UNIVERSIDAD DE ALMERÍA

TRABAJO DE FIN DE GRADO

Genómica comparada de cuatro especies de cultivo huérfanas: caracterización funcional de familias génicas expandidas

Autor: Juan Pablo Marczuk Rojas

Tutor: Lorenzo Carretero Paulet

Co-tutora: María Salinas Navarro

Grado en Biotecnología

Facultad de Ciencias Experimentales

Departamento de Biología y Geología

Área de Genética

Curso académico 2020-2021

Convocatoria de Mayo

Índice de contenido

Resumen ..... **i**

Summary ..... **ii**

1. Introducción **1**

1.1. Antecedentes 1

1.2. Motivación y precedentes del estudio 2

1.3. Presentación de los cultivos huérfanos africanos seleccionados para el estudio 3

1.3.1. Biología 4

1.3.2. Usos y propiedades 5

1.3.3. Características genómicas 8

2. Objetivos **11**

2.1. Anotación funcional genómica 11

2.2. Clasificación de ortogrupos/familias génicas 11

2.3. Modelización de la evolución de ortogrupos/familias génicas a escala genómica 11

2.4. Categorización funcional de los ortogrupos/familias génicas identificadas como expandidas 11

3. Materiales y metodología **12**

3.1. Genomas 12

3.2. Clasificación de ortogrupos/familias génicas 12

3.3. Modelización de las dinámicas evolutivas en ortogrupos/familias génicas 13

3.4. Reconstrucción de un árbol filogenético de especies 15

3.5. Anotación funcional de genomas 15

3.5.1. Identificación de posibles homólogos mediante alineamiento local de secuencias proteicas 16

3.5.2. Mapeo 16

3.5.3. Anotación 17

3.5.4. Ampliación de la anotación funcional 17

3.5.5. Análisis de enriquecimiento funcional de listas de genes 18

3.6. Representaciones gráficas 19

4. Resultados y discusión **20**

4.1. Filtrado previo de los genomas 20

4.2. Anotación funcional de genomas 20

4.3. Clasificación en ortogrupos/familias génicas en los genomas de 11 especies de plantas 22

4.4. Identificación de ortogrupos/familias génicas expandidos y contraídos en cuatro especies del AOCC 26

4.4.1. Árbol de especies 26

4.4.2. Resultados generales del análisis de Badirate 27

4.5. Categorización funcional de los ortogrupos/familias génicas expandidos..... 28
4.5.1. Resultados de los tests de enriquecimiento funcional 28
4.5.2. Funciones biológicas enriquecidas identificadas en las familias expandidas 31
5. Conclusiones.....35
6. Bibliografía 36

Índice de tablas

Tabla 1. Propiedades de *Faidherbia albida* 5
Tabla 2. Propiedades de *Lablab purpureus* 6
Tabla 3. Propiedades de *Sclerocarya birrea* 7
Tabla 4. Propiedades de *Vigna subterranea* 8
Tabla 5. Datos de los genomas secuenciados de las especies 8
Tabla 6. Datos del ensamblaje de los genomas secuenciados de las especies 9
Tabla 7. Completitud de los genomas secuenciados de las especies 9
Tabla 8. Proporción de secuencias repetitivas en los genomas secuenciados de las especies .. 10
Tabla 9. Especies y versiones de sus genomas 12

Índice de figuras:

Figura 1. Especies del AOCC seleccionadas 3
Figura 2. Número de genes presentes en los genomas de las 11 especies antes y después del filtrado 20
Figura 3. Estadísticas por especie de la anotación funcional 22
Figura 4. Estadísticas por especie de la clasificación de OrthoFinder 24
Figura 5. Resultados del análisis de OrthoFinder 25
Figura 6. Árbol filogenético ultramétrico de las 11 especies utilizado para el análisis de Badirate 26
Figura 7. Estadísticas por especie del análisis de Badirate 28
Figura 8. Estadísticas por especie de los resultados de los tests de enriquecimiento funcional 29
Figura 9. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Proceso Biológico asignados a las familias génicas expandidas de cada especie 29
Figura 10. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Componente Celular asignados a las familias génicas expandidas de cada especie 30
Figura 11. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Función Molecular asignados a las familias génicas expandidas de cada especie 30

Resumen

En las últimas décadas, han surgido distintas iniciativas para abordar la falta de recursos agrícolas que permitan satisfacer las necesidades alimentarias y nutricionales de una población mundial creciente (especialmente en África), sorteando las dificultades que plantea el cambio climático y la pérdida de agrodiversidad en la productividad de los sistemas agrícolas. Una de estas iniciativas es el Consorcio de Cultivos Africanos Huérfanos (AOCC por sus siglas en inglés), que promueve la secuenciación de los genomas y transcriptomas de distintas especies de cultivo locales infraestudiados (huérfanos) procedentes del África subsahariana con el objeto de proporcionar recursos agrigenómicos que permitan la mejora de dichas especies. En este estudio, distintas aproximaciones de genómica comparada fueron implementadas en cuatro de las especies seleccionadas por el consorcio cuyos genomas ya han sido presentados, la anacardiácea *Sclerocarya birrea* y las leguminosas *Faidherbia albida*, *Lablab purpureus*, y *Vigna subterranea*. En primer lugar, se anotaron funcionalmente 68433 genes con 315830 términos GO, 17909 genes con 80986 términos EC y 78951 genes con 411776 términos InterPro. A continuación, se obtuvo una clasificación de 17998 ortogrupos o familias génicas en las cuatro especies del AOCC más otras siete representativas de distintos linajes de angioespermas. A partir de dicha clasificación, se llevó a cabo un análisis evolutivo bajo máxima verosimilitud para modelar la variación del tamaño de cada una de las familias génicas en un contexto filogenético con la finalidad de detectar aquellas familias significativamente expandidas y contraídas en los genomas de las especies del AOCC seleccionadas. Finalmente, se identificaron funciones biológicas sobrerrepresentadas entre las familias expandidas, las cuales podrían estar en el origen de adaptaciones biológicas relevantes, así como de rasgos agronómicos de potencial interés, como la resistencia a estrés biótico, la tolerancia a estrés abiótico y la capacidad de formar nódulos fijadores de nitrógeno.

Palabras clave: Genomas, Cultivos Africanos Huérfanos, expansiones y contracciones de familias génicas, Enriquecimiento funcional, Genómica Comparada, caracteres agronómicos

Summary

Over the last decades, different initiatives have surged to address the lack of agricultural resources to meet the alimentary and nutritional requirements of a growing global population (specially in Africa) and bypass the complications posed by climate change and the loss of agrodiversity to the productivity of farming systems. One of these initiatives is the African Orphan Crops Consortium (AOCC) which promotes the sequencing of genomes and transcriptomes of understudied (orphan) local plants coming from Sub-saharan Africa with the aim of providing agrigenomic resources to facilitate their improvement. In this study different comparative genomics approaches were implemented on four of the species selected by the consortium whose genomes have already been reported, the anacardiaceaous *Sclerocarya birrea* and the leguminous *Faidherbia albida*, *Lablab purpureus*, and *Vigna subterranea*. First, a functional annotation was performed resulting in 68433 genes annotated with 315830 GO terms, 17909 genes annotated with 80986 EC terms and 78951 genes annotated with 411776 InterPro terms. Then, a classification of 17998 orthogroups or gene families was obtained in the genomes of the four AOCC species and seven other species representing different angiosperms lineages. From that classification a Maximum Likelihood evolutionary analysis was performed to model the size variation of each gene family in a phylogenetical context in order to detect those that are significantly expanded and contracted in the genomes of the selected AOCC species. Finally, overrepresented biological functions were identified in the expanded families, which could be at the origin of relevant biological adaptations as well as agronomical traits of potential interest such as resistance to biotic stress, tolerance to abiotic stress and the capacity to form nitrogen-fixing nodules.

Keywords: Genomes, African Orphan Crops, gene family expansions and contractions, Functional enrichment, Comparative Genomics, agronomical traits

1. Introducción

1.1. Antecedentes

Desde la invención de la agricultura hasta la actualidad, un número reducido de especies vegetales (alrededor de 30) ha alimentado al grueso de la humanidad de los cuales el arroz, el maíz y el trigo proporcionan la gran mayoría (FAO, 2010), quedando otras especies ignoradas y desaprovechadas (los denominados cultivos "huérfanos"). Hasta el inicio del siglo XXI los esfuerzos en aumentar la productividad agrícola y, con ello, la tasa de alimentación de las poblaciones humanas se ha basado en incrementar el rendimiento de cultivos ampliamente difundidos y estudiados e introducirlos en otras regiones afectadas por la escasez de alimentos y la malnutrición; la llamada Revolución verde. Por ejemplo, en el caso de África oriental se promovió el cultivo de maíz (World Bank, 2009) obteniéndose resultados desfavorables a nivel nutricional debido a su bajo contenido en ácidos grasos, así como otras carencias nutricionales (Sands, et al., 2009).

Como alternativa a esta práctica, para África oriental y el resto del África subsahariana se fundó en 2011 el Consorcio de Cultivos Africanos Huérfanos (mejor conocido por sus siglas en inglés, AOCC), una iniciativa para canalizar recursos hacia la explotación de especies vegetales propias del continente con potencial agrónomo demostrado tal y como atestigua su presencia en los sistemas agrícolas de comunidades locales, pero largamente ignoradas e infraestudiadas. En total, 101 especies huérfanas africanas fueron seleccionadas para la secuenciación de sus genomas y transcriptomas para proveer los recursosa agrigenómicos requeridos (marcadores moleculares y genes relacionados con caracteres agronómicos, principalmente) y, posteriormente, diseñar programas de mejora genética que faciliten su cultivo a gran escala y optimicen sus respectivos contenidos nutricionales (Hendre, et al., 2019) (http://africanorphancrops.org/).

Este grupo de plantas destacan por su tolerancia a condiciones de estrés tales como la sequía y su contenido nutricional heterogéneo (Tadele, 2019). Por estas cualidades, la Organización de las Naciones Unidas de la Alimentación y la Agricultura (ONUAA, o mejor conocida como *Food and Agriculture Organization*, FAO) ha recomendado las "inversiones en la investigación y mejora de la productividad, adaptabilidad y utilización de cultivos abandonados" (http://www.fao.org/news/story/en/item/1032516/icode/), convirtiéndose en una de las estrategias a implementar para el cumplimiento de la Agenda 2030 para el Desarrollo Sostenible, un plan trazado por las Naciones Unidas para eliminar la pobreza que se articula en 17 objetivos. En concreto, el estudio de cultivos ignorados puede enmarcarse en los siguientes objetivos:

- Objetivo 2 (Hambre Cero): plantea como metas acabar con la malnutrición, duplicar la productividad agrícola y promover y mantener la diversidad genética de semillas y plantas entre otras (https://www.un.org/sustainabledevelopment/hunger/).

- Objetivo 13 (Acción Climática): plantea como meta implementar medidas para combatir el cambio climático y sus efectos adversos (https://www.un.org/sustainabledevelopment/climate-change/).

- Objetivo 15 (Biodiversidad, Bosques y Desertificación): plantea como metas proteger, restablecer y promover el uso sostenible de los ecosistemas terrestres, luchar contra la desertificación, detener e invertir la degradación de las tierras y frenar la pérdida de biodiversidad entre otras (https://www.un.org/sustainabledevelopment/biodiversity/).

Por consiguiente, la explotación de cultivos huérfanos se ha convertido en uno de los pilares de una nueva revolución verde donde las prioridades serán la sostenibilidad, el incremento de la agrobiodiversidad y la resiliencia biológica, y la sustitución de monocultivos poco capaces de adaptarse adecuadamente a un mayor estrés climático, con escasas e irregulares precipitaciones, olas extremas de calor e inundaciones, y plagas y enfermedades cada vez más virulentas (Bailey-Serres, et al., 2019).

1.2. Motivación y precedentes del estudio

Tal y como se expuso en el apartado anterior, la Agenda 2030 ha puesto en relieve la urgencia de doblar la producción agrícola para antes del año 2050 debido a la necesidad de satisfacer las crecientes demandas de una población mundial en aumento, los cambios nutricionales y el incremento en el uso de biocombustibles, especialmente en el contexto del cambio climático (Ray, et al., 2013). Por ello, desde la FAO y otros organismos internacionales se ha promovido la búsqueda de cultivos alternativos a los tradicionales como es el caso de los cultivos huérfanos (Jamnadass, et al., 2020).

Sin embargo, por definición, la investigación y los programas de mejora asistidos por marcadores moleculares destinados a adaptar la inmensa diversidad de especies de cultivo huérfanas infrautilizadas a las necesidades de los productores, procesadores y consumidores de África y otras regiones del planeta se ha visto dificultada por la falta de recursos genéticos y genómicos.

En este sentido, el consorcio del AOCC está haciendo disponible un número creciente de genomas de distintas especies de cultivos huérfanos que, unido al desarrollo de nuevos métodos bioinformáticos, permiten análisis comparativos a escala genómica en un contexto evolutivo con el fin de identificar la base genética de las particularidades fenotípicas de dichas especies. Algunas de estas particularidades se manifiestan en forma de señales de especialización funcional génica en especies o linajes específicos que pueden estar en el origen de adaptaciones morfológicas y metabólicas relevantes, así como de rasgos agronómicos deseables con un impacto potencial en propiedades nutricionales, manejo/mejora de cultivos o producción de alimentos. Dichas señales pueden detectarse a nivel de:

I. Acumulación de cambios de aminoácidos bajo selección positiva o negativa relajada resultante en la adquisición de funciones génicas nuevas y/o especializadas.
II. Recableado de redes reguladoras existentes a través de ganancias o pérdidas de elementos en *cis* que conducen a la evolución de nuevas funciones en genes ancestrales.
III. Expansiones o contracciones significativas en el tamaño de familias génicas concretas en especies específicas debidos a cambios en las tasas de duplicación génica, a la adquisición de genes novedosos y a la perdida diferencial de genes.

Estas aproximaciones han sido implementadas en cultivos tan relevantes como las variedades de café arabica y robusta, revelando el origen polifilético de la ruta biosintética de cafeína y su disposición como grupos de genes metabólicos duplicados en tándem, así como expansiones significativas en genes relacionados con la defensa frente a patógenos y enzimas involucradas en la síntesis de alcaloides y flavonoides (Denoeud, et al., 2014), o el aguacate, ayudando a descifrar la genética del desarrollo y la bioquímica de las frutas carnosas y oleaginosas y revelando respuestas génicas adaptativas influenciado por patógenos (Rendon, et al., 2019), y otros tan prometedores como la Moringa, permitiendo detectar cantidades masivas de DNA

plastídico en el genoma nuclear que representan alrededor del 4,71%, la mayor reportada hasta la fecha (Ojeda-López, et al., 2020). Estudios similares han sido realizados también sobre: i) las plantas carnívoras *Utricularia gibba* y *Cephalotus follicularis*, ofreciendo información sobre la genómica evolutiva de adaptaciones morfológicas y metabólicas extremas a ambientes pobres en nutrientes (Fukushima, et al., 2017; Ibarra-Laclette, et al., 2013; Carretero-Paulet, et al., 2015; Carretero-Paulet, et al., 2015); ii) *Amborella trichopoda*, la única especie superviviente del linaje basal a todas las angiospermas, contribuyendo a identificar la base genómica de su explosiva diversificación evolutiva, el conjunto de herramientas genéticas ancestrales de las plantas con flores, o la evolución divergente del sistema vascular en monocots y eudicots (Albert, et al., 2013); o iii) los helechos *Salvinia cucullata* y *Azolla filiculoides*, permitiendo estudiar las bases moleculares de la fijación simbiótica de nitrógeno en estas especies y una de las principales transiciones en la evolución de las plantas terrestres (Li, et al., 2018).

En este estudio se implementó un enfoque similar destinado a identificar las bases moleculares y genéticas de las propiedades biológicas (especialente aquellas relacionadas con distintos caracteres de potencial interés agronómico) de cuatro especies del AOCC seleccionadas para este estudio cuyos genomas ya han sido publicados.

1.3. Presentación de los cultivos huérfanos africanos seleccionados para el estudio

Las cuatro especies vegetales del AOCC seleccionadas fueron las siguientes: *Faidherbia albida*, *Lablab purpureus*, *Sclerocarya birrea* y *Vigna subterranea* (***figura 1***). A continuación, se hará una descripción de cada una basada en una revisión bibliográfica efectuada para la preparación de este estudio. Cabe señalar que durante la revisión se encontró una escasa cantidad de estudios bioquímicos y genéticos que permitan precisar la(s) causa(s) de algunas de las propiedades agronómicas, nutricionales, farmacológicas u otras que se les atribuyen a estas plantas.

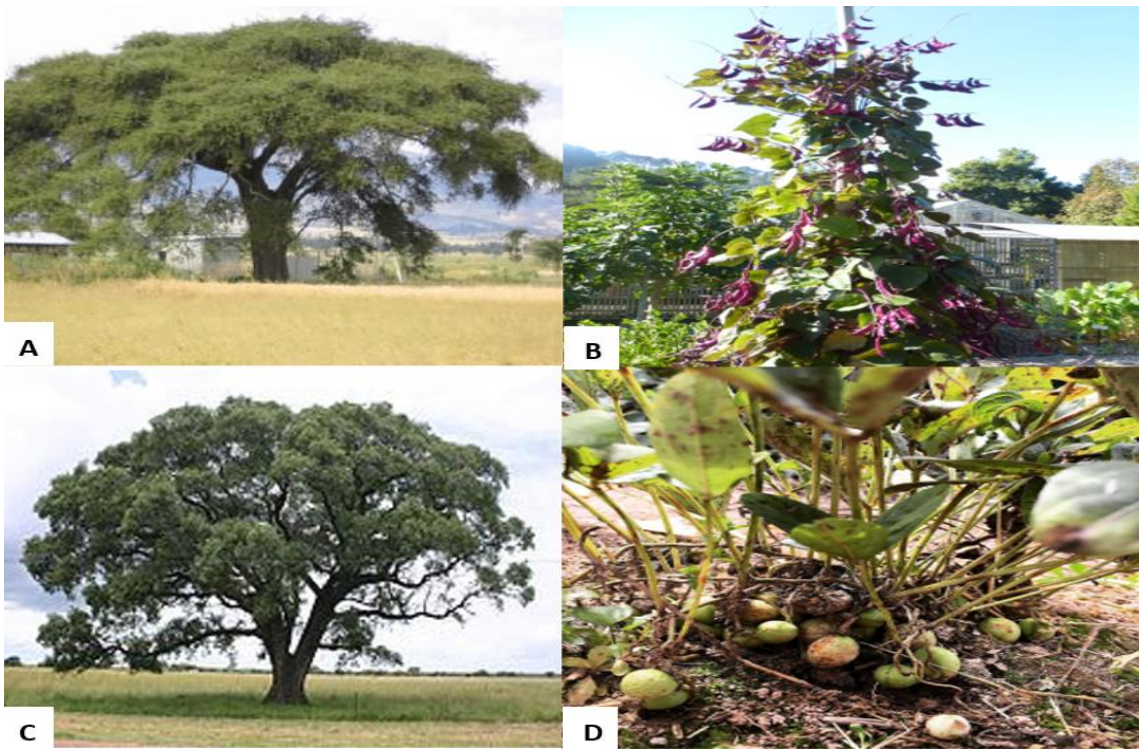

***Figura 1*. Especies del AOCC seleccionadas.** A. *Faidherbia albida*. B. *Lablab purpureus*. C. *Sclerocarya birrea*. D. *Vigna subterranea*.

1.3.1. Biología

Las cuatro especies seleccionadas corresponden a plantas eudicotiledóneas pertenecientes al clado de las llamadas plantas eurrósidas. Una de ellas pertenece a la familia de las anacardiáceas (*Anacardiaceae*) que incluye a diversas especies económicamente relevantes como el mango, el anacardo o el pistacho y caracterizada por el alto contenido en polifenoles de algunos de sus miembros (Schulze-Kaysers, et al., 2015) mientras que las otras tres pertenecen a la familia de las leguminosas (*Fabaceae*), caracterizadas por su capacidad de fijación del nitrógeno atmosférico (y, por consiguiente, fertilizar suelos) lo que ha permitido emplearlas como cultivos de intercalado, su uso como forraje y alimento humano (son una importante fuente de fibra, proteínas crudas y minerales) y su empleo en la preparación de plásticos biodegradables entre otras propiedades (Graham & Vance, 2003).

- *Faidherbia albida*

Su nombre vulgar es espina de invierno. Es una integrante del clado de las fábidas y del orden *Fabales*; pertenece a la familia de las leguminosas (*Fabaceas*) y la subfamilia *Mimosoideae* y constituye el género monotípico *Faidherbia*. Es un árbol espinoso y perenne natural de África oriental y occidental. Se cultiva en Níger y Sudán, entre otros países (Barnes & Fagg, 2003).

- *Lablab purpureus*

Su nombre vulgar es frijol de Dolichos. Es una integrante del clado de las fábidas y del orden *Fabales*; pertenece a la familia de las leguminosas (*Fabaceas*) y la subfamilia *Faboideae* y constituye el género monotípico *Lablab*. Es un árbol perenne productor de frijoles natural de la India que se introdujo en África desde el sureste de Asia durante el siglo VIII. Hoy en día es cultivado en las zonas tropicales de África y otras regiones como Australia (Mass, et al., 2010).

- *Sclerocarya birrea*

Su nombre vulgar es Marula. Es una integrante del orden *Sapindales*; pertenece a la familia *Anacardiaceae*, subfamilia *Spondiadoideae* y es, junto con *S. gilletti*, uno de los dos miembros del género. Es un árbol caducifolio natural de África austral, la cordillera Sudano-saheliana de África occidental y Madagascar. Es considerada una de las especies autóctonas de árbol frutal más importantes del sur de África, aunque ha permanecido silvestre hasta hace relativamente poco. Se cultiva en Sudáfrica y Namibia (Hall, et al., 2002).

- *Vigna subterranea*

Su nombre vulgar es maní de Bambara. Es una integrante del clado de las fábidas y del orden *Fabales*; pertenece a la familia de las leguminosas (*Fabaceas*) y la subfamilia *Faboideae* y es uno de los miembros del género *Vigna*, que incluye a unas 100 especies de las cuales alrededor de 10 han sido domesticadas en Africa y Asia, incluyendo *V. unguiculata* (chícharo) y *V. radiata* (judía Mungo), cultivadas extensamente en ampias áreas tropicales y subtropicales de África, Asia y América (Delgado-Salinas, et al., 2011). Es una planta anual y herbácea que se originó en África occidental y se cultiva en regiones de Tailandia, Malasia e Indonesia. (Hussin, et al., 2020).

1.3.2. Usos y propiedades

- *Faidherbia albida*

Se emplea como cultivo intercalado entre plantas anuales como mijo y maní en áreas secas y densamente pobladas. Se ha reportado que mejora la fertilidad del suelo debido a que transfiere a su superficie materia orgánica rica en nitratos lixiviados y otros nutrientes a través de sus hojas (Shi, et al., 2011). Así, durante la caída de las hojas, *F. albida* deposita un mantillo fino que se descompone rápidamente y enriquece la capa superior del suelo en nutrientes vegetales a la vez que mejora sus condiciones de humedad, su porosidad y la penetración de las raíces (Barnes & Fagg, 2003, p. 230). Por otro lado, también proporciona forraje para mantener rebaños de pequeños rumiantes en forma de follaje cortado y vainas que caen incluso durante períodos de sequía muy severa (Barnes & Fagg, 2003, p. 151).

Tradicionalmente, extractos de la corteza, la goma y las raíces de esta planta se han utilizado para tratar diversas dolencias. Principalmente, se utilizan como broncoemolientes para enfermedades respiratorias (solo la corteza se utiliza para este fin) y como astringentes para tratar trastornos gastrointestinales (Barnes & Fagg, 2003, p. 42). Aparte de estas propiedades, en la tabla 1 están recogidas otras reportadas recientemente:

Tabla 1. Propiedades de *Faidherbia albida*

Propiedad	Fuente	Información adicional
Capacidad citotóxica	Tchoukoua, et al., 2017 Tchoukoua, et al., 2018	Sintetiza saponinas compuestas
Propiedades antidiabéticas, antihiperlipidémicas y antioxidante en ratas	Karoune, et al., 2014	Sintetiza compuestos fenólicos con actividad antioxidante
	Garra, et al., 2020	Sintetiza flavonoles: quercitina, kaempferol, y sus glucósidos.
Capacidad de fijar nitrógeno atmosférico	Degefu et al., 2017	Forma nódulos con especies del género *Bradyrhizobium*
Actividad anticáncer	Guo, et al., 2020	Reportado sobre las líneas celulares HL-60, MCF-7, MDA-MB-231, Hep-2 y Hela
Actividad antitripanosomal sobre infecciones de *Trypanosoma evansi* en ratas	Ndidi, et al.,2015	
Propiedades antipiréticas, antiinflamatorias, antidiarreicas	Tijani, et al., 2008	
Actividades antibacteriana y antipalúdica	Salawu, et al., 2010	
	Usman, et al., 2013	

- *Lablab purpureus*

Se utiliza como cultivo intercalado en sistemas ganaderos. Lablab presenta una considerable diversidad fisiológica (Mass, et al., 2010). En la década de los años 90 se identificó un rango amplio de adaptación a distintos estreses abióticos, como por ejemplo estreses nutricionales derivados del escaso fósforo disponible en los suelos ácidos de Etiopía. También se ha reportado una adaptación altamente diferencial a condiciones semiáridas (Venkatesha, et al., 2010). Debido a su tolerancia a la sequía, Lablab podría ayudar a resolver la falta de recursos hídricos en muchas áreas semiáridas del África subsahariana, incluido el sur de África, y a paliar la disminución en el agua de riego disponible que se prevé experimentarán dichas áreas durante las próximas décadas como consecuencia del cambio climático (Venkatesha, et al., 2010). También ha sido reportada la idoneidad de su uso como forraje en 17 variedades locales obtenidas de los agricultores de Kenia.

Actualmente, los principales programas de mejora para este cultivo se están desarrollando en India y Bangladesh. Hasta ahora se han producido más de 30 variedades mejoradas de Lablab en varias instituciones indias desde que se implementaron los programas de reproducción (Gopalakrishnan, 2007).

Presenta un alto contenido de proteínas (superior al 20%), ácidos grasos esenciales (en concreto, ácido alfa-linolénico), glúcidos y cenizas (una alta presencia de cenizas es indicativo de una mayor presencia de macro y microminerales) y potasio entre otras sustancias de interés humano (Al-Snafi, 2017).

Lablab ha sido señalada durante décadas como una de las especies más agromorfológicamente diversas y una de las especies de leguminosas tropicales más versátiles por sus usos como alimento (judías verdes, vainas, hojas), forraje/abono verde, hierbas medicinales e incluso ornamentales (Maass, 2016; Raghu et al., 2018) y, más recientemente, se han revisado sus propiedades para diversos usos. En la tabla 2 están recogidas dichas propiedades:

Tabla 2. Propiedades de *Lablab purpureus*

Propiedad	Fuente	Información adicional
Efecto antidiabético	Ahmed, et al., 2015	
	Al-Snafi, 2017	
Aumenta el nivel de hierro en sangre	Somulung, et al., 2012	Incremento en el nivel de hemoglobina en ratas
	Al-Snafi, 2017	
Efecto antioxidante	Habib, et al., 2012	
	Al-Snafi, 2017	
Efecto antimicrobiano	Nasrin, et al., 2012	Actividad antifúngica con *Saccharromyces cerevaceaea, Candida albicans* y *Aspergillus niger*
	Priya & Jenifer, 2014	
Efecto insecticida	Janarthanan, et al., 2012	La proteína aislada, Arcelina, mostró actividad insecticida contra *Callosobruchu maculates*
	Al-Snafi, 2017	
Efecto citotóxico	Nasrin, et al., 2012	Efecto citotóxico contra *Artemia salina*

- *Sclerocarya birrea*

Desde la década de los 2000 se han desarrollado los primeros programas de domesticación para mejorar su cultivo a gran escala por lo que hasta el siglo XXI su explotación ha consistido principalmente en el aprovechamiento gastronómico de su fruto el cual es utilizado para preparar jugos y bebidas alcohólicas entre otros (Nyoka, et al., 2015) aunque también se ha reportado el uso de sus hojas para forraje y alimentación del ganado (Kugedera, 2019).

Presenta un alto contenido de aminoácidos esenciales (isoleucina, metionina, cisteína, triptófano y valina), vitamina C y minerales (Hiwilepo-van Hal, et al., 2014) entre otros compuestos. Sus propiedades más destacadas están recogidas en la tabla 3:

Tabla 3. Propiedades de *Sclerocarya birrea*

Propiedad	Fuente	Información adicional
Propiedades antidiabéticas	Dimo, et al., 2007	Se ha reportado actividad hipoglucémica en ratas
	Dieye, et al., 2008	
	Gondwe, et al., 2008	
Modulación de la tasa de filtración glomerular y de la presión arterial media	Garba & Ahmadu, 2006	
Actividades antiplasmodiales y antipalúdicas	Gathirwa, et al., 2008	
Actividad antidiarreica	Belemtougri, et al., 2007	
Efecto tripanocida	Mikail, 2009	Las hojas y la corteza del tallo mostraron una mortalidad completa en *Trypanosoma brucei brucei in vitro*
Actividad antioxidante	Borochov-Neori, et al., 2008	Su jugo posee un alto contenido de flavonoides y compuestos polifenólicos
Efecto antagonista sobre el calcio inducido por cafeína	Musabayane, et al., 2006	
Propiedades anti-inflamatorias	Ojewole, 2003	

- *Vigna subterranea*

Está asociada con la agricultura de subsistencia a pequeña escala. (Tan, et al., 2020). Puede sobrevivir en climas húmedos y suelos pobres en nutrientes (Onimawo, et al., 1998). También se ha reportado que es un cultivo resistente a plagas, enfermedades y sequía capaz de reducir plagas en el campo (Tibe, et al., 2007). Sus semillas son utilizadas principalmente para fabricar harina mientras que su fruto es utilizado como sustituto para el café (Tan, et al., 2020).

Presenta un alto contenido de minerales (K, Ca, Mg, P y Fe), aminoácidos esenciales (lisina y valina), proteínas (Damfami, et al., 2020), glúcidos (son los macronutrientes más abundantes) y fibra dietética (Azman, et al., 2019). Tal y como se puede comprobar en la tabla 4, hasta la fecha, se han reportado pocas propiedades para esta planta.

Tabla 4. Propiedades de *Vigna subterranea*

Propiedad	Fuente	Información adicional
Actividad antioxidante	Harris, et al., 2018	Sintetiza flavonoides (en concreto, antocianinas y polifenoles)
Actividad fijadora de nitrógeno atmosférico	Puozaa, et al., 2017	Forma nódulos con especies del género *Bradyrhizobium*
Actividad antimicrobiana	Udeh, et al., 2020	Se ha reportado actividad bactericida contra bacterias como *Klebsiella pneumoniae* subsp. *pneumoniae* o *Pseudomonas aeruginosa* y actividad fungicida contra levaduras (*Candida albicans*) y mohos (*Aspergillus niger*)

1.3.3. Características genómicas

Hasta ahora, ocho genomas han sido secuenciados y publicados por el consorcio: *Faidherbia albida*, *Lablab purpureus*, *Moringa oleifera*, *Sclerocarya birrea*, *Vigna subterranea* (Chang, et al., 2018), *Solanum aethiopicum* (Song, et al., 2019), *Artocarpus altilis* y *Artocarpus heterophyllus* (Sahu, et al., 2019).

Para este estudio fueron empleados los genomas reportados en Chang, et al., 2018 para las cuatro especies del AOCC seleccionadas. En dicho estudio, se utilizó un análisis de distribución de *k*-mers para estimar el tamaño de los genomas (Tabla 5). La anotación estrucutural de los genomas, y, en particular, la predicción de genes que codifican para proteínas se basó en aproximaciones *ab initio* o *de novo* respaldados en cada caso por datos de expresión basados en RNAseq y en comparaciones con proteínas supuestamente homólogas presentes en las bases de datos (Tabla 5).

Tabla 5. Datos de los genomas secuenciados de las especies

Especies	Tamaño estimado del genoma (Mb)	Tamaño del genoma ensamblado (Mb) y porcentaje	Número de genes codificantes de proteínas predichos
Faidherbia albida	661	654 (98,94%)	28979
Lablab purpureus	423	395 (93,38%)	20946
Sclerocarya birrea	356	331 (92,98%)	18937
Vigna subterranea	550	535 (97,27%)	31707

La contigüidad y calidad de los genomas se determinó a partir de la distribución de longitudes de los fragmentos (*contigs* y *scaffolds*) en los ensamblajes obtenidos, así como medidas de N50, que proporcionan una estimación de la longitud mínima de los contigs o scaffolds, en cada caso, que cubren un 50% del genoma (Tabla 6).

Tabla 6. Datos del ensamblaje de los genomas secuenciados de las especies

Especies	*Contigs* (≥2000 bp)	*Scaffolds* (≥2000 bp)	N50 (*Contigs*)	N50 (*Scaffolds*)	Porcentaje de nucleótidos sin identificar
Faidherbia albida	26459	5758	42029	692039	1,42
Lablab purpureus	15984	4265	21349	335449	2,57
Sclerocarya birrea	22172	4852	64158	4028	2,42
Vigna subterranea	35465	292	19154	640666	4,21

Por su parte, para comprobar la completitud de cada genoma en términos de los genes anotados, se utilizó el programa *Bench-marking Universal Single-Copy Orthologs* (BUSCO) versión 3.0.1 (Seppey, et al., 2019). BUSCO estima la fracción de ortólogos presentes en un genoma problema entre los genes identificados como conservados dentro del clado al que pertenece la especie (*core genes*) (Tabla 7). En este caso, se trata del clado de las embriofitas por lo que la estimación se hizo sobre 1440 genes.

Tabla 7. Completitud de los genomas secuenciados de las especies

Especies	Número y porcentaje de *core genes* detectados (Total: 1440)	Número y porcentaje de *core genes* detectados completos
Faidherbia albida	1315 (91,3%)	1231 (85,5%)
Lablab purpureus	1341 (93,2%)	1258 (87,4%)
Sclerocarya birrea	1384 (96,1%)	1352 (93,9%)
Vigna subterranea	1326 (92,1%)	1244 (86,4%)

Finalmente, las repeticiones de DNA derivadas de la actividad de elementos transponibles fueron identificadas en cada genoma utilizando el programa RepeatMasker versión 4.0.5 (Tarailo-Graovac, et al., 2009) y anotadas mediante la base de datos de secuencias repetitivas RepBase (Tabla 8).

Tabla 8. Proporción de secuencias repetitivas en los genomas secuenciados de las especies

	Porcentaje en genoma			
Secuencia repetitiva	***Faidherbia albida***	***Lablab purpureus***	***Sclerocarya birrea***	***Vigna subterranea***
SINE	<0,01	0,005	0,02	0
LINE	0,91	0,45	0,19	0,25
LTR	44,65	23,78	38,78	19,77
DNA	4	4,76	1,76	7,15
Satélite	0,01	0,02	0	0,01
Repeticiones simples	0,04	0,2	0,04	0,35
Otros	6,48	8,95	5,11	11,94
Total	54,86	37,18	45,18	38,35

2. Objetivos

El presente trabajo se articula en torno a los siguientes cuatro objetivos principales:

2.1. Anotación funcional genómica

El primer objetivo de este trabajo fue el de anotar funcionalmente los genomas de las cuatro especies de cultivo seleccionadas usando ontologías génicas (GO por sus siglas en inglés), rutas bioquímicas y otros procesos celulares disponibles en la Enciclopledia Kyoto de Genes y Genomas (KEGG por sus siglas en inglés), dominios funcionales de proteína presentes en INTERPRO, así como la asignación de los códigos numéricos de enzimas de la *Enzyme Comission* (identificadores EC).

2.2. Clasificación de ortogrupos/familias génicas

Un segundo objetivo fue el de obtener una clasificación de ortogrupos o familias génicas presentes en los genomas de las cuatro especies seleccionadas junto con el de otras siete especies representativas de los distintos linajes de plantas con flores.

2.3. Modelización de la evolución de ortogrupos/familias génicas a escala genómica

La clasificación de ortogrupos obtenida en el objetivo anterior fue usada para modelar las dinámicas evolutivas de recambio génico usando aproximaciones que estiman las tasas de ganancia y pérdida de genes en un entorno de máxima verosimilitud para cada una de las ramas del árbol filogenético que describe las relaciones evolutivas entre las especies a estudio. De esta manera se pueden identificar ortogrupos o familias génicas signficativamente expandidas o contraídas en las ramas correspondientes a las especies a estudio seleccionadas.

2.4. Categorización funcional de los ortogrupos/familias génicas identificadas como expandidas

Finalmente, se utilizará la anotación funcional obtenida en el objetivo 1 para identificar funciones biológicas significativamente sobrerepresentadas entre los genes pertenecientes a las familias génicas identificadas en el objetivo anterior como expandidas en cada una de las cuatro especies a estudio.

3. Materiales y metodología

3.1. Genomas

Para este estudio se utilizaron los genomas de 11 especies vegetales, incluyendo las cuatro especies de cultivo huérfanas seleccionadas así como los de siete especies representativas de los principales linajes de angiospermas: las eudicotiledóneas *Arabidopsis thaliana*, *Vitis vinifera* (uva) y *Moringa oleifera* (otra miembro del AOCC), las monocotiledóneas *Oryza sativa* (arroz) y *Zea mays* (maíz), la magnólida *Persea americana* (aguacate) y el arbusto perennifolio *Amborella trichopoda,* una angiosperma basal. En la tabla 9 aparecen recogidos los genomas seleccionados, así como las versiones utilizadas en cada caso y las abreviaturas asignadas a cada especie.

Tabla 9. Especies y versiones de sus genomas

Especies	Versión del genoma
***Arabidopsis thaliana* (Ath)**	TAIR 10
***Amborella trichopoda* (Atr)**	JGI v1
***Faidherbia albida* (Fal)**	AOCC v1
***Lablab purpureus* (Lpu)**	AOCC v1
***Moringa oleifera* (Mob)**	AOCC v1
***Oryza sativa ssp. Japonica* (Osj)**	JGI v7.0
***Persea americana cv. Hass* (Pah)**	COGE (Genome Id: 29302)
***Sclerocarya birrea* (Sbi)**	AOCC v1
***Vigna subterranea* (Vsu)**	AOCC v1
***Vitis vinifera* (Vvi)**	JGI 12xMarch2010
***Zea mays* (Zma)**	AGP v4.0

Los 11 genomas de las plantas seleccionadas fueron previamente filtrados para i) seleccionar, en cada caso, la isoforma de *splicing* alternativo de mayor longitud si es que hubiera varias, y ii) identificar y eliminar genes con marcos abiertos de lectura incompletos y/o erróneos (por ej., aquellos genes truncados que contenían codones STOP a lo largo de su secuencia) originados probablemente por errores de secuenciación o en la predicción de la estructura de genes durante la anotación estructural del genoma.

Por otra parte, fueron también eliminados aquellos genes que mostraban similitud con elementos transponibles. Dichos genes fueron detectados mediante el algoritmo de búsqueda BLASTN en la base de datos Repbase versión 23.08 (Jurka, et al., 2005). Los parámetros seleccionados fueron los siguientes: valor E < 10^{-5}, *bit score* > 45.

3.2. Clasificación de ortogrupos/familias génicas

El programa de genómica comparada OrthoFinder versión 2.3.3. (Emms, et al., 2019) fue empleado para generar una clasificación de ortogrupos/familias génicas a partir de los genomas de las once especies seleccionadas. OrthoFinder emplea un algoritmo para agrupar genes ortólogos y parálogos en ortogrupos que consta de las siguientes fases:

1. Alineamiento de secuencias de aminoácidos: todas las secuencias proteicas codificadas en cada uno de los genomas representados por ficheros en formato FASTA son comparadas utilizando un algoritmo de alineamiento basado en la doble indexación llamado DIAMOND (Buchfink, et al., 2015). Para no filtrar secuencias muy cortas se empleó umbral conservador (valor E < 10^{-3}) puesto que en las fases posteriores los

falsos positivos serán filtrados utilizando criterios más estrictos y específicos para la construcción de ortogrupos.

2. Normalización de la longitud y la distancia filogenética de los resultados del alineamiento múltiple: en esta fase los resultados del alineamiento múltiple (esto es, la lista de candidatos a homólogos para cada gen) son modelados por cada comparación de pares de especies para identificar y eliminar el sesgo por similitud génica al que están expuestos la longitud de un gen y la distancia filogenética. Esto se hace para que los mejores candidatos a homólogo entre todas las especies reciban las mismas puntuaciones independientemente de la longitud de la secuencia o la distancia filogenética.

3. Delimitación de los umbrales de similitud de secuencia de ortogrupo mediante RBNH: Este paso utiliza información de RBNHs (*Reciprocal Best length-Normalised hits*, mejores candidatos a homólogos recíprocos cuyas longitudes han sido normalizadas) para definir el límite inferior de similitud de secuencia para los genes afines putativos de cada gen problema. Para pasar a la fase 4, un par de genes debe ser un RBNH o tener mejor puntuación que el RBNH con la puntuación más baja (independientemente de la especie) para cualquier gen.

4. Construcción de grafos: los pares de genes afines putativos seleccionados son conectados mediante grafos con ponderaciones dadas por las puntuaciones normalizadas del alineamiento múltiple. Un grafo es un conjunto de ítems llamados vértices o nodos unidos por enlaces llamados aristas o arcos, que permite representar relaciones binarias entre elementos de un conjunto y detectar agrupaciones naturales de ítems que en este caso representan posibles ortogrupos o familias de genes.

5. Agrupamiento de genes en ortogrupos mediante MCL: los grafos generados en la fase anterior son empleados para construir los ortogrupos usando un algoritmo de agrupamiento de grafos denominado *Markov Cluster Algorithm* (MCL). El parámetro de inflación empleado por el algoritmo MCL para definir la granularidad o tamaño de los clústeres fue de 1,5, el valor por defecto.

3.3. Modelización de las dinámicas evolutivas en ortogrupos/familias génicas

La evolución de los ortogrupos resultantes de la clasificación de OrthoFinder fue modelada utilizando el programa Badirate versión 1.35 (Librado, et al., 2012) que emplea diversos modelos estocásticos para estimar las tasas de sustitución o recambio de genes (*turnover rate*) en cada familia génica y especie. Con el objeto de estimar dichas tasas, se ha utilizado en este trabajo el modelo estocástico de GD (*Gain-and-Death*). Bajo este modelo, las tasas de ganancia de genes (*Gain*) incluyen aquellos resultantes de duplicaciones génicas y genómicas (*Birth*), así como la adquisición de nuevos genes por otros mecanismos como la domesticación de elementos transponibles o genes adquiridos por transferencia génica horizontal (*Acquisition*). Por otra parte, las tasas de pérdida de genes (*Death*) modelan la pérdida de genes en una familia génica concreta a lo largo de la evolución (p. ej., pseudogenización).

Badirate usa distintos métodos de estimación de las tasas de sustitución génica: Máxima verosimilitud (*Maximum Likelihood*, ML), Máximo a Posteriori (MAP) y parsimonia. Estas

estimaciones de las tasas de sustitución de genes se realizan en un contexto filogenético, lo que permite testar hipótesis evolutivas concretas. Para ello, el programa utiliza como ficheros de entrada la uno de los ficheros de salida de la clasificación de ortogrupos en las 11 especies (en concreto, el que contiene el número de genes de cada especie asignados a cada ortogrupo) realizada por OrthoFinder y un árbol filogenético ultramétrico representando las relaciones evolutivas entres dichas especies en formato Newick.

Badirate permite testar distintas hipótesis filogenéticas aplicando los siguientes tres modelos competitivos a cada familia génica/ortogrupo:

- Modelo del *Global Ratio*: estima las mismas tasas de *Gain*/*Death* en todas las ramas del árbol. Este modelo encajaría con las familias de genes que se han mantenido estables en términos del número de genes durante la evolución.

- Modelo de *Free Ratio*: estima tasas de *Gain*/*Death* de forma independiente para cada rama del árbol. Es el más adecuado para aquellas familias que han evolucionado estocásticamente.

- Modelo de *Branch Ratio*: estima distintas tasas de *Gain*/*Death* para la rama correspondiente a la especie a estudio y otras para el resto de las ramas en el árbol. Este modelo nos permite identificar ortogrupos o familias génicas expandidas específicamente en una especie.

Para cada modelo se efectuaron cinco iteraciones por familia, donde las estimaciones de las tasas de sustitución fueron calculadas en la primera iteración mediante parsimonia, mientras que en las otras cuatro se calcularon mediante ML. Por tanto, por cada ortogrupo se ejecutaron cinco iteraciones del modelo *Global Ratio*, cinco del modelo *Free Ratio* y cinco iteraciones del modelo *Branch Ratio* para cada una de las cuatro especies a estudio, esto es, un total de 30 iteraciones por ortogrupo. Para cada uno de los modelos se seleccionó la iteración que resultaba en el mejor valor de probabilidad logarítmica (*log-likelihood*). Las mejores iteraciones de cada modelo fueron comparadas entre sí mediante el criterio de información Akaike (AIC) (Akaike, 1992). Si el mejor modelo resultaba en una ratio de AIC 2,7 veces mayor que el segundo mejor modelo, se concluía que aquel era el que mejor se ajustaba de forma significativa al ortogrupo, es decir, era el más adecuado para modelar la evolución de la familia génica correspondiente.

Con el fin de reducir el tiempo de computación se seleccionó un ortogrupo representativo de cada una de las distribuciones en número de genes. Posteriormente, los resultados obtenidos se proyectaron sobre la totalidad de familias con la misma distribución para obtener los resultados globales.

Por último, para calcular el número total de genes ganados en cada linaje o especie se sumaron los genes que ganó cada ortogrupo clasificado como expandido en la rama correspondiente respecto del nodo inmediatamente anterior. Esta operación fue repetida con los genes que perdió cada ortogrupo clasificado como contraído para calcular el número total de genes perdidos en la rama correspondiente respecto del nodo inmediatamente anterior.

3.4. Reconstrucción de un árbol filogenético de especies

Para describir las relaciones evolutivas entre las 11 especies de plantas con flores seleccionadas, se reconstruyó de forma manual un árbol ultramétrico y enraizado utilizando la topología y los tiempos de divergencia entre especies disponibles en la base de datos TimeTree (http://www.timetree.org/) (Kumar, et al., 2017). Para dotar de raíz al árbol, se utilizó como grupo externo (*outgroup*) *A. trichopoda* por su relación evolutiva con los demás miembros del grupo ya que es considerada un ancestro viviente de todas las angioespermas (Albert, et al., 2013). La presencia de un representante de las magnólidas (*P. americana*) entre las 11 especies hizo necesaria una revisión bibliográfica debido a que trabajos anteriores (Chaw, et al., 2019; Moore, et al., 2007; Soltis, et al., 2011; Rendon, et al., 2019) habían planteado hipótesis evolutivas contradictorias con respecto a la posición de las magnólidas dentro de la filogenia de las angiospermas. Finalmente, se utilizó la hipótesis evolutiva propuesta en Rendon, et al., 2019 que favorecía la ubicación taxonómica de las magnólidas como clado hermano al superclado formado por las dicotiledóneas y las monocotiledóneas combinadas en base a los resultados de un análisis de distancias de ortólogos sinténicos entre 14 especies.

3.5. Anotación funcional de genomas

La anotación funcional de un genoma es la etapa en la anotación de un genoma que permite asociar información funcional a cada gen identificado en el genoma para describir distintos aspectos de su identidad biológica (Berardini et al., 2014).

Actualmente, el método más empleado para representar la anotación funcional de los genes que componen un genoma consiste en la utilización de las ontologías génicas (*Gene Ontology*, GO) desarrolladas por el *GO Consortium* (Ashburner et al., 2000). Una ontología génica es un vocabulario controlado y estandarizado de términos numerados (términos GO) organizados en una jerarquía de grafos acíclicos dirigidos que describe las características funcionales del producto de un gen mediante representaciones computacionalmente tratables. Estas características funcionales se agrupan en tres categorías de términos GO:

- Proceso Biológico: describe el proceso (en este contexto proceso se define como un conjunto de actividades moleculares) dentro y/o fuera de la célula en el que participa el producto génico.

- Componente Celular: describe la localización del producto génico dentro de la célula pudiendo ser un orgánulo o una estructura macromolecular.

- Función Molecular: describe la actividad que el producto génico ejecuta a nivel molecular.

Cada término GO tiene asignado un código de evidencia (*evidence code*) que provee información sobre cómo se obtuvo, incluyendo las siguientes fuentes posibles: evidencia experimental, análisis computacional, evidencia filogenética, dictaminado por el autor o por declaración curatorial.

Además de las anotaciones GO genéricas, existen versiones comprimidas específicas de organismos o grupo de organismos concretos que reciben el nombre de GO-slim. GO-slim ofrece una visión general del rango de funciones y procesos que se encuentran codifcados en el genoma de un organismo o de un grupo taxonómico determinados, proporcionado un sentido

general a funciones biológicas clave para los mismos (Gene Ontology Resource, 2020). Una de estas subversiones del GO-slim es el Plant GO Slim, desarrollado por el TAIR (The Arabidopsis Information Resource) y el Consorcio GO, diseñado específicamente para plantas (TAIR, 2020).

El uso más común de las anotaciones GO es la interpretación de experimentos de biología molecular a gran escala, a veces llamados experimentos "ómicos", y en particular genómica, aproximando de esta manera la estructura, función y dinámica de un organismo. La anotación funcional de los genomas de las cuatro especies objeto de este estudio con términos GO genéricos y del Plant GO Slim se llevó a cabo mediante BLAST2GO versión 5.2.5. (Conesa, et al., 2005), un programa integrado en OmicsBox, un paquete bioinformático que integra distintas herramientas destinadas al análisis de genomas, transcriptomas y metagenomas (OmicsBox, 2019). Se describen a continuación los distintos pasos que comprenden la anotación funcional de genomas con BLAST2GO.

3.5.1. Identificación de posibles homólogos mediante alineamiento local de secuencias proteicas

Para identificar secuencias que muestran un porcentaje de identidad significativa que nos permita inferir homología o evolución a partir de una secuencia ancestral común, se efectuaron búsquedas en las bases de dato de *non-redundant* (nr) del *National Center for Biotechnology Information* (NCBI) usando como *queries* las secuencias correspondientes a cada una de las proteínas codificadas en los genomas de las cuatro especies seleccionadas . El resultado es una lista de homólogos putativos (*hits*) a cada uno de los genes presentes en los genomas de las plantas del AOCC seleccionadas. Solo fueron aceptados aquellos *hits* con un valor $E \leq 1e^{-10}$, fijándose 20 como número máximo de *hits*.

Para este paso del protocolo BLAST2GO emplea comúnmente el algoritmo BLAST. Sin embargo, para ahorrar tiempo computacional, en este trabajo se usó el programa DIAMOND versión 0.9.36.137 (Buchfink, et al., 2015) el cual ha demostrado ser hasta 20000 veces más rápido que BLAST y tener un grado de sensibilidad similar.

3.5.2. Mapeo

A continuación, se ejecuta el proceso de mapeo para el cual fueron seleccionados los parámetros por defecto. Este es el proceso de extracción de términos GO asociados a los resultados (es decir, la lista de *hits*) obtenidos tras la búsqueda con DIAMOND. BLAST2GO ejecuta cuatro operaciones durante este proceso para obtener la mayor cantidad de información GO posible para cada *hit*:

1. Las accesiones de los resultados de la búsqueda de DIAMOND son utilizados para recuperar nombres de genes o símbolos haciendo uso de dos archivos de mapeo proporcionados por el NCBI (*gene info*, *gene2accession*). A continuación, los nombres de los genes identificados son buscados en aquellas entradas de la tabla de productos génicos de la base de datos del GO que sean específicas de la especie de interés.

2. Los identificadores GI de un resultado o *hit* (esto es, una serie de dígitos que se asignan consecutivamente a cada registro de secuencia procesado por NCBI) de la búsqueda de DIAMOND son empleados para recuperar identificadores de UniProt mediante un

archivo de mapeo del PIR, una base de datos de proteínas de referencia no redundantes que incluye PSD, UniProt, SwissProt, TrEMBL, RefSeq, GenPept y PDB.

3. Las accesiones son buscadas directamente en la tabla dbxref de la base de datos del GO. */db_xref* es un calificador que sirve como vehículo para vincular los registros de secuencias de ADN a otras bases de datos externas.

4. Las accesiones de los resultados de la búsqueda de DIAMOND son buscadas directamente en la tabla de productos génicos de la base de datos GO.

3.5.3. Anotación

Finalmente, cada uno de los genes en los cuatro genomas fue anotado mediante la asignación de términos funcionales a las secuencias problema a partir de la lista de términos GO obtenidos en el mapeo usando los parámetros por defecto. La asignación de la función depende del vocabulario de la ontología génica. El algoritmo de anotación de BLAST2GO tiene en cuenta la similitud entre las secuencias problema y las obtenidas, la calidad de la fuente de evidencia de los términos GO y la estructura del grafo acíclico dirigido. La anotación GO se realiza por una regla de anotación (*annotation rule*, AR) obtenida en los términos de ontología. Esta regla busca la anotación más específica con un cierto nivel de fiabilidad. Para cada candidato GO, se le asigna una puntuación de anotación (*annotation score*, AS), la cual está compuesta por dos términos:

- Término directo (*direct term*, DT): representa la mayor similitud de su GO determinado por un factor que se corresponde a su código de evidencia (*evidence code*, EC).

- Término de posibilidad de abstracción (AT): se define como la anotación de un nódulo parental cuando varios nodos hijos se encuentran presentes en la colección de términos GO candidatos. Este término multiplica el número total de GOs unificados en el nodo por un valor ponderado de GO, definido por el usuario, que controla la posibilidad y la fuerza de abstracción.

Por último, la AR selecciona el menor término por rama que se encuentra bajo un determinado umbral.

3.5.4. Ampliación de la anotación funcional

Para confirmar y eventualmente completar la anotación funcional de los genomas problema, las siguientes funcionalidades integradas en BLAST2GO fueron empleadas:

- InterProScan: es un programa que permite identificar los dominios funcionales de la base de datos del *InterPro Consortium* (InterPro, 2020) presentes en una proteína a partir de su secuencia, así como su agrupación en clasificaciones precomputadas de familias y superfamilias génicas. Cada uno de los dominios funcionales presentes en InterPro tienen asociados términos GO, los cuáles son transferidos a BLAST2GO para ser fusionados con las anotaciones GO derivadas del alineamiento (Jones, et al., 2014).

- *Kyoto Encyclopedia of Genes and Genomes* (KEGG): es una base de datos de rutas bioquímicas y otros procesos celulares que permite asignar códigos de enzima (*Enzyme Code*, EC) a aquellas proteínas con actividad catalítica (Kanehisa, et al., 2018).

- *Annotation Expander* (ANNEX): es un programa desarrollado por *Gene Ontology Annotation Toolbox* (GOAT) que permite extender las anotaciones GO. Emplea el Concepto de la Segunda Capa (Myhre, et al., 2006) que consiste en utilizar relaciones uni-vocales entre los términos GO de las diferentes categorías GO para agregar anotaciones implícitas.

3.5.5. Análisis de enriquecimiento funcional de listas de genes

Para detectar términos GO (genéricos o del Plant GO Slim) sobrerrepresentados e infrarrepresentados en una lista de genes se realizaron análisis de enriquecimiento funcional mediante el test exacto de Fisher (Fisher, 1922). Este test permite evaluar la significación estadística de las hipotéticas distribuciones diferenciales de términos GO entre los genes pertenecientes a las familias identificadas a través del análisis de Badirate como expandidas o contraídas con respecto al total de genes contenidos en un genoma, permitiendo así la identificación de funciones biológicas favorecidas y desfavorecidas por la evolución en una especie dada.

Dado que los análisis de enriquecimiento funcional con términos GO implican realizar tantos tests exactos de Fisher como términos GO individuales haya presentes en la anotación de un genoma dado, se necesita introducir una corrección para el testeo de hipótesis múltiples de los valores p resultantes que permita controlar la tasa de falsos positivos, esto es, tests que resulten en valores de p significativos por error. Se consideró estadísticamente significativo el enriquecimiento (o empobrecimiento) de un término GO en una lista de genes cuando el test de Fisher correspondiente resultaba en un valor $p \leq 0{,}05$ después de corrección por el llamado método de Bonferroni (Bonferroni, 1936).

Por último, para visualizar los términos GO sobre e infrarrepresentados detectados en los dos conjuntos se emplearon diagramas de burbujas hechos por REVIGO (Supek, et al., 2011), un servidor web (http://revigo.irb.hr/) para la visualización y reducción de listas de términos GO.

La principal ventaja de utilizar REVIGO es la simplificación de largas listas de términos GO cercanos dentro de la jerarquía GO (términos hermanos) o que están relacionados por herencia (términos hijos y parentales). Estas listas redundantes son difíciles de interpretar por lo que REVIGO realiza un procedimiento de agrupamiento simple para identificar un término GO representativo de cada grupo utilizando como criterio de selección una lista de valores p proveída por el usuario. Este procedimiento es en un plano conceptual similar a los métodos de agrupamiento jerárquico (aglomerativo) como el método de unión de vecinos (*Neighbor-Joining Method*).

La longitud de la lista resultante podrá ser fijada por el usuario seleccionando un valor umbral de similitud semántica entre términos representativos. Para calcular el valor de similitud semántica entre cada pareja de términos se utilizará una medida de similitud semántica también elegida por el usuario (REVIGO ofrece las siguientes: SimRel, Lin, Resnik, Jiang and Conrath) quien también puede elegir el método para calcular la frecuencia de cada término que consiste en extraer el porcentaje de genes anotados con el término en una especie de la base de datos UniProt o la base de datos entera (la elección por defecto). En este caso, fueron selecionados el valor y la medida que emplea el programa por defecto (0,7 y SimRel) y la base de datos entera.

3.6. Representaciones gráficas

Las gráficas generadas para visualizar los resultados de los experimentos de este trabajo fueron elaboradas mediante el lenguaje de programación R versión 2020 (R Core Team, 2020) empleando el entorno de desarrollo integrado (o IDE por sus siglas en inglés) RStudio versión 2020 (RStudio Team, 2020) para redactar y editar el código. En concreto, fueron empleadas las siguientes extenciones de R (también llamadas paquetes):

- readxl versión 1.3.1. (Wickham & Bryan 2019): este paquete permitió cargar los archivos Excel que contienen los datos de los experimentos.

- tidyverse (Wickham, et al., 2019): la mayoría de las gráficas (execepto la ***figura 5b***) fueron generadas a partir de este paquete que consta a su vez de otros paquetes de los cuales los únicos utilizados fueron ggplot2 versión 3.3.3. (Wickham, 2016) y tydir versión 1.1.0. (Wickham H. & Henry L. 2020).

- ggpubr versión 0.4.0. (Alboukadel, 2020): la mayoría de las gráficas fueron generadas usando este paquete junto con ggplot2 con las excepciones de las ***figuras 2*** y ***5***.

- RColorBrewer versión 1.1-2. (Neuwirth, 2014): las ***figuras 3, 4, 7*** y ***8*** fueron generadas usando este paquete junto con ggplot2 y ggpubr.

- scales versión 1.1.1. (Wickham & Seidel, 2020): las ***figuras 9, 10*** y ***11*** fueron generadas usando este paquete junto con ggplot2 y ggpubr.

- UpSetR versión 1.4.0. (Conway, et al., 2017): la ***figura 5b*** fue generada a partir de este paquete.

Aparte, para elaborar la ***figura 6*** (esto es, la figura correspondiente al árbol filogenético) se utilizó el programa Figtree versión 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) mientras que para los diagramas de burbujas (***figuras 9, 10*** y ***11***) se utilizaron códigos de R proporcionados por REVIGO que fueron modificados.

4. Resultados y discusión

4.1. Filtrado previo de los genomas

El número genes presentes en los genomas de las 11 especies de plantas seleccionadas era de 306170. Después de seleccionar como representante para cada gen la isoformas de *splicing* alternativo más larga, de eliminar genes incorrectamente predichos o conteniendo errores de secuenciación, así como de descartar aquellos que mostraban identidad de secuencia con elementos repetitivos, este número se redujo a 295716 genes. La distribución del número de genes en cada uno de los 11 genomas antes y después del flitrado se muestra en la ***figura 2***, observándose reducciones considerables en los genomas de *Z. mays* y *O. sativa* y reducciones mínimas en *F. albida*, *L. purpureus*, *P. americana* y *S. birrea* (menos de 100 genes fueron eliminados).

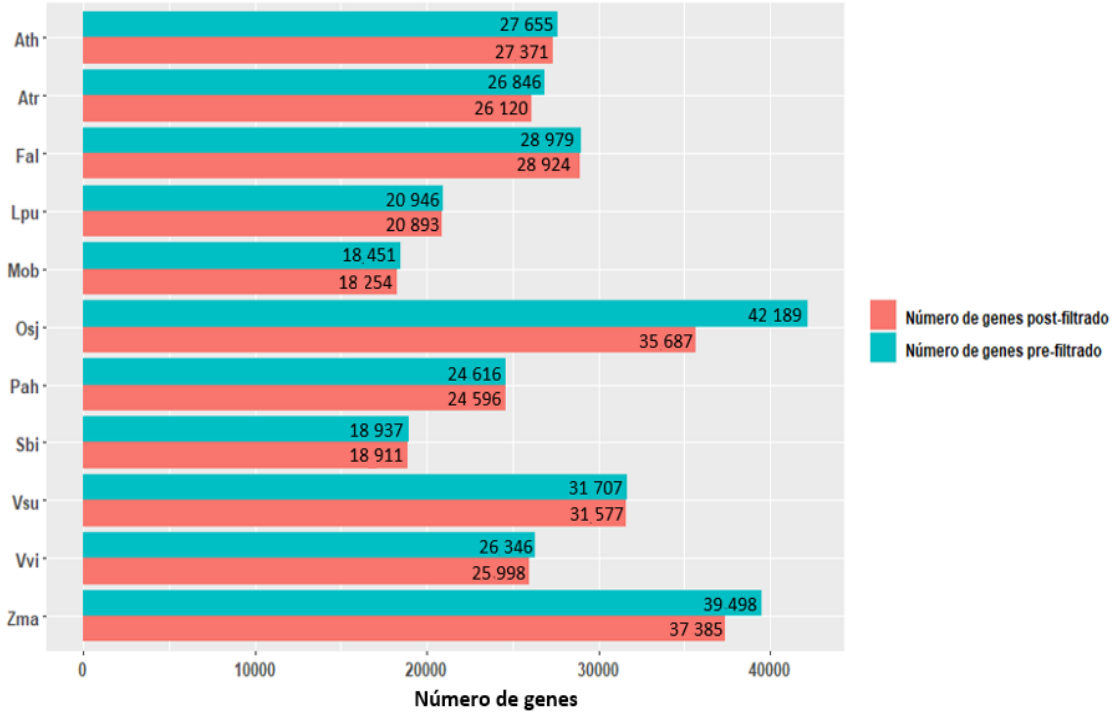

***Figura 2*. Número de genes presentes en los genomas de las 11 especies antes y después del filtrado**

4.2. Anotación funcional de genomas

Con el objeto de obtener la anotación funcional completa de los genomas de las cuatro especies del AOCC seleccionadas para este estudio con términos GO se empleó el programa BLAST2GO (apartado 3.5). Los resultados generales fueron los siguientes:

- Un total de 315830 términos GO genéricos fueron asignados a 68433 genes siendo *V. subterranea* la especie con mayor número de términos asignados (más de 106000 términos) y *S. birrea* la especie con menor número de términos asignados (menos de 30000) (***figura 3a***). Por consiguiente, también son las especies con mayor y menor número de genes anotados respectivamnte (más de 200000 y menos de 12000) (***figura 3b***).

- Las anotaciones GO genéricas cubren más del 50% de los genes de cada especie, superando el 75% en los casos de *V. subterranea* y *L. pupureus* (77 y 81%, respectivamente) (***figura 3c***), observándose una media de genes anotados que oscila entre 2,5 y 6,08 según la especie.

- Un total de 80986 códigos EC fueron asignados a 17909 genes, siendo *V. subterranea* la especie con mayor número de códigos asignados (cerca de 300000) y *S. birrea* la especie con menor número de términos asignados (menos de 10000) (***figura 3d***). Por consiguiente, también son las especies con mayor y menor número de enzimas anotadas respectivamente (más de 7500 y menos de de 2100) (***figura 3e***).

- Las anotaciones EC cubren entre el 10 y el 25% de los genomas de las plantas (***figura 3f***), observándose una media una media de genes anotados que oscila entre 3,28 y 7,88 según la especie.

- Un total de 411776 términos de InterPro fueron asignados 78951 genes, siendo *V. subterranea* la especie con mayor número de términos asignados (más de 118525 términos) y *S. birrea* la especie con menor número de términos asignados (menos de 89000) (***figura 3g***). Por consiguiente, también son las especies con mayor y menor número de genes anotados respectivamente (más de 23000 y menos de de 16100) (***figura 3h***).

- Las anotaciones InterPro cubren más del 75% de los genomas de las plantas, superando el 80% en el caso de *S. birrea* (85%) (***figura 3f***), observándose una media de genes anotados que oscila entre 5 y 5,5 según la especie.

- Las anotaciones resultantes de ejecutar ANNEX supusieron más de 5000 anotaciones extra para cada especie.

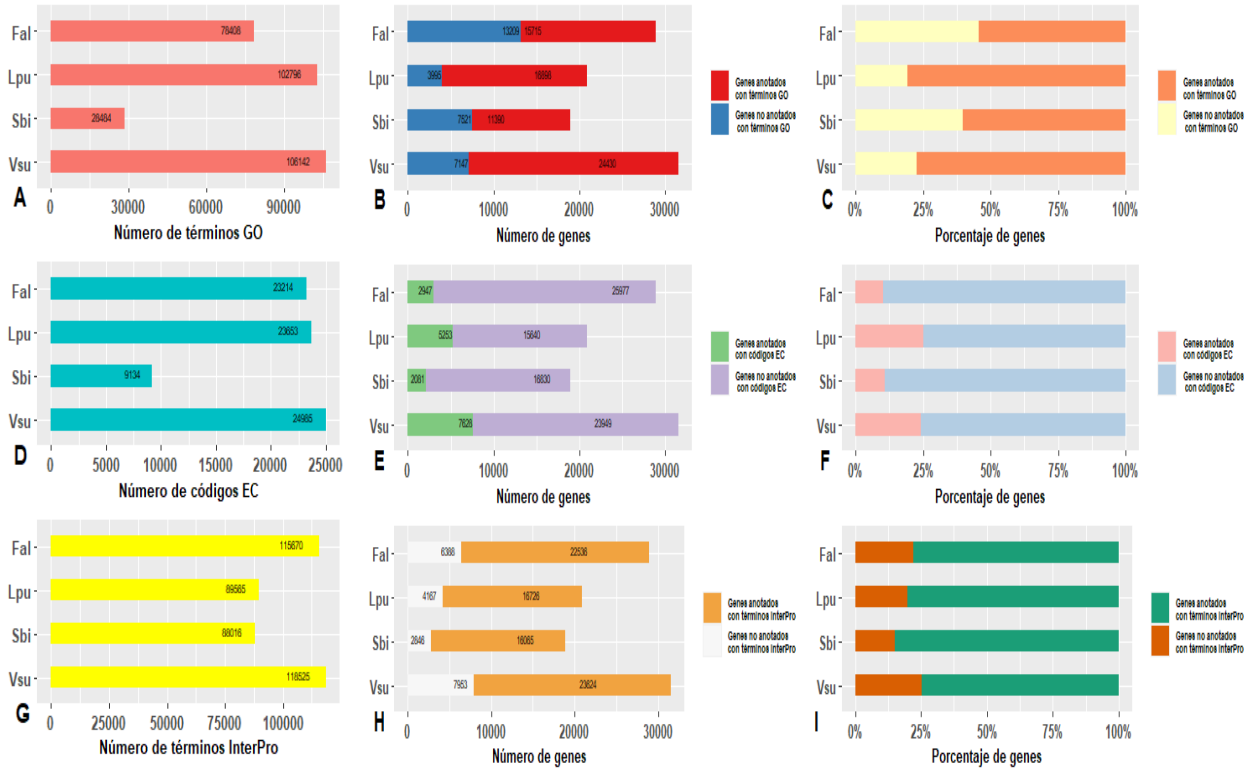

Figura 3. **Estadísticas por especie de la anotación funcional.** *A. Histograma representando por especie el número de términos GO asignados. B. Histograma apilado representando por especie el número de genes anotados con términos GO y no anotados. C. Histograma apilado representando por especie en porcentajes los genes anotados con términos GO y no anotados. D. Histograma representando por especie el número de códigos EC asignados. E. Histograma apilado representando por especie el número de genes anotados con códigos EC y no anotados. F. Histograma apilado representando por especie en porcentajes los genes anotados con términos EC y no anotados. G. Histograma representando por especie el número de términos InterPro asignados. H. Histograma apilado representando por especie el número de genes anotados con términos InterPro y no anotados. I. Histograma apilado representando por especie en porcentajes los genes anotados con términos InterPro y no anotados.*

4.3. Clasificación en ortogrupos/familias génicas en los genomas de 11 especies de plantas

La clasificación por ortogrupos en los genomas de las 11 especies de plantas generada por el protocolo de OrthoFinder versión 2.3.3 arrojó los siguientes resultados generales:

- Los 295716 genes totales presentes en los 11 genomas fueron agrupados en 78345 ortogrupos, de los cuales 17998 (aproximadamente el 23% del total) contenían dos o más genes, agrupando un total de 235369 genes (aproximadamente el 79,6% del total) (***figura 4a***).

- Los restantes 60347 ortogrupos (77% del total de ortogrupos y 20,4% del total de genes, esto es, 60347) estaban formados por un único miembro que el programa no pudo agrupar con otros, esto es, corresponden a genes de copia única específicos de especie, los llamados genes huérfanos que no tienen homólogos dentro de la misma especie (parálogos) ni en otras especies (ortólogos).

- El 50% de todos los genes fueron asignados a ortogrupos de 14 o más miembros. Dichos genes están repartidos en 5911 ortogrupos.

- 5591 ortogrupos que contienen al menos un gen de cada especie fueron generados (***figura 5b***). 677 de estos ortogrupos contenían únicamente genes de una sola copia.

Con respecto a las cuatro especies del AOCC seleccionadas para este estudio, más del 80% de sus genes fueron asignados a algún ortogrupo, siendo *L. purpureus* la especie con mayor porcentaje de genes asignados en ortogrupos (93,2%) de las cuatro y la segunda entre las 11 con mayor porcentaje de genes asignados en ortogrupos, detrás de *M. oleifera* (***figura 4b***). Asimismo, entre el 40 y aprox. el 44% de los genes en ortogrupos se distribuyen a razón de uno por ortogrupo (***figura 5a***). Por otro lado, entre aprox. el 7 y el 18% de los genes de cada planta se encuentran agrupados en 1427 a 5699 ortogrupos compuestos por genes huérfanos de copia única (***figura 4a***). Por otra parte, entre 7 y 44 ortogrupos, los cuales incluyen entre 6 y 202 de los genes (0,1 a 0,7% del total), estaban formados por múltiples copias de genes específicos de cada una de las cuatro especies, es decir, no contenían ortólogos en las especies restantes (***figura 4c***). El origen de los genes huérfanos puede deberse a una de las siguientes causas:

I. Son genes *de novo*, esto es, se originaron en regiones genómicas que no contenían previamente ningún gen o región codificante.
II. Son genes resultantes de la "domesticación" de elementos transponibles.
III. Son genes xenólogos, es decir, se originaron a partir de eventos de transferencia génica horizontal.

Duplicaciones subsiguientes en genes huérfanos podrían resultar en algunos de los ortogrupos específicos de especie o grupos de especies observados como los 211 ortogrupos exclusivos de las especies leguminosas (***figura 5b***) de los cuales 65 fueron detectados como expandidos en este estudio. El posible origen de estos genes podría situarse en la divergencia del ancestro común de todas las leguminosas.

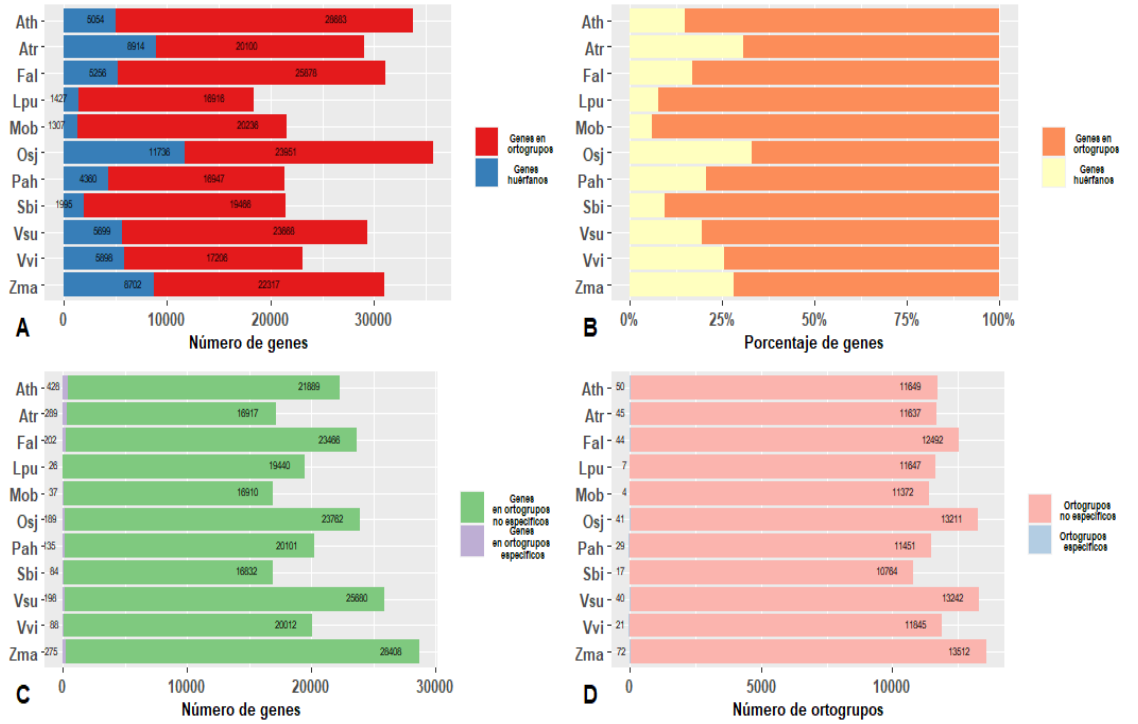

Figura 4. **Estadísticas por especie de la clasificación de OrthoFinder** *A. Histograma apilado representando por especie el número de genes que fueron asignados a ortogrupos y los que no fueron asignados y constituyen familias de un solo miembro (genes huérfanos). B. Histograma apilado representando por especie el porcentaje de genes que fueron asignados a ortogrupos y los que no fueron asignados y constituyen familias de un solo miembro (genes huérfanos). C. Histograma apilado representando por especie el número de genes que fueron asignados a ortogrupos específicos y no específicos. D. Histograma apilado representando por especie el número de ortogrupos formados por genes de dos o más especies y aquellos formados únicamente por genes de una especie.*

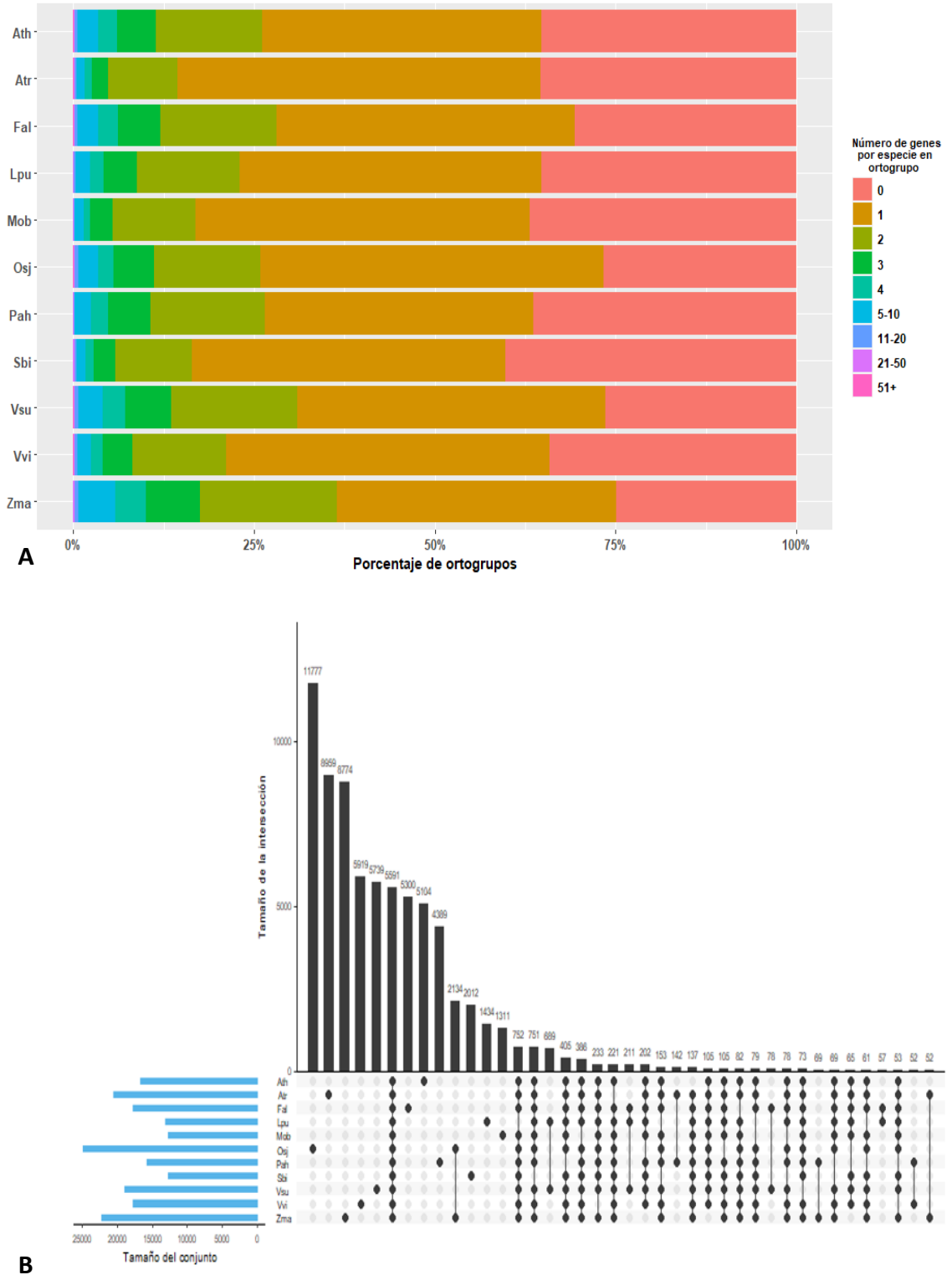

Figura 5. **Resultados del análisis de OrthoFinder.** *A. Histograma apilado representando en porcentajes la distribución de ortogrupos por número de genes y por especie. B. Representación del número de ortogrupos específicos de especie y compartidos por cada combinación de especies. El tamaño de la intersección es el número de ortogrupos compartidos y los puntos negros en el eje de las ordenadas indican qué plantas están presentes en cada intersección. Por ejemplo, la sexta barra muestra que 5591 ortogrupos contienen genes de todas las plantas. Adjunto a la izquierda está un histograma donde está representado el tamaño del conjunto, esto es, el número total de ortogrupos presentes en cada especie.*

4.4. Identificación de ortogrupos/familias génicas expandidos y contraídos en cuatro especies del AOCC

4.4.1. Árbol de especies

En la ***figura 6*** aparece representado el árbol filogenético utilizado para describir las relaciones evolutivas entre las 11 especies e identificar mediante el análisis de Badirate las familias expandidas y contraídas en cada una de las cuatro especies AOCC seleccionadas para este estudio. Como era previsible, las leguminosas fueron agrupadas en el mismo clado tal y como está señalado en la figura.

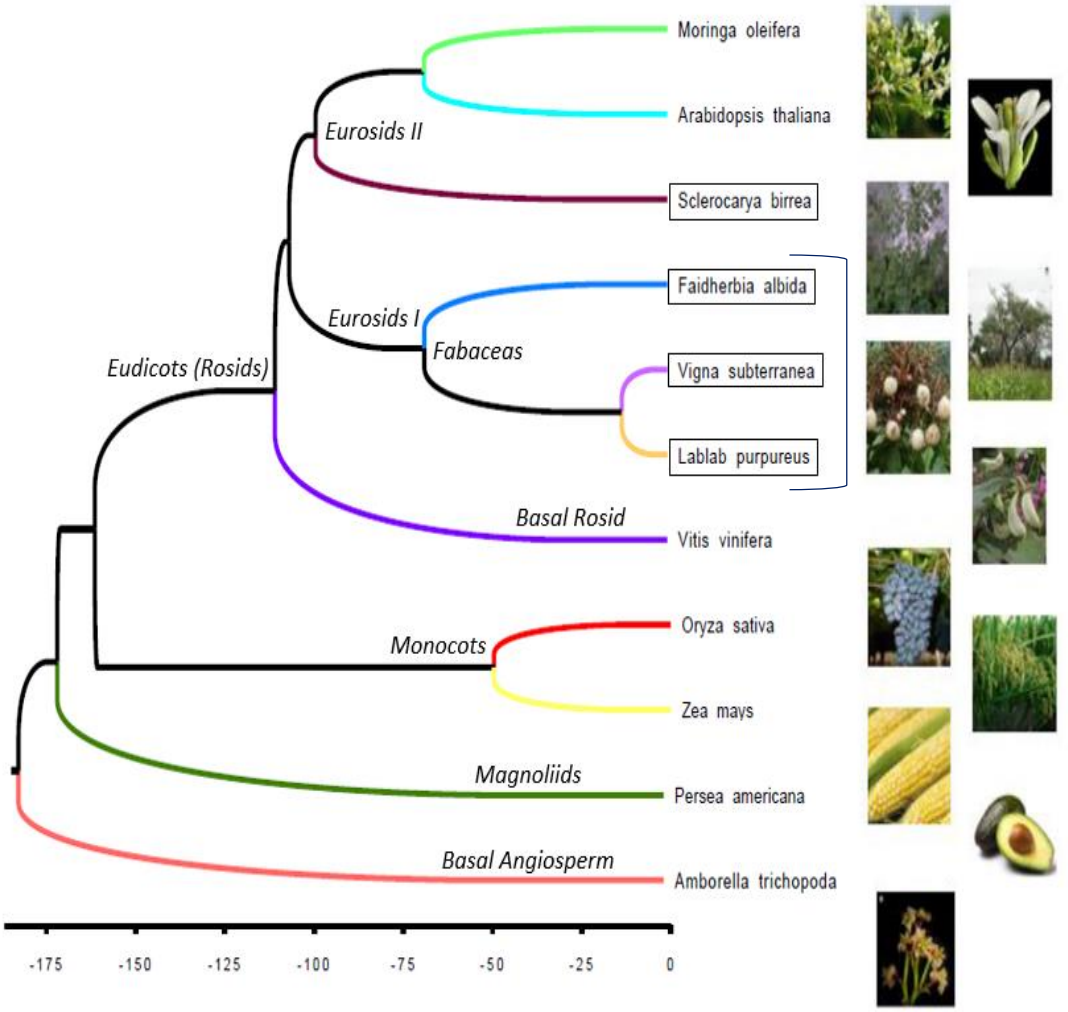

***Figura 6*. Árbol filogenético ultramétrico de las 11 especies utilizado para el análisis de Badirate.** *Sobre los nodos y/o ramas correspondientes aparecen indicados en cursiva los principales clados de angioespermas representados en el árbol. Dentro de los clados, el de las plantas leguminosas (Fabaceas) está destacado. Las longitudes de las ramas reflejan el tiempo evolutivo en millones de años.*

4.4.2. Resultados generales del análisis de Badirate

Con la finalidad de identificar familias significativamente expandidas o contraídas en cada una de las cuatro especies de plantas AOCC seleccionadas para este estudio, se llevó a cabo la modelización de la evolución de los ortogrupos formados por 11 especies de plantas a través de la estimación mediante el programa BadiRate (apartado 3.3) de sus tasas de recambio o sustitución génica en un entorno de máxima verosimilitud. Los resultados generales fueron los siguientes:

- *F. albida* y *V. subterranea* presentan un número mayor de ortogrupos expandidos que contraídos. Por su parte, *V. subterranea* y *L. purpureus* presentan el mayor número de ortogrupos expandidos y contraídos respectivamente (más de 1000 en cada caso) (***figura 7a***).

- *F. albida* y *V. subterranea* también son las únicas especies que presentan un número neto de genes ganados superior a 1000, destacando *V. subterranea* con más de 3000 genes ganados desde su divergencia del último ancestro común con *L. purpureus*. En cambio, *L. purpureus* y *S. birrea* presentan un número neto de genes perdidos superior a 500, considerablemente superior, quizá explicando, al menos parcialmente, la reducción en el número total de genes presentes en sus genomas con respecto a las especies restantes del análisis comparativo (son los genomas más pequeños del grupo junto con el de *M. oleifera*). A su vez, *L. purpureus* es la única especie que presenta un número neto de genes perdidos (más de 1000) mayor que ganados (***figura 7b***).

- *L. purpureus* y *S. birrea* presentan un número de genes en ortogrupos contraídos superior a 10. En el caso de de *L. purpureus*, se tratan de 69 genes en 1023 familias contraídas, un número considerablemente alto. Por otro lado, también presentan un número de genes en ortogrupos expandidos inferior a 1000. Cabe destacar que *V. subterranea*, caracterizada por tener el mayor número de ortogrupos expandidos, presenta también el mayor número de genes en ortogrupos expandidos (más de 4000) (***figura 7c***).

  Además, cuando se compara el número de genes en ortogrupos expandidos con el total de genes presentes en el genoma se observan proporciones que oscilan entre el 4,6 y el 10,02% con la excepción de *V. subterranea* donde representa cerca del 20% de su genoma (***figura 7d***).

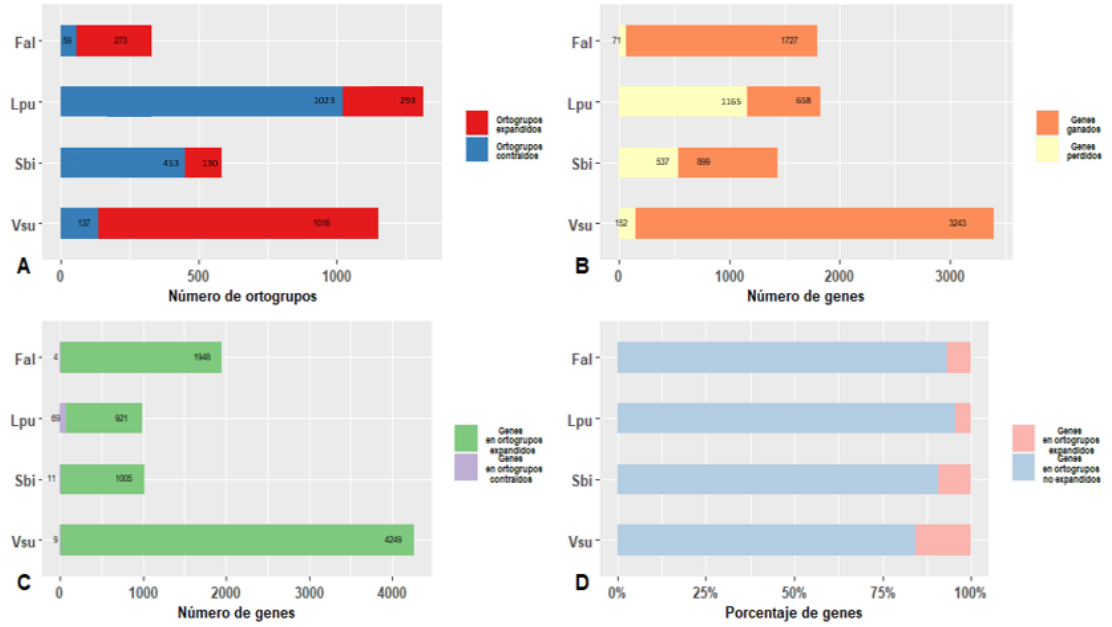

***Figura 7*. Estadísticas por especie del análisis de Badirate.** *A. Histograma apilado representando por especie el número ortogrupos identificados como expandidos y contraídos. B. Histograma apilado representando por especie el número de genes ganados y perdidos. C. Histograma apilado representando por especie el número de genes en ortogrupos identificados como expandidos y contraídos. D. Histograma apilado representando el porcentaje de los genes presentes en ortogrupos identificados como expandidos sobre el total del genoma.*

4.5. Categorización funcional de los ortogrupos/familias génicas expandidos

4.5.1. Resultados de los tests de enriquecimiento funcional

Con el propósito de identificar funciones biológicas significativamente sobrerrepresentadas (o infrarrepresentadas) entre los genes pertenecientes a las familias expandidas en cada una de las cuatro especies, se llevaron a cabo análisis de enriquecimiento funcional de listas de genes mediante tests de Fisher (apartado 3.5.6). Los resultados generales fueron los siguientes:

- De los 18992 términos GO genéricos y los 16884 términos GO Plant Slim totales asignados a 5214 de los 8123 genes contenidos en ortogrupos expandidos en cada una de las cuatro especies, 145 términos GO genéricos y 32 términos GO Plant Slim fueron identificados como sobrerrepresentados. En cuanto a términos infrarrepresentados, 5 términos GO genéricos y 30 términos GO Plant Sim fueron identificados como infrarrepresentados (***figura 8***).

- *F. albida* y *S. birrea* son las especies con mayor número de términos GO genéricos y Plant Slim sobrerrepresentados detectados respectivamente (más de 50 y 10 en cada caso) y *L. purpureus* la especie con menor número (menos de 10 y 3) (***figura 8a***).

- *S. birrea* y *F. albida* las especies con mayor número de términos GO genéricos y GO Plant Slimi nfrarrepresentados detectados respectivamente (3 y más de 10 en cada caso) y L. purpureus la especie con menor número (0 y 1 en cada caso) (***figura 8b***).

- Mientras que todas las especies presentan más términos GO genéricos sobrerrepresentados que infrarrepresentados, solo *F. albida* presenta un menor número de términos GO Plant Slim sobrerrepresentados que infrarrepresentados (***figura 8***).

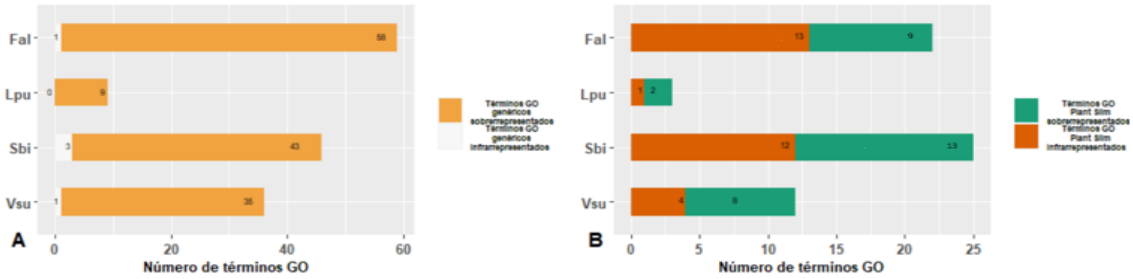

***Figura 8*. Estadísticas por especie de los resultados de los tests de enriquecimiento funcional.** *A. Histograma apilado representando por especie el número de términos GO genéricos sobre e infrarrepresentados. B. Histograma apilado representando por especie el número de términos GO Plant Slim sobre e infrarrepresentados.*

A continuación, los diagramas de burbujas generados por REVIGO para visualizar los términos GO más relevantes de ambos conjuntos en cada especie asignados a los ortogrupos expandidos. En estos diagramas los términos GO no redundantes de una lista son representados por burbujas en un gráfico bidimensional que representa un espacio semántico donde los términos GO pueden aparecer agrupados en clústeres según sus similitudes. El color de cada burbuja indica un valor p asociado (por tanto, en la leyenda del gráfico encontraremos una escala de colores para poder estimarlo) que en este caso proviene del test de enriquecimiento correspondiente y su tamaño indica la frecuencia del término GO en la base de datos GOA (por consiguiente, las burbujas más grandes representan los términos más generales).

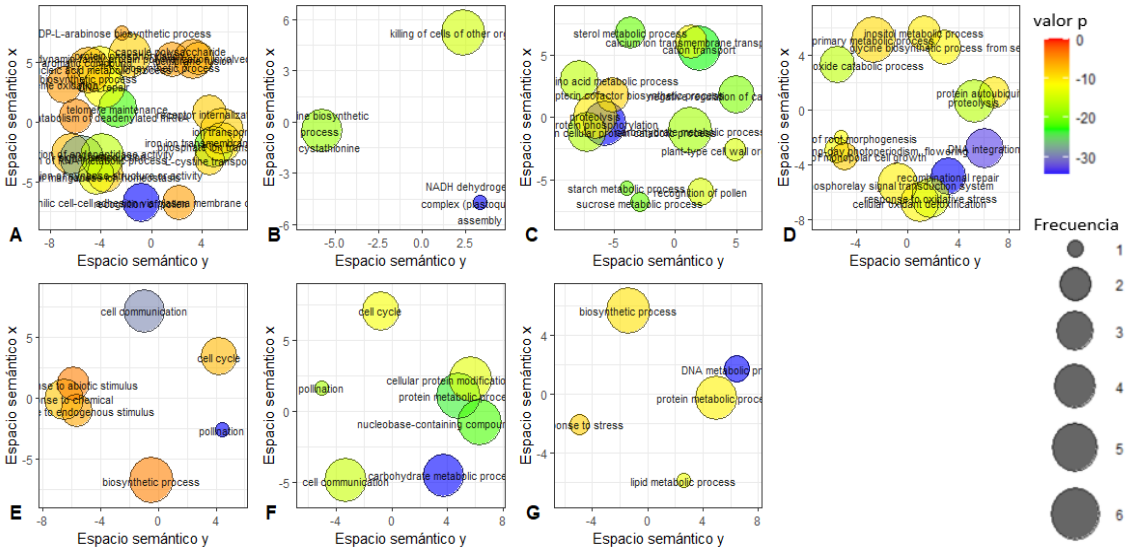

***Figura 9*. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Proceso Biológico asignados a las familias génicas expandidas de cada especie.** *En los paneles superiores están representados los términos GO genéricos mientras que en los paneles inferiores están representados los términos GO Plant Slim. A/E.* Faidherbia albida. *B.* Lablab purpureus. *C/F.* Sclerocarya birrea. *D/G.* Vigna subterranea.

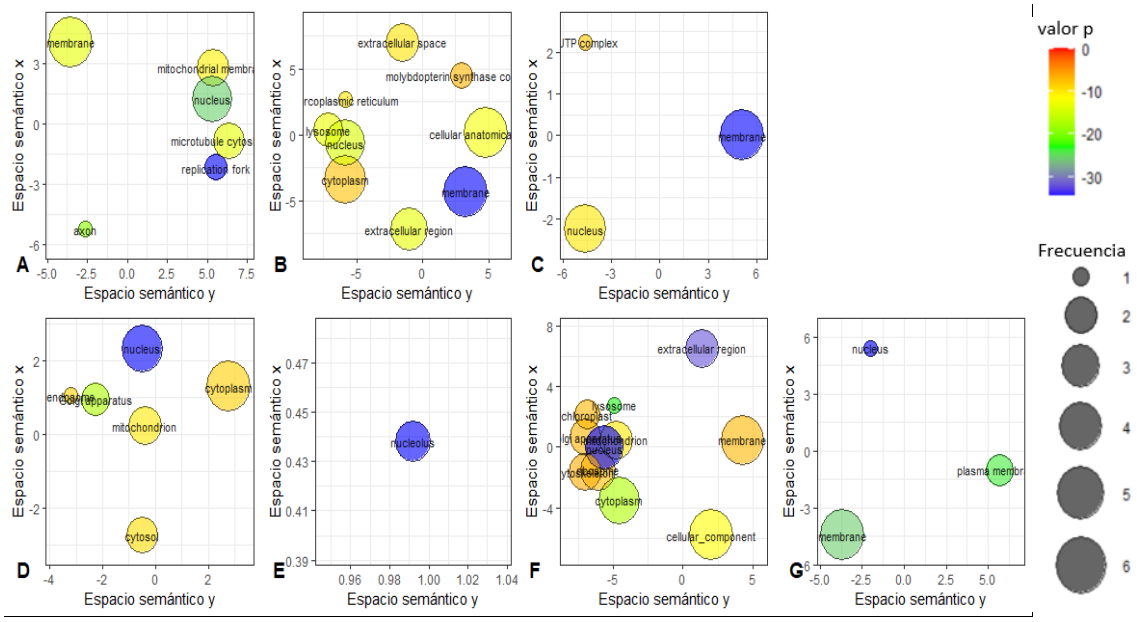

***Figura 10*. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Componente Celular asignados a las familias génicas expandidas de cada especie.** *En los paneles superiores están representados los términos GO genéricos mientras que en los paneles inferiores están representados los términos GO Plant Slim. A/D.* Faidherbia albida*. B/F.* Sclerocarya birrea*. C/G.* Vigna subterranea*. E.* Lablab purpureus*.*

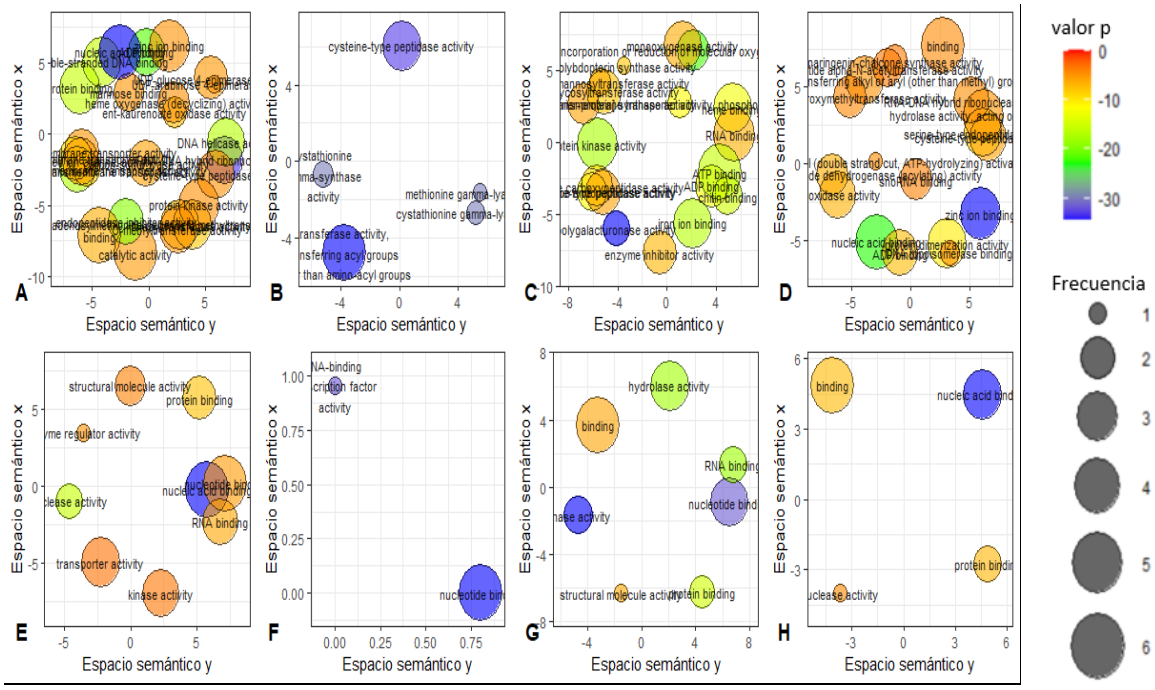

***Figura 11*. Diagramas de burbujas representando los términos GO sobre e infrarrepresentados de la categoría Función Molecular asignados a las familias génicas expandidas de cada especie.** *En los paneles superiores están representados los términos GO genéricos mientras que en los paneles inferiores están representados los términos GO Plant Slim. A/E.* Faidherbia albida*. B/F.* Lablab purpureus*. C/G.* Sclerocarya birrea*. D/H.* Vigna subterranea*.*

4.5.2. Funciones biológicas enriquecidas identificadas en las familias expandidas

En cuanto a las funciones biológicas identificadas como sobrerrepresentadas entre las familias expandidas en cada especie, algunas de las más destacadas estaban relacionadas con distintos aspectos de la nodulación en el caso de las tres especies de leguminosas y de forma genérica con la resistencia y tolerancia a distintos estreses bióticos y abióticos con potencial interés agronómico.

- *Faidherbia albida*

-Nodulación: se han detectado expansiones en tres familias génicas de enzimas de las cuales dos (OG0012003 y OG0012753) están anotadas como 2'-O-metiltransferasa o chalcona-O-metiltransferasa y la otra (OG0013671) está anotada como isoflavona 7-O- metiltransferasa. Las dos primeras son familias de enzimas exclusivas de leguminosas que participan de la ruta de biosíntesis de flavonoides, una familia de compuestos del metabolismo secundario, algunos de los cuáles actúan como inductores de la nodulación (esto es, el proceso de formación de nódulos fijadores de nitrógeno en las leguminosas) (Liu & Murray, 2016).

Estas familias (OG0012003, OG0012753 y OG0013671) experimentaron expansiones de 13 genes en total (cinco, cuatro y cuatro cada una) para *F. albida* y tienen asignados un término GO genérico de la categoría Proceso Biológico (GO:0019438; “aromatic compound”) que puede observarse en la ***figura 9a*** y otro de la categoría Función Molecular (GO:0008171; “O-methyltransferase activity”) que aparece en la ***figura 11a***.

Asimismo, una cuarta familia de genes significativamente expandida en *F. albida* codifica para receptores kinasa tipo LysM, cuyos ortólogos en *Medicago truncatula* y *Lotus japonicus* son receptores de los factores de nodulación (Kouchi, et al. 2010). Los factores de nodulación son producidos por los rizobios, las bacterias fijadoras de nitrógeno, en respuesta a la presencia de flavonoides específicos producidos por la planta. La percepción de estos estos factores de nodulación desencadena una cáscada de transducción en la planta hospedadora que, en la mayoría de leguminosas, da lugar a la formación de estructuras intracelulares especializadas llamadas hilos de infección. Estos hilos de infección actúan como un conducto para proveer acceso a los rizobios en los tejidos internos, donde son englobados por endocitosis en células de nodulos y comienzan a fijar nitrógeno atmosférico (Yuan, et al., 2016).

Esta familia (OG0001286) experimentó una expansión de cinco genes y tiene asignados los términos GO genérico de la categoría Proceso Biológico (GO:0006468; “protein phosphorylation”) que puede observarse en la ***figura 9a*** y otro de la categoría Función Molecular (GO:0004672; “protein kinase activity”) que aparece en la ***figura 11a***.

-Almacenamiento de sustancias: se han detectado expansiones en una familia de genes (OG0000092) de transportadores de membrana para macronutrientes como el fosfato inorgánico y en otra (OG0000504) de transportadores vacuolares para micronutrientes como el hierro y el manganeso (aunque la afinidad por este último es menor) fueron detectados. Mientras que los primeros son esenciales para el crecimiento y desarrollo vegetal (Wang, et al., 2017) el rol de los últimos dentro de la fisiología vegetal ha ganado prominencia por sus distintas funciones: desde mantener la homeostasis de metales hasta contribuir a la formación de nódulos fijadores de nitrógeno (Ram, et al., 2021).

Estas familias (OG0000092 y OG0000504) experimentaron expansiones de 25 genes en total (16 y 9 cada una) para *F. albida* y tienen asignados cinco términos GO genéricos de la categoría

Proceso Biológico de las cuales dos (GO:0006817; "phosphate ion transport" y GO:0034755; "iron ion transmembrane transport") pueden observarse formando un clúster junto con otros términos en la ***figura 9a*** y tres términos GO genéricos de la categoría Función Molecular de los cuales también dos (GO:0005315; "inorganic phosphate transmembrane transporter activity" y GO:0005381; "iron ion transmembrane transporter activity") aparecen formando un clúster con otros términos en la ***figura 11a***.

-Resistencia a patógenos y plagas: fueron detectadas expansiones en cuatro familias de genes (OG000014, OG0001008, OG0001083 y OG0008228) relacionados con la resistencia al virús del moisaico de tabaco (genes N de resistencia. Niemeyer, et al., 2013) y a enfermedades y en otra (OG0000109) a insectos lepidópteros y bacterias, en concreto, inhibidores que actúan a la vez sobre alfa-amilasas, enzimas del sistema sistema digestivo, y subtisilinas, proteasas bacterianas (Yu, et al., 2017). En una de estas familias (OG0000109) también fueron identificados genes de inhibidores de tripsina (una enzima hidrolítica que rompe los enlaces peptídicos de las proteínas) que actúan en contra de insectos y otros animales y están muy difundidos entre las plantas (Shamsi, et al., 2016).

Estas familias (OG000014, OG0001008, OG0001083, OG0008228 y OG0000109) experimentaron expansiones de 95 genes en total (9, 31, 27, 10 y 18 cada una). Por un lado, el grupo de familias de genes N de resistencia y la familia de inhibidores tienen asignados cada uno un término GO genérico de la categoría Proceso Biológico (GO:0007165; "signal transduction" y GO:0010951; "negative regulation of endopeptidase activity") que pueden observarse en la ***figura 9a*** y, por otro, también a cada uno a se les fue asignado un término GO genérico de la categoría Función Molecular (GO:0043531; "ADP binding" y GO:0004866; "endopeptidase inhibitor activity") los cuales aparecen en la ***figura 11a***.

- *Lablab purpureus*

-Resistencia a lepidópteros: fue detectada una expansión en una familias génica relacionada con la resistencia a insectos lepidópteros (sus genes codifican inhibidores de la alfa-amilasa. Parag, et al., 2013).

Esta familia (OG0001585) experimentó una expansión de cinco genes y tiene asignado un término GO genérico de la categoría Proceso Biológico (GO:0031640; "killing of cells of other organism") que puede observarse en la ***figura 9b***.

-Alto contenido en cisteína: se ha detectado una expansión en una familia génica de una enzima (metionín gamma liasa) que forma parte de la ruta de biosíntesis de cisteína a partir de metionina y cuya expresión es constitutiva (Goyer, et al., 2007).

Esta familia (OG0004299) experimentó una expansión de cuatro genes y tiene asignados dos términos GO genéricos de la categoría Proceso Biológico de los cuales uno (GO:0019343; "cysteine biosynthetic process via cystathionine") puede observarse en la ***figura 9b***; y tres términos genéricos de la categoría Función Molecular que pueden observarse en la ***figura 11b*** donde dos de ellos (GO:0018826; "methionine gamma-lyase activity" y GO:000412; "cystathionine gamma-lyase activity") forman un clúster.

- *Sclerocarya birrea*

-Tolerancia al estrés abiótico: fueron identificado familia (OG0000043) expandida de genes relacionados con las proteínas *GsSRK* (*G-type lectin S-receptor-like serine/threonine-protein kinase*) que participan como reguladores positivos de la respuesta al estrés salino (Sun, et al., 2013). También fue detectada otra familia expandida (OG0000278) de genes de expansinas que son sobreexpresados durante sequías para reducir la rigidez de la pared celular aumentando la resistencia a la deshidratación (Tenhaken, 2014).

Estas familias (OG0000043 y OG0000278) experimentaron expansiones de 33 genes en total (23 y 10 cada uno). Mientras que la familia OG0000043 tiene asignados dos términos GO genéricos de la categoría Proceso Biológico (GO:0009069; "serine family amino acid metabolic process" y GO:0006468; "protein phosphorylation") que se pueden observar en la ***figura 9c*** y un término de la categoría Componente Celular (GO:0016020; "membrane") que se puede observar en la ***figura 10b***, la OG0000278 tiene un término GO genérico de la categoría Proceso Biológico (GO:0009664; "plant-type cell wall organization") que aparece en la ***figura 9c*** y tres términos de la categoría Componente Celular de los cuales uno (GO:0016020) aparece en la ***figura 10b***.

-Respuesta inmune: se ha detectado una familia expandida de genes cuyos productos son receptores kinasa ricos en cisteína 10, componentes clave de la respuesta inmune vegetal (Chern, et al., 2016).

Esta familia (OG0000006) experimentó una expansión de 19 genes y tiene asignados dos términos GO genéricos de la categoría Proceso Biológico (GO:0009069; "serine family amino acid metabolic process" y GO:0006468; "protein phosphorylation") que se pueden observar en la ***figura 9c*** y 1 término de la categoría Componente celular (GO:0016020; "membrane") que se puede observar en la ***figura 10b***

- *Vigna subterranea*

-Crecimiento celular elevado: se han detectado un grupo de dos familias génicas expandidas (OG0013622 y OG0002324) relacionadas con la regulación del crecimiento monopolar y una familia expandida (OG0001927) relacionadas con la regulación de la morfogénesis de la raíz cuyos respectivos productos génicos son las proteínas LONGIFOLIA 1 y 2 que promueven el alargamiento longitudinal polar de la célula vegetal (Lee, et al., 2006) y ATPasas plastídicas.

Estas familias (OG0013622, OG0002324 y OG0001927) experimentaron expansiones de 11 genes en total (uno, tres y siete cada uno). Tanto el grupo de familias de proteínas LONGIFOLIA 1 y 2 y la familia de ATPasas plastídicas tienen asignados cada uno un término GO genérico de la categoría Proceso Biológico (GO:0051513; "regulation of monopolar cell growth" y GO:2000067; "regulation of root morphogenesis") que pueden observarse en la ***figura 9d*** formando un clúster con otro término.

-Tolerancia al estrés oxidativo: fueron detectadas familias génicas expandidas cuyos productos son peroxidasas como la peroxidasa 7 que elimina peróxido de hidrógeno y oxida reductores tóxicos entre otras funciones.

Estas familias (OG0000004 y OG0002167) experimentaron expansiones de 18 genes en total (16 y 2 cada una) y tienen asignados un término GO genérico de la categoría Proceso Biológico

(GO:0042744; “hydrogen peroxide catabolic process”) que puede observarse en la ***figura 9d*** y un término de la categoría Función Molecular (GO:0004601; “peroxidase activity”) que aparece en la ***figura 11d***.

-Respuesta inmune: se han detectado tres familias génicas expandidas cuyos productos son las subtilasas (proteasas de la serina), componentes clave de la respuesta inmune vegetal (Figuereido, et al., 2014).

Estas familias (OG0000102, OG0012406 y OG0014382) experimentaron expansiones de 12 genes en total (ocho, dos y dos cada una) y tienen asignados un término GO genérico de la categoría Proceso Biológico (GO:0006508; “proteolysis”) que puede observarse en la ***figura 9d*** y un término de la categoría Función Molecular (GO:0004252; “serine-type endopeptidase activity”) que puede observarse en la ***figura 11d***.

-Alta proliferación celular: se han detectado familias génicas expandidas de una enzima (serina hidroximetiltransferasa) vinculada con el control de la proliferación celular por ser la principal fuente de unidades activadas de un carbono (Yi, et al., 2010).

Estas familias (OG0001767 y OG0003025) experimentaron expansiones de cinco genes en total (tres y dos cada una) y tienen asignados un término GO genérico de la categoría Proceso Biológica (GO:0019264; “glycine biosynthetic process from serine”) que puede observarse en la ***figura 9d*** y un término de la categoría Función Molecular (GO:0004372; “glycine hydroxymethyltransferase activity”) que puede observarse en la ***figura 11d***.

-Resistencia a estreses nodulación: se ha detectado una familia génica expandida de una enzima (chalcona sintasa o naringerín chalcona sintasa) que participa en la respuesta a la radiación ultravioleta e infecciones bacterianas o fúngicas mediante la síntesis de (iso)flavonoides (Dao, et al., 2010). También se ha reportado que esta enzima participa en la ruta de biosíntesis de inductores de la nodulación (en concreto, flavonoides. Liu & Murray, 2016).

Esta familia (OG0000527) experimentó una expansión de tres genes y tiene asignados un término GO genérico de la categoría Función Molecular (GO:0016210; “naringenin-chalcone synthase activity”) que puede observarse en la ***figura 11d***.

5. Conclusiones

Las principales conclusiones de este estudio fueron las siguientes:

1. Se ha obtenido la anotación funcional de 68433 genes de cuatro especies del AOCC con 315830 términos GO, 17909 genes con 80986 términos EC y 78951 genes con 411776 términos InterPro. Mientras que las anotaciones GO cubren entre el 50 y 81% de los genes en cada genoma, las anotaciones InterPro cubren entre el 75 y el 85%.

2. Se ha generado una clasificación de ortogrupos/familias génicas donde 85928 genes de las cuatro especies del AOCC fueron asignados a 17998 ortogrupos/familias. A su vez, unos 510 genes están repartidos en 108 ortogrupos específicos. Asimismo, entre un 81 y un 92% de los genes de cada especie fueron asignados a algún ortogrupo quedando unos 14377 genes huérfanos en total.

3. A partir del modelado de las dinámicas evolutivas de familias génicas, se han identificado un total de 1712 ortogrupos/familias génicas expandidas formados por 8123 genes y que habrían ganado un total de 6527 genes. Por su parte, se han identificado 1672 ortogrupos/familias contraídas formados por 93 genes y que habrían perdido un total de 1925 genes.

4. El análisis de enriquecimiento funcional de las funciones biológicas expandidas ha permitido detectar posibles caracteres de interés favorecidos por la selección como la capacidad de incorporar nitrógeno a través de la simbiosis con bacterias fijadoras de nitrógeno atmosférico o la tolerancia/resistencia a distintos estreses.

Este estudio es un trabajo enfocado en caracterizar a través de herramientas bioinformáticas los genomas de plantas poco estudiadas sirviendo como una primera toma de contacto en la identificación de las bases moleculares y genéticas de las propiedades atribuidas a éstas, así como la detección de otras nuevas. Los resultados obtenidos en este trabajo sobre cuatro especies huérfanas africanas que incluyen la anotación funcional de sus genomas, así como la identificación y caracterización funcional de familias de genes expandidas específicamente en cada uno de sus repectivos linajes, podrán ser empleados como material preliminar previo a la realización de investigaciones bioquímicas y/o genéticas.

6. Bibliografía

Ahmed M, Trisha UK, Shaha SR, Dey AK and Rahmatullah M. (2015). An initial report on the antihyperglycemic and antinociceptive potential of Lablab purpureus beans. World Journal of Pharmacy and Pharmaceutical Sciences; 4(10): 95-105.

Akaike, H. (1992). Information Theory and an Extension of the Maximum Likelihood Principle. En: J. N. Kotz S., ed. Breakthroughs in Statistics. New York: Springer Science Business Media, pp. 610-624.

Al-Snafi, A.E. (2017). The pharmacology and medical importance of Dolichos lablab (Lablab purpureus)- A review. IOSR Journal of Pharmacy, 07, 22-30.

Albert, V., Barbazuk, W., dePamphilis, C., Der, J., Leebens-Mack, J., Ma, H., Palmer, J., Rounsley, S., Sankoff, D., Schuster, S., Soltis, D., Soltis, P., Wessler, S., Wing, R., Ammiraju, J., Chamala, S., Chanderbali, A., Determann, R., Ralph, P., … Wanke, S. (2013). The Amborella Genome and the Evolution of Flowering Plants. Science (American Association for the Advancement of Science), 342(6165), 1467–1467. https://doi.org/10.1126/science.1241089

Azman Halimi, R., Barkla, B., Mayes, S., & King, G. (2019). The potential of the underutilized pulse bambara groundnut (Vigna subterranea (L.) Verdc.) for nutritional food security. Journal of Food Composition and Analysis, 77, 47–59. https://doi.org/10.1016/j.jfca.2018.12.008

B.L. Maass. (2016). Origin, domestication and global dispersal of lablab purpureus (L.) Sweet (fabaceae): current understanding Special Issue on Hyacinth Bean: A gem Among Legumes. Legume Perspect, vol. 13, pp. 5-8.

B.R. Raghu, D.K. Samuel, N. Mohan, T.S. (2018). Aghora Dolichos bean: an underutilized and unexplored crop with immense potential Internat. J. Recent Adv. Multidis Res., 05, pp. 4338-4341.

Bailey-Serres, J., Parker, J., Ainsworth, E., Oloyd, G., & Schroeder, J. (2019). Genetic strategies for improving crop yields. Nature (London), 575(7781), 109–118. https://doi.org/10.1038/s41586-019-1679-0

Barnes, R. D., & Fagg, C. W. (2003). Faidherbia albida: monograph and annotated bibliography. Oxford Forestry Institute, University of Oxford.

Belemtougri, R.G.; Traore, A.; Ouedraogo, Y.; Sanou, S.D.; Sawadogo, L. (2007). Toxicological effects of Sclerocarya birrea (A. Rich) Hochst (Anacardiaceae) and Psidium guajava L. (Myrtaceae) leaf extracts on mice and their pharmacological effects on rat duodenum. Int. J. Pharmacol. 3, 68–73.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Volumen 8, pp. 3-62.

Borochov-Neori, H., Judeinstein, S., Greenberg, A., Fuhrman, B., Attias, J., Volkova, N., Hayek, T., & Aviram, M. (2008). Phenolic Antioxidants and Antiatherogenic Effects of Marula (Sclerocarrya birrea Subsp. caffra) Fruit Juice in Healthy Humans. Journal of Agricultural and Food Chemistry, 56(21), 9884–9891. https://doi.org/10.1021/jf801467m

Buchfink, B., Xie, C., & Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59–60. https://doi.org/10.1038/nmeth.3176

Carretero-Paulet, L., Chang, T., Librado, P., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J., & Albert, V. (2015). Genome-wide analysis of adaptive molecular evolution in the carnivorous plant Utricularia gibba. Genome Biology and Evolution, 7(2), 444–456. https://doi.org/10.1093/gbe/evu288

Carretero-Paulet, L., Librado, P., Chang, T., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J., & Albert, V. (2015). High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant Utricularia gibba. Molecular Biology and Evolution, 32(5), 1284–1295. https://doi.org/10.1093/molbev/msv020

Chaw, S., Y., L., Wu, Y. & al, e, (2019). Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nature Plants, Volumen 5, pp. 63-73.

Chern M, Xu Q, Bart RS, Bai W, Ruan D, Sze-To WH, et al. (2016) Correction: A Genetic Screen Identifies a Requirement for Cysteine-Rich-Receptor-Like Kinases in Rice NH1 (OsNPR1)-Mediated Immunity. PLoS Genet 12(7): e1006182. https://doi.org/10.1371/journal.pgen.1006182

Conesa, A., et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674 - 3676.

Damfami A & Namo OAT. (2020). Bambara Groundnut (Vigna subterranea (L.) Verd.): A Review of Its Past, Present and Future Role in Human Nutrition. J Agric Forest Meteorol Res, 3(1): 274-281.

Dao, T.T.H., Linthorst, H.J.M. & Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. Phytochem Rev 10, 397. https://doi.org/10.1007/s11101-011-9211-7

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

Degefu, T., Wolde-Meskel, E., Woliy, K., & Frostegård, Å. (2017). Phylogenetically diverse groups of Bradyrhizobium isolated from nodules of tree and annual legume species growing in Ethiopia. Systematic and Applied Microbiology, 40(4), 205–214. https://doi.org/10.1016/j.syapm.2017.04.001

Delgado-Salinas, A., Thulin, M., Pasquet, R., Weeden, N., & Lavin, M. (2011). Vigna (Leguminosae) sensu lato: The names and identities of the American Segregate Genera. American Journal of Botany, 98(10), 1694–1715. https://doi.org/10.3732/ajb.1100069

Dimo, T., Rakotonirina, S., Tan, P., Azay, J., Dongo, E., Kamtchouing, P., & Cros, G. (2007). Effect of Sclerocarya birrea (Anacardiaceae) stem bark methylene chloride/methanol extract on streptozotocin-diabetic rats. Journal of Ethnopharmacology, 110(3), 434–438. https://doi.org/10.1016/j.jep.2006.10.020

Dièye, A., Sarr, A., Diop, S., Ndiaye, M., Sy, G., Diarra, M., Rajraji/Gaffary, I., Ndiaye/Sy, A., & Faye, B. (2008). Medicinal plants and the treatment of diabetes in Senegal: survey with patients. Fundamental & Clinical Pharmacology, 22(2), 211–216. https://doi.org/10.1111/j.1472-8206.2007.00563.x

Douglas E. Soltis, Stephen A. Smith, Nico Cellinese, Kenneth J. Wurdack, David C. Tank, Samuel F. Brockington, Nancy F. Refulio-Rodriguez, Jay B. Walker, Michael J. Moore, Barbara S. Carlsward, Charles D. Bell, Maribeth Latvis, Sunny Crawley, Chelsea Black, Diaga Diouf, Zhenxiang Xi, Catherine A. Rushworth, Matthew A. Gitzendanner, Kenneth J. Sytsma, … Pamela S. Soltis. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany, 98(4), 704–730. https://doi.org/10.3732/ajb.1000404

Emms, D., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology, 20(1), 238–238. https://doi.org/10.1186/s13059-019-1832-y

FAO. (2010). The 2. Report on the state of the world's plant genetic resources for food and agriculture. In FAO (Ed.), FAO.

Figueiredo, A., Monteiro, F., & Sebastiana, M. (2014). Subtilisin-like proteases in plant-pathogen recognition and immune priming: a perspective. Frontiers in plant science, 5, 739. https://doi.org/10.3389/fpls.2014.00739

Fisher, R. A. (1922). On the Interpretation of χ² from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85, 87-94.

France Denoeud, Lorenzo Carretero-Paulet, Alexis Dereeper, Gaëtan Droc, Romain Guyot, Marco Pietrella, Chunfang Zheng, Adriana Alberti, François Anthony, Giuseppe Aprea, Jean-Marc Aury, Pascal Bento, Maria Bernard, Stéphanie Bocs, Claudine Campa, Alberto Cenci, Marie-Christine Combes, Dominique Crouzillat, Corinne Da Silva, … Olivier Garsmeur. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science (American Association for the Advancement of Science), 345(6201), 1181–1184. https://doi.org/10.1126/science.1255274

Fukushima, K., Fang, X., Alvarez-Ponce, D., Cai, H., Carretero-Paulet, L., Chen, C., Chang, T., Farr, K., Fujita, T., Hiwatashi, Y., Hoshi, Y., Imai, T., Kasahara, M., Librado, P., Mao, L., Mori, H., Nishiyama, T., Nozawa, M., Pálfalvi, G., … Hasebe, M. (2017). Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. Nature Ecology & Evolution, 1(3), 59–59. https://doi.org/10.1038/s41559-016-0059

Garba, S., Ahmadu, S., & John, I. (2006). The effect of aqueous stem bark extract of Sclerocarya birrea (Hoechst) on alcohol carbon tetrachloride induced liver damage in rats. In Pakistan Journal of Biological Sciences (Vol. 9, Issue 12, pp. 2283–2288). https://doi.org/10.3923/pjbs.2006.2283.2287

Garra, A. H., Eltomy, S. A., Aarrag, A. R. H., & Ahmed, N. M. (2020). Antidiabetic, Antihyperlipidemic and Antioxidant Activities of Acacia albida in Streptozotoc in Induced Diabetes in Rats and its Metabolites. Egyptian Journal of Chemistry, 63(1), 337-348.

Gathirwa, J., Rukunga, G., Njagi, E., Omar, S., Mwitari, P., Guantai, A., Tolo, F., Kimani, C., Muthaura, C., Kirira, P., Ndunda, T., Amalemba, G., Mungai, G., & Ndiege, I. (2008). The in vitro anti-plasmodial and in vivo anti-malarial efficacy of combinations of some medicinal plants used traditionally for treatment of malaria by the Meru community in Kenya. Journal of Ethnopharmacology, 115(2), 223–231. https://doi.org/10.1016/j.jep.2007.09.021

Gene Ontology Resource. Guide to GO subsets. (2020). Obtenido de: http://geneontology.org/docs/go-subset-guide/

Gondwe, M., Kamadyaapa, D., Tufts, M., Chuturgoon, A., & Musabayane, C. (2008). Sclerocarya birrea [(A. Rich.) Hochst.] [Anacardiaceae] stem-bark ethanolic extract (SBE) modulates blood glucose, glomerular filtration rate (GFR) and mean arterial blood pressure (MAP) of STZ-induced diabetic rats. Phytomedicine (Stuttgart), 15(9), 699–709. https://doi.org/10.1016/j.phymed.2008.02.004

Gopalakrishnan, TR. (2007). Legume vegetables. In: Vegetable crops. Horticulture science series 4. New India Publishing Agency, New Delhi, India, chapter 8, pp 169–198

Goyer, A., Collakova, E., Shachar-Hill, Y., & Hanson, A. (2007). Functional characterization of a methionine gamma-lyase in Arabidopsis and its implication in an alternative to the reverse trans-sulfuration pathway. Plant and Cell Physiology, 48(2), 232–242. https://doi.org/10.1093/pcp/pcl055

Graham, P., & Vance, C. (2003). Legumes: Importance and Constraints to Greater Use. Plant Physiology (Bethesda), 131(3), 872–877. https://doi.org/10.1104/pp.017004

Guo, T., Song, Y., Lu, Y., Li, G., Liu, T., Han, W., Song, W., Yang, C., Li, F., & Liu, Q. (2020). Total synthesis and anticancer activity of the natural oleanolic acid bidesmoside saponin, albidoside A, isolated from the roots of Acacia albida. Journal of Chemical Research, 44(1-2), 42–49. https://doi.org/10.1177/1747519819884152

Habib MAM, Hasan R, Nayeem J, Uddin N and Rana S. (2012). Anti-inflammatory, antioxidant and cytotoxic potential of methanolic extract of two Bangladeshi bean Lablab purpureus L. sweet white and purple. IJPSR; 3(3): 776-781.

Hall, J.B., Sinclair, F.L., O'Brien, E.M. (2002). Sclerocarya birrea: a monograph. School of Agricultural and Forest Sciences Publication number 19, University of Wales. Bangor, UK, ISBN:1 84220 049 6. ISSN:0962-7766, 157 pp.

Harris, T., Jideani, V., & Le Roes-Hill, M. (2018). Flavonoids and tannin composition of Bambara groundnut (Vigna subterranea) of Mpumalanga, South Africa. Heliyon, 4(9), e00833–e00833. https://doi.org/10.1016/j.heliyon.2018.e00833

Hendre, P., Muthemba, S., Kariba, R., Muchugi, A., Fu, Y., Chang, Y., Song, B., Liu, H., Liu, M., Liao, X., Sahu, S., Wang, S., Li, L., Lu, H., Peng, S., Cheng, S., Xu, X., Yang, H., Wang, J., … Jamnadass, R. (2019). African Orphan Crops Consortium (AOCC): status of developing genomic resources for African orphan crops. Planta, 250(3), 989–1003. https://doi.org/10.1007/s00425-019-03156-9 Chang, Y., Liu, H., Liu, M., Liao, X., Sahu, S. K., Fu, Y., ... & Hendre, P. S. (2019). The draft genomes of five agriculturally important African orphan crops. GigaScience, 8(3), giy152.

Hiwilepo-Van Hal, P., Bille, P., Verkerk, R., Boekel, v, & Dekker, M. (2014). A review of the proximate composition and nutritional value of Marula (Sclerocarya birrea subsp. caffra). Phytochemistry Reviews, 13(4), 881–892. https://doi.org/10.1007/s11101-014-9352-6

Hussin Hilda, Gregory Peter J., Julkifle Advina L., Sethuraman Gomathy, Tan Xin Lin, Razi Fadhil, Azam-Ali Sayed N. (2020). Enhancing the Nutritional Profile of Noodles With Bambara Groundnut (Vigna subterranea) and Moringa (Moringa oleifera): A Food System Approach. Frontiers in Sustainable Food Systems, 4, 59. https://doi.org/10.3389/fsufs.2020.00059

I.A. Onimawo, A.H. Momoh, A. Usman. (1998). Proximate composition and functional properties of four cultivars of Bambara groundnut (Voandezeia subterranean), Plant Foods Hum. Nutr. 53, 153–158.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C., Carretero-Paulet, L., Chang, T., Lan, T., Welch, A., Juárez, M., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S., Himmelbauer, H., Minoche, A., Xu, S., Lynch, M., … Herrera-Estrella, L. (2013). Architecture and evolution of a minute plant genome. Nature (London), 498(7452), 94–98. https://doi.org/10.1038/nature12132

InterPro. (2020). Obtenido de: http://www.ebi.ac.uk/interpro/about/consortium/

Jake R Conway, Alexander Lex, Nils Gehlenborg. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics, Volume 33, Issue 18, Pages 2938–2940, https://doi.org/10.1093/bioinformatics/btx364

Jamnadass, R., Mumm, R., Hale, I., Hendre, P., Muchugi, A., Dawson, I., Powell, W., Graudal, L., Yana-Shapiro, H., Simons, A., & Van Deynze, A. (2020). Enhancing African orphan crops with genomics. Nature Genetics, 52(4), 356–360. https://doi.org/10.1038/s41588-020-0601-x

Janarthanan, S., Suresh, P., Radke, G., Morgan, T., & Oppert, B. (2008). Arcelins from an Indian Wild Pulse, Lablab purpureus, and Insecticidal Activity in Storage Pests. Journal of Agricultural and Food Chemistry, 56(5), 1676–1682. https://doi.org/10.1021/jf071591g

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England), 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jurka, J., et al. (2005). Repbase Update, a database of eukaryoticrepetitive elements. Cytogenet Genome Research, Volumen 110, pp. 462-467.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Research, 47(D1), D590–D595. https://doi.org/10.1093/nar/gky962

Karoune, S., Falleh, H., Kechebar, M. S. A., Halis, Y., Mkadmini, K., Belhamra, M., ... & Ksouri, R. (2015). Evaluation of antioxidant activities of the edible and medicinal Acacia albida organs related to phenolic compounds. Natural product research, 29(5), 452-454.

Kassambara A. (2020). ggpubr: 'ggplot2' Based. Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kouchi, H., Imaizumi-Anraku, H., Hayashi, M., Hakoyama, T., Nakagawa, T., Umehara, Y., Suganuma, N., & Kawaguchi, M. (2010). How Many Peas in a Pod? Legume Genes Responsible for Mutualistic Symbioses Underground. Plant and Cell Physiology, 51(9), 1381–1397. https://doi.org/10.1093/pcp/pcq107

Kugedera Andrew Tapiwa. (2019). Harvesting and utilization of Marula (Sclerocarya birrea) by Smallholder farmers: A review. JOJ Wildlife & Biodiversity, Juniper Publishers Inc., vol. 1(3), pages 76-79, August.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution, 34(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Lee YK, Kim GT, Kim IJ, Park J, Kwak SS, Choi G, Chung WI. (2006). LONGIFOLIA1 and LONGIFOLIA2, two homologous genes, regulate longitudinal cell elongation in Arabidopsis. Development. 133(21):4305-14. doi: 10.1242/dev.02604. PMID: 17038516.

Li FW, Brouwer P, Carretero-Paulet L, Cheng S, de Vries J, Delaux PM, Eily A, Koppers N, Kuo LY, Li Z, Simenc M, Small I, Wafula E, Angarita S, Barker MS, Bräutigam A, dePamphilis C, Gould S, Hosmani PS, Huang YM, Huettel B, Kato Y, Liu X, Maere S, McDowell R, Mueller LA, Nierop KGJ, Rensing SA, Robison T, Rothfels CJ, Sigel EM, Song Y, Timilsena PR, Van de Peer Y, Wang H, Wilhelmsson PKI, Wolf PG, Xu X, Der JP, Schluepmann H, Wong GK, Pryer KM. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat Plants. 4(7):460-472. doi: 10.1038/s41477-018-0188-8. Epub 2018 Jul 2. PMID: 29967517; PMCID: PMC6786969.

Liu, C., & Murray, J. (2016). The Role of Flavonoids in Nodulation Host-Range Specificity: An Update. Plants (Basel), 5(3), 33–. https://doi.org/10.3390/plants5030033

Maass, B., Knox, M., Venkatesha, S., Angessa, T., Ramme, S., & Pengelly, B. (2010). Lablab purpureus-A Crop Lost for Africa? Tropical Plant Biology, 3(3), 123–135. https://doi.org/10.1007/s12042-010-9046-1

Michael J. Moore, Charles D. Bell, Pamela S. Soltis, & Douglas E. Soltis. (2007). Using Plastid Genome-Scale Data to Resolve Enigmatic Relationships among Basal Angiosperms. Proceedings of the National Academy of Sciences - PNAS, 104(49), 19363–19368. https://doi.org/10.1073/pnas.0708072104

Mikail, H.G. (2009). In vitro trypanocidal effect of methanolic extract of Sclerocarya birrea, Commiphora kerstingii and Khaya senegalensis. Afr. J. Biotech. 8, 2047–2049.

Musabayane, C.T.; Gondwe, M.; Kamadyaapa, D.R.; Moodley, K.; Ojewole, J.A.O. (2006). The effects of Sclerocarya birrea [(A. Rich.) Hochst.] [Anacardiaceae] stem-bark aqueous extract on blood glucose, kidney and cardiovascular function in rats. Endocrine Abstracts 2006, 2, SP36.

Nasrin, F., Bulbul, I., Begum, Y., & Khanum, S. (2012). In vitro antimicrobial and cytotoxicity screening of n-hexane, chloroform and ethyl acetate extracts of Lablab purpureus (L.) leaves. Agriculture and Biology Journal of North America, 3, 43-48.

Ndidi, U., Umar, I., Mohammed, A., Samuel, C., Oladeru, A., & Yakubu, R. (2015). Effects of aqueous extracts of Acacia albida stem bark on Wistar albino rats infected with Trypanosoma evansi. In Natural Product Research (Vol. 29, Issue 12, pp. 1153–1157). https://doi.org/10.1080/14786419.2014.981184

Neo C Mokgolodi, Moffat P Setshogo, Ling-ling Shi, Yu-jun Liu, & Chao Ma. (2011). Achieving food and nutritional security through agroforestry: a case of Faidherbia albida in sub-Saharan Africa. Forestry Studies in China, 13(2), 123–. https://doi.org/10.1007/s11632-011-0202-y

Neuwirth E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Niemeyer, J., Ruhe, J., Machens, F. et al. (2013). Differential expression of the TMV resistance gene N prevents a hypersensitive response in seeds and during germination. Planta 237, 909–915. https://doi.org/10.1007/s00425-012-1832-6

Nyoka, B., Nyoka, B., Chanyenga, T., Chanyenga, T., Mng'omba, S., Mng'omba, S., Akinnifesi, F., Akinnifesi, F., Sagona, W., & Sagona, W. (2015). Variation in growth and fruit yield of populations

of Sclerocarya birrea (A. Rich.) Hochst. Agroforestry Systems, 89(3), 397–407. https://doi.org/10.1007/s10457-014-9774-6

Ojeda-López, J., Marczuk-Rojas, J.P., Polushkina, O.A., et al. (2020). Evolutionary analysis of the Moringa oleifera genome reveals a recent burst of plastid to nucleus gene duplications. Sci Rep 10, 17646. https://doi.org/10.1038/s41598-020-73937-w

Ojewole, J. (2003). Evaluation of the anti-inflammatory properties of Sclerocarya birrea (A. Rich.) Hochst. (family: Anacardiaceae) stem-bark extracts in rats. Journal of Ethnopharmacology, 85(2), 217–220. https://doi.org/10.1016/S0378-8741(03)00019-9

OmicsBox – Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, https://www.biobam.com/omicsbox

P. Librado, F. G. Vieira, J. Rozas. (2012). BadiRate: estimating family turnover rates by likelihood-based methods, Bioinformatics, Volume 28, Issue 2, Pages 279–281, https://doi.org/10.1093/bioinformatics/btr623

Parag Gupta, Anuradha Singh, Gaurav Shukla, Neeraj Wadhwa. (2013). Bio-insecticidal potential of amylase inhibitors. JPR:BioMedRx: An International Journal. Volume 1:449-458.

Puozaa, D., Jaiswal, S., & Dakora, F. (2017). African origin of Bradyrhizobium populations nodulating Bambara groundnut (Vigna subterranea L. Verdc) in Ghanaian and South African soils. PloS One, 12(9), e0184943–e0184943. https://doi.org/10.1371/journal.pone.0184943

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Ram, H, Sardar, S, Gandass, N. (2021). Vacuolar Iron Transporter (Like) proteins: Regulators of cellular iron accumulation in plants. Physiologia Plantarum. 171: 823– 832. https://doi.org/10.1111/ppl.13363

Ray, D., Mueller, N., West, P., & Foley, J. (2013). Yield Trends Are Insufficient to Double Global Crop Production by 2050. PloS One, 8(6), e66428–e66428. https://doi.org/10.1371/journal.pone.0066428

Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., Perez-Torres, C., Chacón-López, A., Hernandez-Guzmán, G., Chang, T., Farr, K., Barbazuk, W., Chamala, S., Mutwil, M., Shivhare, D., Alvarez-Ponce, D., Mitter, N., Hayward, A., Fletcher, S., … Herrera-Estrella, L. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences - PNAS, 116(34), 17081–17089. https://doi.org/10.1073/pnas.1822129116

Russo, D., Miglionico, R., Carmosino, M., Bisaccia, F., Andrade, P., Valentão, P., Milella, L., & Armentano, M. (2018). A Comparative Study on Phytochemical Profiles and Biological Activities of Sclerocarya birrea (A.Rich.) Hochst Leaf and Bark Extracts. International Journal of Molecular Sciences, 19(1), 186–. https://doi.org/10.3390/ijms19010186

S. Myhre, H. Tveit, T. Mollestad, A. Laegreid, Additional gene ontology structure for improved biological reasoning. (2006). Bioinformatics 22, 2020–2027. Medline https://doi:10.1093/bioinformatics/btl334

Sahu, S., Liu, M., Yssel, A., Kariba, R., Muthemba, S., Jiang, S., Song, B., Hendre, P., Muchugi, A., Jamnadass, R., Kao, S., Featherston, J., Zerega, N., Xu, X., Yang, H., Deynze, A., Peer, Y., Liu, X., & Liu, H. (2019). Draft Genomes of Two Artocarpus Plants, Jackfruit (A. heterophyllus) and Breadfruit (A. altilis). Genes, 11(1), 27–. https://doi.org/10.3390/genes11010027

Salawu, O. A., Tijani, A. Y., Babayi, H., Nwaeze, A. C., Anagbogu, R. A., & Agbakwuru, V. A. (2010). Antimalarial activity of ethanolic stem bark extract of Faidherbia Albida (Del) a. Chev (Mimosoidae) in mice. Arch Appl Sci Res, 2(5), 261-268.

Sands, D., Morris, C., Dratz, E., & Pilgeram, A. (2009). Elevating optimal human nutrition to a central goal of plant breeding and production of plant-based foods. Plant Science (Limerick), 177(5), 377–389. https://doi.org/10.1016/j.plantsci.2009.07.011

Schulze-Kaysers, N., Feuereisen, M., & Schieber, A. (2015). Phenolic compounds in edible species of the Anacardiaceae family - a review. RSC Advances, 5(89), 7331–73314. https://doi.org/10.1039/c5ra11746a

Shamsi, T., Parveen, R., & Fatima, S. (2016). Characterization, biomedical and agricultural applications of protease inhibitors: A review. International Journal of Biological Macromolecules, 91, 1120–1133. https://doi.org/10.1016/j.ijbiomac.2016.02.069

Somulung SA, Lucero MA, Niverca MS, Dalin KA, Dejesus R and Domingo ED. (2012). In vivo study on the effect of Dolichos lablab (bataw) beans extract against Iron-deficiency in Rattus norvergicus (Wistar rat). Fatima University Research Journal; 4:112-115.

Sun, X., Yu, Q., Tang, L., Ji, W., Bai, X., Cai, H., Liu, X., Ding, X., & Zhu, Y. (2013). GsSRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress. Journal of Plant Physiology, 170(5), 505–515. https://doi.org/10.1016/j.jplph.2012.11.017

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS One, 6(7), e21800–e21800. https://doi.org/10.1371/journal.pone.0021800

TAIR. GO Slim Help. (2020). Obtenido de: https://www.arabidopsis.org/help/helppages/go_slim_help.jsp

Tadele, Z. (2019). Orphan crops: their importance and the urgency of improvement. Planta, 250(3), 677–694. https://doi.org/10.1007/s00425-019-03210-6

Tan, X., Azam-Ali, S., Goh, E., Mustafa, M., Chai, H., Ho, W., Mayes, S., Mabhaudhi, T., Azam-Ali, S., & Massawe, F. (2020). Bambara Groundnut: An Underutilized Leguminous Crop for Global Food Security and Nutrition. Frontiers in Nutrition (Lausanne), 7, 601496–601496. https://doi.org/10.3389/fnut.2020.601496

Tan, X., Azam-Ali, S., Goh, E., Mustafa, M., Chai, H., Ho, W., Mayes, S., Mabhaudhi, T., Azam-Ali, S., & Massawe, F. (2020). Bambara Groundnut: An Underutilized Leguminous Crop for Global Food Security and Nutrition. Frontiers in Nutrition (Lausanne), 7, 601496–601496. https://doi.org/10.3389/fnut.2020.601496

Tanya Z. Berardini, Suparna Mundodi, Leonore Reiser, Eva Huala, Margarita Garcia-Hernandez, Peifen Zhang, Lukas A. Mueller, Jungwoon Yoon, Aisling Doyle, Gabriel Lander, Nick Moseyko, Danny Yoo, Iris Xu, Brandon Zoeckler, Mary Montoya, Neil Miller, Dan Weems, & Seung Y. Rhee. (2004). Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. Plant Physiology (Bethesda), 135(2), 745–755. https://doi.org/10.1104/pp.104.040071

Tarailo-Graovac M, Chen N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. Chapter 4: Unit 4.10. doi: 10.1002/0471250953.bi0410s25. PMID: 19274634.

Tchoukoua, A., Kuiate Tabopda, T., Konga Simo, I., Uesugi, S., Ohno, M., Kimura, K. I., ... & Ngadjui, B. T. (2018). Albidosides H and I, two new triterpene saponins from the barks of Acacia albida Del.(Mimosaceae). Natural product research, 32(8), 924-932.

Tchoukoua, A., Tabopda, T., Uesugi, S., Ohno, M., Kimura, K., Kwon, E., Momma, H., Horo, I., Çalişkan, Ö., Shiono, Y., & Ngadjui, B. (2017). Triterpene saponins from the roots of Acacia albida Del. (Mimosaceae). Phytochemistry (Oxford), 136, 31–38. https://doi.org/10.1016/j.phytochem.2016.12.019

Tenhaken, R. (2014). Cell wall remodeling under abiotic stress. Frontiers in Plant Science, 5, 771–771. https://doi.org/10.3389/fpls.2014.00771

Tibe, O., & Amarteifio, J. (2010). Trypsin Inhibitor Activity and Condensed Tannin Content in Bambara Groundnut (Vigna Subterranea (L.) Verdc) Grown in Southern Africa. Journal of Applied Science & Environmental Management, 11(2). https://doi.org/10.4314/jasem.v11i2.55021

Udeh, E., Nyila, M., & Kanu, S. (2020). Nutraceutical and antimicrobial potentials of Bambara groundnut (Vigna subterranean): A review. Heliyon, 6(10), e05205–e05205. https://doi.org/10.1016/j.heliyon.2020.e05205

Usman, W. A., Mahmoud, S. J., & Ahmed, Z. H. (2013). Antimicrobial activity of stem bark of Faidherbia albida. Journal of Pharmaceutical Research International, 786-794.

Wang, D., Lv, S., Jiang, P., & Li, Y. (2017). Roles, Regulation, and Agricultural Application of Plant Phosphate Transporters. Frontiers in Plant Science, 8, 817–817. https://doi.org/10.3389/fpls.2017.00817

Wickham H. & Bryan J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham H. & Henry L. (2020). tidyr: Tidy Messy Data. R package version 1.1.0. https://CRAN.R-project.org/package=tidyr

Wickham H. & Seidel D. (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

World Bank. (2009). Eastern Africa - A study of the Regional Maize Market and Marketing Costs. World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/3155 License: CC BY 3.0 IGO.

Yi Zhang, Kehan Sun, Francisco J. Sandoval, Katherine Santiago, Sanja Roje. (2010). One-carbon metabolism in plants: characterization of a plastid serine hydroxymethyltransferase. Biochem J; 430 (1): 97–105. doi: https://doi.org/10.1042/BJ20100566

Yu, J., Yu, J., Li, Y., Li, Y., Xiang, M., Xiang, M., Zhu, J., Zhu, J., Huang, X., Huang, X., Wang, W., Wang, W., Tan, R., Tan, R., Zhou, J., Zhou, J., Liao, H., & Liao, H. (2017). Molecular cloning and characterization of α-amylase/subtilisin inhibitor from rhizome of Ligusticum chuanxiong. Biotechnology Letters, 39(1), 141–148. https://doi.org/10.1007/s10529-016-2227-8

Yuan, S., Li, R., Wang, L., Chen, H., Zhang, C., Chen, L., Hao, Q., Shan, Z., Zhang, X., Chen, S., Yang, Z., Qiu, D., & Zhou, X. (2016). Search for Nodulation and Nodule Development-Related Cystatin Genes in the Genome of Soybean (Glycine max). Frontiers in Plant Science, 7, 1595–1595. https://doi.org/10.3389/fpls.2016.01595