Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Plastid DNA is a major source of nuclear genome complexity and of RNA genes in the orphan crop moringa

Juan Pablo Marczuk-Rojas
University of Almería
Antonio Salmerón
University of Almería
Alfredo Alcayde
University of Almería
Viktor Isanbaev
University of Almería
Lorenzo Carretero-Paulet
lpaulet@ual.es

University of Almería

Research Article

Keywords: Organellar genomes, genome structure and evolution, isrR genes, moringa, NUPTs, Non-coding RNAs, Small RNAs, rRNA genes, tRNA genes

Posted Date: March 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4011695/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background:

Unlike Transposable Elements (TEs) and gene/genome duplication, the role of the so-called nuclear plastid DNA sequences (NUPTs) in shaping the evolution of genome architecture and function remains poorly studied. We investigate here the functional and evolutionary fate of NUPTs in the orphan crop *Moringa oleifera* (moringa), featured by the highest fraction of plastid DNA found so far in any plant genome, focusing on i) any potential biases in their distribution in relation to specific nuclear genomic features, ii) their contribution to the emergence of new genes and gene regions, and iii) their impact on the expression of target nuclear genes.

Results:

In agreement with their potential mutagenic effect, NUPTs are underrepresented among structural genes, although their transcription levels and broadness were only lower when involving exonic regions; the occurrence of plastid DNA did not generally result in a broader expression, except among those affected in introns by older NUPTs. In contrast, we found a strong enrichment of NUPTs among several classes of RNA genes, especially those involved in the protein biosynthetic machinery (*i.e.*, rRNA and tRNA genes) and specific classes of regulatory RNAs; a significant fraction of these is functionally expressed, thus potentially contributing to the nuclear pool.

Conclusions:

Our results complete our view of the molecular factors driving the evolution of nuclear genome architecture and function, and support plastid DNA in moringa as a major source of i) genome complexity and, ii) the nuclear pool of RNA genes.

Background

Transposable elements (TEs) and gene and genome duplications are considered as the main molecular forces behind the evolution of plant genome architecture and function [1]. TEs are by far the largest and most variable part of plant genomes [2]. Because of their mobile nature and their propensity to leave traces in their wake across the genome in the form of interspersed repeated sequences, they have been traditionally considered mutagenic and referred to as junk or selfish DNA. However, recent advances in genomics and phenomics are shifting our view of TEs as great contributors to genetic variation on which selection can operate, producing a wide variety of changes in plant gene expression and function with potential adaptive roles on plant evolution [3]. New genes and gene structures, in turn, are being continuously added, and lost, by genomes in a lineage-specific manner. Newly acquired or rearranged genes can evolve novel and/or specialized gene product and/or regulatory functions, ultimately determining to a large degree phenotypic differences between organisms, populations, and species. Several molecular mechanisms are known to be involved in the creation of new gene and gene structures,

including exon shuffling and duplication, gene fusion and fission, domestication of TEs, horizontal gene transfer, *de novo* gene origination and, prominently, gene and genome duplications [4–6]. Of these, gene and genome duplication are considered the main source of raw genetic material on which mutation and selection, as well as other evolutionary forces, can act upon, ultimately resulting in new and novel gene and gene functions. As a result, the mechanisms determining their retention in the genomes have received much attention [7–10].

However, other genomic sources with potential roles on the origin of evolutionary innovation and adaption are less studied, including the well-known copy of stretches of plastid DNA of different sizes and their subsequent integration in plant nuclear genomes, giving rise to the so-called nuclear plastid DNA sequences (NUPTs) [11]. Although the origins and the evolutionary paths of insertions of organelle DNA into the nuclear genome are probably diverse, they generally involve double-stranded breaks and DNA damage and thus are potentially mutagenic [12, 13]. Furthermore, the uncontrolled proliferative insertion of plastid DNA might lead to the unnecessary “obesity” of the nuclear genome. As a result, organellar DNA insertions are generally expected to be neutral or eventually deleterious and selected against [12, 13]. Indeed, most recent plastid DNA insertions are expected to diverge, decay, rearrange, fragment, or vanish over evolutionary time, a process that appears to occur rapidly and probably involves mutation, TEs, other DNA sequences unrelated to organelle DNA and replication slippage [12–14].

Plastid genes, in turn, are expected to be inactive upon arrival into the nuclear genome because they often do not encode for complete open reading frames and/or lack the regulatory motifs required for proper gene expression in the nucleus [12, 13]. Furthermore, epigenetic regulation, and prominently DNA methylation and histone tail modification, commonly reported to inhibit the activity of mobile DNA and other types of extraneous DNA, has also been associated with the transcriptional repression of integrated organellar DNA [13, 15–17]. As a consequence, plastid genes typically show low or null expression in the nucleus and will likely evolve as non-functional pseudogenes or non-coding sequences [18, 19]. However, as noted in [20–22], expression of NUPT genes in the nucleus do actually occur in plants, so assumptions of non-functionality for nuclear genes of plastid origin must be taken with caution. Indeed, on a few occasions, newly arrived organellar genes have been reported, i) to gain expression capabilities in the nucleus, ii) to reshape nuclear genes by adding extra coding (exon) or regulatory sequences, or iii) to proliferate as tandemly arrayed clusters or in distant regions of the nuclear genome[13, 15, 16] .

As a result of the usually rapid decay and functional inactivation of NUPTs, in most species organelle DNA represents only a small fraction of less than 0.1% of the nuclear genome, with very few showing more than 1% [13]. However, we recently reported a strong enrichment of NUPTs in the nuclear genome of the orphan crop *Moringa oleifera* Lam. (moringa), representing the largest fraction of plastid DNA so far reported for a plant genome [23]. NUPTs in moringa were formed through two events separated in time, namely I and II, with NUPTs from every event showing markedly distinctive features [24]. While younger NUPTs from episode II showed seemingly random origins throughout the chloroplast genome, a wide range of sizes, preferential location in hotspots, a weak negative correlation between sequence identity and size and, when found in clusters, no collinear arrangement with the plastid genome [24], older NUPTs

from episode I were featured by a narrower distribution of sizes, their origin from a few short regions in the chloroplast genome and a preferential collinear arrangement with their plastid ancestors when found in clusters [24].

Therefore, one question that immediately arises is about the molecular and evolutionary forces that may have operated on specific plant lineages promoting the fixation of massive amounts of organellar DNA in the nuclear genome. Here, we explore the functional and evolutionary fate of NUPTs using a chromosome-scale assembly of the moringa genome [25]. We focus our analysis on i) potential biases in their distribution in relation to different nuclear genomic features, ii) their contribution to the emergence of new genes and gene regions, and iii) their impact on the expression of target nuclear genes. Our results support NUPTs as a major source of nuclear genome complexity and of functionally expressed RNA genes and highlight the usefulness of the moringa genome as a model to study the actual impact of NUPTs on the evolution of genome architecture and function.

Results

Biased distribution of NUPTs across moringa nuclear genomic features

In order to examine the distribution of NUPTs across the different sequence features found in the moringa nuclear genome, we employed the NUPTs detected in [24] and the structural annotation of a chromosome-scale assembly, *i.e.*, AOCCv2 [25] completed by further classifying tRNA genes as plastid, mitochondrial or nuclear. 91,06% of the genome could be categorized into 12 genomic features, including structural genes, Transposable Elements (TEs), other repeats, NUPTs plus nine different categories of RNA genes, while the rest was deemed as “other DNA” (**Table S1**). A total of 4,885 genomic features were hit by 4,754 NUPTs (**Table 1**), which we will refer to as NUPT genomic features. The majority of the NUPT genomic features corresponded to TEs (2,317) and plastid tRNA genes (912) (**Table 1** and **Table S1**). 3,716 NUPTs were found in two or more specific genomic features (**Table 1**). The percentage of genomic features affected by NUPTs varied widely, ranging between 0%, for spliceosomal and other RNA genes and more than 99% for plastid tRNA genes (**Table 1**). Apart from nuclear plastid tDNA, two additional categories of RNA genes for which the majority of members were affected by NUPTs were self-splicing intron RNA genes (96.06%) and prokaryotic rRNA genes (93.21%) (**Table 1**). Within every category of NUPT genomic features, the percentage of them fully aligning with plastid DNA was also highly variable, ranging between 5.61% for structural genes to more than 99%, as was the case for mitochondrial and plastid tRNA, as well as self-splicing intron RNA genes (**Table 1**). Some specific classes of NUPT genomic features seemed to be preferentially affected by NUPTs from one episode or another (**Table S2**).

Table 1: Summary of NUPT genomic features.

NUPT Genomic feature	Total number	Total percentage (%) in the genome	Percentage (%) fully covered by plastid DNA	Number of NUPTs[2]
Structural gene[1]	428	1.88	5.61	657
TE	2,317	1.81	40.40	3,858
Other repeats	80	0.10	90	105
Eukaryotic rRNA gene	132	5.19	27.27	264
Prokaryotic rRNA gene	302	93.21	55.96	832
Nuclear tRNA gene	112	18.33	95.54	119
Mitochondrial tRNA gene	48	81.36	100	48
Plastid tRNA gene	912	99.89	99.56	885
Self-splicing intron RNA gene[3]	512	96.06	98.63	633
Regulatory RNA genes[4]	42	11.2	69.05	62
Spliceosomal RNA gene	0	0	0	0
Other RNA genes[3]	0	0	0	0

[1]Including the 1 Kb regions upstream and downstream of the ATG and stop codons, respectively.

[2]Note that there are 3,716 NUPTs overlapping two or more genomic features and so the total number is higher than the actual number of NUPTs.

[3]Group I and group II introns.

[4]Only iron stress repressed RNA (*isrR*) genes.

We next examined any potential spatial biases in the distribution of NUPTs with respect to specific genomic features, *i.e.*, whether specific genomic features were more or less tolerant to host NUPTs. First, visually, by graphically representing the arrangement of NUPTs from every episode plus every other feature found in the moringa genome along the 14 chromosomes using Circos plots (**Fig. 1A**). The frequency of genomic features was represented in each case as density plots in windows of 500 kb in length. We then searched for any putative overlaps in density peaks between genomic regions corresponding to NUPTs from one episode or another and the rest of genomic features. As expected, older

NUPTs from episode I, which were found to be apparently distributed homogenously across chromosomes [24], do not show any apparent peaks in the distribution (**Fig. 1A**). In turn, younger NUPTs from episode II, which were found to be concentrated in hotspots [24], collocated with peaks in the density distribution of different classes of RNA genes, including self-splicing intron, plastid tRNA and prokaryotic rRNA, which in turn corresponded with chromosome regions with low density of TEs and other repeats (**Fig. 1A**).

Next, we tried to statistically substantiate putative biases in the distribution of NUPTs from one episode or another against every specific genomic feature by comparing observed versus expected base pair overlap counts (**Fig. 1B** and **Table S3**). Based on the results from performing Pearson's Chi-squared independence tests with Yates' continuity correction, prokaryotic rRNA genes, tRNA genes, self-splicing intron RNA genes, regulatory RNA genes and other DNA were highly enriched for NUPTs from one and another episode whereas structural genes, other repeats, spliceosomal RNA genes and other RNA genes were highly impoverished (**Fig. 1B** and **Table S3**). Features such as eukaryotic rRNA genes, TEs, NUPTs from another episode and unclassified NUPTs, showed opposite trends of enrichment depending on the NUPTs formation event (**Fig. 1B** and **Table S3**).

NUPTs show differential patterns of retention between structural and RNA genes

As stated above, structural genes were found to be affected by NUPTs from one and another episode less than expected by chance (**Fig. 1B** and **Table S2**), likely reflecting the potential deleterious effects resulting from the insertion of exogenous DNA, especially when affecting coding sequences. Nevertheless, impoverishment in NUPTs across structural genes could be observed regardless of the region of the gene affected (**Fig. S1** and **Table S4**).

A total of 428 structural genes were hit by 657 NUPTs (**Table 1** and **S5**), including 249 affected in intron regions, 71 in the 1 Kb region upstream of the ATG start codon, considered as the promoter region, 76 inserted in the 1 Kb region downstream of the stop codon, deemed as the terminator region, and only 30 affected in exons, including 12 single-exon genes originating from six individual NUPTs from episode II (**Table S5**). Although many NUPT structural genes showed one single NUPT (222), the remaining displayed a variable number ranging from two to up to 10 NUPTs (one single gene, *Morol14g00250*) (**Table S6)**. Additionally, two NUPT structural genes (*Morol07g01780* and *Morol05g07670*) were affected by more than one NUPT found in both promoter and terminator regions.

We explored the spatial arrangement of NUPT structural genes across the moringa nuclear genome. For this purpose, NUPT structural genes from one episode or another, categorized according to the gene regions affected, were graphically represented across the moringa chromosomes in the form of Circos plots (**Fig. 1C**). 151 structural genes were exclusively affected by NUPTs-I, while 153 were exclusively hit by NUPTs-II; 134 structural genes either contain NUPTs from one and another episode and / or unclassified NUPTs (**Fig. 1C**). In general, NUPT structural genes appear to be scattered across all 14

chromosomes and did not show any apparent arrangement in clusters, the only exception being NUPT structural genes affected in exon sequences (**Fig. 1C**).

In contrast to the underrepresentation of NUPTs among structural genes, the opposite trend could be observed for most categories of RNA genes. This observation was especially significant for organellar tRNA, prokaryotic rRNA and self-splicing intron RNA genes present in the nuclear genome, the majority of which were of plastid origin (**Table 1**). We further mapped NUPT RNA genes onto the region of origin in the plastid genome by using a newly obtained annotation of RNA genes in the moringa plastid genome. Our new annotation of the plastid genome detected additional RNA genes, including five eukaryotic rRNA, one regulatory RNA (corresponding to the iron stress repressed RNA genes, *isrR*) plus 22 self-splicing introns (**Fig. 1D**). Furthermore, on top of the 36 tRNA genes found in the original annotation, an additional selenocysteine tRNA gene was detected. tRNA genes found in the plastid genome were further classified as plastid, mitochondrial and nuclear (**Fig. 1D**). In general, NUPT RNA genes matched RNA genes in the plastid genome belonging to the same category. For example, out of the 913 and 59 genes in the moringa nuclear genome annotated as plastid and mitochondrial tRNA, 912 and 48, respectively, corresponded to tRNA genes identically annotated in the plastid genome (**Fig. 1D**). Out of the 611 nuclear tRNA genes found in the nuclear genome, 112 were of plastid origin, 85 out of which were similarly annotated in the plastid genome, while the rest were annotated as plastid tRNA genes (**Fig. 1D**). A similar situation applied to the 324 genes annotated in the nuclear genome as encoding for prokaryotic rRNA, 302 out of which originated from prokaryotic rRNA genes found in the plastid genome. With respect to the 2,541 eukaryotic rRNA genes found in the nuclear genome, only 132 derived from the plastid genome, where they were identically annotated, except 36 genes encoding for 5S rRNA, which corresponded to two 5S rRNA genes annotated as prokaryotic in the plastid. Another category of RNA genes enriched for NUPTs was that of regulatory RNA genes, 42 out of 375 arising from a single gene found in the plastid annotated as *isrR*. Furthermore, every single of the 512 out of 533 genes annotated as NUPT-self-splicing intron in the nucleus proceeded from a gene region identically annotated in the plastid. Finally, the representation of NUPTs RNA genes across the 14 chromosomes of the moringa nuclear genome revealed their preferential arrangement in clusters, in contrast to that observed for structural genes (**Fig. 1D**).

Functional and expression characterization of NUPT structural genes

We examined whether the presence of NUPTs in structural genes could determine differences in their expression, either qualitatively or quantitatively, with respect to the rest of genes in the genome, using RNA-seq data from five tissues, *i.e.*, flower, leaf, root, seed, and stem [26]. Out of the 428 NUPT structural genes, 380 showed significant expression in at least one of the five tissues, a fraction not significantly different to that found among all structural genes according to a Fisher's exact test (**Fig. 2A**). When partitioned by formation episode, only NUPT-II structural genes featured a fraction of expressed genes significantly smaller than expected by chance (**Fig. 2A**). The fraction of expressed versus unexpressed NUPT structural genes showed deviations from non-NUPT ones depending on the region affected by

NUPTs (**Fig. 2A**). While this fraction was significantly greater in the case of NUPT structural genes affected in introns, the opposite situation was observed for structural genes with NUPTs located in promoter or terminator regions, with no significant differences among those genes affected by NUPTs in exons (**Fig. 2A**).

Moreover, the overall expression of NUPT structural genes was not different than that of non-NUPT ones according to a Wilcoxon's rank test, either when considered together, partitioned by formation episode or by the affected gene region, except among those affected in exons, whose overall expression was significantly lower (**Fig. 2B**).We also checked for differences in expression broadness between NUPT structural genes and non-NUPT ones by using the Tau index, calculated using the RNA-seq data from each of the five tissues sampled. Values of the Tau index range from 0, indicating broader unspecific expression, to 1, reflecting narrower specific expression [27]. NUPT structural genes showed a significantly broader expression across the five tissues with respect to the rest of structural genes in the genome according to a Wilcoxon's rank test (**Fig. 2C**), an effect specifically related to structural genes affected by NUPTs-I in intronic regions (**Fig. 2C**). In contrast, NUPT structural genes affected in exonic regions showed a significantly narrower expression across the five tissues with respect to the rest of structural genes in the genome (**Fig. 2C**).

Next, we attempted to get insights about the potential involvement of the 428 NUPT structural genes in specific biological functions. For this purpose, we first used GO terms. Seven GO functional terms related to chloroplast and photosynthesis molecular functions, biological processes or subcellular locations were found to be significantly overrepresented (**Table S7**). Similar enrichment tests were also performed using EC numbers, representing a hierarchical classification of chemical reactions catalyzed by enzymes, and KEGG KO terms, describing molecular functions represented in terms of functional orthologs. One EC term (EC:1, grouping oxidoreductases), was found to be significantly overrepresented, whereas two other EC terms (EC:3.5.1.98, histone deacetylase; EC:1.10.3.9, photosystem II oxidoreductase) were found to be marginally overrepresented (**Table S8**). One KO term (K02704), describing a chlorophyll apoprotein, was found to be enriched among NUPT structural genes, whereas two other KO terms also describing chlorophyll apoproteins (K02690 and K02705) were found to be marginally enriched (**Table S9**). The enrichment found among NUPT-structural genes for biological functions and enzymatic activities related to chloroplast functions is not surprising considering that most of them corresponded to genes entirely of plastid origin (**Tables S7-10**). Indeed, DeepLoc 2.0 predictions of subcellular localization [28], did not find a significantly different number of NUPT structural genes to be imported to the chloroplast (45) than for non-NUPT structural genes (1,982), according to a Fisher's exact test ($P = 0.23$), as could be expected if NUPT-structural genes had preferentially evolved chloroplast-related functions.

Out of the six NUPT-structural genes not fully covered by plastid DNA, *Morol06g13970,* affected by a single NUPT-II spanning the end of intron fifteen and the beginning of exon sixteen, showed the highest expression across all five tissues (**Fig. 3**). *Morol06g13970* was annotated as encoding for the nuclear TPR3-like protein (**Table S10**), featured by a number of WD40 repeated motifs rich in Asp and Trp residues. WD40 motifs are found in a diverse range of proteins covering a wide variety of plant

developmental related functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis [29]. A second gene, *Morol10g09890*, annotated as encoding for the cytochrome P450 CYP72A219-like protein, was affected by a NUPT-II spanning the end of its last exon and the beginning of its terminator region and was only found to be marginally expressed in roots (**Fig. 3** and **Table S10**). *Morol11g02440*, highly expressed in all five tissues and affected by a single NUPT-I spanning the end of exon two and the beginning of intron two (**Fig. 3**), and *Morol02g15190*, only found to be marginally expressed in seeds and affected by an unclassified NUPT spanning its last exon (**Fig. 3**), had no annotated function (**Table S10**). BLASTP searches for putative homologous proteins in the Uniprot database v2023_04 [30] yielded translation initiation factor IF-1, an essential component for the initiation of protein synthesis, as best hit for *Morol11g02440*. In turn, the best retrieved BLAST hit for *Morol02g15190* was annotated as a polyprenol reductase, a key player in the early steps of protein N-linked glycosylation.

NUPT RNA genes are functionally expressed

In a first attempt to assess whether nuclear RNA genes of plastid origin were functional, we examined their expression using our dataset of RNA-seq from five tissues. As for structural genes above, the number and percentage of expressed genes with respect to the total plus the distribution of their expression levels were represented as stacked bar plots and violin plots, respectively, for every class of RNA genes found as enriched for NUPTs (**Fig. 4)**. The expression of homologous RNA genes present in the plastid genome was also shown for comparison. The fraction of expressed NUPT eukaryotic rRNA genes was found to be significantly higher than expected according to a Fisher's exact test, an effect specifically related to NUPTs-I, while no significant differences were found among NUPT prokaryotic rRNA genes (**Fig. 4A, C)**. The overall expression of NUPT rRNA genes was in general greater than that of non-NUPT ones, with differences being significant for eukaryotic rRNA genes in the case of NUPTs-I and for prokaryotic ones in the case of NUPT-IIs (**Fig. 4B, D)**. For the rest of RNA genes in the nuclear genome, the fraction of expressed genes among those of plastid origin was generally lower than expected for most categories (**Fig. 4E, G, I)**, and in contrast to what was observed among rRNA genes, their overall expression was in all cases smaller among those of plastid origin, with differences significant in all cases where a Wilcoxon's rank tests could be implemented (**Fig. 4F, H, J)**.

Discussion

The rich fraction of plastid DNA found in the moringa genome provides an unprecedent opportunity to study the impact of NUPTs on the evolution of nuclear genome architecture and function. Results presented here reveal the biased distribution of NUPTs across different genomic features, likely indicating that some genomic regions tolerate better the insertion of plastid DNA than others. NUPTs have been found associated to TEs, in coherence with the role attributed to TEs and other nuclear DNA sequences unrelated to organelle DNA in promoting the erosion and rearrangement in the nucleus of recently

inserted plastid DNA, although the precise molecular mechanisms involved have not yet been fully elucidated [14, 31]. In moringa, although young NUPTs from episode II were found to overlap with TEs significantly more than expected by chance, older ones from episode I were actually underrepresented among TEs. These observations might indicate that association of NUPTs to TEs, i) is species-specific [14]; ii) is the outcome of some plastid DNA only immediately upon arrival especially into TE-rich regions of nuclear genomes reported to be hotspots of NUPTs, *e.g.*, the (peri)centromeric regions, [32]; and iii) does not necessarily result in the erosion of NUPTs, as revealed the weak, although significant, negative correlation found between size and sequence identity of younger NUPTs-II [24].

Not surprisingly, structural genes were consistently found to be hit by NUPTs less than expected by chance, reflecting their likely deleterious effect especially when integrated in exon coding regions. Indeed, of the 30 moringa NUPT structural genes affected in exons, plastid DNA contributed partially to the target gene coding sequences only in six out of them, while the rest were entirely of plastid origin. Therefore, although the repeated transfer of copies of plastid DNA stretches to the nuclear genome might provide the plant with a source of genetic material to modify pre-existing gene functions and/or acquire novel ones, other molecular mechanisms rather than exonization seem to have operated on the moringa lineage promoting the repeated fixation of massive amounts of plastid DNA in the nucleus. However, it should be noted that most NUPTs are expected to diverge in sequence through evolutionary time resulting in the amelioration of the plastid DNA sequence to the nucleotide composition of its host chromosome, becoming gradually difficult to detect through direct searches of significant identity with the donor plastid genome regions [31]. Thus, NUPTs might still have contributed to ancient functional exon acquisitions more than anticipated [16]. It had also been reported a role for NUPTs in the dissemination of regulatory elements in the promoter or enhancer of specific genes, resulting in a more efficient transcription [33–35]. Although this might be the case for individual NUPTs in moringa, it does not seem to be a general pattern; overall transcription levels of NUPT structural genes were not found to be different than that of their non-NUPTs counterparts, with the exception of those affected at exonic regions, whose expression was found to be lower and more specific, and those affected by NUPTs-I at intronic regions, which displayed a broader expression. Nevertheless, the consistent expression found here, both quantitatively or qualitatively, for NUPT -structural and -RNA genes taken collectively, suggest i) they are not preferential targets for transcription repression through hypermethylation or other epigenetic silencing mechanisms as had been previously claimed from studies in other species [13, 15–17], or ii) transcription repression only affect specific subclasses of NUPT genes, such as exon structural ones.

In contrast to structural genes, most categories of RNA genes considered in our study were consistently found to contain plastid DNA more than expected by chance. Upon arrival to the nuclear genome, and similarly to structural genes, RNA genes are not expected to be functional. However, we found here a significant fraction of NUPT RNA genes from different categories showing functional expression, in some cases at higher levels than the corresponding non-NUPT counterparts. This was the case of nuclear genes of plastid origin annotated as eukaryotic or prokaryotic rRNA, although it remains to be determined whether NUPT rDNA can contribute to the cytosolic pool of rRNA and ribosomes. The hundreds or thousands of rRNA genes commonly found in eukaryotic nuclear genomes are remarkably well conserved

in sequence, with gene conversion and/or concerted evolution through unequal crossover being the major driving force underlying sequence conservation by sweeping away any newly acquired mutations [36]. This provides a suggestive mechanism to explain the amelioration of prokaryotic rDNA sequences of plastid origin to the nuclear genome. Additionally, extra-ribosomal functions have been suggested for repetitive tandems of rDNA, including to: i) evolve as rRNA-derived RNA fragments (rRFs), a novel class of regulatory small noncoding RNAs (sncRNA), whose exact functions have not been elucidated yet [37–39]; ii) contribute to the maintenance of genome stability, being particularly sensitive to genomic stresses and acting as a source of adaptive response [40].

NUPTs also constitute a major source of tRNA genes in the nuclear genome. The occurrence of organellar tRNA genes had been previously observed in the nuclear genomes of different plant species [41–43]. However, NUPT tDNA commonly represents a minor, although highly variable, fraction of the total tDNA content, while in the case of moringa, 67,72% out of the total 1,583 tRNA genes present in the nuclear genome were of plastid origin, representing almost 100% of tRNA genes annotated as plastid. As a result, the total number of tRNA genes encoded by the moringa nuclear genome is significantly higher than the total number of tRNA genes found in other plant genomes, typically ranging between 500 and 600 [43]. As for rRNA genes, we found a significant fraction of complete tRNA genes of plastid origin being functionally expressed. Given the versatility shown by some nuclear tRNAs that are imported and function in the mitochondria [43] or by functional plastid tRNA genes found in the mitochondrial genome [41, 44], it is tempting to speculate at least some plastid tRNAs might also be contributing to the nuclear pool of tRNA involved in cytosolic translation. Indeed, tRNA gene sequences have been shown to evolve rapidly to meet novel translational demands [45]. 27 out of the 112 NUPT nuclear tRNA genes found in the moringa nuclear genome proceeded from tRNA genes annotated as plastid in the plastid genome, which might well indicate the adaptation of their sequences to the new nuclear environment, in a process similar to the concerted evolution of rRNA genes. Two alternative paths for plastid tDNA in the nuclear genome could be to evolve as i) tRNA-derived RNA fragments (tRFs), or ii) tRNA-related short interspersed nuclear elements (SINEs). tRFs are a class of sncRNAs identified in all domains of life, a significant fraction of which originate as cleavage products from mature plastid tRNAs and have been attributed possible regulatory functions within the plant cell as part of signaling pathways [46]. Plastid tRNA genes are also considered a major source of SINEs, a family of small, abundant, and highly heterogeneous mobile non-autonomous elements transcribed by RNA PolIII, which rely on the enzymatic machinery of an autonomous long interspersed elements (LINE) partner for propagation by retrotransposition [43, 47, 48]. Although for most families of SINEs, their functions remain unknown and need to be elucidated, there is increasing evidence of their impact on gene function and genome evolution in plants. Similar roles have been hypothesized for RNA genes annotated in the nuclear genome as self-splicing intron, most of them shown in moringa to be of plastid origin and functionally expressed. The majority of these belonged to group II introns, capable of carrying out both self-splicing and retrotransposition and also suggested for having a profound impact on nuclear genome evolution. Plastid group II introns likely provided the framework for the emergence of spliceosomal introns and other key components of the spliceosome, eukaryotic retroelements, including telomeres, and other machinery that controls genetic variation and

stability [49, 50]. Furthermore, the ability shown by tRNA-derived SINEs and by group II self-splicing introns of plastid origin to experience retrotransposition provides an alternative mechanism to explain their propagation upon arrival in the nuclear genome through repeated duplication.

In addition, a total of 42 *isrR* genes included in the category of regulatory RNA genes were found in the moringa nuclear genome, all of them deriving from a plastid homolog. *IsrR* genes form a class of specific iron-deficiency-responsive antisense RNA genes, whose product binds specifically to the mRNA of the *isiA* gene, which in turn encodes for a protein component of the photosystem, to induce its degradation [51]. Interestingly, *isrR* genes had been previously found only in cyanobacteria [51], although searches in the RFAM database (https://rfam.org/) showed they were present in the nuclear and plastid genomes of other photosynthetic organisms, including plants; in contrast, no homologues of *isiA* genes have been found in plants [52]. As observed for other categories of NUPT RNA genes, many nuclear *isrR* RNA genes are expressed and are therefore likely functional; it remains to be determined what are the actual mRNA targets of nuclear, and plastid, *isrR* RNA genes in moringa.

In summary, plastid DNA in moringa has a profound impact in the evolution of nuclear genome architecture and function as a main contributor to the nuclear pool of RNA genes, especially those involved in the protein biosynthetic machinery (*i.e.*, rRNA and tRNA genes) and specific classes of regulatory RNAs. Furthermore, our results support similar molecular and evolutionary forces would be contributing to the fixation of NUPTs formed in two events separated in time through seemingly disparate mechanisms. An interesting follow-up question is to determine whether these patterns of fixation of NUPTs observed in moringa are species-specific or also apply to other plant species or taxonomic groups.

Materials and methods

Reannotation of RNA genes in the moringa nuclear and plastid genomes

Reannotation of RNA genes in the moringa chloroplast genome [53] was performed by scanning the RFAM v14.10 database of non-coding RNA families [54] using the command *cmscan* from Infernal v1.1.4-1 [55]. The tRNA genes found by Infernal were completed by merging with the original annotation reported in [53] and the results obtained through tRNAscan-SE v 2.0.12 [56] using the options -G and -O to search for tRNA from the three domains (eukarya, prokarya and archaea) and organellar genomes, respectively. tRNA genes found in the moringa nuclear and plastid genomes were further classified as nuclear, mitochondrial or plastid according to the annotation of the best hit resulting from BLASTN v2.12.0+ [57] searches of a database of *Arabidopsis thaliana* tRNA genes retrieved from PltRNAdb [58], selecting word size 11 and E-value 10^{-2} as settings.

Analysis of the distribution of NUPTs across genomic features

The genome assembly and genomic feature (gff3) files for the moringa nuclear were retrieved from [25]. The genome assembly file for the moringa plastid genome was retrieved from the NCBI Reference Sequence (NCBI-RefSeq) database (https://www.ncbi.nlm.nih.gov/nuccore/NC_041432) [53]. The tabular BLAST file containing the alignments between the moringa nuclear and plastid genomes representing individual NUPTs was retrieved from [24]. Circular plot representations of genomic features were obtained using Circos version 0.69 – 8 [59]. To detect NUPTs overlapping genomic features, the intersect subcommand from BEDTools [60], was employed. The R package *GenomicDistributions* v1.8.0 [61] was used for the base pair overlap count analysis between NUPTs and every other genomic feature. The expected counts of overlapping base pairs between NUPTs and each genomic feature were calculated through the *calcExpectedPartitions* function. Since a large genomic feature in terms of the total number of base pairs it encompasses was expected to overlap more NUPTs by chance than a small one, the *bpProportion* option was activated to account for this bias. Subsequently, *calcExpectedPartitions* performs Pearson's Chi-squared independence tests with Yates' continuity correction to assess whether the observed counts of overlapping base pairs between NUPTs and each genomic feature were significantly different from expected.

Gene functional annotation and enrichment analysis

Functional annotation terms attached to structural genes found in the moringa genome, including Gene Ontology (GO), Enzyme Commission (EC) and KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology (KO) were retrieved from [25]. Enrichment analysis for detecting over- and under-represented functional terms among NUPT structural genes was performed by means of Fisher's exact tests [62]. To control for multiple hypotheses testing, the resulting *P* values were corrected according to the Bonferroni test [63], and those < 0.05 were considered significant. Sub-cellular localization of the nuclear structural genes was predicted by means of DeepLoc 2.0 [28], which generates predictions based on protein language models that only use sequence information.

Analysis of RNA-seq expression data

Expression values measured in Transcripts Per Million (TPM) for nuclear and plastid genes were obtained from paired RNA-sequencing (RNA-Seq) data from five tissues, *i.e.*, flower, leaf, root, seed, and stem, generated in a study of the moringa transcriptome [26], and available at the NCBI Reference Sequence Short Read Sequence Archive (NCBI-SRA) (**Table S11**). The pool of paired end RNA-Seq reads for each tissue was aligned to the nuclear and plastid genomes simultaneously using the aligner GSNAP v2021-12-2 [64] with one mismatch allowed. The resulting SAM alignment file was sorted by position using the command sort from SAMtools v1.13 [65], and then used to obtain TPM values by employing StringTie v2.2.1 [66], a program for transcript assembly and quantitation of RNA-Seq data, on the basis of nuclear and plastid gff3 annotation files. Broadness or tissue-specificity of gene expression was calculated using the Tau index for every gene in the nuclear genome employing the method described in [67] and the expression values from each of the five tissues. Tau index ranges from 0, indicating broader unspecific expression, to 1, reflecting narrower specific expression [27]. Significance in the departure of the fraction of expressed versus unexpressed genes from that expected by chance for specific classes of genes was

assessed through Fisher's exact tests [62]. The significance of the differences in the overall expression, as measured by TPM values, or in expression broadness, as measured by Tau indexes, between subsets of genes, was assessed through Wilcoxon rank tests [68].

For the GSNAP alignments, a Single Nucleotide Polymorphism (SNP) file containing editing sites predicted in the plastid transcripts was used to distinguish them from nuclear transcripts and thus ensure read mapping results reflected actual transcription of NUPT genes. The prediction of editing sites was made by REDItools v2.0 [69], a collection of python scripts for RNA editing site prediction, from a sorted BAM file containing alignments between the pool of reads of the five tissues considered together and the plastid genome. The alignments were performed though GSNAP v2021-12-27 [64] which generated a SAM alignment file that was later converted to bam and sorted by position using the commands view and sort from SAMtools v1.13 [65], respectively.

Abbreviations

EC, Enzyme Commission; GO, Gene Ontology; KO, KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology; LINE, long interspersed element; NUPT, nuclear plastid DNA sequence; rRF, rRNA-derived RNA fragment; SINE, short interspersed nuclear element; sncRNA, small noncoding RNA; RNA sequencing, RNA-seq; TE, Transposable Element; tRF, tRNA-derived RNA fragment; TPM, Transcripts Per Million.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This work was supported 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. Partially funded by

grants PID2019-106758GB-C32 and PID2022-139293NB-C31 funded by MCIN/AEI/10.13039/501100011033, “ERDF A way of making Europe”, to AS.

Author contributions

LC-P conceived and designed the project and all research activities. JPM-R performed the analyses. AS contributed to the statistical analysis implemented in the paper. VI and AA contributed to coding scripts used in the paper and provided computational support. All authors contributed to data analysis and interpretation. LC-P wrote and edited the manuscript with substantial contributions from JPM-R. All authors reviewed the manuscript.

Acknowledgements

Not applicable.

References

1. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biology. 2016;17.
2. Lee S-I, Kim N-S. Transposable Elements and Genome Size Variations in Plants. Genomics Inform. 2014;12:87.
3. Lisch D. How important are transposons for plant evolution? Nature Reviews Genetics. 2013;14:49–61.
4. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003;4:865–75.
5. Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. Annu Rev Genet. 2013;47:307–33.
6. Andersson DI, Jerlström-Hultqvist J, Näsvall J. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol. 2015;7.
7. Carretero-Paulet L, Fares MA. Evolutionary Dynamics and Functional Specialization of Plant Paralogs Formed by Whole and Small-Scale Genome Duplications. Mol Biol Evol. 2012;29:3541–51.
8. Defoort J, Van de Peer Y, Carretero-Paulet L. The Evolution of Gene Duplicates in Angiosperms and the Impact of Protein-Protein Interactions and the Mechanism of Duplication. Genome Biol Evol. 2019;11:2292–305.
9. Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of Saccharomyces cerevisiae Genes. PLoS Genet. 2013;9:e1003176.
10. Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. Plant Cell. 2017;29:2766–85.

11. Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A. 2003;100:8828–33.
12. Kleine T, Maier UG, Leister D. DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol. 2009;60:115–38.
13. Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX. Nuclear Integrants of Organellar DNA Contribute to Genome Structure and Evolution in Plants. Int J Mol Sci. 2020;21:15.
14. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: Size, relative age and chromosomal localization. Heredity (Edinb). 2013;111:314–20.
15. Leister D, Kleine T. Role of Intercompartmental DNA Transfer in Producing Genetic Diversity. Int Rev Cell Mol Biol. 2011;291:73–114.
16. Noutsos C, Kleine T, Armbruster U, DalCorso G, Leister D. Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends in Genetics. 2007;23:597–601.
17. Yoshida T, Furihata HY, To TK, Kakutani T, Kawabe A. Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation. Sci Rep. 2019;9.
18. Pinard D, Myburg AA, Mizrachi E. The plastid and mitochondrial genomes of Eucalyptus grandis. BMC Genomics. 2019;20.
19. Zhao N, Grover CE, Chen Z, Wendel JF, Hua J. Intergenomic gene transfer in diploid and allopolyploid Gossypium. BMC Plant Biol. 2019;19.
20. Stegemann S, Bock R. Experimental reconstruction of functional gene transfer from the tobacco plastid geneome to the nucleus. Plant Cell. 2006;18:2869–78.
21. Lloyd AH, Timmis JN. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. Mol Biol Evol. 2011;28:2019–28.
22. Wang D, Qu Z, Adelson DL, Zhu JK, Timmis JN. Transcription of nuclear organellar DNA in a model plant system. Genome Biol Evol. 2014;6:1327–34.
23. Ojeda-López J, Marczuk-Rojas JP, Polushkina OA, Purucker D, Salinas M, Carretero-Paulet L. Evolutionary analysis of the Moringa oleifera genome reveals a recent burst of plastid to nucleus gene duplications. Sci Rep. 2020;10:1–15.
24. Marczuk-Rojas JP, Álamo-Sierra AM, Salmerón A, Alcayde A, Isanbaev V, Carretero-Paulet L. Spatial and temporal characterization of the rich fraction of plastid DNA present in the nuclear genome of Moringa oleifera reveals unanticipated complexity in NUPTs´ formation. BMC Genomics. 2024;25:60.
25. Chang J, Marczuk-Rojas JP, Waterman C, Garcia-Llanos A, Chen S, Ma X, et al. Chromosome-scale assembly of the *Moringa oleifera* Lam. genome uncovers polyploid history and evolution of secondary metabolism pathways through tandem duplication. Plant Genome. 2022;15:e20238.
26. Pasha SN, Shafi KM, Joshi AG, Meenakshi I, Harini K, Mahita J, et al. The transcriptome enables the identification of candidate genes behind medicinal value of Drumstick tree (Moringa oleifera). Genomics. 2020;112:621–8.

27. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.
28. Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res. 2022;50:W228–34.
29. Sharma M, Pandey GK. Expansion and function of repeat domain proteins during stress and development in plants. Frontiers in Plant Science. 2016;6.
30. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523–31.
31. Noutsos C, Richly E, Leister D. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res. 2005;15:616–28.
32. Matsuo M, Ito Y, Yamauchi R, Obokata J. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. Plant Cell. 2005;17:665–75.
33. Mohan V, Pandey A, Sreelakshmi Y, Sharma R. Neofunctionalization of chromoplast specific lycopene beta cyclase gene (CYC-B) in tomato clade. PLoS One. 2016;11.
34. Blanchard JL, Schmidt GW. Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol. 1995;41:397–406.
35. Ott RW, Chua NH. Enhancer sequences from Arabidopsis thaliana obtained by library transformation of Nicotiana tabacum. Mol Gen Genet. 1990;223:169–79.
36. Eickbush TH, Eickbush DG. Finely orchestrated movements: Evolution of the ribosomal RNA genes. Genetics. 2007;175:477–85.
37. Asha S, Soniya E V. The sRNAome mining revealed existence of unique signature small RNAs derived from 5.8SrRNA from Piper nigrum and other plant lineages. Sci Rep. 2017;7:41052.
38. Chen Z, Sun Y, Yang X, Wu Z, Guo K, Niu X, et al. Two featured series of rRNA-derived RNA fragments (rRFs) constitute a novel class of small RNAs. PLoS One. 2017;12:e0176458.
39. Wang L, Yu X, Wang H, Lu Y-Z, de Ruiter M, Prins M, et al. A novel class of heat-responsive small RNAs derived from the chloroplast genome of Chinese cabbage (Brassica rapa). BMC Genomics. 2011;12:289.
40. Lopez FB, McKeown PC, Fort A, Brychkova G, Spillane C. The boys are back in town: Rethinking the function of ribosomal DNA repeats in the genomic era. Molecular Plant. 2023;16:514–6.
41. Tian X, Zheng J, Hu S, Yu J. The discriminatory transfer routes of tRNA genes among organellar and nuclear genomes in flowering plants: a genome-wide investigation of indica rice. J Mol Evol. 2007;64:299–307.
42. Lin X, Kaul S, Rounsley[2] S, Shea[2] TP, Benito M-I, Town CD, et al. Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. 1999.

43. Michaud M, Cognat V, Duchêne AM, Maréchal-Drouard L. A global picture of tRNA genes in plant genomes. Plant Journal. 2011;66:80–93.
44. Bock R. Extranuclear inheritance: Gene transfer out of plastids. In: Progress in Botany. Berlin/Heidelberg: Springer-Verlag; 2006. p. 75–100.
45. Yona AH, Bloom-Ackermann Z, Frumkin I, Hanson-Smith V, Charpak-Amikam Y, Feng Q, et al. Trna genes rapidly change in evolution to meet novel translational demands. Elife. 2013;2013.
46. Cognat V, Morelle G, Megel C, Lalande S, Molinier J, Vincent T, et al. The nuclear and organellar tRNA-derived RNA fragment population in Arabidopsis thaliana is highly dynamic. Nucleic Acids Res. 2017;45:3460–72.
47. Wenke T, Döbel T, Sörensen TR, Junghans H, Weisshaar B, Schmidta T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell. 2011;23:3117–28.
48. Vanburen R, Ming R. Organelle DNA accumulation in the recently evolved papaya sex chromosomes. Molecular Genetics and Genomics. 2013;288:277–84.
49. Novikova O, Belfort M. Mobile Group II Introns as Ancestral Eukaryotic Elements. Trends in Genetics. 2017;33:773–83.
50. Pyle AM. Group II Intron Self-Splicing. Annu Rev Biophys. 2016;45:183–205.
51. Dühring U, Axmann IM, Hess WR, Wilde A. An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. Proceedings of the National Academy of Sciences. 2006;103:7054–8.
52. González A, Fillat MF, Bes M-T, Peleato M-L, Sevilla E. The Challenge of Iron Stress in Cyanobacteria. In: Cyanobacteria. InTech; 2018.
53. Lin W, Dai S, Chen Y, Zhou Y, Liu X. The complete chloroplast genome sequence of Moringa oleifera Lam. (Moringaceae) . Mitochondrial DNA Part B. 2019;4:4094–5.
54. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: An RNA family database. Nucleic Acids Research. 2003;31:439–41.
55. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
56. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49:9077–96.
57. Altschup SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990;215:403–10.
58. Mokhtar MM, Allali AEL. PltRNAdb: Plant transfer RNA database. PLoS One. 2022;17 5 May.
59. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res. 2009;19:1639–45.
60. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

61. Kupkova K, Mosquera JV, Smith JP, Stolarczyk M, Danehy TL, Lawson JT, et al. GenomicDistributions: fast analysis of genomic intervals with Bioconductor. BMC Genomics. 2022;23.
62. Fisher RA. Statistical methods for research workers, 5th ed. Oliver and Boyd: Edinburgh; 1934.
63. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilit `a. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;8:3–62.
64. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
66. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
67. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.
68. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1:80.
69. Lo Giudice C, Tangaro MA, Pesole G, Picardi E. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. Nat Protoc. 2020;15:1098–131.

Figures

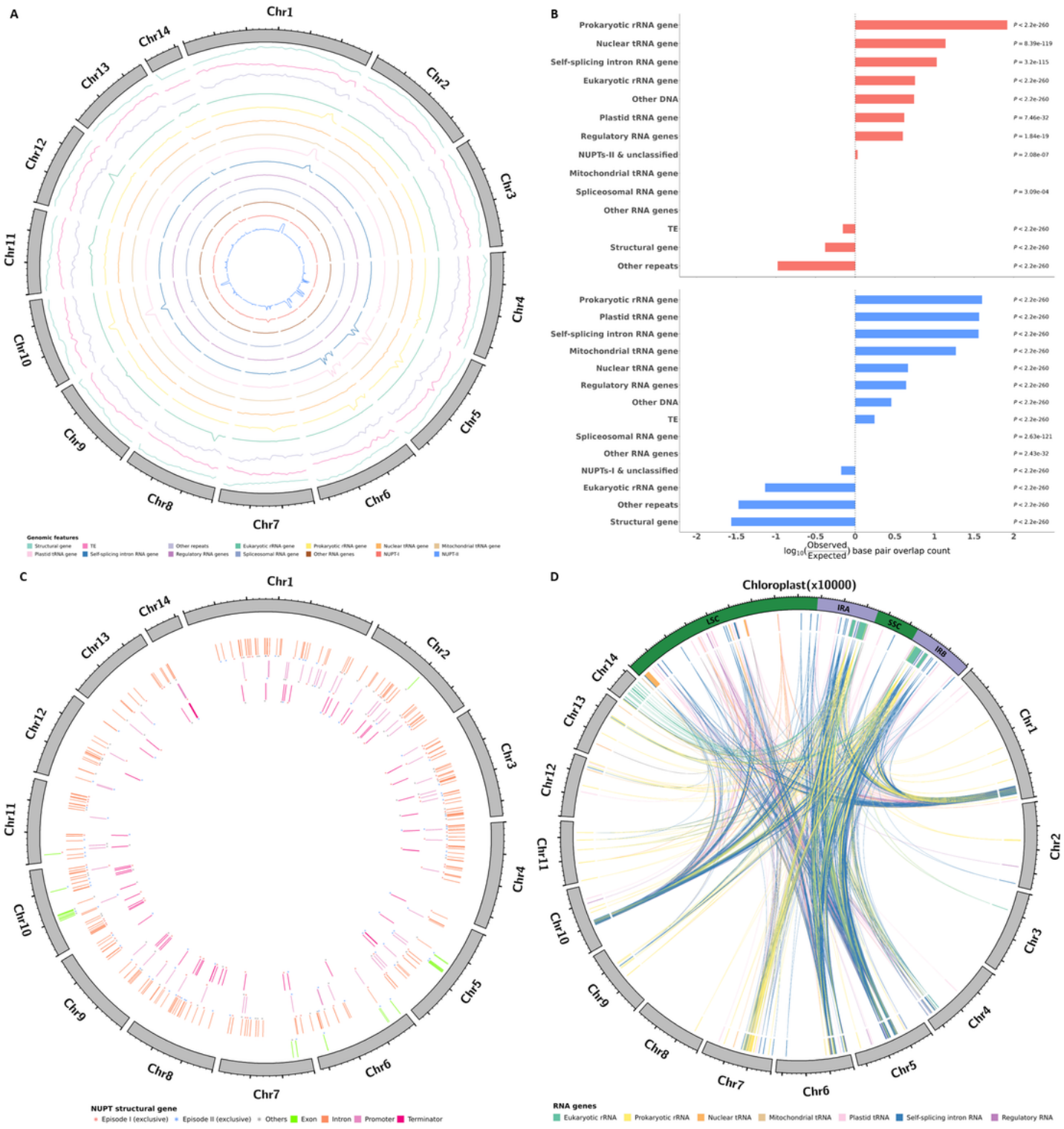

Figure 1

Distribution of NUPTs across genomic features detected in the nuclear genome of moringa. (**A**) Circos plot representation of genomic features detected in the moringa nuclear genome. The 14 nuclear chromosomes are represented as grey filled blocks forming a circle and arranged clockwise. Nuclear chromosomes are drawn to scale, with lengths proportional to size and expressed in Mb. Each inner ring contains a line plot representing the density distribution of specific genomic features in windows of

500,000 bp. Structural genes include the 1 Kb region upstream of the ATG codons, exons, introns and the 1 Kb region downstream of the stop codon. (**B**) Overlap count analysis between NUPTs and every other genomic feature. Above, NUPTs-I (red); below, NUPTs-II (blue). For easier visualization, overlap values are displayed as log10-transformed observed/expected base pair overlap counts. Significant *P*-values resulting from performing Pearson's Chi-squared tests with Yates' continuity correction between observed and expected base pair overlap counts are shown. (**C**) Circos plot representation of NUPT structural genes detected in the moringa nuclear genome. The nuclear genome is depicted as in panel a. NUPT structural genes are shown as tiles colored according to the gene region affected. Colored asterisks indicate the formation episode of the NUPT(s) present in the structural genes affected: I, II and other, *i.e.*, one episode and the another and / or unclassified. (**D**) Circos plot representation of NUPT RNA and plastid RNA genes detected in the moringa nuclear and plastid genomes, respectively. The nuclear genome is depicted as in panel a. The chloroplast chromosome is represented as a green filled block located at 12 o'clock and has been upscaled to occupy a quarter of the image circumference; its size unit was set to 10,000 bp. NUPT RNA and plastid RNA genes are shown as tiles colored according to their category. NUPT RNA genes and their donor gene regions in the chloroplast genome are connected through links whose colors indicate whether they were annotated to the same category or not, in which case they are shown in grey. LSC, Large Single Copy; IRA, Inverted Repeat A; IRB, Inverted Repeat B; SSC, Small Single Copy.

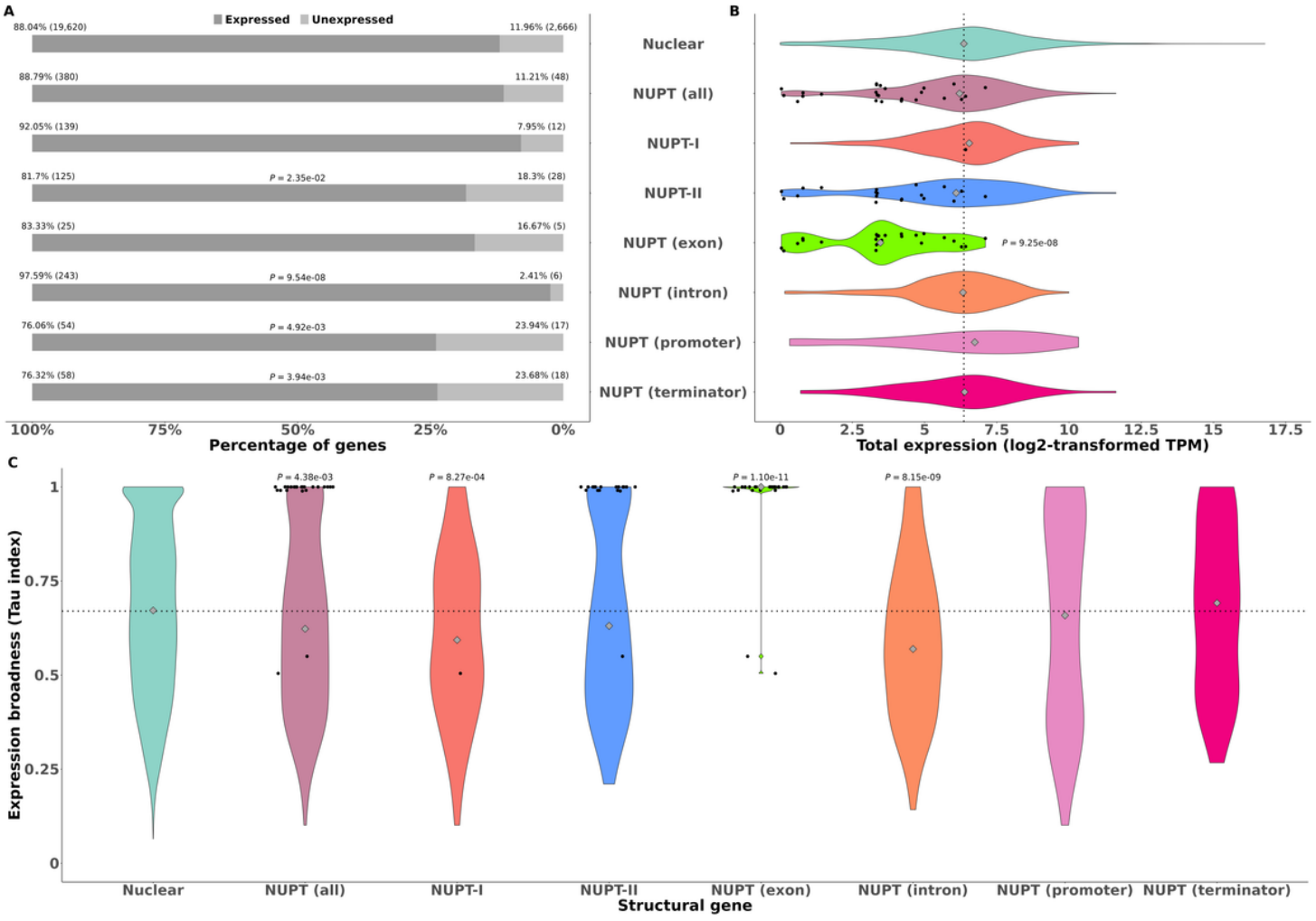

Figure 2

Expression levels and broadness of NUPT structural genes partitioned by formation episode and gene region affected plus non-NUPT ones. (**A**) stacked bar plot representing the percentage of expressed versus unexpressed genes. Significant *P*-values from performing Fisher's exact tests for each category of NUPT structural genes versus the whole set of structural genes are shown. (**B**) violin plot representing log2-transformed expression values measured in Transcripts Per Million. Only expressed structural genes were considered. (**C**) violin plot representing expression broadness measured by means of Tau index for every gene in the category. In all cases, individual NUPT genes affected in exonic regions are depicted as black dots. Grey diamonds inside violin plots in (**B**) and (**C**) represent median expression level and broadness, respectively, for each category of structural genes. Dotted lines in plots (**B**) and (**C**) show the median expression level and broadness, respectively, of non-NUPT structural genes. Significant *P*-values in plots (**B**) and (**C**) resulting from performing Wilcoxon rank tests between each category of NUPT structural genes and non-NUPT ones are shown.

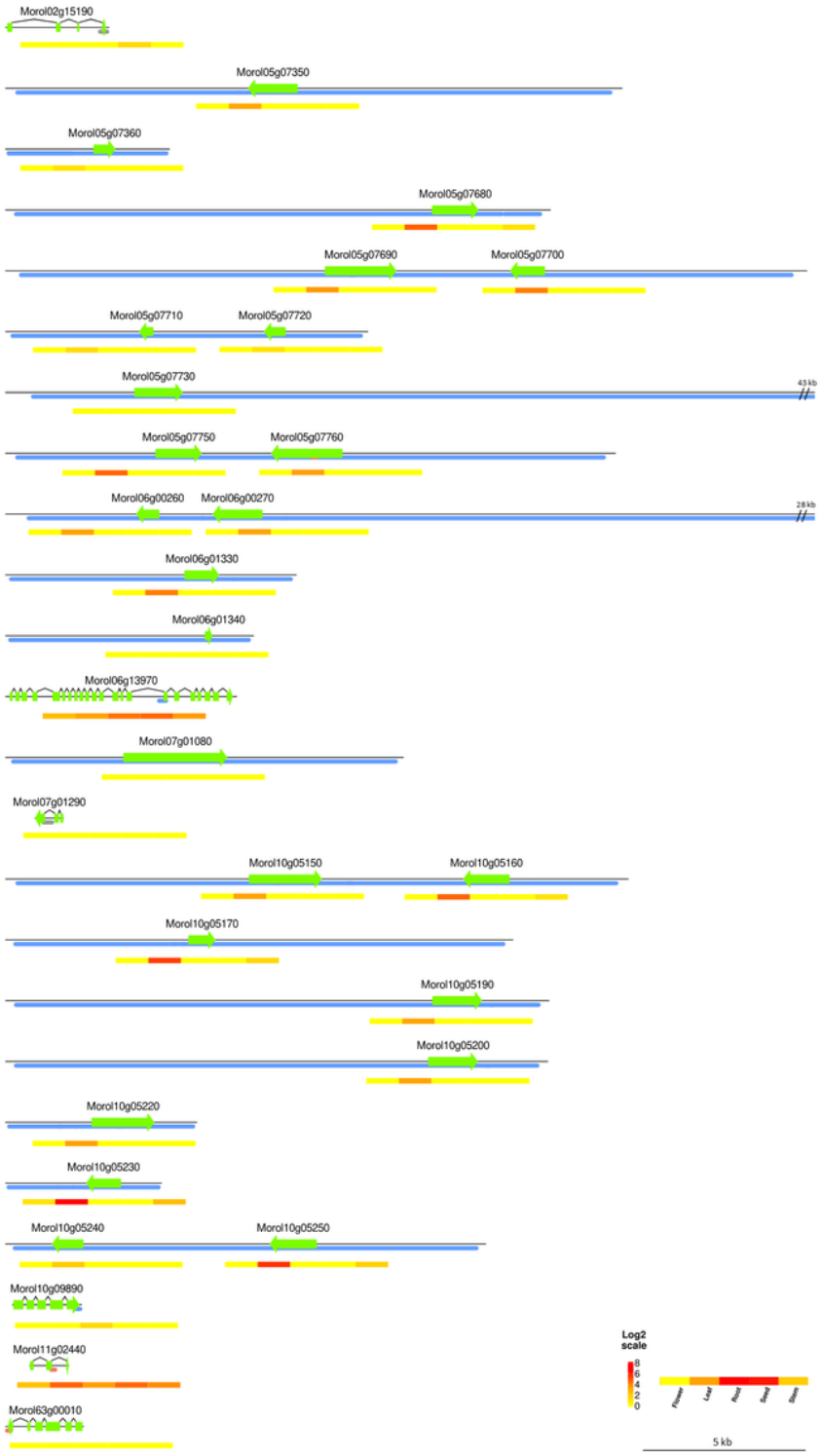

Figure 3

Schematic diagram of the 30 NUPT structural genes affected in exons. Exon and intron regions are represented as green blocks and arrows and curved lines, respectively. NUPTs are depicted as blocks colored according to their formation episode: I (red), II (blue) and unclassified (grey). Below each NUPT structural gene there is a heat map representation of their expression patterns in five tissues. The colors of the heat map represent log2-transformed expression values measured in Transcripts Per Million. The

elements in the diagram are drawn to scale. Some elements were trimmed to adjust the total size of the image; their actual sizes are indicated.

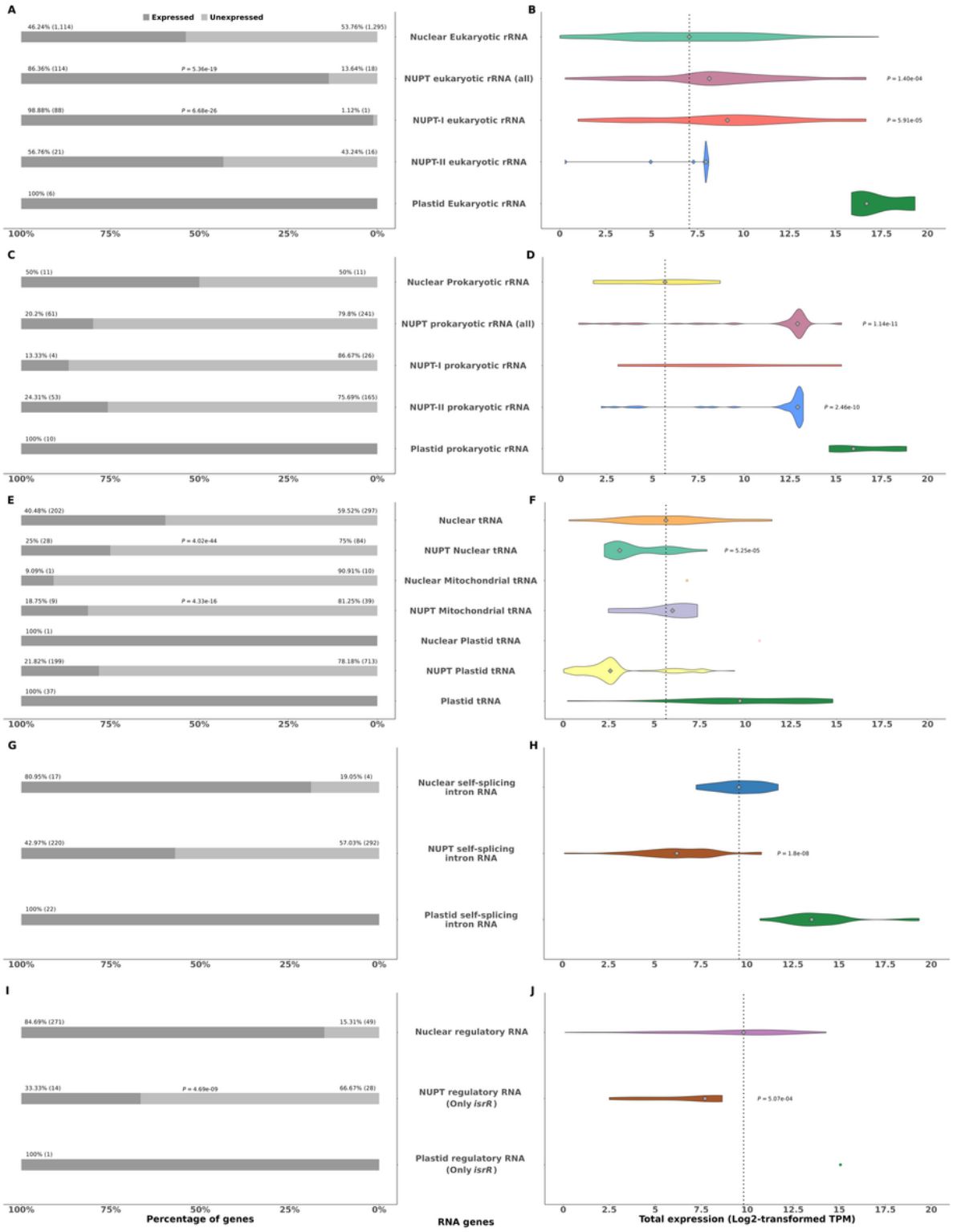

Figure 4

Expression levels of specific classes of NUPT, non-NUPT, and plastid RNA genes. (**A**), (**B**) Eukaryotic rRNA. (**C**), (**D**) Prokaryotic rRNA. (**E**), (**F**), tRNA. (**G**), (**H**) Self-splicing intron RNA. (**I**), (**J**) Regulatory RNA. On the

left panels, stacked bar plots representing the percentage of expressed versus unexpressed genes. Significant *P*-values resulting from performing Fisher's exact tests for each class of NUPT RNA genes versus the whole set of nuclear RNA genes of that class are shown. On the right panels, violin plots representing log2-transformed expression values measured in Transcripts Per Million. Only expressed genes were considered. Groups with fewer than two expressed genes are depicted as colored single dots. Grey diamonds inside violin plots represent median expression level for each class of RNA genes. Dotted lines indicate the median expression of nuclear RNA genes for each category. Significant *P*-values resulting from performing Wilcoxon rank tests between NUPT and nuclear RNA genes of each class are shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiguresPaperV0.pptx
- SupplementaryTablesPaperV0.xlsx