MASTER EN PRODUCCIÓN, PROTECCIÓN Y MEJORA VEGETAL
UNIVERSIDAD DE CÓRDOBA

Identificación y caracterización de la superfamilia génica *ALDH* en garbanzo (*Cicer arietinum*) mediante herramientas bioinformáticas de acceso libre

Rocío Carmona Molero

Departamento de Genética

Córdoba, Octubre 2018

VºB º del Director/es:

Fdo: Dra Teresa Millán Valenzuela

Fdo: Dr Jose V. Die Ramon

Firma del alumno:

Fdo: Rocío Carmona Molero

ÍNDICE

1. Resumen / Abstract
2. Introducción
3. Objetivos
4. Materiales y Métodos
    1. Identificación aldehído deshidrogenasas
        1. 1. Búsqueda proteínas candidatas
        1. 2. Comprobación dominios conservados ALDH
        1. 3. BLASTP contra el genoma de garbanzo
    2. Caracterización aldehído deshidrogenasas
    3. Clasificación aldehído deshidrogenasas
    4. Filogenia
    5. Análisis duplicación
    6. Expresión *in silico*
    7. Disponibilidad del código
5. Resultados y Discusión
    1. Identificación y caracterización aldehído deshidrogenasas
    2. Clasificación aldehído deshidrogenasas
    3. Filogenia
    4. Análisis duplicación
    5. Expresión *in silico*
6. Conclusiones
7. Referencias

RESUMEN

Las aldehído deshidrogenasas (ALDHs) son una superfamilia de proteínas con una función importante en la detoxificación de los aldehídos producidos en respuesta a estreses bióticos/abióticos. La disponibilidad del genoma de referencia de garbanzo (*Cicer arietinum*) da la oportunidad de identificar y caracterizar a los miembros de esta familia en una leguminosa de importancia agronómica. En este estudio, se han identificado 37 ALDHs en el genoma de garbanzo y se ha realizado una caracterización completa de las mismas. Los análisis filogenéticos comparativos con *Medicago* sugieren una gran conservación de la familia entre las dos especies y los análisis de duplicaciones indican una leve incidencia de eventos de duplicación con posterioridad a la especiación. El análisis de expresión *in silico* apoya el papel de la mayoría de miembros de la familia *CaALDH* en la tolerancia al estrés abiótico, con una mayor representación de las secuencias de la familia 18 en librerías EST de tolerancia a sequía. Todos los *scripts* escritos en este trabajo y la secuencia de ejecución para el análisis de las bases de datos están disponibles públicamente en un repositorio online de acceso libre. En resumen, este trabajo proporciona una visión general de la superfamilia génica *ALDH* en garbanzo. Es la primera vez que la familia se estudia en este cultivo. Nuestros resultados respaldan que las proteínas ALDHs están implicadas en una amplia gama de rutas metabólicas y que participan en la respuesta al estrés. Esto proporciona nuevos conocimientos sobre la la presencia y función de la familia en esta especie, lo que puede ser útil para desarrollar estrategias de mejora genética de respuesta a diferentes estreses. Este trabajo también proporciona una base para análisis genómicos comparativos posteriores en el estudio de la evolución de los genes *ALDH* dentro de la familia de las leguminosas.

Palabras clave: ALDH, bioinformática, estrés abiótico, garbanzo, *open science*, R.

ABSTRACT

Aldehyde dehydrogenases (ALDHs) constitute a protein superfamily with an important function in the detoxification of the aldehydes produced in response to biotic/abiotic stresses. The availability of the chickpea (*Cicer arietinum*) reference genome provides an opportunity to identify and characterize the members of this family in a legume of agronomic importance. In this study, 37 ALDHs have been identified in the chickpea genome and a complete characterization of them has been carried out. The comparative phylogenetic analysis with *Medicago* suggests a high conservation of the family between both species and the duplication analysis indicates a slight duplication incidence after the speciation. *In silico* expression analysis supports the abiotic stress tolerance role of most CaALDH family members, showing the family 18 sequences overrepresented in drought tolerance EST libraries. All the code and scripts written for this work are publicly available in an online open access repository. In summary, this work provides the first general overview of the ALDH gene superfamily in this crop. Our results support that ALDH proteins are involved in a wide range of metabolic pathways and they participate in the stress response. This provides new knowledge of the family presence and function in this species, what may be useful to develop chickpea breeding strategies to improve development or stress responses. This work also supplies a basis for further comparative genomic analysis and a framework to study the ALDH genes evolution within the legume family.

Keywords: abiotic stress, ALDH, bioinformatics, chickpea, open science, R.

INTRODUCCIÓN

Las aldehído deshidrogenasas (ALDHs) conforman una superfamilia de enzimas dependientes de $NAD(P)^{+}$ conservadas durante la evolución. Las ALDHs metabolizan una gran variedad de aldehídos alifáticos y aromáticos, tanto endógenos como exógenos. Estos aldehídos son tóxicos debido a su reactividad química o pueden llegar a serlo si sus niveles de estabilidad no son estrictamente regulados (Ayala et al., 2014).

Los aldehídos son intermediarios en una serie de rutas metabólicas fundamentales generadas durante el metabolismo de los aminoácidos, carbohidratos, lípidos, aminas biogénicas, vitaminas y esteroides (Vasiliou et al., 2000). Se pueden producir también en respuesta a ciertos estreses ambientales que alteran el metabolismo, como la deshidratación o salinidad (Barclay & McKersie, 1994; Seki et al., 2007; Stiti et al., 2011), o a estreses bióticos (Ashraf et al., 2009). En respuesta a rayos UV, desecación y salinidad, se producen especies reactivas de oxígeno (ROS) que producen daño celular incluyendo la peroxidación de los lípidos de membrana (Chen et al., 2002). Entre los distintos aldehídos que se pueden formar como productos secundarios de la peroxidación lipídica, encontramos productos mutagénicos y tóxicos, como el malondialdehído (MDA) y 4-hidroxinonenal (4-HNE) (Ayala et al., 2014). Los procesos que limitan el daño celular causado por los aldehídos derivados de la peroxidación de lípidos, como las ALDHs mediante oxidación de sus correspondientes ácidos carboxílicos, representan un mecanismo defensivo para sobrevivir al estrés osmótico y oxidativo. Además, las ALDHs también producen moléculas osmoprotectoras como la glicina betaína (Xing & Rajashekar, 2001) y moléculas que contribuyen a la homeostasis redox como NADPH y NADH (Hou & Bartels, 2014).

Actualmente, se han descrito 24 familias distintas de ALDHs siendo algunas de ellas exclusivas de plantas (ALDH11, ALDH12, ALDH19, ALDH21, ALDH22, ALDH23 y ALDH24; Guo et al., 2017). El número total de genes *ALDH* varía bastante entre especies vegetales y parece aumentar a medida que las plantas son más complejas. Este aumento suele ser el resultado de la expansión de una o más familias (Brocker et al., 2013). La secuenciación del genoma completo de especies vegetales representa una oportunidad para la identificación y caracterización de los genes *ALDH*. Tras haber sido caracterizados inicialmente en *Arabidopsis thaliana* (Kirch et al., 2004), los estudios genómicos recientes incluyen el análisis de la familia en *Vitis vinifera* (Zhang et al., 2012), *Zea mays* (Jimenez-Lopez et al., 2010), *Oryza sativa* (Kotchoni et al., 2010), *Solanum lycopersicum* (Jiménez-López et al., 2016) y varias especies de algodón (*G. arboreum, G. hirsutum* y *G. barbadense*; Guo et al., 2017). En leguminosas se ha caracterizado en *Glycine max* (Kotchoni et al., 2012) y *Lupinus angustifolius* (Jiménez-López et al., 2016).

El garbanzo (*Cicer arietinum* L.) es un cultivo autógamo diploide (2x =2n =16). Es la segunda legumbre de grano más importante a nivel mundial. Tanto su rendimiento como su área cultivada han aumentado en los últimos 10 años (FAOSTAT, 2017). Sin embargo, los estreses abióticos como la sequía, el calor y la alta salinidad son limitantes para su producción (Millán et al., 2015). En España hay 8 bancos de semillas de *Cicer arietinum* (CRF, INIA 2018) localizadas en Madrid (Código FAO: ESP004 y ESP198) con 841 y 3 entradas, respectivamente; Valencia (ESP026) con 7 entradas; Zaragoza (ESP027) con 22 entradas; Córdoba (ESP046) con 687 entradas; Valladolid (ESP109) con 46 entradas; Tenerife (ESP172) con 14 entradas; y Palma de Mallorca (ESP200) con 2 entradas. Estos

recursos fitogenéticos son de extrema importancia al proporcionar el material para la mejora del cultivo. El genoma de garbanzo ha sido secuenciado y publicado en los últimos años (Jain et al., 2013; Varshney et al., 2013). Esto posibilita la identificación de las familias génicas cuyos productos proteicos puedan ser de interés para la mejora molecular y el desarrollo de variedades mejoradas. La aparición del genoma de referencia, el desarrollo de recursos genómicos y la disponibilidad de herramientas bioinformáticas cada vez más complejas pueden contribuir a este avance. Así, recientemente se han identificado y caracterizado varias familias génicas en garbanzo (Chidambaranathan et al., 2017; Die et al., 2018).

En la actualidad, la reproducibilidad se considera un pilar fundamental del método científico. Un estudio científico es reproducible cuando el texto se acompaña del código empleado para generar los datos y, además, éstos se hacen disponibles (Peng, 2011; Marwick, 2017). La existencia de un código ordenado y estructurado ofrece varias ventajas. En primer lugar, permite su reutilización en proyectos posteriores, ahorrando tiempo y esfuerzos al equipo de investigación (Garijo et al., 2013). Además, compartir públicamente el código con el que generamos unos resultados, y en general adoptar prácticas abiertas, puede ayudarnos a identificar errores y abrir nuevas líneas de colaboración (Hampton et al., 2015; McKiernan et al., 2016).

Este trabajo se ha basado en la búsqueda sistemática de secuencias en el banco de datos (GeneBank, NCBI). Para ello hemos aplicado una serie de *scripts* escritos en el lenguaje de programación R con objeto de reducir los pasos manuales tanto como fuera posible y poder así, extraer, filtrar y catalogar los resultados de esa búsqueda de una forma eficiente. También hemos trabajado con un sistema de control de versiones de los *scripts* (*git*) que permite reconstruir la historia del proyecto. Todas las versiones están alojadas y disponibles públicamente en un repositorio de internet (gitHub) (https://github.com/RocioCarmonaMolero/ScriptProteinas).

OBJETIVOS

1. Identificar, caracterizar y clasificar la superfamilia de las aldehído deshidrogenasas en garbanzo.

2. Realizar un análisis de filogenia y de duplicaciones génicas para estudiar los eventos ocurridos antes y después de la especiación, estableciendo su relación con la leguminosa modelo *Medicago truncatula*.

3. Realizar un análisis de expresión *in silico* a partir de librerías EST disponibles en la base de datos del NCBI que ayude a comprender las funciones que estas proteínas pueden estar desarrollando.

4. Poner a libre disposición toda la información tanto del proceso del trabajo como los resultados finales para facilitar su uso en futuros proyectos.

MATERIAL Y MÉTODOS

1. Identificación aldehído deshidrogenasas

1.1. Búsqueda proteínas candidatas

En primer lugar, se realizó una búsqueda basada en palabras clave en la base de datos 'Protein' del National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) especificando *Arabidopsis* como organismo y 'aldehyde dehydrogenase' como título. Con cada una de ellas se ejecutó un BLASTP (Altschul et al., 1990) contra el genoma de *C. arietinum* en la base de datos Reference Protein (refseq_protein). Los umbrales establecidos para seleccionar las ALDHs candidatas de *C. arietinum* fueron: *Query cover* $\geq$ 25%, *E-value* $\leq e^{-25}$, *Identity* $\geq$ 25%. Las secuencias de garbanzo que cumplieron estos parámetros se descargaron en un archivo csv. Para evitar la duplicación de una misma secuencia en archivos csv distintos, se escribió un script en código R. Para asegurar la identificación de la familia génica completa, se repitió el mismo proceso dos veces usando como organismo *Medicago truncatula* y *Glycine max*, respectivamente.

1.2. Comprobación dominios conservados ALDH

Para verificar que las proteínas candidatas son aldehído deshidrogenasas, se comprobó la presencia de los dominios conservados *ALDH-superfamily*: 'PF00171.21', 'PF07368.10' y 'PF05893.13' en Pfam; 'PS00687' y 'PS00070' en ScanProsite; la accesión 'cl11961' en la base de datos de *Conserved Domains* del NCBI; y la accesión '53720' en la base de datos *Superfamily*.

1.3. BLASTP contra el genoma de garbanzo

Se realizó una búsqueda BLASTP contra el propio genoma de garbanzo para incluir cualquier proteína que no tuviera homología con las otras especies y detectar las no predichas.

2. Caracterización aldehído deshidrogenasas

Para la caracterización, nos centramos en las siguientes propiedades: número de aminoácidos, accesión y longitud de su mRNA, identificador del locus, cromosoma donde se encuentra, número de exones, coordenadas de inicio y fin de la proteína, y peso molecular de la proteína. Para extraer esta información de forma eficiente se escribieron dos funciones en lenguaje R con acceso a las bases de datos del NCBI. El punto isoeléctrico teórico se predijo mediante la herramienta *Compute pI/Mw* de la base de datos proteómica ExPASy (http://web.expasy.org/compute_pi/), así como los pesos moleculares de aquellas proteínas sin anotación en el NCBI.

También se usaron herramientas bioinformáticas de acceso libre para la predicción de la localización subcelular y el centro activo. Esto es importante para la clasificación de las proteínas. Para la localización celular se recurrió a DeepLoc 1.0 (Almagro-Armenteros et al., 2017), que diferencia entre 10 localizaciones de proteínas eucarióticas: núcleo, citoplasma, extracelular, mitocondria, membrana celular, cloroplasto, aparato de Golgi, retículo endoplasmático, lisosoma/vacuola y peroxisoma; SLP-Local (Matsuda et al., 2005) y TargetP 1.1 (Emanuelsson et al., 2000). DeepLoc a su vez indica si se trata de una proteína soluble o de membrana; en aquella cuyo resultado fuera proteína de membrana se comprobó el dominio transmembrana mediante SMART (*Simple Modular Architecture Research Tool*, http://smart.embl.de). Para consolidar los resultados que indican localización en el cloroplasto o mitocondria se confirmó la presencia de *chloroplast transit peptides* (cTP) mediante ChloroP 1.1 (Emanuelsson et al., 1999) y de *mitochondrial targeting sequences* (mTP) mediante Mitoprot (Claros & Vincens, 1996). Para el centro activo se analizaron las secuencias proteicas con PROSITE (https://prosite.expasy.org/), buscando la identificación de los centros activos PS00687 (*glutamic acid active site*, E) y PS00070 (*cysteine active site*, C). Para determinar la función molecular se utilizó PROPSEARCH (Hobohm & Sander, 1995, http://abcis.cbs.cnrs.fr/propsearch/). Cada proteína se analizó individualmente y se anotaron los resultados con mayor probabilidad. Como verificación se comprobó la conservación de los residuos funcionales; primero haciendo una búsqueda en BLAST y después mediante un alineamiento entre esos homólogos con el programa Clustal Omega (Sievers et al., 2011, http://www.clustal.org/omega/). La Figura 1 muestra un resumen de las herramientas empleadas durante la caracterización.

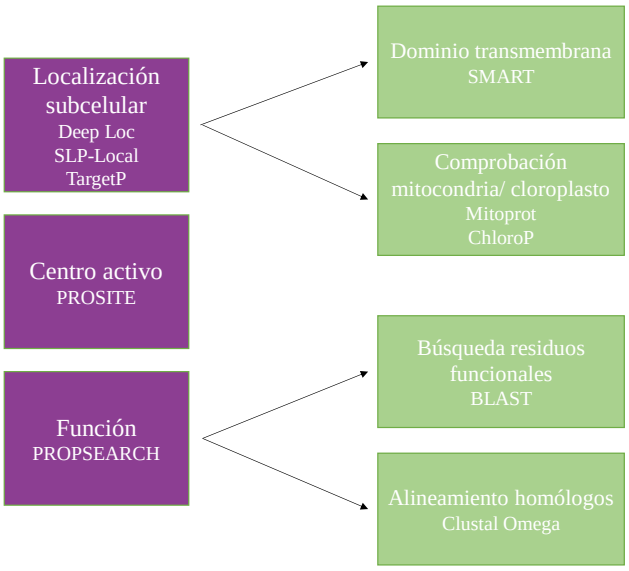

Figura 1. Esquema de las herramientas bioinformáticas utilizadas en la búsqueda.

3. Clasificación aldehído deshidrogenasas

Para la clasificación de las ALDHs se aplicó el criterio establecido por el *ALDH Gene Nomenclature Comittee* (AGNC; http://www.genenames.org/guidelines.html; Vasiliou et al., 1999): dos proteínas pertenecen a la misma familia génica si tienen > 40% de identidad; y a la misma subfamilia si tienen > 60% de identidad. Se anotan de forma que la raíz 'ALDH' vaya seguida de un número descriptor de la familia, una letra mayúscula para describir la subfamilia, un número concretando el gen individual dentro de la subfamilia y una letra minúscula en caso de ser necesario designar variantes.

Los métodos frecuentes para la clasificación de esta familia en plantas están basados en la homología con otras especies vegetales ya descritas. Por esto, se realizó un BLASTP de nuestras 37 CaALDHs contra la base de datos refseq de las leguminosas *Medicago truncatula* y *Glycine max*. Los resultados se descargaron en un fichero y se filtraron mediante una función escrita en lenguaje R para eliminar aquellos con identidad < 40% y cuya longitud de la *query* fuera menor que la longitud del *hit*.

4. Filogenia

Se realizó un alineamiento múltiple de las secuencias de las aldehído deshidrogenasas de *Medicago truncatula* (MtALDH) y *Cicer arietinum* (CaALDH) en formato de un gen por

locus con el programa MUSCLE usando los parámetros por defecto (Edgar, 2004). Para deducir la historia evolutiva se utilizó el método Neighbor-Joining (Saitou & Nei, 1987). El árbol *bootstrap* consenso fue inferido a partir de 1000 réplicas (Felsenstein, 1985). Las distancias evolutivas se calcularon utilizando el método de corrección de Poisson (Zuckerkandl & Pauling, 1965), estando en las unidades del número de sustituciones de aminoácidos por posición. Los análisis evolutivos se realizaron en MEGA6 (Tamura et al., 2013). Consideramos *sister pairs* aquellas proteínas agrupadas en base a valores de *bootstrap* > 65% (Die et al., 2018).

5. Análisis de Duplicación

Para el análisis de duplicación de las CaALDHs se utilizó el programa Circoletto (Darzentas, 2010) siendo tanto el FASTA *query* como el FASTA *database* las secuencias proteicas de CaALDHs; con valores de *E-value* ultra estricto (10^{-180}), usando *score* absoluto/bandas coloreadas y los colores: verde para identidades 90% ≤ 95%, naranja 95% ≤ 99% y rojo > 99%.

6. Expresión *in silico*

Las secuencias codificantes de los genes *CaALDH* se usaron como *query* contra la base de datos de ESTs del NCBI de garbanzo. Los parámetros de búsqueda se establecieron de la siguiente manera: megablast, identidad > 90%, longitud > 180 pb y *E-value* < 10^{-10} (Die et al., 2018).

7. Disponibilidad del código

Los códigos para este TFM se han escrito en el lenguaje de programación R, con el software R (Team R.C., 2017) y la interfaz RStudio (Team R, 2016, http://www.rstudio.com/) de acceso libre. Con estos scripts hemos recopilado datos del NCBI y realizado el análisis de los mismos. Están disponibles en el repositorio: https://github.com/RocioCarmonaMolero/ScriptProteinas. El código se distribuye bajo la licencia MIT de código abierto.

RESULTADOS Y DISCUSIÓN

1. Identificación y caracterización aldehído deshidrogenasas

El proceso de identificación de las CaALDHs se presenta en la Figura 2. Considerando que la búsqueda se ha realizado con la planta modelo *A. thaliana* y las leguminosas *G. max* y *M. truncatula*, podemos concluir que hemos cubierto un espectro amplio de especies que, por homología de secuencias, nos permitirán identificar todas las aldehído deshidrogenasas de *C. arietinum*.

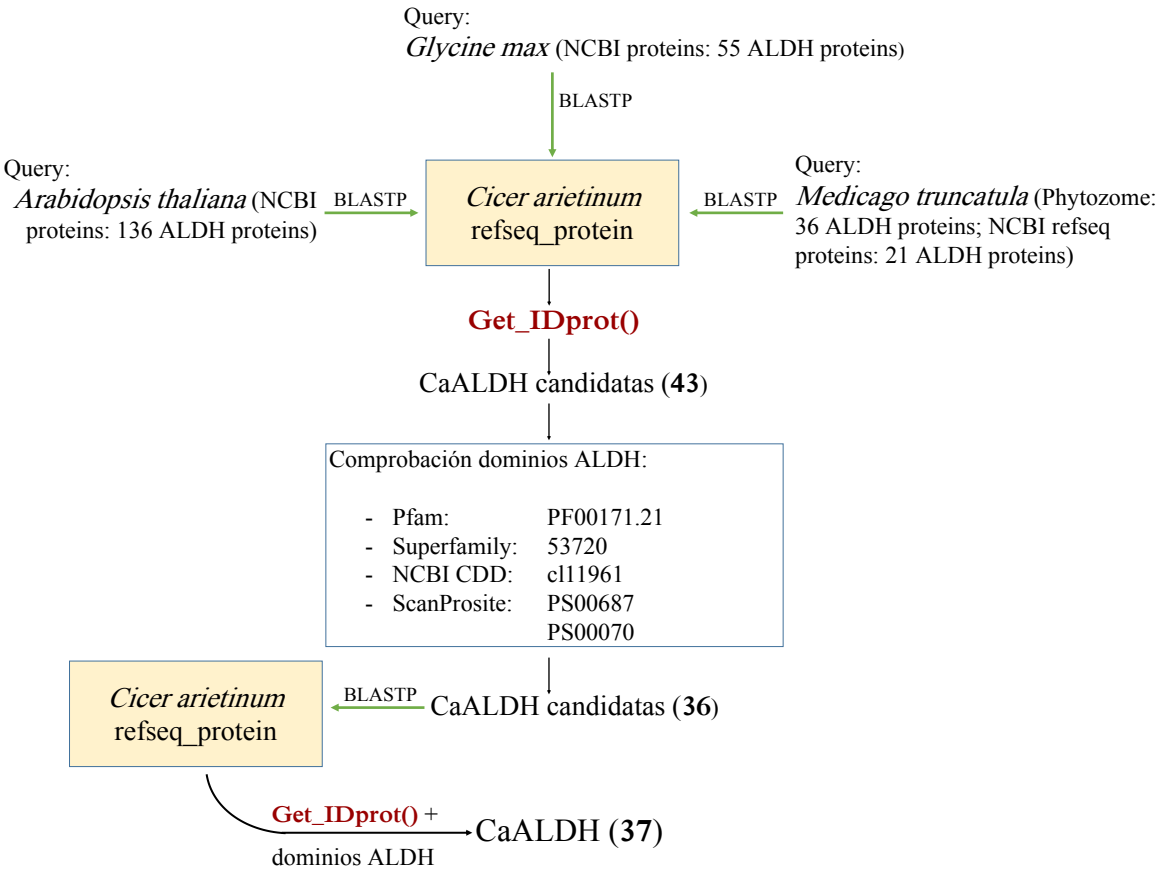

Figura 2. Esquema identificación de CaALDHs. Las funciones propias aparecen en color. Se usaron proteínas de *Arabidopsis*, *G. max* y *Medicago* como *query* en búsquedas BLASTP contra el genoma de referencia del garbanzo. Para evitar repeticiones se limpió el resultado con *scripts* escritos en R. Las proteínas de garbanzo que no tuvieran los dominios ALDH se eliminaron y con las secuencias resultantes se ejecutó otro BLASTP contra el genoma de *Cicer* para detectar las secuencias no predichas y asegurar la identificación de todas las CaALDHs. Esto resultó en 37 aldehído deshidrogenasas codificadas por el genoma de garbanzo.

La compleción y disponibilidad del genoma de garbanzo junto a las búsquedas en bases de datos ha permitido identificar 37 proteínas ALDH en garbanzo codificadas por 29 genes *CaALDH* (Tabla 1). Estos genes codifican proteínas con un rango de 134 aa (CaALDH3H1j) a 755 aa (CaALDH18B3d). El número de exones de los genes *CaALDH* varía de 3 (*CaALDH3H1j)* a 21 (*CaALDH18B3a, CaALDH18B3d* y *CaALDH18B3e*); y los correspondientes pesos moleculares de 15,07 a 81,90 kDa. Los puntos isoeléctricos predichos oscilan entre 4,34 y 9,49 (Tabla 2). El amplio rango de pI sugiere que las proteínas ALDH de garbanzo pueden funcionar en ambientes subcelulares muy diferentes.

Encontramos las familias 5, 6, 11, 12 y 22 definidas por un único gen, similar a *Arabidopsis, O. sativa, S. italica*, *S. bicolor*, *E. parvulum* y *E. salsugineum* (Tabla 3); sugiriendo que estas familias constituyen genes *ALDH* '*house-keeping*', implicados en el metabolismo central de las plantas y la preservación de los niveles de los aldehídos no tóxicos. No hay genes *CaALDH* en las familias 19, 21, 23 y 24. Las familias 21 y 23 contienen solo genes de plantas terrestres primitivas (Chen et al., 2002), mientras que ALDH24 parece ser exclusivo del alga unicelular *C. reinhardtii* (Wood & Duff, 2009). Esto sugiere que estas tres familias podrían haber desempeñado un papel importante en la evolución de plantas inferiores y posteriormente se perdieron en las plantas superiores. La familia 19 solo se ha encontrado en tomate (*S. lycopersicum*), por lo que podría haber evolucionado específicamente en este linaje (Jimenez-Lopez et al., 2016).

Excepto eso, todas las familias ALDH identificadas en plantas superiores están presentes en *C. arietinum* (Tabla 3). El garbanzo, junto al tomate, es la tercera especie con mayor cantidad de genes *ALDH*. Por encima se encuentra el manzano con 39 genes y el algodón con 30 genes. *C. arietinum* parece tener proteínas ALDH adicionales de respuesta al estrés: las familias 3 y 18 son particularmente abundantes, lo que puede ser significativo para llevar a cabo la detoxificación de moléculas de aldehído generadas bajo diferentes tensiones y mantener la homeostasis de equivalentes reductores. Específicamente, la familia 3 (10 genes) y la 18 (6 genes) tienen en garbanzo mayor número de miembros que en el resto de especies vegetales descritas hasta el momento.

Los genes *ALDH18* codifican *Δ1 -pyrroline-5-carboxylate synthetase* (P5CS), definidas como proteínas *ALDH-like* (Sophos & Vasiliou, 2003). Están implicadas en la biosíntesis de prolina (Igarashi et al., 1997), cuya acumulación tiene roles adaptativos en la tolerancia a estreses bióticos/abióticos (Verbruggen & Hermans, 2008). Li et al. (2013) concluyen

que la familia 18 es el grupo que más difiere entre especies. La estructura génica de los miembros de esta familia es distinta a la del resto de familias y presenta dominios adicionales AA-quinasa; careciendo de los sitios activos ALDH conservados (Tabla 1). Además, se cree que el equilibrio entre la biosíntesis y la degradación de prolina es esencial en la determinación de las funciones osmoprotectoras de la misma. En la degradación interviene la *Δ1 -pyrroline-5-carboxylate dehydrogenase* (P5CDH), que es una proteína ALDH12. Esta degradación ocurre en la mitocondria (Verbruggen & Hermans, 2008), lo que coincide con los resultados obtenidos en la predicción de localización subcelular (Tabla 1).

La mayoría de genes de la familia 3 parecen estar regulados por la vía del ácido abscísico (ABA) y su expresión se ha descrito en respuesta a estreses ambientales (Stiti et al., 2011). Las proteínas ALDH3 constituyen uno de los grupos más expandidos y diversos de genes *ALDH* en especies vegetales, lo que ha dado lugar a seis subfamilias: 3E, 3F, 3H, 3I, 3J y 3K (Brocker et al., 2013). Dos de estas subfamilias se encuentran en garbanzo (3F y 3H). La subfamilia 3H1 está muy expandida con 10 miembros codificados por 7 genes, mientras que la 3F1 tiene 3 miembros codificados por 3 genes. Missihoun et al. (2012) han postulado que la abundancia de las proteínas ALDH3 resulta de un patrón de expresión complejo de sus genes regulados por *gene-splicing* o promotores alternativos. Se ha sugerido que las isoformas de ALDH3 han evolucionado como consecuencia de la especialización funcional en tejidos específicos y orgánulos subcelulares (Kirch et al., 2004). La variedad de localizaciones subcelulares predichas para las proteínas de garbanzo pertenecientes a esta familia parecen apoyar la hipótesis (Tabla 1). De las especies vegetales descritas, únicamente las algas *C. reinhardtii* (unicelular) y *V. carteri* (colonial) carecen de la familia 3 (Tabla 3); sugiriendo que esta familia génica surge con la aparición de las plantas terrestres.

Los análisis de ScanProsite muestran que los dominios característicos PS00687 y PF00070 no se encuentran en todas las secuencias ALDH: 12 de las 37 contienen ambos dominios; 5 de ellas contienen solo el dominio PS00687 y 2 de ellas solo el PS00070. Algunas de las proteínas que no contienen estos dominios se encuentran en las subfamilias 3F1, 3H1, y la familia 18. Por esto, fue necesario realizar búsquedas alternativas (empleando la diversidad de bases de datos señaladas en la sección Material y Métodos) para identificar estas ALDHs. En la familia 18 aparecen los dominios PS00902 (*glutamate 5 kinase signature*) y PS01223 (*γ-glutamyl phosphate reductase signature*).

Para entender la distribución cromosómica, los genes *CaALDH* se mapearon al genoma de garbanzo. Basándonos en el ensamblaje del genoma disponible de *C. arietinum*, 23 de los 29 genes se distribuyen en siete de los ocho cromosomas. No pudimos mapear los genes cuyas proteínas tiene corta longitud de secuencia (< 272 aa). Los seis genes *CaALDH* que no se pueden mapear en este genoma (Tabla 2) podrían ser genes mitocondriales; ya que la mitocondria no se encuentra en la distribución del genoma de referencia en el NCBI. Localizamos genes que codifican ALDHs de distintas familias en un mismo cromosoma. Los cromosomas 6 y 7 contienen el 52% de los genes mapeados. En el cromosoma 2 no se encuentra mapeado ningún gen *CaALDH* (Figura 3).

2. **Clasificación aldehído deshidrogenasas**

La clasificación final se encuentra en la Tabla 1. El sistema de clasificación establecido por el AGNC sigue las directrices del *Human Gene Nomenclature* que recomienda su uso para las demás especies. Toda esta información se puede encontrar en la *Aldehyde Dehydrogenase Superfamily Database* (http://www.aldh.org). Al estar toda la información referenciada y regida por el reino animal, la clasificación de las ALDHs en especies vegetales queda subordinada a lo previamente descrito. Según los métodos seguidos en los estudios citados en este trabajo, una proteína pertenecerá a la familia y subfamilia, incluso se le asignará el mismo número génico, de la ALDH ya identificada con la que tenga más homología para facilitar futuros estudios comparativos. Como ejemplo tenemos el caso de la proteína CaALDH6B2 que prueba tanto la subordinación mencionada, ya que en plantas no existe la subfamilia ALDH6A, como la asignación no secuencial del número del gen, ya que en garbanzo solo tenemos un miembro génico pero se le atribuye el número 2. Así, la definición del AGNC parece no coincidir con las prácticas habituales.

Para conseguir una clasificación segura de familias más complejas, como la 18, fue necesario realizar alineamientos entre ellas y mirar el conjunto de propiedades que corresponden con su función y localización subcelular (Tabla 1). En conclusión, de las 24 familias ALDH existentes, las CaALDH se engloban en 10 de las 14 familias ALDH previamente descritas en especies vegetales (2, 3, 5, 6, 7, 10, 11, 12, 18 y 22; Tabla 3).

Tabla 1. Clasificación de las proteínas ALDH de garbanzo.

Familia	Miembro	LOCUS	Proteína ID	ALDH *GLU_Active Site* (PS00687) - *CYS_Active Site* (PS00070)	Localización		Función molecular
2	ALDH2B4a	LOC101490532	XP_004508853	E303 - C337	MITOCONDRIA	SOLUBLE	*ALDH (NAD+)*
	ALDH2B4b		XP_004508854	E190 - C224	CITOSOL	SOLUBLE	*ALDH (NAD+)*
	ALDH2B7a	LOC101492709	XP_004509835	E305 - C339	MITOCONDRIA	SOLUBLE	*ALDH (NAD+)*
	ALDH2B7b		XP_004509834	E306 - C340	MITOCONDRIA	SOLUBLE	*ALDH (NAD+)*
	ALDH2C4c	LOC101493969	XP_004503432	E264 - C298	CITOSOL	SOLUBLE	*ALDH (NAD+)*
	ALDH2C4b	LOC101513875	XP_012574767	E268 - C302	CITOSOL	SOLUBLE	*ALDH (NAD+)*
	ALDH2C4a	LOC101514219	XP_004512967	E270 - C304	CITOSOL	SOLUBLE	*ALDH (NAD+)*
3	ALDH3F1b	LOC101491914	XP_004507095	-	PEROXISOMA, CITOSOL	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3F1a	LOC101497113	XP_004486968	-	PEROXISOMA, CITOSOL	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3F1c	LOC101511819	XP_012573731	-	PEROXISOMA, CITOSOL	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3H1f	LOC101488602	XP_012575132	-	CITOSOL	SOLUBLE	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity*
	ALDH3H1g		XP_004515934	-	CITOSOL	SOLUBLE	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity*
	ALDH3H1h	LOC101497514	XP_012567420	-	CITOSOL	SOLUBLE	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity*
	ALDH3H1i		XP_012567421	-	CITOSOL	SOLUBLE	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity*

	ALDH3H1j	LOC101502106	XP_004514084	-	MITOCONDRIA, CLOROPLASTO	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity*
	ALDH3H1a	LOC101505038	XP_004503899	-	CLOROPLASTO	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3H1b		XP_004503900	-	MITOCONDRIA, CLOROPLASTO	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3H1c	LOC101510937	XP_004502482	E222	PEROXISOMA, CLOROPLASTO	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3H1d	LOC101511680	XP_004502485	E222	PEROXISOMA	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
	ALDH3H1e	LOC101515558	XP_004498346	E222	PEROXISOMA	MEMBRANA	*ALDH [NAD+/NAD(P)+]; Variable substrate ALDH stress-regulated detoxification pathway activity; Fatty aldehyde dehydrogenase (EC 1.2.1.3)*
5	ALDH5F1	LOC101506901	XP_004488550	E299 - C333	MITOCONDRIA	SOLUBLE	*Succinate semialdehyde dehydrogenase (EC 1.2.1.24) (NAD(+)/NAD(P)+)*
6	ALDH6B2	LOC101490310	XP_004504867	C321	MITOCONDRIA	SOLUBLE	*Methylmalonate-semialdehyde dehydrogenase [acylating], mitochondrial*
7	ALDH7A1a	LOC101513733	XP_004488077	E266	CITOSOL	SOLUBLE	*ALDH (NAD+); Turgor-responsive protein 26G (EC 1.2.1.-)*
	ALDH7A1b		XP_012574245	E266	CITOSOL	SOLUBLE	*ALDH (NAD+); Turgor-responsive protein 26G (EC 1.2.1.-)*
10	ALDH10A8b	LOC101506136	XP_004508822	E260 - C294	CITOSOL	SOLUBLE	*Betaine-aldehyde dehydrogenase (EC 1.2.1.8) (BADH).*
	ALDH10A8a	LOC101507930	XP_004501961	E260 - C294	CITOSOL	SOLUBLE	*Betaine-aldehyde dehydrogenase (EC 1.2.1.8) (BADH).*
11	ALDH11A3	LOC101510843	XP_004507722	E264 - C298	CITOSOL	SOLUBLE	*NADP-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.9)*
12	ALDH12A1	LOC101490107	XP_004508769	PS00867 C334	MITOCONDRIA	SOLUBLE	*ALDH (NAD+); Delta 1 -pyrroline-5-carboxylate dehydrogenase (P5CDH)*

18	ALDH18B3a	LOC101490622	XP_012572021	PS00902, PS01223	CITOSOL	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS); Glutamate 5-kinase activity; Glutamate-5-semialdehyde dehydrogenase*
	ALDH18B3f	LOC101495530	XP_012575404	-	CITOSOL, MITOCONDRIA	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS)*
	ALDH18B3b	LOC101499756	XP_004491997	PS00902, PS01223	CITOSOL	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS); Glutamate 5-kinase activity; Glutamate-5-semialdehyde dehydrogenase*
	ALDH18B3c		XP_012568863	PS00902, PS01223	CITOSOL	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS); Glutamate 5-kinase activity; Glutamate-5-semialdehyde dehydrogenase*
	ALDH18B3g	LOC101509592	XP_004514907	-	CITOSOL, MITITOCONDRIA	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS)*
	ALDH18B3d	LOC101512568	XP_004506632	PS00902, PS01223	MITOCONDRIA	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS); Glutamate 5-kinase activity; Glutamate-5-semialdehyde dehydrogenase*
	ALDH18B3e		NP_001296605	PS00902, PS01223	CITOSOL	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS); Glutamate 5-kinase activity; Glutamate-5-semialdehyde dehydrogenase*
	ALDH18B3h	LOC105852801	XP_012574961	PS01223	CITOSOL	SOLUBLE	*Delta 1-pyrroline-5-carboxylate synthetase (P5CS)*
22	ALDH22A1	LOC101512347	XP_004487758	E298 - C332	VIA SECRETORA, RE	MEMBRANA	*NADP-dependent malic enzyme (EC 1.1.1.40) (NADP-ME)*

Tabla 2. Caracterización de las proteínas CaALDH.

Familia	Locus ID	Proteína ID	Nº aminoácidos	RNA ID	Longitud RNA	Chr	Exones	Chr inicio	Chr fin	Peso Molecular (kDa)	pI
ALDH2B4a	LOC101490532	XP_004508853	536	XM_004508796	1911	Ca7	12	9459506	9464853	58,58	7,57
ALDH2B4b		XP_004508854	423	XM_004508797	1909	Ca7	12	9459506	9464853	46,47	6,76
ALDH2B7a	LOC101492709	XP_004509835	538	XM_004509778	1929	Ca7	11	21404992	21399773	57,99	6,58
ALDH2B7b		XP_004509834	539	XM_004509777	1949	Ca7	11	21404992	21399773	58,04	6,58
ALDH2C4c	LOC101493969	XP_004503432	480	XM_004503375	1678	Ca6	10	3278792	3283290	52,33	6,44
ALDH2C4b	LOC101513875	XP_012574767	409	XM_012719313	1345	Ca8	7	14992268	14984301	44,10	5,55
ALDH2C4a	LOC101514219	XP_004512967	503	XM_004512910	1754	Ca8	9	14998357	15002763	54,64	6,19
ALDH3F1b	LOC101491914	XP_004507095	488	XM_004507038	1819	Ca6	10	53416541	53426517	54,56	9,22
ALDH3F1a	LOC101497113	XP_004486968	494	XM_004486911	1718	Ca1	10	11135022	11130597	54,82	8,1
ALDH3F1c	LOC101511819	XP_012573731	488	XM_012718277	1753	Ca7	10	14455250	14450590	54,13	7,99
ALDH3H1f	LOC101488602	XP_012575132	144	XM_012719678	2141	Un	4	203490	206489	15,99	8,83
ALDH3H1g	LOC101497514	XP_004515934	210	XM_004515877	2162	Un	7	363621	368788	23,30	9,16
ALDH3H1h		XP_012567420	210	XM_012711966	2022	Un	7	363621	368788	23,30	9,16
ALDH3H1i		XP_012567421	210	XM_012711967	1508	Un	7	363621	368788	23,30	9,16
ALDH3H1j	LOC101502106	XP_004514084	134	XM_004514027	695	Un	3	165522	167554	15,07	9,49
ALDH3H1a	LOC101505038	XP_004503899	542	XM_004503842	2328	Ca6	12	6829381	6835156	59,76	7,96
ALDH3H1b		XP_004503900	534	XM_004503843	2349	Ca6	12	6829381	6835156	58,88	7,08
ALDH3H1c	LOC101510937	XP_004502482	488	XM_004502425	1819	Ca5	10	44008286	44002223	53,18	7,01
ALDH3H1d	LOC101511680	XP_004502485	486	XM_004502428	1711	Ca5	10	44024240	44016830	52,99	8,33
ALDH3H1e	LOC101515558	XP_004498346	488	XM_004498289	1760	Ca4	10	38313840	38325384	53,06	8,43
ALDH5F1	LOC101506901	XP_004488550	530	XM_004488493	1848	Ca1	20	37645881	37658279	56,59	6,58
ALDH6B2	LOC101490310	XP_004504867	539	XM_004504810	1975	Ca6	19	15177288	15170635	57,63	7,08
ALDH7A1a	LOC101513733	XP_004488077	508	XM_004488020	2131	Ca1	15	23039435	23046658	54,09	5,7
ALDH7A1b		XP_012574245	508	XM_012718791	2229	Ca1	15	23039435	23046658	54,09	5,7
ALDH10A8b	LOC101506136	XP_004508822	503	XM_004508765	2005	Ca7	14	9155131	9150436	54,40	5,37
ALDH10A8a	LOC101507930	XP_004501961	503	XM_004501904	2544	Ca5	15	39963815	39971531	54,53	5,37

ALDH11A3	LOC101510843	XP_004507722	496	XM_004507665	2020	Ca7	9	1260655	1264712	52,81	6,53
ALDH12A1	LOC101490107	XP_004508769	553	XM_004508712	2044	Ca7	16	8738314	8744683	61,30	6,17
ALDH18B3a	LOC101490622	XP_012572021	715	XM_012716567	2505	Ca6	21	1317437	1311762	77,65	6,62
ALDH18B3f	LOC101495530	XP_012575404	262	XM_012719950	1139	Un	9	287491	291250	28,79	7,69
ALDH18B3b	LOC101499756	XP_004491997	714	XM_004491940	2787	Ca3	20	8714210	8700487	77,39	5,96
ALDH18B3c		XP_012568863	717	XM_012713409	2796	Ca3	20	8714210	8700487	77,75	5,96
ALDH18B3g	LOC101509592	XP_004514907	271	XM_004514850	816	Un	9	332228	334535	29,57	6,41
ALDH18B3d	LOC101512568	XP_004506632	755	XM_004506575	2667	Ca6	21	44541900	44527967	81,90	6,8
ALDH18B3e		NP_001296605	715	NM_001309676	2585	Ca6	21	44541900	44527967	77,52	6,76
ALDH18B3h	LOC105852801	XP_012574961	248	XM_012719507	747	Un	8	190120	200777	27,72	4,34
ALDH22A1	LOC101512347	XP_004487758	595	XM_004487701	2962	Ca1	14	17180105	17171566	65,35	6,72

Tabla 3. Miembros de las familias ALDH identificados en plantas, humanos y hongos.

	Familias ALDH																								Nº genes	
Especie/grupo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	***ALDH***	**Referencias**
Human	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-		Jimenez-Lopez et al., 2010
Fungi	+	-	-	+	+	-	-	-	-	+	-	-	-	+	+	+	-	+	-	-	-	-	-	-		Jimenez-Lopez et al., 2010
Lupinus angustifolius	0	4	3	0	1	1	2	0	0	2	2	1	0	0	0	0	0	2	0	0	0	1	0	0	19	Jimenez-Lopez 2016
Glycine max	0	5	1	0	0	0	4	0	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	18	Kotchoni et al., 2012
Arabidopsis thaliana	0	3	3	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	16	Kirch et al., 2004
Chlamydomonas reinhardtii	0	1	0	0	1	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	1	9	Wood & Duff, 2009
Eutrema parvulum	0	3	3	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	16	Hou & Bartels, 2014
Eutrema salsugineum	0	3	4	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	17	Hou & Bartels, 2014
Gossypium raimondii	0	8	6	0	1	3	1	0	0	2	3	1	0	0	0	0	0	4	0	0	0	1	0	0	30	He et al., 2014
Malus domestica	0	13	7	0	2	2	2	0	0	2	3	2	0	0	0	0	0	4	0	0	0	2	0	0	39	Li et al., 2013
Oryza sativa	0	5	5	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	20	Gao & Han, 2009
Ostreococcus tauri	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	6	Wood & Duff, 2009
Physcomitrella patens	0	2	5	0	2	1	1	0	0	1	5	1	0	0	0	0	0	1	0	0	1	0	1	0	21	Wood & Duff, 2009
Populus trichocarpa	0	4	6	0	1	4	2	0	0	2	3	1	0	0	0	0	0	2	0	0	0	1	0	0	26	Tian et al., 2015
Selaginella moellendorffii	0	6	2	0	1	1	1	0	0	1	6	1	0	0	0	0	0	1	0	0	1	1	2	0	24	Brocker et al., 2013
Setaria italica	0	6	4	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	20	Khan et al., 2014
Sorghum bicolor	0	5	4	0	1	1	1	0	0	2	1	1	0	0	0	0	0	2	0	0	0	1	0	0	19	Paterson et al., 2009
Vitis vinifera	0	5	4	0	3	3	2	0	0	2	2	1	0	0	0	0	0	2	0	0	0	1	0	0	25	Zhang et al., 2012
Volvax carteri	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	7	Prochnik et al., 2010
Zea mays	0	6	5	0	2	1	1	0	0	3	1	1	0	0	0	0	0	2	0	0	0	1	0	0	23	Jimenez-Lopez et al., 2010
Solanum lycopersicum	0	8	5	0	2	1	2	0	0	2	4	1	0	0	0	0	0	2	1	0	0	1	0	0	29	Jimenez-Lopez et al., 2016
Cicer arietinum	0	5	10	0	1	1	1	0	0	2	1	1	0	0	0	0	0	6	0	0	0	1	0	0	29	**Este trabajo**

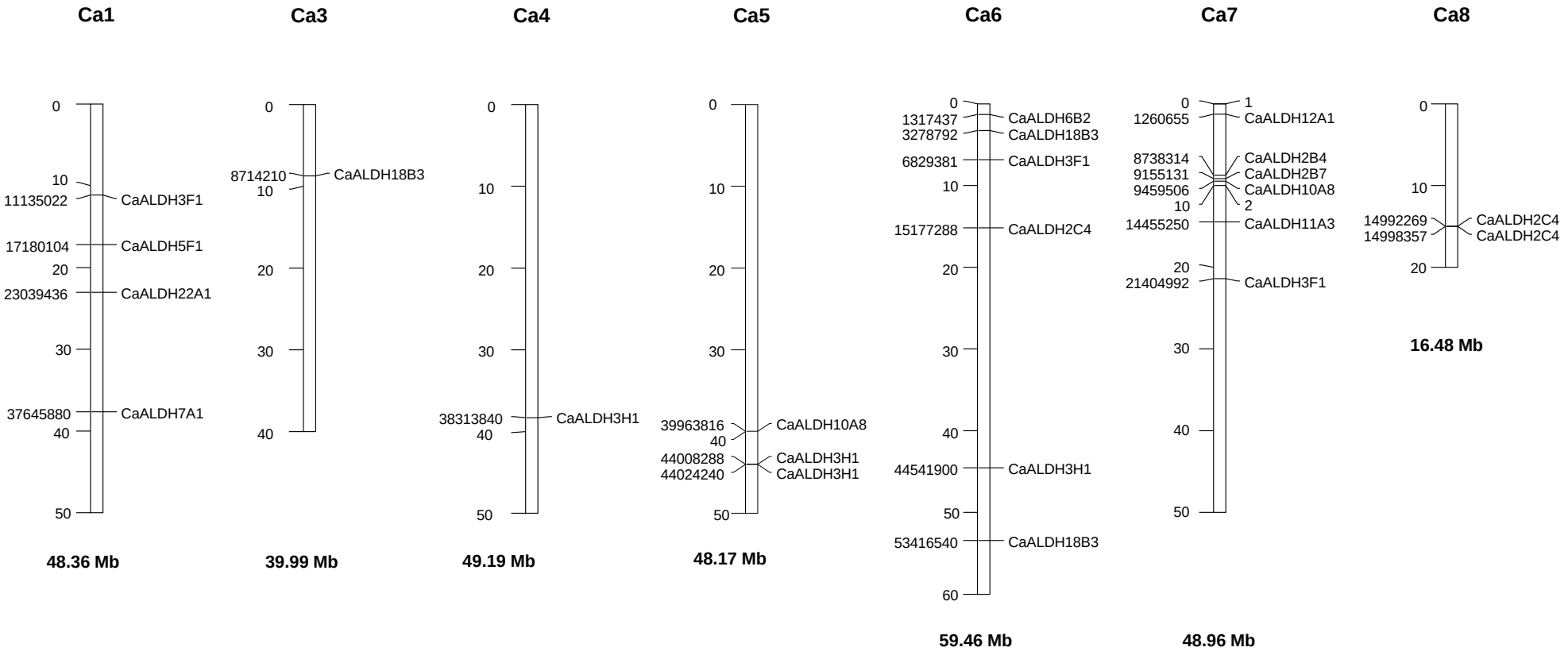

Figura 3. Distribución genómica de los genes *CaALDH* en los cromosomas de garbanzo. El número de cromosoma y los tamaños (Mb) están indicados encima y debajo de cada barra respectivamente. Solo los cromosomas con genes *CaALDH* están representados.

3. **Filogenia**

Con el fin de estudiar las relaciones filogenéticas entre las ALDHs de garbanzo y *Medicago*, realizamos un alineamiento múltiple entre sus secuencias. Ambas especies están estrechamente relacionadas; *C. arietinum* divergió de *M. truncatula* hace ~10-20 millones de años (Varshney et al., 2013). Este análisis filogenético nos da una visión de los cambios ocurridos durante la especiación.

El árbol filogenético (Figura 4) muestra 2 pares (*sister pairs*) entre proteínas de garbanzo y un par entre proteínas de *Medicago*, lo que sugiere duplicaciones recientes después de que las dos especies se separaran resultado de la expansión natural de la familia en cada especie. Por otro lado, encontramos 14 pares CaALDH - MtALDH que indican una gran conservación de estas secuencias.

Las secuencias ALDH18 aparecen en un *cluster* apartado del resto de secuencias. Así la filogenia recoge la diversidad estructural de esta familia señalada por Li et al. (2013). Todas aparecen con sus respectivos pares de *Medicago*, a excepción de las secuencias de las tres proteínas problemáticas en la clasificación (18B3h, 18B3f y 18B3g, de corta longitud de aa) que aparecen en un subgrupo junto a una MtALDH18 y a CaALDH18B3d. Así, se plantea la hipótesis de que esas tres ALDH18 de *C. arietinum* podrían ser fruto de duplicaciones posteriores a la especiación y representan la expansión de esta familia en garbanzo.

A continuación se encuentra el *cluster* que corresponde a las secuencias ALDH12. Considerando la función de las familias 12 y 18 en la misma ruta metabólica, tiene sentido que su secuencia también difiera del resto de secuencias del análisis.

La identidad 100% del par *Cicer* - *Medicago* en las secuencias ALDH7 coincide con la idea de que esta familia es una de las proteínas eucarióticas más conservadas en la evolución (Missihoun et al., 2014).

El resto de secuencias se encuentran englobadas en un *cluster* que a su vez se subdivide. En el subgrupo de las secuencias ALDH3, los pares *Cicer*-*Cicer* y *Medicago*-*Medicago* de la subfamilia 3H1 dan indicios de la expansión diferencial de esta familia en cada especie. En el subgrupo de las secuencias ALDH6 encontramos un caso contrario al postulado para la familia 18: *Medicago* contiene dos MtALDH6B2, y *Cicer* una CaALDH6B2. La presión evolutiva podría haber propiciado la pérdida de una de ellas en

garbanzo después de la especiación; o, más probable, es el caso de una duplicación reciente como parte de la expansión de la familia génica en *Medicago*. Tenemos el mismo caso en los subgrupos de las secuencias ALDH11, ALDH10 (o BADH) y ALDH2B4.

La aldehído deshidrogenasa no caracterizada de *Medicago* aparece asociada al par *Cicer-Medicago* en el *cluster* de las secuencias ALDH22; podría pertenecer a este familia. Sin embargo, la comprobación mediante BLASTP descarta esta opción.

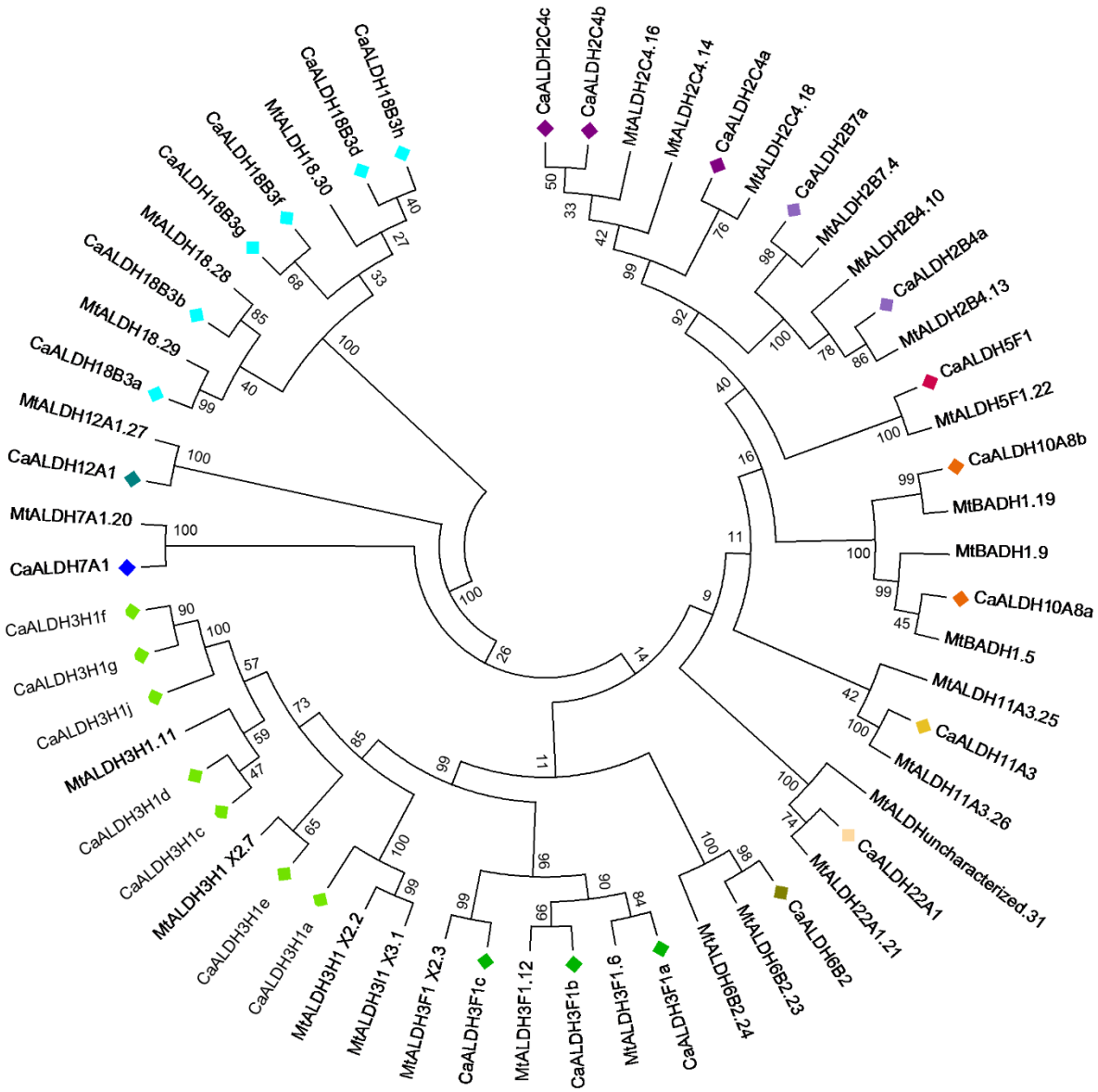

Figura 4. Análisis de las relaciones filogenéticas de ALDH en C. arietinum y M. truncatula. 29 proteínas CaALDH y 28 proteínas MtALDH. El porcentaje de árboles replicados en el que las proteínas se asociaron en un mismo cluster en el test bootstrap (1000 réplicas) está indicado al lado de las ramas. Las CaALDHs aparecen marcadas con un rombo, los colores indican cada familia y subfamilia.

4. **Análisis duplicación**

Para corroborar las hipótesis de duplicaciones indicadas a lo largo del trabajo, se realizó un análisis con la herramienta Circoletto. Las secuencias de las proteínas de la subfamilia 3F1 señalan una identidad ≤ 99% (Figura 5). En el árbol de filogenia, Figura 4, aparecen cada una con su respectivo par de *Medicago*. Considerando ambos resultados, en caso de existir un evento de duplicación en estos genes, se plantean dos hipótesis: este ocurrió en

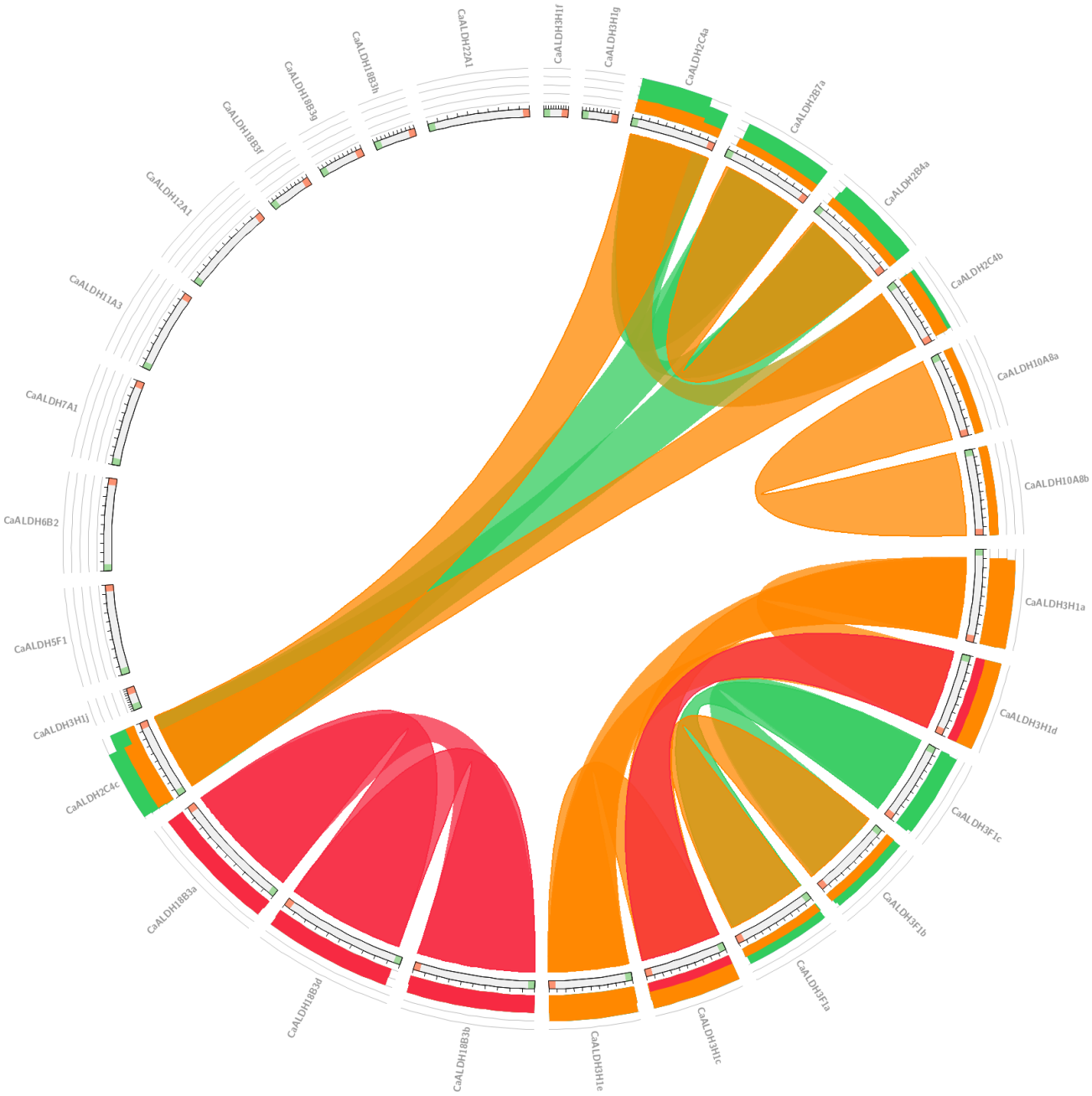

Figura 5. Análisis de duplicaciones entre CaALDHs con la herramienta Circoletto; 1 gen por locus. Colores: identidades 90 <= verde <=95, 95 < naranja <=99, y rojo > 99%.

el genoma de *Medicago* antes del proceso de especiación; o este ocurrió en ambos genomas de manera independiente en el momento de la especiación y la presión de selección ha actuado de igual forma en los resultados proteicos de la duplicación en ambas especies.

En filogenia se agrupa en un mismo *cluster* las secuencias de las proteínas MtALDH3H1.11, CaALDH3H1c y CaALDH3H1d (Figura 4). En la Figura 5 estas dos proteínas de garbanzo muestran una identidad > 99%. Coinciden los resultados de ambos análisis y apoyan la hipótesis del evento de duplicación posterior a la especiación.

5. **Expresión *in silico***

Para estudiar la función putativa de los genes *CaALDH* en garbanzo, hemos analizado sus perfiles de expresión utilizando los conjuntos de datos EST disponibles en el NCBI. Teniendo en cuenta los estrictos umbrales establecidos en la sección Material y Métodos, 19 de las 37 proteínas CaALDH están presentes en un total de 14 librerías EST (Tabla 4). Hay 16 secuencias presentes en estudios de resistencia a sequía: 2 en tejido de hoja y 15 en tejidos de raíz. Una aparece en raíces sensibles a la salinidad; una en raíces expuestas a cadmio, en la pared de la vaina y en la infección de *Fusarium oxysporum*; y otra en la respuesta inducida por un insecto en tejido de hoja.

Las ALDH18s y ALDH3s aparecen muy representadas en estos resultados. La expresión de genes que codifican para las proteínas de la familia 3 en estudios de sequía coincide con los resultados obtenidos para *A. thaliana* (Stiti et al., 2011). La presencia de las secuencias de la familia 18 en librerías bajo distintos estreses y condiciones apoya la conclusión de Gao & Han (2009) que sugiere un rol para esta familia en vías superpuestas al estrés osmótico; ya que muestran respuestas distintas entre ellas, pero todas una gran expresión inducida por la sequía.

El resto de resultados también coinciden con los obtenidos en estudios anteriores: la familia ALDH6, representadas por metilmalonil semialdehído deshidrogenasas, se han visto expresadas en tratamientos prolongados de salinidad y deshidratación (Zhang et al., 2012); ALDH7 se propone como parte de las vías generales de respuesta al estrés (Brocker et al., 2013); y las proteínas ALDH10, también conocidas como betaína aldehído deshidrogenasas (BADHs), han sido ampliamente estudiadas por su papel en las respuestas al estrés y la producción del osmoprotector glicina betaína (GB), cuya

acumulación aumenta para contrarrestar las consecuencias negativas del desequilibrio osmótico (Fitzgerald et al., 2009).

Las proteínas que aparecen en las librerías LIBEST_026410 y LIBEST_026408 son de especial interés por la propia metodología empleada para la construcción de la librería (librerías sustractivas *SSH*); identificando las poblaciones de ARN mensajeros que son específicos de una condición o tratamiento. Esto hace que se pueda sacar una conclusión fiable de los resultados y podamos deducir que estas proteínas desempeñan una función importante en la respuesta y tolerancia a la sequía en *Cicer arietinum*. Futuros trabajos experimentales deberían explorar esta hipótesis.

Tabla 4. Resultados expresión *in silico*, librerías y *loci* en los que aparecen las aldehído deshidrogenasas de garbanzo.

Librería EST	Proteína	Locus	Descripción Librería	Referencia
LIBEST_014674	CaALDH6B2	CK148910 CK149025 CK148985	*SSH root tissue library. ICC4958 as tester and Annigeri as driver, both drought tolerant*	Buhariwalla et al., 2005
LIBEST_022820	CaALDH10A8a	FE671085	*Drought-stressed chickpea leaf library. XJ-209 cultivar, drought tolerant*	Wang et al., 2012
LIBEST_022821	CaALDH3F1c	FE669635	*XJ-209 cultivar (drought tolerant) water-stressed leaf tissue library*	
LIBEST_024779	CaALDH3H1f CaALDH3H1g CaALDH3H1h CaALDH3H1i CaALDH3H1j	GR394613	*ICC1882 (drought sensitive) slow drought stressed root cDNA library*	Varshney et al., 2009
LIBEST_024782	CaALDH18B3e	GR397237	*ICC4958 field drought stressed root cDNA library*	
LIBEST_024785	CaALDH22A1	GR401364 GR401365	*ICCV2 (salinity sensitive) Salinity-stressed chickpea root cDNA library*	
LIBEST_026408	CaALDH7A1a CaALDH7A1b	HO062359	*Chickpea drought stressed root tissue cDNA SSH library AB1. Water-stressed ICC4958 (tolerant as tester and water-stressed ICC1882 (sensitive) as driver*	Deokar et al., 2011
	CaALDH10A8b	HO062466		
	CaALDH18B3a CaALDH18B3g CaALDH18B3d CaALDH18B3h CaALDH18B3e	HO062469		
LIBEST_026410	CaALDH10A8a	HO063270	*Chickpea drought stressed root tissue cDNA SSH library AR1. ICC4958 tolerant cultivar; water-stress as tester and control as driver*	
	CaALDH18B3h CaALDH18B3a CaALDH18B3g	HO063530 HO063265		
LIBEST_026414	CaALDH10A8a	HO067503	*Chickpea drought stressed root tissue cDNA SSH library BULK1. RILs derived from cross between ICC4958 (tolerant) and ICC1882 (sensitive). Water-stressed ICC4958 as tester and water-stressed ICC1882 as driver*	
LIBEST_027775	CaALDH11A3	JK815345	*Cicer arietinum (chickpea) cadmium-exposed root cDNA library. Pusa1105 cultivar*	
LIBEST_028463	CaALDH11A3	JZ714742	*Suppressive Subtracted cDNA of susceptible chickpea cultivar JG 62 after inoculation with Fusarium oxysporum f. sp. ciceris*	

LIBEST_028743	CaALDH11A3	JZ923355	*Cicer arietinum pod wall subtractive cDNA library*
LIBEST_020668	CaALDH10A8a	EH059086 EH059318	*Drought stress up-regulated SSH-cDNA library prepared from flowering stage chickpea plant tissue. Pusa 362 cultivar*
LIBEST_020862	CaALDH3H1e	EL585382	*Pusa-362 cultivar. Chickpea insect-induced leaf tissue SSH library*

CONCLUSIONES

1. En el análisis del genoma de garbanzo identificamos un total de 37 ALDHs codificadas por 29 genes ampliamente distribuidos en el genoma. Los cromosomas 6 y 7 contienen la mayor parte de ellos; en el cromosoma 2 no se encuentra ninguno.

2. *C. arietinum* podría presentar ALDH adicionales de respuesta al estrés: familias 3 y 18 son particularmente abundantes. En la expresión *in silico*, 16 de las 19 familias se encuentran en librerías de tolerancia a sequía, uno de los estreses abióticos que más afectan al rendimiento del garbanzo.

3. Los estudios filogenéticos con su antecesor cercano *Medicago* y el análisis de duplicación muestran la fuerte conservación de la familia génica *ALDH* y posibles duplicaciones posteriores a la especiación.

4. La identificación y caracterización, por primera vez, de esta familia génica tiene un potencial interés agronómico ya que la información que aquí se presenta podría ser usada por el programa de mejora del cultivo.

5. La bioinformática es actualmente una rama fundamental en los estudios genómicos: el método de análisis bioinformático de secuencias, combinado con la anotación manual empleado en este trabajo, es una aproximación poderosa para la identificación de familias génicas en especies con genomas disponibles. Sin el libre acceso a la información de las bases de datos, el trabajo no habría sido posible. Por eso, con este trabajo quiero resaltar la importancia de la ciencia abierta. Acorde con esto, todos los scripts elaborados para el análisis de datos de este TFM, la documentación acompañante de las funciones, así como los principales resultados, están a libre disposición en el sitio web creado para alojar este trabajo.
Repositorio de scripts: https://github.com/RocioCarmonaMolero/ScriptProteinas;
página web del TFM: https://rociocarmonamolero.github.io/TFMweb/.

REFERENCIAS

1. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21), 3387-3395.
2. Ashraf, N., Ghai, D., Barman, P., Basu, S., Gangisetty, N., Mandal, M. K., ... & Chakraborty, S. (2009). Comparative analyses of genotype dependent expressed sequence tags and stress-responsive transcriptome of chickpea wilt illustrate predicted and unexpected genes and novel regulators of plant immunity. *BMC genomics*, 10(1), 1.
3. Ayala, A., Muñoz, M. F., & Argüelles, S. (2014). Lipid peroxidation: production, metabolism, and signaling mechanisms of malondialdehyde and 4-hydroxy-2-nonenal. *Oxidative medicine and cellular longevity*, *2014*.
4. Barclay, K. D., & McKersie, B. D. (1994). Peroxidation reactions in plant membranes: effects of free fatty acids. *Lipids*, 29(12), 877-882.
5. Brocker, C., Vasiliou, M., Carpenter, S., Carpenter, C., Zhang, Y., Wang, X., ... & Nebert, D. W. (2013). Aldehyde dehydrogenase (ALDH) superfamily in plants: gene nomenclature and comparative genomics. *Planta*, 237(1), 189-210.
6. Buhariwalla, H. K., Jayashree, B., Eshwar, K., & Crouch, J. H. (2005). Development of ESTs from chickpea roots and their use in diversity analysis of the *Cicer* genus. *BMC Plant Biology*, 5(1), 16.
7. Chen, X., Zeng, Q., & Wood, A. J. (2002). The stress-responsive *Tortula ruralis* gene ALDH21A1 describes a novel eukaryotic aldehyde dehydrogenase protein family. *Journal of Plant Physiology*, 159(7), 677-684.
8. Chidambaranathan, P., Jagannadham, P. T. K., Satheesh, V., Kohli, D., Basavarajappa, S. H., Chellapilla, B., ... & Srinivasan, R. (2017). Genome-wide analysis identifies chickpea (*Cicer arietinum*) heat stress transcription factors (Hsfs) responsive to heat stress at the pod development stage. *Journal of plant research*, 1-18.
9. Claros, M. G., & Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, *241*(3), 779-786.
10. CRF, INIA (2018). Centro Nacional de Recursos Fitogenéticos, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria.
11. Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, 26(20).
12. Deokar, A. A., Kondawar, V., Jain, P. K., Karuppayil, S. M., Raju, N. L., Vadez, V., ... & Srinivasan, R. (2011). Comparative analysis of expressed sequence tags (ESTs) between drought-tolerant and-susceptible genotypes of chickpea under terminal drought stress. *BMC Plant Biology*, 11(1), 70.
13. Die, J. V., Gil, J., & Millan, T. (2018). Genome-wide identification of the auxin response factor gene family in *Cicer arietinum*. *BMC genomics*, 19(1), 301.

14. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1), 113.
15. Emanuelsson, O., Nielsen, H., & Von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, 8(5), 978-984.
16. Emanuelsson, O., Nielsen, H., Brunak, S., & Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*, 300(4), 1005-1016.
17. Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4), 783-791.
18. Fitzgerald, T. L., Waters, D. L., & Henry, R. J. (2009). Betaine aldehyde dehydrogenase in plants. *Plant biology*, *11*(2), 119-130.
19. FAOSTAT (2017). Food and Agriculture Organization of the United Nations.
20. Gao, C., & Han, B. (2009). Evolutionary and expression study of the aldehyde dehydrogenase (ALDH) gene superfamily in rice (*Oryza sativa*). *Gene*, 431(1), 86-94.
21. Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., & Gil, Y. (2013). Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PloS one*, *8*(11), e80278.
22. Guo, X., Wang, Y., Lu, H., Cai, X., Wang, X., Zhou, Z., ... & Liu, F. (2017). Genome-wide characterization and expression analysis of the aldehyde dehydrogenase (ALDH) gene superfamily under abiotic stresses in cotton. *Gene*, 628, 230-245.
23. Hampton, S. E., Anderson, S. S., Bagby, S. C., Gries, C., Han, X., Hart, E. M., ... & Mudge, J. (2015). The Tao of open science for ecology. *Ecosphere*, *6*(7), 1-13.
24. He, D., Lei, Z., Xing, H., & Tang, B. (2014). Genome-wide identification and analysis of the aldehyde dehydrogenase (ALDH) gene superfamily of *Gossypium raimondii*. *Gene*, *549*(1), 123-133.
25. Hobohm, U., & Sander, C. (1995). A sequence property approach to searching protein databases. *Journal of molecular biology*, 251(3), 390-399.
26. Hou, Q., & Bartels, D. (2014). Comparative study of the aldehyde dehydrogenase (ALDH) gene superfamily in the glycophyte *Arabidopsis thaliana* and *Eutrema* halophytes. *Annals of botany*, 115(3), 465-479.
27. Igarashi, Y., Yoshiba, Y., Sanada, Y., Yamaguchi-Shinozaki, K., Wada, K., & Shinozaki, K. (1997). Characterization of the gene for Δ 1-pyrroline-5-carboxylate synthetase and correlation between the expression of the gene and salt tolerance in *Oryza sativa* L. *Plant molecular biology*, 33(5), 857-865.
28. Jain, M., Misra, G., Patel, R. K., Priya, P., Jhanwar, S., Khan, A. W., ... & Yadav, M. (2013). A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *The Plant Journal*, 74(5), 715-729.
29. Jimenez-Lopez, J. C. (2016). Narrow-leafed lupin (*Lupinus angustifolius* L.) functional identification and characterization of the aldehyde dehydrogenase (ALDH) gene superfamily. *Plant Gene*, 6, 67-76.

30. Jimenez-Lopez, J. C., Gachomo, E. W., Seufferheld, M. J., & Kotchoni, S. O. (2010). The maize ALDH protein superfamily: linking structural features to functional specificities. *BMC structural biology*, 10(1), 43.
31. Jimenez-Lopez, J. C., Lopez-Valverde, F. J., Robles-Bolivar, P., Lima-Cabello, E., Gachomo, E. W., & Kotchoni, S. O. (2016). Genome-Wide Identification and Functional Classification of Tomato (*Solanum lycopersicum*) Aldehyde Dehydrogenase (ALDH) Gene Superfamily. *PloS one*, 11(10), e0164798.
32. Khan, Y., Yadav, A., Bonthala, V. S., Muthamilarasan, M., Yadav, C. B., & Prasad, M. (2014). Comprehensive genome-wide identification and expression profiling of foxtail millet [*Setaria italica* (L.)] miRNAs in response to abiotic stress and development of miRNA database. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 118(2), 279-292.
33. Kirch, H. H., Bartels, D., Wei, Y., Schnable, P. S., & Wood, A. J. (2004). The ALDH gene superfamily of *Arabidopsis*. *Trends in plant science*, 9(8), 371-377.
34. Kotchoni, S. O., Jimenez-Lopez, J. C., Gao, D., Edwards, V., Gachomo, E. W., Margam, V. M., & Seufferheld, M. J. (2010). Modeling-dependent protein characterization of the rice aldehyde dehydrogenase (ALDH) superfamily reveals distinct functional and structural features. *PloS one*, 5(7), e11516.
35. Kotchoni, S. O., Jimenez-Lopez, J. C., Kayodé, A. P., Gachomo, E. W., & Baba-Moussa, L. (2012). The soybean aldehyde dehydrogenase (ALDH) protein superfamily. *Gene*, 495(2), 128-133.
36. Li, X., Guo, R., Li, J., Singer, S. D., Zhang, Y., Yin, X., ... & Wang, X. (2013). Genome-wide identification and analysis of the aldehyde dehydrogenase (ALDH) gene superfamily in apple (*Malus× domestica* Borkh.). *Plant physiology and biochemistry*, 71, 268-282.
37. Marwick, B. (2017). Computational reproducibility in archaeological research: basic principles and a case study of their implementation. *Journal of Archaeological Method and Theory*, *24*(2), 424-450.
38. Matsuda, S., Vert, J. P., Saigo, H., Ueda, N., Toh, H., & Akutsu, T. (2005). A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science*, 14(11), 2804-2813.
39. McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., ... & Spies, J. R. (2016). How open science helps researchers succeed, ELife 5. *See https://doi. org/10.7554/elife*, *16800*.
40. Millán, T., Madrid, E., Cubero, J. I., Amri, M., Castro, P., & Rubio, J. (2015). Chickpea. In *Grain Legumes* (pp. 85-109). Springer, New York, NY.
41. Missihoun, T. D., Hou, Q., Mertens, D., & Bartels, D. (2014). Sequence and functional analyses of the aldehyde dehydrogenase 7B4 gene promoter in *Arabidopsis thaliana* and selected Brassicaceae: regulation patterns in response to wounding and osmotic stress. *Planta*, 239(6), 1281-1298.
42. Missihoun, T. D., Kirch, H. H., & Bartels, D. (2012). T-DNA insertion mutants reveal complex expression patterns of the aldehyde dehydrogenase 3H1 locus in *Arabidopsis thaliana*. *Journal of experimental botany*, 63(10), 3887-3898.

43. Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., ... & Schmutz, J. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, *457*(7229), 551.
44. Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227.
45. Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., ... & Hellsten, U. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science*, *329*(5988), 223-226.
46. Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406-425.
47. Seki, M., Umezawa, T., Urano, K., & Shinozaki, K. (2007). Regulatory metabolic networks in drought stress responses. *Current opinion in plant biology*, 10(3), 296-302.
48. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... & Thompson, J. D. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1), 539.
49. Sophos, N. A., & Vasiliou, V. (2003). Aldehyde dehydrogenase gene superfamily: the 2002 update. *Chemico-biological interactions*, 143, 5-22.
50. Stiti, N., Missihoun, T. D., Kotchoni, S., Kirch, H. H., & Bartels, D. (2011). Aldehyde dehydrogenases in *Arabidopsis thaliana*: biochemical requirements, metabolic pathways, and functional analysis. *Frontiers in plant science*, 2, 65.
51. Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30(12), 2725-2729.
52. Team, R. (2016). RStudio: Integrated Development for R. Boston: RStudio Inc.; 2015.
53. Team, R. C. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
54. Tian, F. X., Zang, J. L., Wang, T., Xie, Y. L., Zhang, J., & Hu, J. J. (2015). Aldehyde dehydrogenase Gene superfamily in *Populus*: organization and expression divergence between Paralogous Gene pairs. *PLoS One*, *10*(4), e0124669.
55. Varshney, R. K., Hiremath, P. J., Lekha, P., Kashiwagi, J., Balaji, J., Deokar, A. A., ... & Siddique, K. H. (2009). A comprehensive resource of drought-and salinity-responsive ESTs for gene discovery and marker development in chickpea (*Cicer arietinum* L.). *BMC genomics*, 10(1), 523.
56. Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., ... & Millan, T. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature biotechnology*, 31(3), 240.
57. Vasiliou, V., Bairoch, A., Tipton, K. F., & Nebert, D. W. (1999). Eukaryotic aldehyde dehydrogenase (ALDH) genes: human polymorphisms, and recommended nomenclature based on divergent evolution and chromosomal mapping. *Pharmacogenetics*, *9*(4), 421-434.

58. Vasiliou, V., Pappa, A., & Petersen, D. R. (2000). Role of aldehyde dehydrogenases in endogenous and xenobiotic metabolism. *Chemico-biological interactions*, *129*(1-2), 1-19.
59. Verbruggen, N., & Hermans, C. (2008). Proline accumulation in plants: a review. *Amino acids*, 35(4), 753-759.
60. Wang, X., Liu, Y., Jia, Y., Gu, H., Ma, H., Yu, T., ... & Zhang, J. (2012). Transcriptional responses to drought stress in root and leaf of chickpea seedling. *Molecular biology reports*, *39*(8), 8147-8158.
61. Wood, A. J., & Duff, R. J. (2009). The aldehyde dehydrogenase (ALDH) gene superfamily of the moss *Physcomitrella patens* and the algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri*. *The Bryologist*, 112(1), 1-11.
62. Xing, W., & Rajashekar, C. B. (2001). Glycine betaine involvement in freezing tolerance and water stress in *Arabidopsis thaliana*. *Environmental and Experimental Botany*, 46(1), 21-28.
63. Zhang, Y., Mao, L., Wang, H., Brocker, C., Yin, X., Vasiliou, V., ... & Wang, X. (2012). Genome-wide identification and analysis of grape aldehyde dehydrogenase (ALDH) gene superfamily. *PloS one*, 7(2), e32153.
64. Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (pp. 97-166).