







ORIGINAL RESEARCH

Chromosome-scale assembly of the *Moringa oleifera* Lam. genome uncovers polyploid history and evolution of secondary metabolism pathways through tandem duplication

Jiyang Chang^{1,2,#} | Juan Pablo Marczuk-Rojas^{3,4,#} | Carrie Waterman⁵  |
 Armando Garcia-Llanos⁶ | Shiyu Chen⁶  | Xiao Ma^{1,2} | Amanda Hulse-Kemp^{7,8}  |
 Allen Van Deynze⁶  | Yves Van de Peer^{1,2,9,10}  | Lorenzo Carretero-Paulet^{3,4} 

¹Dep. of Plant Biotechnology and Bioinformatics, Ghent Univ., Ghent 9052, Belgium

²Center for Plant Systems Biology, VIB, Ghent 9052, Belgium

³Dep. of Biology and Geology, Univ. of Almería, Ctra. Sacramento s/n, Almería 04120, Spain

⁴Centro de Investigación de Colecciones Científicas de la Universidad de Almería (CECOUAL), Univ. of Almería, Ctra. Sacramento s/n, Almería 04120, Spain

⁵Dep. of Nutrition, Univ. of California, Davis, CA 95616, USA

⁶Seed Biotechnology Center, Univ. of California, Davis, CA 95616, USA

⁷Genomics and Bioinformatics Research Unit, USDA-ARS, Raleigh, NC 27695, USA

⁸Dep. of Crop and Soil Sciences, North Carolina State Univ., Raleigh, NC 27695, USA

⁹Dep. of Biochemistry, Genetics and Microbiology, Univ. of Pretoria, Pretoria 0028, South Africa

¹⁰College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural Univ., Nanjing 210095, China

Correspondence

Lorenzo Carretero-Paulet, Dep. of Biology and Geology, Univ. of Almería, Ctra. Sacramento s/n, Almería, Spain, 04120.

Yves Van de Peer, Dep. of Plant Biotechnology and Bioinformatics, Ghent Univ., Ghent, Belgium, 9052.

Allen Van Deynze, Seed Biotechnology Center, Univ. of California, Davis, CA 95616, USA.

Emails: lpaulet@ual.es;
yvpee@psb.vib-ugent.be;
avandeynze@ucdavis.edu

Assigned to Associate Editor Eric von Wettberg.

#These authors have contributed equally to this work and share first authorship.

Abstract

The African Orphan Crops Consortium (AOCC) selected the highly nutritious, fast growing and drought tolerant tree crop moringa (*Moringa oleifera* Lam.) as one of the first of 101 plant species to have its genome sequenced and a first draft assembly was published in 2019. Given the extensive uses and culture of moringa, often referred to as the multipurpose tree, we generated a significantly improved new version of the genome based on long-read sequencing into 14 pseudochromosomes equivalent to $n = 14$ haploid chromosomes. We leveraged this nearly complete version of the moringa genome to investigate main drivers of gene family and genome evolution that may be at the origin of relevant biological innovations including agronomical favorable traits. Our results reveal that moringa has not undergone any additional whole-genome duplication (WGD) or polyploidy event beyond the gamma

Abbreviations: AOCC, African Orphan Crops Consortium; BUSCO, Benchmarking Universal Single-Copy Orthologs; EC, Enzyme Commission; GO, Gene Ontology; Iso-seq, isoform sequencing; KEGG, Kyoto Encyclopedia of Genes and Genomes; KO, KEGG Orthology; Ks, synonymous substitution rate; ML, maximum likelihood; RNA-seq, RNA sequencing; SMGC, secondary metabolite gene cluster; SSD, small-scale duplication; WGD, whole-genome duplication.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

[Correction added on 4 August 2022, after first online publication: in Abstract section “lambda” is replaced by “gamma”.]

WGD shared by all core eudicots. Moringa duplicates retained following that ancient gamma events are also enriched for functions commonly considered as dosage balance sensitive. Furthermore, tandem duplications seem to have played a prominent role 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 and might, at least partially, explain the outstanding phenotypic plasticity attributed to this species. This study provides a genetic roadmap to guide future breeding programs in moringa, especially those aimed at improving secondary metabolism related traits.

1 | INTRODUCTION

The perennial tree crop moringa (*Moringa oleifera* Lam.) is the most widely cultivated species from the 13 species belonging to the genus, which collectively conform the monogeneric taxonomic family Moringaceae (Olson, 2002). In turn, Moringaceae forms a clade sister to the Caricaceae, the family to which papaya (*Carica papaya* L.) belongs, within the Brassicales order (Olson, 2003). Possibly indigenous to sub-Himalayan hilly lowlands in northwestern India (Olson, 2017), moringa has been widely cultivated since ancient times throughout warm and semi-arid tropical and subtropical areas of the world, specifically in Africa, Asia, Latin America, the U.S. state of Florida, the Caribbean, and the Pacific Islands (Gandji et al., 2018). Moringa has been credited with high morphological and biochemical plasticity, which allows the crop to adapt to very different local environments and to tolerate stressful conditions especially drought, heat, and UVB radiation, stressors that are expected to aggravate under global climate change (Araujo et al., 2016; Brunetti et al., 2020; Brunetti et al., 2018). A successful strategy used by moringa to cope with prolonged periods of drought involves long-term investment of resources into multiple secondary metabolites such as isoprenoids and flavonoids, thus limiting severe photo-inhibitory processes and oxidative damages. Furthermore, moringa has a deep root system and is an exceptionally fast-growing tree; it can reach 3 m high in just 3 mo and up to 12–15 m in few years and is often used in agroforestry (Devkota & Bhusal, 2020; Y. Kumar et al., 2017). Nowadays, its culture is spreading to other areas characterized by low soil nutrients and water availability and high annual low temperatures including the Mediterranean basin (Trigo et al., 2020; Vaknin & Mishal, 2017).

Nicknamed the multipurpose tree, different parts of the tree, including leaves, flowers, pods, and seeds, are edible and used in human nutrition and livestock forage for their rich nutritional content (Olson et al., 2016). The leaves contain the greatest amount of nutrients compared with other parts of the plant and are especially abundant in proteins, including

several essential amino acids, providing an alternative protein source to meet the regular demand of malnourished people, as well as vitamins A, B, C, D, and E and minerals such as Ca, Fe, K, Se, Zn, or Mg, among others, satisfying the required dietary amounts of these micronutrients required for proper growth and development (Islam et al., 2021; Leone et al., 2015; Trigo et al., 2020). The seeds yield a high-oleic oil used in cooking, cosmetics, and as a machinery lubricant. After oil extraction, the remaining seed cake can be used to clarify turbid water or to increase protein in animal feed or crop fertilizer. Other uses of leaves, seeds and, or other parts of the moringa plant include but are not limited to biomass production for biodiesel or biogas, domestic cleaning agent, blue dye, fencing, gum, pulp, ornamental, biopesticide, and honey and sugar cane juice clarifier (Gandji et al., 2018; Leone et al., 2015).

Besides these uses, moringa and related species have been used in traditional medicine for millennia as they are featured by the production of a wide and diverse range of secondary metabolites, including carotenoids, steroids and other isoprenoids, saponins, tannins, phenolic acids, chlorogenic acids, and fatty acids, for which diverse pharmacological roles as bioactive compounds are under study (Abd Rani et al., 2018; Leone et al., 2015). Of these, flavonoids, alkaloids, and glucosinolates are of special interest. Flavonoids, a class of polyphenolic secondary metabolites, are present in large amounts in moringa mostly in the flavanol and glycoside form (Abd Rani et al., 2018; Leone et al., 2015). Rhamnetin, apigenin, myricetin, quercetin, and kaempferol are the main flavonoids found in moringa leaves and are responsible of most of the antioxidant activity of the plant (Bennett et al., 2003; Siddhuraju & Becker, 2003). Epidemiological studies have consistently shown that high intake of flavonoids has protective effects against many infectious diseases and against cardiovascular, kidney and neurodegenerative diseases, cancers, and other age-related diseases (Kou et al., 2018; Leone et al., 2015; Siddhuraju & Becker, 2003). The presence of alkaloids, a family of cyclic organic secondary metabolic compounds containing nitrogen, has also been confirmed in

moringa leaves, notably marumosiide A and marumosiide B together with pyrrolemarumine-4"-*O*- α -*L*-rhamnopyranoside (Sahakitpichan et al., 2011), for which cardioprotective, anti-inflammatory and antihypertensive effects have been described (Dangi et al., 2002; Panda et al., 2013; Vudhigiri et al., 2016). Finally, glucosinolates, a family of sulfur-containing secondary metabolites synthesized from amino acids in Brassicaceae species and related families from the Brassicales order, are also produced in all parts of the moringa plant, particularly in the form of aromatic glucosinolates derived from tryptophan and others from phenylalanine (Bennett et al., 2003). Ongoing investigation is providing evidence of very diverse medicinal properties including antioxidant, anti-inflammatory, antibiotic, neuroprotective, cytoprotective, chemoprotective and cancer-suppressing for glucosinolates, and, especially, their isothiocyanate counterparts derived from glucosinolate hydrolysis through the action of specific myrosinase enzymes (Dinkova-Kostova & Kostov, 2012; Fahey et al., 2018; Jaafaru et al., 2019). The most abundant glucosinolates produced by moringa are glucomoringin, glucosoonjnain, and their acetylated derivatives. Given that glucosoonjnain is apparently mostly responsible for the bitter harsh taste of leaves and seeds, it is not surprising that glucosoonjnain levels could have been selected against during domestication of this species (Chodur et al., 2018), suggesting the genetic toolkit involved in the biosynthesis of these compounds could be a valuable target in future breeding programs to alter their relative distribution according to the purpose of the end product (food or source of pharmacologically active ingredients).

The African Orphan Crop Consortium (AOCC) emerged to promote the research and production of neglected or underutilized (orphan) local plants but with great agronomic potential and nutritional content (Jamnadass et al., 2020). For this purpose, the consortium selected 101 orphan species from indigenous crops of the African continent and other naturalized exotic species (e.g., moringa) to sequence their genome and transcriptome. Moringa was one of the first species selected by the AOCC, and a draft of its genome was published in 2019 (Chang et al., 2019), herein called AOCC v1, which came after the publication of the first actual draft moringa genome by an independent group in 2015 (Tian et al., 2015). However, the unambiguous identification of genes of agronomic interest and associated molecular markers and, ultimately, the development of genomics assisted plant breeding programs in moringa requires genome assemblies capturing the full complement and order of genes. To fulfill this demand, we present here a significantly improved, chromosome-scale reference version of the moringa genome based on Oxford Nanopore long reads (AOCC v2). We leveraged this novel version of the moringa genome to uncover its whole-genome duplication (WGD) or polyploidy history, decipher the genetic toolkit involved in diverse secondary

Core Ideas

- We present a chromosome-scale genome sequence of moringa based on long reads.
- Only the gamma polyploidy event was detected in the moringa genome.
- Whole-genome duplicates in moringa are enriched for dosage balance functions.
- We characterized the genomic organization of secondary metabolism genes in moringa.
- Tandem duplications may have contributed to moringa's phenotypic plasticity.

metabolism pathways relevant to the plant, and to investigate the role of tandem duplications in promoting genomic plasticity through the evolution of specific defense response, secondary metabolism, and developmental pathways.

2 | MATERIAL AND METHODS

2.1 | Sample collection, library construction, and sequencing

Genomic DNA was extracted from a moringa accession Mtongwe1, collected in Mtongwe, Kenya on 26 Aug. 2016 by Kenya Forestry Research Institute using OmniPrep for Plant (<https://www.gbiosciences.com/Molecular-Biology/OmniPrep-for-Plant>) and evaluated on pulse field gel electrophoresis for quality. Oxford Nanopore long-read genomic library was constructed as per manufacturer (Oxford Nanopore Technologies, <https://nanoporetech.com/products/kits>). One flow cell was run on Oxford Nanopore Promethion at the University of California–Davis Genome and Biomedical Sciences Facility. It generated 92 Gb sequence data with raw read N50 of 24 Kb. Adapters were trimmed from raw reads using Porechop v0.2.4 (<https://github.com/rrwick/Porechop>). The longest reads with minimum length of 20 Kb and minimum quality of 7 were kept using Filtrlong v0.2.0 (<https://github.com/rrwick/Filtrlong>).

2.2 | Genome assembly and completeness evaluation

We used Shasta-Linux-0.1.0 on the filtered 60× reads to assemble the genome (Shafin et al., 2020). The total of 7.4 Gb Illumina reads, corresponding to 63 million read pairs 2 by 151 bp, generated by BGI-Shenzhen and China National Gene Bank was used for polishing in RACON v1.4.3

(Chang et al., 2019; Vaser et al., 2017). Then the Shasta-assembled Illumina-polished assembly was scaffolded into the Shasta-HiC version using Chicago and HiRise by DovetailGenomics Omni-C proximity ligation. Estimated physical coverage was 196.48× for short range (Chicago) and 197.77× for long range (HiRise). The short-range assembly started with our nanopore assembly, broke five, and joined 129 contigs, and the long-range HiRise assembly started with the short-range assembly and joined 110 contigs. The long-range assembly was then analyzed for loop resolution using Juicer and 3D Assembly software using default parameters to create final Hi-C scaffolds (Dudchenko et al., 2017; Durand et al., 2016). The final assembly was investigated for misjoins and splits and adjusted to create pseudomolecules in JuiceBox (Durand et al., 2016). It was compared with the long-read assembly recently published by Shyamli et al. (2021) using MUMmer4 (Marçais et al., 2018). Genome assembly completeness was assessed with Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.4 (Manni et al., 2021).

2.3 | Genome structural annotation

We adopted a combination of three strategies that included homology-based predictions, ab initio predictions, and transcriptome-assisted predictions to annotate the protein-coding genes in our genome assembly. For the homology-based predictions, the protein sequences of moringa, *Arabidopsis* [*Arabidopsis thaliana* (L.) Heynh.], black cottonwood (*Populus trichocarpa* Torr. & A. Gray ex Hook.) and sorghum [*Sorghum bicolor* (L.) Moench] were used as query sequences to search the reference genome using TBLASTN (v2.6.0) with different e-value thresholds (moringa with the e-value $\leq 1 \times 10^{-10}$, *Arabidopsis*, black cottonwood, and sorghum with the e-value $\leq 1 \times 10^{-5}$) and the regions mapped by these query sequences were subjected to Exonerate (v2.4.0) (Slater & Birney, 2005) to predict gene models. For ab initio predictions, BRAKER2 (v2.1.2) (Bruna et al., 2021) was used and model training was based on RNA sequencing (RNA-seq) and isoform sequencing (Iso-seq) data after the predicted repeats were soft masked within the assembly. To achieve transcriptome-assisted predictions, five libraries of RNA-seq data with accession numbers SRX3011282 (stem), SRX3011281 (root), SRX3011280 (seed), SRX3011278 (leaf), and SRX3011259 (flower) were downloaded from the NCBI Short-Read Sequence Archive and aligned to the new assembly of the moringa genome by Hisat2 (Supplemental Table S1) (Pasha et al., 2020). The RNA-seq data was assembled into transcripts using Trinity (Haas et al., 2013) and, together with isoforms from Iso-seq, were subjected to the PASA pipeline (v2.4.1) (Haas et al., 2003) to improve the gene structures. Open reading frames were then predicted with TransDecoder

(<https://github.com/TransDecoder/TransDecoder>). The number of reads resulting from Iso-seq was 144,327 with a coverage of ~6×. A total of 142,951 transcripts were obtained, of which 111,228 had predicted open reading frames thanks to the high-coverage RNA-seq data (366-fold). To finalize the gene set, EVIDENCEModeler (v1.1.1) (Haas et al., 2008) was employed to combine all the predictions to produce the nonredundant gene set. Specifically, a set of 1,000 incorrect gene models identified by wgd software (Zwaenepoel & Van de Peer, 2019) and was manually curated using the genome browser GenomeView (<http://genomeview.org/>), and the gene annotation results were evaluated by BUSCO hits.

Repeat families in our genome assembly were de novo identified and classified using RepeatModeler (v2.0) (Flynn et al., 2020). Subsequently, the output data from RepeatModeler were used as a custom repeat library for RepeatMasker (v4.1) (<http://www.repeatmasker.org>) to discover and classify repeats in the assembly with the default parameters. Furthermore, transposable elements not classified by RepeatModeler were analyzed using DeepTE (Yan et al., 2020). Transfer RNAs were predicted by tRNAscan-SE (v1.31) (Lowe & Eddy, 1997) with default parameters. The predictions of noncoding RNA, such as microRNAs, small nuclear RNAs, and ribosomal RNAs, were also performed by comparing with known noncoding RNA libraries (Rfam v14.7) using the cmscan program of Infernal (v1.1.2) (Nawrocki & Eddy, 2013).

2.4 | Genome functional annotation

The proteins encoded by the moringa genome were annotated with Gene Ontology (GO) terms through BLAST2GO v6.0.1 (Conesa & Gotz, 2008). BLAST2GO performs a three-step functional annotation: sequence similarity-based inferences of homology with proteins from the NCBI nonredundant database, mapping, and annotation with the GO terms identified in significant hits. BLAST2GO allows expanding the GO terms with INTERPRO functional domains detected by INTERPROSCAN and Enzyme Commission (EC) enzyme codes represented in the Kyoto Encyclopedia of Genes and Genomes (KEGG) biochemical pathways. In this study, Diamond v2.0.11 with an e-value threshold of 1×10^{-10} (Buchfink et al., 2015) and INTERPROSCAN v5.52-86.0 with default settings (Jones et al., 2014), were used to identify hits corresponding to putative homologous proteins and protein functional domains, respectively.

Additionally, moringa proteins were annotated with KEGG orthology (KO) functional orthology numbers through BlastKOALA (<https://www.kegg.jp/blastkoala/>) (Kanehisa et al., 2016). Only the best scoring KO number was assigned to each gene, resulting in one KO term per gene except for those annotated with two or more best KO numbers with

identical score. The same KO can be involved; however, in several KEGG pathways and can also be found annotating more than one gene.

2.5 | Genome evolution

Analysis of the distribution of synonymous substitution rates (Ks) per synonymous site for pairs of homologous genes was performed using the wgd package and the paranome (entire collection of duplicated genes) was obtained with 'wgd mcl' using all-against-all BlastP and MCL clustering (Zwaenepoel & Van de Peer, 2019). The Ks distribution of moringa was then constructed using 'wgd ksd' with default settings using MAFFT (Katoh & Standley, 2013) for multiple sequence alignment, codeml for maximum likelihood (ML) estimation of Ks values (Yang, 2007), and FastTree for inferring phylogenetic trees used in the node weighting procedure (Price et al., 2009). Pairs of anchor paralog genes (duplicates lying in collinear or syntenic regions of the genome) were obtained using i-ADHoRe (Proost et al., 2012) employing the default settings in 'wgd syn'. Next, 'wgd mix' from the wgd package (Zwaenepoel & Van de Peer, 2019) was used to fit Gaussian mixture models using an expectation–maximization algorithm implemented using the scikit-learn python library. MCscan JCVI was then used to do the analysis of syntenic depth ratio (i.e., the number of times a genomic region is syntenic to a region in another genome) by providing the genome protein-coding sequences and annotation file in gff3 format (Tang et al., 2008).

We used the program Duplicate_gene_classifier from the Multiple Collinearity Scan toolkit (MCScanX, <https://github.com/wyp1125/MCScanX>) (Wang et al., 2012) to classify duplicates in the moringa genome per mechanism of duplication (WGD–segmental, dispersed, proximal and tandem). Paralogous pairs of moringa proteins were first identified using an all-to-all search in Diamond with an e-value cutoff of 1×10^{-10} .

2.6 | Orthogroup and gene family definition

The sequences from the complete proteomes of moringa and 10 other plant species (Supplemental Table S2) were firstly compared all-against-all using Diamond (Buchfink et al., 2015) with a relaxed e-value threshold of 1×10^{-3} in order not to filter out very short sequences in these previous steps and then classified into orthogroups using the clustering algorithm implemented in OrthoFinder v2.5.2 under the default settings (Emms & Kelly, 2019). We used the orthogroup classification method based on hierarchical orthogroups inferred at each hierarchical level (i.e., at each node in the species tree) by analyzing rooted gene trees, which is claimed to be far more

accurate than the gene similarity or graph-based approach used by all other methods and used previously by OrthoFinder (Emms & Kelly, 2019).

2.7 | Identification and characterization of secondary metabolism gene clusters in the moringa genome

We searched for potential biosynthetic gene clusters in the moringa genome that are associated with secondary metabolite biosynthesis using the online tool plantiSMASH v1.0 (<http://plantismash.secondarymetabolites.org/>) with default parameters (Kautsar et al., 2017). The arbitrary definition of a metabolic gene cluster requires that the cluster contains genes for at least three biosynthetic genes of two different types (and closely related duplicate genes are counted only once). For selected genes found in biosynthetic gene clusters, we further identified putative orthologous proteins in 10 other plant species by querying the orthogroups classification generated through OrthoFinder. Phylogenetic analysis of selected orthogroups were performed based on multiple protein sequence alignments obtained using MUSCLE (Edgar, 2004) through SeaView v4.6.4 (Gouy et al., 2010). Next, ML phylogenetic trees were obtained using the online v1.6.12 of the IQ-TREE software (<http://iqtree.cibiv.univie.ac.at/>) (Trifinopoulos et al., 2016). Prior to the analysis, IQ-TREE performs the automated selection of the best fitting substitution model under the Bayesian information criterion. JTTDCMut + I + G4, JTT + G4 and JTT + F + I + G4, with JTT referring to the Jones-Taylor-Thornton model (Jones et al., 1992), I to proportion of invariant sites, F to empirical amino acids frequencies and G4 to heterogeneity in substitution rates modelled using a gamma-distribution with four categories, were chosen as the best fitting amino acid substitution models for HOG0001451, HOG0003249, and HOG0000748 orthogroups, respectively. Three independent branch support analyses (ultrafast bootstrap analysis with 1,000 replicates, SH-aLRT branch test, and approximate Bayes test) were employed to assess the reliability of the internal branches.

3 | RESULTS

3.1 | Genome assembly

Our novel moringa genome assembly (AOCC v2) was 236.4 Mb in length and consisted of 748 scaffolds with a scaffold N50 of 15.0 Mb and L90 of 13 pseudochromosomes (Table 1); 219.1 Mb (i.e., 92.7% of the genome assembly) could be anchored and oriented into 14 pseudochromosomes after scaffolding with Dovetail Genomics Hi-C Omni

TABLE 1 Comparison of assembly statistics between the current moringa genome assembly using long reads (African Orphan Crops Consortium [AOCC] v2), the previous AOCC v1, and Shyamli et al. (2021)

Parameters	AOCC v1	AOCC v2	Shyamli
Platform	HiSeq 2000	Oxford Nanopore	Pacbio sequel
Assembly size, bp	216,759,177	236,366,566	281,946,330
Number of scaffolds	22,329	748	915
Scaffold N50, bp	957,246	14,962,574	4,719,167
Scaffold L50	56	7	17
Scaffold N90, bp	57,837	13,210,789	225,696
Scaffold L90	366	13	115
Largest scaffold, bp	4,637,711	30,079,500	13,807,473
Shortest scaffold, bp	150	13	1,056
Mean scaffold length, bp	9,707.52	315,998.1	308,103.9
GC (%)	36.5	35.7	37.82
Total no. of N	3,014,085	346,401	700
N per 100 kb (%)	1.39	0.1	0.0025

TABLE 2 Comparison of annotation statistics between the current moringa genome assembly using long reads (African Orphan Crops Consortium [AOCC] v2) and the previous AOCC v1

Genome annotation	AOCC v1	AOCC v2
Protein coding genes	18,451	22,714
Mean gene length, bp	3,308	3,340
Mean coding sequence length, bp	1,238	1,208
Mean exon per gene	5	5.5
Mean exon length, bp	232	220
Mean intron length, bp	478	476

C proximity ligation strategy (Supplemental Figure S1). The combination of technologies created a chromosome-level assembly with high accuracy for this diploid species $2n = 28$.

3.2 | Genome structural annotation and completeness

In terms of structural gene annotation, a combination of three strategies (ab initio prediction, protein homology, and RNA-seq-based evidence) and manual curation resulted in 22,714 protein-coding genes, 93.1% of which (21,143 genes) are supported by transcriptome data together with isoforms from Iso-seq in this new moringa assembly (AOCC v2), a number substantially higher than that of the previous one (AOCC v1), which represented approximately 77.9% of its estimated genome size and was annotated with a total of 18,451 genes (Chang et al., 2019; Ojeda-Lopez et al., 2020). On average, predicted protein-coding genes are 3,340 bp long and contain 5.5 exons, values that are slightly higher than that observed in the previous assembly of the moringa genome (Chang et al., 2019; Ojeda-Lopez et al., 2020) (Table 2). Five

TABLE 3 Comparison of annotation statistics between the current moringa genome assembly using long reads (African Orphan Crops Consortium [AOCC] v2) and the previous AOCC v1

BUSCO assessment, <i>n</i> = 1614 genes	AOCC v1	AOCC v2
	%	
Complete	83	99.8
Single	82.1	98.8
Duplicated	0.9	1.0
Fragmented	5.3	0.1
Missing	11.7	0.1

libraries of RNA-seq data from different moringa tissues were downloaded from the NCBI Short Read Sequence Archive (Supplemental Table S1) and aligned to the new moringa genome assembly (Pasha et al., 2020). For all tissues examined, mapping rates were >95% (Supplemental Table S1).

The BUSCO v4.1.4 assessment of the moringa genome assembly completeness was performed against the core embryophyte genes database10. Of the 22,714 predicted genes, 99.8% were found as complete BUSCO genes; 98.8% of the BUSCO genes were single copy and 1% was found in duplicate. Of the remaining 0.2% of the BUSCO genes, 0.1% were found as fragmented and 0.1% were missing (Table 3; Supplemental Figure S2). These figures represent a significant improvement over the AOCC v1 of the moringa genome (Chang et al., 2019), with only 88.8% complete BUSCO genes, 1.6% of the BUSCO genes fragmented, and 8.3% missing (Table 3; Supplemental Figure S2).

As an additional check of the quality of the structural annotation, we also compared the lengths of the encoded proteins predicted in the new version of the moringa genome and the previous one using reciprocal best hits

TABLE 4 Repeat annotation in the moringa genome assembly African Orphan Crops Consortium v2

Type	Length	Percentage of sequence
	bp	%
SINEs	451,948	0.19
LINEs		
R2/R4/NeSL	215,884	0.09
L1/CIN4	4,165,443	1.76
LTR elements	7,759,076	3.28
Copia		
Gypsy	19,033,928	8.05
Retroviral	31,213	0.01
DNA transposons	43,138,971	18.25
Rolling-circles	232,418	0.1
Unclassified	5,185,744	2.19
Small RNA	8,293,172	3.51
Simple repeats	2,756,682	1.17
Low complexity	829,129	0.35

Note. SINE, short-interspersed nuclear element; LINE, long-interspersed nuclear element; LTR, long terminal repeat

(Supplemental Figure S3). The vast majority of the predicted coding sequences were the same length in both versions of the genome; however, a significant proportion had a longer sequence in the newly sequenced moringa genome, further supporting the current genome annotation reported here was improved regarding the previous one.

In terms of repetitive sequences, these accounted for ~38.95% of the new genome assembly, 36.76% of which corresponded to known families of transposable elements. DNA transposons, the largest class of transposable elements found in the moringa genome, represented 18.25% of the total (Table 4; Supplemental Figure S4). Screening the new moringa genome against the Rfam v14.7 database identified a total of 5,434 noncoding RNAs including 1,922 ribosomal RNAs, 1,583 transfer RNAs, 165 small nuclear RNAs, and 151 microRNAs (Supplemental Table S3).

3.3 | Genome functional annotation

Searches for putative homologous of moringa genes resulted in 19,374 genes (85.30% of the total) with at least one hit in Diamond searches against the NCBI nonredundant protein database (Supplemental Table S4). Up to 17,736 encoded protein sequences showed at least one INTERPRO functional domain, including a total of 16,334 PFAM domains (Supplemental Table S4). Furthermore, 10,306 EC codes (totaling 1,139 unique EC codes) were assigned to a total of 8,414 moringa enzyme-coding predicted genes (Supplemental Figure S5).

We performed the functional annotation of the moringa genome with GO terms using the BLAST2GO pipeline merging Diamond, INTERPRO, and EC annotations (Conesa & Gotz, 2008) (Supplemental Figure S5). A substantial fraction of the moringa genes, that is, 16,929 genes out of 22,714 (74.53%), were annotated with at least one GO term, amounting to a total of 66,271 GO terms, which represents an average of three GO terms per annotated gene. Similarly, moringa genes were mapped to their corresponding KO functional orthology groups in the KEGG database using the BlastKOALA tool (Kanehisa et al., 2016). A total of 7,878 genes were mapped onto 3,821 KO functional orthology groups (Supplemental Figure S5).

3.4 | The moringa genome shares the ancient core eudicot hexaploidy

To uncover the history of WGD in the moringa genome, we first identified all paralogs (a total of 8,735 were found, which we referred to as the entire paranome) and paralogs retained in collinear regions (a total of 4,445 were found in 284 syntenic genomic regions, which we referred to as anchor or syntenic paralogs). Next, Ks values were computed for every pair of paralogs and syntenic paralogs and the resulting distributions modeled separately using fitted Gaussian mixture models. The mixture model identified a single main peak in both distributions centered around $K_s = 1.65$, likely corresponding to an ancient WGD event (Supplemental Figure S6). To infer when this ancient WGD event occurred relative to the divergence with grapevine (*Vitis vinifera* L.), taxonomically located at the base of the Rosids clade of angiosperms to which moringa belongs, we compared the distribution of Ks values of the moringa syntenic paralogs to the distribution of Ks values of syntenic orthologs between moringa and grapevine. The latter distribution showed a Ks peak around 0.98, located to the left of the main peak found for moringa paralogs (Figure 1a), suggesting that the ancient WGD event detected in the moringa genome occurred prior to the divergence of moringa and grapevine and thus likely corresponds to the Gamma event shared by all core eudicots (Jiao et al., 2012).

In addition, we compared the syntenic depth ratio (that is, the number of times a genomic region is syntenic to a region in another genome) between moringa and *Amborella trichopoda* Baill., the single extant species of a lineage sister to all flowering plants that has not experienced any WGD since the divergence of angiosperms (*Amborella* Genome Project, 2013). As expected, we observed an overall 3:1 syntenic depth ratio between moringa and *A. trichopoda* (Figure 1b), which means that a single *A. trichopoda* region generally aligns to three moringa genomic segments, further supporting that moringa would have undergone the hexaploidy event

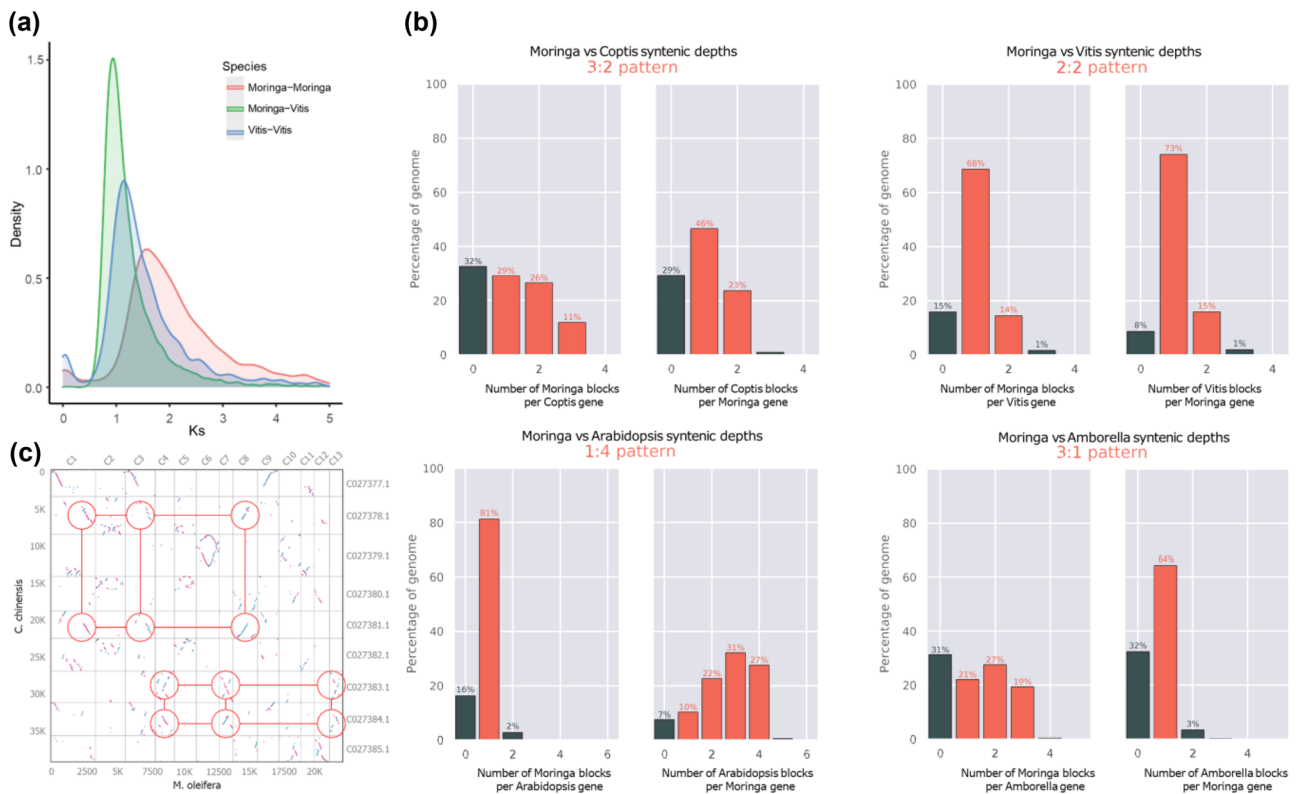


FIGURE 1 WGD analysis of the moringa genome. (a) Density plots resulting from fitting Gaussian mixture models to synonymous substitution rates per synonymous site (Ks) distributions for pairs of syntenic paralogues within the moringa and grapevine genomes as well as of syntenic orthologues between moringa and grapevine. (b) Syntenic depth ratios between moringa and *Amborella trichopoda*, *Coptis chinensis*, grapevine, and *Arabidopsis* genomes. (c) Dot plots of orthologues between moringa and *Coptis chinensis* genomes. The red circles highlight examples of major duplication events suggestive of a 3:2 syntenic relationship

shared by all core eudicots. Furthermore, we also compared moringa with Chinese goldthread (*Coptis chinensis* Franch.), which belongs to an early diverging eudicot lineage in which a single round of WGD was identified but lacking the gamma triplication event (Y. Liu et al., 2021). Consistent with our hypothesis of a single WGD in moringa, we found a 3:2 syntenic depth ratio between moringa and Chinese goldthread (Figure 1b,c). Further comparing moringa to grapevine and Arabidopsis, we found a 1:1 synteny relationship between moringa and grapevine (Figure 1b) and a 1:4 synteny relationship between moringa and Arabidopsis (Figure 1b), again consistent with the grapevine and moringa genomes having a hexaploid origin (The French–Italian Public Consortium for Grapevine Genome Characterization, 2007), and Arabidopsis experiencing two additional rounds of recent WGD events, the so-called α and β WGD events (Blanc et al., 2003; Bowers et al., 2003).

3.5 | Gene family evolution in moringa

To study gene family evolution in moringa, we first obtained a classification of orthogroups in the moringa genome and 10 plant species representing the main angiosperm plant lin-

eages (Figure 2a; Supplemental Tables S2 and S5). These included the Brassicales Arabidopsis and papaya, the legume barrel clover (*Medicago truncatula* Gaertn.), the basal Rosid grapevine, the Asterid tomato (*Solanum lycopersicum* L.), the basal dicot sacred lotus (*Nelumbo nucifera* Gaertn.), the monocots rice (*Oryza sativa* L. subsp. *japonica* Kato) and maize (*Zea mays* L.), the magnoliid avocado (*Persea americana* Mill.), and the early diverging angiosperm *A. trichopoda*, which is sister to the rest of flowering plants. A total of 349,936 protein sequences encoded by the genomes of these species were compared with each other, and, based on this comparison, 299,745 genes (i.e., 85.66% of the total) could be classified into 28,161 orthogroups containing at least two genes; 5,185 of these orthogroups were found across all 11 species (Figure 2b; Supplemental Table S5). The remaining 50,191 genes, including 2,668 moringa genes, were classified as unassigned to any orthogroup, likely corresponding to singleton orphan sequences with no detected homologs in any species (Supplemental Table S5). In the moringa genome, 20,046 genes (nearly 88.25% of the total) were assigned into 13,597 orthogroups (Figure 2b; Supplemental Table S5), of which 148 orthogroups, grouping together 941 genes, were found to be unique to moringa. Moringa

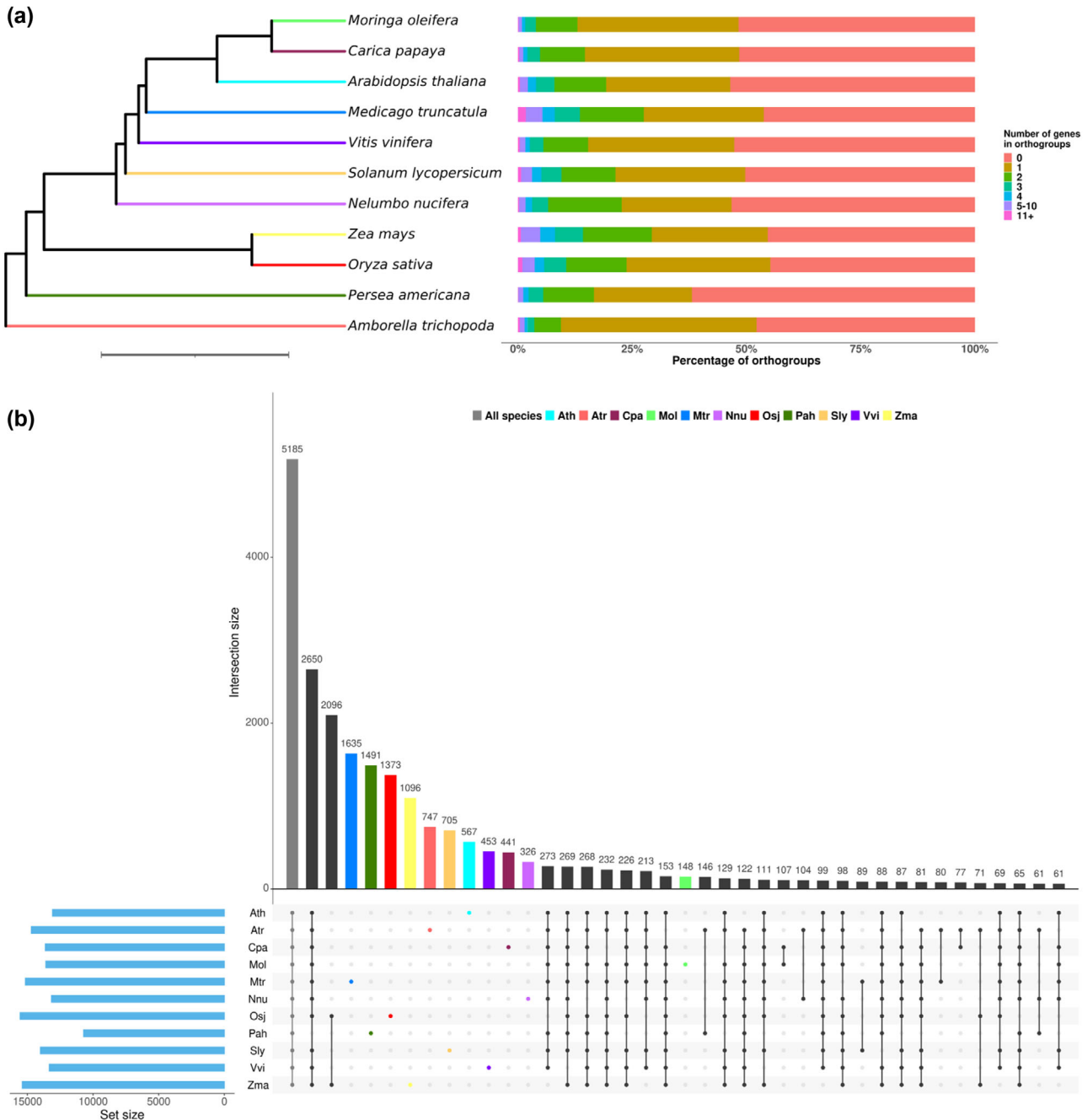


FIGURE 2 Gene family and orthogroup classification in moringa and 10 other plant species representing the main flowering plants lineages. (a) Ultrametric phylogenetic tree depicting the evolutionary relationships among moringa and 10 other plant species (left panel) and histogram representing in percentages the gene distribution per species resulting from the gene family classification (right panel). Tree topology and divergence times were taken from TimeTree (S. Kumar et al., 2017), except for the branching of the Magnoliid avocado, for which clustering as a sister clade to the large clade formed by monocots plus dicots species was favored. Branches in the trees are proportional to evolutionary time, with the scale bar representing 100 million yr. Species names abbreviations: Mol, *Moringa oleifera*; Ath, *Arabidopsis thaliana*; Cpa, *Carica papaya*; Mtr, *Medicago truncatula*; Vvi, *Vitis vinifera*; Sly, *Solanum lycopersicum*; Nnu, *Nelumbo nucifera*; Osj, *Oryza sativa* (rice); Zma, *Zea mays*; Pah, *Persea americana*; and Atr, *Amborella trichopoda*. (b) UpSet plot representing the number of families (bars) containing genes from a specific species or set of species (dots). Only intersections spanning >60 families are displayed. The core orthogroups, that is, the orthogroups formed by members of all 11 species, are shown with grey bars and dots, whereas the bars and dots corresponding to unique orthogroup clusters are colored according to the species color scheme from the tree in panel (a). The histogram on the left represents the total number of gene families for each species

represents the second species, only after lotus, with the smallest percentage of genes found in species-specific orthogroups (Supplemental Table S5).

Although only 108 genes out of the 941 genes clustering in moringa-specific orthogroups were annotated with GO terms, functional enrichment tests identified specific developmental processes including seedling development, meristem growth, and male and female gamete generation, and defense responses such as response to misfolded protein enriched among them (Supplemental Table S6). Interestingly, most of the genes annotated with both sets of GO terms belong to the same orthogroup, HOG0020145, which is composed of three moringa genes annotated as encoding a specific subunit of the 26 proteasome complex (Supplemental Table S6). Next, we investigated whether any mechanism of duplication was prevalent in the expansion of moringa-specific orthogroups. For this purpose, we obtained a classification of gene duplicates in the moringa genome by mechanism of duplication using the MCScanX software: whole-genome or segmental duplications (collinear genes in collinear blocks); different forms of small-scale duplications (SSDs), including tandem; duplications in consecutive regions of the genome, proximal; duplications in nearby chromosomal region but not adjacent or dispersed; duplications of modes other than tandem, proximal, or WGD–segmental (Supplemental Figure S7). Moringa-specific orthogroups were strongly enriched for tandem (200 total genes; Fisher's exact test, $P = 1.1 \times 10^{-29}$) and proximal duplicates (145 total genes; Fisher's exact test, $P = 2 \times 10^{-41}$), while neither enriched nor impoverished for dispersed (271 genes; Fisher's exact test, $P = 1$) or WGD duplicates (106 genes; Fisher's exact test, $P = 1$). Therefore, tandem and proximal SSD duplications appeared to be the prevalent mechanisms behind the expansion of moringa-specific orthogroups.

Previous studies have reported notable differences in the evolutionary and functional fate of duplicates depending on the mechanism or mode of duplication. For example, genes with certain biological functions (e.g., transcriptional regulation, signal transduction, protein transport, and protein modification) are preferentially retained after WGDs, whereas they are rarely retained after SSDs and vice versa (Maere et al., 2005). The so-called dosage balance hypothesis is claimed to predict such biased pattern of loss and retention between WGD and SSD duplicates (Freeling, 2009). To check whether the predictions of the dosage balance hypothesis applied to moringa, we performed GO functional enrichment tests of duplicates categorized by mechanism of duplication. The GO terms most significantly enriched among WGD duplicates were related to transcriptional regulation followed by different forms of protein modification, including phosphorylation and kinase activities or protein dimerization, functions commonly considered as dosage balance sensitive (Figures 3a,b; Supplemental Table S7).

On the other hand, tandem duplicates have been commonly observed to be retained in a lineage-specific fashion and enriched in functional categories related to response to stress or secondary metabolism (Carretero-Paulet & Fares, 2012; Chae et al., 2014; Deneud et al., 2014). Indeed, moringa tandem duplicates were enriched in GO terms corresponding to specific secondary metabolism enzymes such as isoprenoids (lanosterol synthase activity, beta-amyrin synthase activity), alkaloids (reticuline oxidase, strictosidine synthase) phenylpropanoids/flavonoids/glucosinolates (S-adenosylmethionine-dependent methyltransferase), glycosylated flavonols (quercetin 3-O-glucosyltransferase, quercetin 7-O-glucosyltransferase), and glutathione (glutathione transferase) (Figures 3c,d; Supplemental Table S8) or specific secondary metabolism pathways (e.g., anthocyanin-containing compound biosynthetic process, aromatic compound biosynthetic process, triterpenoid biosynthetic process) (Figures 3c,d; Supplemental Table S8). In addition, several GO terms associated with defense response of plants against biotic and abiotic environmental cues were found as enriched among tandem duplicates including defense response, xenobiotic detoxification by transmembrane export across the plasma membrane, response to toxic substance, defense response to other organism, killing of cells of other organism, and response to biotic stimulus (Figures 3c,d; Supplemental Table S8).

To further check whether genes involved in secondary metabolism were enriched for tandem duplicates in moringa, we mapped 395 KO groups corresponding to a total of 1,181 (810 unique) genes in the moringa genome onto 50 secondary metabolism pathways (Supplemental Table S9) using KEGG pathway annotation. Biochemical pathways involved in the metabolism of amino acids, which serve as precursors for a wide range of secondary metabolites including phenolic compounds, alkaloids, or glucosinolates, were also considered. One hundred twenty-one of these genes corresponded to tandem duplicates, which is significantly higher than expected by chance (Fisher's exact test, $P = 1.2 \times 10^{-7}$). Furthermore, dispersed duplicates, accounting for 324 of the secondary metabolism genes, were also found to be in a higher proportion than expected by chance, although less significantly (Fisher's exact tests, $P = 3 \times 10^{-5}$). In contrast, WGD (183 genes; Fisher's exact test, $P = .066$) or proximal (39 genes; Fisher's exact test, $P = .3$) duplicates appeared to be neither underrepresented nor overrepresented among secondary metabolism genes.

3.6 | Secondary metabolite biosynthetic pathways in the moringa genome

Given the importance of secondary metabolites to the nutritional content, organoleptic properties, and pharmacological

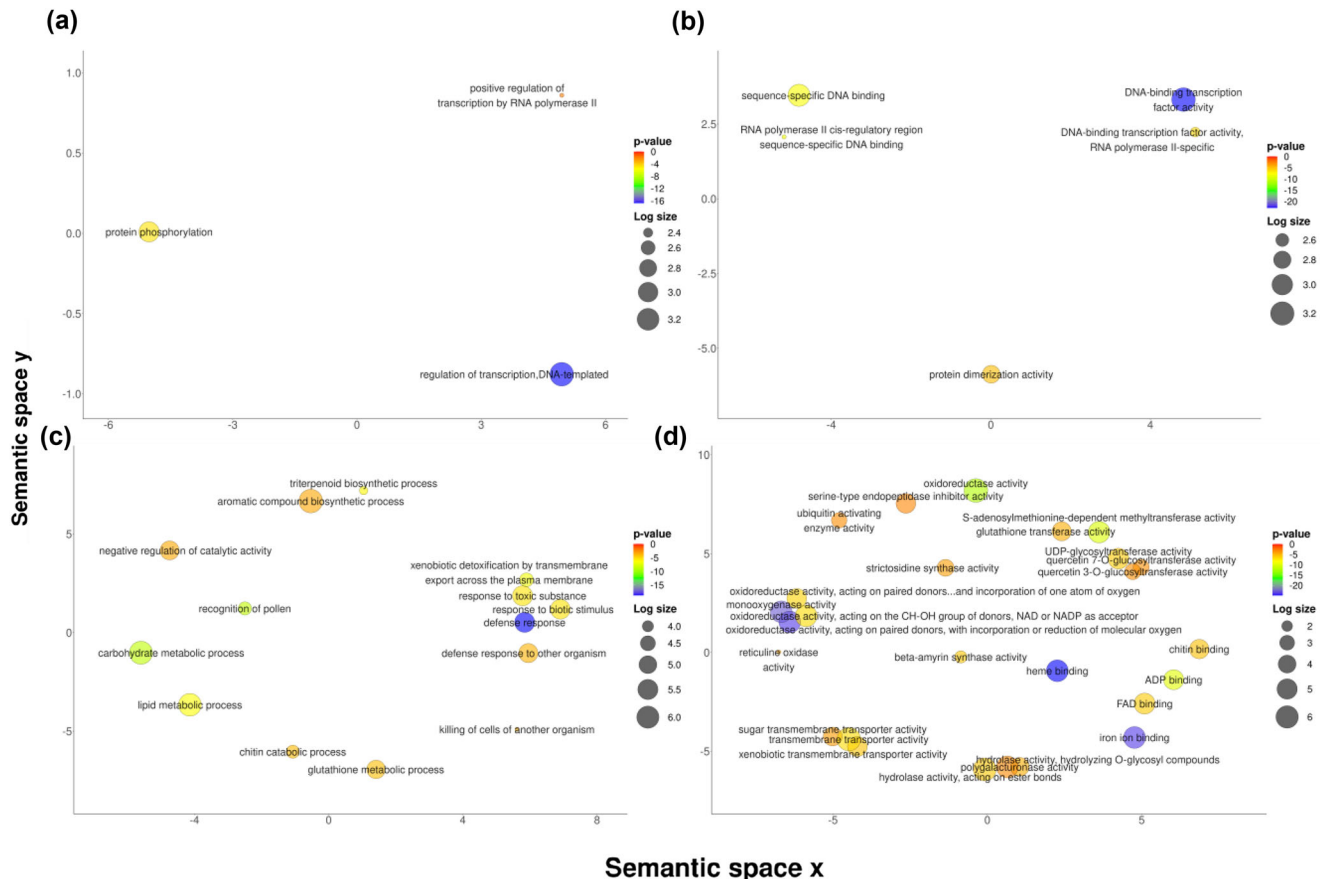


FIGURE 3 ReViGo scatterplot representation of gene ontology (GO) biological processes significantly enriched among selected subsets of moringa genes. Results are shown for genes belonging to moringa (a, b) whole-genome duplication (WGD) duplicates and (c, d) tandem duplicates for (a, c) biological process and (b, d) molecular function GO terms. The GO terms found as significantly enriched are represented as circles, with diameters proportional to the sample size (number of genes with the corresponding term as compared with the background protein database) and the resulting \log_{10} of the p values resulting from Fisher's exact tests corrected after Bonferroni and color encoded. ReViGo groups functionally similar GO terms close in the two-dimensional semantic space. Results were obtained using the ReViGo web server at <http://revigo.irb.hr/> with *Arabidopsis* selected as background protein database and the rest of settings as default (Supek et al., 2011)

activity of moringa leaves and seeds, we performed the identification and molecular evolutionary characterization of gene families putatively involved in secondary metabolism focusing on two main drivers of the evolution of metabolism pathways in plants involving clusters of tandemly duplicated genes: neofunctionalization of tandem gene duplicates and biosynthetic gene clusters (Tohge & Fernie, 2020).

Neofunctionalization of tandem gene duplicates resulting in differential substrate specificity of the gene copies has been identified as a major mechanism driving the evolution of specific secondary metabolism pathways, notably, glucosinolate biosynthesis in *Arabidopsis* (Tohge & Fernie, 2020). By merging relevant GO, KEGG, and EC annotations, we attempted to reconstruct the full complement of genes putatively involved in glucosinolate biosynthesis and regulation, especially those encoding for enzymatic activities, in the moringa genome and the two other Brassicales representatives examined in this work (*Arabidopsis* and papaya) (Supplemental Table S10). Supplemental Figure S8 shows

the EC enzymatic activities found as encoded in the moringa genome mapped onto the glucosinolate biosynthesis KEGG pathway map (map entry 00966). A total of 104 moringa genes likely involved in glucosinolate metabolism were found, which grouped into 28 out of the 33 detected orthogroups including 11 with at least a pair of tandem duplicates (Supplemental Table S10). Two interesting instances were orthogroups HOG0000588 and HOG0009822. Orthogroup HOG0000588 included 15 moringa genes arranged in two clusters of tandem duplicates located on chromosomes 1 and 5, together with 18 and 14 in *Arabidopsis* and papaya, respectively, and zero to 24 (grapevine) in the rest of species. Genes in HOG0000588 were annotated as encoding for indol-3-yl-methylglucosinolate hydroxylases. Orthogroup HOG0009822 clusters seven moringa genes tandemly arranged in consecutive positions of chromosome 10, for five and three in *Arabidopsis* and papaya, respectively, and zero in the rest of species. Genes in HOG0009849 are annotated as encoding for GDSL-type esterases/lipases, a

family of lipid hydrolytic enzymes with multifunctional properties such as broad substrate specificity and regioselectivity involved in diverse secondary metabolism pathways, including glucosinolates (Lai et al., 2017). In contrast, moringa orthologs could not be detected for Arabidopsis glucosinolate-related genes in orthogroups HOG0016593, HOG0022085, HOG0017508, HOG0008572, HOG0004358, and HOG0019714 encoding for branched-chain amino acid aminotransferase (EC:2.6.1.42), homomethionine N-monooxygenase (EC:1.14.14.42), aliphatic glucosinolate S-oxygenase (EC:1.14.13.237), magnesium transporter, beta-glucosidase (EC:3.2.1.21), and myrosinase (E3.2.1.147), respectively (Supplemental Table S10). However, moringa genes annotated with such enzymatic activities were found in other glucosinolate orthogroups, except for homomethionine N-monooxygenase and aliphatic glucosinolate S-oxygenase (Supplemental Table S10). Notably, orthogroup HOG0000577, annotated as encoding for a specific class of beta glucosidases (myrosinases) involved in the production of biologically active glucosinolates, was found as strongly expanded in moringa, with 14 genes, for four and five in Arabidopsis and papaya, respectively (Supplemental Table S10), and zero to six in the rest of species. We also found a single moringa representative in orthogroups HOG0002947 and HOG0006126, encoding for glucosinolate gamma-glutamyl hydrolase (EC:3.4.19.16) and methylthioalkylmalate synthase (EC:2.3.3.17), respectively, both enzyme families reported as examples of diversification in Arabidopsis secondary metabolism through neofunctionalization of tandem duplicates (Kliebenstein et al., 2001; Petersen et al., 2019).

We next used RNA-seq-based data from five tissues to assess expression of the 104 moringa genes identified in our analysis as putatively involved in glucosinolate biosynthesis. Out of the 104 genes, 84 and 88 were found to be expressed in seeds and leaves, respectively, where glucosinolate biosynthesis is more abundant (Figure 4). Moringa tandemly duplicated genes belonging to glucosinolate-related orthogroups showed diversified expression levels across all five tissues, with two genes belonging to orthogroup HOG0009822 (*Morol10g15130* and *Morol10g15170*), another belonging to orthogroup HOG0000577 (*Morol03g00630* and *Morol04g12880*) and HOG0000588 (*Morol01g14150*) displaying notable expression in leaves and seeds (Figure 4).

In turn, there is growing evidence of nonhomologous genes encoding specific biosynthetic enzymes involved in the same secondary metabolism pathway evolving as clusters occupying neighboring regions of plant genomes, as previously observed in bacteria and fungi (Rokas et al., 2018). Plant secondary metabolism genes arranged in clusters would eventually be subjected to similar transcriptional and epigenetic regulation mechanisms (Boutanaev et al., 2015; Nutzman et al., 2016; Zhang et al., 2006). Such secondary metabo-

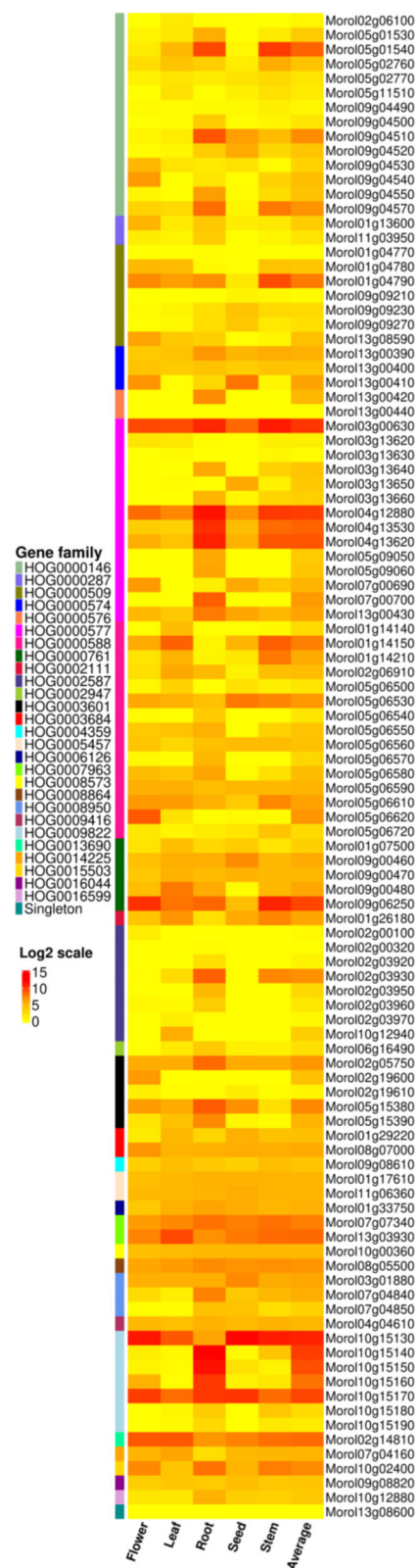


FIGURE 4 Heat map representation of the expression patterns of 104 putative glucosinolate-related moringa genes across five tissues plus average expression. The colors of the heatmap represent \log_2 -transformed expression values measured in transcripts per million. Colored bands on the left indicate the gene family and orthogroup to which each gene belongs

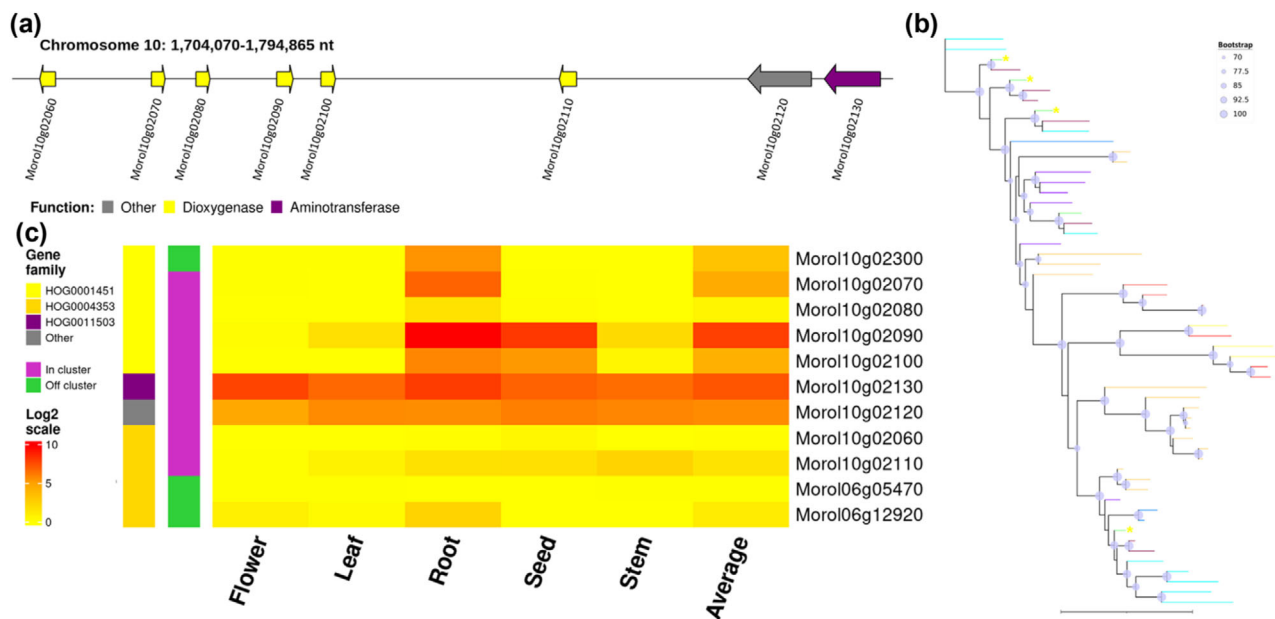


FIGURE 5 Characterization of a secondary metabolism gene cluster one in the moringa genome. (a) Genomic organization of a secondary metabolism-related gene cluster in moringa formed by six dioxygenase, one aminotransferase, and one encoding other enzymatic activity. (b) Maximum likelihood phylogenetic tree of the HOG0001451 dioxygenase gene family. Statistical support values for clades resulting from ultrafast bootstrap analysis are shown next to the corresponding nodes in the form of purple circles with the diameter proportional to the resulting values. Only values >70 are shown. Branches are colored according to the species color scheme from the tree in Figure 2a. Branches in the trees are proportional to evolutionary time, with lengths reflecting the number of amino acid changes. moringa genes included in the cluster are indicated with colored asterisks. (c) Heat map representation of the expression patterns in different tissues plus the average expression of the genes included in the biosynthesis-related gene cluster and paralogous genes identified in the orthogroup classification. The colors of the heatmap represent log₂-transformed expression values measured in transcripts per million. Colored bands on the left indicate the gene family to which gene belongs. Genes included in the cluster (in cluster) or elsewhere in the genome (off cluster) are also indicated

lite gene clusters (SMGCs) in plants have been identified and experimentally validated for a diverse group of secondary metabolism compounds including isoprenoids (Chae et al., 2014), polyketides (Schneider et al., 2016), or modified fatty acids (Jeon et al., 2020). Here, we used PLANTISMASH v1.0 to identify potential SMGC in the moringa genome (Kautsar et al., 2017). PLANTISMASH identified 18 putative SMGCs related to various plant secondary metabolic pathways (Supplemental Table S11), which included two alkaloid, two lignan-polyketide, nine saccharide, three terpene, and two putative gene clusters. The genomic regions of these biosynthetic gene clusters spanned from 28.9 to 339.26 Kb and contained between six and 19 genes. The SMGCs appeared to be evenly distributed across the moringa genome, with every chromosome accommodating at least one SMGC, except for chromosomes 1 and 14 (Supplemental Table S11).

We first focused our attention on SMGC one, which contained eight moringa genes located in chromosome 10. Six out of them were annotated as dioxygenases, four of which belonging to orthogroup HOG0001451 (Figure 5a). Orthogroup HOG0001451 included nine Arabidopsis orthologs, five of which formed a well-supported clade in a ML phylogenetic tree depicting the evolution-

ary relationships among the 56 sequences belonging to the orthogroup (Figure 5b). This clade was embedded within a larger clade containing two papaya genes and one moringa gene (*Moro10g02100*). Interestingly, three out of these five Arabidopsis genes (*AT4G03050*, *AT4G03060*, and *AT4G03070*) had been reported as 2-oxoglutarate-dependent dioxygenase involved in glucosinolate biosynthesis (Kliebenstein et al., 2001), which were not detected in our survey of glucosinolate-related orthogroups. *Moro10g02100* showed expression in roots and, to lesser extent, seeds (Figure 5c).

Next, we examined SMGC five (Supplemental Figure S9A), annotated as of lignan-polyketide type, which included one gene annotated as ketosynthase belonging to orthogroup HOG0003249, which, in turn, included four additional genes tandemly arranged in neighboring positions of the chromosome plus a sixth gene located elsewhere in the genome (Supplemental Figure S9A). The three Arabidopsis orthologs belonging to orthogroup HOG0003249 (*AT5G04530.1*, *AT2G28630.1*, and *AT1G07720.2*) had been reported as members of the 3-ketoacyl-CoA synthase family involved in the biosynthesis of VLCFA (very-long-chain fatty acids) as precursors of wax compounds, participating

in the limitation of nonstomatal water loss during drought adaptation and in the prevention of pathogen attacks (Li-Beisson et al., 2013). With six genes in HOG0003249 for two to five in the remaining 10 plant genomes compared in this study, moringa was found as slightly expanded in this family. Five out of the six genes, including the one found in the SMGC, grouped in a ML phylogenetic tree within a well-supported clade together with counterparts from papaya (Supplemental Figure S9B). Most members of the family in moringa showed low expression, except for *Morol13g04770*, which was strongly expressed in leaves (Supplemental Figure S9C).

Finally, saccharide-biosynthesis-related cluster 10 was formed by 15 genes, including five annotated as UDP-glycosyltransferases, one as methyltransferase, another one as cytochrome P450, another one as dioxygenase, while the rest annotated as encoding for unrelated functions (Supplemental Figure S10A). The five UDP-glycosyltransferases grouped together with 10 additional sequences into orthogroup HOG0000748, including two also detected in the cluster with significantly shorter sequences and annotated as encoding for unrelated functions. Orthogroup HOG0000748 also included 13 and 11 genes in the Brassicales Arabidopsis and papaya, respectively. The ML phylogenetic analysis based on the multiple alignment of the 117 aminoacidic sequences included in HOG0000748 grouped five moringa genes into one well-supported clade, including three of the ones found in the SMGC, sister to a clade formed by seven papaya genes (Supplemental Figure S10B). The resulting clade in turn showed a sister relationship with a clade composed by seven Arabidopsis genes (Supplemental Figure S10B). The remaining two genes in the SMGC were located in disparate positions of the tree. These genes were annotated as encoding for UDP-glucosyltransferases and, specifically, anthocyanidin 3-O-glucosyltransferases (EC:2.4.1.115). Orthologs in Arabidopsis have been reported to be involved in the lycosylation of specific flavonols, but also specific phenylpropanoids, steroids or citoquinines such as zeatin, commonly in response to specific stresses. In general, moringa genes found in the cluster showed, similarly low expression patterns, with the exception of *Morol04g07740*, *Morol04g07730*, and *Morol04g07710*, which showed moderate expression levels (Supplemental Figure S10C).

4 | DISCUSSION

We have presented here a novel assembly of the moringa genome based on Oxford Nanopore long reads, combined with Hi-C. The quality of this assembly is significantly better than the short-read genomes previously published by the AOCC (Chang et al., 2019) and elsewhere (Tian et al., 2015)

both in terms of contiguity and completeness, allowing a more accurate examination of relevant genomic features including some that might be at the origin of interesting phenotypic and agronomic traits of this tree crop plant. Furthermore, it improves on and complements recent long-read genome published by Shyamli et al. (2021) by associating 66 of their contigs into 14 pseudochromosomes in our assembly (Table 1; Supplemental Figure S11). In this figure, we show high congruency between the two genomes.

Given the importance attributed to WGD in promoting phenotypic diversity and speciation (Leitch & Leitch, 2008; Soltis & Soltis, 2009; Van de Peer et al., 2009), in helping to overcome stressful environments and periods of environmental turmoil (Van de Peer et al., 2021; Van de Peer et al., 2017), and in driving the evolution of favorable domestication traits in plants (Salman-Minkov et al., 2016), we leveraged the high-quality assembly of the moringa genome presented here to uncover its duplication evolutionary history. A combination of comparative analyses of moringa and other species based on modelling Ks distributions of syntenic paralogs (anchors) and orthologs together with synteny analysis with selected plant genomes reveals that moringa shares the ancient hexaploidy of all core eudicots, that is, the so-called gamma event (Jiao et al., 2012). Furthermore, and in contrast to their Brassicales counterparts from the Brassicaceae family (Blanc et al., 2003; Bowers et al., 2003), no evidence of additional lineage specific WGD was found in the moringa genome similar to what was reported for papaya (Ming et al., 2008) belonging to the Caricaceae family, the clade sister to the Moringaceae (Olson, 2003).

Moringa gene duplicates resulting from the ancestral eudicot gamma WGD event were functionally characterized using their associated GO terms, revealing an enrichment in transcriptional and protein modification such as transcription factors or protein phosphorylation and dimerization activities. These functions are commonly considered to be dosage balance sensitive (Papp et al., 2003; Tasdighian et al., 2017) and, therefore, are expected to be preferentially retained after WGD in agreement with the dosage balance hypothesis (Freeling, 2009). In turn, SSD duplicates, especially tandem and, to a lower extent, dispersed duplicates, were found to be enriched for specific secondary metabolism enzymes. Therefore, the reciprocal retention pattern of WGD vs. SSD duplicates for functional classes anticipated by the dosage balance hypothesis (Freeling, 2009) can also be verified in moringa.

Apart from secondary metabolism enzymes, tandem duplicates were enriched in several GO terms associated to defense response of plants against biotic and abiotic environmental cues. Indeed, this could help to explain the retention of tandem duplicates in plant genomes, despite that, they are expected to upset dosage balance immediately after duplication—at least when part of multiprotein complexes or intricate gene

regulatory networks—and result in fitness defects. One of the mechanisms proposed to explain the long-term persistence of tandem duplicates involved in secondary metabolism is rapid neofunctionalization resulting in differential substrate specificity of the gene copies (Tohge & Fernie, 2020). Such a mechanism has been experimentally validated for different *Arabidopsis* enzymes involved in the glucosinolate pathway (Kliebenstein et al., 2001; Petersen et al., 2019), for which we have determined the full complement of moringa genes and gene families likely involved in their biosynthesis and regulation. We found at least 11 glucosinolate-related orthogroups including tandemly arranged genes in moringa. Two notable examples are formed by orthogroup HOG000577, annotated as myrosinases, and orthogroup HOG0009822 annotated as encoding for GDSL-type esterases/lipases, both of which were found to be specifically expanded in moringa and in agreement with what had been previously observed (Ojeda-Lopez et al., 2020). Glucosinolates are relatively stable in nature and have no biological activity. When tissue is damaged because of a biotic attack, specific hydrolytic enzymes called myrosinases are activated, leading to the formation of different degradation products such as isothiocyanates, nitriles, oxazolidinethiones, thiocyanate, epithionitriles, and other products, which all exhibit a wide range of biological activity and strongly influence the taste and flavor of the plant (Z. Liu et al., 2021). Furthermore, one of the *Arabidopsis* orthologs in HOG0009822 (*AT3G14210.1*) has been reported as *EPITHIOSPECIFIER MODIFIER1 (ESM1)*, which represses nitrile formation and favors isothiocyanate production during glucosinolate hydrolysis (Zhang et al., 2006). In turn, orthogroup HOG000588 included *Arabidopsis CYP81F2* gene (*AT5G57220.1*), the gene underlying the metabolic quantitative trait loci *Indole Glucosinolate Modifier1 (IGMI)*, which alters the structure of Trp-derived indole glucosinolates (Pfalz et al., 2009). Interestingly, moringa genes belonging to orthogroups HOG000577, HOG0009822, and HOG000588 showed diversified expression patterns across all five tissues examined, with some members featured by significant expression levels in seeds and leaves. It is tempting to speculate rapid functional specialization after tandem duplication within these two gene families could have contributed to the diversity of glucosinolates across tissues between wild-type and domesticated accessions of moringa and among other Moringaceae species (Chodur et al., 2018; Fahey et al., 2018).

Additional enzymes involved in specific secondary metabolism pathways, notably fatty acids, flavonoids, terpenoids, and alkaloids, were found arranged in clusters of homologous genes interspersed with nonhomologous genes possibly involved in the same biochemical pathway. Some of these secondary metabolites are related to plant responses and adaptations to different environmental stresses including water deficit and UVB radiation (Bandurska

et al., 2013). The enrichment in secondary metabolism and defense responses observed among tandem duplicates supports their involvement in rapid adaptations to local environmental stimuli (Hanada et al., 2008) and may be related to the high phenotypic plasticity and adaptability of this species to different environmental constraints, notably water stress (Brunetti et al., 2020; Brunetti et al., 2018) or UVB (Araujo et al., 2016).

Tandem duplications, together with proximal duplications, were also found to be the preferential mode of duplication leading to the expansion of 148 moringa-specific gene families. Interestingly, some of these gene families were also found to be enriched for stress defense responses and developmental processes. These included orthogroup HOG0020145, comprising three moringa genes encoding for regulatory components of the ubiquitin/26S proteasome including ATPase domains (EC:5.6.1.5) specifically involved in channel gating and polypeptide unfolding before proteolysis. This might be indicating the ubiquitin/26S proteasome system in moringa is contributing to the proteomic plasticity required to link plant growth and development with adaptation to environmental stress such as drought, heat, and UV stress (Xu & Xue, 2019).

In summary, the current version and annotation of the moringa genome presented here will facilitate the identification of genes at the origin of biological, agronomical, nutritional, or pharmacological properties in this species and will greatly assist the development of genomics-assisted plant improvement programs especially related to secondary metabolism traits of interest.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during the current study (including genome sequence, annotation, and orthogroup classifications files) are available at the Online Resource for Community Annotation of Eukaryotes (ORCAE) database (https://bioinformatics.psb.ugent.be/gdb/aocc/morolbgi/Moringa_version2/). All data generated or analyzed during this study are included in this published article (and its supplemental information files).

ACKNOWLEDGMENTS

The authors would like to thank Professors Manuel Torres Gil and José Antonio Martínez García from the Department of Informatics at the University of Almería for kindly providing us with access to Cloud-DI-UAL servers. This work was supported in part by a “Proyectos I+D Generación de Conocimiento” grant from the Spanish Ministry of Science and Innovation (grant code: PID2020-113277GB-I00) to LCP and by funds received by the “Sistema de Información Científica de Andalucía” Research Group id BIO359. Funds for sequencing, assembly, and annotation were from the UC Davis Seed Biotechnology Center on behalf of the African Orphan Crops Consortium and by a grant to Carrie

Waterman by NIH. Scaffolding was done in-kind by Dovetail Genomics (Scotts Valley, USA). YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

AUTHOR CONTRIBUTIONS

Jiyang Chang: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-review & editing. Juan Pablo Marczuk-Rojas: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-review & editing. Carrie Waterman: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-review & editing. Armando Garcia-Llanos: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-review & editing. Shiyu Chen: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-review & editing. Xiao Ma: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization. Amanda Hulse-Kemp: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization. Allen Van Deynze: Conceptualization; Funding acquisition; Project administration; Supervision; Validation; Writing-review & editing. Yves Van de Peer: Conceptualization; Funding acquisition; Project administration; Supervision; Validation; Writing-review & editing. Lorenzo Carretero-Paulet: Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Software; Supervision; Validation; Visualization; Writing – original draft; Writing-review & editing.


CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Carrie Waterman  <https://orcid.org/0000-0003-3986-7034>

Shiyu Chen  <https://orcid.org/0000-0003-0102-8112>

Amanda Hulse-Kemp  <https://orcid.org/0000-0001-9670-9433>

Allen Van Deynze  <https://orcid.org/0000-0002-2093-0577>

Yves Van de Peer  <https://orcid.org/0000-0003-4327-3730>

Lorenzo Carretero-Paulet  <https://orcid.org/0000-0001-6697-827X>

REFERENCES

- Abd Rani, N. Z., Husain, K., & Kumolosasi, E. (2018). Moringa genus: A review of phytochemistry and pharmacology. *Frontiers in Pharmacology*, 9, 108. <https://doi.org/10.3389/fphar.2018.00108>
- Albert, V. A., Bradley Barbazuk, W., de Pamphilis, 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., Ammiraju, J. S. S., Chamala, S., Chanderbali, A. S., Determann, R., & Tomsho, L., *Amborella* Genome Project. (2013). The *Amborella* genome and the evolution of flowering plants. *Science*, 342, 1241089. <https://doi.org/10.1126/science.1241089>
- Araujo, 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>
- Bennett, R. N., Mellon, F. A., Foidl, N., Pratt, J. H., Dupont, M. S., Perkins, L., & Kroon, P. A. (2003). Profiling glucosinolates and phenolics in vegetative and reproductive tissues of the multi-purpose trees *Moringa oleifera* L. (horseradish tree) and *Moringa stenopetala* L. *Journal of Agricultural and Food Chemistry*, 51, 3546–3553. <https://doi.org/10.1021/jf0211480>
- Blanc, G., Hokamp, K., & Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research*, 13, 137–144. <https://doi.org/10.1101/gr.751803>
- Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J., & Osbourn, A. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proceedings of the National Academy of Sciences*, 112, E81–E88. <https://doi.org/10.1073/pnas.1419547112>
- Bowers, J. E., Chapman, B. A., Rong, J., & Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422, 433–438. <https://doi.org/10.1038/nature01521>
- Brûna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3, lqaa108. <https://doi.org/10.1093/nargab/lqaa108>
- 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, coaa028. <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*, 38, 1640–1654. <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, 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, 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, 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, giy152. <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, 7995. <https://doi.org/10.1038/s41598-018-26059-3>
- Conesa, A., & Gotz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 2008, 619832. <https://doi.org/10.1155/2008/619832>
- Dangi, S. Y., Jolly, C. I., & Narayanan, S. (2002). Antihypertensive activity of the total alkaloids from the leaves of *Moringa oleifera*. *Pharmaceutical Biology*, 40, 144–148. <https://doi.org/10.1076/phbi.40.2.144.5847>
- 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., Cruzillat, D., Da Silva, C., ... Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, 345, 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, 1805951. <https://doi.org/10.1080/23311932.2020.1805951>
- Dinkova-Kostova, A. T., & Kostov, R. V. (2012). Glucosinolates and isothiocyanates in health and disease. *Trends in Molecular Medicine*, 18, 337–347. <https://doi.org/10.1016/j.molmed.2012.04.003>
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., & Aiden, A. P. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356, 6333. <https://doi.org/10.1126/science.aal3327>
- Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, 3, 95–98. <https://doi.org/10.1016/j.cels.2016.07.002>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 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.). *Science Reports*, 8, 7994. <https://doi.org/10.1038/s41598-018-26058-4>
- Flynn, J. M., Hubble, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117, 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- 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, 433–453. <https://doi.org/10.1146/annurev.arplant.043008.092122>
- The French–Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449, 463467. <https://doi.org/10.1038/nature06148>
- Gandji, K., Chadare, F., Idohou, R., Salako, V., Assogbadjo, A., & Glèlè Kakaï, R. (2018). Status and utilisation of *Moringa oleifera* Lam: A review. *African Crop Science Journal*, 26, 20. <https://doi.org/10.4314/acsj.v26i1.10>
- 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, 221–224. <https://doi.org/10.1093/molbev/msp259>
- Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., & White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31, 5654–5666. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14500829
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8, 1494–1512. <https://doi.org/10.1038/nprot.2013.084>
- Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, 9, R7. <https://doi.org/10.1186/gb-2008-9-1-r7>
- 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, 993–1003. <https://doi.org/10.1104/pp.108.122457>
- Islam, Z., Islam, S. M. R., Hossen, F., Mahtab-Ul-Islam, K., Hasan, M. R., & Karim, R. (2021). *Moringa oleifera* is a prominent source of nutrients with potential health benefits. *International Journal of Food Science*, 2021, 6627265. <https://doi.org/10.1155/2021/6627265>
- Jaafaru, M. S., Nordin, N., Rosli, R., Shaari, K., Bako, H. u. Y., Saad, N., Noor, N. M., & Razis, A., & A, F. (2019). Neuroprotective effects of glucomoringin-isothiocyanate against H₂O₂-induced cytotoxicity in neuroblastoma (SH-SY5Y) cells. *Neurotoxicology*, 75, 89–104. <https://doi.org/10.1016/j.neuro.2019.09.008>
- Jamnadas, R., Mumm, R. H., Hale, I., Hendre, P., Muchugi, A., Dawson, I. K., Powell, W., Graudal, L., Yana-Shapiro, H., Simons, A. J., & Van Deynze, A. (2020). Enhancing African orphan crops with genomics. *Nature Genetics*, 52, 356–360. <https://doi.org/10.1038/s41588-020-0601-x>
- Jeon, J. E., Kim, J. G., Fischer, C. R., Mehta, N., Dufour-Schroif, C., Wemmer, K., Mudgett, M. B., & Sattely, E. (2020). A pathogen-responsive gene cluster for highly modified fatty acids in tomato. *Cell*, 180, 176–187 e119. <https://doi.org/10.1016/j.cell.2019.11.037>
- Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., Rolf, M., Ruzicka, D. R., Wafula, E., Wickett, N. J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter, E. J., Deyholos, M. K., Kutchan, T. M., Chanderbali, A. S., Soltis, P. S., ... Depamphilis, C. W. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biology*, 13, R3. <https://doi.org/10.1186/gb-2012-13-1-r3>

- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, *8*, 275–282. <https://www.ncbi.nlm.nih.gov/pubmed/1633570>
- 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*, 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*, 726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*, 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*. *Plant Cell*, *13*, 681–693. <https://doi.org/10.1105/tpc.13.3.681>
- Kou, X., Li, B., Olayanju, J. B., Drake, J. M., & Chen, N. (2018). Nutraceutical or pharmacological potential of *Moringa oleifera* Lam. *Nutrients*, *10*, 343. <https://doi.org/10.3390/nu10030343>
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*, 1812–1819. <https://doi.org/10.1093/molbev/msx116>
- Kumar, Y., Thakur, T. K., Sahu, M. L., & Thakur, A. (2017). A multi-functional wonder tree: *Moringa oleifera* Lam open new dimensions in field of agroforestry in India. *International Journal of Current Microbiology and Applied Sciences*, *6*, 229–235.
- Lai, C. P., Huang, L. M., Chen, L. O., Chan, M. T., & Shaw, J. F. (2017). Genome-wide analysis of GDSL-type esterases/lipases in *Arabidopsis*. *Plant Molecular Biology*, *95*, 181–197. <https://doi.org/10.1007/s11103-017-0648-y>
- Leitch, A. R., & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science*, *320*, 481–483. <https://doi.org/10.1126/science.1153585>
- 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*, 44. <https://doi.org/10.3390/ijms160612791>
- Li-Beisson, Y., Shorosh, 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. *Arabidopsis Book*, *11*, e0161. <https://doi.org/10.1199/tab.0161>
- Liu, Y., Wang, B., Shu, S., Li, Z., Song, C., Liu, D., Niu, Y., Liu, J., Zhang, J., Liu, H., Hu, Z., Huang, B., Liu, X., Liu, W., Jiang, L., Alami, M. M., Zhou, Y., Ma, Y., He, X., ... Nie, J. (2021). Analysis of the *Coptis chinensis* genome reveals the diversification of protoberberine-type alkaloids. *Nature Communication*, *12*, 3276. <https://doi.org/10.1038/s41467-021-23611-0>
- Liu, Z., Wang, H., Xie, J., Lv, J., Zhang, G., Hu, L., Luo, S., Li, L., & Yu, J. (2021). The roles of Cruciferae glucosinolates in disease and pest resistance. *Plants*, *10*, 1097. <https://doi.org/10.3390/plants10061097>
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964. <https://doi.org/10.1093/nar/25.5.955>
- 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*, 5454–5459. <https://doi.org/10.1073/pnas.0501102102>
- Manni, M., Berkeley, M. R., Seppey, M., Simao, 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*, 4647–4654. <https://doi.org/10.1093/molbev/msab199>
- Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*, e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
- Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., ... Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, *452*, 991–996. <https://doi.org/10.1038/nature06856>
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*, 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>
- Nutzmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters—From genetics to genomics. *New Phytologist*, *211*, 771–789. <https://doi.org/10.1111/nph.13981>
- Ojeda-Lopez, J., Marczuk-Rojas, J. P., Polushkina, O. A., Purucker, D., Salinas, M., & Carretero-Paulet, L. (2020). Evolutionary analysis of the *Moringa oleifera* genome reveals a recent burst of plastid to nucleus gene duplications. *Science Reports*, *10*, 17646. <https://doi.org/10.1038/s41598-020-73937-w>
- Olson, M. E. (2002). Combining data from DNA sequences and morphology for a phylogeny of Moringaceae (Brassicales). *Systematic Botany*, *27*, 55–73. <http://www.jstor.org/stable/3093895>
- Olson, M. E. (2003). Ontogenetic origins of floral bilateral symmetry in Moringaceae (Brassicales). *American Journal of Botany*, *90*, 49–71. <https://doi.org/10.3732/ajb.90.1.49>
- Olson, M. E. (2017). *Moringa* frequently asked questions. *Acta Horticulturae*, *1158*, 19–32. <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*, e0159782. <https://doi.org/10.1371/journal.pone.0159782>
- Panda, S., Kar, A., Sharma, P., & Sharma, A. (2013). Cardioprotective potential of N,α-L-rhamnopyranosyl vincosamide, an indole alkaloid, isolated from the leaves of *Moringa oleifera* in isoproterenol induced cardiotoxic rats: In vivo and in vitro studies. *Bioorganic & Medicinal Chemistry Letters*, *23*, 959–962. <https://doi.org/10.1016/j.bmcl.2012.12.060>

- Papp, B., Pal, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, *424*, 194–197. <https://doi.org/10.1038/nature01771>
- Pasha, S. N., Shafi, K. M., Joshi, A. G., Meenakshi, I., Harini, K., Mahita, J., Sajeevan, R. S., Karpe, S. D., Ghosh, P., Nitish, S., Gandhimathi, A., Mathew, O. K., Prasanna, S. H., Malini, M., Mutt, E., Naika, M., Ravooru, N., Rao, R. M., Shingate, P. N., ... Sowdhamini, R. (2020). The transcriptome enables the identification of candidate genes behind medicinal value of Drumstick tree (*Moringa oleifera*). *Genomics*, *112*, 621–628. <https://doi.org/10.1016/j.ygeno.2019.04.014>
- 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*, <https://doi.org/10.1042/BSR20190446>
- Pfalz, M., Vogel, H., & Kroymann, J. (2009). The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in *Arabidopsis*. *Plant Cell*, *21*, 985–999. <https://doi.org/10.1105/tpc.108.063115>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, *26*, 1641–1650. <https://doi.org/10.1093/molbev/msp077>
- Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., & Vandepoele, K. (2012). i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research*, *40*, e11. <https://doi.org/10.1093/nar/gkr955>
- Rokas, A., Wisecaver, J. H., & Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology*, *16*, 731–744. <https://doi.org/10.1038/s41579-018-0075-3>
- Sahakitpichan, P., Mahidol, C., Disadee, W., Ruchirawat, S., & Kanchanapoom, T. (2011). Unusual glycosides of pyrrole alkaloid and 4'-hydroxyphenylethanamide from leaves of *Moringa oleifera*. *Phytochemistry*, *72*, 791–795. <https://doi.org/10.1016/j.phytochem.2011.02.021>
- Salman-Minkov, A., Sabath, N., & Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, *2*, 16115. <https://doi.org/10.1038/nplants.2016.115>
- Schneider, L. M., Adamski, N. M., Christensen, C. E., Stuart, D. B., Vautrin, S., Hansson, M., Uauy, C., & von Wettstein-Knowles, P. (2016). The *Cer-cqu* gene cluster determines three key players in a beta-diketone synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *Journal of Experimental Botany*, *67*, 2715–2730. <https://doi.org/10.1093/jxb/erw105>
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., ... Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*, 1044–1053. <https://doi.org/10.1038/s41587-020-0503-6>
- 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*, 766999. <https://doi.org/10.3389/fpls.2021.766999>
- Siddhuraju, P., & Becker, K. (2003). Antioxidant properties of various solvent extracts of total phenolic constituents from three different agroclimatic origins of drumstick tree (*Moringa oleifera* Lam.) leaves. *Journal of Agricultural and Food Chemistry*, *51*, 2144–2155. <https://doi.org/10.1021/jf020444+>
- Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. <https://doi.org/10.1186/1471-2105-6-31>
- Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, *60*, 561–588. <https://doi.org/10.1146/annurev.arplant.043008.092039>
- Supek, F., Bosnjak, M., Skunca, N., & Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, *6*, e21800. <https://doi.org/10.1371/journal.pone.0021800>
- Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, *320*, 486–488. <https://doi.org/10.1126/science.1153917>
- 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. *Plant Cell*, *29*, 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*, 627–638. <https://doi.org/10.1007/s11427-015-4872-x>
- Tohge, T., & Fernie, A. R. (2020). Co-regulation of clustered and neofunctionalized genes in plant-specialized metabolism. *Plants*, *9*, 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*, W232–W235. <https://doi.org/10.1093/nar/gkw256>
- Trigo, C., Castello, M. L., Ortola, M. D., Garcia-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*, 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>
- Van de Peer, Y., Ashman, T. L., Soltis, P. S., & Soltis, D. E. (2021). Polyploidy: An evolutionary and ecological force in stressful times. *Plant Cell*, *33*, 11–26. <https://doi.org/10.1093/plcell/koaa015>
- Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*, 725–732. <https://doi.org/10.1038/nrg2600>
- Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*, 411–424. <https://doi.org/10.1038/nrg.2017.26>
- Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*, 737–746. <https://doi.org/10.1101/gr.214270.116>
- Vudhgiri, S., Prasad, R. B. N., Kota, A., Poornachandra, Y., Kumar, C. G., R, S., & RC, R. J. (2016). Synthesis and biological evaluation of Marumoxide A isolated from *Moringa oleifera* and its lipid derivatives. *International Journal of Pharmaceutical Sciences and Research*, *7*, 607–617. [https://doi.org/10.13040/IJPSR.0975-8232.7\(2\).607-17](https://doi.org/10.13040/IJPSR.0975-8232.7(2).607-17)

- 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*, e49. <https://doi.org/10.1093/nar/gkr1293>
- Xu, F. Q., & Xue, H. W. (2019). The ubiquitin-proteasome system in plant responses to environments. *Plant, Cell and Environment*, *42*, 2931–2944. <https://doi.org/10.1111/pce.13633>
- Yan, H., Bombarely, A., & Li, S. (2020). DeepTE: A computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics*, *36*, 4269–4275. <https://doi.org/10.1093/bioinformatics/btaa519>
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Zhang, Z., Ober, J. A., & Kliebenstein, D. J. (2006). The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in *Arabidopsis*. *Plant Cell*, *18*, 1524–1536. <https://doi.org/10.1105/tpc.105.039602>
- Zwaenepoel, A., & Van de Peer, Y. (2019). wgd—simple command line tools for the analysis of ancient whole-genome duplications.

Bioinformatics, *35*, 2153–2155. <https://doi.org/10.1093/bioinformatics/bty915>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chang, J., Marczuk-Rojas, J. P., Waterman, C., Garcia-Llanos, A., Chen, S., Ma, X., Hulse-Kemp, A., Van Deynze, A., Van de Peer, Y., & Carretero-Paulet, L. (2022). Chromosome-scale assembly of the *Moringa oleifera* Lam. genome uncovers polyploid history and evolution of secondary metabolism pathways through tandem duplication. *The Plant Genome*, *15*, e20238. <https://doi.org/10.1002/tpg2.20238>