



UNIVERSIDAD  
DE ALMERÍA

## TRABAJO DE FIN DE MÁSTER

**Ensamblaje a escala cromosómica del genoma de  
*Moringa oleifera* revela evolución de rutas del  
metabolismo secundario/Chromosome-scale assembly  
of the *Moringa oleifera* genome uncovers evolution of  
secondary metabolism pathways**

**Autor:** Juan Pablo Marczuk Rojas

**Tutor:** Lorenzo Carretero Paulet

**Máster en Biotecnología Industrial y Agroalimentaria**

Facultad de Ciencias Experimentales

Departamento de Biología y Geología

Área de Genética

Curso académico 2021-2022

Convocatoria de Mayo



## Índice de contenido

<b>Resumen</b> .....	<b>i</b>
<b>Summary</b> .....	<b>ii</b>
<b>1. Introducción</b> .....	<b>1</b>
1.1. Antecedentes, motivación y enfoque del estudio .....	1
1.2. Presentación de <i>Moringa oleifera</i> .....	2
1.2.1. Biología.....	2
1.2.2. Cultivo, propiedades y usos .....	4
1.2.3. Genoma .....	7
<b>2. Objetivos</b> .....	<b>9</b>
2.1. Realizar la anotación funcional del genoma de Moringa .....	9
2.2. Generar clasificaciones de ortogrupos/familias génicas y de genes duplicados en Moringa .....	9
2.3. Determinar la contribución de las duplicaciones WGD y SSD a la expansión y/o retención de determinadas funciones biológicas .....	9
2.4. Investigar el rol de las duplicaciones en tándem en la evolución de la respuesta de defensa y las rutas del metabolismo secundario vegetal .....	9
<b>3. Materiales y métodos</b> .....	<b>10</b>
3.1. Genoma y transcriptoma .....	10
3.2. Anotación funcional del genoma .....	10
3.3. Clasificación de ortogrupos/familias génicas .....	11
3.4. Clasificación de genes duplicados en base a modos de duplicación.....	12
3.5. Identificación y caracterización de rutas biosintéticas y de clústeres de genes de metabolismo secundario. ....	13
3.6. Representaciones gráficas y tests estadísticos .....	14
<b>4. Resultados y discusión</b> .....	<b>15</b>
4.1. Anotación funcional del genoma de Moringa cubre más del 80% de sus genes .....	15
4.2. Clasificaciones de ortogrupos/familias génicas y de genes duplicados en Moringa abarcan más del 85% y del 65% de sus genes, respectivamente .....	15
4.3. Análisis de la evolución de familias génicas en Moringa.....	17
4.3.1. El patrón diferencial de retención funcional de genes duplicados clasificados por mecanismos de duplicación en Moringa es compatible con la hipótesis de balance de dosis.....	17
4.3.2. Las duplicaciones SSD (en tándem y proximales) son la principal fuente de genes específicos de Moringa .....	19
4.4. Rutas y clústeres biosintéticos de metabolitos secundarios en el genoma de Moringa... ..	20
<b>5. Conclusiones</b> .....	<b>29</b>
<b>6. Anexos</b> .....	<b>30</b>

<b>7. Bibliografia .....</b>	<b>31</b>
------------------------------	-----------

## Índice de tablas

Tabla 1 .....	5
Tabla 2 .....	6
Tabla 3 .....	7
Tabla 4 .....	8
Tabla 5 .....	12
Tabla 6 .....	18
Tabla 7 .....	20

## Índice de figuras

Figura 1 .....	3
Figura 2 .....	4
Figura 3 .....	15
Figura 4 .....	16
Figura 5 .....	17
Figura 6 .....	19
Figura 7 .....	22
Figura 8 .....	23
Figura 9 .....	26
Figura 10 .....	27
Figura 11 .....	28

## Resumen

*Moringa oleifera* (Moringa) es un cultivo arbóreo altamente nutritivo, de rápido crecimiento y tolerante a la sequía, al que se suele denominar el árbol multipropósito. Para este estudio, y dado el amplio uso y cultivo de la Moringa, se generó una versión del genoma de Moringa a escala cromosómica casi completa. Primero, se realizó una anotación funcional completa empleando términos de la Ontología de Genes (GO), códigos de la Comisión de Enzimas (EC), dominios funcionales INTERPRO y números de Ortología KEGG (KO). Esta versión del genoma de Moringa fue aprovechada para investigar los principales mecanismos responsables de la evolución de familias génicas que pueden estar en el origen de innovaciones biológicas relevantes, incluyendo caracteres agronómicos favorables. Para ello, se generaron dos clasificaciones: i) una clasificación de ortogrupos/familias génicas basada en la comparación del conjunto completo de proteínas codificadas por el genoma de la moringa y los de 10 especies que representan los principales linajes de angiospermas; ii) una clasificación de genes duplicados basada en mecanismos de duplicación (genoma completo/WGD, en tándem, proximal, disperso). Los resultados revelan que los duplicados WGD en Moringa están enriquecidos para funciones sensibles al balance de dosis. Además, los duplicados en tándem parecen haber desempeñado un papel destacado en la formación de familias génicas específicas de Moringa y en la evolución de rutas específicas de metabolismo secundario, incluidas las implicadas en la biosíntesis de compuestos glucosinolatos, flavonoides y alcaloides bioactivos, así como de rutas de respuesta de defensa. Algunos de estos genes implicados en el metabolismo secundario se encontraron dispuestos como clústeres génicos formados por genes homólogos y no homólogos que ocupan posiciones vecinas del genoma. Además, la adquisición de estas rutas podría explicar, al menos en parte, la destacada plasticidad fenotípica atribuida a esta especie. Este estudio proporciona una hoja de ruta genética para guiar futuros programas de mejora genética, especialmente los destinados a mejorar los caracteres relacionados con el metabolismo secundario en Moringa.

**Palabras clave:** Moringa, Evolución, Genoma, Genes Duplicados, Duplicaciones en Tándem, Plasticidad, Metabolitos Secundarios

## Summary

*Moringa oleifera* (Moringa) is a highly nutritious, fast growing and drought tolerant tree crop, often referred to as the multipurpose tree. For this study, and given the extensive uses and culture of Moringa, a nearly complete chromosome-scale version of the Moringa genome was generated. Firstly, a complete functional annotation was performed employing Gene Ontology (GO) terms, Enzyme Commission (EC) codes, INTERPRO functional domains and KEGG Orthology (KO) numbers. This version of the Moringa genome was leveraged to investigate main mechanisms of gene family evolution that may be at the origin of relevant biological innovations including agronomical favorable traits. For this purpose, two classifications were generated: i) an orthogroup/gene family classification based on the comparison of the complete set of proteins encoded by the moringa genome and those of 10 species representing the main angiosperm lineages; ii) a gene duplicate classification based on mechanisms of duplication (whole genome, tandem, proximal, dispersed). The results reveal that whole genome duplicates in Moringa are enriched for dosage balance-sensitive functions. Furthermore, tandem duplicates seem to have played a prominent role in the formation of Moringa specific gene families and in the evolution of specific secondary metabolism pathways, including those involved in the biosynthesis of bioactive glucosinolate, flavonoid and alkaloid compounds, as well as of defense response pathways. Some of these genes involved in secondary metabolism were found arranged as gene clusters formed by homologous and non-homologous genes occupying neighboring positions of the genome. Moreover, the acquisition of these pathways might, at least partially, explain the outstanding phenotypic plasticity attributed to this species. This study provides a genetic roadmap to guide future breeding programs, especially those aimed at improving secondary metabolism related traits in Moringa.

**Keywords:** Moringa, Evolution, Genome, Gene Duplicates, Tandem Duplications, Plasticity, Secondary Metabolites



## 1. Introducción

### 1.1. Antecedentes, motivación y enfoque del estudio

*Moringa oleifera* (Moringa en lenguaje vulgar) es un cultivo que ha ganado relevancia en las últimas décadas por sus múltiples cualidades y propiedades, las cuales han llegado a ser promocionadas por la *Food and Agriculture Organization* (FAO) para diversos usos (<https://www.fao.org/traditional-crops/moringa/en/>). Es considerado un cultivo altamente nutritivo y de rápido crecimiento; se le atribuye una gran plasticidad morfológica y bioquímica que le permite adaptarse a entornos locales muy diferentes y tolerar distintas condiciones de estrés (Araújo et al., 2016; Brunetti et al., 2018, 2020).

Un primer ensamblaje de su genoma fue publicado en 2015 (Tian et al., 2015). Posteriormente, se publicaron otras dos versiones más en los años 2019 y 2021, respectivamente (Chang et al., 2019; Shyamli et al., 2021). Sin embargo, la identificación inequívoca de genes de interés agronómico y de marcadores moleculares asociados y, en última instancia, el desarrollo de programas de mejora vegetal asistidos por genómica enfocados en *Moringa*, requiere de ensamblajes genómicos que capturen la completitud de su genoma. Para satisfacer esta demanda, se generó una nueva versión a escala cromosómica casi completa del genoma de *Moringa* (cuyo nombre oficial es AOCC v2) que será caracterizada en el presente estudio.

Asimismo, el estudio de los mecanismos evolutivos que moldean la estructura y función de un genoma es esencial para explicar el origen de adaptaciones a factores ambientales y caracteres fenotípicos relevantes que, a nivel genético, supone la formación de nuevos genes y nuevas familias génicas que originan funciones especializadas (Panchy et al., 2016). Los más importantes son las duplicaciones del genoma completo (*Whole Genome Duplications*, WGD) y las duplicaciones a pequeña escala (*Small Scale Duplications*, SSD). Son la principal fuente de nuevos genes nucleares. Aunque la mayoría de los genes duplicados son eliminados o se convierten en pseudogenes (es decir, genes que han perdido la capacidad de codificar para un producto funcional), algunos duplicados se subespecializan (subespecialización) o adquieren nuevas funciones (neofuncionalización) mientras que otros retienen la función de sus respectivos parálogos (Panchy et al., 2016). Para esta última situación se ha reportado un patrón diferencial de pérdida y retención de genes duplicados según el mecanismo de duplicación que ha intentado ser explicado con distintos modelos como la hipótesis del balance de dosis (Freeling, 2009; Tasdighian et al., 2017).

En la literatura se puede encontrar una amplia variedad de estudios de genómica que, aplicando distintas metodologías, han podido estudiar estos mecanismos evolutivos (duplicaciones WGD y SSD) como es el caso de (Denoëud et al., 2014), el cual reveló el origen polifilético de la ruta biosintética de la cafeína y su disposición como grupos de genes metabólicos duplicados en tándem (un tipo de duplicación SSD) y la existencia de expansiones significativas en genes relacionados con la defensa frente a patógenos y enzimas involucradas en la síntesis de alcaloides y flavonoides.

Por ello, también se investigarán mediante aproximaciones de genómica evolutiva y comparativa los principales mecanismos responsables de la evolución de familias génicas (duplicaciones WGD y SSD), los cuales pueden estar en el origen de innovaciones biológicas relevantes y de caracteres agronómicos favorables en *Moringa*.

## 1.2. Presentación de *Moringa oleifera*

### 1.2.1. Biología

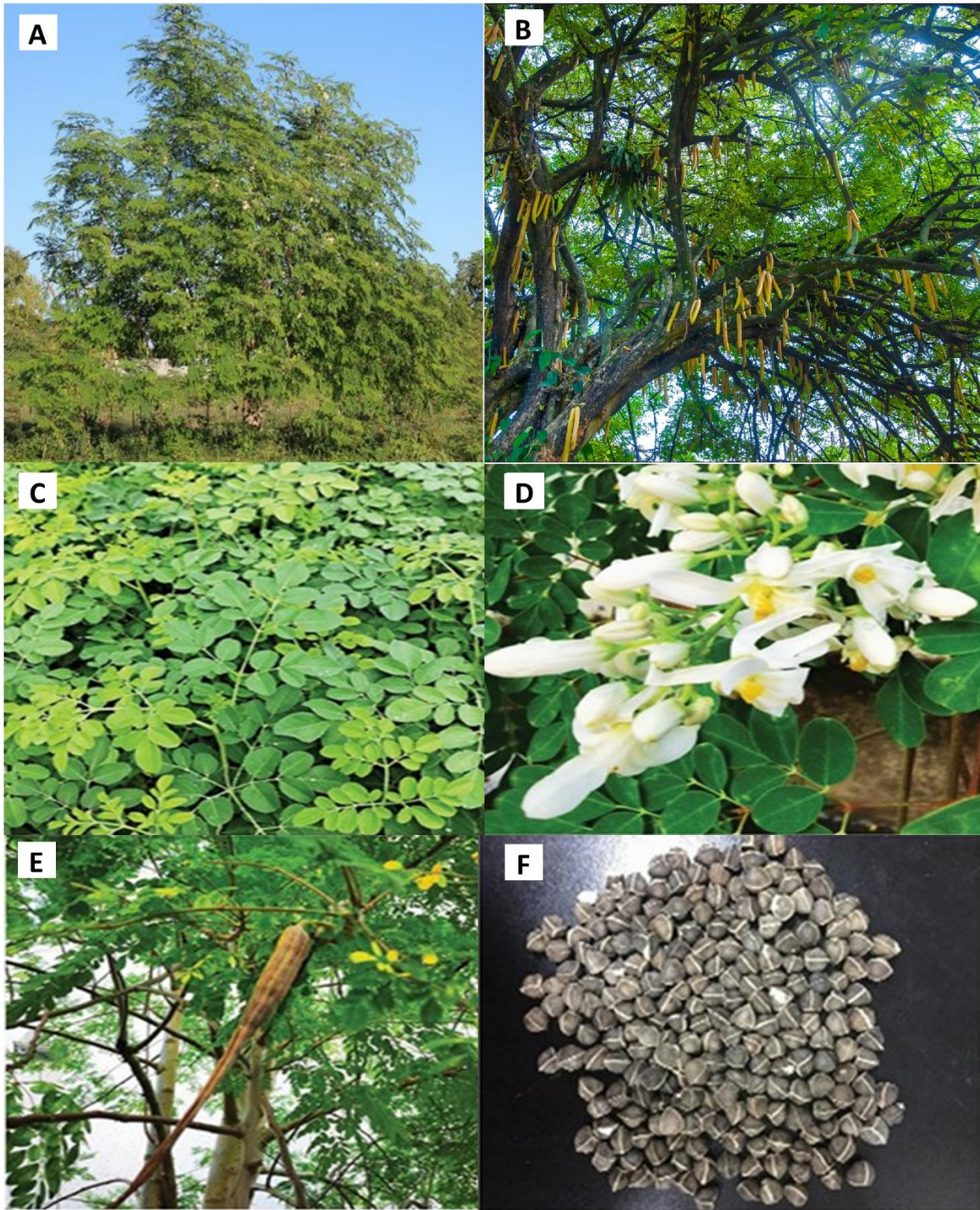
*Moringa* es un árbol perenne que suele presentar un hábito de ramificación baja con una copa extendida (**Figura 1A**); puede identificarse por sus hojas parcial o totalmente tripinnadas (aunque también produce hojas alternas o bipinnadas) (**Figuras 1B y 1C**) y sus largos frutos (**Figura 1B**). La corteza del árbol es lisa, el tronco exuda una goma blanca opaca cuando está herido.

La inflorescencia de *Moringa* es un racimo terminal de 10-15 cm de longitud, que nace en la base de la hoja y se dispersa en los extremos de las ramas. Las flores son fragantes y hermafroditas; se producen en forma de racimos extendidos o caídos sobre tallos delgados y peludos y con cinco sépalos y cinco pétalos desiguales doblados o curvados (**Figura 1D**). La floración comienza seis meses después de la plantación. Cuando hay un patrón estacional uniforme de temperaturas y precipitaciones, el pico de floración se produce dos veces al año mientras que en climas fríos solo se produce una vez, aunque en algunos lugares se ha reportado la formación de flores durante todo el año (Raja et al., 2013).

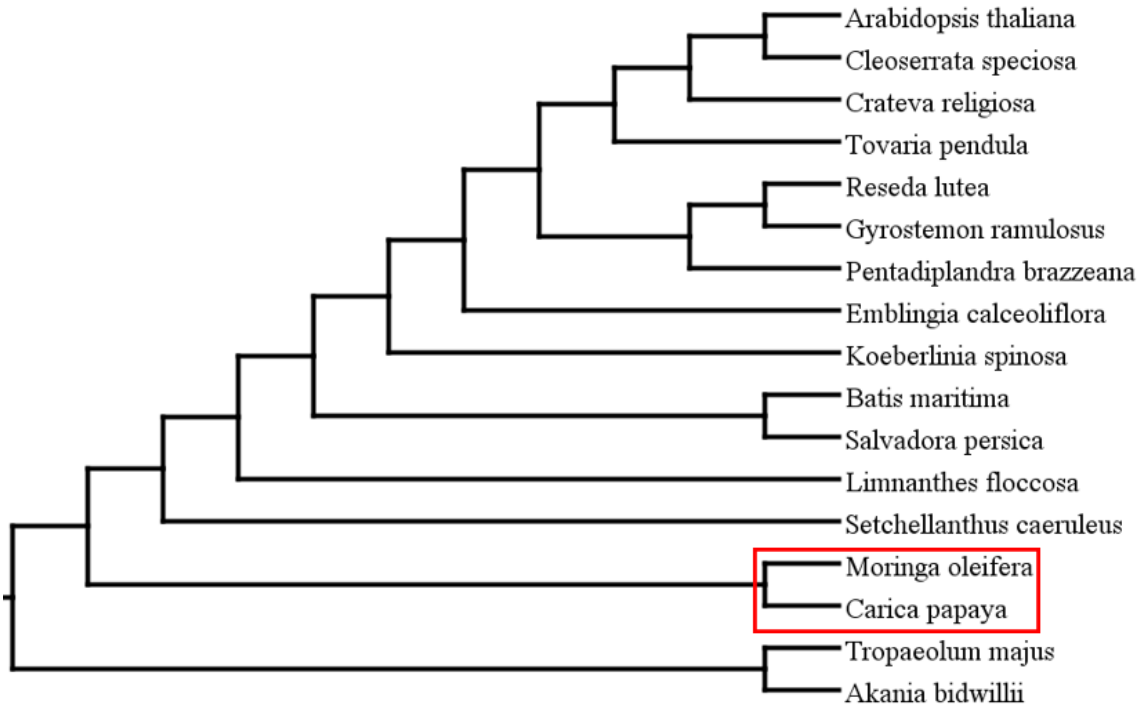
Los frutos de *Moringa* son colgantes, alargados (miden entre 30 y 50 cm), lineales con una sección transversal triangular y su color puede ser marrón o beige a grisáceo cuando están completamente maduros (**Figura 1E**). Las semillas son esféricas, negras (a veces marrones), de 7-8 mm de diámetro, con tres o cuatro alas que facilitan la dispersión por viento y agua (**Figura 1F**).

*Moringa* es posiblemente originaria de las tierras bajas de las colinas subhimalayas en el noroeste de la India (Olson, 2017). Junto a las otras 12 especies pertenecientes al género *Moringa*, conforma la familia taxonómica monogenérica Moringaceae (Olson, 2002). A su vez, Moringaceae forma un clado hermano con Caricaceae (**Figura 2**), la familia a la que pertenece *Carica papaya* (papaya), en base a análisis filogenéticos y ontogenéticos (Edger et al., 2018; Olson, 2003). Ambas familias forman parte del orden Brassicales (**Figura 2**) en base a los dos estudios anteriormente citados y por su capacidad de sintetizar glucósidos de aceite de mostaza (Rodman et al., 1998).

Se ha reportado una extensa variabilidad genética en *Moringa* como consecuencia de la polinización cruzada, las diversas condiciones agroecológicas a las que ha estado expuesta, la migración de material genético debido a la deriva genética, el flujo de genes, la introducción/intercambio de reservas genéticas (*genetic stocks*) a nivel nacional e internacional, y de la selección artificial adoptiva e intensiva. Por consiguiente, presenta una alta heterogeneidad acompañada de una alta diversificación en muchos caracteres (como la floración) lo cual implica una amplia diversidad de propiedades morfológicas y bioquímicas que sirven de recursos para su mejora genética (Godino et al., 2017). En la India ha sido posible identificar hasta nueve ecotipos (A. Kumar et al., 2014), ocho cultivares (Ram et al., 2020) y 36 genotipos los cuales fueron agrupados en cinco grupos según el contenido en vitamina C, proteínas, nitrógeno, fósforo, potasio, calcio, hierro y magnesio (Tak & Maurya, 2017).



**Figura 1. Aspecto general de Moringa y de sus partes.** A. Vista del árbol desde lejos. B. Vista del árbol desde su base. C. Hojas. D. Flores. E. Fruto (vaina). F. Semillas.



**Figura 2. Filogenia del orden Brassicales.** Árbol filogenético ultramétrico de 17 especies basado en los resultados de Edger et al., 2018 donde se analizaron 72 genes plásticos de distintas especies de cada familia perteneciente al orden Brassicales. El clado formado por *M. oleifera* y *C. papaya* está indicado.

### 1.2.2. Cultivo, propiedades y usos

*Moringa* es cultivada ampliamente desde la antigüedad en todas las zonas tropicales y subtropicales cálidas y semiáridas del mundo, concretamente en África, Asia, América, el Caribe y el Pacífico (Gandji et al., 2018). Asimismo, su cultivo se está extendiendo a otras zonas caracterizadas por la escasa disponibilidad de nutrientes y agua en el suelo y las altas temperaturas, incluida la cuenca del mediterráneo (Trigo et al., 2020; Vaknin & Mishal, 2017).

Conocida por distintos apodos (árbol multipropósito, árbol milagroso, árbol de baquetas...), diferentes partes de la planta de *Moringa* (hojas, flores, vainas, semillas) poseen un alto contenido nutricional (Olson et al., 2016), destacando las hojas que contienen varios aminoácidos esenciales, proporcionando una fuente de proteínas alternativa para satisfacer la demanda habitual de las personas desnutridas; vitaminas A, B, C, D y E, suministrando las cantidades dietéticas requeridas y minerales como Ca, Fe, K, Se, Zn o Mg entre otros (**Tabla 1**) (Islam et al., 2021; Leone et al., 2015; Trigo et al., 2020). Por ello, tiene una alta valoración como alimento para consumo humano y animal (**Tabla 2**).





Además, la composición química de sus distintas partes también incluye una amplia y diversa gama de metabolitos secundarios para los que se están estudiando diversas funciones farmacológicas como compuestos bioactivos, encontrándose flavonoides, carbamatos, fenoles, carotenoides, esteroides, y glucosinolatos. Otros compuestos de interés son alcaloides, isoprenoides, saponinas, taninos, ácidos fenólicos, ácidos clorogénicos y ácidos grasos (**Tabla 1**). Esta extensa colección de metabolitos secundarios le confiere a *Moringa* (especialmente los glucosinolatos) actividades farmacológicas muy diversas (**Tabla 2**).

**Tabla 1.** Componentes químicos y contenido nutricional de Moringa. Fuente: (Liu et al., 2022)

Parte	Componentes químicos	Macronutrientes	Micronutrientes
	Proteínas, carbohidratos, aminoácidos, minerales, flavonoides, carbamatos, fenoles, glucosinolatos, carotenoides, terpenoides, esteroides, ácidos orgánicos, aceites volátiles, alcaloides.	Proteínas (35%), fibra (23%), grasas (16%), cenizas (12%), carbohidratos (7%), humedad (7%).	Aminoácidos: histidina (0,730% ± 0,03%), isoleucina (1,155% ± 0,03%), leucina (2,070% ± 0,15%), lisina (1,540% ± 0,14%), metonina...  Vitaminas: β-caroteno, VA, VB-1 (0,006%), VB-2 (0,005%), VB-3 (0,08%), VB-5, VB-6, VB-9, VB-12, VC (22%), VE (44,8%).  Minerales: Ca (11 g/kg), Fe (132 mg/kg), K (29,60 g/kg), Zn (30,10 mg/kg), Se (1,570 mg/kg), Mg (3,73 g/kg)...
	Proteínas, vitaminas, aminoácidos, minerales, flavonoides, alcaloides, polifenoles, taninos, terpenoides (β-amirina), esteroides (β-sitosterol), saponinas, inhibidores de la tripsina, ácidos orgánicos.	Proteínas, ácidos orgánicos (6,42% ± 0,01%).	Aminoácidos (31%): alanina, argenina, ácido glutámico, glicina, serina, treonina, valina, lisina.  Vitaminas: VC, VA.  Minerales: Ca, K, antioxidantes de calcio (α- y γ-tocoferol).
	Proteínas, vitaminas, minerales, aminoácidos, flavonoides, carbamatos, fenoles, glucosinolatos, taninos, esteroides, carotenoides, ácidos orgánicos, carbohidratos, fibras, grasas.	Proteínas, ácidos orgánicos (6,42% ± 0,01%).	Aminoácidos (31%): alanina, arginina, ácido glutámico, glicina, serina, treonina, valina, lisina.  Vitaminas: VC, VA.  Minerales: Ca, K, antioxidantes del calcio (α- y γ-tocoferol).
	Proteínas, carbohidratos, vitaminas, aminoácidos, minerales, grasa bruta, fibra, cenizas, aceite de behen, ácidos orgánicos, flavonoides, carbamatos, fenoles, glucosinolatos, esteroides, carotenoides, alcaloides.	Proteínas (35%), hidratos de carbono (25%), grasas (16%), fibra (16%), humedad (7%), cenizas (1%).	Aminoácidos: ácido glutámico (3,724% ± 0,18%), ácido aspártico (3,059% ± 0,02%), leucina (2,898% ± 0,22%), arginina (2,548% ± 0,08%), alanina (2,247% ± 0,46%)...  Vitaminas: VB-1 (0,005%), provitamina A (~2%), VB-2 (0,006%), VB-3 (0,02%), VC (0,45% ± 0,017%), VE (75,167% ± 0,441%), tocoferoles (α-, β-, γ- y δ)

Aparte de sus usos para la alimentación humana, el forraje y como fuente de compuestos farmacológicos activos, otras muchas utilidades son atribuidas a alguna parte o producto de Moringa: desde fertilizante hasta productor de biodiésel pasando por biopesticida o lubricante para maquinaria fina (Tabla 2).

Tabla 2. Usos atribuidos a Moringa. Fuente: (Liu et al., 2022)

Parte	Alimentación	Usos médicos	Usos en la agricultura y en la industria
	Alimento humano: verduras, sopas, condimentos, complementos alimenticios/fortificación de alimentos (galletas, yogur), alimentos de destete. Forraje.	Anemia, artritis, fiebre, cicatrización de la piel, bronquitis, antibacteriano, antifúngico, tumores, inflamaciones, helmintiasis, purgante, infecciones oculares y de oído...	Fertilizante, biopesticida, biogás, crema hidratante y acondicionador de la piel, bálsamo labial, cremas, promotores del crecimiento (zeatina), agente de limpieza doméstico, loción hidratante a base de hierbas/tóner facial/jabón.
	Alimento humano: vegetal, pasteles, miel (néctar de flor), enriquecimiento de alimentos, alimentos de destete.	Artritis, garganta, infección, inhibidor de la tripsina, antiparasitario, antiinflamatorio, antihipertensivo, diurético...	Biopesticida.
	Alimento humano: vegetal, fortificación de alimentos/complemento alimenticio.	Artritis, antiparasitario, tonifica el bazo y elimina la humedad, antiviral, antitumoral, hipotensor, hipocolesterolemia, obesidad, diarrea y dolor articular, antiinflamatorio, antioxidante y desintoxicante.	
	Alimento humano: vegetal, aceite de cocina, alimentos funcionales/fortificación de alimentos (pan), nutraceuticos. Forraje.	Tonifica el bazo y elimina la humedad, antipirético, antihipertensivo, antiinflamatorio, antitumoral, anticardiovascular.	Fertilizante orgánico, biodiésel, biosorbente, perfumes, loción para la piel, peluquería, cosméticos, lubricante para maquinaria fina, purificación del agua, clarificador de miel y zumo de caña de azúcar, agentes insecticidas.

*Moringa* puede crecer de forma excepcionalmente rápida a temperaturas que oscilan entre los 25 y 35°C; cuenta con un sistema radicular profundo y puede alcanzar hasta los tres metros de altura en solo tres meses y hasta 12-15 metros en pocos años (Devkota & Bhusal, 2020; Y. Kumar et al., 2017). Crece bien en suelos semisecos, cálidos y húmedos o franco-arenosos (Prajapati et al., 2022). Es especialmente tolerante a la sequía, el calor y la radiación UV-B, factores de estrés que se agravarán previsiblemente con el cambio climático global (Araújo et al., 2016; Brunetti et al., 2018, 2020).

### 1.2.3. Genoma

*Moringa* es una especie diploide; su genoma está formado por 14 pares de cromosomas (2n=28). Se estima que su tamaño total oscila entre 278 y 315 Mpb en base a análisis de distribución de *k*-mers (Chang et al., 2019) y de citometría de flujo (Tian et al., 2015) por lo que se trata de un genoma pequeño dentro del grupo de plantas arbóreas. La versión del genoma desarrollada para este estudio (AOCC v2) es la primera de las cuatro versiones existentes en la que ha sido posible ensamblar 14 pseudomoléculas equivalentes a sus cromosomas que, en conjunto, representan el 92,8% del ensamblaje genómico (en total, se ensamblaron 748 *scaffolds*). El ensamblaje tiene un tamaño de 236 Mpb, representando entre un 84,89 y un 74,92% del tamaño total estimado.

En este ensamblaje se identificaron mediante tres estrategias (predicción *ab initio*, búsqueda por homología, evidencia basada en experimentos de *RNA-seq*) 22.714 genes estructurales (es decir, genes que codifican para proteínas), 5.121 genes no estructurales (es decir, genes que codifican para moléculas de RNA distintas a mRNA) y 213.796 secuencias repetitivas (mayoritariamente, elementos transponibles) que representan un 32,09, un 1,46 y un 38,72% del mismo, respectivamente. Cuenta con unos indicadores de calidad mucho más altos que los de las tres versiones anteriores tal y como se puede observar en la **Tabla 3**.

**Tabla 3.** Estadísticas de los ensamblajes del genoma de *Moringa*. Los ensamblajes están ordenados de izquierda a derecha según su año de publicación.

Parámetro	Tian et al., 2015	AOCC v1 (Chang et al., 2019)	Shyamli et al., 2021	AOCC v2
Plataforma de secuenciación	Illumina HiSeq2500TM	Illumina HiSeq 2000	Pacbio sequel	Oxford Nanopore
Tamaño del ensamblaje, pb	289.241.074	216.759.177	281.946.330	236.366.566
Número de <i>scaffolds</i>	33.332	22.329	915	748
<i>Scaffold</i> N50, pb	114.476	957.246	4.719.167	14.962.574
<i>Scaffold</i> L50	61	56	17	7
<i>Scaffold</i> N90, pb	5.792	57.837	225.696	13.210.789
<i>Scaffold</i> L90	1.382	366	115	13
<i>Scaffold</i> más grande, pb	6.788.971	4.637.711	13,807,473	30.079.500
<i>Scaffold</i> más corto, pb	200	150	1.056	13
Longitud media de los <i>scaffolds</i> , pb	8.678	9.707,52	308.103,9	315.998,1
Contenido en GC (%)	36,5	36,5	37,82	35,7
Número total de nucleótidos no identificados (N)	1.821.349	3.014.085	700	346.401

En la **Tabla 3** puede observarse, por ejemplo, un valor de N50 (un valor que proporciona la longitud mínima necesaria del *contig* o *scaffold* a partir del cual es posible cubrir el 50% del genoma juntando *contigs* o *scaffolds* de igual o mayor tamaño que este) muy alto para AOCC v2 lo cual indica que su grado de fragmentación es bajo. Otro valor útil es el L90 que indica el menor número de *contigs* o *scaffolds* necesarios para cubrir el 90% del genoma que en el caso de AOCC v2 es 13 correspondiendo a 13 de los 14 cromosomas.

Por otro lado, una medida útil para comprobar el grado de completitud de un genoma secuenciado en términos de anotación estructural (esto es, la identificación de genes y otros elementos genómicos) consiste en examinar la completitud de los genes predichos. Para ello, se empleó el programa *Benchmarking Universal Single-Copy Orthologs* (BUSCO) v4.1.4 (Manni et al., 2021) que llevó a cabo una comparación con una base de datos de *core genes* (genes homólogos de una sola copia que están presentes en la mayoría de las especies conocidas de un grupo de organismos) de embriofitas. Las cifras registradas por BUSCO representan una mejora significativa en comparación a las otras versiones (**Tabla 4**).

**Tabla 4.** Estadísticas de la calidad de las anotaciones estructurales de los ensamblajes del genoma de *Moringa*. Los ensamblajes están ordenados de izquierda a derecha según su año de publicación.

<b>Core genes de BUSCO (1.440)</b>	<b>Tian et al., 2015</b>	<b>AOCC v1 (Chang et al., 2019)</b>	<b>Shyamli et al., 2021</b>	<b>AOCC v2</b>
Completos (%)	-	83	95,9	99,8
Únicos (%)	-	82,1	94,4	98,8
Duplicados (%)	-	0,9	1,5	1,0
Fragmentados (%)	-	5,3	1,3	0,1
No encontrados (%)	-	11,7	2,8	0,1



## 2. Objetivos

El presente estudio está articulado en cuatro objetivos:

### 2.1. Realizar la anotación funcional del genoma de Moringa

Se llevará a cabo la asignación de hipotéticas funciones moleculares y biológica a cada uno de los genes estructurales de Moringa identificados en AOCC v2 mediante términos funcionales para caracterizar sus respectivas identidades biológicas.

### 2.2. Generar clasificaciones de ortogrupos/familias génicas y de genes duplicados en Moringa

Se estudiará la evolución de familias génicas mediante una clasificación de ortogrupos que agrupa genes estructurales del genoma Moringa (AOCC v2) y de los genomas de otras 10 especies vegetales que representan los principales linajes evolutivos de plantas angiospermas. Asimismo, dado que la principal fuente de nuevos genes y nuevas familias génicas son genes resultantes de duplicaciones, también se generará una clasificación de genes duplicados en base a mecanismos de duplicación.

### 2.3. Determinar la contribución de las duplicaciones WGD y SSD a la expansión y/o retención de determinadas funciones biológicas

El destino evolutivo y funcional de un gen duplicado varía notablemente según el mecanismo de duplicación. Se estudiarán las funciones moleculares de los genes resultantes de duplicaciones WGD y SSD para determinar qué tipos de funciones biológicas son expandidas y/o retenidas según el tipo de evento de duplicación que los origina.

### 2.4. Investigar el rol de las duplicaciones en tándem en la evolución de la respuesta de defensa y las rutas del metabolismo secundario vegetal

Diversos estudios (Carretero-Paulet & Fares, 2012; Chae et al., 2014; Denoeud et al., 2014) han reportado el rol de las duplicaciones en tándem en la promoción de la plasticidad genómica. Por ello, se examinará su papel en la evolución de la respuesta de defensa al estrés y del metabolismo secundario, relevantes para explicar la destacada plasticidad fenotípica de Moringa.

### 3. Materiales y métodos

#### 3.1. Genoma y transcriptoma

El material de partida de este estudio fue el propio ensamblaje genómico AOCC v2, el cual fue generado en el *Seed Biotechnology Center*, afincado en la Universidad de Davis (California), por miembros del grupo de investigación dirigido por el Director Científico del Consorcio de Cultivos Africanos Huérfanos (<https://africanorphancrops.org/>), el Dr. Allen Van Deynze (<http://africanorphancrops.org/allen-van-deynze/>), quien ejerció de supervisor. El DNA genómico fue extraído de una accesión de *Moringa* (la accesión Mtongwe1) recolectada en Mtongwe (Kenia) el 26-8-16 por el Instituto de Investigación Forestal de Kenia. La plataforma de secuenciación empleada fue Oxford Nanopore (**Tabla 3**).

El otro material de partida utilizado fueron datos de anotación estructural y datos de expresión génica en cinco tejidos (tallo, hoja, flor, semilla y raíz) medidos en transcritos por millón (TPM) y obtenidos a partir de experimentos de *RNA-seq* almacenados en un repositorio público administrado por el *National Center for Biotechnology Information* (NCBI) (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA394193>). Estos datos fueron generados en el Centro para Biología de Sistemas Vegetales (<https://vib.be/vib-ugent-center-plant-systems-biology>), afincado en Gante (Bélgica), por miembros del grupo de investigación dirigido por el Prof. Dr. Yves Van de Peer, quien ejerció de supervisor. Los datos se encuentran disponibles en la base de datos *Online Resource for Community Annotation of Eukaryotes* (ORCAE) a través del siguiente enlace: [https://bioinformatics.psb.ugent.be/gdb/aocc/morolbgi/Moringa\\_version2/](https://bioinformatics.psb.ugent.be/gdb/aocc/morolbgi/Moringa_version2/). En esta base de datos también se encuentra depositado el material generado durante la realización de este estudio.

#### 3.2. Anotación funcional del genoma

La anotación funcional de los genes estructurales identificados en AOCC v2 fue realizada mediante análisis comparativo (alineamiento) de secuencias proteicas empleando bases de datos de acceso público que resultó en la asignación de los siguientes términos funcionales:

- Términos de la Ontología de Genes (*Gene Ontology*, GO): la Ontología de Genes es un vocabulario controlado y estandarizado de términos numerados 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, Componente Celular, Función Molecular).
- Términos INTERPRO: son los identificadores que emplea la base de datos InterPro, administrada por el *InterPro Consortium* (<http://www.ebi.ac.uk/interpro/>) (Mitchell et al., 2019). Describen los dominios funcionales de una proteína; también se usan para agrupar proteínas en clasificaciones precomputadas de familias y superfamilias.
- Códigos de la Comisión de Enzimas (*Enzyme Commission*, EC): son códigos numéricos que describen distintos aspectos de una enzima como su sustrato.
- Números de Ortología del KEGG (*KEGG Orthology*, KO): son los identificadores que emplea la base de datos *KEGG Orthology*, administrada por la Enciclopedia Kyoto de Genes y Genomas (*Kyoto Encyclopedia of Genes and Genomes*, KEGG). Describen funciones moleculares representadas en términos de ortólogos funcionales definidos a partir de genes y proteínas caracterizados experimentalmente en organismos específicos.

Las proteínas codificadas por el genoma de *Moringa* fueron anotadas con términos GO mediante BLAST2GO versión 6.0.1 (Conesa & Götz, 2008). BLAST2GO genera una anotación funcional en tres pasos: inferencia de homología basada en la similitud de la secuencia de consulta (la proteína problema) con respecto a las proteínas de la base de datos *non-redundant* (nr) administrada por el NCBI, la extracción de los términos GO asociados a las hipotéticas proteínas homólogas (*hits*) identificadas (mapeo) y la anotación de dichos términos GO a la proteína problema. BLAST2GO permite ampliar los términos GO con dominios funcionales INTERPRO y códigos EC representados en rutas bioquímicas identificadas en el KEGG. En este estudio, se utilizaron Diamond versión 2.0.11 con un valor E de  $1 \times 10^{-10}$  como umbral (Buchfink et al., 2015) e INTERPROSCAN versión 5.52-86.0 con los parámetros por defecto (P. Jones et al., 2014) para identificar los *hits* y los dominios funcionales, respectivamente.

Para la anotación con números KO, se utilizó el programa BlastKOALA (<https://www.kegg.jp/blastkoala/>) (Kanehisa et al., 2016). Solo se asignó a cada gen el número KO con mejor puntuación, lo que dio como resultado un máximo de un término de KO por gen, excepto para aquellos anotados con dos o más números KO que obtuvieron la puntuación más alta. Sin embargo, el mismo número KO puede estar implicado en varias rutas KEGG y también puede encontrarse anotando más de un gen.

### 3.3. Clasificación de ortogrupos/familias génicas

Las secuencias proteicas de *Moringa* y de otras 10 especies de plantas angiospermas (**Tabla 5**) fueron comparadas todas contra todas empleando Diamond (Buchfink et al., 2015) con un umbral bajo (valor E de  $1 \times 10^{-3}$ ) para no filtrar secuencias muy cortas. A continuación, se clasificaron en ortogrupos utilizando el algoritmo de agrupación implementado en OrthoFinder versión 2.5.2 bajo la configuración predeterminada (Emms & Kelly, 2019). Se utilizó el método de clasificación de ortogrupos basado en la inferencia de ortogrupos jerárquicos a partir de análisis de árboles enraizados de genes usando un árbol ultramétrico de especies como referencia, el cual es considerado mucho más preciso que el método basado en similitud de genes/grafos utilizado anteriormente por OrthoFinder (Emms & Kelly, 2019).

Cada planta seleccionada representa un linaje evolutivo dentro del grupo de las plantas con flores. Fueron incluidas las brassicales *Arabidopsis thaliana* (*Arabidopsis*) y *Carica papaya* (papaya), la legumbre *Medicago truncatula*, la rósida basal *Vitis vinifera* (uva), la astérida *Solanum lycopersicum* (tomate), la dicot basal *Nelumbo nucifera*, las monocots *Oryza sativa* cv. Japonica (arroz) y *Zea mays* (maíz), la magnólida *Persea americana* (aguacate) y el arbusto perennifolio *Amborella trichopoda*. Para dotar de raíz al árbol de especies, se utilizó como grupo externo (*outgroup*) a *A. trichopoda* por su condición de angiosperma basal a todas las plantas con flores (Albert et al., 2013). La topología y los tiempos de divergencia del árbol filogenético se obtuvieron de TimeTree (S. Kumar et al., 2017), a excepción de la posición de la rama correspondiente a *P. americana* que fue asignada manualmente debido a la existencia de hipótesis evolutivas contradictorias con respecto a la posición de las magnólidas dentro de la filogenia de las plantas angiosperma. Se utilizó la hipótesis evolutiva propuesta en (Rendón-Anaya et al., 2019) que favorecía la ubicación taxonómica de las magnólidas como clado hermano al superclado formado por dicotiledóneas y monocotiledóneas.

**Tabla 5.** Versiones de los genomas de las especies seleccionadas para generar la clasificación de ortogrupos. A cada especie se le asignó una abreviatura.

Espece	Versión del genoma	Tamaño del ensamblaje genómico (Mpb)	Número de genes identificados
<i>Arabidopsis thaliana</i> (Ath)	Araport11	120	27.655
<i>Amborella trichopoda</i> (Atr)	JGI v1.0	706	26.846
<i>Carica papaya</i> (Cpa)	ASGPB v0.4	342	27.768
<i>Moringa oleifera</i> (Mol)	AOCC v2	236	22.714
<i>Medicago truncatula</i> (Mtr)	JGI Mt4.0 v1	412	50.894
<i>Nelumbo nucifera</i> (Nnu)	LOTUS-DB v1.1	1.075	26.685
<i>Oryza sativa ssp. Japonica</i> (Osj)	JGI v7.0	374	42.189
<i>Persea americana cv. Hass</i> (Pah)	COGE (Genome Id: 29302)	913	24.616
<i>Solanum lycopersicum</i> (Sly)	ITAG v2.4	824	34.725
<i>Vitis vinifera</i> (Vvi)	JGI v2.1	486	26.346
<i>Zea mays</i> (Zma)	AGP v4.0	2.135	39.498

### 3.4. Clasificación de genes duplicados en base a modos de duplicación

El programa *Duplicate\_gene\_classifier* del kit de herramientas *Multiple collinearity Scan* (MCScanX, <https://github.com/wyp1125/MCScanX>) (Wang et al., 2012) fue empleado usando los parámetros por defecto para clasificar los genes duplicados en el genoma de *Moringa* según su posición lo que permite identificar a su vez el mecanismo de duplicación: WGD/segmental o SSD. A su vez, las duplicaciones SSD agrupan duplicaciones en tándem, proximales y dispersas.

- Duplicación WGD/segmental: todo el genoma experimentó una duplicación o grandes bloques genómicos experimentaron duplicaciones. Cualquiera de estas dos situaciones da lugar a genes colineales en bloques sinténicos.
- Duplicación en tándem: es el resultado de entrecruzamientos desiguales de alelos no homólogos, entre otros mecanismos. El gen duplicado ocupa una posición adyacente con respecto a su parálogo.
- Duplicación proximal: se originó posiblemente a través de actividades de transposones localizados o a partir de antiguos duplicados en tándem que fueron separados por la inserción de otros genes. El gen duplicado ocupa una posición cercana pero no adyacente con respecto a su parálogo (20 genes o menos los separan; este umbral es arbitrario y puede ser fijado en MCScanX).
- Duplicación dispersa: ocurrió a través de patrones impredecibles y aleatorios por mecanismos aun no esclarecidos. El gen no ocupa una posición adyacente o cercana a su parálogo.

Para generar la clasificación, primero hubo que identificar pares de proteínas parálogas. La identificación se realizó mediante una búsqueda de todos contra todos (esto es, se llevaron a cabo alineamientos entre todas las proteínas de Moringa) en Diamond con un valor E de corte de  $1 \times 10^{-10}$ . Una vez obtenida, se realizaron tests de enriquecimiento de conjuntos de genes mediante prueba exacta de Fisher en los genes duplicados con el objeto de encontrar funciones moleculares y biológicas sobrerrepresentadas, así como detectar sobrerrepresentación en ortogrupos específicos y genes del metabolismo secundario de Moringa. Los valores p ( $P$ ) resultantes fueron corregidos con método de Bonferroni (Bonferroni, 1936) para resolver el problema más común del testeo de múltiples hipótesis (*family-wise error rate*): que la hipótesis nula sea verdadera para todas las comparaciones simultáneamente o, alternativamente, que la hipótesis nula sea falsa para al menos una prueba. Solo se consideraron estadísticamente significativos aquellos  $P$  corregidos por método de Bonferroni con un valor inferior o igual a 0,05.

### 3.5. Identificación y caracterización de rutas biosintéticas y de clústeres de genes de metabolismo secundario.

Se intentó reconstruir el conjunto completo de genes presuntamente implicados en la biosíntesis de glucosinolatos y su regulación, especialmente aquellos que codifican actividades enzimáticas, en el genoma de Moringa y los otros dos representantes de Brassicales examinados en este estudio, es decir, Arabidopsis y papaya. Para ello, se fusionaron los términos GO, KEGG y EC obtenidos de la anotación funcional de los genes estructurales de Moringa que están relacionados con el metabolismo de glucosinolatos.

Asimismo, para detectar posibles clústeres de genes biosintéticos en el genoma de Moringa que estuvieran asociados a la biosíntesis de metabolitos secundarios, se utilizó la herramienta online plantiSMASH v1.0 (<http://plantismash.secondarymetabolites.org/>) con los parámetros por defecto (Kautsar et al., 2017). La definición arbitraria de un clúster de genes metabólicos requiere que el clúster contenga genes de al menos tres genes de dos tipos diferentes (los genes duplicados estrechamente relacionados solo son contados una vez). Posteriormente, ciertos genes encontrados en clústeres biosintéticos que codifican para reacciones clave en la biosíntesis de metabolitos secundarios importantes para las propiedades farmacológicas y nutricionales de Moringa fueron seleccionados para estudiar los ortogrupos que formaron junto con los hipotéticos genes ortólogos de las otras 10 especies vegetales escogidas para la clasificación de OrthoFinder (en concreto, los ortogrupos HOG0001451, HOG0003249 y HOG0000748). Por un lado, se examinaron los genes ortólogos de Arabidopsis y, por otro, se realizó un análisis filogenético de los ortogrupos seleccionados a partir de alineamientos múltiples de secuencias proteicas ejecutados mediante el algoritmo MUSCLE (Edgar, 2004) a través del programa SeaView versión 4.6.4 (Gouy et al., 2010). A continuación, se obtuvieron árboles filogenéticos hechos mediante un método de máxima verosimilitud usando la versión online 1.6.12 del programa IQ-TREE (<http://iqtree.cibiv.univie.ac.at/>) (Trifinopoulos et al., 2016). Antes del análisis, IQ-TREE realiza una selección automatizada del modelo de sustitución de aminoácidos que mejor se ajusta según el criterio de información bayesiano. JTTDCMut+I+G4, JTT+G4 y JTT+F+I+G4 fueron elegidos como los modelos de sustitución de aminoácidos que mejor se ajustan para los ortogrupos HOG0001451, HOG0003249 y HOG000074882, respectivamente (D. T. Jones et al., 1992). JTT se refiere a los autores de los modelos (Jones-Taylor-Thornton); I a la proporción de sitios invariantes; F a las frecuencias empíricas de aminoácidos y G4 a la heterogeneidad en las tasas de sustitución modelada mediante una distribución gamma con cuatro categorías. Se emplearon tres análisis independientes de

soporte de rama (análisis bootstrap ultrarrápido con 1.000 réplicas, prueba de ramas SH-aLRT y prueba de Bayes aproximada) para evaluar la fiabilidad de las ramas internas.

### 3.6. Representaciones gráficas y tests estadísticos

El lenguaje de programación R versión 4.2.0 (R Core Team, 2020) fue utilizado para generar la mayoría de las figuras. El código para generar cada figura fue redactado y editado en el entorno de desarrollo integrado RStudio versión 2022.02.0+443 (RStudio Team, 2019). Asimismo, fueron empleadas las siguientes extensiones de R (también llamadas paquetes):

- ggplot2 (Wickham, 2016): **Figuras 3, 4A** (subpanel derecho), **5 y 6**.
- ggpubr (Kassambara, 2020): **Figura 6**.
- scales (Wickham & Seidel, 2020): **Figura 6**.
- UpSetR (Gehlenborg, 2019): **Figura 4B**.
- ComplexHeatmap (Gu et al., 2016): **Figuras 8, 9C, 10C y 11C**.

Para los diagramas de burbuja de la **Figura 6**, el código base fue proporcionado por REVIGO (Supek et al., 2011), un servidor web (<http://ReViGo.irb.hr/>) para la visualización y reducción de listas de términos GO. REVIGO realiza 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. 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 y Conrath) quien también puede seleccionar 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 opción por defecto). En este caso, fueron seleccionados el valor y la medida que emplea el programa por defecto (0,7 y SimRel) y las entradas de UniProt para Arabidopsis como base de datos.

Para editar los árboles de especies y de genes de las **Figuras 4A** (subpanel izquierdo), **9B, 10B y 11B** se utilizó el programa online iTol versión 5 (<https://itol.embl.de/>) (Letunic & Bork, 2021) mientras que para generar el árbol filogenético de especies de la **Figura 2** se empleó Mesquite versión 3.70 (<https://github.com/MesquiteProject/MesquiteCore>) (Maddison & Maddison, 2021). Por otro lado, las representaciones gráficas de clústeres de genes (**Figuras 9A, 10A y 11A**) fueron hechas en la herramienta online Gene Graphics (<https://katlabs.cc/genegraphics/>) (Harrison et al., 2018).

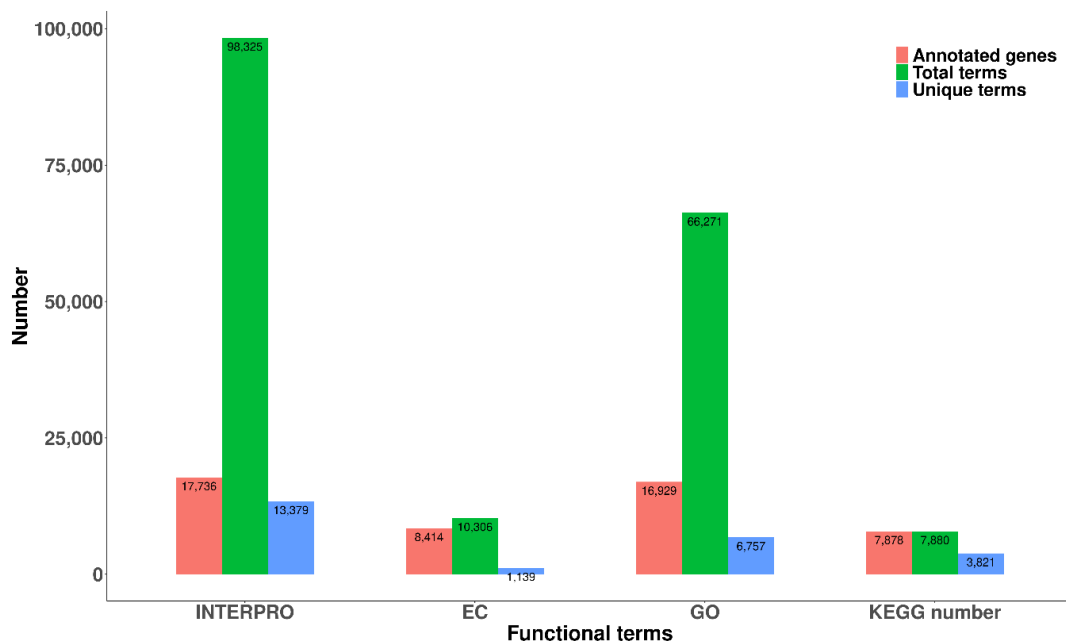
Por último, los tests de enriquecimiento para funciones moleculares y biológicas en genes duplicados se realizaron con códigos del lenguaje de programación PERL versión 5 (Wall et al., 2000) mientras que los otros tests para detectar enriquecimiento de duplicados en genes del metabolismo secundario y ortogrupos específicos de Moringa se realizaron en R versión 4.2.0 (R Core Team, 2020) mediante el paquete GeneOverlap (Shen & Icahn School of Medicine at Mount Sinai, 2021)

## 4. Resultados y discusión

### 4.1. Anotación funcional del genoma de *Moringa* cubre más del 80% de sus genes

Las búsquedas en la base de datos de proteínas nr del NCBI de hipotéticos genes homólogos con respecto a los 22.714 genes estructurales identificados en AOCC v2 mediante Diamond permitió identificar al menos un *hit* para 19.374 genes, el 85,30%. En cuanto a la anotación con términos funcionales, 18.460 genes (81,27%) fueron anotados con alguno de los siguientes términos:

- 17.736 proteínas (78,08%) mostraron al menos un dominio funcional INTERPRO sumando un total de 98.325 términos INTERPRO (de los cuales 13.779 son únicos) (**Figura 3**) lo que representa un promedio de seis términos INTERPRO por proteína.
- A 8.414 genes (37,04%) que codifican enzimas se les fue asignado al menos un código EC de un total de 10.306 (de los cuales 1.139 son únicos) (**Figura 3**), lo que representa un código EC por enzima.
- 16.929 genes (74,53%) fueron anotados con al menos un término GO, sumando un total de 66.271 términos GO (de los cuales 6.757 son únicos) (**Figura 3**), lo que representa un promedio de tres términos GO por gen anotado.
- 7.878 genes (34,68%) fueron mapeados en un total de 3.821 grupos de ortología funcional (números KO) del KEGG únicos (**Figura 3**).

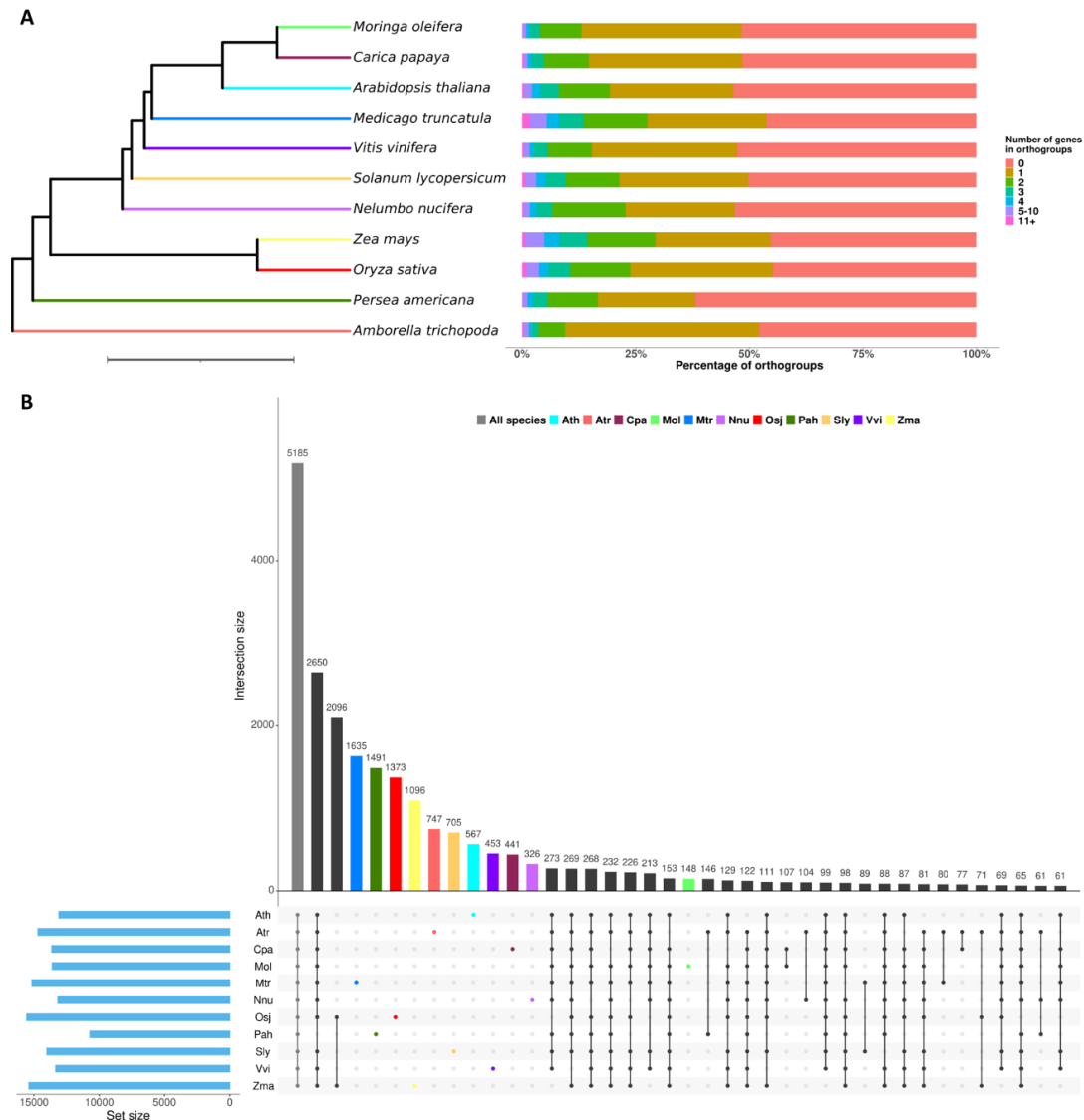


**Figura 3. Resumen de los resultados de la anotación funcional del genoma de *Moringa*.**

### 4.2. Clasificaciones de ortogrupos/familias génicas y de genes duplicados en *Moringa* abarcan más del 85% y del 65% de sus genes, respectivamente

Para estudiar la evolución de familias génicas en *Moringa*, se generó una clasificación de ortogrupos para AOCC v2 y los genomas de otras 10 plantas angiospermas (**Tabla 5**) (**Figura 4A**). 299.745 proteínas de un total de 349.936 procedentes de *Moringa* y de las 10 especies vegetales seleccionadas (**Figura 4A**) pudieron clasificarse en 28.161 ortogrupos que contenían al menos dos genes (**Anexo 1**); 5.185 de estos ortogrupos agruparon genes de las 11 especies (**Anexo 1**) (**Figura 4B**). Los 50.191 genes restantes, incluidos 2.668 genes de *Moringa*, se clasificaron como

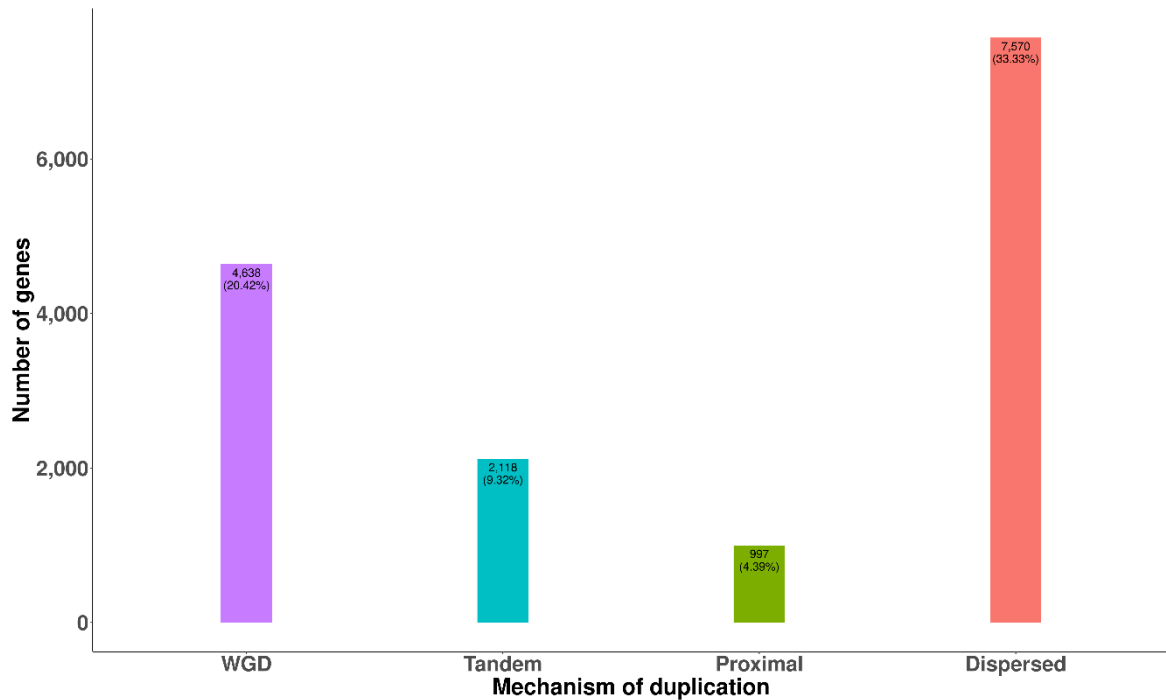
no asignados a ningún ortogrupo (**Anexo 1**), correspondiendo probablemente a secuencias huérfanas únicas sin homólogos detectados en ninguna especie. 20.046 genes del genoma de la *Moringa*, es decir, casi el 88,25% del total, fueron asignados a 13.597 ortogrupos (**Anexo 1**) (**Figura 4B**) de los cuales 148 ortogrupos, que agrupan 941 genes, son exclusivos de *Moringa* (**Anexo 1**) (**Figura 4B**).



**Figura 4. Clasificación de ortogrupos/familias génicas en *Moringa* y otras 10 especies de plantas que representan los principales linajes de plantas con flores.** A. Árbol ultramétrico que representa las relaciones evolutivas entre *Moringa* y otras 10 especies de plantas (panel izquierdo) e histograma que representa en porcentajes la distribución de genes por especie resultante de la clasificación de familias génicas (panel derecho). B. Gráfico UpSet que representa el número de familias (barras) que contienen genes de una especie o un conjunto de especies específico (puntos). Las especies tienen asignados identificadores que corresponden a las abreviaturas usadas en la **Tabla 5**. Sólo se muestran las intersecciones que abarcan más de 60 familias. Los ortogrupos formados por miembros de las 11 especies se muestran con barras y puntos grises, mientras que las barras y los puntos correspondientes a los ortogrupos específicos de especie están coloreados según el esquema de colores del árbol de especies del panel A. El histograma de la izquierda representa el número total de familias de genes para cada especie.



Asimismo, también se generó una clasificación de genes duplicados en base a mecanismos de duplicación donde un total de 15.323 genes (el 67,46% de los genes estructurales de *Moringa*) fueron identificados como duplicados: 4.638 son duplicados WGD, 2.118 son duplicados en tándem, 997 son duplicados proximales y 7.570 son duplicados dispersos (**Figura 5**).



**Figura 5. Distribución de los duplicados de genes por mecanismo de duplicación en el genoma de *Moringa*.**

### 4.3. Análisis de la evolución de familias génicas en *Moringa*

4.3.1. El patrón diferencial de retención funcional de genes duplicados clasificados por mecanismos de duplicación en *Moringa* es compatible con la hipótesis de balance de dosis

Estudios previos han reportado que genes con ciertas funciones biológicas (regulación transcripcional, transducción de señales, transporte de proteínas y la modificación de proteínas) se conservan preferentemente después de duplicaciones WGD, mientras que rara vez se conservan después de duplicaciones SSD (en tándem, proximales y dispersas), y viceversa (Maere et al., 2005) puesto que estas funciones involucran redes de regulación de genes y complejos multiproteicos (entre otros sistemas complejos) que solo son operativas si se expresa el conjunto completo de genes responsables. Entre los diferentes modelos propuestos para explicar el patrón diferencial de pérdida y retención de genes observado después de una duplicación, solo la hipótesis del balance de dosis predice tal reciprocidad entre duplicados WGD y SSD (Freeling, 2009; Tasdighian et al., 2017). Para comprobar si las predicciones de esta hipótesis son aplicables a *Moringa*, se realizaron tests de enriquecimiento funcional de términos GO en genes duplicados categorizados por mecanismo de duplicación. Los términos GO más significativamente enriquecidos entre los duplicados WGD estaban relacionados con la regulación transcripcional, seguidos de diferentes formas de modificación de proteínas, incluyendo las actividades de fosforilación/quinasa o la dimerización de proteínas, funciones comúnmente consideradas como sensibles al equilibrio de dosis (**Anexo 2**) (**Figuras 6A, B**).

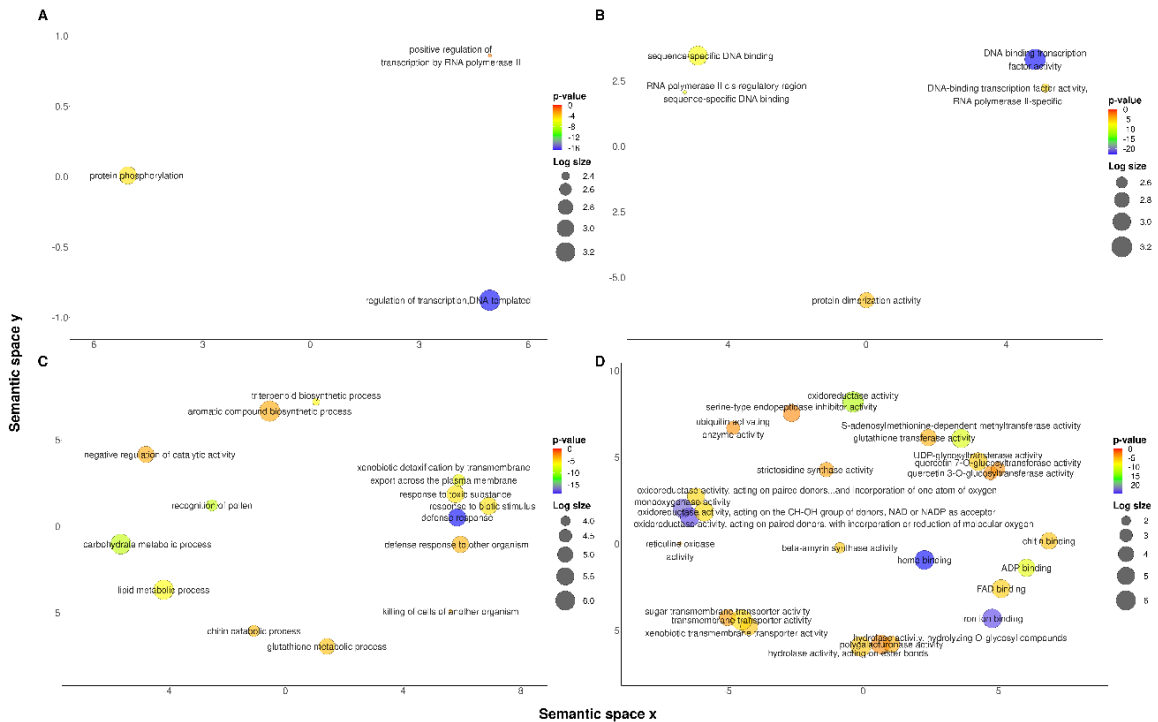
Asimismo, los duplicados en tándem de Moringa estaban enriquecidos en términos GO correspondientes a enzimas específicas del metabolismo secundario, como isoprenoides (actividad lanosterol sintasa, actividad beta-amirina sintasa), alcaloides (reticulina oxidasa, estrictosidina sintasa) fenilpropanopidos/flavonoides/glucosinolatos (metiltransferasa dependiente de S-adenosilmetionina), flavonoles glicosilados (quercetina 3-O-glucosiltransferasa, quercetina 7-O-glucosiltransferasa) y glutatión (glutatión transferasa) (**Anexo 3**) (**Figuras 6C, D**) o rutas específicas del metabolismo secundario (por ejemplo, proceso biosintético de compuestos que contienen antocianina, proceso biosintético de compuestos aromáticos, proceso biosintético de triterpenoides) (**Anexo 3**) (**Figuras 6C, D**). Para reforzar este resultado (también anticipado por la hipótesis del balance de dosis), se mapearon 395 grupos de ortología del KEGG correspondientes a un total de 1.181 genes (de los cuales 810 son únicos) en el genoma de Moringa en 50 rutas de metabolismo secundario (**Anexo 4**) utilizando la anotación de rutas KEGG (números KO). También se consideraron las rutas bioquímicas implicadas en el metabolismo de los aminoácidos, puesto que sirven como precursores de una amplia gama de metabolitos secundarios, incluyendo compuestos fenólicos, alcaloides o glucosinolatos. 121 de estos genes correspondían a duplicados en tándem, lo que es significativamente mayor de lo esperado por el azar (**Tabla 6**). Además, los duplicados dispersos, que representan 324 de los genes del metabolismo secundario, también se encontraron en una proporción mayor de la esperada por el azar, aunque de forma menos significativa. En cambio, los duplicados genómicos (183 genes) o proximales (39 genes) no parecían estar ni infrarrepresentados ni sobrerrepresentados entre los genes del metabolismo secundario (**Tabla 6**). Por consiguiente, las predicciones hechas por la hipótesis del balance de dosis se cumplen en Moringa.

**Tabla 6.** Resultados del test de enriquecimiento para genes duplicados clasificados en base a mecanismos de duplicación en genes del metabolismo secundario.

Tipo de gen duplicado	Genes duplicados presentes en rutas del metabolismo secundario	Genes duplicados ausentes en rutas del metabolismo secundario	Genes no duplicados presentes en rutas del metabolismo secundario	Genes no duplicados ausentes en rutas del metabolismo secundario	P resultante de prueba exacta de Fisher
WGD	183	4.455	627	17.449	0,066
En tándem	121	1.997	689	19.907	$1,2 \times 10^{-7}$
Proximal	39	958	771	20.946	0,3
Disperso	324	7.246	486	14.658	$3 \times 10^{-5}$

Aparte, varios términos GO asociados con la respuesta de defensa de las plantas contra las señales ambientales bióticas y abióticas se encontraron como enriquecidos entre los duplicados en tándem, incluyendo la respuesta de defensa, la desintoxicación de xenobióticos por la exportación transmembrana a través de la membrana plasmática, la respuesta a la sustancia tóxica, la respuesta de defensa a otro organismo, la muerte de las células de otro organismo y la respuesta al estímulo biótico (**Anexo 4**) (**Figuras 5C, D**). Este aparente enriquecimiento de funciones biológicas relacionadas con la respuesta al estrés podría ayudar a explicar la retención de duplicados en tándem en genomas vegetales, a pesar de que se espera que alteren el balance de dosis inmediatamente después de la duplicación -al menos cuando forman parte de redes de regulación de genes o complejos multiproteicos- y den lugar a defectos de aptitud (*fitness*).

## Ensamblaje a escala cromosómica del genoma de “Moringa oleifera” | Juan Pablo Marczuk Rojas



**Figura 6. Diagramas de burbuja hechos en ReViGo para representar funciones biológicas significativamente enriquecidas entre duplicados WGD y en tándem de Moringa.** Los resultados se muestran para los genes resultantes de duplicaciones genómicas (A, B) y los duplicados en tándem (C, D) para los términos GO de proceso biológico (A, C) y función molecular (B, D). Los términos GO encontrados como significativamente enriquecidos están representados como círculos, con diámetros proporcionales al tamaño de la muestra (número de genes con el término correspondiente en comparación con la base de datos de proteínas de referencia escogida) y valores p resultantes de tests de Fisher corregidos con método de Bonferroni, transformados en logaritmos de base 10 y codificados en una escala de colores. REVI GO agrupa los términos GO funcionalmente similares en un espacio semántico bidimensional.

### 4.3.2. Las duplicaciones SSD (en tándem y proximales) son la principal fuente de genes específicos de Moringa

Comúnmente, se ha observado que los genes resultantes de duplicaciones en tándem se conservan de forma específica para cada linaje y se enriquecen en categorías funcionales relacionadas con la respuesta al estrés y con el metabolismo secundario (esto último ya comprobado para Moringa en el apartado anterior) (Carretero-Paulet & Fares, 2012; Chae et al., 2014; Denoeud et al., 2014).

Aunque sólo 108 de los 941 genes agrupados en los ortogrupos específicos de Moringa fueron anotados con términos GO, tests de enriquecimiento identificaron como enriquecidos procesos de desarrollo específicos, incluyendo el desarrollo de las plántulas, el crecimiento del meristemo y la generación de gametos masculinos y femeninos, y respuestas de defensa como la respuesta a proteínas mal plegadas (**Anexo 5**). Asimismo, la mayoría de los genes anotados con ambos conjuntos de términos GO pertenecen al mismo ortogrupo HOG0020145, que está compuesto por tres genes de Moringa anotados como codificantes para componentes reguladores del sistema ubiquitina/26S proteosoma (**Anexo 5**) incluyendo dominios ATPasa (EC:5.6.1.5),

específicamente implicados en la apertura de canales y el despliegue de polipéptidos antes de la proteólisis. Esto podría indicar que el sistema ubiquitina/26S proteosoma en *Moringa* está contribuyendo a la plasticidad proteómica necesaria para vincular el crecimiento y el desarrollo de la planta con la adaptación al estrés ambiental, como la sequía, el calor y el estrés UV (Xu & Xue, 2019).

A continuación, se investigó si algún mecanismo de duplicación prevalecía en la expansión de los ortogrupos específicos de *Moringa*. De acuerdo con los tests de enriquecimiento, los ortogrupos específicos de *Moringa* estaban fuertemente enriquecidos en duplicados en tándem (200 genes) y proximales (145 genes), mientras que no estaban ni enriquecidos ni empobrecidos en duplicados dispersos (271 genes) y duplicados WGD (106 genes) (**Tabla 7**). Por lo tanto, las duplicaciones en tándem y proximales parecen ser los mecanismos prevalentes detrás de la expansión de los ortogrupos específicos de *Moringa*.

**Tabla 7.** Resultados del test de enriquecimiento para genes duplicados clasificados en base a mecanismos de duplicación en genes de ortogrupos específicos de *Moringa*.

Tipo de gen duplicado	Genes duplicados presentes en ortogrupos específicos	Genes duplicados ausentes en ortogrupos específicos	Genes no duplicados presentes en ortogrupos específicos	Genes no duplicados ausentes en ortogrupos específicos	P resultante de prueba exacta de Fisher
WGD	106	4.532	835	17.241	1
En tándem	200	1.918	741	19.855	$1,1 \times 10^{-29}$
Proximal	145	852	796	20.921	$2 \times 10^{-41}$
Disperso	271	7.299	670	14.474	1

#### 4.4. Rutas y clústeres biosintéticos de metabolitos secundarios en el genoma de *Moringa*

Dada la importancia de los metabolitos secundarios para el contenido nutricional, las propiedades organolépticas y la actividad farmacológica de las hojas y las semillas de *Moringa*, se realizó la identificación y la caracterización molecular evolutiva de familias génicas presuntamente implicadas en el metabolismo secundario, dirigiendo el enfoque a los dos motores principales de la evolución de las rutas del metabolismo vegetal que implican grupos de genes duplicados en tándem, es decir, la neofuncionalización de los duplicados en tándem y los clústeres de genes biosintéticos (Tohge & Fernie, 2020).

La neofuncionalización de genes duplicados en tándem se ha identificado como un mecanismo importante que impulsa la evolución de las rutas específicas del metabolismo secundario debido a que podría proveer una especificidad de sustrato diferencial a las copias. Ese fue el caso de la ruta de biosíntesis de glucosinolatos en *Arabidopsis* (Tohge & Fernie, 2020). Por ello, se procedió a reconstruir todas las rutas de relacionadas con el metabolismo de glucosinolatos en *Moringa* usando términos funcionales relacionados que fueron asignados a los genes estructurales detectados en su genoma (AOCC v2), así como a los genes de *Arabidopsis* y papaya (**Anexo 6**). Diversas rutas para la biosíntesis de determinados tipos de glucosinolatos fueron identificadas en *Moringa*, así como actividades enzimáticas para la síntesis de ciertos precursores de otros tipos (**Figura 7**). En total, se encontraron 104 genes de *Moringa* probablemente implicados en el metabolismo de glucosinolatos, los cuales fueron agrupados en 28 de los 33 ortogrupos detectados, incluyendo 11 con al menos un par de duplicados en tándem (**Anexo 6**). Dos casos interesantes fueron los ortogrupos HOG0000588 y HOG0009822. El ortogrupo HOG0000588

incluía 15 genes de *Moringa* dispuestos en dos grupos de duplicados en tándem localizados en los cromosomas 1 y 5, junto con 18 y 14 en *Arabidopsis* y papaya, respectivamente, y de cero a 24 (uva) en el resto de especies. Los genes de HOG0000588 fueron anotados como codificantes para hidroxilasas de indol-3-ilmetilglucosinolato. El ortogrupo HOG0009822 agrupa siete genes de *Moringa* dispuestos en tándem en posiciones consecutivas del cromosoma 10, por cinco y tres en *Arabidopsis* y papaya, respectivamente, y cero en el resto de especies. Los genes de HOG0009849 están anotados como codificantes para esterasas/lipasas de tipo GDSL, una familia de enzimas hidrolíticas de lípidos con propiedades multifuncionales (como una amplia especificidad de sustrato y regioespecificidad) que está implicada en diversas rutas del metabolismo secundario, incluidas las rutas de glucosinolatos (Lai et al., 2017). En cambio, no se pudieron detectar ortólogos de *Moringa* para los genes de *Arabidopsis* relacionados con glucosinolatos en los ortogrupos HOG0016593, HOG0022085, HOG0017508, HOG0008572, HOG0004358 y HOG0019714, los cuales codifican para la aminotransferasa de cadena ramificada (EC:2.6.1.42), la homometionina N-monooxigenasa (EC:1.14.14.42), la glucosinolato alifático S-oxigenasa (EC:1.14.13.237), el transportador de magnesio, la beta-glucosidasa (EC:3.2.1.21) y la mirosinasa (E3.2.1.147), respectivamente (**Anexo 6**). Sin embargo, los genes de *Moringa* anotados con tales actividades enzimáticas se encontraron en otros ortogrupos de glucosinolatos, excepto la homometionina N-monooxigenasa y la glucosinolato alifático S-oxigenasa (**Anexo 6**). En particular, el ortogrupo HOG0000577, anotado como codificante para una clase específica de beta glucosidasas (mirosinasas) implicadas en la producción de glucosinolatos biológicamente activos, se encontró como fuertemente expandido en *Moringa*, con 14 genes, por cuatro y cinco en *Arabidopsis* y papaya, respectivamente, y de cero a seis en el resto de especies. También se encuentra un único representante de *Moringa* en los ortogrupos HOG0002947 y HOG0006126, los cuales codifican para la gamma-glutamil hidrolasa de glucosinolato (EC:3.4.19.16) y la metiltilmalato sintasa (EC:2.3.3.17), respectivamente (**Anexo 6**); ambas familias de enzimas habían sido reportadas como ejemplos de diversificación en el metabolismo secundario de *Arabidopsis* a través de la neofuncionalización de duplicados en tándem (Kliebenstein et al., 2001; Petersen et al., 2019).

A continuación, se utilizaron datos de expresión basados en *RNA-seq* de cinco tejidos para examinar la expresión de los 104 genes de *Moringa* identificados en el análisis como presuntamente implicados en la biosíntesis de glucosinolatos. De los 104 genes, 84 y 88 se expresaron en las semillas y las hojas, respectivamente, donde la biosíntesis de glucosinolatos es más abundante (**Figura 8**). Los genes de *Moringa* duplicados en tándem y pertenecientes a los ortogrupos relacionados con los glucosinolatos mostraron niveles de expresión diversificados en los cinco tejidos, con un gen del ortogrupo HOG0000588 (Morol01g14150), dos genes pertenecientes al ortogrupo HOG0009822 (Morol10g15130 y Morol10g15170) y otros dos pertenecientes al ortogrupo HOG0000577 (Morol03g00630, Morol04g12880) mostrando una notable expresión en hojas y semillas (**Figura 8**). Es tentador especular que la rápida especialización funcional después de la duplicación en tándem detectada en estas dos familias génicas (HOG0009822 y HOG0000577) podría haber contribuido a la diversidad de glucosinolatos entre los tejidos y entre la variedad silvestre y las accesiones domesticadas de *Moringa*, así como entre esta y otras especies de la familia Moringaceae (Chodur et al., 2018; Fahey et al., 2018).

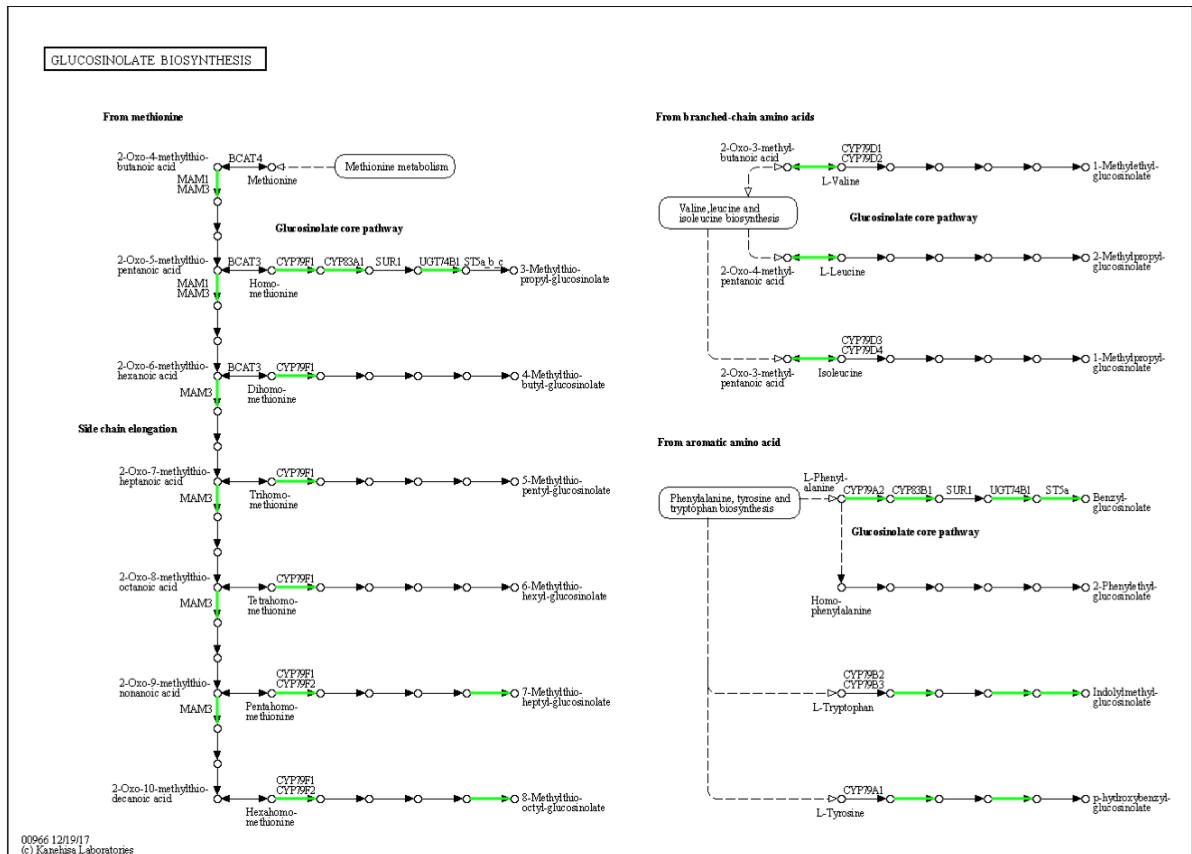
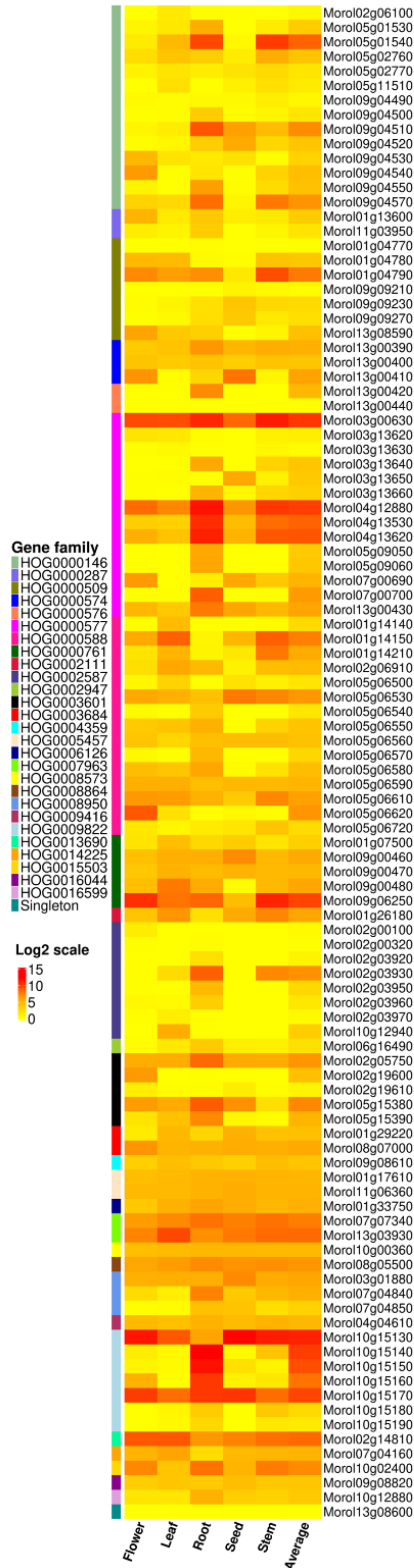


Figura 7. Ruta del KEGG para la biosíntesis de glucosinolatos (entrada del mapa 00966). Las actividades enzimáticas que se encuentran codificadas en el genoma de la Moringa están mapeadas como flechas verdes sobre la reacción bioquímica correspondiente.



**Figura 8.** Representación en forma de mapa de calor de los patrones de expresión de 104 hipotéticos genes de Moringa relacionados con los glucosinolatos en cinco tejidos, además de la expresión media. Los colores del mapa de calor representan los valores de expresión medidos en transcritos por millón y transformados en logaritmos en base 2. Las bandas de colores en la izquierda indican la hipotética familia génica/ortogrupo a la que pertenece cada gen.

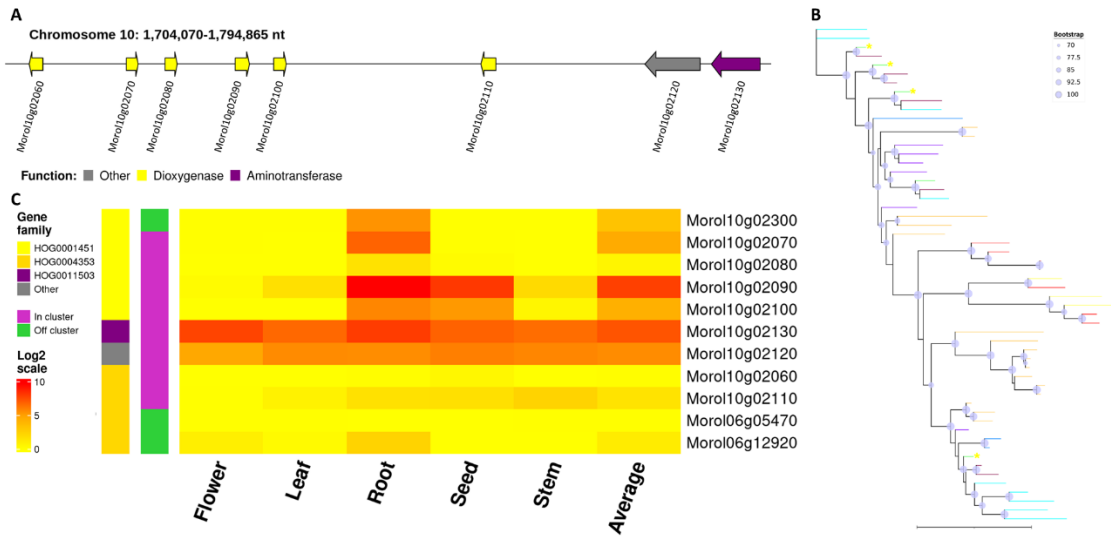
Asimismo, hay cada vez más pruebas de que los genes no homólogos que codifican enzimas biosintéticas específicas implicadas en la misma ruta de metabolismo secundario evolucionan como clústeres que ocupan regiones vecinas de los genomas vegetales, de forma similar a lo observado anteriormente en bacterias y hongos (Rokas et al., 2018). Por ello, se realizó una búsqueda de clústeres de genes de metabolitos secundarios (CGMSs).

Se identificaron 18 hipotéticos clústeres relacionados con varias rutas metabólicas secundarias de las plantas, que incluían dos alcaloides, dos lignan-policétidos, nueve sacáridos, tres terpenos y dos clústeres sin función biosintética específica (**Anexo 7**). Algunos de estos metabolitos secundarios podrían estar relacionados con las respuestas y adaptaciones de las plantas a diferentes estreses ambientales, incluyendo el déficit de agua y la radiación UV-B (Bandurska et al., 2013). Las regiones genómicas que ocupan estos clústeres de genes biosintéticos abarcaban desde 28,9 hasta 339,26 Kpb y contenían entre seis y 19 genes. Estos CGMSs parecían estar distribuidos uniformemente a lo largo del genoma de *Moringa*, con cada cromosoma albergando al menos un CGMS, excepto los cromosomas 1 y 14 (**Anexo 7**). Tres CGMSs recibieron especial atención:

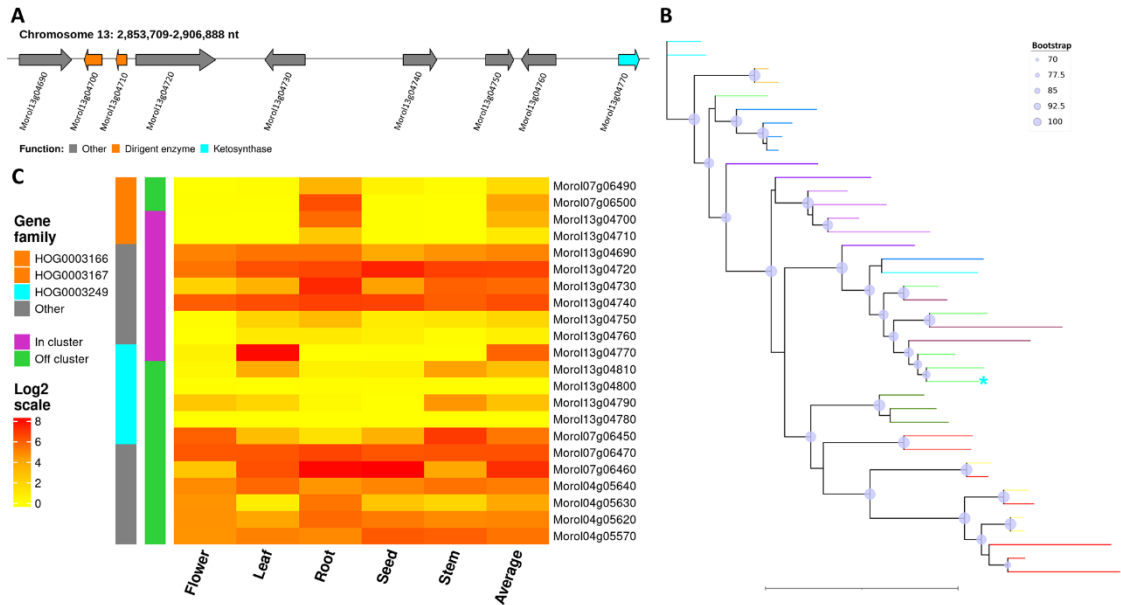
- CGMS uno (**Figura 9A**): contiene ocho genes de *Moringa* situados en el cromosoma 10. Seis de ellos estaban anotados como dioxigenasas, cuatro de los cuales pertenecían al ortogrupo HOG0001451. HOG0001451 incluye nueve ortólogos de *Arabidopsis*, cinco de los cuales formaban un clado bien soportado en un árbol filogenético que representaba las relaciones evolutivas entre las 56 secuencias pertenecientes al ortogrupo (**Figura 9B**). Este clado está incluido en un clado mayor que contenía dos genes de papaya y un gen de *Moringa* (Morol10g02100). Tres de estos cinco genes de *Arabidopsis* (AT4G03050, AT4G03060, AT4G03070) habían sido reportados como dioxigenasas dependientes de 2-oxoglutarato involucradas en la biosíntesis de glucosinolatos (Kliebenstein et al., 2001), los cuales no fueron detectados en el estudio de los ortogrupos relacionados con los glucosinolatos. Morol10g02100 mostró expresión en las raíces y, en menor medida, en las semillas (**Figura 9C**).
- CGMS cinco (**Figura 10A**): anotado como de tipo lignan-policético, incluye un gen anotado como cetosintasa perteneciente al ortogrupo HOG0003249, que a su vez incluía cuatro genes adicionales dispuestos en tándem en posiciones vecinas del cromosoma más un sexto gen localizado en otra parte del genoma (**Figura 10A**). Los tres ortólogos de *Arabidopsis* pertenecientes al ortogrupo HOG0003249 (AT5G04530.1, AT2G28630.1 y AT1G07720.2) habían sido reportados como miembros de la familia de la 3-cetoacil-CoA sintasa involucrada en la biosíntesis de ácidos grasos de cadena muy larga como precursores de compuestos de cera, participando en la limitación de la pérdida de agua no estomática durante la adaptación a la sequía y en la prevención de ataques de patógenos (Li-Beisson et al., 2013). Con seis genes en HOG0003249 por dos a cinco en los restantes 10 genomas vegetales comparados en este estudio, *Moringa* fue encontrada como ligeramente expandida en esta familia. Cinco de los seis genes, incluyendo el encontrado en el SMGC, fueron agrupados en un árbol filogenético dentro de un clado bien apoyado junto con sus homólogos de la papaya (**Figura 10B**). La mayoría de los miembros de la familia en *Moringa* mostraron una baja expresión, excepto Morol13g04770, que se expresa fuertemente en las hojas (**Figura 10C**).



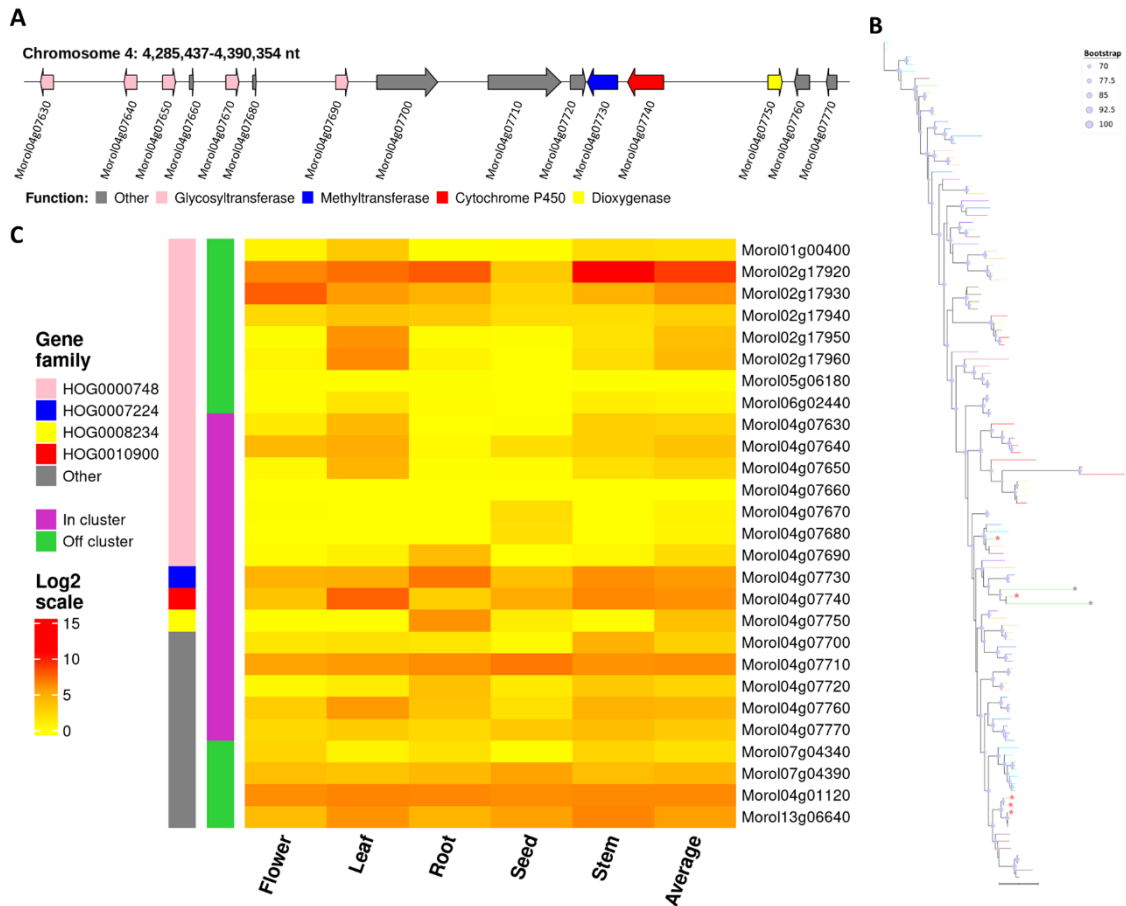
- CGMS 10 (**Figura 11A**): relacionado con la biosíntesis de sacáridos, está formado por 15 genes, entre los que se encontraban cinco anotados como UDP-glicosiltransferasas, uno como metiltransferasa, otro como citocromo P450, otro como dioxigenasa, mientras que el resto estaba anotado como codificantes para funciones no relacionadas (**Figura 11A**). Las cinco UDP-glicosiltransferasas se agruparon junto con 10 secuencias adicionales en el ortogrupo HOG0000748, incluyendo dos también detectadas en el clúster con secuencias significativamente más cortas y anotadas como codificantes para funciones no relacionadas. HOG0000748 también incluía 13 y 11 genes en las Brassicales Arabidopsis y papaya, respectivamente. El análisis filogenético de máxima verosimilitud basado en el alineamiento múltiple de las 117 secuencias de aminoácidos incluidas en HOG0000748 agrupó cinco genes de Moringa en un clado bien soportado, incluyendo tres de los encontrados en el CGMS, el cual es hermano de un clado formado por siete genes de papaya (**Figura 11B**). El clado resultante mostró a su vez una relación de hermandad con un clado compuesto por siete genes de Arabidopsis (**Figura 11B**). Los dos genes restantes del CGMS se encontraban en posiciones dispares en el árbol. Estos genes fueron anotados como codificantes para UDP-glucosiltransferasas y, específicamente, antocianidinas 3-O-glucosiltransferasas (EC:2.4.1.115). Se ha reportado que los ortólogos en Arabidopsis están implicados en la glicosilación de flavonoles específicos, pero también de fenilpropanoides específicos, esteroides o citoquininas como la zeatina, comúnmente en respuesta a estreses específicos. En general, los genes de Moringa encontrados en el clúster mostraron, de forma similar, patrones de expresión bajos, con la excepción de Morol04g07740, Morol04g07730 y Morol04g07710, que mostraron niveles de expresión moderados (**Figura 11C**).



**Figura 9. Caracterización del clúster unos de genes del metabolismo secundario en el genoma de Moringa.** A, Organización genómica de un clúster de genes relacionado con el metabolismo secundario en Moringa formado por seis dioxigenasas, una aminotransferasa y un gen que codifica para otra actividad enzimática. B, Árbol filogenético de máxima verosimilitud de la familia de genes de dioxigenasa HOG0001451. Los valores de soporte estadístico para los cladros resultantes del análisis bootstrap ultrarrápido se muestran junto a los nodos correspondientes en forma de círculos púrpura, con el diámetro proporcional a los valores resultantes. Sólo se muestran los valores superiores a 70. Las ramas están coloreadas según el esquema de colores del árbol de especies de la **Figura 4A**. Las ramas de los árboles son proporcionales al tiempo evolutivo, con longitudes que reflejan el número de cambios de aminoácidos. Los genes de Moringa incluidos en el clúster están indicados con asteriscos de colores. C, Representación en mapa de calor de los patrones de expresión en diferentes tejidos más la expresión media de los genes incluidos en el clúster y los genes parálogos identificados en la clasificación de ortogrupos. Los colores del mapa de calor representan los valores de expresión medidos en transcritos por millón y transformados en logaritmos de base 2. Las bandas coloreadas de la izquierda indican la familia de genes a la que pertenecen. También se indican los genes incluidos en el clúster (*in cluster*) o en otra parte del genoma (*off cluster*).



**Figura 10. Caracterización del clúster cinco de genes del metabolismo secundario en el genoma de Moringa.** A, Organización genómica de un clúster de genes relacionados con la biosíntesis de lignan-policétidos en Moringa, formado por dos enzimas dirigentes, una cetosintasa y seis genes que codifican para otras actividades enzimáticas. B, Árbol filogenético de máxima verosimilitud de la familia de genes de cetosintasa HOG0003249. Los valores de soporte estadístico para los clados resultantes del análisis bootstrap ultrarrápido se muestran junto a los nodos correspondientes en forma de círculos morados, con el diámetro proporcional a los valores resultantes. Sólo se representan los valores superiores a 70. Las ramas están coloreadas según el esquema de colores del árbol de especies de la **Figura 4A**. Las ramas de los árboles son proporcionales al tiempo evolutivo, con longitudes que reflejan el número de cambios de aminoácidos. Los genes de Moringa incluidos en el clúster están indicados con asteriscos de colores. C, Representación en mapa de calor de los patrones de expresión en diferentes tejidos más la expresión media de los genes incluidos en el clúster y los genes parálogos identificados en la clasificación de ortogrupos. Los colores del mapa de calor representan los valores de expresión medidos en transcritos por millón y transformados en logaritmos de base 2. Las bandas coloreadas de la izquierda indican la familia de genes a la que pertenecen. También se indican los genes incluidos en el clúster (*in cluster*) o en otra parte del genoma (*off cluster*).



**Figura 11. Caracterización del clúster 10 de genes del metabolismo secundario en el genoma de la Moringa.** A, Organización genómica de un grupo de genes relacionados con la biosíntesis de sacáridos en *Moringa*, que incluye cinco glucosiltransferasas, una metiltransferasa, un citocromo P450, una dioxigenasa y siete genes que codifican para otras actividades enzimáticas. B, Árbol filogenético de máxima verosimilitud de la familia de genes de UDP-glicosiltransferasa HOG0000748. Los valores de soporte estadístico para los clados resultantes del análisis bootstrap ultrarrápido se muestran junto a los nodos correspondientes en forma de círculos morados, con el diámetro proporcional a los valores resultantes. Sólo se representan los valores superiores a 70. Las ramas están coloreadas según el esquema de colores del árbol de especies de la **Figura 4A**. Las ramas de los árboles son proporcionales al tiempo evolutivo, con longitudes que reflejan el número de cambios de aminoácidos. Los genes de *Moringa* incluidos en el clúster están indicados con asteriscos de colores. C, Representación en mapa de calor de los patrones de expresión en diferentes tejidos más la expresión media de los genes incluidos en el clúster los genes parálogos identificados en la clasificación de ortogrupos. Los colores del mapa de calor representan los valores de expresión medidos en transcritos por millón y transformados en logaritmos de base 2. Las bandas coloreadas de la izquierda indican la familia de genes a la que pertenecen. También se indican los genes incluidos en el clúster (*in cluster*) o en otra parte del genoma (*off cluster*).

## 5. Conclusiones

Las principales conclusiones de este estudio fueron:

1. Se obtuvo una anotación funcional completa del genoma de Moringa donde, de un total de 22.714 genes estructurales, 18.460 (81,27%) fueron anotados con al menos un término funcional. Asimismo, las anotaciones KEGG, EC, GO e INTREPRO cubrieron entre el 34,68 y el 78,08% de los genes abarcando entre 7.780 y 98.325 términos totales.
2. Se generó una clasificación de ortogrupos/familias génicas donde, de un total de 22.714 genes estructurales de Moringa, 20.046 fueron asignados a 13.597 ortogrupos/familias (88,25%) quedando unos 2.668 genes huérfanos (11,75%). Además, unos 941 genes fueron repartidos en 148 ortogrupos específicos de Moringa.

También se obtuvo una clasificación de genes duplicados que representan más de la mitad de los genes estructurales del genoma de Moringa (67,46%), siendo los más abundantes los duplicados dispersos (7.570) seguidos por duplicados WGD (4.638), representando un 33.33 y un 20.42%, respectivamente.

3. La caracterización funcional de duplicados WGD en Moringa mediante términos GO asociados revela un enriquecimiento en la regulación transcripcional y la modificación de proteínas, como factores de transcripción o actividades de fosforilación/dimerización de proteínas. Estas funciones son comúnmente consideradas sensibles al balance de dosis (Papp et al., 2003; Tasdighian et al., 2017) y, por lo tanto, se espera que se retengan preferentemente después de eventos de duplicación WGD (Freeling, 2009). A su vez, los duplicados SSD, especialmente los duplicados en tándem y, en menor medida, los duplicados dispersos, se encontraron enriquecidos para enzimas específicas del metabolismo secundario. Por lo tanto, el patrón diferencial de retención recíproco de duplicados WGD contra duplicados SSD para clases funcionales anticipado por la hipótesis del balance de dosis (Freeling, 2009) también puede ser verificado en Moringa.
4. Las duplicaciones en tándem, junto con las duplicaciones proximales, son el modo preferente de duplicación que condujo a la expansión de 148 familias de genes específicos de Moringa. Asimismo, el enriquecimiento en funciones del metabolismo secundario y de respuestas de defensa observado en las duplicaciones en tándem corrobora su implicación en las adaptaciones rápidas a estímulos ambientales locales (Hanada et al., 2008) y puede estar relacionado con la alta plasticidad fenotípica y la adaptabilidad de Moringa a diferentes limitaciones ambientales, especialmente el estrés hídrico (Brunetti et al., 2020; Brunetti et al., 2018) o la radiación UV-B (Araujo et al., 2016).

En resumen, la versión actual y la anotación del genoma de Moringa que se presenta en este estudio facilitará la identificación de los genes que están en el origen de las propiedades biológicas, agronómicas, nutricionales o farmacológicas en esta especie y ayudará en gran medida al desarrollo de programas de mejora de plantas asistidos por genómica, especialmente aquellos relacionados con los caracteres del metabolismo secundario de interés.

## 6. Anexos

**Anexo 1. Resumen de las estadísticas de la clasificación de ortogrupos de Orthofinder para el genoma de Moringa y los de otras 10 especies de plantas con flores.**

**Anexo 2. Categorización funcional con términos GO genéricos asignados a los genes duplicados WGD en Moringa.** Se muestran los términos GO significativamente sobrerrepresentados tras la corrección de Bonferroni. A continuación, se muestra la lista de genes duplicados WGD clasificados por pertenencia a un ortogrupo junto con sus anotaciones GO y EC.

**Anexo 3. Categorización funcional con términos GO genéricos asignados a los genes duplicados en tándem en Moringa.** Se muestran los términos GO significativamente sobrerrepresentados tras la corrección de Bonferroni. La lista de genes duplicados en tándem clasificados por pertenencia a un ortogrupo junto con sus anotaciones GO y EC se muestra a continuación.

**Anexo 4. Número de genes de Moringa, Arabidopsis y papaya implicados en las rutas de metabolismo secundario del KEGG resultantes de las anotaciones hechas por BlastKOALA.**

**Anexo 5. Categorización funcional con términos GO genéricos asignados a los genes pertenecientes a las familias específicas de Moringa.** Se muestran los términos GO significativamente sobrerrepresentados tras la corrección de Bonferroni. La lista de familias de genes específicos de Moringa se muestra a continuación junto con sus anotaciones GO y EC.

**Anexo 6. Resumen de los ortogrupos de Moringa involucrados en la biosíntesis de glucosinolatos y su regulación en Moringa, Arabidopsis y papaya.** Resumen de los ortogrupos de Moringa involucrados en la biosíntesis de glucosinolatos y su regulación en Moringa, Arabidopsis y papaya. Los identificadores de genes en negrita indican su agrupación dentro del correspondiente grupo de ortología funcional KEGG como resultado de la anotación BlastKOALA (número KO) en el caso de los genes de Moringa y papaya. Los términos GO y los códigos EC en negrita indican la anotación GO y las enzimas del correspondiente grupo de ortología funcional KEGG, mientras que el resto corresponde a la anotación GO y EC de BLAST2GO para los genes de Moringa. Los grupos de ortología de Moringa que incluyen al menos un representante de duplicados en tándem están resaltados con una sombra gris.

**Anexo 7. Resumen de los 18 clústeres de genes de metabolitos secundarios identificados en el genoma de Moringa.**

## 7. Bibliografía

- Albert, V. A., Barbazuk, W. B., dePamphilis, C. W., Der, J. P., Leebens-Mack, J., Ma, H., Palmer, J. D., Rounsley, S., Sankoff, D., Schuster, S. C., Soltis, D. E., Soltis, P. S., Wessler, S. R., Wing, R. A., Albert, V. A., Ammiraju, J. S. S., Barbazuk, W. B., Chamala, S., Chanderbali, A. S., ... Tomsho, L. (2013). The *Amborella* Genome and the Evolution of Flowering Plants. *Science*, 342(6165). <https://doi.org/10.1126/science.1241089>
- Araújo, M., Santos, C., Costa, M., Moutinho-Pereira, J., Correia, C., & Dias, M. C. (2016). Plasticity of young *Moringa oleifera* L. plants to face water deficit and UVB radiation challenges. *Journal of Photochemistry and Photobiology B: Biology*, 162, 278–285. <https://doi.org/10.1016/j.jphotobiol.2016.06.048>
- Bandurska, H., Niedziela, J., & Chadzinikolau, T. (2013). Separate and combined responses to water deficit and UV-B radiation. *Plant Science*, 213, 98–105. <https://doi.org/10.1016/j.plantsci.2013.09.003>
- Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze*, 8, 3–62.
- Brunetti, C., Gori, A., Moura, B. B., Loreto, F., Sebastiani, F., Giordani, E., & Ferrini, F. (2020). Phenotypic plasticity of two *M. oleifera* ecotypes from different climatic zones under water stress and re-watering. *Conservation Physiology*, 8(1). <https://doi.org/10.1093/conphys/coaa028>
- Brunetti, C., Loreto, F., Ferrini, F., Gori, A., Guidi, L., Remorini, D., Centritto, M., Fini, A., & Tattini, M. (2018). Metabolic plasticity in the hygrophyte *Moringa oleifera* exposed to water stress. *Tree Physiology*. <https://doi.org/10.1093/treephys/tpy089>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Carretero-Paulet, L., & Fares, M. A. (2012). Evolutionary Dynamics and Functional Specialization of Plant Paralogs Formed by Whole and Small-Scale Genome Duplications. *Molecular Biology and Evolution*, 29(11), 3541–3551. <https://doi.org/10.1093/molbev/mss162>
- Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic Signatures of Specialized Metabolism in Plants. *Science*, 344(6183), 510–513. <https://doi.org/10.1126/science.1252076>
- Chang, Y., Liu, H., Liu, M., Liao, X., Sahu, S. K., Fu, Y., Song, B., Cheng, S., Kariba, R., Muthemba, S., Hendre, P. S., Mayes, S., Ho, W. K., Yssel, A. E. J., Kendabie, P., Wang, S., Li, L., Muchugi, A., Jamnadass, R., ... Liu, X. (2019). The draft genomes of five agriculturally important African orphan crops. *GigaScience*, 8(3). <https://doi.org/10.1093/gigascience/giy152>
- Chodur, G. M., Olson, M. E., Wade, K. L., Stephenson, K. K., Nouman, W., Garima, & Fahey, J. W. (2018). Wild and domesticated *Moringa oleifera* differ in taste, glucosinolate composition, and antioxidant potential, but not myrosinase activity or protein content. *Scientific Reports*, 8(1), 7995. <https://doi.org/10.1038/s41598-018-26059-3>
- Conesa, A., & Götz, S. (2008). Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics*, 2008, 1–12. <https://doi.org/10.1155/2008/619832>

- Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., Aury, J.-M., Bento, P., Bernard, M., Bocs, S., Campa, C., Cenci, A., Combes, M.-C., Crouzillat, D., da Silva, C., ... Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, 345(6201), 1181–1184. <https://doi.org/10.1126/science.1255274>
- Devkota, S., & Bhusal, K. K. (2020). *Moringa oleifera* : A miracle multipurpose tree for agroforestry and climate change mitigation from the Himalayas – A review. *Cogent Food & Agriculture*, 6(1), 1805951. <https://doi.org/10.1080/23311932.2020.1805951>
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Edger, P. P., Hall, J. C., Harkess, A., Tang, M., Coombs, J., Mohammadin, S., Schranz, M. E., Xiong, Z., Leebens-Mack, J., Meyers, B. C., Sytsma, K. J., Koch, M. A., Al-Shehbaz, I. A., & Pires, J. C. (2018). Brassicales phylogeny inferred from 72 plastid genes: A reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *American Journal of Botany*, 105(3), 463–469. <https://doi.org/10.1002/ajb2.1040>
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>
- Fahey, J. W., Olson, M. E., Stephenson, K. K., Wade, K. L., Chodur, G. M., Odee, D., Nouman, W., Massiah, M., Alt, J., Egner, P. A., & Hubbard, W. C. (2018). The Diversity of Chemoprotective Glucosinolates in Moringaceae (*Moringa* spp.). *Scientific Reports*, 8(1), 7994. <https://doi.org/10.1038/s41598-018-26058-4>
- Freeling, M. (2009). Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology*, 60(1), 433–453. <https://doi.org/10.1146/annurev.arplant.043008.092122>
- Gandji, K., Chadare, F. J., Idohou, R., Salako, V. K., Assogbadjo, A. E., & Kakaï, R. L. G. (2018). Status and utilisation of *Moringa oleifera* Lam: A review. *African Crop Science Journal*, 26(1), 137. <https://doi.org/10.4314/acsj.v26i1.10>
- Gehlenborg, N. (2019). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. <https://CRAN.R-project.org/package=UpSetR>
- Godino, M., Arias, C., & Izquierdo, M. I. (2017). *Moringa oleifera* : potential areas of cultivation on the Iberian Peninsula. *Acta Horticulturae*, 1158, 405–412. <https://doi.org/10.17660/ActaHortic.2017.1158.46>
- Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, 27(2), 221–224. <https://doi.org/10.1093/molbev/msp259>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.



- Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., & Shiu, S.-H. (2008). Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiology*, *148*(2), 993–1003. <https://doi.org/10.1104/pp.108.122457>
- Harrison, K. J., Crécy-Lagard, V. de, & Zallot, R. (2018). Gene Graphics: A genomic neighborhood data visualization web application. *Bioinformatics*, *34*(8), 1406–1408. <https://doi.org/10.1093/bioinformatics/btx793>
- Islam, Z., Islam, S. M. R., Hossen, F., Mahtab-ul-Islam, K., Hasan, Md. R., & Karim, R. (2021). Moringa oleifera is a Prominent Source of Nutrients with Potential Health Benefits. *International Journal of Food Science*, *2021*, 1–11. <https://doi.org/10.1155/2021/6627265>
- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>
- 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*, *30*(9), 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, *428*(4), 726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Kassambara, A. (2020). *ggpubr: “ggplot2” Based Publication Ready Plots*. <https://CRAN.R-project.org/package=ggpubr>
- Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*(W1), W55–W63. <https://doi.org/10.1093/nar/gkx305>
- Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J., & Mitchell-Olds, T. (2001). Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis. *The Plant Cell*, *13*(3), 681–693. <https://doi.org/10.1105/tpc.13.3.681>
- Kumar, A., Prabhu, M., Ponnuswami, V., Lakshmanan, V., & Nithyadevi, A. (2014). Scientific seed production techniques in Moringa. *Agricultural Reviews*, *35*(1), 69. <https://doi.org/10.5958/j.0976-0741.35.1.009>
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (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>
- Kumar, Y., Thakur, T. K., Sahu, M. L., & Thakur, A. (2017). A Multifunctional Wonder Tree: Moringa oleifera Lam Open New Dimensions in Field of Agroforestry in India. *International Journal of Current Microbiology and Applied Sciences*, *6*(8), 229–235. <https://doi.org/10.20546/ijcmas.2017.608.031>

- Lai, C.-P., Huang, L.-M., Chen, L.-F. O., Chan, M.-T., & Shaw, J.-F. (2017). Genome-wide analysis of GDSL-type esterases/lipases in Arabidopsis. *Plant Molecular Biology*, 95(1–2), 181–197. <https://doi.org/10.1007/s11103-017-0648-y>
- Leone, A., Spada, A., Battezzati, A., Schiraldi, A., Aristil, J., & Bertoli, S. (2015). Cultivation, Genetic, Ethnopharmacology, Phytochemistry and Pharmacology of Moringa oleifera Leaves: An Overview. *International Journal of Molecular Sciences*, 16(12), 12791–12835. <https://doi.org/10.3390/ijms160612791>
- Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296. <https://doi.org/10.1093/nar/gkab301>
- Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., Baud, S., Bird, D., DeBono, A., Durrett, T. P., Franke, R. B., Graham, I. A., Katayama, K., Kelly, A. A., Larson, T., Markham, J. E., Miquel, M., Molina, I., Nishida, I., ... Ohlrogge, J. (2013). Acyl-Lipid Metabolism. *The Arabidopsis Book*, 11, e0161. <https://doi.org/10.1199/tab.0161>
- Liu, R., Liu, J., Huang, Q., Liu, S., & Jiang, Y. (2022). Moringa oleifera: a systematic review of its botany, traditional uses, phytochemistry, pharmacology and toxicity. *The Journal of Pharmacy and Pharmacology*, 74(3), 296–320. <https://doi.org/10.1093/jpp/rgab131>
- Maddison, W. P., & Maddison, D. R. (2021). *Mesquite: a modular system for evolutionary analysis* (3.70). <http://www.mesquiteproject.org>.
- Maere, S., de Bodt, S., Raes, J., Casneuf, T., van Montagu, M., Kuiper, M., & van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences*, 102(15), 5454–5459. <https://doi.org/10.1073/pnas.0501102102>
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10), 4647–4654. <https://doi.org/10.1093/molbev/msab199>
- Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H. Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., ... Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1), D351–D360. <https://doi.org/10.1093/nar/gky1100>
- Olson, M. E. (2002). Combining Data from DNA Sequences and Morphology for a Phylogeny of Moringaceae (Brassicales). *Systematic Botany*, 27(1), 55–73. <https://doi.org/https://doi.org/10.1043/0363-6445-27.1.55>
- Olson, M. E. (2003). Ontogenetic origins of floral bilateral symmetry in Moringaceae (Brassicales). *American Journal of Botany*, 90(1), 49–71. <https://doi.org/10.3732/ajb.90.1.49>
- Olson, M. E. (2017). Moringa frequently asked questions. *Acta Horticulturae*, 1158(1158), 19–32. <https://doi.org/https://doi.org/10.17660/ActaHortic.2017.1158.4>

- Olson, M. E., Sankaran, R. P., Fahey, J. W., Grusak, M. A., Odee, D., & Nouman, W. (2016). Leaf Protein and Mineral Concentrations across the “Miracle Tree” Genus Moringa. *PLOS ONE*, *11*(7), e0159782. <https://doi.org/10.1371/journal.pone.0159782>
- Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*(4), 2294–2316. <https://doi.org/10.1104/pp.16.00523>
- Papp, B., Pál, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, *424*(6945), 194–197. <https://doi.org/10.1038/nature01771>
- Petersen, A., Hansen, L. G., Mirza, N., Crocoll, C., Mirza, O., & Halkier, B. A. (2019). Changing substrate specificity and iteration of amino acid chain elongation in glucosinolate biosynthesis through targeted mutagenesis of *Arabidopsis* methylthioalkylmalate synthase 1. *Bioscience Reports*, *39*(7). <https://doi.org/10.1042/BSR20190446>
- Prajapati, C., Ankola, M., Upadhyay, T. K., Sharangi, A. B., Alabdallah, N. M., Al-Saeed, F. A., Muzammil, K., & Saeed, M. (2022). Moringa oleifera: Miracle Plant with a Plethora of Medicinal, Therapeutic, and Economic Importance. *Horticulturae*, *8*(6), 492. <https://doi.org/10.3390/horticulturae8060492>
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>
- Raja, S., Bagle, B. G., & More, T. A. (2013). Drumstick (*Moringa oleifera* Lamk.) improvement for semiarid and arid ecosystem: Analysis of environmental stability for yield. *Journal of Plant Breeding and Crop Science*, *5*(8), 164–170. <https://doi.org/10.5897/jpbcs12.029>
- Ram, H. H., Kushwaha, S., & Dubey, R. K. (2020). A glimpse of indigenous and minor vegetables of India. *Indian Horticulture*, *65*(3), 25–29. <https://www.researchgate.net/publication/343576872>
- Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., Perez-Torres, C. A., Chacón-López, A., Hernandez-Guzmán, G., Chang, T. H., Farr, K. M., Brad 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 of the United States of America*, *116*(34), 17081–17089. <https://doi.org/10.1073/pnas.1822129116>
- Rodman, J. E., Soltis, P. S., Soltis, D. E., Sytsma, K. J., & Karol, K. G. (1998). Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American Journal of Botany*, *85*(7), 997–1006. <https://doi.org/10.2307/2446366>
- Rokas, A., Wisecaver, J. H., & Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology*, *16*(12), 731–744. <https://doi.org/10.1038/s41579-018-0075-3>
- RStudio Team. (2019). *RStudio: Integrated Development Environment for R*. <http://www.rstudio.com/>
- Shen, L., & Icahn School of Medicine at Mount Sinai. (2021). *GeneOverlap: Test and visualize gene overlaps*. <http://shenlab-sinai.github.io/shenlab-sinai/>

- Shyamli, P. S., Pradhan, S., Panda, M., & Parida, A. (2021). De novo Whole-Genome Assembly of *Moringa oleifera* Helps Identify Genes Regulating Drought Stress Tolerance. *Frontiers in Plant Science*, 12. <https://doi.org/10.3389/fpls.2021.766999>
- Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, 6(7), e21800. <https://doi.org/10.1371/journal.pone.0021800>
- Tak, S., & Maurya, I. B. (2017). Genetic diversity of *Moringa oleifera* Lam. in Rajasthan, India. *Acta Horticulturae*, 1158, 71–78. <https://doi.org/10.17660/ActaHortic.2017.1158.9>
- Tasdighian, S., van Bel, M., Li, Z., van de Peer, Y., Carretero-Paulet, L., & Maere, S. (2017). Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. *The Plant Cell*, 29(11), 2766–2785. <https://doi.org/10.1105/tpc.17.00313>
- Tian, Y., Zeng, Y., Zhang, J., Yang, C., Yan, L., Wang, X., Shi, C., Xie, J., Dai, T., Peng, L., Zeng Huan, Y., Xu, A., Huang, Y., Zhang, J., Ma, X., Dong, Y., Hao, S., & Sheng, J. (2015). High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. *Science China Life Sciences*, 58(7), 627–638. <https://doi.org/10.1007/s11427-015-4872-x>
- Tohge, T., & Fernie, A. R. (2020). Co-regulation of Clustered and Neo-functionalized Genes in Plant-Specialized Metabolism. *Plants*, 9(5), 622. <https://doi.org/10.3390/plants9050622>
- Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44(W1), W232–W235. <https://doi.org/10.1093/nar/gkw256>
- Trigo, C., Castelló, M. L., Ortolá, M. D., García-Mares, F. J., & Desamparados Soriano, M. (2020). *Moringa oleifera*: An Unknown Crop in Developed Countries with Great Potential for Industry and Adapted to Climate Change. *Foods*, 10(1), 31. <https://doi.org/10.3390/foods10010031>
- Vaknin, Y., & Mishal, A. (2017). The potential of the tropical “miracle tree” *Moringa oleifera* and its desert relative *Moringa peregrina* as edible seed-oil and protein crops under Mediterranean conditions. *Scientia Horticulturae*, 225, 431–437. <https://doi.org/10.1016/j.scienta.2017.07.039>
- Wall, L., Christiansen, T., & Orwant, J. (2000). *Programming perl*. “O’Reilly Media, Inc.”
- Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. -h., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7), e49–e49. <https://doi.org/10.1093/nar/gkr1293>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., & Seidel, D. (2020). *scales: Scale Functions for Visualization*. <https://CRAN.R-project.org/package=scales>
- Xu, F., & Xue, H. (2019). The ubiquitin-proteasome system in plant responses to environments. *Plant, Cell & Environment*, 42(10), 2931–2944. <https://doi.org/10.1111/pce.13633>