Genome-wide identification of conserved RNA structure in plastids

Dolly Mehta[1], Jingmin Hua[1], Rodrigo S. Reis[1,*]

[1] Institute of Plant Sciences, University of Bern, Switzerland

* Correspondence: rodrigo.reis@unibe.ch

Abstract

The systematic discovery of functional RNA structures directly from sequence remains a central challenge in genomics. Plastid genomes, with their deep evolutionary history and extensive data availability, provide an ideal system to address this challenge. Here, we developed a stringent, covariation-based pipeline that integrates well-established and novel tools to perform an unbiased screen for conserved secondary structures across ~14,000 plastid genomes. Our analysis identified a repertoire of 51 conserved RNA motifs. Beyond recovering known functional classes, we uncover novel structures within the UTRs and introns of key photosynthetic genes, including psbN and atpF, pointing to previously unknown regulatory hubs. This work provides a stringent atlas of conserved plastid RNA structures, reveals an extensive layer of post-transcriptional regulation, and establishes a robust framework for structural genomics in other systems, including nuclear genomes.

Introduction

Plastids evolved from endosymbiotic cyanobacteria likely ~1.5 billion years ago (Bya) in a common ancestor shared by land plants and algae[1–3]. In land plants, chloroplast genomes are conserved in structure and organisation, typically ~150 kb circular quadripartite DNA sequence with two copies of an inverted repeat (IR) separated by a large and small single copy (LSC and SSC) regions, and 100-120 genes, mostly encoding for subunits of the photosystems, electron transport chain, and gene expression machinery[4]. Substantial differences are, however, observed in some plant clades, from the absence of "essential" genes in grasses[5] and lack of an IR copy in Fabeae and related legumes[6,7], to extreme cases where chloroplast lost its photosynthetic function, such as in parasitic plants[8]. The latter comprises the smallest plastid genomes, reaching as short as ~11 kb, while few genomes are larger than 200 kb[4]. Despite some variations, plastid genome in land plants is far more conserved than mitochondrial and nuclear genomes, possibly because of lower mutation rate in plastids[9], indicating stronger evolutionary pressure on a likely optimal sequence composition shared among land plants.

Plants have evolved to rely on plastids for various functions beyond photosynthesis, including for development and environmental adaptation[10]. Compared with the cytoplasm, plastid provides cellular compartments with unique properties, such as distinct pHs (i.e., ~8.0 and ~5.5 in stroma and thylakoid lumen, respectively), reducing environment, and high levels of O_2, CO_2, Mg^{2+}, and ATP[11]. Many metabolites are primarily or exclusively produced in plastids, including starch, chlorophylls, carotenoids, glutathione, fatty acids, and some amino acids and hormones[12]. In fact, as opposed to mitochondria, the same plastid genome (ptDNA) is shared by diverse and specialised organelles, such as chloroplast and non-photosynthetic amyloplast (storage of starch), elaioplast (fatty acid), chromoplast (pigments), proteinoplast (protein), gerontoplast (aging chloroplast), and etioplast (non-photosynthetic chloroplast in dark-grown plants)[12]. Nuclear and plastid gene expression determine such specialisation—a process permissive to interconversion in response to environmental cues, developmental stage, and tissue type[13]. For instance, chloroplast to chromoplast conversion is responsible for the green to red transition in tomato fruit during ripening[14], while potato tubers exposed to light have their amyloplasts converted into chloroplasts in peripheral cell layers, causing tuber greening via accumulation of chlorophyll[15]. Among the different plastids, chloroplast has a broader role in plant development and adaptation, not only for the importance of photosynthesis, but also for the conserved retrograde signalling pathways, by

which chloroplast regulate nuclear gene expression via selective release of specific signals in response to cellular and environmental cues[16–18].

Evolution of plastid gene expression regulation is central to plant biology. Indeed, there has been a wealth of conservation studies on plastid genes and proteins[19–21], revealing, for instance, deeply conserved prokaryotic-type RNA polymerase[22,23] and ribosome[24,25] similar to those in cyanobacteria. Most plastid genes are organised as operons, and chloroplast has evolved specialised post-transcriptional RNA processing to ensure transcript diversity and proper stability and translation of mature transcripts, such as RNA editing, splicing, 5′ and 3′ trimming, and intercistronic cleavage[26]. Regulation of RNA processing, stability, and translation often require the presence of RNA structure within the primary or mature transcript[27]. However, whether and how these regulatory elements are conserved are still poorly understood, and most well-characterized examples are limited to common cases for RNA structure, such as tRNA[28], rRNA[29], and self-splicing intron[26,30], or functional analysis of RNA structure in one or few plant species[31–35]. For instance, presence of stable stem-loop structure at the 3′ end of many primary transcripts is well-known, as this is a critical motif for precise 3′ end processing and RNA stability[36], while much less is understood about their conservation, including nucleotide pairing covariation and stem-loop length.

In this work, we present an unbiased genome-wide identification of conserved RNA secondary structures in plastids. We developed a computational approach for efficient, robust, and stringent covariation analysis using multiple sequence alignments (MSAs) produced with sliding windows covering the entire plastid genome of thousands of plant species. We compared pairing probability matrices for each sequence in each MSA to identify the shared secondary structure in the alignments, i.e., no a priori structure was used (e.g., minimal free energy structure). We then retained only MSAs associated with structures ranked above stringent criteria for accepting motifs using RNA-SCoRE, recently developed in our group (unpublished). We further applied R-scape[37] to determine covariation and, thus, retain RNA motifs with well-supported evidence for conservation. We could identify high-confidence 51 RNA motifs conserved across ~14k plastid genomes, including known functional motifs (e.g., tRNA, group II intron, and 3′ end stem-loop) and novel uncharacterized motifs, such as within the UTR of psbI, rps4 and psbN, and intronic region of atpF. Our work establishes a framework for reliable identification of conserved RNA structure within large genomic datasets, and unveils novel putative regulatory elements in plastid RNA maturation, stability and/or translation.

Results and discussion

Genome-wide identification of conserved RNA secondary structures in plastids

Identification of conserved RNA structure requires multiple sequence alignment, typically, built using pre-specified target sequences[38–40]. Tools for screening of entire genomes, such as RNAz[41], are based on calculation of thermodynamic stability that mainly favours minimum free energy structures, which commonly leads to spurious predictions and are, thus, not suitable for identification of motifs. A powerful approach is to leverage the existing repertoire of conserved RNA structures available from the RNA family (Rfam) database[42]. For this, we used the cmsearch program of Infernal[43] to screen ~14k plastid genomes (i.e., currently available sequences from NCBI), which identified conserved motifs for tRNAs, rRNAs, group I and group II intron, and the isrR-antisense RNA. The latter was found in nearly all plastid genomes and mapped to an antisense region within the psbC gene. In iron-deficient cyanobacteria, IsiA (iron stress-induced protein A) forms a giant ring structure around photosystem I, a process post-transcriptionally regulated by the expression of an internal IsiA antisense noncoding RNA, IsrR (iron stress-repressed RNA)[44]. IsrR-antisense RNA is conserved across photosynthetic species, suggesting that this noncoding RNA is a central, yet uncharacterized regulator of photosynthesis in most plant species. Therefore, search using the Rfam database proved valuable, but limited to the identification of a single putative novel motif in plastids. We, thus, developed a pipeline for automated, fast, stringent, and reliable screening of thousands of genomes for the identification of conserved RNA motifs. This was applied to plastid genomes, albeit it is in principle applicable to any set of genomes.

Seed evolution ~370 Mya was a pivotal innovation in plants, with seed plants (i.e., gymnosperms and angiosperms) accounting for over 90% of living species[45]. Their current success involved diversification of photosynthetic strategies to support environmental adaptation, such as the evolution of C3, C4, and CAM photosynthesis[46]. We, thus, obtained 3823 manually curated plastid genomes from seed plants[47], and extracted sliding window sequences of 250 nucleotide (nt) in length, with 75 nt of overlap, covering the entire genomes (~5.6M windows) (Fig. 1a). Sequences were processed to remove redundancy and clustered based on a minimum of 50% sequence identity using MMSeqs2[48] (Fig. 1a), resulting in >10,000 clusters, each with a minimum of 10 sequences. A pre-screening was performed with the fast-performing RNALalifold[49], which calculates locally stable secondary structures, to identify clusters with structures sharing at least 2 base pairs. This resulted in 3136 clusters (Fig. 1b), for which each sequence was folded and realigned using the

probabilistic mode of LocARNA[50], resulting in the identification of 1827 cluster where sequences share secondary structure, even if only partially. We recently developed RNA-SCoRE, a tool to evaluate whether each sequence within an alignment adopts the consensus structure and to produce clean alignment containing only the sequences forming the predicted structure (unpublished). RNA-SCoRE evaluation resulted in 667clusters (out of 1827) with a minimum of 7 sequences. We used these cleaned alignment as seeds to find homolog sequences using Infernal[43], which were then processed to obtain clusters including the new hits and those that passed the RNA-SCoRE evaluation were retained, thus producing improved alignments with better performance as seed. The improved alignments were evaluated for covariation using R-scape[37] to identify conserved RNA structures beyond phylogenetic expectations. R-scape identified 145 clusters (out of 667) with at least 2% base pairs showing statistically significant covariation (Fig. 1b). Alignments obtained with LocARNA are often biased towards position containing structure, hence, we further evaluated the alignments using nhmmer[51] and CaCoFold[52] to realign the sequences without structure information and evaluate the LocARNA consensus structure, respectively, to retain only alignments that support the corresponding predicted structure by LocARNA. Out of 145 clusters, 51 passed this evaluation, i.e., RNA motifs with covariation for which we established high confidence predicted structure and high alignment reliability, hereafter, conserved RNA motifs in plastids (Fig. 1a-b). We then used the multiple sequence alignment for each conserved RNA motif as covariance models to search for structure homologs across ~14k plastid genomes, including non-seed plants and evolutionarily distant unicellular flagellates (Fig. 1a). We mapped the genomic context of each motif and observed that ~70% encompass tRNAs, exon-intron junctions, or 5' or 3' UTR, while genic context of the remaining ~30% is unclear (Fig. 1b).

tRNA function is based on folding into specific conformation and it is, therefore, expected to display covariation in our analysis. We obtained eight models mapping to annotated plastid tRNAs, all of which were also identified using the Rfam model for tRNA (RF00005) (illustrated in Fig. 1c; listed in Supplementary Table 1). In the Rfam model, there is covariation in all base pairs, while fewer were identified in our tRNA models, likely explained by higher sequence identity among tRNAs in plastids, as compared to phylogenetic distant genomes used in the Rfam model[53]. On the other hand, tRNA models in our study mapped to different types of tRNAs, each specific to an amino acid, indicating that our pipeline captures functionally related structures with relatively subtle differences that are important for functions. Furthermore, our models identified tRNA in plastids across the entire plant phylogenetic tree (Fig. 1d), evidencing that our pipeline is effective and reliable.

The other identified conserved structures were found in intron or untranslated region (UTR) associated with functionally diverse genes, such as photosystem II (psbA, psbI, psbJ, psbM and psbN), ATP synthase (atpF), NADH dehydrogenase (ndhA, ndhB and ndhD), cytochrome b/f complex (petD), maturase (matK), ribosome subunit (rpl16), c-type cytochrome biogenesis (ccsA), and photosystem I assembly (ycf3), (Fig. 1d). In most cases, conservation was restricted to plastid in seed plants, with the except of a motif in the psbN 5'UTR that was found conserved across the entire phylogenetic tree.

Conservation in group II intron

Plastid genome has an origin in cyanobacterium and has retained self-splicing intron within a subset of their genes, albeit plastid evolved to require protein cofactors (mostly nuclear-encoded) necessary for efficient and proper splicing[54]. They belong to two main classes of self-splicing RNAs, i.e., group I and II introns, and their catalytic activity is intrinsically linked to their ability to fold into specific and highly conserved conformation—required for both the ribozyme activity and RNA interaction with protein cofactors[54]. In land plants, ~20 group II introns and one group I intron have been reported in plastid genomes[55]. Here, we identified conserved structures in group II intron mapping to NADH dehydrogenase (ndhA and ndhB), ribosome subunit (rpl16), photosystem I assembly (ycf3), and tRNA (trnK).

Group II intron is defined by a characteristic and conserved secondary structure, typically consisting of six major stem-loop domains, designated D1, D2, D3, D4, D5, and D6 (Fig. 2a). These domains radiate from a central core region (multiloop junction), forming an arrangement that is essential for bringing distant sequence elements, including the 5' and 3' splice sites and the branch-point nucleotide, into the precise spatial orientation required for splicing[56]. D5, a key catalytic domain, is the most conserved domain (structurally and in sequence), typically forms a short stem-loop structure (~34 base pairs), and contains a conserved catalytic nucleotide triad (partially forming the AC bulge; Fig. 2a) and a GNRA tetraloop involved in D5 stabilisation and interaction with divalent metal ions and other regions of the intron. In fact, D5 shares the same structural conservation between group II intron and U6 spliceosomal RNA (Fig. 2b), indicating an evolutionary link between self-spliced group II introns and splicing in eukaryotes[57]. The stem-loop domain 6 (D6), on the other hand, harbours and unpaired bulged nucleotide, typically an adenosine, that serves as branch point for lariat formation (Fig. 2a), in which the adenosine's 2'-hydroxyl group initiates the first transesterification reaction of splicing.

Our analysis identified conserved D5 in ndhA, ndhB, rpl16, and ycf3 (Fig. 2c), as well as in trnK (Fig. 2d), all of which known to harbour group II intron[58]. We also found a conserved

motif that resembles D5 in petD (Fig. 2d), also known to have group II intron; however, this motif diverges in the loop (AAAA[59,60], instead of canonical GNRA tetraloop) and within the catalytic nucleotide triad (GU, instead of AC bulge). petD has a single group II intron processed by similar proteins required for processing of ndhA, ndhB, rpl16, ycf3–int1, among others[61], suggesting that the differences in D5 between petD and aforementioned transcripts (Fig. 2c) might have a regulatory function, but group II intron in petD likely retained similar processing. In accordance, petD is transcribed as part of the psbB operon (psbB-psbT-psbH-petB-petD), which has two group II introns (psbB and petD), and null mutation in ALBINO EMBRYO AND SEEDLING (AES) gene results in similar intron retention in psbB and petD[62].

We also identified a conserved structure in the group II intron domain D1 of atpF (Fig. 2e). Typically, D1 is the largest and most variable of the six domains, which is further organised in subdomains that include conserved sequence elements known as exon-binding site 1 (EBS1) and EBS2. EBSs base pair with corresponding complementary intron-binding site 1 (IBS1) and IBS2 in the upstream exon to ensure accurate 5' splice site recognition. D1 is also critical for global stabilisation of the intron RNA via tertiary interactions (e.g., δ-δ', a-a', and b-b')[56]. The conserved motif identified here is the stem-loop containing EBS1 and the guanine (δ) that makes tertiary interaction with the first 3′ exon nucleotide (δ′) adjacent to D6[63] (Fig. 2e). EBS1–IBS1 pairing has been shown to be essential for splicing, whereas EBS2–IBS2 pairing is necessary for efficient splicing but is not required[63]. In the same work, the authors showed that δ-δ′ interaction is not involved in efficient splicing and likely does not contribute to the 3′ splice-site selection, indicating a mechanistic difference between self-splicing group II intron and protein-mediated group II intron splicing in higher plant chloroplasts.

Conservation in 3′ end stem-loop structure

Plastid transcription is typically pervasive and termination is inefficient, producing transcripts that require processing[36]. Transcript processing often involves the presence of an inverted repeat (IR) in 3′ UTR to evade 3′-to-5′ exonuclease degradation and define precise 3′ end[36]. Although terminal IRs are well-documented functional motifs in plastids, their conservation in sequence and structure is less well understood. Here, we identified conserved terminal IR motifs associated with genes involved in photosystem II (psbA, psbM, and psbJ), cytochrome c biogenesis (ccsA), and NDH complex (ndhD) (Fig. 3a). IR-dependent RNA maturation and stability have been described for psbA[36,66] and ndhD[67], both of which with similar IR sequences as those identified here. Structural predictions have identified IR in the 3′ UTR of ccsA, psbJ, and psbM in various plant species[68,69], also similar to psbA and ndhD. Conserved IR motifs displayed 19-24 base pairs in the main stem-loop, including a bulge in

psbA and psbJ, while there were no unpaired nucleotides in the IR for ccsA, psbJ, and psbM. Furthermore, there were other shorter stem-loops as part of the conserved motifs in psbA and psbJ, both upstream the IR, and downstream the IR in psbM; however, we only identified covariation in the stem-loops involved in 3′ processing.

We found the conserved motif in psbM across the phylogenetic tree. There is evidence for widespread IR formation in 3′ ends via small inversions in plastid genomes[68], which is expected to produce stems with perfect base pairing, while mismatches likely arise during evolutionary selection, as observed for stem-loop in plant miRNA precursors[70]. However, the consensus structure for psbM stem shows perfect base pairing, even though it has been under selection for over 1 billion years in plastid evolution, suggesting that such regulatory motif is under strong evolutionary selection. The significance of this observation will require further experimental analysis beyond this work, but we speculate that psbM terminal stem-loop might reveal fundamental rules on the evolution of 3′ end processing in plastids.

Stem-loop stability typically shows positive correlation with GC content[71]. Hence, we assessed GC content distribution in aligned sequences for psbA, psbM, psbJ, ccsA, and ndhD covariation models (Fig. 3b). We observed a broad range of median GC content, from 27% in psbM to 44% in psbJ, with relatively tight spread in all motifs except for ccsA and ndhD, where GC content showed difference of over 10% between the 25% and 75% percentile, as opposed to ~4% difference in the other motifs. It is, thus, possible that IRs display distinct stability with regulatory function, such as weaker evasion of 3′-to-5′ exonuclease degradation under conditions that destabilise IRs with lower GC (e.g., heat stress).

In addition to typically long terminal stem-loop, we identified a conserved terminal motif formed by two short stem-loops in tandem (Fig. 3c). This motif is found at the 3′ of psbI and rps4 in most seed plant species, but also at the 3′ of ycf3 is some seed species and, in fewer cases, at the 3′ of chlN in evolutionarily older species, such as single-cell green alga *Chlamydomonas reinhardtii*.

However, compared to psbM IR motif, this tandem terminal stem-loops motif is not well-conserved among plastids in non-seed plants, suggesting a possible convergent evolution to produce related regulation of 3′ processing in some plastid genomes, likely different from typical IR-protected 3′ UTR to evade 3′-to-5′ exonuclease degradation. Furthermore, this tandem terminal motif has been detected in rps4 by 3' RACE, where the authors showed evidence for protein interaction as a major mechanism to define 3' transcript termini in

plastids[72], albeit no detailed characterisation was performed for this particular tandem terminal motif.

Conservation in 5′ UTR and coding sequence

The 5' UTR in plastid mRNAs contains various *cis*-regulatory elements that typically interact with proteins to control translation, stability, and processing in response to developmental and environmental cues[73–76]. Here, we identified a conserved stem-loop (11-nt in length) in the psbN 5' UTR (Fig. 4a), for which we are unaware of reported function. Unlike most plastid 5' UTRs, psbN 5' UTR lacks a Shine-Dalgarno (SD)-like sequence and, instead, requires protein binding for translation initiation, e.g., binding to a 18-nt site adjacent to the START codon in tobacco psbN 5' UTR[77]. However, the identified conserved motif is often not within 50 nucleotides from the START codon, hence, it is unlikely to play a direct role on translation initiation. Identification of primary and mature transcripts in barley revealed that the conserved motif is retained in the mature psbN (i.e., 5' UTR), but only in a subset of transcripts[78], suggesting a possible association between alternative 5' processing and this conserved structure.

Many plastid genomes contain two inverted repeats (IRa and IRb) that separate a single copy section (LSC) from a short single copy section (SSC), hence, genes in inverted repeat are typically found duplicated in the genome. We identified a conserved motif in the 5' UTR of both copies of ycf2 (Fig. 4b). This motif, however, is associated with a tRNA in the opposite DNA strand, and selective pressure is likely on the tRNA structure, rather than on ycf2 5' UTR. Although such motifs are usually discarded from RNA conservation analyses, it is possible that tRNA-like structure in the 5' UTR has regulatory function, particularly because it’s association with ycf2 has been retained in both gene copies. For instance, the psbDCZ operon has a similar tRNA-like structure between psbC and psbZ, which has been proposed to function in polycistronic mRNA maturation via tRNA-induced RNase cleavage, yielding mono-cistronic species[79]. Similarly, in mitochondria, tRNA-induced RNase cleavage has been shown to regulate maturation of transcripts containing tRNA-like structure[80]. It is possible that the tRNA-like motif identified in ycf2 5' UTR is required in mRNA maturation to define precise 5' end; however, this requires further experimental evaluation.

The large ribosomal subunit rpl22 is essential in most seed plants[81]. Here, we identified a conserved motif in the rpl22 coding sequence (CDS), formed by a short stem-loop, likely uncharacterized for its function in plastids (Fig. 4c). rpl22 gene is often within a large operon with other ribosomal protein genes (e.g., rpl2, rps19, rpl23, and rps3), for which maturation requires activity of endonuclease at multiple sites[82], while splicing and editing seems either

less prevalent or non-existing in rpl22 across plant species. Given the tight regulation of ribosomal subunits at multiple levels, it is possible that the conserved motif within rpl22 CDS plays a regulatory role possibly by interacting with specific proteins.

Discussion

This study presents a novel genome-wide, unbiased pipeline for the identification of conserved RNA secondary structures across thousands of plastid genomes. Our computational framework, which relies on stringent covariation analysis, identified 51 conserved RNA motifs. The methodology was validated by the recovery of well-characterized functional structures, including tRNAs, canonical domains of group II introns, and 3' end processing stem-loops. In addition, we discovered novel and uncharacterized motifs, such as those within the 5' UTR of psbN, the coding sequence of rpl22, and the intronic regions of atpF. These findings significantly expand the known repertoire of putative *cis*-regulatory RNA elements, evidencing that conserved RNA structure is a pervasive feature of plastid gene regulation.

The functional significance of these novel motifs requires experimental investigation. The deep conservation of the psbM 3' inverted repeat, which shows perfect base pairing across more than a billion years of evolution, may point to a fundamental mechanism of 3' end processing in plastids. In contrast, the tandem terminal stem-loops found in psbI and rps4 appear conserved mainly in seed plants, suggesting a more recently evolved regulatory mechanism. Future work should prioritize functional analysis of these novel structures. Furthermore, these conserved RNA motifs likely function by recruiting nuclear-encoded proteins to mediate RNA maturation. Identifying these protein interactors is a critical next step toward elucidating the molecular pathways these structures regulate.

Method

Plastid genomic sequences

An initial list of 3823 plastid genomes from seed plants (i.e., spermatophyta) was obtained from the CpGDB database[47]. This list was complemented with plastid genomes obtained directly from NCBI representing all available photosynthetic organisms, resulting in a total of 14034 plastid genomes.

Genome-wide identification of conserved RNA structure

The pipeline developed in this work is available from GitHub https://github.com/RodrigoReisLab/RNA_conservation_chloroplast) and the tools used in the pipeline made available as a container at Docker Hub Container Image Library (https://hub.docker.com/r/dollycm/rnatools). First, we produced sequence windows of 250 nt with an overlap of 75 nt covering 3823 plastid genomic sequence (seed plants), totalling ~5.6M windows. Windows were processed to remove sequences comprised of a single nucleotide type (e.g., A, T, G, or C), containing regions with unsequenced nucleotides (i.e., marked as 'N'), or redundant sequences (i.e., 100% identical). Sequences were then clustered using MMSeqs2 package (easy-cluster)[48] with 95% sequence identity to identify highly similar sequences, which were then removed with only one representative sequence retained per cluster. Representative sequences were further clustered with a minimum of 50% sequence identity, resulting in a total of 10431 clusters. Clustered sequences were aligned with Clustal Omega[83] and RNALalifold[49] was used to predict folding for each sequence, which was performed with temperature set to 21°C / 294.15K (-T option) and no lone pairs (--noLP) to identify sequences in a cluster that share a minimum of 2 base pairs. Approximately 30% of the clusters (3136/10431) passed this screening step. A progressive RNA structure prediction was performed using LocARNA package (mlocarna --probabilistic)[50] on the unaligned sequences obtained from MMSeqs2 with parameters set to --local-progressive and --rnafold-temperature=21°C. We identified 1827 clusters with structures sharing a minimum of 3 base pairs.

Alignments produced with LocARNA had a predefined length of 250 nt (i.e., input window size), hence, they were trimmed to only include the identified motif and additional 20 nt in each extremity using an in-house PERL program. Trimmed alignments were evaluated using our recently developed evaluator program, RNA-SCoRE (unpublished; available at https://github.com/RodrigoReisLab/RNA-SCoRE), with the following parameters: hairpin threshold (-t) of 0.75, motif threshold (--mt) of 0.5, GC/CG base-pair threshold (--gc) of 0.30, and removal of redundant sequences (-d 0). Only clusters with reliability defined as 'High' or 'Mid' were considered further. Cleaned alignments were converted into covariance models using cmbuild and cmcalibrate (Infernal v1.1.4)[43] against the 3823 plastid genomes from seed plants and then, used to find homologs with cmsearch and cmalign programs of Infernal v1.1.4. Output of Infernal were processed using previously described PERL scripts to reformat Infernal output[84]. RNA-SCoRE with the same parameters as above, except for --gc set to 0.15, was used to evaluate the identified homologs and a produce a new alignment including homolog that passed evaluation. Only clusters capable of finding homologs and reported as High or Mid by RNA-SCoRE were considered further. Covariation analysis was performed using R-scape[37] with two-tailed test and --cacofold to identify statistically

covarying base-pairs, indicative of conservation of structure beyond phylogeny, and any additional base-pairing missed out by LocARNA. Only the clusters with at least 2% covariation were considered as reliable alignments. The structures obtained with LocARNA and R-scape --cacofold were compared qualitatively and, if the predicted structures were completely different, the LocARNA predicted structures were considered, since it is the more biologically relevant structure. To further check the sequence-structure compatibility, nhmmer[51] was used with default parameters, for which the clustered sequences were de-aligned and re-aligned independently without structure information to correct for sequence misalignment, which may arise due to structure positions bias during LocARNA simultaneous fold and align. Structure consensus predicted with LocARNA were then added to nhmmer alignments and subjected to evaluation with R-scape (--cacofold), i.e., only alignments for which predicted structures with LocARNA and R-scape showed consistent overall architecture and at least one base pair covariation were considered reliable and high confidence predictions. This approach resulted in the discovery of 51 conserved motifs in plastid genomes. Alignments were then used as seed for each conserved motifs to obtain covariance models built with cmbuild and cmcalibrate programs of Infernal[43]. Covariance models were used against all 14034 plastid genomes available from NCBI with e-value 0.01 to search for homologs using cmsearch program of Infernal v1.1.4. The obtained hits were processed as previously described and further analysed using RNA-SCoRE (-t 0.75 --mt 0.5 --gc 0.15) to evaluate each identified homolog.

Our pipeline can be readily applied to any genome or set of sequences, including selective usage at any analysis step, composable, and adaptable to user needs.

Genomic context and phylogenetic analyses of conserved RNA structures

To identify the genomic context, RNA coordinates were mapped to gene annotation (*.gff3*) for the corresponding plastid genome using bedtools[85] closest with options -s (gene and the RNA are on the same strand) -D b (reports if genes are upstream or downstream to the RNAs). bedtools outputs were analysed using in-house PERL and BASH scripts and only RNAs within 500 nt of a gene were considered as homologs and further used to categorize RNA motifs. We employed the following broad categories for genome context: tRNA, 5' UTR, 3' UTR, and exon-intron junction. To obtain the phylogenetic distribution of each conserved RNA motif, NCBI taxonomic information (taxon ID and taxonomic lineage) were obtained using edirect utilities from provided NCBI genome IDs. The mapping was obtained for genera across 55 plant families from which a dataset compatible with iTOL[86] was generated containing percentage of genera within a plant family carrying RNA motifs. The dataset was mapped to the NCBI *Viridiplantae* tree in iTOL (https://itol.embl.de/) for visualization.

RNA structure search using Rfam database

Covariance models (.cm) of 4178 known RNA families were obtained from Rfam database (v15.0) and cmsearch (from Infernal v1.1.4 package)[43] with --ga threshold was performed against 14034 plastid genome sequences and any hits found were considered as RNA structure homologs.

References

1. Shih, P. M. & Matzke, N. J. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences* **110**, 12355–12360 (2013).
2. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences* **108**, 13624–13629 (2011).
3. Archibald, J. M. Genomic perspectives on the birth and spread of plastids. *Proceedings of the National Academy of Sciences* **112**, 10147–10153 (2015).
4. Tonti-Filippini, J., Nevill, P. G., Dixon, K. & Small, I. What can we do with 1000 plastid genomes? *The Plant Journal* **90**, 808–818 (2017).
5. Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L. & Jansen, R. K. Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. *J Mol Evol* **70**, 149–166 (2010).
6. Palmer, J. D. & Thompson, W. F. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* **29**, 537–550 (1982).
7. Wu, C.-S., Wang, Y.-N., Hsu, C.-Y., Lin, C.-P. & Chaw, S.-M. Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny. *Genome Biology and Evolution* **3**, 1284–1295 (2011).
8. Wicke, S. *et al.* Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. *Proceedings of the National Academy of Sciences* **113**, 9045–9050 (2016).
9. Smith, D. R. Mutation Rates in Plastid Genomes: They Are Lower than You Might Think. *Genome Biology and Evolution* **7**, 1227–1234 (2015).
10. Bobik, K. & Burch-Smith, T. M. Chloroplast signaling within, between and beyond cells. *Front. Plant Sci.* **6**, (2015).
11. Kirchhoff, H. Chloroplast ultrastructure in plants. *New Phytologist* **223**, 565–574 (2019).
12. Neuhaus, H. E. & Emes, M. J. Nonphotosynthetic metabolism in plastids. *Annual Review of Plant Biology* **51**, 111–140 (2000).
13. Choi, H., Yi, T. & Ha, S.-H. Diversity of Plastid Types and Their Interconversions. *Front. Plant Sci.* **12**, (2021).
14. Powell, A. L. T. *et al.* Uniform ripening Encodes a Golden 2-like Transcription Factor Regulating Tomato Fruit Chloroplast Development. *Science* **336**, 1711–1715 (2012).
15. Tanios, S., Eyles, A., Tegg, R. & Wilson, C. Potato Tuber Greening: a Review of Predisposing Factors, Management and Future Challenges. *Am. J. Potato Res.* **95**, 248–257 (2018).
16. Richter, A. S. *et al.* Retrograde signaling in plants: A critical review focusing on the GUN pathway and beyond. *Plant Communications* **4**, 100511 (2023).
17. Chan, K. X., Phua, S. Y., Crisp, P., McQuinn, R. & Pogson, B. J. Learning the Languages of the Chloroplast: Retrograde Signaling and Beyond. *Annu. Rev. Plant Biol.* **67**, 25–53 (2016).
18. Lee, K. P. & Kim, C. Photosynthetic ROS and retrograde signaling pathways. *New Phytologist* **244**, 1183–1198 (2024).
19. Martín, M. & Sabater, B. Plastid ndh genes in plant evolution. *Plant Physiology and Biochemistry* **48**, 636–645 (2010).

20. Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **76**, 273–297 (2011).
21. Guo, X. *et al.* Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* **18**, 176 (2017).
22. Vergara-Cruces, Á. *et al.* Structure of the plant plastid-encoded RNA polymerase. *Cell* **187**, 1145-1159.e21 (2024).
23. Wu, X.-X. *et al.* Cryo-EM structures of the plant plastid-encoded RNA polymerase. *Cell* **187**, 1127-1144.e21 (2024).
24. Graf, M. *et al.* Cryo-EM structure of the spinach chloroplast ribosome reveals the location of plastid-specific ribosomal proteins and extensions. *Nucleic Acids Research* **45**, 2887–2896 (2017).
25. Bieri, P., Leibundgut, M., Saurer, M., Boehringer, D. & Ban, N. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. *The EMBO Journal* **36**, 475–486 (2017).
26. Stern, D. B., Goldschmidt-Clermont, M. & Hanson, M. R. Chloroplast RNA Metabolism. *Annu. Rev. Plant Biol.* **61**, 125–155 (2010).
27. Zoschke, R. & Bock, R. Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. *Plant Cell* **30**, 745–770 (2018).
28. Lohan, A. J. & Wolfe, K. H. A Subset of Conserved tRNA Genes in Plastid DNA of Nongreen Plants. *Genetics* **150**, 425–433 (1998).
29. Tiller, N. & Bock, R. The Translational Apparatus of Plastids and Its Role in Plant Development. *Molecular Plant* **7**, 1105–1120 (2014).
30. Vogel, J. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *The EMBO Journal* **21**, 3794–3803 (2002).
31. Zhou, F., Karcher, D. & Bock, R. Identification of a plastid intercistronic expression element (IEE) facilitating the expression of stable translatable monocistronic mRNAs from operons. *The Plant Journal* **52**, 961–972 (2007).
32. Scharff, L. B., Childs, L., Walther, D. & Bock, R. Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites. *PLOS Genetics* **7**, e1002155 (2011).
33. Scharff, L. B. *et al.* Shine-Dalgarno Sequences Play an Essential Role in the Translation of Plastid mRNAs in Tobacco. *The Plant Cell* **29**, 3085–3101 (2017).
34. Prikryl, J., Rojas, M., Schuster, G. & Barkan, A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proceedings of the National Academy of Sciences* **108**, 415–420 (2011).
35. Hammani, K., Cook, W. B. & Barkan, A. RNA binding and RNA remodeling activities of the half-a-tetratricopeptide (HAT) protein HCF107 underlie its effects on gene expression. *Proceedings of the National Academy of Sciences* **109**, 5651–5656 (2012).
36. Stern, D. B. & Gruissem, W. Control of plastid gene expression: 3′ inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. *Cell* **51**, 1145–1157 (1987).
37. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods* **14**, 45–48 (2016).
38. Stav, S. *et al.* Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol* **19**, 66 (2019).
39. Brewer, K. I. *et al.* Comprehensive discovery of novel structured noncoding RNAs in 26 bacterial genomes. *RNA Biology* **18**, 2417–2432 (2021).
40. Gao, W., Jones, T. A. & Rivas, E. Discovery of 17 conserved structural RNAs in fungi. *Nucleic Acids Research* **49**, 6128–6143 (2021).
41. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences* **102**, 2454–2459 (2005).
42. Ontiveros-Palacios, N. *et al.* Rfam 15: RNA families database in 2025. *Nucleic Acids Research* **53**, D258–D267 (2025).
43. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
44. Dühring, U., Axmann, I. M., Hess, W. R. & Wilde, A. An internal antisense RNA regulates expression of the photosynthesis gene isiA. *Proceedings of the National Academy of Sciences* **103**, 7054–7058 (2006).

45. Baroux, C. & Grossniklaus, U. Chapter Twenty-One - Seeds—An evolutionary innovation underlying reproductive success in flowering plants. in *Current Topics in Developmental Biology* (ed. Grossniklaus, U.) vol. 131 605–642 (Academic Press, 2019).
46. Edwards, E. J. Evolutionary trajectories, accessibility and other metaphors: the case of C4 and CAM photosynthesis. *New Phytologist* **223**, 1742–1755 (2019).
47. Singh, B. P., Kumar, A., Kaur, H., Singh, H. & Nagpal, A. K. CpGDB□: A Comprehensive Database of Chloroplast Genomes. *Bioinformation* **16**, 171–175 (2020).
48. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
49. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
50. Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012).
51. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
52. Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLOS Computational Biology* **16**, e1008387 (2020).
53. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
54. Small, I., Melonek, J., Bohne, A.-V., Nickelsen, J. & Schmitz-Linneweber, C. Plant organellar RNA maturation. *The Plant Cell* **35**, 1727–1751 (2023).
55. Asakura, Y. & Barkan, A. A CRM Domain Protein Functions Dually in Group I and Group II Intron Splicing in Land Plant Chloroplasts. *Plant Cell* **19**, 3864–3875 (2007).
56. Lambowitz, A. M. & Zimmerly, S. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harb Perspect Biol* **3**, a003616 (2011).
57. Shukla, G. C. & Padgett, R. A. A Catalytically Active Group II Intron Domain 5 Can Function in the U12-Dependent Spliceosome. *Molecular Cell* **9**, 1145–1150 (2002).
58. Wang, X., Wang ,Jingyi, Li ,Simin, Lu ,Congming & and Sui, N. An overview of RNA splicing and functioning of splicing factors in land plant chloroplasts. *RNA Biology* **19**, 897–907 (2022).
59. Korotkova, N. *et al.* Phylogeny of the eudicot order Malpighiales: analysis of a recalcitrant clade with sequences of the petD group II intron. *Plant Syst Evol* **282**, 201–228 (2009).
60. Borsch, T., Korotkova, N., Raus, T., Lobin, W. & Löhne, C. The petD group II intron as a species level marker: utility for tree inference and species identification in the diverse genus Campanula (Campanulaceae). *will* **39**, 7–33 (2009).
61. Asakura, Y., Bayraktar, O. A. & Barkan, A. Two CRM protein subfamilies cooperate in the splicing of group IIB introns in chloroplasts. *RNA* **14**, 2319–2332 (2008).
62. An, H. *et al.* ALBINO EMBRYO AND SEEDLING is required for RNA splicing and chloroplast homeostasis in Arabidopsis. *Plant Physiology* **193**, 483–501 (2023).
63. Inaba-Hasegawa, K., Ohmura, A., Nomura, M. & Sugiura, M. Development of an In Vitro Chloroplast Splicing System: Sequences Required for Correct pre-mRNA Splicing. *Plant and Cell Physiology* **62**, 1311–1320 (2021).
64. Cannone, J. J. *et al.* The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2 (2002).
65. Duffy, A. M., Kelchner, S. A. & Wolf, P. G. Conservation of selection on *matK* following an ancient loss of its flanking intron. *Gene* **438**, 17–25 (2009).
66. Adams, C. C. & Stern, D. B. Control of mRNA stability in chloroplasts by 3′ inverted repeats: effects of stem and loop mutations on degradation of psbA mRNA in vitro. *Nucleic Acids Research* **18**, 6003–6010 (1990).
67. Zandueta-Criado, A. Surprising features of plastid ndhD transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. *Nucleic Acids Research* **32**, 542–550 (2004).
68. Kim, K.-J. & Lee, H.-L. Widespread Occurrence of Small Inversions in the Chloroplast Genomes of Land Plants. *Molecules and Cells* **19**, 104–113 (2005).
69. Jo, S., Kim, Y.-K., Cheon, S.-H., Fan, Q. & Kim, K.-J. Characterization of 20 complete plastomes from the tribe Laureae (Lauraceae) and distribution of small inversions. *PLOS ONE* **14**, e0224622 (2019).
70. Axtell, M. J. & Bowman, J. L. Evolution of plant microRNAs and their targets. *Trends in plant science* **13**, 343–349 (2008).
71. Becskei, A. & Rahaman, S. The life and death of RNA across temperatures. *Computational and Structural Biotechnology Journal* **20**, 4325–4336 (2022).

72. Zhelyazkova, P. *et al.* Protein-mediated protection as the predominant mechanism for defining processed mRNA termini in land plant chloroplasts. *Nucleic Acids Research* **40**, 3092–3105 (2012).
73. Zerges, W., Auchincloss, A. H. & Rochaix, J.-D. Multiple Translational Control Sequences in the 5′ Leader of the Chloroplast psbC mRNA Interact With Nuclear Gene Products in Chlamydomonas reinhardtii. *Genetics* **163**, 895–904 (2003).
74. Hirose, T. & Sugiura, M. Multiple elements required for translation of plastid atpB mRNA lacking the Shine-Dalgarno sequence. *Nucleic Acids Research* **32**, 3503–3510 (2004).
75. Higgs, D. C., Shapiro ,Risa S., Kindle ,Karen L. & and Stern, D. B. Small cis-Acting Sequences That Specify Secondary Structures in a Chloroplast mRNA Are Essential for RNA Stability and Translation. *Molecular and Cellular Biology* **19**, 8479–8491 (1999).
76. Mayfield, S. P., Cohen, A., Danon, A. & Yohn, C. B. Translation of the psbA mRNA of Chlamydomonas reinhardtii requires a structured RNA element contained within the 5' untranslated region. *Journal of Cell Biology* **127**, 1537–1545 (1994).
77. Kuroda, H. & Sugiura, M. Processing of the 5'-UTR and existence of protein factors that regulate translation of tobacco chloroplast psbN mRNA. *Plant Mol Biol* **86**, 585–593 (2014).
78. Zhelyazkova, P. *et al.* The Primary Transcriptome of Barley Chloroplasts: Numerous Noncoding RNAs and the Dominating Role of the Plastid-Encoded RNA Polymerase. *The Plant Cell* **24**, 123–136 (2012).
79. Gawroński, P., Pałac, A. & Scharff, L. B. Secondary Structure of Chloroplast mRNAs In Vivo and In Vitro. *Plants* **9**, 323 (2020).
80. Okui, S., Ushida, C., Kiyosawa, H. & Kawai, G. Sequence and structure analysis of a mirror tRNA located upstream of the cytochrome oxidase I mRNA in mouse mitochondria. *The Journal of Biochemistry* **159**, 341–350 (2016).
81. Fleischmann, T. T. *et al.* Nonessential Plastid-Encoded Ribosomal Proteins in Tobacco: A Developmental Role for Plastid Translation and Implications for Reductive Genome Evolution[W][OA]. *Plant Cell* **23**, 3137–3155 (2011).
82. Daniell, H., Lin, C.-S., Yu, M. & Chang, W.-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* **17**, 134 (2016).
83. Sievers, F. & Higgins, D. G. The Clustal Omega Multiple Alignment Package. in *Multiple Sequence Alignment: Methods and Protocols* (ed. Katoh, K.) 3–16 (Springer US, New York, NY, 2021). doi:10.1007/978-1-0716-1036-7_1.
84. Predicting Riboswitch Regulation on a Genomic Scale | SpringerLink. https://link.springer.com/protocol/10.1007/978-1-59745-558-9_1.
85. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
86. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **52**, W78–W82 (2024).

Acknowledgements

The authors thank Elena Rivas (Harvard University) for valuable discussions and feedback. We also thank Michael Crawford and Rachappa Balkunde (Bayer Crop Science) for guidance on applicability for crop protection. This work was supported by the Swiss National Science Foundation (PCEFP3_203328) and the Grants4Ag Program (Bayer Crop Science).

Author contributions

RSR and DM conceived and designed this work. DM performed all bioinformatic analyses. JH contributed to preliminary analyses. RSR and DM wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

The pipeline developed in this work is available from GitHub https://github.com/RodrigoReisLab/RNA_conservation_chloroplast) and tools used in the pipeline are available as a container at Docker Hub Container Image Library (https://hub.docker.com/r/dollycm/rnatools). RNA-SCoRE is available from GitHub https://github.com/RodrigoReisLab/RNA-SCoRE). All novel conserved RNA motifs have been submitted to Rfam (identifiers to be provided). For correspondence, contact RSR (rodrigo.reis@unibe.ch).

Figure legend

Figure 1. Identification of conserved RNA structure in plastid genomes. **(a)** Pipeline to predict and evaluate conservation of secondary structure in plastid genomes. **(b)** Bar plot showing the cluster number at each pipeline step (left), resulting in the identification of 51 high confidence conserved RNA motifs. Pie chart showing the genomic context for the conserved RNA motifs (right). **(c)** Consensus structure for tRNA obtained from Rfam database (left) and this work (right). Covarying base pairs (E-value ≤ 0.05) are highlighted in green. **(d)** Distribution of conserved RNA motifs in the phylogenetic tree of Viridiplantae. Circle size (outer layer) represents the percentage distribution of RNA motif in given plant clade. Grey circle represents previously known RNA motifs (available in Rfam) and yellow represents novel motifs (not in Rfam).

Figure 2. Conserved RNA motifs in group-II intron. **(a)** Simplified model for group II intron showing the typical six domains (D1-6) and key nucleotides involved in splicing. **(b)** Consensus structure for domain 5 (D5) available in Rfam (Clan CL00102). Highly conserved GNRA loop and AY bulge highlighted with red line. **(c)** Consensus structure for two motifs identified in this work mapping to intronic regions. Conserved D5 with GNRA loop and AY bulge highlighted with red line. **(d)** Consensus structure for motif identified in petD intron resembling D5, except for variation in loop and bulge sequence identity. **(e)** Consensus motif for D1 in atpF involved in tertiary interactions. EBS1, exon-binding site 1.

Figure 3. Conserved structure in 3' terminal region. **(a)** Consensus structure for representative motifs identified in the 3' end of given genes. **(b)** Violin plot for GC content percentage of sequences within each model (one-way ANOVA, p-value < 0.0001). **(c)** Consensus structure for a short motif in the 3' end of psbI, ycf3, and chlN.

Figure 4. Conserved structure in the 5' untranslated region. Consensus structure for motif identified in the 5' UTR of psbN **(a)**, ycf2 **(b)**, and rpl22 **(c)**.

Statistical information

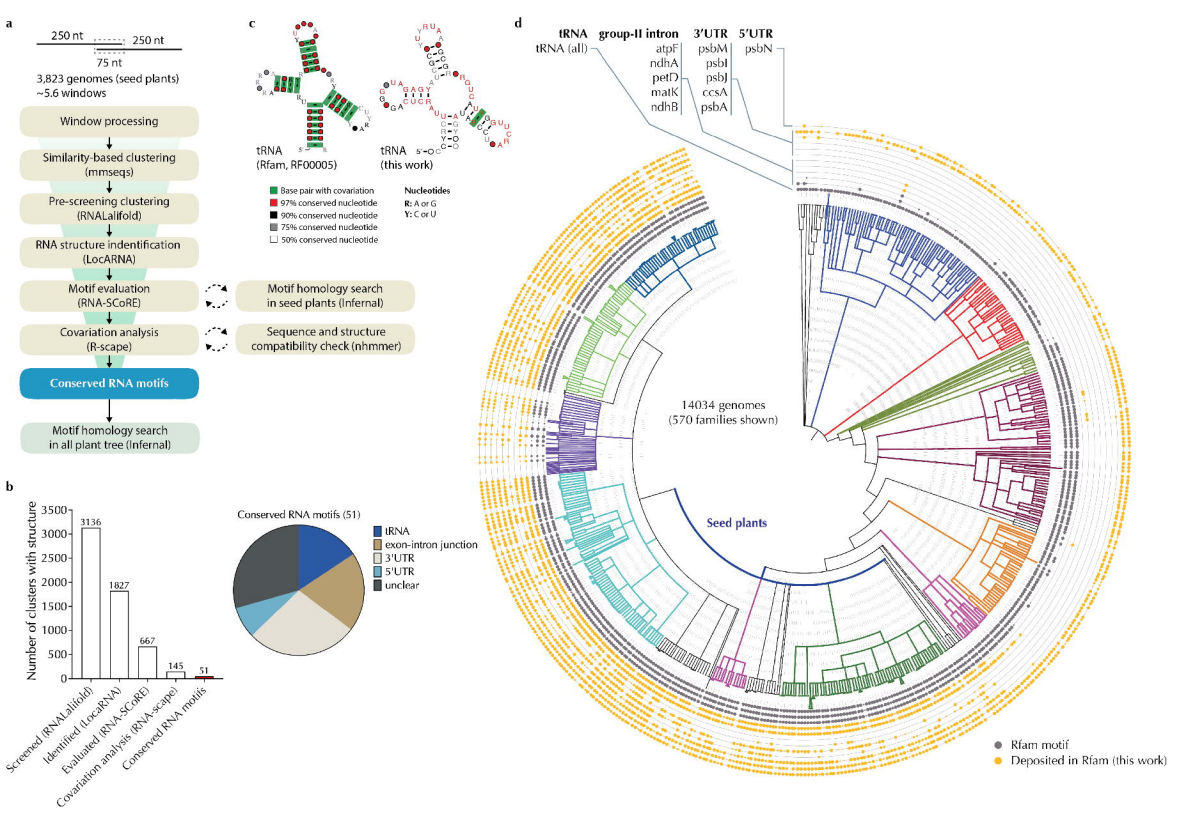
Statistical analyses were performed by computational packages for RNA processing, as described in Methods.

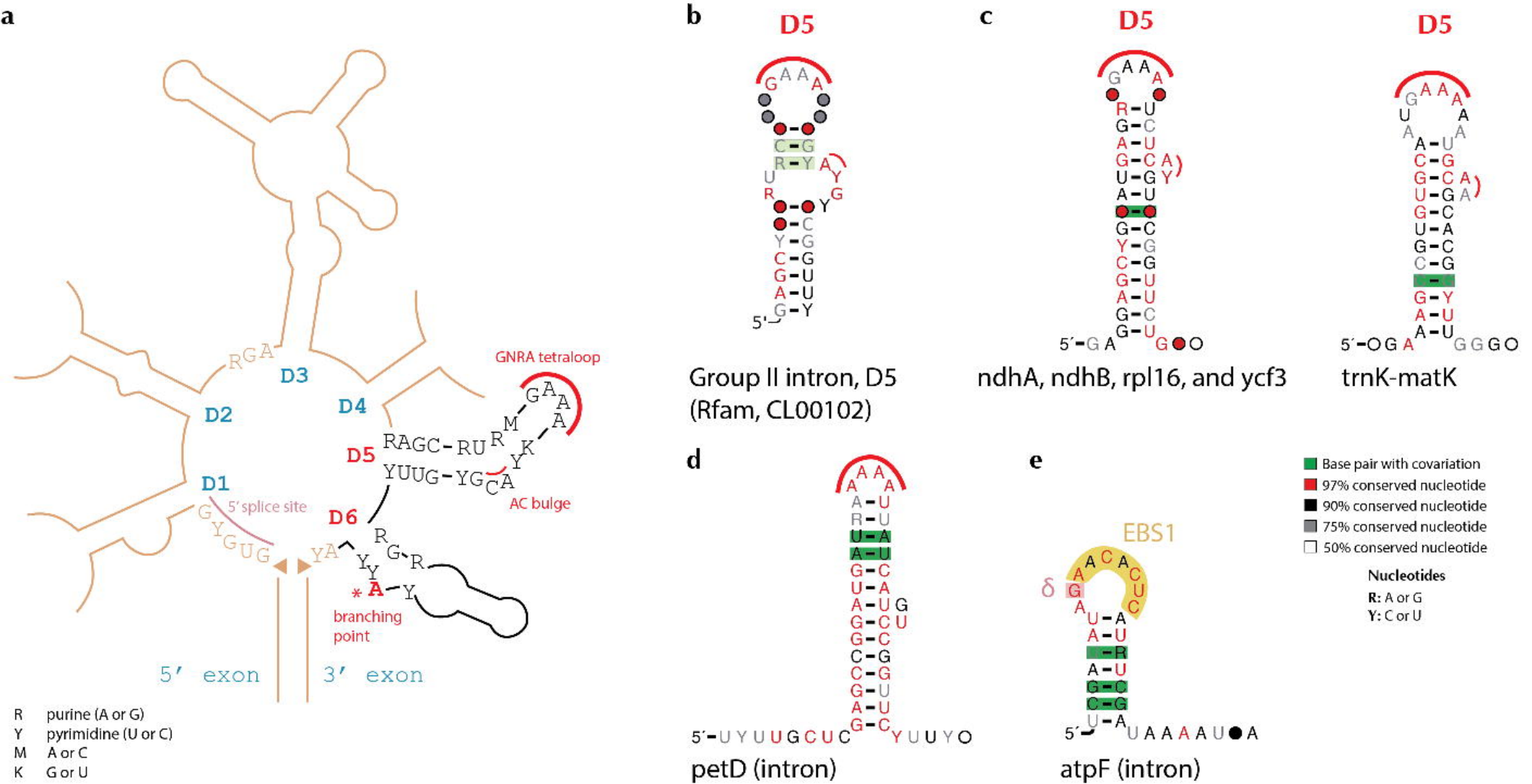
Supplementary information

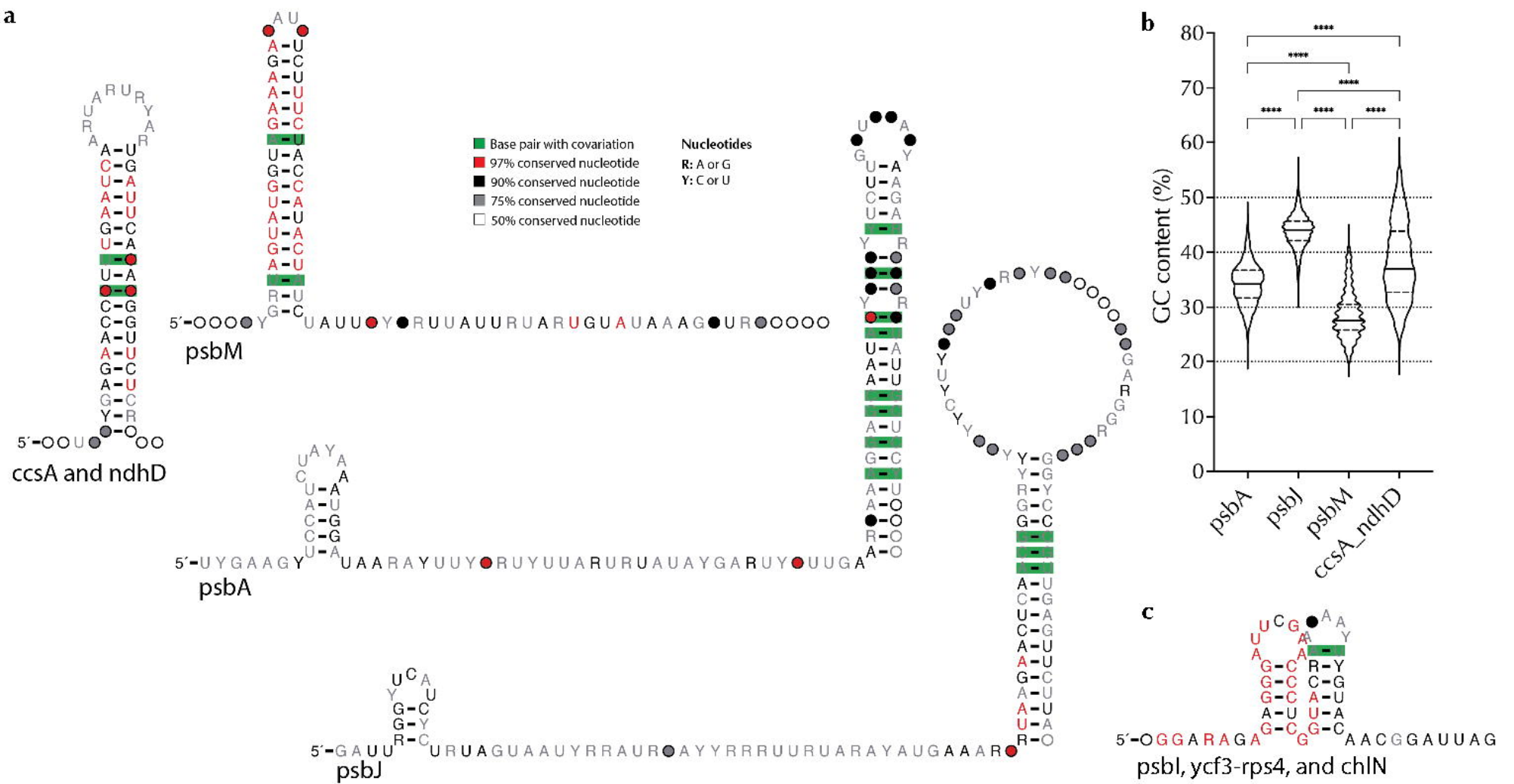
Supplementary Table 1.

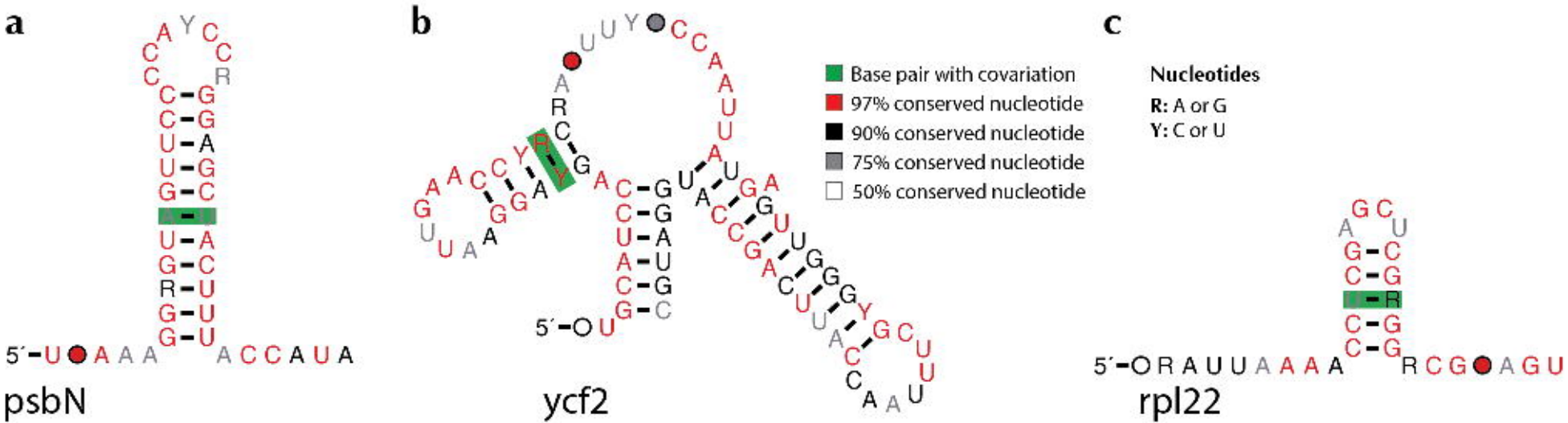
Data availability

All novel conserved RNA motifs have been submitted to Rfam (identifiers to be provided).


a
250 nt
250 nt
75 nt
3,823 genomes (seed plants)
~5.6 windows
Window processing
Similarity-based clustering (mmseqs)
Pre-screening clustering (RNALalifold)
RNA structure indentification (LocARNA)
Motif evaluation (RNA-SCoRE)
Covariation analysis (R-scape)
Conserved RNA motifs
Motif homology search in all plant tree (Infernal)
Motif homology search in seed plants (Infernal)
Sequence and structure compatibility check (nhmmer)
b
Number of clusters with structure
3136
1827
667
145
51
Screened (RNALalifold)
Identified (LocARNA)
Evaluated (RNA-SCoRE)
Covariation analysis (RNA-scape)
Conserved RNA motifs
Conserved RNA motifs (51)
tRNA
exon-intron junction
3'UTR
5'UTR
unclear
c
tRNA (Rfam, RF00005)
tRNA (this work)
Base pair with covariation
97% conserved nucleotide
90% conserved nucleotide
75% conserved nucleotide
50% conserved nucleotide
Nucleotides
R: A or G
Y: C or U
d
tRNA
group-II intron
3'UTR
5'UTR
tRNA (all)
atpF
ndhA
petD
matK
ndhB
psbM
psbI
psbJ
ccsA
psbA
psbN
14034 genomes
(570 families shown)
Seed plants
Rfam motif
Deposited in Rfam (this work)


a
D1
D2
D3
D4
D5
D6
GNRA tetraloop
AC bulge
5′ splice site
branching point
5′ exon
3′ exon
R purine (A or G)
Y pyrimidine (U or C)
M A or C
K G or U
b
D5
Group II intron, D5
(Rfam, CL00102)
c
D5
ndhA, ndhB, rpl16, and ycf3
D5
trnK-matK
d
petD (intron)
e
EBS1
δ
atpF (intron)
Base pair with covariation
97% conserved nucleotide
90% conserved nucleotide
75% conserved nucleotide
50% conserved nucleotide
Nucleotides
R: A or G
Y: C or U


a
Base pair with covariation
97% conserved nucleotide
90% conserved nucleotide
75% conserved nucleotide
50% conserved nucleotide
Nucleotides
R: A or G
Y: C or U
ccsA and ndhD
psbM
psbA
psbJ
psbI, ycf3-rps4, and chlN
b
GC content (%)
psbA
psbJ
psbM
ccsA_ndhD
c


a
psbN
b
ycf2
c
rpl22
Base pair with covariation
97% conserved nucleotide
90% conserved nucleotide
75% conserved nucleotide
50% conserved nucleotide
Nucleotides
R: A or G
Y: C or U