Functional Characterization of Selected Chloroplast RNA-Binding Proteins from *Arabidopsis thaliana*

by

Jan P. Kuciński

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular and Developmental Biology)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Andrzej T. Wierzbicki, Chair
Professor Steven E. Clark
Assistant Professor Peter L. Freddolino
Professor Lyle A. Simmons

Jan P. Kucinski

janpk@umich.edu

ORCID iD: 0000-0003-0550-5705

DEDICATION

For my wife and children.

ACKNOWLEDGEMENTS

I wanted to express my gratitude to Andrzej for being my mentor over past six years. I have surely learned a lot and became a better scientist and possibly a better person. Thank you, Andrzej.

To all Wierzbicki's lab members. Specifically, Gudrun Boehmdorfer for being extremely helpful during my first years in the lab. She was not only a great teacher and mentor but also a supportive friend. Shriya Sethuraman and Hafiz Rothi for being fantastic friends over past five years. It was a blast working with you! Undergraduate students I had an honor to work with: Natalie Blackwood, Olivia Metcalf, Pragya Khurana and Matthew Dunn. Thank you for helping me with my projects and improving my teaching skills.

To my Thesis Committee, I have enjoyed hearing your feedback and receiving your support and guidance throughout my candidacy. Thank you for the critical but supportive role you have played in my development as a scientist.

To my dearest Family, you have always been my greatest support. If that was not for you I would not be anywhere close to where I am now. Specifically, to my Grandfather who have always believed in me and provided the best professional advices that I have ever gotten.

TABLE OF CONTENTS

LIST OF FIGURES

Figure:

ABSTRACT

Plastids are indispensable, plant-specific organelles of prokaryotic origin. They serve a plethora of biological functions crucial for plant metabolism. Plastids contain their own genomes, which are remnants of their cyanobacterial ancestor. However, mechanisms regulating expression of genes encoded in plastid DNA remain poorly understood. It is hypothesized that RNA processing plays a predominant role in this regulation but exact proteins involved in this process as well as their mode of action are not fully understood. Using a combination of phylogenetic, genetic, biochemical and cell biology approaches, I have expanded the knowledge about two groups of RNA-associated proteins, which act in *Arabidopsis thaliana* plastids. First, I focused on RNase H1, a conserved enzyme responsible for digestion of the RNA strand of RNA:DNA hybrids. Through the phylogenetic analysis of this protein I have shown that the common ancestor of two largest groups of land plants, monocots and dicots, contained only one RNase H1. Subsequent gene duplication occurred independently in monocots and dicots and resulted in the presence of at least two RNase H1 paralogs in most angiosperm species. Additionally, I have shown that the *Arabidopsis thaliana* genome contains three RNase H1 genes encoding four RNase H1 proteins which display the canonical RNase H1 activity. Furthermore, I have demonstrated that these proteins localize to the nucleus, mitochondria and chloroplasts and the presence of at least one organellar (mitochondrial or plastid) RNase H1 is required for proper embryo

development. I have also shown that plants deficient in plastid RNase H1 accumulate higher levels of plastid DNA and display elevated sensitivity to replicative stress compared to wild type plants. Altogether, these results suggest that the canonical RNase H1 activity is crucial for proper nucleic acid metabolism in plastids and for plant embryonic development. Subsequently, I have performed characterization of another family of proteins that has been previously implicated in plastid RNA processing, the Defective in Chloroplasts and Leaves (DeCL) protein family. Through phylogenetic analysis I have shown that DeCL proteins are present mostly in photosynthetic organisms and are probably of cyanobacterial origin. These proteins group into five sub-families, which share a well-defined domain of unknown function, the DeCL domain. Two of these sub-families contain subunits of plant-specific DNA-dependent RNA polymerases involved in Transcriptional Gene Silencing (RNA Pol IV and Pol V), one contains nuclear-localized rRNA binding proteins (DOMINO1) and two remaining sub-families contain organellar proteins, DeCL1 and DeCL2. I have demonstrated that in *Arabidopsis thaliana* both DeCL1 and DeCL2 associate with chloroplasts membranes and the presence of at least one of those proteins is required for plant viability under normal growth conditions. DeCL1 interacts with a subset of plastid-encoded mRNAs, mostly binding to their 5’ or 3’ ends. rRNAs are also bound by DeCL1 and their processing is affected in the *decl1* mutant, which is consistent with previous data implicating DeCL proteins in rRNA processing. Altogether, these results provide evidence of direct or indirect DeCL-RNA interaction and suggest at least partial redundancy between DeCL1 and DeCL2. This identifies the establishment of the molecular mechanism of DeCL function as an important goal for future research. Altogether, this work provides important, novel insights into the processes involved in

plastid RNA metabolism. It answers several important questions about two plastid RNA-associated proteins and opens new avenues for studying regulation of plastid-encoded genes at the RNA level.

CHAPTER I

INTRODUCTION

Plastids

Plastids are endosymbiotic organelles specific to plants. These indispensable organelles are enclosed within a double membrane and serve a plethora of functions. Plastids are mostly known for performing photosynthesis, but they are also involved in fatty acid biosynthesis, pigment production as well storage of multiple energy sources. These organelles, like mitochondria, have retained their own genome which encodes some elements of plastid metabolism. Throughout this chapter I describe plastid metabolism and give an overview of current knowledge regarding regulation of the expression of plastid-encoded genes.

Endosymbiotic origin of plastids

Plastids originate from an incorporation of cyanobacteria-like prokaryote into an eukaryotic cell in the process of endosymbiosis[1], which occurred approximately 1.5 billion years ago[2]. The endosymbiotic organelles retained certain essential features of their cyanobacterial ancestors like the membrane system and independent genome. Otherwise, they have been severely reduced and specialized in metabolism and especially photosynthesis. It is worth noting that the eukaryote which initially obtained plastids already contained a different type of endosymbiotic organelles – mitochondria[3]. These, however, originate from α-proteobacteria-like prokaryote and therefore are clearly distinct from plastids in their origin, structure and function[4].

It is hypothesized that plastids were obtained once during evolution and the first photosynthetic eukaryote is a common ancestor of all photosynthesizing algae and plants[5]. However, rising evidence suggests at least two more recent examples of established or ongoing primary endosymbiosis. Approximately 60 million years ago an amoeba *Paulinella chromatophora* had gained photosynthetic symbionts similar to cyanobacteria[6,7]. These symbiotic organelles, known as chromatophores, possess peptidoglycan wall like their cyanobacterial ancestor and are also capable of photosynthesis[8,9]. The second example of additional endosymbiotic origin of plastids comes from sea slugs *Elysia chlorotica*[10]. These animals sequester plastids from algae into a dedicated compartment in their gut in a process known as kleptoplastidy[11]. It is worth noting that these plastids provide products of photosynthesis to the slug and the slug itself nurtures acquired plastids[12].

Despite the abovementioned exceptions it is a consensus that a single primary endosymbiosis between a mitochondria-containing eukaryote and a cyanobacteria-like prokaryote gave rise to all plants and algae.

Plastid development

All plant cells originate from differentiation of meristematic (undifferentiated) cells. These cells contain approximately 10-20 plastids of only one type – proplastids[13–15]. Proplastids have the ability to differentiate into any specific type of plastid depending on the combination of environmental and cellular signals, a process involving extensive communication between the nucleus and plastids. Depending on the interplay of these factors, proplastids differentiate into one of several types of mature plastids and serve a specialized biological function[16].

The main environmental signal regulating plastid differentiation is light[17]. It is perceived by cytoplasmic photoreceptors PhyA and PhyB which, upon light illumination, undergo conformational change and in their active forms are translocated into the nucleus[18,19]. There they regulate the activity of transcription factors responsible for the transition to light-mediated growth. The downstream components of this regulation include proteins involved in chlorophyll synthesis as well as plant hormones synthesis and signaling[20].

Chloroplasts are the most studied type of plastids, carrying the primary plastid function – photosynthesis. This type of plastids is found in all photosynthetic organs and tissues. Chloroplasts contain internal membrane structures, thylakoids, stacked on top of each other in large features, the grana. Thylakoid membranes are the anchor sites for chlorophyll pigments and light reactions of photosynthesis[21]. Light exposure induces multiple hormonal regulatory pathways which finetune chloroplast development. Plant hormone, gibberellic acid (GA), has been shown to inhibit greening of *Arabidopsis* seedlings grown in the dark as well as regulate chlorophyll and carotenoid synthesis[22]. It was also demonstrated to affect plastid division and stacking thylakoids into grana – process hypothesized to determine photosynthesis efficiency.

Apart from chloroplasts, plants contain three main types of plastids fulfilling different roles. Chromoplasts are responsible for synthesis and storage of carotenoids, including β-carotene, lycopene, neoxanthin, violaxanthin and lutein[23]. Instead of thylakoids, chromoplasts contain plastoglobules which mainly serve as storage sites for lipids and proteins but also are a site of a number of enzymatic reactions[24–26]. Leucoplasts are another type of plastids not involved in photosynthesis, and include three groups of plastids of different functions: amyloplasts (primarily involved in starch biosynthesis,

storage and degradation)[27,28], elaioplasts (specialized in lipid biosynthesis and storage)[29], and poorly characterized proteinoplasts (proposed to facilitate occurrence of specific enzymatic reactions)[30]. The last major type of plastids are gerontoplasts. These organelles develop from chloroplasts in senescent leaves and are characterized by the breakdown of chlorophyll pigments as well as degradation of the thylakoid membranes[31,32].

Plastid division

Due to their endosymbiotic origin plastids, similarly to mitochondria, cannot be synthesized *de novo* inside the cells. Instead, they need to undergo division and be segregated to the daughter cells during cytokinesis. This complex biological process needs to be tightly controlled and coordinated[33,34].

In order to multiply, plastids propagate through binary fission of a pre-existing plastid using protein machinery of chimeric origin that produces three major plastid-division (PD) ring structures: inner and outer PD rings and a dynamin ring[35–38]. First, the inner PD ring is formed by polymerization of FtsZA-FtsZB heterodimer in plastid stroma[39]. The information about localization of the FtsZ ring is transferred by a number of proteins[40–43] and determines the binding of PDR1 protein of glycosyltransferase activity to the outer plastid membrane. In the next step, PDR1 synthesizes polyglucan filament ring. This polyglucan filament is then bound by a dynamin-related GTPase protein DRPB5 which cross-links the PD ring filaments. Formed PD-dynamin ring generates contractile force at the plastid division site and pinches off the membranes. Resulting daughter plastids are separated and division machinery is disassembled[39].

It is important to note that cell and plastid division cycles need to be precisely coordinated. For example, in red alga *Cyanidioschyzon merolae* and in glaucophyte alga *Cladophora paradoxa*, it was demonstrated that cell division is arrested in prophase if plastid division is blocked before the formation of FtsZ ring. In these cells the increase of major cell-cycle regulator, cyclin B, is blocked[44]. It has been also shown that expression of genes encoding plastid division machinery in algae is cell-cycle regulated[45]. Additionally, multiple examples of coordination between nuclear DNA replication and plastid division has been shown and it is hypothesized that plastid-produced tetrapyrroles play a major role in this regulation[46–49]. However, the central mechanism driving cell and plastid division coordination remains undiscovered.

Plastid genome

Plastids, similar to mitochondria, contain their own DNA descended from the cyanobacterial ancestor. Therefore, plastid equivalent of chromosome, the nucleoid, resembles a bacterial genome – it is genetically circular with most of the genes organized in polycistronic units. However, the presence of linear nucleoids has been strongly argued by some researchers[50–52]. Interestingly, during the course of evolution, the plastid genome underwent significant reduction and many genes of cyanobacterial origin are now encoded in the nuclear genome. From the initial hypothesized several thousands of genes in cyanobacterial ancestor of plastids[1], *Arabidopsis* plastid genome encodes only 87 protein coding and 41 rRNA and tRNA genes[53]. In other plant species the number of genes present in the plastid genome does not exceed 200, while the plastid proteome consists of around 3000 proteins[54,55]. Interestingly, genes which remained in the plastid genome encode mostly proteins involved in photosynthesis and

this subset of genes remains relatively well conserved between plant species[56,57]. However, a significant portion of the photosynthetic apparatus proteins is encoded in the nuclear DNA. It has been shown that plastid encoded genes may still be transferred to the nuclear chromosomes and, interestingly, gene transfer from nucleus to the plastid genome is also possible[58]. The exact reason why plastids, similar to mitochondria, still contain their own DNA rather than fully depend on the import of nuclear encoded proteins, remains unknown but several hypotheses has been offered. One possibility is that plastids retained their genomes to allow for precise regulation of photosynthesis by controlling stoichiometry of protein complexes involved in these processes on the level of individual plastids rather than the entire cells[56]. Another hypothesis, Co-location for Redox Regulation (CoRR), originally proposed by John Allen[59] suggests that plastids retained protein coding genes in the nucleoid in order to control their expression by the redox state of protein products of these genes. The CoRR hypothesis proposes that the status of the photosynthetic machinery affects the redox state of plastid proteins. This state may then be used to regulate the expression levels of plastid-encoded genes in a plastid-autonomous manner. In *Arabidopsis* the Chloroplast Sensor Kinase (CSK) has been proposed to serve as a redox sensor. This plastid-localized protein is encoded in the nuclear genome and has homologs in organisms from all major groups of photosynthetic organisms. This protein was also shown to be autophosphorylated in a redox-dependent manner[60]. Despite being extremely interesting, the CoRR hypothesis still remains controversial and awaits confirmation.

A third speculative explanation of the retention of the organellar genomes is based on the observation that they mainly encode hydrophobic proteins. It was proposed that the import of these proteins into plastids or mitochondria would be a major challenge and

therefore these proteins need to be synthesized *in situ*[61–63]. However, this hypothesis is also controversial and it remains possible that all current explanations are partially correct.

Plastid genome organization

The plastid genome encodes a subset of proteins needed for chloroplast metabolism and several proteins involved in plastid gene expression. It also encodes a complete set of rRNAs and tRNAs, since there is no known mechanism of RNA transport into plastids (REF). The plastid genome is typically divided into four major regions: the Large Singe Copy (LSC), Small Single Copy (SSC) and two Inverted Repeat Regions (IRR). LSC and SSC contain protein coding as well as some tRNA genes. The IRRs are nearly identical in sequence and contain rRNA the rest of tRNA genes and some protein-coding genes[64]. Interestingly, plastid genomes are very AT-rich, typically ~60%, especially in protein coding genes[65]. Recent report of gene sequences from plastids of *Balanophora reflexa* and *B. laxiflora*, non-photosynthetic plants, demonstrated AT contents of 88.4% and 87.8%, respectively, for some genes reaching 98%. This strong AT bias is also represented in amino acid composition of proteins encoded in these genomes. Specifically, ~80% of the proteome is represented by only six amino acids (asparagine, isoleucine, leucine, lysine, phenylalanine and tyrosine). All of these amino acids may be encoded by codons lacking C or G. Despite the lack of photosynthetic abilities, plastids from these plants are metabolically active, especially in fatty acid synthesis[54,66]. The exact reason for the high AT content of plastid genomes is still unknown. Although, it was observed that GC content is correlated with plastid genome size and proposed that higher AT content allows for more structural rearrangements and higher mutation rate[67].

The single plastid genome size is typically between 100-200kbp, however exceptions do exist. For example marine unicellular alga *Acetabularia* has been reported to contain chloroplast genome up to 1.5Mbp[68], on the contrary different marina alga *Osteococcus taurus* has a plastid genome tightly packed into 72kbp[69]. Diversity in plastid genomes has also been reported within single genera. For example, algae from the genus *Chlamydomonas* vary in plastid genome size from 187kbp up to 292kbp, depending on the species[70].

Plastid nucleoid

Plastid DNA is packaged into a three dimentional nucleoprotein structure known as the nucleoid. Plastids contains multiple nucleoids, each containing multiple copies of the genome. Therefore, the copy number of the plastid genome greatly exceeds the copy number of the nuclear genome and chloroplasts may contain the majority of total cellular DNA content[71].

Upon staining with a nucleic acid stain such as DAPI (4’,6’-diamidino-2-phenylindole), nucleoids are visible as granules of various sizes localized in the plastid stroma. They are often associated with plastid membranes such as thylakoid and inner envelope membranes. DNA in plastid nucleoids is packaged with several nucleoid-associated proteins[71,72]. During evolution land plants lost bacterial nucleoid-associated proteins like HU and H-NS, which are responsible for chromosome packaging in cyanobacteria. Plastids also do not contain histones or other typical nuclear chromatin proteins. Instead, plastid nucleoid is organized by unique set of DNA-binding proteins[73]. Additionally, proteomic analysis of maize nucleoids showed association of plastid DNA with various proteins involved in DNA replication and repair as well as in RNA metabolism:

transcription, RNA processing, translation and ribosome assembly[73,74]. Importantly, presence of those proteins on plastid DNA may suggest that plastid RNA processing and translation may occur co-transcriptionally.

Plastid DNA replication

In order to remain fully functional, plastids need to replicate their DNA before organelle division. First plastid protein exhibiting DNA polymerase activity was isolated from tobacco chloroplast and was demonstrated to be similar in size to the Klenow fragment of Polymerase I from *E. coli*[75,76]. Interestingly, activity of this enzyme resembled this of bacterial polymerases but no apparent sequence or structural similarity to cyanobacterial enzymes was found[77]. Due to this unclear origin, plastid Polymerase I is called POP – Plant and protist Organellar DNA Polymerase. Two genes encoding POPs were found in *Arabidopsis* – PolIa and PolIb. It was demonstrated that protein products of both genes localize to chloroplasts and mitochondria and act redundantly in organellar DNA replication[78]. Maize, in turn, contains only one gene encoding POP and loss of this enzyme leads to dramatic decrease in cpDNA copy number and seedling lethality[79]. This indicates that POP servers an important role in replication of plastid DNA.

Mechanism of plastid DNA replication is believed to follow the rolling circle model, where one of the strands of circular nucleoid is nicked, displaced from the complementary (not-nicked) strand which serves as a template for DNA replication while the nicked strand is elongated[80,81]. It was shown that the tobacco plastid genome contains two origins of replication (*ori*), located in the IRR section of the plastid genome, and at least one of them is needed to maintain plastid DNA[82].

An additional mechanism of plastid DNA replication was proposed, the Recombination-Dependent Replication (RDR)[50]. It is hypothesized that in this process, the 3'-OH end of ssDNA overhang of one of the genome copies invades another copy of the genome at a homologous site. It then primes DNA replication of the leading strand and eventually complete replication fork machinery is established[83,84]. This leads to the formation of a four-stranded DNA structure, the Holliday junction. It was shown that proper resolving of this structure is crucial for correct nucleoid segregation[85]. Importantly, if RDR originates in one of the IRRs in the plastid genome it may lead to inversion of the single copy regions, a phenomenon commonly observed in plastid genomes[86]. It was also hypothesized that the ssDNA overhang which initiates homologous recombination might be a substrate for C-to-T deamination and therefore, in combination with the presence of RNA editing enzymes, be an explanation of high AT-content in plastid genomes[87].

Interestingly, it was recently suggested that R-loops, three-stranded nucleic acid structures with RNA:DNA hybrids, may also serve as DNA replication initiation sites in plastids, similarly to mitochondrial DNA replication, where D-loops, three-stranded DNA, were shown to be DNA replication initiation sites[50,88]. However, the exact mechanism of the replication initiation by R-loops and direct evidence of this process remain undiscovered. Importantly, we need to remember that the central element of the R-loop, RNA:DNA hybrid, if not properly removed stalls DNA replication resulting in DNA breaks and genome instability[89,90].

Additionally, plastid DNA replication does not seem to be coupled with either organelle or cell division[91]. Especially, that plastids by default contain variable numbers of genome copies. Instead, it is believed that nucleoid molecules are tightly associated

with thylakoid membranes and therefore can be uniformly propagated to the daughter organelles[92].

Plastid RNA metabolism

Organization of plastid-encoded genes

Despite over 95% of chloroplast proteome being encoded in the nucleus and imported from the cytoplasm, proteins encoded in plastid DNA serve functions crucial to organelle, cell and whole organism survival[93]. Therefore, expression of plastid-encoded genes must be tightly regulated and coordinated with the metabolism of the entire cell.

Genes in the plastid genome are organized in operons, polycistronic transcriptional units, a characteristic of prokaryotic organisms. However, bacterial operons usually encode functionally related genes while those in plastids sometimes contain genes encoding proteins of unrelated functions[94,95]. Additionally, some plastid-encoded genes contain introns, which need to be co-transcriptionally spliced[96].

Transcription

Transcription of genes encoded in the plastid genome is orchestrated by a combination of two DNA-dependent RNA polymerase complexes. Interestingly, one of these complexes is encoded in the nuclear genome, the Nuclear-Encoded Polymerase (NEP), while the genes encoding components of the second one, the Plastid-Encoded Polymerase (PEP), are present in the plastid chromosome[97]. It is believed that these two RNA polymerase complexes function sequentially, but on a different subset of operons and specificity is achieved through different promoters being recognized by NEP and PEP[96,98]. NEP is thought to be the most active in proplastids where it transcribes plastid-encoded genes for PEP subunits, then PEP transcribes genes crucial

for mature plastid development. Interestingly, NEP is also active in mature plastids and may rescue PEP function upon the latter one being absent or defective[99].

NEP was initially found in spinach chloroplasts as a 110kDa polypeptide exhibiting RNA polymerase activity[100]. Interestingly, the number of genes encoding NEP is variable between plant species. This variability correlates with the number of localization sites for these enzymes. Most dicots (e.g. *Arabidopsis, Tobacco*) contain three NEP genes, products of which localize to plastids, mitochondria while the third one exhibits dual targeting to both of these organelle types. Monocots (e.g. *Rice, Wheat, Maize*) contain only two NEPs, localized to either plastids or mitochondria, and basal plants (e.g. green algae, *Selaginella*) encode only one NEP[96]. NEP is the only transcriptional machinery for *rpoB* operon (which contain genes encoding PEP components) as well as single genes *accD* and *ycf2*, which encode acetyl-CoA carboxylase and ATPase of unknown function, respectively[101].

The first evidence of RNA synthesis in isolated broad bean chloroplasts was obtained in 1960s[102], a discovery followed by characterization and purification of the PEP complex from maize[103], pea[104] and spinach[105] chloroplasts. Interestingly, all of these enzymes were found to be composed of five subunits - two α subunits, and one of each β, β’ and β’’, architecture resembling that of the bacterial RNA polymerase complex[106] but the ω subunit is absent. Protein products of genes *rpoC1* and *rpoC2* which encode plastid β’ and β’’ subunits, respectively, can be aligned with the N- and C-terminal fragments of bacterial β’ subunit, suggesting duplication of this gene in the cyanobacterial ancestor of plastids or in the common ancestor of all photosynthesizing eukaryotes[107].

Transcription initiation

DNA binding and transcription initiation by bacterial RNA polymerase was demonstrated to be dependent on the activity of a protein named σ (sigma) factor. Initially, a 70 kDa protein was discovered in *E. coli* and named sig70[108,109] followed by discoveries of other sigma factors in *E. coli* and almost all known bacteria[110]. Prokaryotic σ factors are divided into two independent families, named after their first representatives, namely SIG70 and SIG54. Families are then divided into groups, with group I containing essential σ factors and groups II-IV containing alternative or specialized sigma factors[111]. It was demonstrated that σ factor binding to the bacterial core RNA polymerase leads to enhanced affinity of the holoenzyme to the gene promoter and increased transcription[112].

Considering high similarity between bacterial RNA polymerase and PEP it is not surprising that plastid RNA polymerase machinery uses sigma factors as well. Initially discovered in the red alga and soon described from many higher plant species, eukaryotic sigma factors were found to share high sequence similarity of their C-terminal fragments with the members of bacterial SIG70 family[113,114]. Interestingly, the N-terminal portion of plastid sigma factors is shared among all photosynthesizing eukaryotes but is absent from the bacterial ones and its biological function remains unknown.

The number of plastid sigma factors varies from one in green alga *Chlamydomonas reinhardtii*[115] up to seven in California poplar tree with typical number of six in majority plants, including *Arabidopsis thaliana* (SIG1-6)[114,116]. The expression of sigma factors depends on the environmental signals, cell type as well as developmental stage. These transcription factors are known to either regulate expression of specific subset of

genes[117–122] or broadly regulate expression of plastid encoded genes in response to particular stimulus[123–125].

Intercistronic transcript processing

Plastid-encoded genes are transcribed as long, polycistronic mRNA molecules[96]. However, contrary to what is observed in bacteria, polycistronic mRNAs are subjected to slicing, known as intercistronic transcript processing[126]. It has been shown that processed, monocistronic transcripts are translated more effectively than polycistronic ones[127]. However, it was also demonstrated that this processing is not required for translation and, additionally, that it is not crucial for over-expression of transgenes, which may be efficiently translated from unprocessed transcripts[128,129].

Little is known about specific proteins involved in intercistronic cleavage of plastid transcripts. Two Pentatricopeptide Repeat Proteins (PPR) have been demonstrated to bind some plastid-encoded transcripts. *Arabidopsis* HCF152 and maize CRP1 proteins recognize and bind specific RNA sequence. It was demonstrated that this binding determines 5'-ends of processed transcripts through protecting it from 5'->3' exonuclease digestion[130]. The *Arabidopsis* HCF107 protein is another identified factor which was proposed to be involved in this process. Plants deficient in HCF107 fail to accumulate processed mRNA encoding PsbH protein, a component of the photosynthetic machinery, which leads to lethality of mutant seedlings[131]. Importantly, it cannot be excluded that HCF107 stabilizes processed mRNA allowing for its' accumulation.

Splicing of plastid-encoded mRNAs

Genes encoded in the plastid genome contain two types of introns, group I and group II[126]. Splicing of both is considered RNA-catalyzed because some members of these groups are capable of self-splicing *in vitro*. However, many of organelle introns require protein machinery for *in vivo* splicing.

Group I introns are spliced in a two-step process. In the first step, the *trans*-esterification, guanine performs nucleolytic attack on the 5'-end of the intron. This step is followed by another nucleophilic attack by the same guanine on the now exposed 3'OH group of the upstream exon[132]. Interestingly, land plants retained only one group I intron, which has lost the self-splicing ability. However, all five group I introns from green algae *Chlamydomonas* exhibit autocatalytic splicing activity *in vitro*[133].

Group II introns are characterized by the conserved structure of six helical domains connected to a central core[134]. These introns are usually spliced in a two-step *trans*-esterification reaction. First, the 2'OH of adenosine nucleotide bulging in the sixth helix domain performs the nucleophilic attack on the 5'-end of the intron. This results in a branched structure, and after second *trans*-esterification, the intron is released as a lariat. Alternatively, the water molecule may serve as a nucleophile during first step of this splicing[135]. Plastid genome of *Arabidopsis* contains twenty group II introns with one unusual example of *trans*-splicing of *rps12-1* gene[136,137].

Splicing factors involved in plastid pre-mRNA maturation are encoded in both plastid and nuclear genomes[126]. The main plastid-encoded splicing factor, MatK, shares sequence similarity with canonical proteins involved in intron splicing (maturases). It was demonstrated to bind intron RNA *in vitro* and to be required for proper splicing of some group II introns[138–140].

The main family of nucleus-encoded splicing factors is the CRM (Chloroplast RNA and ribosome Maturation) domain family[141]. They contain up to four RNA-binding domains. Interestingly, in prokaryotes, proteins containing single CRM domain are involved in ribosome maturation. Majority of plant CRMs have been predicted or shown to localize to either plastids or mitochondria[142]. Five of them participate in the splicing of specific group II introns in plastids. Interestingly, no redundancy was detected between them. One of the CRM domain proteins, CFM2 was shown to be required for a single group I intron in plastids[143,144].

The Pentatricopeptide Repeat (PPR) is another protein family shown to participate in plastid splicing. PPR4 was found to be essential for *trans*-splicing of *rps12-1* intron in *Arabidopsis* and maize[137]. Loss of a different PPR protein, OTP51, causes substantial defect in splicing of *ycf3-2* intron as well as partial defects in splicing of other group II introns[145]. PPR5 protein was shown to protect *trnG* pre-mRNA from degradation but is also suggested to be directly involved in splicing due to its' ability to bind intron RNA *in vitro*[146,147].

WTF1 protein was found to bind RNA *in vitro* as well as associate with a specific subset of introns in maize chloroplasts. Loss of WTF1 leads to defects in splicing of these introns[148]. Additionally, another protein from this family, WTF9, was found to be involved in group II intron splicing in *Arabidopsis* mitochondria[149].

Intron splicing from plastid-encoded mRNAs uses a number of mechanisms and specific proteins which often are involved in processing of only a few genes. Additionally, only a small number of plastid-encoded genes contain introns. This raises an important question of whether introns play a role in regulation of plastid gene expression. It was shown that deletion of group I introns in *C. reinhardtii psbA* and rRNA genes did not

cause any obvious effect on cell growth[150–152]. Additionally, it was shown that *trans*-splicing mutants can be rescued with intron-less *psaA* sequence[126]. Therefore, it is possible that plastid introns are not important for plastid gene expression but rather are relics from their bacterial ancestors. However, conclusive results from flowering plants are still missing.

Transcription termination

Properly controlled transcription termination prevents formation of antisense RNAs, interference with the activity of other RNA polymerase complexes as well as secures a pool of RNA polymerase for new transcription events[153]. In bacteria two major mechanisms of transcription termination have been described - Rho-dependent and intrinsic (Rho-independent).

Despite the bacterial origin of plastids, no homologs of prokaryotic Rho factors have been identified in chloroplasts so far. Additionally, in plastids it is mainly RNA maturation rather than transcription termination that determines 3'-ends of transcripts[130,154]. However, *Arabidopsis* genome sequence analysis revealed the presence of several proteins predicted to localize to plastids, which contained an RNA-binding domain similar to the one present at the N-terminus of bacterial Rho factors. These proteins were named RHON and proposed to be involved in plastid RNA metabolism. Specifically, RHON1 protein was demonstrated to participate in transcription termination of *rbcL* operon. Loss of RHON1 caused read-through of this operon leading to the production of a large precursor transcript consisting of *rbcL* as well as multiple downstream operons[155]. Biochemical characterization of RHON1 function suggests that it is able to terminate *rbcL* transcription in a similar way to bacterial Rho factors.

Specifically, RHON1 specifically binds to the 3'UTR of *rbcL* mRNA as well as single-stranded DNA encoding this region. *In vitro* transcription termination studies showed that RHON1 terminates transcription of *rbcL* and that this process was dependent on its' ATPase activity[156]. However, the actual mechanism of this termination may be different from the one observed in bacteria. The mRNA sequence bound by RHON1 does not show similarity to the Rho binding site, specifically C-content is significantly lower. Additionally, in *E. coli* Rho binds naked mRNA molecules which are not being translated or bound to RNA-binding proteins[157]. In contrast, RHON1 binds nascent mRNAs as part of a large RNA-protein complex *in vivo*[155].

An interesting feature of plastid transcriptional units is the presence of inverted repeat sequences in their 3'-ends[158,159]. These sequences are capable of folding into stem-loop structures, similar to those observed in intrinsic transcription terminators in *E. coli*. Therefore, it was suggested that inverted repeat sequences may serve as intrinsic transcription terminators in plastids[160]. *In vitro* transcription termination experiment in spinach chloroplasts showed that some of the inverted repeats were partially effective as intrinsic transcription terminators. However, based on the partial effect, it was proposed that the inverted repeats at the 3'-ends of plastid transcripts form stem-loops in order to stabilize those transcripts, preventing 3'->5' mRNA degradation or that these structures serve as platforms for RNA-binding proteins which protect nascent mRNAs from degradation[161]. However, whether this mechanism truly applies to plastid transcription termination, still remains to be determined.

RNA processing

Regulation of plastid-encoded gene expression is believed to be regulated mainly at the post-transcriptional level, contrary to bacterial regulation affecting primarily (but not only) transcription initiation[162]. RNA molecules produced in plastids undergo extensive processing affecting their sequence, structure, stability, size and functions. Protein machinery responsible for RNA processing represents a significant fraction of the plastid proteome[74] and therefore a thorough understanding of its functions remains a key step in discovering all regulatory mechanisms orchestrating the plastid transcriptome.

RNA editing

An unusual feature of organellar RNA metabolism in plants is RNA editing. It most likely emerged approximately 450 million years ago in early land plants. Interestingly, evidence of RNA editing have not been found in green algae and liverworts but have been discovered in animals, suggesting convergent evolution of this process in land plants and animals[163]. However, RNA editing in animals occurs in both the nucleus and mitochondria, while in land plants it is restricted to plastid and mitochondrial RNA maturation.

The most common event of RNA editing is conversion of cytidine to uridine (C-to-U) while adenosine to inosine (A-to-I) conversion is limited to animals[164]. The number of editing sites in chloroplast transcripts varies from zero to thousands between plant species, with flowering plants exhibiting usually between 20 and 60 editing sites[165–168]. This process seems to be more common in plant mitochondria where the number of editing sites is usually between 400 and 600[169–172]. RNA editing is also more common in basal plants (e.g. lycophytes and ferns) in both plastids and mitochondria exhibiting approximately 10 times more editing sites than their flowering plant counterparts[173–175].

C-to-U conversion can be explained by cytidine deamination. However, plant enzymes directly involved in this reaction have not been identified. Multiple sequence-specific PPR proteins have been found to bind in close proximity to editing sites[126]. Also, DYW domain-containing proteins were implicated in RNA editing. The DYW domains contain the zinc-binding motif also found in cytidine deaminases from different organisms. However, *in vitro* editing assays did not prove enzymatic activity of these proteins and loss of them does not affect RNA editing *in vivo*[176,177].

It is hypothesized that RNA editing serves as post-transcriptional correction mechanism and acts as a buffer for T-to-C mutations in the coding sequences[178]. However, why it is limited to organelles in plants and why only one type of mutation can be corrected by this mechanism remains unknown.

RNA-binding proteins

PPR proteins

The most thoroughly characterized group of plastid RNA-binding proteins is the Pentatricopeptide repeat (PPR) protein family. PPR proteins contain between 2 and 30 helical repeats composed of 35 amino acids. Each of the repeats is composed of two α-helices, which were proposed to specifically recognize RNA bases. These repeats stack together and form an extended surface recognizing a specific RNA sequence[179].

Genes encoding PPR proteins have been found in every sequenced plant genome and were predicted to localize mainly to plastids and mitochondria. Maize genome encodes almost 600 PPR proteins and *Arabidopsis* genome was found to encode 450 PPR proteins, out of which 41 and 63 are targeted to plastids and mitochondria, respectively[180].

PPR proteins affect various aspects of RNA metabolism. They were implicated in splicing, maturation, editing, stabilization as well transcription and translation in the organelles[181]. These proteins also affect a number of biological processes including photosynthesis, respiration, gametogenesis and organelle development[128,182–184].

Members of this family have been shown to bind and promote splicing of specific introns[137,185]. It has been proposed that in the case of *cis*-splicing, PPR proteins prevent the formation of an RNA hairpin which would prevent efficient splicing[146,147]. Additionally, PPR proteins protect some introns, including those being a subject of *trans*-splicing, from nucleolytic cleavage and therefore allowing for proper splicing to occur[181].

This family of proteins also plays an important role in determination of both ends of mature RNAs. PPR proteins have been shown to bind 5'- and 3'-ends and protect them from exonucleolytic cleavages and it was demonstrated that the PPR protein binding site determines the end of mature transcripts[154,186]. Importantly, these proteins affect ends of transcription units but also termini which result from intercistronic cleavages[126]. It has been proposed that the RNA-stabilizing function of PPR proteins is its major role and that it determines the complexity of the plastid transcriptome[181].

Members of the PPR protein family also affect translation of plastid-encoded genes. It was shown PPR protein binding specifically enhances translation of mRNAs. It was demonstrated that in the absence of PPR10 an RNA hairpin is formed at the ribosome-binding site of *atpH* mRNA. Upon PPR10 binding the hairpin is resolved and translation may occur. Importantly, at the same time PPR10 protects the 5'-end of this mRNA from exonucleolytic cleavage. It was proposed that translation of other plastid-encoded mRNAs is stimulated in a similar manner by other PPR proteins[187,188].

In addition to their RNA-metabolism related role, PPR proteins have been shown to associate with a number of protein complexes in plastids. They directly interact with elements of photosynthetic machinery, electron transfer as well as carbon metabolism and protein degradation complexes[181]. However, the exact role of those interactions remains unknown, it is tempting to speculate that they may affect plastid gene expression at the levels beyond RNA-metabolism.

RNA-stabilizing proteins

Apart from PPR proteins recognizing specific RNA sequences other plastid RNA-binding proteins exist and participate in transcript stabilization. Large group of cpRNPs (Chloroplast Ribonucleoproteins) consists of chloroplast stroma-localized proteins containing RNA-Recognition Motifs (RRM)[189]. These proteins bind various mRNA molecules which are not engaged in the translation process as well as pre-tRNAs[190]. It was demonstrated that upon cpRNPs depletion from stromal extracts, RNA was degraded rapidly, which was reversed upon addition of recombinant cpRNPs[191]. These results indicate that cpRNPs are key proteins responsible for chloroplast RNA stability. Specifically, it was shown *in vitro* that spinach 28RNP is important for stabilization of 3'-ends of several chloroplast mRNAs[192]. It was also shown that *Arabidopsis* cpRNPs, CP31A and CP29A, bind 3'-ends of multiple sense and antisense transcript protecting them from degradation under cold stress condition[193]. Additionally, CP33A protein from *Arabidopsis* was discovered to serve as global regulator of chloroplast RNAs, crucial for plant development and chloroplast biogenesis[194].

Multiple additional plastid RNA-stabilizing proteins have been discovered so far. HCF145 protein contains two transcript-binding motif (TMR) domains and was shown to

bind and stabilize 5'-end of polycistronic *psaA* transcript in *Arabidopsis* chloroplasts. HCF107 is another protein shown to be involved in plastid mRNA stabilization. This thylakoid-bound protein exists as part of high molecular weight complexes and binds RNA in a sequence-dependent manner[195]. It was demonstrated that HCF107 binds 5'-UTR of *psbH* mRNA protecting it from 5'->3' exonucleolytic digestion *in vitro*. It was also proposed that this binding affects the structure of mRNA which allows for ribosome binding and more efficient translation[196,197].

RNA ends modifications and degradation

The most thoroughly characterized enzyme responsible for processing of 5'-ends of plastid mRNAs is RNase J. Initially characterized in bacteria *B. subtilis*, RNase J usually contains metallo-β-lactamase (MBL) domain followed by β-CASP as well as RNA recognition motif domains. It was found to act as a dimer or tetramer and exhibit 5'->3' exonucleolytic as well as endonucleolytic activity which is dependent on the presence of two zinc ions in its' catalytic center[198]. Plant RNase J polypeptide chains are typically longer than their bacterial counterparts and contain N-terminal transit peptide responsible for plastid localization as well as C-terminal GT1-like DNA-binding domain, which exact properties remain unknown but it is not essential for RNA degradation *in vitro*. *Arabidopsis* RNase J was found to be phosphorylated and form high molecular weight complex with other plastid exonucleases PNPase and RNase E[199]. Analysis of *Arabidopsis* mutants deficient in RNase J revealed that this enzyme is crucial for embryo development. Specifically, *rnj* homozygous mutant embryos ceased development at the globular stage and exhibited impaired chloroplast development, suggesting a crucial role for RNase J in plastid RNA metabolism[200]. Based on *in vitro* experimental results, it was

proposed that RNase J digests mRNA from its' 5'-end until it is stopped by a PPR protein protecting a specific sequence within a transcript[199]. Interestingly, virus-induced decrease of RNase J expression in tobacco leaves caused significant accumulation of antisense transcripts, suggesting high specificity of RNase J toward these[201]. However, exact mechanism of this recognition remains unknown but it is tempting to speculate that RNase J non-selectively digests RNA molecules and RNA-stabilizing proteins are required for the protection of specific mRNA species.

Maturation of 3'-ends of plastid-encoded mRNAs is believed to be coordinated by the activity of two exoribonucleases: polynucleotide phosphorylase (PNPase) and RNase R (RNR1)[126]. PNPase contains N-terminal transit peptide, responsible for plastid localization, two RNase PH-like core domains as well as KH and S1 domains responsible for RNA binding[202]. Interestingly, as suggested by the crystal structure of its' bacterial counterpart, only one RNase PH-like domain of PNPase is sufficient for its' 3'->5' exonucleolytic activity[203]. Analysis of *Arabidopsis* mutants deficient in PNPase showed that loss of this enzyme leads to chlorosis of leaves as well as slower growth of these plants suggesting a role of PNPase in proper chloroplast development and function. It was demonstrated that PNPase polyadenylates 3'-ends of mRNAs and acts as a hexamer, while the bacterial enzyme functions as a homotrimer. This activity is shared with specialized chloroplast poly(A) polymerase (cpPAP). Notably, plastid transcript polyadenylation, contrary to nuclear polyadenylation, destabilizes mRNAs and leads to their degradation[202]. It was also shown that plant PNPase exonucleotically degrades mRNA in 3'->5' direction and that this activity is stopped upon encountering 3'-end stem loop structure or RNA-stabilizing proteins[204].

RNR1 protein contains RNase II-like (RNB) as well as C-terminal S1 RNA-binding domains and exhibits dual targeting to both mitochondria and chloroplasts in *Arabidopsis*[205]. Loss of RNR1 in *Arabidopsis* plants leads to dwarfism and significant decrease in chlorophyll accumulation. It was demonstrated that the major function of plant RNR1 is processing and degradation of rRNA species, similar to the activity of bacterial RNR1 enzyme[206]. However, thorough analysis of *Arabidopsis* double *rnr1/pnpase* mutants revealed an additional role of RNR1. It was proposed that RNR1 works together with PNPase on mRNA molecules and contributes to 3'-ends generation. It was also demonstrated that RNR1 is not capable of digesting structured dsRNA, a feature typical for RNase II rather than RNase R[205]. Altogether, these results suggest that RNR1 and PNPase work together in 3'-end formation and that their activity is controlled by the presence of RNA secondary structures such as stem loops.

Endonucleolytic RNA processing

Plastid-encoded mRNA molecules are not only processed exonucleolytically but are also digested endonucleotically. In bacteria, RNase E (RNE) is a well characterized enzyme involved in regulation of gene expression through initiating RNA degradation and processing. In *E. coli* this protein works as a part of degradosome, high molecular weight complex, together with PNPase, Rhl B protein as well as glycolytic enzyme enolase. Bacterial RNE is composed of a large N-terminal catalytic domain and C-terminal scaffold for degradosome assembly[207]. Its plant counterpart however, lacks the C-terminal portion and contains transit peptide upstream of the catalytic domain followed by a long linker peptide. *In vitro* analysis showed that plant RNE forms homo-oligomers. Altogether these results suggest that plastids do not contain a bacterial-type

degradosome[208]. A study in *Arabidopsis* showed that loss of RNE causes severe developmental defects, chlorosis of the mutant's leaves, decreased size and defective development of chloroplast. It was also shown to accumulate precursors of multiple RNA species. Interestingly, all aforementioned features were similar to those observed in plants deficient in RHON1 protein, indicated in mRNA maturation and transcription termination. It was found that RNE forms a complex with RHON1 in the chloroplast stroma where they process both rRNA and mRNA precursors[155].

Plastid-encoded pre-tRNAs are thought to be mainly processed at their 5'- and 3'-ends by RNase P and RNase Z enzymes, respectively. Surprisingly, plant RNase P does not exhibit similarity to the bacterial enzyme, specifically it does not contain the catalytic RNA molecule, but does have RNA-binding PPR motifs. Yet, it is still able to perform pre-tRNA processing *in vitro*, suggesting a different mode of action for this protein[209].

In *B. subtilis* RNase Z was shown to process 3'-ends of tRNA precursors which lack the CCA motif which is essential for aminoacylation and interaction with the ribosome[210]. Cleavage performed by RNase Z generates a substrate for addition of this motif by tRNA nucleotidyltransferase[211]. Interestingly, *Arabidopsis* genome encodes four RNase Z genes. Three of them encode proteins localized to the cytoplasm, nucleus and mitochondria and their loss does not cause lethality. However, the fourth RNase Z localizes to chloroplasts and is required for plant viability, which suggests its key role in plastid tRNA metabolism[212].

CSP41 is an RNA-binding protein unique to photosynthetic organisms. Interestingly, cyanobacteria contain only one gene encoding this protein, while photosynthetic eukaryotes encode two CSP41 proteins[213]. These proteins contain Rossman fold

domains, similar to bacterial epimerase/dehydratase proteins, and were proposed to be involved in non-specific RNA binding and processing[214]. Analysis of *Arabidopsis* mutants revealed that the loss of CSP41A does not cause any observable phenotype. However, plants deficient in CSP41B exhibited delayed growth, chlorosis, and both chloroplast and embryo development defects. Decreased accumulation of PEP-transcribed RNAs was also observed. CSP41B was shown to directly interact with PRIN2, forming a DNA-binding complex *in vitro*. Therefore, it was proposed that CSP41B together with PRIN2 regulate the activity of PEP but the exact molecular mechanism remains unknown[215].

RNase H proteins

Another group of enzymes likely to be involved in plastid RNA metabolism are RNase H1 proteins. RNase H1 proteins hydrolyze the phosphodiester bonds of the RNA strand in a RNA:DNA hybrid. This enzymatic activity requires at least four consecutive ribonucleotide in the substrate[216]. Human RNase H1 protein requires Mg^{2+} ions for its catalytic activity[217] but enzymes from other organisms have been shown to use Mn^{2+} ions as co-factors[218].Genes encoding RNase H1 proteins are highly conserved and were found in genomes of almost all eukaryotes and the majority of prokaryotes. These enzymes contain N-terminal transit peptide followed by catalytic domain and RNA:DNA hybrid binding domain at the C-terminus[216,219]. Additionally, eukaryotic RNase H1s often contain N-terminal mitochondrial targeting sequence (MTS). In this case a single gene expresses two splice variants, where one containing the MTS localizes to the mitochondria, while the second one does not contain MTS and localizes to the nucleus[220,221].

Canonical substrates of the RNase H1 enzymes, RNA:DNA hybrids, are formed during replication, transcription but also proposed to serve regulatory functions in gene expression[222–224]. Importantly, RNA:DNA hybrids need to contain at least four consecutive ribonucleotides in order to be recognized and cleaved by RNase H1[225]. The best characterized role of RNase H1 in eukaryotes is its involvement in removing RNA primers during mitochondrial DNA replication in mouse[89,220]. Additionally, it was demonstrated that RNase H1 binds telomeres and digests RNA:DNA hybrids formed between long non-coding RNA TERRA and telomeric DNA fragments[226]. Loss of RNase H1 leads to increased homologous recombination (HR) rates, replicative stress and rapid telomere loss. Conversely, overexpression of RNase H1 caused reduced HR, diminished replicative stress and telomere shortening[226]. In yeast RNase H1 has been shown to be involved in genome maintenance but also to be partially redundant with RNase H2 complexes[227]. It has been shown that loss of *Arabidopsis* chloroplast localized RNase H1 (RNH1C) leads to chlorosis, stunted growth and chloroplast development defects and that this protein interacts with gyrase and is involved in plastid DNA replication[228]. However, several essential questions about plastid RNase H1 and its nuclear and mitochondrial paralogs remained unanswered. These include the evolutionary origin of multiple RNase H1 genes in plants, biochemical activity of plant RNase H1-like proteins, role of RNase H1 in plant development and DNA replication. These questions are answered in Chapter II.

In addition to RNase H1 enzymes, plants also contain RNase H2[229]. RNases H2 are nucleus-localized heterotrimeric complexes with one catalytic subunit and two subunits serving mainly structural roles. These complexes are responsible for the removal of single ribonucleotides incorporated into dsDNA and therefore have been implicated in

DNA repair. However, they have been also shown to digest longer RNA:DNA hybrids on a genome-wide scale as well as to be involved in Okazaki fragment maturation. Importantly, RNases H2 have been proposed to serve as a main machinery resolving RNA:DNA hybrids in the genome, while RNase H1s are proposed to work in locus-specific manner[216,219,227,230]. In plants, loss of RNase H2 leads to increased homologous recombination rates as well as higher frequency of ribonucleotides in the nuclear DNA. Additionally, it was demonstrated that plants deficient in RNase H2 are more susceptible to replication stress. Interestingly, loss of WEE1 cell cycle checkpoint kinase reversed this phenotype suggesting that WEE1 and RNase H2 interact and together affect cell cycle progression. It was proposed that in *Arabidopsis* RNase H2 may act as a sensor of DNA damage which may cause cell cycle arrest if the level of ribonucleotides in DNA is too high[229,231]. However, direct evidence confirming this speculation remains to be found.

DeCL proteins

Another group of proteins hypothesized to be involved in plastid RNA metabolism are DeCL (Defective Chloroplasts and Leaves) proteins. They form an interesting but poorly characterized protein family. Initially described in tomato plants, these proteins contain one Domain of Unknown Function (DUF3223) which has a bacterial, uncharacterized, counterpart[232]. It was reported that loss of DeCL proteins leads to embryo lethality and that these proteins are involved in rRNA maturation, but presented data is at least partially inconclusive[233,234].

In addition to DeCL proteins at least three plant proteins have been found to contain DUF3223 but both of them localize to the nucleus and not to plastids. DOMINO1 has

been shown to localize to the nucleolus and its loss leads to accumulation of rRNA precursors as well as decondensation of the nucleolus. Therefore it was proposed to process ribosomal RNA species[235]. Another nuclear protein containing DUF3223 is the largest subunit of plant-specific RNA Polymerases IV and V (Pol IV and Pol V)[236,237]. These polymerases are central enzymes involved in transcriptional gene silencing in plants, known as RNA-directed DNA methylation (RdDM)[238,239]. This mechanism is mainly responsible for repression of transposon expression through DNA methylation but has been also shown to regulate expression of protein-coding genes[240,241]. The exact role of DUF3223 in Pol IV and V remains unknown but it was shown to be required for transcription *in vivo* as well as for interaction of Pol V with the RRP6L exonuclease[242]. We hypothesize that elucidating molecular and biological roles of DeCLs in plastids will allow for better understanding of nuclear DUF3223-containing proteins.

Functional analysis of *Arabidopsis* DeCL proteins is described in Chapter III. In order to better understand the role of DeCL proteins we performed a thorough phylogenetic analysis of these proteins. Additionally, using genetic, biochemical and microscopy tools we aimed to determine the molecular function as well as specific substrates of organellar DeCLs. Finally, we propose a speculative model explaining the role of this protein family.

CHAPTER II

RNase H1 is required for *Arabidopsis thaliana* embryonic development

The content of this chapter is currently under review in Plant Cell and Physiology journal. Sebastian Chamera and Aleksandra Kmera performed *in vitro* assay for enzymatic activity of RNase H1s. Pragya Khurana performed analysis of nuclear, mitochondrial and plastid DNA content.

Abstract

RNase H1 is an endonuclease specific towards the RNA strand of RNA:DNA hybrids. Members of this protein family are present in most living organisms and are essential for removing RNA that base pairs with DNA. It prevents detrimental effects of RNA:DNA hybrids and is involved in several biological processes. We show that *Arabidopsis thaliana* contains at least three RNase H1-like proteins originating from two gene duplication events and alternative splicing. These proteins have the canonical RNase H1 activity, which requires at least four ribonucleotides for enzymatic activity. One of those proteins is nuclear, one is localized to plastids, one is localized to mitochondria. While the nuclear RNase H1 is dispensable for development under normal growth conditions, the presence of at least one organellar RNase H1 is required for embryonic development. The plastid protein RNH1C affects plastid DNA copy number and sensitivity to replicative stress. This suggests that three genomes present in each plant cell are served by at least one specialized RNase H1 protein.

Significance

We integrate phylogenetic, genetic, molecular and physiological approaches to characterize RNase H1-like genes in *Arabidopsis* and their protein products. We provide detailed evolutionary history of all plant RNase H1s and for the first time report biochemical activity of plant RNase H. We present direct evidence of RNase H1 requirement for embryonic development as well as demonstrate that plastid-localized RNase H1 is involved in DNA replication. Our results provide fundamental insight into the biochemical function of plant RNase H1s as well as biological processes they are involved in.

Introduction

Double stranded nucleic acids, which contain deoxyribonucleotides on one strand and one or more ribonucleotides on the other strand are known as RNA:DNA hybrids. They are common byproducts of replication, transcription and other processes. Ribonucleotides within RNA:DNA hybrids are specifically removed by a class of endonucleases known as RNases H[216]. RNases H2 are multisubunit complexes capable of removing even individual ribonucleotides incorporated in double stranded DNA and have been studied in various eukaryotes, including plants[229,231]. RNases H1 are monomers[243] and require at least four ribonucleotides incorporated into double stranded DNA to bind and digest the substrate[244].

Among the substrates of RNases H1 are R-loops[245] (RNA:DNA hybrid and a displaced ssDNA strand) which are often formed during transcription and replication[223,246,247]. These structures have also been implicated in DNA repair[222,248,249],

telomere maintenance[226], IgG class-switch recombination[250] and regulation of gene expression[224,247]. RNase H1 digests ribonucleotides within its substrate in a metal ion-dependent manner[225], leading to single stranded DNA formation. The main feature of RNases H1 is the presence of the catalytic domain[225,251]. Additionally, some bacterial and most eukaryotic RNase H1 proteins contain a RNA:DNA hybrid binding domain (HBD)[252].

Genes encoding RNase H1 proteins are present in the vast majority of living organisms, including Archea[253,254], Bacteria[219] and all kingdoms of Eukarya[216]. They are not essential in prokaryotes and lower eukaryotes but are required for survival in higher eukaryotes[216]. Eukaryotic genomes usually contain unique RNase H1 genes[216,219], which may however be subject to alternative splicing[221]. *Arabidopsis thaliana* RNases H1 are encoded by three different genes with different predicted subcellular localizations[228].

Among the three RNase H1 proteins in *Arabidopsis thaliana,* only the chloroplast-localized paralog has been studied so far[228]. It has been shown to be essential for proper plastid development by maintaining the integrity of chloroplast DNA. It works with its interacting partner, DNA gyrase, to resolve transcription-replication conflicts and prevent DNA damage[228]. The role of the remaining two proteins remains unknown beyond the presumption that they resolve R-loops, which are relatively common in the *Arabidopsis* genome[255].

Here, we characterize all RNase H1 proteins detectable in the *Arabidopsis* genome. We identify two ancient gene duplication events, which led to the formation of RNase H1 proteins targeted to various cellular compartments in monocots and dicots. A more recent duplication and alternative splicing produced four RNase H1 proteins in

Arabidopsis thaliana, one of which is targeted to the nucleus, one to the chloroplast, one to the mitochondria and its splice variant which surprisingly does not localize to the nucleus upon the loss of MTS, contrary to the phenomenon observed in mammals[220,221]. The proteins localized to endosymbiotic organelles are required for proper embryonic development. These proteins exhibit canonical RNase H1 activity and are involved in nucleic acid metabolism.

Results

Origin of angiosperm RNases H1

Plant genomes have been shown to encode multiple RNase H1-proteins with different subcellular localizations[228]. Previous phylogenetic analysis indicated that these paralogs originate from recent gene duplication events[228], however the exact timing and order of those events remained unresolved. To determine the evolutionary origins of plant RNase H1-like proteins we performed their in-depth phylogenetic analysis. We first used BLAST to identify all plant proteins, which display sequence similarity to *Arabidopsis thaliana* AtRNH1C[228] and contain both RNase H1 and RNA:DNA hybrid binding (HBD) domains. The identified proteins were subject to a simultaneous Bayesian alignment and phylogenetic analysis implemented using BAli-Phy package in order to integrate over the uncertainty in both the phylogeny and alignment[256]. In parallel, we predicted the subcellular localization of each identified protein using the TargetP prediction tool[257]. RNase H1-like proteins from angiosperms grouped into four distinct clades (Fig. 2.1, Fig. S2.1). Within both monocots and dicots there are proteins with mostly nuclear and mostly organellar (chloroplast or mitochondrial) predicted localization (Fig. 2.1, Fig. S2.1, Fig. 2.2A). Phylogenetic relationships between these proteins

indicate that the common ancestor of monocots and dicots had one RNase H1-like protein. Two independent gene duplication events in early evolution of monocots and dicots led to both acquiring at least two RNase H1 proteins with distinct subcellular localizations.

Origin of four RNases H1 in *Arabidopsis*

Arabidopsis thaliana has been shown to contain three genes encoding RNase H1-like proteins: RNH1A (AT3G01410), RNH1B (AT5G51080) and RNH1C (AT1G24090). Products of these genes localize to the nucleus, mitochondria and chloroplasts, respectively[228]. While the split into nuclear and organellar proteins occurred early in dicot evolution (Fig. 2.1), the origin of two organellar proteins remains unknown. To determine the relationship between RNH1B and RNH1C we analyzed the phylogeny of RNase H1-like proteins in *Brasicaceae* (Fig. 2.2B). The phylogenetic tree identified one clade of nuclear and two distinct clades of organellar RNase H1-like proteins (Fig. 2.2B). This indicates that a second diversification event occurred in the common ancestor of *Brasicaceae*, which led to the formation of two organellar proteins. While in *Arabidopsis thaliana* these proteins are localized to mitochondria and chloroplasts[228], these specific localizations cannot be conclusively predicted for other *Brasicaceae*.

Multiple subcellular localizations of proteins produced from unique RNase H1 genes in animals are commonly determined by alternative splicing[220]. Araport11 genome annotation[258] suggests a similar mechanism in *Arabidopsis thaliana*, where three splice variants of RNH1B (AT5G51080) have been identified (Fig. 2.2C). We partially confirmed these annotations using 5'RACE (Fig. 2.2D). Subcloning of RACE products followed by Sanger sequencing confirmed the presence of transcripts very similar to

splice variants AT5G51080.2 and AT5G51080.3 (Fig. S2.2). Although AT5G51080.1 has not been identified by sequencing, it is predicted to encode a protein identical to AT5G51080.2 and the difference is limited to the presence of an intron in 5'-UTR. The third splice variant (AT5G51080.3) has a truncated mitochondrial presequence region (Fig. 2.2E). TargetP predicted that this variant does not localize to mitochondria or chloroplasts, suggesting nuclear localization. In order to confirm this predicted localization, we transiently expressed GFP-tagged AtRNH1B.3 under control of a strong, constitutive 35S promoter in *A. thaliana* protoplasts (Fig. 2.2FG). While AtRNH1A shows the expected nuclear localization (Fig. 2.2G), for AtRNH1B.3 we did not observe signal typical to nuclear localization (Fig. 2.2F). Instead, the fusion protein localized in small distinct foci, similar to the signal reported for AtRNH1B.1-GFP[228]. We conclude that *Arabidopsis thaliana* contains at least three RNase H1-like proteins originating from two independent gene duplication events. Additionally, one of those proteins may be present in two splice variants.

RNase H1-like proteins from *Arabidopsis* have canonical RNase H1 activity

Although *Arabidopsis thaliana* RNase H1-like proteins have extensive sequence similarity with RNase H1 proteins, their exact enzymatic activity remains unknown. RNH1C has been shown to remove binding sites of S9.6 antibody from chloroplast DNA[228]. This however, does not conclusively show RNase H1 activity of this protein. To determine if RNH1A and RNH1B are indeed RNases H1, we expressed their truncated versions, fragments from the beginning of HBD until STOP codon, and incubated the recombinant proteins with oligonucleotide substrates (Fig. 2.3AC). RNH1B.1/2 and RNH1B.3 differ only in the presence of the N-terminal pre-sequence, which is expected

to be proteolytically removed from the preprotein during protein import into the mitochondria[259] and would not be included in our recombinant protein. Therefore, only one form of RNH1B was subject to the enzymatic assay. We also expressed a truncated version of RNH1A protein carrying mutation in its predicted catalytic active site (Fig. 2.3B) as well as human RNase H1 (Fig. 2.3D) known to possess the canonical RNase H1 activity[251]. Double stranded DNA and RNA was not digested by the recombinant proteins (Fig. 2.3A-D two right-most columns). The same was true for oligonucleotides containing one or two ribonucleotides, which differentiates RNase H1 from RNase H2 (Fig. 2.3A-D two left-most columns). Oligonucleotides containing four ribonucleotides were however digested by both RNH1A and RNH1B, as well as human RNase H1 (Fig. 2.3ACD center column). Active site mutant of RNH1A did not exhibit any detectable catalytic activity (Fig. 2.3B). We did not test RNH1C or catalytic mutant of RNH1B because despite significant efforts we were unable to produce soluble recombinant proteins. These results indicate that RNH1A and RNH1B have the canonical RNase H1 activity. Because of the extensive sequence similarity (Fig. S2.3A), RNH1C is likely to have the same enzymatic activity.

RNH1A and RNH1B do not affect development

RNH1C has been shown to be required for proper chloroplast development[228]. To determine the roles of all three RNase H1 encoding genes, we obtained T-DNA mutants in RNH1A (SALK_150285C) and RNH1B (SAIL_1174_C11) as well as the previously published RNH1C (SAIL_97_E11). Single mutants *atrnh1a* and *atrnh1b* did not exhibit any obvious developmental phenotypes (Fig. 2.4A-C). *atrnh1c* mutant had the expected

pale leaf and dwarf phenotype (Fig. 2.4D). The pale green phenotype attributed to the loss of RNH1C was confirmed by chlorophyll content quantification (Fig. 2.4J).

To determine if the studied T-DNA mutants may express truncated proteins, we performed RT-PCR with primers upstream or downstream of the T-DNA insertion sites (Fig. 2.4K). All mutants had strongly and significantly reduced RNA accumulation downstream of T-DNA insertions (Fig. 2.4LM). This indicates that these mutants are unlikely to produce truncated proteins containing the C-terminal RNase H1 domain. RNA accumulation upstream of T-DNA was reduced in *atrnh1a*, unchanged in *atrnh1c* and strongly increased in at*rnh1b* (Fig. 2.4LM). This indicates that a truncated N-terminal fragment is unlikely to be produced in *atrnh1a*. Truncated N-terminal fragments may be produced in *atrnh1b* and *atrnh1c*, however *atrnh1b* is not expected to contain full length RNase H1 domain. On the other hand, *atrnh1c* may produce a truncated protein including the RNase H domain (Fig. 2.4K), which indicates that the C-terminal part of the protein is important for its function. The strong increase of upstream RNA accumulation in *atrnh1b* may indicate the presence of an autoregulatory mechanism within *RNH1B*.

Presence of RNH1B or RNH1C is required for viability

Because RNases H1 are required for viability in animals[220], we tested the phenotypes of all combinations of RNH1 double mutants. The *atrnh1a*, *atrnh1b* double mutant did not show any visible developmental phenotypes (Fig. 2.4E) and the phenotype of the *atrnh1a*, *atrnh1c* double mutant was similar to the *atrnh1c* single mutant (Fig. 2.4F). Phenotypic similarity of *atrnh1c* single mutant and *atrnh1a*, *atrnh1c* double mutant was confirmed by chlorophyll content quantification (Fig. 2.4J). We were unable to generate *atrnh1a, atrnh1b, atrnh1c* triple mutants but, interestingly plants

deficient in RNH1A and RNH1B and heterozygotic for RNH1C did not exhibit any obvious developmental phenotypes (Fig. 2.4G). Additionally, we generated triple *atrnh1a, atrnh1b, atrnh1c* mutant rescued with RNH1C-GFP under control of its native promoter which also did not exhibit any obvious phenotype (Fig. 2.4HI).

Despite several attempts we were unable to obtain the *atrnh1b*, *atrnh1c* double mutant. Among 211 F1 plants obtained from self-pollination of *atrnh1b*$^{-/-}$*atrnh1c*$^{+/-}$ we found no viable double homozygous mutant ($p < 5*10^{-17}$, Chi square test). However, self-pollination of *atrnh1a/b/c*$^{+/-}$ resulted in 25.5% of seeds not developing properly (Fig. 2.5A-D), which is consistent with 25% embryo lethality ($p < 0.9$, Chi square test) and significantly more than expected without a lethal phenotype ($p < 5*10^{-6}$, Chi square test). This indicates that most likely double mutants in *RNH1B* and *RNH1C* are embryo lethal. Importantly, triple mutant plants rescued with RNH1C::RNH1C-GFP did not produce aborted seeds (Fig. 2.5CD). These improperly developing seeds contain aborted embryos which stop developing at approximately heart-shape stage of embryonic development (Fig. 2.5EF). These results suggest that while the nuclear RNH1A is not required for viability, the presence of at least one organellar protein RNH1B or RNH1C is essential.

RNH1C is involved in chloroplast nucleic acid metabolism

Because RNH1C has been shown to be involved in DNA maintenance[228], we hypothesized that RNases H1 expressed from the three *Arabidopsis* genes affect DNA copy numbers in various cellular compartments. To test this hypothesis, we performed real time PCR with multiple primers specific to sequences throughout the nuclear, mitochondrial and chloroplast genomes. Average signal from three to five primer pairs

specific to each genome was used as an estimate of their relative copy number[260]. *atrnh1c* mutant and *atrnh1a, atrnh1c* double mutant contained approximately three times more chloroplast DNA than wild type mutant plants (Fig. 2.6A-C). This result indicates that RNH1C protein is involved in controlling DNA copy number in chloroplasts, which is consistent with its postulated role in plastid genome maintenance[228]. Surprisingly, *atrnh1a*, *atrnh1b* mutants and the *atrnh1a*, *atrnh1b* double mutant did not affect the relative content of the three genomes (Fig. 2.6A-C). This indicates that only RNH1C affects content of the plastid genome relative to nuclear and mitochondrial genomes, which may be explained by a partial redundancy of RNases H1, RNase H2 complex[261], topoisomerases[262] or other factors.

The increased amount of chloroplast DNA in *atrnh1c* mutants might be a result of changes in DNA replication. To test this possibility, we treated wild type plants and RNase H1 mutants with replication stress and measured root length. To induce replication stress, we applied 2mM hydroxyurea[229–231], which is known to inhibit replication by inhibiting rNTP reductase and therefore decreasing the amount of available dNTPs[263]. In parallel, we used 40μM aphidicolin which inhibits replication through a different mechanism, specifically by inhibiting DNA polymerase ability to bind dCTP[264]. Treatment with 2mM hydroxyurea caused a decrease of root length in wild type plants to approximately 40% of root length in untreated plants (Fig. 2.6DF). *rnh1a* and *rnh1b* mutants had no effect on hydroxyurea sensitivity (Fig. 6G. However, roots of *atrnh1c* mutant and *atrnh1a*, *atrnh1c* double mutant grew to less than 30% of root length in untreated plants (Fig. 2.6G). Importantly, *atrnh1a, atrnh1b, atrnh1c* triple mutant expressing RNH1C::RNH1C-GFP did not exhibit increased sensitivity to HU (Fig. 2.6DG). Similarly to hydroxyurea, 40μM aphidicolin treatment caused a decrease in root

length of wild type plants to approximately 40% (Fig. 2.6EH). Loss of RNH1A had no effect on root length upon replicative stress caused by aphidicolin treatment and roots of plants deficient in RNH1C grew to approximately 30% of the root length of untreated plants. Interestingly, loss of RNH1B lead to the increase of root length to approximately 50% of length observed in untreated plants (Fig. 2.6H). Likewise, *atrnh1a, atrrnh1b, atrnh1c* RNH1C::RNH1C-GFP plants did not exhibit higher aphidicolin sensitivity than wild type plants. Altogether, these results suggest that RNH1C is likely to affect DNA replication.

Discussion

RNase H1 proteins are known to digest the RNA component of RNA:DNA hybrids and this substrate must contain at least four sequential ribonucleotides[225]. In contrast, RNase H2 complexes require only a single ribonucleotide incorporated into double-stranded DNA for activity[216,265]. Therefore, RNase H2 has a broader substrate range and is partially redundant with RNase H1[227]. We demonstrate for the first time that *Arabidopsis thaliana* RNH1A and RNH1B have the canonical RNase H1 activity. Because the presence of RNase H2 complexes has been previously shown[229,231], this indicates that plants, like other eukaryotes, have both RNase H1 and H2. Our results suggest that all three RNase H1 genes identified in the *Arabidopsis thaliana* genome encode canonical *bona fide* RNase H1 proteins.

Arabidopsis thaliana contains at least three RNase H1 proteins. Their evolutionary origin may be traced down to three events. The first event occurred in the common ancestor of dicots, where nuclear and organellar paralogs have been formed. An independent event in the common ancestor of monocots (or possibly the common

ancestor of all angiosperms) led to the formation of similar, yet evolutionarily independent nuclear and organellar paralogs in monocots. Nuclear and organellar RNases H1 do not display any obvious structural differences beyond the presence or absence of a transit peptide/presequence. Therefore, it is unknown if they are functionally distinct beyond having different subcellular localizations and expression patterns.

The second event in evolution of RNase H1 proteins in *Arabidopsis thaliana* occurred in the common ancestor of *Brassicaceae*. The ancestral organellar protein diversified into two proteins, which are represented by RNH1B and RNH1C in *Arabidopsis*. Although these proteins are localized to mitochondria and chloroplasts[228], their orthologs in *Brassicaceae* do not have consistent predicted localizations and the functional impact of this event remains unknown. The third event is alternative splicing of RNH1B, which likely results in translation of proteins with distinct N-termini which is reminiscent of metazoan RNases H1[221]. Because we did not observe a nuclear localization pattern of the RNH1B.3 splice variant, the mechanism of its localization remains unknown and may possibly be attributed to internal MTS-like signals (iMTS-Ls)[266]. Similarly, biological importance of RNH1B.3 remains to be elucidated. Overall, our results are consistent with recent evidence that each genome within *Arabidopsis thaliana* cells has at least one RNase H1[228]. Additionally, most eukaryotes, including *Arabidopsis thaliana* contain RNase H2 complexes, which are nuclear localized and at least partially redundant with RNases H1[227,229,265]. This may suggests that the nuclear genome uses a combination of several activities resolving RNA:DNA hybrids.

Our results indicate that at least one organellar RNase H1 is needed for proper embryonic development. This is consistent with data from metazoans, where RNase H1

is crucial for proper mitochondrial DNA replication during embryo development[89,90,220]. Viability of single *atrnh1b* and *atrnh1c* mutants may speculatively be explained by dual (mitochondrial and plastid) localization of those proteins, which is not an uncommon phenomenon[267,268]. Alternatively, defects in both mitochondria and chloroplasts may have a synergistic effect in embryonic development.

RNase H1 has been shown to be required for genome maintenance in chloroplasts[228]. Our observation that plastid DNA copy number is substantially increased in *atrnh1c* is counterintuitive, especially since this mutant has a reduced number of chloroplasts[228]. It may indicate that genome instability in *atrnh1c* leads to overamplification of the entire genome. Because the plastid genome is copied by a variety of mechanisms, including recombination[269] and RNase H1 substrates, R-loops, were shown to increase recombination rates[248,250], this overamplification may rely on homologous recombination. Alternatively, it is possible that RNA:DNA hybrids may be used by plastid DNA polymerase as primers for replication. Interestingly, loss of mitochondrial RNH1B does not affect the level of mitochondrial DNA. This may suggest that, contrary to animals, plant mitochondria contain enzymes at least partially redundant with RNH1s.

The reported role of RNH1C in release of replication-transcription conflicts and chloroplast DNA integrity[228] is consistent with our observations that *atrnh1c* mutation increases sensitivity to hydroxyurea. This may be interpreted as evidence of disrupted DNA replication in *atrnh1c*. This result should however be interpreted carefully because hydroxyurea inhibits synthesis of deoxyribonucleotides and increases misincorporation of ribonucleotides[230,263], which may have a replication-independent effect on RNase H1-dependent processes. Additionally, hydroxyurea may cause oxidative stress [270–272],

which could also have a replication-independent effect. However, the results of replicative stress experiment with use of aphidicolin, drug which inhibits DNA replication through different mechanism than HU[264], strongly suggest that RNH1C is indeed involved in DNA replication.

Our results raise questions about the roles of nuclear and mitochondrial RNases H1 in *Arabidopsis thaliana*. Single mutations in genes encoding these proteins do not result in visible phenotypes, which may likely be attributed to protein redundancy and localization to multiple cellular compartments. Resolving the roles of those proteins remains an important goal for future studies.

Materials and Methods

Plant material and oligonucleotides

Plant lines used in this study: Col-0 (CS 70000), *atrnh1a* (SALK_150285C), *atrnh1b* (SAIL_1174_C11), *atrnh1c* (SAIL_97_E11) and crosses between abovementioned. Oligonucleotides used for expression analysis: P1: 5'-TTTAGTTTGGGTTGATGGGTTCC-3' + 5'-CACACTCATTGCAGGATGTGATAC-3'; P2: 5'-ATCTCTTAGACGGGGAAGATTTGT-3' + 5'-AATGCAAGTCATGTCAAAGATGAT-3'; P3: 5'-TCAATAATGCAAGTTTCATATGAGGT-3' + 5'-GTTTGGAGCTCTTACACCTTGTCT-3'; P4: 5'-ATCCCTTATAAACGCTAACTGGAG-3' + 5'-TCAAGTTGGATCTTCGGTTTATG-3'; P5: 5'-TACACCATGTCTTTTCCAGGAG-3' + 5'-GATATAGAAGCTGAAGGAAGTTGATCT-3'; P6: 5'-AAAGATCCGGAGTTACACACTAGC-3' + 5'-CATTCTGGTTCTTCACCAGTTTCT-3'. Sequences of oligonucleotides used for organellar DNA content quantification were previously described by Kim et al[260].

Phylogenetic analysis

Sequences of putative plant RNH1 proteins were retrieved from Phytozome[273], CoGE[274], Ensembl Plants[275], PLAZA[276] and 1KP[277] through BLAST search and aligned in MAFFT[278]. Only sequences containing full length hybrid binding domain (HBD), RNASEH1-like domain and four amino acids (DEDD) crucial for catalytic activity[251], were kept. BAli-Phy[279] with default settings on CIPRES[280] platform was used for phylogenetic analysis. Consensus tree was generated using default burn-in values and visualized using FigTree. Subcellular localization was predicted by TargetP[257].

Protein expression and purification

cDNAs encoding fragments from the beginning of HBD until STOP codon of wild type and catalytic mutants of *Arabidopsis thaliana* RNH1A and RNH1B were cloned into pET28-SUMO vector. Coding sequence of *Homo sapiens* RNase H1 was cloned into pET15b vector.

All *A. thaliana* proteins were produced in *E. coli* BL21* cells induced with 0.4 mM 1-thio-β-D-galactopyranoside at 18°C. The cells were then harvested and suspended in 40 mM Tris-HCl (pH=8.0), 200 mM NaCl, 5% glycerol, 10 mM imidazole and 1.4 mM β-mercaptoethanol and incubated on ice in the presence of 1 mg/ml lysozyme and protease inhibitor cocktail. Following sonication, the cleared lysate was loaded onto a HisTrap column (GE Healthcare) equilibrated with 40 mM Tris-HCl (pH=8.0), 500 mM NaCl, 5% glycerol, 10 mM imidazole and 1.4 mM β-mercaptoethanol. After a wash step with 60 mM imidazole, the protein was eluted with 300 mM imidazole. The eluted fraction was dialyzed overnight against 40 mM Tris-HCl (pH=8.0), 500 mM NaCl, 5%

glycerol, 1 mM dithiothreitol (DTT) with SUMO protease to remove SUMO-His tag (due to extreme difficulties with purification of wild type AtRNH1B we decided not to remove the tag) and loaded onto a second HisTrap (GE Healthcare) equilibrated with the buffer containing 40 mM Tris-HCl (pH=8.0), 500 mM NaCl, 5% glycerol, 10 mM imidazole and 1.4 mM β- mercaptoethanol. Proteins were further purified on a Superdex 200 column (GE Healthcare) equilibrated with 40 mM Tris (pH 8.0), 500 mM NaCl, 5% glycerol, 0.5 mM EDTA and 0.5 mM DTT. Selected fractions were pooled and concentrated.

H. sapiens RNase H1 was expressed in *E. coli* BL21* cells induced with 0.4 mM 1-thio-β-D-galactopyranoside at 18°C. The cells were then harvested and suspended in 40 mM NaH_2PO_4 (pH=7.0), 1M NaCl, 5% glycerol, 10 mM imidazole and 2.8 mM β- mercaptoethanol and incubated on ice in the presence of 1 mg/ml lysozyme and protease inhibitor cocktail. Following sonication, the cleared lysate was loaded onto a HisTrap column (GE Healthcare) equilibrated with 40 mM NaH_2PO_4 (pH=7.0), 1M NaCl, 5% glycerol, 10 mM imidazole and 2.8 mM β-mercaptoethanol. After a wash step with 60 mM imidazole, the protein was eluted with 300 mM imidazole. The eluted fraction was dialyzed overnight against 40 mM NaH_2PO_4(pH=7.0), 100mM NaCl, 5% glycerol, 0.5 mM EDTA. His-tag was removed by thrombin digestion and then protein was loaded onto second HisTrap column. Protein was further purified on a Superdex 75 column (GE Healthcare) that was equilibrated with 40 mM NaH_2PO_4 (pH=7.0), 500 mM NaCl, 5% glycerol, 0.5 mM EDTA and 1 mM DTT. Selected fractions were pooled and concentrated.

In vitro activity of RNH1s

Reaction was conducted for 30min at 37°C in buffer containing 75mM NaCl, 20mM HEPES pH=7.0, 5% glycerol, 1mM DTT, 5mM MgCl2 with increasing enzyme:substrate ratio (black bar) indicated on top of the gels and stopped by addition of EDTA to final concentration of 40mM. First lane of each gel is a control (no enzyme added). Products were analyzed on 15% denaturing TBE-urea polyacrylamide gels containing 20% formamide and visualized by fluorescence readout.

Chlorophyll content measurement

Chlorophyll content was determined as described by Lichtenthaler[281]. Briefly, 1g of frozen 2-3 week old seedlings was ground in liquid nitrogen and chlorophyll was extracted with 100% acetone. Samples were centrifuged at 10000 x g for 10min at 4^0C and absorbance was measured at 645 and 662nm. Chlorophyll concentration was calculated using following formula: C_{a+b}(µg/ml) = 18.09 x A645 + 7.05 x A662.

Microscopic observations

Protoplasts were isolated from young leaves with protocol by Yoo et al[282]. Briefly, *A. thaliana* leaves were finely sliced and digested with macerozyme and cellulase for 2 hours in the dark. Isolated protoplasts were transformed in 20% PEG 4000 with 10µg of plasmid pEG103 carrying cDNA encoding RNH1B.3.

Mature siliques of Col-0 and *rnh1b*$^{-/-}$ / *rnh1c*$^{+/-}$ plants were destained in 70% ethanol and photographed under preparative microscope. For embryos observations siliques produced by *rnh1b*$^{-/-}$ / *rnh1c*$^{+/-}$ mutant were dissected, seeds were split based on

morphology and cleared for 2h at room temperature with Visikol[283]. Embryos inside seeds were visualized using Nomarski optics.

5'RACE

5'RACE was performed using Invitrogen™ GeneRacer™ SuperScript™ TA Cloning™ TOPO™ GeneRacer Kit with SuperScript III RT and TOPO TA Cloning™ Kit for Sequencing. All procedures were performed as described by manufacturer. Cloning products were subjected to Sanger sequencing and results were aligned to *A. thaliana* mRNA collection in BLAST.

Generation of transgenic plants

The genomic DNA sequence including promoter region of RNH1C was cloned into pENTR/D-TOPO vector (Invitrogen) to produce clones as described by manufacturer. Resulting plasmid was incubated with the destination vector pMDC107[284] and LR Clonase™ II Enzyme Mix to obtain *RNH1C::RNH1C-GFP*. Constructed plasmid was introduced into the GV3101 strain of *Agrobacterium tumefaciens* and transformed into *atrnh1c* plants by the floral dip method[285]. Plants carrying the transgene were selected based on their resistance to hygromycin and crossed with *atrnh1a, atrnh1b* double mutant to generate *atrnh1a*$^{+/-}$*, atrnh1b*$^{+/-}$*, atrnh1c* $^{+/-}$ RNH1C::RNH1C-GFP. Resulting progeny was self-pollinated and desired *atrnh1a, atrnh1b, atrnh1c* RNH1C::RNH1C-GFP were identified in the next generation by genotyping.

IP-Western blot

2 grams of young leaves were ground in liquid nitrogen, resuspended in 7ml of the Protein Extraction Buffer (20mM Tris-HCl pH=8.0, 300mM NaCl, 5mM $MgCl_2$, 5mM DTT, 1% Plant Protease Inhibitor) and centrifuged at 16,000 x g for 15min at 4^0C. Resulting supernatant was immunoprecipitated with 50µl of GFP-Trap at 4^0C overnight. After IP the beads were washed four times with wash buffer (200mM Tris-HCl pH=8.0, 300mM NaCl, 5mM $MgCl_2$, 5mM DTT, 1mM PMSF, 0.5% IGE-PAL 630). Proteins were eluted in 2x Laemmli buffer at 99^0C for 10min, separated in 12% SDS-PAGE and subjected to Western Blot. Detection was performed with primary mouse α-GFP antibody (SIGMA, cat. #: 11814460001) and secondary goat α-mouse antibody Dylight 800 conjugate (Fisher, cat. #: SA535521).

Organellar DNA content quantification and expression analysis by qPCR

Total DNA was isolated from 2-3 weeks old seedlings with DNeasy Plant Mini Kit (QIAGEN). 1ng of DNA per qPCR reaction was used as template. For expression analysis total RNA was extracted with RNeasy Plant Mini Kit (QIAGEN) from 2-3 weeks old seedlings and 500ng was treated with DNaseI (Ambion) and used for reverse transcription. 0.5ng of cDNA was used as template in qPCR.

DNA content quantification was performed as described[260], except that a subset of three primer pairs were used to quantify nuclear DNA. qPCR signals were averaged over three technical replicates, three biological replicates and three or five primer pairs. Values for all tested genotypes are plotted relative to Col-0 wild type.

qPCR data were analyzed using a Bayesian hierarchical model, implemented using the brms package[286]. We centered the observed Ct values, and then fitted the re-

centered Ct values as a sum of population-level terms for the primer pair, genotype, and primer/genotype interactions, plus a group level term for each primer/biological replicate pair. Student-distributed errors were assumed, and all priors were left at the BRMS defaults. Fits were performed for 6 Monte Carlo chains with 2000 iterations per chain. Convergence was assessed based on the observed Rhat values (all <1.01) and manual inspection of the posterior predictive distribution. Inferences stated in the text arise from inspection of the posterior 95% credible intervals of the pertinent parameters.

Replicative stress assay

Seeds were germinated on ½ MS plates supplemented with 1% sucrose and 0.75% agar. After 1 week seedlings were transferred to liquid ½ MS + 1% sucrose with 2mM hydroxyurea (HU), 40μM aphidicolin (Aph) or without any drugs and grown in continuous light. After 2 weeks root length was measured. Experiment was performed in three biological replicates, in each replicate 48 plants/genotype/treatment were used.

Acknowledgements

We want to thank Steven Smith for his help with phylogenetic analysis, Peter Freddolino for performing statistical analysis of qPCR data as well as Karolina Górecka and Małgorzata Figiel for their help with protein purification and biochemical activity assays. This work was supported by the National Science Foundation grant MCB 1120271 and the National Science Centre, Poland grant Polonez 2015/19/P/NZ/03619. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665778.

Figures

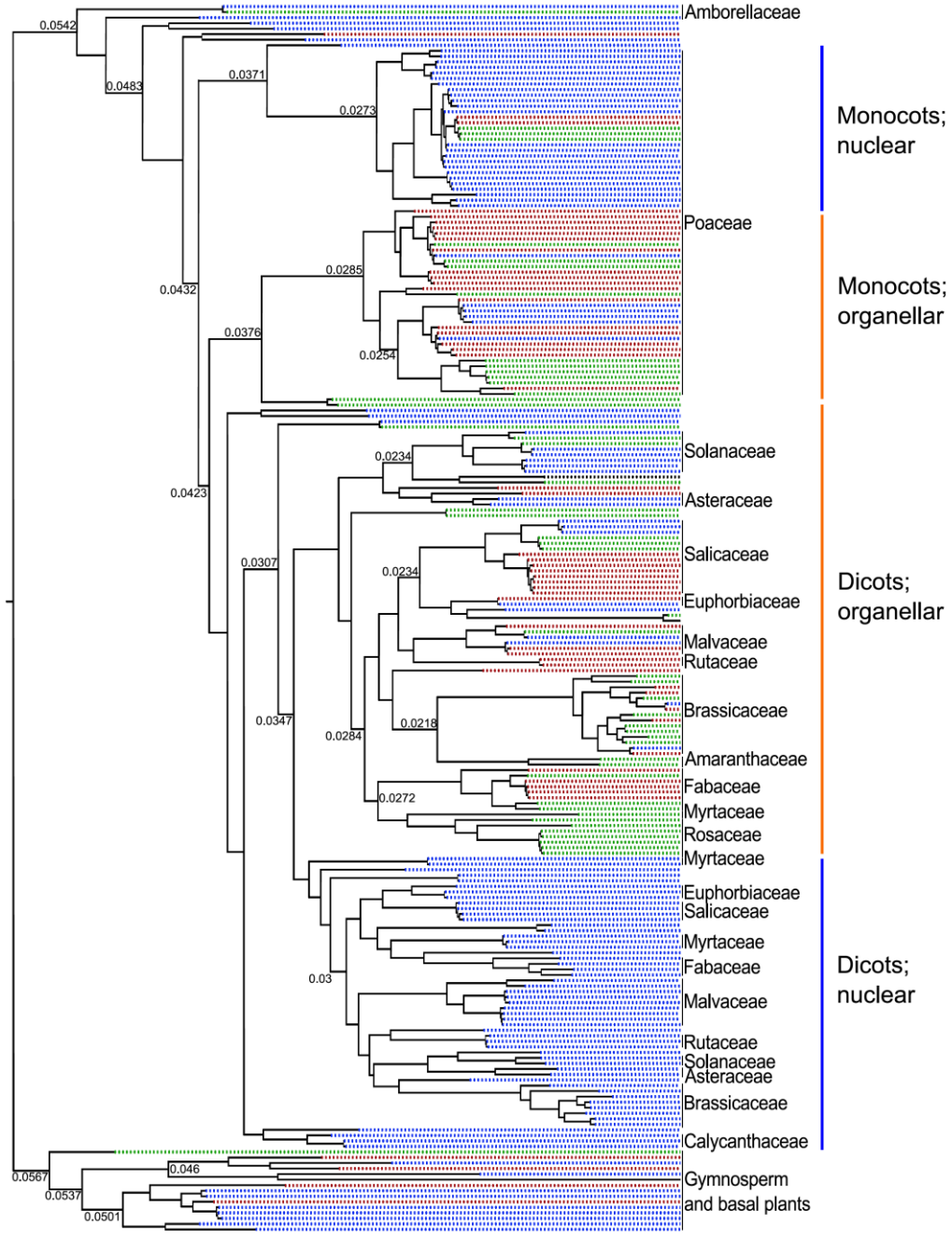

Figure 2.1. Evolutionary origin of plant RNase H1 proteins.Phylogenetic tree of all full length predicted RNase H1 proteins. Support values at tree branches are posterior probability scores, which integrate over the uncertainty in both the alignment and the phylogeny. Predicted protein localizations obtained using TargetP are marked with colors. Mitochondria – red, chloroplast – green, other – blue. A version of this phylogenetic tree with all species names, sequence IDs and support values is shown in Fig. S2.1.

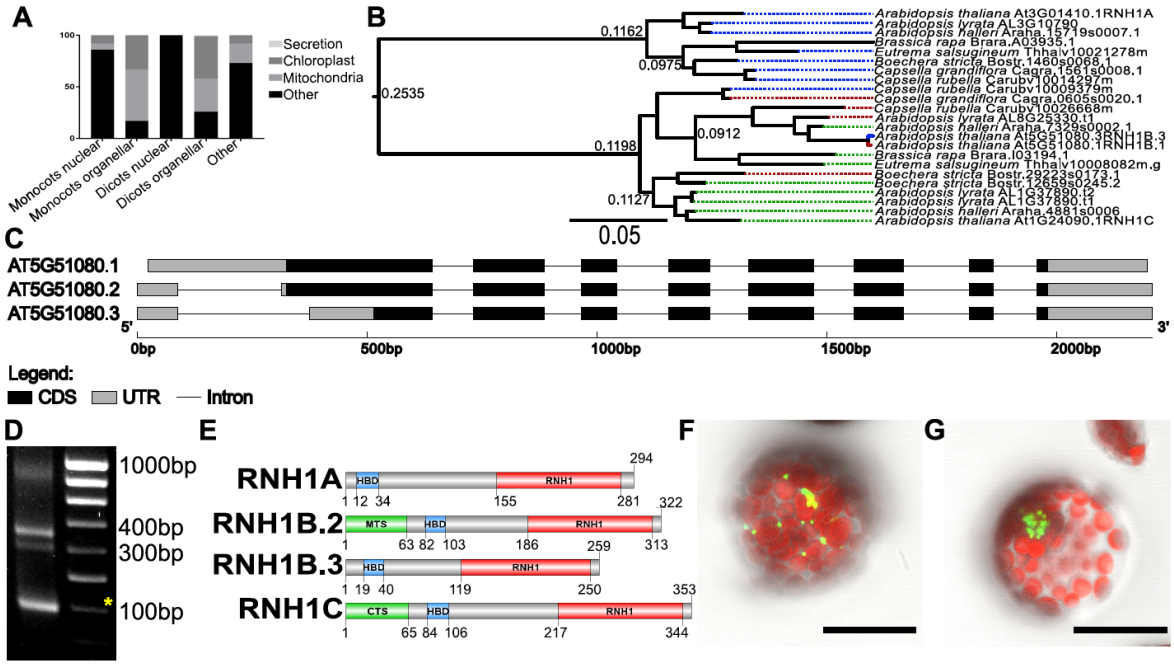

Figure 2.2. Origin of four RNase H1 proteins in *Arabidopsis thaliana*. **(A)** Predicted localization of RNase H1 proteins from distinct clades identified in Fig 1. Predictions were performed using TargetP. **(B)** Detailed phylogenetic tree of all RNase H1 proteins in *Brassicaceae*. Support values and color-coded predicted localization are used as described for Fig 1. **(C)** Splice variants of *RNH1B* predicted in Araport11 (AT5G51080.1) and identified by 5' RACE (AT5G51080.2 and AT5G51080.3). **(D)** 5' RACE of *RNH1B*. Asterisk indicates a band, which is likely to be non-specific. **(E)** Domain composition of four RNase H1 proteins in *Arabidopsis thaliana*. **(F)** Subcellular localization of GFP-tagged RNH1B.3 in transiently transformed *Arabidopsis* protoplasts. **(G)** Subcellular localization of GFP-tagged RNH1A in transiently transformed *Arabidopsis* protoplasts. Scale bars indicate 25μm.

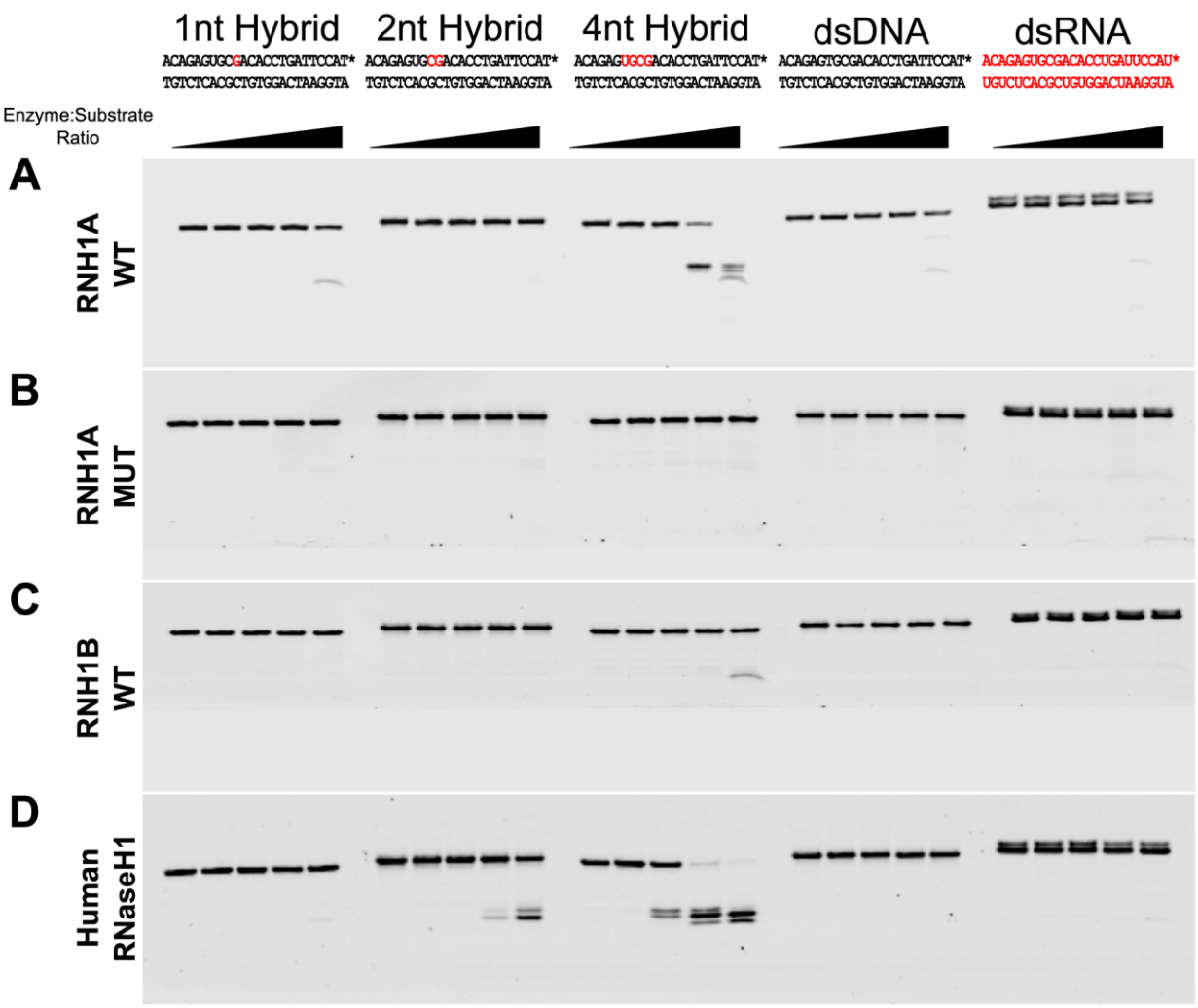

Figure 2.3. *In vitro* activity assays for RNH1A and RNH1B proteins from *Arabidopsis thaliana*. Wild type RNH1A **(A)**, catalytic mutant of RNH1A **(B)**, wild type RNH1B **(C)** and wild type human RNase H1 **(D)** (positive control) were incubated with fluorescently labeled oligonucleotide substrates containing different combinations of deoxyribonucleotides (black) and ribonucleotides (red). The triangle indicates increasing protein:substrate molar ratio. The reaction products were analyzed on TBE-urea PAGE and scanned for fluorescence. The experiment was repeated three times.

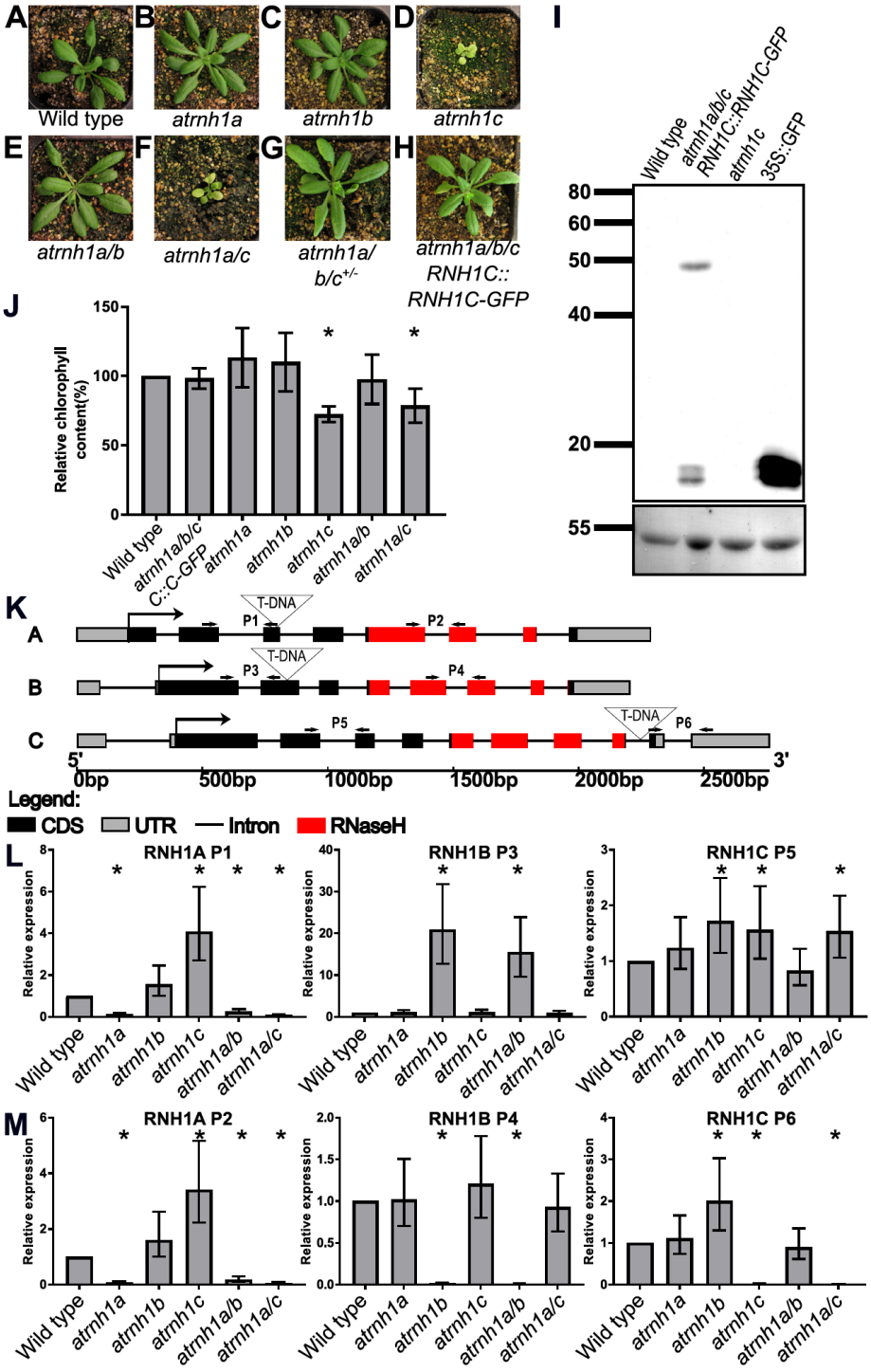

Figure 2.4. Effects of mutations in genes encoding RNase H1 proteins. Approximately 3-week-old plants of Col-0 wild type **(A)**, *atrnh1a* **(B)**, *atrnh1b* **(C)**, *atrnh1c* **(D)**, *atrnh1a, atrnh1b* double mutant **(E)**, *atrnh1a, atrnh1c* double mutant **(F)**, *atrnh1a, atrnh1b* double mutant heterozygous for *atrnh1c* **(G)** and *atrnh1a, atrnh1b, atrnh1c* triple mutant expressing RNH1C::RNH1C-GFP **(H)**. **(I)** Immunoprecipitation and detection of GFP-tagged RNH1C expression in triple *atrnh1a, atrnh1b, atrnh1c* mutant background.

Bottom panel shows non-specific staining with Ponceau S. **(J)** Relative chlorophyll content of Col-0 wild type and mutants in genes encoding RNase H1 proteins. Error bars indicate standard deviation from three biological replicates. Asterisks indicate $p < 0.05$ obtained from t-test. **(K)** Schematic representation of RNH1 genes in *A. thaliana*. Arrows indicate START codons, triangles indicate T-DNA insertion positions. Red boxes indicate positions of the conserved RNase domain. **(L, M)** Expression of genes encoding RNases H1 measured **(L)** upstream of T-DNA insertion and **(L)** downstream of T-DNA insertion. Primer pairs are marked in **(K)**. Error bars indicate 95% confidence intervals. Asterisks indicate $p < 0.05$ in comparison to Col-0 wild type.

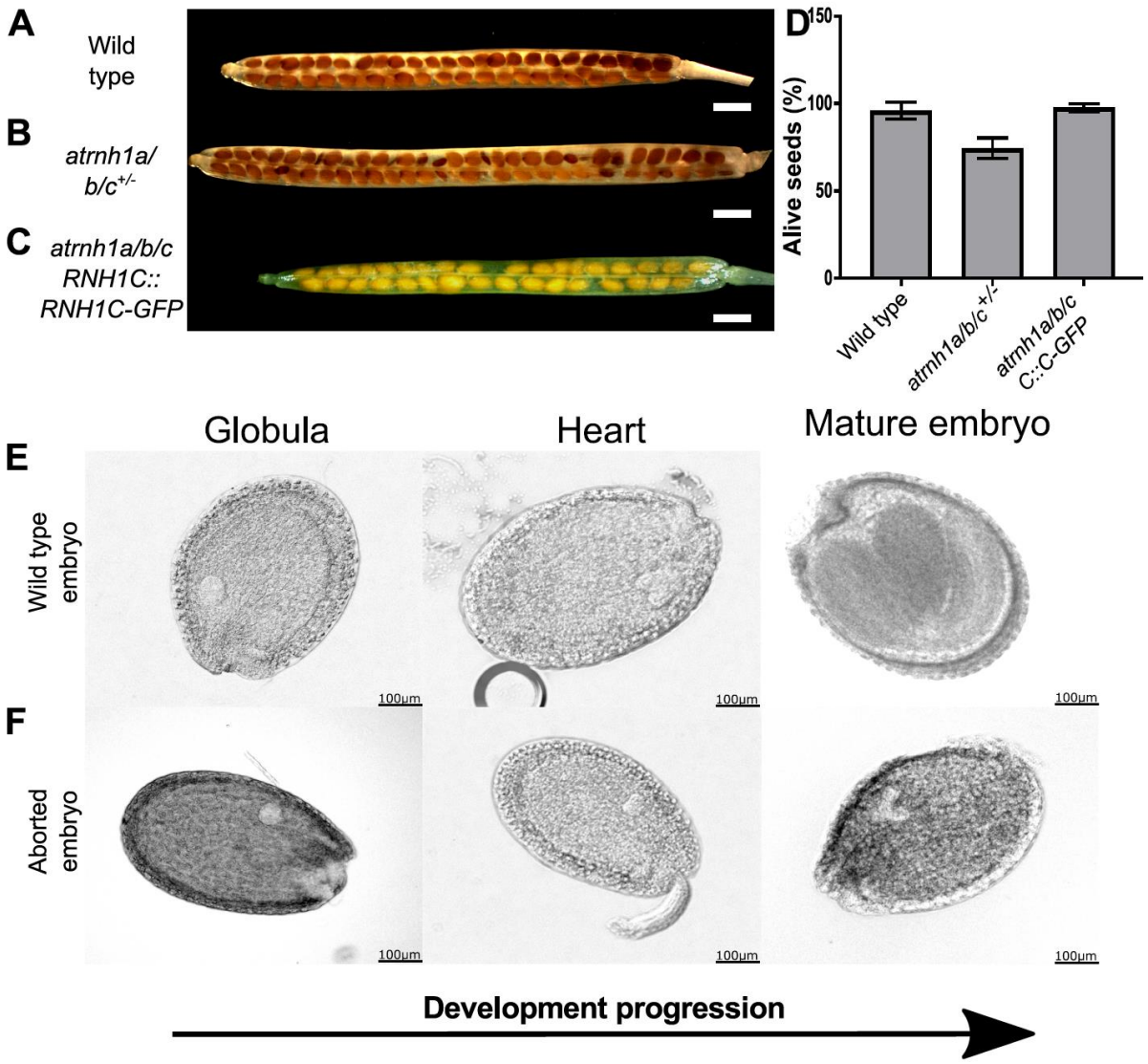

Figure 2.5. At least one organellar RNase H1 is required for embryo development**. (A)** Mature silique of Col-0 wild type. **(B)** *atrnh1a, atrnh1b* double mutant heterozygous for *atrnh1c* (*atrnh1a/b/c*$^{+/-}$). **(C)** *atrnh1a, atrnh1b, atrnh1c* triple mutant expressing RNH1C::RNH1C-GFP. **(D)** Percentage of properly developing seeds in siliques. Error bars indicate standard deviation from at least six independent siliques. **(E)** Development of a embryos in normally developing seeds obtained from *atrnh1b*$^{-/-}$*atrnh1c*$^{+/-}$ parents. **(F)** Development of a embryos in abnormally developing seeds obtained from *atrnh1b*$^{-/-}$ *atrnh1c*$^{+/-}$ parents.

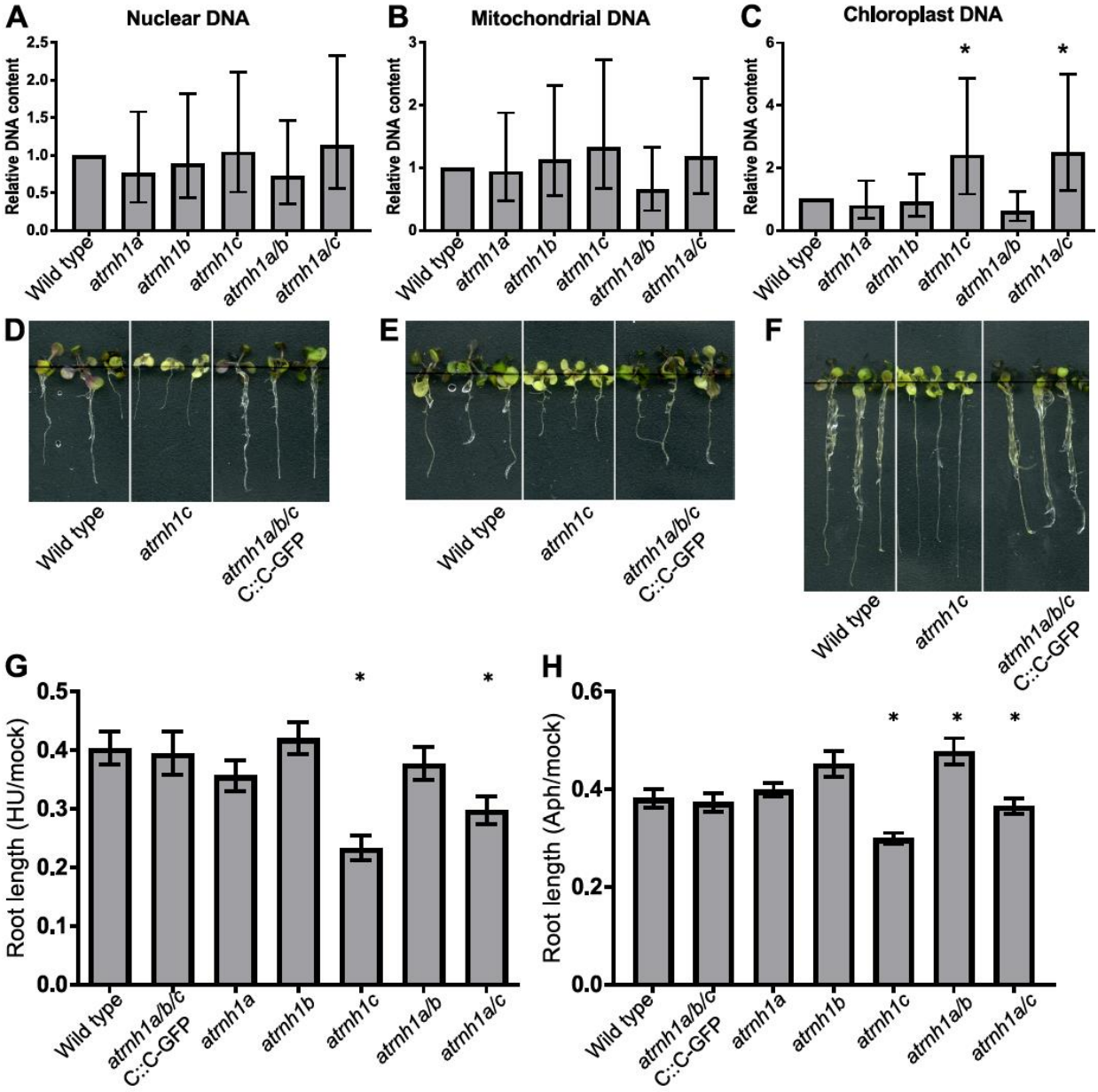

Figure 2.6. Loss of RNH1C leads to increased chloroplast DNA content and hypersensitivity to replication stress. Relative levels of nuclear **(A)**, mitochondrial **(B)** and chloroplast **(C)** DNA in Col-0 wild type and *atrnh1* mutants. Error bars indicate 95% confidence intervals. Asterisks indicate $p < 0.05$ in comparison to Col-0 wild type. **(D-F)** Phenotype of plants subjected to Hydroxyurea **(D)**, Aphidicolin **(E)** and mock **(F)** treatment. **(G, H)** Relative root length of Col-0 wild type and *atrnh1* mutants upon 2mM hydroxyurea **(F)** and 40µM aphidicolin **(G)** treatment. Asterisks indicate $p<0.0001$determined using 2-way ANOVA.

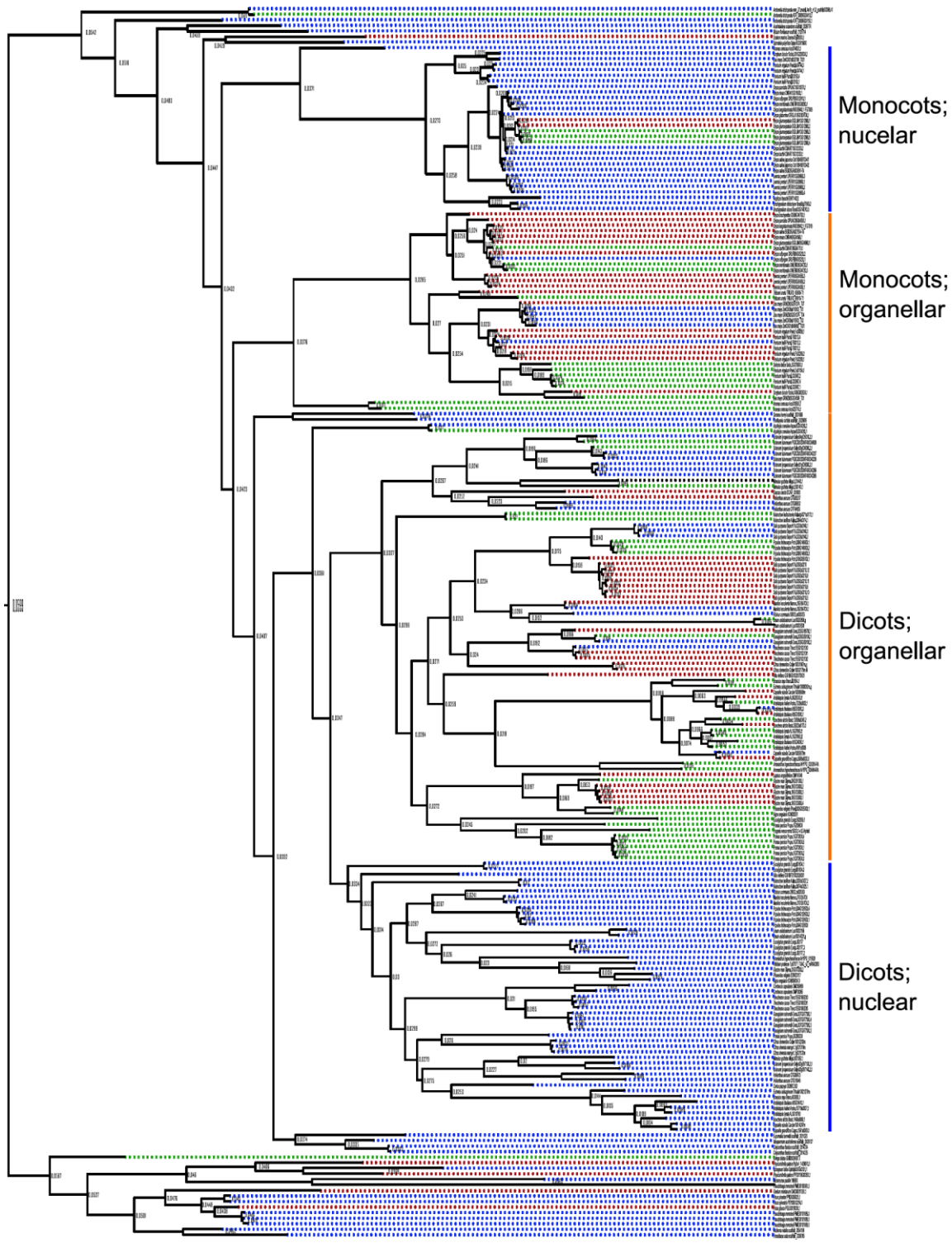

Figure S2.1. Evolutionary origin of plant RNase H1 proteins. A version of the phylogenetic tree shown in Fig. 1 with all species names, sequence IDs and support values.

Figure S2.2. Sequences of 5' ends of *RNH1B* splice variants identified by 5' RACE shown in Fig. 2.2D. Arrow indicates position of the RACE primer. AT5G51080.1 is predicted by Araport11 annotation. AT5G51080.2 and AT5G51080.3 have been identified by sequencing.

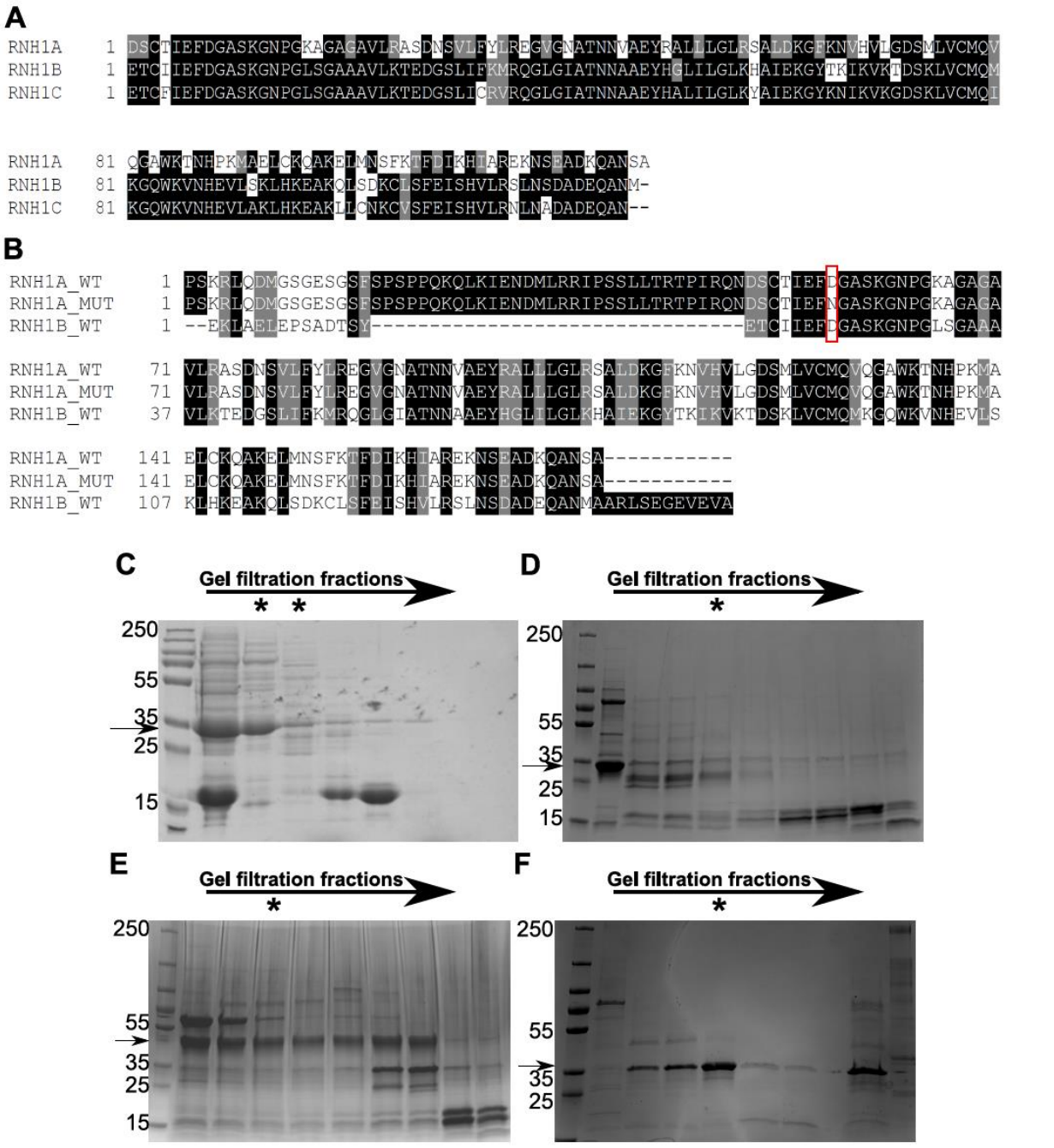

Figure S2.3. (A) Sequences of the catalytic domains of RNH1A, RNH1B and RNH1C. **(B)** Sequences of wild type RNH1A, wild type RNH1B and the catalytic mutant of RNH1A **(C-F)** Purification of recombinant RNase H1 proteins used for in vitro activity assays shown in Fig. 2.3. Stars indicate fractions used for the activity assays. Arrows indicate specific bands. Panels show RNH1A **(C)**, RNH1A catalytic mutant **(D)**, RNH1B **(E)** and human RNase H1 **(F)**.

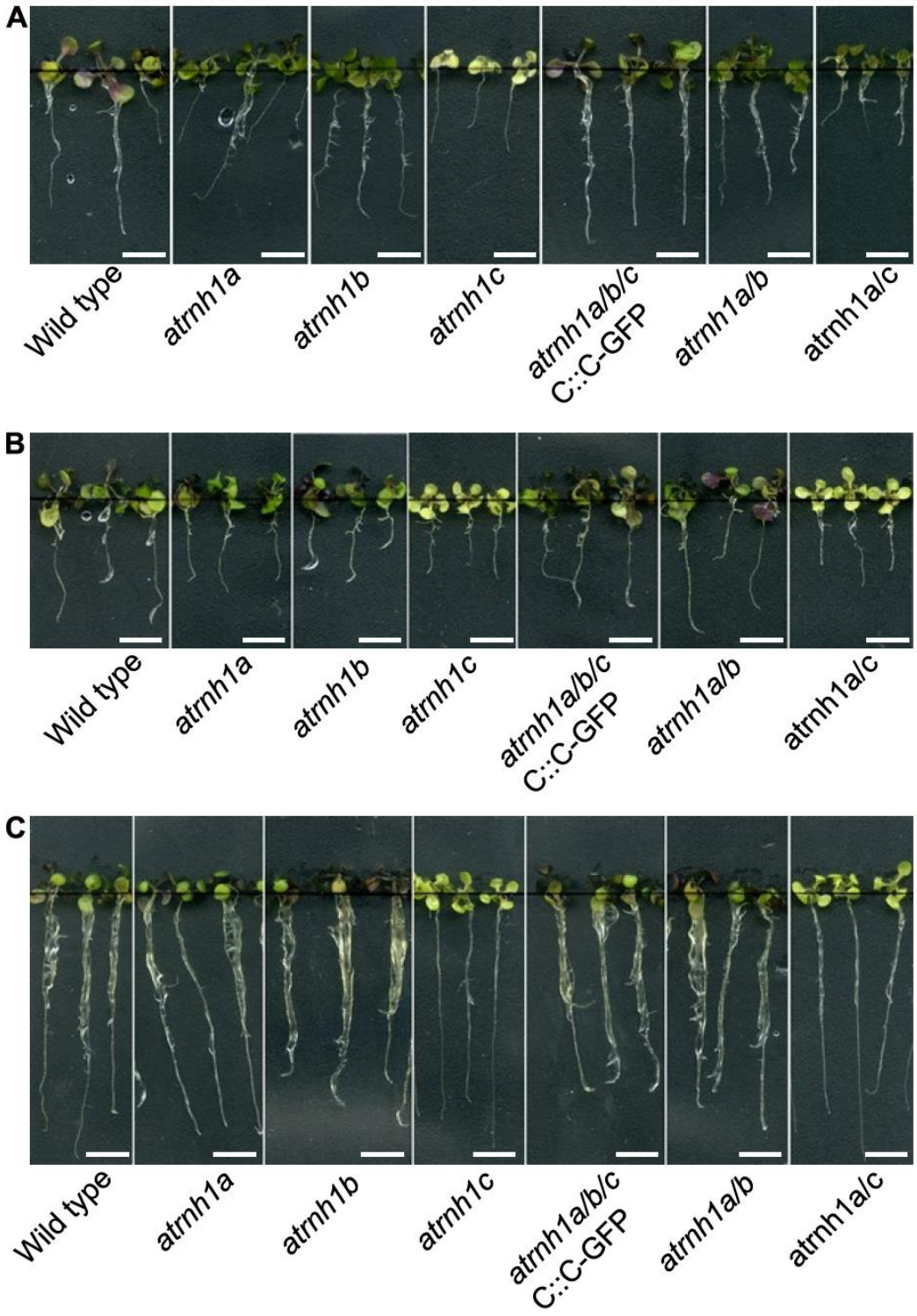

Figure S2.4. Pictures of plants subject to hydroxyurea **(A)**, aphidicolin **(B)** and mock **(C)** treatments. Scale bars indicate 10mm. Pictures of wild type, *rnh1c* and *rnh1a/b/c* expressing RNH1C-GFP are also shown in Fig. 2.6 D-F.

CHAPTER III

Functional characterization of DeCL proteins from *Arabidopsis thaliana*

cpRIP-seq was performed by Anoumid Vaziri. Wei-Yu Liu selected homozygous *atdecl* lines and generated transgenic lines. We wish to thank Fei Zhou for his help with protoplast transformation protocol optimization.

Abstract

RNA-binding proteins serve several essential roles in nucleic acid metabolism and gene regulation. They are present throughout the eukaryotic cell, including endosymbiotic organelles, chloroplast and mitochondria. Their roles in metabolism of organellar-encoded RNA are not fully understood. Here, we focus on characterization of an understudied group of RNA-associated proteins specific to the photosynthetic organisms, Defective in Chloroplasts and Leaves (DeCL) proteins. We show that in *Arabidopsis thaliana* these proteins accumulate in plastids and are expressed throughout plant development, also in non-photosynthesizing cells. We characterize *Arabidopsis* mutants in genes encoding the DeCL proteins and propose that two *Arabidopsis* DeCLs are partially redundant. Finally, we show that at least one of *Arabidopsis* DeCLs, DeCL1 specifically interacts with multiple plastid-encoded RNAs but loss of either DeCL does not lead to changes in RNA accumulation levels. Our preliminary RNA-seq data indicate that DeCL1 binding to rRNA may be associated with

enhanced processing by exonucleases. We hypothesize that DeCL proteins bind plastid-encoded RNAs and mediate recruitment of exonucleases or are involved in resolving higher-order RNA structures, which affect RNA stability.

Significance

We integrate genetic and molecular analysis of *Arabidopsis* plastid DeCL proteins to understand their functions in plastid RNA metabolism as well as two nuclear DeCL domain-containing proteins. We decipher phylogeny of DeCL proteins and aim to determine their biological functions. We provide the first evidence of DeCL-RNA interactions and show that they are limited mainly to the 5'- and 3'-ends of the transcripts, segments undergoing extensive processing and maturation. Our results provide fundamental insight into the biological and molecular functions of the plastid DeCL protein family and contribute potential advances into understanding of its nuclear counterpart.

Introduction

RNA processing is a crucial element in regulation of gene expression and therefore is one of the main mechanisms establishing cell homeostasis. In plant cells RNA is produced mainly in three cellular compartments: the nucleus, mitochondria and plastids[126]. While metabolism of nuclear-encoded RNA is well characterized, processing of RNAs in endosymbiotic organelles remains less understood. Large numbers of proteins have been proposed to be involved in plastid RNA metabolism, based on their predicted subcellular localization, but their specific roles and exact mechanisms are

unknown. Plastid-encoded RNA species have been shown to undergo extensive processing and modifications[162]. Additionally, it was shown that plastid transcription start sites and transcription termination sites rarely correspond with actual termini of mature transcripts. Both 5'- and 3'-ends of mature transcripts are determined by the processing mechanisms, and the activity of RNase J[287] and a combination of RNR1 and PNPase[288,289], respectively, rather than correspond to the ends of nascent RNAs being transcribed. The termini of mature RNAs are often marked by PPR proteins binding which protects them from further nucleolytic degradation[181]. Additionally, often polycistronic plastid transcripts are subjected to cleavage by HCF107 protein[131], and mRNAs occasionally contain introns which are spliced with assistance of plastid splicing factors, such as CRM, CFM and some PPR proteins[143,181]. On the other hand, a different subset of PPR proteins is involved in RNA editing[179]. Similarly to the nuclear rRNA, plastid-encoded rRNA species also undergo extensive maturation by a combination of exo- and endonucleolytic cleavages. Specifically, DEAD-box RNA helicases have been shown to be involved in 23S and 4.5S maturation, YbeY being responsible for proper maturation of both ends of 16S, 23S and 4.5S maturation[290], and a number of other, unknown proteins was proposed to process all pre-rRNA species[291]. One other family of proteins, DeCL (Defective Chloroplasts and Leaves), was proposed to, directly or indirectly, affect maturation of 4.5S rRNA[233]. However, despite their discovery over two decades ago, DeCL proteins remain poorly characterized and their molecular role is still unknown.

The first gene encoding a protein from this family was found in tomato and initially named DCL[232], however, in order to avoid being potentially mistaken for another protein

DCL (Dicer-like), it is currently often referred to as DeCL[242]. The structural characteristic of these small, usually 20-25kDa proteins, is the presence of a domain of unknown function DUF3223, later referred to as the DeCL domain, as well as the transfer peptide localized at the N-terminus[234]. Interestingly, the DeCL domain is also found in some bacterial proteins, suggesting its endosymbiotic origin in eukaryotes[292]. Initial research showed that tomato DeCL protein is required for proper development of chloroplasts as well as palisade cells in tomato leaves. Additionally, leaf sectors transiently depleted of DeCL displayed chlorosis and efforts to establish stable tomato line deficient in this protein failed, suggesting an essential role for DeCL in tomato. Furthermore, it was shown that upon DeCL loss, the mRNA encoding one of the Plastid Encoded Polymerase (PEP) subunits RpoB accumulated at a higher level[232].

Further characterization of tomato DeCL proteins demonstrated that upon DeCL loss, accumulation of all investigated plastid-encoded proteins was decreased while their corresponding mRNA levels remain unchanged. Interestingly, levels of some, but not all, nuclear-encoded chloroplast proteins were also decreased, which was suggested to be an effect of retrograde signaling between improperly developing chloroplasts and the nucleus. These results suggested that DeCL may be involved in plastid protein biosynthesis, possibly at the level of translation. It was shown that upon DeCL loss, plastid 4.5S rRNA precursor was not efficiently processed and that this leads to improper polysome assembly. Importantly, cytosolic ribosomes were not affected[233].

Finally, the *Arabidopsis* ortholog of tomato DeCL was identified. Similar in size and domain composition, it was shown to be expressed in all tested types of tissues and to localize in plastids. Interestingly, it was shown that stable overexpression of DeCL-

GFP fusion protein caused leaf chlorosis and retarded growth compared to wild-type plants. Furthermore, plants overexpressing the fusion protein accumulated plastid 4.5S rRNA precursor. Surprisingly, plants expressing an antisense DeCL transgene, presumably knock-down mutation lines, exhibited phenotypes similar to plants overexpressing DeCL-GFP fusion protein, as well as accumulated plastid-encoded proteins at lower levels compared to wild type plants. It was concluded that overexpression of DeCL-GFP fusion protein acts as a dominant negative mutation and that *Arabidopsis* DeCL protein serves similar function in plastid 4.5S rRNA maturation as its counterpart in tomato. However, no direct evidence was shown[234].

Four additional proteins containing the DeCL domain have been identified so far in *Arabidopsis thaliana*. DOMINO1 is a small, 22kDa protein localized in the nucleus. It was shown that embryos deficient in DOMINO1 develop significantly slower than wild-type plants and die during early embryo development, when the seed already desiccates. It was also shown, that these embryos display abnormally large nucleoli, suggesting DeCL involvement in rDNA organization. Additionally, presence of a gene encoding potential mitochondrial DeCL in *Arabidopsis* genome was reported but no further characterization of this protein was provided[235].

Finally, the presence of the DeCL domain was reported in the C-terminal domains (CTD) of largest subunits (NRPD1 and NRPE1) of plant-specific Polymerase IV (Pol IV) and Polymerase V (Pol V), respectively[236]. These polymerases are best known for their crucial role in Transcriptional Gene Silencing (TGS) process – RNA-directed DNA Methylation (RdDM)[293]. This mechanism is mainly responsible for silencing expression of transposons but also has been shown to regulate expression of protein-coding

genes[239,241]. While Pol IV is mainly responsible for the generation of siRNA precursors, the core of this process is long (usually ~200bp) non-coding RNA (ncRNA) produced by Pol V[294]. This ncRNA remains in the nucleus where it is bound by an ARGONAUTE (AGO) protein. AGO proteins are imported from the cytoplasm into the nucleus upon binding of Pol IV-produced siRNA. It is believed that it is the base pairing between siRNA and ncRNA produced by Pol IV and Pol V, respectively, that provides the specificity of AGO binding[293]. Upon binding to ncRNA, AGO protein is believed to interact with, directly or indirectly, and recruit DNA methyltransferase machinery which methylates DNA strands. DNA methylation in plants is mainly associated with the repression of transcription. This allows for silencing of transposon expression which might be detrimental to genome stability[295]. Additional, RdDM-regulated gene control mechanisms, such as nucleosome remodeling[296] and chromatin looping[297], have also been reported. Importantly, after recruitment of DNA methyltransferases, AGO is believed to endonucleotically cleave (slice) ncRNA[298], so that its 3'-end is still bound by the Pol V while 5'-end portion of the transcript is subjected to degradation. It was shown that DeCL domain of NRPE1 interacts with 3'->5' exonuclease RRP6L, which was demonstrated to digest sliced ncRNAs *in vitro*. Additionally, it was shown that the DeCL domain is required for Pol V transcription *in vivo* as well as for DNA methylation[242]. However, no specific function of DeCL domain of NRPE1 was proposed. Similarly, the function and importance of NRPD1 DeCL domain remains unknown. Importantly, DeCL domain is the main structural difference between plant Polymerase II and Pol IV and Pol V[292]. Why would this domain be required for long non-coding RNA production but not for mRNA transcription, remains an intriguing question.

Here, we show that DeCL proteins are present mostly in photosynthetic organisms, we provide characterization of two *Arabidopsis* DeCL proteins and show that both localize to chloroplasts. We also investigate their role in regulation of gene expression and provide the first evidence for association of these proteins with RNA but not DNA.

Results

DeCL proteins are present mostly in photosynthetic organisms

In order to identify potential DeCL proteins we used the BLAST search tool, using sequence of a previously identified *Arabidopsis* DeCL domain. We found that DeCL proteins are present in all flowering plant species that we examined, as well as green algae, mosses and cyanobacteria (Fig. 3.1). Often, more than one DeCL protein was identified in each of the species. Interestingly, we did not detect the presence of a DeCL domain in animals or fungi but also in majority of bacterial species that we investigated (Table 3.1). Specifically, we did not identify this domain in Rickettsiales, a group of bacteria believed to be ancestors of mitochondria[4]. Low resolution phylogenetic analysis of obtained hypothetical DeCL protein sequences from plants and cyanobacteria indicated presence of five major groups, which we refer to as DeCL1, DeCL2, DOMINO and Pol IV/Pol V. Bacterial sequences have formed a distinct group on the unrooted phylogenetic tree (Fig. 3.1A).

Arabidopsis contains two plastid-localized DeCLs

We performed a high resolution phylogenetic analysis of DeCL1 and DeCL2 protein sequences from a selected subset of plant species and found that plants often

encode both DeCL1 and DeCL2 and that proteins from both groups are present in all flowering plant species that we tested (Fig. 3.1B). We hypothesized that a gene duplication that lead to split of DeCL into two groups is due to DeCL proteins acting in various cell compartments. We found that in addition to DeCL domain, these proteins often contain N-terminal transfer peptides, presumably required for chloroplast or mitochondrial localization (Fig. 3.2B). In order to test our hypothesis we used TargetP[257] localization prediction tool and found that the vast majority of DeCL1 proteins displayed predicted chloroplast localization, consistent with published work[234], while most DeCL2 proteins were predicted to localize to mitochondria (Fig. 3.2A). To further validate our hypothesis that DeCL proteins localize to various cellular compartments, we cloned *Arabidopsis* genes encoding DeCL1 and DeCL2, AT1G45230 and AT3G46630, respectively, fused them with GFP followed by transient overexpression in *Arabidopsis* protoplasts (Fig. 3.2C). We confirmed previously reported chloroplast localization of DeCL1 but, unexpectedly, we found that DeCL2 also localizes to the chloroplasts and no GFP signal was visible outside of these organelles. Interestingly, we observed strong, focal GFP signal for both fusion constructs, similar to the appearance of stained nucleoids. In order to further confirm this result, we generated *Arabidopsis* lines stably expressing DeCL proteins fused with GFP under control of their respective promoters. Biochemical purification of chloroplast followed by Western Blot confirmed chloroplast localization of both *Arabidopsis* DeCL proteins (Fig. 3.2D). In order to further dissect suborganellar localization of DeCLs we lysed chloroplast containing GFP-tagged DeCL1 and DeCL2 in a hypotonic buffer. Interestingly, both DeCL1 and DeCL2 were in the insoluble fraction suggesting membrane localization. Therefore, in order to elute weakly membrane-bound proteins we treated the insoluble fractions with high salt

concentration. This treatment should break protein-protein interactions allow only transmembrane proteins to remain in the membranes. Interestingly, both GFP-tagged DeCL proteins remained in the membrane fraction even after 2M NaCl treatment, suggesting a strong association of these proteins with chloroplast membranes. Importantly, *Arabidopsis* DeCL proteins do not contain any obvious transmembrane domains in their predicted structure.

Plastid DeCLs are broadly expressed

In order to better understand the biological function of *Arabidopsis* DeCL proteins, we looked at their expression patterns throughout development using publicly available RNA-seq data[299]. We found that both DeCL1 and DeCL2 are expressed in all organs tested, including roots, stems, flowers, young and old leaves as well as fruits (siliques) (Fig. 3.3AB). Similarly, both genes are transcribed at all stages of development, including male and female gametophytes, dry seeds, developing embryo and germinating seedling. Highest mRNA accumulation of both genes was observed in mature embryos, cotyledons and leaves. Interestingly, mRNAs encoding both DeCL1 and DeCL2 were found in non-photosynthetic cells of roots and flowers as well as non-photosynthesizing cells of dry seeds. This suggests that function of organellar DeCLs is not limited to photosynthesis-related processes.

DeCL1 and DeCL2 are required for plant survival

We obtained *Arabidopsis* mutants deficient in DeCL1 and DeCL2 proteins. Both mutant lines contained T-DNA insertions in the protein-coding sequences of

corresponding genes (Fig. 3.4A). We found that both single mutants are viable but simultaneous loss of both DeCL1 and DeCL2 leads to plant lethality shortly after germination on soil ($p < 2.26*10^{-13}$, Chi square test; n = 170 plants). Interestingly, plants deficient in DeCL1 displayed lower chlorophyll level and retarded growth, similarly to knock-down line described previously. However, mutation of *DeCL2* did not cause any visible phenotype. Importantly, we were able to obtain viable *atdecl1/atdecl2* seedlings from seeds germinated on ½ MS agar plates supplemented with 1% sucrose (Fig. 3.4B). These plants were significantly smaller and displayed stronger chlorosis than single *atdecl* mutants grown in parallel. We were able to reverse visible phenotype of *atdecl1* by complementation with DeCL1-GFP fusion construct, suggesting specificity of the *atdecl1* phenotype (Fig. 3.4C). To determine if the lower chlorophyll content is caused by a reduction in the abundance of the plastid genome, we quantified their plastid DNA content, similarly to an approach used in Chapter 2 (Fig. 3.4DEF). We did not identify any significant difference in plastid DNA content between tested lines, which may suggest that the copy number of plastid chromosomes is not affected by the loss of DeCL proteins.

DeCL1 interacts with plastid-encoded mRNAs

In order to decipher the molecular role of DeCL proteins we performed cpRNA-immunoprecipitation, followed by sequencing (cpRIP-seq), of GFP-tagged DeCL1 protein in two biological replicates (Fig. 3.5A-C, top panel). We found that DeCL1 interacts with multiple RNA molecules in chloroplasts. Specifically, we found that this protein often associates with 5'-ends of polycistronic transcripts as well as 3'-ends of

transcripts produced from monocistronic units. Interestingly, transcripts associated with DeCL1 almost exclusively encode elements of photosynthetic machinery, ribosomal proteins and very rarely tRNAs and subunits of ATPases. We did not observe any significant interaction between DeCL1 and mRNAs encoding NADH dehydrogenases. Furthermore, interaction sites appeared to be limited to PEP-transcribed genes. Importantly, our results were comparable between biological replicates. To confirm specificity of DeCL1-RNA interaction we performed locus-specific Chromatin Immunoprecipitation followed by qPCR (ChIP-qPCR) (Fig. 3.5A-CD, bottom panel). We did not detect significant DeCL1 interaction with any of the tested loci, which indicates that DeCL1 is unlikely to associate with DNA.

Previous reports showed that loss of DeCL1 causes lower accumulation of chloroplast-encoded proteins, possibly due to affected translation, but no changes in accumulation of mRNAs of specific genes. In order to verify this observation on a plastome-wide scale we performed cpRNA-seq in wild-type as well as plants deficient in DeCL1 and DeCL2. Similar to previous reports[233,234], we did not identify any major changes in accumulation levels of plastid-encoded mRNAs. However, we did observe some changes in rRNA accumulation levels between *atdecl1* and wild type plants, specifically in fragments of pre-rRNA which are believed to be post-transcriptionally processed (Fig. 3.5E, bottom panel). These changes in *atdecl1* are consistent with an increased accumulation of RNA fragments, which are typically removed by the activity of exonucleases[155,212,289,300–303]. cpRIP-seq indicated that these sequences are normally bound by DeCL1 (Fig. 3.5E, top panel). These results, although preliminary, may allow speculating that DeCL1 binds rRNA ends and mediates the recruitment of

exonucleases. Alternatively, it may be involved in resolving higher-order RNA structures, which affect RNA stability.

Discussion

Here, we characterize a novel family of RNA-associated proteins, the DeCL family. We found that DeCL proteins are present in all tested flowering plants as well as basal plants, algae and cyanobacteria (Fig. 3.1A). We did not find any direct evidence for presence of these proteins in non-photosynthetic eukaryotes or bacteria, specifically not in ancestors of mitochondria (Table 3.1). Presence of a bacterial protein in eukaryotes may be the result of an endosymbiotic event or horizontal gene transfer. However, specificity of DeCLs presence and their evolutionary conservation specifically in photosynthetic organisms strongly indicates endosymbiotic origin of these proteins. Additionally, despite some DeCL proteins being predicted to localize to mitochondria, there is no evidence of DeCLs presence in Rickettsiales, the mitochondria ancestors[4], indicating that this localization was acquired later during evolution. Altogether, these data strongly suggest cyanobacterial origin of DeCL proteins. Therefore, it may be hypothesized that their primary function is crucial for photosynthesis-related metabolism regulation.

Phylogenetic analysis of DeCL proteins from flowering plants suggests presence of four subfamilies. Two of them, DOMINO and Pol IV/Pol V, contain proteins localized in the nucleus. The remaining subfamilies, DeCL1 and DeCL2, contain proteins localized in endosymbiotic organelles (Fig. 3.1B, Fig. 3.2A). We found that plant genomes often encode more than one DeCL protein and that in many cases one of these proteins is

predicted to localize to chloroplasts (DeCL1) and the other to mitochondria (DeCL2). Sequence similarities between DeCL1and DeCL2 protein from different species suggest that duplication which lead to the presence of multiple DeCLs occurred relatively early in flowering plant evolution and that the genome of the last common ancestor of monocots and dicots encoded two DeCL proteins.

Interestingly, both microscopy observations as well as biochemical fractionation approaches demonstrated chloroplast localization of both DeCL proteins from *Arabidopsis* (Fig. 3.2CD). Our results also suggest that these proteins are anchored in chloroplast membranes, despite, DeCL proteins not containing any predicted transmembrane domains. Therefore, it is tempting to speculate that they are associated with plastid membranes through post-translational modifications. Additionally, both DeCL1-GFP and DeCL2-GFP fusion proteins form strong foci inside chloroplasts which resemble signal characteristic for plastid nucleoid or nucleoid-associated proteins. Therefore, it is possible that DeCLs association with plastid membranes is mediated by nucleic acids. However, neither of DeCL domain-containing proteins were found to directly associate with plastid nucleoids in mass spectrometry analysis[73,74]. Membrane localization and close proximity with cpDNA are consistent with potential role of DeCLs in co-transcriptional RNA processing.

Analysis of DeCLs expression showed that genes encoding these proteins are transcribed at virtually all developmental stages and in all organs (Fig. 3.3AB). Interestingly, DeCL genes were expressed in non-photosynthetic parts of the plant, such as roots, flowers and seeds. This indicates a role of DeCL proteins not limited to photosynthesis or its regulation. However, the highest expression levels for both DeCLs

were observed in photosynthesizing tissues, such as leaves and cotyledons. This may indicate that DeCL proteins are required in higher amounts in plastids actively processing multiple metabolic processes at high pace, e.g. photosynthesis and fatty acid metabolism. Alternatively, it is possible that DeCL proteins are not required for non-photosynthesizing plastids but their respective mRNAs accumulate in all plastid types to allow for quick response and photosynthesis regulation upon light detection. Yet another possibility is that the activity of DeCLs is related to the activity of PEP polymerase, which is most active in chloroplasts but less pronounced in other plastid types.

We obtained and characterized *Arabidopsis* mutants deficient in DeCL proteins. We found that only loss of DeCL1 but not DeCL2 leads to a visible phenotype – retarded growth and chlorosis (Fig. 3.4BC). Importantly, we found that plants deficient in both DeCLs die shortly after germination on soil but survive on media supplemented with sucrose. Altogether, these results, together with plastid localization of both DeCLs, indicate partial redundancy between DeCL1 and DeCL2 functions. Additionally, we showed that chlorosis caused by the loss of DeCL1 is not due to DNA maintenance defect and most likely not due to differences in mRNA accumulation levels in these plants, but rather deficiencies in photosynthesis machinery (Fig. 3.4DEF). Interestingly, loss of a single DeCL protein in tomato leads to severe chloroplast development defects and embryo lethality. This may suggest that tomato contains only one functional plastid DeCL protein. Surprisingly, we found that tomato genome encodes at least two organellar DeCLs (Fig. 3.1B). Whether only one is functional or if they act in different cellular compartments remains an open question. Viability of *atdecl1/atdecl2* mutant

embryos may indicate that DeCL proteins in tomato and *Arabidopsis* do not serve the exact same functions.

Using RIP-seq, we found that the DeCL1 protein associates with multiple plastid-encoded RNAs, which is consistent with its possible role in regulation of plastid gene expression (Fig. 3.5A-D). We found, that DeCL1 often interacts with transcribed 5'-ends of operons as well as mRNA 3'-ends of genes which are not parts of polycistronic units. Additionally, our results suggest that the role of DeCL1 is limited to PEP-transcribed genes, indicating its direct or indirect interaction with this polymerase complex. Interestingly, we found that out of the DeCL1-RNA association sites, most were present in close proximity to genes encoding elements of photosynthetic machinery and ribosomal proteins. These observations are consistent with the hypothesis that DeCL1 is responsible for the regulation of photosynthesis but also suggest specificity of this protein toward PEP-produced transcripts. We identified sequences directly bound by DeCL1 and subjected them to MEME sequence motif search[304]. Interestingly, we did not find any sequence preference for DeCL1 binding. Additionally, we did not detect any interaction between cpDNA and DeCL1 which suggests that this protein interacts only with RNA but not DNA, does not interact directly with RNA polymerase and potentially associates with RNA post-transcriptionally. These negative results confirm mass spectrometry analysis of *Arabidopsis* nucleoid-associated proteins[73,74], where neither of the DeCL proteins was found to directly associate with plastid DNA.

Previous studies have suggested the role DeCL proteins in ribosomal rRNA maturation[233]. Our cpRNA-seq data show that plants deficient in DeCL1 protein accumulate higher levels of pre-rRNA transcripts (Fig. 3.5E). Specifically, we observe

increased RNA accumulation at the pre-rRNA processing sites. This results confirms previously proposed role of DeCL1 in plastid rRNA maturation, however the exact function remains unknown. It is possible that DeCL1 possesses nucleolytic activity and directly digests nascent transcript. Alternatively, DeCL1 may bind specific RNA fragments and recruit RNA processing enzymes. However, we need to remember that this experiment requires additional biological replicates, therefore current results need to be interpreted carefully.

It was previously shown that loss of the DeCL1 protein caused decrease in specific plastid-encoded proteins accumulation but did not affect respective mRNAs levels[233,234]. Consistent with previously reported results[234], in our cpRNA-seq results we did not observe any substantial changes in plastid-encoded mRNA accumulation upon loss of either of DeCL proteins. This result suggests that DeCLs affect expression of plastid genes at the step of translation.

Altogether, these results increase our knowledge of plant DeCL proteins. However, their exact biological and molecular roles remain mysterious, similarly to the functions of nuclear DOMINO protein and DeCL domain of Pol IV and Pol V. Based on our results, it is tempting to speculate that proteins containing the DeCL domain specifically bind RNA and mediate recruitment of specific RNA processing enzymes. Alternatively, it is possible that DeCL domain-containing proteins are able to directly resolve specific type of higher-order RNA structure which allows for more efficient ribosome binding, RNA processing or degradation.

Materials and Methods

Phylogenetic analysis

Sequences of DeCL proteins were identified through BLAST search using amino acid sequence of DUF3223 domain of the previously described protein DeCL1. Specific protein sequences were downloaded from Phytozome[273] and NCBI[305] and aligned in MAFFT[278]. Phylogenetic analysis was performed with BALI-Phy[279] on the CIPRES[280] platform with default settings. Consensus greedy tree was constructed with the burn-in value of 100,000 and visualized in FigTree. Subcellular localization was predicted for the amino acid sequences using TargetP[257] online tool.

Protoplast preparation and imaging

Arabidopsis protoplasts were prepared using a published protocol[282]. Twenty leaves from young *Arabidopsis* plants were finely chopped and submerged in 10ml of enzyme solution (20mM MES pH=5.7, 1.5% cellulase R10, 0.4% macerozyme R10, 0.4M mannitol, 20mM KCl), vacuum infiltrated 3 times for 30 seconds and incubated at room temperature in the dark for 2 hours. Following digestion, equal volume of W5 solution (2mM MES pH=5.7, 154mM NaCl, 125mM $CaCl_2$, 5mM KCl) and large leaf fragments were removed by filtering through 100µm nylon mesh. Liquid containing released protoplasts was centrifuged at 100 x g for 2 minutes with slow acceleration and deceleration. Pelleted protoplasts were resuspended in 5ml of W5 solution by gentle swirling and incubated in ice for 30 minutes. After incubation, supernatant was gently removed and protoplasts were resuspended in 3ml of MMg solution (4mM MES pH=5.7, 0.4M mannitol, 15mM $MgCl_2$). For each transformation 10µl plasmid DNA (1µg/µl;

pEG103 plasmid containing DeCL1 or DeCL2 cDNA sequence fused to the N-terminus of GFP under control of 35S promoter) was mixed with 100µl of protoplasts and 110µl of PEG solution (40% PEG4000, 0.2M mannitol, 100mM $CaCl_2$) in 2-ml centrifuge tubes. Transfection mixture was incubated at room temperature in the dark for 15 minutes. To stop the transfection process, 440µl of W5 solution was added, mixed gently and mixtures were centrifuged 100 x g for 2 minutes with slow acceleration and deceleration. Protoplasts were resuspended in 300µl of W5 solution and incubated in 24-well plate overnight in the dark at room temperature. On the next day protoplasts were visualized under confocal Leica SP5 microscope.

Biochemical fractionation of chloroplasts

Arabidopsis chloroplasts were isolated using a published protocol[306]. Approximately 2 grams of young *Arabidopsis* leaves were ground in Cold Isolation Buffer (20mM Tris-HCl, pH=8.0, 1.25M NaCl, 5mM $MgCl_2$, 5mM DTT, 1% Plant Protease Inhibitor) and centrifuged for 10min at 3000 x g at 4^0C. Chloroplast pellet was washed three times with Cold Isolation Buffer. Chloroplasts were lysed by incubation in 10ml of Low Salt Buffer (20mM Tris-HCl, pH=8.0, 20mM NaCl, 5mM $MgCl_2$, 5mM DTT, 1% Plant Protease Inhibitor) for 30 minutes and centrifuged for 15min at 16 000 x g at 4^0C. Supernatant containing soluble chloroplast proteins was saved and pellet containing membrane fraction was resuspended in 10ml of High Salt Buffer (20mM Tris-HCl, pH=8.0, 2M NaCl, 5mM $MgCl_2$, 5mM DTT, 1% Plant Protease Inhibitor), incubated for 30min at room temperature and centrifuged for 15min at 16 000 x g at 4^0C. Resulting supernatant containing membrane-associated proteins was saved and pellet containing

insoluble, tightly membrane-associated proteins, was resuspended in 1x Laemmli buffer[307] and incubated at room temperature for 2 hours. Soluble fractions were mixed with Laemmli buffer and boiled at 99^{0}C for 10min. Protein samples were separated on 12% SDS-PAGE, subjected to Western Blot and detection with α-GFP antibody.

Plant material and oligonucleotides

Plant lines used in this study: Col-0 (CS 70000), *atdecl1* (SALK_02883), *atdecl2* (GABI_370H12). Transgenic lines were generated using abovementioned mutant lines transformed with pMDC107 plasmid containing genomic DNA sequence of respective gene at the N-terminus of GFP, under control of its native promoter sequence. Oligonucleotides used in ChIP-qPCR: ndhcF: 5'-AGCAGAAACATAGACGAACTCCT-3'; ndhcR: 5'-CTTGCCCAATCCACTCCGAT-3'; atpbF: 5'-AGCGAATTCGAAACGGAACTTT-3'; atpbR: 5'-AAATGGGACGCATAACCGGA-3'; petndF: 5'-CGCATGGGCTGCTTTAATGG-3'; petndR: 5'-GCCGAACGGTCTAGAAAACG-3'; psaIF: 5'-GATCGGCTGAGACCGAATCAT-3'; psaIR: 5'-CCAAATGGGTCTTATCGAATCGAAG-3'. Sequences of oligonucleotides used for organellar DNA content quantification were previously described by Kim et al[260].

Chlorophyll content analysis

Chlorophyll content was analyzed as described previously[281]. Briefly, 1 gram of *Arabidopsis* leaves was ground in liquid nitrogen and chlorophyll extracted with 100% acetone. Samples were centrifuged for 10min at 10,000 x g at 4^{0}C and absorbance was

measured at 645 and 662nm. Chlorophyll content was calculated using following formula: C_{a+b}(µg/ml) = 18.09 x A645 + 7.05 x A662.

DNA content analysis and RNA-seq

Total DNA was isolated from 2-3 week old *Arabidopsis* seedlings with DNeasy Plant Mini Kit (QIAGEN). 1ng of DNA per qPCR reaction was used as a template. Primers used were described previously. For RNA-seq total RNA was isolated from 2-3 week old *Arabidopsis* seedlings with RNeasy Plant Mini Kit (QIAGEN) and submitted for Illumina sequencing.

Nucleic acid IP

For cpRIP-seq, 2 grams of young *Arabidopsis* leaves were ground in Cold Isolation Buffer[306] and centrifuged at 3,000 x g for 10min at 4^0C. Resulting chloroplast pellet was washed three times with Cold Isolation Buffer. Purified chloroplasts were resuspended in the lysis buffer (50mM Tris-HCl, pH=8.0, 10mM EDTA, 1%SDS, 1mM PMSF, 1% Plant Protease Inhibitors) and sonicated. RNA-protein complexes were immunoprecipitated overnight with GFP-trap. Beads were washed and proteins digested as described previously. Purified RNA was submitted for sequencing. Total DNA ChIP was performed as described by Hanaoka et al[308,309]. Briefly, *Arabidopsis* tissue was crosslinked with 1% formaldehyde and ground in liquid nitrogen. Powdered tissue was resuspended in the lysis buffer (50mM Hepes-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% Plant Protease Inhibitor and 10%

glycerol), filtered through two layers of Miracloth and resulting slurry sonicated. The sample was centrifuged and resulting supernatant was immunoprecipitated with 25µl of GFP-Trap at 4^0C overnight. The DNA was extracted from the beads and used for qPCR analysis.

Figures and Tables

Table 3.1: Number of proteins containing DUF3223 in various organisms

	Organism/group of organisms	**Number of proteins containing DUF3223**
	Animals	0
	Fungi	0
Bacteria	*Escherichia coli*	0
	Bacillus sp.	0
	Rickettsiales	0
	Cyanobacteria	8
Plants	*Selaginella moellendorffii*	5
	Amborella trichopoda	3
	Vitis vinifera	6
	Glycine max	8
	Solanum tuberosum	5
	Solanum lycopersicum	5
	Arabidopsis thaliana	5
	Populus trichocarpa	7
	Oryza sativa	8
	Sorghum bicolor	7
	Zea mays	5
	Zostera marina	5

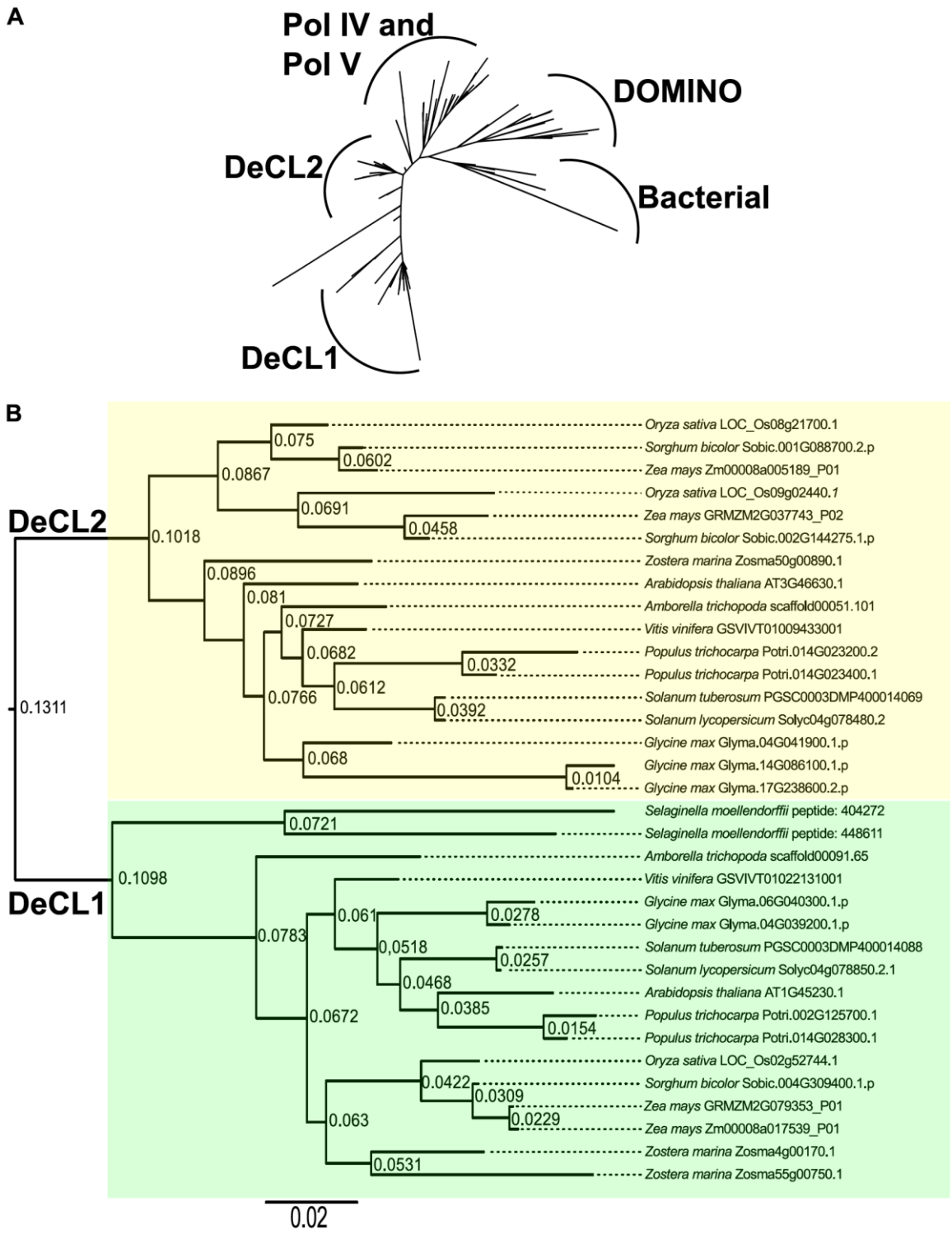

Figure 3.1. Evolutionary history of DeCL proteins. **(A)** Phylogenetic tree of all predicted full length DeCL proteins from selected organisms. **(B)** High resolution phylogenetic tree of DeCL1 and DeCL2 proteins. Green box marks DeCL1s, yellow box marks DeCL2 proteins. Species names and transcript IDs are provided.

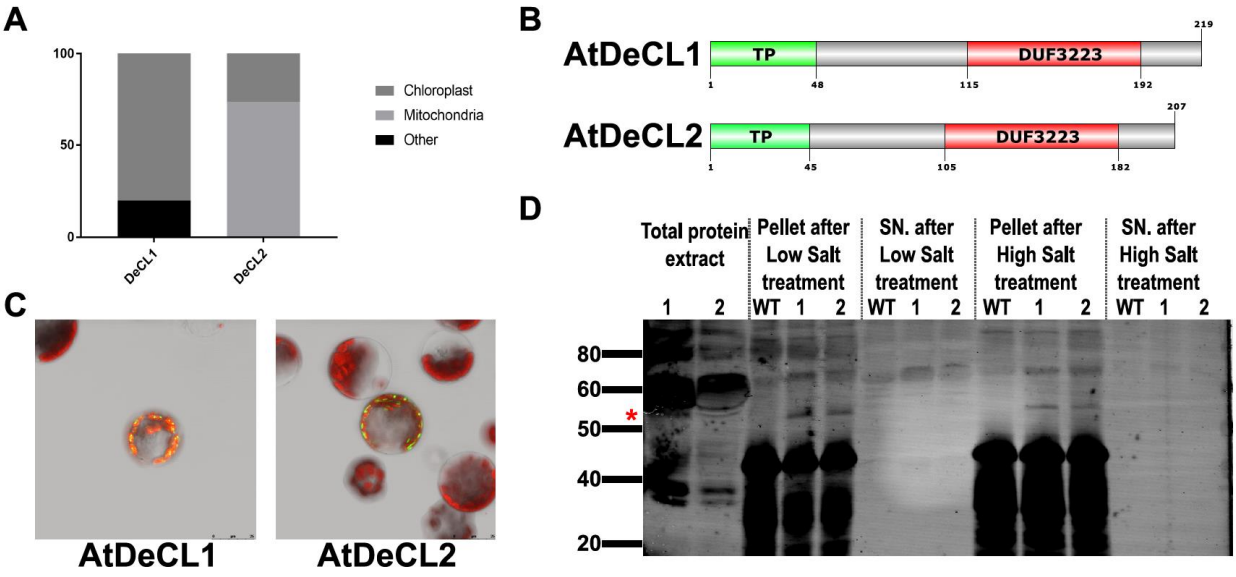

Figure 3.2. Subcellular localization of DeCL proteins. **(A)** Predicted localization of proteins from distinct DeCL groups identified in Fig. 3.1. **(B)** Domain composition of *Arabidopsis* DeCL1 and DeCL2. TP – Targeting Peptide **(C)** DeCL1-GFP (left) and DeCL2-GFP (right) localization in *Arabidopsis* protoplasts. **(D)** α-GFP Western Blot analysis of biochemically fractionated chloroplasts. WT – wild type, 1 – DeCL1-GFP, 2 – DeCL2-GFP. Red asterisk indicates expected size for GFP-tagged DeCL proteins.

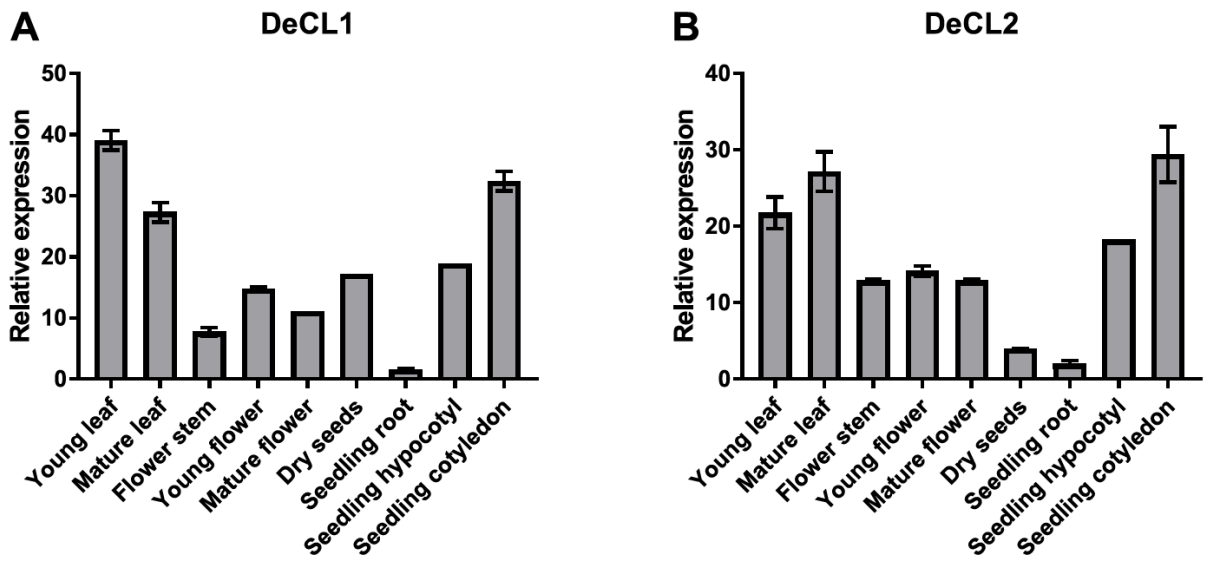

Figure 3.3. Expression of genes encoding **(A)** DeCL1 and **(B)** DeCL2 proteins in *Arabidopsis*. Expression in various organs and at developmental stages was determined by RNA-seq[299] and displayed in relative units provided by the authors. Error bars indicate standard deviations from two biological replicates.

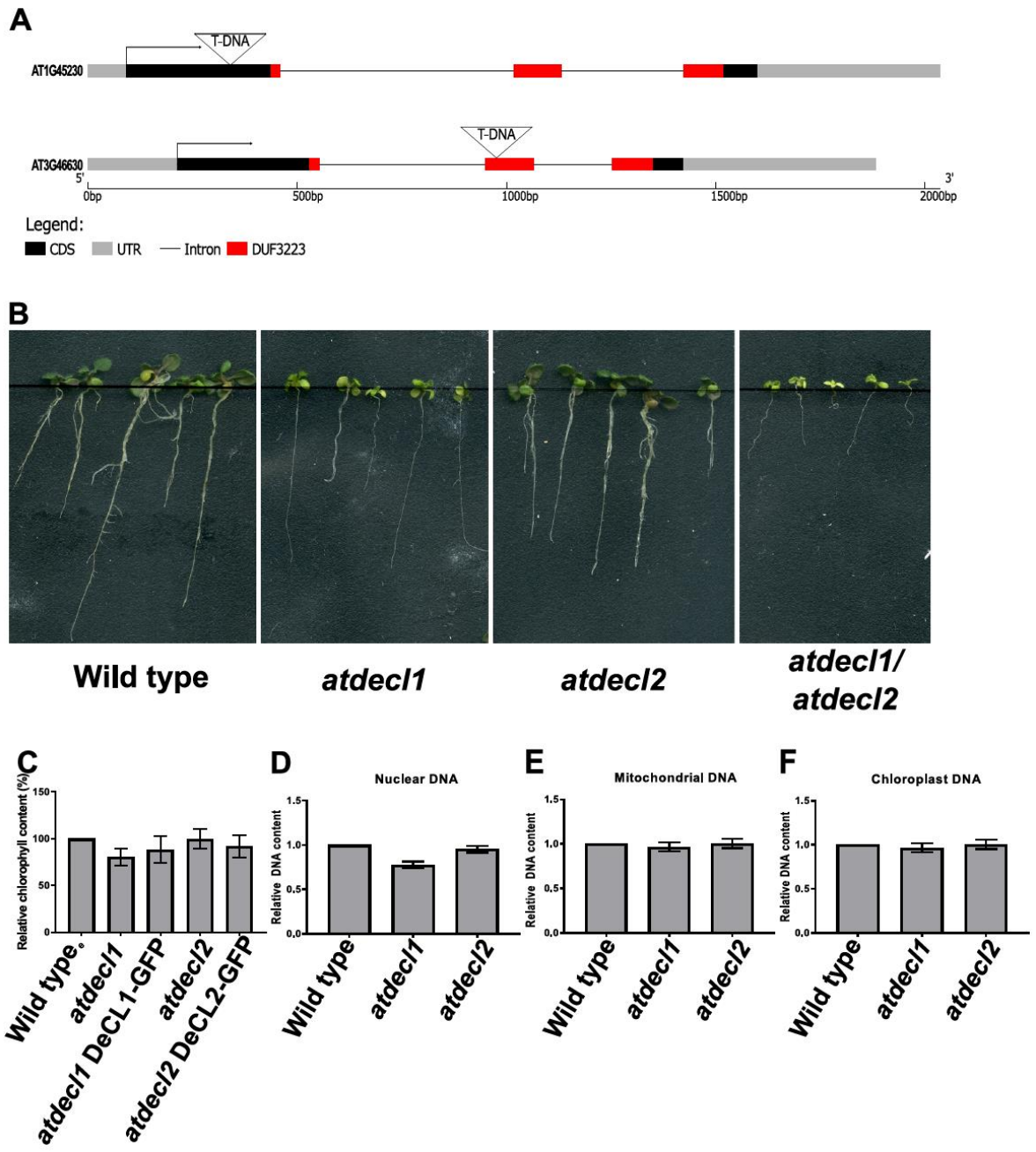

Figure 3.4. Effects of mutations in DeCL genes in *Arabidopsis*. **(A)** Schematic representation of DeCL genes in *Arabidopsis*. Arrows indicate START codon and transcription direction, triangles indicate T-DNA insertion positions. Red boxes indicate positions of DUF3223 domain. **(B)** Approximately 3-week-old plants of wild type (left), *atdecl1* (center) and *atdecl2* (right). **(C)** Relative chlorophyll content of wild type, mutants in genes encoding DeCL proteins and mutant complemented with GFP-tagged DeCLs. Error bars indicate standard deviation from three biological replicates. **(D-F)** Relative levels of nuclear **(D)**, mitochondrial **(E)** and plastid **(F)** DNA in wild type plants, *atdecl* mutants and mutants complemented with GFP-tagged DeCLs.

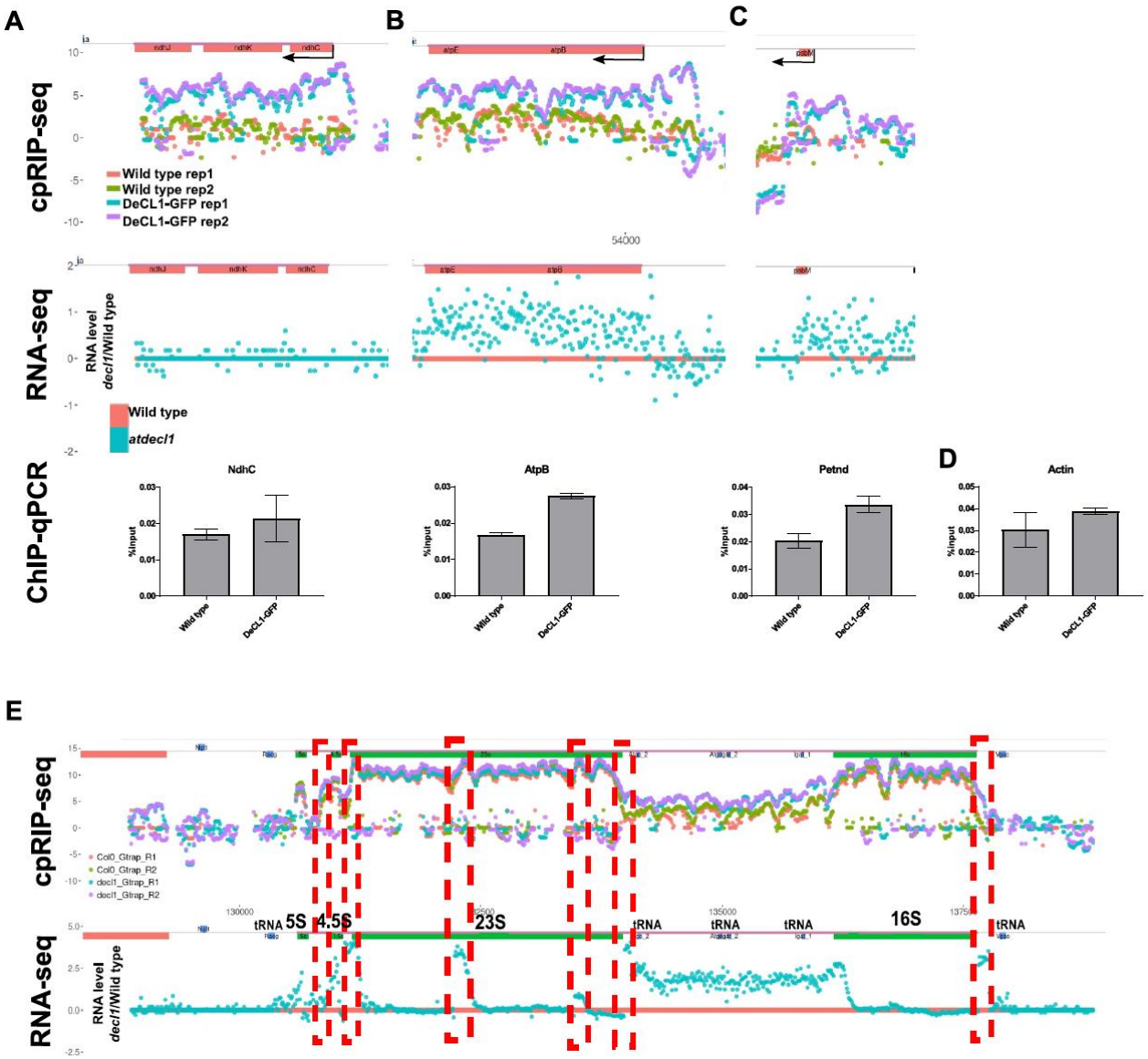

Figure 3.5. DeCL1 interacts with cpRNA but not cpDNA. **(A-C)** Browser screen-shots of cpRIP-seq (top), cpRNA-seq (middle) and ChIP-qPCR (bottom) analysis of corresponding DNA locus. Arrows indicate START codon and transcription direction. **(D)** ChIP-qPCR analysis at nuclear locus (Actin) – negative control. **(E)** Browser screen-shot of cpRIP-seq (top) and cpRNA-seq (bottom) at the rRNA locus. Red dotted lines indicate pre-rRNA processing sites.

CHAPTER IV

Concluding Remarks and Future Directions

Introduction

Throughout the course of this dissertation, plastid RNA metabolism has been discussed in detail (Chapter I), investigation of the *Arabidopsis* RNase H1 proteins functions *in vivo* and *in vitro* have been discussed (Chapter II), and an *in vivo* analysis of *Arabidopsis* DeCL proteins have been performed (Chapter III). This chapter focuses on a brief discussion of unanswered questions within the field of plastid RNA metabolism, biological functions of RNase H1 and DeCL proteins as well as attempts to place obtained results in a broader context. Future directions for investigation of these topics will be acknowledged, accentuating ongoing work on DeCL proteins functions.

Plastid RNA metabolism

Regulation of plastid-encoded gene expression has long been simplified to a model in which it is the RNA processing and translation efficiency that dictates the amount of protein produced, while transcriptional regulation is believed to have little effect on gene expression[162,310]. Relatively much attention has been given to the research on mRNA editing[163,179], a process which has its counterpart in the animal kingdom, but not a lot of interest has focused on other aspects of the plastid-encoded RNA metabolism. Therefore, many questions in this field remain open.

We need to remember that plastids employ a unique combination of molecular mechanisms of prokaryotic and eukaryotic origin. These mechanisms have been evolving through the course of millions of years of co-evolution of endosymbiotic organelles in eukaryotic cells[5]. However, in some cases neither of these mechanisms is employed suggesting the presence of potential new solutions used by endosymbiotic organelles. For example, plastids of land plants have lost HU, H-NS and other nucleoid-associated proteins of cyanobacterial origin[92] but did not take advantage of nuclear histones to organize plastid chromosomes, suggesting different mechanisms responsible for DNA packaging in these organelles. Therefore, we cannot simply assume that any step of plastid biology follows its nuclear or bacterial counterparts. However, thorough characterization of mechanisms employed by plastids may provide foundations for better understanding of processes occurring outside of endosymbiotic organelles as well potentially opening new perspectives for synthetic biology, such as increasing efficiency of photosynthesis or use of plastids as protein overexpression platforms.

It has been shown that plastid-encoded RNA molecules are actively processed by an orchestra of enzymes recognizing either specific ribonucleotide sequence or a secondary RNA structure[126,181,311]. Interestingly, the largest family of proteins interacting with RNAs in plastid and mitochondria, the PPR family, contains enzymes recognizing often a single editing site in the transcriptome[179]. This provides specificity of regulation, however, simultaneously maintaining such large family of proteins, each specific towards a single location, seems to be a substantial expense of energy. Therefore, it is tempting to speculate that precise regulation and processing of RNA molecules in plastids

provides a significant evolutionary advantage. Alternatively, it may be advocated that the low frequency of editing in plastids suggests that this process gives no measurable advantage and might be considered an evolutionary dead end.

Large number of plastid-localized RNA-binding proteins has been identified so far[209]. However, in many cases this localization is predicted based on the presence of N-terminal transfer peptides. Despite many studies aiming to determine the total plastid proteome[24,28,73,74], precise determination of proteins specifically associated with plastid-encoded RNAs remains an open task. Out of the proteins with confirmed plastid localization very few have been thoroughly characterized. In many cases, published research is limited to mutant plant morphology analysis and limited *in vitro* characterization, as it is the case for RNH1C[228]. However, the *in vivo* evidence of direct interaction between a protein and its targets, despite being absolutely crucial for precise characterization, is often missing. Similarly, apart from proteins binding specific RNA sequences, for example PPR proteins, mechanisms of target recognition by RNA processing enzymes remain unclear[126]. It has been shown that RNase J[199] and RNase E[155] initiate and are mainly responsible for intercistronic cleavage of plastid-encoded mRNAs. However, it is unknown how their specificity is achieved. Another example of plastid RNA metabolism is that of secondary RNA structures which were shown to inhibit activity of certain nucleases[126], but how specificity is achieved remains an open question. Likewise, plastid machinery responsible for RNA degradation remains poorly characterized. In bacteria this function is fulfilled by the high molecular weight complex – degradosome[207]. Some plant orthologs of bacterial degradosome components have been shown to interact. However, the presence of the degradosome complex itself has

not been shown[126,209]. Therefore, the exact mechanism responsible for RNA degradation in plastids remains unclear.

Finally, the sub-organellar localization of plastid RNA-associated machinery is often unknown. Plastid nucleoid is believed to be strongly associated with plastid membranes[92], therefore it is hypothesized that in plastids transcription and co-transcriptional translation occur on the surface of plastid membranes[312]. RNA processing is also believed to be mainly co-transcriptional[126]. However, a number of RNA-associated proteins, including RNA-stabilizing factors, have been found in the plastid stroma – the liquid fraction of plastids[311]. An explanation of this intriguing phenomenon remains to be found.

RNase H1 proteins functions

Despite their biochemical activity being known for decades[219], the exact biological role of RNase H1 proteins remains rather mysterious. They have been demonstrated to digest RNA component of RNA:DNA hybrids *in vitro* and *in vivo*[216,251] and shown to be crucial for mitochondrial DNA replication in animals[220]. These results seem to be consistent with its reported function in plastids, where it resolves RNA:DNA hybrids and facilitates DNA replication[228]. However, loss of the RNase H1 protein in mice leads to decreased mtDNA and mRNA levels as well as retention of RNA primers which lead to the conclusion that this protein is responsible for RNA primer removal during DNA replication and prevents DNA polymerase stalling upon encountering RNA:DNA hybrids[89,220]. Interestingly, our results show a significant increase in plastid DNA content upon loss of RNH1C and we did not detect any significant effect of RNase H1 loss on

mRNA accumulation. These discrepancies may suggest different mechanisms of plastid RNase H1 action than the one observed in mitochondria of animals. It has been proposed that RNA:DNA hybrids formed in plastids of plants deficient in RNH1C serve as primers for DNA replication, leading to excessive plastid DNA amplification[228] – a hypothesis consistent with our results. However, lack of misregulation of mRNA accumulation is surprising. It is tempting to speculate, that higher DNA content will correlate with increased mRNA accumulation. Alternatively, nascent mRNA may be quickly degraded, suggesting tight control of transcript accumulation. However, how this precise control may be achieved remains unknown. It may be also hypothesized that the accumulation of RNA:DNA hybrids in plastids of *atrnh1c* plants will result in a decrease of mRNA accumulation due to more frequent transcription – replication machinery collisions. Lack of observable effects on transcript quantity may indicate that such collisions do not happen more often than in wild-type plants and that only a well-defined subset of DNA molecules in plastids undergoes transcription. How such precise determination could be achieved, remains an open question.

Despite a proposed role of *Arabidopsis* RNH1C in plastid DNA replication no direct evidence of interaction between RNH1C and origins of replication or RNA primers for Okazaki fragments has been shown. However, it was shown that RNH1C interacts with DNA gyrase in DNA- and RNA-dependent manner. Interestingly, it was also demonstrated that RNH1C and DNA gyrase are partially redundant in RNA:DNA hybrid removal[228]. This may confirm the proposed involvement of RNH1C in plastid DNA replication. However, it is unknown if RNH1C is a canonical component of DNA replication complex in plastids or if it is recruited only upon DNA polymerase complex

encountering RNA:DNA hybrids. Additionally, little is known about direct targets of RNase H1 proteins. It was shown that RNH1C digests substrates recognized by S9.6 antibody, presumably RNA:DNA hybrids[228]. However, this antibody was shown to efficiently bind ssRNA as well as dsRNA molecules and exhibit preference toward specific sequences[313,314]. Therefore, all results generated with this tool need to be interpreted carefully. Although alternative approaches for RNA:DNA hybrid detection (for example nucleic acid pull-down with catalytically inactive human RNase H1 protein) do exist they were not tested in the case of plant RNase H1 proteins. Additionally, it was found that in *Saccharomyces cerevisiae* RNase H1 enzyme associates with RNA:DNA hybrids genome-wide but acts enzymatically only at a small subset of loci[227]. In contrast, RNase H2 complex binds and digests RNA:DNA hybrids genome-wide. It was proposed that the enzymatic activity of RNase H1 in yeast is regulated at a step after hybrid binding[227]. Whether this is the case for *Arabidopsis* RNase H1 proteins and what triggers, or inhibits, the enzymatic activity of this enzyme, remains to be found.

Loss of mitochondrial RNase H1 in mice leads to embryonic lethality and, as mentioned earlier, decreased mtDNA content[89,220]. Interestingly, loss of RNH1B in *Arabidopsis* does not cause any visible phenotype or changes in mtDNA content. This suggests, at least partial, redundancy between plastid and mitochondrial enzyme. Dual localization is not uncommon for plant plastid and mitochondrial proteins but has not been reported for *Arabidopsis* RNase H1 proteins[267,268]. This potential dual localization may also explain lethality of *atrnh1b/atrnh1c* double mutants, while single mutants do not exhibit this feature. Alternatively, this lethality might be an additive effect of mitochondria and plastid malfunctioning. In this scenario loss of RNH1B in a single

atrnh1b mutant could be complemented by a redundant activity of other, unknown nucleases. Likewise, direct targets of RNH1B, its regulators as well as biological function remain to be discovered.

Similarly to its mitochondrial counterpart, the function of nuclear RNH1A remains unknown. In yeast it was shown to work on a subset of RNase H2 targets, mainly the most highly abundant RNA:DNA hybrids resulting from aberrant transcription[227,315]. However, in plants no function of RNase H2 complex in co-transcriptional RNA:DNA hybrid removal have been shown. Instead, it was shown that in *Arabidopsis* the RNase H2 complex is responsible for the nuclear genome stability and acts as a sensor of DNA damage. Plants deficient in subunits of this complex were less sensitive to replication stress but exhibited increased ribonucleotide incorporation in DNA as well as increased frequency of small base pair deletions[229,231]. Whether in plants the nuclear RNase H1 works together with RNase H2 complex, or if their activities are limited to separate processes, remains unknown and substantial genome-wide data is required to conclusively answer this question.

Despite recent advances in understanding the biological roles of RNase H1 proteins, the list of their direct targets is limited. It is possible that this group of enzymes does not have specific targets defined by the sequences but rather scan the genome in search of most abundant RNA:DNA hybrids[227]. Interestingly, potential RNase H1 targets, R-loops, have been found to be involved in regulation of gene expression[224]. This may suggest that RNase H1 proteins are involved in the regulation of expression of specific genes, however direct evidence has not been shown so far. Additionally, *in vitro* activity of RNase H1 proteins is not dependent on the presence of additional proteins[251].

Therefore, it is possible that *in vivo* enzymatic activity of RNase H1s is controlled by additional regulatory proteins, which allow it to act only when RNase H2 complex, or other redundant protein (such as topoisomerase) activity is not sufficient. Whether this is the case and what actually regulates RNase H1 activity, remains to be discovered.

Recent genome-wide identification of RNA:DNA hybrids in *Arabidopsis* may provide a strong basis for further characterization of plant nuclear RNase H proteins[316]. It was shown that the presence of potential substrates of RNase H1 or RNase H2 is a common feature of nuclear chromatin. These substrates are equally split between Watson and Crick strands of DNA and are usually ~200bp long. RNA:DNA hybrids were found to be enriched at gene promoters and depleted in the intergenic regions. They are also present in gene bodies and UTRs. Out of approximately 28,000 annotated genes, 24 000 were found to contain at least one RNA:DNA hybrid. Interestingly, these structures were found in all chromatin contexts, and regions transcribed by all five plant polymerases[316]. Altogether, this data suggests that potential RNase H substrates are commonly found in *Arabidopsis* nucleus. Therefore, precisely regulated metabolism of these features seems indispensable and RNase H1 proteins are potentially very important element of this mechanism.

DeCL proteins functions

DeCL proteins have been proposed to be involved in rRNA maturation in tomato and *Arabidopsis* based on the observation that plants deficient in this enzyme accumulate rRNA precursors[232–234]. Additionally, nuclear counterparts of DeCLs, DOMINO proteins, have been found to be involved in rDNA condensation and nucleolar

dominance[235]. Interestingly, we found that at least one of *Arabidopsis* DeCL proteins binds multiple plastid-encoded mRNA species. This suggests a direct role of DeCLs in regulation of plastid-encoded genes expression. However, the biochemical function of these proteins remains unknown. It was proposed that DeCLs process pre-rRNAs but no direct evidence was shown[233]. Therefore, it is also possible that these proteins bind transcripts and either stabilize and protect them from degradation or recruit other factors involved in mRNA processing. Importantly, we have found no major effects of DeCL loss on mRNA accumulation. It is possible that DeCLs, either directly or indirectly, stabilize nascent mRNAs and facilitate their translation. This hypothesis will be validated using the Ribosome Profiling approach which allows measuring translation efficiency. Moreover, this hypothesis is consistent with the proposed role of DeCL-like domain of nuclear RNA Polymerase V[242]. Importantly, the DeCL-like domain is the only structured domain distinguishing RNA Pol IV and Pol V from Polymerase II[292]. It was proposed that the main function of this domain is to bind ncRNA which is further processed by RRP6L exonuclease[242]. However, the reason for why a prokaryotic factor would be required for eukaryotic DNA polymerase complex remains an open question. Additionally, no evidence was shown for the necessity of DeCL domain presence in the largest subunit of Pol IV so far. Likewise, specificity of organellar DeCLs toward a certain subset of transcripts is an interesting feature. We found no specific sequence or position within mRNA that would be preferred by DeCL1. It is possible that other proteins play a role in DeCL specificity but they yet remain to be discovered.

Interestingly, sequence-based subcellular localization prediction of DeCL proteins, except for the nuclear DOMINO and Pol IV and Pol V subunits, suggests that

the majority of plants have at least two DeCLs localizing to plastids and mitochondria. This leads to a hypothesis that both DeCL proteins serve the same molecular function but in different subcellular compartments. However, both biochemical fractionation as well as microscopy observations suggest plastid localization of both DeCL proteins from *Arabidopsis*. Whether it is the case for DeCLs from other plants remains an interesting question. Likewise, functional separation of *Arabidopsis* proteins is possible but the reason for it is unknown. Based on lethality of double *atdecl1/atdecl2* mutants at least partial redundancy between these two proteins can be proposed. An additional question arises regarding the mitochondrial counterpart of plastid DeCL. Considering prokaryotic origin of both plastids and mitochondria as well as millions of years of their co-evolution in eukaryotic cells, one would hypothesize that on a molecular level the majority of processes involving nucleic acids will be similar between nucleus, plastids and mitochondria. Nucleus contains DOMINO protein which may serve similar function to that carried by plastid DeCL. Therefore, which mitochondrial protein fulfills molecular function of plastid DeCL remains an intriguing question. Alternatively, it is possible that DOMINO only structurally resemble DeCL proteins but have been adapted to serve different molecular function and DeCLs are required only for plastid-specific RNA metabolism. Another interesting question is the sub-organellar localization of DeCL proteins from *Arabidopsis*. Neither of these proteins contains a trans-membrane domain but biochemical fractionation of chloroplasts suggests membrane localization for both. Also, microscopy observations, resulting in a signal typical for nucleoid-associated proteins, suggest membrane localization of these proteins. It is possible that this colocalization allows for co-transcriptional processing of specific mRNAs. But if this is truly the case remains an open question.

Determination of DeCLs functions will potentially increase the knowledge about DOMINO proteins as well as the DeCL domain of Pol V. The exact function of DOMINO was not demonstrated but it is possible that it affects nuclear rDNA and rRNA in a different manner than anticipated. It is possible that it interacts with rRNA and either stabilizes it or serves as a scaffold for other proteins. Likewise, DOMINO may act on a different subset of RNA species, like mRNAs and regulate rRNAs indirectly. The exact role of the C-terminal DeCL domain of the largest subunits Pol IV and Pol V complexes remains speculative. It was shown to be required for Pol V transcription *in vivo* but not *in vitro* as well as downstream effects of Pol V activity but no direct function was demonstrated[242]. The role in non-sequence specific RNA binding seems to be the most plausible hypothesis but does not explain why it is required for transcription itself. It is possible that in the absence of DeCL domain Pol V-produced transcripts are quickly degraded by non-specific exonucleases. Alternatively, it is possible that nascent RNA produced by Pol V forms a higher order structure which obstructs transcription elongation and leads to premature transcription termination. Altogether, it can be hypothesized that DeCL domain serves a dual function. First, it binds and stabilizes nascent transcript and when downstream RdDM machinery is already recruited, it interacts with exonuclease RRP6L which specifically degrades Pol V-produced transcripts. However, whether this is true, remains to be determined. It seems that the only functional similarity between DOMINOs, DeCL domains of Pol IV and Pol V as well as organellar DeCL proteins is their hypothesized role in RNA metabolism. However, direct interactions between abovementioned proteins and RNA molecules remain to be shown. Although our RIP-seq experiment strongly supports such interactions between DeCL1 and plastid-encoded mRNAs, it is still possible that there are other proteins

involved in this process, especially that no obvious RNA-binding domain can be identified in the sequence of DOMINOs or organellar DeCL proteins.

Altogether, it is possible that DeCLs serve at least one of the hypothesized biochemical roles. First, it might be responsible for protein binding and serve as a docking platform for enzymes modifying nucleic acids. In this scenario, DeCLs would interact with RNA or DNA indirectly. Alternatively, it is possible that these proteins directly bind nucleic acids and recruit enzymatic machinery to specific sites of RNA or DNA molecules or protect these sites from degradation by nucleases. Yet another possibility is that DeCLs exhibit enzymatic activity and are able to digest or modify nucleic acids without other enzymatic factors. Finally, we cannot exclude that DeCLs play multiple biochemical roles in nucleic acid metabolism.

Future directions

New mechanisms of RNA metabolism await further exploration. Specifically, discovery of direct targets of *Arabidopsis* RNH1C protein as well as its interactors, apart from DNA gyrase[228], will provide substantial knowledge toward understanding the precise roles of this enzyme. Similarly, determination of protein partners of nuclear and mitochondrial RNase H1s and RNA:DNA hybrids that they resolve remains an important goal. Double mutant lacking RNH1A and RNH1B displays no obvious phenotype. This strongly indicates that both nuclear and mitochondrial genomes possess alternative pathways for resolving RNA:DNA hybrids. Finding these mechanisms, at least partially redundant with RNase H1s, remains an important goal toward understanding the metabolism of RNA:DNA hybrids. Likewise, it is an important task to find a molecular

basis of embryonic lethality of *atrnh1b/atrnh1c* mutants. Importantly, mutants deficient in only one of these two RNase H1s do not die during embryo development. This suggests that RNH1B and RNH1C are partially redundant or that this lethality is a result of additive effect of plastids and mitochondria misfunctioning. However, the molecular phenotype of RNase H1 loss in mitochondria remains to be found. Additionally, we found that the expression of genes encoding RNH1s is often increased upon the loss of another RNH1 protein. This suggests the existence of autoregulatory mechanism of RNH1s expression. What is the exact mode of action of this mechanism and if RNH1s can exhibit dual localization is an intriguing and important question. Also the biological relevance of splice variant of RNH1B, the RNH1B.3 remains to be tested. It is particularly intriguing that even upon loss of the mitochondrial presequence, this protein, contrary to the phenomenon observed in animals[220], does not exhibit a nuclear localization. Whether this protein is produced at all and if and where it is important for nucleic acid metabolism remains an important question for future research. Finally, it was proposed that yeast RNase H1 protein can only digest the substrate upon activation by another protein[227]. Our *in vitro* data suggest that for *Arabidopsis* RNase H1 enzymes such activator is not required. But it cannot be ruled out that *in vivo* catalytic activity of these enzymes is regulated by an inhibitor. Whether this is the case and what protein, or perhaps different molecule, might serve this function remains an intriguing question to test.

Similarly to RNase H1s, many questions remain open in the biology of DeCL proteins. First, it needs to be determined if these proteins directly interact with RNAs, possibly through *in vitro* assays, and if specific sequences or RNA structures are their

preferred substrates. Additionally, extensive *in vitro* characterization of these proteins and their enzymatic activities need to be performed. Furthermore, actual *in vivo* targets of both DeCL1 and DeCL2 need to be found through RIP-seq. Additionally, sub-organellar localization of DeCL proteins remains to be found. Preliminary results suggest that both DeCL1 and DeCL2 are tightly associated with plastid membranes. Whether this is the case and the underlying mechanism of this localization needs to be found. We speculate that organellar DeCL proteins regulate expression of plastid-encoded proteins at the level of translation and will test this hypothesis using Ribosome Profiling approach. Furthermore, it is important to determine the exact mechanism of DeCLs action and possibly other biological processes they might be involved in. Additionally, it is important to determine proteins interacting with DeCLs which might reveal potential regulators of their activity. Finally, results obtained from organellar DeCL proteins may allow for extrapolating roles served by DeCL domains present in the structures of nuclear Pol IV and Pol V as well as DOMINO protein. Importantly, transcripts produced by abovementioned polymerases do not undergo translation. Therefore, any conclusions need to be drawn very carefully and any potential roles need to be independently and carefully tested in the context of nuclear RNA metabolism.

BIBLIOGRAPHY

1. McFadden, G. I. Origin and Evolution of Plastids and Photosynthesis in Eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6**, (2014).
2. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **14**, 255–274 (1967).
3. Zimorski, V., Ku, C., Martin, W. F. & Gould, S. B. Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* **22**, 38–48 (2014).
4. Gray, M. W. Mitochondrial Evolution. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
5. Gould, S. B., Waller, R. F. & McFadden, G. I. Plastid Evolution. *Annu. Rev. Plant Biol.* **59**, 491–517 (2008).
6. Marin, B., Nowack, E. C. M. & Melkonian, M. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156**, 425–432 (2005).
7. Archibald, J. M. The puzzle of plastid evolution. *Curr. Biol. CB* **19**, R81-88 (2009).
8. Howe, C. J., Barbrook, A. C., Nisbet, R. E. R., Lockhart, P. J. & Larkum, A. W. D. The origin of plastids. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363**, 2675–2685 (2008).
9. Maréchal, E. Primary Endosymbiosis: Emergence of the Primary Chloroplast and the Chromatophore, Two Independent Events. *Methods Mol. Biol. Clifton NJ* **1829**, 3–16 (2018).
10. Rumpho, M. E., Pelletreau, K. N., Moustafa, A. & Bhattacharya, D. The making of a photosynthetic animal. *J. Exp. Biol.* **214**, 303–311 (2011).
11. Pillet, L. The role of horizontal gene transfer in kleptoplastidy and the establishment of photosynthesis in the eukaryotes. *Mob. Genet. Elem.* **3**, (2013).
12. Rumpho, M. E. *et al.* Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug Elysia chlorotica. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 17867–17871 (2008).
13. Juniper, B. E. & Clowes, F. a. L. Cytoplasmic Organelles and Cell Growth in Root Caps. *Nature* **208**, 864–865 (1965).
14. Cran, D. G. & Possingham, J. V. Variation of plastid types in spinach. *Protoplasma* **74**, 345–356 (1972).
15. Lyndon, R. F. & Robertson, E. S. The quantitative ultrastructure of the pea shoot apex in relation to leaf initiation. *Protoplasma* **87**, 387–402 (1976).
16. Liebers, M. *et al.* Regulatory Shifts in Plastid Transcription Play a Key Role in Morphological Conversions of Plastids during Plant Development. *Front. Plant Sci.* **8**, (2017).
17. Chen, M. *et al.* Arabidopsis HEMERA/pTAC12 Initiates Photomorphogenesis by Phytochromes. *Cell* **141**, 1230–1240 (2010).
18. Chen, M., Tao, Y., Lim, J., Shaw, A. & Chory, J. Regulation of Phytochrome B Nuclear Localization through Light-Dependent Unmasking of Nuclear-Localization Signals. *Curr. Biol.* **15**, 637–642 (2005).
19. Waters, M. T. *et al.* GLK Transcription Factors Coordinate Expression of the Photosynthetic Apparatus in Arabidopsis. *Plant Cell* **21**, 1109–1128 (2009).
20. Castillon, A., Shen, H. & Huq, E. Phytochrome Interacting Factors: central players in phytochrome-mediated light signaling networks. *Trends Plant Sci.* **12**, 514–521 (2007).
21. Pribil, M., Labs, M. & Leister, D. Structure and dynamics of thylakoids in land plants. *J. Exp. Bot.* **65**, 1955–1972 (2014).
22. Pogson, B. J., Ganguly, D. & Albrecht-Borth, V. Insights into chloroplast biogenesis and development. *Biochim. Biophys. Acta BBA - Bioenerg.* **1847**, 1017–1024 (2015).
23. Sun, T. *et al.* Carotenoid Metabolism in Plants: The Role of Plastids. *Mol. Plant* **11**, 58–74 (2018).
24. Grennan, A. K. Plastoglobule Proteome. *Plant Physiol.* **147**, 443–445 (2008).

25. Austin, J. R., Frost, E., Vidi, P.-A., Kessler, F. & Staehelin, L. A. Plastoglobules are lipoprotein subcompartments of the chloroplast that are permanently coupled to thylakoid membranes and contain biosynthetic enzymes. *Plant Cell* **18**, 1693–1703 (2006).
26. van Wijk, K. J. & Kessler, F. Plastoglobuli: Plastid Microcompartments with Integrated Functions in Metabolism, Plastid Developmental Transitions, and Environmental Adaptation. *Annu. Rev. Plant Biol.* **68**, 253–289 (2017).
27. Balmer, Y., Vensel, W. H., DuPont, F. M., Buchanan, B. B. & Hurkman, W. J. Proteome of amyloplasts isolated from developing wheat endosperm presents evidence of broad metabolic capability. *J. Exp. Bot.* **57**, 1591–1602 (2006).
28. Hurkman, W. J., Vensel, W. H., Dupont, F. M., Altenbach, S. B. & Buchanan, B. B. Endosperm and Amyloplast Proteomes of Wheat Grain. in *Plant Proteomics* 207–222 (John Wiley & Sons, Ltd, 2008). doi:10.1002/9780470369630.ch14.
29. Zhu, M. *et al.* A comprehensive proteomic analysis of elaioplasts from citrus fruits reveals insights into elaioplast biogenesis and function. *Hortic. Res.* **5**, 6 (2018).
30. Vigil, E. L. & Ruddat, M. Development and enzyme activity of protein bodies in proteinoplasts of tobacco root cells. *Histochemistry* **83**, 17–27 (1985).
31. Matile, P., Hortensteiner, S., Thomas, H. & Krautler, B. Chlorophyll Breakdown in Senescent Leaves. *Plant Physiol.* **112**, 1403–1409 (1996).
32. Hörtensteiner, S. & Kräutler, B. Chlorophyll breakdown in higher plants. *Biochim. Biophys. Acta* **1807**, 977–988 (2011).
33. Miyagishima, S. -y. Mechanism of Plastid Division: From a Bacterium to an Organelle. *PLANT Physiol.* **155**, 1533–1544 (2011).
34. Hwang, I. Plastid biogenesis and homeostasis. *Plant Cell Rep.* (2019) doi:10.1007/s00299-019-02437-7.
35. Osteryoung, K. W. & Nunnari, J. The division of endosymbiotic organelles. *Science* **302**, 1698–1704 (2003).
36. Kuroiwa, T. *et al.* Vesicle, mitochondrial, and plastid division machineries with emphasis on dynamin and electron-dense rings. *Int. Rev. Cell Mol. Biol.* **271**, 97–152 (2008).
37. Miyagishima, S., Nakanishi, H. & Kabeya, Y. Structure, regulation, and evolution of the plastid division machinery. *Int. Rev. Cell Mol. Biol.* **291**, 115–153 (2011).
38. Yoshida, Y., Miyagishima, S., Kuroiwa, H. & Kuroiwa, T. The plastid-dividing machinery: formation, constriction and fission. *Curr. Opin. Plant Biol.* **15**, 714–721 (2012).
39. Yoshida, Y. *et al.* Chloroplasts divide by contraction of a bundle of nanofilaments consisting of polyglucan. *Science* **329**, 949–953 (2010).
40. Glynn, J. M., Froehlich, J. E. & Osteryoung, K. W. Arabidopsis ARC6 coordinates the division machineries of the inner and outer chloroplast membranes through interaction with PDV2 in the intermembrane space. *Plant Cell* **20**, 2460–2470 (2008).
41. Glynn, J. M. *et al.* PARC6, a novel chloroplast division factor, influences FtsZ assembly and is required for recruitment of PDV1 during chloroplast division in Arabidopsis. *Plant J. Cell Mol. Biol.* **59**, 700–711 (2009).
42. Zhang, M., Chen, C., Froehlich, J. E., TerBush, A. D. & Osteryoung, K. W. Roles of Arabidopsis PARC6 in Coordination of the Chloroplast Division Complex and Negative Regulation of FtsZ Assembly. *Plant Physiol.* **170**, 250–262 (2016).
43. Wang, W. *et al.* Structural insights into the coordination of plastid division by the ARC6-PDV2 complex. *Nat. Plants* **3**, 17011 (2017).
44. Sumiya, N., Fujiwara, T., Era, A. & Miyagishima, S.-Y. Chloroplast division checkpoint in eukaryotic algae. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7629–E7638 (2016).

45. Miyagishima, S.-Y., Suzuki, K., Okazaki, K. & Kabeya, Y. Expression of the nucleus-encoded chloroplast division genes and proteins regulated by the algal cell cycle. *Mol. Biol. Evol.* **29**, 2957–2970 (2012).
46. Raynaud, C. *et al.* Cell and plastid division are coordinated through the prereplication factor AtCDT1. *Proc. Natl. Acad. Sci.* **102**, 8216–8221 (2005).
47. Kobayashi, Y. *et al.* Tetrapyrrole signal as a cell-cycle coordinator from organelle to nuclear DNA replication in plant cells. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 803–807 (2009).
48. Kobayashi, Y., Imamura, S., Hanaoka, M. & Tanaka, K. A tetrapyrrole-regulated ubiquitin ligase controls algal nuclear DNA replication. *Nat. Cell Biol.* **13**, 483–487 (2011).
49. Miyagishima, S. *et al.* Translation-independent circadian control of the cell cycle in a unicellular photosynthetic eukaryote. *Nat. Commun.* **5**, 3807 (2014).
50. Oldenburg, D. J. & Bendich, A. J. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *J. Mol. Biol.* **335**, 953–970 (2004).
51. Oldenburg, D. J. & Bendich, A. J. Changes in the structure of DNA molecules and the amount of DNA per plastid during chloroplast development in maize. *J. Mol. Biol.* **344**, 1311–1330 (2004).
52. Shaver, J. M., Oldenburg, D. J. & Bendich, A. J. The structure of chloroplast DNA molecules and the effects of light on the amount of chloroplast DNA during development in Medicago truncatula. *Plant Physiol.* **146**, 1064–1074 (2008).
53. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. & Tabata, S. Complete structure of the chloroplast genome of Arabidopsis thaliana. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **6**, 283–290 (1999).
54. Smith, D. R. Evolution: A Plant Plastid Genome that Has Forsaken Guanine and Cytosine. *Curr. Biol. CB* **29**, R99–R101 (2019).
55. Kleine, T., Maier, U. G. & Leister, D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu. Rev. Plant Biol.* **60**, 115–138 (2009).
56. Allen, J. F. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10231–10238 (2015).
57. Archibald, J. M. Genomic perspectives on the birth and spread of plastids. *Proc. Natl. Acad. Sci.* **112**, 10147–10153 (2015).
58. Eckardt, N. A. Genomic Hopscotch: Gene Transfer from Plastid to Nucleus. *Plant Cell* **18**, 2865–2867 (2006).
59. Allen, J. F. The CoRR hypothesis for genes in organelles. *J. Theor. Biol.* **434**, 50–57 (2017).
60. Puthiyaveetil, S. *et al.* Transcriptional control of photosynthesis genes: the evolutionarily conserved regulatory mechanism in plastid genome function. *Genome Biol. Evol.* **2**, 888–896 (2010).
61. von Heijne, G. Why mitochondria need a genome. *FEBS Lett.* **198**, 1–4 (1986).
62. Daley, D. O. & Whelan, J. Why genes persist in organelle genomes. *Genome Biol.* **6**, 110 (2005).
63. Björkholm, P., Harish, A., Hagström, E., Ernst, A. M. & Andersson, S. G. E. Mitochondrial genomes are retained by selective constraints on protein targeting. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10154–10161 (2015).
64. Barbrook, A. C., Howe, C. J., Kurniawan, D. P. & Tarr, S. J. Organization and expression of organellar genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 785–797 (2010).
65. Smith, D. R. & Keeling, P. J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10177–10184 (2015).
66. Su, H.-J. *et al.* Novel genetic code and record-setting AT-richness in the highly reduced plastid genome of the holoparasitic plant Balanophora. *Proc. Natl. Acad. Sci.* **116**, 934–943 (2019).
67. Wicke, S. & Naumann, J. Chapter Eleven - Molecular Evolution of Plastid Genomes in Parasitic Flowering Plants. in *Advances in Botanical Research* (eds. Chaw, S.-M. & Jansen, R. K.) vol. 85 315–347 (Academic Press, 2018).

68. Simpson, C. L. & Stern, D. B. The Treasure Trove of Algal Chloroplast Genomes. Surprises in Architecture and Gene Content, and Their Functional Implications. *Plant Physiol.* **129**, 957–966 (2002).
69. Robbens, S. *et al.* The Complete Chloroplast and Mitochondrial DNA Sequence of Ostreococcus tauri: Organelle Genomes of the Smallest Eukaryote Are Examples of Compaction. *Mol. Biol. Evol.* **24**, 956–968 (2007).
70. Boudreau, E. & Turmel, M. Extensive gene rearrangements in the chloroplast DNAs of Chlamydomonas species featuring multiple dispersed repeats. *Mol. Biol. Evol.* **13**, 233–243 (1996).
71. Sato, N., Terasawa, K., Miyajima, K. & Kabeya, Y. Organization, developmental dynamics, and evolution of plastid nucleoids. *Int. Rev. Cytol.* **232**, 217–262 (2003).
72. Kuroiwa, T. Review of cytological studies on cellular and molecular mechanisms of uniparental (maternal or paternal) inheritance of plastid and mitochondrial genomes induced by active digestion of organelle nuclei (nucleoids). *J. Plant Res.* **123**, 207–230 (2010).
73. Majeran, W. *et al.* Nucleoid-enriched proteomes in developing plastids and chloroplasts from maize leaves: a new conceptual framework for nucleoid functions. *Plant Physiol.* **158**, 156–189 (2012).
74. Melonek, J., Oetke, S. & Krupinska, K. Multifunctionality of plastid nucleoids as revealed by proteome analyses. *Biochim. Biophys. Acta* **1864**, 1016–1038 (2016).
75. Sakai, A. *et al.* Comparative analysis of DNA synthesis activity in plastid-nuclei and mitochondrial-nuclei simultaneously isolated from cultured tobacco cells. *Plant Sci.* **140**, 9–19 (1999).
76. Sakai, A. In vitro Transcription/DNA Synthesis Using Isolated Organelle-nuclei: Application to the Analysis of the Mechanisms that Regulate Organelle Genome Function. *J. Plant Res.* **114**, 199–211 (2001).
77. Ono, Y. *et al.* NtPolI-like1 and NtPolI-like2, bacterial DNA polymerase I homologs isolated from BY-2 cultured tobacco cells, encode DNA polymerases engaged in DNA replication in both plastids and mitochondria. *Plant Cell Physiol.* **48**, 1679–1692 (2007).
78. Morley, S. A. & Nielsen, B. L. Chloroplast DNA Copy Number Changes during Plant Development in Organelle DNA Polymerase Mutants. *Front. Plant Sci.* **7**, (2016).
79. Udy, D. B., Belcher, S., Williams-Carrier, R., Gualberto, J. M. & Barkan, A. Effects of reduced chloroplast gene copy number on chloroplast gene expression in maize. *Plant Physiol.* **160**, 1420–1431 (2012).
80. Kolodner, R. D. & Tewari, K. K. Chloroplast DNA from higher plants replicates by both the Cairns and the rolling circle mechanism. *Nature* **256**, 708–711 (1975).
81. Nielsen, B. L., Cupp, J. D. & Brammer, J. Mechanisms for maintenance, replication, and repair of the chloroplast genome in plants. *J. Exp. Bot.* **61**, 2535–2537 (2010).
82. Kunnimalaiyaan, M. & Nielsen, B. L. Chloroplast DNA Replicatie'n Mechanism, Enzymes and Replication Origins. 7.
83. Manchekar, M. *et al.* DNA Recombination Activity in Soybean Mitochondria. *J. Mol. Biol.* **356**, 288–299 (2006).
84. Backert, S. & Börner, T. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). *Curr. Genet.* **37**, 304–314 (2000).
85. Kobayashi, Y. *et al.* Holliday junction resolvases mediate chloroplast nucleoid segregation. *Science* **356**, 631–634 (2017).
86. Maréchal, A. & Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytol.* **186**, 299–317 (2010).
87. Krishnan, N. M. & Rao, B. J. A comparative approach to elucidate chloroplast genome replication. *BMC Genomics* **10**, 237 (2009).
88. Wu, M., Lou, J. K., Chang, D. Y., Chang, C. H. & Nie, Z. Q. Structure and function of a chloroplast DNA replication origin of Chlamydomonas reinhardtii. *Proc. Natl. Acad. Sci.* **83**, 6761–6765 (1986).

89. Holmes, J. B. *et al.* Primer retention owing to the absence of RNase H1 is catastrophic for mitochondrial DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 9334–9339 (2015).
90. Lima, W. F. *et al.* Viable RNaseH1 knockout mice show RNaseH1 is essential for R loop processing, mitochondrial and liver function. *Nucleic Acids Res.* **44**, 5299–5312 (2016).
91. Oldenburg, D. J. & Bendich, A. J. DNA maintenance in plastids and mitochondria of plants. *Front. Plant Sci.* **6**, (2015).
92. Sakamoto, W. & Takami, T. Chloroplast DNA Dynamics: Copy Number, Quality Control and Degradation. *Plant Cell Physiol.* **59**, 1120–1127 (2018).
93. Barbrook, A. C., Howe, C. J. & Purton, S. Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci.* **11**, 101–108 (2006).
94. Barkan, A. Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs. *EMBO J.* **7**, 2637–2644 (1988).
95. Westhoff, P. & Herrmann, R. G. Complex RNA maturation in chloroplasts. The psbB operon from spinach. *Eur. J. Biochem.* **171**, 551–564 (1988).
96. Ortelt, J. & Link, G. Plastid gene transcription: promoters and RNA polymerases. *Methods Mol. Biol. Clifton NJ* **1132**, 47–72 (2014).
97. Maliga, P. Two plastid RNA polymerases of higher plants: an evolving story. *Trends Plant Sci.* **3**, 4–6 (1998).
98. Liere, K. & Maliga, P. In vitro characterization of the tobacco rpoB promoter reveals a core sequence motif conserved between phage-type plastid and plant mitochondrial promoters. *EMBO J.* **18**, 249–257 (1999).
99. Hess, W. R. & Börner, T. Organellar RNA Polymerases of Higher Plants. in *International Review of Cytology* (ed. Jeon, K. W.) vol. 190 1–59 (Academic Press, 1999).
100. Lerbs-Mache, S. The 110-kDa polypeptide of spinach plastid DNA-dependent RNA polymerase: single-subunit enzyme or catalytic core of multimeric enzyme complexes? *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5509–5513 (1993).
101. Liere, K., Kaden, D., Maliga, P. & Börner, T. Overexpression of phage-type RNA polymerase RpoTp in tobacco demonstrates its role in chloroplast transcription by recognizing a distinct promoter type. *Nucleic Acids Res.* **32**, 1159–1165 (2004).
102. Kirk, J. T. Studies on RNA synthesis in chloroplast preparations. *Biochem. Biophys. Res. Commun.* **16**, 233–238 (1964).
103. Bottomley, W., Smith, H. J. & Bogorad, L. RNA Polymerases of Maize: Partial Purification and Properties of the Chloroplast Enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 2412 (1971).
104. McKown, R. L. & Tewari, K. K. Purification and properties of a pea chloroplast DNA polymerase. *Proc. Natl. Acad. Sci.* **81**, 2354–2358 (1984).
105. Lerbs, S., Briat, J.-F. & Mache, R. Chloroplast RNA polymerase from spinach: purification and DNA-binding proteins. *Plant Mol. Biol.* **2**, 67–74 (1983).
106. Sutherland, C. & Murakami, K. S. An Introduction to the Structure and Function of the catalytic core enzyme of Escherichia coli RNA polymerase. *EcoSal Plus* **8**, (2018).
107. Shiina, T., Tsunoyama, Y., Nakahira, Y. & Khan, M. S. Plastid RNA polymerases, promoters, and transcription regulators in higher plants. *Int. Rev. Cytol.* **244**, 1–68 (2005).
108. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. Factor stimulating transcription by RNA polymerase. *Nature* **221**, 43–46 (1969).
109. Bervoets, I. *et al.* A sigma factor toolbox for orthogonal gene expression in Escherichia coli. *Nucleic Acids Res.* **46**, 2133–2144 (2018).
110. Wösten, M. M. Eubacterial sigma-factors. *FEMS Microbiol. Rev.* **22**, 127–150 (1998).
111. Gruber, T. M. & Gross, C. A. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57**, 441–466 (2003).

112. Ishihama, A. Functional modulation of Escherichia coli RNA polymerase. *Annu. Rev. Microbiol.* **54**, 499–518 (2000).
113. Tanaka, K. *et al.* Nuclear encoding of a chloroplast RNA polymerase sigma subunit in a red alga. *Science* **272**, 1932–1935 (1996).
114. Chi, W., He, B., Mao, J., Jiang, J. & Zhang, L. Plastid sigma factors: Their individual functions and regulation in transcription. *Biochim. Biophys. Acta* **1847**, 770–778 (2015).
115. Bohne, A.-V., Irihimovitch, V., Weihe, A. & Stern, D. B. Chlamydomonas reinhardtii encodes a single sigma70-like factor which likely functions in chloroplast transcription. *Curr. Genet.* **49**, 333–340 (2006).
116. Fujiwara, M., Nagashima, A., Kanamaru, K., Tanaka, K. & Takahashi, H. Three new nuclear genes, sigD, sigE and sigF, encoding putative plastid RNA polymerase sigma factors in Aarabidopsis thaliana. *FEBS Lett.* **481**, 47–52 (2000).
117. Tozawa, Y. *et al.* The plastid sigma factor SIG1 maintains photosystem I activity via regulated expression of the psaA operon in rice chloroplasts. *Plant J.* **52**, 124–132 (2007).
118. Kanamaru, K. *et al.* An arabidopsis sigma factor (SIG2)-dependent expression of plastid-encoded tRNAs in chloroplasts. *Plant Cell Physiol.* **42**, 1034–1043 (2001).
119. Favory, J.-J. *et al.* Specific function of a plastid sigma factor for ndh F gene transcription. *Nucleic Acids Res.* **33**, 5991–5999 (2005).
120. Zghidi, W., Merendino, L., Cottet, A., Mache, R. & Lerbs-Mache, S. Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the psb N gene in plastids. *Nucleic Acids Res.* **35**, 455–464 (2007).
121. Zghidi-Abouzid, O., Merendino, L., Buhr, F., Malik Ghulam, M. & Lerbs-Mache, S. Characterization of plastid psb T sense and antisense RNAs. *Nucleic Acids Res.* **39**, 5379–5387 (2011).
122. Loschelder, H., Schweer, J., Link, B. & Link, G. Dual Temporal Role of Plastid Sigma Factor 6 in Arabidopsis Development. *Plant Physiol.* **142**, 642–650 (2006).
123. Onda, Y., Yagi, Y., Saito, Y., Takenaka, N. & Toyoshima, Y. Light induction of Arabidopsis SIG1 and SIG5 transcripts in mature leaves: Differential roles of cryptochrome 1 and cryptochrome 2 and dual function of SIG5 in the recognition of plastid promoters. *Plant J.* **55**, 968–978 (2008).
124. Shimizu, M. *et al.* Sigma factor phosphorylation in the photosynthetic control of photosystem stoichiometry. *Proc. Natl. Acad. Sci.* **107**, 10760–10764 (2010).
125. Nagashima, A. *et al.* The multiple-stress responsive plastid sigma factor, SIG5, directs activation of the psbD Blue Light-Responsive Promoter (BLRP) in Arabidopsis thaliana. *Plant Cell Physiol.* **45**, 357–368 (2004).
126. Stern, D. B., Goldschmidt-Clermont, M. & Hanson, M. R. Chloroplast RNA metabolism. *Annu. Rev. Plant Biol.* **61**, 125–155 (2010).
127. Hirose, T. & Sugiura, M. Both RNA editing and RNA cleavage are required for translation of tobacco chloroplast ndhD mRNA: a possible regulatory mechanism for the expression of a chloroplast operon consisting of functionally unrelated genes. *EMBO J.* **16**, 6804–6811 (1997).
128. Barkan, A., Walker, M., Nolasco, M. & Johnson, D. A nuclear mutation in maize blocks the processing and translation of several chloroplast mRNAs and provides evidence for the differential translation of alternative mRNA forms. *EMBO J.* **13**, 3170–3181 (1994).
129. Quesada-Vargas, T., Ruiz, O. N. & Daniell, H. Characterization of heterologous multigene operons in transgenic chloroplasts: transcription, processing, and translation. *Plant Physiol.* **138**, 1746–1762 (2005).
130. Pfalz, J., Bayraktar, O. A., Prikryl, J. & Barkan, A. Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. *EMBO J.* **28**, 2042–2052 (2009).
131. Felder, S. *et al.* The nucleus-encoded HCF107 gene of Arabidopsis provides a link between intercistronic RNA processing and the accumulation of translation-competent psbH transcripts in chloroplasts. *Plant Cell* **13**, 2127–2141 (2001).

132. Herrin, D. L. Chapter 27 - Chloroplast RNA Processing and Stability. in *The Chlamydomonas Sourcebook (Second Edition)* (eds. Harris, E. H., Stern, D. B. & Witman, G. B.) 937–965 (Academic Press, 2009). doi:10.1016/B978-0-12-370873-1.00035-6.
133. Goldschmidt-Clermont, M. Chapter 26 - Chloroplast RNA Splicing. in *The Chlamydomonas Sourcebook (Second Edition)* (eds. Harris, E. H., Stern, D. B. & Witman, G. B.) 915–935 (Academic Press, 2009). doi:10.1016/B978-0-12-370873-1.00034-4.
134. Fedorova, O. & Zingler, N. Group II introns: structure, folding and splicing mechanism. *Biol. Chem.* **388**, 665–678 (2007).
135. ODOM, O. W., SHENKENBERG, D. L., GARCIA, J. A. & HERRIN, D. L. A horizontally acquired group II intron in the chloroplast psbA gene of a psychrophilic Chlamydomonas: In vitro self-splicing and genetic evidence for maturase activity. *RNA* **10**, 1097–1107 (2004).
136. Schmitz-Linneweber, C. & Barkan, A. RNA splicing and RNA editing in chloroplasts. in *Cell and Molecular Biology of Plastids* (ed. Bock, R.) 213–248 (Springer Berlin Heidelberg, 2007). doi:10.1007/4735_2007_0233.
137. Schmitz-Linneweber, C. *et al.* A Pentatricopeptide Repeat Protein Facilitates the trans-Splicing of the Maize Chloroplast rps12 Pre-mRNA. *Plant Cell* **18**, 2650–2663 (2006).
138. Liere, K. & Link, G. RNA-binding activity of the matK protein encoded by the chloroplast trnK intron from mustard (Sinapis alba L.). *Nucleic Acids Res.* **23**, 917–921 (1995).
139. Vogel, J., Börner, T. & Hess, W. R. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res.* **27**, 3866–3874 (1999).
140. Jenkins, B. D., Kulhanek, D. J. & Barkan, A. Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classes with distinct requirements for splicing factors. *Plant Cell* **9**, 283–296 (1997).
141. Ostheimer, G. J. *et al.* Group II intron splicing factors derived by diversification of an ancient RNA-binding domain. *EMBO J.* **22**, 3919–3929 (2003).
142. Barkan, A. *et al.* The CRM domain: an RNA binding module derived from an ancient ribosome-associated protein. *RNA N. Y. N* **13**, 55–64 (2007).
143. 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).
144. 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).
145. de Longevialle, A. F. *et al.* The pentatricopeptide repeat gene OTP51 with two LAGLIDADG motifs is required for the cis-splicing of plastid ycf3 intron 2 in Arabidopsis thaliana. *Plant J. Cell Mol. Biol.* **56**, 157–168 (2008).
146. Beick, S., Schmitz-Linneweber, C., Williams-Carrier, R., Jensen, B. & Barkan, A. The Pentatricopeptide Repeat Protein PPR5 Stabilizes a Specific tRNA Precursor in Maize Chloroplasts. *Mol. Cell. Biol.* **28**, 5337–5347 (2008).
147. Williams-Carrier, R., Kroeger, T. & Barkan, A. Sequence-specific binding of a chloroplast pentatricopeptide repeat protein to its native group II intron ligand. *RNA* **14**, 1930–1941 (2008).
148. Kroeger, T. S., Watkins, K. P., Friso, G., van Wijk, K. J. & Barkan, A. A plant-specific RNA-binding domain revealed through analysis of chloroplast group II intron splicing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 4537–4542 (2009).
149. Francs-Small, C. C. des *et al.* A PORR domain protein required for rpl2 and ccmF(C) intron splicing and for the biogenesis of c-type cytochromes in Arabidopsis mitochondria. *Plant J. Cell Mol. Biol.* **69**, 996–1005 (2012).
150. Johanningmeier, U. & Heiss, S. Construction of a Chlamydomonas reinhardtii mutant with an intronless psbA gene. *Plant Mol. Biol.* **22**, 91–99 (1993).

151. Holloway, S. P. & Herrin, D. L. Processing of a Composite Large Subunit rRNA: Studies with Chlamydomonas Mutants Deficient in Maturation of the 23S-like rRNA. *Plant Cell* **10**, 1193–1206 (1998).
152. Miyamoto, T., Obokata, J. & Sugiura, M. Recognition of RNA editing sites is directed by unique proteins in chloroplasts: biochemical identification of cis-acting elements and trans-acting factors involved in RNA editing in tobacco and pea chloroplasts. *Mol. Cell. Biol.* **22**, 6726–6734 (2002).
153. Ji, D., Manavski, N., Meurer, J., Zhang, L. & Chi, W. Regulated chloroplast transcription termination. *Biochim. Biophys. Acta Bioenerg.* **1860**, 69–77 (2019).
154. Zhelyazkova, P. *et al.* Protein-mediated protection as the predominant mechanism for defining processed mRNA termini in land plant chloroplasts. *Nucleic Acids Res.* **40**, 3092–3105 (2012).
155. Stoppel, R. *et al.* RHON1 is a novel ribonucleic acid-binding protein that supports RNase E function in the Arabidopsis chloroplast. *Nucleic Acids Res.* **40**, 8593–8606 (2012).
156. Chi, W. *et al.* RHON1 Mediates a Rho-Like Activity for Transcription Termination in Plastids of Arabidopsis thaliana. *Plant Cell* **26**, 4918–4932 (2014).
157. Mitra, P., Ghosh, G., Hafeezunnisa, Md. & Sen, R. Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* **71**, 687–709 (2017).
158. Sugita, M. & Sugiura, M. Nucleotide sequence and transcription of the gene for the 32,000 dalton thylakoid membrane protein from Nicotiana tabacum. *MGG Mol. Gen. Genet.* **195**, 308–313 (1984).
159. Kirsch, W., Seyer, P. & Herrmann, R. G. Nucleotide sequence of the clustered genes for two P700 chlorophyll a apoproteins of the photosystem I reaction center and the ribosomal protein S14 of the spinach plastid chromosome. *Curr. Genet.* **10**, 843–855 (1986).
160. Blowers, A. D., Klein, U., Ellmore, G. S. & Bogorad, L. Functional in vivo analyses of the 3' flanking sequences of the Chlamydomonas chloroplast rbcL and psaB genes. *MGG Mol. Gen. Genet.* **238**, 339–349 (1993).
161. 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).
162. del Campo, E. M. Post-Transcriptional Control of Chloroplast Gene Expression. *Gene Regul. Syst. Biol.* **3**, 31–47 (2009).
163. Ichinose, M. & Sugita, M. RNA Editing and Its Molecular Mechanism in Plant Organelles. *Genes* **8**, (2016).
164. Sommer, B., Köhler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11–19 (1991).
165. Corneille, S., Lutz, K. & Maliga, P. Conservation of RNA editing between rice and maize plastids: are most editing events dispensable? *Mol. Gen. Genet. MGG* **264**, 419–424 (2000).
166. Maier, R. M., Neckermann, K., Igloi, G. L. & Kössel, H. Complete Sequence of the Maize Chloroplast Genome: Gene Content, Hotspots of Divergence and Fine Tuning of Genetic Information by Transcript Editing. *J. Mol. Biol.* **251**, 614–628 (1995).
167. Ruwe, H., Castandet, B., Schmitz-Linneweber, C. & Stern, D. B. Arabidopsis chloroplast quantitative editotype. *FEBS Lett.* **587**, 1429–1433 (2013).
168. Hein, A., Polsakiewicz, M. & Knoop, V. Frequent chloroplast RNA editing in early-branching flowering plants: pilot studies on angiosperm-wide coexistence of editing sites and their nuclear specificity factors. *BMC Evol. Biol.* **16**, 23 (2016).
169. Notsu, Y. *et al.* The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol. Genet. Genomics* **268**, 434–445 (2002).
170. Giegé, P. & Brennicke, A. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *Proc. Natl. Acad. Sci.* **96**, 15324–15329 (1999).

171. Bentolila, S., Oh, J., Hanson, M. R. & Bukowski, R. Comprehensive High-Resolution Analysis of the Role of an Arabidopsis Gene Family in RNA Editing. *PLOS Genet.* **9**, e1003584 (2013).
172. Grimes, B. T., Sisay, A. K., Carroll, H. D. & Cahoon, A. B. Deep sequencing of the tobacco mitochondrial transcriptome reveals expressed ORFs and numerous editing sites outside coding regions. *BMC Genomics* **15**, 31 (2014).
173. Oldenkott, B., Yamaguchi, K., Tsuji-Tsukinoki, S., Knie, N. & Knoop, V. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte Selaginella uncinata. *RNA* **20**, 1499–1506 (2014).
174. Wolf, P. G., Rowe, C. A. & Hasebe, M. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern Adiantum capillus-veneris. *Gene* **339**, 89–97 (2004).
175. Hecht, J., Grewe, F. & Knoop, V. Extreme RNA Editing in Coding Islands and Abundant Microsatellites in Repeat Sequences of Selaginella moellendorffii Mitochondria: The Root of Frequent Plant mtDNA Recombination in Early Tracheophytes. *Genome Biol. Evol.* **3**, 344–358 (2011).
176. Okuda, K. *et al.* Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. *Plant Cell* **21**, 146–156 (2009).
177. Nakamura, T. & Sugita, M. A conserved DYW domain of the pentatricopeptide repeat protein possesses a novel endoribonuclease activity. *FEBS Lett.* **582**, 4163–4168 (2008).
178. Maier, R. M. *et al.* RNA editing in plant mitochondria and chloroplasts. *Plant Mol. Biol.* **32**, 343–365 (1996).
179. Rovira, A. G. & Smith, A. G. PPR proteins - orchestrators of organelle RNA metabolism. *Physiol. Plant.* **166**, 451–459 (2019).
180. Lurin, C. *et al.* Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**, 2089–2103 (2004).
181. Barkan, A. & Small, I. Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* **65**, 415–442 (2014).
182. Wang, Z. *et al.* Cytoplasmic Male Sterility of Rice with Boro II Cytoplasm Is Caused by a Cytotoxic Peptide and Is Restored by Two Related PPR Motif Genes via Distinct Modes of mRNA Silencing. *Plant Cell* **18**, 676–687 (2006).
183. Yamazaki, H., Tasaka, M. & Shikanai, T. PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in Arabidopsis. *Plant J. Cell Mol. Biol.* **38**, 152–163 (2004).
184. Zsigmond, L. *et al.* Overexpression of the mitochondrial PPR40 gene improves salt tolerance in Arabidopsis. *Plant Sci. Int. J. Exp. Plant Biol.* **182**, 87–93 (2012).
185. Longevialle, A. F. de *et al.* The Pentatricopeptide Repeat Gene OTP43 Is Required for trans-Splicing of the Mitochondrial nad1 Intron 1 in Arabidopsis thaliana. *Plant Cell* **19**, 3256–3265 (2007).
186. Ruwe, H. & Schmitz-Linneweber, C. Short non-coding RNA fragments accumulating in chloroplasts: footprints of RNA binding proteins? *Nucleic Acids Res.* **40**, 3106–3116 (2012).
187. Zoschke, R., Watkins, K. P. & Barkan, A. A Rapid Ribosome Profiling Method Elucidates Chloroplast Ribosome Behavior in Vivo. *Plant Cell* **25**, 2265–2275 (2013).
188. Prikryl, J., Rojas, M., Schuster, G. & Barkan, A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proc. Natl. Acad. Sci.* **108**, 415–420 (2011).
189. Ohta, M., Sugita, M. & Sugiura, M. Three types of nuclear genes encoding chloroplast RNA-binding proteins (cp29, cp31 and cp33) are present in Arabidopsis thaliana: presence of cp31 in chloroplasts and its homologue in nuclei/cytoplasms. *Plant Mol. Biol.* **27**, 529–539 (1995).
190. Nakamura, T., Ohta, M., Sugiura, M. & Sugita, M. Chloroplast ribonucleoproteins are associated with both mRNAs and intron-containing precursor tRNAs. *FEBS Lett.* **460**, 437–441 (1999).
191. Nakamura, T., Ohta, M., Sugiura, M. & Sugita, M. Chloroplast ribonucleoproteins function as a stabilizing factor of ribosome-free mRNAs in the stroma. *J. Biol. Chem.* **276**, 147–152 (2001).

192. Schuster, G. & Gruissem, W. Chloroplast mRNA 3' end processing requires a nuclear-encoded RNA-binding protein. *EMBO J.* **10**, 1493–1502 (1991).
193. Kupsch, C. *et al.* Arabidopsis chloroplast RNA binding proteins CP31A and CP29A associate with large transcript pools and confer cold stress tolerance by influencing multiple chloroplast RNA processing steps. *Plant Cell* **24**, 4266–4280 (2012).
194. Teubner, M., Fuß, J., Kühn, K., Krause, K. & Schmitz-Linneweber, C. The RNA recognition motif protein CP33A is a global ligand of chloroplast mRNAs and is essential for plastid biogenesis and plant development. *Plant J. Cell Mol. Biol.* **89**, 472–485 (2017).
195. Sane, A. P., Stein, B. & Westhoff, P. The nuclear gene HCF107 encodes a membrane-associated R-TPR (RNA tetratricopeptide repeat)-containing protein involved in expression of the plastidial psbH gene in Arabidopsis. *Plant J. Cell Mol. Biol.* **42**, 720–730 (2005).
196. 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. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5651–5656 (2012).
197. Levey, T., Westhoff, P. & Meierhoff, K. Expression of a nuclear-encoded psbH gene complements the plastidic RNA processing defect in the PSII mutant hcf107 in Arabidopsis thaliana. *Plant J. Cell Mol. Biol.* **80**, 292–304 (2014).
198. Mathy, N. *et al.* 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA. *Cell* **129**, 681–692 (2007).
199. Halpert, M., Liveanu, V., Glaser, F. & Schuster, G. The Arabidopsis chloroplast RNase J displays both exo- and robust endonucleolytic activities. *Plant Mol. Biol.* **99**, 17–29 (2019).
200. Chen, H., Zou, W. & Zhao, J. Ribonuclease J is required for chloroplast and embryo development in Arabidopsis. *J. Exp. Bot.* **66**, 2079–2091 (2015).
201. Sharwood, R. E., Halpert, M., Luro, S., Schuster, G. & Stern, D. B. Chloroplast RNase J compensates for inefficient transcription termination by removal of antisense RNA. *RNA* **17**, 2165–2176 (2011).
202. Germain, A. *et al.* Mutational analysis of Arabidopsis chloroplast polynucleotide phosphorylase reveals roles for both RNase PH core domains in polyadenylation, RNA 3'-end maturation and intron degradation. *Plant J. Cell Mol. Biol.* **67**, 381–394 (2011).
203. Nurmohamed, S., Vaidialingam, B., Callaghan, A. J. & Luisi, B. F. Crystal structure of Escherichia coli polynucleotide phosphorylase core bound to RNase E, RNA and manganese: implications for catalytic mechanism and RNA degradosome assembly. *J. Mol. Biol.* **389**, 17–33 (2009).
204. Castandet, B., Hotto, A. M., Fei, Z. & Stern, D. B. Strand-specific RNA sequencing uncovers chloroplast ribonuclease functions. *FEBS Lett.* **587**, 3096–3101 (2013).
205. Germain, A., Kim, S. H., Gutierrez, R. & Stern, D. B. Ribonuclease II preserves chloroplast RNA homeostasis by increasing mRNA decay rates, and cooperates with polynucleotide phosphorylase in 3' end maturation. *Plant J. Cell Mol. Biol.* **72**, 960–971 (2012).
206. Kishine, M. *et al.* Ribosomal RNA processing and an RNase R family member in chloroplasts of Arabidopsis. *Plant Mol. Biol.* **55**, 595–606 (2004).
207. Carpousis, A. J. The RNA degradosome of Escherichia coli: an mRNA-degrading machine assembled on RNase E. *Annu. Rev. Microbiol.* **61**, 71–87 (2007).
208. Schein, A., Sheffy-Levin, S., Glaser, F. & Schuster, G. The RNase E/G-type endoribonuclease of higher plants is located in the chloroplast and cleaves RNA similarly to the E. coli enzyme. *RNA* **14**, 1057–1068 (2008).
209. Stoppel, R. & Meurer, J. The cutting crew - ribonucleases are key players in the control of plastid gene expression. *J. Exp. Bot.* **63**, 1663–1673 (2012).
210. Pellegrini, O., Nezzar, J., Marchfelder, A., Putzer, H. & Condon, C. Endonucleolytic processing of CCA-less tRNA precursors by RNase Z in Bacillus subtilis. *EMBO J.* **22**, 4534–4543 (2003).

211. Betat, H., Rammelt, C. & Mörl, M. tRNA nucleotidyltransferases: ancient catalysts with an unusual mechanism of polymerization. *Cell. Mol. Life Sci. CMLS* **67**, 1447–1463 (2010).
212. Canino, G. *et al.* Arabidopsis encodes four tRNase Z enzymes. *Plant Physiol.* **150**, 1494–1502 (2009).
213. Schuster, G. & Stern, D. RNA polyadenylation and decay in mitochondria and chloroplasts. *Prog. Mol. Biol. Transl. Sci.* **85**, 393–422 (2009).
214. Baker, M. E., Grundy, W. N. & Elkan, C. P. Spinach CSP41, an mRNA-binding protein and ribonuclease, is homologous to nucleotide-sugar epimerases and hydroxysteroid dehydrogenases. *Biochem. Biophys. Res. Commun.* **248**, 250–254 (1998).
215. Qi, Y. *et al.* Arabidopsis CSP41 proteins form multimeric complexes that bind and stabilize distinct plastid transcripts. *J. Exp. Bot.* **63**, 1251–1270 (2012).
216. Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: the enzymes in eukaryotes. *FEBS J.* **276**, 1494–1505 (2009).
217. Wu, H., Lima, W. F. & Crooke, S. T. Investigating the structure of human RNase H1 by site-directed mutagenesis. *J. Biol. Chem.* **276**, 23547–23553 (2001).
218. Goedken, E. R. & Marqusee, S. Co-crystal of Escherichia coli RNase HI with Mn2+ ions reveals two divalent metals bound in the active site. *J. Biol. Chem.* **276**, 7266–7271 (2001).
219. Tadokoro, T. & Kanaya, S. Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. *FEBS J.* **276**, 1482–1493 (2009).
220. Cerritelli, S. M. *et al.* Failure to produce mitochondrial DNA results in embryonic lethality in Rnaseh1 null mice. *Mol. Cell* **11**, 807–815 (2003).
221. Suzuki, Y. *et al.* An upstream open reading frame and the context of the two AUG codons affect the abundance of mitochondrial and nuclear RNase H1. *Mol. Cell. Biol.* **30**, 5123–5134 (2010).
222. Hamperl, S. & Cimprich, K. A. The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability. *DNA Repair* **19**, 84–94 (2014).
223. Stuckey, R., García-Rodríguez, N., Aguilera, A. & Wellinger, R. E. Role for RNA:DNA hybrids in origin-independent replication priming in a eukaryotic system. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5779–5784 (2015).
224. Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N. J. & Dean, C. R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. *Science* **340**, 619–621 (2013).
225. Nowotny, M., Gaidamakov, S. A., Crouch, R. J. & Yang, W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* **121**, 1005–1016 (2005).
226. Arora, R. *et al.* RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. *Nat. Commun.* **5**, 5220 (2014).
227. Zimmer, A. D. & Koshland, D. Differential roles of the RNases H in preventing chromosome instability. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12220–12225 (2016).
228. Yang, Z. *et al.* RNase H1 Cooperates with DNA Gyrases to Restrict R-Loops and Maintain Genome Integrity in Arabidopsis Chloroplasts. *Plant Cell* **29**, 2478–2497 (2017).
229. Kalhorzadeh, P. *et al.* Arabidopsis thaliana RNase H2 deficiency counteracts the needs for the WEE1 checkpoint kinase but triggers genome instability. *Plant Cell* **26**, 3680–3692 (2014).
230. Arudchandran, A. *et al.* The absence of ribonuclease H1 or H2 alters the sensitivity of Saccharomyces cerevisiae to hydroxyurea, caffeine and ethyl methanesulphonate: implications for roles of RNases H in DNA replication and repair. *Genes Cells Devoted Mol. Cell. Mech.* **5**, 789–802 (2000).
231. Eekhout, T., Kalhorzadeh, P. & De Veylder, L. Lack of RNase H2 activity rescues HU-sensitivity of WEE1 deficient plants. *Plant Signal. Behav.* **10**, (2015).
232. Keddie, J. S., Carroll, B., Jones, J. D. & Gruissem, W. The DCL gene of tomato is required for chloroplast development and palisade cell morphogenesis in leaves. *EMBO J.* **15**, 4208–4217 (1996).

233. Bellaoui, M., Keddie, J. S. & Gruissem, W. DCL is a plant-specific protein required for plastid ribosomal RNA processing and embryo development. *Plant Mol. Biol.* **53**, 531–543 (2003).
234. Bellaoui, M. & Gruissem, W. Altered expression of the Arabidopsis ortholog of DCL affects normal plant development. *Planta* **219**, 819–826 (2004).
235. Lahmy, S. *et al.* DOMINO1, a member of a small plant-specific gene family, encodes a protein essential for nuclear and nucleolar functions. *Plant J. Cell Mol. Biol.* **39**, 809–820 (2004).
236. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat. Rev. Mol. Cell Biol.* **12**, 483–492 (2011).
237. Huang, Y. *et al.* Ancient Origin and Recent Innovations of RNA Polymerase IV and V. *Mol. Biol. Evol.* **32**, 1788–1799 (2015).
238. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**, 635–648 (2008).
239. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat. Genet.* **41**, 630–634 (2009).
240. Rowley, M. J., Avrutsky, M. I., Sifuentes, C. J., Pereira, L. & Wierzbicki, A. T. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genet.* **7**, e1002120 (2011).
241. Zheng, Q. *et al.* RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J. Cell Mol. Biol.* **73**, 179–189 (2013).
242. Wendte, J. M. *et al.* Functional Dissection of the Pol V Largest Subunit CTD in RNA-Directed DNA Methylation. *Cell Rep.* **19**, 2796–2808 (2017).
243. Stein, H. & Hausen, P. Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. *Science* **166**, 393–395 (1969).
244. Lima, W. F. *et al.* Human RNase H1 discriminates between subtle variations in the structure of the heteroduplex substrate. *Mol. Pharmacol.* **71**, 83–91 (2007).
245. Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.* **16**, 583–597 (2015).
246. Boque-Sastre, R. *et al.* Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5785–5790 (2015).
247. Cloutier, S. C. *et al.* Regulated Formation of lncRNA-DNA Hybrids Enables Faster Transcriptional Induction and Environmental Adaptation. *Mol. Cell* **61**, 393–404 (2016).
248. Stirling, P. C. & Hieter, P. Canonical DNA Repair Pathways Influence R-Loop-Driven Genome Instability. *J. Mol. Biol.* **429**, 3132–3138 (2017).
249. Sollier, J. *et al.* Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol. Cell* **56**, 777–785 (2014).
250. Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442–451 (2003).
251. Nowotny, M. *et al.* Structure of Human RNase H1 Complexed with an RNA/DNA Hybrid: Insight into HIV Reverse Transcription. *Mol. Cell* **28**, 264–276 (2007).
252. Evans, S. P. & Bycroft, M. NMR structure of the N-terminal domain of Saccharomyces cerevisiae RNase HI reveals a fold with a strong resemblance to the N-terminal domain of ribosomal protein L9. *J. Mol. Biol.* **291**, 661–669 (1999).
253. Ohtani, N., Yanagawa, H., Tomita, M. & Itaya, M. Identification of the first archaeal Type 1 RNase H gene from Halobacterium sp. NRC-1: archaeal RNase HI can cleave an RNA-DNA junction. *Biochem. J.* **381**, 795–802 (2004).
254. Ohtani, N., Yanagawa, H., Tomita, M. & Itaya, M. Cleavage of double-stranded RNA by RNase HI from a thermoacidophilic archaeon, Sulfolobus tokodaii 7. *Nucleic Acids Res.* **32**, 5809–5819 (2004).

255. Xu, W. *et al.* The R-loop is a common chromatin feature of the Arabidopsis genome. *Nat. Plants* **3**, 704–714 (2017).
256. Suchard, M. A. & Redelings, B. D. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinforma. Oxf. Engl.* **22**, 2047–2048 (2006).
257. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
258. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J. Cell Mol. Biol.* **89**, 789–804 (2017).
259. Backes, S. & Herrmann, J. M. Protein Translocation into the Intermembrane Space and Matrix of Mitochondria: Mechanisms and Driving Forces. *Front. Mol. Biosci.* **4**, (2017).
260. Kim, M., Lee, U., Small, I., des Francs-Small, C. C. & Vierling, E. Mutations in an Arabidopsis Mitochondrial Transcription Termination Factor–Related Protein Enhance Thermotolerance in the Absence of the Major Molecular Chaperone HSP101[W]. *Plant Cell* **24**, 3349–3365 (2012).
261. O'Connell, K., Jinks-Robertson, S. & Petes, T. D. Elevated Genome-Wide Instability in Yeast Mutants Lacking RNase H Activity. *Genetics* **201**, 963–975 (2015).
262. Cerritelli, S. M., Chon, H. & Crouch, R. J. A New Twist for Topoisomerase. *Science* **332**, 1510–1511 (2011).
263. Yarbro, J. W. Mechanism of action of hydroxyurea. *Semin. Oncol.* **19**, 1–10 (1992).
264. Baranovskiy, A. G. *et al.* Structural basis for inhibition of DNA replication by aphidicolin. *Nucleic Acids Res.* **42**, 14013–14021 (2014).
265. Chon, H. *et al.* RNase H2 roles in genome integrity revealed by unlinking its activities. *Nucleic Acids Res.* **41**, 3130–3143 (2013).
266. Backes, S. *et al.* Tom70 enhances mitochondrial preprotein import efficiency by binding to internal targeting sequences. *J. Cell Biol.* **217**, 1369–1382 (2018).
267. Carrie, C., Giraud, E. & Whelan, J. Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.* **276**, 1187–1195 (2009).
268. Silva-Filho, M. C. One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. *Curr. Opin. Plant Biol.* **6**, 589–595 (2003).
269. Oldenburg, D. J. & Bendich, A. J. DNA maintenance in plastids and mitochondria of plants. *Front. Plant Sci.* **6**, 883 (2015).
270. Wood, K. C. & Granger, D. N. Sickle Cell Disease: Role of Reactive Oxygen and Nitrogen Metabolites. *Clin. Exp. Pharmacol. Physiol.* **34**, 926–932.
271. Burkitt, M. J. Nitric oxide generation from hydroxyurea: significance and implications for leukemogenesis in the management of myeloproliferative disorders. *Blood* **107**, 2219–2222 (2006).
272. Sakano, K., Oikawa, S., Hasegawa, K. & Kawanishi, S. Hydroxyurea induces site-specific DNA damage via formation of hydrogen peroxide and nitric oxide. *Jpn. J. Cancer Res. Gann* **92**, 1166–1174 (2001).
273. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
274. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J. Cell Mol. Biol.* **53**, 661–673 (2008).
275. Kersey, P. J. *et al.* Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).
276. Van Bel, M. *et al.* PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **46**, D1190–D1196 (2018).
277. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**, 17 (2014).
278. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

279. Suchard, M. A. & Redelings, B. D. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinforma. Oxf. Engl.* **22**, 2047–2048 (2006).
280. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. in *2010 Gateway Computing Environments Workshop (GCE)* 1–8 (2010). doi:10.1109/GCE.2010.5676129.
281. Lichtenthaler, H. K. [34] Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. in *Methods in Enzymology* vol. 148 350–382 (Academic Press, 1987).
282. Yoo, S.-D., Cho, Y.-H. & Sheen, J. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* **2**, 1565–1572 (2007).
283. Villani, T. S., Koroch, A. R. & Simon, J. E. An improved clearing and mounting solution to replace chloral hydrate in microscopic applications1. *Appl. Plant Sci.* **1**, (2013).
284. Curtis, M. D. & Grossniklaus, U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol.* **133**, 462–469 (2003).
285. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J. Cell Mol. Biol.* **16**, 735–743 (1998).
286. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
287. Barkan, A. Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. *Plant Physiol.* **155**, 1520–1532 (2011).
288. Yehudai-Resheff, S., Hirsh, M. & Schuster, G. Polynucleotide Phosphorylase Functions as Both an Exonuclease and a Poly(A) Polymerase in Spinach Chloroplasts. *Mol. Cell. Biol.* **21**, 5408–5416 (2001).
289. Bollenbach, T. J. *et al.* RNR1, a 3'–5' exoribonuclease belonging to the RNR superfamily, catalyzes 3' maturation of chloroplast ribosomal RNAs in Arabidopsis thaliana. *Nucleic Acids Res.* **33**, 2751–2763 (2005).
290. Liu, J. *et al.* The Conserved Endoribonuclease YbeY Is Required for Chloroplast Ribosomal RNA Processing in Arabidopsis1. *Plant Physiol.* **168**, 205–221 (2015).
291. Palm, D. *et al.* Plant-specific ribosome biogenesis factors in Arabidopsis thaliana with essential function in rRNA processing. *Nucleic Acids Res.* **47**, 1880–1895 (2019).
292. Wang, Y. & Ma, H. Step-wise and lineage-specific diversification of plant RNA polymerase genes and origin of the largest plant-specific subunits. *New Phytol.* **207**, 1198–1212 (2015).
293. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).
294. Böhmdorfer, G. *et al.* Long non-coding RNA produced by RNA polymerase V determines boundaries of heterochromatin. *eLife* **5**, (2016).
295. Böhmdorfer, G. *et al.* RNA-directed DNA methylation requires stepwise binding of silencing factors to long non-coding RNA. *Plant J. Cell Mol. Biol.* **79**, 181–191 (2014).
296. Zhu, Y., Rowley, M. J., Böhmdorfer, G. & Wierzbicki, A. T. A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. *Mol. Cell* **49**, 298–309 (2013).
297. Rowley, M. J., Rothi, M. H., Böhmdorfer, G., Kuciński, J. & Wierzbicki, A. T. Long-range control of gene expression via RNA-directed DNA methylation. *PLoS Genet.* **13**, e1006749 (2017).
298. Qi, Y. *et al.* Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443**, 1008–1012 (2006).
299. Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J. Cell Mol. Biol.* **88**, 1058–1070 (2016).
300. Gobert, A. *et al.* A single Arabidopsis organellar protein has RNase P activity. *Nat. Struct. Mol. Biol.* **17**, 740–744 (2010).

301. Luro, S., Germain, A., Sharwood, R. E. & Stern, D. B. RNase J participates in a pentatricopeptide repeat protein-mediated 5' end maturation of chloroplast mRNAs. *Nucleic Acids Res.* **41**, 9141–9151 (2013).
302. Hotto, A. M. *et al.* Arabidopsis chloroplast mini-ribonuclease III participates in rRNA maturation and intron recycling. *Plant Cell* **27**, 724–740 (2015).
303. Bollenbach, T. J., Sharwood, R. E., Gutierrez, R., Lerbs-Mache, S. & Stern, D. B. The RNA-binding proteins CSP41a and CSP41b may regulate transcription and translation of chloroplast-encoded RNAs in Arabidopsis. *Plant Mol. Biol.* **69**, 541–552 (2009).
304. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
305. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
306. Milligan, B. G. Purification of chloroplast DNA using hexadecyltrimethylammonium bromide. *Plant Mol. Biol. Report.* **7**, 144–149 (1989).
307. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685 (1970).
308. Hanaoka, M., Kato, M., Anma, M. & Tanaka, K. SIG1, a sigma factor for the chloroplast RNA polymerase, differently associates with multiple DNA regions in the chloroplast chromosomes in vivo. *Int. J. Mol. Sci.* **13**, 12182–12194 (2012).
309. Zhang, Y. *et al.* A nuclear-encoded protein, mTERF6, mediates transcription termination of rpoA polycistron for plastid-encoded RNA polymerase-dependent chloroplast gene expression and chloroplast development. *Sci. Rep.* **8**, 11929 (2018).
310. Chotewutmontri, P., Stiffler, N., Watkins, K. P. & Barkan, A. Ribosome Profiling in Maize. *Methods Mol. Biol. Clifton NJ* **1676**, 165–183 (2018).
311. Manavski, N., Schmid, L.-M. & Meurer, J. RNA-stabilization factors in chloroplasts of vascular plants. *Essays Biochem.* **62**, 51–64 (2018).
312. Tiller, N. & Bock, R. The Translational Apparatus of Plastids and Its Role in Plant Development. *Mol. Plant* **7**, 1105–1120 (2014).
313. König, F., Schubert, T. & Längst, G. The monoclonal S9.6 antibody exhibits highly variable binding affinities towards different R-loop sequences. *PloS One* **12**, e0178875 (2017).
314. Hartono, S. R. *et al.* The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J. Mol. Biol.* **430**, 272–284 (2018).
315. Wahba, L., Amon, J. D., Koshland, D. & Vuica-Ross, M. RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol. Cell* **44**, 978–988 (2011).
316. Xu, W. *et al.* The R-loop is a common chromatin feature of the Arabidopsis genome. *Nat. Plants* **3**, 704–714 (2017).