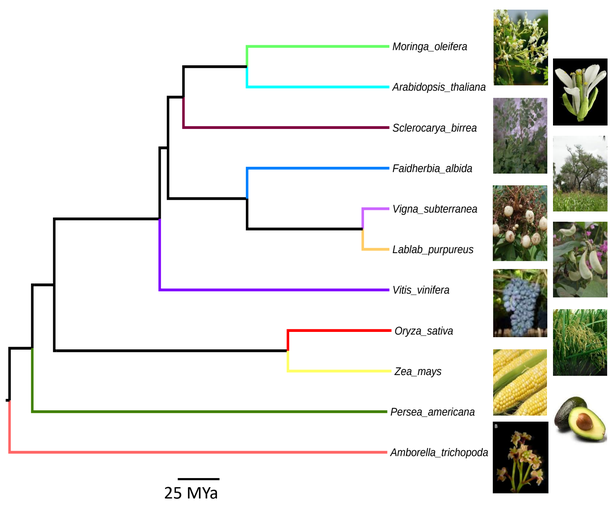
Evolutionary analysis of the *Moringa oleifera* genome reveals a recent burst of plastid to nuclear gene duplications

Juan Pablo Marczuk-Rojas, José Ojeda-López, Oliver Aleksandrei Polushkina, Darius Purucker, María Salinas-Navarro*, Lorenzo Carretero-Paulet*
E-mails : jmr386@inlumine.ual.es (JP M-R, Presenting author); msalinas@ual.es, lpaulet@ual.es (M S-N & L C-P*, Corresponding author).
Address: Area of Genetics. Department of Biology and Geology. University of Almería. Ctra. Sacramento s/n 04120 Almería. Spain.

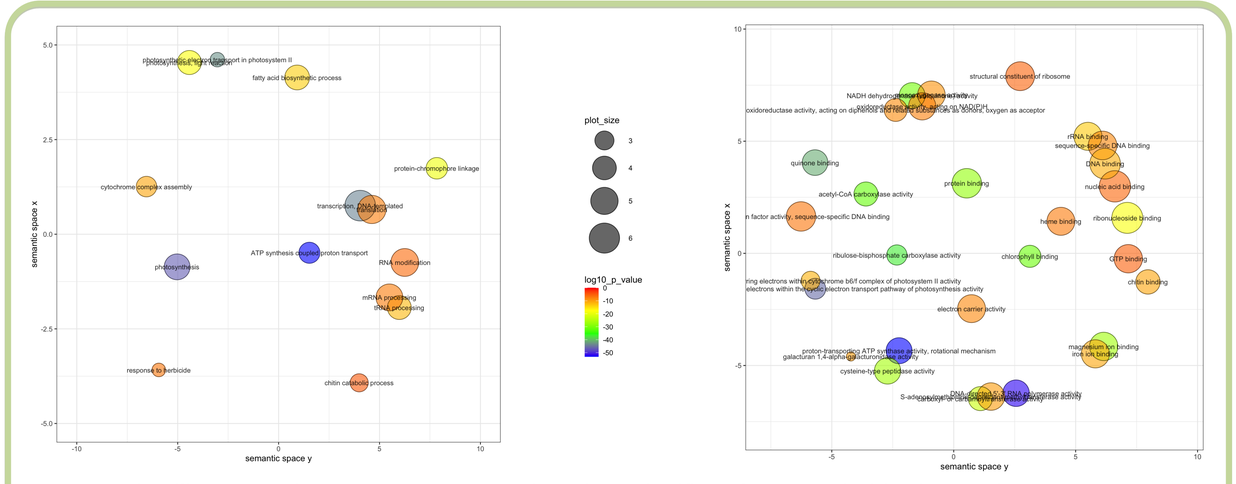
It is necessary to identify suitable alternative crops to ensure the nutritional demands of a growing global population. The genome of *Moringa oleifera*, a fast-growing drought-tolerant orphan crop with highly valuable agronomical, nutritional and pharmaceutical properties, has recently been reported. We model here gene family evolution in Moringa as compared with ten other flowering plant species. Despite the reduced number of genes in the compact Moringa genome, 101 gene families, grouping 957 genes, were found as significantly expanded. Expanded families were highly enriched for chloroplastidic and photosynthetic functions. Indeed, almost half of the genes belonging to Moringa expanded families grouped with their Arabidopsis thaliana plastid encoded orthologs. Microsynteny analysis together with modeling the distribution of synonymous substitutions rates, supported most plastid duplicated genes originated recently through a burst of simultaneous insertions of large regions of plastid DNA into the nuclear genome. These, together with abundant short insertions of plastid DNA, contributed to the occurrence of massive amounts of plastid DNA in the Moringa nuclear genome, representing 4.71%, the largest reported so far. Our bioinformatic study provides key genetic resources for future breeding programs and highlights the potential of plastid DNA to impact the structure and function of nuclear genes and genomes.

Evolutionary tree of Moringa and 10 plants species used in this study



We first obtained a classification of gene families in the Moringa genome and 10 plant species representing the main angiosperm lineages using the software Orthofinder. In order to identify Moringa-specific expanded and contracted gene families, the 11-species tree and gene family classification in 17,998 orthogroups were then used to evaluate the fit of different Maximum Likelihood gain and death stochastic models of gene family evolution implemented in the BadiRate program.

Scatterplot representation of enriched GO terms in Moringa-expanded gene families (Biological Process and Molecular Function)

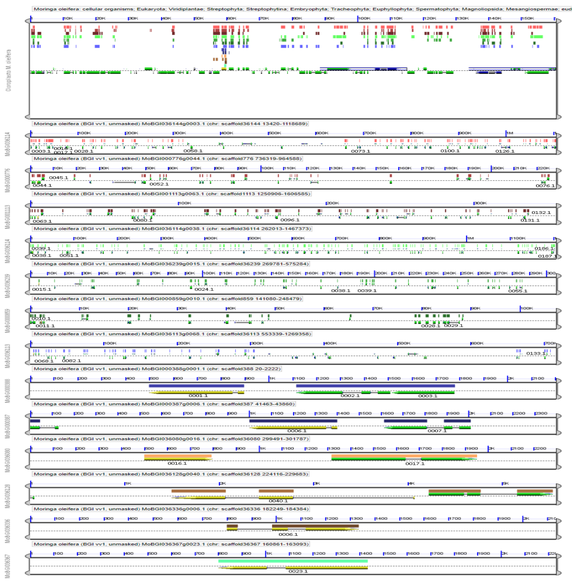


101 gene families, grouping 957 genes, were found as significantly expanded in the Moringa genome. **Most overrepresented GO terms found among Moringa-expanded families corresponded to plastid and, especially, chloroplast associated functions**. Indeed, 27 out of the 101 families (grouping a total of 457 Moringa genes) corresponded to orthogroups that included at least one Arabidopsis orthologous gene encoded by the plastid genome.

ML Phylogenetic Analysis of the Moringa-expanded *RBCL* gene family in 11 plant species and multiple protein sequence alignment

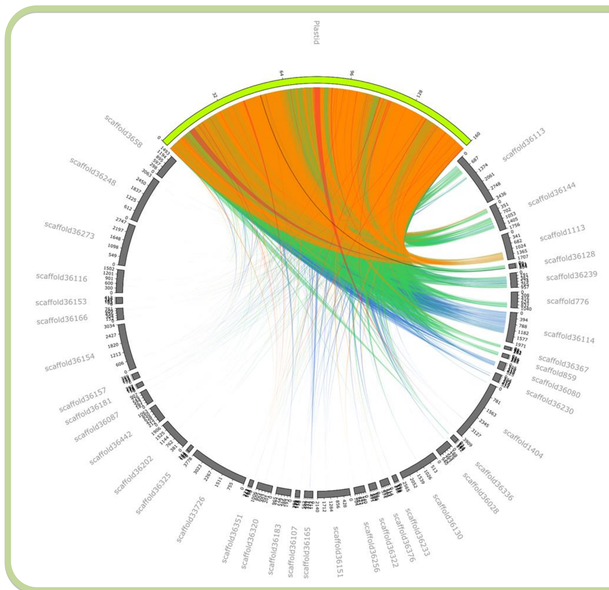
Despite the low phylogenetic signal of the alignment to solve deep evolutionary relationships, **the general picture emerging from the tree reflects recent events of plastid gene duplication at the origin of the observed Moringa-specific expansion.**

Microsynteny analysis of the Moringa chloroplast genome and 13 nuclear genomic regions containing *RBCL* genes



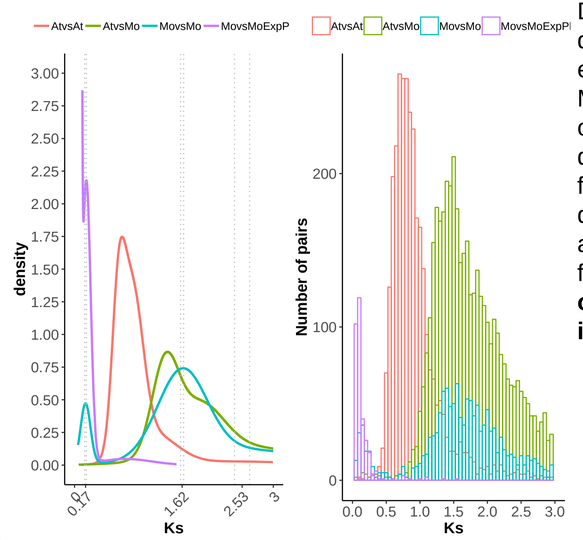
Pairwise genomic comparisons were performed using the Moringa chloroplast genome, located on top, as reference. High-scoring sequence pairs (HSPs) between protein-coding sequences are marked by short coloured vertical bars on top of the corresponding gene models in the Moringa chloroplast genome. Collinear series of HSPs across genomic regions indicates a syntenic relationship between the regions concerned. **The strong syntenic signal shown by most *RBCL* genomic regions reveals that plastid gene duplicates originated through insertions of large regions of the plastid genome.**

Circos plot representation of plastid DNA insertions in the Moringa nuclear genome



Local nucleotide alignments resulting from BLASTN comparisons between the Moringa chloroplast genome (represented in green) used as a query and nuclear genome scaffolds (represented in dark grey) used as sequence database are represented as ribbons. Only the 39 nuclear genome scaffolds with a length equal or higher than that of the plastid genome and returning BLAST alignments between them with at least 90 % of sequence identity over a region of minimum 2000 kb are shown. **Altogether, the fraction of plastid DNA found in the Moringa nuclear genome represented 4.71%, the largest so far reported for a plant genome.**

Modeling genome duplications in Moringa



Density plots from fitting Gaussian mixture models to distributions of *Ks* values (synonymous substitutions) estimated from pairs of syntenic paralogues within the Moringa and Arabidopsis genomes, of syntenic orthologues between both genomes as well as the duplicated genes belonging to the 27 plastid gene families found as expanded by Badirate. The peak in the distribution of Ks values shared by syntenic paralogues and duplicates belonging to plastid expanded gene families reveals that **most plastid duplicated genes originated through a recent burst of insertions involving large regions of the plastid genome.**

African Orphan Crops Consortium AOCC

IX Simposio de investigación en Ciencias Experimentales

ciencias experimentales

UNIVERSIDAD DE ALMERÍA