



Centro de Investigación Científica de Yucatán, A.C.

Posgrado en Ciencias Biológicas

**Identificación de genes relacionados con la biosíntesis
de isoprenoides en transcriptomas de *Pentalinon
andrieuxii***

Tesis que presenta

MARCOS DAVID COUOH CAUICH

En opción al título de

MAESTRO EN CIENCIAS

(Ciencias Biológicas: Opción Bioquímica y Biología Molecular)

Mérida, Yucatán, México

2022

CENTRO DE INVESTIGACIÓN CIENTÍFICA DE YUCATÁN, A. C.

POSGRADO EN CIENCIAS BIOLÓGICAS



RECONOCIMIENTO

Por medio de la presente, hago constar que el trabajo de tesis de **Marcos David Couoh Cauch** titulado **Identificación de genes relacionados con la biosíntesis de isoprenoides en transcriptomas de *Pentalinon andrieuxii***, fue realizado en la **Unidad de Bioquímica y Biología Molecular** del **Centro de Investigación Científica de Yucatán, A.C.** bajo la dirección del **Dr. Gregorio Godoy Hernández** y la **Dra. Elsa Góngora Castillo**, dentro de la opción de Bioquímica y Biología Molecular, perteneciente al Programa de Posgrado en Ciencias Biológicas de este Centro.

Atentamente



Dra. Cecilia Hernández Zepeda

Directora de Docencia

Mérida, Yucatán, México, a 20 de diciembre de 2021

DECLARACIÓN DE PROPIEDAD

Declaro que la información contenida en la sección de Materiales y Métodos, los Resultados y Discusión de este documento proviene de las actividades de investigación realizadas durante el período que se me asignó para desarrollar mi trabajo de tesis, en las Unidades y Laboratorios del Centro de Investigación Científica de Yucatán, A.C., y que a razón de lo anterior y en contraprestación de los servicios educativos o de apoyo que me fueron brindados, dicha información, en términos de la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, le pertenece patrimonialmente a dicho Centro de Investigación. Por otra parte, en virtud de lo ya manifestado, reconozco que de igual manera los productos intelectuales o desarrollos tecnológicos que deriven o pudieran derivar de lo correspondiente a dicha información, le pertenecen patrimonialmente al Centro de Investigación Científica de Yucatán, A.C., y en el mismo tenor, reconozco que si derivaren de este trabajo productos intelectuales o desarrollos tecnológicos, en lo especial, estos se regirán en todo caso por lo dispuesto por la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, en el tenor de lo expuesto en la presente Declaración.



Marcos David Couoh Cauich

Este trabajo se llevó a cabo en la Unidad de Bioquímica y biología molecular del Centro de Investigación Científica de Yucatán A.C., y forma parte del proyecto titulado Análisis del transcriptoma relacionado con la biosíntesis de isoprenoides en *Pentalinon andrieuxii* (Apocynaceae). Proyecto CONACyT: 257915, bajo la dirección del Dr. Gregorio Godoy Hernández y la Dra. Elsa Góngora Castillo.

AGRADECIMIENTOS

Al Centro de Investigación Científica de Yucatán por aceptarme y permitirme realizar mi posgrado.

Al CONACYT por la beca otorgada (número 748211).

Al Dr. Gregorio Godoy Hernández por aceptarme en su grupo, guiarme y brindarme las herramientas y materiales necesarios para realizar este trabajo.

A la Dra. Elsa Góngora Castillo por aceptarme en su grupo de bioinformática, sus consejos y su tiempo dedicado a mi aprendizaje. Su pasión y experiencia compartida en los seminarios me motivó a querer continuar aprendiendo bioinformática.

A mi comité tutorial y revisores de tesis: Dr. Yair Cárdenas Conejo, gracias por los consejos y sugerencias en cada tutorial. Al Dr. Luis Manuel Peña Rodríguez y al Dr. Felipe Vázquez Flota por sus comentarios y observaciones para mejorar mi tesis.

A la M.C. Elidé Avilés Berzunza por sus consejos para desarrollar la parte experimental, estar siempre pendiente de que tengamos lo necesario y por su amistad.

A mis compañeros, Alexa, Samuel y Lucía por hacer de esta experiencia única, por ser mis amigos, escucharme y aconsejarme dentro y fuera del laboratorio.

A mis padres por apoyarme siempre.

DEDICATORIAS

A mi familia: a mis padres por inmensurable apoyo en cada etapa o proyecto de mi vida. Son un gran ejemplo de dedicación tanto en el trabajo como en la familia; a mi tía y a mis hermanitas por crecer juntos.

ÍNDICE

INTRODUCCIÓN	1
CAPÍTULO I	3
ANTECEDENTES.....	3
1.1.2 Usos	4
1.1.5 ÁCIDO BETULÍNICO.....	6
1.2. TERPENOS	7
1.2.1. BIOSÍNTESIS DE PRECURSORES DE TERPENOS EN PLANTAS	8
1.2.2. LA RUTA DEL MEVALONATO	10
1.2.3. RUTA DEL METIL ERITRITOL 4-FOSFATO (MEP).....	13
1.2.4 DIVERSIFICACIÓN DE TERPENOIDES.....	14
1.2.5 FAMILIA TRANS-PRENILTRANSFERASAS	14
1.2.6 FAMILIA TERPENOS SINTASAS.....	15
1.3 HERRAMIENTAS BIOINFORMÁTICAS PARA LA IDENTIFICACIÓN DE GENES	17
1.3.1 ENSAMBLADO <i>DE NOVO</i> Y ANÁLISIS DE TRANSCRIPTOMAS COMO HERRAMIENTA PARA LA IDENTIFICACIÓN DE GENES	17
1.3.2 HERRAMIENTAS PARA LA IDENTIFICACIÓN DE SECUENCIAS HOMÓLOGAS.....	22
1.3.2.1 BLAST	22
1.3.2.2 MODELOS OCULTOS DE MARKOV	23
1.2.3.3 ANÁLISIS FILOGENÉTICO.....	23
1.2.3.4 CUANTIFICACIÓN DE LA EXPRESIÓN <i>IN SILICO</i>	24

RECAPITULACIÓN DE ANTECEDENTES	25
JUSTIFICACIÓN	27
HIPÓTESIS.....	27
OBJETIVO GENERAL.....	28
OBJETIVOS ESPECÍFICOS.....	28
ESTRATEGIA EXPERIMENTAL	29
CAPÍTULO II	31
MATERIALES Y MÉTODOS.....	31
2.1. IDENTIFICACIÓN DE GENES INVOLUCRADOS EN LA BIOSÍNTESIS DE ISOPRENOIDES EN EL TRANSCRIPTOMA DE <i>PENTALINON ANDRIEUXII</i>	31
2.1.1 ENSAMBLADO DEL TRANSCRIPTOMA DE HOJAS Y RAÍCES DE <i>PENTALINON ANDRIEUXII</i> ..	32
2.1.2 IDENTIFICACIÓN DE SECUENCIAS DE TRANSCRITOS RELACIONADOS CON LA BIOSÍNTESIS DE TERPENOS EN LOS TRANSCRIPTOMAS DE <i>P. ANDRIEUXII</i>	33
2.2 CLASIFICACIÓN FILOGENÉTICA DE SECUENCIAS	35
2.3 ANÁLISIS DE EXPRESIÓN <i>IN SILICO</i> DE GENES IDENTIFICADOS.	35
2.4 VALIDACIÓN DEL ENSAMBLADO DEL TRANSCRIPTOMA MEDIANTE EL AISLAMIENTO Y SECUENCIACIÓN DE TRANSCRITOS.....	36
2.4.1 EXTRACCIÓN DE ARN Y SÍNTESIS DE ADNC	36
2.4.2 AMPLIFICACIÓN POR PCR PUNTO FINAL DE LOS TERPENOS SINTASA	38
CAPÍTULO III.....	41
RESULTADOS	41
3.1 ENSAMBLADO DE NOVO DEL TRANSCRIPTOMA DE <i>PENTALINON ANDRIEUXII</i>	41

3.2 IDENTIFICACIÓN DE SECUENCIAS DE TRANSCRITOS RELACIONADOS CON LA BIOSÍNTESIS DE TERPENOS EN LOS TRANSCRIPTOMAS DE <i>P. ANDRIEUXII</i>.....	47
3.2.1 CLASIFICACIÓN FILOGENÉTICA DE LAS SECUENCIAS IDENTIFICADAS	55
3.3 ANÁLISIS DE EXPRESIÓN <i>IN SILICO</i>	62
3.4 VALIDACIÓN DEL ENSAMBLADO DEL TRANSCRIPTOMA MEDIANTE EL AISLAMIENTO Y SECUENCIACIÓN DE TRANSCRITOS.....	67
3.4.1 CLONACIÓN.....	71
CAPÍTULO IV	74
3.1. DISCUSIÓN	74
CAPÍTULO V	85
CONCLUSIONES Y PERSPECTIVAS.....	85
5.1 CONCLUSIONES	85
5.2 PERSPECTIVAS.....	86
BIBLIOGRAFÍA	87
ANEXOS.....	106

LISTADO DE FIGURAS

Figura 1.1. Características de <i>Pentalinon andrieuxii</i> (Müll. Arg.) Hansen & Wunderlin. a) Inflorescencia. b) Vainas. c) semillas con vilano. d) semillas desprovistas de su vilano.....	3
Figura 1.2. Estructura química del urechitol A y B.....	6
Figura 1.3. Estructura y ruta de biosíntesis del ácido Betulínico.....	7
Figura 1.4. Compartimentación subcelular de las vías MVA y MEP en células vegetales.	10
Figura 1.5. Ruta metabólica del mevalonato (MVA).....	11
Figura 1.6. Ruta metabólica del 2-C-metil-D-eritritol-4-fosfato (MEP)	13
Figura 1.7. Número de genes asociados al metabolismo isoprenoide en plantas dicotiledóneas de las cuales su genoma se ha secuenciado..	17
Figura 1.8 Diagrama general de la estrategia experimental.....	29
Figura 3.1. Distribución de la calidad de las secuencias obtenidas de la plataforma Illumina para el tejido de raíz joven de <i>P. andrieuxii</i> calculado con el programa FastQC.	42
Figura 3.2. Contenido por base obtenidas de la plataforma Illumina para el tejido de raíz joven de <i>P. andrieuxii</i> calculado con el programa FastQC.....	43
Figura 3.3. Distribución de la calidad de las secuencias filtradas con la herramienta Trimmomatic para el tejido de raíz joven de <i>P. andrieuxii</i> calculado con el programa FastQC.	45
Figura 3.4. Alineamiento de las dos secuencias HMGR identificadas de <i>Pentalinon andrieuxii</i> junto a HMGR conocidas de otras especies.....	56
Figura 3.5. Análisis filogenético de las proteínas DXS.....	57
Figura 3.6. Análisis filogenético de la familia trans-preniltransferasas.	59
Figura 3.7. Árbol filogenético de la familia terpeno sintasa, incluyendo las secuencias identificadas de <i>Pentalinon andrieuxii</i>	60

Figura 3.8. Motivos conservados de las TPS de <i>Pentalinon andrieuxii</i> identificadas..	61
Figura 3.9. Expresión <i>in silico</i> de genes de la biosíntesis de IPP..	63
Figura 3.10. Expresión <i>in silico</i> de los genes de la familia trans-preniltransferasa de <i>Pentalinon andrieuxii</i> .	64
Figura 3.11. Expresión <i>in silico</i> de los genes de la familia terpenos sintasas de <i>Pentalinon andrieuxii</i> .	66
Figura 3.12. Amplificación de las sesquiterpenos sintasas TPS2, TPS7, TPS8 y TPS10.	68
Figura 3.13. Amplificación del monoterpeno sintasa TPS9 y de los diterpenos sintasas TPS4, TPS6 y TPS11.	69
Figura 3.14. Amplificación de los monoterpenos sintasas TPS1 y TPS3.	70
Figura 3.15. Amplificación de las secuencias tps1, tps2 y tps3 mediante PCR usando ADNc sintetizado a partir de ARN proveniente de tejido de hoja adulta de <i>P. andrieuxii</i> .	71
Figura 3.16. Verificación del ADNc clonado mediante restricción enzimática y PCR.	72

LISTADO DE TABLAS

Tabla 1.1. Enzimas y genes de la ruta del mevalonato identificados en <i>Arabidopsis thaliana</i> (Modificado de Vranová <i>et al.</i> , 2013).	12
Tabla 2.1. Lecturas generadas por condición	32
Tabla 2.3. Condiciones usadas para la PCR.	38
Tabla 3.1. Lecturas antes y después del filtrado de calidad con Trimmomatic.	44
Tabla 3.2. Estadísticas del transcriptoma ensamblado.	46
Tabla 3.3. Resultado de la herramienta BUSCO para el transcriptoma de <i>Pentalinon andrieuxii</i> ensamblado.	47
Tabla 3.4. Número de secuencias de cada base de datos usadas para la identificación de secuencias con blastx.	48

Tabla 3.5. Resultado de la metodología aplicada para la identificación de genes relacionados con la biosíntesis de isoprenoides.	48
Tabla 3.6. Comparación de herramientas de traducción de secuencias.	49
Tabla 3.7. Secuencias identificadas de la ruta del mevalonato.	50
Tabla 3.8. Secuencias identificadas de la ruta MEP.	51
Tabla 3.9. Secuencias identificadas de la familia trans-preniltransferasa.....	52
Tabla 3.10. Secuencias identificadas de la familia terpeno sintasa.....	53
Tabla 3.11. Porcentaje de alineamiento obtenido usando la herramienta Salmon.	62

ABREVIATURAS

ADN: Ácido desoxirribonucleico

ADNc: Ácido desoxirribonucleico complementario

ARN: Ácido ribonucleico

ARNm: Ácido ribonucleico mensajero

ARN-seq: Secuenciación de ARN

BLAST: Basic Local Alignment Search Tool

BLASTX: Basic Local Alignment Search Tool, modo para buscar en bases de datos de proteínas utilizando una consulta de nucleótidos traducida

Cds: Región de codificación

DMAPP: Pirofosfato de dimetilalilo

E.C.: Enzyme Commission number

HMM: Modelos ocultos de Márkov

IPP: Isopentenil difosfato

MEP: Vía del 2-C-metil-D-eritritol 4-fosfato

MVA: Vía del mevalonato

PCR: Reacción en cadena de la Polimerasa

PT: Preniltransferasa

TPM: Transcritos por millón

TPS: Terpeno sintasa.

RESUMEN

Los terpenos, también conocidos como terpenoides o isoprenoides, constituyen la familia de productos naturales más diversa. Actualmente se han reportado más de 80000 estructuras diferentes, además, están presentes en todos los organismos vivos, en los cuales cumplen funciones primarias y en algunos casos, como en las plantas, cumplen funciones especializadas. Sin embargo, el mayor interés por estudiar estos compuestos deriva en que son utilizados en la industria alimentaria y en la medicina.

En las plantas, los precursores universales de los isoprenoides, el IPP y DMAPP se sintetizan en dos rutas diferentes, cada una ubicada en compartimentos diferentes, la ruta citosólica se conoce como ruta del mevalonato, mientras que la ruta cloroplastídica es llamada ruta MEP. La diversidad tan grande de estos compuestos se debe a la participación de varias familias de genes, sin embargo, los genes preniltransferasa y terpenos sintasa son los responsables de la formación de la gran mayoría de los isoprenoides.

En este estudio, se realizó el ensamblado de *novo* del transcriptoma de *Pentalinon andrieuxii* que servirá de referencia para futuras investigaciones. Además, se desarrolló una metodología para identificar secuencias completas que correspondan a genes de la ruta MVA, de la ruta MEP y de las familias preniltransferasa y terpenos sintasas, lo cual resultó un desafío debido a la alta redundancia de secuencias y la presencia de intrones en la mayoría de los contigs.

De las once secuencias de la familia terpenos sintasas identificadas, tres sesquiterpenos sintasas y una monoterpene sintasa mostraron patrones de expresión específicos de tejido, estas secuencias podrían ser un punto de partida para lograr elucidar la ruta biosintética del urechitol, que debido a su inusual estructura y su dinámica de acumulación surge la incógnita sobre su función en la planta.

ABSTRACT

Terpenes, also known as terpenoids or isoprenoids, constitute the most diverse family of natural products. Currently more than 80,000 different structures have been reported, in addition, they are present in all living organisms, in which they perform primary functions and in some cases, as in plants, they fulfill specialized functions. However, the greatest interest in studying these derivative compounds is that they are used in the food industry and in medicine.

In plants, the universal precursors of isoprenoids, IPP and DMAPP, are synthesized in two different pathways, each located in different compartments, the cytosolic pathway is known as the mevalonate pathway, while the chloroplastid pathway is called MEP pathway. The great diversity of these compounds is due to the participation of several gene families, however, the prenyltransferase and terpenes synthase genes are responsible for the formation of the vast majority of isoprenoids.

In this study, the de novo assembly of the *Pentalinon andrieuxii* transcriptome was carried out, which will serve as a reference for future research, and a methodology was developed to identify complete sequences that correspond to genes of the MVA and MEP pathway, and the prenyltransferase and terpene synthases families, which was challenging due to the high sequence redundancy and the presence of introns in most contigs.

Of the 11 sequences of the terpene synthase family identified, three sesquiterpene synthases and a monoterpene synthase showed tissue-specific expression patterns. These sequences could be a starting point to elucidate the biosynthetic pathway of urechitol, which due to its unusual structure and its accumulation dynamics raises the unknown about its biosynthesis pathway and its function in the plant.

INTRODUCCIÓN

Pentalinon andrieuxii, es una planta utilizada en la medicina tradicional maya para el tratamiento de la leishmaniasis. Estudios fitoquímicos de esta planta permitieron el aislamiento de dos nuevos trinosesquiterpenos, los urechitoles A y B (1 y 2) que presentan un esqueleto novedoso denominado campechano (Yam-Puc *et al.*, 2009), de estos compuestos se desea conocer su ruta de biosíntesis y el órgano u órganos donde ello ocurre. Otro metabolito aislado es el ácido betulínico, un triterpeno pentacíclico con esqueleto de tipo lupano, el cual presenta numerosas actividades biológicas, siendo las más importantes la anticancerígena y antiviral (Moghaddam *et al.*, 2012). La investigación actual en esta planta en nuestro grupo, se centra en una parte en aumentar el contenido de ácido betulínico y por otro lado identificar los genes responsables de la biosíntesis de los urechitoles, así como el lugar y condiciones fisiológicas de su biosíntesis. A pesar que se han realizado varias investigaciones en torno a los urechitoles, poco se ha avanzado debido a que estos metabolitos solo se logran cuantificar en bajísimas concentraciones en raíces de plantas con desarrollo intermedio a adulto (Hiebert-Giesbrecht *et al.*, 2016) y en los cultivos *in vitro* muestreados por diferentes estudios no produjeron urechitoles.

La importancia de estudiar el metabolismo isoprenoide es debido a que esta clase de moléculas, desempeñan funciones importantes en el metabolismo primario (biosíntesis de clorofilas) y secundario (atraer polinizadores, establecer simbiosis, repeler herbívoros), además presentan propiedades de interés científico e industrial (anticancerígenos, colorantes) (Guerriero *et al.*, 2018). Por su complejidad muchos terpenoides no son posibles de obtener por síntesis química debido a su gran número de centros quirales, por lo tanto, la producción biológica sigue siendo el método preferido para la producción industrial de estos compuestos. Existe un gran interés en aumentar la producción o expresión de las vías biosintéticas de los metabolitos de interés, lo cual requiere la elucidación de su ruta de biosíntesis (Ikram *et al.*, 2015).

La biosíntesis de los isoprenoides comienza con la producción del IPP y el DMAPP mediante dos rutas; la ruta del ácido mevalónico (MVA) y la ruta del metil eritritol fosfato (MEP). Seguidamente, las enzimas preniltransferasas se encargan de unir estas unidades para formar estructuras lineales que posteriormente serán cicladas y sus enlaces reordenados por las enzimas terpeno sintasas para producir los esqueletos que serán la base para la formación de una gran diversidad de isoprenoides (Booth *et al.*, 2017).

Este trabajo de investigación tiene como principal objetivo generar un transcriptoma de *Pentalinon andrieuxii* que servirá para identificar los genes de ambas rutas MEV y MEP, de las familias preniltransferasas y terpenos sintasas, con la perspectiva de elucidar la ruta de biosíntesis del urechitol. Las estrategias y recursos generados se podrán aplicar para el estudio de los demás metabolitos novedosos de esta especie como el pentalinonsterol o conocer más sobre la biosíntesis del ácido betulínico.

CAPÍTULO I

ANTECEDENTES

1.1 *Pentalinon andrieuxii* (Müll. Arg) Hansen & Wunderlin

Pentalinon andrieuxii, llamado comúnmente bejuco de la víbora, bejuco guaco o contrayerba, es una planta trepadora leñosa, alcanza hasta 6 metros de alto; posee hojas opuestas ovales, de 5 a 10 cm de largo; tal como se observa en la figura 1.1, presenta inflorescencias en forma de cima escorpioide arqueada, flores de color amarillo, de hasta 7 cm de largo y en forma de trompeta, sus frutos son dehiscentes, pueden medir hasta 25 cm de largo, cafés en forma de vaina, los cuales en su interior tienen numerosas semillas color café claro, de aproximadamente 7 mm de largo y 1 mm de ancho en cuyo extremo se ubica un penacho de pelos denominado vilano, de hasta 2.5 cm de largo (Rzedowski y Calderón, 1998). Se distribuye principalmente en la península de Yucatán, florece de junio a septiembre (Juárez *et al.*, 2007).



Figura 1.1. Características de *Pentalinon andrieuxii* (Müll. Arg.) Hansen & Wunderlin. a) Inflorescencia. b) Vainas. c) semillas con vilano. d) semillas desprovistas de su vilano.

1.1.2 Usos

Pentalinon andrieuxii es utilizada en la medicina tradicional maya para el tratamiento de la leishmaniasis cutánea localizada, comúnmente llamada "úlceras del chiclero". Los curanderos tradicionales mayas recomiendan lavar la lesión con una infusión de raíz de *P. andrieuxii*, y luego aplicar la raíz seca en polvo sobre el área (Chan-Bacab *et al.*, 2003). Esta planta también se usa para tratar mordeduras de serpiente, para aliviar disturbios nerviosos y dolores de cabeza (Argüeta *et al.*, 1994).

1.1.3 Estudios realizados en *Pentalinon andrieuxii*

Los primeros estudios fitoquímicos reportaron actividad antidepresiva y antiinflamatoria del extracto de raíz de la planta (Jiu, 1966); igualmente, actividad antiprotozoaria en los extractos de plantas contra promastigotes de *Leishmania mexicana* (Viscencio *et al.*, 1995).

La leishmaniasis provoca lesiones desfigurantes en la piel, o en órganos internos, causando infecciones letales, esta enfermedad es producida por protozoos del género *Leishmania* los cuales parasitan células del sistema inmunológico, se transmite mediante la picadura de la mosca de arena infectada. Aproximadamente 12 millones de personas se encuentran infectadas por *Leishmania* (Getti *et al.*, 2009). Los fármacos empleados para tratar esta enfermedad presentan elevada toxicidad y requieren de una prolongada administración, por lo tanto, existe una continua búsqueda de metabolitos efectivos contra esta condición (Murray, 2001).

La humedad parece jugar un papel importante en la actividad antiprotozoaria en *Pentalinon andrieuxii*, un estudio realizado en diferentes poblaciones silvestres de esta especie demostró que las plantas que crecen con mayor humedad ambiental presentan extractos con mayor actividad antileishmanial (Chan-Bacab *et al.*, 2003).

Del extracto metanólico de hojas de *P. andrieuxii* se ha identificado el taraxasterol (Getti *et al.*, 2009), el cual es antialérgico (Liu *et al.*, 2013), antiinflamatorio y también se sabe que exhibe una actividad contra el veneno de serpiente (Mors *et al.*, 2000).

Los metabolitos con mayor actividad antileishmanicida se encuentran en la raíz de *Pentalinon andrieuxii* (Lezama-Dávila *et al.*, 2007), los cuales no habían sido reportados antes y son el pentalinonsterol (colest-4,20,24-trien-3-one), el pentalinonside (14,16-14,21-15,20-triepoxy-14,15secopregnan-5-en-3-ol3-O-β-D-diginopyranoside) y el más potente es el compuesto 6,7-

dihydroneeridienone, todos derivados de los esteroides, además otros 18 compuestos conocidos, incluidos 14 esteroides, tres cumarinas y un triterpeno han podido detectarse (Pan *et al.*, 2012).

El pentalinosterol (PEN, colest-4,20,24-trien-3ona) ha recibido especial atención debido a su facilidad de producirse químicamente a bajo costo y a que presenta una elevada actividad antileishmanicida selectiva, a baja concentración y nula toxicidad tanto *in vivo* e *in vitro* (Gupta *et al.*, 2015). Además de sus propiedades antiparasitarias, esta sustancia exhibe actividad inmunomoduladora, potenciando las respuestas inmunitarias del huésped lo cual tiene aplicaciones en el tratamiento de enfermedades infecciosas e inmuno-asociadas, así como en el diseño de vacunas (Oghumu *et al.*, 2017).

1.1.4 Urechitoles

En el extracto metanólico de la raíz de *P. andrieuxii* se ha aislado e identificado trinorsesquiterpenoides (C₁₂) estructuralmente inusuales y biológicamente no activos, nombrados urechitol (U) A y B (figura 1.2), de los cuales el UA se encuentra en mayor cantidad que el UB, presenta una insaturación en el C₈ y un H menos en la posición C₅ que el UB. Las estructuras se identificaron mediante la interpretación de datos espectroscópicos y cristalografía de rayos X. Este compuesto presenta un nuevo esqueleto de sesquiterpeno nombrado “campechano” (Yam-Puc *et al.*, 2009). El urechitol A se detecta principalmente en raíces de plantas de desarrollo intermedio y adulto, aumentando su concentración durante la floración (Hiebert-Giesbrecht *et al.*, 2016), se ha sugerido que toda la planta es necesaria para su biosíntesis (Hiebert-Giesbrecht *et al.*, 2021). Actualmente se desconoce su ruta biosintética, experimentos de marcado isotópico de ¹³CO₂ en plantas enteras de *P. andrieuxii* han resultado infructuosos debido a la producción de urechitol A marcado (Peña-Rodríguez *et al.* 2014), la limitante en el estudio de este metabolito, es su concentración en condiciones específicas de desarrollo de la planta, sin embargo, un estudio reciente reportó un aumento de hasta 3.35 veces la concentración de este metabolito en plantas transformadas genéticamente con los genes rol de *A. rhizogenes* (Hiebert-Giesbrecht *et al.*, 2021).

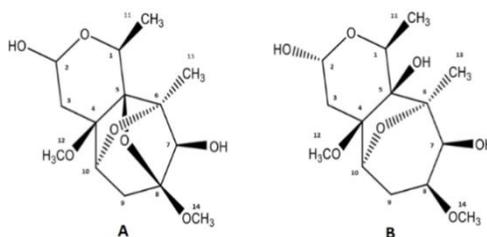


Figura 1.2. Estructura química del urechitol A y B (Modificado de Yam-Puc *et al.*, 2009).

1.1.5 Ácido betulínico

El ácido betulínico (AB) es un triterpeno pentacíclico con esqueleto de tipo lupano, su estructura química puede verse en la figura 1.3, el cual se sintetiza a partir de precursores de la ruta del mevalonato (Peña-Rodríguez *et al.*, 2014). El AB se encuentra ampliamente distribuido en el reino vegetal. Entre las actividades biológicas descritas para el ácido betulínico, destaca su actividad anti-melanómica, anti-neuroblastómica, anti-leucémica, anti-viral VIH, anti-malárica, entre otras (Cano-Flores, 2013; Moghaddam *et al.*, 2012).

Este compuesto se encuentra en la corteza de varias especies de plantas, principalmente el abedul blanco *Betula pubescens* (Tan *et al.*, 2003) de la que recibe su nombre. La corteza de la planta *Betula platyphylla* es la principal fuente de AB para el mercado farmacéutico. En China, se consumen alrededor de 230,000 toneladas de las cortezas de *Betula platyphylla* por año para la extracción de AB (Zhou *et al.*, 2016). Se ha reportado un contenido máximo de 20.22 % de betulina y 1.86% de ácido betulínico en trozos de corteza seca en una población silvestre de esta especie (Zhao *et al.*, 2007), mientras que en invernadero se reporta que su acumulación es de aproximadamente 35 mg/g de peso seco (Ying *et al.*, 2013).

La ruta biosintética de AB ha sido elucidada (Huang *et al.*, 2012); requiere la ciclación del 2,3-oxidoescualeno por parte de la enzima lupeol sintasa para producir lupeol, el cual es el paso comprometido inicial hacia la biosíntesis de AB. El lupeol es oxidado sucesivamente en su posición C₂₈ para producir AB mediante una enzima de la familia citocromo P450, la lupeol C-28 oxidasa (LO) (Zhou *et al.*, 2016). Se ha reportado que una amirina oxidasa de *Catharanthus roseus* (CrAO) mostró la actividad LO (Huang *et al.*, 2012). Mediante la expresión combinatoria de una CrAO y una lupeol sintasa de *Arabidopsis thaliana* (AtLup1), se ha logrado la producción

de AB en *Saccharomyces cerevisiae* a través de la producción endógena de 2,3-oxidosqualeno en la levadura (Huang *et al.*, 2012). Además, la actividad de NADPH reductasa se requiere para la actividad de las enzimas P450 de la planta (Urban *et al.*, 1997).

El AB se encuentra principalmente en hojas de *Pentalinon andrieuxii* a lo largo de toda la vida de la planta y la iluminación presenta cierta influencia en su concentración. La concentración media reportada es de 0.5 mg/g de peso seco en las hojas, pudiendo aumentar hasta 11,37 veces su concentración en plantas transformadas con los genes rol de *A. rhizogenes* (Hiebert-Giesbrecht *et al.*, 2016; Almeyda-Cen, 2017, Hiebert-Giesbrecht *et al.*, 2021).

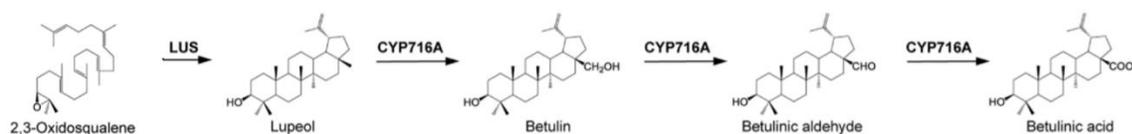


Figura 1.3. Estructura y ruta de biosíntesis del ácido Betulínico (modificado de Suzuki *et al.*, 2018).

1.2. Terpenos

Los terpenos, también conocidos como terpenoides o isoprenoides, constituyen la familia de productos naturales más diversa y están presentes en todos los organismos vivos (Pichersky y Raguso, 2018). Actualmente cuenta con más de 80000 miembros que también incluyen esteroides y carotenoides (Buckingham *et al.*, 2016; Christianson, 2017). Muchos isoprenoides son de interés económico para la producción de caucho, así como para nutracéuticos, aromas, fragancias, pigmentos, agroquímicos y desinfectantes (Bohlmann y Keeling, 2008); además, muestran efectos farmacológicos significativos, como antivirales, antibacterianos, antimaláricos, antiinflamatorios, actividades anticancerígenas (Lovett, 2005), anti-protozoaria (Pan *et al.*, 2012), entre otras.

La biosíntesis de isoprenoides es esencial en todos los organismos vivos, y su ausencia impone un parasitismo intracelular obligado (Boucher y Doolittle, 2000). En las plantas, los isoprenoides, cumplen importantes funciones primarias en la fotosíntesis (clorofilas y carotenoides), transporte de electrones (cadenas laterales de ubiquinona y plastoquinona), regulación del crecimiento y desarrollo (giberelinas, ácido abscísico, estrigolactonas, brasinoesteroides, citocininas), en la glucosilación de proteínas (dolicoles), o como elementos de la estructura y función de la membrana (fitoesteroides) (Vranová *et al.*, 2013). Por otra parte, existen metabolitos terpenoides

especializados (que solo se encuentran en una o un grupo de especies en particular) que participan en interacciones alelopáticas entre patógenos y plantas para proteger a las plantas contra herbívoros y patógenos, y también se producen para atraer polinizadores y animales dispersores de semillas, entre otras interacciones bióticas por encima y por debajo del suelo. (Loreto *et al.*, 2014; Tholl, 2015; Vranová *et al.*, 2013).

1.2.1. Biosíntesis de precursores de terpenos en plantas

A pesar de su diversidad funcional y química, todos los isoprenoides están relacionados biosintéticamente por un precursor común de cinco carbonos, el isopentenil difosfato (IPP) y su isómero, el dimetilalil pirofosfato (DMAPP) (Enfissi *et al.*, 2005). Los isoprenoides se clasifican con base al número de unidades que conforman su estructura química, estos se pueden dividir en monoterpenos (10 carbonos), sesquiterpenos (15 carbonos) y diterpenos (20 carbonos). Los terpenos más grandes incluyen triterpenos, tetraterpenos y politerpenoides, con 30, 40 y >40 carbonos, respectivamente. Algunos terpenos lineales (C_{10} , C_{15} , C_{20} , C_{30} y C_{40}) se someten a ciclación, oxidación (de P450) y otras reacciones enzimáticas para producir más de 80 000 terpenoides (Hamberger y Bak, 2013; Misawa, 2011).

Las plantas sintetizan IPP y DMAPP a través de dos rutas metabólicas independientes (ver figura 1.4) y que se encuentran en compartimentos celulares diferentes. Los precursores de isoprenoides citosólicos y mitocondriales se sintetizan a través de la vía del ácido mevalónico (MVA) mientras que la vía del 2-C-metil-D-eritritol 4-fosfato (MEP), también conocida como vía 1-desoxi- d -xilulosa 5-fosfato (DXP), produce IPP y DMAPP en los plástidos. El hecho de que ambas rutas metabólicas se encuentren en compartimentos diferentes, permite la utilización rentable y la canalización de recursos necesarios para el crecimiento o defensa de la planta, lo que asegura que las plantas sobrevivan en un entorno dinámico, aumentando así la aptitud y la capacidad competitiva (Kliebenstein, 2004). La vía MVA proporciona predominantemente los precursores de la biosíntesis citosólica de sesquiterpenoides, poliprenoles, fitoesteroles, brasinoesteroides y triterpenoides, y para la biosíntesis de terpenoides en mitocondrias (p. Ej., ubiquinonas, poliprenoles), y los precursores derivados de la vía MEP son preferiblemente utilizados para la biosíntesis de hemiterpenoides (p. ej., isopreno), monoterpenoides, diterpenoides, tetraterpenos (carotenoides), reguladores de crecimiento como el ácido abscísico, citocininas, giberelinas y brasinoesteroides, proteínas preniladas, plastoquinonas, compuestos poliprenoides no vitamínicos (dolicoles), vitaminas (vitamina A) y fitoesteroles (vitamina K) y sus productos de descomposición, clorofila, tocoferoles. Sin embargo, experimentos de marcado

sugieren que algunos isoprenoides pueden producirse a partir de precursores producidos por ambas vías (Hemmerlin *et al.*, 2012), en especial en la biosíntesis de sesquiterpenos (Wu, 2006). Aunque un flujo cruzado limitado de IPP y prenil difosfatos (5C a 15C) puede tener lugar entre los plástidos y el citosol de ciertas especies, el efecto causado por la inhibición de una de las dos rutas de síntesis de IPP, ya sea farmacológico o genético, no puede ser compensado por los productos suministrados por la otra vía metabólica en condiciones normales de crecimiento (Laule *et al.*, 2003).

dos enzimas, primeramente la acetoacetyl-CoA tiasa (AACT; E.C. 2.3.1.9) es la responsable de condensación de dos AcCoA para formar una molécula de acetoacetyl-coenzima A (AcAc-CoA); seguidamente la 3-hidroxi-3-metilglutaril-CoA sintasa (HMGS; EC 2.3.3.10) da lugar al 3-hidroxi-3-metilglutaril-CoA mediante la condensación del AcAc-CoA con la tercera molecular de CoA. El siguiente paso es irreversible, la HMG-CoA es convertida en mevalonato mediante dos pasos de reducción los cuales requieren NADPH, esta reacción está mediada por la enzima HMG-CoA reductasa (HMGR; E.C. 1.1.1.34) (Tholl y Lee, 2011). Los pasos restantes hacia IPP comprenden dos reacciones de fosforilación para convertir MVA en mevalonato 5-difosfato (MVADP), catalizado por mevalonato cinasa (MK; E.C. 2.7.1.36) y fosfomevalonato cinasa (PMK; E.C. 2.7.4.2), seguido de una descarboxilación dependiente de ATP del mevalonato bifosforilado (MVADP), catalizada por la difosfo-mevalonato descarboxilasa (PPMD; E.C. 4.1.1.33) para formar el IPP. Finalmente, la isopentenil-difosfato isomerasa (IDI; E.C. 5.3.3.2) cataliza la conversión reversible de IPP a DMAPP (Hemmerlin *et al.*, 2012), esta ruta biosintética se representa en la figura 1.5.

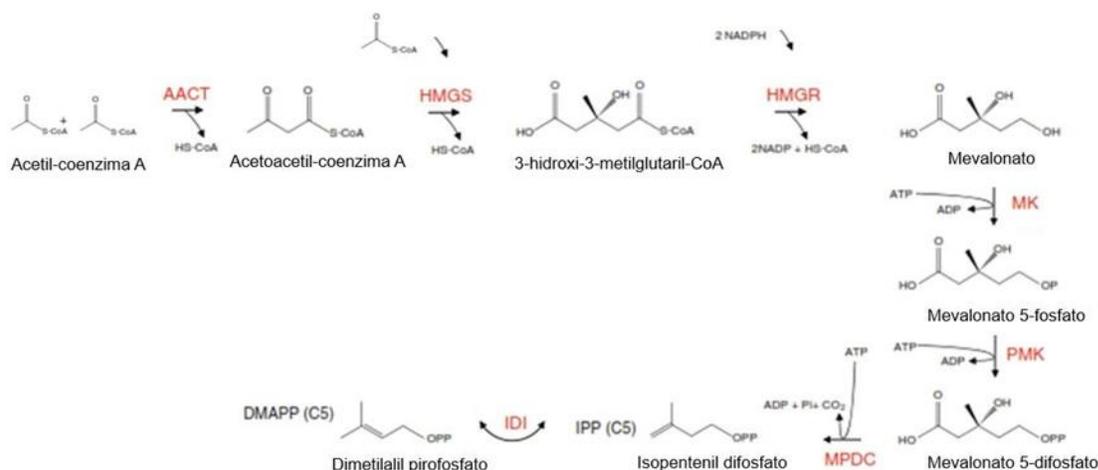


Figura 1.5. Ruta metabólica del mevalonato (MVA) (Modificado de Hemmerlin *et al.*, 2012).

En la planta modelo *Arabidopsis thaliana*, se ha identificado que el número de genes que codifican para cada enzima puede ser variado. En la Tabla 1.1, se resumen los nombres de los genes de *A. thaliana* que participan en la ruta MVA y algunas de sus características.

Tabla 1.1. Enzimas y genes de la ruta del mevalonato identificados en *Arabidopsis thaliana* (Modificado de Vranová *et al.*, 2013).

Enzima		Gen		
Abreviación	Nombre (E.C.)	Identificador	Isoformas	Localización
AACT	Acetoacetil-CoA tiolasa (2.3.1.9)	At5g47720	5	Citosol/peroxisomas
		At5g48230	2	Citosol
HMGS	3-hidroxi-3-metilglutaril-CoA sintasa (2.3.3.10)	At4g11820	2	Citosol
HMGR	3-hidroxi-3-metilglutaril-CoA reductasa (1.1.1.34)	At1g76490	2	Retículo endoplasmático
		At2g17370	1	Retículo endoplasmático
MK	Mevalonato cinasa (2.7.1.36)	At5g27450	3	Citosol
PMK	fosfomevalonato cinasa (2.7.4.2)	At1g31910	2	Peroxisomas
PPMD	difosfo-mevalonato decarboxilasa (4.1.1.33)	At2g38700	1	Desconocido
		At3g54250	1	Desconocido

A diferencia de las enzimas que se enlistan en la Tabla 1.1 que participan únicamente en la vía MVA, la enzima IPPI participa en ambas rutas de síntesis de IPP.

A. thaliana tiene dos genes que codifican la enzima IPPI (At3g02780, At5g16440), ambos genes producen dos variantes de la proteína, las variantes largas de los genes AtIPPI1 and AtIPPI2 se localizan en las mitocondrias (AtIPPI1L) y plástidos (AtIPPI1L, AtIPPI2L). Las versiones cortas de las proteínas AtIPPI1 and AtIPPI2, se localizan en los peroxisomas y producen los precursores para los isoprenoides citosólicos (Okada *et al.*, 2008; Phillips *et al.*, 2008; Sapir-Mir *et al.*, 2008). Muy contrastantemente la especie *Catharanthus roseus* tiene un único gen que codifica a la enzima IPPI, lo que supone una triple localización celular similar a AtIPPI1 (Guirimand *et al.*, 2012).

1.2.3. Ruta del metil eritritol 4-fosfato (MEP)

Las enzimas de la ruta de MEP son codificadas por los genes nucleares y sus productos proteicos son dirigidos a los plástidos. La reacción inicial de la vía MEP (figura 1.6) comienza con la condensación del piruvato y el D-gliceraldehído-3-fosfato (G3P) para generar 1-desoxi-D-xilulosa-5-fosfato (DXP) y O_2 , esta reacción es catalizada por la enzima 1-desoxi-D-xilulosa-5-fosfato sintasa (DXS; E.C. 2.2.1.7). La DXP es reducida por la 1-desoxi-D-xilulosa 5-fosfato reductoisomerasa (DXR; E.C. 1.1.1.267) para generar 2-C-metil-D-eritritol 4-fosfato (MEP). El Tercer paso de la ruta conlleva la conversión de MEP en 2-metileritritol-2,4-ciclodifosfato (ME-CDP), una reacción dependiente de citidina trifosfato (CTP) y catalizada por la enzima 4-difosfocitidil-2-C-metil-D-eritritol sintasa (MCT; E.C. 2.7.7.60), posteriormente, ME-CDP es nuevamente fosforilado por la 4-difosfocitidil-2-C metilD-eritritol cinasa (CMK; E.C. 2.7.1.148), una reacción dependiente de ATP que da como producto el compuesto 4-difosfocitidil-2-C-metil-D-eritritol-2-fosfato (ME2P-CDP); ME2P-CDP es convertido en 2-C-metil-D-eritritol 2,4-ciclodifosfato (MEcPP; E.C. 4.6.1.12) mediante la pérdida de CMP catalizada por 2-C-metil-D-eritritol-2,4ciclodifosfato sintasa (MDS; E.C. 4.6.1.12). El MEcPP es el sustrato para la enzima 4-hidroxi-3-metil-2-butenil-difosfato sintasa (HDS; EC1.17.7.1) que genera 4-hidroxi-3-metil-2-enil difosfato (HMBPP) el cual será reducido por la enzima 4-hidroxi-3-metil-2-enil difosfato reductasa (HDR; E.C. 1.17.1.2) a IPP y DMAPP en una proporción de 5 a 6:1, esta razón es ajustada a 7:3 por la enzima isopentenil difosfato isomerasa (IDI). (Tholl y Lee, 2011; Bach y Rohmer, 2014; Rodríguez, 2010).

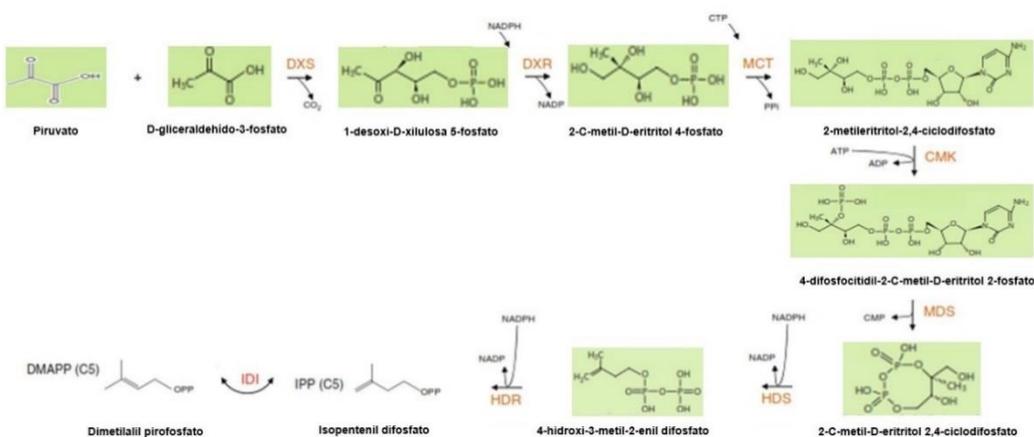


Figura 1.6. Ruta metabólica del 2-C-metil-D-eritritol-4-fosfato (MEP) (Modificado de Hemmerlin *et al.*, 2012).

En la mayoría de las especies de plantas, todas las proteínas de la ruta MEP residen en el estroma del plástido y están codificadas por genes de copia única (Joyard *et al.*, 2009), siendo la única excepción las enzimas DXS que están codificadas por una pequeña familia de genes parálogos, aunque no todos los miembros de esta familia parecen estar involucrados en la biosíntesis de IPP (Estévez *et al.*, 2000; Phillips *et al.*, 2008; Zhang *et al.*, 2018).

1.2.4 Diversificación de terpenoides

Una vez sintetizado el IPP y el DMAPP, los terpenoides se forman a continuación mediante la condensación de restos de IPP adicionales a través de las preniltransferasas. Los monoterpenoides se derivan de geranil pirofosfato (GPP, 10C), los sesquiterpenoides, del farnesil pirofosfato (FPP, 15C), y los diterpenoides del geranilgeranil pirofosfato (GGPP, 20C). Incluso terpenoides mas grandes como los triterpenoides (30C) se forman por condensación de dos moléculas de FPP, y los tetraterpenos (C40) se forman en gran parte mediante la condensación de dos moléculas de GGPP (Roberts, 2007). Después de la formación de los bloques estructurales de terpenoides acíclicos, las enzimas terpeno sintasa actúan para generar el esqueleto de carbono del terpeno principal. Transformaciones adicionales que a menudo implican oxidación, reducción, isomerización y enzimas de conjugación decoran o alteran el esqueleto principal con grupos funcionales variados para producir la familia terpenoide ampliamente diversa de compuestos (Booth *et al.*, 2017).

1.2.5 Familia trans-preniltransferasas

La diversidad de los isoprenoides se determina en primer lugar por las cis- o trans-preniltransferasas que median la condensación consecutiva de IPP con diversos sustratos de PP alílicos, incluida la condensación con DMAPP en cis o trans-estereoisómeros (Jia y Chen *et al.*, 2016).

En las plantas, han caracterizado seis tipos de trans-PT, incluida geranil-PP (GPP, C10) sintasa (GPS), farnesil-PP (FPP, C15) sintasa (FPS), geranilgeranil-PP (GGPP, C20) sintasa (GGPS), geranilfarnesil-PP (GFPP, C25) sintasa (GFPS), solanesil-PP (SPP, C45) sintasa (SPS) y poliprenil-PP (PPP, C>50) sintasa (PPS) (You *et al.*, 2020).

Una de las clasificaciones más recientes de las trans-PT propone 3 grupos (You *et al.*, 2020), el primer grupo quedan incluidos las enzimas GPS, GGPS y GFPS, este grupo se subdivide en seis grupos: el subgrupo A lo conforman las subunidades largas de GGPS homodimérico las cuales

producen GGPP y tienen un motivo de interacción con los miembros de la familia E y F; el subgrupo B de las enzimas GFPS lo forman únicamente proteínas de *Arabidopsis thaliana* que producen geranilfarnesil difosfato (C25), este tipo de terpenos solo se han identificado en miembros de la familia Brassicaceae (Wang *et al.*, 2016); el subgrupo C se encuentra conformado por G(G)PS homoméricas que utilizan FPP para producir GGPP y PPP;

1.2.6 Familia terpenos sintasas

Los precursores formados por las preniltransferasas son sustrato para las enzimas terpenos sintasas (TPS), la reacción realizada por las enzimas TPS es el punto de mayor diversificación de los isoprenoides, debido a que algunas enzimas como la δ -selineno sintasa y la γ -humuleno sintasa de la especie *Abies grandis* pueden producir hasta 34 y 52 diferentes sesquiterpenos, respectivamente (Steele, *et al.*, 1998).

Según el sustrato que utilicen, las TPS se pueden clasificar en monoterpenos sintasas si actúan sobre el GPP, si transforman el FPP son denominadas sesquiterpenos sintasas, los diterpenos sintasas son capaces de utilizar GGPP. Los genes de las enzimas terpenos sintasas codifican proteínas de 550 a 850 aminoácidos, en general, las monoterpenos sintasas son entre 600 y 650 aa de longitud y son más grandes que las sesquiterpenos sintasas por 50-70 aa, esta diferencia es en gran parte el resultado de los péptidos de tránsito N-terminales requeridos para el direccionamiento plastidial de las monoterpenos sintasas; la mayoría de las diterpenos sintasas son aproximadamente 210 aa más largas que las monoterpenos sintasas (Bohlmann *et al.*, 1998).

El análisis de varios genomas de plantas que se han secuenciado y anotado indica que, con la excepción del musgo *Physcomitrella patens*, que tiene un solo gen TPS funcional. La familia de genes TPS es una familia de tamaño medio, con números de genes que van desde aproximadamente 20 a 100 (Hofberger *et al.*, 2015), en la figura 1.7 se ilustra un recuento del número de genes asociados a los diferentes módulos del metabolismo isoprenoide de diferentes plantas dicotiledóneas las cuales su genoma se ha secuenciado.

Los genes TPS derivan de un gen bifuncional CPS/KS ancestral, que se fue duplicando, perdiendo la actividad en uno de sus dos dominios, subfuncionalizando y neofuncionalizando, lo que resultó CPS, KS y TPS del metabolismo especializado tanto en las gimnospermas como en las angiospermas (Chen *et al.*, 2011).

Según el mecanismo catalítico y el producto formado, los genes TPS también se clasifican en dos clases: clase I y clase II. Los TPS de clase I y II tienen motivos de aminoácidos conservados únicos que son esenciales para la catálisis (Chen *et al.*, 2011; Gao *et al.*, 2012).

Las enzimas TPS de clase I tienen un dominio C-terminal (también denominado dominio α o pliegue clase I) que cataliza la ionización del sustrato mediada por un catión divalente. Esta ionización del sustrato dependiente del metal puede conducir a ciclaciones, cambios de hidruros y reordenamientos estructurales para producir el producto final. El dominio α presente en esta clase de TPS adopta el pliegue de proteína α -helicoidal y contiene dos motivos de unión de metal 'DDXXD' altamente conservado y un motivo 'NSE/DTE' menos conservado colocado en hélices opuestas cerca de la entrada del sitio activo (Kumar *et al.*, 2018).

Las enzimas terpenos sintasas de clase II comprenden un N-terminal funcional (dominio β), con un tercer dominio y de "inserción" que forma un pliegue vestigial de clase II. Las enzimas de esta clase contienen un motivo conservado funcional 'DXDD' que reside en un dominio β separado y responsable de la ciclación iniciada por la protonación del sustrato (Gao *et al.*, 2012; Thoma *et al.*, 2004) El dominio γ lleva un motivo similar a EDXXD altamente ácido, que contribuye a la actividad de los TPS de clase II. Muchas enzimas terpenos sintasas también tienen un motivo RR[X]₈W altamente conservado cerca del péptido de tránsito N-terminal, mientras que este motivo no se requiere para la actividad de monoterpene sintasa (Cao *et al.*, 2010).

Además, los genes TPS se clasifican en ocho subfamilias: TPS-a a TPS-h en función de las propiedades de secuencia y las características funcionales. La familia de TPS de Clase I contiene TPS-a, -b, -e / f, y -g, donde la clase II contiene TPS-c solamente que comprende genes para las enzimas diterpeno sintasa relacionadas con el copatilo difosfato. La familia TPS-d se compone principalmente de enzimas bifuncionales que son capaces de la ciclación iniciada por protonación o por ionización y se encuentran principalmente en las gimnospermas. La subfamilia TPS-h es específica de la espiguilla *Selaginella moellendorffii* (Chen *et al.*, 2011).

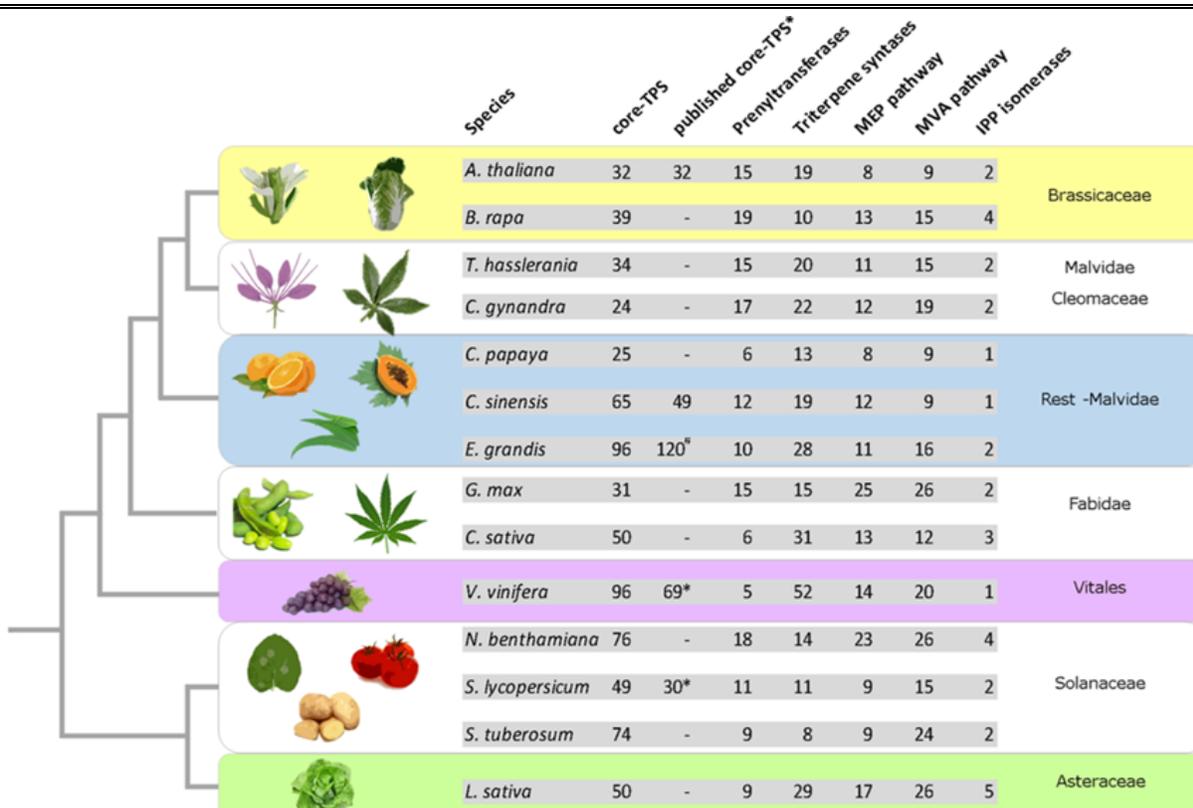


Figura 1.7. Número de genes asociados al metabolismo isoprenoide en plantas dicotiledóneas de las cuales su genoma se ha secuenciado. Modificado de Hofberger *et al.*, 2015.

1.3 Herramientas bioinformáticas para la identificación de genes

1.3.1 Ensamblado *de novo* y análisis de transcriptomas como herramienta para la identificación de genes

Un transcriptoma es el conjunto completo de transcripciones (moléculas de ARN) presentes dentro de una célula o tejido, entre los tipos de ARN presentes en un transcriptoma se pueden encontrar los ARN no codificantes y codificantes de proteínas junto a todas sus formas de procesamiento alternativo, poliadenilamiento alternativo y transcripciones editadas con ARN, estos genes en conjunto reflejan que se expresa activamente en un tejido en particular (Lu *et al.*, 2013).

El ensamblaje de transcriptomas es el proceso de reconstruir las secuencias completas de todas las transcripciones y sus variantes que se expresan en una muestra determinada (Lu *et al.*, 2013).

Un transcriptoma ensamblado es un conjunto de secuencias contiguas (contig) que representan regiones de transcripción (Li *et al.*, 2014). En ausencia de un genoma de referencia para una especie vegetal, se deberá utilizar un ensamblaje *de novo* del transcriptoma para construir las transcripciones completas (Miller *et al.*, 2010).

Se conoce como ensamblaje de transcriptoma *de novo* al proceso mediante el cual se combinan lecturas de secuenciación de ARN superpuestas para reconstruir secuencias de transcripción sin ningún conocimiento previo de un genoma de referencia, característica que lo hace atractivo particularmente cuando el costo y la complejidad de generar un genoma eucariota son prohibitivos (Moreton *et al.*, 2016).

En general, el transcriptoma ensamblado *de novo* se puede utilizar como referencia para alinear las lecturas de secuencias del mismo o de otro experimento para determinar la expresión génica diferencial y explorar la diversidad genética. De manera particular, en organismos modelo que cuentan con su genoma secuenciado tales como *Arabidopsis*, el arroz y el maíz, generar conjuntos de transcripciones *de novo* son útiles para descubrir nuevas isoformas de transcripción de genes anotados existentes, eventos de procesamiento alternativos y transcritos de genes novedosos de una variedad vegetal, o en respuesta a un tratamiento específico (Geniza y Jaiswal, 2017).

A pesar de todas las ventajas y aplicaciones del ensamblado y análisis de transcriptomas *de novo*, no debe verse como una alternativa rápida y se deben tomar a consideración unas pautas específicas que se citan a continuación.

Plataforma de secuenciación. Para el ensamblaje de transcriptomas *de novo*, se recomienda secuenciar mediante de las plataformas de Illumina, esta plataforma cuenta con dos formatos principales de lecturas, las lecturas en formato single-end (SE) que son apropiadas para el ensamblaje de transcriptomas *de novo* cuando existen recursos secuenciados de referencia de la misma especie o de especies estrechamente relacionadas, en caso contrario, se recomienda secuencias las lecturas en formato paired-end (PE) porque conservan la información sobre la direccionalidad de la transcripción (Levin *et al.*, 2010).

Control de calidad de las lecturas sin procesar antes del ensamblaje del transcriptoma. Un estudio de RNA-Seq puede producir cientos de millones de lecturas por muestra, en el caso de la plataforma Illumina, las lecturas son comúnmente secuencias de 25-250 nucleótidos de longitud producidas por una reacción cíclica de terminación reversible donde se asociada señales

colorimétricas específicas de cada base dentro de la máquina de secuenciación (Del Fabbro *et al.*, 2013). Estas señales colorimétricas se traducen en llamadas de base mediante un software interno de Illumina (CASAVA), representado en el formato FASTQ (Cock *et al.*, 2010). Cada nucleótido está asociado a un número de calidad codificado en ASCII correspondiente a una puntuación Phred (Q) (Ewing y Green, 1998), que se traduce directamente en la probabilidad p de que la llamada de base correspondiente sea incorrecta, el valor de p se define mediante la siguiente ecuación:

$$p = 10^{(-Q/10)}$$

Por ejemplo, una puntuación Q de 30 (Q30) a una base, es equivalente a la probabilidad de que exista una llamada de base incorrecta cada 1000 bases secuenciadas, la tasa de error aceptada para la plataforma Illumina es aproximadamente del 1% o 1/100 bases. Por tanto, los datos necesitan más controles de calidad y filtrado (Hartley y Mullikin, 2015; Li *et al.*, 2015), por lo tanto, los investigadores pueden optar por recortar bases potencialmente llamadas incorrectamente en las lecturas utilizando herramientas como Trimmomatic (Bolger *et al.*, 2014).

Herramientas para el ensamblado de *novo*. Los ensambladores de transcriptomas de *novo* suelen utilizar una estrategia que implica la construcción de gráficos de Bruijn, que consiste en localizar dentro de cada lectura todas las subsecuencias de longitud k , conocidos como "k-mers", que se encuentren, para crear un gráfico de De Bruijn se utiliza todos los k-mers únicos como nodos, con bordes de conexión que representan k-mers inmediatamente superpuestos (Moreton *et al.*, 2016), "es decir, si una subcadena k-mer se desplaza por una base de secuencia y se superpone a otra k-mer (por $k-1$ bases), entonces se dibuja un borde entre los nodos asociados con esos k-mers. Una cadena lineal de nodos k-mer se comprime en un solo nodo cuando es posible (donde los dos nodos están unidos por un solo borde único). Las variantes de la transcripción se pueden ensamblar recorriendo las diferentes rutas del gráfico" (Martin y Wang, 2011). Actualmente, no existe una herramienta de ensamblaje óptima para todos los conjuntos de datos de RNA-Seq. En general, las herramientas de ensamblado como Trinity, SPAdes y Trans-ABYSS, superan a otras herramientas y deberían ser las preferidas, dentro de las cuales, logró Trinity obtuvo la mejor puntuación de métrica general, sin embargo, las necesidades computacionales de esta herramienta, suelen ser muy altos (Hölzer y Marz, 2019).

Evaluación del ensamblaje del transcriptoma de *novo*. Después del ensamblaje del transcriptoma *de novo* se suelen calcular unas métricas de calidad para los ensamblajes como

el "ExN50" que examina las transcripciones más expresadas que representan el 50% del total de datos de expresión normalizada. El ExN50 toma en cuenta la naturaleza dinámica de los transcriptomas y requiere una estimación de la abundancia de las transcripciones para poder calcularlas. Además, es importante calcular el porcentaje de lecturas alineadas al transcriptoma ensamblado, en general entre el 70-90% de todas las lecturas deben asignarse al ensamblado (Moreton *et al.*, 2016).

Los autores Hölzer y Marz, 2019 recomiendan para evaluar ensamblajes las herramientas como BUSCO (Sim *et al.*, 2015), TransRate (Smith-Unna *et al.*, 2016) y DETONATE (Li *et al.*, 2014).

Los conjuntos de datos de la herramienta BUSCO (Benchmarking Universal Single-Copy Ortholog) comprenden genes que evolucionan bajo "control de copia única" (Waterhouse *et al.* 2011), es decir, "...dentro de cada linaje están presentes casi universalmente como ortólogos de una sola copia. La completitud se cuantifica en términos de este contenido génico esperado mediante la evaluación del estado de la ortología de los genes predichos utilizando perfiles de secuencia BUSCO" (Waterhouse *et al.*, 2018).

Problemáticas presentes en el análisis de transcriptomas.

Presencia de intrones.

Los estudios de transcriptomas de mamíferos que utilizan ARN-seq muestran que, aunque la mayoría de las lecturas de secuencias están asociadas con exones en genes conocidos, muchas son intrónicas (Kapranov *et al.*, 2011; van Bakel *et al.*, 2010; Wetterbom *et al.*, 2010). Además, sin estudios más detallados, no se puede demostrar si las lecturas intrónicas se originan a partir de transcripciones independientes ubicadas dentro de intrones, o si representan transcripciones inmaduras que aún no se han empalmado (Ameur *et al.*, 2011). Se ha identificado un mecanismo denominado splicing co-transcripcional, que es el proceso por el cual la maquinaria de empalme trabaja detrás de la ARN polimerasa para formar productos empalmados a medida que la polimerasa avanza con la transcripción (Bentley, 2005; Goldstrohm *et al.*, 2001; Kornblihtt *et al.*, 2004), este mecanismo fue recientemente identificado en *Arabidopsis* (Li *et al.*, 2020). Los autores Ameur *et al.*, 2011 demostraron que el patrón de cobertura de lectura de secuencia intrónica se explica por la transcripción naciente en combinación con el empalme co-transcripcional.

Los experimentos realizados por Sultan *et al.*, 2014 sugieren que la gran mayoría de las lecturas intrónicas identificadas en experimentos de ARN-seq, corresponden a transcripciones nucleares no procesadas en lugar de a unidades transcripcionales independientes, además, algunos métodos de extracción de ARN basados en el uso de TRIzol en conjunto con protocolos de secuenciación de ARN basados en el agotamiento del ARNr retienen mayor cantidad de especies de ARN nuclear y ARN inmaduros, lo cual plantea problemas con respecto a la estimación de los niveles de expresión de los genes codificantes.

AUGUSTUS (Stanke y Waack, 2003; Stanke *et al.*, 2008; Keller *et al.*, 2011) es una herramienta para encontrar genes que codifican proteínas y su estructura exón-intrón en secuencias genómicas, para realizar esto requiere parámetros específicos de la especie para su modelo de Markov oculto subyacente (HMM) o campo aleatorio condicional (CRF), en su página web (<http://bioinf.uni-greifswald.de/augustus/submission.php>) se encuentran modelos de diferentes especies preestablecidos, sin embargo, usarlo para analizar datos de otras especies restará precisión. Los ARNm inmaduros al contener intrones se parecen más a su secuencia genómica y no a su secuencia codificante (cds), por lo tanto, con AUGUSTUS se podría predecir las cds correspondiente de cada ARNm inmaduro.

Isoformas redundantes.

Idealmente, cada contig corresponde a una determinada isoforma de transcripción. En eucariotas, cada locus potencialmente puede producir varias transcripciones (isoformas) debido a eventos de empalme alternativo (Conesa *et al.*, 2016). Las lecturas cortas derivadas de 1 exón pueden ser parte de múltiples rutas en el gráfico de ensamblaje. Por lo tanto, la estructura del gráfico puede ser ambigua y las isoformas representadas pueden ser difíciles de resolver. Derivado de esto, herramientas como Trinity pueden generar múltiples isoformas con secuencias redundantes (Hölzer y Marz, 2019).

Cuando se ensambla un transcriptoma *de novo*, puede ser difícil identificar isoformas ensambladas de manera redundante debido a los diferentes niveles de expresión. CD-HIT-EST (Li y Godzik, 2006) agrupa secuencias similares en un transcriptoma ensamblado y genera un conjunto de secuencias representativas no redundantes. En resumen, CD-HIT-EST toma la secuencia canónica de un grupo producido para eliminar secuencias por encima de un umbral de identidad especificado y corregir el sesgo dentro de un transcriptoma ensamblado dado

(Geniza y Jaiswal, 2017), además el paquete de herramientas de CD-HIT cuenta con una versión para proteínas que opera bajo el mismo principio.

1.3.2 Herramientas para la identificación de secuencias homólogas

1.3.2.1 BLAST

Los alineamientos de secuencias a menudo proporcionan la primera conexión entre el ADN o la proteína recién secuenciada y las secuencias ya categorizadas. La herramienta BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1997) es una de las opciones más populares para buscar relaciones entre dos secuencias. BLAST toma una secuencia de nucleótidos o aminoácidos como entrada y la busca en una base de datos de secuencias de nucleótidos o proteínas. BLAST utiliza heurística para acelerar las búsquedas. BLAST también proporciona estadísticas que estiman la probabilidad de que ocurra una coincidencia por casualidad (Boratyn *et al.*, 2013).

El algoritmo BLAST está ajustado para encontrar dominios funcionales que a menudo se repiten dentro de la misma proteína, así como en diferentes proteínas de diferentes especies o tramos cortos de similitud de secuencia. Al consultar una secuencia, BLAST primero hace una tabla de búsqueda de todas las "palabras" (subsecuencias cortas, que para proteínas el valor predeterminado es tres letras) y "palabras vecinas", es decir, palabras similares en la secuencia de consulta. Luego, la base de datos de secuencias se escanea en busca de estos "puntos calientes". Cuando se identifica una coincidencia, se utiliza para iniciar la extensión del alineamiento que pueden tener espacios o no, a cada coincidencia, sustitución, inserción o deleción se le asigna una puntuación y se les asigna un valor estadístico, llamado "Valor esperado". El "valor esperado" es el número de veces que se esperaría que ocurriera por casualidad una alineación tan buena o mejor que la encontrada por BLAST (Madden, 2005; Wheeler y Bhagwat, 2007).

La función de una proteína también se puede inferir de su relación evolutiva con proteínas de función conocida, siempre que la relación se inspeccione adecuadamente. Las proteínas ortólogas en diferentes especies a menudo comparten la función, pero la paralogía (es decir, la divergencia después de la duplicación del gen original) no garantiza una función común (Loewenstein *et al.*, 2009).

Algoritmos como BLAST detectan similitudes de secuencia (Altschul *et al.*, 1990). Sin embargo, muchos impactos BLAST insignificantes (por valor *e*) son homólogos. Se puede obtener una mayor sensibilidad utilizando métodos más sofisticados, como algoritmos de búsqueda más precisos como PSI-BLAST y modelos ocultos de Markov (HMM) (Boekhorst y Snel, 2007).

1.3.2.2 Modelos ocultos de Markov

Un modelo de Markov oculto (HMM) es un modelo estadístico que se puede utilizar para describir la evolución de eventos observables que dependen de factores internos, que no son directamente observables. Llamamos al evento observado un "símbolo" y al factor invisible subyacente a la observación un "estado". Un HMM consta de dos procesos estocásticos, a saber, un proceso invisible de estados ocultos y un proceso visible de símbolos observables. Los estados ocultos forman una cadena de Markov y la distribución de probabilidad del símbolo observado depende del estado subyacente (Rabiner, 1989)

Normalmente, una secuencia biológica consta de subestructuras más pequeñas con diferentes funciones, y diferentes regiones funcionales a menudo muestran distintas propiedades estadísticas. Por ejemplo, es bien sabido que las proteínas generalmente constan de múltiples dominios. Dada una nueva proteína, sería interesante predecir los dominios que la constituyen (correspondientes a uno o más estados en un HMM) y sus ubicaciones en la secuencia de aminoácidos (observaciones). Además, es posible que también deseemos encontrar la familia de proteínas a la que pertenece esta nueva secuencia de proteínas (Yoon, 2009).

1.2.3.3 Análisis filogenético

El análisis filogenético de los datos de la secuencia de proteínas es una parte integral de la anotación de genes, la predicción de la función de los genes, la identificación y construcción de familias de genes y el descubrimiento de genes. Los árboles filogenéticos son estructuras matemáticas que representan la historia evolutiva de un grupo de organismos o genes (Rokas, 2011). El objetivo de los árboles filogenéticos es representar relaciones históricas (es decir, evolutivas), y no el grado de semejanza. Claramente, el grado de similitud de secuencia no es igual al grado de relación evolutiva (Sternier *et al.*, 2006).

Inferencia de homología. En los análisis filogenéticos de datos moleculares se asume que las proteínas en estudio son homólogas; es decir, que están relacionadas por descendencia con la misma proteína ancestral. Es solo después de que se hayan inferido (o asumido) los homólogos

que se puede realizar el análisis filogenético. La inferencia de homólogos se realiza normalmente mediante la realización de análisis de búsqueda de similitud con algoritmos de búsqueda de alineación local, como BLAST (Rokas, 2011).

Tipos de homología. Dos genes son homólogos si se heredan de un gen ancestral. La ortología es un tipo especial de homología en el que los genes de diferentes especies han divergido entre sí debido a la especiación (Koonin, 2005; Fitch, 1970). Otras formas de homología incluyen la paralogía se refieren a homólogos dentro de un linaje y se deben, por ejemplo, a duplicaciones en tándem, transposición o genoma completo (Ohno, 1970; Freeling, 2009).

1.2.3.4 Cuantificación de la expresión *in silico*

Cuando se está investigando los fundamentos moleculares de un estímulo, una enfermedad, una mutación genética o cualquier otra perturbación, la cuantificación de la expresión génica es la parte más importante de un experimento de RNA-Seq. Debido a que brinda información sobre cómo se regulan los genes y revela detalles de la biología de un organismo. También puede ayudar a inferir las funciones de genes previamente no anotados (Lowe *et al.*, 2017; Koen *et al.*, 2019).

Para vincular la abundancia de lecturas secuenciadas a la expresión de un gen particular, las secuencias de transcripción se alinean con un genoma o transcriptoma de referencia. Debido a la complejidad de los genes eucariotas, la cuantificación de una transcripción se puede medir a nivel gen o a nivel de isoforma, sin embargo, la expresión total de un gen es la suma de la expresión de sus isoformas y se pierde información sobre la participación de una isoforma en específico en un proceso biológico y se recomienda cuantificar a nivel isoforma (Lowe *et al.*, 2017; Koen *et al.*, 2019).

Para asignar cada lectura, el alineador debe determinar el punto de origen probable de la lectura con respecto a la referencia, muchos alineadores utilizan un índice de genoma para reducir rápidamente la lista de ubicaciones de alineación candidatas (Langmead y Salzberg, 2012), sin embargo, este proceso puede aumentar el tiempo de ejecución y los requerimientos computacionales de las herramientas, por lo que actualmente se han desarrollado herramientas como Salmon (Patro *et al.*, 2017) que realiza un mapeo ligero sin crear índices, seguidamente estima los niveles de expresión iniciales que posteriormente son refinados en una tercera fase. Los resultados obtenidos con Salmon se han comparado con otras herramientas que requieren un mapeo inicial, demostrando que no tienen diferencias significativas, por lo tanto, para

optimizar recursos y tiempo de ejecución, Salmon es una alternativa destacable (Schaarschmidt *et al.*, 2020; Zhang *et al.*, 2017).

Para un gen dado, el número de lecturas mapeadas no solo depende de su nivel de expresión y longitud del gen, sino también de la profundidad de secuenciación y no se pueden comparar directamente entre muestras, para normalizar estas dependencias, medidas como TPM (transcripciones por millón) se utilizan para medir los niveles de expresión de genes o transcripciones (Zhao *et al.*, 2020). Sin embargo, la variación en la preparación de la biblioteca o la composición del ARN entre muestras también contribuye a la variabilidad entre muestras y debe tenerse en cuenta (Robinson y Oshlack, 2010).

“Para una muestra de ARN determinada, si tuviera que secuenciar un millón de transcripciones completas, un valor de TPM representa la cantidad de transcripciones que habría visto para un gen o isoforma determinados” (Zhao *et al.*, 2020) y se calcula de la manera siguiente:

$$TPM = 10^6 * \frac{\text{lecturas mapeadas a una transcripción} / \text{largo de de la transcripción}}{\text{sum(lecturas mapeadas a una transcripción} / \text{largo de de la transcripción)}}$$

Recapitulación de antecedentes

Los terpenos, también conocidos como terpenoides o isoprenoides, constituyen la familia de productos naturales más diversa, actualmente cuenta con más de 80000 miembros (Buckingham *et al.*, 2016; Christianson, 2017). Muchos isoprenoides son de interés económico debido a que poseen un amplio rango de actividades biológicas como antivirales, antibacterianas, antiparasitarias o para la industria alimentaria como nutracéuticos, aromas, fragancias, pigmentos (Bohlmann y Keeling, 2008; Lovett, 2005).

Todos los isoprenoides están relacionados biosintéticamente por un precursor común de cinco carbonos, el isopentenil difosfato (IPP) y su isómero, el dimetilalil pirofosfato (DMAPP) (Enfissi *et al.*, 2005). Las plantas producen el IPP y DMAPP a través de dos rutas metabólicas independientes, los precursores de isoprenoides citosólicos y mitocondriales se sintetizan a través de la vía del ácido mevalónico (MVA) mientras que la vía del 2-C-metil-D-eritritol 4-fosfato (MEP), también conocida como vía 1-desoxi-d-xilulosa 5-fosfato (DXP), produce IPP y DMAPP en los plástidos (Hemmerlin *et al.*, 2012).

Una vez sintetizado el IPP y el DMAPP, los terpenoides se forman a continuación mediante la condensación de DMAPP y diferentes unidades de IPP a través de las enzimas preniltransferasas (Jia y Chen *et al.*, 2016). Los precursores formados por las preniltransferasas son sustrato para las enzimas terpenos sintasas (TPS), la reacción realizada por las enzimas TPS es el punto de mayor diversificación de los isoprenoides, debido a que algunas enzimas como pueden producir decenas de compuestos diferentes (Steele, *et al.*, 1998).

La búsqueda de plantas medicinales con actividades biológicas importantes, condujo al estudio de *Pentalinon andrieuxii*, una especie utilizada en la medicina tradicional maya para el tratamiento de las lesiones cutáneas provocadas por la *Leishmania*. Derivado de la búsqueda de metabolitos con propiedades para combatir la *Leishmania*, se identificaron el pentalinonsterol, el ácido betulínico y los urechitoles A y B, perteneciendo los tres a la familia de los isoprenoides (Pan *et al.*, 2012; Hiebert-Giesbrecht *et al.*, 2016). El AB se encuentra principalmente en hojas de *Pentalinon andrieuxii* a lo largo de toda la vida de la planta, mientras que el urechitol A se detecta principalmente en raíces de plantas de desarrollo intermedio y adulto, aumentando su concentración durante la floración (Hiebert-Giesbrecht *et al.*, 2016). En los últimos años se ha intentado encontrar la ruta biosintética de los urechitoles y estrategias para aumentar la producción del ácido betulínico.

Para especies no modelo, para las cuales no se cuenta con su genoma secuenciado, el análisis bioinformático de transcriptomas ensamblados *de novo*, ha permitido la identificación de transcritos potencialmente involucrados en algún proceso biológico como la biosíntesis de un metabolito en particular (Góngora-Castillo y Buell, 2013). La cuantificación de la expresión génica es la parte más importante de un experimento de RNA-Seq. Debido a que brinda información sobre cómo se regulan los genes y revela detalles de la biología de un organismo. También puede ayudar a inferir las funciones de genes previamente no anotados (Lowe *et al.*, 2017; Koen *et al.*, 2019).

El análisis de los transcriptomas de hojas y raíces de plantas jóvenes y adultas de *Pentalinon andrieuxii*, en especial los patrones de expresión de los genes que codifican enzimas clave en la biosíntesis de isoprenoides como las enzimas de la ruta MVA, MEP, de las familias preniltransferasa y terpeno sintasa permitiría la identificación de genes involucrados en metabolitos específicos de esta especie como el urechitol A.

JUSTIFICACIÓN

Los isoprenoides o terpenoides desempeñan funciones importantes en las plantas, tanto a nivel primario, como en su interacción con el ambiente. Además, muchos isoprenoides presentan actividades biológicas importantes en seres humanos, lo que ha motivado el estudio de estos compuestos.

En *Pentalinon andrieuxii* se han identificado derivados isoprenoides, por lo que, en los transcriptomas de hoja y raíz de plantas adultas y jóvenes, se pretende identificar, genes de la ruta del ácido mevalónico (MVA) y del 2-C-metil-D-eritritol 4-fosfato (MEP), ambas encargadas de producir IPP y DMAPP, la primera en el citosol, la segunda en los cloroplastos. También se identificarán los genes de prenil transferasas y terpenos sintasas, ambas familias de genes codifican enzimas son las responsables de la gran diversidad de isoprenoides, de los cuales se conocen al menos 80 000 estructuras diferentes.

Mediante la realización de este trabajo de investigación se obtendrá un transcriptoma que servirá como referencia para estudios posteriores relacionados con *Pentalinon andrieuxii* y especies cercanas, anteriormente ninguna secuencia se ha reportado para la especie, contar con esta herramienta permitirá realizar estudios moleculares para estudiar diferentes aspectos de este organismo, además se pretende identificar genes involucrados en la biosíntesis de mono, sesqui y diterpenos, dichas secuencias posteriormente se podrán usar como base para estudios de qPCR, clonación y caracterización de las proteínas correspondientes.

HIPÓTESIS

La variación en la producción de isoprenoides en los tejidos de hojas y raíces de plantas adultas y jóvenes de *Pentalinon andrieuxii* se debe los diferentes patrones de expresión de los genes que codifican para enzimas de las rutas de biosíntesis de IPP, prenil transferasas y terpeno sintasas, implicados en la biosíntesis de dichos isoprenoides.

OBJETIVO GENERAL

Analizar los patrones de expresión de los genes de la ruta mevalonato, metileritrol fosfato, de las familias trans-preniltransferasas y terpenos sintasas implicados en la biosíntesis de isoprenoides en *Pentalinon andrieuxii*.

OBJETIVOS ESPECÍFICOS

1. Identificar genes de la ruta del mevalonato y del 2-C-metil-D-eritritol 4-fosfato, de las familias trans-preniltransferasas y terpenos sintasas en el transcriptoma de hojas y raíces de *Pentalinon andrieuxii* mediante homología de secuencias.
2. Clasificar las secuencias identificadas de los genes DXS, de las familias trans-preniltransferasa y terpenos sintasas a través de un análisis filogenético
3. Analizar los niveles de expresión *in silico* de los genes de ambas rutas de biosíntesis de IPP, de las familias prenil transferasa y terpenos sintasas identificados en los transcriptomas de *Pentalinon andrieuxii*.

ESTRATEGIA EXPERIMENTAL

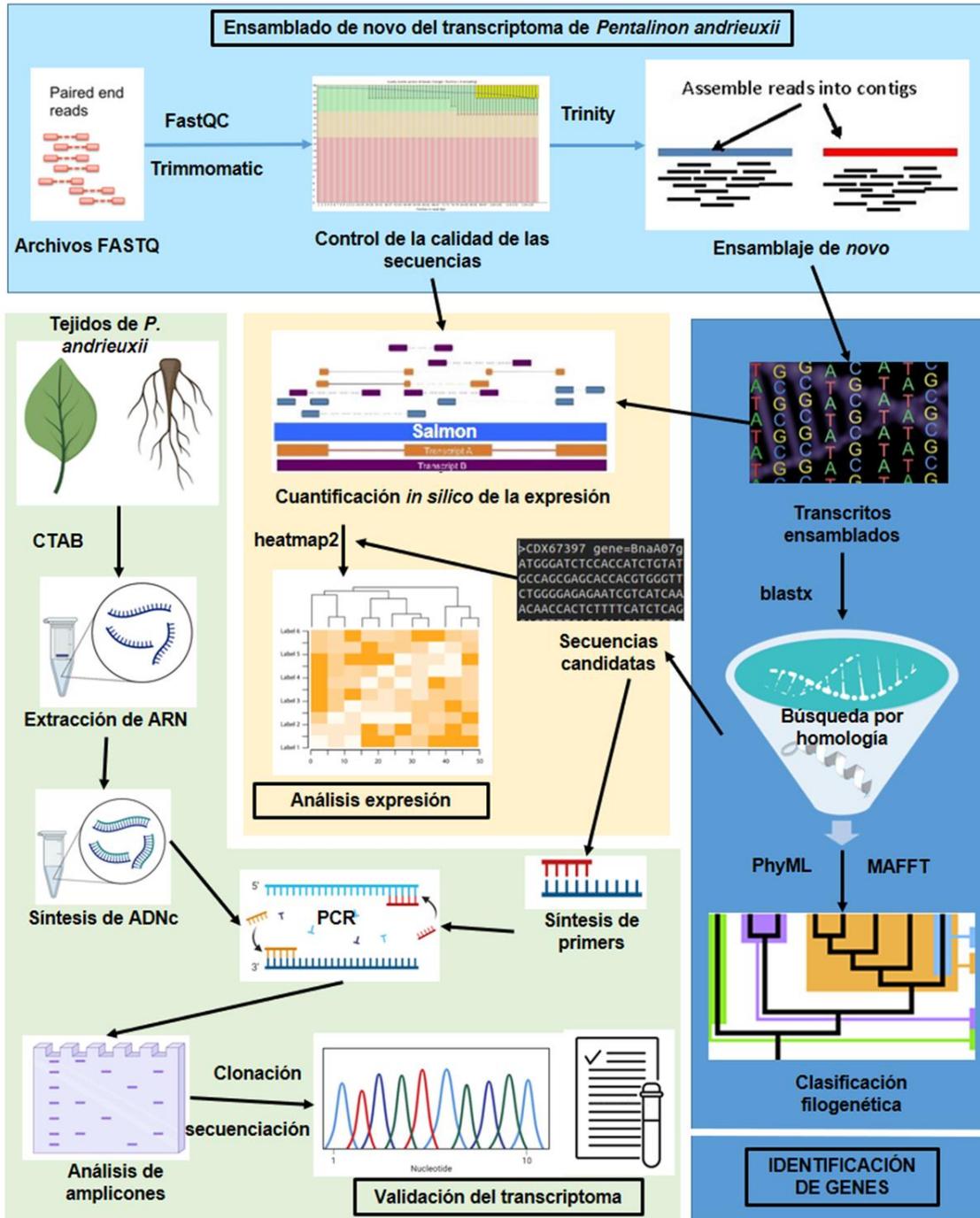


Figura 8.8 Diagrama general de la estrategia experimental

Este trabajo pretende identificar los genes principales, así como sus patrones de expresión *in silico*, implicados en la biosíntesis de isoprenoides en *Pentalinon andrieuxii*. El trabajo de desarrollo en dos etapas, en la primera, se obtendrá un transcriptoma ensamblado a partir de 4 muestras de ARN secuenciadas, dicho transcriptoma se usará para identificar mediante búsqueda por homología, transcritos ensamblados que corresponden a genes de las rutas de biosíntesis de IPP, de las familias preniltransferasas y terpenos sintasas. Posteriormente se cuantificará la expresión mediante los datos normalizados de las lecturas mapeadas a los transcritos identificados, lo que permitirá conocer que genes están expresados en cada condición. En la segunda etapa se validará el ensamblado del transcriptoma mediante la amplificación de ADNc obtenido a partir de muestras de ARN de hojas y raíces de *Pentalinon andrieuxii*, se usará primers específicos diseñados a partir de las secuencias de TPS identificadas, las cuales se pretenden clonar y secuenciar para comparar las secuencias resultantes con las secuencias obtenidas del transcriptoma.

CAPÍTULO II

MATERIALES Y MÉTODOS.

2.1. Identificación de genes involucrados en la biosíntesis de isoprenoides en el transcriptoma de *Pentalinon andrieuxii*

En el laboratorio del Dr. Gregorio Godoy se desarrolló un método de extracción de ARN por Jorge Xool (ver sección 2.4.1 Extracción de ARN y síntesis de ADNc) basado en el método descrito por Gambino *et al.*, 2008 que emplea el reactivo CTAB. El reactivo CTAB es un detergente catiónico que ayuda a liberar el contenido interno de la célula, promueve la separación de proteínas y polisacáridos de los ácidos nucleicos. Se usa en conjunto con altas concentraciones de cloruro de sodio para prevenir la formación de complejos CTAB-ácido nucleico y para crear un ambiente en el que los ácidos nucleicos puedan precipitar, pero los polisacáridos permanezcan solubles, estos últimos serán eliminados (Barbier *et al.*, 2019).

El laboratorio, además, cuenta con cuatro muestras de ARN total secuenciadas mediante la plataforma Illumina. Las muestras secuenciadas corresponden a ARN extraído de hojas y raíces de plantas silvestres jóvenes y adultas de *Pentalinon andrieuxii*, éstas últimas en floración. El material colectado se seleccionó según los criterios propuestos por Hiebert-Giesbrecht *et al.*, 2017 para etiquetar las muestras como plantas jóvenes o adultas, se determinó como un organismo joven si la planta se puede mantener erecta por sí misma y si su estatura es menor a 50 cm, mientras que la floración es la principal característica de una planta adulta. Para la secuenciación de las muestras se contrataron los servicios del CINVESTAV-Irapuato, los cuales fueron la generación de cuatro bibliotecas (una por muestra) a partir de ARN, una corrida de secuenciación en la plataforma Illumina NextSeq500, el cual tiene la capacidad de generar aproximadamente 110 Gb y alrededor de 400 Millones de lecturas pareadas, formato 2x150, y análisis de los datos. El número de lecturas obtenidas por cada muestra se incluye en el Tabla 2.1.

Tabla 2.1. Lecturas generadas por condición

Muestra	Nombre	Lecturas Totales	Bases totales
Raíz Joven	GG1TP4SS01	206,737,002	30,472,911,072
Hoja Joven	GG1TP4SS02	279,856,240	41,269,468,159
Raíz Adulta	GG1TP4SS03	240,748,812	35,288,133,876
Hoja Adulta	GG1TP4SS04	266,935,878	38,897,351,714

2.1.1 Ensamblado del transcriptoma de hojas y raíces de *Pentalinon andrieuxii*

Para ensamblar un transcriptoma, primeramente, se debe evaluar la calidad de la secuenciación y recortar las lecturas de baja calidad en caso de ser necesario. Para la evaluación de la calidad de secuenciación, los archivos en formato fastq, provenientes de la secuenciación de las muestras de *Pentalinon andrieuxii* (ver sección 1.3.4), se cargaron en la plataforma de Galaxy (<https://usegalaxy.org/>) y se les realizó una prueba de calidad con la herramienta FastQC versión Galaxy 0.72+ (Andrews, 2010). Para eliminar las lecturas de baja calidad y otros errores detectados en los gráficos de FastQC, se utilizó la herramienta Trimmomatic (Bolger et al., 2014) versión 0.38.0, con los siguientes parámetros:

ILLUMINACLIP inical: TruSeq3 (paired ended, for MiSeq and HiSeq, Crop: 130, Headcrop: 14, SLIDINGWINDOW 4:20, MINLEN:50

Los parámetros no especificados se dejaron con valores predeterminados.

Como resultado de este paso, por cada muestra (par de archivos R1, R2), se generaron cuatro archivos, sin embargo, solo se tomaron en cuenta los archivos con terminación “paired” en los siguientes pasos.

Para el ensamblaje de novo del transcriptoma se utilizó la herramienta Trinity versión galaxy 2.9.1 (Grabherr *et al.*, 2011), como entrada se seleccionaron los archivos “paired” generados por Trimmomatic, los parámetros modificados fueron:

Strand specific data: false

Run in silico normalization of reads: True

Minimum Contig Length: 200

Use the genome guided mode?: no

Como salida de esta herramienta se obtuvo un archivo de secuencias de transcritos ensamblados.

Para obtener la secuencia de aminoácidos correspondiente de cada contig ensamblado, se utilizó la herramienta TransDecoder versión Galaxy 5.5.0+ (Haas *et al.*, 2013), la cual, además, produce otro archivo con la región codificante de cada contig.

Para evaluar la integridad del ensamblado realizado con Trinity se ejecutó la herramienta BUSCO versión Galaxy 5.2.2+ (Simão *et al.*, 2015), entre los parámetros se seleccionó el modo transcriptoma con el linaje Embryophyta. Además, para comparar la versión realizada en este trabajo, se ejecutó BUSCO para evaluar los transcriptomas ensamblados de manera individual realizados por el CINVESTAV Irapuato.

2.1.2 Identificación de secuencias de transcritos relacionados con la biosíntesis de terpenos en los transcriptomas de *P. andrieuxii*.

Con el fin de facilitar el análisis de las numerosas secuencias que se esperaba obtener, el metabolismo isoprenoide se dividió en 3 módulos, el primero contempla los genes de las rutas MVA y MEP hasta la biosíntesis de IPP y DMAPP, el segundo y tercer módulo corresponden a las familias trans-preniltransferasa y terpenos sintasas, respectivamente.

La identificación de contigs correspondientes a transcritos de genes del metabolismo isoprenoide, se realizó mediante búsqueda por homología con la herramienta BLAST+ (Camacho *et al.*, 2009). El primer paso para utilizar BLAST+ es la construcción de bases de datos que contienen secuencias anotadas (subject) contra la cual nuestras secuencias de interés (query) se van a comparar, se decidió que las bases de datos serían de proteínas.

Para obtener las secuencias que conformaría cada base de datos, se realizó una búsqueda bibliográfica, se recopilaron los identificadores de las secuencias citadas en los artículos, cada identificador se utilizó como entrada en la base de datos de nucleótidos y proteínas del portal <https://www.ncbi.nlm.nih.gov/>, en el caso que el identificador recopilado pertenezca a una secuencia de nucleótidos, se seleccionó la opción “buscar secuencia relacionada en la base de datos de aminoácidos, lo que provoca que el identificador sea diferente a los citados en el artículo.

Una vez creada las tres bases de datos, se ejecutó blastx versión 2.9.0+ con el transcriptoma ensamblado con Trinity como entrada, para cada módulo se utilizó la base de datos (-db) correspondiente, se usaron los siguientes parámetros, -num_alignments 1 -evaluate 1e-10 -outfmt 6.

De cada archivo de salida, se recuperaron los identificadores de las secuencias del transcriptoma que hicieron hit a las secuencias de la base de datos y teniendo los identificadores se recuperó la secuencia original, la secuencia de aminoácidos y su cds de los archivos generados por TransDecoder.

La presencia de intrones cambia el marco de lectura de una secuencia, en especial si la herramienta usada para la traducción a aminoácidos no contempla la presencia de estos. Para predecir los exones e intrones presentes en cada secuencia, las secuencias filtradas con ayuda del paso realizado con blastx, se subieron a la plataforma online de la herramienta Augustus (Sommerfeld *et al.*, 2009) <http://bioinf.uni-greifswald.de/augustus/submission.php>, se usó el modelo de intrones-exones del organismo de referencia *Arabidopsis thaliana* y se seleccionó la opción Alternative transcript: none, como resultado se obtuvo la región codificante de cada contig sin la presencia de intrones (si los tuviera) y su correspondiente secuencia de aminoácidos.

Para eliminar la redundancia en los datos obtenidos para cada módulo, se concatenaron las secuencias de aminoácidos obtenidas con TransDecoder y Augustus y se utilizó el archivo resultante como entrada para la herramienta cd-hit (Li y Godzik, 2006) versión 4.8.1, se utilizó -c 0.95 (porcentaje de identidad) como parámetro de esta herramienta.

A las secuencias resultantes se les realizó un segundo blastx versión Galaxy 2.10.1+ ejecutado en el servidor Galaxy de Europa (<https://usegalaxy.eu/>), las bases de datos seleccionadas fueron swissprot_2018-01-22, refseq_protein_2018-01-22, el límite del valor esperado fue de 1e-10, con solo un hit por consulta y formato de salida tabular (extendido, 25 columnas). Se descartaron las

secuencias que su putativa función no correspondía con las esperadas para cada módulo, además, se calculó el porcentaje que representa el alineamiento a la longitud completa de la secuencia de la base de datos, porcentajes por debajo del 50% fueron descartados.

En caso que TransDecoder y Augustus generaran una secuencia de proteína diferente para un mismo contig, se comparaban los hits obtenidos con blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) y se observaba que el alineamiento no produzca gaps largos y que la longitud del query y el subject no difieran por muchos aminoácidos.

Finalmente, se determinó la posible ubicación de las secuencias obtenidas mediante el programa online Wolf-PSORT (Horton *et al.*, 2007 <https://wolfpsort.hgc.jp/>).

2.2 Clasificación filogenética de secuencias

La clasificación de las secuencias de posibles genes DXS, terpenos sintasas y preniltransferasas obtenidas, se realizó en base a la rama del árbol filogenético en la cual quedaron ubicadas. En el servidor online del MAFFT (Kato y Standley, 2013) <https://mafft.cbrc.jp/alignment/server/> se realizó el alineamiento múltiple de las secuencias recuperadas de la bibliografía y las identificadas en los transcriptomas de *P. andrieuxii*, se usó los ajustes predeterminados del programa, el alineamiento se descargó en formato Phylip, este último archivo se cargó en el servidor de PhyML 3.0 (<http://www.atgc-montpellier.fr/phyml/>) (Guindon *et al.*, 2005; Guindon *et al.*, 2010) conservando los parámetros por default.

2.3 Análisis de expresión *in silico* de genes identificados.

Para la cuantificación de la expresión *in silico*, se utilizó la herramienta Salmon (Patro *et al.*, 2017), esta herramienta asigna cada lectura al probable punto de origen de la lectura con respecto a la referencia mediante un mapeo ligero sin crear índices, seguidamente estima los niveles de expresión iniciales que posteriormente son refinados en una tercera fase. La versión disponible en la plataforma Galaxy se llama salmon quant (versión 1.5.1) y realiza estimaciones de la abundancia de transcripciones basadas en mapeo o lecturas de fase dual a partir de lecturas de RNA-seq, se seleccionó el transcriptoma ensamblado con Trinity como transcriptoma de referencia al cual se va a mapear los reads, los cuales son los archivos “paired” generados por la herramienta Trimmomatic descritas en la sección 2.1.1, se seleccionó la opción paired end y se activó la opción `-validateMappings`, los demás parámetros se quedaron con valores

predeterminados. Salmon presenta dentro de los resultados el número de lecturas que se mapearon a un contig en específico, para normalizar estos valores y se puedan comparar entre muestras, transforma dichos valores a TPM mediante la siguiente fórmula:

$$TPM = 10^6 * \frac{\text{lecturas mapeadas a una transcripción / largo de de la transcripción}}{\text{sum(lecturas mapeadas a una transcripción / largo de de la transcripción)}}$$

Esta herramienta produjo cuatro archivos en formato tabular con los resultados de cada conjunto de datos (uno por muestra secuenciada), se recuperaron las filas que corresponden a los contigs identificados para cada módulo. Se generó un archivo con cinco columnas, en la primera con el identificador de cada contig y las cuatro columnas restantes corresponden a los valores TPM de cada contig en una muestra específica.

Los resultados de la cuantificación de las secuencias identificadas se representaron en un mapa de calor generado con la herramienta de Galaxy heatmap2 versión 3.0.1 con los parámetros:

-Data transformation: Log2(value+1), -Clustering columns and rows: Cluster columns and not rows, -Data scaling: Do not scale my data

2.4 Validación del ensamblado del transcriptoma mediante el aislamiento y secuenciación de transcritos.

Esta actividad se realizó adicionalmente a los objetivos presentados.

Para comprobar el ensamblado de los transcriptomas y la metodología usada para identificar las secuencias de terpenos sintasas, se amplificaron ADN correspondiente a ADNc de transcritos de genes terpenos sintasa de *Pentalinon andrieuxii* mediante PCR.

2.4.1 Extracción de ARN y síntesis de ADNc

Se colectaron tejidos de hoja y raíz de plantas jóvenes (6 meses de germinación) de *Pentalinon andrieuxii*, se lavaron con agua destilada estéril y se metieron inmediatamente en nitrógeno líquido, el peso seco del tejido colectado fue pesado y el ARN se extrajo mediante el protocolo de Xool, empleado para la obtención del ARNm y enviado para la secuenciación de los transcriptomas (datos no publicados), los pasos de la extracción se presentan a continuación:

- I. El material colectado se maceró en morteros con nitrógeno líquido y se transfieren hasta 0.2 g en cada tubo eppendorf. A cada tubo se le adicionaron 700 µl de

-
- amortiguador con CTAB con β -mercaptoetanol al 2% y se incubaron a 65°C por 10 min.
- II. Se les adicionó 500 μ l de cloroformo:alcohol isoamilico (24:1), las muestras se mezclaron perfectamente y se dejaron en reposo por 3 min a temperatura ambiente, posteriormente se centrifugaron los tubos a 12,000 x g a 4°C por 10 min. El sobrenadante de cada tubo se transfirió a tubos Eppendorf nuevos
 - III. Se repite el paso 2 nuevamente con el sobrenadante recuperado.
 - IV. Se agregó a cada tubo 1 volumen de LiCl 8 M y se mezcló perfectamente. La mezcla anterior se dejó precipitando toda la noche a 4°C. Al día siguiente se centrifugaron los tubos a 15,000 x g por 30 min a 4°C.
 - V. Se desechó el sobrenadante, se deben ver un sedimento blanco al fondo de los tubos, a la cuales se les adicionó 500 μ l de amortiguador SSTE y se incubaron a 65°C por 10 min
 - VI. Se le añadió 500 μ l de cloroformo:alcohol isoamilico (24:1), la mezcla anterior se homogenizó y se dejó en reposo por 3 min a temperatura ambiente, posteriormente se centrifugaron los tubos a 12,000 x g por 10 min a 4°C
 - VII. Se recuperó el sobrenadante y se le adicionó un volumen de isopropanol, la mezcla se homogenizó y se dejó precipitar a 20°C por 2 horas. Una vez transcurrido el tiempo, los tubos se centrifugaron a 15,000 x g por 30 min a 4°C.
 - VIII. Las pastillas precipitadas se lavaron con 1 ml de etanol grado biología molecular al 70% dos veces y una tercera vez con etanol al 100%. Después de haber lavado las pastillas se dejaron secar. Finalmente, se resuspendieron en 30 μ l de agua desionizada con DEPC.

Para obtener mayor concentración de ARN, se mezcló el contenido de varios tubos que pertenezcan a la misma muestra, debido a que se hará una segunda purificación con un kit y todas las muestras tendrán una concentración menos por el ADN eliminado y ARN perdido.

Para eliminar el ADN genómico se utilizó el paquete DNase I, RNase-free (Código de catálogo EN0521), se utilizó 8 unidades de enzima, 1x del buffer de reacción con $MgCl_2$, el volumen final depende de la cantidad de muestras juntas, no se agregó agua, se incubaron las muestras a 37° por 30 min.

Para purificar el ARN se utilizó el paquete PureLink™ RNA Mini Kit (Código de catálogo 12183020), el volumen de la mezcla anterior se completó a 300 μ l con la solución de lisis, se

homogenizó y se cargó cada muestra en una columna del kit, los siguientes pasos se realizaron de acuerdo con las especificaciones del fabricante. El ARN se eluyó de la columna con 40 µl de buffer de elución, finalmente se cuantificó cada muestra.

Para la síntesis de ADNc se empleó el producto RevertAid H Minus First Strand cDNA Synthesis Kit (Código de catálogo K1632). Las condiciones particulares en la síntesis de ADNc fueron 1 µg de ARN purificado, 1 µl de oligo dT, los demás componentes se agregaron de siguiendo las instrucciones del fabricante. Se diluyó cada muestra a 500 ng/µl.

2.4.2 Amplificación por PCR punto final de los terpenos sintasa

Se diseñaron primers de los transcritos de terpenos sintasa identificados, la región codificante de cada secuencia se utilizó como entrada para la plataforma de Primer3Plus (<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>), se seleccionó el parámetro clonning para diseñar primer que amplifiquen la longitud completa de la región codificante.

Para amplificar cada TPS, se utilizó el ADNc sintetizado del tejido donde se encontró con mayor expresión según datos *in silico*. La mezcla para la PCR se preparó con 1x de DreamTaq 10X Buffer, 0.2 mM de cada DNTP, 0.4 mM de cada cebador, 1 µl de DNAC, 1.5 unidades de DreamTaq DNA Polymerase (Código de catálogo EP0701) y agua hasta alcanzar un volumen de 20 µl. Las condiciones de PCR se describen en la tabla 2.3.

Tabla 2.2. Condiciones usadas para la PCR.

Fase	Temperatura (°c)	Tiempo (minutos)	Ciclos
Desnaturalización	95	4:00	1
Desnaturalización	95	0:30	30
Alineamiento	Específica de cada primer (ver anexo IV)	0:30	
Extensión	72	1:00 por cada 1000 pb	
Extensión	72	5	1

Los productos de PCR fueron inyectados en un gel de 40 ml de agarosa al 1% con 0.5 µl de SYBR safe y se corrieron en una cámara electroforética a 90 V por 35 min. La visualización se realizó través de un transiluminador

2.4.3 Clonación

Para clonar la tps7, primeramente, se preparó una mezcla para PCR para un volumen de 40 µl con primers específicos para esta secuencia, el programa del termociclador se describe en la sección anterior. El total de la reacción se corrió en un gel de electroforesis y se visualizó en un transiluminador, seguidamente la banda que corresponde al producto de PCR de la tps7 se cortó y se depositó en un tubo Eppendorf previamente pesado. Para purificar el fragmento se utilizó el kit Zymoclean Gel DNA Recovery, los pasos se describen a continuación:

- I. Se agregó 1 volumen (peso:volumen) del bufer ADB al tubo donde se encuentra la muestra y seguidamente se calentó a 55°C por 10 minutos o hasta que esté completamente disuelto el gel de agarosa.
- II. Se transfirió toda la mezcla en la columna insertada en el tubo de recolección provistos por el kit, se centrifuga 60 segundos y se descarta el líquido del tubo de recolección.
- III. Se le agrega a la columna 200 µl del bufer ADB y se centrifuga nuevamente.
- IV. Se lava la columna 3 veces con 200 µl del buffer de lavado
- V. finalmente, se pasa la columna a un tubo Eppendorf nuevo y se le agrega 12 µl de agua GBM, se deja reposar por 2 minutos y se centrifuga 60 segundos. La muestra se cuantifica en el nanodrop.

Para la reacción de ligación se usó el kit pGEM®-T Easy, se preparó la reacción con 5 µl de 2X Rapid Ligation Buffer, 1 µl pGEM®-T Easy Vector, producto purificado a una relación molar 3:1 inserto:vector, 1 µl T4 DNA Ligase, agua hasta alcanzar los 10 µl de volumen final, se incubó la muestra a 22°C por una hora, posteriormente se dejó toda la noche en el refrigerador.

Para transformar las células competentes, se usó competentes comerciales de Escherichia coli JM109 que vienen en conjunto con el kit de pGEM®-T Easy, se usó el protocolo de la sección 4.A. del manual del fabricante.

Las colonias crecidas se picaron con un palillo estéril y se inocularon en 5 ml de medio LB con ampicilina, cada muestra se pasó a tubos Eppendorf nuevos y se centrifugo 1.5 ml a la vez. Para purificar el ADN plasmídico, se usó el kit ZymoPURE Plasmid Miniprep Kit (cat D4209), siguiendo

el protocolo del proveedor y usando agua GBM en lugar del bufer de elusión, esto como requisito para mandar la muestra a secuenciar, para lo cual se contrató los servicios de MacroGen (<https://dna.macrogen.com/#>), las muestras enviadas se prepararon con 5 µl del ADN plasmidico a una concentración de 50ng/µl más 5 µl a 5 nmol/µl del primer forward o reverse.

Para comprobar que la clonación fue exitosa, se realizó una digestión enzimática con las enzimas HindIII (cat. FD050) y Sall (cat. FD0644), que se eligieron previamente ingresando la cds de la secuencia de tps7 en el portal del programa restrictionMapper (<http://www.restrictionmapper.org/>), igualmente se amplificó la secuencia por PCR, finalmente los resultados se corrieron por electroforesis, se visualizaron y se analizaron.

CAPÍTULO III

RESULTADOS

3.1 Ensamblado de novo del transcriptoma de *Pentalinon andrieuxii*

Con la finalidad de eliminar los errores de secuenciación, se utilizó el programa FastQC para evaluar la calidad de la secuenciación. Se observó que el módulo de “Calidad de secuencia por base”, el cual muestra una descripción general del rango de valores de calidad Phred (Q) de todas las bases en cada posición en el archivo FastQ presentaba advertencias en todas las muestras secuenciadas. La herramienta FastQC emite advertencias en este módulo si los valores de calidad Q del cuartil inferior de cualquier base es menor que 10, o si la mediana de cualquier base es menor que 25. El valor de calidad Phred se traduce directamente en la probabilidad de que la llamada de base correspondiente sea incorrecta, una puntuación Q de 30 (Q30) a una base, es equivalente a la probabilidad de que exista una llamada de base incorrecta cada 1000 bases secuenciadas. La figura 3.1 presenta un gráfico o diagrama de cajas y bigotes (BoxWhisker) con los resultados para el módulo anteriormente mencionado del archivo de fastq de la muestra de hoja joven, en las demás muestras, se observa un comportamiento similar y se podrá encontrar sus respectivos gráficos en el anexo I. En los gráficos de cajas y bigotes, en cada posición del eje x, la línea roja central representa el valor de la mediana, el cuadro amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos del 10% y el 90%, mientras que la línea azul representa la calidad media. En la figura 3.1 se puede observar que todos los indicadores de las primeras 70 bases, representados por el eje x, se encuentran en la zona donde la calidad es considerada buena, es decir, con valores Phred de 28 a 36 (zona verde) en el eje de la y, sin embargo, en el resto de las posiciones, los bigotes inferiores se encuentran en la zona de calidad media (valores Q de 20 a 28) y mala ($Q < 20$), en las últimas 20 bases, incluso los cuadros amarillos se encuentran en la zona de calidad media llegando a la zona de calidad baja en la última posición. Antes de continuar al ensamblaje del transcriptoma, se deben eliminar las bases que tengan valores Q menores a 20.

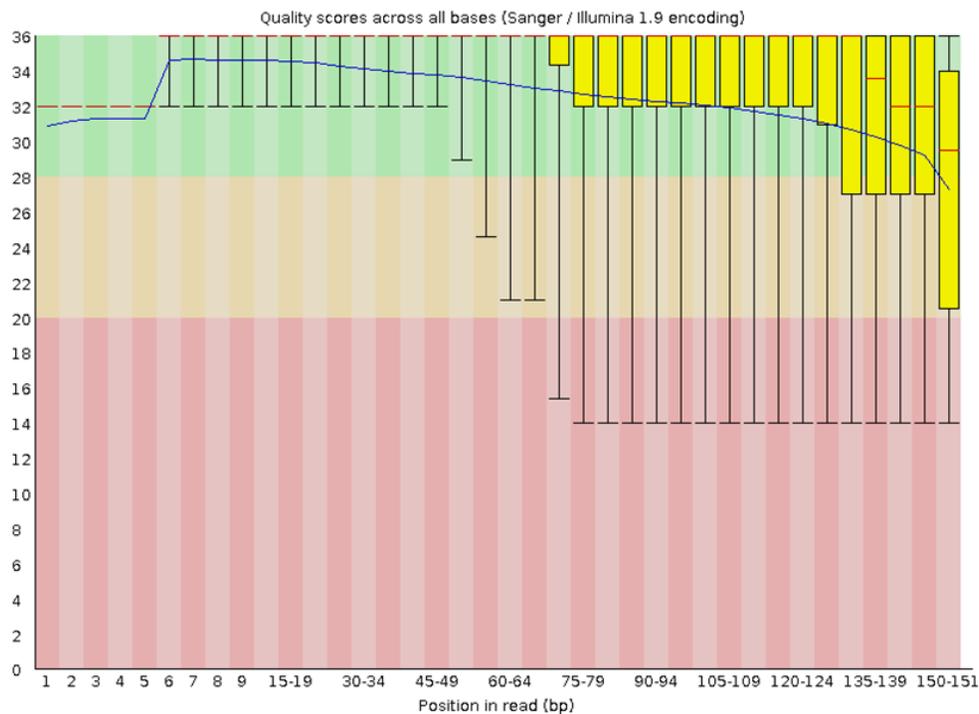


Figura 3.1. Distribución de la calidad de las secuencias obtenidas de la plataforma Illumina para el tejido de raíz joven de *P. andrieuxii* calculado con el programa FastQC. El eje x corresponde a los pares de bases de las secuencias, en el eje y se muestra los puntajes de calidad Phred. La línea roja central representa el valor de la mediana, el cuadro amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos del 10% y el 90%, mientras que la línea azul representa la calidad media.

Para corregir los errores observados en la figura 3.1, se decidió recortar las secuencias hasta la posición 145. Se utilizó la función slidingwindow 4:20 de la herramienta Trimmomatic para eliminar lecturas con calidad $Q < 20$ que no se hayan conservado tras el paso anterior. Además, se recortaron las primeras 15 bases de todas las secuencias debido a que el porcentaje de cada base en esas posiciones contrastaba con el porcentaje del resto de las secuencias, este comportamiento se puede observar en la figura 3.2.

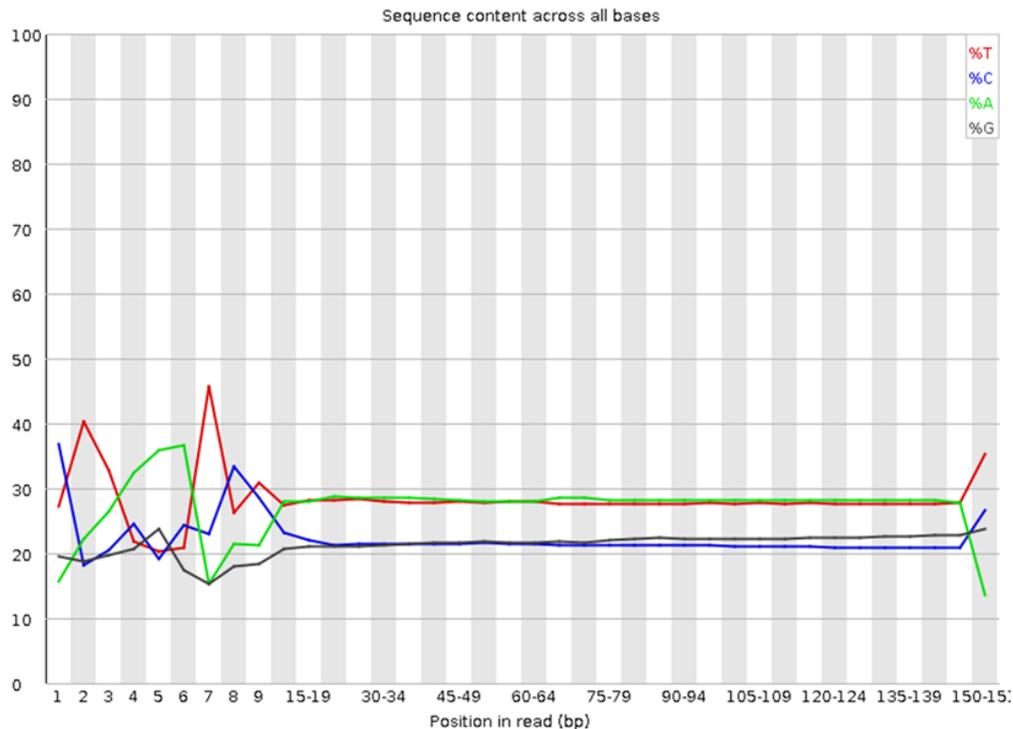


Figura 3.2. Contenido por base obtenidas de la plataforma Illumina para el tejido de raíz joven de *P. andrieuxii* calculado con el programa FastQC. En la figura se muestra el porcentaje de cada base en cada posición de las lecturas secuenciadas.

En el módulo “secuencias sobrerrepresentadas” solo se detectaron adaptadores de Illumina o largas “N”, las cuales indican lecturas de mala calidad, para eliminar los adaptadores, se eligió aplicar el paso inicial ILLUMINACLIP de Trimmomatic.

Por último, se aplicó el filtro “minlen: 50” para que únicamente las secuencias mayores o iguales a 50 nucleótidos se conserven.

Una vez ejecutado Trimmomatic, cada archivo obtuvo dos salidas denominadas paired y unpaired, en el primer caso se conservaron las lecturas que cumplieron los requisitos previamente mencionados y que, además, se conservó su par en el archivo fastq complementario, en el segundo caso, las lecturas pasaron los filtros de calidad, pero su par respectivo fue descartado, por consiguiente, únicamente se trabajó con los archivos paired. En la tabla 3.1 se presenta el número de secuencias que se mantuvieron luego de ejecutar Trimmomatic con los parámetros mencionados en la sección 2.1.

Tabla 3.1. Lecturas antes y después del filtrado de calidad con Trimmomatic.

Muestra	Secuencias crudas	Filtro de calidad aplicado	% de recuperación
	Reads R1	Reads R1_paired	
Raíz joven	103,368,501	80,136,798	77.525
Hoja joven	139,928,120	109,352,667	78.462
Raíz adulta	120,374,406	93,208,478	77.432
Hoja adulta	133,467,939	106,536,604	79.821

Para determinar si los filtros de calidad aplicados con la herramienta Trimmomatic fueron suficientes para eliminar las lecturas de mala calidad observados en la figura 3.1, se evaluó la calidad de los datos filtrados haciendo uso del programa FastQC. Como resultado se presenta el gráfico obtenido para el módulo “Calidad de secuencia por base” (figura 3.3), en la que se puede observar que tanto la media, la mediana y las barras amarillas (representa los valores que abarca desde el primer cuartil hasta el tercero de los datos ordenados por calidad en orden ascendente) se encuentran en la zona de calidad buena ($Q > 28$), solo la parte inferior de los valores representados por las barras negras (rango del 10% al 90% de los valores de calidad ordenados de forma ascendente) se encuentran en la parte más alta de la zona de la calidad media ($28 > Q > 20$), por lo tanto se consideró adecuados los filtros de calidad aplicados. Como en el caso anterior, solo se muestra el gráfico de un conjunto de datos, los demás son similares y de adjuntan en la sección de anexos I.

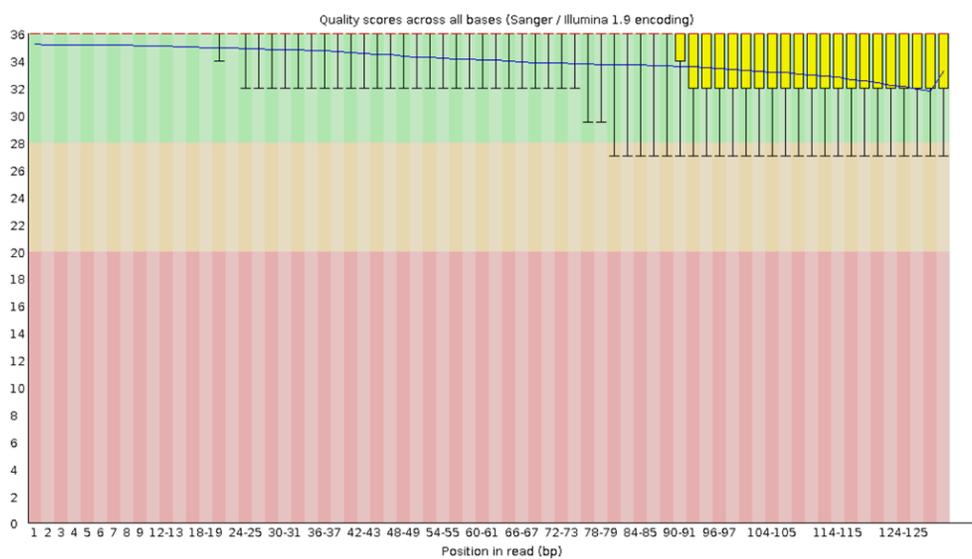


Figura 3.3. Distribución de la calidad de las secuencias filtradas con la herramienta Trimmomatic para el tejido de raíz joven de *P. andrieuxii* calculado con el programa FastQC. El eje x corresponde a los pares de bases de las secuencias, en el eje y se muestra los puntajes de calidad Phred. La línea roja central representa el valor de la mediana, el cuadro amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos del 10% y el 90%, mientras que la línea azul representa la calidad media.

El comportamiento observado en la figura 3.2 ya no estuvo presente en los reads procesados, tampoco se encontraron adaptadores ni secuencias sobrerrepresentadas, por lo tanto, se continuó con el ensamblado *de novo* del transcriptoma con la herramienta Trinity.

Una vez ensamblado el transcriptoma se procedió a calcular sus estadísticas mediante el comando `perl Trinity_Stats.pl`. La tabla 3.2 muestra las estadísticas del ensamblado. La herramienta Trinity denomina a un gen como un grupo de transcritos relacionados que comparten segmentos de una secuencia, a cada variante del gen se le denomina isoforma, al tomar en cuenta únicamente la isoforma más larga por gen las estadísticas cambian notablemente debido a que el programa Trinity genera demasiadas isoformas, especialmente para las transcripciones más largas. La medida N50 indica que el 50% de los nucleótidos ensamblados se encuentran en contigs con una longitud de al menos el valor resultante para N50. Los valores de N50 se basan únicamente en la longitud de los contigs y no toman en cuenta la abundancia de las mismas, por lo tanto, se calculó el valor de E50N50, esta medida se obtiene de manera similar al valor N50

pero se limita a los genes más expresados que representan el 50% de los datos de expresión normalizados totales.

Tabla 3.2. Estadísticas del transcriptoma ensamblado.

Estadísticas del transcriptoma ensamblado

Total de genes	160909
Total de transcritos	282627
Porcentaje de GC	40.26

Estadísticas basadas en todas las transcripciones

Contig N50	3094
Longitud mediana del contig	500
Promedio de longitud del contig	1378.65
Total de bases ensambladas	389643597

Estadísticas basadas únicamente en la isoforma más larga por gen

Contig N50	1472
Longitud mediana del contig	285
Promedio de longitud del contig	650.79
Total de bases ensambladas	104718634

Estadísticas basadas en las transcripciones más expresadas

E50N50	2026
--------	------

Para analizar la integridad del transcriptoma ensamblado se utilizó la herramienta BUSCO (Benchmarking Universal Single-Copy Orthologues), la cual analiza la integridad de los ensamblajes indicando el número total de ortólogos conservados dentro del linaje Embryophyta presentes en el transcriptoma, en este linaje se consideran 1614 genes (columna 6 de la tabla 3.3). Se define como gen “completo” cuando el tamaño del contig corresponde al tamaño esperado para ese gen dentro del linaje seleccionado, los contigs menores a la longitud esperada se clasifican como “fragmentados” y los genes faltantes se consideran como “Faltantes”. Debido a que se espera que ortólogos considerados por la herramienta BUSCO se encuentren como copia única, las isoformas pudieran ser clasificados como genes “Duplicados”.

Los valores obtenidos por la herramienta BUSCO para evaluar el ensamblaje del transcriptoma de *Pentalinon andrieuxii* se muestran en la tabla 3.3, los datos indican que el 99% de los ortólogos lograron ensamblarse de manera completa, sin embargo, el 92% se encuentra duplicado.

Tabla 3.3. Resultado de la herramienta BUSCO para el transcriptoma de *Pentalinon andrieuxii* ensamblado.

Completos C	Completos de copia única S	Completos duplicados D	Fragmentados F	Faltantes M	Número de genes utilizados
1598	112	1486	6	10	1614

3.2 Identificación de secuencias de transcritos relacionados con la biosíntesis de terpenos en los transcriptomas de *P. andrieuxii*

Con el objetivo de crear bases de datos específicas para cada módulo de la biosíntesis de isoprenoides, que serán usados por la herramienta blastx para identificar homólogos en las secuencias del transcriptoma ensamblado, se realizó una búsqueda bibliográfica para recuperar secuencias que hayan sido caracterizadas experimentalmente, así evitar errores derivados de la anotación automática de secuencias. El número de secuencias incluidas en las bases de datos se muestra en la tabla 3.4.

Tabla 3.4. Número de secuencias de cada base de datos usadas para la identificación de secuencias con blastx.

Módulo	Secuencias no redundantes*	Referencias
MVA y MEP	157	Hemmerlin <i>et al.</i> , 2012 (tablas 2, 3 y 4)
PT	118	Zhou <i>et al.</i> 2017; Wang y Dixon, 2009; You <i>et al.</i> , 2020
TPS	492	Chen <i>et al.</i> , 2011; Durairaj <i>et al.</i> , 2019

Como resultado de la aplicación de la metodología descrita en la sección 2.2, se logró identificar en total 42 secuencias que podrían ser de transcritos de genes que participen en la biosíntesis de isoprenoides. En la tabla 3.5 se resume los resultados que la metodología diseñada generó en cada paso.

Tabla 3.5. Resultado de la metodología aplicada para la identificación de genes relacionados con la biosíntesis de isoprenoides.

Módulo	Primer filtro			Segundo filtro			
	blastx	Trans Decoder	Augustus	Secuencias nr	Alineamiento > 50%	Función correcta	secuencias finales
MVA y MEP	251	84	81	104	72	44	19
PT	56	31	20	31	27	19	10
TPS	159	42	38	55	43	41	13
Total	466	157	139	190	142	104	42

La diferencia entre el número de secuencias iniciales y las secuencias finales se ve afectado principalmente por segmentos de la secuencia que son reconocidos como intrones por el programa Augustus, sin embargo, al no contar con el genoma de *P. andrieuxii* o un modelo cercano, usar un modelo de predicción basado en *Arabidopsis thaliana*, puede afectar la secuencia resultante. Por su parte TransDecoder al no poder reconocer dichas secuencias, seguirá traduciendo hasta encontrar un codón de paro. En comparativa, tal como se observa en

la tabla 3.6, ninguna herramienta por sí sola pudo generar el total de las secuencias completas, ambas, en conjunto con cd-hit permitieron obtener mayor información sobre las secuencias.

Tabla 3.6. Comparación de herramientas de traducción de secuencias.

Módulo	Secuencias finales	TransDecoder	Augustus	secuencias completas
MVA y MEP	19	15	16	19
PT	10	8	8	10
TPS*	13	8	7	11
Total	42	31	31	40

Para anotar la posible función de las secuencias identificadas, se realizó nuevamente una búsqueda por homología mediante la plataforma online blastp del NCBI. Como resultado, se presentan en las tablas 3.7, 3.8 y 3.9, el código asignado a cada secuencia, su identificador dado por Trinity, su longitud en aminoácidos, así como las principales características de la secuencia con la cual tuvieron su mejor alineamiento. Los datos obtenidos indican que se pudo identificar al menos una isoforma completa de todos los genes de la ruta MVA y MEP, diez preniltransferasas y once terpenos sintasas.

La secuencia del contig nombrado terpeno sintasa 5 (tps5) generó 2 clusters que comparten un porcentaje de similaridad del 87.2%, aun así, tienen la misma longitud, por tal motivo al considerarlas variantes se les definió como “a” y “b”, en cuanto a la tps5x comparte un segmento con las otras tps5, sin embargo, en el inicio se le detectó un péptido señal al cloroplasto y a las demás no.

La secuencia codificante (cds) de los contig que corresponden a genes DXS, HMGR y TPS identificados se encuentran se pueden consultar en el anexo V.

Tabla 3.7. Secuencias identificadas de la ruta del mevalonato.

Código: Identificador	qlen ¹	Descripción - Organismo	Accesión	slen	piden ²	Evalue
AACT1: TRINITY_DN1093_c0_g1_i101:g37.t1	415	acetyl-CoA acetyltransferase - <i>Arabidopsis thaliana</i>	Q9FIK7.1	416	80	0
AACT2: TRINITY_DN11351_c0_g1_i14.p1	406	acetyl-CoA acetyltransferase - <i>Arabidopsis thaliana</i>	Q9FIK7.1	415	80	0
HMGS: TRINITY_DN14264_c0_g1_i6.p1	464	Hydroxymethylglutaryl-CoA synthase - <i>Arabidopsis thaliana</i>	P54873.2	461	82	0
HMGR1: TRINITY_DN4167_c0_g1_i3.p1	577	3-hydroxy-3-methylglutaryl coenzyme A reductase 1 - <i>Panax ginseng</i>	A0A0A1C3I2.1	573	78	0
HMGR2: TRINITY_DN249_c0_g1_i3:g193.t1	599	3-hydroxy-3-methylglutaryl-coenzyme A reductase - <i>Catharanthus roseus</i>	Q03163.1	601	85	0
MK: TRINITY_DN1601_c0_g1_i5.p1	388	Mevalonate kinase - <i>Arabidopsis thaliana</i>	P46086.1	378	68	0
PMK: TRINITY_DN16663_c0_g1_i21.p1	498	Phosphomevalonate kinase - <i>Arabidopsis thaliana</i>	Q9C6T1.1	505	71	0
PPMD: TRINITY_DN6581_c0_g1_i3.p1	419	Diphosphomevalonate decarboxylase, peroxisomal - <i>Arabidopsis thaliana</i>	F4JCU3.1	419	80	0
IDI: TRINITY_DN4092_c0_g1_i8.p1	349	Isopentenyl-diphosphate Delta-isomerase II - <i>Camptotheca acuminata</i>	O48965.1	309	82	5E-156

¹ qlen: Longitud de la secuencia de consulta

² pident: Porcentaje de identidad entre la secuencia de consulta y la proteína más cercana

Tabla 3.8. Secuencias identificadas de la ruta MEP.

Código: Identificador	qlen	Descripción	Organismo	Accesión	piden	slen	Evalue
DXS1: TRINITY_DN7236_c0_g1_i1.p1	729	1-deoxy-D-xylulose-5-phosphate synthase 2	<i>Oryza sativa</i>	Q6YU51.1	79	713	0
DXS2: TRINITY_DN14633_c0_g2_i7.p1	718	1-deoxy-D-xylulose-5-phosphate synthase	<i>Capsicum annuum</i>	O78328.1	88	719	0
DXS3: TRINITY_DN272_c0_g1_i14.p1	715	1-deoxy-D-xylulose-5-phosphate synthase 2	<i>Oryza sativa</i>	Q6YU51.1	82	713	0
DXS4: TRINITY_DN4693_c0_g1_i28:g4.t1	713	1-deoxy-D-xylulose-5-phosphate synthase 1	<i>Oryza sativa</i>	O22567.2	59	720	0
DXR: TRINITY_DN2680_c0_g1_i63.p1	478	1-deoxy-D-xylulose 5-phosphate reductoisomerase	<i>Mentha x piperita</i>	Q9XES0.2	85	470	0
MCT: TRINITY_DN3403_c1_g1_i17.p1	320	2-C-methyl-D-erythritol 4-phosphate cytidyltransferase	<i>Arabidopsis thaliana</i>	P69834.1	76	302	8E-128
CMK: TRINITY_DN12871_c0_g1_i3.p1	407	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	<i>Solanum lycopersicum</i>	P93841.1	80	401	0
MDS: TRINITY_DN1746_c0_g1_i10.p2	239	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	<i>Catharanthus roseus</i>	Q9M4W3.1	85	236	2E-132
HDS: TRINITY_DN2095_c0_g1_i55.p1	741	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (ferredoxin)	<i>Arabidopsis thaliana</i>	F4K0E8.1	85	741	0
HDR: TRINITY_DN11792_c0_g1_i20.p1	466	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	<i>Arabidopsis thaliana</i>	Q94B35.1	76	466	0

Tabla 3.9. Secuencias identificadas de la familia trans-preniltransferasa.

Código: Identificador	qlen	Loc ³	Descripción- Organismo	Accesión	piden	slen	Evalue
PT1:TRINITY_DN11006_c0_g1_i1	299	Cit	Geranylgeranyl pyrophosphate synthase, chloroplastic - <i>Hevea brasiliensis</i>	Q94ID7.1	33	370	4E-40
PT2:TRINITY_DN12179_c0_g1_i1	377	Mit	Geranylgeranyl diphosphate synthase, chloroplastic - <i>Leucoscepttrum canum</i>	A0A0U3BRC5	63	364	4E-150
PT3:TRINITY_DN148217_c0_g1_i1	385	Clo	Geranylgeranyl pyrophosphate synthase, chloroplastic - <i>Capsicum annuum</i>	P80042.1	69	369	0
PT4:TRINITY_DN2128_c1_g1_i1	349	Cit	Farnesyl pyrophosphate synthase 1 - <i>Lupinus albus</i>	P49351.1	82	342	0
PT5:TRINITY_DN2963_c1_g1_i4	422	Clo	Solanesyl diphosphate synthase 1, chloroplastic - <i>Arabidopsis thaliana</i>	Q8S948.1	71	406	0
PT6:TRINITY_DN4834_c0_g1_i1	343	Cit	Farnesyl pyrophosphate synthase - <i>Panax ginseng</i>	Q4JHN6.1	84	342	0
PT7:TRINITY_DN5099_c0_g1_i1	351	-	Geranylgeranyl pyrophosphate synthase, chloroplastic - <i>Hevea brasiliensis</i>	Q94ID7.1	69	370	8E-148
PT8:TRINITY_DN5760_c0_g1_i1	210	-	Geranylgeranyl pyrophosphate synthase SSU, chloroplastic- <i>Arabidopsis thaliana</i>	Q39108.2	63	326	2E-81
PT9:TRINITY_DN6650_c0_g1_i22	422	Clo	Solanesyl diphosphate synthase 3, chloroplastic/mitochondrial - <i>Arabidopsis thaliana</i>	Q5HZ00.1	67	422	0
PT10:TRINITY_DN8686_c0_g1_i1	381	Clo	Geranylgeranyl pyrophosphate synthase, chloroplastic <i>Catharanthus roseus</i>	Q42698.1	86	357	0

³ Loc: Predicción de la localización, según WoLF-PSORT (Cit: citoplasma, Mit: mitocondria, Clo: cloroplasto).

Tabla 3.10. Secuencias identificadas de la familia terpeno sintasa.

Código: Identificador	qlen	cTP ⁴	Descripción	Organismo	Accesión	pident	slen	Evalue
tps1: TRINITY_DN17403_c0_g1_i1	382	-	Tricyclene synthase EBOS, chloroplastic	<i>Lotus japonicus</i>	Q672F7.1	51	595	1E-144
tps2: TRINITY_DN119011_c0_g1_i1	568	-	(-)-germacrene D synthase	<i>Vitis vinifera</i>	Q6Q3H3.1	56	557	0
tps3: TRINITY_DN25972_c0_g1_i13	566	Si	(3S,6E)-nerolidol synthase 1, chloroplastic	<i>Fragaria vesca</i>	P0CV96.1	50	580	0
tps4: TRINITY_DN590_c0_g1:g1	772	Si	Ent-kaurene synthase TSP4, chloroplastic	<i>Vitex agnus-castus</i>	A0A2K9RFY0.1	54	789	0
tps5x: TRINITY_DN6190_c0_g1:g1	200	Si	(-)-alpha-terpineol synthase	<i>Vitis vinifera</i>	Q6PWU2.1	47	590	2E-34
tps5a: TRINITY_DN127_c0_g1_i43	608	-	(-)-alpha-terpineol synthase	<i>Vitis vinifera</i>	Q6PWU2.1	57	590	0
tps5b: TRINITY_DN127_c0_g1_i62	608	-	(-)-alpha-terpineol synthase	<i>Vitis vinifera</i>	Q6PWU2.1	57	590	0
tps6: TRINITY_DN15212_c0_g1_i2	834	Si	(E,E)-geranylinalool synthase	<i>Arabidopsis thaliana</i>	Q93YV0.1	46	877	0

⁴ cTP: Indica si la secuencia tiene péptido de tránsito al cloroplasto según la herramienta ChloroP 1.1 (<http://www.cbs.dtu.dk/services/ChloroP/>)

tps7: TRINITY_DN2222_c0_g1_i8	566	-	(-)-germacrene D synthase	<i>Vitis vinifera</i>	Q6Q3H3.1	56	557	0
tps8: TRINITY_DN4534_c0_g1_i9	599	-	Vetispiradiene synthase 1	<i>Solanum tuberosum</i>	Q9XJ32.1	49	566	0
tps9: TRINITY_DN10529_c0_g1_i9	595	Si	Alpha-terpineol synthase, chloroplastic	<i>Magnolia grandiflora</i>	B3TPQ7.1	44	592	2E-164
tps10: TRINITY_DN2434_c0_g1_i29	562	-	(-)-germacrene D synthase	<i>Vitis vinifera</i>	Q6Q3H3.1	54	557	0
tps11: TRINITY_DN4739_c0_g1_i10	825	Si	Ent-copalyl diphosphate synthase, chloroplastic	<i>Salvia miltiorrhiza</i>	A0A0U3L Q20.1	62	793	0

3.2.1 Clasificación filogenética de las secuencias identificadas

Debido a que la clasificación de las secuencias de la familia de las HMGR no aporta mayor información sobre su posible participación en el metabolismo de la planta, no se realizó su árbol filogenético correspondiente, en su lugar, solo se realizó un alineamiento con parálogos caracterizados. En la figura 3.4 se observa parte del alineamiento de las dos secuencias HMGR de *Pentalinon andrieuxii* junto a otras HMGR de *Arabidopsis thaliana* (At), *Vitis vinífera* (Vv), *Nicotiana tabacum* (Nt) y *Catharanthus roseus* (Cr), se puede apreciar que los motivos de unión a HMGR-CoA y NADP(H) se encuentran presentes, por lo tanto, se espera que la secuencia represente a una enzima funcional. A lo largo de todo el alineamiento se puede apreciar que no solo los motivos catalíticamente importantes se encuentran conservados, en caso de haber una sustitución, en la mayoría de los casos el aminoácido presente, tiene características similares a los que se encuentran en las demás secuencias en la misma posición. La clave para la variación que preserva la función de la proteína es la retención de nucleótidos específicos en el codón de ADN que codifican aminoácidos con polaridad o hidrofobicidad similar (Castro-Chavez, 2010).

Otra característica identificada en las dos secuencias HMGR de *P. andrieuxii*, es la presencia de dos regiones transmembranales en el extremo N-terminal (ver anexo II), rasgo que comparte con las HMGR de plantas.

Con la finalidad de clasificar los genes DXS encontrados en el transcriptoma, se llevó a cabo un análisis filogenético que incluyó 26 secuencias de DXS que han sido caracterizadas en plantas y tres DXS de bacterias. El análisis mostró que los genes DXS detectados en el transcriptoma se agrupan en tres clados. Estos resultados nos indican que los genes identificados son distintos entre sí, como puede observarse en la figura 3.5, en el clado I se agrupó PaDXS2 y las DXS de plantas de tipo I, en el clado II se encuentran PaDXS1 y PaDXS3 junto con las secuencias DXS tipo II, por último, en el clado III se agrupó PaDXS4 junto con DXS de tipo III.

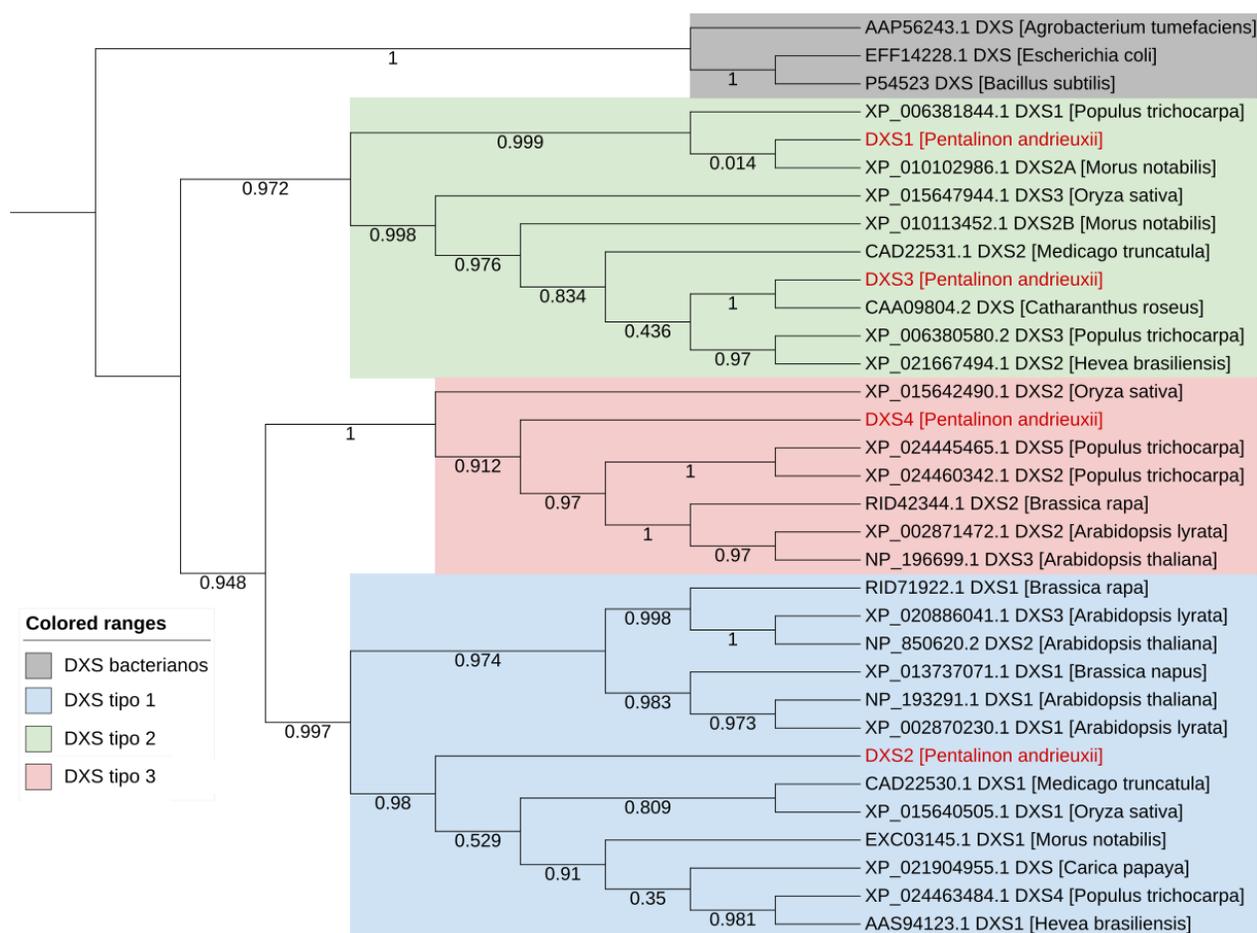


Figura 3.5. Análisis filogenético de las proteínas DXS. El árbol filogenético se infirió utilizando el modelo BIONJ y el modelo de sustitución JTT. El análisis involucró 29 DXS caracterizadas y las 4 DXS identificadas en el transcriptoma de *Pentalinon andrieuxii*.

Para el análisis filogenético de la familia prenilttransferasa, se infirió un árbol filogenético que incluyó 35 secuencias previamente caracterizadas en la literatura más las secuencias identificadas en el transcriptoma. Como resultado, se logró agrupar todas las PT de *P. andrieuxii* dentro de las subfamilias establecidas por You et al., 2020. En la figura 3.6 se muestra el árbol filogenético obtenido. Como se puede observar, PT4 y PT6 quedaron dentro del clado FPS junto con cuatro FPS de otras plantas, PT1 y PT8 quedaron en los clados GGPS-SSU-I y GGPS-SS-II respectivamente, ser catalíticamente inactivas en su forma monomérica, es la principal característica de las secuencias que quedaron en estos clados, en cambio necesitan formar complejos con las secuencias del clado GGPS-LSU, del cual PT3 y PT10 quedaron agrupados, por otra parte, PT2 y PT7 se agruparon en el clado G(G)PS, finalmente PT5 y PT9 se agruparon en los clados SPS y PPS respectivamente, los cuales no participan en la formación de mono, sesqui y diterpenos.

La construcción del árbol filogenético correspondiente a los terpenos sintasas se basó en la más reciente clasificación de las subfamilias TPS realizado por Chen *et al.*, 2011, no se agregó secuencias de la subfamilia tps-h debido a que es exclusiva de la especie *Selaginella moellendorffii*, tampoco se agregaron secuencias de la subfamilia tps-d que pertenecen a TPS de gymnoespermas.

En la figura 3.7 se puede observar el resultado del árbol filogenético de TPS, todas las secuencias de *P. andrieuxii* quedaron agrupados en los clados formados, las secuencias tps1, tps5a y tps9 se agruparon en el clado tps-b formado por monoterpenos sintasas, la secuencia tps3 se agrupó en el clado TPS-g el cual su principal característica es la formación de productos acíclicos, las secuencias TPS2, TPS7, TPS8 y TPS10 se agruparon en el clado de las sesquiterpenos sintasas denominado TPS-a, el clado TPS-e, agrupa únicamente enzimas KS, en el cual se ubicó la TPS4, tal como lo sugería el resultado de BLASTp, de manera similar, el clado TPS-c, se forma únicamente de enzimas CPS, en este caso la TPS11 se agrupó en esta subfamilia, finalmente la TPS6 se ubicó en el clado TPS-f el cual está compuesto principalmente de diterpenos sintasas.

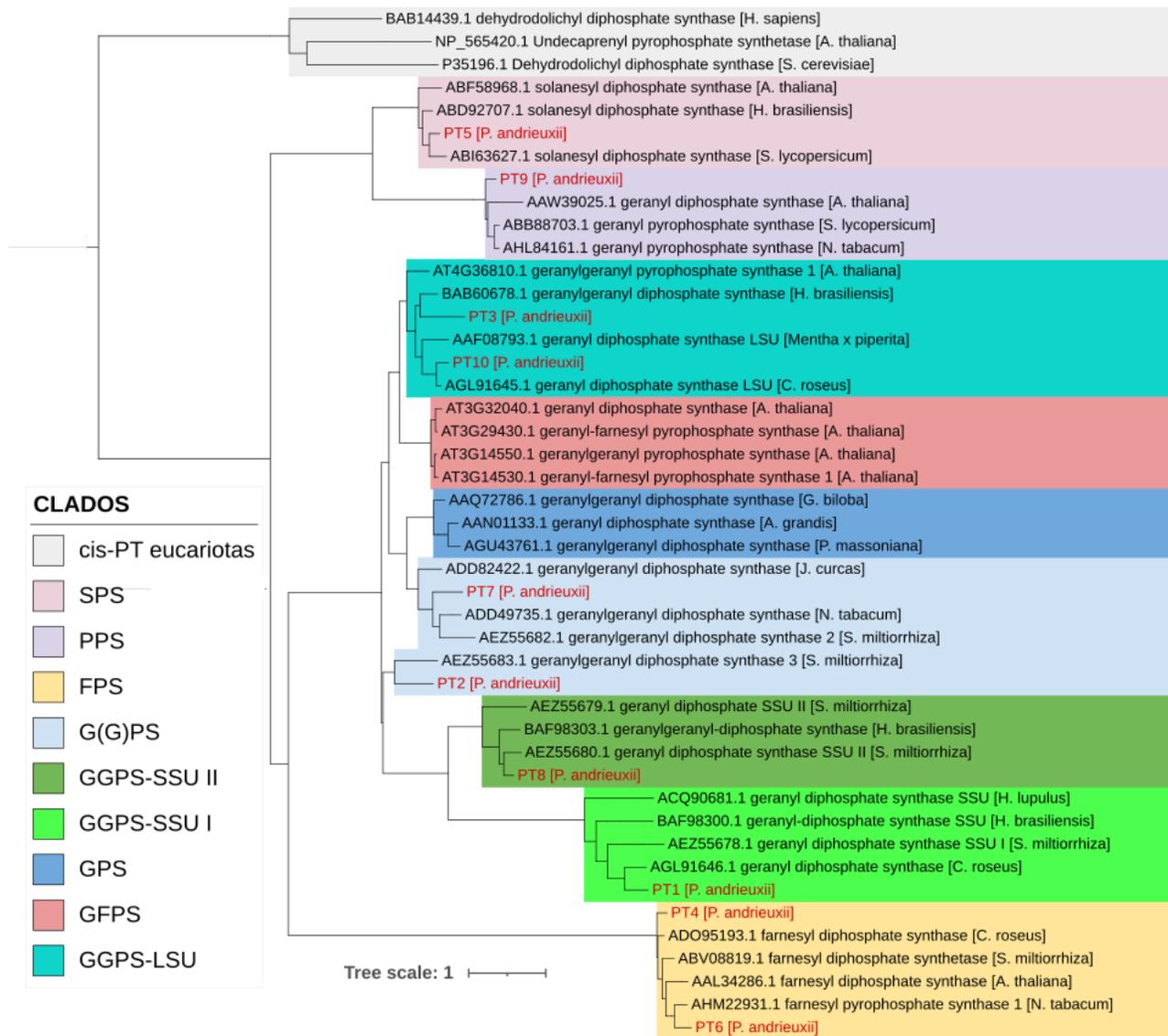


Figura 3.6. Análisis filogenético de la familia trans-preniltransferasas. El árbol filogenético se infirió utilizando el modelo BIONJ y el modelo de sustitución JTT. El análisis involucró 35 secuencias identificadas en la literatura y 10 secuencias identificadas en el transcriptoma de *P. andrieuxii* (señaladas de color rojo). Los números de accesoión, la actividad enzimática y la especie a la cual pertenecen se encuentran descritos en las ramas del árbol.

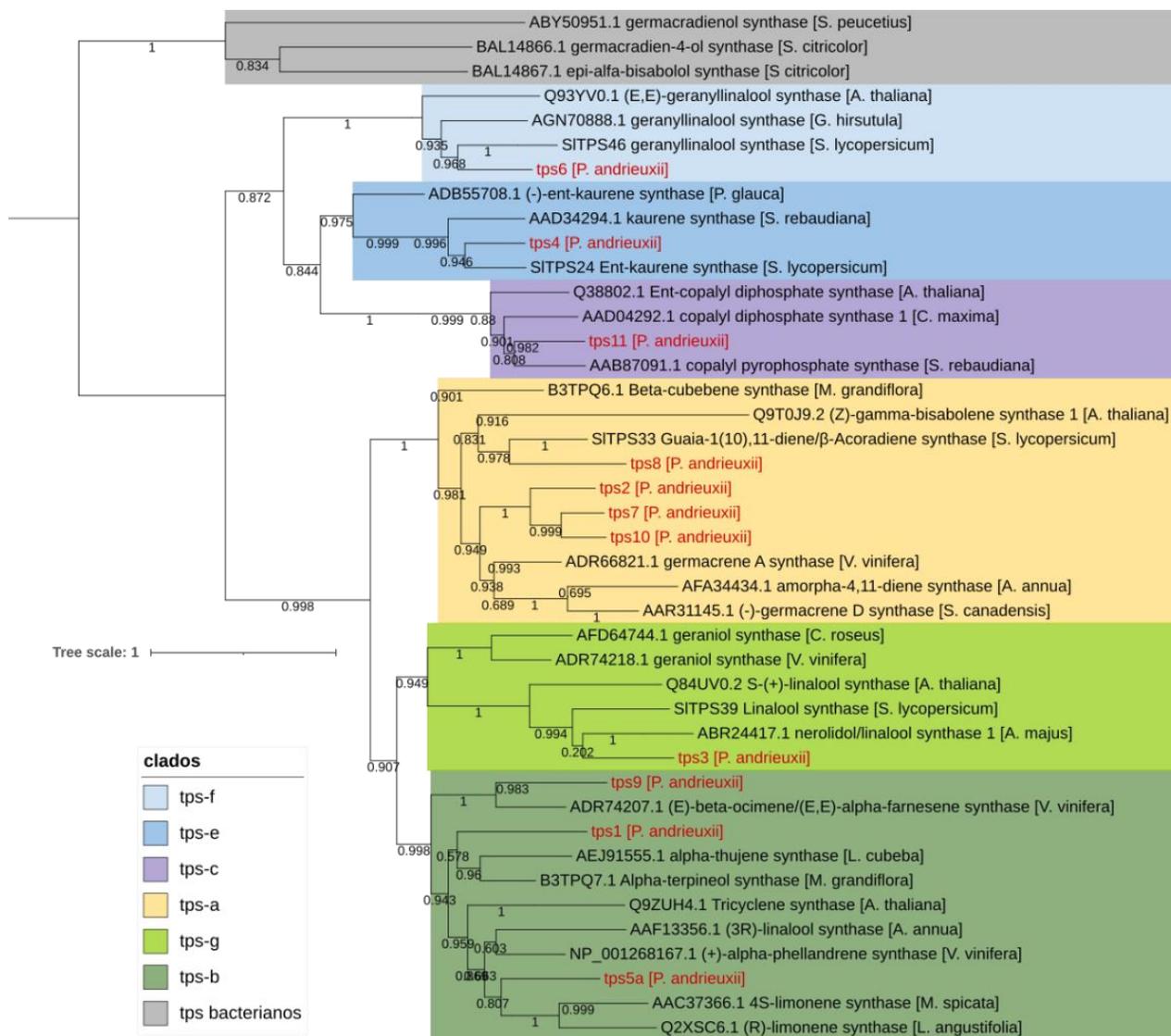


Figura 3.7. Árbol filogenético de la familia terpeno sintasa, incluyendo las secuencias identificadas de *Pentalinon andrieuxii*. El árbol filogenético se infirió utilizando el modelo BIONJ y el modelo de sustitución JTT. El análisis involucró 31 TPS caracterizadas y las 11 de *P. andrieuxii* que contienen el dominio catalítico.

La presencia del motivo conservado DXDD es fundamental en las TPS de clase II, mientras que el motivo DDXXD es fundamental para las TPS de clase I, solo la presencia de los motivos mencionados indica que sea catalíticamente activo. Para identificar la presencia de los motivos conservados mencionados anteriormente, se realizó un alineamiento múltiple con la herramienta online MAFFT, dicho alineamiento incluyó únicamente las doce secuencias de TPS de *P.*

andrieuxii. Como se puede observar en la figura 3.8, a excepción de la TPS11 y la tps5x, todas las demás TPS son de clase I, es decir, inician su reacción por ionización del sustrato mediado por metales, mientras que la TPS11 es de clase II, en esta clase de enzimas un aminoácido es el encargado de ionizar el sustrato para iniciar la reacción.

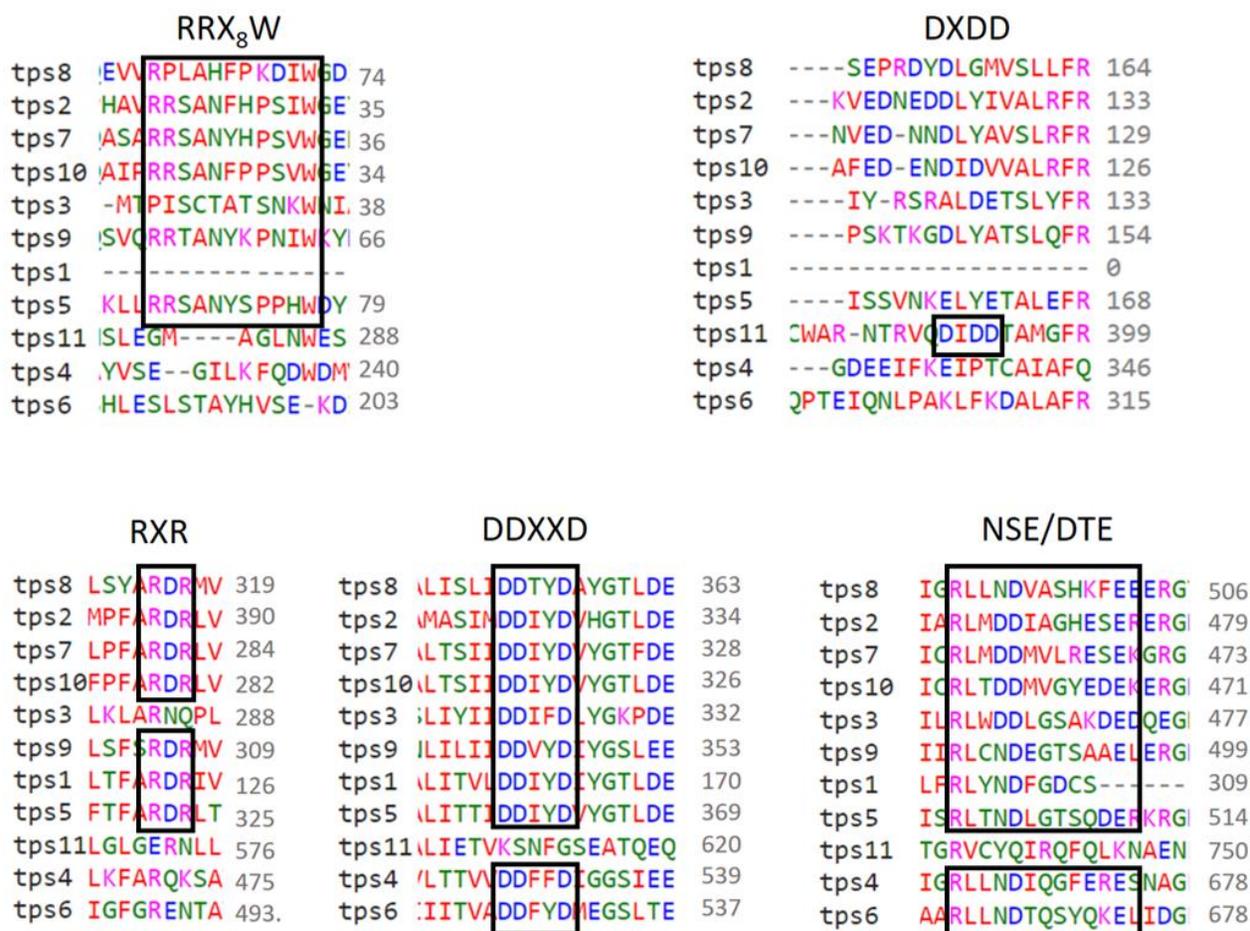


Figura 3.8. Motivos conservados de las TPS de *Pentalinon andrieuxii* identificadas. Los motivos conservados clave se señalan con cuadros negros. Los aminoácidos hidrofóbicos se encuentran resaltados en color rojo, los aminoácidos con carga negativa o ácidos están de color azul, los que tienen carga positiva o básicos se representan de color magenta, mientras que los aminoácidos polares se encuentran de color verde.

3.3 Análisis de expresión *in silico*

Las lecturas previamente filtradas descritas en las secciones 2.1 y 3.1 se utilizaron para cuantificar la expresión *in silico* utilizando el programa Salmon. El porcentaje de alineamiento obtenido de las lecturas filtradas al transcriptoma de referencia se encuentra dentro de los parámetros recomendados (mayor al 80%). En la tabla 3.11 se presenta el porcentaje de alineamiento obtenido para cada conjunto de datos correspondiente a las muestras secuenciadas.

Tabla 3.11. Porcentaje de alineamiento obtenido usando la herramienta Salmon.

Muestra	Porcentaje de mapeo
Isoformas	
GG1TP4SS01 Raíz Joven	97.8725
GG1TP4SS02 Hoja adulta	98.3067
GG1TP4SS03 Raíz adulta	98.1059
GG1TP4SS04 Hoja adulta	98.3741
Supertranscritos	
GG1TP4SS01 Raíz Joven	79.0629
GG1TP4SS02 Hoja adulta	78.2868
GG1TP4SS03 Raíz adulta	81.0241
GG1TP4SS04 Hoja adulta	79.0222

Salmon, además, nos da en una tabla, el número de lecturas alineadas a cada contig y su valor normalizado en TPM. Con los identificadores de las secuencias halladas de los tres módulos de biosíntesis de los isoprenoides, se recuperaron los valores de TPM de cada uno de ellos y se graficó con la herramienta heatmap2 de Galaxy, el resultado se muestra en las figuras 3.9, 3.10 y 3.11, cada uno corresponde a un solo módulo.

Se decidió usar la cuantificación a nivel gen debido a que algunos contigs ensamblados tenían muchas isoformas y en algunas ocasiones la isoforma más expresada no era la misma en todos los tejidos o no estaba completa.

Como se puede observar en la figura 3.9 correspondiente a la expresión de los genes de las rutas MVA y MEP, cada fila representa un gen y cada columna representa los valores de expresión del gen en cada muestra secuenciada. Las muestras secuenciadas carecen de réplicas biológicas, por lo tanto, los estudios de expresión diferencial no se pudieron realizar, sin

embargo, se puede observar que los datos de expresión de los genes en raíz adulta y raíz joven se agrupan juntos, mientras que los datos de hoja adulta y joven forman otro grupo, lo que quiere decir que los niveles de expresión de los genes en raíz son similares a pesar del estadio de desarrollo de la planta, el comportamiento es similar en las muestras de hoja.

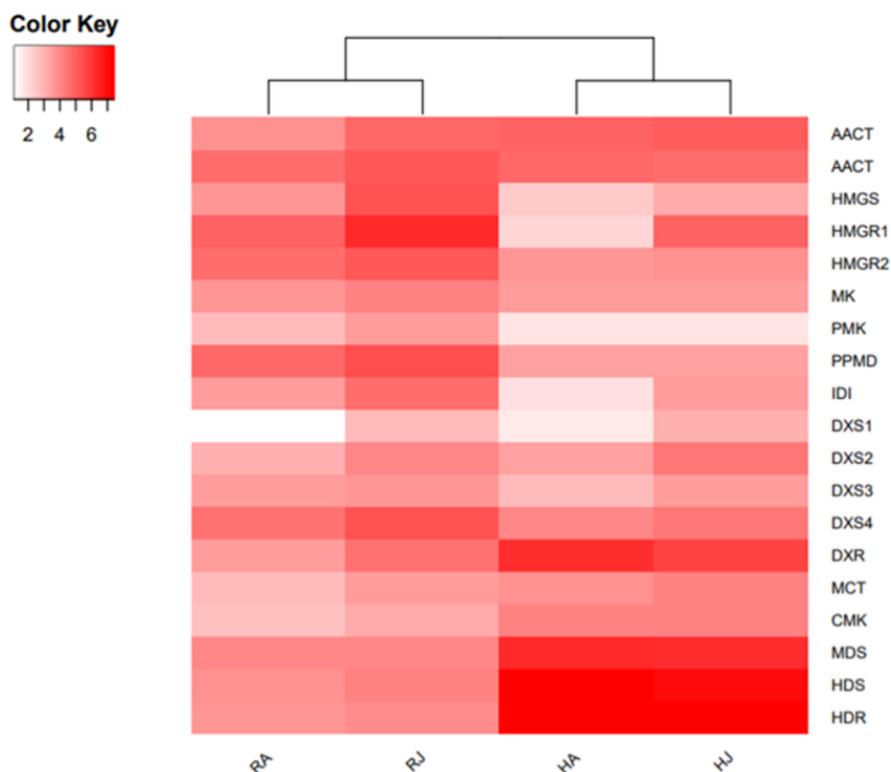


Figura 3.9. Expresión *in silico* de genes de la biosíntesis de IPP. Mapa de calor generado con la herramienta heatmap2, se utilizó la función $\text{Log}_2(\text{valor}+1)$ para transformar los valores de expresión *in silico* (en TPM) de cada gen.

La expresión de los miembros de la familia de los genes claves de ambas rutas, HMGR para la ruta MVA y DXS para la ruta MEP, se expresan de manera diferente, en general, la expresión del gen HMGR1 tuvo valores más altos de TPM en relación con HMGR2, a excepción en la muestra de hoja adulta, sin embargo, al no tener mayor evidencia, no se podría relacionar uno de los dos genes con la producción de metabolitos primarios o secundarios. Por el contrario, según la clasificación de las secuencias DXS, DXS2 perteneciente a la clase I estaría relacionada con la producción de clorofilas y carotenoides, mientras que los genes DXS1 y DXS3 que se

agruparon en subfamilia clase II, que podrían estar involucrados en el aporte de precursores para la biosíntesis de terpenos especializados o de defensa, por último, DXS4 que se clasificó como parte de la subfamilia de clase III, no se le podría asignar una función en particular. De los cuatro genes DXS, se esperaba que DXS1 tuviera una expresión más alta, por su participación en la producción de metabolitos primarios, sin embargo, fue DXS4 que obtuvo los valores más altos de TPM (ver anexo III).

Por otra parte, el mapa de calor correspondiente a la expresión *in silico* de las secuencias de la familia PT identificados se muestra en la figura 3.10, como se puede apreciar, a excepción de las FPS, las demás preniltransferasas están más expresadas en las muestras de hoja que en las de raíz y en general, todas las trans-PT tienen mayor expresión en muestras de plantas jóvenes que en muestras de plantas adultas.

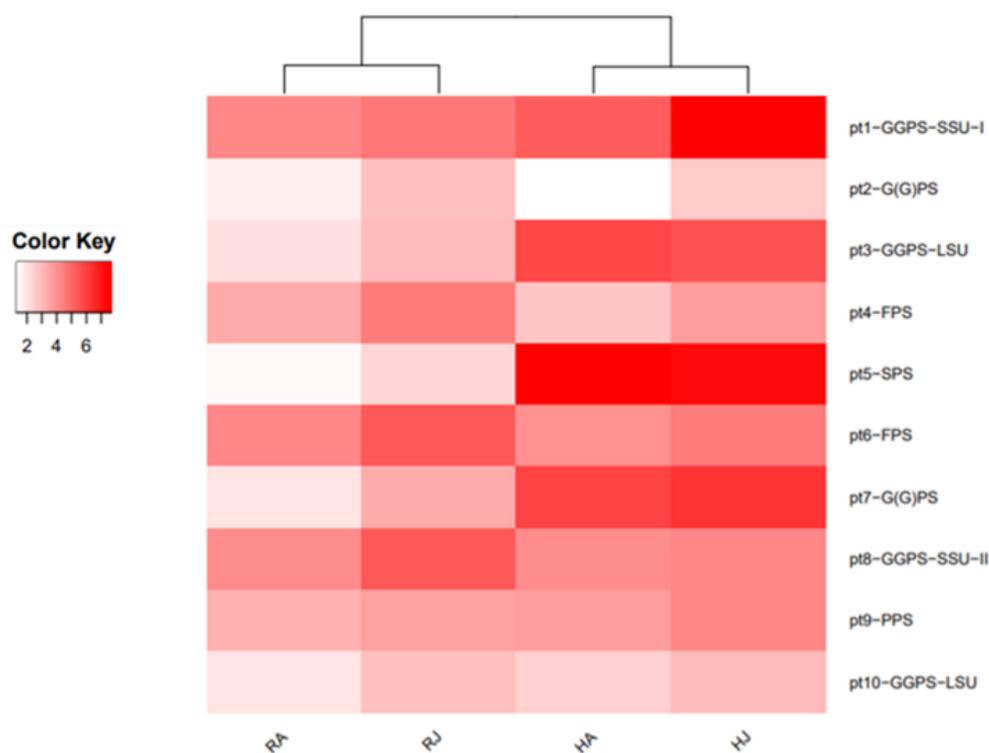


Figura 3.10. Expresión *in silico* de los genes de la familia trans-preniltransferasa de *Pentalinon andrieuxii*. Mapa de calor generado con la herramienta heatmap2, se utilizó la función $\text{Log}_2(\text{valor}+1)$ para transformar los valores de expresión *in silico* (en TPM) de cada gen.

Por último, se presenta en la figura 3.11 el mapa de calor correspondiente a los niveles de expresión *in silico* de los genes de la familia TPS. Como se puede apreciar, en este módulo en particular, si se encontraron genes que se expresan únicamente en hoja o en raíz, independientemente del estadio de desarrollo. Las tps1, tps2, tps6, tps9 y tps10 se expresan únicamente en hoja, mientras que las tps8 y tps11 se expresan únicamente en raíz, finalmente las tps3, tps4, tps5 y tps7 se expresan en ambos tejidos, se consideró que un gen se expresa si su valor de TPM es mayor de 0.3, debido a que se pudo amplificar la secuencia de la tps2 en tejido de hoja adulta, la cual tiene un valor de 0.367152. No se puede asegurar la expresión de genes con menor valor de TPM. Los valores numéricos en TPM de cada tps, se pueden consultar en el anexo III. Ninguna TPS se expresó en una condición en específico que podría suponer su participación en la biosíntesis de los urechitoles, en particular, la secuencia tps10 que corresponde a una sesquiterpeno sintasa, tuvo el mayor valor de expresión en tejido de hoja adulta y joven, seguido de la tps7, una monoterpene sintasa, mientras que la sesquiterpeno tps8 y un fragmento que se alinea a esta, tuvieron los valores más altos en tejido de raíz. La secuencia tps5x es un fragmento que se alinea con la tps5. Sin embargo, solo se encuentra expresado en raíz, sin embargo, no se encontró su secuencia completa y solo una región del extremo N-terminal.

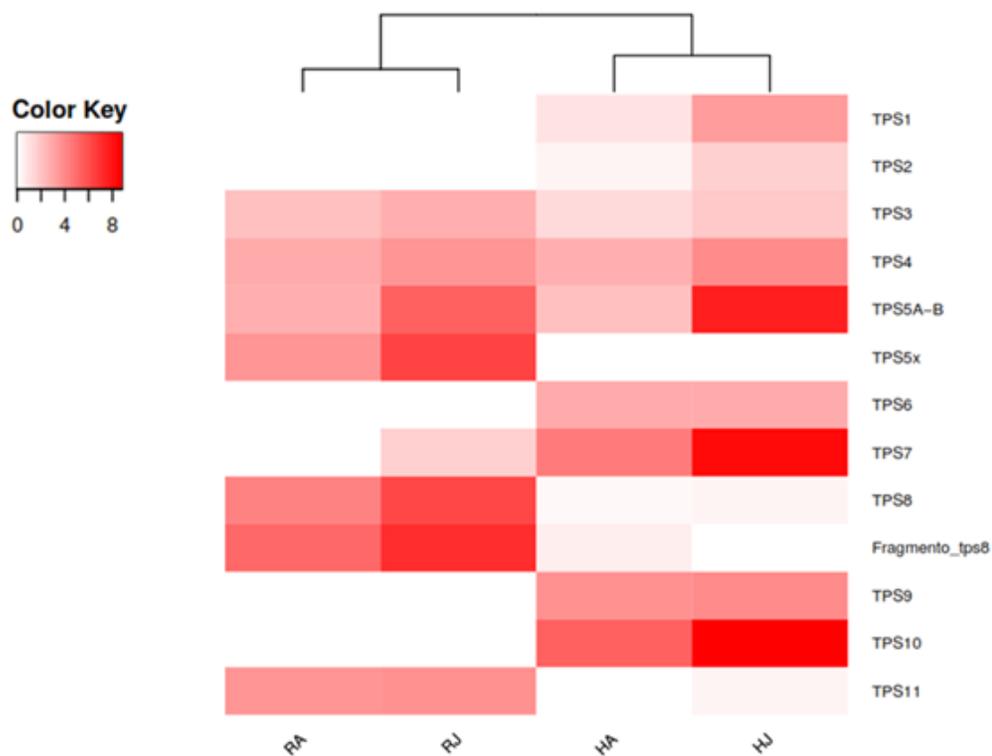


Figura 3.11. Expresión *in silico* de los genes de la familia terpenos sintasas de *Pentalinon andrieuxii*. Mapa de calor generado con la herramienta heatmap2, se utilizó la función $\text{Log}_2(\text{valor}+1)$ para transformar los valores de expresión *in silico* (en TPM) de cada gen.

3.4 Validación del ensamblado del transcriptoma mediante el aislamiento y secuenciación de transcritos

Al final de las rutas del mevalonato y MEP únicamente se producen IPP y DMAPP, mientras que las preniltransferasas producen los esqueletos precursores de los isoprenoides, sin embargo la diversidad de estos precursores es muy limitada, son las enzimas terpenos sintasas las que determinan en mayor medida la diversidad de terpenoides que puede sintetizar una especie, por lo tanto, el estudio de estas últimas podría aportar información sobre la biosíntesis de metabolitos novedosos o únicos que produce *P. andrieuxii*, por tal motivo, solo se diseñaron primers específicos para las secuencias identificadas de esta familia. En el anexo III se presenta la secuencia de cada primer, la temperatura de alineamiento según el proveedor y la posición dentro de la secuencia en la que se localiza cada primer.

De acuerdo a los datos de expresión *in silico*, se seleccionó los tejidos en los cuales la expresión de cada TPS sea más alta, por lo tanto, solo se extrajo ARN de hojas y raíces jóvenes de *P. andrieuxii*. El ADNc obtenido de hojas de plantas jóvenes se usó como templado para amplificar las secuencias TPS11, TPS2, TPS4, TPS5, TPS6, TPS7, TPS9 y TPS10, para las secuencias restantes, se usó ADNc obtenido de raíz de plantas jóvenes, lo anterior, no indica que en los demás tejidos no se exprese el gen, sino, que, en el tejido seleccionado para cada secuencia, se espera una mayor probabilidad de poder amplificar la secuencia completa de cada transcrito.

En la siguiente figura 3.12 se observa los productos de PCR de las posibles sesquiterpenos sintasas según lo indicado en la tabla 3.10 y la figura 3.7. A pesar que se diseñó un primer para amplificar la longitud completa del TPS2, no se pudo amplificar con éxito en todas las temperaturas probadas, mientras que el primer Forward del TPS8 amplifica hasta el nucleótido 117, el cual codifica para la segunda metionina, esto es debido a que se mandaron a diseñar los primers antes de haber identificado la longitud completa del transcrito. Las dos secuencias restantes se lograron amplificar completas exitosamente.

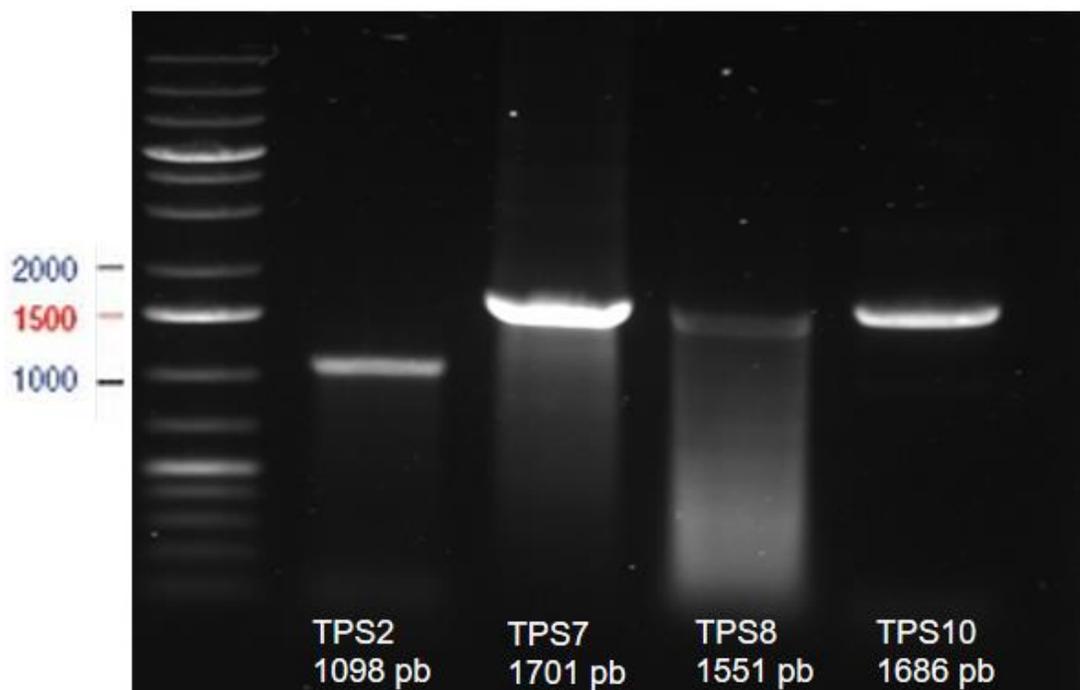


Figura 3.12. Amplificación de las sesquiterpenos sintasas TPS2, TPS7, TPS8 y TPS10.

En cuanto a las demás secuencias de TPS identificados, de manera completa, se pudieron amplificar los transcritos TPS9, TPS4, TPS6 y TPS11, el primero corresponde a una monoterpene sintasa, los tres restantes, fueron identificados como diterpenos sintasas. En la figura 3.13 se presentan los productos de PCR de las secuencias anteriormente mencionadas, similar a lo realizado con las sesquiterpenos de la figura 3.12, se utilizó muestra obtenida del tejido con expresión *in silico* mas alta según corresponda a cada TPS.

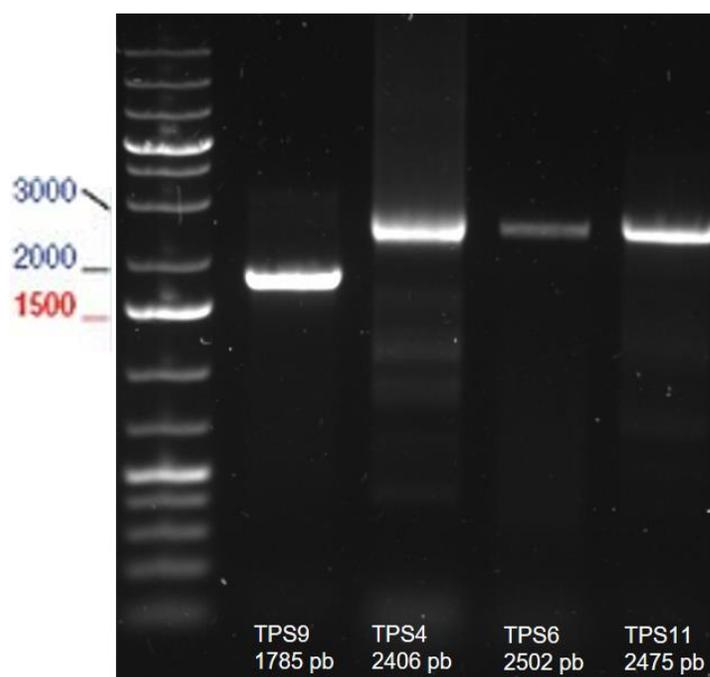


Figura 3.13. Amplificación del monoterpene sintasa TPS9 y de los diterpenos sintasas TPS4, TPS6 y TPS11.

Como se puede observar en la figura 3.13, el patrón electroforético de las bandas de cada TPS corresponde al tamaño previsto de las secuencias obtenidas en el análisis del transcriptoma de *P. andrieuxii*.

De los TPS restantes, únicamente se lograron amplificar parcialmente los transcritos TPS1 y TPS3, ambos monoterpene sintasas, sin embargo, como se puede observar en la figura 3.14, en los carriles correspondientes al TPS1 se amplificaron tres bandas que tienen un comportamiento similar usando dos primers Forward diferentes, en ambos casos una de las bandas corresponde al tamaño de amplicón esperado. El amplicón obtenido del TPS3 corresponde al tamaño del fragmento esperado para el cual se diseñó los primers, sin embargo, no se pudo amplificar la secuencia completa.

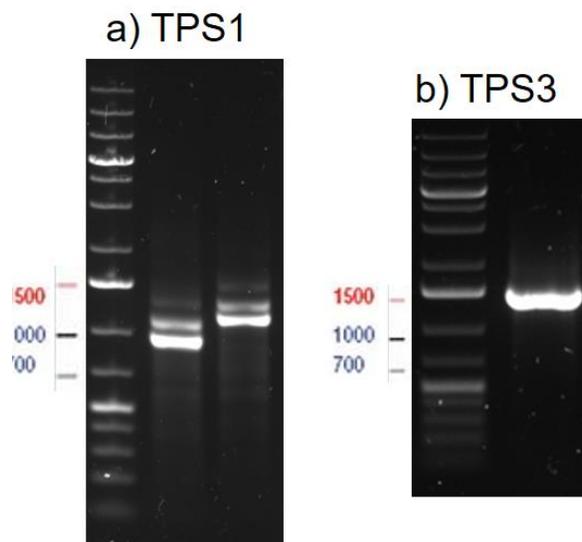


Figura 3.14. Amplificación de los monoterpenos sintasas TPS1 y TPS3.

De los terpenos sintasas anteriormente presentados se logró amplificar en cuatro casos al menos un fragmento de la secuencia y de los otros seis casos se obtuvo la secuencia completa, sin embargo, en ningún caso se logró amplificar el TPS5, el cual presenta en los resultados *in silico* varias isoformas no redundantes, además, a pesar de que los mejores alineamientos obtenidos con blastp resultan con secuencias de monoterpenos sintasas, la secuencia TPS5 no presenta péptido de tránsito al cloroplasto, sin embargo, el fragmento parcial 5' nombrado como TPS5x se alinea con las demás secuencias TPS5 en la extremo N-terminal y si se le detectó un péptido señal.

Para comprobar si los genes que presentaron bajos valores de expresión *in silico* como la secuencia tps2 (0.3671 TPM en hoja adulta) se expresan, se amplificó a partir de ADNc proveniente de ARN de tejido de hoja adulta las primeras TPS. Como resultado, en la figura 3.15 se puede observar que las secuencias tps1, tps2 y tps3 lograron amplificarse, indicando que valores de expresión menores a 1 representan genes activos.

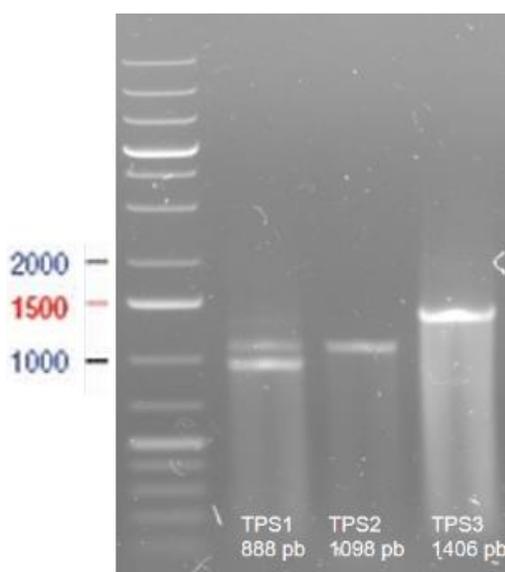


Figura 3.15. Amplificación de las secuencias tps1, tps2 y tps3 mediante PCR usando ADNc sintetizado a partir de ARN proveniente de tejido de hoja adulta de *P. andrieuxii*.

3.4.1 Clonación

Debido a que la secuencia TPS7 fue la primera sesquitepeno sintasa en lograr amplificarse de manera completa, se clonó en el vector pGEM t-easy. Para comprobar que la clonación fue exitosa, se analizó los patrones electroforéticos de las bandas de ADN obtenida tras la digestión enzimática de los plásmidos purificados. La enzima Sall corta el plásmido pgem-t-easy en el nucleótido 91, mientras que la enzima HindIII corta la secuencia tps7 en el nucleótido 1134 de los 1701 pb de longitud total del cds de este transcrito, lo que nos daría un fragmento de aproximadamente 600 pb en caso de que el codón que inicio se haya ligado en dirección contraria al sitio de reconocimiento de la enzima Sall. En la figura 3.16 del lado izquierdo se puede apreciar dos fragmentos de ADN resultantes de la digestión enzimática con las enzimas Sall y HindIII, los fragmentos de aproximadamente 4134 pb corresponden a los 3015 pb del vector más los 1134 pb del fragmento de tps7, los fragmentos más pequeños corresponden a los 567pb restantes de la secuencia tps7 más los nucleótidos del sitio de clonación presentes entre la posición 64 y 90 del vector. Del lado derecho de la figura 3.16 se presenta los amplicones resultantes de la

reacción de PCR usando los primers específicos de la secuencia tps7 y como templado se usó el plásmido purificado de las colonias transformadas, las posiciones de los carriles numerados del 1 al 3 en ambos lados de la figura 3.16 corresponden a la misma muestra de ADN plasmídico purificado.

Los resultados de la secuenciación del gen transcrito TPS7 clonado, arrojó un porcentaje de similitud del 99.1% con respecto a la secuencia identificada en el transcriptoma, la secuencia de la muestra mandada a secuenciar se encuentra en el anexo VI. Es importante mencionar que al utilizar primers específicos de la secuencia del transcrito para secuenciar, las primeras y últimas bases de la secuencia no se lograron secuenciar y solo se alineó el segmento donde inició la secuenciación.

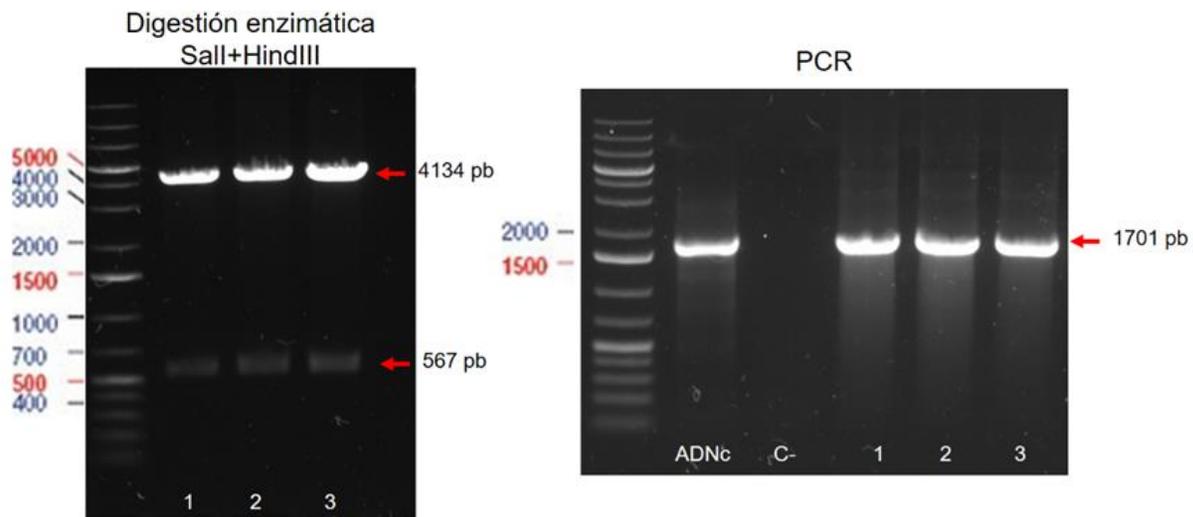


Figura 3.16. Verificación del ADNc clonado mediante restricción enzimática y PCR.

Recapitulación de resultados

Este trabajo estuvo orientado a la identificación de genes potencialmente implicados en la biosíntesis de isoprenoides de *Pentalinon andrieuxii*, en especial del urechitol. Para este propósito, se analizaron los transcriptomas de hojas y raíces y plantas jóvenes y adultas de *P. andrieuxii*, este objetivo se dividió en 4 etapas:

En la primera etapa se realizó un análisis de calidad de las lecturas secuenciadas que corresponden a las muestras anteriormente mencionadas, posteriormente las lecturas de baja calidad (con valor Phred menores a 20) fueron eliminadas. En total se obtuvo un porcentaje de recuperación de 78.3, que corresponden a alrededor de 778 millones de lecturas. Seguidamente se realizó el ensamblado de *novο* del transcriptoma de *P. andrieuxii* mediante la herramienta Trinity que generó 282,627 transcritos de 160,909 genes con promedio de longitud de los contig de 1378.65 pb.

En la segunda etapa se realizó una búsqueda por homología mediante la herramienta BLAST que permitió la identificación de 40 secuencias que pertenecen a genes que codifican enzimas de la ruta del mevalonato y metileritrol fosfato, de las familias preniltransferasa y terpenos sintasas. Se identificaron diez preniltransferasas, de las cuales dos son FPPS, dos GGPPS y cuatro secuencias que requieren caracterización bioquímica para determinar su función como GPPS o GGPPS. De la familia terpeno sintasa se logró identificar cuatro monoterpénos sintasas, cuatro sesquiterpénos sintasas, tres diterpénos sintasas, las cuales dos son del metabolismo de las giberelinas.

En la tercera etapa se realizó un análisis de expresión *in silico* mediante la herramienta Salmon, de las secuencias identificadas, tres sesquiterpénos sintasas y dos monoterpénos sintasas mostraron patrones de expresión *in silico* específicos de tejido. Las tps1, tps2, tps9 y tps10 se expresan únicamente en hoja, mientras que la tps8 se expresan únicamente en raíz. Se propone estudiar más a detalle las sesquiterpénos sintasas tps2, tps8 y tps10 para correlacionar su expresión con la biosíntesis de urechitol.

Finalmente, para validar el transcriptoma ensamblado se diseñaron primers para amplificar cada una de los terpenos sintasas mediante PCR. Como resultado se logró amplificar 10 de los 11 terpenos sintasas identificados y uno se logró clonar y secuenciar resultando un 99.1% de identidad con la secuencia identificada en el transcriptoma, por lo que se concluye que el transcriptoma es funcional para la identificación de secuencias.

CAPÍTULO IV

3.1. DISCUSIÓN

En la investigación de productos naturales, en específico la deducción de rutas metabólicas se ha visto favorecida por el desarrollo y mejoramiento de las tecnologías de secuenciación de ARN (ARN seq) que ha permitido el descubrimiento de genes en especies de las cuales no se encuentra disponible su genoma de referencia (Góngora-Castillo y Buell, 2013).

El transcriptoma representa la porción expresada del genoma y ofrece una visión general de los genes transcritos. Es una herramienta poderosa en el descubrimiento de genes y en la comprensión de las vías bioquímicas involucradas en las respuestas fisiológicas, los patrones de expresión espacio-temporal, así como la abundancia de la expresión, pueden ser tomados en cuenta como evidencia de la función de un gen o un conjunto de genes (Gahlan *et al.*, 2012).

En este trabajo se llevó a cabo el ensamblaje *de novo* de las lecturas cortas derivadas de la secuenciación de cuatro transcriptomas de *Pentalimon andrieuxii*, la identificación de secuencias relacionadas con la biosíntesis de isoprenoides, así como la validación del ensamblaje mediante la amplificación de genes por PCR, la clonación de uno de ellos y su posterior secuenciación.

Si bien no existe un flujo de trabajo óptimo y universal para la diversidad de aplicaciones y análisis en los que se puede utilizar las tecnologías de ARN-seq, en una revisión realizada por Conesa *et al.*, 2016, se hace una serie de recomendaciones sobre las etapas principales de un experimento típico de ARN-seq, en la cual se recomienda que una vez realizada la secuenciación de los transcriptomas, se revise la calidad de las lecturas y se eliminen los errores que pudieran haber surgido durante la secuenciación, tales como la calidad de las lecturas, lecturas sobre representadas, presencia de adaptadores, duplicación excesiva de lecturas, entre otros, entre las herramientas recomendadas para estos pasos, se encuentran FastQC para evaluar la calidad de la secuenciación y Trimmomatic para corregir los errores detectados, seguido de un ensamblado *de novo* usando la herramienta Trinity, además, se recomienda la reducción del número de lecturas para muestras secuenciadas en profundidad. Para análisis comparativos entre muestras, es aconsejable combinar todas las lecturas de múltiples muestras en una sola entrada para obtener un conjunto consolidado de contigs (transcripciones), seguido de mapeo de las lecturas cortas para la estimación de expresiones (Haas *et al.*, 2013). Todos los aspectos

anteriores fueron tomados en cuenta, lo que derivó al ensamblaje del transcriptoma *de novo* de *Pentalinon andrieuxii*.

Para evaluar la calidad de las secuencias ensambladas se usó la herramienta BUSCO (Simao *et al.*, 2015), dicha herramienta proporciona valores numéricos sobre cómo se lograron encontrar genes que se sabe se encuentren en una sola copia en los genomas de un linaje en particular, en este caso se decidió usar el linaje de las embryophytas, debido a que dicho linaje fue usado para evaluar transcriptomas ensamblados en especies de plantas como *Salvia hispanica* (Wimberley *et al.*, 2020), *Ferula assafoetida* (Amini *et al.*, 2019) y *Camellia sinensis* (Qiao *et al.*, 2019) que comparten características con *P. andrieuxii*. La evidente mejoría en calidad del ensamblaje al combinar todas las lecturas para obtener un único conjunto de contigs, reduce la probabilidad de que las transcripciones de baja expresión sean imposibles de ensamblar debido a que carecen de cobertura suficiente, en este caso, solo si dichas transcripciones se expresan en más de una muestra.

A pesar del masivo incremento de las secuencias en las diferentes bases de datos y la necesidad de anotar la enorme cantidad de datos obtenidos por las tecnologías de secuenciación modernas, BLAST sigue siendo una herramienta clave en la anotación funcional de genes y familias de genes (Buchfink *et al.*, 2021; Glover *et al.*, 2019). BLAST es capaz de conducir una búsqueda de similitud de secuencia para una secuencia de interés contra una base de datos de secuencia. BLAST se basa en la identificación de regiones de similitud entre las secuencias de interés contra una base de datos de secuencia a través de una alineación local basada en semillas y extensiones (Astul *et al.*, 1990; Dash *et al.*, 2021). La ejecución de BLAST puede llegar a ser muy demandante y la cantidad de consultas, así como el tamaño de la base de datos de referencia pueden afectar significativamente el tiempo de ejecución, además los parámetros usados para obtener el valor esperado de e depende de la cantidad de secuencias presente en las bases de datos (Dash *et al.*, 2021), por lo tanto, se decidió a crear bases de datos pequeñas que contengan únicamente secuencias de parálogos de los genes de la ruta MVA, MEP, preniltransferasas y terpenos sintasas, para poder así hacer frente a las limitaciones computacionales. Dado que las secuencias de enzimas terpenos sintasas comprenden múltiples tipos de dominios conectados por enlazadores parcialmente conservados, el enfoque BLAST puede dar como resultado falsos positivos debido a pares de secuencias de puntuación más alta (HSP) cortos pero altamente conservados en partes funcionalmente no relevantes (es decir, estructurales) de la proteína (Hofberger *et al.*, 2015), además, se puede optar por otros enfoques como la utilización de HMM específicos para cada familia de genes puede ser una alternativa

viable y menos demandante computacionalmente que además ha tenido buenos resultados para la identificación de TPS (Hofberger *et al.*, 2015; Kumar *et al.*, 2018), sin embargo, las enzimas de la ruta MVA y MEP no pertenecen todos a la misma familia, por el contrario, cada uno realiza una función específica y a nivel global, la secuencia tiende a ser más conservada para lo cual el enfoque de búsqueda por similitud sería igual o más eficiente, por lo tanto, en orden de ser más homogéneos con las herramientas usadas, se decidió usar BLAST para identificar cada conjunto de secuencias, el valor esperado de E, se asignó arbitrariamente debido a que no existe un valor preestablecido para todas las tareas a realizar y para resolver los falsos positivos, se ejecutó BLAST en dos fases, primeramente usando bases de datos locales creadas específicamente para cada módulo, la segunda ronda se ejecutó en el servidor Galaxy usando las bases de datos de Refseq y Swissprot del año 2018, de esta forma se realizó un descarte manual de secuencias con tamaño menores al 50% del tamaño de su mejor hit, además sirvió para eliminar secuencias que no tengan relación a cada módulo establecido o que no sean de especies vegetales.

Como resultado del análisis de transcriptomas de *Pentalinon andrieuxii* se identificó genes candidatos implicados en las rutas MEV y MEP que proporcionan los precursores IPP y DMAPP necesarios para la biosíntesis de todos los isoprenoides. A pesar de no contar con secuencias genómicas de *P. andrieuxii*, las isoformas identificadas de la mayoría de los genes de ambas rutas, permiten suponer que provienen de genes de copia única, a excepción de los genes *aact*, *hmgr* y *dxs*, lo cual es similar para la especie *A. thaliana* (Vranová *et al.*, 2013). En el caso de la familia de genes HMGR, se encontró que las dos copias se encuentran expresadas en los tejidos de hoja y raíz de *P. andrieuxii*, en donde el transcrito *PaHMGR1* se expresa en mayor medida que *PaHMGR2* menos en hoja adulta, además, ambas HMGR tienen mayor expresión en raíz que en hoja. Por otra parte, en este trabajo, se identificaron cuatro DXS no redundantes, de las cuales, según las relaciones filogenéticas con DXS de otras especies, el transcrito *PaDXS2* estaría involucrado en la biosíntesis de precursores para isoprenoides fotosintéticos, mientras que *PaDXS1* y *PaDXS2* podría estar comprometido en la biosíntesis de metabolitos especializados, por último, el transcrito *PaDXS4* perteneciente a la subfamilia DXS tipo III, tuvo la expresión más alta de todas las DXS, sin embargo, a la subfamilia III no se le relaciona con algún proceso en particular. En la especie *Scoparia dulcis* se encontró resultados similares en el número de genes que codifica para cada enzima de las rutas MEV y MEP, se detectaron dos genes *AACT*, dos genes *HMGR*, cuatro *DXS* en transcriptomas de hoja, raíz y hojas tratadas con MeJa (Yamamura *et al.*, 2017).

La ruta MVA tiene un efecto regulatorio en el crecimiento celular debido a que el crecimiento celular depende de los derivados de la ruta MVA como los fitoesteroles que son componentes estructurales de la membrana celular (Leivar *et al.*, 2011; Ha *et al.*, 2001). Dentro de la ruta MVA, la reacción catalizada por la enzima HMGR es el primer paso comprometido de la vía del mevalonato (Bach, 1995), tanto los genes HMGR como las enzimas a las cuales codifican, se encuentran regulados en múltiples niveles (Hemmerlin *et al.*, 2012), sin embargo, los efectos del nivel transcripcional de estos genes y la producción de isoprenoides ha sido de los aspectos más estudiados, encontrándose que plantas transgénicas de *Solanum lycopersicum* que sobreexpresan el gen HMGR del melón tienen un efecto en el tamaño final del fruto, logrando obtener aumentos de tamaño y peso (Omura *et al.*, 2007), resultado que concuerda con la rápida acumulación de la proteína HMGR en las etapas tempranas del desarrollo del fruto del melón (*Cucumis melo*), específicamente en el pericarpio, lo que determina el tamaño del fruto (Kobayashi *et al.*, 2002). Al usar un súper promotor (Ni *et al.*, 1995) para regular la expresión del gen HMGR en *Hevea brasiliensis* se obtienen transcritos de hasta 160 veces más abundantes que logra aumentar el contenido de látex en la planta hasta cuatro veces más comparado con plantas silvestres, sin embargo, el mayor contenido de látex se obtuvo en una línea transgénica con niveles no tan altos de sobreexpresión del gen, estos datos indican que existe una correlación positiva entre los niveles de expresión génica de los genes HMGR y la producción de isoprenoides aunque este incremento no es directamente proporcional, dejando en claro que otros niveles de regulación también están influyendo (Jayashree *et al.*, 2018).

En cuanto al número de copias de genes HMGR y su participación en la biosíntesis de isoprenoides se ha reportado un estudio realizado por Li *et al.*, 2014 que incluyó el análisis de los genomas de 14 especies de plantas terrestres, se encontró que, dependiendo de la especie, es posible identificar de una a nueve copias de genes HMGR, en donde diez de ellas tienen tres o menos genes HMGR. *A. thaliana* y *Citrus sinensis* contienen dos copias, mientras que el mayor número de genes HMGR se encontró en el genoma de *Gossypium raimondii* con un total de nueve (Li *et al.*, 2014). En *Arabidopsis thaliana*, la cual solo tiene dos genes *hmgr* en su genoma, el gen *Athmgr1* se expresa en niveles similares en todos los órganos de la planta a lo largo de su desarrollo, a su vez, el gen *hmgr1* se expresa en promedio 2.2 veces más que el gen *Athmgr2* a excepción de los tejidos gametofitos masculinos (Vranová *et al.*, 2013). Por el contrario, las siete *hmgr* de *Zea mays* se expresaban de manera contrastante en tejidos específicos, por ejemplo, el gen *ZmHMG6* presenta sus niveles más altos de expresión en las semillas, mientras que los genes *ZmHMG2*, *ZmHMG3* y *ZmHMG5* tienen una expresión remarcable en raíces (Li *et*

al., 2014). En *Salvia miltiorrhiza* los genes SmHMG1 y SmHMG4 están altamente expresados en las flores, mientras que SmHMG2 está expresado principalmente en hojas y tallo, el gen SmHMG3 se expresa en todos los tejidos a excepción de las flores (Ma *et al.*, 2012). Al parecer, la duplicación y mayor número de copias trae consigo la especialización de cada una de ellas en ciertos procesos del metabolismo, en contrario a la expresión constante encontrada en especies que tienen menos número de copias como lo es *Pentalinon andrieuxii*, sin embargo, al no tener replicas, datos transcriptómicos de otros tejidos y de más estadios de desarrollo, complementado con la cuantificación de isoprenoides, no se puede establecer una relación directa entre los niveles de expresión de los genes de la ruta MVA y la producción de metabolitos como los urechitoles, el ácido betulínico o fitoesteroles como el pentalinoesterol, los cuales se cree que obtienen sus precursores de la ruta MVA. Los datos de expresión *in silico* de ambas isoformas de los genes HMGR podrían indicar una participación de HMGR1 en el crecimiento de la planta, proceso que supone una mayor necesidad de precursores, debido a que en hoja y raíz jóvenes se obtuvieron valores de TPM más altos comparados con las muestras adultas, por el contrario, no se observaron diferencias muy grandes en los niveles de expresión de HMGR2 entre muestras jóvenes y adultas, esta isoforma podría estar participando en la producción de precursores para otro tipo de isoprenoides que no tengan funciones en el desarrollo, donde experimentos con elicitores junto con cuantificación por qPCR podrían aportar más información para ambos genes.

La identificación de dominios y motivos clave en la actividad enzimática de las proteínas HMGR y TPS, permite confirmar que la mayoría de ellas corresponden a enzimas funcionales y que el transcriptoma ensamblado es funcional al menos para la identificación de genes relacionados con la biosíntesis de isoprenoides.

Trasladándonos a la ruta MEP, se ha reportado que, dependiendo de la especie de la planta u órgano, las enzimas DXS y DXR participan de manera diferente en la regulación de la ruta MEP, por ejemplo, en las hojas de *Arabidopsis*, tanto AtDXS como AtDXR funcionan como enzimas clave en la biosíntesis de carotenoides y clorofila de las hojas (Carretero-Paulet *et al.*, 2006; Estévez *et al.*, 2001), de manera similar la expresión heteróloga de genes DXS o DXR de *A. thaliana*, aumentó la cantidad del diterpeno abietano en las raíces peludas de *Salvia sclarea* (Vaccaro *et al.*, 2014). En contraste, solo AtDXS y no AtDXR, provocó un aumento de carotenoides y clorofilas en hojas y raíces de *Daucus carota* (Simpson *et al.*, 2016), al igual que en *Lavandula latifolia*, solo AtDXS causó un aumento significativo en el contenido de monoterpenos (Munoz-Bertomeu *et al.*, 2006; Mendoza-Poudereux *et al.*, 2014), por otro lado,

en hojas de *Nicotiana tabacum* la sobreexpresión de genes DXR causa un aumento en el contenido de carotenoides y clorofilas (Hasunuma *et al.*, 2008; Yang *et al.*, 2012).

La familia de genes DXS ha sido clasificada en 3 grupos. Los del grupo DXS-I tienen un papel esencial en la biosíntesis de terpenos fotosintéticos como clorofilas y carotenoides (Córdoba *et al.*, 2011; Kim *et al.*, 2005; Walter *et al.*, 2002). El grupo tipo DXS-II juega un papel secundario y ecológico en la producción de metabolitos terpenoides funcionales, como el ginkgólido en el *Ginkgo biloba* (Kim *et al.*, 2006), fitoalexinas en arroz (Okada *et al.*, 2007), apocarotenoides en raíces micorrízicas de *Medicago* (Floss *et al.*, 2008) y carotenoides en granos amarillos de maíz (Córdoba *et al.*, 2011). Las enzimas de tipo DXS-III no tienen una función específica (Carretero-Paulet *et al.*, 2013; Córdoba *et al.*, 2011; You *et al.*, 2020). La enzima DXS2 de *Arabidopsis*, la cual pertenece a la subfamilia DXS-III carece de actividad de tipo DXS; entre algunos miembros de esta subfamilia, se puede destacar que la expresión del gen *dxs3* del maíz, se detectó en diversos tejidos analizados, pero los niveles más altos se observaron en hojas maduras, mazorcas inmaduras y cáscaras, además, estudios de complementación en *E. coli* demostró la funcionalidad de esta enzima; en arroz, la expresión del gen *dxs3* se detectó en todos los tejidos analizados, sin embargo presentó los niveles más bajos de expresión comparado con sus parálogos (Carretero-Paulet *et al.*, 2013; Córdoba *et al.*, 2011; You *et al.*, 2020). Se tiene evidencia que la secuencia de las enzimas DXS de diferentes clados podría tener elementos regulatorios que permiten realizar su función en procesos específicos, sumado a los elementos regulatorios de sus promotores, la transformación de plantas de *A. thaliana* que sobreexpresan los genes DXS1 y DXS2A de *Morus notabilis* los cuales pertenecen a los clados DXS I y DXS II respectivamente, la sobreexpresión de MnDXS1 en *Arabidopsis thaliana* aumentó el contenido de ácido giberélico y resultó en una floración temprana, mientras que la sobreexpresión de MnDXS2A mejoró el crecimiento de las raíces y aumentó el contenido de clorofila y carotenoides, dichos efectos se observaron a pesar de estar bajo el control del mismo promotor, en este caso el promotor 35s de virus del mosaico de la coliflor (Zhang *et al.*, 2020) mientras que en plantas silvestres de *Morus notabilis*, DXS1 podría funcionar como gen de mantenimiento, mientras que DXS2A y DXS2B incrementan su expresión en contacto con el gusano de seda y en tratamientos de jasmonato de metilo, en el árbol filogenético presentado PaDXS2 se agrupó con MnDXS1 de *M. notabilis*, mientras que PaDXS1 y PaDXS3 se agruparon con MnDXS2A y MnDXS2B.

Se encontró que todos los genes de la ruta MEP a excepción de DXS3 y DXS4 se encontraron expresados en mayor medida en las muestras de hoja adulta y hoja joven en comparación con sus contrapartes de raíz, en especial los últimos dos genes específicos de la ruta (HDS y HDR).

En relación con los genes que se tiene conocimiento que tienen un efecto mayor en la regulación de esta ruta, tenemos que las cuatro isoformas no redundantes de genes DXS se expresan de diferente manera. DXS2 que se agrupó en la subfamilia I que se sabe que participa en la biosíntesis de precursores para la biosíntesis de isoprenoides fotosintéticos, se expresó de manera similar en ambos tejidos del mismo estadio de desarrollo, aun cuando se esperaba que la expresión sea mayor en hojas al ser tejido fotosintético, de las dos secuencias que se agruparon en la subfamilia II, DXS1 presentó los valores más bajos de expresión *in silico*, mientras que DXS3 obtuvo valores de expresión más altos en ambas muestras de raíz, sabiendo que en la subfamilia II se encuentran genes los cuales su expresión se correlacionó con la producción de metabolitos especializados, el gen DXS2 podría estar involucrado en la producción de metabolitos especializados en raíz. Por último, se encuentra el gen DXS4 el cual tuvo los valores más altos de expresión entre los DXS, sin embargo, del grupo al cual pertenece, no se tiene evidencia de su función. Se espera que los genes DXR de copia única se expresen altamente, en especial en tejidos fotosintéticos, tal como sucede en la especie *Oryza sativa* (You *et al.*, 2020) y en los datos obtenidos en este trabajo, el gen DXR se expresó altamente en muestras de hoja y en menor medida en muestras de raíz, aunque en mayor medida que los genes DXS a excepción de DXS4.

La presencia de una única isoforma no redundante del gen IDI indica que en *P. andrieuxii* ocurre un proceso similar al descrito para la especie *Catharanthus roseus* que cuenta únicamente con una copia de este gen, lo que supone una localización múltiple de la enzima IDI en peroxisomas, mitocondrias y cloroplastos (Guirimand *et al.*, 2012).

Al final de la ruta MVA y MEP únicamente se tienen dos productos, el IPP y su isómero DMAPP. La diversidad tan grande de isoprenoides comienza con la condensación de DMAPP con diferentes unidades de IPP para producir esqueletos de isoprenoides de diferente longitud, los cuales servirán como base para las modificaciones consecutivas producto de las TPS, la condensación de IPP y DMAPP para elongar la cadena de carbonos es producida por las preniltransferasas (Jia *et al.*, 2016). Debido a que las enzimas trans-PT funcionan como uno de los principales puntos de ramificación metabólica en el metabolismo terpenoide, no se consideró en este estudio a las enzimas cis-PT, además este trabajo se limitó a la identificación de secuencias que participan en la biosíntesis de isoprenoides de 10 a 20 carbonos, por lo tanto, entre las trans-PT identificadas, PT5 y PT9 que corresponden a SPS y PPS correspondientemente, no se tomaron en cuenta. La clasificación de las trans-PT identificadas, se llevó a cabo de acuerdo al modelo propuesto por You *et al.*, 2020A, en la cual, el grupo I se

nombró GGPP_LSU, en él se agrupan PaPT3 y PaPT9 y enzimas que en su forma homomérica funcionan como GGPS pero cuando forman complejos con los miembros del grupo GGPS_SSU tienen actividad de GPS, el grupo GFPS es específico de *A. thaliana* y produce GFPP de 25 carbonos; el grupo G(G)PS donde se encuentran PaPT2 y PaPT7 es el menos entendido, pero podría ser un tipo de la familia PT que comúnmente usa FPP como sustratos alílicos; los miembros del grupo GGPS-SSU-II, donde se agrupó la secuencia PaPT8, son catalíticamente inactivos pero capaces de unirse con los miembros del grupo GGPS-LSU y dirigir la producción de GGPP a GPP o aumenta la eficiencia catalítica (Zhou *et al.*, 2017), no hay diferencia clara entre los miembros de la subfamilia GGPS-SSU-I y GGPS-SSU-II pudiendo aumentar la eficiencia catalítica de tipo GGPS o promover la función de GPS; los miembros de la subfamilia GPS se encuentran en forma homomérica y producen GPP. Los miembros de la subfamilia FPS han sido bien caracterizados y producen FPP, entre grupo se encuentra PaPT4 y PaPT6. En todos los casos es importante determinar la localización de cada PT por medio de experimentos en sistemas biológicos debido a que los programas bioinformáticos no logran predecir eficazmente la localización de las proteínas, si bien la secuencia de aminoácidos es suficiente para saber la posible función de los miembros de esta familia, la localización determina hacia qué tipo de metabolitos se dirigirán los esqueletos sintetizados (You *et al.*, 2020A), por lo tanto, no puede determinarse con los datos obtenidos la participación de las diferentes PT identificadas con la producción de algún metabolito en específico.

El número de secuencias de terpenos sintasa encontrada en los transcriptomas de *Pentalinon andrieuxii* es bajo comparado con el número de secuencias de las especies de la figura 1.7, sin embargo, se debe tener en cuenta que solo se consideraron los transcriptomas de dos tejidos en dos estadios de desarrollo de la planta, de manera similar, en *Arabidopsis thaliana*, en tejidos de raíz se expresan únicamente cinco TPS de las 32 con las que cuenta en su genoma, mientras que en la hoja, solo se detecta la expresión de tres diferentes TPS, en flor se ha detectado la expresión de cuatro TPS, sin embargo, una de ellas también se expresa en hoja, por lo tanto entre los tres tejidos se expresan once TPS (Chen *et al.*, 2011). Se espera que en el genoma de *Pentalinon andrieuxii* se encuentren más genes TPS.

A pesar de contar con la secuencia de aminoácidos de los TPS identificados, no se puede predecir el producto final de cada enzima, Durairaj *et al.*, 2019 realizó un estudio con 262 sesquiterpenos sintasas caracterizadas y no encontró relación entre la secuencia de aminoácidos y su posible producto final, en el caso de esta familia de enzimas la única forma de determinar su producto final es mediante pruebas de actividad enzimática *in vitro*.

Existe evidencia de que varias TPS son enzimas multisustratos, en estudios in vitro, son capaces de sintetizar terpenos de diferente longitud de cadena dependiendo de los sustratos suministrados, sin embargo, dentro de las células, la disponibilidad de los diferentes sustratos se encuentra compartimentalizado y difícilmente una sesquiterpeno sintasa citosólica podría tener acceso a GPP para producir monoterpenos, o al contrario, una monoterpeno sintasa del cloroplasto tendría restricciones para disponer de FPP para producir diferentes sesquiterpenos, aun así, en algunas plantas se ha encontrado que podría estar presente depósitos de GPP en el citosol (Davidovich-Rikanati *et al.*, 2008 ; Gutensohn *et al.*, 2013 ; Pazouki *et al.*, 2015), lo que se puede observar en plantas transgénicas de *Nicotiana tabacum* que expresan un gen de *Perilla frutescens* que codifica a una enzima limoneno sintasa (LS) dirigida al citosol, se encontró la acumulación de limoneno, aunque en menor medida que las plantas que expresan LS dirigido a los cloroplastos (Ohara *et al.*, 2003).

La multifuncionalidad de las TPS debe ser tomado en cuenta cuando se tienen secuencias como la TPS5A y TPS5B las cuales carecen de péptido de tránsito al cloroplasto, a pesar de haberse agrupado en la subfamilia b, grupo que pertenece a monoterpenos sintasa e isopreno sintasa de angiospermas, de las que se espera que contengan un péptido señal en el extremo N-terminal que las dirija a los cloroplastos. La posibilidad de que una monoterpeno sintasa como TPS5 carente de péptido de tránsito se encuentre en el citosol, le permitiría actuar como sesquiterpeno sintasa teniendo disponibilidad de FPP, además de producir monoterpenos en caso de haber depósitos de GPP citosólicos. También cabe la posibilidad de que no se haya podido reconstruir la secuencia completa de TPS5 o que esté relacionado de alguna forma con la secuencia TPS5x debido a que en un segmento de la secuencia se alinean y comparten una similaridad arriba del 95% en dicho segmento del alineamiento, sin embargo, la expresión de la secuencia TPS5x solo se detectó en tejidos de hoja pudiendo ser genes diferentes.

Una característica de las enzimas TPS es que inician su reacción formando un carbocatión altamente reactivo en el sustrato que se convierte rápidamente en diferentes productos intermedios, lo que típicamente da lugar a múltiples productos terpénicos (Bohlmann y Keeling, 2008; Christianson, 2008). La especificidad del producto de diferentes terpenoides sintasas es muy variable y depende principalmente de qué tan bien se pueda estabilizar el carbocatión del sustrato en el centro activo de la enzima (Bohlmann y Keeling, 2008; Christianson, 2008). Tanto los sitios activos de monoterpenos como de sesquiterpenos sintasas contienen un conjunto de residuos completamente conservados que, cuando mutan, resultan en una pérdida completa de la actividad enzimática, el resto del sitio activo consta de posiciones menos conservadas, cuya

mutación puede conducir a un cambio en el perfil del producto de la enzima (Greenhagen *et al.*, 2006; Little *et al.*, 2002; Srividya *et al.*, 2015; Xu *et al.*, 2017). A pesar que dos TPS comparten los mismos aminoácidos en el sitio activo de la enzima, no necesariamente producirán los mismos productos, este hecho es observado en las TPS4 y TPS nueve de *Cannabis sativa* que, además, son 97% idénticos a nivel de proteína, pero TPS4 produce principalmente aloaromadendreno con productos secundarios que incluyen α -humuleno, mientras que TPS9 produce β -cariofileno y α -humuleno (Booth *et al.*, 2017), por el contrario, dos sesquiterpenos sintasas, una de *Arabidopsis lyrata* y otra *Zea perennis* comparten menos del 30% de identidad, sin embargo, ambas producen β -cariofileno (Durairaj *et al.*, 2019). Retomando el caso de las secuencias TPS5A y TPS5B, existen diferencias a lo largo de toda la secuencia cuando se comparan entre sí, por lo tanto, es seguro que produzcan diferentes productos. Para intentar explicar la diversidad de ambas secuencias, se teoriza que podría ser un gen polimórfico, sin embargo es posible encontrar dichas variaciones entre muestras de hoja y raíz adulta las cuales provienen del mismo individuo, lo mismo se repite para las muestras de planta joven, por lo tanto, otra alternativa podría ser un gen de reciente duplicación que conservan gran parte de la secuencia, o bien, tales diferencias podrían deberse a errores de secuenciación o ensamblado, aunque no se detectaron casos similares en las secuencias identificadas.

La identificación de secuencias de TPS caracterizadas de otras especies de plantas que compartan un alto grado de identidad con las TPS de *P. andrieuxii*, no asegura que produzcan los mismos productos, la única característica que se ha podido establecer para tener un indicio de la función de una secuencia de TPS es proporcionada por sus relaciones filogenéticas, dependiendo en la familia a la que se agrupe indicaría si podría ser un mono, sesqui o diterpeno sintasa, lo anterior, a pesar de ser cierto para la mayoría de los casos, se han identificado TPS de *Solanum lycopersicum* clasificadas en la subfamilia TPS-b que se encuentran en el citosol en lugar de los plástidos, de hecho se ha demostrado que de las TPS con que cuenta esta especie, seis monoterpenos sintasas son plastídicas y cuatro son citosólicas; las sesquiterpenos sintasas son casi todas citosólicas, con la excepción de una que se encuentra en las mitocondrias; y tres diterpenos sintasas se encuentran en los plástidos, uno en el citosol y dos en las mitocondrias (Zhou *et al.*, 2020).

Con la información analizada hasta el momento y partiendo del supuesto de que los urechitoles deriven de los sesquiterpenos, las secuencias candidatas que podrían estar involucradas en la biosíntesis de los urechitoles son TPS10 si el lugar de la biosíntesis son las hojas, por el contrario, si el lugar de biosíntesis es la raíz, el candidato sería TPS8, en caso de no ser ninguna de estas,

se debe considerar las isoformas de la enzima TPS5, así como TPS2 y TPS7. En el caso que los urechitoles se sinteticen en los cloroplastos, únicamente TPS1 y TPS9 podrían ser candidatos debido a que los miembros de la subfamilia tps-g son conocidos por producir terpenos lineales y la única evidencia sobre la biosíntesis de un trisnorsesquiterpenos sugiere que se necesita un alcohol sesquiterpénico cíclico además de la participación adicional de una enzima de la familia de los citocromos P450 (Chen *et al.*, 2019). Por lo tanto, se debe complementar este estudio con la identificación de secuencias cyp P450 y caracterizar las enzimas identificadas de los diferentes TPS y en combinación con las cyp P450. Se considera que el sistema de *N. benthamiana* es altamente eficiente para la caracterización de TPS (Lau *et al.*, 2020) por lo tanto, podría ser un buen modelo para la identificación de la biosíntesis del urechitol.

A pesar de que se han identificado numerosos metabolitos de *Pentalinon andrieuxii* como el taraxasterol, el pentalinonsterol, el pentalinonside, el ácido betulínico y el urechitol (Pan *et al.*, 2012; Yam-Puc *et al.*, 2009), únicamente de los últimos dos se tiene estudios sobre su dinámica de producción espacio-temporal (Hiebert-Giesbrecht *et al.*, 2016). El ácido betulínico se produce principalmente en hojas durante todos los estadios de desarrollo de la planta, mientras que el urechitol A se encuentra principalmente en raíces de plantas de edad intermedia y adulta, sin embargo, también se detectó en raíces y tallos en esas mismas condiciones. El mayor contenido de urechitol A se obtuvo de raíces de plantas adultas (Hiebert-Giesbrecht *et al.*, 2016), por lo tanto, se esperaba encontrar alguna TPS que se encuentre expresada en mayor medida en la muestra de raíz de planta adulta, sin embargo, ninguna TPS tuvo una mayor expresión *in silico* en la muestra anteriormente mencionada. Tomando en consideración la biosíntesis del discodieno, un trisnorsesquiterpeno, se necesita la participación de una enzima de la familia CYP (Chen *et al.*, 2019), dicha familia no se incluyó en análisis de este trabajo.

CAPÍTULO V

CONCLUSIONES Y PERSPECTIVAS

5.1 CONCLUSIONES

Las tecnologías de ARN-seq son una herramienta útil y actualmente la mejor alternativa para identificar genes potencialmente involucrados en algún proceso biológico específico, como la biosíntesis de un metabolito novedoso en una especie de la cual no se disponga su genoma de referencia como lo es *Pentalinon andrieuxii* que biosintetiza el urechitol, por otra parte, el enfoque de ensamblado *de novo* de transcriptomas, permitió ensamblar un conjunto único de transcritos que servirá como transcriptoma de referencia para esta especie.

La metodología desarrollada para identificar los genes de interés de *Pentalinon andrieuxii* resultó efectiva para la identificación de secuencias relacionadas con la biosíntesis de terpenos en los transcriptomas de *Pentalinon andrieuxii*, pero que además puede ser empleada para identificar cualquier gen o familia de genes diferentes a los tratados en este trabajo.

Se lograron identificar todos los genes que codifican para todas las secuencias de las enzimas de la ruta MVA y MEP, la mayoría posiblemente sean de copia única o solo generen una isoforma, los únicos que no cumplen con esta característica son las secuencias de AACT de las cuales se encontraron dos secuencias diferentes, HMGR la cual posee dos secuencias y DXS el cual posee cuatro secuencias.

Se identificaron diez preniltransferasas, de las cuales dos son FPPS, dos GGPPS y cuatro secuencias que requieren caracterización bioquímica para determinar su función como GPPS o GGPPS, además determinar la localización celular de cada una de las preniltransferasas ayudaría a tener mayor información sobre su participación en el metabolismo isoprenoide.

Se logró identificar cuatro monoterpenos sintasas, cuatro sesquiterpenos sintasas, tres diterpenos sintasas, las cuales dos son del metabolismo de las giberelinas. En particular la secuencia TPS5 a pesar de ser clasificada como monoterpeno sintasa, no se le detectó péptido señal al cloroplasto además de múltiples isoformas que requieren ser estudiadas más a detalle.

Los patrones de expresión de los genes estudiados en este trabajo no se relacionan con la biosíntesis de los isoprenoides como el urechitol y el ácido betulínico en *P. andrieuxii*, debido a que no se encontró un gen que tuviera mayor expresión en los tejidos donde estos metabolitos

presentan los niveles más altos de biosíntesis, se requiere complementar este trabajo con los estudios de otras familias de genes como los CYP.

Factores a nivel postraduccional podrían estar regulando en mayor medida la biosíntesis de los isoprenoides estudiados, por lo tanto, los genes TPS identificados no se deben descartar como candidatos para la biosíntesis de urechitol. El gen TPS más expresado en raíz corresponde a TPS8 que es un sesquiterpeno sintasa, suponiendo que el lugar de biosíntesis de los urechitoles sea la raíz, esta secuencia sería el un candidato potencial, por el contrario, si se sintetiza en la hoja, las secuencias tps2 y tps10 que corresponden a sesquiterpenos sintasas podrían ser candidatas.

5.2 PERSPECTIVAS

Se recomienda clonar y secuenciar la longitud completa de cada TPS para confirmar el ensamblado de cada transcrito y si algún gen cuenta con isoformas derivadas del procesamiento alternativo. La secuencia completa de TPS1 se podría obtener mediante la tecnología de clonación RACE.

Con la información del transcriptoma ensamblado de *Pentalimon andrieuxii* se pueden identificar genes de la familia cyp P450 y de otras familias que se tenga evidencia que puedan modificar los esqueletos isoprenoides que generen las enzimas TPS. Algunos candidatos identificados se deberían clonar para su posterior expresión heteróloga en parejas TPS-CYP P450 para ver si alguna logra producir urechitol, además se podrían identificar metabolitos aun no elucidados que las enzimas de *P. andrieuxii* tiene el potencial de producir.

El patrón de expresión obtenidos de cada gen identificado se debe corroborar mediante qPCR, esta técnica podría proveer el soporte estadístico necesario para identificar si algún gen está expresado de manera diferencial en alguna condición y si se correlaciona con la producción de metabolitos.

BIBLIOGRAFÍA

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol*. Nov 6;18(12):1435-40.

Amini, H., Naghavi, M. R., Shen, T., Wang, Y., Nasiri, J., Khan, I. A., Fiehn, O., Zerbe, P., & Maloof, J. N. (2019). Tissue-specific transcriptome analysis reveals candidate genes for terpenoid and phenylpropanoid metabolism in the medicinal plant *Ferula assafoetida*. *G3* (Bethesda, Md.), 9(3), 807–816.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> [Acceso 10 septiembre 2021].

Argüeta, A., Cano, L. y Rodarte, M. (1994). Atlas de las plantas de la medicina tradicional mexicana. México: Instituto Nacional Indigenista; vol. 2.

Bach, T. J. (1995). Some new aspects of isoprenoid biosynthesis in plants—a review. *Lipids* 30: 191–202.

Bach, J. y Rohmer, M. (2014). Isoprenoid biosynthesis in plants and microorganisms. New concepts and experimental approaches. *Springer Science*. Université de Strasbourg, France. 441.

Barbier, F. F., Chabikwa, T. G., Ahsan, M. U., Cook, S. E., Powell, R., Tanurdzic, M., & Beveridge, C. A. (2019). A phenol/chloroform-free method to extract nucleic acids from recalcitrant, woody tropical species for gene expression and sequencing. *Plant methods*, 15, 62.

van Bakel, H., Nislow, C., Blencowe, B. J., y Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biology*, 8(5), e1000371.

Bentley, D. L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell. Biol*, Jun;17(3):251-6.

Boekhorst, J., y Snel, B. (2007). Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinformatics*, 8, 356.

Bohlmann, J. y Keeling, C. I. (2008). Terpenoid biomaterials. *Plant J.* 54:656–69

Bohlmann, J., Meyer-Gauen, G. & Croteau, R (1998). Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* 95, 4126–4133.

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

Booth, J., Page, J. y Bohlmann, J. (2017). Terpene synthases from *Cannabis sativa*. *PLoS One*, 12: e0173911.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, 41. W29–W33

Boucher, Y. y Doolittle, W. F. (2000). The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol.* 37:703–16

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368.

Buckingham, J., Cooper, C. y Purchase, R. (2016). Natural Products Desk Reference; CRC Press, Taylor & Francis Group: Boca Raton.

Camacho, C., Coulouris, G., Avagyan, V., Ning, Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 421.

-
- Campos, N., & Boronat, A. (1995). Targeting and topology in the membrane of plant 3-hydroxy-3-methylglutaryl coenzyme A reductase. *The Plant Cell*, 7(12),
- Cano-Flores A. (2013). Biotransformación de triterpenos con diferentes microorganismos. *Revista Mexicana de Ciencias Farmacéuticas, Asociación Farmacéutica Mexicana, A.C. Distrito Federal, México*. 44(2): 7-16.
- Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., Mead, M. E., Prisic, S., Wang, K., Lin, F. Y., Chang, T. K., Peters, R. J., & Oldfield, E. (2010). Diterpene cyclases and the nature of the isoprene fold. *Proteins*, 78(11),
- Carretero-Paulet, L., Cairo, A., Talavera, D., Saura, A., Imperial, S., Rodriguez-Concepcion, M., Campos, N., Boronat, A. (2013). Functional and evolutionary analysis of DXL1, a non-essential gene encoding a 1-deoxy-D-xylulose 5-phosphate synthase like protein in *Arabidopsis thaliana*. *Gene*, 524(1):40–53.
- Carretero-Paulet, L., Cairo, A., Botella-Pavia, P., Besumbes, O., Campos, N., Boronat, A., Rodriguez-Concepcion, M. (2006). Enhanced flux through the methylerythritol 4-phosphate pathway in *Arabidopsis* plants overexpressing deoxyxylulose 5-phosphate reductoisomerase. *Plant Mol Biol*, 62(4–5):683–95.
- Chan-Bacab, M., Balanza, E., Deharo, E., Muñoz, V., Durán-García, R., y Peña-Rodríguez, L. (2003). Variation of leishmanicidal activity in four populations of *Urechites andrieuxii*. *J Ethnopharmacol*, (86), 243–247.
- Chen, F., Tholl, D., Bohlmann, J., Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J*. 66: 212–229. pmid:21443633.
- Chen, X., Köllner, T. G., Jia, Q., Norris, A., Santhanam, B., Rabe, P., Dickschat, J. S., Shaulsky, G., Gershenzon, J., & Chen, F. (2016). Terpene synthase genes in eukaryotes beyond plants and fungi: Occurrence in social amoebae. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43), 12132–12137.
- Chen, X., Luck, K., Rabe, P., Dinh, C. Q., Shaulsky, G., Nelson, D. R., Gershenzon, J., Dickschat, J. S., Köllner, T. G., & Chen, F. (2019). A terpene synthase-cytochrome P450 cluster in *Dictyostelium discoideum* produces a novel trisnorsesquiterpene. *eLife*, 8, e44352.

Christianson, D. W. (2017). Structural and chemical biology of terpenoid cyclases. *Chem Rev* 117:11570–11648.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 13.

Cordoba, E., Porta, H., Arroyo, A., San Roman, C., Medina L., Rodriguez-Concepcion, M., Leon, P. (2011). Functional characterization of the three genes encoding 1-deoxy-D-xylulose 5-phosphate synthase in maize. *J. Exp. Bot*, 62(6):2023–38.

Cui, G., Duan, L., Jin, B., Qian, J., Xue, Z., Shen, G., Snyder, J. H., Song, J., Chen, S., Huang, L., Peters, R., Qi, X. (2015). Functional divergence of diterpene syntheses in the medicinal plant *Salvia miltiorrhiza* Bunge. *Plant Physiology*, 169: 1607–1618.

Darabi, M., Masoudi-Nejad, A., Nemat-Zadeh, G. (2012). Bioinformatics study of the 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGR) gene in Gramineae. *Mol Biol Rep*, 39: 8925–8935.

Dash, S., Rahman, S. R., Hines, H. M., Feng, W. (2021) iBLAST: Incremental BLAST of new sequences via automated e-value correction. *PLoS ONE* 16(4): e0249410.

Domínguez-Carmonam D., Escalante-Erosam F., García-Sosam K., Ruiz-Pinellm G., Gutierrez-Yapum D., Chan-Bacab M., Giménez-Turbam A. y Peña-Rodríguez L. (2010). Antiprotozoal activity of betulinic acid derivatives. *Phytomedicine* 17: 379–382.

Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., and van Dijk, A. D. (2019). An analysis of characterized plant sesquiterpene synthases. *Phytochemistry*, 158, 157–165.

Enfissi, E., Fraser, P., Lois, L., Boronat, A., Schuch, W. y Bramley, P. (2005). Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health promoting isoprenoids in tomato. *Plant Biotechnol J*. 3: 17-27.

Estévez, J. M., Cantero, A., Reindl, A., Reichler, S., Leon, P. (2001). 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. *J Biol Chem*, 276(25):22901–9.

Estévez, J. M., Cantero, A., Romero, C., Kawaide, H., Jiménez, L. F., Kuzuyama, T., Seto, H., Kamiya, Y., y León, P. (2000). Analysis of the expression of CLA1, a gene that encodes the 1-deoxyxylulose 5-phosphate synthase of the 2-C-methyl-D-erythritol-4-phosphate pathway in *Arabidopsis*. *Plant Physiology*, 124(1), 95–104.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* 19: 99–113. pmid:5449325.

Floss, D.S., Hause, B., Lange, P.R., Kuster, H., Strack, D., Walter, M.H. (2008) Knock-down of the MEP pathway isogene 1-deoxy-D-xylulose 5-phosphate synthase 2 inhibits formation of arbuscular mycorrhiza-induced apocarotenoids, and abolishes normal expression of mycorrhiza-specific plant marker genes. *Plant J*, 56(1):86–100.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant. Biol.*, 60: 433–453. pmid:19575588.

Gahlan, P., Singh, H. R., Shankar, R., Sharma, N., Kumari, A., Chawla, V., Ahuja, P. S., & Kumar, S. (2012). *De novo* sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics*, 13, 126.

Gambino, G., Perrone, I., Gribaudo, I.A. (2008) Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. *Phytochem Anal*, 19(6):520-5.

Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Natural Product Reports*, 29(10), 1153–1175.

Geniza, M., Jaiswal, P. (2017). Tools for building *de novo* transcriptome assembly. *Curr Plant Biol*, 11:41–5.

Getti, G., Durgadoss, P., Domínguez-Carmona, D., Martín-Quintal, Z., Peraza-Sánchez, S., Peña-Rodríguez, L.M., Humber, D. (2009). Leishmanicidal activity of Yucatecan medicinal plants on *Leishmania* species responsible for cutaneous leishmaniasis. *J Parasitol*, 95(2):456-60.

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S. K., Gabaldón, T., Huerta-Cepas, J., Martín, M. J., Muffato, M., Patricio, M., Pereira, C., da Silva, A. S., Wang, Y., Sonnhammer, E., y Thomas,

P. D. (2019). Advances and applications in the quest for orthologs. *Molecular Biology and Evolution*, 36(10), 2157–2164.

Goldstrohm, A.C., Greenleaf, A.L., Garcia-Blanco, M.A. (2001). Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene*. 2001 Oct 17;277(1-2):31-47.

Góngora-Castillo, E., Buell, C.R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Nat Prod Rep*, 30:490–500.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–652.

Greenhagen, B. T., O'Maille, P. E., Noel, J. P., & Chappell, J. (2006). Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proceedings of the National Academy of Sciences of the United States of America*, 103(26), 9826–9831.

Guerriero, G., Berni, R., Muñoz-Sanchez, J.A., Apone, F., Abdel-Salam, E.M., Qahtan, A.A., Alatar, A.A., Cantini, C., Cai, G., Hausman, J.-F., Hernández-Sotomayor, S.M.T., Faisal, M. (2018). Production of plant secondary metabolites: examples, tips and suggestions for biotechnologists. *Genes*, 9, 309.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307-21,

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, 33(Web Server issue), W557–W559.

Guirimand, G., Guihur, A., Phillips, M.A., Oudin, A., Glévarec, G., Melin, C., Papon, N., Clastre, M., St-Pierre, B., Rodríguez-Concepción, M., Burlat, V., Courdavault, V. A. (2012) single gene

encodes isopentenyl diphosphate isomerase isoforms targeted to plastids, mitochondria and peroxisomes in *Catharanthus roseus*. *Plant. Mol. Biol.* 2012 Jul;79(4-5):443-59.

Gupta, G., Peine, K. J., Abdelhamid, D., Snider, H., Shelton, A. B., Rao, L., Kotha, S. R., Huntsman, A. C., Varikuti, S., Oghumu, S., Naman, C. B., Pan, L., Parinandi, N. L., Papenfuss, T. L., Kinghorn, A. D., Bachelder, E. M., Ainslie, K. M., Fuchs, J. R., & Satoskar, A. R. (2015). A novel sterol isolated from a plant used by mayan traditional healers is effective in treatment of visceral *leishmaniasis* caused by *Leishmania donovani*. *ACS Infectious Diseases*, 1(10), 497–506.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., ... Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512.

Ha, S.H., Lee, S.W., Kim, Y.M., Hwang, Y.S. (2001) Molecular characterization of Hmg2 gene encoding a 3-hydroxymethylglutaryl-CoA reductase in rice. *Mol. Cell.* 11: 295–302.

Hamberger, B., y Bak, S. (2013). Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1612), 20120426.

Hartley, S.W., y Mullikin, J.C. (2015). QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics*, 16(1), 224.

Hasunuma, T., Takeno, S., Hayashi, S., Sendai, M., Bamba, T., Yoshimura, S., Tomizawa, K., Fukusaki, E., Miyake, C. (2008). Overexpression of 1-Deoxy-D-xylulose-5-phosphate reductoisomerase gene in chloroplast contributes to increment of isoprenoid production. *J. Biosci. Bioeng*, 105(5):518–26.

Hayashi, K., Kawaide, H., Notomi, M., Sakigi, Y., Matsuo, A. and Nozaki, H. (2006) Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *Physcomitrella patens*. *FEBS Lett.* 580, 6175–6181.

Hemmerlin, A. (2013). Post-translational events and modifications regulating plant enzymes involved in isoprenoid precursor biosynthesis. *Plant Sci* 203:41–54.

Hemmerlin, A., Harwood, J. y Bach, T. (2012). A raison d'être for two distinct pathways in the early steps of plant isoprenoid biosynthesis?. *Prog Lipid Res*, 51:95-148.

Hiebert-Giesbrecht, M.R., C.Y. Novelo-Rodríguez, G.R. Dzib, L.M. Calvo-Irabién, G. von Arx, y L.M. Peña-Rodríguez. (2017). Herb-chronology as a tool for determining the age of perennial forbs in tropical climates. *Botany* 96 (1):73–78.

Hiebert-Giesbrecht, M., Escalante-Erosa, F., García-Sosa, K., Dzib, G., Calvo-Irabién, L. y Peña-Rodríguez, L. (2016). Spatio-temporal variation of terpenoids in wild plants of *Pentalinon andrieuxii*. *Chem. Biodiversity*, 13: 1521-1526.

Hiebert-Giesbrecht, M.R., Avilés-Berzunza, E., Godoy-Hernández, G., Peña-Rodríguez, L. (2021). Genetic transformation of the tropical vine *Pentalinon andrieuxii* (Apocynaceae) via *Agrobacterium rhizogenes* produces plants with an increased capacity of terpenoid production. *In Vitro Cell.Dev.Biol.-Plant* 57, 21–29.

Hofberger, J.A., Ramirez, A.M., Bergh, E.v., Zhu, X., Bouwmeester, H.J., Schuurink, R.C., y Schranz, M.E. (2015). Large-scale evolutionary analysis of genes and supergene clusters from terpenoid modular pathways provides insights into metabolic diversification in flowering plants. *PLoS one*, 10(6), e0128808.

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, 8(5), giz039.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35, W585–W587.

Huang, L., Li, J., Ye, H., Li, C., Wang, H., Liu, B. y Zhang, Y. (2012). Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in *Catharanthus roseus*. *Planta*. 236: 1571–1581.

Ikram, N.K.B.K., Zhan, X., Pan, X.-W., King, B.C. y Simonsen, H.T. (2015). Stable heterologous expression of biologically active terpenoids in green plant cells. *Front. Plant Sci.* 6, 129.

Jayashree, R., Nazeem, P., Rekha, K., Sreelatha, S., Thulaseedharan, A., Krishnakumar, R., Kala, R., Vineetha, M., Leda, P., Jinu, U. y Venkatachalam, P. (2018). Over-expression of 3-hydroxy-3-methylglutaryl-coenzyme A reductase 1 (hmg1) gene under super-promoter for

enhanced latex biosynthesis in rubber tree (*Hevea brasiliensis* Muell. Arg. *Plant Physiology and Biochemistry*, 127: 414-424

Jia, Q. Chen, F. (2016). Catalytic functions of the isoprenyl diphosphate synthase superfamily in plants: A growing repertoire. *Mol. Plant*, 9, 189–191.

Jiu J. (1966). A survey of some medicinal plants of Mexico for selected biological activities. *Lloydia* 29, 250–259.

Joyard, J., Ferro, M., Masselon, C., Seigneurin-Berny, D., Salvi, D., Garin, J., Rolland, N. (2009). Chloroplast proteomics and the compartmentation of plastidial isoprenoid biosynthetic pathways. *Mol Plant*, 2(6):1154-80.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., & Triche, T. J. (2011). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biology*, 9, 86.

Katoh, K., y Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772-780.

Keller, O., Kollmar, M., Stanke, M., Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 15;27(6):757-63

Kim, B.R., Kim, S.U., Chang, Y.J. (2005). Differential expression of three 1-deoxy-D-xylulose-5-phosphate synthase genes in rice. *Biotechnol Lett*, 27(14):997–1001.

Kim, S.M., Kuzuyama, T., Chang, Y.J., Song, K.S., Kim, S.U. (2006). Identification of class 2 1-deoxy-D-xylulose 5-phosphate synthase and 1-deoxy-D-xylulose 5-phosphate reductoisomerase genes from *Ginkgo biloba* and their transcription in embryo culture with respect to ginkgolide biosynthesis. *Planta Med*. 2006;72(3):234–40.

Kliebenstein, D.J. (2004). Secondary metabolites and plant/environment interactions: a view through *Arabidopsis thaliana* tinted glasses. *Plant Cell Environ*. 27:675–84

Kobayashi, T, Kato-Emori, S, Tomita, K, Ezura, H. (2002). Detection of 3-hydroxy-3-methylglutaryl-coenzyme A reductase protein Cm-HMGR during fruit development in melon (*Cucumis melo* L.). *Theor Appl Genet*, 104(5):779–785.

-
- Koen, V.D.B., Katharina, M.H., Charlotte, S., Simone, T., Lieven, C., Michael, I.L., Rob, P., Mark, D.R. (2019). RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.*, 2:139–173.
- Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics 1. *Annu Rev Genet* 39: 309–338. pmid:16285863.
- Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., & Nogues, G. (2004). Multiple links between transcription and splicing. *RNA*, 10(10), 1489–1498.
- Kumar, Y., Khan, F., Rastogi, S., Shasany A.K. (2018). Genome-wide detection of terpene synthase genes in holy basil (*Ocimum sanctum* L.). *PLoS ONE*, 13(11): e0207097.
- Langmead, B. y S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359.
- Lau, K. H., Bhat, W. W., Hamilton, J. P., Wood, J. C., Vaillancourt, B., Wiegert-Rininger, K., Newton, L., Hamberger, B., Holmes, D., Hamberger, B., y Buell, C. R. (2020). Genome assembly of *Chiococca alba* uncovers key enzymes involved in the biosynthesis of unusual terpenoids. *DNA Research*, 27(3), dsaa013.
- Laule, O., Fürholz, A., Chang, H., Zhu, T., Wang, X., Heifetz, P., Grisse, W. y Lange, M. (2003). Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 100:6866–71.
- Leivar, P., Antolín-Llovera, M., Ferrero, S., Closa, M., Arró, M., Ferrer, A., Boronat, A., & Campos, N. (2011). Multilevel control of *Arabidopsis* 3-hydroxy-3-methylglutaryl coenzyme A reductase by protein phosphatase 2A. *The Plant Cell*, 23(4), 1494–1511.
- Leivar, P., González, V. M., Castel, S., Trelease, R. N., López-Iglesias, C., Arró, M., Boronat, A., Campos, N., Ferrer, A., & Fernández-Busquets, X. (2005). Subcellular localization of *Arabidopsis* 3-hydroxy-3-methylglutaryl-coenzyme A reductase. *Plant Physiology*, 137(1), 57–69.
- Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., & Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, 7(9), 709–715.

-
- Lezama-Dávila, C., Isaac-Márquez, A., Zamora-Crescencio, P., Úc-Encalada, M., Justiniano-Apolinar, S., del Angel-Robles, L., Satoskar, A. y Hernández-Rivero L. (2007). Leishmanicidal activity of *Pentalinon andrieuxii*. *Fitoterapia*, 78(3), 255-257.
- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15(12), 553. <https://doi.org/10.1186/s13059-014-0553-5>
- Li, S., Wang, Y., Zhao, Y., Zhao, X., Chen, X., & Gong, Z. (2020). Global Co-transcriptional Splicing in *Arabidopsis* and the Correlation with Splicing Regulation in Mature RNAs. *Molecular Plant*, 13(2), 266–277.
- Li, X., Nair, A., Wang, S., Wang, L. (2015). Quality control of RNA-seq experiments. *Methods Mol Biol.*,1269:137-46.
- Li, W., Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 1;22(13):1658-9.
- Li, W., Liu, W., Wei, H., He, Q., Chen, J., Zhang, B., Zhu, S. (2014). Species-specific expansion and molecular evolution of the 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) gene family in plants. *PLoS One*, 10;9(4):e94172.
- Liu, J., Xiong, H., Cheng, Y., Cui, C., Zhang, X., Xu, L., Xu, L. y Zhang, X. (2013). Effects of taraxasterol on ovaalbumin induced allergy in mice. *J Ethnopharmacol.* 2013;148: 787–93.
- Llovet, J. (2005). Updated treatment approach to hepatocellular carcinoma. *J Gastroenterol* 40: 225–235.
- Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., & Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biology*, 10(2), 207.
- Loreto, F., Pollastri, S., Fineschi, S. y Velikova, V. (2014). Volatile isoprenoids and their importance for protection against environmental constraints in the Mediterranean area. *Environ. Exp. Bot.* 103, 99–106.

-
- Lu, B., Zeng, Z., Shi, T. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci. China Life Sci.* 56, 143–155.
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, 13(5), e1005457.
- Ma, Y., Yuan, L., Wu, B., Li, X., Chen, S., & Lu, S. (2012). Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. *Journal of Experimental Botany*, 63(7), 2809–2823. <https://doi.org/10.1093/jxb/err466>
- Madden, T. The BLAST sequence analysis tool. In NCBI Handbook, ed. J. McEntyre and J. Ostell (National Library of Medicine, Bethesda, MD, 2005).
- Martin, J.A., Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682.
- Mendoza-Poudereux, I., Munoz-Bertomeu, J., Arrillaga, I., Segura, J. (2014). Deoxyxylulose 5-phosphate reductoisomerase is not a rate-determining enzyme for essential oil production in spike lavender. *J Plant Physiol*, 171(17):1564–70.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315–327.
- Misawa, N. (2011). Pathway engineering for functional isoprenoids. *Current Opinion in Biotechnology*, 22(5), 627–633.
- Moghaddam, M., Ahmad, F. y Samzadeh, A. (2012). Biological activity of betulinic acid: A review. *Pharmacology & Pharmacy*, 3:119-123.
- Moreton, J., Izquierdo, A., & Emes, R. D. (2016). Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. *Frontiers in Genetics*, 6, 361.
- Mors, W., Nascimento, M., Pereira, B., Pereira, N. (2000). Plant natural products active against snake bite, the molecular approach. *Phytochem*; 55: 627–42.

Munoz-Bertomeu, J., Arrillaga, I., Ros, R., Segura, J. (2006). Up-regulation of 1-deoxy-D-xylulose-5-phosphate synthase enhances production of essential oils in transgenic spike lavender. *Plant Physiol*, 142(3):890–900.

Murray, H. (2001). Clinical and experimental advances in treatment of visceral leishmaniasis. *Antimicrobial Agents and Chemotherapy* 45: 2185–2197.

Oghumu, S., Varikuti, S., Saljoughian, N., Terrazas, C., Huntsman, A., Parinandi, N. y Satoskar, A. (2017). Pentalinosterol, a constituent of *Pentalinon andrieuxii*, possesses potent immunomodulatory activity and primes T cell immune responses. *Journal of Natural Products*, 80(9), 2515–2523.

Ohara, K., Ujihara, T., Endo, T., Sato, F., Yazaki, K. (2003). Limonene production in tobacco with *Perilla limonene* synthase cDNA. *J. Exp. Bot.*, Dec;54(393):2635-42.

Ohno, S. (1970) Evolution by gene duplication. New York: Springer Publishing Group. 160 p

Okada, A., Shimizu, T., Okada, K., Kuzuyama, T., Koga, J., Shibuya, N., Nojiri, H., Yamane, H. (2007). Elicitor induced activation of the methylerythritol phosphate pathway toward phytoalexins biosynthesis in rice. *Plant. Mol. Biol.*, Sep;65(1-2):177-87.

Okada, K., Kasahara, H., Yamaguchi, S., Kawaide, H., Kamiya, Y., Nojiri, H., Yamane, H. (2008). Genetic evidence for the role of isopentenyl diphosphate isomerases in the mevalonate pathway and plant development in *Arabidopsis*. *Plant. Cell. Physiol.*, Apr;49(4):604-16.

Omura, T., Watanabe, S., Iijima, Y., Aoki, K., Shibata, D., Ezura, H. (2007). Molecular and genetic characterization of transgenic tomato expressing 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Plant Biotechnology* 24: 107–115.

Pan, L., Lezama-Dávila, C., Isaac-Marquez, A., Calomeni, E., Fuchs, J., Satoskar, A. y Kinghorn, A. (2012). Sterols with antileishmanial activity isolated from the roots of *Pentalinon andrieuxii*. *Phytochemistry*.82:128-135.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417–419.

Peña-Rodríguez, L.M., Yam-Puc, A., Knispel, N., Schramek, N., Huber, C., Graßberger, C., Ramírez-Torres, F.G., Escalante-Erosa, F., García-Sosa, K., Hiebert-Giesbrecht, M.R., Chan-

Bacab, M.J., Godoy-Hernández, G., Bacher, A., Eisenreich, W. (2014). Isotopologue profiling of triterpene formation under physiological conditions. Biosynthesis of lupeol-3-(3'-R-hydroxy)-stearate in *Pentalinon andrieuxii*. *J. Org. Chem*, Apr 4;79(7):2864-73.

Phillips, M.A., D'Auria, J.C., Gershenzon, J., Pichersky, E. (2008). The *Arabidopsis thaliana* type I isopentenyl diphosphate isomerases are targeted to multiple subcellular compartments and have overlapping functions in isoprenoid biosynthesis. *Plant Cell* 20:677–97.

Phillips, M.A., León, P., Boronat, A., Rodríguez-Concepción, M. (2008). The plastidial MEP pathway: unified nomenclature and resources. *Trends Plant Sci.* 13:619–23

Pichersky, E. y Raguso, R. (2018). Why do plants produce so many terpenoids compounds?. *New Phytol.* 220(3):692-702.

Qiao, D., Yang, C., Chen, J., Guo, Y., Li, Y., Niu, S., Cao, K., & Chen, Z. (2019). Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (*Camellia sinensis*). *Scientific Reports*, 9(1), 2709.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286.

Robinson, M.D., y Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), R25.

Rokas A. Phylogenetic analysis of protein sequence data using the Randomized Accelerated Maximum Likelihood (RAXML) Program. *Curr Protoc Mol Biol.* 2011 Oct;Chapter 19:Unit19.11.

Rzedowski, J. y Calderon, G. (1998). *Apocynaceae*. En flora del bajo y de regiones adyacentes. Fascículo 70. Pátzcuaro, Michoacán: Instituto de Ecología.

Sapir-Mir, M., Mett, A., Belausov, E., Tal-Meshulam, S., Frydman, A., Gidoni, D., & Eyal, Y. (2008). Peroxisomal localization of *Arabidopsis* isopentenyl diphosphate isomerases suggests that part of the plant isoprenoid mevalonic acid pathway is compartmentalized to peroxisomes. *Plant Physiology*, 148(3), 1219–1228.

Schaarschmidt, S., Fischer, A., Zuther, E., y Hinch, D. K. (2020). Evaluation of seven different rna-seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *International Journal of Molecular Sciences*, 21(5), 1720.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D. y Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Simpson, K., Quiroz, L.F., Rodriguez-Concepcion, M., Stange, C.R. (2016). Differential contribution of the first two enzymes of the MEP pathway to the supply of metabolic precursors for carotenoid and chlorophyll biosynthesis in carrot (*Daucus carota*). *Front. Plant. Sci.* ;7:1344.

Smith-Unna, R., Bournnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26(8), 1134–1144.

Sommerfeld D., Lingner T., Stanke M., Morgenstern B., Richter H. (2009) AUGUSTUS at MediGRID: adaption of a bioinformatics application to grid computing for efficient genome analysis. *Future Gener Comput Syst.* 25, 337 – 345.

Srividya, N., Davis, E.M., Croteau, R.B., Lange, B.M. (2015). Functional analysis of (4 S)-limonene synthase mutants reveals determinants of catalytic outcome in a model monoterpene synthase. *Proc Natl Acad Sci.*, 112(11):3332–7. pmid:25733883.

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, Mar 1;24(5):637-44.

Stanke, M., Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, Oct;19 Suppl 2:ii215-25

Steele, C.L., Crock, J., Bohlmann, J., Croteau, R. (1998). Sesquiterpene synthases from grand fir (*Abies grandis*). Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of delta-selinene synthase and gamma-humulene synthase. *J Biol Chem*, 23;273(4):2078-89.

Sterner, K.N., Raaum, R.L., Zhang, Y.P., Stewart, C.B., Disotell T.R. (2006). Mitochondrial data support an odd-nosed colobine clade. *Mol. Phylogenet. Evol.* 40(1):1-7.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., & Yaspo, M. L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*, 15(1), 675.

Suzuki, H., Fukushima, E. O., Umemoto, N., Ohyama, K., Seki, H., & Muranaka, T. (2018). Comparative analysis of CYP716A subfamily enzymes for the heterologous production of C-28 oxidized triterpenoids in transgenic yeast. *Plant Biotechnology* (Tokyo, Japan), 35(2), 131–139.

Tan, Y., Yu, R. y Pezzuto, J. (2003). Betulinic acid-induced programmed cell death in human melanoma cells involves mitogen-activated protein kinase activation. *Clinical Cancer Research* 9(7): 2866-75.

Tholl, D. 2015. Biosynthesis and biological functions of terpenoids in plants. *Adv Biochem Eng Biot.* 148: 63–106.

Thoma, R., Schulz-Gasch, T., D'Arcy, B., Benz, J., Aebi, J., Dehmlow, H., Hennig, M., Stihle, M., Ruf, A. (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature*, 432(7013):118-22.

Urban, P., Mignotte, C., Kazmaier, M., Delorme F. y Pompon, D. (1997). Cloning, yeast expression, and characterization of the coupling of two distantly related *Arabidopsis thaliana* NADPH-cytochrome P450 reductases with P450 CYP73A5. *J. Biol. Chem.* 272: 19176–19186.

Vaccaro, M., Malafronte, N., Alfieri, M., Tommasi, N., Leone, A. (2014). Enhanced biosynthesis of bioactive abietane diterpenes by overexpressing AtDXS or AtDXR genes in *Salvia sclarea* hairy roots. *Plant Cell Tissue Organ Cult.*, 119:65–77.

Van Bel, M., Bucchini, F., Vandepoele, K. (2019). Gene space completeness in complex plant genomes. *Curr. Opin. Plant. Biol.*, 48:9-17.

Viscencio de la Rosa G., Tamay-Segovia P., Issac-Márquez A. y Lezama-Dávila C. (1995). Toxicidad *in vitro* de extractos de *Urechites andrieuxii* Muell.-Arg. en contra de *L. mexicana*. Memorias de la III Reunión de Investigación Química en el Sureste de México, Mérida, Yucatán, p. 93

Vranová, E., Coman, D. y Gruissem, W. (2013). Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annual Review of Plant Biology*, 64(1), 665-700.

Wagner, G.P., Kin, K. y Lynch, V.J. (2013). A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci.* 132, 159–164.

Walter, MH, Hans, J, Strack, D. (2002). Two distantly related genes encoding 1-deoxy-d-xylulose 5-phosphate synthases: differential regulation in shoots and apocarotenoid-accumulating mycorrhizal roots. *Plant J.*, 31(3):243–54.

Wang, C, Chen, Q, Fan, D, Li, J, Wang, G, Zhang, P. (2016). Structural analyses of short-chain prenyltransferases identify an evolutionarily conserved gfpss clade in *Brassicaceae plants*. *Mol Plant*, 9(2):195-204.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35(3), 543–548.

Waterhouse, R.M., Zdobnov, E.M., Kriventseva, E.V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biol Evol.* 3:75–86.

Wetterbom, A., Ameer, A., Feuk, L., Gyllensten, U., y Cavelier, L. (2010). Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. *Genome Biology*, 11(7), R78.

Wheeler, D., Bhagwat, M. (2007). BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In: Bergman NH, editor. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press. Chapter 9.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (2nd ed.). New York, NY: Springer.

Wimberley, J., Cahill, J., & Atamian, H. S. (2020). *De novo* sequencing and analysis of *salvia hispanica* tissue-specific transcriptome and identification of genes involved in terpenoid biosynthesis. *Plants* (Basel, Switzerland), 9(3), 405.

Wu, S., Schalk, M., Clark, A., Miles, B., Coates, R. y Chappell, J. (2006). Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. *Nat. Biotechnol.* 24, 1441–1447.

Xu, J, Ai, Y, Wang, J, Xu, J, Zhang, Y, Yang, D. Converting S-limonene synthase to pinene or phellandrene synthases reveals the plasticity of the active site. *Phytochemistry*. 2017;137:34–41. pmid:28215610

Yamamura, Y., Kurosaki, F. y Lee, J.B. (2017). Elucidation of terpenoid metabolism in *Scoparia dulcis* by RNA-seq analysis. *Sci. Rep.* 7, 43311.

Yam-Puc, A., Escalante, F., Pech, M., Chan J., Athimoolam, A., Ola, W., Olov, S. y Peña L. (2009). Trinosesquiterpenoids from the extracts of *Pentalinon andrieuxii*. *Journal of Natural Products* 72:745-748.

Yam-Puc, A., Chee-González, L., Escalante-Erosa, F., Chan-Bacab, M., Arunachalampillai, A., Wend, O., Sterner, O., Peña-Rodríguez, L. y Godoy-Hernández G. (2012b). Steroids from the root extract of *Pentalinon andrieuxii*. *Phytochemistry Letters*. 5: 45-48.

Yang, J, Adhikari, MN, Liu, H, Xu, H, He, G, Zha,n R, Wei, J, Chen, W. (2012). Characterization and functional analysis of the genes encoding 1-deoxy-D-xylulose-5-phosphate reductoisomerase and 1-deoxy-D-xylulose-5-phosphate synthase, the two enzymes in the MEP pathway, from *Amomum villosum* Lour. *Mol. Biol. Rep.*, 39(8):8287–96.

Yin, J., Ma, H., Gong, Y., Xiao, J., Jiang, L., Zhan, Y., Li, C., Ren, C. y Yang, Y. (2013). Effect of MeJA and light on the accumulation of betulin and oleanolic acid in the saplings of white birch (*Betula platyphylla* Suk.). *American Journal of Plant Sciences* 4: 7–15.

Yoon, B.J. (2009). Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6), 402–415.

You, M. K., Lee, Y. J., Yu, J. S., & Ha, S. H. (2020). The predicted functional compartmentation of rice terpenoid metabolism by trans-prenyltransferase structural analysis, expression and localization. *International Journal of Molecular Sciences*, 21(23), 8927.

You, M. K., Lee, Y. J., Kim, J. K., Baek, S. A., Jeon, Y. A., Lim, S. H., & Ha, S. H. (2020). The organ-specific differential roles of rice DXS and DXR, the first two enzymes of the MEP pathway, in carotenoid metabolism in *Oryza sativa* leaves and seeds. *BMC plant biology*, 20(1), 167.

Zhang, F., Liu, W., Xia, J., Zeng, J., Xiang, L., Zhu, S., Zheng, Q., Xie, H., Yang, C., Chen, M., & Liao, Z. (2018). Molecular characterization of the 1-deoxy-d-xylulose 5-phosphate synthase gene family in *Artemisia annua*. *Frontiers in Plant Science*, 9, 952.

Zhang, S, Ding, G, He, W, Liu, K, Luo, Y, Tang, J, He, N. (2020). Functional characterization of the 1-deoxy-d-xylulose 5-phosphate synthase genes in *Morus notabilis*. *Frontiers Plant Sci.* 24;11:1142.

Zhang, C., Zhang, B., Lin, L. L., & Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 18(1), 583.

Zhao, S., Ye, Z., y Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, 26(8), 903–909.

Zhao, G., Yan, W. y Cao, D. (2007). Simultaneous determination of betulin and betulinic acid in white birch bark using RP-HPLC J. *Pharm. Biomed. Anal.*, 43 pp. 959-962.

Zhou, C., Li, J., Li, C. y Zhang, Y. (2016). Improvement of betulinic acid biosynthesis in yeast employing multiple strategies. *BMC Biotechnol.*16, 59.

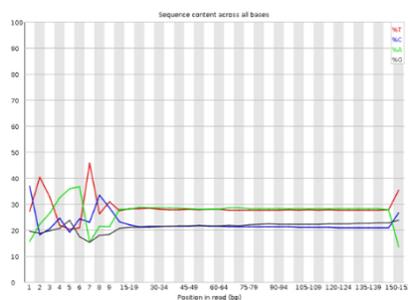
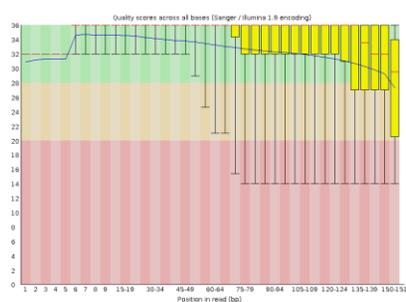
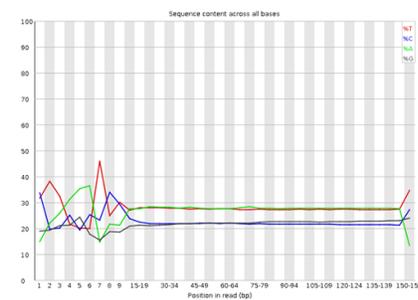
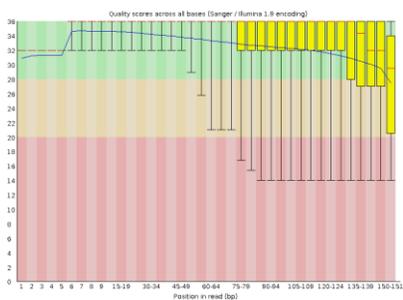
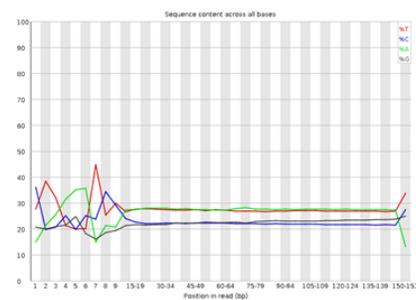
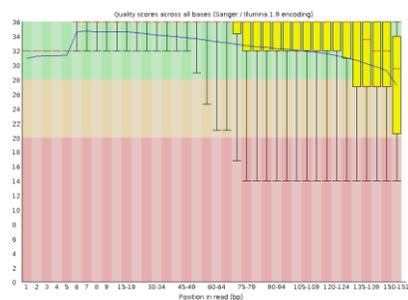
Zhou, F., y Pichersky, E. (2020). The complete functional characterisation of the terpene synthase family in tomato. *The New phytologist*, 226(5), 1341–1360.

Zhou, F.; Wang, C.Y.; Gutensohn, M.; Jiang, L.; Zhang, P.; Zhang, D.; Dudareva, N.; Lu, S. A recruiting protein of geranylgeranyl diphosphate synthase controls metabolic flux toward chlorophyll biosynthesis in rice. *Proc. Natl. Acad. Sci. USA* 2017, 114, 6866–6871.

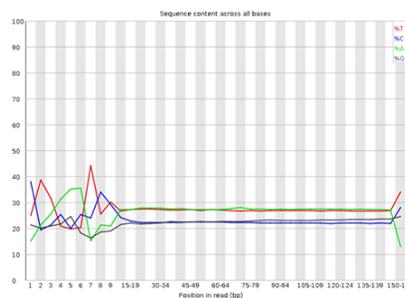
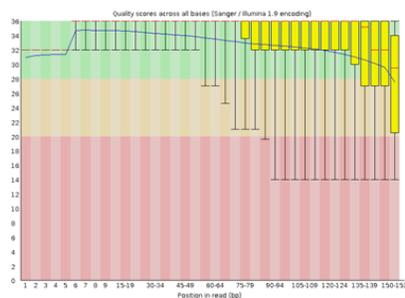
ANEXOS

Anexo I. Gráficos de calidad y contenido de bases por secuencia de los datos de los transcriptomas crudos y filtrados.

Reporte de FastQC de datos crudos, a la izquierda reporte del módulo Per base sequence quality, a la derecha reporte del módulo Per base sequence content.

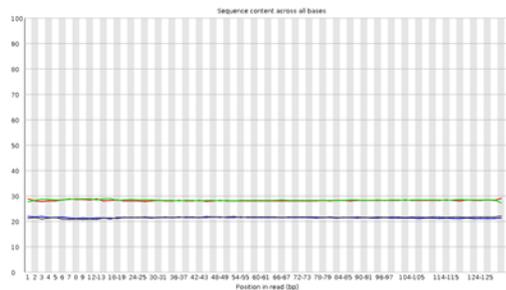
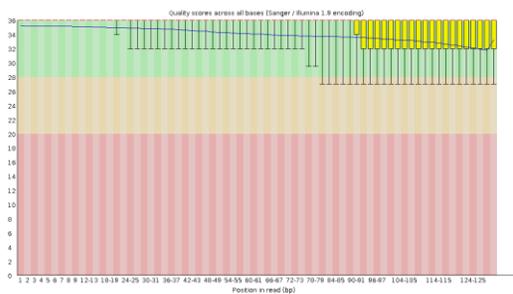
GG1TP4SS01_R1 Raíz Joven**GG1TP4SS02 Hoja Joven****GG1TP4SS03_R1 Raíz Adulta**

GG1TP4SS04_R1 Hoja Adulta

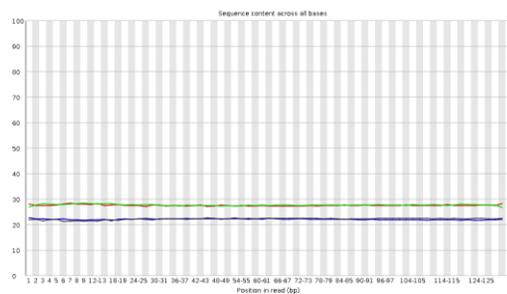
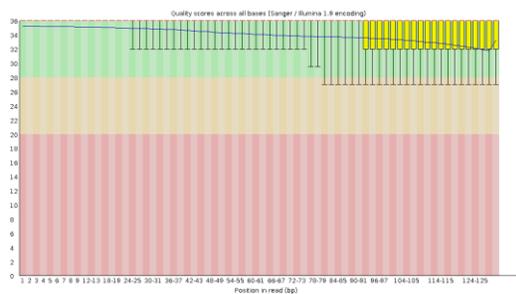


Reporte de FastQC de datos filtrados con Trimmomatic, a la izquierda reporte del módulo Per base sequence quality, a la derecha reporte del módulo Per base sequence content.

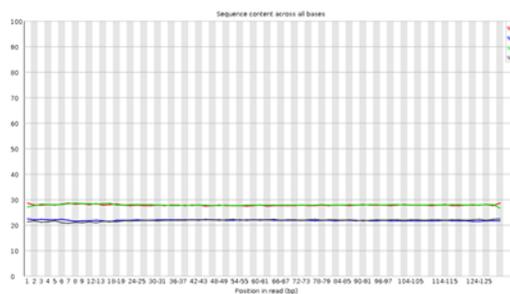
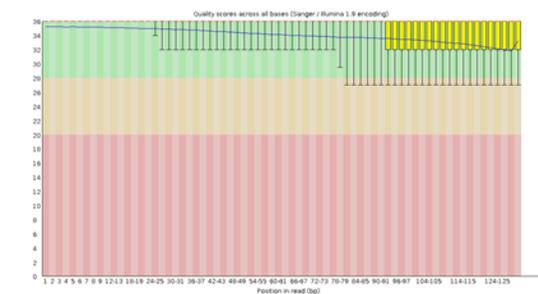
GG1TP4SS01_R1 Raíz Joven



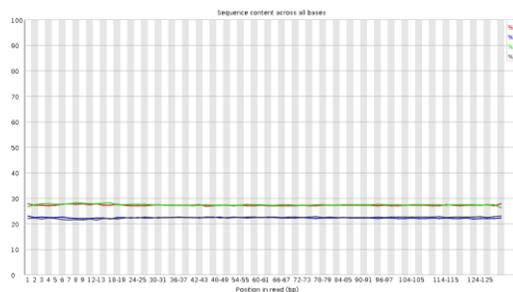
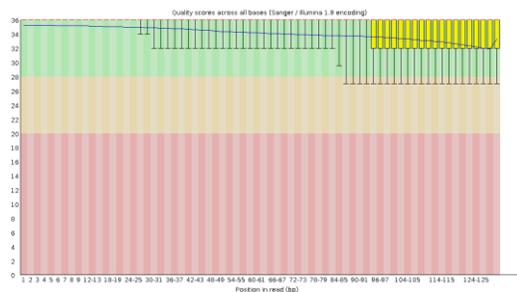
GG1TP4SS02 Hoja Joven



GG1TP4SS03_R1 Raíz Adulta

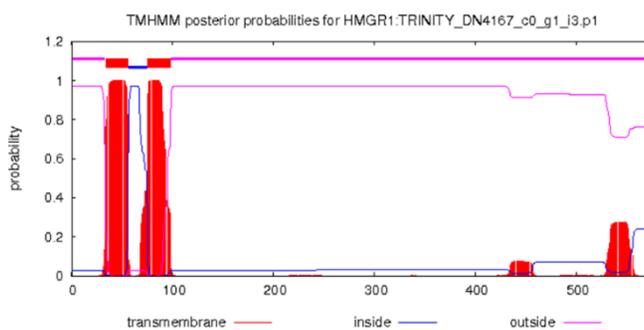


GG1TP4SS04_R1 Hoja Adulta

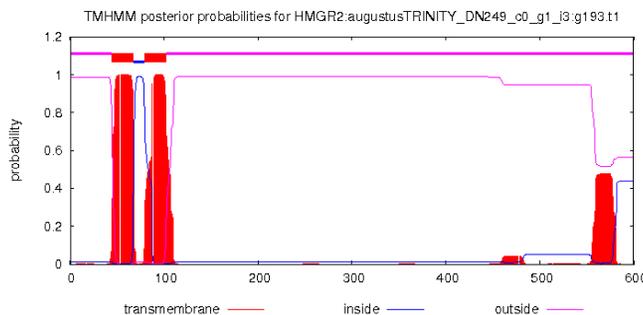


Anexo II. Regiones transmembranales identificadas en las secuencias HMGR usando el programa TMHMM - 2.0 (<https://services.healthtech.dtu.dk/service.php?TMHMM-2.0>)

HMGR1:TRINITY_DN4167_c0_g1_i3.p1



HMGR2:augustusTRINITY_DN249_c0_g1_i3:g193.t1



 Anexo III. Valores obtenidos en TPM de los genes identificados

Módulo de la biosíntesis de IPP

ID	HA	HJ	RA	RJ
AACT	29.37827	34.20963	13.304454	27.58139
AACT	27.110905	25.799228	25.079093	35.068988
HMGS	4.395577	8.131183	11.976074	39.153108
HMGR1	3.807275	30.910157	31.480212	82.347688
HMGR2	12.027106	13.800165	25.158307	37.848967
MK	11.196096	10.829691	11.720387	18.228432
PMK	2.594103	2.692042	6.43399	10.603195
PPMD	10.359096	10.335678	26.817525	42.180896
IDI	2.924347	10.800879	10.990101	26.35422
DXS1	2.153736	7.644098	1.221456	6.156448
DXS2	9.973476	20.554945	7.526678	15.627989
DXS3	6.172257	10.800213	10.684487	12.607014
DXS4	16.543515	21.308823	22.977317	38.8055
DXR	74.23461	53.980571	11.27056	24.09342
MCT	13.537432	17.359086	6.098626	10.672281
CMK	16.844089	18.271109	5.596456	8.079677
MDS	83.121442	73.340804	15.640216	16.576961
HDS	162.317845	131.335367	12.962549	17.998375
HDR	168.930681	156.398797	12.342606	14.099788

Módulo de las prenilttransferasas

ID	HA	HJ	RA	RJ
pt1-GGPS-SSU-I	40.014731	201.625098	19.268162	23.971497
pt2-G(G)PS	1.487663	5.231703	2.250023	6.687215
pt3-GGPS-LSU	56.847753	46.718462	3.277801	7.355001
pt4-FPS	5.753989	13.443941	9.358933	22.26279
pt5-SPS	184.600189	155.256452	1.803323	4.314032
pt6-FPS	15.478689	22.814948	18.450832	40.61245
pt7-G(G)PS	62.118123	75.971571	3.199833	9.659394
pt8-GGPS-SSU-II	17.091489	18.929012	16.827662	41.477732
pt9-PPS	13.199221	19.009355	8.782424	11.575986
pt10-GGPS-LSU	4.563927	7.116904	2.884152	6.444929

Módulo de la familia terpeno sintasa

ID	HA	HJ	RA	RJ
TPS1	1.019135	9.835444	0.021262	0.044115
TPS2	0.367152	2.085053	0	0
TPS3	1.376702	2.539329	3.607538	6.02353
TPS4	5.88252	14.853204	6.215303	10.749863
TPS5A-B	3.708332	207.379065	5.248899	42.00806
TPS5x	0	0	11.836738	82.712739
TPS6	6.417551	6.948725	0	0
TPS7	21.073769	318.637628	0.030295	2.142358

TPS8	0.238905	0.340574	19.984496	72.711949
Fragmento_tps8	0.521124	0.026117	34.138735	131.413339
TPS9	12.823012	14.024248	0.006822	0.103489
TPS10	39.097707	447.716301	0.103514	0.110927
TPS11	0.065513	0.295105	11.885681	13.333177

Anexo IV. Lista de primers diseñados para amplificar las TPS mediante PCR

Nombre	Forward			Reverse			Resultados	
	Secuencia	T° c	Posición	Secuencia	T° c	Posición	T° rx	Longitud
tps1	GCTTCACATCTTGCA TTTG	50	1	CTCCAACATCTGAAC GGG	53	1110	52	1110
	GCCAAGCTCGATTT AACAT	53. 8	258	–	–	–	53	582
tps2	ATGGATGCGTTTATA CAAC CAT	53. 8	1	TTATATTGCAGGGTC AAC GAGC	55. 2	1704	–	–
	–	–	–	GTCCAACAAAGCCCG ATAGA	54. 8	1098	53	1098
tps3	ATGGCAATGTCTTT GCTTC	51. 8	39	TTACATCATAAAGCA TAGA CTTC	48. 2	1651	–	–
	–	–	–	TAGGAACCATCATGC CCTTC	54. 4	1406	52	1367
tps4	ATGGCAACGCTACTG CTCTC	57. 6	1	TCAGAACTCATTAGG GACT GGC	56. 3	2406	56. 4	2406
tps5	ATGGTTGGTAGAGTT TCTTC	49. 5	1	TCTTTGTAAGAGGAA TGGAC	49. 3	1824	–	–
tps6	ATGGAGTCCTTAATC TCTTC	48. 5	1	CTAAGCAAACCTTCAT TAGCA	48. 1	2502	48	2502
tps7	ATGGAAGTTTGTGCT CAAGG	53. 2	1	TCATAATTTGACAGG CTCGA	51. 3	1701	50. 1	1701
tps8	ATGTCCATAACCAAA GAAGT	49. 0	117	TTAAGGATCTACAAG CATG GCAA	54. 4	1668	49. 3	1551

tps9	ATGTATTGCGCTTTC TCTAG	49. 9	1	TTAGTTGAGAGCTAG TGGAT	49. 9	1785	49	1785
tps10	ATGGTCCGCATTCAA GCATC	55. 9	1	TTATCTGATTGGATC AACAA GCAG	53. 3	1686	53. 6	1686
tps11	ATGTCTTCTTCTTCTT CTGT	48. 3	1	TCAAAGTACTCGCTC AAAGA	51. 2	2475	52	2475

Anexo V. Secuencia codificante de las transcritos identificados

>dxs1:trinity_dn7236_c0_g1_i1.p1

atgtcacctgtggagttataatgagttgtggcagcttctgctcattgctccatcccaagatcttctcagtaaccaggttagctccaacactattgcaactccgtcaaga
agagaaagttactggagtgtagcagctgctcaccgggacagtaatgtagttgatgaaaacgtaacattgcagaaacggacagccccgcaaatatctgaactttcc
ggggagaaacctctactccaattttgtagtctcaattatcccattcacgtgaagaatctctctgtagaggaactaaagcttggccgacgaacttcgagaagaaa
ttgtctacaaagtgtcaaaaacaggtggccatctgagttcaagcttgggggtgactgagctcactgttctctcatcatgtgttcaatacccaaaagataaatcatctg
ggatgttggtcaccagacatagcgcataaaatctgacaggaaggagatctaggatgcatacaattcgacagacttggactttcgggatttctaaagggatga
aagtttgcatgatgctttggagttggcacagctctaccagatttcaagctggcttagggatggcggctgtagagatttactgggaagaacaacctatgaatcatcag
gataggagatggagccatgacaggaggacaagcatacgaagcaatgaataatgctggtcttctgattccaatctcataatcatcttgaacgataacagacaagttc
cctccccactgccacagttgatgtctgctgcccgttggagcactaagcagagctttgagtaggctcaatcaagccgaaaattcgcttctccgggaggcagca
aaggggtgacaaaagcaactggtagcgaagctcatgaattggctgctagagttgatacataatgagagggatggtaatattagccacggggcttctctatttgaag
aattaggattgactatcgaccagctgatggccacaatagaaatctgtacacattctgaagcaggtgaagctatgccagaccaggacctgtgctcattcac
atcgftacagaaaaaggaaggttaccctccagctgaaggcgtgctggataaaatgcatgggggtgaaatttgagccgaagacagggaaaacaattcaagtc
aaagtcaagtacaaagtcataactacttacttctgctgacgcctgattgctgaagctgagaaagatgataaaattgtgctattcatgctgctatgggaggaggaaactg
gcctcaattgttccagaagcgtccctaataatagattttgatgttggatagcagagcaacatgcagtcacttttctgctcaggtctagctactgaaggtctcaagccattc
tgtctatttattcttcttctacaaaagggttatgatcaggtggttcatgatgtggacctgcaaaaggctaccagtgagggttgccttggatagacagggcttgttggggca
gatgttccaactcattgtgtgacattgacacaacttatatggcctgttgcctaataatggtggtcatggtccagcagatgaagctgaactgatgcacatgattgcaaca
gctgccccattgatgacagccctgctgctttagataccccagggaaacggcatcgggtccatgcttccgcaaaacaagaaggaactcctttagagattggttaa
gggaagaatactgcccggagggtaccagagttggcattctgggatatggaacaatagtaagaactgtttagaagctgctcggcttctcaggttgttgaatacggcc
acagtcgccgatgcacgattctgaaacctctgatggcgattgatcagacgggtggcccaagaacatgaagttctataactgttgaagaaggatccactggaggatt
tagctctcacgtctcacaattcttggattgaatggattactgatggaaaactgaagtgagaccaatgatgcttctgatagatacattgatcatggagatcaaagag
accagactgaagaggctggcctaacaatcaagcatattgcaggcaccgtactatcggtggctggcaagaacaaggacagcctccatcttgaatatgtag

>dxs2:trinity_dn14633_c0_g2_i7.p1

atggctctctgtgcatftacatttctggtggaattgaccaagccttggtcagattcttaaaaaacagcctttctgttctaattgggctctatggctcagatctaccttctca
tttcaaccccaaaaacaaccaggttaccaaaaggccaatgggtgctggcgctcactctgaaatgggggagatttctcgcagaagccaccgactcctctcttggac
actatcaattatccaattcatatgaaaatcttctgctaaggaactcaacaacttcagacgaattgcgttcggatcatttttaattgttcaagactggcggctcatctg
ggtccagcttctgttggtagctcactgtggtcttctcattatgtttcaattgccacaagataaaattcttggatgttggccatcaggttaccctcataagattttgact
gggagaagagacaagatgccaaccatgagacagacaatgggtatctgtttcaccaaacgctcagaaagtgaatagactgcttggcactggccacagttcta
ccaccatctcggctggcctaggaatggctgtgggaaggatctgaaaggaagaaaaacaatgttgtgctgcatagggcagtgagctatgacagcaggtcagggc
atacagaggccatgaataatgctggctatctgattctgatattgtattctcaatgacaacaacaagctccttggcaactgcaacgttggatggctcctgctccccg
gttggagctctgagcaggtttagcaagttgcaatcaacagcctcttagagagttgaagagaagttgccaaggggtaaccaagcagataggtggacaagtc

atgagcttcagcaaaagttgatgaatatgctcgtggattgattagtgatctggatccacattattgaagagcttgattctattacattggctctgtgatggccacagc
attgatgatctagtgccattctaaagaggttaagagtactagaacaacaggtcccgtgctaaccatggtgactgagaaaggcagggctatccatgatgctgaaaa
agctgcagacaagtatcacggagttgtaagttgatccagctacaggaaagcaattcaagtcagtgccaaaactcagcttatacaacatactttgagaggctctg
attgctgaagcagagggcagacaaaaacattggtctattcatgcagcaatgggaggtggaacaggaatgaacctttctcactcgtcgtcccaagacgatgctttgatg
ttgggatagcagaacaacatgctgtaactttgctgcaggattggcgtgtaaggcctaaaacctttgtgctatctactctctttcatgcaaggcttatgatcaggtag
tgcattgatggacctccagaagttgctgtcgtttgcaatggacagagccgggctggttgagcagatggccctacacattggtgctttc gatgtcacattatggc
atgacctcaaacatggtagtgatggctcctcagatgaatctgagctattcacatggttgcaactgcagctgtagatgatagaccaagctgtttccgataccctagg
ggaaatgggataggtgtagagctgccatccgaaacaaaggcattcctctgaggttgaaaggctggatactgatagaggagagagagtgccactctgggct
atggaacagcagttcaatgctgtttggctgcagctgctttagtagagcctcgtggttacgactaacagttgcagatgcacgctttgcaagccactggatcgtgcgctcat
ccgcagcctggcaagtcgcacgaggtcttgatcactgttgaagaaggatcaattggggctttggatcacatggttccaattatggccctggatggactcttgatg
gcaattgaagtgaggccgttggtcttccgatagatacatagaccatggatctcctcatgaccaactggcagaagctggtcttacaccatctcatattgcagcaact
gttttaacatactggcaacaagaagctttggaagtaatgcataa

>dxs3:trinity_dn272_c0_g1_i14.p1

atggcgtttcatgggctgtcattgggctaaaccaaccaatctccattcctgacagcccaaggctcaatacagtgaaagaacagtttctgtagagcatctg
gtaatggttcagatggcaagacggaggggaaatgtataaagaagataaagatggcgttgaagatcgatttctcaggagaaaagccacttaccctact
agatacgcataactatccgcttcatatgaagaatctcaacacaagatctgaacagcttactgcagaacttagagcagaaattgttacactatagccaagactggc
ggacatcttagtgcaagtttaggtgtagtagtaactgtggccctgcaccatgtttcaatacacctgaagataaaattatctgggatgctggctcatcagacatccgc
ataagattttgacaggaagaaggtctaaatgcatacagctcgaactctggttggcaggattcccaagagggatgagagcatctatgatacattggagcag
gacatagctcaacaagcatalctgctggtcttgatggctattgcgagggatcttttaggcaagaacaacaatgctattctgctatcgggatgggctatgacggcag
gacaagcatatgagccatgaacaatgcaggtttctcagatacactgattgctggttgaatgataacaagcaagtttctaccactgctacccttgatggccag
caacacctgttgggactcagcagcacttaagtaaacccaagcaagtcacaaattagacaactacgtgaagctgcaaaaagataaccaagcatattggacc
acaagcagatgaagttgcagcaaaagtgatgaatgctagaggtatgtaagtgctgctggatcaactctttagggagctcggattgactatattggaccagt
gatgggcataatattgaagacttagtaactatctccagaagtaaaagctatgccagcactggaccagtttaattcatattgtaacagagaagggaaaaggatc
ccctgctgagatggcagctgataaaatgcatggggttgaagttgatcccaactggaagcagtttaagcaaaaatctcaacacttcatatacacagattttc
cagaatccctgattaaagaagccgaaatagataacaagattgtgctatccatgctgccatgggtggtggaacagccctcaattatctccagaagagattccagatc
atgctttgatggtgattgcagagcagcatgctgtcacattgcagctggttagccacagaaggctcaaacattctgtgcaatctatcatattctgcaacgaggct
atgaccagggtgcatgatgtgatctcagaagttacctgttcgatttgctatggatcgagctggtctcgttggcagatgggctactcactggtgcatgttgattac
atacatggcctgtttccgaacatggtggtatggctccatcagatgaagcagaattgatgcatatggttgaacagcagcaacaataggtgatagaccatgctgctc
aggttccgagggggaacgggtgttgagcagctctcctcccaataacaaaggaaacgcctctgagattgggaaaggaagaattcttggagggtactagagttgc
aatactaggatattggtcagatgccaacaatggttgaagctgcgaaatgctgaaatcccagaacatctcggcaacagtagcagatgccagattctgcaaacctt
ggatggcgacctatcaaaggactaacaagaagagcatgaaatcttaactcgtggaagaaggctcagttggaggtttggatccatgtaactcactacctaacttg
acaggaatttagacggaccattaagttgagatcaatggtctcctgatagatacattgaccatggggcaccctgatcaatagaagcagcaggtctatcatcaa
gacacattgtccactgtctatcactattaggaagccaagagaagccatcaactcaatga

>hmgr1:trinity_dn4167_c0_g1_i3.p1

atggatgctcgtcggaggcaacctaacctctggtccatcggagaagcctctgtgaaatatgacaaccgttctccaaggctcggatgctctactcctgtccctctctac
tgacgaacggcattttctcatctctctctcctgctcactatctcctcctcgtggcgcgataagatccgcagctccactccgctgcacgttgaatctttctgagctc
ttgccattgctccttatcgcctcttcttattttgtggctctttggcattgattttgctcagtcattcattgctcgtacttctaattgatgctcgggagctgtgaagaacgtt
gtcatggacgaagatcgccgccaccctgtcctgtcctcggccctcagttgccctctcccagcccttctgcctctcaagagtagtgatcctcttggcggcagc

```
agccttcttgcgaagatgaggaactaatcaaatcagttgtcaacggccaaattccgtctgattcgttgaatccagtctcggtgattgctacagagcggttctatacgc
gagagcactgcagaggttactggaaggcaatgtcggcctcccttgatggattgattatggctccatctggggcagtgctgtgagatgccggttgcttgtac
aaattcccgtaggagttgcaggtccttattgctcaacgggtgcgaatacacagttccaatggcgacaaccgaaggctgttggtagccagtagcaacagaggtgca
aggctatatgtcatctggagcgctaccagtgcttctcagagacgggatgtccagagctccagttgtcaggttctccacggctaaaaggaccgctgagctgaaatt
cttctggaggaccgccaaatttggatccttggcagtgcttcaacaagcaagtagattcgcgaaactccaggttgcgaatgctctattgctggaaaaatctctat
atccgacttctgctcagcactggtgatgctatgggatgaacatggtgctaaagggttcagaatgtttggatttctcagagtgaaatccctgacatggatattatggc
atttctgggaatttctgtcgaagaacacctgctgcagtaattggattgaaggacgaggaagtcagttgttggaggcaacaactggaaggtagtgaaaga
cgtattgaaaccagcgtgccagcactgttgagcttaacatgcttaagaacctactggatctgctgttgcgggtgctctgggtggttcaatgccatgcagccaatgtt
gttctgcaattttatagccactggcaggatcctgcacaaaatcgaaagctcactgcattacaatgatggaagctgtcaatgatggagaggatctcacatctca
gtgactatgcttccattgaggttgaactataggaggaggaactcaactgcttcaatcagctgtctcaacctgcttggcgtgaaggggtcacaagaggagtcacc
tgatcaaatgctaggtcctgcccgcattgtagccggttccgttctagctggggagctatcttggatgggagctattgcatccgggcagctgttaatatgcatcaatgaa
tacaacaggtaagccgggatataaccgtaacctcctcagaattaa
```

```
>hmgr2:augustrinity_dn249_c0_g1_i3:g193.t1
```

```
atggatgttcgcccgcacctgcaactggtacgccaatccccgcggcagaagtcgcccggcgcaacagacattgaaaccacaaaaggccagaacagcag
ccttctcacctaaagctcagatgcgttgccttgccttatttattgacgaatgggttattctcactctgttcttctgtgatgtacttctctgaccaggtggcgtgagaagat
ccgtaattactcctcctccacgtcgtcaccttctgagcttgcgcttggcctcattgatgcctcctcatctacctttggggttcttgggattgacttgtcagctcctgatt
tgcaagcctaataatgaagttgggaaatgatgatgagattcctatggaggaagatagccgcaaggagctgcaccacctgggtgctgcttctcctccatctgc
tcccaaatgccctcaatggttccacaacagccctcaagatggctatgaccattacagagaagcctgcacctctcattacccaacaagattcggaggaggacgaa
gagattataaaggcgttgttccggcacaactccttcttctctgagtcgaagctcggcgactgcaagagagcagctgcaattcgcaggaggcgtccaaggagg
atcacagggaaagtcctggagggttactcttggagggtttgattacgggtctatfttgggcaatgctgtgagatgccagtcgggtatgtccagttgctgtgggaata
gctgggcttgttctgatgggcaagaatacatggttccaatggcaacaacagaaggatgttgggtggccagcactaacaggggtgcaagccatcttctcatctg
gtggtgcaaccagcgtgctgacgagagatggtatgactagagctccggtcgttaggttggctccgctgagagcagctgcaactgaaagtctcagttgaggatcctgt
gaaacttgagtcttctcactagcttcaacaaatcgagcagattcctagattacaaggcatccagtgatgcaatagcaggaaagaattgtacatgagattagctcag
cactggtgatgcaatggggatgaatattggtgcaaaagggtgtcaaaaacttcttatttctccaggacgatttctctgatattggtgttggcatttctgaaactctgc
gctgacaagaaaccagcagcagtaactggattgaaggaaggagaaaatctgtatttgcaggcaatcataaaggaggagattgtaagaagggttgaagactg
aagttgctgccctgttgaactaacatgcttaagaatctcactggatctgcaatggccggagccctggggattcaatgccatgccagcaatatagtctctgccgat
acattgagcactggcgaagatccggctcagaatgtggagagctctcagtgatcaccaatgatggaggccgtcaatgatggcaagatctcatactctgctcagatgcc
cgctattgaggtgggtcagttggtggcactcaactggatctcagtcagctgttgaactgcttggagtgaagggtccagcaaaagagacaccaggggcca
atgcaagactctggccaccattgttctgctggtcagttcggctggggaactctctcatggctgcatctctgctggccaactcgtgaaagacatataaatacaacc
gatctagcagggatgtaaccaagatgtcctcttag
```

```
>dxs4:trinity_dn4693_c0_g1_i28:g4.t1
```

```
atgggtactctgtgtgaataaccatttattataatgctcattctacaagaattcgaggaattgtctccaaaaggaggattctcaactgtaaaatttcttctactgtggag
tttcaaaactcaacttaatccgagttctgcaatcagctatactccaaagagtcagctgcaaacacaggctgtacctgaccagagaaggccgcaacgccata
ttagatacaaatgaaagtccttgcacttaaagaatctgctcacaaggaggtgagacaatggctgatgatccactctgaattatcattatgcaaaagactcaca
aaccatttaagtcagctcggcagttgtggagctaacagtgccattcattatgtttccatgctcaattgacaagatacttgggatgtcgaagaacatacatatgcaca
taaacttctactggaaggaggccctcctgcaaatgaatggagaacaaacaatctatcaggcttgcactctgagagcatgtttgaccaattggtgctggcaggtt
gcaacagtatatctgctgttagcatggctgtagctcgagatataaaggaaaacgagatcgattgttctgtcattgcaatcgaactactatggctggtcaggtt
acgaggcaatgagtaatgaggttatttagattctgatattgttatactaaatgacagccagcattcctacatccaaagctggaggagacacctaaccgccaatt
```

aatgctttatctagtctttaagcaagcttcagcaagtaaatcttcagaagattagagaatctgccaagggtttaactaagcgaattggcaggggtatgcatgagtg
cgccaaagtgatgagatgcacgtggttaataggtcctccaggatcaactctcttgaagagcttgctgtattatattggccctgtggatggacataataggat
cttattgttttacatgaagtgcatcctggagtcctggctgtttgtttcatgtcataacaaaggaagatcaagaggaacaacatgatgaatcaagaaaaa
tcaaatgccatgacaaagattctgtatgtctgattcattgccctccactggcagcctcgaacctatagtagctgtttgtggaggcttattctacatgctgagagaga
caaagatagttgcagttcacggggaatgggaatggagcctccctaaatattcaaggatagattccagataagttctttgatgtgggatggctgaacaacatg
cagtcacattctcagctgtctgtctgtggaggttgaagccctttgtataattccttcagctttctcagagagcttatgaccaggtagttcatgatgtagaccgcaaag
gattccagtcgattgtaatacaagtcaggattagttggtctgatggccaacacacictgtctattgacatcacatatatgctgtttgccaacatgattgtag
gcaccatctgatgagattgagcttcacataatggtggccactgcagccagattgaggataggcctgtttgctttcgtatcccagaggggccattgctaagataagca
atacttaggatatggaattcaatagaggttgtaaaggaaactcctgtgaggaaaagatgttcctactgtgtatggatctatggtccaaaattgtcttagagct
caattactctgcaaagctggtgttgaagtaacagttgctgatcaagattctcaagccactgacgttgatctcatcaggcacttatgtgcaaaccatgggtctgat
tacagttgaggaggtgctattggaggttggatcgatgttgcgaattcttctcaatggactgctgatgcagggataaagtggtccaaattaccctaccaga
cacctacatcgagcgtgatcaccacatgaacaactgtctgtagcaggcttgactggaaaccacattgctgctacggcactgagttgctgtgctcgcactcgcaggct
ctccttctgatctgctag

>tps1_augustus_trinity_dn17403_c0_g1_i1:g2.t1

atggacgtcgaaggaatgtaagttggtacgaagcttcacatctgcatltaagggagaggtcactgacgcagccaaagcattcacaagatgaatcttgaaaatt
ccaaagagattattgaaccaaataattctcaagaagtaaagcatgctctagaattccgtatcaccgtagaatgcaaagattagaatctaggtgaaaattgaagcat
acaaaagaagaaaggagaaaaataaagtgtacttgaactagccaagctcgatttcaacattgtccagctactctccaagaagatctcaagatgttcaagggtgt
ggaaagctttgatattagcaggcaaattgaccttgaagagacagaatcgtgaaagctctttgggctgtaggaatgatattgaacctcagctagcaaatgccga
ataggactagccaaagtagcagccctataacagttcttgatgataatataatgatctatggcaccctggatgagcttgaacaattcaccatagctgttgaaggtggat
cttgattatgtaaaaatctcctgactatataagttgttctccttcttacaacaccgtcaatgattggactatgacacccttaaggaaaaggggaaaaatggtcgtc
cctcactgagaaaagcttggccgattatgcagatcttcatgagagaagcagatgaaatacatgaacatcaccccaagttttagaataatattgagaatgat
gggtctcggcatctggtgctgttcttaactatgcttacttctagtcactgaaaatacagaggaggcaatacaatgcctagaaaatggtcataatgctgttattgg
ccttccaaacttttcgctatacaacgatttcggagattgctcgtcgatattaacaactgattgaagaatgttgaagaagatgaacatggaactaatcaatgattca
ccttgcgaaaagccttcatcagattgctattaatctgctcgaattgctctatgccaataaccagaacggagatgcaaatagtctcctgatgttgcgcaagaaccac
acttgttggtactaatcacccttcagatgttgagaattaa

>tps2_trinity_dn119011_c0_g1_i1.p1

atggatgctttatacaaccattaatttctaccactagtagtactgatcatgatccacatgctgttcgtctgccaattttcacctagattttggggagaatacttctgc
actcactctctgaagcggagaaagtagttttcccaaggagattgaaaaaacatgaattgaagaattaagagaagaggttaggcagattctgttatcaactccag
acagctctccacaaaagctggactgatcagttctatccaacgtctagggtgtcttaccattttgggactgagatcacagcatcgttacaagcatatctaactccacat
cgatcagtacaaagttgaagataatgaagatgacctgtatagttgctctcgtttctgattattgaggcagcaagggctttatgctccagtgatgttcaacagattca
gaaacttagaagggaaatgaagaaatcattggtgggtgatccgaggaatgtaagctgttggaggcagcacattacagagttcatgataggacattttgatga
ggcattgatattcacctactatcctgactctgtggtcccaatttgagcaatattcccttgaacacaaagttaatcaagcgctagagatgccaatccacaaaacttg
acgaggataggagcaagaagatatgctcattaccaacaagatgaatcacatacaaaagcattactcaaatttcgaaattggattcaataggtgcagaagttgc
atcagaaggagctaagtgattgacaatatggtggaacaaattataaattgcgaaaaaatatgccctttgcaagagacagactgtggagtgctattttggatattg
ggtgaatactttgagcctcagattgctcgcagaaagattccacaaaagttattgcatggcctccatcatggatgacatttatgatgtacatggcacccttgatgaac
tcttattttcacgaaatgcaattgaaagatgggataccgtaccatagatcagctgccaccctacatgtctacttctatcgggcttgttggacgtgactctgaaatggag
gtagatttggccaaagaaggcaacaaagtctagttactatggaaggcagaatgaaaaagttatccagggcgattttgcaagaggcttactggttcataatgg
atattaccaacatgaagagtagatgactgactcactccagcggatcatgatgttagtagtactctcattggttggcatgggagaggttagtaacaaaagaag

tccttgattgggcaacgagtgaaaccttgattgtcaaatctgcagcaatgattgccagattaatggacgacatagctggacatgagtcgagagggaaaggagagac
gtcgcctctgccgtggaatgctatatgaatcaatatggtgccgcaaaggaagaagcatatcatgagttacaggaacaagtaacaagtcctggaagaagataaac
aaagaatgcctggaagtagtaagctaccctgctgagcgaggttgaatcttgcaaaagtaatcgatcttctataaaagatggagatggctacacaaattccacatc
caaatctaaagatttatcaccacgggtgctgctgaccctgcaatataa

>tps3_trinity_dn25972_c0_g1_i13.p1

ctctctttgggtatcatcttactccctgctccatccaatggcaatgtctttgcttactatttgaatgactccaatatcatgtacagccacaagcaacaagtggacatcg
cccagaatctcactacggtttctaactctcgatatttctgaaaattacaaggagacttatccagcaaacatgagacgaagataatggatgaaggcacatgctgga
agtgaagggaaaacagccactgaaatgcatggtatggtcgtatgccctccagaggctggcattgactatcattccaggaggatattgaagcagagctttggagaat
gcacgtgacaacgggaaatgcattacagatctcagctcttgatgaaactctttatttcaggctttgagacaacaaggttactatgtgatgcagatgatttaaca
tttcaagaataaagaagggaggttgacaagactctgacaacgatactggtcagtgagggaattctatgaagcagcacagcttagtattgaaggagaacatatac
tggacgaagccgtaaaactttagtgccaactcctcggtgaatggatcagaaatggtgatgataaaagctgcaataattgtaatacactgagacatccttatcgaa
aagtttagctacattccataccaagaacttaactgcatgtctgctggcaacaatacctggaaaggcagcttgcaagagctttcatgcatgattcatttctgcacgaag
cgtacgtcgggaagagctacgccatgtttcaaatggtggaacctcttgctggcgaaggaaatgaagctgtaggaaccagccgctgaaatggtacacttggcc
gatggctatcctcactgatcctgtctcagatgaaagaatcgagctcacaagctgatctcttaatttaccatcatagatgatattctcagatctctatgtaagccggatga
attgagctctctcgcagaagccgtaaatagatgggaattagatgcagctgatcaattaccagactacatgaagatgagttcagggtctctatgacatcataaatgaa
actgccagaagattgaaagaagcacggctctaactctgtgtatctttacaaaaacgtgggcaagttgtgctgatcatttctagcagaatcaaaaatggttctcgc
acggaaatgccatgctcagaacagctacctgaaaaatggatgtaagctcgggagctcatgtggttagttcacttgttttctgtgggttaggggtacagatgaaa
gaacaacaatgctggaagatattcgggtatcatctctgtggtcacaattctctgctttgggacgacttaggaagtccaaggacgagatcaagaagggcatgat
ggttctataataattgtatcatgaatgagcaccagggtctcagttgaattgtcaagagcagctcttaaaatgattcaaatgaatggaacgactaaacagaga
atgtctcatctgaaccgatctccgctcattccagaaaggggctcgaacctgagggatggttccatcatgtacgactacgagccaccaaggctccattct
tgagcagtatatgaagtctatgctttatgatgaaatttaa

>tps4_trinity_dn590_c0_g1:g1.t1

atggcaacgctactgctctccaattcccgcatcaccatgtcttccctctatcgctcacaagttcccgcctttcgtataatcaagcccccaacggctctctgcttcagc
gttctctttggaaaaggactgaaagagcaacatcaactgattcatctgacctgcttgaagagaccaagcaagaattaggaaactattacatgatgttatttct
tctgtctcgtatgacactgcatggtagctatggtccctctccacattcttaagctaccttctcccaagtgcatagattggttctgacaatcaactcaagatggtc
ctgactcctctcatcaccatccctgctactaaagaagcttctctctacgctagcagctgttctgactcaaaagatgggaaatggtgaacaaattattgacaag
ggtcttcatattatagattgaaatgttgaactgacccaaatcagctcactccaattggctttgacataataatcccggaatgctgaaatgcaagagattggatc
taaacctcagctagaatctgcatccttgatgactgttccataaaagacatctagagctcaaaagatgccaagaaagtcagttcaggggagtgatgatacctgct
atacgtatccgagggaaatctaaagttcaggactgggatggttatgacgtatcagagaaagaatggatcactattcaactcaccttctacaacggcagctgctctat
gcatgtccaaaatcctgctttgttaacttacactcagctcttagagaagttggctgtgctgttccaacagtttaccctgaatgtgtacgcccgcctctgctgattgaca
atattgagaggcttggtattgtcggcatttctggaatgatattcagagtggttggagatacttacaggctttggctgaggggtgatgaagaatatttaaggaaatccg
acatgtctatagcattcagttattgcgaaagcatggatgatgtctccttaaatccattagataagctcttaaaagatgatctctattcaatcctttgatggacatttaa
ggatataatttactcagattacatcgggcatcagaatgattatagctgtgatgaatcagctctgggaaactaaatctatggtccaaagatttactgagagaga
aagtattcgggggttattcaaatggagtcactagctcaaatctgcaaaagaggtagatgatgctcctcaagttccctatttcaacaatgtcaacttatacacttgaagaact
gaaagaactagaaaggtgggtcatcgagaacagattggacaagctaaaatttgcgaggcagaaatctgcatactgttactttctgctgcagcaaccatatttctcca
gagctatctgatgctcagatgcatggccaaaaatggtgtcctcactactgtggtgatgatttcttgacatcgggggttctatagaggaaatgaagaactaatcctttat
ttaaaggtgggatgtaacactacactgtctgtcagagcaagttgatattatattctgcactcgcagcacaatcattgagattggagacacagcatttgagcacc
aagcagctgatgtgacggccacataattgagattcaattaacatggttactgacgaggagctgtatctctgtttacagtggttagattgtcaattctatggtgaaggaa

gctgaatggactagaaatgtccgtgccacaatggatgaatatgacaaatggtatatactcttgcctgggaccaatgtcctcctgctctttattccttggctca
agcttctgaggaaactgttcggcattctgaatatcaatctcttttaaaactatgagcaccattggcgctgctaatgatccagggtttgagagggaaatcaaagct
ggtaagttaaatgctctatcattgcacatgattcacactggtggcaacataaccgaagcagctgctattggacacatgaagagtttgattgcagatcaaaggagagaa
ctgctgagattgctcagagcgaacatgaggtggtccaaggcatgagagatcttttgaatatgatcagagtgctgaatcagttctacaaggaggatgatgg
attcacctcagacgagatgatggagctgataaaagcaattatgtataagccagtcctaatgagttctga

>tps5x_trinity_dn6190_c0_g1:g1.t1

atggttgtagagtttctccatttgaagtcaatggcagcttcatcaatggactgtcatagccaactaaagccttctcacttcacagcatagattgatcctaaatccc
gagggtctatacttcatcctcctcgcgctgtgcttctcaggtgacacattcagggaaccaacaggtagtaaattctcacacacaaaaacattgaggaggtccgga
aactacgagcctccagattggattatgagatgctgctgtctcagaaatgaatgcaacagataaattcatccaggaccgtgtacattgaacgatctagtgtct
aaaggaacagataagaatcatgctagatgaagagcagaagccactgttcaactagatgtaattgataactgcaagactggatatacttaccatttcgaggacaa
aatccagtcagtttgagaagaatatacaacgaaaacaacagtcacataaggctttaaagaaagaactatacgaacagctctgtaattagactcttagacagca
ccatttcatgtccctcaagagggtttgattgttcaagtcagaaatg

>tps5a_trinity_dn127_c0_g1_i43:g33.t1

atgttacaggagaagtgaacagcgccgattaaggatggaggaacaccacctccgtgttgggatcactctatggtgtaaaagtgatcccaaacacggaggtgg
tgttctccatccttaatcgccgctgttacacttctcctgtaacatgctggacacaactggttctaagttctcggagaaaaagctctaaggcggctgcaactactctc
ctccacattgggattatgagcggctgctgactcaaaaatgaggatgagcaggtgtatagagcgggtccaggttctgaaggaaaaagtcaagatttctgctg
atgaagaggagaagtcgctgaagcaactagaatgaatgatcactgcaaaagactggcgtgtcttaccatttggagaaaaatcaatcaattttacgaagaataca
caagaagaacagcaacaacatctctgtaaacaaaagagctatatgaacagcactgtaatttagacttcttagacagcacaatttctgtccctcaagagggtttg
atagttcaagtgtgaaaatggcactttcgggctagcctctgtaaagaccggaaggactgctgtactgtttgaagcttcttataaccgaaaacgaaagcaact
ggagttgcaagggaaattagtgcaagaatctcaagaaaatctggcggaaaaaggctaatccagaactctatgactagttcaaatgctgggagcttctctg
cattggaggatgccgctgagggcaaggtggtcatagaagagatgagaaaaagaaagacatgaatcctgatatactggaattggcaaaactggactcaaca
ttgttcaggcaactcatcagaagatctgaatgcatcatggtggtggaataatacaggtttcagaaaaagttacatttctcgggacagattgacgaaagttttc
tggaccataggattgatattatcctggatacggatattgcagaagaatgatagcaaaagatcactgctgataactacaatagatgatattatgatgtttacggaacat
tggacgaattggagctcttactgatccatagaaaagggtggatacaaaaactattgaccttccagactcttgaaaaatgttatcttgagttacaactcctgta
gaaatggcttatgtttctaaaagaacgaggagtaacatctccacaactaagcaaggtggtgggagatatttctgagcctactgcaggaagcgaatggtatt
tcaatggtgacactcctgactgcaagaatacctcaaaaatgccagattcaatatcagctcctgcaataactagaacatgcttattctctgtagcaaccgatcaag
gaggaggctctagaatgctggacaataatacaatattctgttgcagcatgattcaaggcttcaaatgacctcggaaacatcacaggatgagcggaaaaaga
ggagatgtccaaaatcaatacaatgtfacgtgaaagaaacgggtcttctgaagaagaggctcagagatacataaggatgttgattgctggagacctggaagaa
atcaacaaaagaacgaatataagaatctccttctcaaaagattataggtatggcaatgaacgttctaggtatggcacagtcctgtaccagaatggagatggacat
ggccaccaggaccctgtactaaggatcgcatctccacaattctttttagccattcctctgcaagacttaa

>tps5b_trinity_dn127_c0_g1_i62:g23.t1

atgttacaggagaagtgaacagcgccgattaaggatggaggaacaccacctccgtgttgggatcactctatggtgtaaaagtgatcccaaacacggaggtgg
tgttctccatccttaatcgccgctgttacacttctcctgtaacatgctggacacaactggttctaagttctcggagaaaaagctctaaggcggctgcaactactctc
ctccacattgggattatgagcagctgctgactcaaaaatgaggatgcaatccagggtgtatagagcgggtccaggttctaaagggaaaaagcaagatttctgctg
gaagaggagaagtcactgaggcaactagaatgaatgataactgcaagactggcatalcttaccatttcgaggaaaaaatccagcaatttggagaaatatac
aatgagaacagcaccaacatctctgtaaacaaaaggctatatgaaactgcttgaatttagacttcttagacagcaccatttcatgtccctcaagagatgttcgatt
gtttcaagtgtaaaatgacactttcgggcttgcctctgtgaagaccggaagggtgctgtacttattgatgcttctatcttcaacaaaaatgaaggtaccttggagt

tagcaaggaattacaaccaagaatctcaagaaaatctggaggagaaaaggcttgatccagaactcaccgattagttcaacattccttgaactccgctgcatt
ggaggatgcttcgattggaggcaagtggttcttgcagggtatgagaaaaaacagacatgaatcctaataactgaatttgcgaaactgactcaacatcgttcaa
acaactcatcaaaaagatcttgaatagcattatgggtggaagaatacaggtttgcagaaaagttcagatttgcctgggacaggttgggaaaacttttctgaca
acaggattatattaatcctgaatcgggtattgcagaagaatgtaacaaaggctcatgttctgtaactacaatagatgattttatgatgtttacggaacattggacga
attggagctcttactgatccatagaaagtggtgatcaaaaactattgaccatctccagactcttggaaaatatgttatcttgcagtttacaactccgtaatgaaatgg
cttatgtttctcaaaaagacgaggagtaacatcattccacaactaagcaaagtggtgggagatatttgcgagcctacttgcaggaagcgaatggtattcaatggt
gacacccgctcactgcaagaatacctcaaaaatgccagatttcaatatcagctcctgcaatactagaacatgcttatttctctgtagcaacccgatcaaggaggagg
ctctagaatgctggacaataatacaatatttctgttctcagccatgatttcaaggcttcaaatgacctcggaaacatcaccggatgagcggaaaagaggagatgt
gccaaaatcaatacaatgttacgtgaaggaaactgtgcttctgaggaagaggctcgagaatacaaggttggattagtcagacttgaatgacatcaacaaag
aacgagtcagagaatctcccttctcaaaagattatagcaatagcaatgaatcttagtaggatggcacagtcagttaccagaacggagatggacatggtcatcaag
actccactactaaggatcgactttcacactcttcaagtcattctcttacaagacttaa

>tps6_trinity_dn15212_c0_g1_i2.p1

atggagtccttaactctctccgtccaatctctagttgcccaagtaagaaagaattgttctcaggacatgagaatgtgcacgcatttctccccatctgcatacgaacag
cttgctagccatggtccccgacaccgagaagtctgaccaaccaatgttcacgaattgttaacttgatactcaacaacaaaacaatgaaggttctggggagaa
cacattggggagaatctccaaccattgatacttctcgaaccctgtctgcatggtggcactcagaaaatggaacctgggggaagaaaacattagaagaggggtg
gatttcattattccaaggcggagatcattctcaaaataaattgtcatcacctccacgttggttaccattgctttccctgcaatgattgaactgccaacaagcaggttag
atgctgttttctcaaggatcaacgaggttctctgtgataatttctgcaagaggcagaaaatactgaaacggagaatctcgtagttgaaaagcaataccgcaatcctt
attatcgacacctggagagcttgcagacataatcatgtctctgaaaaggatatagtaaagcatttaaggaagatggctcaatctccagtcactatcagccacagcat
gtgctgacatggctactgaaatctgcaatgcttgaagtacctcaagtctgttatcgaagatgtcctaattggagtaccgcaaaagtaccatgtggacgaagagtaat
atatctctgcatggttactgtacaaaagtttagtctgctgaacatttgggtgaagagattgaacggatttgaagcaagtagcacagaagcaaaagaaactacgactt
ccagccaacagaaatacagaatttaccagcgaagttattcaaatgattggtgcttgcggttctgagatttcaagggtatgatgtaagccatgcaatttgggttct
tatctgatgttataatcatgaatcatctggaacagaactgtgaaagtttaccagtgctatacgcataatataaggaacagacctgcatcataggtgagatgaact
ggaggaggcagatcgcttcccggaactgcttgagaaatcaatggcttgaatagcagagctgatgatattgtaatttccaggtctcacaattgataaagcaag
attgactctccctggattactcggatggaacatcttgatcataggtatggatcgaagaaaataaatacaaacctctggaccggaaggatcctctcagcagttat
cctgtctgcataacgagaagttgatggaactggctgcaagaactacaactccggcagtcactacgccaaggagttggaggagttaaaaagatggtccagcaa
aagtgttactgataattgggttggacgagagaacaccgcataattgctactcgcgaattcagcaagcactgcctgcctcatgattcctctcaggttactcattgcac
agagtgcaataattacaggttctgatgattttacgacatggaaggttccctgaccgagttggaactttgactgaggccgtacagagatggatggttaaggatctt
acggtcatgtaagacgatcttcaatgccctcgatgatttagtaacacaacagcagcaaaagtaccaacatcaacacgggagcatttctgagagaacttagagat
atatggaaagaaacatttaccctcatggatggtggaagaacatggagtgattcaggctacataccaacctgatgaatatctagacaccggaatgatctccatcgc
gatacacacatagttctccagcttctgttgaagtcaggttaccgtttcaaaagataaagcctccagaatgcaaggatcaccgaattgctcatgactgcagct
cgttgtgaaatgacactcaagttatcagaaggaactgatcgatgggaaaatgaacctgtatcgctgcacctaataaagaaatcctcagtcctatagaagattcaa
ttgcttatgtgaaacatattctgaacgagaagacgagagagtttctgagcttgtttgatggacgttaatggactggatgaatatcaaaaacaaaaatcattgaaacaa
ctccatctcttgcataaaggtttcatatgttttcaactctaccaacctatttgacaccaaggccgaactgctcagacatcaagcaagcaatgatgtcccttaccatt
ctcaacctcagatcttgcataaagctcccaccattaaccactccaaaaacgaaggagatcgcaaaaacctctgcgtgtttggatctcaacatcagaggtcggcttg
gcaaggaatgatggcaaaaagttaccgatcagcattccatgactccaacattgctaataagtttgcctag

>tps7_augustus_trinity_dn2222_c0_g1_i8:g85.t1

atggaagtttgctcaaggttcagttctgattcactcaacaccaatgaatgctcaagcatctcccagcgtctgccaattatcatccgagtggttggggagaacatt
cctcagatagctactgatctcaaggaacctttactgccaagaagagatagaacatgcacgactgaaagaagagatcagaaaaatgcttatcggattcccgata

aatgcttcagaaactgactgattgacgccatccaactgctaggctgtcttaccatgtgaaagtgagataaaggcatcattgctcagtatatataatgctacggcc
atgagctaacgtgaagacaataatgatctttatgagttctctacgtttccgactcctcagacaaaatgggcattatatctctgtgacgtgtcgatgaattcaaggact
cgaaagggaaattcagagcatcactgaaaatgactgaaagggctgtgagctgtacgagggcctcaaatttcagagttcatggcgagaccatttggaggaagcct
taatgttactaccactcaactgagtcctgatgacctgagaggttctcaggaactaaagttaaacacgcactcagatgccaatccagaaaaacttggacaag
gctgggaggattacagtatctaactgtatacaaggaagcgaatcatttgatgacacttaaaacttgcataactcgattcaatgattgagaagatgcacca
atggagctcggcaaaactaacaaatgggtgaaagaattggactgtccagaacattacctttgcaagagacagattagtagaatgctactctggatattgggagtg
atgtgagccccattatggcctgcccagatgggttcttctaaagtattgccttgacatccatcatcgacgacatctatgatgtgatggaaccttggatgaactattcttca
cggatgcaatgaaaaggtgggataaagggccgcaaatcaattaccgtcactgtgacatctttaccaagttctgttagatgtttataacgaaatggaggaggagtg
gcaaaagaggcatatctggcagagtgactatgcaaaagaaaagtgaacaattagcgagagcactccaagaagcttctgggagcacaacggatagc
accaacattgacgagtagatgagggctcactcgtgctggggatataatgatgctgaccgacgctcttggcggcatgtagacatagcaactgaggaagttg
gcttgatgaccaatgaaccattgatcctcgcgacatcagtaatttgatgataatggatgacatggtttgcgtgagagtgagaaggggaaggagaagtggcct
cagctgtggcatgctacatgaaggaataggtgttcaagaaagaggcgttgatgaactcataaacagggtgacagatgctggaaggatataaacaagaacg
ctaaacgtccatgactgaagatcaatgccatcctgagcagctcgaatctgcaaggggataaatctactacaagatgaagattgtatacaaatctac
aacatagtaaagaactcatcctctgtactcgtcgtatcctgcaaatatga

>tps8_trinity_dn4534_c0_g1_i9.p1

atgaacagatccgctctctattataaatatggatcatctggactgtattttcatcattcccctcatcagaaaaataccagaaagctcaagccaatcaagcactcac
caaaatgtccataagcaaaagattatggcaaagccttctgaacagcactgataagcaggaggtgttcgcccgttagctcactccccaaagacatctggggtgat
atcttactgcatcttctttgatgatcaggtttatggcatgctatgacggagatgaaacactcaagaagaagtaaagattttacagagacaagaaatgatgtg
atggagaaatgaagttgatagacacacttgaacgtcttggtatcatatcactcaaggaaaaaattgaagaggaattagaccaaatfacaacaatgcaactcc
aaaagtgaaccacgtgactatgatctggcatggttagctttgttctgactactacggcaacatggttccggtatctcctctgatctttgacaatttcgatgccaatg
ggaactcaaaagactttatgcaatgacataaagggttactgaactttatgaagcagcacaactgaggacaaaagaagatcatatagaggactccctagcat
gtgcatcactcatcaagtcaagcagccttgcctcctgattcttacttggtaaacaagtaagcatgcccctgagcagccactcaataaaggccattcgagaacaga
agctcgttattatgtcttctacgaggaagatgactcaaaagaatgaaatacttcaaagttgcaaaattagattataagatgttcagatgttcataaacaagagctt
ctcaattatgaggtgggtggcgggaagtgaatctgtctcaaaacttctatgcaaggatagaatgggtgaatgtattttgggctctagcatgttactgtgacccatg
ctctcgacatgcaatctcctacaaaacctagcacttattcttaatagatgatacatatgatgcttggaaactcttgatgaactagagatctcaccgaagcaattg
aaaggtgggatctcaaggggattgataacctgccagactacctgacactgtctataaggctctctagatctcagccacgaaatgaagaagagttggtcatccaag
aaaagcttatgcaattgattattataacgaaaagtggaaggatattgtaggtttcaacattcaatcaaaagttctcaaaaaggttctcaaaagttcagatgactac
ttgaaggttgattgtttacaagcagctacaatctgatcgtagcaggatccttgatgcttgaattatgcaaccagggaaggtttgattggctgcatctactcaaaaat
aatgggtggccacatgatgattggtcttacttaatgatgttgcactcacaagttgaggaagaagaggtactgcaactgcaatccactttatgaacaattatagt
ataccagaagaagaagcaatgatgaaactagaagagcttcagaggatgcatggaaggattaaatgaggagtgatcaagccaacaactgtccaagggaa
gttctgtcctcgattctcggtgcacgcttctgatgtttgctacaacatgatgtggatgatacactactcctatcaaatattgaagtccacatcattgcatgcttga
gatcctattattatctaa

>tps9_trinity_dn10529_c0_g1_i9.p1

atgtattgcttctctagtagttttcttccatgtatccttfaatcaagaaccgcagaaactgagcttcccactctggaagaatatcacctctagaagccaaaaact
actgaagcagcaacaactcgtgatacggctgatcaaaagtccaaaggcgaacagctaattacaagccaaatatctggaatatgatctctacagtccttacca
acaacttttcgaaaacaagatgcaaaagaagctgagaaactggaagagcatgtgagacggatgctgctggcacaatgatctgattccaagctggagcttata
gacagcatttcaaacctagggtcaccagctactttgagaaggaaatcaagacgaactggatgcccattgtctttgtgagacgagctcctgaagacaaaaggaga
ccttatgctacttctacagtttaggtcttaaggcaagaaggctataatccatcacaagatgcttccgggagttcatggacaaagtggaatcatgctgattcca

gtttcaaatgtcaaagggatgctgaactctttcaagcctcaaactctatgctcagaaggagagaacatattgaaggaggccagattatttactctaaacatctcaaggg
ggttaccatgaattcaaaagatcataaacttaaacaaatcttagtgattaagtctaccaatgcactgaaagtgaatggatggcatcaggaacatattgatgatc
acgaagctgagattaaaacaagctccgagttggtgagctggcaaagatcaatttaatactgccaagctgcacaccaggaagatctcaaggagttgcaaggtggt
ggataaatcttgagtaactaaaagttgagcttccagggatcggaatgctgaaagtttctgttgcaggggaatcgctttagcctcagcatcgaattctaagg
aaatggctggccaaagtattaactgatactgataatagatgatgtttatgatatttaccggtccttgaagaattagaatgcttactgatgccctacagagatggaagc
ctgagcaaataaatgaattacctgagtgatgaagatctgttctgggcattatacaactacagaggagatagctgctgaaattcagaaagaaaaggctcaac
acagtattaccgtatctacagaaagcatgggcagattttgcaattcctgctcacggaagctaagtggtttaacaaggggtatacaccatcactaaagagattttggac
aacgctgggtatcatcctcagggcctctactgcccgtcacgtaatactggaataggaaatcaaacatccaaggatattcagaactctgaaagataaccgtgactg
attactatacatcactcataattcggctatgcaatgatgaggaactctcggcggaaactagagagaggagatgctcctctgctgattctgtgtacatcgagaagc
aaatgttcagaagaggaggttagagaacatgttcggaccatattaacaaacatgggataaatcaacagtcattgcatgaatagtgctaattggtacatttctgca
agaacctgtgaaaaaacatataataaacacagcgagaacagcacatttcatgtaccaaaatggagatggattggaatacaagatgtagagactcgggaacagat
actctccagttgattgatccactagctctcaactaa

>tps10_trinity_dn2434_c0_g1_i29.p1

atggtccgcatcaagcatcagtttctccataactcagaggcaccaaagcaagcaatccctcgtcgtcagccaatttctccaagcgtttggggagaatattttctc
acatccgttctgatccgaaggaatctttacagctgcagaagagatagagcatcccacttaagaagaggttgcaaatatgcttactggaattcccgacaaatca
tggcgaagcttgactgattgacaagatacaacgtctagggtctcttacctttgaaaatgagatagaggcatcactgctcaatataccatgcttacaatgacggt
gcctttgaagatgagaatgatattgatgtagctctctgcttctcagacaaagcggccattacgcttctgcatcgtgcaataaattcaagaactcagaa
gggaaattccaagaatccatggttagtgacctgaaggaatattgagctgtacgaggccttaattacagagttcagggagaattacttagaggaagcattgacg
gtfactactctcagctgaaatccatgcttctaagttgagtgattccctgctactcaggttaacaagctctcaaacatccaagccaaaagcgttgagaaggctggga
ggattacagtacatatccatataccaagaagatgaatcacacaatgatgtattactaaattggcaaaacttgattcaatagattgagaagttgcatcagatggagct
acgtgaaatagccctatggtgaaagaattggattgaaagaaactccctttgagagagatagattagtagaatgctctttgtgcatcggcagtgatttcgagcc
ccagatggccttgcagaagtttctaaccaaaattggggcctgacttccatcattgatgacattacgatgctatggaactctgatgaactcatttctacagatga
atagaaagttgggacataaaggctacaaatgaattaccaacatacctgagatagctttatcaggttctattagatgtttatgccgaaatggaggaagattggccaaga
aaggaaaatcagacagaattaactatgcaaaggaagagatgaaaaaattagtagcagcatatataccaagaagctcattggtgacagggcctatataaccaccatt
ggaggagtgatgaaggtctgctgctctcttgcggatacatgatgctagcaactacagccatggtcggcatgggggatttagcaaccaaggaatgtttgattgggtg
actaatgagccaccgattgtacgagcggcatcagtaattgtagattaaccgatgacatggttgatagaggatgaaaggaagaggagatggcctttctgttg
aaggttacatgaaagaataggtgtgcaaaagcaacaggcgttccatgaactcagaacaggcctggaaggatataaacagagaatgtttaaagtc
aactgcagcaccaatgccatcctggagcgaacctgaatctagcaaaggtgatagatcttctgtagagatgaagatgggtatacaaatccaaaaccattgcta
aagaactcatcacctgctgctgttgatccaatcagataa

>tps11_trinity_dn4739_c0_g1_i10.p1

atgtcttcttcttctgtgtagcctatcatcatcaacttctccctttctccggtagcagcaaatcgtcactttatataccatttctccggctattaatggacctgccgatact
tttaacggatctgcccactggaagggcctgactttacctttgccttcagtgtagtactgttcgagtcctgacgcaggaatacaagatcaagaagtgccca
actggcgtgccagtaatacaaaagtgccagagacactgaaagaggccgtcgaagaaagaaactgcagagccgctgctcattaaacgtgctgagtcataaga
tcaatgctccgttccatggacgatggagtaataagcatalcggcttacgacacagcctgggttccctcgtggaagatgtaattggaagcggcactccacagtttctc
cagcttgaatggatagccaacaatcagcttctgatggctgctggggcagacatacatcttttagctcagatcgcacatcaacaccttagcctgtgtgttgatta
aaatctggaacatgcatctgacaaaagagaaaaaggttatctttatcagagataacatctgcaagctggagatgagaatattgagcagatgccattggtttga
agtagcatttctttaaattgataggcaagaagttaggaatcgacttgcagatgattctcatcccattctgcaagagatatacgaagggagaaatctaagctcaa
aaggataccgaaggacgttctgcaaacagtcccaccacattctcatagtttagaaggaatggcaggccttaactgggaaagtctactgaaactgaagttgctg

atggttcgtctctgttttccctcctcactgcccttgcgctcatggagactaaagatccagattgctcggatataactaaaattgtcaaaaattcagtgggggagtcc
gaatgtctatccagtggaacttattcgagcacattgggtgttgatcgactgcaaagactgggatttctcgatattccagacagaaattgaggaatgtgcaattatgtca
agagggcattggacagacaaaaggtatctgctgggcaagaaatacccgggtcaggacattgatgacacagctatgggttttaggattcaagggtgcatggtcacaatg
tttccagaggtgtccgacattttgagaaaggtggtgaattcttgcctcacgggacagtcgaaccaagcggtgaccgggatataacctgtacagggcttctgag
attatgttccggggaggagattctgtgatccaagaaattctcgtccgacttctacaagaaaagcgagcaaaataatgagctttagacaatggatcataacca
aggattgcccggggagggtcggatgactagatgtccgtggtacgccagcttacctcgcgtagaaacaagattctattggaacaatacggtgccgaggatgacg
tttgattggcaagacattgtacaggatgtctaattgtaataacaacatctacctcaactaggcaactagattacaacaattgccaggaatacatcaactgagtgg
aaactcattcaaaagtgtatgtggactgtggcctgggagacctaggattaggtgaaagaaatcttactagcattttatctggcggcagcaagcatattcgaaccag
gaagatcaggagaacggctgcctggccaaaactgcagcgttatcgagacggtaaatctaatttggcagtgaggcaaccaagagcagaagaccgcttct
ccgggaattcgagcatggcagtttctgccctatgaaaaagcggaaggcacaagcaaaagactcgtcgggacttacttggaaacgctaaatcaattcacgctg
gacgcaatagtggtgcacggcagagatattcatcaacgctgcgtatgctgggcaaagtgttgtaactatgcaagctggagatgacataggcagggtcaggg
tgaagcagaactatagtcgcacgctgaatctatgtccggagatggcaaatatggatcggagaatcaattgctgtctcatataagatcaacaactcatggatc
actggtagagtttctaccaaattcgcaatttcagctcaagaatgcagagaaccgattggaataacatggtggcgtcacaagcattctataaattcagacatgcaa
gaattgtcaagtagtctttccagtagtattctctcagaagatctgatgctgagatcaagcagagatttcttagtgccaagagcttactatactgcacactgc
aatccggggacaatcaactccatagctaaagtctcttgagcagactttga

Anexo VI. Secuencia del amplicón clonado.

Secuencia de la muestra enviada a analizar

>H211028-067_L21_1FF6ZAC098_premix.ab1

acaccaattaatgctcaagcatctcccagcctctccaattatcatccgagtggttggggagaacattcctcagatgctactgatctcaaggaacctttactgccg
aagaagagatagaacatgcacgactgaaagaaggatcagaaaaatgctatcgggattcccagataatgctgcagaaactgacttgattgacccatccaacg
tctagggtgcttaccattttgaaagtgagataagggcattgctcagatataatgctcagccatgagctaacgttgaagacaataatgactttatgacgtttc
tctacgttccgactcctcagacaaaatggcattatctctgtgacgtgttcgatgaattcaaggactcgaagggaattcagagcacttgaaaatgactgaa
agggtgtgagctgtacgagggcctcaaatcagagttcatggcagaccattttggaggaagccttaattgttactaccactcaactgagctcatgatgcctaactg
agaggttctcgagcaactaaagttaaacacgcactcagatgccaatccagaaaacttgacaaggctgggaggattacagatctactgtatacaaggaagacg
aatcattgatgatgacttaactaaacttgcaaaactcgattcaatagattgcagaagatgcaccaaatggagctcggcaactaaccaaatggtggaagaattgg
actgtgccagaacattcctttgcaagagacagattagtagaatgctacttctcgatattggagtgattttgagccccattatggccttccagatggtttctgtcaaag
ttattgccttgacatccatcatcgacgacatctatgatgtatggaacctttgatgaactcattttcagcgatgcaatagaaagggtggatataagggccgcaaatca
attaccgtcactgtgagacattttaccaagttctgttagattttataacgaaatggaggaggagttggccaaaagaggcatatctggcagagttgactatgcaaaaag
aaaagttgaaacaattagcggggcacttccaagaagcttgcctgggagcacaacggatagccaacatttgacgagatcatgagggtcacactcgtgtctgc
gggatataatgatgcttgcacacgctccttggcgtggtgagacatagcaactgaggaaagttggccttgatgaccaatgaaccattgatccttcgagcagatca
gtaatttgatagattaatggatgacatggtttgctgagagtgagaaggaaagaggagaagtggcctcagctgtggcatgctacatgaaggaataggtgtttcaaaag
aagaggcgttgatgaactcataaacaggtgacagatgctggaaggatataaacaagaacgcttaaacgtccaagcactgagatcaatgccatccttga
cggagctcctgaatctggcaagggtgataaatctctactacaaagatgaagattgtatacaaatctacaacctagtaaagaactcat

