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 structure 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 57 conserved RNA motifs. Beyond recovering known functional classes, we uncovered *cis*-acting novel structures within the UTRs and introns of key photosynthetic genes, including psbN, clpP, and atpF, as well as *trans*-acting antisense RNAs to petB and psbT, pointing to previously unknown regulatory elements. 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 (https://github.com/RodrigoReisLab/RNA-SCoRE). 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 57 RNA motifs conserved across ~14k plastid genomes, including known functional motifs (e.g., 16S rRNA, tRNA, group II intron, and 3′ end stem-loop) and novel uncharacterized motifs, such as within the UTR of psbI, rps4, rps18, clpP, rpoB, and psbN, and intronic region of atpF and rpl16. In addition to *cis*-acting RNA motifs, we identified conserved structure in antisense RNAs to petB, petD, rps16, and psbT-psbB. Our work establishes a framework for reliable identification of conserved RNA structures 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[42,43]. A powerful approach is to leverage the existing repertoire of conserved RNA structures available from the RNA family (Rfam) database[44]. For this, we used the cmsearch program of Infernal[45] 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 (Supplementary Table 1). 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)[46]. 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[47]. Their current success involved diversification of photosynthetic strategies to support environmental adaptation, such as the evolution of C3, C4, and CAM photosynthesis[48]. We, thus, obtained 3823 manually curated plastid genomes from seed plants[49], 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). We removed redundant windows and produced clusters with a minimum of 10 windows and 50% sequence identity (MMSeqs2[50]), resulting in >10,000 clusters. A pre-screening was performed with the fast-performing RNALalifold[51], which calculates locally stable secondary structures, to identify clusters with structures

sharing at least 2 base pairs. This resulted in 3136 clusters, for which each sequence was refolded and realigned using the probabilistic mode of LocARNA[52], producing 1827 clusters where sequences share secondary structure (Fig. 1b). 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 (https://github.com/RodrigoReisLab/RNA-SCoRE). RNA-SCoRE evaluation resulted in 667 clusters (out of 1827) with a minimum of 7 sequences. We used these cleaned alignment as seeds to find homolog sequences using Infernal[45], which were then processed to obtain clusters including the new hits. Sequences that passed the RNA-SCoRE evaluation were retained, thus producing improved alignments with better performance as seed. We then evaluated these alignments 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[53] and CaCoFold[54] to realign the sequences without structure information and evaluate the LocARNA consensus structure, respectively, and, then, retained only the alignments that support the corresponding predicted structure by LocARNA. Out of 145 clusters, 76 passed this evaluation, thus, corresponding to RNA motifs with covariation for which we established high confidence predicted structure and high alignment reliability. We then used the multiple sequence alignment for each conserved RNA motif as covariance models to search for structure homologs across ~14k plastid genomes, from non-seed plants to evolutionarily distant unicellular photosynthetic organisms (Fig. 1a). We reasoned that conserved structure displaying consistent context (e.g., similar position in a gene) across plastid genomes is more likely to be a true conserved regulatory element. Therefore, we mapped the genomic context for each motif and identified 57 motifs (out of 76) with regulatory potential, while the remaining 19 motifs were intergenic without readily identifiable putative function. We thus focused on those 57 conserved motifs with regulatory potential, hereafter, conserved RNA motifs in plastids (Fig. 1a-b). Most of these motifs (~85%) are associated with noncoding (rRNA, tRNAs, and antisense RNA) or coding (UTR, exon-intron junctions, and CDS) genes, while genic and functional association for ~15% is unclear (Fig. 1c).

tRNA function is based on folding into specific conformation and it is, therefore, expected to display covariation in our analysis. We obtained six models mapping to annotated plastid tRNAs, all of which also identified using the Rfam model for tRNA (RF00005) (Fig. 1d, Supplementary Fig. 1a). The Rfam model displays 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[55]. While two of our tRNA models (model1 and model4) mapped to different types of tRNAs, the other four models showed amino acid specificity (Fig. 1e), indicating that our pipeline captures functionally related structures with relatively subtle differences that are important for function. We also identified a conserved 16S rRNA structure motif mapping to bacterial helix 8, matching the Rfam model for 16S rRNA RF00177 (Supplementary Fig. 1b and Supplementary Table 1). Our models identified tRNA and 16S rRNA in plastids across the entire plant phylogenetic tree with high sensitivity, i.e. covering over 99% of plant species compared to Rfam models (Fig. 2), evidencing that our pipeline for the identification of conserved structures across genomes is effective and reliable.

Conservation in motifs within 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[56]. 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[56]. In land plants, ~20 group II introns and one group I intron have been reported in plastid genomes[57]. Here, we identified conserved structures in group II intron mapping to NADH dehydrogenase (ndhA and ndhB), ribosome subunit (rpl16 and rps12), photosystem I assembly (ycf3), photosystem II (petB and petD), ATP synthase (atpF) and tRNA (trnK and trnG) (Supplementary Fig. 2a).

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. 3a). 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[58]. 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. 3a) 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. 3b), indicating an evolutionary link between self-spliced group II introns and splicing in eukaryotes[59]. 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. 3a), in which the adenosine's 2'-hydroxyl group initiates the first transesterification reaction of splicing.

Our analysis identified conserved D5 represented by five models mapping to ndhA, ndhB, petB, atpF, rps12, rpl16, rpoC1, and ycf3 (Fig. 3c and Supplementary Fig. 2b), and another two different models mapping to D5 in trnK (Fig. 3c and Supplementary Fig. 2c). We also found a conserved motif in petD that resembles D5 (Fig. 3d); however, this motif diverges in the loop (AAAA[60,61], 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[62], suggesting that the differences in D5 between petD and aforementioned transcripts (Fig. 3c) 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 (petB and petD), and null mutation in ALBINO EMBRYO AND SEEDLING (AES) gene results in similar intron retention in petB and petD[63]. All mapped genes are known to harbour group II intron[64], and our findings indicate strong structural conservation for D5 and regulatory function associated with a D5 variant.

We also identified a conserved structure in the group II intron domain D1 of atpF, ycf3, rpl16, and clpP and D3 of trnK (Fig. 3e and Supplementary Fig. 3d-e). 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')[58]. 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[65] (Fig. 3e). EBS1–IBS1 pairing has been shown to be essential for splicing, whereas EBS2–IBS2 pairing is necessary for efficient splicing but is not required[65]. 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. We also found a structure in the 5' end of rpl16 as part of the group II intron D1, albeit it includes EBS2 binding site as previously reported for Japanese black pine (*Pinus thunbergii*)[66], instead of EBS1 found in the atpF motif (Fig. 3f). We found another motif in the group II intron D1 in ycf3 and clpP (Fig. 3g) based on its sequence position within the intron, while its position within the structured D1 is unclear. Furthermore, we identified a motif in trnK likely

part of the group II intron D3 (Supplementary Fig. 2e), as compared with a trnK intron sequence in *Nicotiana tabacum* and relatively mapped to D3 of group II in trnK intron (RNAcentral: URS0002349F48/4097)[67]. Altogether, we produced conservation models for group II intron D5, similar to those already available in Rfam, but also models for motif within D1 and, possibly, D3 that have not been previously described in Rfam.

Conservation in 3′-termini processing motif

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 13 motifs representing conserved terminal IR motifs associated with genes involved in photosystem II (psbA, psbM, and psbJ), cytochrome c biogenesis (ccsA), and NDH complex (ndhD) (Fig. 4a and Supplementary Fig. 3a-c). These conservation models are mainly associated with three genes and, given that they identify corresponding motif across the phylogenetic tree, the models were combined into three clans: psbA, psbJ, and ccsA-ndhD (Supplementary Fig. 3d). IR-dependent RNA maturation and stability have been described for psbA[36,68] and ndhD[69], 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[70,71], 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; however, we only identified covariation in the stem-loops involved in 3′ processing.

We found the conserved motif in psbM across the phylogenetic tree, including in plastid of early flowering plant species such as *Amborella trichopoda*. There is evidence for widespread IR formation in 3′ ends via small inversions in plastid genomes[70], 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[72]. Stem-loop stability typically shows positive correlation with GC content[73]. Hence, we assessed GC content distribution in aligned sequences for psbA, psbM, psbJ, ccsA, and ndhD representative covariation models (Fig. 4b). We observed a broad range of median GC content, from 26% in psbM to 44% in psbJ, with relatively tight spread in the psbJ motif, suggesting that its higher GC content has been under evolutionary pressure. Selection of GC content in the psbM motif, on the other hand, appears weak and permissive to stem-loop

with likely poor stability. 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. 4c). This motif is found in the 3′ of psbI and rps4 in most seed plants, but also in the 3′ of ycf3 is some seed species and, in fewer cases, in the 3′ of chlN in evolutionarily distant species, such as single-cell green alga *Chlamydomonas reinhardtii*. 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[74], albeit no detailed characterisation was performed for this particular motif. Moreover, we identified several other shorter stem-loop motifs in the 3' UTR of ndhJ, rps18, rpl20, and rps12, where the rpl20 and rps12 are formed by tandem stem-loops, indicating a role in transcript maturation likely different from the longer, canonical IRs (Supplementary Fig. 3e).

Conservation of putative transcription terminator motif

Generally, plastid transcripts obtain a defined 3′ termini via post-transcriptional RNA processing[74], while, transcription termination is commonly defined by stem-loop formation in nascent RNA in bacteria, such as rho-independent termination[75]. We identified a conserved stem-loop motif that can be formed in both strands at the intergenic region between petD and rpoA operons expressed from the sense and anti-sense strand, respectively (Fig 4d-e). Nearly 40 years ago, a report on spinach plastid showed that the conserved stem-loop motif identified here might be important for transcription termination[76], possibly helping avoid RNA polymerase collision during transcription of both strands, as shown for bacteria[77]. However, transcriptional termination via RNA polymerase stalling by stem-loop structure in nascent transcript has not been shown for plastids yet.

Conservation in 5′ UTR motifs

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[78–81]. Here, we identified a conserved stem-loop (11-nt in length) in the psbN 5' UTR (Fig. 5a), 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[82]. However, the identified conserved motif is often within 80-250nt (Supplementary Fig. 4a), 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 5' UTR of mature psbN transcript, but only in a subset of transcripts[83], 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 large 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. 5b). 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, and the tRNA-like and ycf2 5' UTR association 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[84]. Similarly, in mitochondria, tRNA-induced RNase cleavage has been shown to regulate maturation of transcripts containing tRNA-like structure[85]. 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. In addition, we identified a novel conserved motifs in the 5' UTR of a protease complex subunit (clpP) and a chloroplast-encoded RNA polymerase subunit (rpoB) (Supplementary Fig. 4b), for which we are also unaware of described function. Altogether, we have identified three novel conserved motifs located in the 5' UTR of diverse transcripts, likely involved in post-transcriptional regulation.

Conservation in motifs within coding region and antisense noncoding RNA

The large ribosomal subunit rpl22 is essential in most seed plants[86]. 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. 5c), as well as a similar stem-loop identified in the rpl2 CDS (Supplementary Fig. 4c). rpl22 and rpl2 genes are often part of a large operon with other ribosomal protein genes (e.g., rps19, rpl23, and rps3), for which maturation requires activity of endonuclease at multiple sites[87], while splicing and editing seems either less prevalent or non-existing in rpl2 and rpl22 across plant species. Given the tight regulation of ribosomal subunits at multiple levels, it is possible that the conserved motif within rpl2 and rpl22 CDS play a regulatory role possibly by interacting with specific proteins.

We identified a novel conserved motif in the psaA CDS, near the 3' (Supplementary Fig. 4c), and another in psbT CDS (Fig 5d), both of which yet uncharacterized for their function. psbT is part of the psbB operon, which has psbN on the opposite genomic strand situated between psbT and psbH genes. It has been shown that psbT expression is controlled by two antisense RNAs originated from psbN transcription by the nuclear-encoded polymerase (NEP), in which the short antisense RNA hybridizes to psbT CDS, thereby stabilizing psbT transcript while blocking its translation during photooxidative stress[88]. It is thus possible that the conserved motif in psbT CDS is involved in the regulation by sense-antisense interaction with the psbN-derived noncoding RNA. Interestingly, we identified a conserved motif mapping to the long, but not the short antisense RNA derived from psbN transcription (Fig 5d), suggesting a complex and fine-tuned regulation of psbT and, potentially the entire psbB operon, involving RNA motifs conserved in seed plants, fern, moss, and green algae (Supplementary Fig. 4d).

We identified another conserved motif antisense to a coding gene; however, in the opposite strand to group II intron D5 and D6 in petB, petD, rps16 (Fig. 5e). In barley, RNA was also found in the antisense region of petB intron[83], suggesting a regulatory function for conserved motif in such noncoding antisense RNAs. Furthermore, beyond the conserved motifs associated with genic context, we identified eight conserved motifs likely associated with noncoding RNA for which evidence for transcription is unclear (Supplementary Fig. 5). It is thus possible that various functional noncoding RNAs still remain undiscovered in plastids.

Conclusion

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 57 conserved RNA motifs, most of which associated with genomic context, and still not reported in Rfam. We successfully validated our discovery pipeline by the recovery of various well-characterized RNA structures, including rRNA, tRNAs, group II intron domain, and 3' end processing stem-loops. In addition, we discovered novel and uncharacterized motifs, such as those within the 5' UTR of psbN, coding sequence of psbT and rpl22, intronic regions of atpF, rpl16, ycf3, and clpP, and anti-sense RNAs to psbT, petB, petD, and rps16. These findings significantly expand the known repertoire of putative *cis*- and *trans*-regulatory RNA elements, evidencing that conserved RNA structure is a pervasive feature of plastid gene

regulation. The functional significance of novel conserved motifs will require experimental investigation, ideally, involving multiple and evolutionarily distinct plant species.

Method

Plastid genomic sequences

An initial list of 3823 plastid genomes from seed plants (i.e., spermatophyta) was obtained from the CpGDB database[49]. This list was complemented with plastid genomes obtained directly from NCBI (July 11, 2024) 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) and 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)[50] 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[89] and RNALalifold[51] 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)[52] 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 and available on GitHub: 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)[45] 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[90]. 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 significant 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[53] was used with default parameters, for which the clustered sequences were de-aligned and are 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 CaCoFold showed consistent overall architecture and at least one base-pair with statistically significant covariation were considered reliable and high confidence predictions. This approach resulted in the discovery of 57 conserved motifs in plastid genomes. CaCoFold alignments were then used as seed for each conserved motifs to obtain covariance models built with cmbuild and cmcalibrate programs of Infernal[45]. Covariance models were used against all 14034 plastid genomes 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[91] 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. Additional curation was carried out for tRNA motif homologs with bitscore threshold used was half of the highest scoring RNA homolog and with a distance of 0 (overlapping with genes). If multiple motif map to the same RNA, the highest scoring RNA is retained as a homolog. 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 586 plant families from which a dataset compatible with iTOL[92] 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)[45] 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. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLOS Computational Biology* **3**, e65 (2007).
43. Ziesel, A. & Jabbari, H. Unveiling hidden structural patterns in the SARS-CoV-2 genome: Computational insights and comparative analysis. *PLOS ONE* **19**, e0298164 (2024).
44. Ontiveros-Palacios, N. *et al.* Rfam 15: RNA families database in 2025. *Nucleic Acids Research* **53**, D258–D267 (2025).
45. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
46. 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).
47. 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).
48. Edwards, E. J. Evolutionary trajectories, accessibility and other metaphors: the case of C4 and CAM photosynthesis. *New Phytologist* **223**, 1742–1755 (2019).
49. Singh, B. P., Kumar, A., Kaur, H., Singh, H. & Nagpal, A. K. CpGDB□: A Comprehensive Database of Chloroplast Genomes. *Bioinformation* **16**, 171–175 (2020).
50. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

51. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
52. 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).
53. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
54. Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLOS Computational Biology* **16**, e1008387 (2020).
55. 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).
56. Small, I., Melonek, J., Bohne, A.-V., Nickelsen, J. & Schmitz-Linneweber, C. Plant organellar RNA maturation. *The Plant Cell* **35**, 1727–1751 (2023).
57. 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).
58. Lambowitz, A. M. & Zimmerly, S. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harb Perspect Biol* **3**, a003616 (2011).
59. 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).
60. 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).
61. 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).
62. 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).
63. An, H. *et al.* ALBINO EMBRYO AND SEEDLING is required for RNA splicing and chloroplast homeostasis in Arabidopsis. *Plant Physiology* **193**, 483–501 (2023).
64. 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).
65. 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).
66. Kelchner, S. A. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *American Journal of Botany* **89**, 1651–1669 (2002).
67. 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).
68. 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).
69. 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).
70. 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).
71. 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).
72. Axtell, M. J. & Bowman, J. L. Evolution of plant microRNAs and their targets. *Trends in plant science* **13**, 343–349 (2008).
73. Becskei, A. & Rahaman, S. The life and death of RNA across temperatures. *Computational and Structural Biotechnology Journal* **20**, 4325–4336 (2022).

74. 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).
75. Chen, J., Morita, T. & Gottesman, S. Regulation of Transcription Termination of Small RNAs and by Small RNAs: Molecular Mechanisms and Biological Functions. *Front. Cell. Infect. Microbiol.* **9**, (2019).
76. Sijben-Müller, G., Hallick, R. B., Alt, J., Westhoff, P. & Herrmnann, R. G. Spinach plastid genes coding for initiation factor IF-1, ribosomal protein S11 and RNA polymerase α -subunit. *Nucleic Acids Res* **14**, 1029–1044 (1986).
77. Wang, L., Watters, J. W., Ju, X., Lu, G. & Liu, S. Head-on and co-directional RNA polymerase collisions orchestrate bidirectional transcription termination. *Molecular Cell* **83**, 1153-1164.e4 (2023).
78. 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).
79. 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).
80. 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).
81. 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).
82. 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).
83. 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).
84. Gawroński, P., Pałac, A. & Scharff, L. B. Secondary Structure of Chloroplast mRNAs In Vivo and In Vitro. *Plants* **9**, 323 (2020).
85. 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).
86. 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).
87. Daniell, H., Lin, C.-S., Yu, M. & Chang, W.-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* **17**, 134 (2016).
88. Zghidi-Abouzid, O., Merendino, L., Buhr, F., Malik Ghulam, M. & Lerbs-Mache, S. Characterization of plastid psbT sense and antisense RNAs. *Nucleic Acids Res* **39**, 5379–5387 (2011).
89. 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.
90. Predicting Riboswitch Regulation on a Genomic Scale | SpringerLink. https://link.springer.com/protocol/10.1007/978-1-59745-558-9_1.
91. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
92. 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 initial experiments and planning. 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 (github.com/RodrigoReisLab/RNA_conservation) and the 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 number of clusters at each pipeline step, resulting in the identification of 57 high confidence conserved RNA motifs (red bar). **(c)** Pie chart showing the genomic context for the conserved RNA motifs. **(d)** Consensus structure for tRNA obtained from Rfam database (left) and this work (right). Covarying base pairs (e-value ≤ 0.05) are highlighted in green, as shown in the legend. **(e)** Stacked bar plot showing the number of tRNAs identified across 14034 plastid genomes with six tRNA models.

Figure 2. 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 (coloured branches). Grey circle represents RNA motifs available in Rfam (i.e., known

motifs) and yellow represents novel conserved motifs. RNA motifs are grouped based on genomic context and cellular functions. Seed plants are highlighted in blue.

Figure 3. 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 representative structure model for 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, frequently observed in ndhA, ndhB, petB, atpF, and ycf3 introns, while AA bulge was found in trnK intron. **(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. **(f)** Consensus structure for D1 in rpl16 highlighting the conserved EBS2. **(g)** Consensus structure for D1 in ycf3 identified based on sequence position.

Figure 4. Conserved structure in 3' terminal region. **(a)** Consensus structure for representative motifs involved in 3' termini processing. **(b)** Violin plot for GC content percentage of sequences within each model shown in **(a)** (one-way ANOVA, p-value < 0.0001). **(c)** Consensus structure for a short tandem motif in the 3' end of psbI, ycf3, and chlN. **(d)** Consensus structure for a putative transcription terminator. **(e)** Genomic context of putative transcription terminator **(d)** in the *Arabidopsis thaliana* plastid.

Figure 5. Conserved structure in 5' untranslated region (UTR), coding sequence and antisense RNA. Consensus structure for motif identified in the 5' UTR of psbN **(a)** and ycf2 **(b)**, coding sequence of rpl22 **(c)**. **(d)** Consensus structure for motif identified in psbT coding sequence (psbT sense, left) and in psbT long antisense RNA (middle). Genomic context of the psbT sense RNA motif (blue) and antisense RNA motif (red) in *Arabidopsis thaliana* plastid (right). **(e)** Consensus structure of antisense RNA motif to petB, petD, and rps16 (left). Genomic context of the antisense RNA (red) in *Hordeum vulgare* mapping to group II intron of petB and rps16 (right).

Statistical information

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

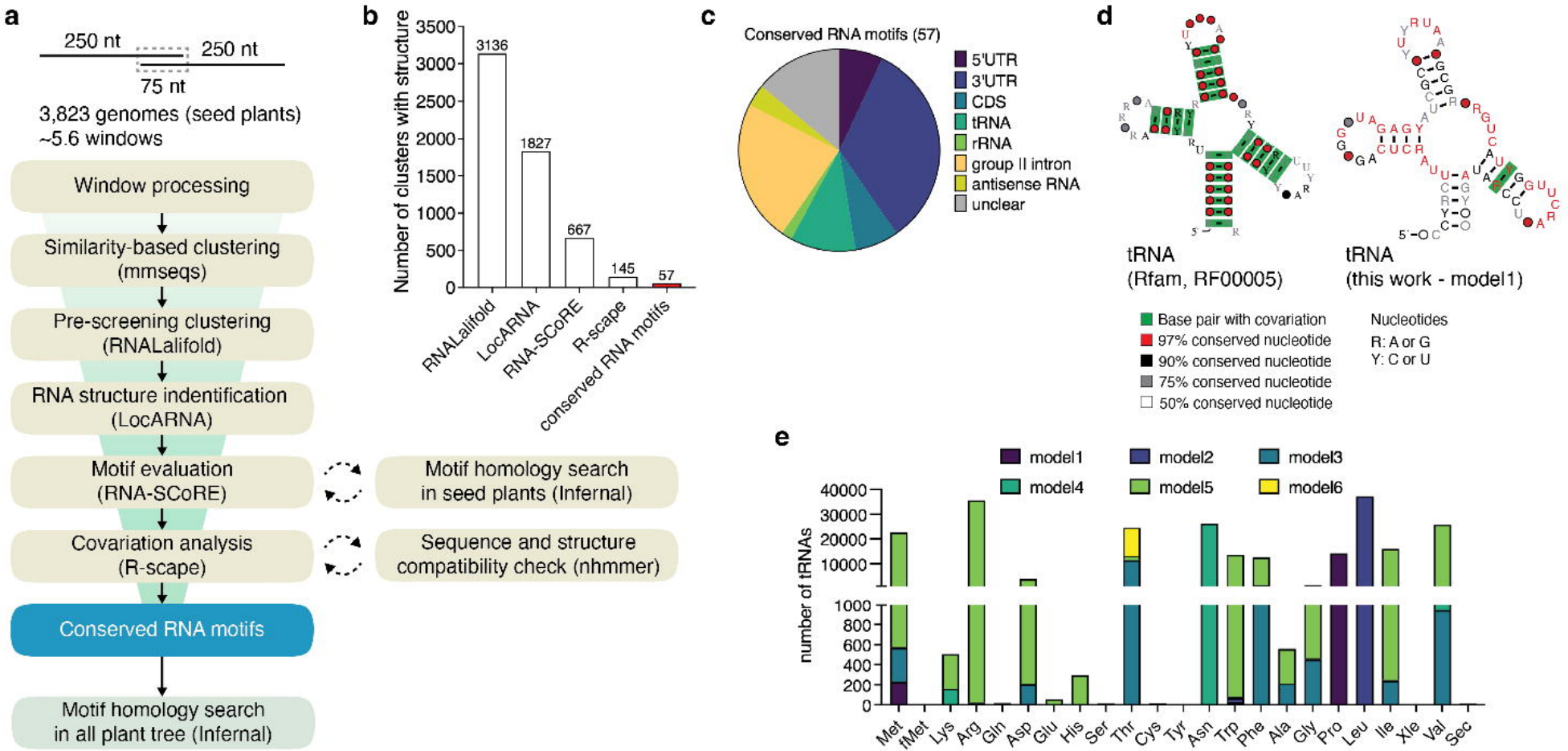
Supplementary information

Supplementary Table 1.

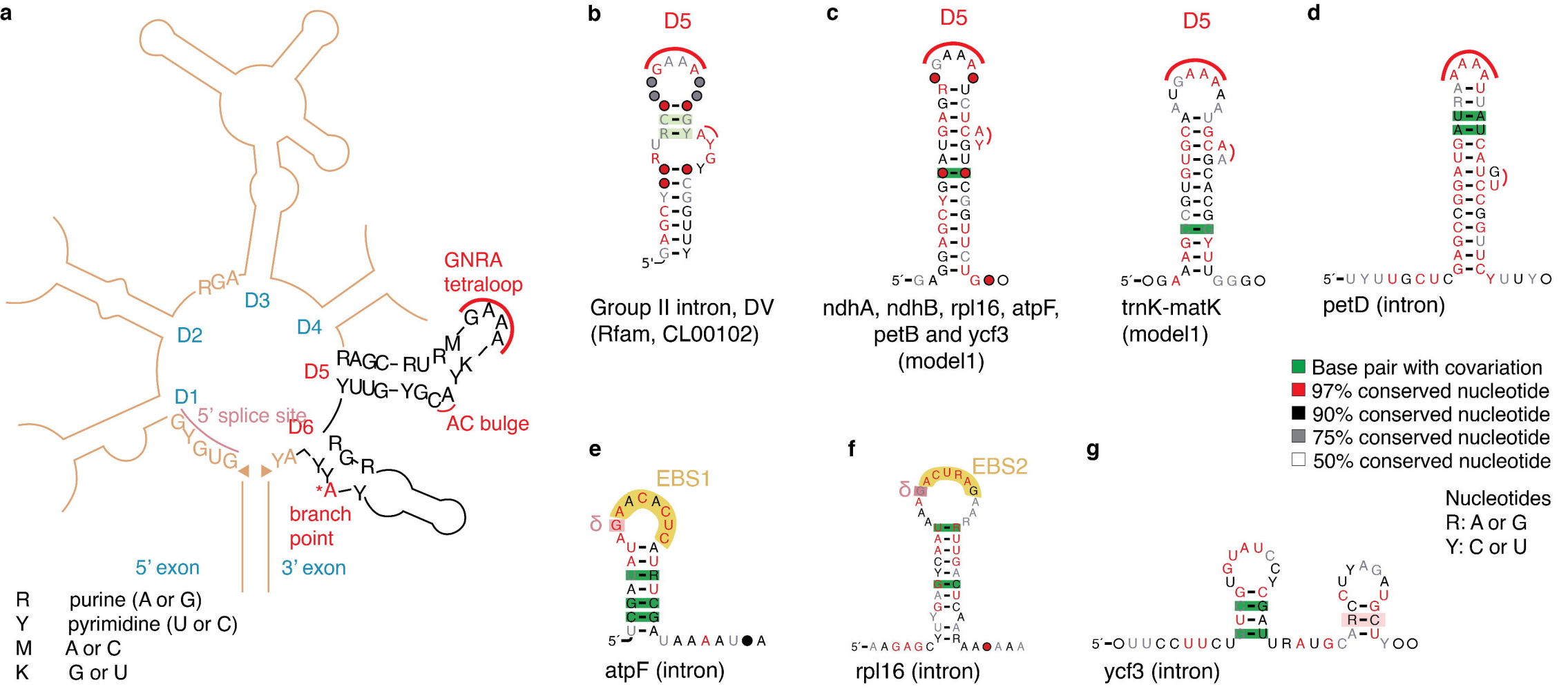
Supplementary Figure 1-5.

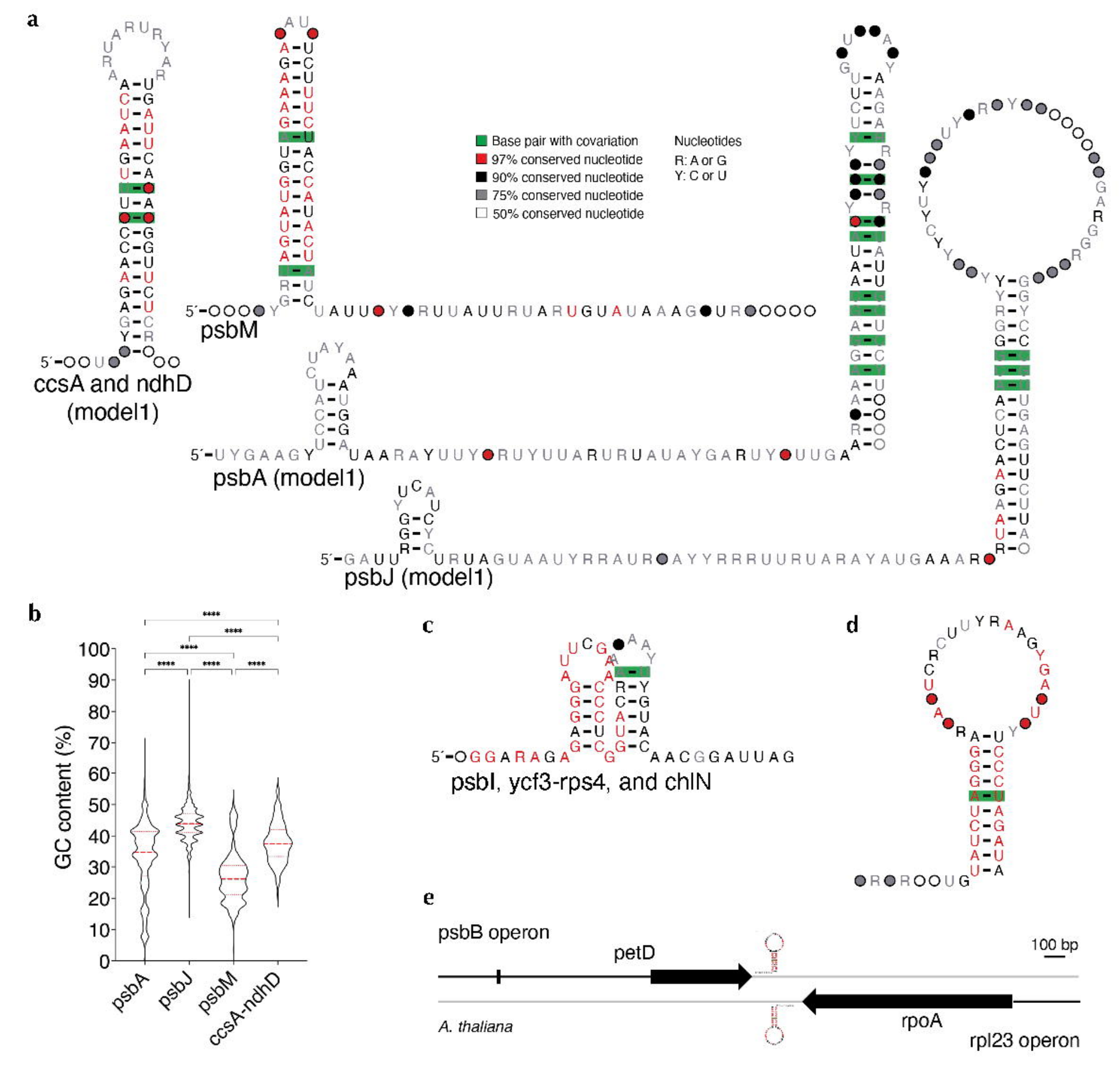
Data availability

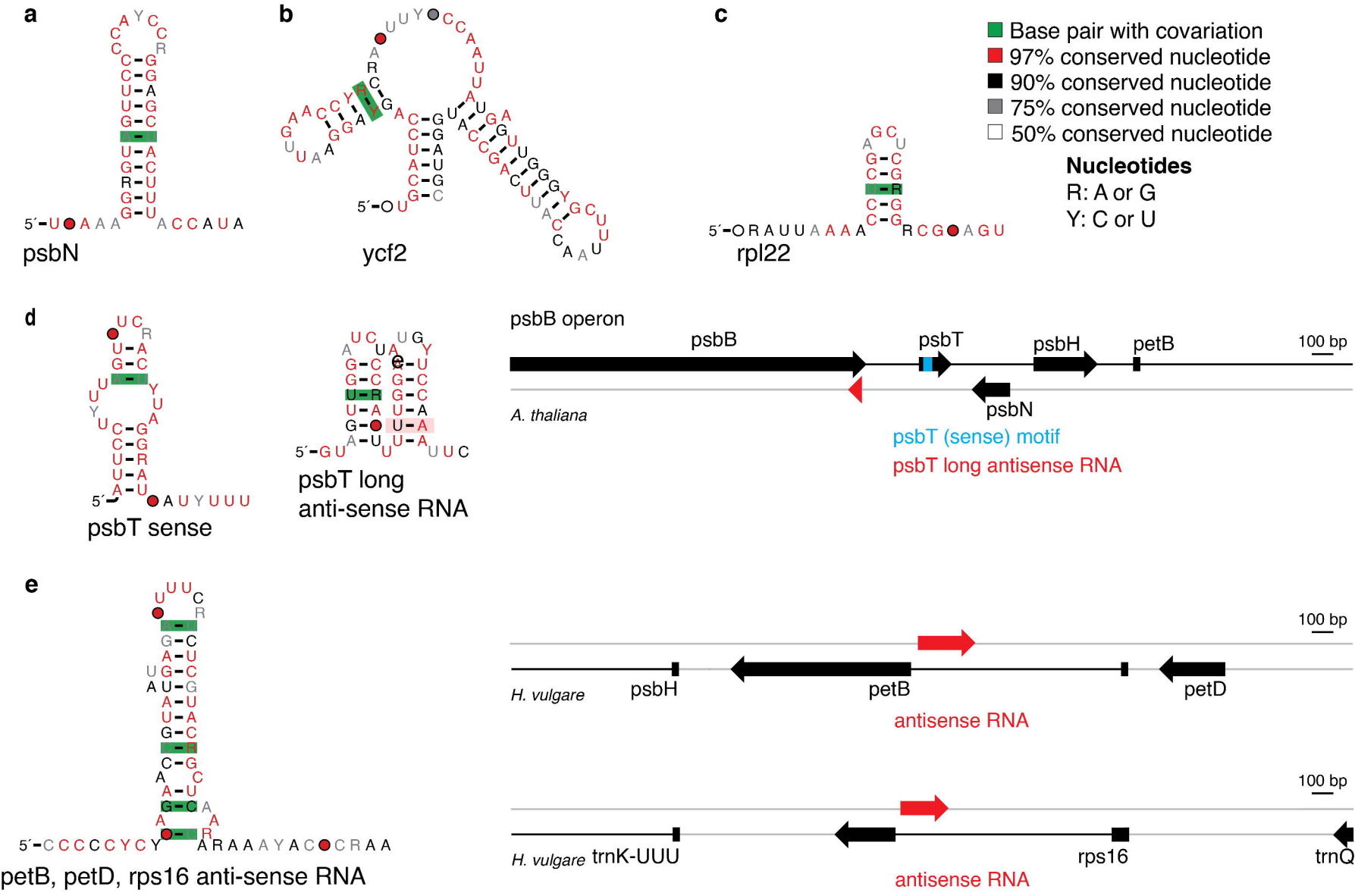
All novel conserved RNA motifs have been submitted to Rfam (identifiers to be provided). Novel motif family alignments and RNA motif annotations can be found at https://bitbucket.org/rodrigoreislab/chloroplast_motif_families/src/main/.


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)
Motif homology search in seed plants (Infernal)
Covariation analysis (R-scape)
Sequence and structure compatibility check (nhmmer)
Conserved RNA motifs
Motif homology search in all plant tree (Infernal)
b
Number of clusters with structure
3136
1827
667
145
57
RNALalifold
LocARNA
RNA-SCoRE
R-scape
conserved RNA motifs
c
Conserved RNA motifs (57)
5'UTR
3'UTR
CDS
tRNA
rRNA
group II intron
antisense RNA
unclear
d
tRNA (Rfam, RF00005)
tRNA (this work - model1)
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
e
model1
model2
model3
model4
model5
model6
number of tRNAs
Met
fMet
Lys
Arg
Gln
Asp
Glu
His
Ser
Thr
Cys
Tyr
Asn
Trp
Phe
Ala
Gly
Pro
Leu
Ile
Xle
Val
Sec

antisense RNA
CDS
5′ UTR
3′ UTR
group II intron (DI and DIII)
group II intron (DV)
tRNA
16S rRNA (helix 8)
14034 genomes
(586 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
branch 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, DV (Rfam, CL00102)
c
D5
ndhA, ndhB, rpl16, atpF, petB and ycf3 (model1)
D5
trnK-matK (model1)
d
petD (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
e
EBS1
δ
atpF (intron)
f
EBS2
δ
rpl16 (intron)
g
ycf3 (intron)


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
(model1)
psbM
psbA (model1)
psbJ (model1)
b
GC content (%)
psbA
psbJ
psbM
ccsA-ndhD
c
psbI, ycf3-rps4, and chlN
d
e
psbB operon
petD
100 bp
A. thaliana
rpoA
rpl23 operon


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
d
psbT sense
psbT long
anti-sense RNA
psbB operon
psbB
psbT
psbH
petB
100 bp
A. thaliana
psbN
psbT (sense) motif
psbT long antisense RNA
e
petB, petD, rps16 anti-sense RNA
100 bp
H. vulgare
psbH
petB
petD
antisense RNA
100 bp
H. vulgare
trnK-UUU
rps16
trnQ
antisense RNA