UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica

etsinf

DIPLOMA DE ESPECIALIZACIÓN EN

BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

TRABAJO DE FIN DE DIPLOMA

Análisis de RNA-Seq, ensamblaje de transcriptoma *de novo*, y anotación funcional de dos cultivares de tomate de árbol (*Solanum betaceum* Cav.)

Presentado por:

Juan Pacheco

Valencia, enero 2022

Director: Dr. Jaime Prohens Tomás

Cotutores: Rubén Grillo Risco, Francisco García García

Contenido

Resumen

El tomate de árbol o tamarillo (*Solanum betaceum* Cav.) Es un cultivo frutal subutilizado originario de la región andina y filogenéticamente cercano al tomate y la papa. Los frutos del tomate de árbol tienen un alto contenido en nutrientes y compuestos bioactivos como carotenoides, antocianinas, flavonoides y vitaminas. Debido a su potencial de desarrollo como nuevo cultivo, el tomate de árbol se ha introducido en varias regiones y países fuera de su región de origen. Sin embargo, hasta el momento no hay estudios a nivel de genoma o transcriptoma para esta especie. En este estudio, realizamos el ensamblaje *de novo* y la anotación del transcriptoma para dos accesiones de tomate de árbol, una con frutos de color púrpura (A21) y la otra con frutos de naranja (A23). Se obtuvieron más de 38 y 54 millones de lecturas de alta calidad de los tejidos de las hojas y los botones florales de A21 y A23 mediante secuenciación de ARN. Se han ensamblado un total de 174,252 (A21) y 194,417 (A23) transcritos con una longitud promedio de 851 y 849 pb. Una cobertura significativa (> 98%) del Benchmarking Universal Single-Copy Orthologs (BUSCO) indicó que los transcriptomas se han ensamblado con un alto grado de integridad. Según la búsqueda de similitud de secuencia, 34,636 (A21) y 36,224 (A23) transcritos mostraron una similitud significativa con proteínas conocidas en la base de datos Swiss-Prot. De manera similar, se identificaron y anotaron con éxito los péptidos señal 1,623 (A21) y 1,745 (A23), 6,899 (A21) y 7,216 (A23) transmembrana y 22,954 (A21) y 23,637 (A23) Pfam. Entre los unigenes anotados, 22.096 (A21) y 23.095 (A23) se asignaron a la anotación de término de Gene Ontology (GO) y se encontró que 14.035 (A21) y 14.540 (A23) tenían clasificaciones de términos Clusters of Orthologous Group (COG). Además, se asignaron un total de 22.096 (A21) y 23.095 (A23) transcritos a 155 y 161 (A23) rutas metabólicas de la Kyoto Encyclopedia of Genes and Genomes (KEGG). Nuestros datos mostraron que los términos GO del proceso biosintético de carotenoides se enriquecieron significativamente en la accesión A21. Finalmente, se

identificaron in silico 68.647 single nucleotide variations (SNV) y casi 2 millones de SNV interespecíficas (1.973.023 con *S. tuberosum* y 1.809.264 con *S. lycopersicum*). En general, los resultados de este estudio proporcionan una gran cantidad de datos genómicos para mejorar nuestra comprensión de los genes involucrados en las rutas metabólicas clave para dirigir los esfuerzos futuros en la mejora genética del tomate de árbol. El elevado número de SNV identificados será muy valioso para la mejora asistida por marcadores y otros estudios genéticos en tomate de árbol.

Introducción

El tomate de árbol o tamarillo (*Solanum betaceum* Cav.) Es un cultivo de solanáceas originario de la región andina (Duarte, O., & Paull, 2015; Orqueda et al., 2017). El tomate de árbol está relacionado filogenéticamente con la papa y el tomate, formando parte del mismo clado (Särkinen et al., 2013). La planta de tomate de árbol es un árbol pequeño, aunque algunos cultivares pueden crecer hasta cuatro metros, con un sistema radicular de rápido crecimiento, poco profundo y un desarrollo reproductivo y vegetativo simultáneo (Ramírez & Kallarackal, 2019). En los últimos años, el tomate de árbol ha llamado la atención de los productores y de la industria gracias a sus atractivos frutos comestibles, que pueden ser consumidos en ensaladas o como fruta de postre, o procesados para la elaboración de mermeladas, yogures, jugos o bebidas alcohólicas entre otros (Chen et al., 2020). De ser un cultivo desatendido, con un interés local en las granjas de subsistencia (Acosta-Quezada et al., 2011), se ha convertido en un cultivo frutal prometedor, habiendo sido introducido en varios países de Oceanía, Sudeste Asiático, Europa y África (Diep et al., 2020). Aparte de los países de América del Sur, Nueva Zelanda es el mayor productor y exportador y donde la palabra comercial "tamarillo" se acuñó del término maorí "tama" que significa liderazgo con la palabra española "Amarillo" o la palabra "Tomatillo" que significa tomate pequeño (Diep et al., 2020). El interés por el tomate

de árbol también radica en sus considerables cantidades de antioxidantes, vitaminas y carotenoides presentes en la fruta. Las porciones estándar de tomate de árbol proporcionan 67-75% de la ingesta dietética recomendada (IDR) de ácido ascórbico, 16-23% IDR de α-tocoferol y 9-20% IDR de β-caroteno (Diep et al., 2020). Sin embargo, el perfil fitoquímico del tomate de árbol varía entre cultivares y condiciones ambientales (Diep et al., 2020). Los principales grupos de cultivares (anaranjado, anaranjado punton, morado, rojo y rojo cónico) se diferencian por el color y la forma del fruto, con diferentes rangos de variación morfológica y genética entre ellos (Acosta-Quezada et al., 2011, 2012). A pesar del gran potencial del tomate de árbol como un importante nuevo cultivo frutal, no se han realizado estudios genéticos o genómicos de alto rendimiento para esta especie. Los avances recientes en RNA next-generation sequencing (RNA-Seq) y de los recursos bioinformáticos facilitan los estudios transcriptómicos incluso para especies de plantas no modelo donde no se dispone de genomas de referencia (Huang et al., 2016; Ward et al., 2012). De hecho, RNA-Seq se ha realizado con éxito y cada vez más para descifrar el transcriptoma vegetal de especies de plantas desatendidas (Xia et al., 2011). Sin embargo, la secuenciación de ARN ofrece muchas otras características interesantes como evaluación de la expresión génica, descubrimiento de polimorfismos, filogenómica, descubrimiento de variantes de empalme, entre otras (Herraiz et al., 2016). En este estudio, realizamos la secuenciación del transcriptoma y el ensamblaje de dos accesiones de tomate de árbol con diferentes colores de fruta (morado y anaranjado) seguido de su anotación estructural y funcional integral. Además, se identificaron polimorfismos intraespecíficos entre los dos cultivares e interespecíficos con tomate y papa. Los transcriptomas y la información generada en el presente estudio serán un recurso útil para futuros estudios genómicos y moleculares y serán una herramienta genómica clave para ayudar a los programas de mejoramiento del tomate de árbol.

Materiales y Métodos

Material vegetal

Para el presente estudio se utilizó una accesión de tomate de árbol de fruta morada (A21, con epicarpio morado y peso medio del fruto de 108,8 g) y una accesión de tomate de árbol de color anaranjado (A23, con epicarpio anaranjado con un peso medio de fruta de 75,1 g) (Acosta-Quezada et al., 2011) (Figura 1), obtenidos del banco de germoplasma de la Universitat Politècnica de València (UPV). Las semillas de cada accesión fueron germinadas siguiendo el protocolo de Ranil et al., (2015). Posteriormente, las plantas fueron cultivadas en un invernadero en la UPV, España (coordenadas GPS: latitud, 39° 28' 55" N; longitud, 0° 20' 11" O; 7 m sobre el nivel del mar). De cada accesión, el tejido fue muestreado de hojas jóvenes y de botones florales. Todas las muestras recogidas se congelaron inmediatamente en nitrógeno líquido y se almacenaron a −80 °C para su uso posterior.

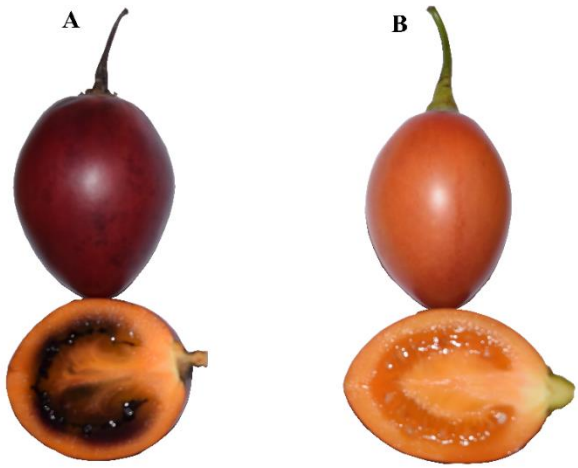

Figura 1. Frutos de las accesiones de tomate de árbol A21 (A) y A23 (B).

Extracción de ARN, construcción de bibliotecas y secuenciación de ARN

El ARN total se aisló de cada tejido utilizando el Mini spin kit (Macherey-Nage, Dueren, Alemania). La integridad del ARN fue determinada por electroforesis del gel de la agarosa

1,0% (w/v) y la cuantificación del ARN fue realizada por el fluorómetro de Qubit 2,0 (Thermo Fisher Scientific, Waltham, MA, los E.E.U.U.). Para cada una de las dos accesiones, combinamos cantidades equivalentes de ARN de los dos tejidos muestreados en un pool. Un total de 2 μg de ARN total para cada grupo fue enviado a Novogene (Cambridge, Reino Unido) para la preparación y secuenciación de la biblioteca. Las bibliotecas paired-end de cDNA de 150 bp (tamaño de inserción de 250 ~ 300 bp) se construyeron de acuerdo con las instrucciones de Illumina. El ARNm de cada muestra se purificó del ARN total mediante el uso de Sera-mag Magnetic Oligo (dT), luego se dividió en fragmentos cortos utilizando el búfer de fragmentación. Usando estos fragmentos como plantillas, la primera hebra de cDNA fue sintetizada. La segunda hebra de ADNc se sintetizó utilizando el tampón que contenía dNTPs, ARNasa H y ADN polimerasa I. Fragmentos cortos (200 ± 20 pb) se conectaron a los adaptadores de secuenciación y fragmentos adecuados se extirparon de un gel de agarosa mediante un kit de extracción de gel. Luego, la biblioteca fue secuenciada usando el secuenciador Illumina Hiseq-2000.

Procesamiento de secuencias de ADN y ensamblaje de transcriptomas de novo

La calidad de las lecturas se evaluó mediante FastQC v0.11.8 (Andrews, 2010). Las secuencias del adaptador, las lecturas de baja calidad (Phred <30) y las lecturas con una longitud media inferior a 135 pb se recortaron utilizando Trimmomatic v0.36 (Bolger et al., 2014). Las dos accesiones se ensamblaron por separado utilizando el software Trinity v2.10 (Grabherr et al., 2013) con un tamaño k-mer predeterminado de 25. Los contigs idénticos o casi idénticos se agruparon en un solo contig por la herramienta CD-HIT-EST v 4.8.1 (Fu et al., 2012) con una identidad de más del 80%. La calidad y la integridad de los ensamblajes se evaluaron en primer lugar con Bowtie2 v2.3.2 2 (Langmead & Salzberg, 2012) para evaluar el número de lecturas que estaban presentes en los transcritos ensamblados, luego la longitud de contig del transcrito Ex90N50 (el valor de contig N50 basado en el conjunto de transcritos que representan el 90%

de los datos de expresión) se calculó utilizando contig ExN50 statistic.pl script incluido con Trinity. Finalmente, se evaluó la integridad de los ensamblajes utilizando BUSCO v4.1.1 (Simão et al., 2015; Waterhouse et al., 2018) utilizando un conjunto de genes eucariotas como base de datos (https://busco-data.ezlab.org/v5/data/lineages/eukaryota_odb10.2020-09-10.tar.gz).

Anotación estructural y funcional

Los marcos de lectura abiertos (ORF) se predijeron utilizando Transdecoder v5.5.0 (http://transdecoder.sourceforge.net/) utilizando los transcriptos ensamblados como entrada. Después de que los ORF fueron extraídos del ensamblaje, los contigs redundantes con más del 90% de identidad fueron eliminados usando CD-HIT-EST. La anotación funcional de los transcritos ensamblados se realizó utilizando el software OmicsBox v 1.4.11 (Götz et al., 2008) y el pipeline Trinotate v3.2.1 (https://trinotate.github.io/) (Bryant et al., 2017). Tanto los transcritos de nucleótidos como las secuencias de proteínas se alinearon contra la base de datos UniProtKB/Swiss-Prot (uniprot_sprot.trinotate_v2.0.pep.gz), utilizando NCBI-BLASTx y BLASTp v2.10.1+ (-evalue 1e−3 -max_target_seqs 1 -outfmt 6). Los dominios funcionales se identificaron utilizando la base de datos de dominios Pfam (Pfam-A.hmm.gz) utilizando HMMER v3.3.1 (Wheeler & Eddy, 2013). Los péptidos de señal potenciales se identificaron utilizando la herramienta SignalP v4.1 (Petersen et al., 2011). El programa OmicBox (https://www.biobam.com/omicsbox/) se utilizó para anotar aún más los transcritos utilizando la función de anotación funcional del software Blast2GO para predecir términos de ontología génica (GO), código de enzimas EC, identificar posibles vías KEGG (Kyoto Encyclopedia of Genes and Genomes) y relaciones de ortología utilizando las bases de datos eggNOG v5.0 (Götz et al., 2008; Huerta-Cepas et al., 2019; Kanehisa & Goto, 2000). El análisis de enriquecimiento de GO se realizó utilizando el paquete de bioconductores 'topGO (http://www.bioconductor.org/packages/release/bioc/html/topGO.html).

Variaciones de nucleótido único (SNVs)

Para los SNVs intraespecíficos, las lecturas limpias de cada accesión se mapearon por separado al transcriptoma ensamblado A23 que actuó como referencia utilizando BWA v0.7.17 (Li & Durbin, 2009), mientras que para los SNVs interespecíficos las lecturas se mapearon contra el genoma de referencia de tomate (Heinz 1706 version SL4.0, Hosmani et al., 2019) y papa (DM 1-3 516 R44 v6.1; Pham et al., 2020). Posteriormente, SAMtools v1.10 (Li et al., 2009) se utilizó para convertir SAM a formato BAM, mientras que las lecturas duplicadas se eliminaron de las respectivas secuencias de alineación utilizando Picard-tools v2.23.8 (http://picard.sourceforge.net). Las variantes fueron llamadas por FreeBayes v1.3.4 (Garrison & Marth, 2012) para identificar polimorfismos intra e interespecíficos que se filtraron utilizando VcfFilter v0.2 (https://github.com/biopet/vcffilter) basado en un minQualScore de 30, minTotalDepth de 40 y un minSampleDepth de 20. Finalmente, los efectos de impacto variable se predijeron utilizando SnpEff v5.0 (Cingolani et al., 2012).

Resultados y Discusión

Secuenciación y ensamblaje del transcriptoma

La secuenciación de ARN de las dos accesiones de tomate de árbol generó 100,919,310 (14.68 Gb) y 113,802.281 (15.84 Gb) lecturas paired-end para A21 y A23, respectivamente (Tabla 1). Después del recorte inicial y el estricto filtrado de calidad para eliminar adaptadores y datos de baja calidad, se obtuvieron 38,411,167 (4.25 Gb) de lecturas limpias paired-end para A21 y 54,474,055 (5.97 Gb) para A23 (Tabla 1).

Tabla 1. Resumen de estadísticas de lecturas antes y después del procesamiento, ensamblajes *de novo* e integridad de BUSCO para las accesiones de tomate de árbol A21 y A23. (Tabla de Pacheco et al., 2021)

	Accessions	
Statistics	**A21**	**A23**
Total raw reads	100,919,310	113,802,281
Total raw reads data size (Gb)	14.68	15.84
G/C (%)	42.2	42.2
Total clean reads	38,411,167	54,474,055
Total clean reads data size (Gb)	4.25	5.97
Number of transcripts	174,252	194,417
Total nucleotide length	148,352,996	165,074,290
Average transcript length	851.37	849.07
Maximum transcript length	17,046	16,865
N50	1,494	1,503
G/C (%)	38.8	38.6
Overall alignment rate (%)	99.09	99.21
BUSCO (%)	98.4	98.8

Las lecturas limpias se ensamblaron independientemente en los transcriptomas usando Trinity. Para la accesión A21, el transcriptoma ensamblado consistió en 174,252 transcritos comprendidos por 148,352,996 pb, con una longitud media de transcrito de 851.37 pb (Tabla 1). El valor de N50 fue de 1,494 pb y el contenido de GC de 38.8% (Tabla 1). Por otro lado, el accession A23 fue ensamblado en 194,417 transcritos con una longitud total de 165,074,290 pb y una longitud media de 849.07 pb (Tabla 1). El valor de N50 para este último fue de 1,503 pb y el contenido de GC de 38.6% (Tabla 1). Las longitudes de secuencia ensambladas variaron desde el valor de corte de 200 pb hasta una longitud máxima de transcrito de 17,046 pb para A21 y 16,865 pb para A23 (Tabla 1). La mayoría de las secuencias ensambladas estaban en los rangos de 200 pb a 500 bp y 501 a 1,000 bp. El número de transcritos de estas accesiones fue mayor que los obtenidos en estudios anteriores de transcriptoma en otras Solanáceas relacionadas como el tomate, la papa y el pepino (Herraiz et

al., 2016; Moon et al., 2018; Scarano et al., 2017), pero fue similar a otras obtenidas en especies vegetales de la misma familia, como *S. commersonii* Dunal y *S. aculeatissimum* (Yang et al., 2017; Zuluaga et al., 2015), sugiriendo la alta calidad de nuestro ensamble.

Para evaluar la calidad de los ensamblajes, las lecturas limpias se mapearon de nuevo al transcriptoma ensamblado. Las tasas de alineación general utilizando el software de alineación Bowtie2 fueron de 99.09% para A21 y 99.21% para A23 (Tabla 1). BUSCO se empleó para evaluar la precisión y la integridad de nuestro ensamblaje de transcriptomas. Al comparar el conjunto de genes con el genoma, encontramos que la proporción de BUSCO fue de 98,4% para A21 y 98,8% para A23 (Tabla 1), estos valores son comparables o incluso superiores a los de otros transcriptomas recientes de *Solanum* que exhibieron valores de 97% para *S. tuberosum* y 93% para *S. chilense* (Petek et al., 2020; Stam et al., 2019).

Anotación estructural y funcional

Se utilizó el software TransDecoder para identificar los marcos de lectura abiertos (ORF) de los transcritos ensamblados y sus funciones asociadas, prediciendo 27,441 ORFs y 34,636 potenciales proteínas para A21 y 28,336 ORFs y 3,224 636 potenciales proteínas para A23 (Tabla 2), los resultados estuvieron de acuerdo con los observados para los genes codificantes de proteínas en otras especies de *Solanum* como el tomate (35,35), papa (39,290), y berenjena (30,630 y 34,231) (Aversano et al., 2015; Gramazio et al., 2016; Wang et al., 2020). Posteriormente, los transcritos únicos y las proteínas identificadas fueron anotadas realizando búsquedas de Blast contra varias bases de datos utilizando el pipeline Trinotate. Un total de 57,422 (33.0%) y 60,772 unigenes (31.3%) mostraron una homología significativa cuando se realizó Blastx y 24,311 (14.0%) y 25,054 secuencias de proteínas (12.3%) cuando se realizan búsquedas Blastp en la base de datos UniProtKB/Swiss-Prot (valor E de corte de 1e-3) para A21 y A23, respectivamente (Tabla 2).

Tabla 2. Resumen de la anotación funcional por homología de transcriptomas para las accesiones de tomate de árbol A21 y A23. (Tabla de Pacheco et al., 2021)

	Accessions	
Statistics	**A21**	**A23**
Predicted ORFs	27,441	28,336
Predicted proteins	34,636	36,224
sprot_Top_BLASTX_hit	57,422	60,772
sprot_Top_BLASTP_hit	24,311	25,054
Pfam	22,954	23,637
SignalP	1,623	1,745
TmHMM	6,899	7,216
GO terms	196,800	204,090
EC numbers	15,828	16,668
Kegg	14,035	14,540

Además, se han identificado 22,954 y 23,637 motivos únicos de proteínas Pfam, 1,623 y 1,745 secuencias de proteínas con péptidos de señal (SignalP) y 6,899 y 7,216 transcritos con al menos un dominio transmembrana (TmHMM) para A21 y A23, respectivamente (Tabla 2), Estos porcentajes fueron superiores a los obtenidos en otras especies vegetales de la familia Solanaceae como *S. trilobatum* y *S. sisymbriifolium* (Lateef et al., 2018; Wu et al., 2019). La distribución de especies mostró que la mayoría de las secuencias exhibieron alta similitud principalmente con las de *Arabidopsis thaliana* (17,602 para A21 y 18,117 para A23), grupo *Oryza sativa* japonica (1,039 y 1,066), *Nicotiana tabacum* (613 y 653), *S. lycopersicum* (487 y 486) y *S. tuberosum* (267 y 276) (Figura 2).

Figura 2. Distribución de unigenes de especies anotados para las accesiones A21 y A23 de tomate de árbol. (Figura de Pacheco et al., 2021)

La clasificación funcional basada en GO para los ensamblajes de transcriptomas A21 y A23 recuperó un total de 196,800 términos GO para A21 y 204,090 para A23 de 22,096 y 23,095 transcripciones, respectivamente (Tabla 2). El mayor número de términos GO (75.2%) se anotó en secuencias con una longitud entre 100 y 500 pb (Figura 3).

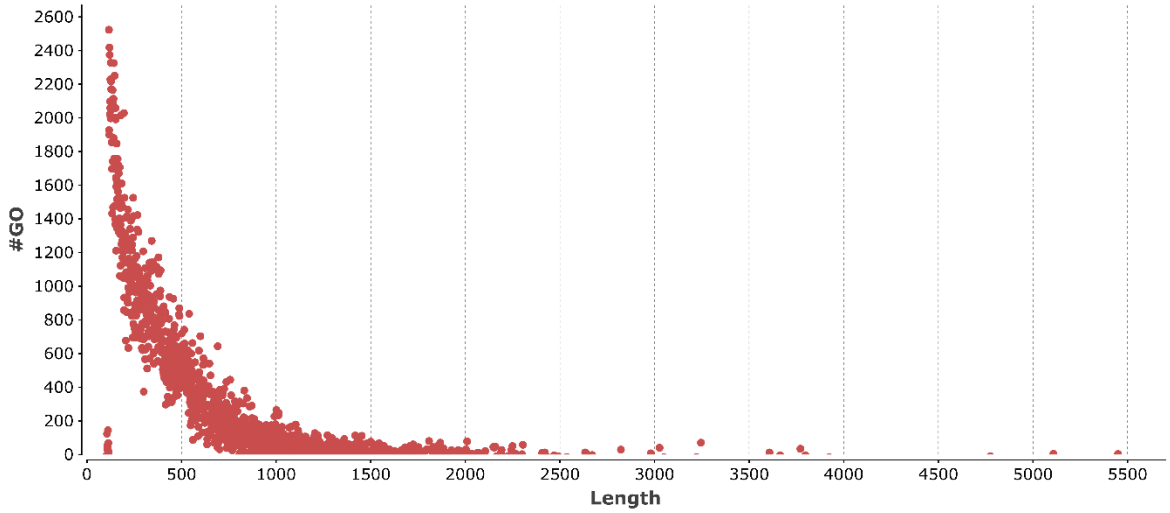

Figura 3. Números de términos GO relativos a la longitud de las secuencias en los transcriptomas de las accesiones A21 y A23 de tomate de árbol. (Figura de Pacheco et al., 2021)

Ambos ensamblajes tenían una distribución GO similar para cada categoría; de cuatro a nueve términos en la categoría de proceso biológico (BP), de tres a nueve en función molecular (MF) y de cuatro a ocho en la categoría de componentes celulares (CC) (Figura 4). Los niveles de GO que oscilaron entre 5 y 15, fueron del 88.9 % para los procesos biológicos, del 69.8 % para la función molecular y del 88.2 % para los componentes celulares, lo que indica que la precisión de la anotación fue buena (Figura 4) y que se muestreó una amplia diversidad de genes en nuestros transcriptomas.

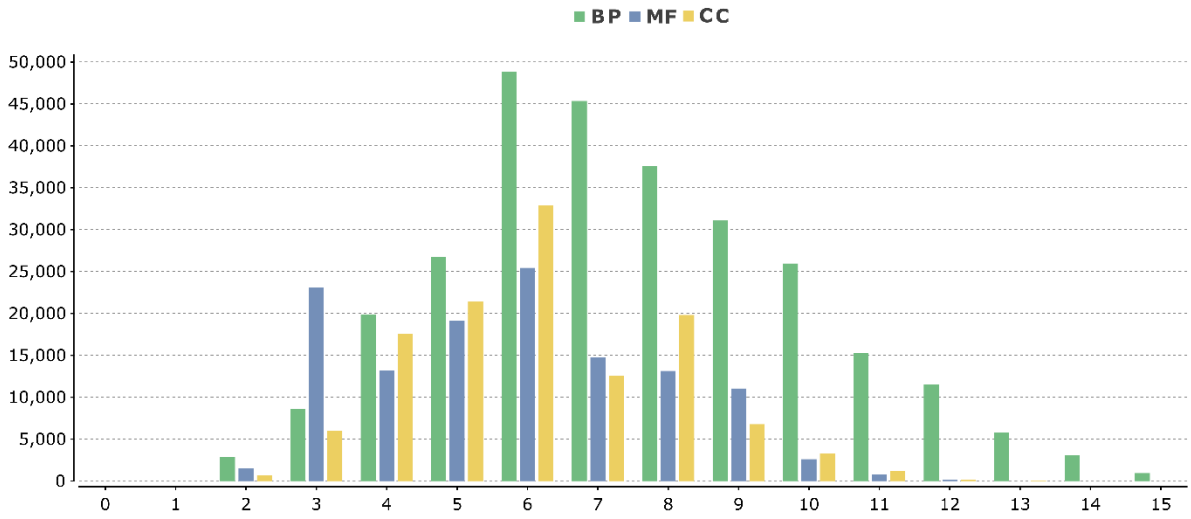

Figura 4. Distribución del nivel de GO en cada categoría para los unigenes de tomate de árbol anotados. El eje X representa el nivel GO y el eje Y el número de unigenes anotados. BP = Proceso biológico, MF = Función molecular, CC = Componente celular. (Figura de Pacheco et al., 2021)

Entre todos los términos GO extraídos, 137, 333 (69.8%) para A21 y 140,193 (68.7%) para A23 fueron asignados a la categoría de proceso biológico, 35,153 (17.9%) y (38,464) 18.9% a clase de función molecular y 24,314 (12.4%) y 25,233 (12.5%) a los componentes celulares, respectivamente (Figura 5). Para la categoría de proceso biológico, las tres principales subcategorías fueron proceso celular con 19,220 (14.0%) secuencias para A21 y 20,660 (14.8%) para A23, proceso metabólico con 15,954 (11.6%) y 17,273 (12.3%) y respuesta al

estímulo con 13,134 (9.6%) y 13,400 (9.6%) secuencias (Figura 5). Para la categoría de función molecular, la gran mayoría de las secuencias pertenecían a dos subcategorías, binding (15,824; 45.0% para A21 y 18,204; 47.3% para A23) y la actividad catalítica (11.940; 34,0% y 13.566; 35,3%) secuencias (Figura 5). Finalmente, para la categoría de componentes celulares, la mayoría de las secuencias se clasificaron en dos subcategorías, entidad anatómica celular (20,315 secuencias; 83.6% y 21,109; 83.0%) y complejo que contiene proteínas (4,315; 16.4% y 3,998; 17.0%) (Figura 5).

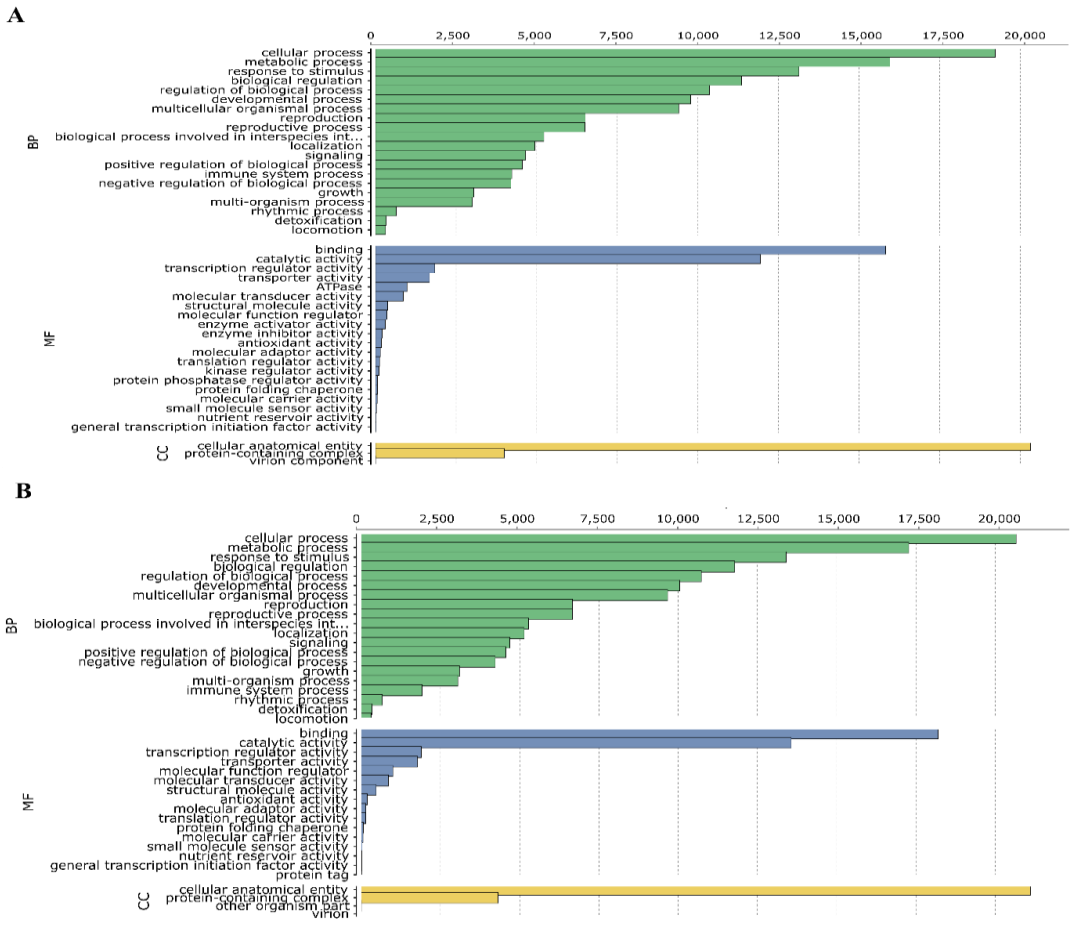

Figura 5. Clasificación funcional de ontología génica (GO) de los transcriptomas A21 (A) y A23 (B) del tomate de árbol. Histogramas de transcripciones anotadas en categorías específicas de GO; BP = proceso biológico, MF = funciones moleculares y CC = componentes celulares están representados por barras verdes, azules y amarillas, respectivamente. (Figura de Pacheco et al., 2021)

Para A21, el análisis de enriquecimiento del término GO indicó términos significativos de GO asociados con la respuesta de defensa (GO:0006952), la proteólisis (GO:0006508), el proceso metabólico celular y lipídico (GO:0006629, GO:0044255), el proceso metabólico de carotenoides (GO:0016116) y el proceso biosintético de carotenoides (GO:0016117) (Figura 7). Por el contrario, en A21 los términos significativos de GO enriquecido de A23 fueron localización de proteínas (GO:0008104), morfogénesis radicular (GO:0010015), desarrollo radicular (GO:0048364), morfogénesis de órganos vegetales postembrionarios (GO:0090697), regulación del proceso catabólico (GO:0009894), regulación del proceso catabólico celular (GO:0031329) (Figura 6).

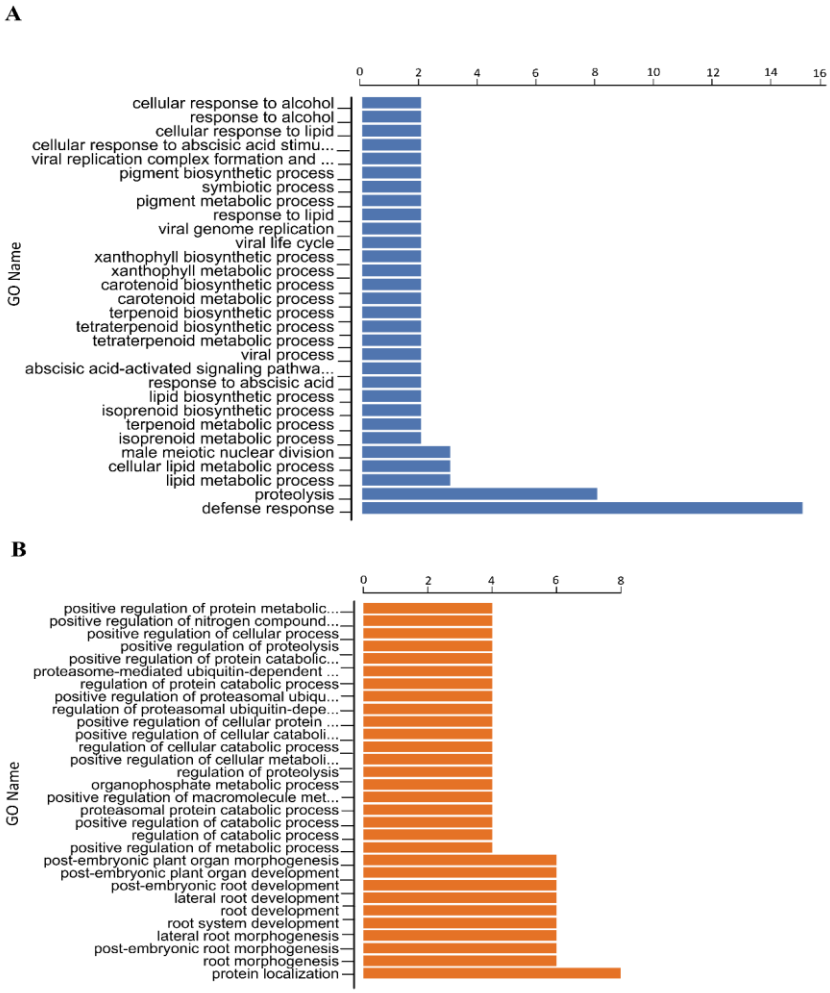

Figure 6. Análisis de enriquecimiento de GO en transcriptomas A21 (A) y A23 (B) de tomate de árbol. (Figura de Pacheco et al., 2021)

Los números de enzyme commission (EC) se asignaron a 15,828 para A21 y 16,668 para A23 unigenes (Tabla 2). Las enzimas más representadas fueron las hidrolasas (5,489 unigenes en A21 y 5,562 en A23), las transferasas (5,430 y 5.857), las oxidorreductasas (1,766 y 2,031) y las translocases (1,708 y 1,680) (Figura 7). Otras clases de enzimas como liasas, isomerasas y ligasas fueron representadas en menor grado.

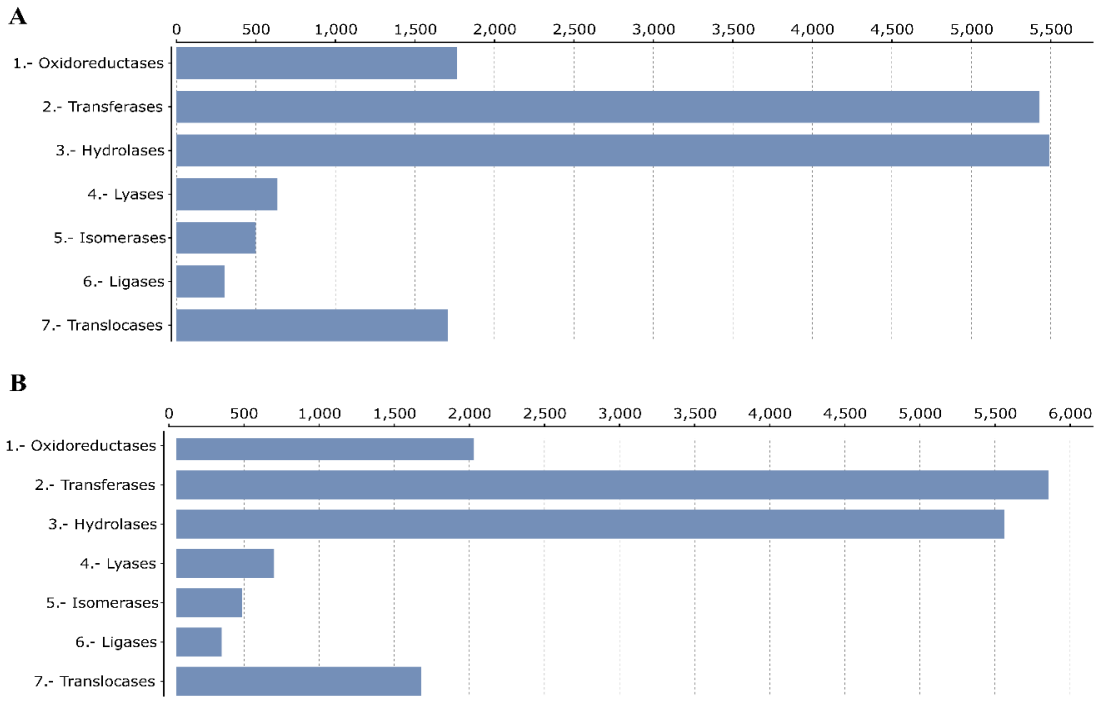

Figura 7. Número de unigenes para cada categoría de la comisión enzimática (CE) para los transcriptomas A21 (A) y A23 (B) del tomate de árbol. (Figura de Pacheco et al., 2021)

El análisis de KEGG fue realizado para identificar los mecanismos y las rutas potenciales representados en los unigenes identificados. Se asignaron un total de 14,035 uni genes para A21 y 14,540 uní genes para A23 a 155 y 161 vías KEGG respectivamente (Tabla 2). Las rutas más representadas en cuanto al número de transcripciones homólogas fueron el metabolismo de la purina (map00230, 58 secuencias), el metabolismo de la cisteína y la metionina (map00270, 58 secuencias), el metabolismo del azúcar amino y del azúcar

nucleótido (map00520, 46 secuencias), la biosíntesis terpenoide (map00900, 33 secuencias), el metabolismo de fármacos (map00983, 20 secuencias), la biosíntesis flavonoide (map00941, 17 secuencias) y la biosíntesis de carotenoides (map00906, 17 secuencias). El aumento de la acumulación de flavonoides y carotenoides en los cultivos frutales mejora sus valores comerciales y saludables (Karasawa & Mohan, 2018). Entre las características biológicas, la propiedad más reconocida de los flavonoides y carotenoides son sus efectos antioxidantes, que a menudo son mucho más altos que los de la vitamina E y la vitamina C (Britton, 2020; Nabavi et al., 2020). Nuestros resultados transcripcionales confirmaron la presencia de genes y enzimas conocidos en rutas relacionadas con la síntesis de flavonoides y carotenoides. Estos resultados están de acuerdo con estudios anteriores que informaron que el tomate de árbol es una fuente abundante de carotenoides, antocianinas, flavonoides y compuestos fenólicos y tiene una mayor actividad antioxidante que otras frutas ricas en antioxidantes como el kiwi o la uva (Diep et al., 2020). En la accesión A21, nuestros datos mostraron que el proceso biosintético del carotenoide GO términos fue enriquecido significativamente. Esto está de acuerdo con los resultados anteriores (Acosta-Quezada et al., 2015; Vasco et al., 2009) que informaron que el cultivar morado tenía niveles más altos de caroteneides en comparación con los cultivares amarillos o anaranjados. Nuestros resultados sugieren que los genes relacionados con la vía de biosíntesis de flavonoides y carotenoides están bastante bien conservados en el tomate de árbol en comparación con el tomate (Ye et al., 2015). Las variantes de secuencia en esos genes entre las variedades de tomate de árbol podrían utilizarse como marcadores funcionales para el mejoramiento asistido por marcadores para obtener nuevas variedades de tomate de árbol con valores nutricionales mejorados.

Clasificación COG

Un grupo ortólogo de clúster (COG) se define como un clúster de tres o más secuencias homólogas que divergen del mismo evento de especiación. Los grupos ortólogos fueron

anotados funcionalmente utilizando la base de datos EggNog (evolutionary genealogy of genes: Non-supervised Orthologous Groups). En total, 97, 437 para A21 y 99,471 para A23 GO se asignaron a 14,530 y 14,928 secuencias únicas respectivamente (Figura 8). El grupo más numeroso está representado por el clúster de procesos celulares y señalización (CPS) (6,311; 21.4% y 6,443; 21.0%), seguido del metabolismo (MB) (6,052; 20.5% y 6,417; 20.9%), almacenamiento y procesamiento de información (ISP) (6,040; 20.4% y 6,396; 20.9%) (Figura 8). Dentro de la categoría CPS, la mayor proporción fue asignada a, mecanismos de transducción de señales (T) (2,359 para A21 y 2,392 para A23) y modificación postraduccional, recambio de proteínas, chaperonas (O) (2,051 y 2,104), dentro de la categoría MB: transporte y metabolismo de aminoácidos (E) (1,232 y 1,342) y transporte y metabolismo de carbohidratos (G) (1,249 y 1,314) y dentro de la categoría ISP, la mayoría fueron asignados a la replicación , recombinación y reparación (L) (2,043 y 2,254) y transcripción (K) (1,997 y 2,043) (Figura 8).

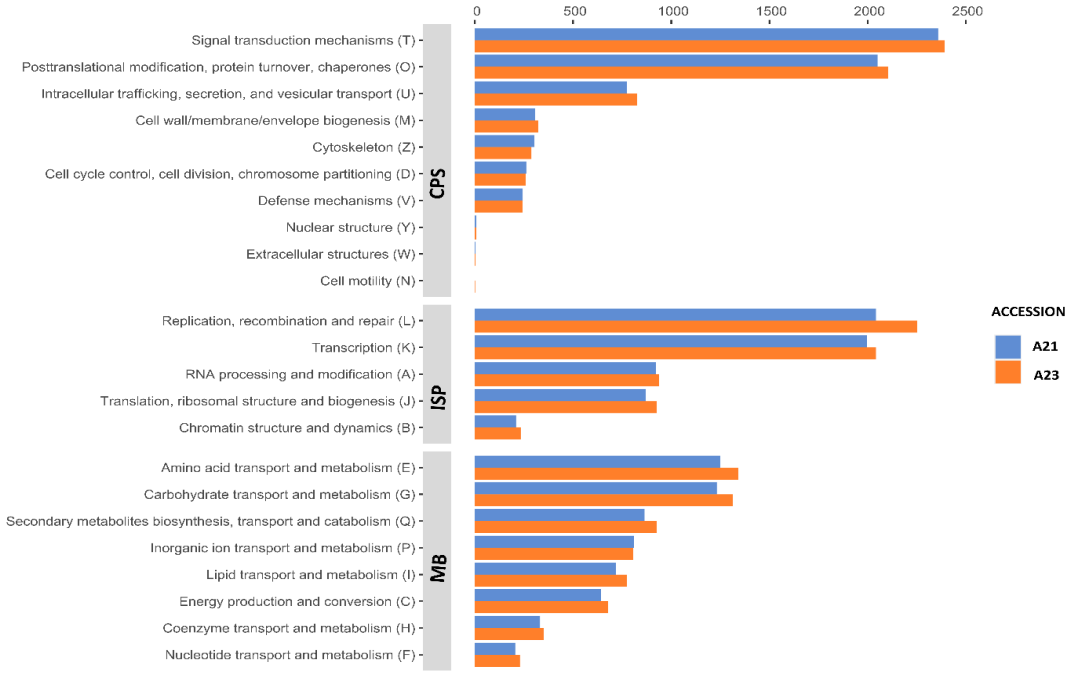

Figura 8. Categorías del COG en los transcriptomas de las accesiones A21 y A23 de tomate de árbol. (Figura de Pacheco et al., 2021)

Identificación y Caracterización de SNVs

Se identificaron polimorfismos intra e interespecíficos tanto en accesiones como entre los genomas del tomate y la papa. El número de SNVs intraespecíficos fue significativamente mayor en A23 (49,530) que en la accesión A21 (19,117) (Tabla 3). De ellos, 14,837 (77.6%) en A21 y 38,183 (77.1%) en A23 fueron SNPs, 3,283 (17.2%) y 8,213 (16.6%) polimorfismos de nucleótidos múltiples (MNP), 767 (4.0%) y 2,391 (4.8%) InDels y 227 (1.2%) y 726 (1.5%), nucleótido múltiple y un InDel (MIXTO) (Tabla 3).

Tabla 3. Estadísticas de polimorfismo para los transcriptomas A21 y A23 del tomate de árbol.

Statistics	SNPs	MNP	INDELs	MIXED	Total SNVs
SNVs intraspecific variations					
A21	14,837	3283	767	227	19,117
A23	38,183	8213	2391	726	49,530
SNVs interspecific variations					
A21 and *S. tuberosum*	619,626	174,982	28,2835	23,115	1,973,023
A23 and *S. tuberosum*	805,997	242,484	42,142	36,352	
A21 and *S. lycopersicum*	624,503	194,857	23,407	20,788	1,809,264
A23 and *S. lycopersicum*	684,775	218,205	27,102	24,627	

Los SNVs reportados aquí son más altos que los SNVs identificadas en otros estudios transcriptómicos de Solanaceae, como los 17,000 SNVs encontrados en el tomate (Scarano et al., 2017); sin embargo, en el caso de la papa se ha reportado un número similar de SNPs 69,011 (Hamilton et al., 2011).

Entre los SNPs, el número de transiciones (10,687 en A21 y 25,925 en A23) fue mayor que el número de transversiones (5,758 y 13,810), con una relación transición/transversión (Ts/Tv) de 1.86 y 1.88, respectivamente (Tabla 4). Para la sustitución de transición, los más abundantes fueron C/T (17.4% en A21 y 17.0% A23) seguidos de G/A (16.8% y 16.6%), A/G (15.7% y 16.1%) y T/C (14.2% y 15.4%) (Tabla 4). En caso de sustitución por transversión, la frecuencia de ocurrencia de los SNPs fue A/T (6.2% y 5.4%) seguido de T/A (5.6% y 5.4%), G/T (4.6% y 4.8%), C/A (4.3% y 4.5%), A/C (4.1% y 4.3%), G/C (3.0% y 2.9%) y C/G (2.7% y 2.6%) (Tabla 4). La frecuencia media de variación genómica de SNPs e InDels fue de uno en 242 pb

en A21 y uno en 204 pb en A23. En ambas accesiones, el número y la proporción de variantes heterocigotas fueron mayores,74% en A21 y 79% en A23, que las homocigóticas (Tabla 4). Esto último podría deberse al hecho de que, a pesar de que algunos cultivares de tomate de árbol se consideran autocompatibles y autógamos, las flores son visitadas con frecuencia por insectos polinizadores que pueden conducir a la polinización cruzada (Ramírez & Kallarackal, 2019).

Tabla 4. Número de (transiciones / transversiones), tasa de variante y variante homocigota y heterocigota en las accesiones A21 y A23 de tomate de árbol. (Tabla de Pacheco et al., 2021)

	A21	**A23**
Transitions	10,687	25,925
C/T (%)	17.4	17.0
G/A (%)	16.8	16.6
A/G (%)	15.7	16.1
T/C (%)	14.2	15.4
Transversions	5,758	13,810
A/T (%)	6.2	5.4
T/A (%)	5.6	5.4
G/T (%)	4.6	4.8
C/A (%)	4.3	4.5
A/C (%)	4.1	4.3
G/C (%)	3.0	2.9
C/G (%)	2.7	2.6
Ts/Tv ratio	1.86	1.88
Variant rate (bp)	242	204
homozygous (%)	74	79
heterozygous (%)	26	21

La gran mayoría de las variantes (12,095; 50.4% en A21 a 38,632; 61.3% en A23), clasificadas según SNPeff, fueron predichas como “modifier” (es decir, las variantes se localizaron en regiones intergénicas o intrónicas, o en un exón de una transcripción no codificante), lo que indica que no hay evidencia de impacto o que sus predicciones son difíciles de evaluar (Tabla 5). Los segundos efectos de impacto más abundantes previstos fueron “low” (6,507; 27.1% y

11,732; 18.7%), que son en su mayoría variantes inofensivas o poco probables que cambien el comportamiento de las proteínas (Tabla 5). Los terceros fueron los que se predijo que tenían efectos de impacto "moderate" (5,39; 21.4% y 11,637; 18.6%), es decir, variantes no disruptivas, como la inserción/deleción de codones o la sustitución de codones, que podrían cambiar la efectividad de las proteínas (Tabla 5). Finalmente, la clase de impacto menos abundante correspondió a los efectos de variación "high" (262; 1,1% y 808; 1,3%), que se consideran de impacto disruptivo sobre la proteína como truncamiento o pérdida de función causada por deleción/deleción de exones (Tabla 5).

Tabla 5. Número de efectos por impacto en las accesiones A21 y A23 de tomate de árbol. (Tabla de Pacheco et al., 2021)

	MODIFIER	**LOW**	**MODERATE**	**HIGH**
Number of effects intraspecific				
A21	12,095	6,507	5,139	262
A23	38,362	11,732	11,637	808
Number of effects interpecific				
A21 and *Solanum tuberosum*	1,325,386	547,527	461,527	8,686
A23 and *Solanum tuberosum*	1,861,338	644,515	568,120	15,010
A21 and *Solanum lycopersicum*	5,568	377,062	291,925	5,568
A23 and *Solanum lycopersicum*	1,110,318	403,147	319,944	6,700

Las principales categorías de variantes fueron en regiones de exones (48% para A21 y 37%, para A23), región intergénica (21% y 30%), variante 3' UTR (17%), variante UTR 5' (14% y 16%) y variante sinónima (25% y 17%) (Tabla 6). Con respecto a los efectos sobre la función proteica, en promedio, (58% en A21 y 52% en A23) de las variantes se predijo que producirían un efecto silencioso, (41% y 47%) un impacto sin sentido y (1%) un producto proteico sin sentido (Tabla 6).

Tabla 6. Porcentaje de efectos por región y clase funcional en las accesiones A21 y A23 de tomate de árbol. (Tabla de Pacheco et al., 2021)

	A21	**A23**
Exon	48	37
Intergenic	21	30
3' UTR variant	17	17
5' UTR variant	14	16
synonymous variant	17	25
Silent	58	52
Missense	41	47
Nonsense	1	1

En cuanto a los SNVs interespecíficos, el mayor número de SNVs se identificó con la papa (1,973,023) y un poco menos con el tomate (1,809,264), confirmando que el tomate de árbol está filogenéticamente más cerca de este último (Olmstead et al., 2008). De ellos, 1,425,623 (72.3%) con papa y 1,309,278 (72.0%) con tomate eran SNPs; 417,416 (21.2%) y 413,062 (22.7%) eran MNP; 70,427 (3.6%) y 50,509 (2.8%) eran InDels, y 59,507 (3.0%) y 45,415 (2,5%) fueron MIXED (Tabla 7). La accesión A23 exhibió un mayor número de variantes interespecíficas que A21 (Tabla 7).

Tabla 7. Estadísticas de polimorfismo para los transcriptomas A21 y A23 del tomate de árbol. (Tabla de Pacheco et al., 2021)

Statistics	**SNPs**	**MNP**	**INDELs**	**MIXED**	**Total SNVs**
SNVs intraspecific variations					
A21	14,837	3,283	767	227	19,117
A23	38,183	8,213	2,391	726	49,530
SNVs interspecific variations					
A21 and *S. tuberosum*	619,626	174,982	28,2835	23,115	1,973,023
A23 and *S. tuberosum*	805,997	242,484	42,142	36,352	
A21 and *S. lycopersicum*	624,503	194,857	23,407	20,788	1,809,264
A23 and *S. lycopersicum*	684,775	218,205	27,102	24,627	

En contraste con los SNVs intraespecíficos, la proporción de variantes homocigóticas fue mayor (más del 95%) que los heterocigotos. Se observaron diferencias considerables en el número medio de polimorfismos entre los cromosomas, con diferencias de más de 2 veces entre el cromosoma 1 (259,267 en papa y 237,098 en tomate) y el cromosoma 12 (127,595 y 113,202) en ambas accesiones (Tabla 8).

Tabla 8. Distribución cromosómica de variantes de tomate de árbol con papa (S. tuberosum) y tomate (S. lycopersicum). (Tabla de Pacheco et al., 2021)

	Species	
Chromosome	***S. tuberosum***	***S. lycopersicum***
1	259,267	237,496
2	200,827	90,558
3	205,561	92,045
4	176,113	78,104
5	133,442	59,508
6	164,694	72,583
7	152,942	67,774
8	140,397	62,569
9	148,610	64,651
10	130,071	58,183
11	133,138	59,040
12	127,595	54,634

El impacto de 3,186,724 SNVs (58,7%) en papa y 2.095.805 (59,9%) en tomate se clasificó como “modifier”, 1,192,042 (21.9%) y 728,209 (22.3%) se clasificó como "low", 1,029,629 (19.0%) y 611,869 (17.5%) se clasificó como “moderate”, y el impacto de los 23,696 restantes (0.4%) y 12,268 (0.4%) Los SNPs fueron "hight" (Tabla 5). La mayoría de categorías de las variantes estaba en los exones (el 39% a el 43%), la región intergénica de la variante downstream del gen (el 25% y el 28%), la variante upstream del gen (el 15% y el 19%), la

variante de 3' UTR (el 5% y el 8%), la variante del intrón (el 2% y el 6%), la región intergénica (el 2% y el 3%), y 5' variante de UTR (el 2% y el 3%). La identificación de las variantes intraespecíficas e interespecíficas fomentará varias aplicaciones, incluyendo el mapeo genético, identificación de genotipos, selección asistida por marcadores, mejora, y la comprensión del control genético de los rasgos adaptativos en el tomate de árbol (He et al., 2014)

Conclusiones

En este trabajo, ensamblamos secuencias de transcriptomas de alta calidad de dos cultivares de tomate de árbol, un cultivo frutal estrechamente relacionado con el tomate y la papa, con gran potencial en regiones subtropicales. La exhaustiva anotación aportó información extensa y detallada que facilitará la disección de rasgos de interés agronómico, como el contenido en compuestos bioactivos o la respuesta a estrés entre otros. Además, este es el primer estudio en tomate de árbol donde se ha identificado un elevado número de polimorfismos, tanto intraespecíficos como con especies estrechamente relacionadas que podrían ser utilizadas en análisis de diversidad genética, mapeo cualitativo y cuantitativo de rasgos y programas de mejora en tomate de árbol. Esta información constituye un recurso valioso para los programas de mejoramiento de tomates, estudios de diversidad genética y ayudará en la mejora del tomate de árbol y su introducción exitosa en otras regiones y países.

Referencias

Acosta-Quezada, P. G., Martínez-Laborde, J. B., & Prohens, J. (2011). Variation among tree tomato (Solanum betaceum Cav.) accessions from different cultivar groups: Implications for conservation of genetic resources and breeding. *Genetic Resources and Crop Evolution*, *58*(6), 943–960. https://doi.org/10.1007/s10722-010-9634-9

Acosta-Quezada, P. G., Raigón, M. D., Riofrío-Cuenca, T., García-Martínez, M. D., Plazas, M., Burneo, J. I., Figueroa, J. G., Vilanova, S., & Prohens, J. (2015). Diversity for chemical composition in a collection of different varietal types of tree tomato (Solanum betaceum Cav.), an Andean exotic fruit. *Food Chemistry*, *169*, 327–335. https://doi.org/10.1016/j.foodchem.2014.07.152

Acosta-Quezada, P. G., Vilanova, S., Martínez-Laborde, J. B., & Prohens, J. (2012). Genetic diversity and relationships in accessions from different cultivar groups and origins in the tree tomato (Solanum betaceum Cav.). *Euphytica*, *187*(1), 87–97. https://doi.org/10.1007/s10681-012-0736-7

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Molin, A. D., Avanzato, C., Ferrarini, A., Delledonne, M., Sanseverino, W., Cigliano, R. A., Capella-Gutierrez, S., Gabaldón, T., Frusciante, L., Bradeen, J. M., & Carputo, D. (2015). The Solanum commersonii genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell*, *27*(4), 954–968. https://doi.org/10.1105/tpc.114.135954

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Britton, G. (2020). Carotenoid research: History and new perspectives for chemistry in

biological systems. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, *1865*(11), 158699. https://doi.org/10.1016/j.bbalip.2020.158699

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T. H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, *18*(3), 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Chen, X., Quek, S. Y., Fedrizzi, B., & Kilmartin, P. A. (2020). Characterization of free and glycosidically bound volatile compounds from tamarillo (Solanum betaceum Cav.) with considerations on hydrolysis strategies and incubation time. *Lwt*, *124*(November 2019), 109178. https://doi.org/10.1016/j.lwt.2020.109178

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Diep, T. T., Rush, E. C., & Yoo, M. J. Y. (2020). Tamarillo (Solanum betaceum Cav.): A Review of Physicochemical and Bioactive Properties and Potential Applications. *Food Reviews International*, *00*(00), 1–25. https://doi.org/10.1080/87559129.2020.1804931

Duarte, O., & Paull, R. (2015). *Exotic fruits and nuts of the New World. Cabi.*

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. 1–9. http://arxiv.org/abs/1207.3907

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. https://doi.org/10.1093/nar/gkn176

Grabherr, M. G. ., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N., & Friedman, and A. R. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883.Trinity

Gramazio, P., Blanca, J., Ziarsolo, P., Herraiz, F. J., Plazas, M., Prohens, J., & Vilanova, S. (2016). Transcriptome analysis and molecular marker discovery in Solanum incanum and S. aethiopicum, two close relatives of the common eggplant (Solanum melongena) with interest for breeding. *BMC Genomics*, *17*(1), 1–17. https://doi.org/10.1186/s12864-016-2631-4

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., De Jong, W. S., Douches, D. S., & Buell, C. R. (2011). Single nucleotide polymorphism discovery in elite north American potato germplasm. *BMC Genomics*, *12*(June). https://doi.org/10.1186/1471-2164-12-302

He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H. K., & Li, Z. (2014). Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, *5*(SEP), 1–8. https://doi.org/10.3389/fpls.2014.00484

Herraiz, F. J., Blanca, J., Ziarsolo, P., Gramazio, P., Plazas, M., Anderson, G. J., Prohens, J., & Vilanova, S. (2016). The first de novo transcriptome of pepino (Solanum muricatum):

Assembly, comprehensive analysis and comparison with the closely related species S. caripense, potato and tomato. *BMC Genomics*, *17*(1), 1–17. https://doi.org/10.1186/s12864-016-2656-8

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *BioRxiv*, *2012*, 1–21. https://doi.org/10.1101/767764

Huang, X., Chen, X. G., & Armbruster, P. A. (2016). Comparative performance of transcriptome assembly methods for non-model organisms. *BMC Genomics*, *17*(1), 1–14. https://doi.org/10.1186/s12864-016-2923-8

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Karasawa, M. M. G., & Mohan, C. (2018). Fruits as Prospective Reserves of bioactive Compounds: A Review. *Natural Products and Bioprospecting*, *8*(5), 335–346. https://doi.org/10.1007/s13659-018-0186-6

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lateef, A., Prabhudas, S. K., & Natarajan, P. (2018). RNA sequencing and de novo assembly of Solanum trilobatum leaf transcriptome to identify putative transcripts for major

metabolic pathways. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-33693-4

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

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

Moon, K. B., Ahn, D. J., Park, J. S., Jung, W. Y., Cho, H. S., Kim, H. R., Jeon, J. H., Park, Y. Il, & Kim, H. S. (2018). Transcriptome profiling and characterization of drought-tolerant potato plant (Solanum tuberosum L.). *Molecules and Cells*, *41*(11), 979–992. https://doi.org/10.14348/molcells.2018.0312

Nabavi, S. M., Šamec, D., Tomczyk, M., Milella, L., Russo, D., Habtemariam, S., Suntar, I., Rastrelli, L., Daglia, M., Xiao, J., Giampieri, F., Battino, M., Sobarzo-Sanchez, E., Nabavi, S. F., Yousefi, B., Jeandet, P., Xu, S., & Shirooie, S. (2020). Flavonoid biosynthetic pathways in plants: Versatile targets for metabolic engineering. *Biotechnology Advances*, *38*(June). https://doi.org/10.1016/j.biotechadv.2018.11.005

Olmstead, R. G., Bohs, L., Migid, H. A., Santiago-Valentin, E., Garcia, V. F., & Collier, S. M. (2008). A molecular phylogeny of the Solanaceae. *Taxon*, *57*(4), 1159–1181. https://doi.org/10.1002/tax.574010

Orqueda, M. E., Zampini, I. C., Torres, S., Alberto, M. R., Pino Ramos, L. L., Schmeda-Hirschmann, G., & Isla, M. I. (2017). Chemical and functional characterization of skin, pulp and seed powder from the Argentine native fruit mistol (Ziziphus mistol). Effects of phenolic fractions on key enzymes involved in metabolic syndrome and oxidative stress. *Journal of Functional Foods*, *37*, 531–540.

https://doi.org/10.1016/j.jff.2017.08.020

Pacheco, J., Vilanova, S., Grillo-Risco, R., Garcia-Garcia, F., Prohens, J., & Gramazio, P. (2021). De novo Transcriptome Assembly and Comprehensive Annotation of Two Tree Tomato Cultivars (Solanum betaceum Cav.) with Different Fruit Color. In *Horticulturae* (Vol. 7, Issue 11). https://doi.org/10.3390/horticulturae7110431

Petek, M., Zagorščak, M., Ramšak, Ž., Sanders, S., Tomaž, Š., Tseng, E., Zouine, M., Coll, A., & Gruden, K. (2020). Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. *Scientific Data*, *7*(1), 1–15. https://doi.org/10.1038/s41597-020-00581-4

Petersen, T. N., Brunak, S., Von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & Robin Buell, C. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience*, *9*(9), 1–11. https://doi.org/10.1093/gigascience/giaa100

Ramírez, F., & Kallarackal, J. (2019). Tree tomato (Solanum betaceum Cav.) reproductive physiology: A review. *Scientia Horticulturae*, *248*(October 2018), 206–215. https://doi.org/10.1016/j.scienta.2019.01.019

Ranil, R. H. G., Niran, H. M. L., Plazas, M., Fonseka, R. M., Fonseka, H. H., Vilanova, S., Andújar, I., Gramazio, P., Fita, A., & Prohens, J. (2015). Improving seed germination of the eggplant rootstock Solanum torvum by testing multiple factors using an orthogonal array design. *Scientia Horticulturae*, *193*, 174–181. https://doi.org/10.1016/j.scienta.2015.07.030

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for

evolutionary study of the nightshades (Solanaceae): A dated 1000-tip tree. *BMC Evolutionary Biology*, *13*(1). https://doi.org/10.1186/1471-2148-13-214

Scarano, D., Rao, R., & Corrado, G. (2017). In Silico identification and annotation of noncoding RNAs by RNA-seq and de Novo assembly of the transcriptome of Tomato Fruits. *PLoS ONE*, *12*(2), 1–16. https://doi.org/10.1371/journal.pone.0171504

Simão, 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*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Stam, R., Nosenko, T., Hörger, A. C., Stephan, W., Seidel, M., Kuhn, J. M. M., Haberer, G., & Tellier, A. (2019). The de novo reference genome and transcriptome assemblies of the wild tomato species solanum chilense highlights birth and death of NLR genes between tomato species. *G3: Genes, Genomes, Genetics*, *9*(12), 3933–3941. https://doi.org/10.1534/g3.119.400529

Vasco, C., Avila, J., Ruales, J., Svanberg, U., & Kamal-Eldin, A. (2009). Physical and chemical characteristics of golden-yellow and purple-red varieties of tamarillo fruit (Solanum betaceum Cav.). *International Journal of Food Sciences and Nutrition*, *60*(sup7), 278–288. https://doi.org/10.1080/09637480903099618

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., Zhang, J., Mainiero, S., Strickler, S. R., Catala, C., Martin, G. B., Mueller, L. A., Vrebalov, J., Giovannoni, J. J., Wu, S., & Fei, Z. (2020). Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19682-0

Ward, J. A., Ponnala, L., & Weber, C. A. (2012). Strategies for transcriptome analysis in nonmodel plants. *American Journal of Botany*, *99*(2), 267–276.

https://doi.org/10.3732/ajb.1100334

Waterhouse, R. M., Seppey, M., Simao, 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. https://doi.org/10.1093/molbev/msx319

Wheeler, T. J., & Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489. https://doi.org/10.1093/bioinformatics/btt403

Wu, L., Du, G., Bao, R., Li, Z., Gong, Y., & Liu, F. (2019). De novo assembly and discovery of genes involved in the response of Solanum sisymbriifolium to Verticillium dahlia. *Physiology and Molecular Biology of Plants*, *25*(4), 1009–1027. https://doi.org/10.1007/s12298-019-00666-4

Xia, Z., Xu, H., Zhai, J., Li, D., Luo, H., He, C., & Huang, X. (2011). RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. *Plant Molecular Biology*, *77*(3), 299–308. https://doi.org/10.1007/s11103-011-9811-z

Yang, X., Liu, F., Zhang, Y., Wang, L., & Cheng, Y. fu. (2017). Cold-responsive miRNAs and their target genes in the wild eggplant species Solanum aculeatissimum. *BMC Genomics*, *18*(1), 1–13. https://doi.org/10.1186/s12864-017-4341-y

Zuluaga, A. P., Solé, M., Lu, H., Góngora-Castillo, E., Vaillancourt, B., Coll, N., Buell, C. R., & Valls, M. (2015). Transcriptome responses to Ralstonia solanacearum infection in the roots of the wild potato Solanum commersonii. *BMC Genomics*, *16*(1), 1–16. https://doi.org/10.1186/s12864-015-1460-1