In *Arabidopsis thaliana* distinct alleles encoding mitochondrial RNA PROCESSING FACTOR 4 support the generation of additional 5' termini of *ccmB* transcripts

Katrin Stoll¹, Christian Jonietz¹, Sarah Schleicher¹, Catherine Colas des Francs-Small², Ian Small² and Stefan Binder^{1*}

¹Institut Molekulare Botanik, Universität Ulm, D-89069 Ulm, Germany

²Australian Research Council 40 Centre of Excellence in Plant Energy Biology, University of Western Australia, Crawley WA 6009, Australia

Key words: *Arabidopsis thaliana*, mitochondria, RNA processing, 5' end polymorphisms, pentatricopeptide repeat proteins

Corresponding author:

Stefan Binder

Institut Molekulare Botanik, Universität Ulm, Albert-Einstein-Allee 11, 89069 Ulm, Germany

Phone: ++49 731 502 2625

Fax: ++49 731 502 2626

E-mail: stefan.binder@uni-ulm.de

Abstract

In plant mitochondria, the 5' ends of many transcripts are generated post-transcriptionally. We show that the pentatricopeptide repeat (PPR) protein RNA PROCESSING FACTOR 4 (RPF4) supports the generation of extra 5' ends of *ccmB* transcripts in Landsberg *erecta* (*Ler*) and a number of other *Arabidopsis thaliana* ecotypes. RPF4 was identified in *Ler* applying a forward genetic approach supported by complementation studies of ecotype Columbia (*Col*), which generates the *Ler*-type extra *ccmB* 5' termini only after the introduction of the *RPF4* allele from *Ler*. Studies with chimeric RPF4 proteins composed from RPF4 proteins from *Ler* and *Col* identified differences in the N-terminal and central PPR motifs that explain ecotype-specific variations in *ccmB* processing. These results fit well with binding site predictions in *ccmB* transcripts based on the known determinants of nucleotide base recognition by PPR motifs.

Introduction

Plant mitochondria encode approximately 60 genes whose expression is essential for the function of these organelles, and thus for plant cell viability (Hammani and Giege 2014, Kubo and Newton 2008). Mitochondrial genes are transcribed into precursor RNAs that go through several processing steps to obtain their mature forms. These processes include the site-specific conversion of cytidines to uridines, called RNA editing, the removal of group II introns and the processing of 5' and 3' ends. Fundamental to all of these processes is the highly specific interaction of the proteinaceous processing machinery with the target RNAs (Binder et al. 2010, Hammani and Giege 2014). Pentatricopeptide repeat (PPR) proteins play a crucial role in the RNA metabolism in plant mitochondria (Barkan and Small 2014, Schmitz-Linneweber and Small 2008). These proteins are characterized by tandem arrays of repeated motifs, which according to slight variations of their lengths and sequences can be classified into ten distinct variants, which group into two major classes (Cheng et al. 2016). Sequence-specific binding of PPR proteins to their target sites depends on the specific interaction of certain amino acids within each repeat with the aligned nucleotide, in a relationship that can be described by a simple one repeat to one nucleotide code. Positions 5 and 35 (using the numbering of the canonical P motif) are the primary determinants of specificity since the amino acids at these positions can form specific hydrogen bonds with the aligned nucleotide

bases. (Barkan et al. 2012, Fujii et al. 2011, Shen et al. 2016, Takenaka et al. 2013, Yagi et al. 2013a). The amino acid side chain at position 2 intercalates between the aromatic rings of the bases (Shen et al. 2016, Yin et al. 2013), and may also influence specificity (Fujii et al. 2011, Yagi et al. 2013a). Thus, the sequence-specific interaction between a PPR protein and its cognate binding site seems to depend on relatively few amino acids.

As mentioned before, PPR proteins can be divided into two classes (Lurin et al. 2004). PLS-class proteins contain variations of P, L and S motifs and specific C-terminal E and/or DYW motifs. This class of proteins is predominantly involved in RNA editing, where they specify the cytidines to be modified (Chateigner-Boutin and Small 2010, Takenaka et al. 2014, Yagi et al. 2013b). There is increasing evidence that the DYW domain contains a cytidine-deaminase-like active site (Boussardon et al. 2014, Hayes et al. 2015, Hayes et al. 2013), extending the function of these proteins beyond passive sequence-specific interaction with the target RNA.

P-class PPR proteins are mainly composed of canonical P motifs and do not have extra domains apart from a small group of proteins containing C-terminal small MutS-related (SMR) domains (Liu et al. 2013). P-class PPR proteins including PPR-SMR are involved in splicing, translation and the post-transcriptional maturation of 5' and 3' ends both in chloroplasts and mitochondria (Barkan and Small 2014, Binder et al. 2013, Haili et al. 2016, Wu et al. 2016, Zoschke et al. 2016). The exact function of most of these proteins in these processes in mitochondria is largely unknown. A reasonable function seems to be evident for the MITOCHONDRIAL STABILITY FACTOR 1 supposed to block progression of 3' exonucleolytic degradation thereby defining the mature 3' terminus of the *nad4* transcript similar to is the general exonuclease-dependent 3' processing mechanism observed in chloroplasts (Haili et al. 2013). In the latter organelles, 5' processing of RNAs follows an analogous mechanism by PPR proteins building barriers to 5' exonucleolytic degradation, which determines the 5' termini (Pfalz et al. 2009). In contrast to plastids, 5' processing of transcripts in plant mitochondria seems to follow a different mode of action. Though many features are still unclear, there is indirect evidence that P-class PPR proteins involved in this process bind to the 5' leader sequences upstream of the mature 5' ends of the transcripts (Binder et al. 2013, Hauler et al. 2013, Stoll et al. 2015). These so-called RNA processing factors (RPFs) are suggested to recruit a 5' processing machinery to the cleavage site. The

identity of any additional proteins and the exact mode of cleavage are still unknown, although two potential mitochondrial nucleases have been described recently and proteinaceous RNase P (PRORP) has been found to cleave *orf291* RNA targeted by RFL2 (Fujii et al. 2016, Stoll and Binder 2016). In addition, the functional importance of the frequently observed 5' processing events is still mysterious.

To obtain a clearer view of post-transcriptional 5' processing of plant mitochondrial transcripts, we have analyzed a previously observed 5' end polymorphism of *ccmB* mRNAs in *Arabidopsis thaliana* (Forner et al. 2008). Using a linkage analysis and complementation assays, we identified RNA PROCESSING FACTOR 4 (RPF4-Ler), a P-class PPR protein encoded by At1g62910, as required for the formation of several 5' ends approximately 200 to 231 nucleotides upstream of the *ccmB* open reading frame in accession Landsberg *erecta* (Ler). These transcript termini are detected in several other accessions but are absent in Columbia (Col) and most other ecotypes, demonstrating that the *RPF4* alleles encoded in Col and the other lines do not support the generation of these additional termini. In order to identify the region essential for the function of RPF4-Ler in *ccmB* processing, we analyzed the potential of several chimeric proteins composed of different parts of RPF4-Ler and RPF4-Col to support the generation of *ccmB* transcripts with 5' ends around positions -200 and -231 in accession Col.

Results

Additional 5' termini of *ccmB* transcripts are formed in various *Arabidopsis* ecotypes.

In a previous study of mitochondrial transcripts in different *Arabidopsis thaliana* (*Arabidopsis*) accessions, we have identified polymorphic 5' termini of *ccmB* mRNAs (Forner et al. 2008). In the accession Columbia (Col) the major 5' ends were mapped to positions around -140 and -347/6 (Fig. 1a), indicated by circularized RNA RT-PCR (CR-RT-PCR) products of 260 and 470 bp (Fig. 1b, products 5 and 1, Supplemental Fig. S1). In Landsberg *erecta* (Ler) and other accessions, two additional 5' termini around -200 and -231 were identified (Fig. 1a) (Forner et al. 2008, Stoll et al. 2012). In a CR-RT-PCR analysis, the latter 5' ends were represented by products of approximately 310 and 340 bp (Fig. 1b, products 4 and 3). A primer extension analysis of *ccmB* mRNAs from Col and Ler confirmed the previously

determined ends (Fig. 1c, lanes b, signals 5 (≈ 400 nt) and 7 (≈ 190 nt), corresponding to 5' ends at -347/6 and -140). This primer extension experiment also resolved at least four extra ends in *Ler*, two of which correspond to the 5' termini around -200 and -231 (Fig. 1c, *Ler*, lane b, signals 6i (≈ 290 nt) and 6iii (≈ 260 nt)). In addition, several larger extension products indicated further upstream located 5' ends, which were not detected by the CR-RT-PCR analysis shown in Fig. 1b. Transcripts with these ends partially accumulated to different amounts in *Col* and *Ler*, respectively (Fig. 1c, lanes b, signals 2 to 4). A primer extension analysis of RNAs treated with Terminator exonuclease (TEx) revealed that the extra 5' ends at -200 and -231 in *Ler* did not originate from transcription initiation suggesting that they are generated by post-transcriptional processes. In *Col*, transcripts starting further upstream are enriched after TEx digestion suggesting that the 5' ends of these RNAs derived from transcript initiation (Fig. 1c, lanes a, signals 1 to 3). An investigation of the parental accessions and the corresponding reciprocal F_1 hybrids confirmed the biparental, i.e. nuclear inheritance of the *Ler ccmB* phenotype, which appeared to be dominant (Fig. 1b).

Mapping of the *RPF4* locus

To map the gene required for the generation of the additional 5' termini around positions -200 and -231 of the *ccmB* transcripts in *Ler*, we used an F_2 mapping population that was previously established for the identification of the *RPF1* gene (Hölzle et al. 2011). From 198 F_2 plants analyzed, 50 showed the *Col*-type *ccmB* mRNA, confirming the *Ler*-type *ccmB* transcript (found in 148 individuals) to be dominant (Supplemental Table S1). The 1:3 ratio between these transcript phenotypes strongly suggested a single nuclear gene to be responsible for the polymorphic *ccmB* RNAs. We called this gene *RNA PROCESSING FACTOR 4* (*RPF4*).

A linkage analysis with markers covering all chromosomes localized the *RPF4* locus on the lower arm of chromosome 1 (Supplemental Fig. S2). The examination of further markers delimited the region containing the *RPF4* gene to a 0.55 Mbp interval between markers CER 429086 and CER 428577 (Supplemental Fig. S3). For additional fine mapping, we used several recombinant inbred lines (RILs) that have been characterized previously (Lister and Dean 1993). The genotypes of these RILs suggested them to contain recombinations within the identified genomic region. Phenotyping of the *ccmB* transcripts and further high-

resolution mapping using three independent lines (N1929, N1962 and N1997) localized the *RPF4* gene to a region spanning 80 kbp between markers CER 429086 and CER 472531 (Supplemental Fig. S4). This region contains 28 annotated genes (TAIR10) between At1g62780 and At1g62981, including three genes and a pseudogene encoding pentatricopeptide repeat (PPR) proteins. Since PPR proteins were shown to be involved in 5' processing of other mitochondrial transcripts (Arnal et al. 2014, Binder et al. 2013, Hauler et al. 2013, Hölzle et al. 2011, Jonietz et al. 2011, Jonietz et al. 2010, Stoll et al. 2014, Stoll et al. 2015), At1g62910, At1g62914 and At1g62930 were considered to be the best candidates to encode the *RPF4* gene. At1g62860 was excluded from further analysis since it was suggested to encode a pseudogene (Lurin et al. 2004).

In *Arabidopsis* accession Landsberg *erecta* *RPF4* is encoded by At1g62910

The *Ler* alleles of the candidate genes were cloned and independently transformed into Col wild type. Due to difficulties in attempting to amplify DNA fragments from *Ler* with some primers designed from the Col sequence, we completely sequenced the alleles of At1g62910 and At1g62914 from *Ler* (LN997834 and LN997840). After transformation into Col wild-type and selection of transformed plants, *ccmB* transcripts were investigated by CR-RT-PCR analyses. The introduction of the *Ler* alleles of At1g62914 and At1g62930 into Col wild-type plants did not change the *ccmB* transcript pattern indicating that these genes do not encode *RPF4*. In contrast, Col plants transformed with the *Ler* allele of At1g62910 showed a CR-RT-PCR product pattern identical to *Ler*, where two products of 310 and 340 bp indicate the presence of the additional *ccmB* transcripts with 5' ends between -200 and -231 (Fig. 2a). A northern hybridization confirmed this result (Fig. 2b). In Col and Sorbo, another ecotype with the Col-type *ccmB* transcripts, two mRNA species of about 1050 and 850 nucleotides were detected (Fig. 2b, transcripts 1 and 3). These mRNAs were also found in *Ler* and En-1, however, in these accessions and in a Col plant containing the *Ler* allele of At1g62910, additional transcripts of approximately 950 nucleotides were seen (Fig. 2b, transcript 2).

Taken together, both the results of the CR-RT-PCR and of the northern analysis demonstrated that the generation of *ccmB* transcripts with 5' termini between -200 and -231 depends on the function of the *Ler* allele of At1g62910. This data unambiguously showed that this gene in *Ler* encodes *RPF4*. To discriminate it from the previously identified

At1g62910 allele from Col, which encodes RESTORER OF FERTILITY-LIKE PROTEIN 9 (RFL9) involved in the generation of a 5' truncated *rps3-rpl16* dicistronic mRNA and in processing of *orf240a* transcripts (Arnal et al. 2014), we designate this gene *RPF4-Ler*.

Multiple differences between At1g62910 alleles

To identify the region responsible for the function of an At1g62910 allele as *RPF4-Ler*, we compared the sequence of this gene from various accessions using the Genome Browser Web-Tool <http://signal.salk.edu/atg1001>. However, the numerous gaps or differences in the amino acid sequences in the corresponding sequences did not correlate with the distinct functions of the alleles. Therefore we sequenced the At1g62910 alleles in various ecotypes with the *Ler*- (En-1, Got-7, and Br-0) or Col-type (Ws, Tsu-1, Shakdara, Sorbo) *ccmB* mRNA phenotype (Supplemental Fig. S1). A comparison of the amino acid sequences deduced from the distinct At1g62910 alleles did not reveal positions where amino acid identities are consistently different between the *Ler*- and Col-type *RPF4* alleles (Supplemental Fig. S5). Only the potential mitochondrial targeting sequence and the N-terminal half of the first PPR motif showed substantial variability whereas the sequences in motifs 8 to 16 are highly conserved. This includes also almost all of the amino acid residues at positions 5 and 35 suggested to be crucial for the interaction with specific nucleotide identities (Barkan et al. 2012, Cheng et al. 2016).

We also investigated the 5' ends of the *rps3-rpl16* dicistronic RNAs in the above mentioned accessions. The *rps3* reading frame internal 5' termini at positions +17/+15 were detected in Col and Sha but not in *Ler* and the other ecotypes. In Sha only a very weak product was observed (Supplemental Fig. S6).

To get further information about amino acid positions of the *RPF4-Ler* protein relevant for the function in *ccmB* processing, we identified a potential binding site of the *RPF4-Ler* protein on the *ccmB* precursor transcript. Following the rules of the amino acid/nucleotide code for the interaction of PPR proteins with the RNA targets, a putative binding site was predicted between nucleotide position -267 and -252 relative to the translation start codon (NATG, N = -1, Fig. 1a). The potential binding upstream of the processing sites at -200 and -231 is consistent with predictions of recognition sites for other RPFs, which were all found 5'

to the cleavage sites (Binder et al. 2013) and with indirect evidence from the analysis of the RNA recognition sites of RPF3 and RPF6 (Stoll et al. 2015). The putative interaction of RPF4-Ler did not involve the first PPR motif but indicated a perfect match between the relevant amino acids of the other repeats and the corresponding nucleotide identities (Fig. 3a). When the 5 and 35 positions of the At1g62910 protein from Col (At1g62910-Col) were aligned with the predicted RNA binding site, four positions did not match with the suggested rules. Three of these mismatches were seen between the N-terminal half of the protein and the binding site, whereas the C-terminal repeats 12 to 16 showed a nearly perfect match, as did the same part of the protein from Ler.

We also identified the putative binding site on the *rps3-rpl16* mRNA located between position -30 to -15 upstream of the translation start codon. The prediction indicated binding of RPF4-Col to fit much better to the one repeat to one nucleotide code than RPF4-Ler (Supplemental Fig. S6).

N-terminal repeats are important for the RPF4-dependent generation of *ccmB* 5' ends around position -200 and -231.

To experimentally identify regions crucial for the specific function of RPF4-Ler in *ccmB* mRNA 5' end formation, a number of chimeric *RPF4* genes was generated by combining different parts of the alleles from Col (*RPF4-Col*, non-functional in terms of *ccmB* processing) and Ler (*RPF4-Ler*, functional in terms of *ccmB* processing) (Fig. 3a and b). After their stable introduction into Col plants, RT-PCR analyses confirmed transcription of the transgenes (Supplemental Fig. S7). Subsequently, the ability of these chimeric genes to support *ccmB* processing identical to Ler wild type was investigated by CR-RT-PCR analysis in the corresponding transformants (Fig. 3c). Consistent with the high similarity between Ler- and Col-type amino acid sequences in the C-terminal part of At1g62910, the exchange of the corresponding region downstream of repeat 11 from Ler by the one from Col still allowed the chimeric protein to support Ler-type *ccmB* processing (Fig. 3b and c, RPF4-c1). However, exchanges of sequences upstream of repeat 9 led to a failure to support Ler-type *ccmB* processing (Fig. 3b and c, RPF4-c4 and RPF4-c6). These results indicated that this region is crucial for the function of RPF4-Ler in *ccmB* processing.

We also exchanged sequences encoding just the N-terminal part of the At1g62910 protein. When regions upstream of position 5 of repeats 4 and 3, respectively, were changed from *Ler* to *Col* the corresponding chimeric protein (containing *Ler* sequence downstream of repeats 3 and 4, respectively) did not support *Ler*-type *ccmB* mRNA processing in *Col* (Fig. 3b and c, RPF4-c2 and RPF4-c7). Thus, the N-terminal part is crucial for *Ler*-type *ccmB* processing, consistent with the much better match to the RNA sequence of the protein part from *Ler* in comparison to that from *Col*. Finally, we exchanged regions in the central part of the protein by replacing the *Ler* sequence between repeat 4 and repeat 9 by the corresponding sequence derived from *Col*. As for all the composite RPF4 proteins, this chimeric protein was also incapable of supporting the *Ler*-type processing of *ccmB* transcripts (Fig. 3b and c, RPF4-c8).

Collectively these data revealed that parts upstream of PPR motif 12 contain sequence information required for the generation of the extra 5' ends of *ccmB* mRNA typically observed in *Ler*. Further CR-RT-PCR analyses excluding the -140 terminus, the -200 and -231 5' ends and the -346 5' extremities did not reveal any influence of the chimeric RPF4 genes on other *ccmB* 5' termini suggesting that RPF4-*Ler* is specifically required for the generation of the *Ler*-type -200 to -231 5' ends (Supplemental Fig. S8).

Binding of RFP4 to the predicted recognition site located upstream of the *Ler*-specific *ccmB* 5' ends

We hypothesized that the differential ability of the *Col*- and *Ler*-type Atg162910 proteins to support *Ler*-type *ccmB* transcript processing is linked to the different binding properties of these proteins to *ccmB* precursor RNAs. To test this assumption experimentally, we performed RNA electrophoretic mobility shift assays (REMSAs) with an RNA oligonucleotide containing the predicted *ccmB* binding site and recombinant RPF4-*Col* and RPF4-*Ler* proteins partially purified after overexpression in *E. coli*. Both proteins were found to bind to this RNA fragment, however, competition experiments performed with poly(C) oligonucleotides indicated a slightly stronger binding of RPF4-*Ler* in comparison to RPF4-*Col*. Perhaps the C-terminal parts, almost identical between both proteins, explain binding of both proteins *in vitro*. This is supported by the binding predictions and by the *in vivo* domain-swapping

experiments. Thus the different binding efficiencies might contribute to the selective function of both proteins in *ccmB* transcript processing.

Discussion

RNA PROCESSING FACTOR 4-Ler (RPF4-Ler) is encoded by At1g62910

In *Arabidopsis thaliana* several RESTORER of FERTILITY (RF)-like PPR proteins have been found to be involved in 5' processing of mitochondrial transcripts. The majority of the RF-like PPR proteins, including At1g62910, are encoded in a gene cluster on the lower arm of chromosome 1. Two compelling lines of evidence unambiguously show At1g62910 to encode RPF4-Ler. First, the linkage analysis using a Col x Ler F₂ mapping population of 50 plants and several lines from a previously established and characterized recombinant inbred line connected the formation of the Ler-type *ccmB* 5' processing to a genomic region encoding three genes for PPR proteins including At1g62910 (Supplemental Fig. S2 to S4) (Lister and Dean 1993). Second, when transformed into Col plants At1g62910 supports the generation of the Ler-type *ccmB* 5' processing pattern while the other PPR genes did not (Fig. 2). Thus in Ler, the formation of the *ccmB* 5' ends located between positions -200 and -231 is linked to a distinct allele of At1g62910 encoding one of the fast-evolving P-class PPR proteins with high similarity to RF proteins (Desloire et al. 2003, Fujii et al. 2011). Interestingly, the protein encoded by At1g62910, called RFL9, has previously been identified to be involved in cleavage of *rps3-rpl16* and *orf240a* transcripts in a small subset of ecotypes including Col. In both transcripts, cleavage occurs within the reading frames discriminating these events from classical 5' processing in regions upstream of the reading frames (Arnal et al. 2014).

Proteins encoded by At1g62910 have distinct ecotype-specific functions

The complementation assay identified the RPF4-Ler allele to be competent in the formation of the Ler-specific *ccmB* 5' termini, whereas the allele of the same gene from Col (*RPF4*-Col) does not support this processing event provoking the question how the functional diversification is established. The comparison of several functional and non-functional alleles of *RPF4* from other accessions did not reveal a clear link between a distinct amino acid

identities and the function of an allele as *RPF4* (Supplemental Fig. S5). Thus it is possible that in the non-functional *RPF4* alleles, amino acid identities at different positions are incompatible with the function in *ccmB* processing. This is different to the observation made during the analysis of *RFL9*. *RFL9* alleles show activity towards *rps3-rpl16* and *orf240a* transcripts only in Col and closely related accessions. In Pr-0, Mz-0b, Or-0 and El-0, *rps3-rpl16* and *orf240a* transcripts are unaffected although they encode *RFL9* alleles similar to Col and other ecotypes, in which *rps3-rpl16* transcripts were processed. When the processing phenotypes were correlated with the *RFL9* sequences, the processing deficiency of the *RFL9* alleles in the four accessions could be linked to the exchange of an aspartate residue by serine at position 1 in the fourth repeat (Arnal et al. 2014), now position 35 in repeat 3 according to the new definition (Cheng et al. 2016). Interestingly, the CR-RT-PCR analysis of Sha revealed a weak product indicating *rps3-rpl16* mRNA processing although this ecotype contains a serine at this position. Unlike the group of accessions with intact *RFL9* alleles, accessions with *RPF4* alleles functional in *ccmB* processing group into different ecotype clusters which exhibit larger genetic diversity (Simon et al. 2012), which explains the difficulties to correlate certain amino acid identities with the function *RPF4-Ler* in *ccmB* processing.

Collectively, our data and the previous study confirm the fast evolving nature of the *RPF4/RFL9* gene creating ecotype-specific gene functions. In addition, our data support the previously considered possibility that the functional diversification of *RPF4/RFL9* alleles is fixed by genetic drift since both functions of *RPF4/RFL9* alleles seem to be of rather limited biological significance. The sequence variation leads to new binding specificities towards different RNA species. This observation is supported by our REMSAs indicating a slightly better binding of *RPF4-Ler* to the *ccmB* binding sites than *RPF4-Col* (Fig. 4). In turn, the *RPF4* protein encoded by the allele in Col exhibits a better fit to the predicted binding site on the *rps3-rpl16* mRNA.

The C-terminal PPR motifs do not contribute to functional differences

To identify the repeats contributing to the functional specification of the *RPF4-Ler* allele, we tested the ability of several chimeric genes assembled from different parts of the functional *RPF4-Ler* and the non-functional *RPF4-Col* for their ability to support *Ler*-type-specific

generation of *ccmB* 5' ends in Col. The transcripts of the chimeric genes accumulated to comparable levels, suggesting that expression was functionally sufficient as shown by the test with construct RPF4-c1 (Supplemental Fig. S7 and Fig. 3). Complementation assays with the chimeric gene containing repeats 1 to 11 from *Ler* and 12 to 16 from Col showed 5' processing identical to the *Ler* wild-type allele demonstrating that repeats 12 to 16 do not contribute to the specific function of RPF4-*Ler* in *ccmB* processing (Fig. 3a and b, RPF4-c1). This is consistent with the predicted nearly perfect match of this part of both RPF4-*Ler* and its counterpart from Col with the *ccmB* binding site (Fig. 3a). Replacement of any other part of the RPF4-*Ler* protein with the corresponding region from RPF4-Col abolished the activity of the chimeric proteins on the *ccmB* transcript (Fig. 3b and c, RPF4-c2, RPF4-c4, RPF4-c6, RPF4-c7 and RPF4-c8). These results are also in agreement with binding predictions as all of the non-functional chimeras exhibit mismatches to the RNA sequence. By successively exchanging the C-terminal, central and N-terminal segments of the protein we showed that PPR motifs throughout the protein might interact specifically with the RNA, in contrast to previous suggestions that the N- and C-terminal motifs and not the central motifs are essential for the specific binding of maize PPR10 and other P-class PPR proteins (Barkan et al. 2012, Yin et al. 2013).

A single factor provokes the generation of four additional 5' ends

RPFs have been linked to the generation of 5' ends of one, two or three different transcripts. For a given transcript, an RPF was required for the formation of only a single 5' terminus, sometimes scattering over a few nucleotides, except for RPF2, which seems to affect two different 5' termini of the *cox3* transcripts (Binder et al. 2013, Jonietz et al. 2010). Now we found that RPF4-*Ler* is involved in the generation of four additional 5' ends around -200 and -231 (Fig. 1c). This observation raises the question of how a single factor provokes the generation of four different 5' termini. Considering the distinct signals seen in the primer extension analysis, it seems to be rather unlikely that the different ends are generated exonucleolytically from the largest additional *Ler*-type specific *ccmB* transcript. Instead, distinct endonucleolytic cuts seem to be responsible for the generation of the four ends. Also the RPF4-*Ler* recognition site is unique in the *Ler* and Col mitochondrial genome sequences rendering binding to alternative sites on the *ccmB* transcript rather unlikely. We

have previously hypothesized that RNA secondary structure might be important for endonucleolytic 5' processing and therefore we favor the possibility that binding of RPF4-Ler leads to the formation of different higher order structures which undergo endonucleolytic cleavage at different sites. Admittedly, this is a speculative explanation, however, it is in support with the general idea that RNA secondary structures are important for the selection of the cleavage sites for the generation of 5' transcript termini (Binder and Kühn 2015, Binder et al. 2013). This notion is substantiated by the observation that RFL2, another RF-like PPR protein, directs proteinaceous RNase P to cleave *orf291* mRNA (Fujii et al. 2016). Cleavage by this endonuclease requires RNA secondary structures resembling acceptor stems of precursor tRNAs, which could be formed by intramolecular base pairing or in a bimolecular manner (Derksen et al. 2015). The formation of such structure could be influenced by PPR proteins.

Materials and methods

Nucleic acid analysis

Arabidopsis thaliana plants and cell suspension cultures of accessions Col and Ler were cultivated as described previously (Forner et al. 2007, Köhler et al. 2010). Extraction of total cellular RNA and mitochondrial RNA, northern blot hybridization as well as CR-RT-PCR analysis were done as outlined before (Forner et al. 2007). The DNA probe used for the northern hybridization was amplified with the primer pair Atccb2NS.H/AtccbNS.R in a standard PCR. Primer extension was done with oligonucleotide Atccb2-1 and 5 µg mitochondrial RNA following standard protocols (Sambrook and Russell 2001). Rosette leaves of 20- to 30-day old plants were used to prepare total genomic DNA with the DNeasy plant mini kit (Qiagen) or applying a previously established protocol (Edwards et al. 1991).

Complementation studies and generation of chimeric *RPF4* genes

For transformation of the candidate genes into Col, At1g62910, At1g62914 and At1g62930 were amplified with oligonucleotide pairs At1g62910Ler-Kompl.1/At1g62910.Kompl.R2, At1g62915Kompl.H/At1g62915Kompl.R and At1g62930Kompl.H/At1g62930Kompl.R on total

DNA from *Ler*. After digestion with *PacI* and *Ascl*, the PCR products were cloned into the respective sites in pMDC123 (Curtis and Grossniklaus 2003). Transformation was done as described (Clough and Bent 1998) and *ccmB* transcripts were analyzed by CR-RT-PCR analysis in selected transformants as outlined before (Forner et al. 2007).

Chimeric *RPF4* genes were obtained by an overlap PCR strategy. The 5' and 3' parts were separately amplified by the following oligonucleotide pairs on total DNA from *Col* and *Ler*:

RPF4-c1 (Ler, At1g62910.Ler.kompl.1/At1g62910Ler.Teil1.1; Col, At1g62910.Teil2/At1g62910-kompl.R2), RPF4-c2 (Col, At1g62910-kompl.H/At1g62910Col.Teil1; Ler, At1g62910Ler.Teil2/At1g62910-kompl.R2), RPF4-c4 (Ler, At1g62910.Ler.kompl.1/At1g62910Konstrukt4.Ler; Col, At1g62910Konstrukt4.Col/At1g62910-kompl.R2), RPF4-c5 (Ler, At1g62910.Ler.kompl.1/At1g62910Col.Teil1; Col, At1g62910Konstrukt5.Col/At1g62910-kompl.R2), RPF4-c6 (Ler, At1g62910.Ler.kompl.1/At1g62910Col.Teil1; Col, At1g62910Ler.Teil2/At1g62910-kompl.R2), RPF4-c7 (Col, at1g62910-kompl.H/At1g62910Konstrukt7.Col; Ler, At1g62910Konstrukt7.Ler/At1g62910-kompl.R2). RPF4-c8 was generated using DNA from construct RPF4-c6 and total DNA from *Ler* as template (RPF4-c6, At1g62910Konstrukt4.Ler/At1g62910Ler.kompl.1; Ler, At1g62910Konstrukt4.Col/At1g62910-kompl.R2). PCR products were inspected by agarose gel electrophoresis and recovered from the gel using the GenElute™ Gel Extraction Kit (Sigma-Aldrich). Subsequently the 5' and 3' products were joined by overlap PCR with primers At1g62910Ler.kompl.1/At1g62910-kompl.R2 in the case of RPF4-c1, RPF4-c4, RPF4-c6 and RPF4-c8, respectively, or At1g62910kompl.H/At1g62910-kompl.R2 in the case of RPF4-c2 and RPF4-c7. After digestion with *PacI* and *Ascl* the PCR products were cloned into the respective sites in pMDC123. Finally constructs were checked by DNA sequence analysis.

Sequence analysis

For the sequence analysis of At1g62910 and At1g62914 the genes were amplified using Phusion High Fidelity DNA polymerase (NEB and Thermo Scientific). At1g62910 was amplified using primers cornichon.1/At1g62910-kompl.R2 on total DNA extracted from *Ler*, *En-1*, *Shakdara*, *Sorbo*, *Tsu-1*, *Col-0* and *Ws* or primers At1g62910Ler-kompl.1/At1g62910-kompl.R2 on total DNA extracted from *Got-7* and *Br-0*. At1g62914 was amplified with

primers cornichon.1/At1g62915-kompl.R. Sequencing was done by primer walking with oligonucleotides At1g62910Ler.1 (*Ler*, En-1, Shakdara, Sorbo, Tsu-1, Ws, Got-7, Br-0), At1g62910Ler.2 (*Ler*, En-1, Shakdara, Sorbo, Tsu-1, Col-0, Ws, Got-7, Br-0), At1g62910Ler.3 (*Ler*, En-1, Shakdara, Sorbo, Tsu-1, Col-0, Ws, Got-7), At1g62910Ler.4 (*Ler*, En-1, Shakdara, Sorbo, Tsu-1, Col-0, Ws), At1g62910Ler.6 (*Ler*, En-1, Shakdara, Sorbo, Tsu-1, Col-0, Ws, Got-7, Br-0), At1g62910Col.Teil1 (Br-0, Got-7), At1g62910Br-0.1 (Br-0), At1g62910Ler.12 (Br-0), At1g62910.Br-0.R1 (Br-0) and At1g62910Ler-kompl.1 (Got-7) (At1g62910) and oligonucleotides At1g62915Ler.1, At1g62915Ler.2 and At1g62915Ler.3 (At1g62914). Sequencing was done commercially (GATC, www.gatc-biotech.com/de/index.html, Seqlab, www.seqlab.de).

Linkage analysis

Rough mapping was done with insertion/deletion and SNP markers with 50 plants showing the recessive *ccmB* Col phenotype. Markers and primers used are listed in Supplemental Tables S2 and S3. PCRs and digestion with restriction enzymes were done according to standard procedures (Sambrook and Russell 2001). Previously established Col/*Ler* recombinant inbred lines N1929, N1949, N1962, N1967, N1971, N1984, N1985 and N1997 were obtained from the European Arabidopsis Stock Centre (Lister and Dean 1993).

RNA electrophoretic mobility shift assays (REMSA)

For protein expression and purification in bacteria, the coding sequences of RPF4-*Ler* and RPF4-Col (devoid of N-terminal mitochondrial targeting peptides) were amplified from genomic DNA from *Arabidopsis* ecotypes *Ler* and Col, respectively. The expressed proteins contained an N-terminal 6xHis tag, which was used for purification. A Coomassie stained gel showing different fractions obtained during protein purification is shown in Supplemental Fig. S9. The amounts of RPF4-*Ler* and RPF4-Col T in the lysates used for the binding studies are xx and xx %. Purities of the RPF4-*Ler* and RPF4-Col fractions seem to be comparable, Protein purification and expression as well as REMSAs were performed as described previously (Kindgren et al. 2015), with a few modifications. Briefly, 10 µL binding buffer consisting of 1× THE (34 mM Tris, 66 mM HEPES, and 0.1 mM EDTA, pH 8.3) with 200 mM NaCl, 5 mM DTT, 5

mg/mL heparin, 0.1 mg/mL BSA, and 8 units RNaseOUT (Invitrogen) were mixed with 5 μ L protein dilution and incubated at room temperature for 10 min. 5'-Fluorescein- or Cy5-labeled probes (Sigma-Aldrich) were heated at 94 °C for 2 min and cooled on ice for 5 min. Denatured probes (final concentration 500 pM each) were then added to the binding reaction for a total reaction volume of 25 μ L. The reactions were incubated at 25°C for 15 min and 20 μ L were loaded onto a 5 % native acrylamide gel (in 1 $\times$ THE) and run at 4°C. The gels were imaged with a Typhoon Trio imager (GE Healthcare). Fluorescein- and Cy5-labeled oligonucleotides were excited at 488 nm and 633 nm respectively, and detected through a 520-nm and 670-nm band-pass filter respectively. The fraction of oligonucleotide bound was determined with ImageQuant (GE Healthcare).

Accession numbers

Sequence data from this article can be found in the EMBL/Genbank data libraries under the accession numbers: LN997834 (RPF4/At1g62910 from Ler), LN997835 (At1g62910 from En-1), LN997836 (At1g62910 from Shakdara), LN997837 (At1g62910 from Sorbo), LN997838 (At1g62910 from Ws), LN997839 (At1g62910 from Tsu-1), LT622251 (At1g62910 from Br-0), LT622250 (At1g62910 from Got-7) and LN997840 (At1g62914 from Ler).

Acknowledgement

We are very grateful to Conny Guha and Uli Tengler for excellent technical assistance. This work was supported by DFG grant Bi 590/10-2.

Author contributions

K.S., C.J., S.S. C.CdF-S and I.S. performed experiments and/or analyzed the data; S.B. conceived the project, designed and supervised experiments and wrote the article with contributions of all the authors.

Supplementary material

Supplemental Fig. S1. CR-RT-PCR analysis of *ccmB* RNA from various *Arabidopsis* ecotypes

Supplemental Fig. S2. Rough mapping of the *RPF4* locus in the *Arabidopsis* nuclear genome

Supplemental Fig. S3. Fine mapping of the *RPF4* locus on chromosome 1

Supplemental Fig. S4. Fine mapping of the *RPF4* locus using Col x *Ler* recombinant inbred lines (RIL)

Supplemental Fig. S5. Clustal Omega alignment of amino acid sequences deduced from different *Ler*- (*Ler*, En-1, Got-7 and Br-0) and Col-type (Shakdara (Sha), Sorbo, Tsu-1, Col, Ws) alleles of At1g62910.

Supplemental Fig. S6. Function and putative binding of RPF4 to *rps3-rpl16* RNA

Supplemental Fig. S7. Transcripts derived from of chimeric constructs

Supplemental Fig. S8. Investigation of *ccmB* 5' ends in Col plants containing chimeric RPF4 constructs

Supplemental Fig. S9. Partial purification of RPF4-*Ler* and RPF4-Col

Supplemental Table S1. Phenotypes of the *ccmB* transcripts in an F₂ mapping population.

Supplemental Table S2. Marker used for the linkage analysis

Supplemental Table S3. Oligonucleotides

References

- Arnal N, Quadrado M, Simon M, Mireau H (2014) A restorer-of-fertility like pentatricopeptide repeat gene directs ribonucleolytic processing within the coding sequence of *rps3-rpl16* and *orf240a* mitochondrial transcripts in *Arabidopsis thaliana*. *Plant J* 78: 134-145
- Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet* 8: e1002910
- Barkan A, Small I (2014) Pentatricopeptide Repeat Proteins in Plants. *Annu Rev Plant Biol* 65: 415-442
- Binder S, Hölzle A, Jonietz C (2010) RNA processing and RNA stability in plant mitochondria. In: Kempken F (ed) *Plant Mitochondria*, pp. 107-130. Springer Science and Business Media, New York, NY, USA
- Binder S, Kühn K (2015) Determining mitochondrial transcript termini for the study of transcription start sites and transcript 5' end maturation. In: Whelan J, Murcha MW (eds) *Plant Mitochondria: Methods and Protocols*, pp. 13 - 30. Humana Press, New York
- Binder S, Stoll K, Stoll B (2013) P-class pentatricopeptide repeat proteins are required for efficient 5' end formation of plant mitochondrial transcripts. *RNA Biol* 10: 1322-1330

- Boussardon C, Avon A, Kindgren P, Bond CS, Challenor M, Lurin C, Small I (2014) The cytidine deaminase signature HxE(x)_n CxC of DYW1 binds zinc and is necessary for RNA editing of ndhD-1. *New Phytol* 203: 1090-5
- Chateigner-Boutin AL, Small I (2010) Plant RNA editing. *RNA Biol* 7: 213-219
- Cheng S, Gutmann B, Zhong X, Ye Y, Fisher MF, Bai F, Castleden I, Song Y, Song B, Huang J, Liu X, Xu X, Lim BL, Bond CS, Yiu SM, Small I (2016) Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J* 85: 532-47
- Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16: 735-742
- Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol* 133: 462-469
- Derksen M, Mertens V, Pruijn GJ (2015) RNase P-Mediated Sequence-Specific Cleavage of RNA by Engineered External Guide Sequences. *Biomolecules* 5: 3029-3050
- Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, Cattolico L, Falentin C, Giancola S, Renard M, Budar F, Small I, Caboche M, Delourme R, Bendahmane A (2003) Identification of the fertility restoration locus, Rfo, in radish, as a member of the pentatricopeptide-repeat protein family. *EMBO Rep* 4: 588-594
- Edwards K, Johnstone C, Thompson C (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res* 19: 1349
- Forner J, Hölzle A, Jonietz C, Thuss S, Weber B, Schwarzländer M, Meyer RC, Binder S (2008) Mitochondrial mRNA polymorphisms in different *Arabidopsis* accessions. *Plant Physiol* 148: 1106-1116
- Forner J, Weber B, Thuss S, Wildum S, Binder S (2007) Mapping of mitochondrial mRNA termini in *Arabidopsis thaliana*: t-elements contribute to 5' and 3' end formation. *Nucleic Acids Res* 35: 3676-3692
- Fujii S, Bond CS, Small ID (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proc Natl Acad Sci U S A* 108: 1723-1728
- Fujii S, Suzuki T, Giege P, Higashiyama T, Koizuka N, Shikanai T (2016) The Restorer-of-fertility-like 2 pentatricopeptide repeat protein and RNase P are required for the processing of mitochondrial orf291 RNA in *Arabidopsis*. *Plant J* 86: 504-13
- Haili N, Arnal N, Quadrado M, Amiar S, Tcherkez G, Dahan J, Briozzo P, Colas des Francs-Small C, Vrielynck N, Mireau H (2013) The pentatricopeptide repeat MTSF1 protein stabilizes the nad4 mRNA in *Arabidopsis* mitochondria. *Nucleic Acids Res* 41: 6650-6663
- Haili N, Planchard N, Arnal N, Quadrado M, Vrielynck N, Dahan J, Francs-Small CC, Mireau H (2016) The MTL1 Pentatricopeptide Repeat Protein Is Required for Both Translation and Splicing of the Mitochondrial NADH DEHYDROGENASE SUBUNIT7 mRNA in *Arabidopsis*. *Plant Physiol* 170: 354-66
- Hammani K, Giege P (2014) RNA metabolism in plant mitochondria. *Trends Plant Sci*
- Hauler A, Jonietz C, Stoll B, Stoll K, Braun H-P, Binder S (2013) RNA PROCESSING FACTOR 5 is required for efficient 5' cleavage at a processing site conserved in RNAs of three different mitochondrial genes in *Arabidopsis thaliana*. *Plant J* 74: 593-604
- Hayes ML, Dang KN, Diaz MF, Mulligan RM (2015) A conserved glutamate residue in the C-terminal deaminase domain of pentatricopeptide repeat proteins is required for RNA editing activity. *J Biol Chem* 290: 10136-42
- Hayes ML, Giang K, Berhane B, Mulligan RM (2013) Identification of two pentatricopeptide repeat genes required for RNA editing and zinc binding by C-terminal cytidine deaminase-like domains. *J Biol Chem* 288: 36519-36529
- Hölzle A, Jonietz C, Törjek O, Altmann T, Binder S, Forner J (2011) A RESTORER OF FERTILITY-like PPR gene is required for 5'-end processing of the nad4 mRNA in mitochondria of *Arabidopsis thaliana*. *Plant J* 65: 737-744

- Jonietz C, Forner J, Hildebrandt T, Binder S (2011) RNA PROCESSING FACTOR 3 is crucial for the accumulation of mature *ccmC* transcripts in mitochondria of *Arabidopsis thaliana* accession Columbia. *Plant Physiol* 157: 1430-1439
- Jonietz C, Forner J, Hölzle A, Thuss S, Binder S (2010) RNA PROCESSING FACTOR2 Is Required for 5' End Processing of *nad9* and *cox3* mRNAs in Mitochondria of *Arabidopsis thaliana*. *Plant Cell* 22: 443-453
- Kindgren P, Yap A, Bond CS, Small I (2015) Predictable alteration of sequence recognition by RNA editing factors from *Arabidopsis*. *Plant Cell* 27: 403-416
- Köhler D, Schmidt-Gattung S, Binder S (2010) The DEAD-box protein PMH2 is required for efficient group II intron splicing in mitochondria of *Arabidopsis thaliana*. *Plant Mol Biol* 72: 459-467
- Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. *Mitochondrion* 8: 5-14
- Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J* 4: 745-750
- Liu S, Melonek J, Boykin LM, Small I, Howell KA (2013) PPR-SMRs: Ancient proteins with enigmatic functions. *RNA Biol* 10: 1312-1321
- Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16: 2089-2103
- Pfalz J, Bayraktar OA, Prikryl J, Barkan A (2009) Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *EMBO J* 28: 2042-2052
- Sambrook J, Russell DW (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends Plant Sci* 13: 663-670
- Shen C, Zhang D, Guan Z, Liu Y, Yang Z, Yang Y, Wang X, Wang Q, Zhang Q, Fan S, Zou T, Yin P (2016) Structural basis for specific single-stranded RNA recognition by designer pentatricopeptide repeat proteins. *Nat Commun* 7: 11285
- Simon M, Simon A, Martins F, Botran L, Tisne S, Granier F, Loudet O, Camilleri C (2012) DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. *Plant J* 69: 1094-1101
- Stoll B, Binder S (2016) Two NYN domain containing putative nucleases are involved in transcript maturation in *Arabidopsis* mitochondria. *Plant J* 85: 278-88
- Stoll B, Stoll K, Steinhilber J, Jonietz C, Binder S (2012) Mitochondrial transcript length polymorphisms are a widespread phenomenon in *Arabidopsis thaliana*. *Plant Mol Biol* 81: 221-233
- Stoll B, Zendler D, Binder S (2014) RNA processing factor 7 and polynucleotide phosphorylase are necessary for processing and stability of *nad2* mRNA in *Arabidopsis* mitochondria. *RNA Biol* 11: 968-976
- Stoll K, Jonietz C, Binder S (2015) In *Arabidopsis thaliana* two co-adapted cyto-nuclear systems correlate with distinct *ccmC* transcript sizes. *Plant J* 81: 247-257
- Takenaka M, Verbitskiy D, Zehrmann A, Härtel B, Bayer-Császár E, Glass F, Brennicke A (2014) RNA editing in plant mitochondria —Connecting RNA target sequences and acting proteins. *Mitochondrion* 19 191-197
- Takenaka M, Zehrmann A, Brennicke A, Graichen K (2013) Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PLoS One* 8(6): e65343
- Wu W, Liu S, Ruwe H, Zhang D, Melonek J, Zhu Y, Hu X, Gusewski S, Yin P, Small ID, Howell KA, Huang J (2016) SOT1, a pentatricopeptide repeat protein with a small MutS-related domain, is required for correct processing of plastid 23S-4.5S rRNA precursors in *Arabidopsis thaliana*. *Plant J* 85: 607-21
- Yagi Y, Hayashi S, Kobayashi K, Hirayama T, Nakamura T (2013a) Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS One* 8: e57286

Yagi Y, Tachikawa M, Noguchi H, Satoh S, Obokata J, Nakamura T (2013b) Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA Biol* 10: 1236-1242

Yin P, Li Q, Yan C, Liu Y, Liu J, Yu F, Wang Z, Long J, He J, Wang HW, Wang J, Zhu JK, Shi Y, Yan N (2013) Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* 504: 168-71

Zoschke R, Watkins KP, Miranda RG, Barkan A (2016) The PPR-SMR protein PPR53 enhances the stability and translation of specific chloroplast RNAs in maize. *Plant J* 85: 594-606

Figure legends

Fig. 1. Polymorphic *ccmB* mRNAs in *Arabidopsis* accessions. **a** Schema of the *ccmB* gene (grey box) and the previously detected 5' (vertical lines) and 3' (vertical line with a dot) ends. Positions of the termini are given with respect to the translation start codon (5' ends, NATG, N = -1) and the translation stop codon (3' end, TAAN, N = +1). The localization of the potential RPF4 binding site from positions -267 to -252 is indicated. Oligonucleotides are indicated by vertical arrows. **b** CR-RT-PCR analysis of *ccmB* RNAs in Col and *Ler* and in reciprocal F₁ hybrids. Major products are indicated in the right margin. Oligonucleotides used for PCR are given below the image. Product sizes are mentioned in the text. The sequence analysis of product 2 did not reveal an alternative 5' end of *ccmB* RNAs. Sizes of DNA marker fragments are given in base pairs [bp]. **c** Primer extension analysis of *ccmB* RNAs using Terminator exonuclease (TEx) treated (+) and untreated (-) RNAs from Col and *Ler* (given above the image). Major products are indicated in the right margin. The image has been assembled to remove spaces and to change the order of the lanes, but all image parts derive from the same X-ray film. The oligonucleotide used for initiation of cDNA synthesis is given below the image. Sizes of relevant products are given in the text. Sizes of DNA marker fragments are given in nucleotides [nt].

Fig. 2. Identification of *RPF4*. **a** CR-RT-PCR analysis of *ccmB* RNAs from Col and *Ler* wild-type plants and three Col plants transformed with the *Ler* allele of At1g62910 (Col + At1g62910 (*Ler*), #2, #3 and #8). Products are obtained with primer pair Atccb2-6/Atccb2-3 and are identical to those described in Fig. 1B. **b** Northern blot analysis of *ccmB* RNA in four different wild-type accessions (given above the image) and a Col plant containing the *Ler* allele of

At1g62910 (Col+). Three different *ccmB* transcript species with approximate sizes of 1050 nt (1), 950 nt (2) and 850 nt (3) are indicated in the right margin.

Fig. 3. Processing competence of chimeric RPF4 proteins. **a** Diagram of RPF4 from *Ler* (RPF4-*Ler*, dark grey) and *Col* (RPF4-*Col*, light grey) showing the 16 repeats and the corresponding amino acids at positions 5 and 35 of each repeat and the predicted target site in the *ccmB* transcript. Perfect matches (:), poor matches (.) and mismatches (x) to the one repeat to one nucleotide recognition code are indicated. **b** Diagram of different chimeric RPF4 (RPF4-c1 to 8) composed of different parts of the gene from *Ler* and *Col*, indicated by the color code given in (a). **c** CR-RT-PCR analysis of the *ccmB* transcripts in *Col* plants containing the chimeric genes indicated in (b). Primer pair used is Atccb2-3/Atccb2-6. *Col* and *Ler* wild-type cDNAs are used as control.

Fig. 4. Binding studies of RFP4. RNA electrophoretic mobility shift assays performed with partially purified recombinant RPF4 protein from *Col* (a, RPF4-*Col*, left) or *Ler* (c, RPF4-*Ler* right) and an RNA oligonucleotide containing the potential *ccmB* binding site (given in the bottom parts) labeled with fluorescein (flc) (1 nM). Basic protein input varies from 15.6 to 500 nM. Complexes (marked by arrows) are formed with both proteins. Competition assays with unlabeled poly(C) oligonucleotides are shown in b for RPF4-*Col* and d for RPF4-*Ler*.

Stoll et al Fig. 1

(A)

(B)

Stoll et al Fig. 2

(A)

RPF4-Ler

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ESSNNGNSSSNTNNNN
35 SNDDDDNDNNNNTDND
.:.:.:.:.:.:.:.
uaauugcgaacaucuc

RPF4-Col

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ENSSTGNSSSSTNNNN
35 SDDDDDDNDNNNDNND
x.xx:....:x:....
uaauugcgaacaucuc

(B)

RPF4-c6

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ESSNTGNSSSSTNNNN
35 SNDDDDNDNNNDNND
.:.x:....:x:....
uaauugcgaacaucuc

RPF4-c4

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ESSNNGNSSSSTNNNN
35 SNDDDDNDNNNDNND
.:.:.:.:.x:....
uaauugcgaacaucuc

RPF4-c1

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ESSNNGNSSSNTNNNN
35 SNDDDDNDNNNDNND
.:.:.:.:.:.:.:.
uaauugcgaacaucuc

RPF4-c7

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ENSNNGNSSSNTNNNN
35 SDDDDNDNNNNTDND
x.:.:.:.:.:.:.
uaauugcgaacaucuc

RPF4-c2

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ENSSNGNSSSNTNNNN
35 SDDDDNDNNNNTDND
x.x:.:.:.:.:.:.
uaauugcgaacaucuc

RPF4-c8

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5 ESSNTGNSSSNTNNNN
35 SNDDDDNDNNNNTDND
.:.x:.:.:.:.:.:.
uaauugcgaacaucuc

(C)

Fig. 4. Binding studies of RFP4. RNA electrophoretic mobility shift assays performed with partially purified recombinant RFP4 protein from Col (**a**, RPF4-Col, left) or Ler (**c**, RPF4-Ler right) and an RNA oligonucleotide containing the potential *ccmB* binding site (given in the bottom parts), labeled with fluorescein (flc) (1 nM). Basic protein input varies from 15.6 to 500 nM. Complexes (marked by arrows) are formed with both proteins. Competition assays with unlabeled poly(C) oligonucleotides are shown in **b** for RPF4-Col and **d** for RPF4-Ler.

Supplemental Fig. S1. CR-RT-PCR analysis of *ccmB* RNA from various *Arabidopsis* ecotypes. Major products are indicated in the right margin. Sizes of the relevant products are given in the text. Amplifications were done with primer pair Atccb2-6/Atccb2-3.

Supplemental Fig. S2. Rough mapping of the *RPF4* locus in the *Arabidopsis* nuclear genome. Plants showing the recessive Col *ccmB* phenotype were genotyped with insertion/deletion (INDEL) markers, which are indicated at their approximate location on each of the chromosomes (numbers of the chromosomes are given in the top). The percentage of Col alleles found at the different INDEL markers is given. The rough mapping identified the *RPF4* locus within a 4 Mbp region on chromosome 1. Details on markers see Supplemental Table S2.

Supplemental Fig. S3. Fine mapping of the *RPF4* locus on chromosome 1. Plants with heterozygous genotypes at the chromosome 1 markers were used for fine mapping with additional INDEL or single nucleotide polymorphism (SNP) markers. Markers used are: CER 449181 (20.93 Mbp), CER 460336 (21.91 Mbp), CER 448685 (22.59 Mbp), CER 450204 (22.95 Mbp), CER 429364 (23.07 Mbp), CER 429086 (23.25 Mbp), CER 472531 (23.33 Mbp), CER 469696 (23.41 Mbp), CER 468588 (23.64 Mbp), CER 428577 (23.80 Mbp), CER 479819 (23.9 Mbp) and CER 464694 (24.01 Mbp). Details see Supplemental Table 2.

(A)

Data from NASC	A Col	B Ler								
	Marker Site (Mbp; Chr.1)	Marker Name	CS/N1929 (115)	CS/N1948 (194)	CS/N1962 (257)	CS/N1967 (279)	CS/N1971 (295)	CS/N1984 (356)	CS/N1985 (358)	CS/N1997 (398)
	23,0	*F19K23a	B	B	B	A	A	B	A	-
		*g4026	B	B	B	A	A	B	A	-
		*agp1e	B	B	B	A	A	B	A	A
	23,15	*ATHGENEA	B	B	-	-	-	-	-	-
		*T3P18a	B	A	B	A	A	-	B	B
		*AS33	B	B	-	-	B	A	A	-
		*mi324	-	B	B	A	-	A	A	-
		*mi353	A	B	B	A	B	A	A	-
*g2488b		-	-	A	A	B	A	A	A	
*F5I14a		A	B	A	A	B	A	A	-	
23,83	*mi424	A	B	A	A	B	A	A	-	
	*cyp86a2	-	-	-	A	-	-	-	-	
	*TOPP3	A	B	A	A	B	A	A	-	
	*m305	A	B	A	B	B	A	-	A	
	*m315	A	B	A	A	B	A	B	A	
23,89	*SGCSNP40	-	-	-	-	B	-	-	-	
	*SGCSNP213	-	B	-	A	B	A	A	-	
ccmB phenotype			Ler	Ler	Ler	Col	Ler	Col	Col	Col/Ler

Supplemental Fig. S4. Fine mapping of the *RPF4* locus using Col x Ler recombinant inbred lines (RIL). a RIL lines showing a genotype of interest are CS/N1929, CS/N1948, CS/N1962, CS/N1967, CS/N1971, CS/N1984, CS/N1985, and CS/N1997 (Lister and Dean, 1993). Genotypes are given in orange (Col) or yellow (Ler). The *ccmB* phenotypes were determined by CR-RT-PCR analyses with primer pair Atccb2-6/Atccb2-3.

(B)

<i>ccmB</i> phenotype		Ler	Ler	Ler	Col	Ler	Col	Col	Col/Ler
		CS/N1929 (115)	CS/N1948 (194)	CS/N1962 (257)	CS/N1967 (279)	CS/N1971 (295)	CS/N1984 (356)	CS/N1985 (358)	CS/N1997 (398)
Location on chr. 1	Marker name								
23,07	CER 429364	B		B		-	-	-	A+B
23,25	CER 429086	B		B		-	-	-	A+B
23,33	CER 472531	B		B		-	-	-	A
23,41	CER 469696	A		B		-	-	-	A
23,59	CER 470164	A		A		-	-	-	A

(C)

(D)

Supplemental Fig. S4. Fine mapping of the *RPF4* locus using Col x Ler recombinant inbred lines (RIL) (continued). **b** Three RILs CS/N1929, CS/N1962, and CS/N1997 were further genotyped with markers located in the genomic interval between 23.25 and 23.80 Mbp. These PCRs finally identified the *RPF4* locus within an 80 kbp genomic region between 23.25 and 23.33 Mbp. Genotypes: A = Col, B = Ler, A+B = Col/Ler. **c** Phenotype of *ccmB* transcripts in Col, Ler and two Col/Ler RILs determined by CR-RT-PCR with primer pair Atccb2-6/Atccb2-3. **d** Fine mapping of the *RPF4* locus by genotyping at four different markers (given below the images). Lines used are RIL N1929, N1962, N1997, Ler and Col.

Supplemental Fig. S5. Clustal Omega alignment of amino acid sequences deduced from different Ler- (Ler, En-1, Got-7 and Br-0) and Col-type (Shakdara, Sorbo, Tsu-1, Col, Ws) alleles of At1g62910. Positions with crucial functions for the interaction with distinct nucleotides (amino acid/nucleotide interaction code) are highlighted in yellow (5) or blue (35). Pentatricopeptide repeats are marked by dashed lines and numbered in red.

(A)

(B)

Supplemental Fig. S6. Function and putative binding of RPF4 to *rps3-rpl16* RNA. **a** CR-RT-PCR analysis of *rps3-rpl16* transcripts with primers Atrps3-2 and Atrpl16-3. The 400 bp PCR product indicates the -163 and +67 5' and 3' ends detected previously (Forner et al., 2007). The 220 bp product, amplified from Col and to minor amounts also in Sha indicates the +17/+15 and +67 5' and 3' termini. **b** Alignment of amino acids at positions 5 and 35 in 16 PPR motifs of RPF4-Col and RPF4-Ler, respectively, with a putative binding on *rps3-rpl16* mRNA (identical with the site in *orf240a*). Consistent with the divergent functions of these proteins binding of RPF4-Col exhibits a better fit to this potential binding site that the proteins from Ler.

Supplemental Fig. S7. Transcripts derived from chimeric constructs. Semi quantitative RT-PCR experiments confirmed transcription of the chimeric genes in Col. Reactions were done with 32 and 35 cycles, respectively. Primers used were given below the images.

Supplemental Fig. S8. Investigation of *ccmB* 5' ends in Col plants containing chimeric RPF4 constructs. CR-RT-PCR analysis using different 5' primers to detect distinct 5' ends in the upstream region of the *ccmB* gene. Primer pairs used are indicated on the left. Known products or sequenced products are indicated on the right. Position of the 5' termini are given with respect to the ATG (NATG, N = -1). P1: Atccb2-NS.H, P2: Atccb2-6, P3: Atccb2-1, P4: Atccb2-3, P5: Atccb2-NS.R, P6: Atccb2-10, P7: Atccb2-7, P8: Atccb2-12.

Supplemental Fig. S9. Partial purification of RPF4-Ler and RPF4-Col. Fractions obtained during His-affinity purification are: FT: flow through; W1 to W3: wash steps; E1 to E3: eluates 1 to 3. The arrow indicates the position of the recombinant RPF4 proteins. Sizes of marker proteins are given in [kDa].

SUPPLEMENTAL TABLES

Supplemental Table 1. Phenotypes of the *ccmB* transcripts in an F₂ mapping population.

	No. of plants with the Col <i>ccmB</i> phenotype	No. of plants with the Ler <i>ccmB</i> phenotype	Total no. of plants analyzed	Percentage of plants with the Col <i>ccmB</i> phenotype
Col X Ler	28	71	99	28,3 %
Ler X Col	22	77	99	22,2 %
Total	50	148	198	25,3 %

Phenotyping of the *ccmB* mRNA in 198 F₂ plants from Col X Ler and Ler X Col crossings. In 50 F₂ plants the Col phenotype was found while 148 F₂ plants showed the Ler phenotype, which is dominant. The 1:3 ratio indicates that a single nuclear gene is responsible for the *ccmB* polymorphism.

Supplemental Table 2: Marker used for the linkage analysis

Marker	Chromosome / Location [Mbp]	Oligos	Fragment length Col [bp]	Fragment length Ler [bp]
CER 452601	1/ 13.4	CER452601.H/ CER452601.R	288	213
CER 458022	1/ 16.4	CER458022.H/ CER458022.R	164	141
CER 479928	1/ 20.6	CER479928.H/ CER479928.R	190	148
CER 449181	1/ 20,93	CER449181.H/ CER 449181.R	244	198
CER 460336	1/ 21.91	CER460336.H/ CER460336.R	333	276
CER 448685	1/ 22.59	CER 448685.H/ CER 448685.R	196	170
CER 450204	1/ 22.95	CER450204.H/ CER 450204.R	634	353
CER 429364	1/ 23.07	CER429364.H/ CER429364.R	Eco RV 560	Eco RV 326, 234
CER 429086	1/ 23.25	CER429086.H/ CER429086.R	RsaI 591	RsaI 402, 189
CER 472531	1/ 23.33	CER472531.H/ CER 472531.R	AluI 310, 199, 15	AluI 240, 199, 69, 15
CER 469696	1/ 23.41	CER469696.H/ CER469696.R	537	154
CER 468588	1/ 23.64	CER468588.H/ CER468588.R	MunI 684	MunI 395, 289
CER 428577	1/ 23.80	CER428577.H/ CER428577.R	DraI 529, 155	DraI 381, 155, 148

CER 479819	1/ 23.9	CER479819.H/ CER479819.R	DraI 597	DraI 339 +259
CER 464694	1/ 24.01	CER464694.H/ CER464694.R	238	207
CER 449068	1/ 24.095	CER449068.H/ CER449068.R	236	202
CER 449079	1/ 24.140	CER 449079.H/ CER 449079.R	519	445
CER 452585	1/ 24.328	CER452585.H/ CER452585.R	307	252
CER 464632	1/ 24.6	CER464632.H/ CER464632.R	172	146
CER 449403	2/ 1.8	CER449403.H/ CER449403.R	230	183
CER 452047	2/ 9.2	CER452047.H/ CER452047.R	396	330
CER 464890	3/ 7.3	CER464890.H/ CER464890.R	216	162
CER 470109	3/ 22.8	CER470109.H/ CER470109.R	183	150
CER 460302	4/ 13.0	CER460302.H/ CER460302.R	218	185
CER 452226	4/ 1.5	CER452226.H/ CER452226.R	577	495
CER 456657	5/ 5.9	CER456657.H/ CER456657.R	633	535
CER 456556	5/ 13.7	CER456556.H/ CER456556.R	220	181

Supplemental Table 3. Oligonucleotides

Analysis of <i>ccmB</i>		
	name	sequence (5'-3')
	Atccb2-1	GGT GTG GAG GAG AAG ATC AG
	Atccb2-3	CAA TTT CGG TCT CGA TTA
	Atccb2-6	GTC GTC TCA TTT CCT TAC TTC
	Atccb2-8	AAA GGT GCA GGA GGA AAG G
	Atccb2-NS.H	GCA GCA GTT TAG TGT ATA AGC
	Atccb2-NS.R	CCT TCT TTC TAA TTC CTC CTA
Mapping of <i>RPF4</i>		
	name	sequence (5'-3')
	CER428577.H	ATG TTC ACC GTG TGA TCT GTG A
	CER428577.R	TGA TAT GGT TCC TAT TGT CCT GAG
	CER429086.H	TGT GTG CAC GCA TAG TGA TG
	CER429086.R	AGC CCG AAA GCT ATC TCA AG
	CER429364.H	ATG GTG GTG GAA AGT GAA TGT AG
	CER429364.R	CTA CTT CCA ATA TTC ATT GAC ATG AG

	CER448685.H	ACA AAC ACA GAA CAA AGG GTC
	CER448685.R	TTG AGT AGT TTA GTC GTT GGG TG
	CER449181.H	TAG GAG TTT GTC CCA GCT TTA G
	CER449181.R	ACC GAT GCA AGT ACA ATT GAC
	CER449403.H	TGA TTC CAG CGG GTA GGC ATG ACC
	CER449403.R	CGA AAC ATT CAA AAA GTG GTT ATG GGA GAG
	CER450204.H	TGA GCG AGA GTC ATC ATG TTG
	CER450204.R	TGG TTG AGA CTG GTA ACA AGG
	CER452047.H	CAA AAG CCA CAC TAA GGG ACA ACA ATC C
	CER452047.R	TGC ACC ATT GAG ACA AGC TCT ATA CAA TCT G
	CER452226.H	GAG TTC GAC GGT GGA GGA GAT TGC
	CER452226.R	TGT TGA GTT ACT TTA AGC TCT GTG TAC ATT CTC CAG
	CER452601.H	CCA TTG AAG CAA GAA ATT TGA GGA AGA TGG
	CER452601.R	TGG ATG CCT AAT ACC GGG ACT ATT TCC TG
	CER456556.H	CGG TTT TTT AAT GGG CTT CAC GTG TTG
	CER456556.R	CAC ACT CTG CTA CGA ATT TTC TCG AAC TCA G
	CER456657.H	CCT TCA GAA TCA CCA ACT GTG GAG TCC
	CER456657.R	GAA AGA AAC CAT ATG CAT CAA ATC TAC CTT CG
	CER458022.H	TCA TTC TTA TAG CAA AGC TTT CGA TCC ATA
	CER458022.R	GCA GTT GGG AGT CTC ATG TAT TCT ATG ATA GGC
	CER460302.H	GTC GGC TTA CAA ACG CGT TTA TGT CC
	CER460302.R	CTG CTA AGT CGA TCA AAG CCA GAT CG
	CER460336.H	ACA TGT GGC TGA GAC TAT TGC
	CER460336.R	TCA ACT CGG TCC AAA CTT TCG
	CER464632.H	TCT GCG AAG TGG AAG AGA AAA AAG AAC TC
	CER464632.R	TTC ATT AGT TTT GCC GGT GAT TGT GAC C
	CER464694.H	TGA GAA ACA TCA TCG GAT GTT GAT AGT ACA TCG
	CER464694.R	GCT CTT AGT TGT TGT AAT GAA TAT GAA ATC TTA GTC G
	CER464890.H	CGG TAC TGC TCG AAG GAA CGA GAC TG
	CER464890.R	GCC TCT GTG GCT GCT ACT GAA TTC G
	CER468588.H	GCT GAC CTA GCA ATT CGA GC
	CER468588.R	TCA TAA CAA ATG CTC ACC GC
	CER469696.H	ATC TGC TTT TGG AAT TGT TGT G
	CER469696.R	CTA CAT GTG TGT ATT TGG TTG C
	CER470109.H	TTT TGT GAT TCT ATG TTG GCA TCA GTG G
	CER470109.R	CAA ACC GTA GTG CAA ATT AGA GCA TAG TTT CC
	CER472531.H	ACG GAT GAT CCA AAA TCG AA
	CER472531.R	GCC ACT AGC GGT GTT ACC TT
	CER479819.H	ATG AAG AAG GTC AAA CCG ACC AG
	CER479819.R	TCA TCT ACC GAA TTA GGC ATC G
	CER479928.H	CAG TTC CCA ATG AGC AAT GTA CAC ATG G
	CER479928.R	CCT TGT TGG ATG TAA AAT AGT TTC GGG TTC C
Complementation constructs and sequence analysis		
	At1g62860-Kompl.Ler.H	CAG CTC GCC GCG CGG TAT CGT GTG GCA CGT GC
	At1g62860-Kompl.R	GCT TAA TTA ACT CGC GTA GAA TCA TGT TCG
	At1g62910Col.Teil1	GGT TTA TAT CCC ATT TCG ACC
	At1g62910Col.Teil2	GGT ATG GAA CTC TTC CGC
	At1g62910-kompl.H	AGG CGC GCC TCG TAC AGT CGT ACT TGC GGA
	At1g62910-kompl.R2	TGG GTC TTA ATT AAT GCT CGA TTC CTT CCT CGC A
	At1g62910Konstrukt4.Col	TTT CAG CGC ATT GAT TGA TGC
	At1g62910Konstrukt4.Ler	TCG TAC AAT TTC TCA GCC
	At1g62910Konstrukt5.Col	ACC TAC AGC TCC CTC ATA AGT TGC C
	At1g62910Konstrukt7.Col	AAG ATG ATG AAA CTC GGC

	At1g62910Konstrukt7.Ler	AAG CGT GAC AAT ATC GGG
	At1g62910Ler.1	TGG AAG AGG CCC ATC TTC
	At1g62910Ler.2	GTC GTA CAA TTT CTC AGC C
	At1g62910Ler.3	CCA TTT GAT CAA CCA AAG C
	At1g62910Ler.6	GCA GGG AAG GTG GAA G
	At1g62910Ler.7	CCT TCC ATC ATT GAG TTC AGC
	At1g62910Ler.8	TGG TTG ACA CCC TCT GGC
	At1g62910Ler.9	ATC TGC TTA GCC TCA CCA AGG
	At1g62910Ler.10	TGC TTG AGA AGA ATA TCA ACC CTG
	At1g62910Ler.11	TAA GTA TAG TGA TTA TGA GGC
	At1g62910Ler.12	TGC AAC TGT TTC TCC CTC
	At1g62910Ler.kompl.1	ATG CAG TCG AGC GGC GCG CCA AGT TTG GAG AAC CCT TAC G
	At1g62910Ler.Teil1.1	CTG TTT GAA AAC CAT TTG GGC
	At1g62910Ler.Teil1.2	GCA TCA CTC CAT CTT CCG
	At1g62910Ler.Teil2	AGG ATT TCT GAT GCC G
	At1g62910.Br-0.R1	ATG AAT CGG GCT GAT ATC CC
	At1g62915-kompl.H	TCG AGC GGC GCG CCT GCT ATT TAT CTC ATT TTC AGT ATC
	At1g62915-kompl.R	TGG GTC TTA ATT AAG TTT GTC CAG CCA CAC AC
	At1g62915Ler.1	ACG AAT CAC ATC GTT CTC
	At1g62910Ler.2	TGG AAG AGT CCC ATC TTC
	At1g62910Ler.3	GCT TCA GAA GCA GTG GC
	At1g62930-kompl.H	ATG CAG TCG AGC GGC GCG CCC CTG AAC AAA AGA GAC ATT AAC TGA AC
	At1g62930-kompl.R	GAC TTC GAG TTT AAT TAA TTC GAA CCG ACA AAT TGT TTG TC
Other primers		
	name	sequence (5'-3')
	At18S-Adapter.H3	CTT ATT GGC GCA TAC CAC GGT GG
	At18S-Adapter.H4	CCT AGC TCC CCG AGA ACA ACG TTC G
	cornichon.1	CTC GCT TAC ATT ATC CTC AC
	UBC9-H.2	CTA CTT CAT GTA GCG CAG GAC
	At4g27960_2	TCA TCT GGG TTT GGA TCC GT
RNA probes for REMSA		
	ccmB	GUA AUU GCG AAC AUC UCU CUU
	Poly(C)	CCC CCC CCC CCC CCC CCC CC