coinciden debido a que se ha perdido parte de la secuencia e I indica que no coinciden debido a que se han insertado bases adicionales. Por ejemplo, en la primera muestra la cadena CIGAR es 23M1I13M1D63M, lo que quiere decir que las primeras 23 bases coinciden, 1 base es una inserción, las siguientes 13 bases coinciden, 1 base se ha perdido y las últimas 63 bases coinciden. En el resto de las muestras de la tabla se han dado 100 *Match*.

Las columnas MAPQ y QUAL miden el nivel de calidad Phred del mapeo y de la asignación de cada base del segmento, de forma respectiva. El nivel de calidad Phred (Q) se define en base a la probabilidad de error de la asignación de base durante la secuenciación (P).

$$Q = -10\,P$$

El valor que aparece en la columna MAPQ es el valor de Q, en este caso para Q=50, tenemos que la probabilidad de que el mapeo sea erróneo es de 10^{-5}.

Al trabajar con datos de RNA-Seq es necesario dar un valor para cada base de los *reads,* se utiliza el código ASCII para asignar a cada posible nivel de calidad un carácter y así poder visualizar mejor la información. Puesto que los primeros 33 caracteres de ASCII no tienen una grafía asociada, es habitual usar el valor de Q más 33 cuando se espera visualizar los datos y no solo procesarlos con otro programa. Así pues, un Phred

de 0 se correspondería con el carácter "!", y un Phred de 32 se correspondería con el carácter "A". Como puede deducirse de la fórmula del Phred, un valor de 30 se corresponde a una probabilidad de 0.001 de haber llamado de forma errónea la base.

Cada una de las bases de la secuencia génica de la columna SEQ se corresponde con el carácter que ocupa la misma posición en la columna QUAL. En la primera muestra las primeras tres bases son “GCT” y su calidad correspondiente es “CCC”, esto quiere decir que la probabilidad de que estas bases se hayan asignado de forma errónea es de 10^{-34}.

Una forma más concisa de mostrar parte de estos resultados es utilizando un navegador genómico para alinear sobre el genoma de *Solanum lycopersicum* los *reads* contenidos en el archivo BAM. Por ejemplo, utilizando el navegador genómico de Sol Genomics Network pueden visualizarse los mismos genes de la Tabla 1, como puede verse en la Figura 10. Esta visualización permite ver rápidamente dónde se solapan múltiples *reads* y también permite la secuencia del genoma analizado, puede verse que de hecho las bases de los *reads* eran correctas y por eso la calidad era alta. Como puede verse en la Figura 11, estos navegadores se usan para ver la gran cantidad de datos que se están tratando de forma concisa y así decidir qué zonas del genoma pueden ser de interés en el estudio.

En la Figura 12 pueden verse los *reads* totales de cada muestra junto con los *reads* alineados solamente una vez por *TopHat*. A pesar de las diferencias en los números de *reads* y de *reads* alineados, puede verse que la proporción es similar.

Hasta este punto se han visto distintos modos de analizar cada muestra de forma individual, esto es importante ya que es necesario asegurar la calidad de los datos iniciales antes de introducirlos en las siguientes etapas del análisis, pero el verdadero objetivo del proceso es realizar un análisis de expresión diferencial entre las distintas muestras, ya sea para ver si existen diferencias en la expresión de genes entre las etapas del desarrollo floral o entre los genotipos silvestre y mutante. Para llevar a cabo este análisis de expresión diferencial, en primer lugar, se deben normalizar los datos para que las distintas muestras puedan ser comparadas entre sí.

RPKM (*Reads Per Kilobase Million*, *reads* por millón de kilobases) es una forma de normalizar los resultados de un experimento de RNA-seq teniendo en cuenta la cobertura de secuenciación y la longitud de los genes. Para calcularlo se cuenta el total de *reads* en cada muestra y se divide entre un millón, el valor obtenido es el factor de escala. Luego se divide el número de *reads* de cada gen entre el factor de escala, normalizando así para la cobertura de secuenciación, y finalmente este valor se divide entre la longitud en *kilobases* del gen correspondiente, normalizando el valor obtenido con la longitud de cada gen.

Tabla 1. *Reads* de la muestra “mut_AD.1” alineados con TopHat

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	SEQ	QUAL
2206:1465:87362	0	SL2.50ch00	23496	50	23M1I13M1D63M	GCTTCTTAAAGCTTTTAAT	CCCFFFFFHGHHHJJJJJJJ
2313:13671:56681	16	SL2.50ch00	549058	50	100M	GAAAAAATTGGTAGAGTTT	CDDDDDDCDCCEEDDFFFFF
1101:13143:99575	16	SL2.50ch00	549081	50	100M	TTCTTTTTGGTTCGTATAA	CCBCCCCBC>=B?:FEHHHH
1116:16058:19253	16	SL2.50ch00	549081	50	100M	TTCTTTTTGGTTCGTATAA	DDDDBBBBDCFDFDHHEHFH
2102:4953:17043	16	SL2.50ch00	549081	50	100M	TTCTTTTTGGTTCGTATAA	DDDBDDDDDDFFFFHHGHGH

Abreviaturas: QNAME: Nombre de la entrada, FLAG: los bits que hacen de *flag*, RNAME: nombre de la secuencia de referencia, MAPQ: calidad de mapeo, CIGAR: *Concise Idiosyncratic Gapped Alignment Report*, SEQ: secuencia del segmento, QUAL: Calidad de cada base del segmento. Se ha acortado el QNAME eliminando la parte común de las cinco entradas (“HWI-ST863:347:C55KDACXX:5:”) y se han acortado SEQ y QUAL.

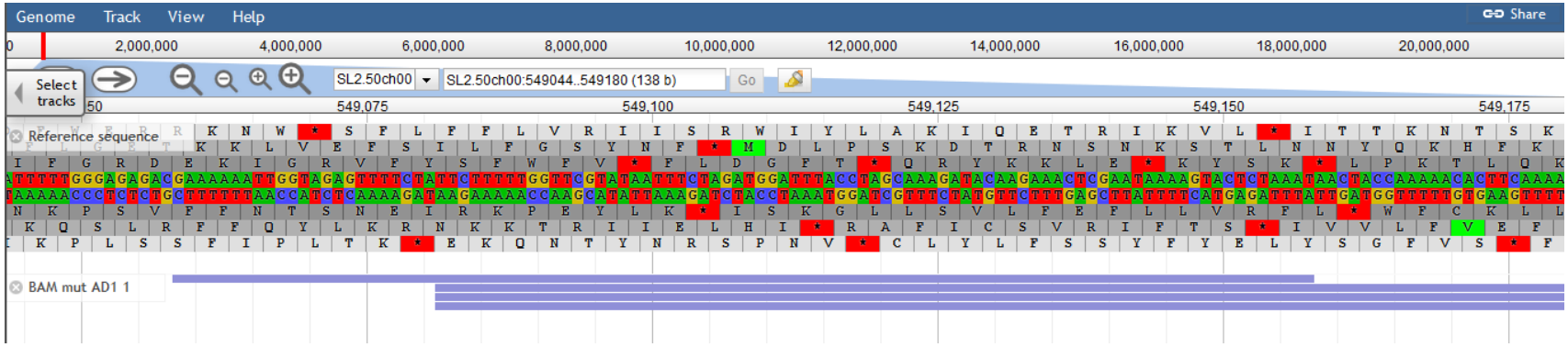

Figura 10. Vista de cuatro *reads* alineados sobre el genoma SL2.50 de *Solanum lycopersicum* usando el navegador genómico de Sol Genomics Network, basado en JBrowse (Buels et al., 2016).

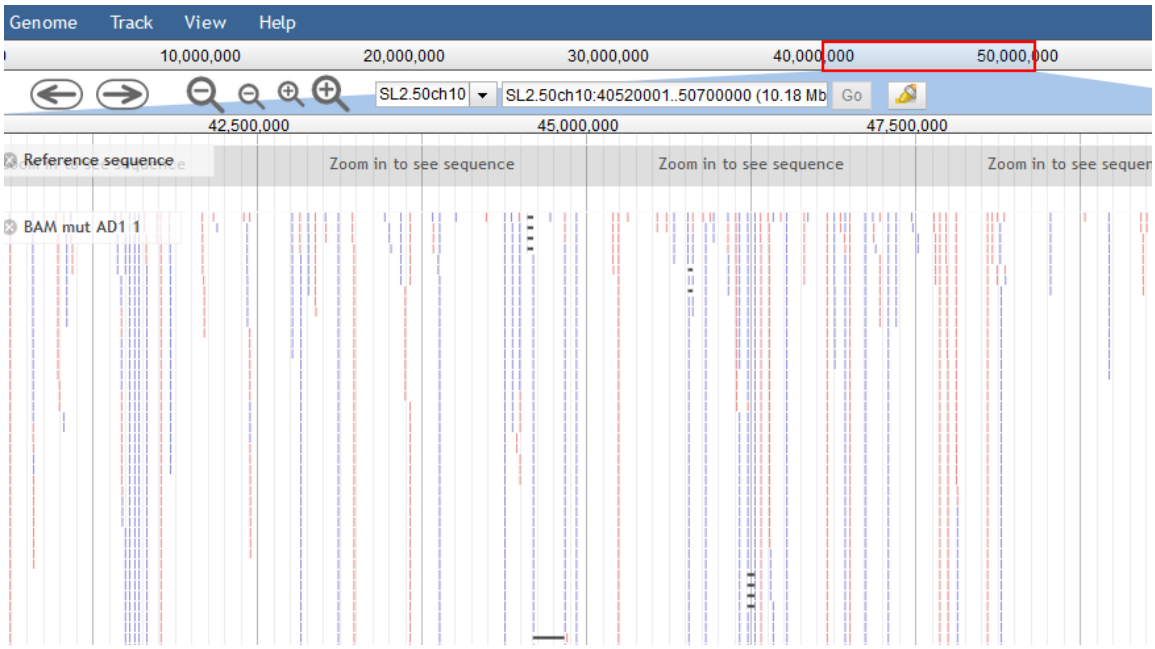

Figura 11. *Reads* de la muestra mut_AD.1 alineados sobre el cromosoma 10, genoma SL2.50, usando el navegador genómico de Sol Genomics Network, basado en JBrowse (Buels et al., 2016).

Figura 12. *Reads* totales y *reads* alineados por *TopHat* una vez. Los puntos indican el porcentaje de *reads* alineados.

Para llevar a cabo esta normalización se ha elaborado un script que en primer lugar cuanta el número de *reads* que hay en cada muestra, luego extrae el nombre y la longitud de los genes de la especie usando un archivo gff3 y tras esto compara el nombre de los genes con los *reads* de las muestras para así tener el número de *reads* de cada gen. Con esta información se puede calcular el número de RPKM con la siguiente expresión:

$$RPKM = \frac{N\acute{u}mero\ de\ reads\ del\ gen}{\frac{Longitud\ del\ gen\ (bases)}{10^3}\ \frac{N\acute{u}mero\ total\ de\ reads}{10^6}}$$

Como ejemplo, en las Tablas 2 y 3 pueden verse los RPKM de varios genes que se han expresado diferencialmente en los diferentes estados del desarrollo floral evaluados en plantas silvestres y mutantes.

4.2 Análisis de expresión diferencial

4.2.1 Control de calidad

Se han encontrado 4953 genes que se han expresado de forma diferencial entre las muestras de estudio. Un primer enfoque a la hora de analizar los resultados es observar los resultados en bruto. En la Tabla 4 puede verse parte de los resultados obtenidos para las muestras de AD con un valor de α=0.05, los resultados aparecen ordenados en función de su *p-value.*

En la Tabla 4 pueden verse varias muestras en las que el valor del *log_2 fold change* tiene un valor absoluto mayor que 2 lo que nos indica que hay cuatro veces más *reads* de ese gen en las plantas silvestres (si es negativo) o en mutantes (si es positivo). El uso de este valor facilita la representación gráfica de los resultados. El estadístico de Wald nos permite rechazar la hipótesis nula "no hay diferencia en la expresión" cuando toma valores que se alejan de 0, pero es más sencillo interpretar directamente el *p-value* ajustado que se calcula a partir de él.

Tabla 2. RPKM de genes expresados diferencialmente en las plantas mutantes.

Gen y exón	mut BF0.1	mut BF0.2	mut BF0.3	mut BF1.1	mut BF1.2	mut BF1.3	mut AD.1	mut AD.2	mut AD.3
Solyc08g081480.2.1	17.3170	173.1696	8.658	125.7916	167.7221	83.8611	439.3532	650.2427	8.7881
Solyc09g075350.2.1	247.7245	19.0557	190.5573	16.6988	581.1191	333.9765	155.2695	522.6144	378.7061
Solyc07g043000.2.1	1062.7097	212.5419	318.8129	861.3189	906.6514	951.9840	173.8806	3477.6118	340.8060
Solyc08g078670.2.1	67.3598	50.3624	62.9531	1935.9105	39.9726	23.5133	581.7413	570.1065	657.3677
Solyc06g006080.2.1	339.4296	170.5676	28.9965	430.8460	444.1711	53.3005	331.0442	304.5607	595.8795
Solyc09g059550.1.1	69.3333	0.0000	53.3333	0.0000	153.6111	30.7222	93.3918	311.3060	127.6355
Solyc06g072700.2.1	122.7851	196.4561	16.8022	299.0993	170.1127	74.7748	556.3526	124.6997	358.1120
Solyc10g086150.1.1	464.9957	318.1550	261.0502	268.3138	294.0034	313.9842	120.1129	404.2826	292.9584

Tabla 3. RPKM de genes expresados diferencialmente en las plantas silvestres.

Gen y exón	wt BF0.1	wt BF0.2	wt BF0.3	wt BF1.1	wt BF1.2	wt BF1.3	wt AD.1	wt AD.2	wt AD.3
Solyc08g081480.2.1	103.1032	66.9795	75.2578	140.2102	184.4871	92.2436	61.5492	74.5069	79.7709
Solyc09g075350.2.1	153.6680	82.9319	243.9174	235.0701	193.8723	161.6410	0.0000	86.1584	99.9437
Solyc07g043000.2.1	654.0623	424.9018	477.4177	308.1648	421.6992	405.4800	547.0259	241.6031	455.8549
Solyc08g078670.2.1	79.9804	48.7880	29.5927	39.1063	49.7277	45.3826	129.0849	203.8183	168.1501
Solyc06g006080.2.1	142.9893	185.7004	239.5535	490.3314	329.8593	410.0953	486.1878	736.6482	914.9171
Solyc09g059550.1.1	67.5415	49.5668	55.5584	34.7010	83.6169	53.6821	38,3034	72,5967	21,4333
Solyc06g072700.2.1	166.8405	123.7957	83.4203	75.3192	46.7498	259.7213	119.7647	207.2851	78.3077
Solyc10g086150.1.1	496.5322	439.2400	401.0452	361.6724	247.1428	286.3240	611.2862	464.5775	579.4993

Tabla 4. Resultados obtenidos para las muestras de AD con un valor de α=0.05, los resultados aparecen ordenados en función de su p-value.

Secuencia	base Mean	log2fold change	lfcSE	stat	pvalue	padj
Solyc08g081480.2.1	737.999	2.247	0.179	12.554	3.78e-36	2.36e-31
Solyc09g075350.2.1	640.516	2.495	0.219	11.372	5.79e-30	1.80e-25
Solyc07g043000.2.1	2437.089	1.560	0.147	10.612	2.61e-26	5.42e-22
Solyc08g078670.2.1	698.071	1.834	0.176	10.393	2.66e-25	4.14e-21
Solyc06g006080.2.1	1470.574	-0.902	0.094	-9.559	1.18e-21	1.23e-17
Solyc09g059550.1.1	220.589	1.954	0.206	9.471	2.76e-21	2.46e-17
Solyc06g072700.2.1	929.239	1.309	0.141	9.253	2.18e-20	1.36e-16
Solyc10g086150.1.1	1312.203	-1.074	0.116	-9.254	2.17e-20	1.36e-16
Solyc07g054860.1.1	177.241	2.837	0.308	9.216	3.09e-20	1.75e-16
Solyc08g083210.2.1	275.035	1.727	0.195	8.876	6.94e-19	3.61e-15
Solyc10g080010.1.1	143.146	2.098	0.240	8.732	2.51e-18	1.12e-14
Solyc01g006540.2.1	557.002	-1.381	0.159	-8.702	3.26e-18	1.27e-14
Solyc09g014350.2.1	1267.345	1.497	0.172	8.709	3.06e-18	1.27e-14
Solyc05g009470.2.1	3913.452	-0.815	0.094	-8.633	5.99e-18	2.19e-14
Solyc09g075210.2.1	3086.369	2.839	0.340	8.355	6.55e-17	2.27e-13
Solyc03g083770.1.1	8560.672	-0.915	0.111	-8.222	2.00e-16	6.57e-13
Solyc08g076250.1.1	301.721	2.381	0.297	8.021	1.05e-15	3.27e-12

Abreviaturas: *base Mean:* media de los conteos normalizados de todas las muestras, *lfcSE*: error estándar del $\log_2$ fold change, *stat*: estadístico de la prueba de Wald, *pvalue:* p-value de la prueba de Wald, *padj*: p-value ajustado mediante Benjamini-Hochberg.

Combinando el valor de *$\log_2$ fold change* y el *p-value* ajustado se puede ver qué diferencia de expresión existe y si hay suficiente información como para aceptar que realmente esa diferencia de expresión no se debe al azar.

Como ejemplo, el gen Solyc08g081480.2.1 (en negrita) se encuentra aproximadamente 4.75 veces más en la variedad mutante que en la variedad silvestre, y como el *p-value* ajustado es menor que el nivel de confianza (0.05) podemos aceptar que esa diferencia en la expresión no se debe al azar. Pero para poder interpretar el significado biológico de estos resultados es necesario disponer del identificador GO de *Gene Ontology*, se redactó un *script* que coteja el nombre de la secuencia con su identificador SGN de *Sol Genomics Network* y obtiene el identificador GO correspondiente. Luego se usa en *Gene Ontology* para obtener la función biológica del producto analizado. Resulta de utilidad comparar la función que le asigna *Gene Ontology* con la función que le asigna *UniProt*. En la Tabla 5 se muestran el identificador GO y las funciones asignadas por *Gene Ontology* y *UniProt*.

Podemos ver que a la primera secuencia le corresponden las funciones “proteína similar a la poligalacturonasa” y “pectina liasa”, las poligalacturonasas son pectinas liasas que se encargan de hidrolizar los enlaces α-1,4-glucosídicos que unen los residuos de ácido galacturónico. Esos enlaces se encuentran en la pared celular de las plantas. Las poligalacturonasas endógenas de las plantas hacen que los frutos se durante las primeras fases del proceso de maduración (Tucker et al., 1980).

En la Figura 13 puede verse una captura del gráfico MA correspondiente a las muestras AD. En el eje de ordenadas se representa el valor del *log_2 fold change,* en el eje de abscisas se representa la media del conteo de *reads* normalizada. Los puntos que aparecen en rojo son aquellos que tienen un *p-value* ajustado menor que 0.05. Las dos líneas celestes marcan los valores de *log_2 fold change* 1 y -1, es decir, a partir de qué punto el gen se expresa el doble en la planta mutante o silvestre, respectivamente. Con esta representación pueden verse rápidamente cuantas muestras cumplen los parámetros establecidos para considerar que existe expresión diferencial de un gen. Esta representación es en realidad un mapa interactivo en R, gracias a la función *Identify* de *DESeq2* es posible identificar cada uno de los puntos de la gráfica, lo que permite ir rápidamente a aquellos puntos de mayor interés. En este caso se ha utilizado *shrinkage* para eliminar el ruido asociado a genes con un número bajo de copias.

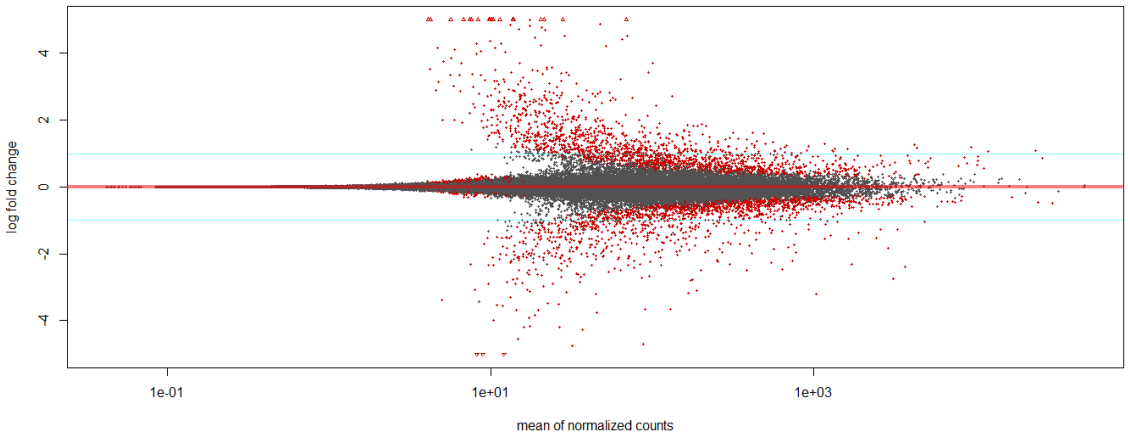

Figura 13. Gráfico MA de las muestras AD.

Cuando se quiere visualizar los resultados o hacer análisis de *clustering* es recomendable trabajar con versiones transformadas de los datos de conteo de *reads.* La transformación más simple consiste en desplazar los logaritmos sumando una constante positiva al conteo para evitar problemas con aquellos genes de los que hay cero *reads*. A veces es necesario usar otro tipo de transformaciones para eliminar la dependencia de la varianza sobre la media. Con estas transformaciones se espera que disminuya la desviación estándar del experimento en global, aplanándose la curva. Aquellos genes cuya varianza esté por encima de esa desviación estándar son los que luego se pueden agrupar en diferentes grupos durante el *clustering.*

Tabla 5. Identificador GO y función génica correspondientes a los genes de las muestras AD que aparecen de la Tabla 4.

Secuencia	Identificador GO	Función	Función de la proteína según Uniprot
Solyc08g081480.2.1	GO:0005975 (carbohydrate metabolic process)	Polygalacturonase-like protein	Pectin lyase fold
Solyc09g075350.2.1	GO:0005618 (cell wall), GO:0030599 (pectinesterase activity)	Pectinesterase	Pectinesterase, catalytic
Solyc07g043000.2.1	*	Unknown Protein	*
Solyc08g078670.2.1	GO:0005975 (carbohydrate metabolic process)	Polygalacturonase-like protein	Pectin lyase fold
Solyc06g006080.2.1	GO:0009228 (thiamine biosynthetic process)	Phosphomethylpyrimidine synthase	Thiamine biosynthesis protein ThiC
Solyc09g059550.1.1	*	Unknown Protein	*
Solyc06g072700.2.1	GO:0030001 (metal ion transport)	Metal ion binding protein	Heavy metal transport/detoxification
Solyc10g086150.1.1	GO:0000166 (nucleotide binding), GO:0003676 (nucleic acid binding)	Single-stranded DNA binding protein	RNA recognition motif, RNP-1
Solyc07g054860.1.1	GO:0006520 (amino acid metabolic process), GO:0019752 (carboxylic acid metabolic process)	Aromatic amino acid decarboxylase	Phenylacetaldehyde synthase
Solyc08g083210.2.1	GO:0003824 (catalytic activity), GO:0005975 (carbohydrate metabolic process)	Endoglucanase 1	Glycoside hydrolase, family 9
Solyc10g080010.1.1	GO:0016757 (glycosyltransferase activity)	Glycosyltransferase	Glycosyltransferase AER61
Solyc01g006540.2.1	GO:0006096 (glycolytic process), GO:0046872 (metal ion binding)	Lipoxygenase	Lipoxygenase, plant
Solyc08g076250.1.1	GO:0020037 (heme binding)	Cytochrome P450	*

En la Figura 14 pueden verse las tres transformaciones que se han comparado, la transformación mediante el desplazamiento de los logaritmos (*shifted logarithm transformation)*, la transformación regularizada de los logaritmos (*regularized log transformation)* y la transformación estabilizadora de la varianza (*variance stabilizing transformation*), respectivamente.

En el panel A puede observarse que la desviación estándar es mayor que las otras dos para prácticamente todos los valores de las medias, lo que la hace menos útil para realizar análisis de *clustering*. El panel B presenta un artefacto matemático en torno a los valores de media de conteos de 50000, además de utilizar un método de cálculo más lento que las otras dos transformaciones. Por último, el panel C presenta los valores más bajos de desviación estándar para todas las medias de conteo del experimento. Debido a todo esto se ha decidido utilizar la transformación estabilizadora de la varianza para el análisis de *clustering*.

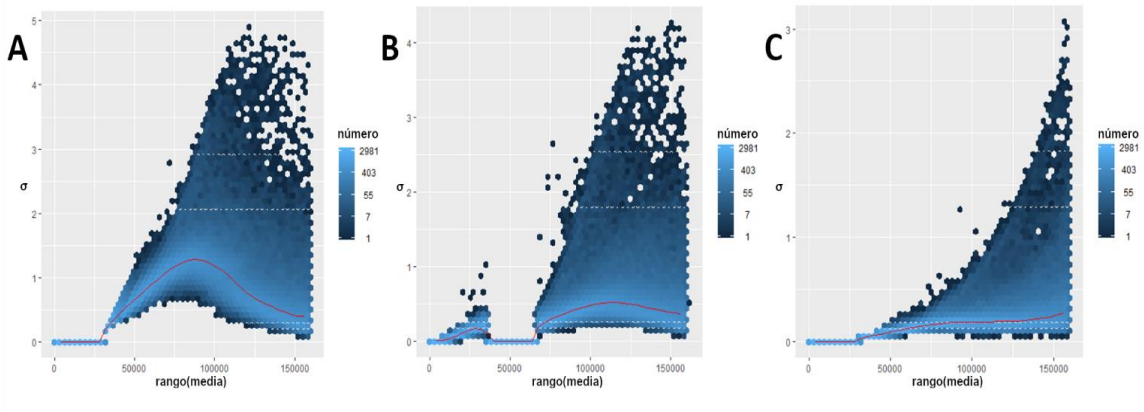

Figura 14. Efecto sobre la varianza de la transformación mediante desplazamiento de los logaritmos (A). Efecto sobre la varianza de la transformación regularizada (B). Efecto sobre la varianza de la transformación estabilizadora de la varianza (C).

Una vez elegida la transformación, se analiza el *heatmap* de la matriz de conteos. Primero se ha generado un *heatmap* que representa la distancia euclídea entre las muestras. En la Figura 15 puede apreciarse que ninguna de las muestras se ha agrupado en un grupo erróneo, en el caso de las muestras correspondientes a BF0 y BF1 puede verse una separación clara entre las muestras de plantas silvestres y mutantes, mientras que para AD las muestras se solapan entre sí.

En la Figura 16 se representa el *heatmap* correspondiente a veinte genes expresados diferencialmente. Para realizar la comparación se ha corregido el tamaño de las distintas muestras usando un factor de tamaño. Para identificar a que muestra pertenece cada columna se utiliza el color de la primera fila. Las filas restantes se corresponden cada una con una muestra, cuanto más rojo es el color de la celda mayor es el número de conteos.

En la Figura 16 puede verse que existe una mayor similitud entre las muestras BF0 y BF1, tanto silvestres como mutantes, que con la muestra AD. Por ejemplo, en el caso del gen Solyc08g74630, las muestras de AD presentan un número menor de conteos que las de BF0 y BF1. Por otra parte, también pueden apreciarse diferencias en la expresión de las muestras AD mutante y silvestre en el caso del gen Solyc03g098780, que se ha expresado más en la variedad silvestre. Finalmente, también es fácil observar la diferencia entre las muestras BF0 y BF1 pues puede verse que el gen Solyc07g007250 se ha expresado en menor medida en las muestras BF0 que en las muestras de BF1 y AD.

En la Figura 17 puede observarse el gráfico de componentes principales (PCA) de las muestras analizadas, en primer lugar, puede observarse que las muestras están separadas en tres grupos en función de las dos componentes principales. La componente con una mayor contribución permite explicar el 76 % de la varianza y separa claramente a las muestras de AD de las muestras de BF0 y BF1, mientras que la segunda componente explica un 13 % de la variación y separa a BF0 de BF1 sin afectar especialmente a AD. Puede verse que hay una muestra mutante y una muestra silvestre de AD muy cercanas entre sí, lo que resulta compatible con el problema que se encontró durante el *clustering.*

.

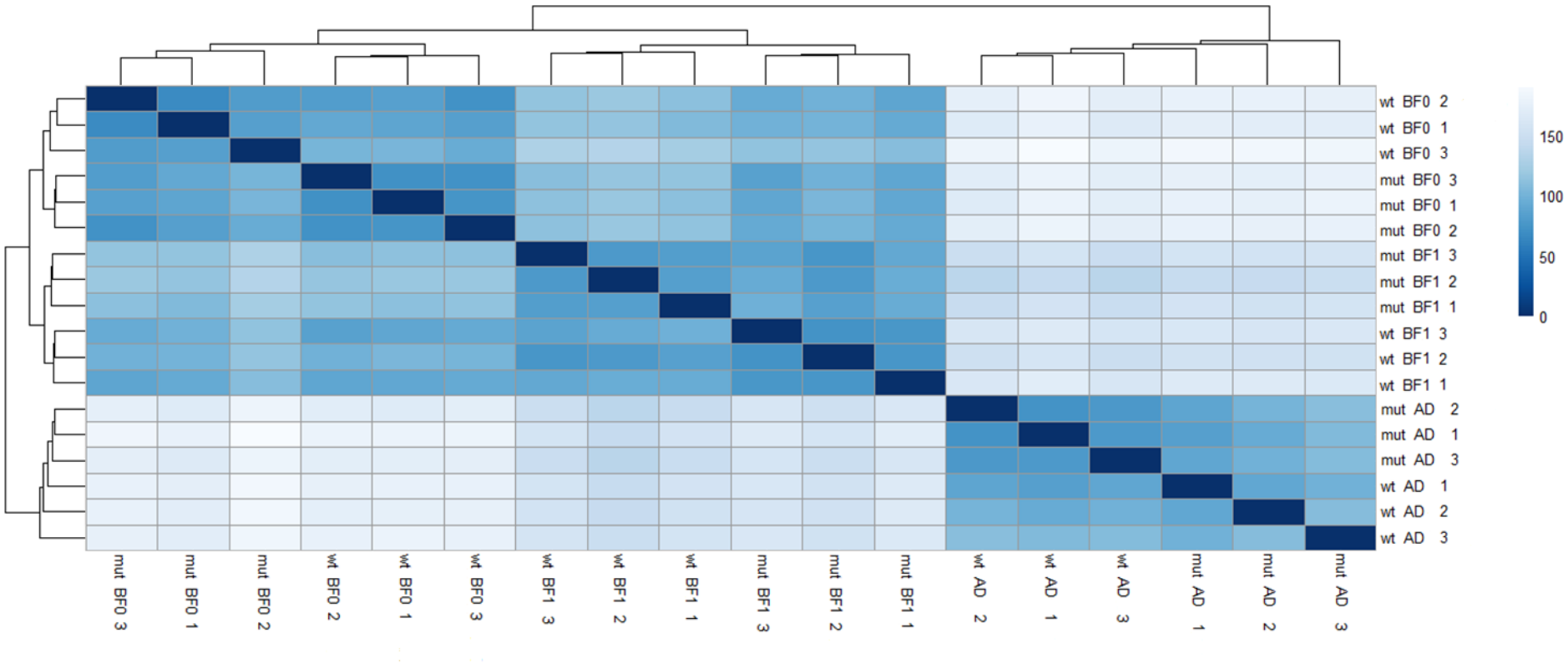

Figura 15. *Heatmap* del análisis *clustering* de las diferentes muestras evaluadas utilizando distancias euclídeas. Botones florales de 3 a 6 mm (BF0) y de 6 a 12 mm (BF1) de longitud, así como flores en antesis (AD) de plantas silvestres (wt) y mutantes (mut).

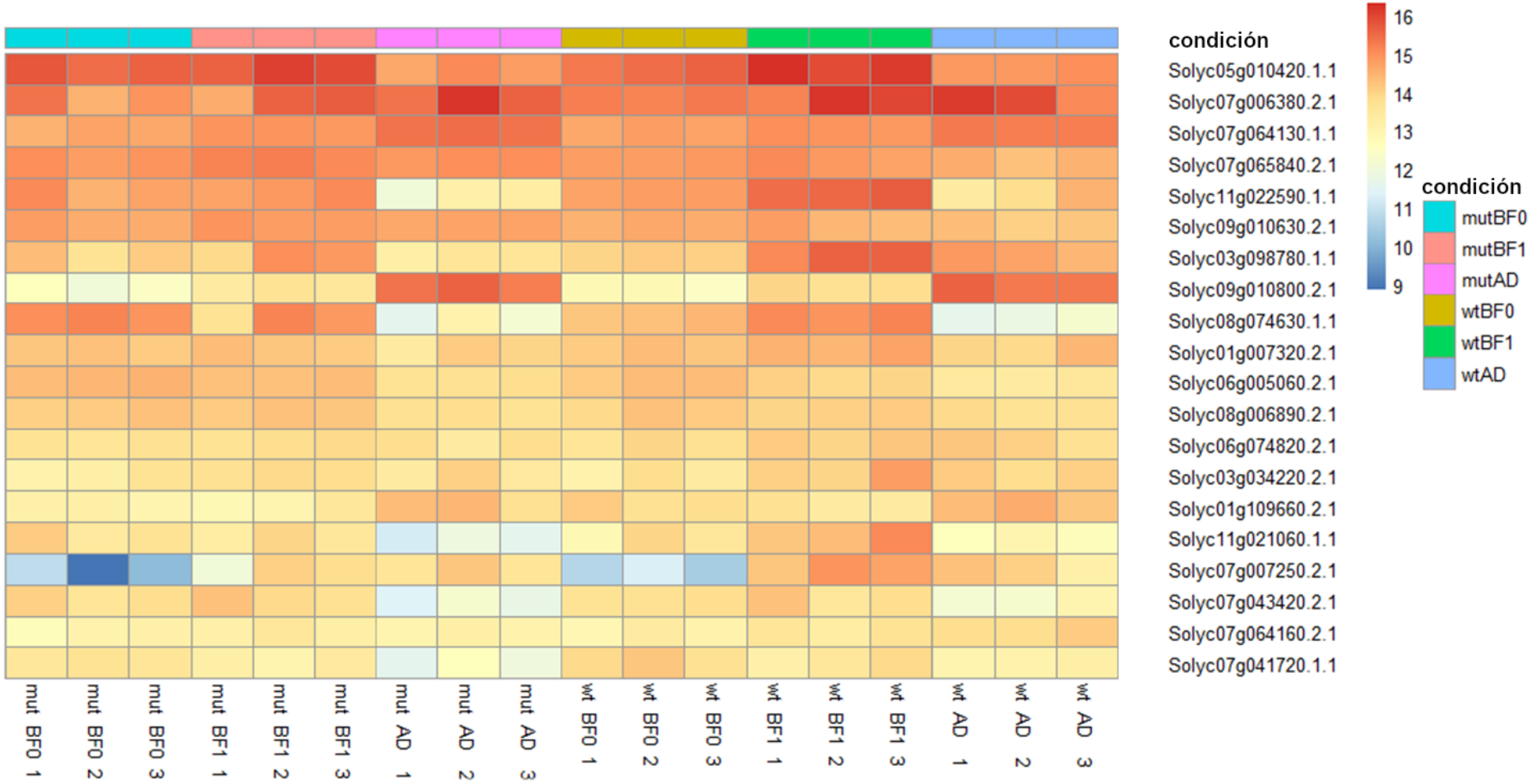

Figura 16. *Heatmap* de los 20 genes (con Log2FC > | 1 |, p-value ajustado < 0.05) en las diferentes muestras evaluadas. La escala de color representa los valores de expresión normalizados usando la transformación estabilizadora de la varianza. Botones florales de 3 a 6 mm (BF0) y de 6 a 12 mm (BF1) de longitud, así como flores en antesis (AD) de plantas silvestres (wt) y mutantes (mut).

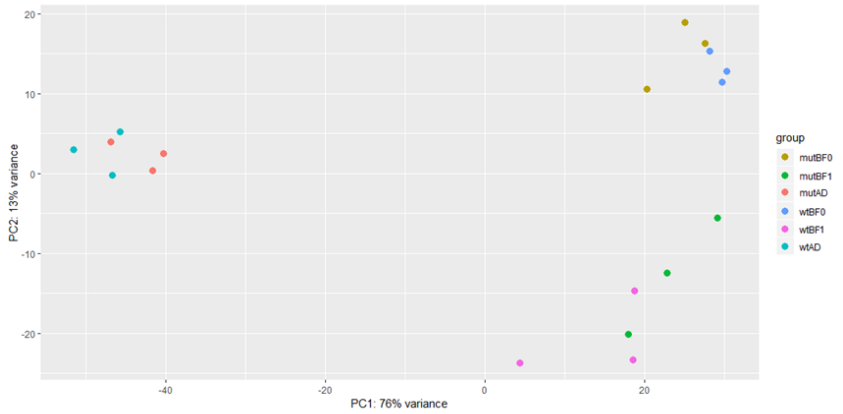

Figura 17. Gráfico PCA de las muestras de estudio. Botones florales de 3 a 6 mm (BF0) y de 6 a 12 mm (BF1) de longitud, así como flores en antesis (AD) de plantas silvestres (wt) y mutantes (mut).

Para contrastar el PCA anterior, se ha generado un PCA generalizado (Figura 18) utilizando el paquete "*glmpca*", aunque el PCA sea distinto ya que se genera directamente con los datos iniciales, puede verse que separa correctamente los tres grupos, dejando de nuevo a AD más separado del resto. De nuevo puede verse que existe una mayor variabilidad interna en los grupos de mut BF0 y mut BF1.

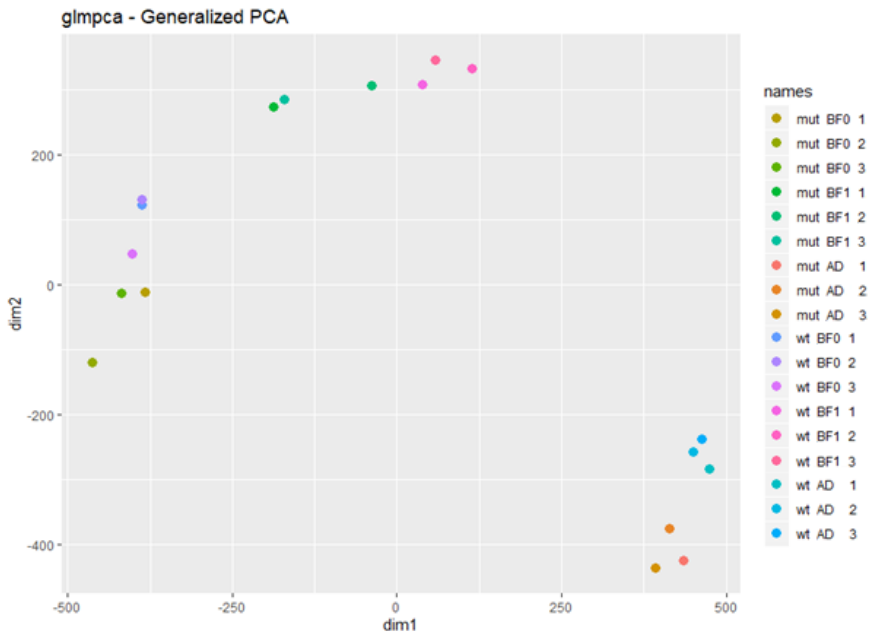

Figura 18. PCA generalizado de las muestras evaluadas. Botones florales de 3 a 6 mm (BF0) y de 6 a 12 mm (BF1) de longitud, así como flores en antesis (AD) de plantas silvestres (wt) y mutantes (mut).

Por otra parte, es también necesario analizar los gráficos de dispersión para confirmar que no hubiese anomalías en el análisis debidas a falsos positivos causados por la falta de datos de cada gen individual. En la Figura 19 puede verse el gráfico de dispersión, el cual representa los estimadores de la dispersión de cada gen frente al nivel de expresión de dicho gen (es decir, su media de los conteos normalizados).

Los puntos de color negro representan los estimadores de máxima similitud obtenidos usando la información disponible de cada gen. A estos puntos se ajusta la curva roja, que muestra la tendencia de la dependencia entre la dispersión y el promedio de conteos normalizados. Ese valor es utilizado como una nueva media a la hora de realizar una segunda estimación de la dispersión, el resultado de esta son los puntos azules que representan los resultados de la estimación máxima a posteriori de los estimadores de dispersión.

Los puntos negros que aparecen abajo del gráfico se corresponden a aquellos genes cuya varianza observada es menor que la varianza esperada al usar un modelo de Poisson, por lo tanto, su estimador de máxima similitud es cero, pero aparecen en el gráfico con el valor de 10^{-8} como sustituto. Los puntos negros que aparecen redondeados en azul son aquellos valores cuya dispersión es mucho mayor que la esperada con el ajuste (más de dos veces la desviación estándar sobre la curva), por lo que se utilizan en lugar de sus correspondientes valores ajustados.

La razón por la que se lleva a cabo este *shrinkage* es para reducir la interferencia entre los estimadores individuales de cada gen con respecto al consenso representado por su media.

En la Figura 19 puede apreciarse que tras la contracción se ha reducido significativamente la variación de la dispersión sobre la media, aunque también puede verse que hay una gran cantidad de genes que aparecen como esos puntos con una expresión mucho mayor que el resto.

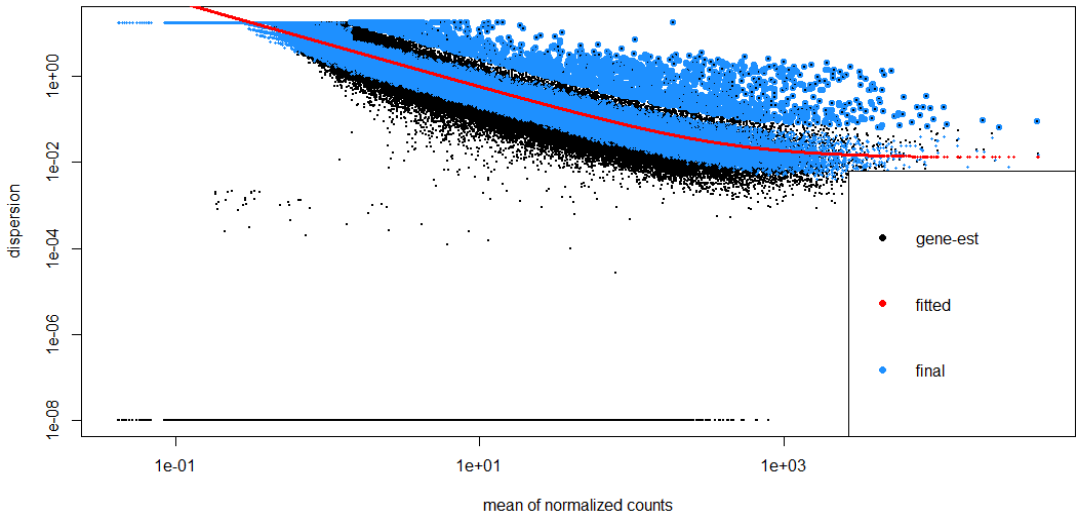

Figura 19. Gráfico de los estimadores dispersión usados.

Tras interpretar estos gráficos se ha considerado que el análisis se ha realizado correctamente, ninguna de las métricas parece indicar que exista un error crítico en cuanto al tratamiento de las muestras usadas, el *clustering* es el esperado, las distancias entre las muestras de estudio son coherentes, la transformación utilizada para interpretar como se agrupan las muestras es la que mejor valores de ajuste presenta, obteniéndose una gran cantidad de valores que se encuentran dentro del rango de interés. En base a esto, se procede al análisis detallado de los resultados.

4.2.2 Expresión diferencial entre los tres estadios

Se han encontrado 4958 genes que se han expresado diferencialmente entre las plantas silvestres y mutantes. En la Figura 20 puede verse el diagrama de Venn de aquellos genes expresados diferencialmente en cada etapa, con un *log_2 fold change* mayor que 1 en el caso de las plantas mutantes o menor que -1 en el caso de las plantas silvestres. Se han encontrado 2423 genes sobreexpresados y 2321 genes reprimidos en las plantas mutantes.

Entre los genes diferencialmente sobreexpresados en las plantas mutantes, se han encontrado 5 genes que se expresan de forma diferencial en las tres etapas de desarrollo floral evaluadas, 27 que se expresan en BF0 y BF1, 11 que se expresan en BF1 y AD y 137 que se expresan en BF0 y AD. La cantidad de genes expresados diferencialmente durante BF0 y AD es similar, mientras que durante BF1 es aproximadamente la mitad.

Respecto a los genes diferencialmente reprimidos en las plantas mutantes, se han encontrado 22 genes que se expresan de forma diferencial en las tres etapas, 41 que se expresan en BF0 y BF1, 65 que se expresan en BF1 y AD y 80 que se expresan en BF0 y AD.

Con respecto a los genes sobreexpresados diferencialmente en las plantas mutantes, de los cinco genes de la triple intersección, dos son desconocidos, dos codifican proteínas de tipo OVATE y uno es un gen que produce una proteína implicada en el transporte polar de auxinas de eflujo.

Los genes de la familia OVATE (*fruit development*, GO:0010154) está relacionado con la forma del fruto del tomate. Estos genes generan proteínas que funcionan como represores de la transcripción, regulando varios aspectos del desarrollo de la planta. Sus mutaciones producen un alargamiento del fruto. En el caso del tomate se debe a una mutación simple, recesiva, que resulta en un codón de stop prematuro que conlleva la eliminación del dominio C-terminal de una proteína OVATE, lo que le hace perder su función. Este gen, se expresa principalmente en órganos reproductivos, sus transcritos se pueden detectar en las flores diez días antes de la antesis y hasta 8 días después de la antesis (Liu et al., 2002).

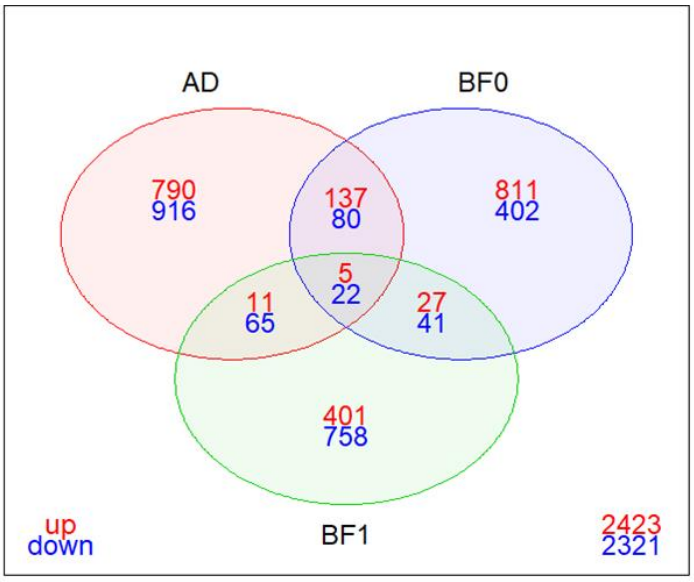

Figura 20. Diagrama de Venn de los genes expresados diferencialmente en cada etapa. En rojo aparecen los genes sobreexpresados en plantas mutantes y en azul aquellos que están reprimidos.

Los transportadores de auxina de eflujo (*transporter activity*, GO:0005215; *flower development,* GO:0009908; *inflorescence development*, GO:0010229) o *PIN1* (Huang et al., 2010) se encuentran en el extremo basal de las células que forman parte del sistema de transporte de auxinas. Son tejido-específicos. Influyen en muchos procesos como la determinación de la polaridad embrionaria, el crecimiento y las respuestas fototrópica y gravitrópica (Benková et al. 2003).

Se hizo un BLAST (*Basic Local Alignment Search Tool)* con las secuencias de los genes desconocidos para ver si presentaban homología con genes de otras especies.

El primero coincide con la secuencia del gen *FLC*, *Flowering Locus C* de *Arabidopsis thaliana (floral meristem determinacy*, GO:0010582; *RNA polymerase II transcription regulatory region sequence-specific DNA binding,* GO:0000977). Este gen codifica una proteína MADS-box, un factor de transcripción que funciona como represor de la transición floral (Sheldon et al., 2000).

El segundo es similar al gen *GRP5*, *Glycine-Rich Protein 5* (*positive regulation of organ growth*, GO:0046622; *response to abscisic acid*, GO:0009737; *structural constituent of cell wall*, GO:0005199) que codifica una proteína que se expresa de forma abundante en las semillas. Parece estar asociado con la respuesta al ácido abscísico y al ácido salicílico (Park et al., 2008).

En cuanto a los genes reprimidos diferencialmente en las plantas mutantes, se han detectado 22 genes que se expresan en los tres estadios de desarrollo. Entre ellos destacan:

El gen *AAT2* que sintetiza hidroxicinamoil transferasa (*fruit ripening, climacteric*, GO:0009836; *response to ethylene*, GO:0009723), esta enzima está implicada en la respuesta al etileno y en la maduración del fruto (Ban et al., 2010).

Se han encontrado factores de transcripción de tipo bLHL (*DNA-binding transcription factor activity*, GO:0003700) y uno de tipo MADS-box perteneciente a la familia APETALA3/PISTILLATA, TPI (*DNA-binding transcription factor activity, RNA polymerase II-specific*, GO:0000981; *RNA polymerase II cis-regulatory region sequence-specific DNA binding*, GO:0000978; *specification of floral organ identity*, GO:0010093).

Se han encontrado proteínas estructurales como proteínas de la pared celular y proteínas de transporte como las acuaporinas que forman canales para el agua o los transportadores no específicos de lípidos.

Hay proteínas que son fundamentales para las funciones celulares como proteínas ribosomales, los factores de despolimerización de actinas, que tienen la capacidad de cortar los filamentos de actina o las chaperonas que son fundamentales para llevar a cabo el pliegue de las proteínas.

Con respecto a los genes diferencialmente sobreexpresados en las plantas mutantes en las intersecciones se han encontrado los factores de transcripción CRABS CLAW, *CRC* (*DNA-binding transcription factor activity*, GO:0003700; *carpel development*, GO:0048440; *floral meristem determinacy, GO:0010582*) en la intersección BF0-BF1 y BF0-AD. En la intersección de BF0-BF1 se ha encontrado el factor de transcripción tipo AP2/ERF, *APETALA2/Ethylene Responsive Factor* (*DNA-binding transcription factor activity*, GO:0003700; *flower development*, GO:0009908; *maintenance of shoot apical meristem identity,* GO:0010492*)*, un factor de transcripción homólogo de *APETALA2-like protein 5* de *Oryza sativa (DNA-binding transcription factor activity*, GO:0003700; *flower development*, GO:0009908; *regulation of floral organ abscission*, GO:0060860) y un factor de transcripción que contiene un dominio GRAS (*DNA-binding transcription factor activity*, GO:0003700; *response to gibberellin*, GO:0009739) . En BF0-BF1 se han expresado factores de transcripción con un dominio C2H2L, un factor de transcripción TDR6, un factor de transcripción *WOX4* (*WUSCHEL-related homeobox 4)* que se relaciona con el desarrollo de los tejidos vasculares de los meristemos (*DNA-binding transcription factor activity*, GO:0003700; *procambium histogenesis,* GO:0010067*)* y un factor de transcripción que contiene un dominio *AT-hook (sequence-specific DNA binding*, GO:0043565; *minor groove of adenine-thymine-rich DNA binding*, GO:0003680).

También se han encontrado reguladores de la transcripción como los factores de respuesta a la auxina ARF5 en BF0 y AD(*transcription cis-regulatory region binding*, GO:0000976; *response to auxin,* GO:0009733; *flower development*, GO:0009908) y ARF8 en BF0 y BF1 (*transcription cis-regulatory region binding*, GO:0000976; *response to auxin*, GO:0009733; *flower development*, GO:0009908; *meristem development*, GO:0048507), *PHD* y *Zinc Finger family proteins* en BF0 y AD (*transcription cis-regulatory region binding*, GO:0000976), la proteína de unión al promotor de *SQUAMOSA* en BF0 y AD (*DNA-binding transcription factor activity*, GO:0003700; *regulation of transcription DNA-templated*, GO:0006355) y el gen *NSD3* en BF1 y AD (*transcription regulator activator activity*, GO:0140537; *positive regulation of transcription, DNA-templated*, GO:0045893; *histone methylation*, GO:0016571). Se han encontrado activadores de transcripción como *SHORT INTERNODE RELATED SEQUENCE* en BF0 y BF1(*auxin*

biosynthetic process, GO:0009851; *auxin-activated signaling pathway*, GO:0009734; *multicellular organism development,* GO:0007275) y el transactivador de respuesta a calcio CREST en BF0 y AD *(DNA-binding transcription factor activity, GO:0003700; transcription coactivator activity GO:0003713).*

Entre estos genes se han encontrado genes implicados en el transporte y la respuesta a la auxina, en la síntesis y respuesta al etileno y en la respuesta al ácido abscísico. En las muestras BF0 y AD se ha expresado el homólogo de *FRIGIDA*, *FRL1* (*cell differentiation*, GO:0030154; *flower development*, GO:0009908), que es requerido para la activación y el mantenimiento de la concentración de *FLC* (Michaels et al., 2004), el cual se ha expresado en las tres muestras mutantes.

En cuanto a los genes diferencialmente reprimidos en las plantas mutantes, la gran mayoría han sido genes relacionados con la fotosíntesis y la respiración. En los tres estadios se reprime la expresión del factor de transcripción MADS-box AP3/PI (*DNA-binding transcription factor activity, RNA polymerase II-specific*, GO:0000981; *RNA polymerase II cis-regulatory region sequence-specific DNA binding*, GO:0000978; *specification of floral organ identity*, GO:0010093), en las muestras BF0 y BF1 se encuentra reprimido otro factor de transcripción MADS-box AP3 (*RNA polymerase II transcription regulatory region sequence-specific DNA binding*, GO:0000977; *DNA-binding transcription factor activity,* GO:0003700; *specification of floral organ identity*, GO:0010093; *flower development*, GO:0009908) y el factor de transcripción SLL1 (*transcription cis-regulatory region binding,* GO:0000976; *inflorescence development,* GO:0010229). En las muestras BF0 y AD se reprime la expresión de los factores de transcripción CYCLOIDEA, CYC (*DNA-binding transcription factor activity, GO:0003700; flower development, GO:0009908)* y bLHL79 (*DNA-binding transcription factor activity, GO:0003700)*.

Entre los genes reprimidos se han encontrado también genes relacionados con la síntesis de giberelinas (*Ent-Kaurene synthase*) y regulados por ácido giberélico. También se ha expresado GH3, y un producto de GH3, relacionados con los niveles de ácidos salicílico y de auxinas respectivamente. Se han expresado enzimas relacionadas con la síntesis de ácido jasmónico (lipoxigenasas) y de etileno (1-aminociclopropano-1-carboxilato oxidasa).

4.2.3 Expresión diferencial en cada estadio

Los genes diferencialmente expresados dentro de cada estadio se han ordenado de forma ascendente según su *p-value* ajustado, eliminando aquellos genes que tengan un log_2 *fold change* entre -1 y 1.

En las plantas mutantes, en BF0 se encontraron diferencialmente sobreexpresados genes como el gen Solyc04g080490.2.1, que es homólogo de los genes ZHD de *Arabidopsis thaliana* que producen *Zinc-finger homeodomain protein* (*regulation of transcription, DNA-templated*, GO:0006355; *long-day photoperiodism, flowering,* GO:0048574). Las enzimas que contienen estos homeodominios tienen un papel

esencial en el desarrollo y se expresan de forma predominante en el tejido floral (Tan y Irish, 2006).

El gen Solyc07g056650.2.1 que es homólogo de *ICR1*, *INTERACTOR OF CONSTITUTIVELY ACTIVE ROP1* (*pollen tube tip*, GO:0090404; *regulation of auxin polar transport*, GO:2000012). *ICR1* interactúa con ROP1, facilitando la estabilización de las proteínas de la familia RHO en la membrana plasmática de los tubos polínicos (Li, S. et al., 2008). Estas proteínas RHO regulan direccionalmente el transporte de auxinas, son requeridos para estabilizar la concentración de auxinas y crear gradientes de auxina durante la embriogénesis, organogénesis y en la actividad meristemática (Hazak et al., 2010).

Los genes Solyc11g066130.1.1, Solyc02g077390.1.1, Solyc04g054880.2.1 y Solyc06g082930.2.1 también están sobreexpresados en BF0 mutantes. El gen Solyc11g066130.1.1 codifica una proteína de tipo taumatina. En el tomate esta proteína se llama NP24, que parece estar relacionada tanto en la respuesta de defensa contra hongos como con variaciones en la salinidad. Se cree que puede estar implicada en el desarrollo y maduración del fruto (*fruit ripening*, GO:0009835*; fruit developmen*t, GO:0010154) (Pressey, 1997).

El gen Solyc02g077390.1.1 produce una proteína homeobox-3 (*DNA-binding transcription factor activity*, GO:0003700; *plant organ development*, GO:0099402; *flower development*, GO:0009908; *stipule development*, GO:0010865), un factor de transcripción que en el tomate parece estar implicado en el desarrollo de los carpelos y en la organización estructural de los meristemos (Li, X. et al., 2018). Sus homólogos en otras especies interactúan con WUSCHEL (Shimizu et al., 2009).

El gen Solyc04g054880.2.1 produce un factor de transcripción BZIP, probablemente homólogo a HY5 (*DNA-binding transcription factor activity*, GO:0003700, *red or far-red light signaling pathway*, GO:0010017; *response to far red light*, GO:0010218) un tipo de factor de transcripción que desencadena la fotomorfogénesis y regula de forma positiva la pigmentación del fruto y su calidad nutricional (Liu et al., 2004).

El gen Solyc06g082930.2.1 produce una proteína FRIGIDA, FRI (*cell differentiation*, GO:0030154; *flower development*, GO:0009908), que es requerida para la activación y el mantenimiento de la concentración de *FLC* (Choi et al., 2011). Además, otros de los genes que se encuentran sobreexpresados BF0 sintetizan factores de transcripción homeobox y factores reguladores del crecimiento.

Entre los genes diferencialmente reprimidos en BF0 destaca el gen Solyc10g080840.1.1 que sintetiza un citocromo de la familia P450. Los citocromos están implicados en una gran cantidad de procesos. En el caso del tomate, en los frutos hay elevadas concentraciones de enzimas P450 de la subfamilia CYP78A, el aumento de su cantidad se relaciona con un aumento del peso del fruto debido al crecimiento del pericarpio y del septo (Chakrabarti et al., 2013). Además, se encontró el gen

Solyc02g065230.2.1 es otro gen que sintetiza citocromo P450, que se expresa en gran cantidad en las flores de tomate.

En las Figuras 21 y 22 pueden verse un resumen de las funciones génicas más representadas, de acuerdo con la clasificación de Gene Ontology entre los genes diferencialmente expresados en BF0.

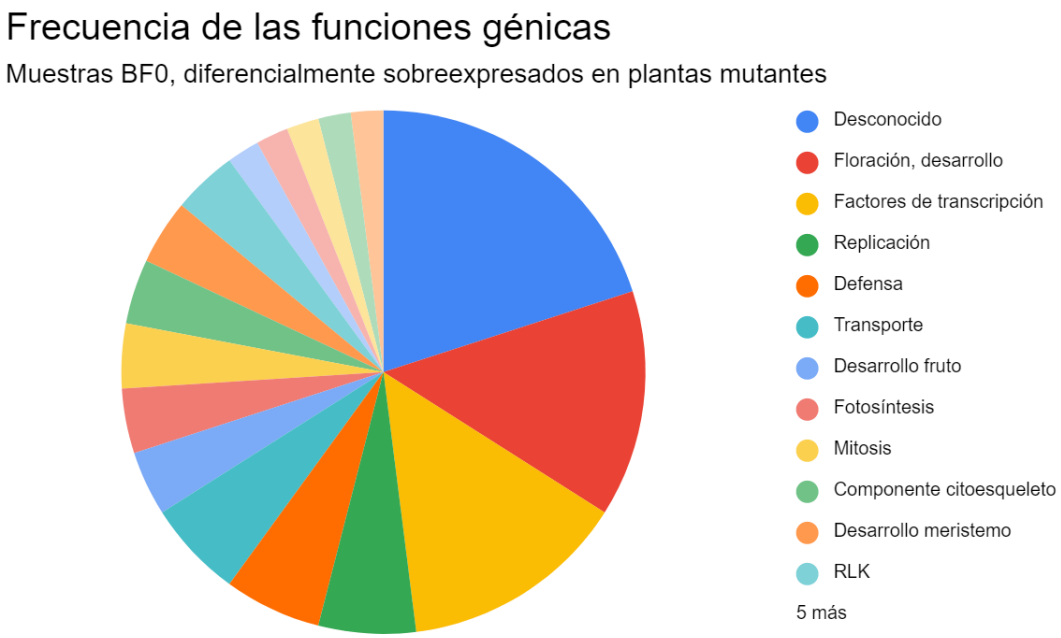

Figura 21. Gráfico circular que muestra las funciones génicas más representadas entre los genes diferencialmente sobreexpresados en las muestras BF0 mutantes.

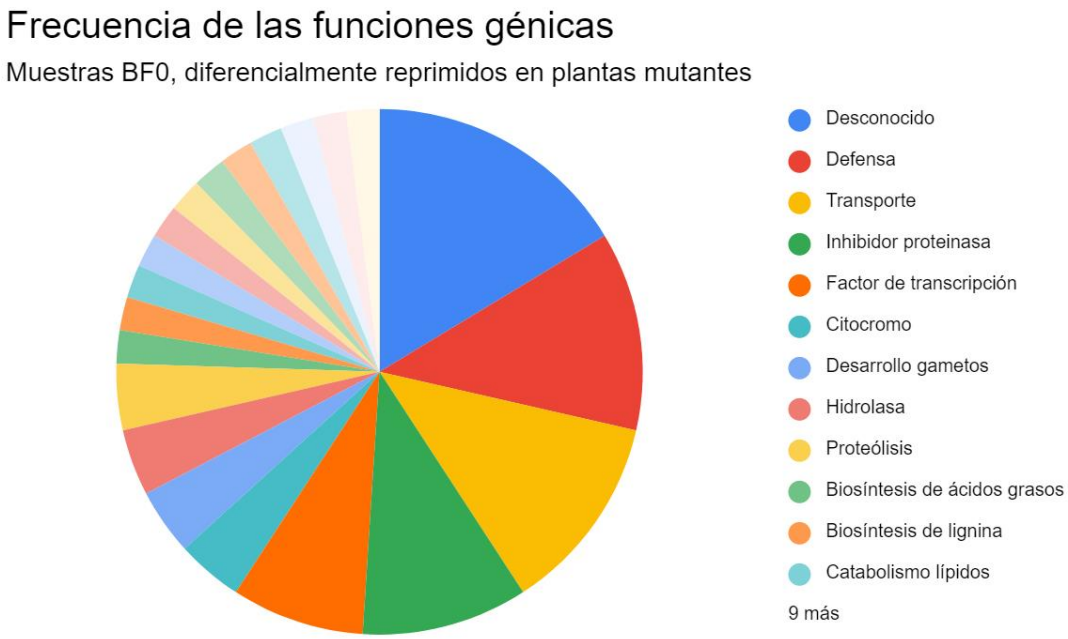

Figura 22. Gráfico circular que muestra las funciones génicas más representadas entre los genes diferencialmente reprimidos en las muestras BF0 mutantes.

Entre los genes diferencialmente sobreexpresados en BF1 mutantes se encontró el gen Solyc08g080470.2.1, que sintetiza una proteína que contiene un motivo IQ (*calcium ion binding*, GO:0005509; *regulation of flower development*, GO:0009909), un motivo extremadamente básico formado por 23 aminoácidos. Este motivo sirve como sitio de unión para diferentes proteínas mano EF, incluyendo algunas esenciales y reguladoras como cadenas ligeras de miosina, calmodulina (CaM) y proteínas tipo CaM. Muchas proteínas con motivo IQ son sitios de fosforilación de la proteína quinasa C. Esta en concreto parece ser un activador de la transcripción que se une a la secuencia consenso 5'-[ACG]CGCG[GTC]-3'. La regulación de la calmodulina juega un papel importante en el desarrollo del fruto del tomate (Yang et al., 2014).

Respecto a los genes diferencialmente reprimidos en BF1 mutantes se encontraron los genes Solyc08g082770.2.1 y Solyc02g093580.2.1. El gen Solyc08g082770.2.1 que sintetiza un transportador bidireccional de azúcar tipo MtN3 (*sugar transmembrane transporter activity*, GO:0051119; *embryo development ending in seed dormancy*, GO:0009793; *seed maturation*, GO:0010431), esta familia de transportadores son proteínas de membrana que presentan el dominio MtN3/Saliva. Estas proteínas participan en múltiples procesos biológicos como el desarrollo de nódulos, la senescencia, la respuesta a estrés abiótico o la interacción huésped-patógeno (Yuan y Wang, 2013). El gen Solyc02g093580.2.1 sintetiza una pectato liasa (*pectate lyase activity,* GO:0030570*; pollen tuve,* GO:0090406*)*, un tipo de enzima que participa en el desarrollo del tejido y eventos relacionados con la polinización, así como en el crecimiento de los tubos polínicos. Elimina alfa-D-galacturona para dejar a los oligosacáridos con grupos 4-deoxi-alfa-D-galact-4-enuronosil en sus extremos no reductores.

En las Figuras 23 y 24 puede verse un resumen las funciones génicas más representadas de acuerdo con la clasificación de Gene Ontology, entre los genes diferencialmente expresados en BF1.

En las muestras AD, entre los genes diferencialmente expresados destacan los genes Solyc08g081480.2.1, Solyc08g078670.2.1, Solyc09g075350.2.1, Solyc07g054860.1.1, Solyc09g014350.2.1, Solyc09g075210.2.1, Solyc08g076250.1.1, Solyc11g066130.1.1, Solyc01g010600.2.1 y Solyc09g091510.2.1. Los genes Solyc08g081480.2.1 y Solyc08g078670.2.1 sintetizan una pectina liasa similar a las poligalacturonasas (*polygalacturonase activity*, GO:0004650; *anther dehiscence,* GO:0009901; *fruit ripening*, GO:0009835; *fruit dehiscence*, GO:0010047), un tipo de pectinas que hidrolizan los enlaces entre residuos araquidónicos que forman parte de la pared celular de las plantas, lo que permite reblandecer los frutos al principio del proceso de maduración (Tucker et al., 1980). En el caso concreto del tomate se sabe que esta enzima trabaja junto con la pectín esterasa durante el proceso de maduración, su inactivación mediante temperatura y presión puede modificar la reología del jugo del tomate (Watson et al., 1994).

Figura 23. Gráfico circular que muestra las funciones génicas más representadas entre los genes diferencialmente sobreexpresados en las muestras BF1 mutantes.

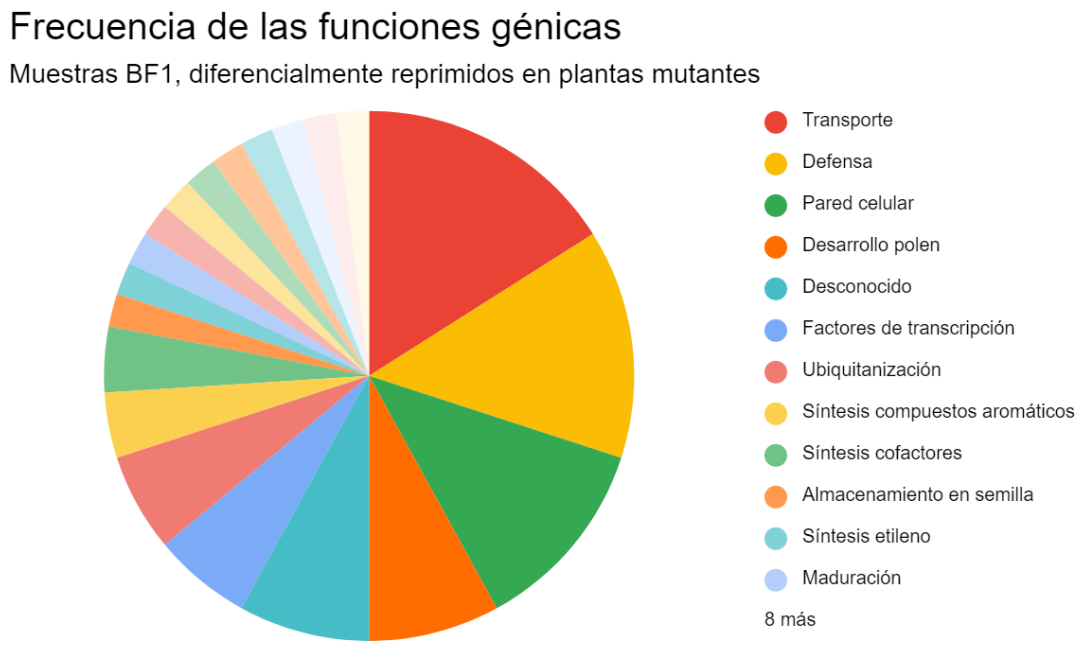

Figura 24. Gráfico circular que muestra las funciones génicas más representadas entre los genes diferencialmente reprimidos en las muestras BF1 mutantes.

El gen Solyc09g075350.2.1 sintetiza una pectín esterasa (*pectinesterase activity*, GO:0030599; *cell wall modification*, GO:0042545; *pollen development*, GO:0009555). Este tipo de hidrolasas se encarga de modificar las paredes celulares mediante la eliminación de ésteres de metilo de la pectina de la pared celular. En *Arabidosis thaliana* destaca su rol en el crecimiento de los tubos polínicos de la flor femenina puesto que al modificar las paredes celulares mejora la comunicación de estos con el resto de los tejidos de la flor (Jiang et al., 2005).

El gen Solyc07g054860.1.1 sintetiza una fenilacetaldehído sintasa (*phenylacetaldehyde synthase activity*, GO:1990055; *L-phenylalanine catabolic process,* GO:0006559) esta enzima cataliza la decarboxilación de L-fenilalanina a 2-feniletilamina, que luego se oxida para formar 2-fenilacetaldehído, un componente del aroma de las flores. El 2-fenilacetaldehído es también el precursor del 2-feniletanol, otro constituyente del aroma floral (Gutensohn et al., 2011).

El gen Solyc09g014350.2.1 codifica una glicerol-3-fosfato aciltransferasa (*fatty acid biosynthetic process*, GO:0006633; *glycerol-3-phosphate O-acyltransferase activity*, GO:0004366; *seed oilbody biogenesis*, GO:0010344; *pollen maturation*, GO:0010152). Las glicerol-3-fosfato aciltransferasas poseen actividad aciltransferasa con alta especificidad por acil coenzima A, lo que desencadena la biosíntesis de lípidos y forman parte de la ruta de Kennedy de síntesis de glicerolípidos. Cataliza la acumulación de triacilglicerol implicada en la membrana lipídica y en la síntesis de aceites, especialmente en las semillas (Shockey et al., 2016). Contribuye a la biosíntesis de lípidos polares y triacilglicerol en las hojas en desarrollo, además de en la producción de lípidos en granos de polen. No contribuye en la síntesis de lípidos superficiales como ceras o cutina (Singer et al., 2016).

El gen Solyc09g075210.2.1 codifica una proteína abundante en las últimas etapas de la embriogénesis (Late embryogenesis abundant protein, LEA). Las LEA (*leaf senescence*, GO:0010150; *regulation of leaf senescence*, GO:1900055; *regulation of photoperiodism, flowering*, GO:2000028) son proteínas cuya función aún no se conoce con exactitud, están implicadas en múltiples procesos. Parecen minimizar el efecto del estrés externo, fomentando el desarrollo radicular y evitando el envejecimiento prematuro (Mowla et al., 2006).

El gen Solyc08g076250.1.1 codifica un citocromo de la familia P450. El aumento de su cantidad se relaciona con un aumento del peso del fruto debido al crecimiento del pericarpio y del septo (Chakrabarti et al., 2013).

El gen Solyc11g066130.1.1, que también se expresa diferencialmente BF0, sintetiza una proteína similar a la taumatina. Se cree que puede estar implicada en el desarrollo y maduración del fruto (*fruit ripening*, GO:0009835*; fruit developmen*t, GO:0010154) (Pressey, 1997).

El gen Solyc01g010600.2.1 parece ser una secuencia homeobox que sintetiza una proteína capaz funcionar como factor de transcripción. Debido a que este tipo de secuencias están conservadas, presenta varias proteínas posibles. Es homólogo de los

genes *HAT* de *Arabidopsis thaliana* (*DNA-binding transcription factor activity, RNA polymerase II-specific,* GO:0000981*; developmental process involved in reproduction,* GO:0003006*; floral meristem determinacy,* GO:0010582*; fruit septum development,* GO:0080127; *gynoecium development,* GO:0048467*).* Otras dos que aparecen documentadas son las proteínas correspondientes a los genes *WUS* y *LET6*. La proteína WUS (*transcription cis-regulatory region binding*, GO:0000976; *axillary shoot meristem initiation*, GO:0090506; *anther development*, GO:0048653) es un factor de transcripción que parece tener un papel fundamental durante procesos del desarrollo como la embriogénesis, la floración y el desarrollo de los meristemos al regular la expresión de genes específicos (Xu et al., 2015).

El gen Solyc09g091510.2.1 sintetiza 4,2',4',6'-tetrahidroxichalcona o naringenina-chalcona (*flavonoid biosynthetic process*, GO:0009813; *naringenin-chalcone synthase activity*, GO:0016210), en condiciones específicas puede producir naringenina, la cual se acumula en la piel del tomate. Estas enzimas participan en la síntesis de flavonoides.

En las Figuras 25 y 26 puede verse un resumen de las frecuencias de las funciones génicas más representadas, de acuerdo con la clasificación de Gene Ontology entre los genes diferencialmente expresados en AD.

Figura 25. Gráfico circular que muestra las funciones génicas que más se han encontrado en las muestras AD, mayor expresión en la variedad mutante.

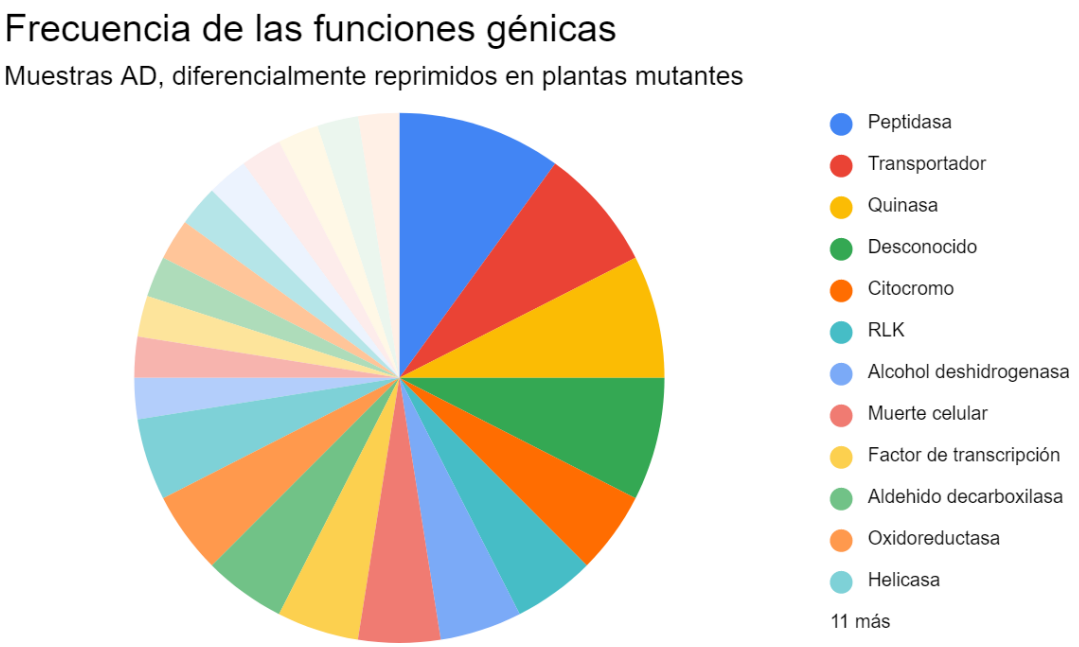

Figura 26. Gráfico circular que muestra las funciones génicas que más se han encontrado en las muestras AD, mayor expresión en la variedad mutante.

4.3 Interpretación de los resultados del análisis de expresión diferencial

Como puede verse en el apartado 4.2.1 los datos iniciales usados eran apropiados para realizar el análisis de expresión diferencial y no se han observado problemas durante el *clustering* ni durante las operaciones internas de los distintos programas.

En el apartado 4.2.2 se ha visto que en las plantas mutantes, durante los tres estadios del desarrollo analizados, existe una clara separación entre los genes sobreexpresados y reprimidos implicados en la determinación de la identidad del meristemo floral (GO:0010582), en el desarrollo vascular de los meristemos (GO:0010067), en el mantenimiento de la identidad de los meristemos apicales (GO:0010492), especificación de la identidad del órgano floral (GO:0010093), en el desarrollo de la flor (GO:0009908), en el desarrollo de la inflorescencia(GO:0010229), en el desarrollo de los carpelos (GO:0048440) y en el desarrollo del fruto(GO:0010154) y en la maduración del fruto (GO:0009836). Estos aparecen resumidos en la Tabla 6.

En cuanto a la especificación de la identidad del órgano floral y el desarrollo de la flor, puede apreciarse como los genes de tipo *AP2*, *AP2/ERF*, *FRI* y *FRL* han sido sobreexpresados mientras que los genes *TPI*, *AP3* y *AP3/PI* han sido reprimidos. También se han encontrado sobreexpresados genes que interactúan entre sí como *FLC*, *FRI* y *FRL1*.

Tabla 6. Comparación de los genes sobreexpresados y reprimidos en los mutantes para las funciones biológicas relacionadas con el desarrollo del fruto, flor y meristemo.

Función	Sobreexpresados	Reprimidos
GO:0010582	*FLC, CRC, HAT*	
GO:0010067	*WOX4*	
GO:0010492	*AP2/ERF*	
GO:0010093		*TPI, AP3/PI, AP3*
GO:0009908	*PIN1, AP2/ERF, APETALA2-like, FRL1, FRI, ARF5, ARF8*	*AP3, CYC*
GO:0010229	*PIN1*	*SLL1*
GO:0048440	*CRC*	
GO:0010154	*OVATE*	
GO:0009836		*AAT2*

En el apartado 4.2.3 se ha comprobado que también existen diferencias dentro de cada etapa del desarrollo floral analizada. En BF0 entre los genes sobreexpresados en las plantas mutantes destacan los que están implicados en la floración y los factores de transcripción, habiendo en menor medida grupos de genes implicados en la replicación, la defensa contra patógenos y el transporte. Entre los genes reprimidos han destacados genes implicados en la defensa, el transporte, los inhibidores de proteasas y los factores de transcripción.

Por otra parte, se ha podido ver que en BF1 se encuentran sobreexpresados en plantas mutantes genes relacionados con el transporte de sustancias, los factores de transcripción, la regulación de la transcripción, la defensa contra patógenos y la regulación de la traducción. En este caso han destacado especialmente los transportadores de calcio, los cuales se relacionan con el desarrollo de los frutos y con su calidad (Hernández-Perez et al., 2020). Por otra parte, entre los genes reprimidos en BF1 se han encontrado genes relacionados con el transporte, la defensa contra patógenos, la formación de la pared celular, factores de transcripción, síntesis de compuestos aromáticos y desarrollo del polen. En este caso, son transportadores de oligopéptidos y azúcares principalmente.

Finalmente, se ha podido ver que entre los genes diferencialmente expresados en muestras AD mutantes destacan aquellos que están implicados en procesos como la modificación y ensamblaje de la pared celular (algunos incluso funcionan de forma sinérgica), la maduración del fruto, la síntesis de compuestos fenólicos, el crecimiento del fruto y la síntesis de compuestos asociados con el sabor. Esto parece coincidir con los cambios fenotípicos observados. Por otra parte, entre los genes reprimidos no se han encontrado funciones que destaquen sobre el resto, aunque existe un mayor número de genes que producen peptidasas, genes que participan en el transporte y genes que producen quinasas.

5 Conclusiones

Se han podido observar diferencias claras entre los genotipos silvestres y mutantes y entre las distintas etapas del desarrollo. Se han encontrado 4958 genes que se han expresado diferencialmente. No se han observado problemas durante el procesamiento de los datos. Se ha confirmado que la mutación *eno* produce un cambio en los genes expresados que están implicados en los procesos del desarrollo meristemático, de la flor y del fruto en las tres etapas del desarrollo floral analizadas.

6 Bibliografía

Anders, S., Pyl, P. T., y Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Ban, Y., Oyama-Okubo, N., Honda, C., Nakayama, M., & Moriguchi, T. (2010). Emitted and endogenous volatiles in 'Tsugaru' apple: The mechanism of ester and (E,E)-α-farnesene accumulation. *Food Chemistry*, *118*(2), 272–277. https://doi.org/10.1016/j.foodchem.2009.04.109

Benjamini, Y., y Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Benková, E., Michniewicz, M., Sauer, M., Teichmann, T., Seifertová, D., Jürgens, G., & Friml, J. (2003). Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell*, *115*(5), 591–602. https://doi.org/10.1016/s0092-8674(03)00924-3

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., & Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*, *17*, 66. https://doi.org/10.1186/s13059-016-0924-1

Chakrabarti, M., Zhang, N., Sauvage, C., Muños, S., Blanca, J., Cañizares, J., Diez, M. J., Schneider, R., Mazourek, M., McClead, J., Causse, M., & van der Knaap, E. (2013). A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(42), 17125–17130. https://doi.org/10.1073/pnas.1307313110

Choi, K., Kim, J., Hwang, H. J., Kim, S., Park, C., Kim, S. Y., & Lee, I. (2011). The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors. *The Plant cell*, *23*(1), 289–303. https://doi.org/10.1105/tpc.110.075911

El sector del tomate – Observatorio del tomate para industria. (2018, 19 de noviembre). Recuperado el 4 de octubre de 2019, de https://observatoriotomate.com/sector/.

FAOSTAT (2017). Recuperado el 19 de septiembre de 2019, de http://www.fao.org/faostat/es/#rankings/countries_by_commodity.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., . . . Mueller, L. A. (2014). The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research, 43(D1)*. doi:10.1093/nar/gku1195

Gutensohn, M., Klempien, A., Kaminaga, Y., Nagegowda, D. A., Negre-Zakharov, F., Huh, J. H., Luo, H., Weizbauer, R., Mengiste, T., Tholl, D., & Dudareva, N. (2011). Role of aromatic aldehyde synthase in wounding/herbivory response and flower scent production in different Arabidopsis ecotypes. *The Plant journal : for cell and molecular biology*, *66*(4), 591–602. https://doi.org/10.1111/j.1365-313X.2011.04515.x

Hazak, O., Bloch, D., Poraty, L., Sternberg, H., Zhang, J., Friml, J., & Yalovsky, S. (2010). A rho scaffold integrates the secretory system with feedback mechanisms in regulation of auxin distribution. *PLoS biology*, *8*(1), e1000282. https://doi.org/10.1371/journal.pbio.1000282

Hernández-Pérez, O. I., Valdez-Aguilar, L. A., Alia-Tejacal, I., Cartmill, A. D., & Cartmill, D. L. (2020). Tomato Fruit Yield, Quality, and Nutrient Status in Response to Potassium: Calcium Balance and Electrical Conductivity in the Nutrient Solution. *Journal of Soil Science and Plant Nutrition, 20(2)*, 484–492. https://doi.org/10.1007/s42729-019-00133-9

Huang, F., Zago, M. K., Abas, L., van Marion, A., Galván-Ampudia, C. S., & Offringa, R. (2010). Phosphorylation of conserved PIN motifs directs Arabidopsis PIN1 polarity and auxin transport. *The Plant cell*, *22*(4), 1129–1142. https://doi.org/10.1105/tpc.109.072678

Informe sobre "el sector hortofrutícola en Andalucía para su internacionalización" (2017). Recuperado el 4 de octubre de 2019, de https://www.extenda.es/wp-content/uploads/2018/01/EXT_Hortofrut%C3%ADcola_2017publicado.pdf.

ITIS Standard Report Page: Solanum lycopersicum (2019). Recuperado el 10 de septiembre de 2019, de Integrated Taxonomic Information System on-line database, https://www.itis.gov.

Jiang, L., Yang, S., Xie, L., Puah, C. S., Zhang, X., Yang, W., . . . Ye, D. (2005). VANGUARD1 Encodes a Pectin Methylesterase That Enhances Pollen Tube Growth in the Arabidopsis Style and Transmitting Tract. *The Plant Cell, 17*(2), 584-596. doi:10.1105/tpc.104.027631

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., y Haussler, A. D. (2002). The Human Genome Browser at UCSC. *Genome Research, 12*(6), 996-1006. doi:10.1101/gr.229102

Lemmon, Z. H., Park, S. J., Jiang, K., Van Eck, J., Schatz, M. C., & Lippman, Z. B. (2016). The evolution of inflorescence diversity in the nightshades and heterochrony during meristem maturation. *Genome research*, *26*(12), 1676–1686. https://doi.org/10.1101/gr.207837.116

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., y 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Gu, Y., Yan, A., Lord, E., & Yang, Z. B. (2008). RIP1 (ROP Interactive Partner 1)/ICR1 marks pollen germination sites and may act in the ROP1 pathway in the control of polarized pollen growth. *Molecular plant*, *1*(6), 1021–1035. https://doi.org/10.1093/mp/ssn051

Li, X., Hamyat, M., Liu, C., Ahmad, S., Gao, X., Guo, C., Wang, Y., & Guo, Y. (2018). Identification and Characterization of the WOX Family Genes in Five *Solanaceae*

Species Reveal Their Conserved Roles in Peptide Signaling. *Genes*, *9*(5), 260. https://doi.org/10.3390/genes9050260

Lippman, Z. B., Cohen, O., Alvarez, J. P., Abu-Abied, M., Pekker, I., Paran, I., Eshed, Y., & Zamir, D. (2008). The Making of a Compound Inflorescence in Tomato and Related Nightshades. PLoS Biology, 6(11), e288. https://doi.org/10.1371/journal.pbio.0060288

Liu, J., Van Eck, J., Cong, B., & Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13302–13306. https://doi.org/10.1073/pnas.162485999

Liu, Y., Roof, S., Ye, Z., Barry, C., van Tuinen, A., Vrebalov, J., Bowler, C., & Giovannoni, J. (2004). Manipulation of light signal transduction as a means of modifying fruit nutritional quality in tomato. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9897–9902. https://doi.org/10.1073/pnas.0400935101

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

Lozano, R., Gimenez, E., Cara, B., Capel, J., & Angosto, T. (2009). Genetic analysis of reproductive development in tomato. The International Journal of Developmental Biology, 53(8–9–10), 1635–1648. https://doi.org/10.1387/ijdb.072440rl

Michaels, S. D., Bezerra, I. C., & Amasino, R. M. (2004). FRIGIDA-related genes are required for the winter-annual habit in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 3281–3285. https://doi.org/10.1073/pnas.0306778101

Mowla, S. B., Cuypers, A., Driscoll, S. P., Kiddle, G., Thomson, J., Foyer, C. H., & Theodoulou, F. L. (2006). Yeast complementation reveals a role for an Arabidopsis thaliana late embryogenesis abundant (LEA)-like protein in oxidative stress tolerance. *The Plant journal: for cell and molecular biology*, *48*(5), 743–756. https://doi.org/10.1111/j.1365-313X.2006.02911.x

Park, J. H., Suh, M. C., Kim, T. H., Kim, M. C., & Cho, S. H. (2008). Expression of glycine-rich protein genes, AtGRP5 and AtGRP23, induced by the cutin monomer 16-hydroxypalmitic acid in Arabidopsis thaliana. *Plant physiology and biochemistry: PPB*, *46*(11), 1015–1018. https://doi.org/10.1016/j.plaphy.2008.06.008

Park, S. J., Jiang, K., Schatz, M. C., y Lippman, Z. B. (2012). Rate of meristem maturation determines inflorescence architecture in tomato. *Proceedings of the National Academy of Sciences*, *109*(2), 639–644. doi: 10.1073/pnas.1114963109

Pérez, A. (2018). Tomates: producción por autonomía en España 2017. Recuperado el 1 de octubre de 2019, de https://es.statista.com/estadisticas/510892/produccion-de-tomates-en-espana-por-comunidad-autonoma/.

Pressey R. (1997). Two isoforms of NP24: a thaumatin-like protein in tomato fruit. *Phytochemistry*, *44*(7), 1241–1245. https://doi.org/10.1016/s0031-9422(96)00667-x

Sheldon, C. C., Rouse, D. T., Finnegan, E. J., Peacock, W. J., & Dennis, E. S. (2000). The molecular basis of vernalization: the central role of FLOWERING LOCUS C (FLC).

Proceedings of the National Academy of Sciences of the United States of America, *97*(7), 3753–3758. https://doi.org/10.1073/pnas.060023597

Shimizu, R., Ji, J., Kelsey, E., Ohtsu, K., Schnable, P. S., & Scanlon, M. J. (2009). Tissue specificity and evolution of meristematic WOX3 function. *Plant physiology*, *149*(2), 841–850. https://doi.org/10.1104/pp.108.130765

Shockey, J., Regmi, A., Cotton, K., Adhikari, N., Browse, J., & Bates, P. D. (2016). Identification of Arabidopsis GPAT9 (At5g60620) as an Essential Gene Involved in Triacylglycerol Biosynthesis. *Plant physiology*, *170*(1), 163–179. https://doi.org/10.1104/pp.15.01563

Singer, S. D., Chen, G., Mietkiewska, E., Tomasi, P., Jayawardhane, K., Dyer, J. M., & Weselake, R. J. (2016). Arabidopsis GPAT9 contributes to synthesis of intracellular glycerolipids but not surface lipids. *Journal of experimental botany*, *67*(15), 4627–4638. https://doi.org/10.1093/jxb/erw242

Stark, R., Grzelak, M., y Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. doi: 10.1038/s41576-019-0150-2

Tan, Q. K., & Irish, V. F. (2006). The Arabidopsis zinc finger-homeodomain genes encode proteins with unique biochemical properties that are coordinately expressed during floral development. *Plant physiology*, *140*(3), 1095–1108. https://doi.org/10.1104/pp.105.070565

The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. (2018). *Nucleic Acids Research, 47*(D1). doi:10.1093/nar/gky1055

The UniProt Consortium, UniProt: A worldwide hub of protein knowledge. (2018). *Nucleic Acids Research, 47*(D1). doi:10.1093/nar/gky1049

Tomato, Food and Agriculture Organization of the United Nations (2019). Recuperado el 10 de octubre de 2019 de http://www.fao.org/land-water/databases-and-software/crop-information/tomato/en/.

Trapnell, C., Pachter, L., y Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Tucker, G. A., Robertson, N. G., y Grierson, D. (1980). Changes in Polygalacturonase Isoenzymes during the 'Ripening' of Normal and Mutant Tomato Fruit. *European Journal of Biochemistry, 112*(1), 119-124. doi:10.1111/j.1432-1033.1980.tb04993.x

Watson, C. F., Zheng, L., & DellaPenna, D. (1994). Reduction of tomato polygalacturonase beta subunit expression affects pectin solubilization and degradation during fruit ripening. *The Plant cell*, *6*(11), 1623–1634. doi:10.1105/tpc.6.11.1623

Xu, C., Liberatore, K. L., MacAlister, C. A., Huang, Z., Chu, Y. H., Jiang, K., Brooks, C., Ogawa-Ohnishi, M., Xiong, G., Pauly, M., Van Eck, J., Matsubayashi, Y., van der Knaap, E., & Lippman, Z. B. (2015). A cascade of arabinosyltransferases controls shoot meristem size in tomato. Nature genetics, 47(7), 784–792. https://doi.org/10.1038/ng.3309

Yang, T., Peng, H., & Bauchan, G. R. (2014). Functional analysis of tomato calmodulin gene family during fruit development and ripening. *Horticulture research*, *1*, 14057. https://doi.org/10.1038/hortres.2014.57

Yuan, M., & Wang, S. (2013). Rice MtN3/saliva/SWEET family genes and their homologs in cellular organisms. *Molecular Plant, 6(3)*, 665—674. https://doi.org/10.1093/mp/sst035

Yuste-Lisbona, F. J., Fernández-Lozano, A., Pineda, B., Bretones, S., Ortíz-Atienza, A., García-Sogo, B., Müller, N. A., Angosto, T., Capel, J., Moreno, V., Jiménez-Gómez, J. M., & Lozano, R. (2020). ENO regulates tomato fruit size through the floral meristem development network. *Proceedings of the National Academy of Sciences*, *117*(14), 8187 LP – 8195. https://doi.org/10.1073/pnas.1913688117