uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Thomas Hazle

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Mitochondrial S1 Ribosomal Protein Genes in Legumes :
Expression and Transfer to the Nucleus During Evolution

TITRE DE LA THÈSE / TITLE OF THESIS

Linda Bonen

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Robert W. Lee

Doug Johnson

Christiane Charest

Myron Smith

Lynn Gillespie

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

MITOCHONDRIAL S1 RIBOSOMAL PROTEIN GENES IN LEGUMES: EXPRESSION AND TRANSFER TO THE NUCLEUS DURING EVOLUTION

Thomas Hazle

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfillment of the requirements for the
Ph.D. degree in the

Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des études supérieures et postdoctorales
Université d'Ottawa
en vue de l'obtention du doctorat

L'institut de biologie d'Ottawa-Carleton

Your file Votre référence
ISBN: 978-0-494-61249-1
Our file Notre référence
ISBN: 978-0-494-61249-1

MITOCHONDRIAL S1 RIBOSOMAL PROTEIN GENES IN LEGUMES: EXPRESSION AND TRANSFER TO THE NUCLEUS DURING EVOLUTION

ABSTRACT

The gene encoding the mitochondrial S1 ribosomal protein (*rps1*) is regarded as being among the most dynamic with respect to gene transfer from the mitochondrion to the nucleus during flowering plant evolution. To provide a detailed characterization of the *rps1* gene, including transcriptional status and transcript editing, as well as to address issues surrounding its transfer to the nucleus, I examined mitochondrial- and nuclear-located *rps1* sequences in closely-related legume species. I found that this gene is nuclear-located in alfalfa and its very close relatives, and is a pseudogene in their mitochondria. In contrast, the functional *rps1* is mitochondrial-located in legume species that are slightly more distantly-related. The nuclear-located copies have retained mitochondrial-type upstream sequences and these provide part of the *rps1* 5' UTR, and the 3' splice site of an upstream intron. This is the first such documented case and illustrates that sequences of mitochondrial origin can be used as nuclear-specific expression elements. From comparative sequence analysis, I inferred that this single transfer event occurred recently in legume evolution. As such, the conservation among *rps1*-associated sequences permitted the evaluation of adaptive changes that occurred following integration into the nuclear genome.

By sequencing *rps1* (and *ψrps1*) flanking regions in the mitochondria of these legumes, I found that *rps1* is linked downstream to the first two exons of *nad5*, yet in each lineage, the genomic environment greater than ~60 nt upstream of *rps1* differs. Northern analyses suggest that the lineage-specific upstream sequences contribute to the variation in transcript profiles, which are complex in part due to the *cis*- and *trans*-arrangement of the five exons of *nad5*. Variation among homologous sequences immediately preceding mitochondrial-located *rps1* in legumes was observed, and using a bioinformatics approach, a broad range of sequence conservation (including non-homology) was seen upstream of protein-coding genes located in the seven completely-sequenced flowering plant mitochondrial genomes. This along with the absence of an RNA-binding domain from the mitochondrial S1 ribosomal protein, and the divergence

of the 3' end of the SSU rRNA from that in bacteria point to the use of a non-classical-bacterial type of ribosome binding in initiator codon recognition. The duplication and recruitment of mitochondrial upstream sequences by multiple genes, as well as the available copies of such sequences that are dispersed in intergenic regions may be an indication of how plant mitochondrial gene expression systems tolerate (or exploit) the highly-recombinogenic genome.

LES GÈNES MITOCHONDRIAUX DE LA PROTÉINE RIBOSOMALE S1 DANS LES LÉGUMINEUSES : EXPRESSION ET TRANSFERT VERS LE NOYAU PENDANT L'ÉVOLUTION

RÉSUMÉ

Le gène codant pour la protéine ribosomale mitochondriale S1 (*rps1*) est perçu comme un des plus dynamiques concernant le transfert du gène de la mitochondrie vers le noyau pendant l'évolution des plantes à fleurs. Dans l'objectif de réaliser une caractérisation détaillée du gène *rps1*, incluant l'état transcriptionel et l'édition du transcrit, et de répondre à certaines questions concernant le transfert au noyau, j'ai examiné les séquences de *rps1* dans les mitochondries et les noyaux d'espèces étroitement reliées de légumineuses. Les résultats ont montré que ce gène se trouve dans le noyau de la luzerne ainsi que chez des espèces lui étant proches . Par contre, ce gène réside dans la mitochondrie d'espèces de légumineuses plus éloignées. Les copies nucléaires ont gardé les séquences en amont de type mitochondrial et celles-ci fournissent une partie de la région 5' non-traduite de *rps1*, ainsi que le site 3' d'épissage de l'intron en amont. C'est la première fois qu'un tel cas est décrit, et révèle que les séquences d'origine mitochondriale peuvent être utilisées comme éléments spécifiques à l'expression dans le noyau. Les analyses comparatives entre ces espèces de *rps1* et de ses séquences environnantes m'ont permis de déduire que ce transfert vers le noyau est un événement unique et s'est passé récemment dans l'évolution des légumineuses. Ainsi, la conservation entre les séquences associées à *rps1* a permis l'évaluation des changements adaptatifs qui se sont passés suite à l'intégration dans le génome nucléaire.

En séquençant les régions environnantes de *rps1* dans les mitochondries des légumineuses, j'ai remarqué que *rps1* est lié en aval aux deux premiers exons de *nad5*, bien que dans chaque lignée, l'environnement génomique >~60nt en amont différait. Des analyses Northern ont suggéré que les séquences en amont de lignées spécifiques contribuent aux variations des profils de transcription observés entre les espèces, et que ces profils sont complexes en partie à cause de l'arrangement *cis* et *trans* des cinq exons de *nad5*. Des variations entre les séquences homologues immédiatement en amont de *rps1* dans les mitochondries des légumineuses ont été observées et, en utilisantune approche bioinformatique, il fut révélé qu'une large gamme de conservation de la séquence (allant même jusqu'à la non-homologie) existe en amont des gènes codant pour

des protéines dans les sept génomes mitochondriaux de plantes à fleurs complètement séquencés. Ce fait, avec la perte du domaine de liaison d'ARN de la protéine ribosomale mitochondriale S1 dans les plantes, et la disparité du bout 3' de la petite sous-unité de l'ARN ribosomal avec celui des bactéries, pointent vers l'utilisation d'un type bactérien non-classique de liaison du ribosome pendant la reconnaissance du codon initiateur. La duplication et le recrutement de séquences en amont par plusieurs gènes, ainsi que la disponibilité de copies dispersées dans les régions intergéniques nous fournissent un indice quant à la manière dont les systèmes d'expression des gènes mitochondriaux de plantes tolèrent (ou exploitent) leur génome grandement recombinant.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the many people who have contributed to the success of this project. Most importantly, I would like to express appreciation to my supervisor Dr. Linda Bonen for her guidance and direction. I would also like to acknowledge the members of my advisory committee, Dr. Myron Smith, Dr. Guy Drouin and Dr. Lynn Gillespie, as well as a former committee member, Dr. Donal Hickey.

Other members of the lab, past and present, have been an invaluable source of support and helpful discussions. Very special appreciation goes to Dr. Jennifer Li-Pook-Than, Sophie Calixte, B. Young and J. Hardy; and to members of the Johnson lab, Dr. Doug Johnson and Dr. C. Webb. I would also like to thank Kerry Mellett for her help in gathering bean mitochondrial *rps1* data. Finally, I would like to acknowledge my wife, Dr. Tracy Burton for her support and insights into plant evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ψrps1	pseudogene of the S1 ribosomal protein gene
C-terminal	carboxy-terminal
CMS	cytoplasmic male sterile/sterility
chl	chloroplast
EST	expressed sequence tag
LSU	large ribosomal subunit
mtDNA	mitochondrial DNA
mtRNA	mitochondrial RNA
Mya	million years ago
N-terminal	amino-terminal
nad	genes encoding NADH dehydrogenase subunits
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PPR	pentatricopeptide repeat (protein)
RACE	rapid amplification of cDNA ends
rpl	ribosomal protein (large subunit) gene
rps	ribosomal protein (small subunit) gene
SSU	small ribosomal subunit
UTR	untranslated region

CHAPTER 1: GENERAL INTRODUCTION

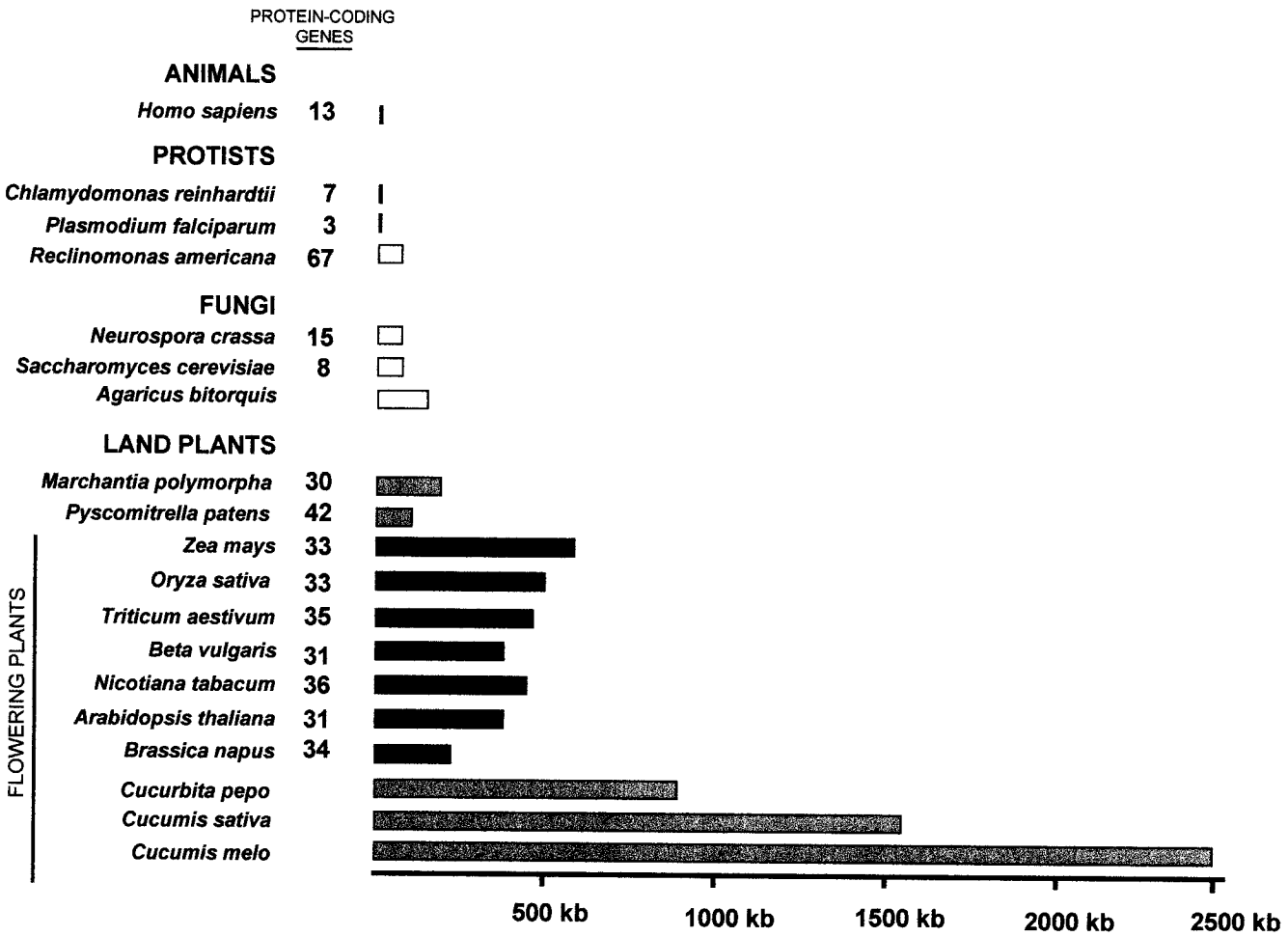
1.1 The origin of mitochondria

The endosymbiont hypothesis explains that mitochondria in eukaryotes originated as a free-living α-proteobacteria-like organism "engulfed" by an ancestor of present-day eukaryotes (Timmis et al. 2004, Gray 1992). Following this event, genes that were essential for free-living were lost from the endosymbiont, while other genes required for functions such as respiration were retained. Although gene loss due to functional replacement was (and still is) a potential fate, many of the remaining genes were transferred to the nucleus of the host. They retained their function, as the encoded proteins would be targeted back to the endosymbiont, which was then reduced as the mitochondrion of the host cell. Since their origin, mitochondrial genomes in various lineages have diverged considerably, and while they all contain only remnants of their bacterial progenitors, the major lineages have evolved many distinctive features. For example, mitochondrial genome sizes range from ~16 kb in humans to hundreds of kb in plants (Fig. 1.1), and these size differences do not correlate with differences in gene content. This is illustrated by comparisons of the human and *Arabidopsis* mitochondrial genomes, which contain a relatively similar number of protein coding genes (13 vs. 31) considering their ~20-fold difference in size (see Fig. 1.1). Other differences include rate of nucleotide substitution (e.g. high in animal mitochondria, low in plant mitochondria; Lynch et al. 2006), use of genetic code (for example, the use of UGA to encode tryptophan rather than as a stop codon in many non-plant mitochondria) and aspects of their expression systems.

1.2 Flowering plant mitochondrial genomes

Plant mitochondrial genomes are not only large relative to other lineages, but also highly variable in size, ranging from about 200 kb to >2000 kb (Fig. 1.1). The gene content among species is similar (50-60), and these consist of ~35 protein coding genes mostly encoding proteins of complexes I-V of the electron transport chain and ribosomal subunits as well as 3 ribosomal RNAs and ~15-20 tRNAs. Thus, most of the differences in genome size are accounted for by variation in intergenic regions, which can increase due to repeated regions (which are thought to often be scrambled to the point of non-recognition), and to a

Figure 1.1 Comparison of genome sizes (bars) and protein coding gene content (where known due to complete genome sequence data) of mitochondria of representative species of the major eukaryotic lineages. Plant genomes are represented in green and flowering plant species with completely-sequenced mitochondrial genomes are shown in red. Adapted from Gray (1988).



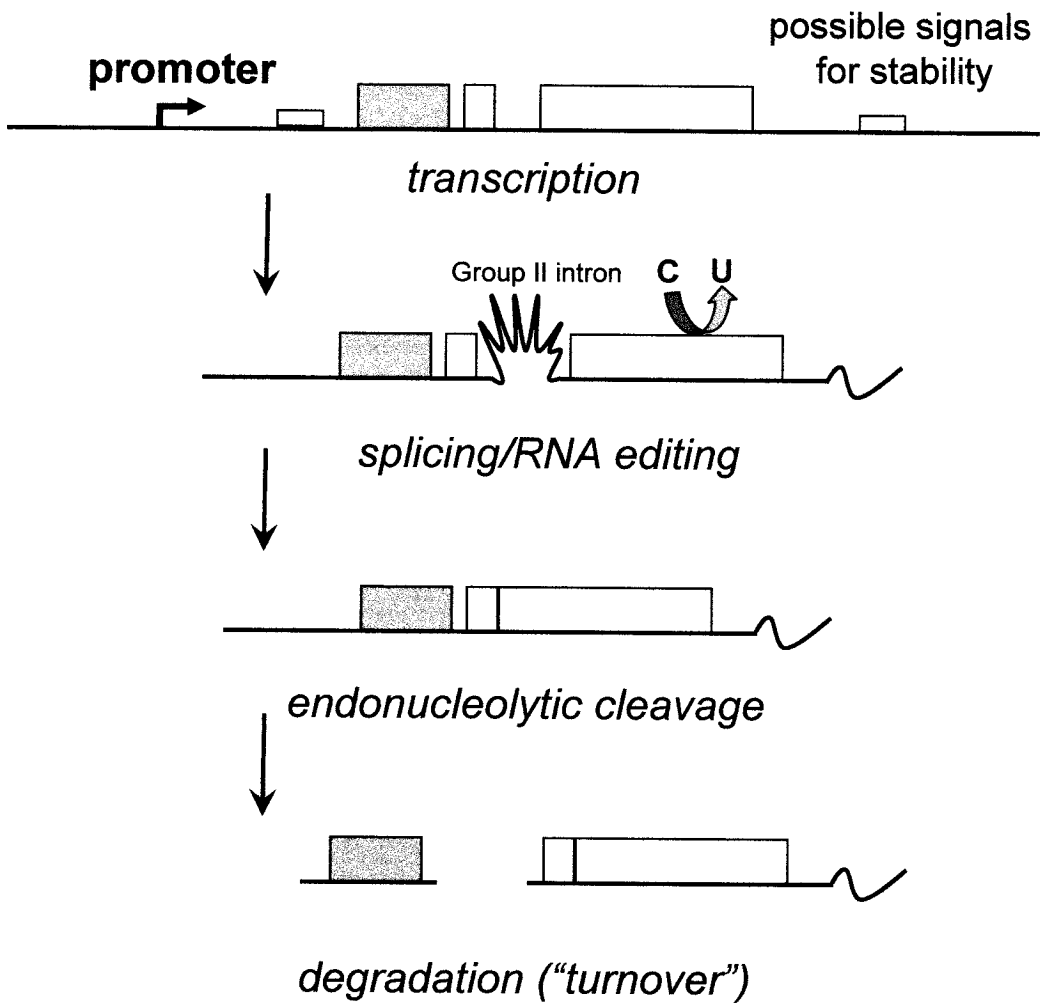
adapted from Gray (Biochem. Cell Biol. 1988)

lesser extent, the uptake of foreign (e.g. nuclear or chloroplast) DNA (Kubo and Mikami 2007). In contrast to the low rate of nucleotide substitution, flowering plant mitochondrial genomes are highly recombinogenic (Kubo and Mikami 2007). Recombination across repeated regions of the genome results in various isomeric and subgenomic forms of the deduced master chromosome. For example, in the wheat mitochondrial genome, Ogihara et al. (2005) detected 16 repeats, while in the *Brassica* mitochondrial genome, Handa (2003) found only a single duplicated region of ~2 kb. Over evolutionary time, intra- or intermolecular recombination has resulted in different genomic organizations, even among closely related species (Kubo and Mikami 2007).

1.3 Plant mitochondrial gene expression

Relative to other lineages little is known about gene expression in plant mitochondria. As in animals and fungi, transcription is driven by a T7 bacteriophage type RNA polymerase (Kuhn et al. 2007), and in plants the promoters have very loosely-conserved sequence features (Hoffman et al. 2001). Reflecting their eubacterial ancestry, genes in plant mitochondria are often co-transcribed into polycistronic transcriptional units. In flowering plants some bacterial-type gene linkages have been maintained, such as *rpl2-rps19* and *rps3-rpl16* in rice (Subramanian et al. 2001). Virtually nothing is known about transcription termination. It appears that eubacterial-like transcription terminators are not used in plant mitochondria, and it is unclear if any analogous signals exist (Hoffman et al. 2001). Plant mitochondrial transcripts can undergo complex processing to generate mRNAs (Fig. 1.2). When genes are co-transcribed, endonucleolytic cleavage often, but not always, generates monocistronic mRNAs. Some plant mitochondrial genes contain group II introns, which range in number from about 20-24 in flowering plants, and must be spliced out. Most of these introns are located in *nad* genes, which encode proteins of complex I of the electron transport chain. A small number of these genes, namely *nad1*, *nad2* and *nad5* contain group II introns that require trans-splicing as exons are located (and transcribed) in different genomic regions. It is unclear as to what elements confer stability of many plant mitochondrial transcripts, however stem-loop structures have been implicated in some cases (Hoffman et al. 2001, Forner et al. 2007). In contrast to eukaryotic mRNAs, polyadenylation

Figure 1.2 Model of the major steps in the processing of mitochondrial transcripts in flowering plants. Co-transcription of two genes (shown as gray and open blocks) from a common promoter are processed into monocistronic mRNAs by group II intron splicing, RNA editing and endonucleolytic cleavage. Signals for stability can reside in 5' or 3' UTRs, and polyadenylation has been implicated in transcript turnover. Figure adapted from Binder and Brennicke (2003).


promoter
possible signals
for stability
transcription
Group II intron
C
U
splicing/RNA editing
endonucleolytic cleavage
degradation ("turnover")

of plant mitochondrial transcripts has been implicated in degradation or 'turnover', and this is similar to what is observed in eubacteria.

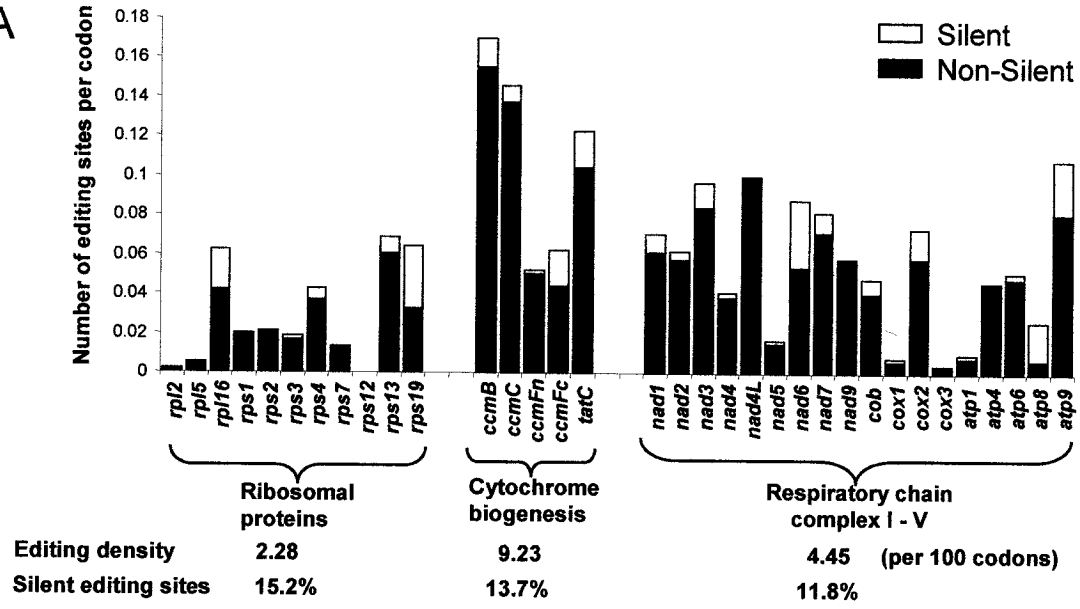
Another interesting feature of plant mitochondrial gene expression is C to U type RNA editing. In this, certain cytosine residues of the transcript are modified to uracil residues, and the modification usually restores a codon which encodes a conserved amino acid. However, editing sometimes occurs at silent positions and in UTRs, and occasionally occurs at unpredicted sites (i.e. results in a non-conserved amino acid) (Fig.1.3). In the mitochondria of rice 491 editing sites were reported (Notsu et al. 2002) but these are not uniformly distributed. For example, genes encoding ribosomal proteins appear to contain fewer editing sites than genes encoding respiratory-chain or cytochrome-biogenesis proteins (Fig. 1.3), and this is consistent with observations made for *Arabidopsis* (Giegé and Brennicke 1999). The editing status of a particular site can differ among closely-related species with the replacement of a cytosine with a genomically-encoded uracil. In *Arabidopsis*, the extent of editing (i.e. the proportion of transcripts showing editing at a particular site) and the number of edited sites was shown to vary between ecotypes and tissue types, and those sites that remained unedited in one of the ecotypes were mostly at silent positions (Bentolila et al. 2007). Editing also occurs in chloroplasts (although to a much lesser extent), and in both organelles experimental evidence suggests that nuclear-encoded proteins are involved in editing site recognition, although the actual enzyme responsible for editing has not been identified (Bentolila et al. 2007).

1.3.1 Translation initiation in plant mitochondria

Virtually nothing is known about mechanisms of the initiation of translation and initiator codon recognition in plant mitochondria. As their genomes have certain features reflecting their endosymbiont origin, it might be expected that initiator codon recognition is aided by bacterial-like Shine-Dalgarno ribosome binding (reviewed in Marintchev and Wagner 2005). In this, base pairing between the 3' end of the 16S (small subunit) rRNA and a purine-rich 'Shine-Dalgarno' sequence preceding initiator codons on bacterial mRNAs aids in the stabilization of the ribosome in the correct location for translation initiation (Fig.1.4A). In addition, the RNA-binding domain of the S1 ribosomal protein interacts with AU-rich sequences on the mRNAs, and also aids in stabilization (Komarova et al. 2005). However,

Figure 1.3 [A, B] Distribution of editing sites in the rice mitochondrial genome (compiled from Notsu et al. 2002; Hazle and Bonen, unpublished data) [A] Frequency of silent and non-silent editing sites by gene, summarized for gene types below. [B] Distribution of editing sites within ribosomal protein genes showing non-silent and silent edits, as well as editing occurring at unexpected locations ("Unexpected edit") and absence of editing at predicted sites ("Absent"). Open triangles indicate locations of introns.

A



Editing density	**2.28**	**9.23**	**4.45**	**(per 100 codons)**
Silent editing sites	**15.2%**	**13.7%**	**11.8%**	

B

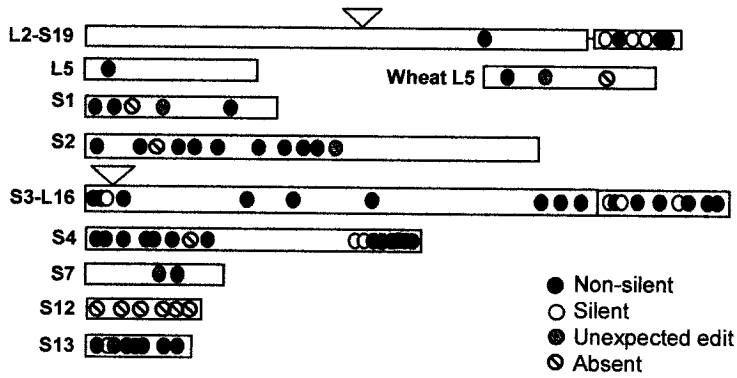
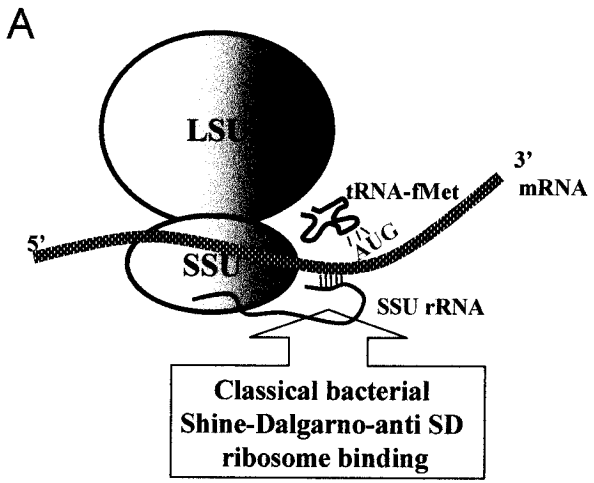
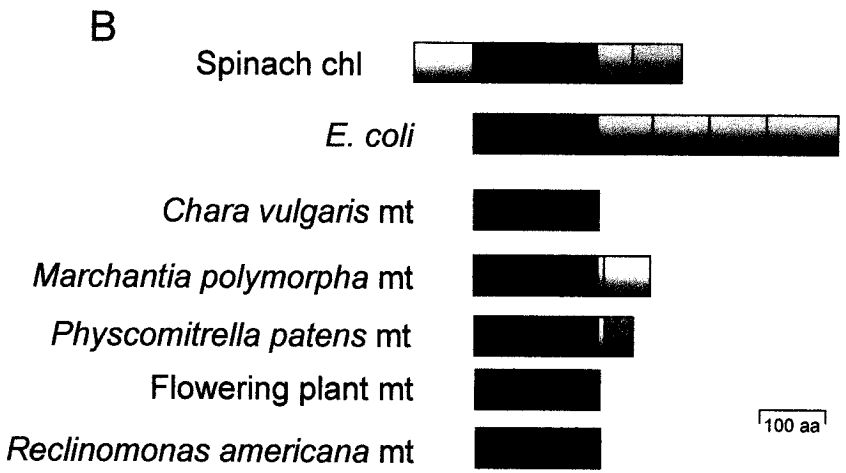
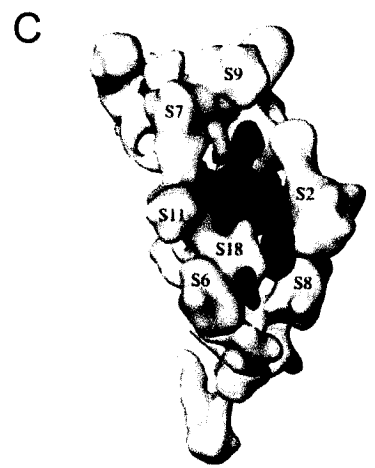


Figure 1.4 [A] Diagram of Shine-Dalgarno (bacterial) ribosome binding showing the interaction of the 3' end of the 16S rRNA with the Shine-Dalgarno sequence on a bacterial mRNA. SSU and LSU indicate the small and large ribosomal subunits. [B] Schematic of the S1 ribosomal protein in *E. coli*, various mitochondrial lineages (mitochondrial-encoded) and the chloroplast of spinach (nuclear-encoded). Ribosome binding domains shown as red blocks and RNA-binding domains shown as blue blocks. Colours indicate homology. [C] Position of the S1 ribosomal protein (red) in the 16S rRNA in *E. coli* relative to other ribosomal proteins (green), adapted from Sengupta et al. (2001).


A
LSU
tRNA-fMet
3'
mRNA
5'
AUG
SSU
SSU rRNA
Classical bacterial
Shine-Dalgarno-anti SD
ribosome binding


B
Spinach chl
E. coli
Chara vulgaris mt
Marchantia polymorpha mt
Physcomitrella patens mt
Flowering plant mt
Reclinomonas americana mt
100 aa


C
S9
S7
S2
S11
S18
S6
S8

the 3' end of the SSU rRNA in the mitochondria of land plants is unlike that in bacteria and the mitochondrial S1 ribosomal protein has no apparent RNA-binding domain (Fig. 1.4B, C). Thus, it is unclear how plant mitochondrial initiator codon recognition is achieved. However, it appears that plant mitochondria use a system unlike that in animal mitochondria, in which mRNAs contain either no leaders, or ones that are very short (Boore 1999). It would also seem unlikely that plant mitochondria use something analogous to the eukaryotic scanning model of initiator codon recognition (reviewed in Marintchev and Wagner 2005) considering things like co-transcription in plant mitochondria, the absence of 5' caps on their mRNAs, and the occurrence of AUG triplets in the 5' UTRs. In the mitochondria of yeast, translation initiation is aided by gene specific nuclear-encoded machinery (reviewed in Costanzo and Fox 1990), and something similar may be occurring in plants.

1.4 Transfer of plant mitochondrial genes to the nucleus

During early mitochondrial evolution, most of the genes that were retained by the symbiont were transferred to the nucleus of the host. In some lineages however, gene transfer from the mitochondria was stopped with the evolution of a different mitochondrial genetic code, and this would preclude correct translation of transferred genes in the cytosol. In plant mitochondria (which use the standard genetic code) gene transfer to the nucleus is still occurring at a surprisingly high rate (Bonen 2006, Adams and Palmer 2003). It is unclear if early transfer events were DNA or RNA-mediated (or both). It is observed for recent transfers in plants that translocated sequences do not contain group II introns, and C-to-U editing sites appear as if edited. This supports the idea that transfers are mediated by a mature mRNA molecule, and this includes the possibility that sequences are translocated as DNA sequence following recombination with a cDNA intermediate (reviewed in Timmis et al. 2004).

Transfer in flowering plants mostly involves genes encoding ribosomal proteins. As illustrated in Table 1.1, only subsets of the 16 ribosomal protein genes present in the mitochondria of the liverwort *Marchantia polymorpha* are still found in the mitochondria of flowering plant lineages, and this sporadic nature is thought to reflect multiple gene transfer events (Bonen and Calixte 2005). Using Southern hybridization experiments, Adams et al. (2002b) inferred the number of independent losses (and likely nuclear relocation) of genes

Table 1.1 Ribosomal protein genes[1] in the completely-sequenced mitochondrial genomes of seven flowering plants, *Marchantia polymorpha* (liverwort) and *Physcomitrella patens* (moss).

	Maize	Rice	Wheat	Sugar Beet	Tobacco	*Arabidopsis*	*Brassica*	*Marchantia*	*Physcomitrella*
rps1	+	+	+	-	+	-	-	+	+
rps2	+[2]	+	+	-	-	-	-	+	+
rps3	+[2]	+	+	+	+	+	+	+	+
rps4	+	+	+	+	+	+	+	+	+
rps7	+	+	+	+	-	+	+	+	+
rps8	-	-	-	-	-	-	-	+	-
rps10	-	-	-	-	+	-	-	+	-
rps11	-	ψ	-	-	-	-	-	+	+
rps12	+	+	+	+	+	+	+	+	+
rps13	+	+	+	+	+	-	-	+	+
rps14	-	ψ	-	-	ψ	ψ	+	+	+
rps19	-	+	ψ	-	+	ψ	-	+	+
rpl2[3]	-	+	ψ	-	5'	5'	5'	+	+
rpl5	-	+	+	+	+	+	+	+	+
rpl6	-	-	-	-	-	-	-	+	+
rpl16	+	+	+	-	+	+	+	+	+

[1] +, - and ψ = present, absent and pseudogene

[2] two copies of *rps2*, and two copies of exon 1 of *rps3* are present in the maize mitochondrial genome

[3] *rpl2* was split into 2 genes in early eudicot evolution (cf. Adams et al. 2001), and only the 5' region is mitochondrial- (vs. nuclear-) located in tobacco, *Arabidopsis* and *Brassica.*

from the mitochondria during flowering plant evolution, and the most dynamic ribosomal protein genes appear to be *rps1*, *rpl2*, *rps7* and *rps19*, having over 30 inferred independent losses each.

1.4.1 The steps of gene transfer

The basic steps for successful gene transfer include the escape of sequences from the mitochondria, sequence incorporation into the nuclear genome, and the acquisition of the appropriate regulatory sequences, such as promoters, or elements found in the untranslated regions (UTRs) of mRNA (Brennicke et al 1993) (Fig. 1.5). The original regulatory elements would be mitochondria-specific and unlikely effective during translation in the cytosol. For example, promoters in mitochondria are unlike those in bacteria, the nucleus of eukaryotes, and even chloroplasts (Hoffmann et al 2001). Usually, a presequence must also be acquired that encodes a targeting peptide, as the protein must be correctly targeted to the mitochondria. In plants, these targeting peptides are usually abundant in hydrophobic, hydroxylated and positively-charged amino acids, and have the potential to form amphiphilic α-helices (Mackenzie 2005). They otherwise show little sequence conservation even in their length, which averages about 40 amino acids (Peeters and Small 2001). This variability is consistent with the idea that such sequences can be derived *de novo*, or targeting signals can reside within the core protein sequence (Adams and Palmer 2003). Some ways in which translocated sequences can acquire the necessary regulatory sequences (and thus be 'activated') have been identified, and many of these involve the recombination of the migrated mitochondrial sequence into a region already containing regulatory information, often including a presequence that encodes a target peptide. Reported cases for ribosomal protein genes are summarized in Table 1.2 and many of these transfers involved the integration of a translocated sequence into a region containing copies of a gene that encode a mitochondrial protein. Interestingly in other cases, the origin of regulatory sequences is unknown. Another case involves regulatory sequence acquisition through alternative splicing.

1.4.2 The transition stage

Following the activation of the nuclear-located copy, there is a period during which a functional gene co-occurs in both cellular compartments. Such a 'transition stage' has been documented for *cox2* in legumes and *atp9* in *Neurospora* (reviewed in Adams and Palmer

Figure 1.5 Steps in the successful transfer of mitochondrial genes to the nucleus including migration of a mitochondrial sequence to the nucleus, integration into the nuclear genome and acquisition of regulatory sequences (which in this case, is shown to include an N-terminal targeting peptide [yellow block]), and the loss of the mitochondrial copy.

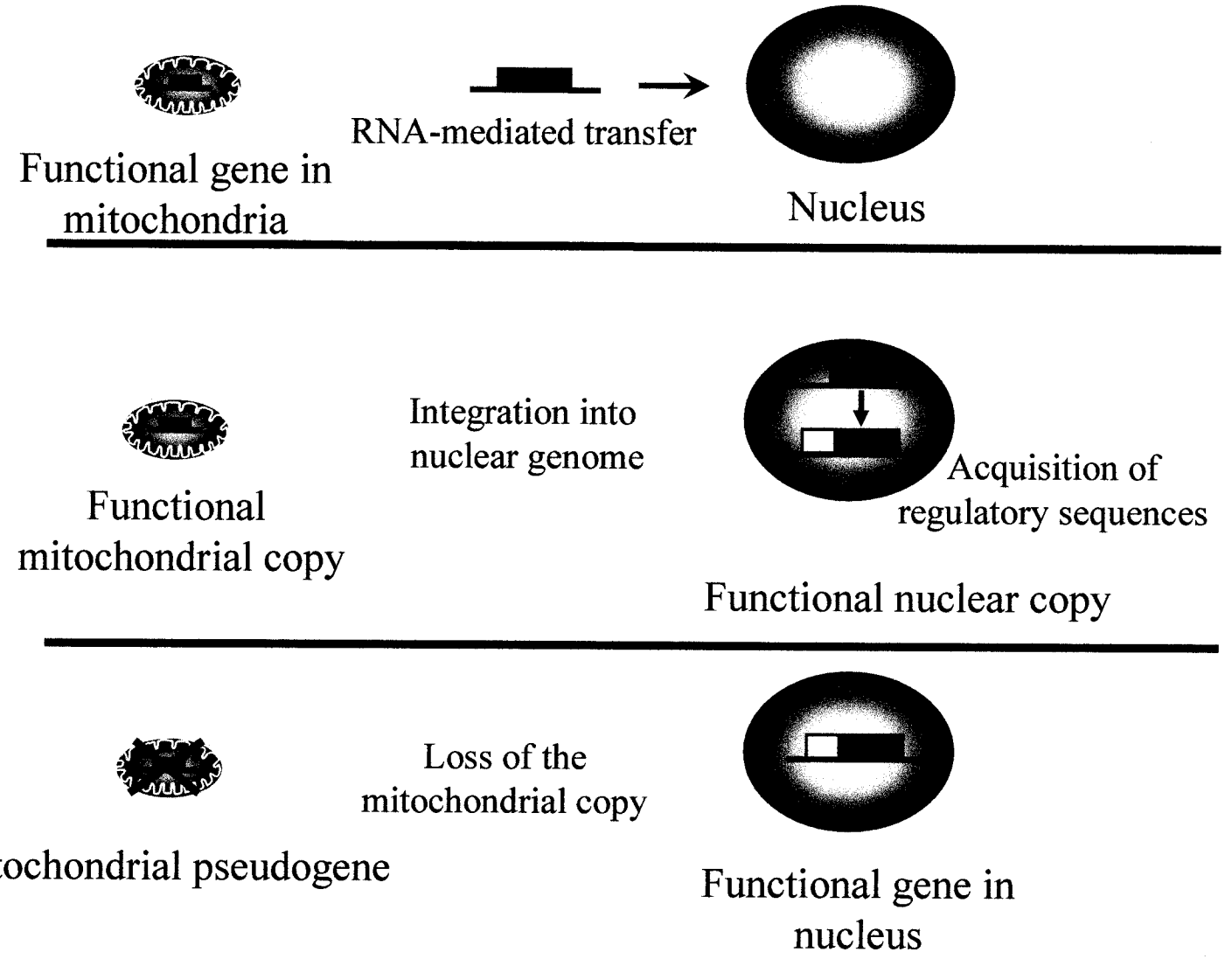

Functional gene in mitochondria
RNA-mediated transfer
Nucleus
Functional mitochondrial copy
Integration into nuclear genome
Acquisition of regulatory sequences
Functional nuclear copy
Mitochondrial pseudogene
Loss of the mitochondrial copy
Functional gene in nucleus

Table 1.2 Reported cases of mitochondrial ribosomal protein genes that have been transferred to the nucleus in flowering plants.

GENE	SPECIES	N-TERMINAL EXTENSION	COMMENTS [1]	REFERENCE
rps2	*Populus*	NO	Introns in 3' UTR	Choi et al. 2006
rps10	*Fuchsia*	YES	N-terminal extension similar to that encoded by *hsp70*. 2 introns in sequence encoding N-terminal extension.	Adams et al. 2000
rps10	carrot	YES	N-terminal extension similar to that encoded by *hsp22*. 1 intron in sequence encoding N-terminal extension	Adams et al. 2000
rps10	*Arabidopsis*	YES (U) [2]	Intron between sequences encoding S10 and N-terminal extension	Wischmann and Schuster 1995
rps10	lettuce	YES	In-frame insertion within non-mitochondrial metalloprotease. N-terminal part not cleaved after mitochondrial import	Adams et al. 2000
rps10	spinach, *Oxalis*, maize	NO	In maize, intron in the 5' UTR	Adams et al. 2000
rps10	rice	NO	2 nuclear copies. 5' UTR is similar to that in several other rice genes of various functions. Intron in 5' UTR	Kubo et al. 2000
rps10	*Populus*	YES (U)	2 Introns in sequence encoding N-terminal extension	Choi et al. 2006
rps11	rice	YES	2 nuclear copies; N-terminal extensions acquired from duplicated *atpB* and *coxVb* sequences. Intron in sequences encoding N-terminal extension	Kadowaki et al. 1996
rps11	*Populus*	YES (U)	Intron in sequence encoding N-terminal extension	Choi et al. 2006
rps12	*Oenothera*	YES (U)	No intron	Grohman et al. 1992
rps14	*Arabidopsis*	YES (U)	Intron in 3' UTR	Figueroa et al. 1999
rps14	maize	YES	shares N-terminal extension with *sdh2* via alternative splicing	Figueroa et al. 1999,
rps14	rice	YES	shares N-terminal extension with *sdh2* via alternative splicing	Kubo et al. 1999
rps14	wheat	YES	shares N-terminal extension with *sdh2* via alternative splicing	Sandoval et al. 2004
rps14	*Populus*	NO	Introns in 3' UTR	Choi et al. 2006
rps19	*Arabidopsis*	YES	N-terminal extension similar to RNP-CS type RNA binding protein. 3 introns in sequence encoding RNP-CS motif.	Sánchez et al. 1996
rps19	maize, barley, wheat	YES	N-terminal extension similar to that encoded by *hsp70*	Fallahi et al. 2005
rps19	*Populus*	YES	N-terminal extension similar to RNA-binding protein. Intron separating sequences encoding S10 and extension.	Choi et al. 2006
5' *rpl2*[3]	soybean	NO	No introns	Adams et al. 2001
3' *rpl2*	*Arabidopsis*, cotton, tomato, soybean	YES (U)	Intron in sequence encoding N-terminal extension.	Adams et al. 2001
3'*rpl2*	*Populus*	YES (U)	Intron in sequence encoding N-terminal extension.	Choi et al. 2006
rpl2 (entire)	maize	YES (U)	No introns	Adams et al. 2001
rpl5	maize	YES (U)	Intron separating *rpl5* from sequence encoding N-terminal extension.	Sandoval et al. 2004
rpl5	wheat	YES	Intron separating *rpl5* from sequence encoding N-terminal extension.	Sandoval et al. 2004

[1] Introns refer to spliceosomal introns acquired in the nucleus
[2] (U) = origin of N-terminal extension is unknown
[3] *rpl2* was split into 2 genes in early eudicot evolution (cf. Adams et al. 2001)

2003), and more recently for *rpl5* in wheat (Sandoval et al. 2004) and *sdh4* in *Populus* (Choi et al. 2006). For *cox2* it was inferred that following the activation of the nuclear-located copy in legumes, in some of the subsequently-diverging lineages it was the nuclear-located copy that was inactivated, illustrating that it is not always the mitochondrial copy that becomes lost (Adams et al. 1999). Hypotheses have been proposed that predict selective pressures that promote gene transfer (reviewed in Adams and Palmer 2003). For example, the hypothesis of Muller's ratchet predicts that nuclear location is favoured because of the possible accumulation of detrimental mutations in asexual mitochondrial genomes. As Adams and Palmer (2003) noted, this is unlikely to be a major factor in plants due to the low substitution rate in their mitochondria. Based largely on this, they favour the hypothesis of Blanchard and Lynch (2000) that states that nuclear-location allows for the fixation of beneficial mutations by recombination between homologous chromosomes.

1.4.3 Loss of the mitochondrial copy following translocation

Following the transition stage, transfer is complete with the inactivation of the mitochondrial copy or alternatively, transfer can fail with the inactivation of the nuclear copy. Differences between the nuclear and mitochondrial genomes likely influence how this inactivation occurs. For example, in the nucleus, where the rate of point mutation is relatively high, the sequence may acquire a nonsense or frame-shift mutation, or even accumulate deleterious missense mutations. In the mitochondria, on the other hand, the rate of point mutation is relatively low (Lynch et al. 2006), but the frequency of rearrangement is high, and therefore a sequence might be expected to be truncated, fragmented or deleted. However, a closely linked gene that is under functional constraint may restrict recombination events in the region of the mitochondrial copy, especially if the genes are co-transcribed and share expression signals (see Fig. 1.2). A striking example is that of *ψrps14* in the mitochondria of grasses, in which its transcription is estimated to have been maintained for at least 50 million years possibly due to its location immediately (1 nt) downstream of functional *rpl5* (Ong and Palmer 2006).

1.5 The S1 ribosomal protein gene in flowering plants

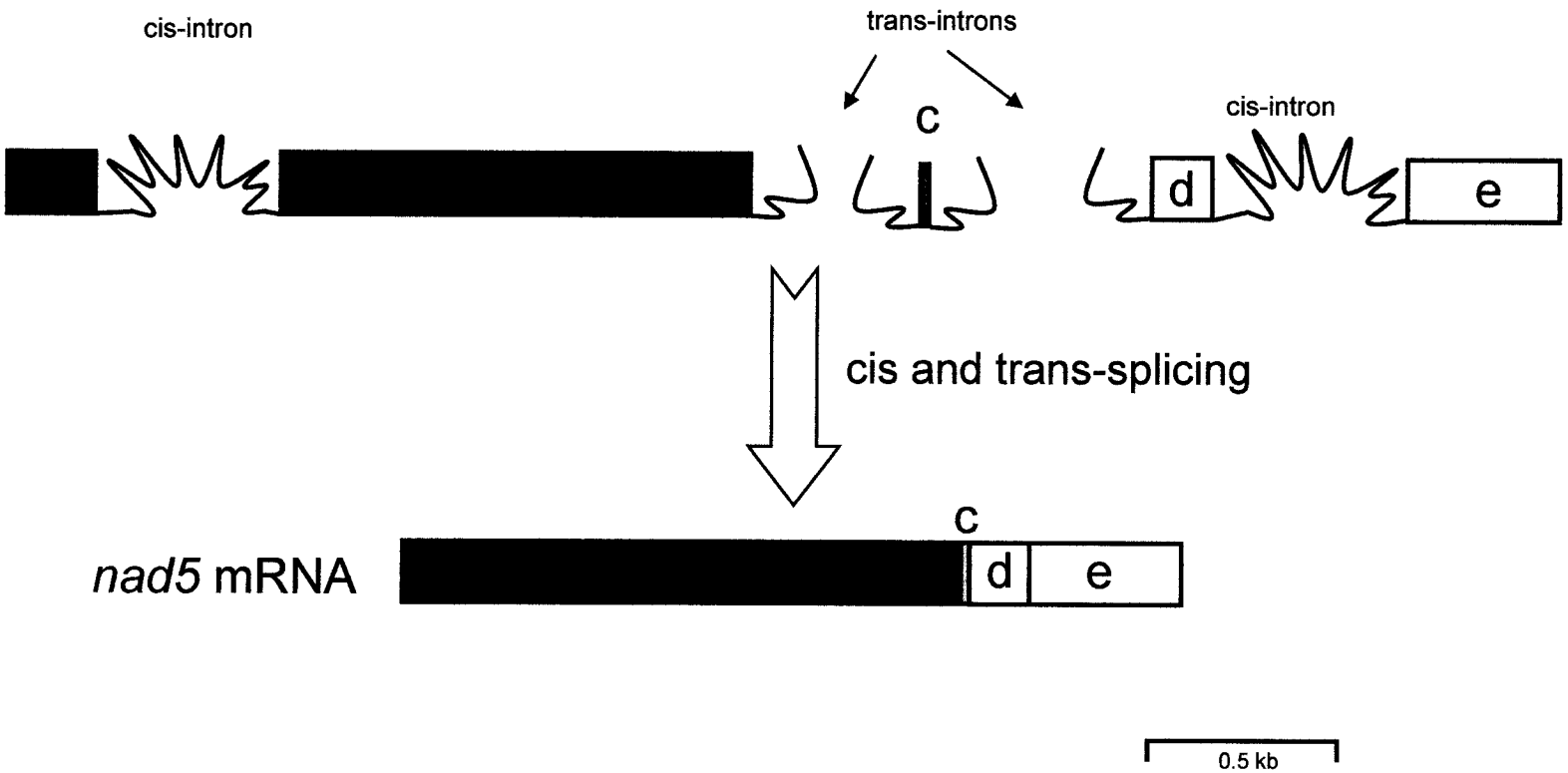
The high number of inferred losses of the *rps1* gene from flowering plant mitochondria (33 times, cf. Adams et al. 2002b) suggested that it would be a good model gene for studies of transfer to the nucleus. Using Southern hybridizations, Adams et al.

(2002b) detected *rps1* sequences in the mitochondria of the three legumes species that they examined, including *Medicago*. However, by examining EST databank sequences for barrel-medic (*Medicago truncatula*) they postulated that a nuclear copy exists. This raised the possibility of a recent transfer of *rps1* to the nucleus in legumes.

In flowering plants, detailed studies of *rps1* have only been done on wheat (Gonzalez et al. 1993) and *Oenothera* (Mundel and Schuster 1996), and in these species, *rps1* was found to be seemingly functional at the DNA level and transcribed. However, its functional status in *Oenothera* was questioned because no editing was observed in the *rps1* sequence despite the fact that there are four sites predicted. The completely-sequenced mitochondrial genomes of rice (Notsu et al. 2002), maize (Clifton et al. 2004) and tobacco (Sugiyama et al. 2005) contain an apparently functional *rps1* gene. In contrast, this gene is absent from the mitochondria of sugar beet (Kubo et al. 2000), *Brassica* (Handa 2003) and *Arabidopsis* (Unseld et al. 2005). In the course of my investigation into *rps1*, I found that this gene was located upstream of the first two exons (a and b) of trans-spliced *nad5* in the mitochondria of all legume species I examined. The five exons of *nad5* are located in three different genomic locations, and multiple *cis*- and *trans*-splicing events generate the *nad5* mRNA (Fig. 1.6, Knoop et al. 1997). In contrast to this stable linkage, the genomic environment upstream of *rps1* varies among most of the species I examined.

The ribosomal protein S1 of the small ribosomal subunit is important for translation in *E. coli*, and while present in most bacteria, is absent from archaea and eukaryotes (Lecompte et al. 2002). In *E. coli*, this protein has 557 amino acids, and approximately the first one-third contains the amino- (N-)terminal ribosome-binding domain. The other two-thirds contain a carboxy- (C-)terminal RNA-binding domain, which is composed of four repeating homologous stretches of ~75 amino acids (Fig. 1.4B) (Subramanian 1983). Among mitochondrial lineages, this protein is absent from yeast and animals (Gan et al. 2002), but can be found in protists (such as *Reclinomonas americana*), green algae (such as *Chara vulgaris*) and land plants (Lang et al. 1997, Turmel et al. 2003, Mundel and Schuster 1996, respectively). The RNA-binding domain present in the S1 ribosomal protein in *E. coli* is not found in the mitochondrial of flowering plants, *Reclinomonas*, *Chara*, the liverwort *Marchantia polymorpha* (Oda et al. 1992), or the moss *Physcomitrella patens* (Terasawa et al. 2007) (Fig. 1.4 B). The loss of its C-terminal RNA-binding domain in plant mitochondria

Figure 1.6 Schematic showing cis- and trans-splicing of the five exons of *nad5* (one of which, exon c, is only 22 nt in length) which are transcribed at three different genomic locations in the mitochondria of flowering plants cf. Knoop et al. (1997). Introns are not shown to scale.


cis-intron
trans-introns
c
cis-intron
d
e
cis and trans-splicing
c
nad5 mRNA
d
e
0.5 kb

raises questions as to the nature of initiator codon recognition. In *Marchantia* and *Physcomitrella* mitochondria, their S1 proteins have non-homologous (to each other) C-terminal regions that extend about 100 and 80 amino acids beyond the bacterial-type ribosome binding region. These do not appear to contain an RNA-binding domain. There is also a chloroplast-type S1 ribosomal protein, and interestingly, this contains a C-terminal RNA-binding domain, albeit much reduced compared to the bacterial-type (Fig. 1.4B) (Merendino et al. 2003, Franzetti et al. 1992).

1.6 Objectives

My primary research objectives focus on the status of the S1 ribosomal protein gene (*rps1*) in the mitochondria of legumes and its transfer to the nucleus as this gene appears to be frequently involved in transfer events in flowering plants. In addition, evidence points to a nuclear-located copy in *Medicago*. Objectives relating to this were met by addressing the following questions:

1) What is the status of rps1 in the mitochondria of members of the legume family?

There have been few detailed reports on the structure and expression of this gene in the mitochondria of flowering plants, and of these, the functional status of *rps1* in the mitochondria of *Oenothera* is unclear. Examination of *rps1* genes in the mitochondria of legumes including features of their RNA-level expression (e.g. transcript profiles and editing) will allow for a more comprehensive understanding of this gene in flowering plants as well as address issues regarding gene transfer.

2) What is the origin of regulatory sequences of transferred rps1 in the Medicago lineage of the legume family?

Given the apparent limitations at various steps, the odds of successful mitochondrial gene transfer to the nucleus seem remote. Yet, in flowering plants, such events appear to be occurring at a relatively high frequency, which suggests there is a diversity of ways in which translocated genes can acquire the signals necessary for nuclear expression and mitochondrial targeting. The examination of nuclear-located *rps1* sequences in legumes can further our knowledge about this step in the transfer process, and the *rps1* gene is an interesting model for this type of study as it appears to be among the more dynamic with

respect to nuclear relocation, perhaps pointing to inherent features within its sequence that facilitate such events.

3) What is the fate of the redundant mitochondrial rps1 copies in legumes?

Mitochondrial gene transfer is complete (and 'fixed') with inactivation of the mitochondrial copy, and thus the chances of successful transfer can be influenced by events impacting on the redundant mitochondrial copy. The examination of *rps1* pseudogenes in the mitochondria of legumes that contain a functional nuclear copy can shed light on how mitochondrial copies of transferred genes are likely (or alternatively, unlikely) to become inactivated.

During the course of my investigation into the status of rps1 in legumes, I discovered that the regions upstream of mitochondrial-located rps1 are different among even closely-related species. In addition to my primary objectives, I was able to address the following questions:

4) How does mitochondrial RNA-level regulation vary among closely-related species?

Plant mitochondria have genomes that frequently undergo rearrangement events and this raises questions about how their gene expression systems function in light of this. Transcript analyses were conducted for genes within the *rps1* locus in the mitochondria of closely-related legume species to evaluate how transcript features vary with the differing upstream genomic environments.

5) Are there features consistent with RNA-level regulatory signals, including those involved in initiation codon recognition in sequences preceding plant mitochondrial protein-coding genes?

The variation upstream of *rps1* in the mitochondria of various legumes raised questions about the nature of sequences preceding a broader range of flowering plant mitochondrial protein coding genes. To date, the mitochondrial genomes of seven flowering plants have been completely sequenced (namely maize, rice, wheat, sugar beet, tobacco, *Arabidopsis* and *Brassica*), and this allows for a comprehensive comparative analysis of such sequences. Also, by focusing on regions very close to the initiation codon, it will be possible to gain insight into how mitochondrial start codon recognition is achieved in flowering plants.

CHAPTER 2: MATERIALS AND METHODS

This Chapter outlines the combined materials and methods used in Chapters 3 and 4.

2.1 DNA and RNA isolation

Mitochondrial DNAs and RNAs were isolated from etiolated 6-day seedlings of soybean (*Glycine max* cv. Maple Arrow), bean (*Phaseolus vulgaris* cv. Tender Green), pea (*Pisum sativum* cv. Sugar Snap), alfalfa (*Medicago sativa*) and sweet clover (*Melilotus officinalis*) using procedures as previously described (Subramanian et al., 2001). In the case of rest harrow (*Ononis spinosa*), a limited seed supply made growing etiolated seedlings unfeasible, therefore mitochondrial DNA and RNA were isolated from the leaves of plants that were greenhouse-grown for 1-3 months under conditions of 14-h days at 23°-25 °C.

Total DNA was isolated from 6-day seedlings of alfalfa and its close relatives, sweet clover and fenugreek (*Trigonella foenum-graecum*) as described by Ausubel et al. (1990) with modifications in that pulverized tissue was incubated in 0.1 M Tris-HCl pH 8, 0.1 M EDTA, 0.2 M NaCl, containing 100 µg/ml proteinase K and 1% sarkosyl at 55 °C for 2 hrs. Following centrifugation, DNA was precipitated with 0.6 volumes of isopropanol, and after 30 min at -20 °C, the centrifuged pellet was resuspended in 10 mM Tris-HCl pH 8, 1 mM EDTA. DNA was then re-precipitated using the CTAB method as described (Ausubel et al., 1990). Total RNA was isolated from 6-day alfalfa seedlings using Trizol® (Invitrogen, Carlsbad, CA) according to the manufacturer's specifications.

2.2 Cloning and sequencing of DNA and cDNA

DNA sequences were obtained from cloned restriction fragments using standard methods (Sambrook et al. 1989) and from cloned PCR products inserted into a pGEM®-T Easy vector (Promega, Madison, WI) following gel-purification using UltraClean™ 15 (MoBio Laboratories Inc., Carlsbad, CA). Automated sequencing was performed by the Ottawa Health Research Institute DNA sequencing facility and the McGill University/Genome Québec Centre. Nuclear *rps1* sequences in alfalfa were obtained using the gene-specific primers 5'-GCATCTACTTGAGTCGACTG-3' and 5'-ATCATTTGATATCCTCCGTC-3' from databank information for barrel-medic (AC119412). Intron sequences in alfalfa, sweet clover and fenugreek were obtained using the

intron-specific primer 5'-CTGGTGGGGTGATTGAGATAAAAGC-3' with the gene-specific primer 5'-AGTGGTTCACCGGCCACTAG-3'. Direct sequencing was done using Sequenase version 2.0 (US Biochemicals) and electrophoresed on 7% polyacrylamide sequencing gels. See Table 2.1 for the complete set of primers used in this study.

Sequence data were deposited in the NCBI databank under the following accession numbers: soybean mitochondrial *rps1* (EU161945), bean mitochondrial *rps1* (EU161946), pea mitochondrial *rps1* (EU161947), rest harrow mitochondrial *rps1* (EU161948), alfalfa mitochondrial *ψrps1* (EU161949), fenugreek mitochondrial *ψrps1* (EU161950), sweet clover mitochondrial *ψrps1* (EU161951), alfalfa nuclear *rps1* mRNA (EU161952), sweet clover nuclear *rps1* (EU161953) fenugreek nuclear *rps1* (EU161954), alfalfa nuclear *rps1* (EU161955). The designated *rps1* initiation codon in the mitochondria of legumes is the 5'-most located, in-frame ATG that is conserved among all legume species, as well as tobacco (NC_006581), wheat (AP008982), rice (BA000029) and maize (AY506529). Our sequence data for mitochondrial-located *rps1* in wheat, along with that of Ogihara et al. (2005) differ from Gonzalez et al. (1993) (X69205), and extend the coding sequence by 36 codons at the 5' end. This brings the initiation codon in wheat (and thus rice and maize) in accordance with that in legumes.

The mapping of the 5' end of alfalfa cytosolic *rps1* transcripts was carried out using the 5' RACE System Version 2.0 (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol using the gene specific primers 5'-ATCATTTGATATCCTCCGTC-3' and 5'-AGTGGTTCACCGGCCACTAG-3'. The 3' end was mapped using the gene-specific 5'-AAAGAACGAGCAGCTGCCAG-3' and oligo-d(T) as primers. Approximately 10 μg of alfalfa total RNA was used. The position of each end was supported by two independent clones.

2.3 Analysis of RNA editing

RT-PCR products were generated for the *rps1* region in rest harrow, pea, soybean and bean using a *nad5b*-specific primer 5'-GACCAAGCTACTTATGAATG-3' in conjunction with a species-specific primer located 58-171 bp upstream of *rps1* (namely, rest harrow, 5'-TCCCTCTTTGTTTATGCTGC-3'; pea, 5'-AGCTTAACTCTTTCTACTGG-3'; soybean, 5'-CCGATTTGAATCGAACACTC-3'; bean, 5'-AGTACAATCCGACCGATGCC-3').

Table 2.1 Primers used in this study.

#	SEQUENCE 5'-3'	S/A[1]	GENE/REGION[2]
LB243	TCGACTCCAATAGATGCTCA	A	*rps1*
LB233	ATTATGTAGTGGAAACGCCT	S	*rps1*
LB 336	AGTGGTTCACCGGCCACTAG	A	*rps1*
LB 386	CCTGGTGGCTCGGTTGATTG	A	*rps1*
LB 244	AGTAGTGAAACCCGCGATGG	A	*rps1*
LB 451	TTGTTGTGAGAACGGAATGG	A	*rps1*
LB 452	GGTTAATGCTCTCAATGGTG	A	*rps1*
LB 453	ATCTGCCGCTGTTAGAACAC	A	*rps1*
LB 367	GAAGTCACCGGATGAATTCG	S	*nad5* intron 1
LB 368	CAGAAGTGAATTACGAGTCG	A	*nad5* intron 1
LB 5	GACCAAGCTACTTATGAATG	A	*nad5b*[3]
LB 317	AGCTTAACTCTTTCTACTGG	S	pea *rps1* upstream[4]
LB 318	CCGATTTGAATCGAACACTC	S	soy *rps1* upstream[4]
LB 435	GCGATAGAGCTCTATTAAGG	A	bean *rps1* upstream
LB 436	AGTACAATCCGACCGATGCC	S	bean *rps1* upstream[4]
LB 471	GCTGAAGCTGAACCAGCTGG	A	rest harrow *rps1* upstream
LB 483	TTCTGGATCAACCAACCAGC	S	rest harrow *rps1* upstream
LB 484	TCCCTCTTTGTTTATGCTGC	S	rest harrow *rps1* upstream[4]
LB 69	GTCTATTAAGGAGAATTCCC	A	*nad4L*
LB 369	GGTTCAGGTAGCTCAGCTGG	S	alfalfa *trnF/P* intergenic
LB 295	AGCAGCTCGGTCTATGAGTC	S	alfalfa *rps1* upstream
LB 371	AAGGTGACAGGATTCGAACC	A	alfalfa *trnP/rps1* intergenic
LB 325	GCATCTACTTGAGCTGACTG	S	nuclear *rps1*
LB 326	ATCATTTGATATCCTCCGTC	A	nuclear *rps1*
LB 327	CGACGGCGGCGATTCTTTCT	S	nuclear *rps1* upstream
LB 491	CTGGTGGGGTGATTGAGATAAAAGC	S	nuclear *rps1* intron
LB 316	TATGTCGTGAGGGATGTGAC	A	*rps4*
LB 361	CGCATTCTCCGAAGATTGAG	S	*rps4*
LB 442	ATGCCAATGTTGGTGACTGG	S	*nad5b*
LB 443	GTATCCTACAAAGAGACTCC	A	*nad5b*

[1] Sense/Antisense

[2] Refers to mitochondrial copy unless otherwise specified.

[3] Used for all RT reactions

[4] Used for RT-PCR

Editing sites were detected by comparing two to four cloned RT-PCR sequences with the genomic sequence.

2.4 Northern blot analysis

Mitochondrial RNAs (approx. 5μg) were electrophoresed on 1.2% agarose/formaldehyde gels and after membrane transfer, hybridizations were performed with ^{32} P-end-labelled oligomer gene-specific probes (*rps1* 5'-TTGTTGTGAGAACGGAATGG-3', 5'-GGTTAATGCTCTCAATGGTG-3', 5'-CCTGGTGGCTCGGTTGATTG-3'; *nad5b* 5'-GACCAAGCTACTTATGAATG-3'; *nad4L* 5'-GTCTATTAAGGAGAATTCCC-3') and ^{32} P-random-labelled PCR products such as a 0.41 kb one containing *trnF* and *trnP* in alfalfa. Hybridization analyses were conducted using standard methods (Sambrook et al., 1989). Blots were reproduced using independently isolated mtRNA preparations. Where possible, hybridization experiments for a particular transcript region were done using both end- and random-labelled probes.

2.5 S1 nuclease protection assay

To determine the 5' terminus of *nad5* transcripts in alfalfa, ^{32}P end- or uniformly-labelled PCR products (~0.48 kb; LB337/LB419) were hybridized to 5-15 μg of alfalfa mtRNA, which was dissolved in 5 μl of water and 3 μl of 5X PIPES buffer (0.2 M PIPES pH 6.5, 5 mM EDTA, 2 M NaCl) and dried. The pellet was resuspended in 12 μl deionized formamide and 3 μl (~10 ng) of the denatured probe was added. Following incubation at 85 °C for 15 min, the mixture cooled to 47 °C and was left overnight to hybridize. The RNA/DNA hybrids were cooled to 43 °C over ~45 min and left on ice. 300 μl of cold S1 buffer (0.25 M NaCl, 30 mM NaOAc pH 5.5, 1 mM zinc acetate, 20 μg/ml denatured herring sperm DNA) was added. This mix was added to 0, 25, 50 and 100 U (or 0, 50, 200 and 400 U) of S1 nuclease (Invitrogen, Carlsbad, CA) and incubated for 30 min at 30 °C. The reaction was stopped by adding 25 μl stop solution (4M NH_4OAc, 50 mM EDTA, 0.05 μg/μl tRNA). The protected products were precipitated by adding 240 μl of 95% ethanol and incubating at -20 °C overnight. After centrifugation, the pellets were washed once with 95% ethanol and dried for ~10 min at 60 °C. The protected products were resuspended in 4 μl TE and 8 μl loading buffer. 6 μl were electrophoresed on a 7% polyacrylamide gel.

2.5.1 Preparation of end-labelled probes

Approximately 200 ng of PCR products were incubated at 37 °C for 45 min with 40 μCi of γ-^{32} P-ATP (3000 Ci/mM, Amersham) and 5 units of T4 polynucleotide kinase in 1x kinase buffer (50 mM Tris, pH 9.5, 10mM $MgCl_2$, 5mM DTT) in a final volume of 25 μl. The reaction was stopped by adding 25 μl TE (10 mM Tris, pH 7.5, 1 mM EDTA), and the reaction products were spun through a Sephadex G-50 column.

2.5.2 Preparation of uniformly-labelled probes

Uniformly-labelled PCR products were generated by amplifying 2-3 ng of plasmid DNA containing the alfalfa mitochondrial *rps1-nad5ab* region in a reaction containing 1X PCR buffer (Invitrogen), 1.5 μl of 50 mM $MgCl_2$, 5.0 μl of 0.1 mM dNTPs (note: dATP, dGTP, dCTP and dTTP), 100 ng each of primers LB337 and LB419, 1.5 μCi α-^{32}P-dATP and 0.5 units of Taq polymerase (Invitrogen). Amplification conditions were 1 cycle at 94 °C for 10 min, 30 cycles of 94 °C for 30 sec, 52 °C for 30 sec, 72 °C for 1 min, followed by 1 cycle at 72 °C for 10 min. The amplification products were spun through a Sephadex G-50 column.

2.6 Primer extension analysis

To determine the 5' terminus of *rps1* transcripts in pea and *nad5* transcripts in alfalfa, 5-10 μg of mtRNA was dissolved in 6 μl low TE with 5 μl (~10 ng) of ^{32}P end-labelled primer (LB243, pea; LB473, alfalfa), and incubated at 70 °C for 10 min. After adding 4 μl 5X first strand buffer (Invitrogen, Carlsbad, CA), 2 μl 100 mM DTT, 1 μl 10 mM dNTPs and 1 μl (40 U) RNasin® (Promega, Madison, WI), the mixture was incubated for 37 °C for 2 min, and 1 μl (200 U) M-MLV Reverse Transcriptase was added. The reaction was carried out for 1.5 hrs at 37 °C, and stopped by adding 10 μl of loading dye. 5-10 μl were electrophoresed on a 7% polyacrylamide gel.

CHAPTER 3: STATUS OF GENES ENCODING THE MITOCHONDRIAL S1 RIBOSOMAL PROTEIN IN CLOSELY-RELATED LEGUMES

Thomas Hazle and Linda Bonen

3.0 Rationale

A survey of mitochondrial genomes using Southern hybridization (Adams et al. 2002b) suggested that *rps1* has undergone many independent transfers to the nucleus during flowering plant evolution, including one in the *Medicago* lineage of the legume family. Therefore, I examined *rps1* sequences in the mitochondria and nucleus of closely-related legume species to better characterize this gene in flowering plants, as well as investigate the ways in which successful transfer of this mitochondrial gene to the nucleus has occurred.

This chapter has been accepted in the journal *Gene* pending minor modifications. Note that Materials and Methods for this chapter have been combined with those for Chapter 4, and appear in Chapter 2.

Addendum 12 December 2007: this chapter has been published in the journal *Gene.* Thomas Hazle and Linda Bonen, **Status of genes encoding the mitochondrial S1 ribosomal protein in closely-related legumes.** Gene 2007 405:108-16

3.1 Abstract

The *rps1* gene, which encodes ribosomal protein S1 of the mitochondrial ribosome in flowering plants, is located in the mitochondrion of some but not all species, and this is assumed to reflect multiple gene transfers to the nucleus. We investigated its status in legumes and found that in alfalfa, sweet clover and fenugreek, the mitochondrial-located *rps1* is a pseudogene, in contrast to intact, transcribed and edited *rps1* genes in the mitochondria of rest harrow, pea, soybean and bean. Among these lineages, the genomic environment upstream of (*ψ*)*rps1* differs, and this contrasts with a stable downstream linkage with the first two exons of the trans-split *nad5* gene. Consequently, the *rps1* transcript profiles differ for each of these closely-related species, and typically do not include monocistronic *rps1* or *nad5* mRNAs. In alfalfa, sweet clover and fenugreek, the functional *rps1* gene is located in the nucleus and it is still flanked by residual non-coding mitochondrial sequences. Notably, the upstream ones provide part of the 5' UTR as well as

the 3' splice site of an intron preceding *rps1*. This exploitation of non-coding mitochondrial sequences in nuclear gene activation adds to a growing list of mechanisms by which successful transfer of mitochondrial genes is achieved.

3.2 Introduction

The large and recombinogenic mitochondrial genomes of flowering plants contain about 35 protein coding genes, mostly specifying respiratory chain components and ribosomal proteins (reviewed in Knoop, 2004). The gene content among lineages varies due to ongoing gene transfer to the nucleus, and this usually involves ribosomal protein genes (reviewed in Adams and Palmer 2003, Bonen 2006). The *rps1* gene, which encodes ribosomal protein S1 of the small ribosomal subunit is regarded as being among the most dynamic, as it has been inferred to have been independently lost from the mitochondrion 33 times during angiosperm evolution, most likely reflecting relocation to the nucleus in most of those lineages (Adams et al., 2002b). In *E. coli*, the S1 ribosomal protein is involved in the recognition and binding of mRNAs by the 30S ribosomal subunit, and although it is believed to be important in translation initiation in most bacteria (reviewed in Marintchev and Wagner, 2005), it is absent from archaeal and cytosolic ribosomes (reviewed in Lecompte et al., 2002). Relative to the bacterial-type, the mitochondrial S1 ribosomal protein in both vascular and non-vascular land plants has retained only the amino-terminal one-third of its sequence (cf. Gonzalez et al., 1993; Mundel and Schuster, 1996), which contains only the ribosome binding domain, and in the non-vascular plants *Marchantia polymorpha* and *Physcomitrella patens* they have unrelated C-terminal extensions (of about 60 and 100 amino acids respectively; Oda et al., 1992; Terasawa et al., 2007) that lack evident RNA binding features. This, along with the apparent absence of base-pairing between small subunit rRNA and mRNA, points to a shift away from a classical bacterial-type (Shine-Dalgarno) initiation codon recognition (Hazle and Bonen, 2007). Notably, the mitochondrial-encoded S1 ribosomal protein in the protist *Reclinomonas americana* is also missing the bacterial-type RNA binding domain, yet potential Shine-Dalgarno sequences have been observed for many of its mitochondrial mRNAs (Lang et al., 1997; Hazle and Bonen, 2007).

Relatively little has been reported about the structure or expression of mitochondrial *rps1* genes in flowering plants. Early on, it was examined in wheat (Gonzalez et al., 1993)

and *Oenothera* (Mundel and Schuster, 1996) mitochondria, and in the latter species, its functionality was questioned because of the absence of editing at predicted sites. More recently, completely-sequenced mitochondrial genomes have revealed its presence in rice (Notsu et al., 2002), maize (Clifton et al., 2004) and tobacco (Sugiyama et al., 2005), whereas it is absent in *Arabidopsis* (Unseld et al., 1997), sugar beet (Kubo et al., 2000), and *Brassica* (Handa, 2003). Interestingly, in the completely-sequenced nuclear genome of *Arabidopsis*, no mitochondrial-type sequence is apparent (Arabidopsis Genome Initiative, 2000). This suggests that the mitochondrial *rps1* has been functionally replaced, the most likely candidate being the nuclear-located chloroplast-type *rps1* (NM_180572) or a divergent duplicated chloroplast copy (NP_177317), somewhat analogous to what has been observed for mitochondrial *rps13* (Adams et al., 2002a).

We have now examined the status of *rps1* in the legume species alfalfa, sweet clover, fenugreek, rest harrow, pea, soybean and bean, all of which belong to the Papilionoideae sub-family, and they most recently shared a common ancestor about 55 Mya (Lavin et al., 2005). Of these, the first four are most closely-related, and diverged from the pea lineage approximately 25 Mya. Our study was in part prompted by an earlier survey of mitochondrial gene sequences in diverse angiosperms, in which Adams et al. (2002b) detected *rps1* sequences in the mitochondrion of *Medicago* using Southern hybridization, and in addition by examining EST databank entries, they postulated that a nuclear-located copy is present in barrel-medic (*Medicago truncatula*), which is congeneric with alfalfa (*Medicago sativa*). This raised the possibility of a recent transfer of *rps1* to the nucleus. We have found that in the mitochondria of alfalfa, sweet clover and fenugreek *rps1* is a pseudogene, and a functional copy is located in the nucleus from a seemingly recent transfer event. In the other legume species that we examined, the mitochondrial-located *rps1* appears functional, and this locus exhibits a history of upstream rearrangements in contrast to a stable downstream linkage with the first two exons of the *nad5* gene.

3.3 Results

3.3.1 A functional *rps1* gene in the mitochondria of rest harrow, pea, soybean and bean

Our examination of genomic sequence information in the NCBI databank (AC145156) suggested that in the mitochondrion of barrel-medic *rps1* is a pseudogene due to two nonsense mutations, and this would be consistent with the proposal of Adams et al. (2002b) that there is a functional *rps1* copy in the nucleus of that species. To assess its status in the mitochondria of closely-related legumes, namely alfalfa, rest harrow, pea, soybean and bean (whose phylogenetic relationships are shown in Fig.3.1), we sequenced cloned mitochondrial DNA restriction fragments or PCR-generated products of *rps1*-containing regions, and found a potentially functional mitochondrial gene in the latter four species and a pseudogene in alfalfa. Amino acid sequence alignments of the functional copies are shown in Figure 3.2A along with those for wheat and tobacco, and the *rps1* pseudogene in alfalfa is discussed below. The deduced S1 ribosomal protein sequence consists of 205 (pea and rest harrow) or 206 (soybean and bean) amino acids, and as in wheat and *Oenothera*, the predicted protein in these legumes lacks the carboxy-terminal RNA-binding domain present in *E. coli* (cf. Gonzalez et al., 1993; Mundel and Schuster, 1996). Among these legume species, the amino acid sequence identity ranges from 95-99%, as expected from the low rate of nucleotide substitution seen in plant mitochondria (Wolfe et al., 1987). Interestingly, although bean and soybean are closer relatives than are pea and rest harrow (~19 Mya vs. ~25 Mya divergence times, cf. Lavin et al., 2005), they have more amino acid substitutions (3 vs. 1) and indels (3 vs. 0). This appears to reflect differences in evolutionary rates for *rps1* among legume lineages. It is worth noting that in the mitochondria of the legumes examined, *rps1* is present as a single-copy as determined by Southern analysis, with the exception that in bean there is an additional pseudogene copy, which contains a 129 bp deletion and is truncated by 162 bp at the 3' end (data not shown).

In all four legume species having an intact mitochondrial-located *rps1*, editing was observed in their transcripts (Fig.3.2), while in alfalfa mitochondria, *rps1* transcripts were not detected by northern analyses or RT-PCR (see below). The number of editing sites within the coding region ranged from three in bean to one in pea (Fig.3.2B, black dots) and of these, one common silent position was observed to be edited in bean and rest harrow (boxed in

Figure 3.1 Schematic showing mitochondrial *rps1* gene and neighbouring regions in legumes and other selected plants. This includes eight legumes (alfalfa, barrel-medic, sweet clover, fenugreek, rest harrow, pea, soybean, and bean), as well as tobacco and three cereals (wheat, rice and maize). Note that the *ψrps1* sequence in bean is not represented here. Genes are represented by blocks in black (*rps1*), white (*Ψrps1*) and grey (genes flanking *rps1*). The estimated time of *rps1* transfer to the nucleus is delimited by wide grey bar on the tree. Phylogenetic relationships are according to Lavin et al. (2005) for the legumes, and Gaut (2002) for the cereals.

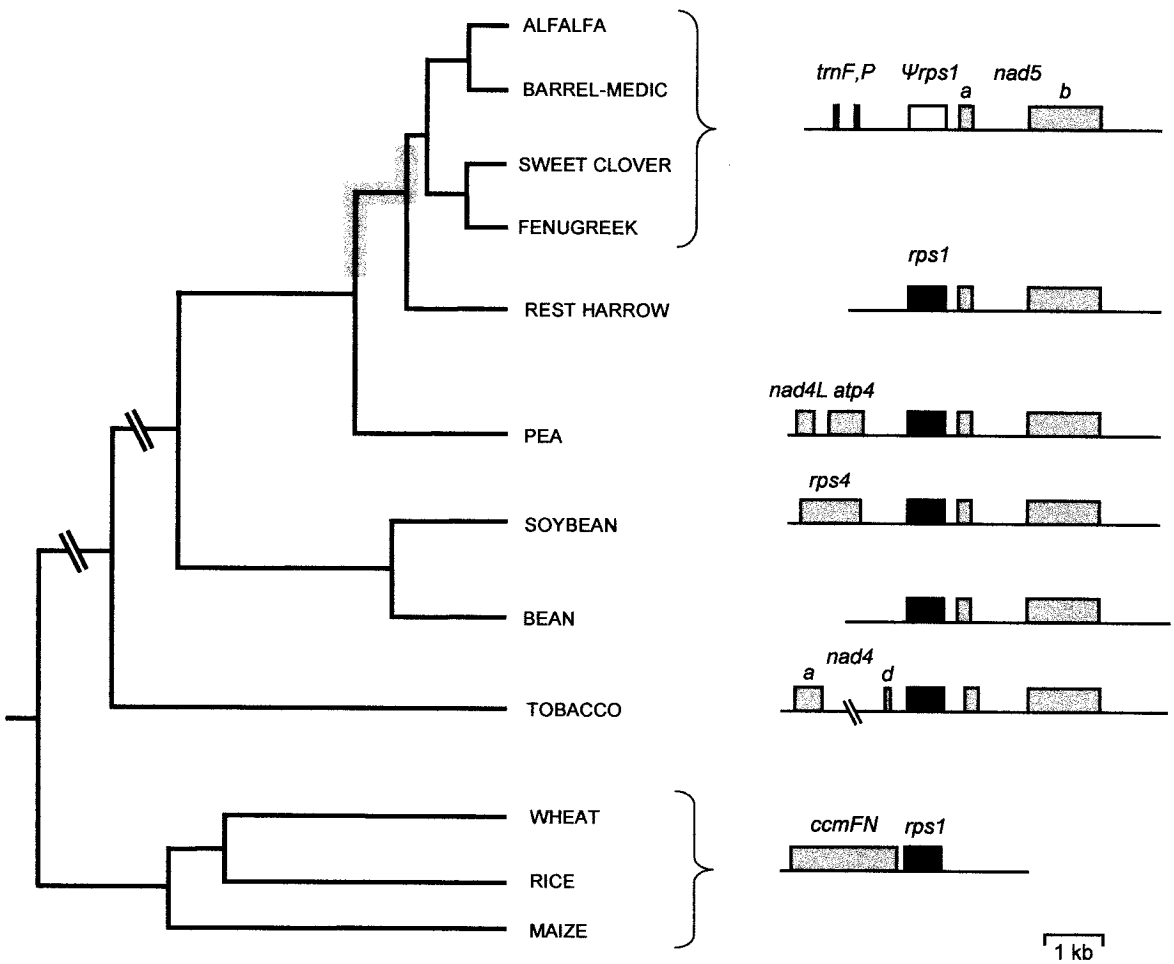

ALFALFA
BARREL-MEDIC
SWEET CLOVER
FENUGREEK
REST HARROW
PEA
SOYBEAN
BEAN
TOBACCO
WHEAT
RICE
MAIZE
trnF,P
Ψrps1
a
nad5
b
rps1
nad4L atp4
rps4
nad4
a
d
ccmFN
rps1
1 kb

Fig.3.2A) but not in pea and soybean. However, editing was not detected at two predicted positions (Fig.3.2A, arrowheads), which are conserved leucine codons in wheat and tobacco (and genomically-encoded ones even in *Marchantia* mitochondria). Interestingly, for one of these codons the nuclear-located *rps1* in alfalfa also specifies leucine. However, it is unclear if this was due to an (edited) RNA-mediated translocated sequence (with editing of the mitochondrial site having since been lost) or substitutions following transfer. Editing was also observed in the 5' untranslated region (UTR) in bean, pea and rest harrow (underlined in Figure 3.2B), and such a phenomenon has not often been reported. The region very near the initiation codon differs due a bean-specific edit at position -3, as well as indels (Fig. 3.2B). This might be expected to impact on start codon recognition, but variation has been observed preceding homologous flowering plant mitochondrial protein-coding sequences (Hazle and Bonen, 2007).

3.3.2 Mitochondrial *rps1* transcript profiles are impacted by upstream DNA rearrangements among legumes

We conducted northern hybridization experiments using mitochondrial RNA isolated from 6-day seedlings (or mature leaves for rest harrow) to evaluate the transcript profiles of *rps1* and its neighbouring genes in these legumes (Fig.3.3A-E). In pea, rest harrow, bean and soybean, *rps1* transcript profiles are complex, and show lineage-specific differences reflecting each unique upstream genomic environment. Conversely, the downstream linkage of *rps1* with *nad5ab* is conserved not only among legumes, but also in tobacco, and this is consistent with this organization predating the divergence of the asterids and rosids (~125 Mya, cf. Wikström et al., 2001). In pea, soybean and alfalfa/sweet clover/fenugreek, *rps1* is located less than 1 kb downstream of known genes, namely *nad4L-atp4*, *rps4* and *trnF-trnP* respectively, whereas this is not the case in bean or rest harrow (Fig.3.1). The linkage of *rps1* to *trnF-trnP* in the mitochondria of sweet clover and fenugreek was determined by PCR analysis (data not shown). In contrast to the dynamic upstream environment in legumes, a *ccmFN-rps1* co-transcriptional unit is present in the three grasses wheat, rice and maize, which shared a common ancestor ~70 Mya (Gaut, 2002). The short intergenic region of ~135 bp for *ccmFN-rps1*, which is rather comparable to the length of the *rps1-nad5ab* intergenic region in legumes (~200 bp) may limit the frequency of mitochondrial DNA rearrangement

Figure 3.2 [A] Alignment of deduced amino acid sequences of mitochondrial-encoded S1 ribosomal protein in four legume species, as well as tobacco, wheat, and nuclear-encoded S1 in alfalfa. Amino acid sites altered by editing are inverse highlighted (shown as edited form), and those that undergo silent editing are boxed. In wheat, an edit generates a stop codon (asterisk; Gonzalez et al. 1993). Predicted sites not edited in legumes are indicated by arrowheads. [B] Editing sites shown as dots for four legumes in same order as in alignment below. Sites that show editing among legume species in the *rps1* coding sequence as well as 5' UTRs are shown by black dots, genomically-encoded uracil residues by grey dots, and unedited sites as white dots. Edited codons along with encoded amino acids are shown above the block with the editing site underlined. In the alignment below the block, edited sites are underlined and the initiation codon is in bold.

A

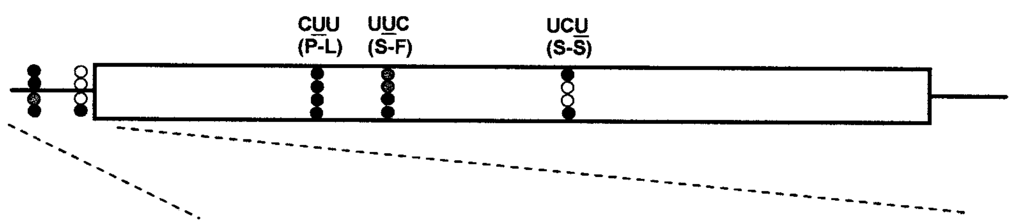
```
                    ▽                                       ▽
ALFALFA NUC       MSIYLSRLFPRCNSSSFLCSGKALQSEVLRLG--KETFLVDAGPGTPKNCTRDELTRVP
REST HARROW MT   MMSIYLSRSFPRCNSSLFLCSGKALQSEVLRLG--EEMFLVDAGPGTPRNCMQDELTGVP
PEA MT           MMSIYLSRSFPRCNSSLFLCSGKALQSEVLRLG--EEMFLVDAGPGTPRNCMQDELTGVP
SOYBEAN MT       MMSIYLSRSFPRSNSSLFLCSGKALQSEVLRLG--EEMFLVDAGPGTPRICMQDELTGVP
BEAN MT          MMSIYLSRSFPRSNSSLFLCSGKALQSEVLRLG--EEMFLVDAGPGTPRICMQDELTGVP
TOBACCO MT       MMSIYLSRLFPRSNSSFFLCSGNALQSEVLRLR--EEMFLVDAGLGTPRICMQDELTGVP
WHEAT MT         MMSIYWSRLFPRSNSSFLLCSGNALQSSVLRLRLREEMFLVDAGLGTPKICMQDELTGLP

ALFALFA NUC      INPATRFENKVGFLDRAAGETHIRKKNLERLFIDLVAGEPLIKERAAARFNDMAGSTDVV
REST HARROW MT   INRATRFENKVGFLDLVAGESLIKKKILERLFIDLVAGESLIKERAAARFNDLVGSTDVV
PEA MT           INRATRFENKVGFLDLVAGESLIKKKILERLFIDLVAGESLIKERAAARFNDLVGSTDVV
SOYBEAN MT       INRATRFENKVGFLDLVAGESLIKKKILERFFIDLVAGESLIKERAAARFNDLVGSTDVV
BEAN MT          INRATRFENKVGFLDLVAGESLIKKKILERLFIDLVAGESLIKERAAARFNDLVGSTDVV
TOBACCO MT       INRATRFENKVGFLDLVAGESLIKEQILERFFIDLVAGESLIKERAAARFNDLVGSTDVV
WHEAT MT         IKRATRFENKVGFKNVVAGESLIKKRIFERFFIDLVAGESLIKERAAARFNDFVGSLDVA

ALFALFA NUC      ADEPLLLLPRRFRQDRAWMKLNKIWRTNTKVKGFIIDKVR-GGYSVAIAGFIAYLPFRSH
REST HARROW MT   AGEPLLLLPRRFRQNRAWMKLNKIWRTNTKVKGFIIDKVK-GGYSVAIAGFITFLPFRSH
PEA MT           AGEPLLLLPRRFRQNRAWMKLNKIWRTNTKVKGFIIDKVK-GGYSVAIAGFITFLPFRSH
SOYBEAN MT       AGEPLLLLPRRFRQNLAWMELNKIWRTNTKVKGFIIDKVKKGGYSVAIAGFITFLPFRSH
BEAN MT          AGEP-LLLPRRFRQNLAWMKLNKIWRTNTKVKGFILDKVK-GGYSVAIAGFITFLPFRSH
TOBACCO MT       AGEPLLLLPRRFRQNRAWMELNKIWRTNTKVKGFIIEKVK-GGYSVAIAGFITFLPFR--
WHEAT MT         AGEP-LLLPQRFRQNRAWIELKKIWRTKKKVKGFIIKKIK-GGYLVAIAGFTTFLPFK--

ALFALFA NUC      SKRQRRRISNDQFTIESINPK--NKSIVVF
REST HARROW MT   NKRKKRRISNDQFTIESINPK--RTNIVVF
PEA MT           NKRKKRRISNDRFTIESINPK--RTNIVVF
SOYBEAN MT       NKRRRKKISNDRFTIESINPK--RTNIVVF
BEAN MT          NKRRRKKISNDRFTIESINPKSKRTNIVVF
TOBACCO MT       --RRRKRISNDRFTIENINPK--KTNIVVF
WHEAT MT         KALLKKRIANDRFTIDSINPK--RRDIVIIAAD*TRT
```

B




```
REST HARROW MT ...GAUGCUUUUGAUCAAUAGAACAGGAAGAGUAGGAAAAAAAAA--GCUU AUG AUG AGC...
PEA MT         ...GAUGCUUUUGAUCAAUAGAACAGGAAGAGGAGGAAAAAAAA---GCUU AUG AUG AGC...
SOYBEAN MT     ...UAUGUUUUUGAUCAAUAGAACAGGAAGAGGAGGAAAAAAAA---GCUU AUG AUG AGC...
BEAN MT        ...GAUGCUUUUUAUCAAUAGAACGGGAAGAGUAGGAAAAAAACAUAUUUU AUG AUG AGC...
```

events. The ~60 bp region upstream of *rps1* is well-conserved among mitochondrial copies and possesses a purine-rich stretch characteristic of sequences preceding a start codon (Hazle and Bonen, 2007). In soybean, this region (plus 27 bp of coding sequence) is also located 1.3 kb upstream of *atp1* (Chanut et al., 1993) and at two additional places in the tobacco mitochondrial genome (Sugiyama et al., 2005), and thus appears to be active in rearrangement events.

In the legume species with a functional *rps1* mitochondrial copy, it is co-transcribed with *nad5ab* and this contributes to the profile complexity as the five exons of *nad5* undergo multiple cis- and trans-splicing events to generate the *nad5* mRNA (reviewed in Knoop, 2004). In the case of pea and soybean, the large co-transcripts include *nad4L-atp4* and *rps4* respectively (Fig.3.3B, C asterisks). Notably, abundant monocistronic *rps1* mRNAs were only observed in rest harrow (Fig.3.3D lane 1, black arrowhead) and monocistronic *nad5* mRNAs were only seen in alfalfa (Fig.3.3E lane 3, grey arrowhead). However, in all four legumes except alfalfa, dicistronic *rps1-nad5* mRNAs were observed (Fig.3.3A; B, C lanes 2, 3; D open arrowhead). Transcripts representing intermediate stages of RNA processing are also evident. For example, a 5.4 kb transcript in pea contains *rps1* and *nad5ab* with intronic sequences (Fig.3.3C, lanes 2, 3; hybridizations using intron-specific probes, data not shown), yet without the upstream *nad4L-atp4* (which is present as a dicistronic 1.3 kb mRNA; Fig.3.3C lane 1), and this is consistent with an endonucleolytic cleavage event. Similarly, in soybean there are transcripts of about 1.4 kb containing *rps4* but not *rps1-nad5ab* (Fig.3.3B). For alfalfa, hybridization with a random-labelled probe containing *trnF* and *trnP* revealed the abundant mature tRNAs, but no precursor *trnF-trnP-rps1-nad5ab* co-transcripts (Fig.3.3E), nor were stable *ψrps1* transcripts evident despite an abundant *nad5* mRNA.

3.3.3 *rps1* is a pseudogene in the mitochondria of alfalfa, sweet clover and fenugreek

In alfalfa, the *ψrps1* sequence was seen to be identical to that in its congeneric species barrel-medic (therefore is a pseudogene), and their flanking sequences have a high identity (>99%) as expected. In contrast, *rps1* is functional in rest harrow mitochondria, and this prompted us to determine its status in the mitochondria of sweet clover and fenugreek (Fig.3.4) as the radiation of these species occurred more recently than did rest harrow. In

Figure 3.3 [A-E] Northern blot analysis of mitochondrial RNA from bean [A], soybean [B], pea [C], rest harrow [D] and alfalfa [E] hybridized with probes as indicated below each lane. Arrowheads indicate *rps1-nad5* dicistronic mRNAs (white), monocistronic *rps1* (black) and monocistronic *nad5* (grey) mRNAs. [B, C]. Precursor transcripts from pea containing *nad4L-atp4-rps1-nad5ab*, and from soybean containing *rps4-rps1-nad5ab* are illustrated by asterisks. Blot shown in panel E, lane 2 was exposed for a longer period than those in lanes 1 and 3. Note that gels were electrophoresed for different lengths of time. The *rps1* gene linkages for these legumes are shown in Figure 3.1.

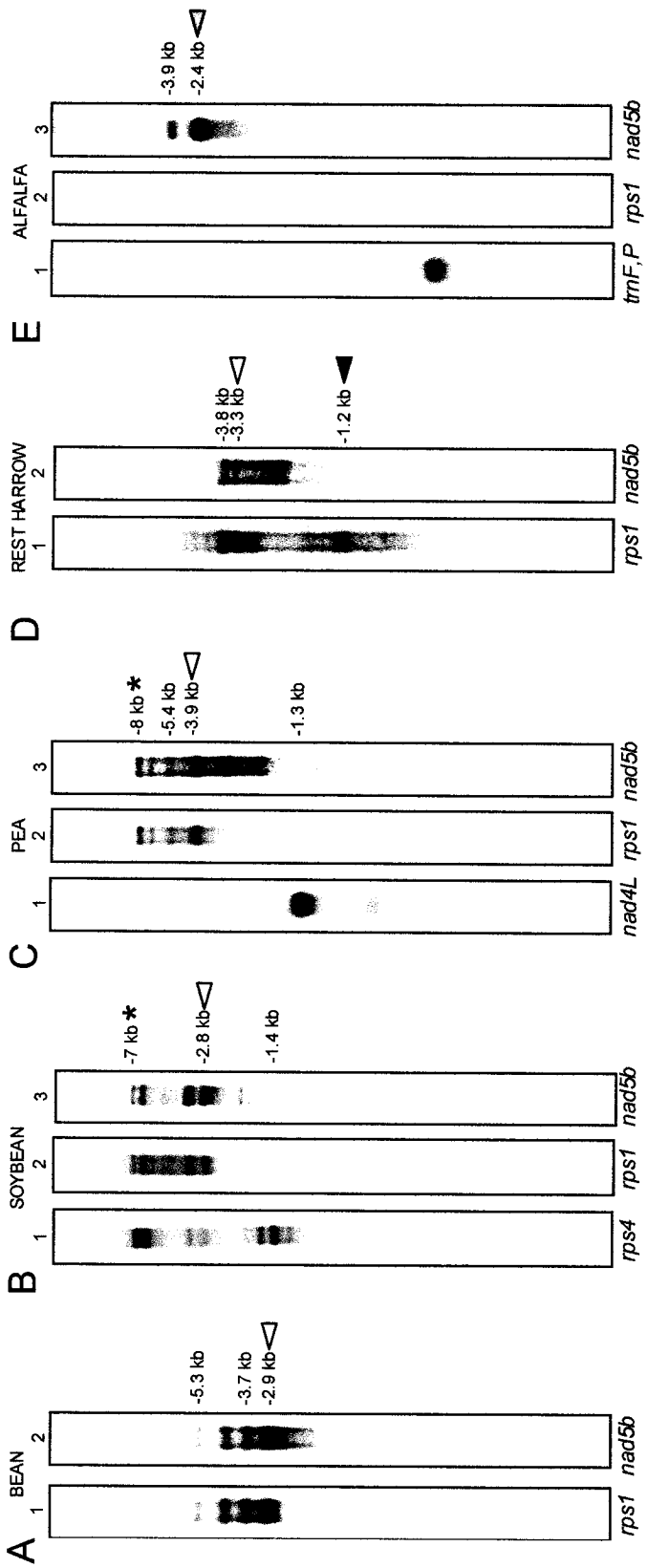
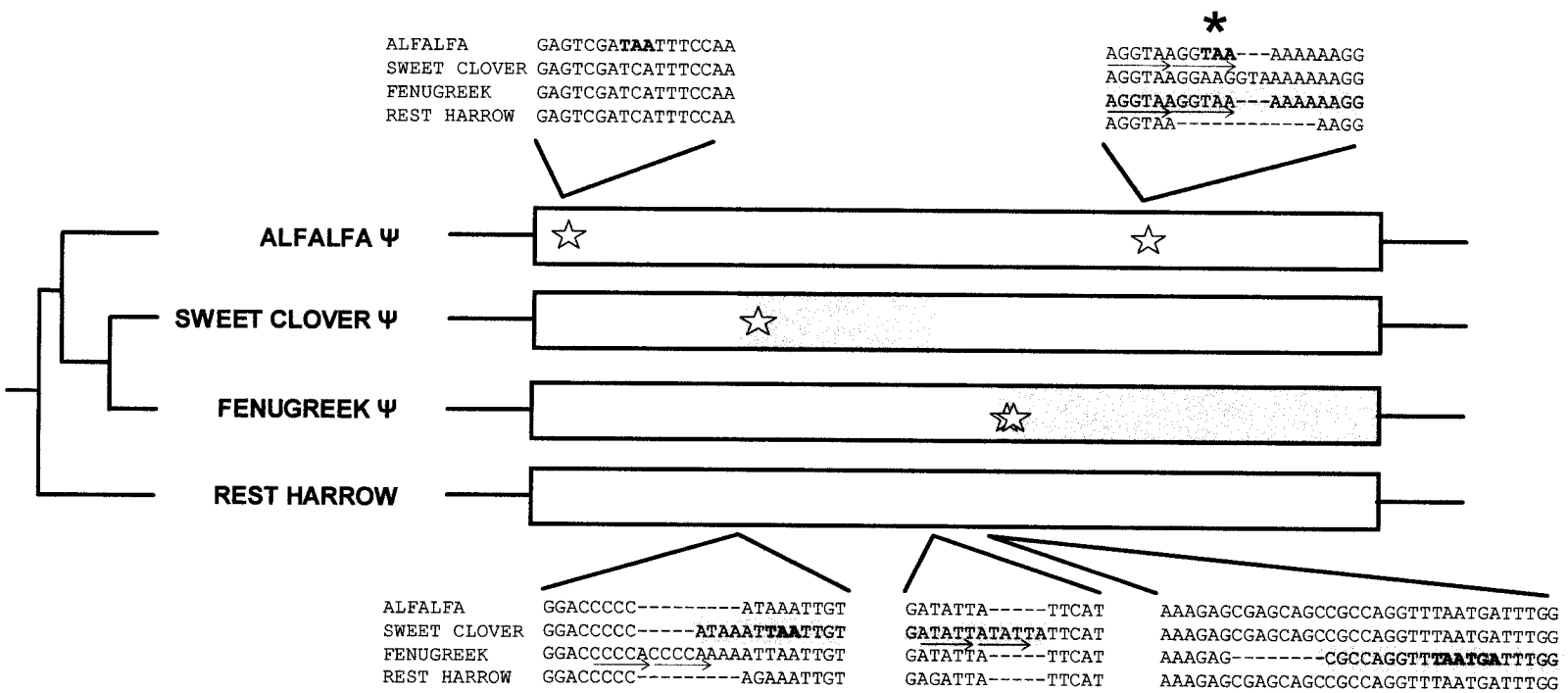

A
BEAN
1
2
rps1
nad5b
-5.3 kb
-3.7 kb
-2.9 kb
B
SOYBEAN
1
2
3
rps4
rps1
nad5b
-7 kb *
-2.8 kb
-1.4 kb
C
PEA
1
2
3
nad4L
rps1
nad5b
-8 kb *
-5.4 kb
-3.9 kb
-1.3 kb
D
REST HARROW
1
2
rps1
nad5b
-3.8 kb
-3.3 kb
-1.2 kb
E
ALFALFA
1
2
3
trnF,P
rps1
nad5b
-3.9 kb
-2.4 kb

Figure 3.4 Schematic of mitochondrial *rps1* pseudogenes in alfalfa, sweet clover and fenugreek, and functional *rps1* in rest harrow highlighting regions containing nonsense and frame-shift mutations. In-frame stop codons are indicated by stars and by bold in alignments. The frame-shifted sequences in sweet clover and fenugreek (shaded) contain multiple stop codons, of which the 5'-most ones are shown. Direct repeat-type indels are illustrated by arrows. Black asterisk indicates a stop codon within a shared insertion, which may have caused the initial inactivation. Phylogenetic relationships are as shown by Lavin et al. (2005).


ALFALFA GAGTCGATAATTTCCAA
SWEET CLOVER GAGTCGATCATTTCCAA
FENUGREEK GAGTCGATCATTTCCAA
REST HARROW GAGTCGATCATTTCCAA
*
AGGTAAGGTAA---AAAAAAGG
AGGTAAGGAAGGTAAAAAAAGG
AGGTAAGGTAA---AAAAAAGG
AGGTAA----------AAGG
ALFALFA Ψ
SWEET CLOVER Ψ
FENUGREEK Ψ
REST HARROW
ALFALFA GGACCCCC---------ATAAATTGT GATATTA-----TTCAT AAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGG
SWEET CLOVER GGACCCCC-----ATAAATTAATTGT GATATTATATTATTCAT AAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGG
FENUGREEK GGACCCCCACCCCAAAAATTAATTGT GATATTA-----TTCAT AAAGAG--------CGCCAGGTTTAATGATTTGG
REST HARROW GGACCCCC---------AGAAATTGT GAGATTA-----TTCAT AAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGG

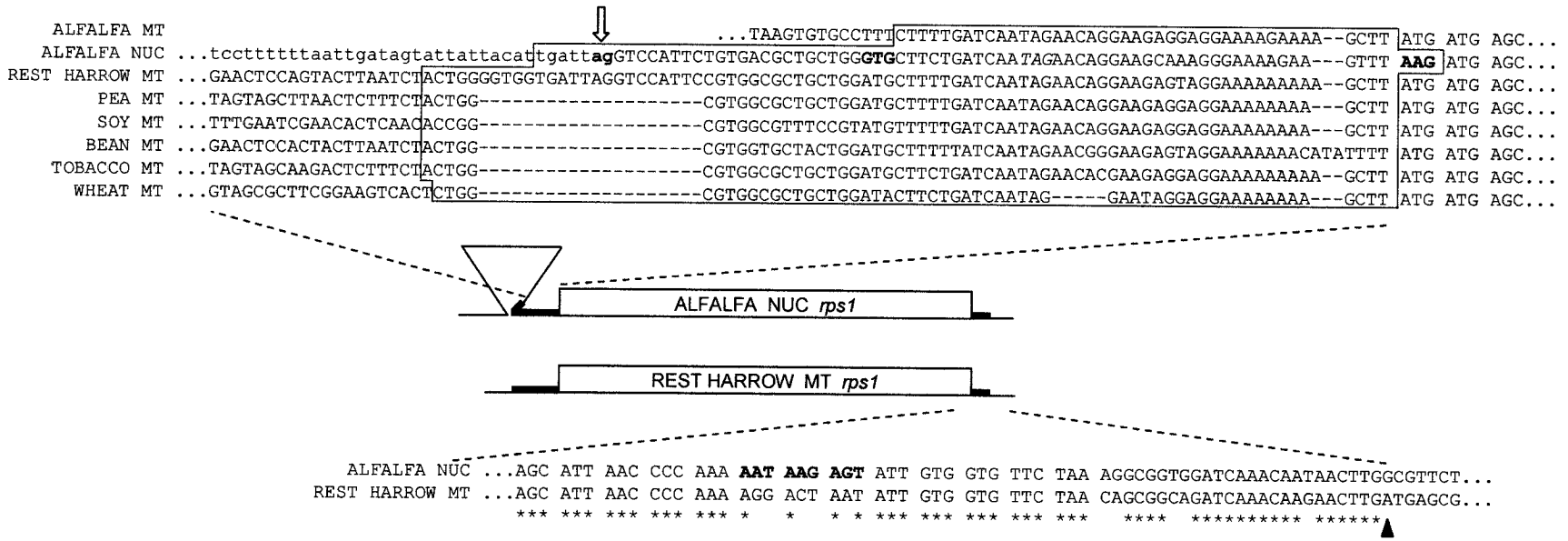
both sweet clover and fenugreek, the mitochondrial *rps1* sequences contain nonsense and frame-shift mutations (Fig.3.4 stars, shading) yet as with alfalfa and barrel-medic, the pseudogene sequences are not truncated and have a high nucleotide identity (93-95%) with rest harrow. Most of the pseudogene variation is the result of indel mutations that are at five locations (four of which involve frame-shifts and are shown in Figure 3.4) and vary from 5-12 bp in length. Most of the insertions contain short direct repeats (Fig.3.4, arrows). One insertion relative to pea and rest harrow is common to all pseudogene sequences (Fig.3.4, asterisk), and the in-frame stop codon within it (which is still in-frame in alfalfa) may well have been the initial inactivating mutation.

3.3.4 The functional *rps1* gene in alfalfa (and its close relatives) is located in the nucleus

Since *rps1* is a pseudogene in the mitochondria of alfalfa and closely-related species, we searched for a functional nuclear-located *rps1* using PCR strategies. The PCR/RT-PCR and RACE products obtained from alfalfa DNA and RNA were cloned and sequenced. We identified a functional nuclear *rps1* copy, which encodes a 204 amino acid S1 ribosomal protein, and has a transcript of 1031 nt (159 nt 5' UTR and 257 nt 3' UTR). The nuclear copy has the same stop codon as the functional mitochondrial version, but the coding sequence is shorter at the 5' end by one codon, as the ancestral AUG was lost through point mutation (Fig.3.5 shaded). An in-frame UAG triplet slightly upstream (Fig.3.5, italics) also indicates the absence of sequences encoding an amino-terminal targeting peptide, similarly noted for certain other recently-transferred genes, such as *rps10* in spinach, maize and *Oxalis* (Adams et al., 2000). Bonen and Calixte (2006) also found that approximately 25% of transferred mitochondrial ribosomal protein genes of bacterial origin appear to lack sequences that encode such targeting peptides. Indeed, the nuclear-encoded S1 ribosomal protein in alfalfa is predicted to inherently contain mitochondrial-targeting type features by Mitoprot (Claros and Vincens, 1996), TargetP 1.1 (Emanuelsson et al., 2007) and Predotar v.1.03 (Small et al., 2004) with probabilities of 0.96, 0.64 and 0.33 respectively. Similar observations were made for barrel-medic by Adams et al. (2002b) and Sugiyama et al. (2005) based on EST data.

We obtained genomic sequence information (AC119412) for the nuclear-located copy of *rps1* in barrel-medic from the NCBI databank and these data indicate the presence of

Figure 3.5 Comparison of sequences flanking *rps1* nuclear and mitochondrial copies among legumes. In the schematic, the intron is depicted by an open triangle and non-coding regions shared by nuclear and mitochondrial copies are shown by black bars. Nucleotide alignment of sequences preceding *rps1* above schematic show conserved region (boxed) and loss of mitochondrial-type ATG (shaded triplets). In alfalfa nuclear sequence, the intron sequence is in lower case with the 3' splice site shown in bold and indicated with an arrow, and in-frame TAG is shown in italics. Region at 3' end of *rps1* is shown below schematic and black arrowhead indicates break in homology between nuclear copy and mitochondrial copy in rest harrow. Shading indicates codons that contain nuclear *rps1*-specific nonsynonymous substitutions (also see Fig. 2). Note that sequence homology upstream of conserved block differs from phylogenetic relationships (cf. pea/tobacco vs. bean/rest harrow) and presumably reflects independent rearrangement events.


ALFALFA MT ...TAAGTGTGCCTTTCTTTTGATCAATAGAACAGGAAGAGGAGGAAAAGAAAA--GCTT ATG ATG AGC...
ALFALFA NUC ...tccttttttaattgatagtattattacattgattagGTCCATTCTGTGACGCTGCTGGGTGCTTCTGATCAATAGAACAGGAAGCAAAGGGAAAAGAA---GTTT AAG ATG AGC...
REST HARROW MT ...GAACTCCAGTACTTAATCTACTGGGGTGGTGATTAGGTCCATTCCGTGGCGCTGCTGGATGCTTTTGATCAATAGAACAGGAAGAGTAGGAAAAAAAAA--GCTT ATG ATG AGC...
PEA MT ...TAGTAGCTTAACTCTTTCTACTGG----------CGTGGCGCTGCTGGATGCTTTTGATCAATAGAACAGGAAGAGGAGGAAAAAAAA---GCTT ATG ATG AGC...
SOY MT ...TTTGAATCGAACACTCAACACCGG----------CGTGGCGTTTCCGTATGTTTTTGATCAATAGAACAGGAAGAGGAGGAAAAAAAA---GCTT ATG ATG AGC...
BEAN MT ...GAACTCCACTACTTAATCTACTGG----------CGTGGTGCTACTGGATGCTTTTTATCAATAGAACGGGAAGAGTAGGAAAAAAACATATTTT ATG ATG AGC...
TOBACCO MT ...TAGTAGCAAGACTCTTTCTACTGG----------CGTGGCGCTGCTGGATGCTTCTGATCAATAGAACACGAAGAGGAGGAAAAAAAAA--GCTT ATG ATG AGC...
WHEAT MT ...GTAGCGCTTCGGAAGTCACTCTGG----------CGTGGCGCTGCTGGATACTTCTGATCAATAG-----GAATAGGAGGAAAAAAAA---GCTT ATG ATG AGC...
ALFALFA NUC rps1
REST HARROW MT rps1
ALFALFA NUC ...AGC ATT AAC CCC AAA AAT AAG AGT ATT GTG GTG TTC TAA AGGCGGTGGATCAAACAATAACTTGGCGTTCT...
REST HARROW MT ...AGC ATT AAC CCC AAA AGG ACT AAT ATT GTG GTG TTC TAA CAGCGGCAGATCAAACAAGAACTTGATGAGCG...

a large intron of 2.7 kb (vs. typical length <150 nt; Wendel et al., 2002; Lorković et al., 2000) located 69 bp upstream of the *rps1* initiation codon, and this in turn is preceded by a non-coding exon of 91 bp. The *duf221* gene of unknown function is located approximately 2 kb upstream of the first *rps1* exon, and a hypothetical gene of 230 codons is located about 5 kb downstream of *rps1*. We determined the presence of a homologous intron at the same location in alfalfa, sweet clover and fenugreek from PCR-derived sequence analysis (data not shown), and this is consistent with a single transfer event of mitochondrial *rps1* to this nuclear genomic location.

3.3.5 Mitochondrial sequences preceding the nuclear-located copy of *rps1* provide a 3' splice site and part of the 5' UTR

In addition to the full length of the mitochondrial *rps1* coding sequence, the nuclear copy has retained flanking mitochondrial sequences which contribute to both the 5' and 3' UTRs. These residual sequences extend 76 bp upstream and 25 bp downstream, and the former shares the longest stretch of homology with the rest harrow mitochondrial counterpart (Fig.3.5) although all of legumes show a rather similar breakpoint in homology (Fig.3.5 boxed, see also Fig. 1). We found that in addition to contributing almost 70 nt to the *rps1* 5' UTR, the upstream mitochondrial-type sequence also provides a short stretch of the intron including the 3' splice site (Fig.3.5 bold with arrow), which conforms to the consensus sequence proposed for plant introns (cf. ugYAG/gu; reviewed in Lorković et al., 2000). Other features conferring expression (including a promoter) were presumably acquired in the nucleus, yet their specific origin is unknown as the region upstream of the mitochondrial-type is unrelated to databank sequences.

We also examined the retained mitochondrial-type upstream sequence for features that are typical of eukaryotic 5'UTRs and compatible with translation in the cytosol. We found that the mitochondrial-type (ancestral) start codon lacks an adenosine at position -3, a feature that has been demonstrated to be important in start codon recognition in eukaryotes (reviewed in Marintchev and Wagner, 2005). However in the nucleus this AUG has been altered to AAG (Fig.3.5, shaded) and the adjacent AUG initiator is preceded by an adenosine at position -3. An additional upstream AUG, which is present in the ancestral mitochondrial sequence has been converted to GUG in the nuclear copy in alfalfa (Fig.3.5, shaded), so is

more compatible with the scanning model of translation initiation in that eukaryotic 5' UTRs show a bias against AUG triplets (Marintchev and Wagner, 2005).

3.4 Discussion

Our identification of functional *rps1* genes in the mitochondrion of some legumes and in the nucleus of others that are closely-related has revealed features that allow us to assess steps involved in its successful transfer. While sequences of unknown origin (presumably acquired following migration from the mitochondrion) provide some expression elements such as a promoter and 5' intron splice site, native mitochondrial non-coding sequence contributes other elements, including most notably, an intron 3' splice site. The coding sequence itself may also provide signals for mitochondrial import, as the S1 ribosomal protein lacks an amino-terminal targeting peptide (and this perhaps relates to the frequent relocations of the *rps1* gene to the nucleus). While this *rps1* transfer event occurred recently enough that evolutionary changes have not obscured the native mitochondrial features, we could still observe changes that appear adaptive such as the loss of AUG triplets in the 5' UTR.

Comparative analysis of *rps1*-associated sequences has also allowed us to infer the approximate timing of this transfer event. Because the sequence upstream of *rps1* in the nucleus has rest harrow-specific mitochondrial features (gained after the pea/rest harrow split), then the earliest time of transfer must post-date the divergence of the pea lineage from other legumes, a date which is believed to be approximately 25 Mya (Lavin et al., 2005). Moreover, the nuclear environment shared among the functional copies of *rps1* in alfalfa, barrel-medic, sweet clover and fenugreek is consistent with a single transfer to the nucleus, and therefore the minimum timing of this must pre-date the divergence of these species from a common ancestor. We estimate that this occurred ~12 Mya based on chloroplast *matK* sequence data in conjunction with its calculated evolutionary rate for that clade of legumes (0.00108 ± 0.00010 sub/site/My; cf. Lavin et al., 2005). Therefore, we conclude that the transfer of *rps1* to the nucleus occurred prior to ~12 Mya but more recently than ~25 Mya (Fig. 3.1, wide grey bar on tree).

Interestingly, the *rps1* transfer appears to have occurred roughly at the same time as the transfer of mitochondrial *cox2* in the Phaseoleae lineage, which includes both soybean

and bean (Adams et al., 1999). In soybean, COX2 is nuclear-encoded and there is an apparently silent copy of *cox2* in the mitochondrion (Adams et al., 1999). These *cox2* copies show 89% nt identity (85% amino acid identity), which is similar to that seen with *rps1*, in that the alfalfa nuclear-encoded *rps1* has 91% nt identity (86% amino acid identity) with the mitochondrial-encoded counterpart in rest harrow. Furthermore, the high proportion of nonsynonymous substitutions in *rps1* (53%; three shown in Fig.3.5, shaded) is consistent with rapid amino acid change during adaptation to the new environment. It will be interesting to learn more about the nature of adaptive changes in early stages following the relocation of mitochondrial genes to the nucleus.

While this is the first documented case of the exploitation of mitochondrial-type upstream non-coding sequence for use as an intron 3' splice site, exon shuffling-type processes have been implicated in other reported cases of mitochondrial gene transfer (reviewed in Adams and Palmer 2003). Unlike with this event, the origin of immediate flanking sequences is often unknown, perhaps partly due to the age of the transfers that were under examination. In some cases, upstream regulatory sequences have been identified as partial copies of known nuclear genes encoding mitochondrial proteins, such as *hsp22*, *atpB* and *coxVb*. Another strategy for activation is expression by alternative splicing as seen for *rps14* in rice and maize (reviewed in Adams and Palmer 2003). It is clear that flowering plants make use of a variety of mechanisms that contribute to the successful transfer of mitochondrial genes, and the recruitment of mitochondrial sequences as nuclear/cytosol-specific regulatory signals illustrates another means by which translocated genes can become active.

3.5 Changes in codon usage of nuclear-located *rps1* following transfer

This analysis was not included in the submitted manuscript because of the low number of non-synonymous substitutions available for examination.

To evaluate if the codon usage of nuclear-located *rps1* has become better suited for translation in the cytosol (cf. Bonen 2006), we examined codons that have non-synonymous substitutions that are specific to the nuclear-located copy of *rps1* in alfalfa, and compared the frequency of the use of the ancestral- (mitochondrial) type codons (i.e. the codon used at the time of transfer) with the frequency of use of the codon seen in present day alfalfa cytosol (Table 3.1, Fig. 3.6). The codon usage table for alfalfa cytosol was obtained from the Codon Usage Database (Nakamura et al. 2000; http://www.kazusa.or.jp/codon/). We examined 15 codons (two were omitted due to ambiguity as to the ancestral type), and found that the non-synonymous substitution in 12 of them resulted in a codon that was more frequently used in the cytosol of alfalfa than was the mitochondrial-type (Table 3.1), albeit at two of these (codon positions 10 and 57) the difference is seemingly insignificant. Curiously, one substitution (codon position 182 of 205) results in a codon that is very infrequently used (2.2 times/thousand). Most of the substitutions seen in the nuclear *rps1* sequence in alfalfa are consistent with adaptation of *rps1* to its new location, however the small data set makes it difficult to determine if these changes actually reflect adaptation.

Table 3.1*A* Usage of codons at non-synonymous sites in nuclear-located *rps1* in alfalfa (see Fig. 3.6).

POSTION IN S1 SEQUENCE	AMINO ACID	# MEMBERS IN CODON FAMILY	CODON		
			AT TIME OF TRANSFER [freq. of use in alfalfa cytosol (%)]	PRESENT-DAY [freq. of use in alfalfa cytosol (%)]	Δ
10	P	4	CCA [2.07]	CCU [2.19]	+ 0.12
24	L	6	UUA [1.06]	UUG [2.33]	+ 1.27
29	L	6	UUA [1.06]	UUG [2.33]	+ 1.27
57	P	4	CCA [2.07]	CCU [2.19]	+ 0.12
71	L	6	CUG [0.86]	UUG [2.33]	+ 1.47
90	I	3	AUC [0.44]	AUU [2.98]	+ 2.54
95	G	4	GGC [0.94]	GGU [2.92]	+ 1.98
121	E	2	GAG [2.69]	GAA [3.71]	+ 1.02
124	A	4	GCC [1.09]	GCU [3.09]	+ 2.00
141	P	4	CCG [0.52]	CCA [2.07]	+ 1.55
145	L	6	CUU **[2.65]**	CUC **[1.14]**	**- 1.51**
150	R	6	AGA **[1.46]**	AGG **[1.11]**	**- 0.35**
159	N	2	AAC [1.77]	AAU [3.15]	+ 1.38
168	V	4	GUA [1.07]	GUG [1.63]	+ 0.56
182	R	6	AGG **[1.11]**	CGG **[0.22]**	**- 0.89**

Table 3.1*B* Codon usage table for alfalfa (cytosol) from the Codon Usage Database[1]

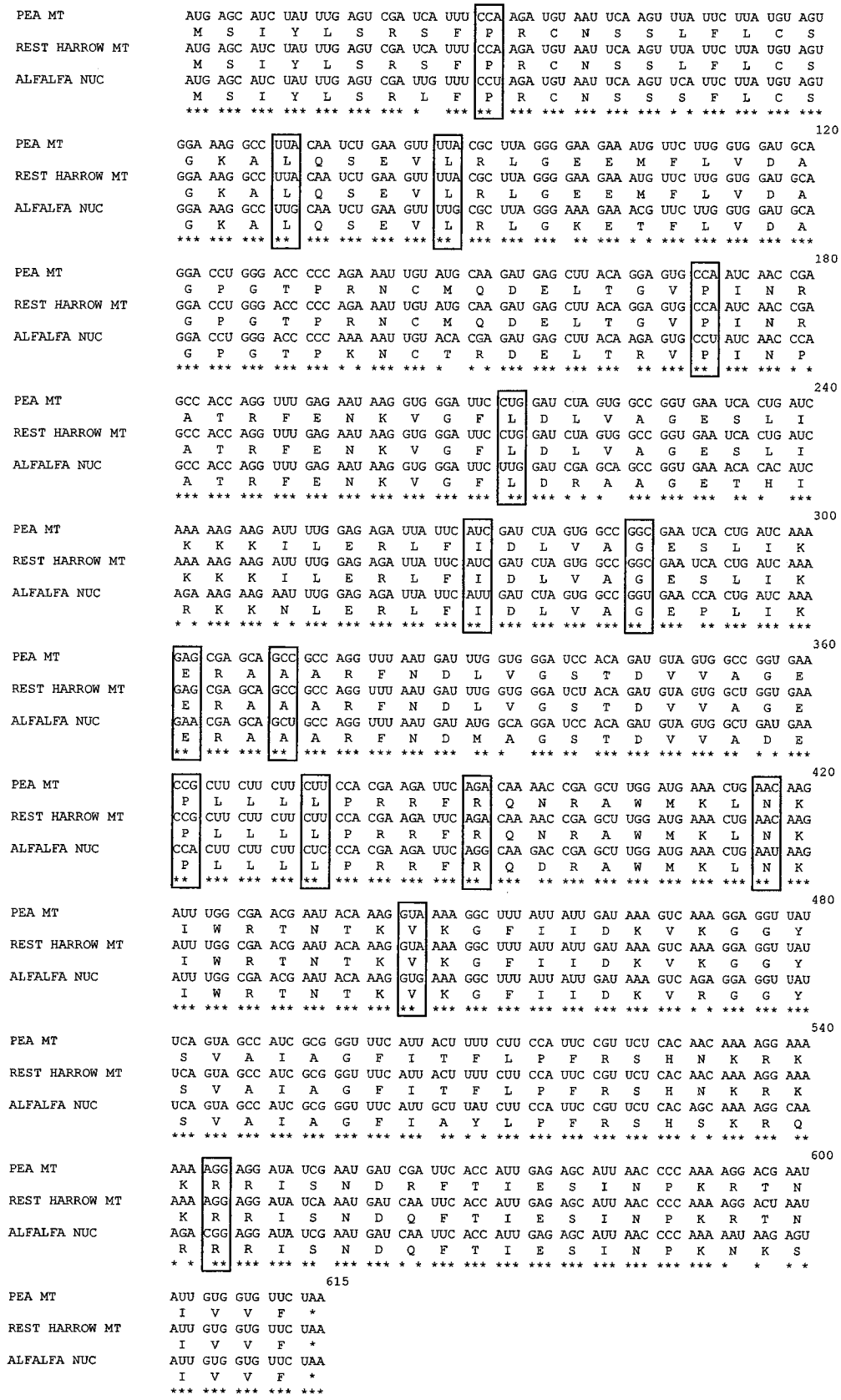
***Medicago sativa* [gbpln]: 268 CDS's (96823 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

UUU	24.6(2385)	UCU	19.6(1896)	UAU	16.2(1567)	UGU	10.8(1041)
UUC	17.4(1688)	UCC	8.5(822)	UAC	13.0(1255)	UGC	5.9(568)
UUA	10.6(1029)	UCA	18.1(1757)	UAA	1.3(125)	UGA	0.7(70)
UUG	23.3(2255)	UCG	4.0(389)	UAG	0.8(73)	UGG	11.0(1063)
CUU	26.5(2561)	CCU	21.9(2123)	CAU	14.8(1437)	CGU	8.9(863)
CUC	11.4(1101)	CCC	5.6(546)	CAC	8.8(850)	CGC	3.9(381)
CUA	8.8(856)	CCA	20.7(2007)	CAA	24.1(2330)	CGA	4.4(426)
CUG	8.6(835)	CCG	5.2(500)	CAG	12.4(1201)	CGG	2.2(217)
AUU	29.8(2882)	ACU	21.0(2031)	AAU	26.3(2549)	AGU	14.1(1363)
AUC	14.4(1393)	ACC	11.8(1139)	AAC	18.8(1823)	AGC	9.0(867)
AUA	12.1(1171)	ACA	17.7(1717)	AAA	32.8(3178)	AGA	14.6(1410)
AUG	22.9(2218)	ACG	3.2(312)	AAG	32.8(3175)	AGG	11.1(1074)
GUU	32.5(3145)	GCU	30.9(2994)	GAU	40.6(3934)	GGU	29.2(2823)
GUC	9.1(877)	GCC	10.9(1058)	GAC	16.0(1551)	GGC	9.4(908)
GUA	10.5(1015)	GCA	24.7(2394)	GAA	37.1(3592)	GGA	26.9(2608)
GUG	16.2(1572)	GCG	4.4(430)	GAG	26.9(2605)	GGG	8.2(798)

[1]http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=Medicago+sativa+[gbpln]

Figure 3.6 Nucleotide alignment of mitochondrial-located *rps1* in pea and rest harrow, and nuclear-located *rps1* in alfalfa (shown as codons with encoded amino acid below). Codons sites having nuclear-specific non-synonymous substitutions are boxed and the frequencies of the use in alfalfa cytosol for both codons at each site are shown in Table 3.1.


PEA MT
REST HARROW MT
ALFALFA NUC
120
180
240
300
360
420
480
540
600
615

CHAPTER 4: TRANSCRIPT ANALYSIS OF THE *RPS1-NAD5AB* LOCUS IN THE MITOCHONDRIA OF LEGUMES

4.0 Rationale

The *rps1* gene in the mitochondria of legumes uses RNA-level regulatory elements that were 'recruited' with the lineage-specific *rps1* upstream regions. To better understand the composition of RNA-level signals associated with this gene, I conducted a more detailed transcript analysis of the *rps1* loci in alfalfa, rest harrow, pea, soybean and bean using northern and end-mapping analyses.

Note that Materials and Methods for this chapter have been combined with those for Chapter 3, and appear in Chapter 2. This chapter will be submitted as a manuscript pending the completion of unfinished experiments.

4.1 Abstract

We previously found that the mitochondrial-located S1 ribosomal protein gene is located closely upstream of *nad5ab* in alfalfa (in which *rps1* is a pseudogene), rest harrow, pea, soybean and bean, and in each case their co-transcript profiles differ due at least in part to different genomic environments upstream of *rps1*. Here, we report the detailed characterization of precursor transcripts in these species, as well as the presence of stable excised introns of *nad5*. In pea, soybean and rest harrow, this includes the *cis*-splicing intron 1, and in the latter two species, the 5' end of the *trans*-splicing intron 2. Stable excised introns were not detected in alfalfa or bean, which are closely-related to the other species. We also observed that in alfalfa, the 5' UTR of *nad5* extends 12 nt into the *ψrps1* sequence, and this region is highly conserved with the other legumes. The occurrence of this feature only in alfalfa but in none of the other species examined raises questions about the role of species-specific nuclear-encoded machinery (or perhaps long-range RNA-RNA interactions) in addition to *cis*-elements in RNA-level regulation. Features of *rps1-nad5ab* transcripts that vary even among closely-related species illustrate the dynamic nature of plant mitochondrial regulatory elements.

4.2 Introduction

Plant mitochondrial transcripts undergo complex processing events to generate mature mRNAs. Such events typically include endonucleolytic cleavage of polycistronic transcripts (containing two or more genes) into monocistronic ones, C to U type RNA editing, and splicing of group II introns. The latter can be complicated, as *nad1, nad2* and *nad5* (encoding NADH dehydrogenase subunits of complex I in the electron transport chain) have exons that are dispersed, and transcribed at different genomic locations. The split introns must undergo *trans*-splicing to generate the mRNAs, which are typically stable (relative to their bacterial counterparts) even though they lack stabilizing poly(A) tails or a 5' cap that are characteristic of eukaryotic mRNAs. In fact, polyadenylation has been implicated in endonucleolytic degradation of mitochondrial RNAs (Hoffman et al. 2001). Little is known about features that confer RNA stability in plant mitochondria, however inverted repeats (forming stem-loop structures) have been correlated with the 3' termini of some but not all transcripts (Forner et al. 2007, Hoffman et al. 2001).

The vast majority of machinery aiding in mitochondrial RNA expression is nuclear-encoded. Among such components, pentatricopeptide repeat (PPR) proteins have been of particular interest because of their apparent function in a variety of aspects of organelle gene expression at the RNA level (Andrés et al. 2007). Genes encoding these proteins, which are present in all eukaryotic lineages, occur in unusually high numbers in plants. For example, this protein family in *Arabidopsis* consists of at least 442 members, almost half of which are predicted to be mitochondrial-targeted (Andrés et al. 2007).

The *nad5* gene in the mitochondria of flowering plants consists of five exons located at three different genomic regions; the first containing exons a and b, the second containing exon c, and the third containing exons d and e (Goth-Malonek et al. 2005). Both *cis*- (introns 1 and 4) and *trans*-splicing (introns 2 and 3) are required to generate the mature *nad5* mRNA. *Cis*-arranged homologues of *nad5* introns 2 and 3 have been detected in the mitochondria of early-diverging plants, suggesting they originated prior to the divergence of mosses (intron 2) and hornworts (intron 3) from other plant lineages (Knoop et al. 1997). The *trans*-arrangement of these introns appears to have occurred prior to the evolution of the angiosperm lineage.

While examining the status of the *rps1* gene in the mitochondria of the closely-related legumes alfalfa, rest harrow, pea, soybean and bean, we found that it was located 0.2 kb upstream of the first two exons of *nad5* in all of these species. This contrasted with the dynamic genomic environment upstream of *rps1*, which differs among each lineage following a conserved block of about 60 nt (see Chapter 3), and this is correlated with the different transcript profiles for this locus. Here, we report that in addition to the mRNAs, precursor transcripts for *rps1-nad5ab* (and neighbouring genes where they exist) are stable as are excised introns in some of the species. The 5' ends of *nad5*-containing transcripts are located 12 nt into the *ψrps1* sequence in alfalfa but not the other legumes despite the conserved nature of this region among these species. As such, it seems unlikely that this specificity is due to a recognition motif in the immediate area.

4.3 Results

4.3.1 *rps1-nad5* co-transcripts in legumes have stable precursors and processing intermediates

To determine the *rps1-nad5ab* transcript profiles in these closely-related legume species (which are each associated with a unique upstream genomic context), we conducted northern analyses using mitochondrial RNA (mtRNA) isolated from 6-day grown seedlings, or from leaves of mature plants (for rest harrow). It was noted in Chapter 3 that high molecular weight precursor transcripts were observed in pea and soybean that contain *rps1-nad5ab* with its upstream neighbouring genes (*nad4L-atp4* and *rps4*, respectively; Fig. 4.1 H, I). Using additional intron-specific probes, we determined that the slightly smaller transcripts of ~7 kb (pea) and ~6 kb (soybean) do not include *nad5* introns 1 and the 5' region of intron 2 (Fig. 4.1 H, I vs. Fig. 4.1 B, C, asterisks). The absence of the 5' end of intron 2 indicates that *trans*-splicing has joined exons b and c, and additional hybridization experiments are in progress to determine if all *nad5* exons (i.e. *nad5* mRNA) are present. The absence of these introns in transcripts that contain *nad4L-atp4* (pea) and *rps4* (soybean) suggests that at least one processing pathway exists for these transcripts that results in intron splicing prior to endonucleolytic cleavage upstream of *rps1*. While it cannot be excluded that *rps1-nad5* transcripts are generated from a promoter located within the intergenic region, in pea for example, the size of the *nad4L-atp4* dicistronic messenger in conjunction with the 5'

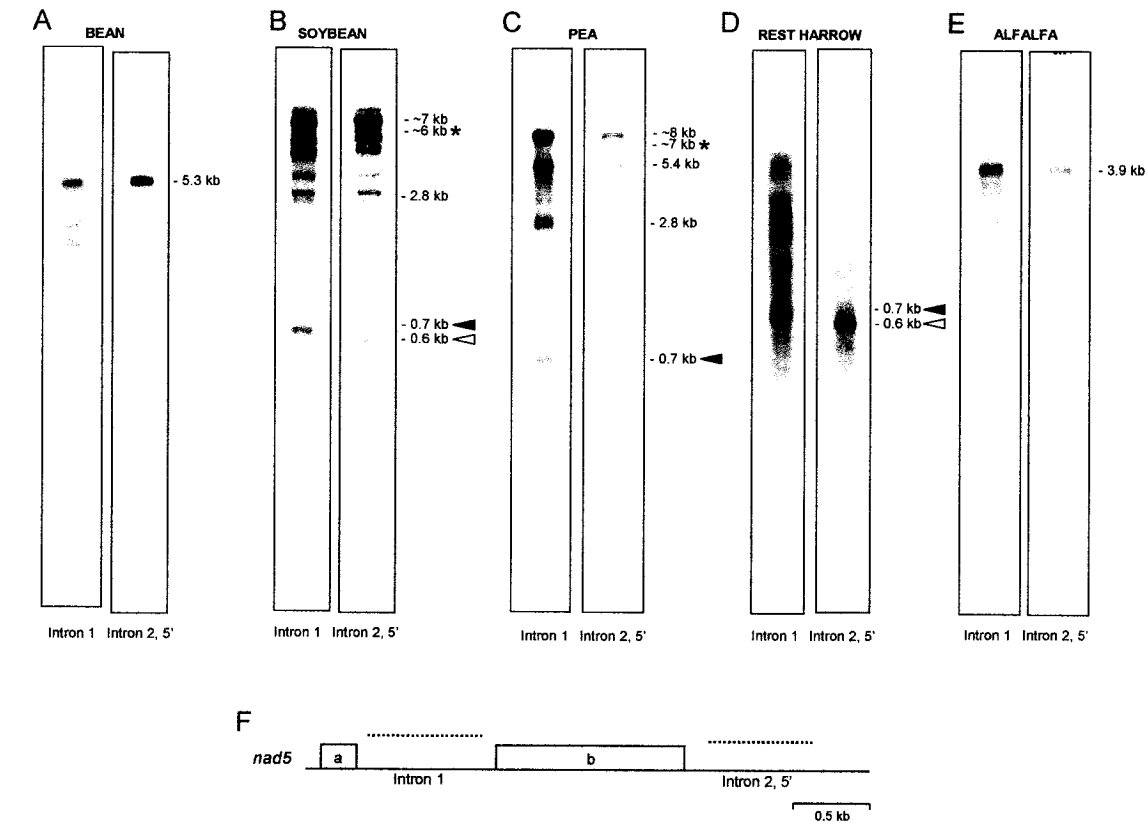
transcript termini located upstream of *rps1* as determined by primer extension analysis (indicating 5' UTRs of 156 nt and 231 nt; Fig. 4.2, open arrowheads) are consistent with endonucleolytic cleavage (Chapter 3). It is of interest to note the low levels of the *rps1-nad5ab* transcripts relative to the *nad4L-atp4* dicistronic transcripts (Fig. 4.1 I), which appear to be transcribed via a common promoter. Thus, differences in the transcript levels appear to be due to differences in transcript stability rather than promoter activity.

In addition to stable precursor and processing intermediate forms, stable excised *nad5* introns were detected in some of these legumes (Fig. 4.1 H-K). Excised forms of intron 1 were observed in soybean, pea and rest harrow (Fig. 4.1 B-D, lane 1; black arrowhead), and in contrast, these were not seen in either bean or alfalfa (Fig. 4.1 G, K; lane 1). There is high nucleotide identity among plant mitochondrial group II intron sequences. For example, *nad5* intron 1 in barrel-medic (AC145156) shares about 97% nucleotide identity with that in *Vicia faba* (in the pea lineage; Y12731). Thus, the presence of stable excised forms in some species and their absence in other closely-related ones might be due to lineage-specific differences in the nuclear-encoded machinery conferring stability/degradation rather than *cis*-elements embedded within the intron sequence. The physical form of the excised intron also likely influences its stability. Intron 1 in barrel-medic (AC145156) contains an adenosine 7 nt upstream of its 3' end suggesting that it excises as a lariat structure, and the susceptibility of this type of molecule to exonuclease degradation could in part depend on the activity of debranching enzymes (cf. Murray et al. 2001). Excised forms of intron 2, 5'-end were also seen, but only in soybean and rest harrow (Fig. 4.1 B, D; lane 2). This intron is *trans*-splicing and the excised physical form is thought to be 'Y'-shaped (or linear, depending on the splicing biochemistry) rather than lariat, and thus it might be that stability is conferred by proteins or perhaps secondary structures protecting the RNA ends.

4.3.2 The 5' ends of *nad5* transcripts are located in the *Ψrps1* sequence in alfalfa

In alfalfa, transcripts that contain full-length *Ψrps1* sequences were not detected by either northern analysis or RT-PCR, however, stable transcripts were seen for *nad5ab* precursors (containing both introns), as well as *nad5* mRNA (Fig. 3.3 E, lanes 2, 3; Fig. 4.1 E), suggesting that their stability is not dependent on *rps1* or its upstream sequence. Using S1 nuclease protection experiments, primer extension analyses, and northern analyses using probes specific to various regions along the length of this locus, we determined that the 5'

Figure 4.1 [A-K] Northern blot analysis of mitochondrial RNA from bean [A], soybean [B], pea [C], rest harrow [D] and alfalfa [E] hybridized with *nad5* intron-specific random-labeled probes as indicated below each lane. Asterisks in [B] and [C] indicate the location of additional transcripts seen with exon probes (see Fig. 3.3 B, C). Arrowheads indicate excised intron 1 (black) and intron 2 (5') (open). Note that gels were electrophoresed for different lengths of time. [F] Schematic of *nad5ab* locus indicating locations of ^{32}P-probes used (dotted lines; intron 1, LB367/LB368 PCR product, intron 2, 5', *Sal*I to *Eco*RI pea mtDNA clone by S. Bird in the Bonen Lab). [G-K] Northern blot analysis of mitochondrial RNA from bean [G], soybean [H], pea [I], rest harrow [J] and alfalfa [K] hybridized with gene-specific probes as indicated below each lane. Note that G-K are also shown in Figure 3.3.


A
BEAN
- 5.3 kb
Intron 1
Intron 2, 5'
B
SOYBEAN
- ~7 kb
- ~6 kb *
- 2.8 kb
- 0.7 kb
- 0.6 kb
Intron 1
Intron 2, 5'
C
PEA
- ~8 kb
- ~7 kb *
- 5.4 kb
- 2.8 kb
- 0.7 kb
Intron 1
Intron 2, 5'
D
REST HARROW
- 0.7 kb
- 0.6 kb
Intron 1
Intron 2, 5'
E
ALFALFA
- 3.9 kb
Intron 1
Intron 2, 5'
F
nad5
a
b
Intron 1
Intron 2, 5'
0.5 kb

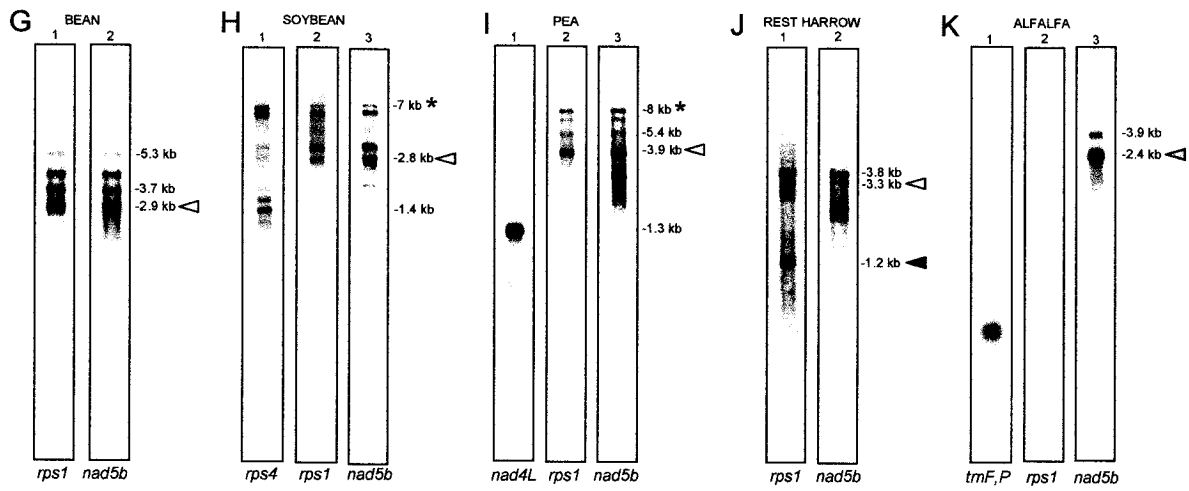
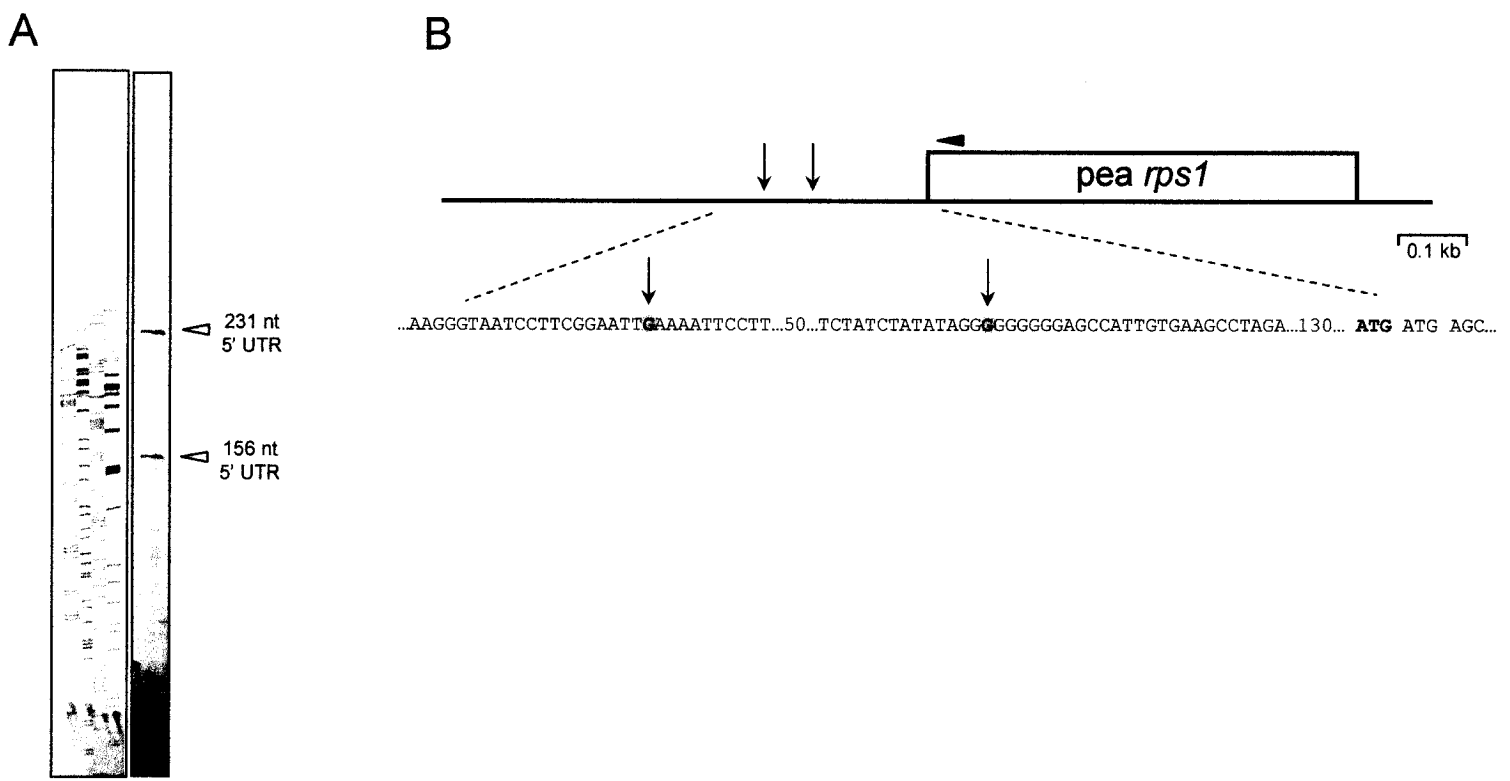

G
BEAN
1
2
-5.3 kb
-3.7 kb
-2.9 kb
rps1
nad5b
H
SOYBEAN
1
2
3
-7 kb *
-2.8 kb
-1.4 kb
rps4
rps1
nad5b
I
PEA
1
2
3
-8 kb *
-5.4 kb
-3.9 kb
-1.3 kb
nad4L
rps1
nad5b
J
REST HARROW
1
2
-3.8 kb
-3.3 kb
-1.2 kb
rps1
nad5b
K
ALFALFA
1
2
3
-3.9 kb
-2.4 kb
tmF,P
rps1
nad5b

Figure 4.2 [A] Primer extension analysis to map the 5' termini of mitochondrial *rps1* transcripts from pea. Extension products are indicated by open arrowheads. [B] Schematic of pea *rps1* showing locations of mapped 5' ends (arrows) and the position of the primer (LB243; black arrowhead). Preliminary sequence data for this region is shown below with 5' termini shaded.


A
231 nt
5' UTR
156 nt
5' UTR
B
pea rps1
0.1 kb
...AAGGGTAATCCTTCGGAATTGAAAATTCCTT...50...TCTATCTATATAGGGGGGGGGAGCCATTGTGAAGCCTAGA...130... ATG ATG AGC...

ends of the *nad5* transcripts extend 12 nt into the *Ψrps1* sequence (Fig. 4.3). The sequence in this region is well-conserved among the legumes examined with no alfalfa-specific substitutions or indels >40 nt in either direction (Fig. 4.3 D), and there is no indication of a *nad5* 5' end at this location in the other species. This suggests that recognition of this site in alfalfa for processing (or transcription initiation) is conferred by nuclear-encoded machinery and/or RNA structural features caused by long-range interactions rather than an alfalfa-specific *cis*-element at this location. Northern blot analysis indicates that a similar phenomenon might be occurring at this locus in bean although the 5' ends of *nad5* transcripts appear to be located at a different region within *rps1* sequence. A 2.9 kb transcript hybridizes with an *rps1* 3' probe (Fig. 3.3 A), but not one located further upstream (Fig. 4.4 A, lane 2). Additionally, a 3.7 kb transcript can be seen using a centrally-located *rps1* probe, but not using one ~180 nt upstream of the initiation codon, pointing to the presence of another 5' terminus, either near the 5' end of the *rps1* sequence or farther upstream (Fig. 4.4 A, black arrowhead). Given the absence of a *nad5* transcript 5' end within the *rps1-nad5ab* intergenic region of any of the legume species studied here, it would be interesting to know if stability is conferred by sequences within the *rps1-nad5ab* intergenic region. However, we noted previously that *rps1* monocistronic transcripts were seen in rest harrow, (Chapter 3, Fig. 3.3 D, lane 1) suggesting the presence of a rest harrow-specific processing site in the *rps1-nad5ab* intergenic region. Nucleotide identity among these legumes in the intergenic sequence ranges from approximately 88-98% (Fig. 4.5,) and the variable sites seem to be clustered into three regions (Fig. 4.5, sites showing variation are boxed). The significance of this is unclear, but if sequences within the intergenic region are involved in higher-order RNA folding, then the clusters of variable sites may reflect compensatory changes. The most distinctive rest harrow-specific feature is a 4 nt direct repeat insertion 8 nt upstream of the *nad5* initiation codon (Fig. 4.5). Experiments to map the 3' ends of *rps1* transcripts in rest harrow are being conducted, and it will be interesting to learn if the location of these ends relate to this insertion.

4.4 Discussion

In the mitochondria of closely-related legume species alfalfa, rest harrow, pea, soybean and bean, precursor transcripts from the *rps1-nad5ab* locus are stable, and of these

Figure 4.3 [A] S1 nuclease protection assay to map the 5' terminus of *nad5* transcripts from alfalfa. The concentrations of S1 nuclease used (0 U, 25 U, 50 U, 100 U) are shown above the lanes, and the protected product (corresponding to 5' UTR of 204 nt) is indicated by an open arrowhead. Sequence surrounding the 5' terminus (G; shaded) shown at right. [B] Primer extension analysis to map the 5' terminus of *nad5* transcripts in [A]. Extension product indicated by an open arrowhead. Short exposure of the marker shown on left. [C] Northern blot analysis of mitochondrial RNA from alfalfa hybridized with *rps1* probes as shown in schematic in [D]. The *nad5* mRNA is indicated by an asterisk. [D] DNA alignment and schematic of *ψrps1* in alfalfa showing positions of the ^{32}P-labelled PCR product used in S1 nuclease experiments (solid line), the primer used in primer extension experiments (LB473; open arrowhead), the random-labeled PCR probe used in [C] lane 1 (LB233/LB244; dotted line), and the ^{32}P-labelled oligoprobes used in [C] lanes 2-4 (LB451-LB453; lines under numbers, which correspond to the lanes in [C]). Location of 5' terminus indicated by arrows. Schematic of 5' region of *nad5* transcripts in alfalfa shown below. Nucleotide sequence alignment below the schematics shows the region of the 5' end of the alfalfa *nad5* transcripts (G, shaded) among the legumes studied. The *rps1* sequences separated from the intergenic sequence by a space and the stop codon is in bold.

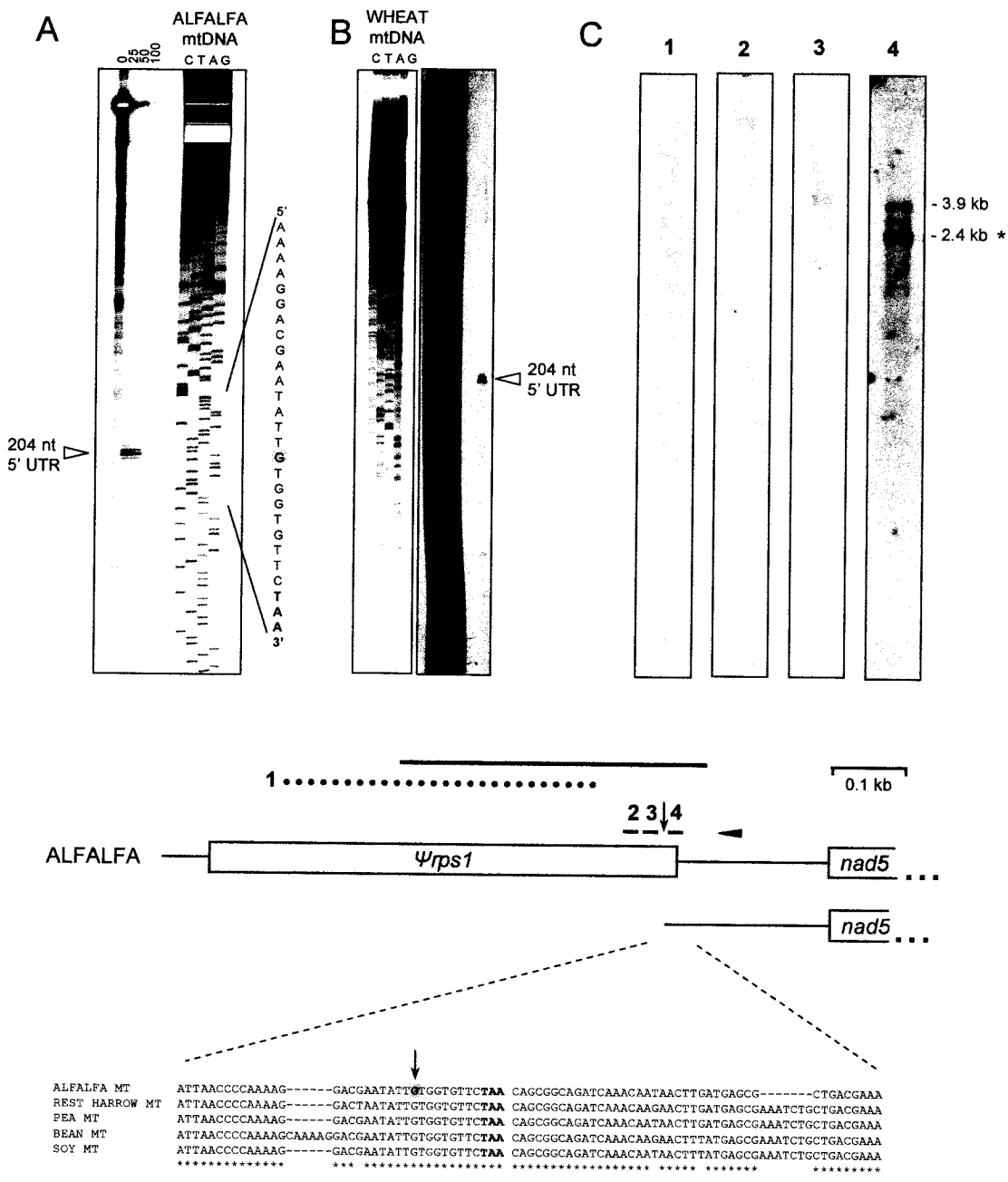

A
ALFALFA
mtDNA
C T A G
0.5
2.5
10.0
5'
A A A A G G A C G A A T A T T G T G G T G T T C T A A
3'
204 nt
5' UTR
B
WHEAT
mtDNA
C T A G
204 nt
5' UTR
C
1
2
3
4
- 3.9 kb
- 2.4 kb *
D
1
0.1 kb
2 3 4
ALFALFA
Ψrps1
nad5
nad5
ALFALFA MT ATTAACCCCAAAAG------GACGAATATTGTGGTGTTCTAA CAGCGGCAGATCAAACAATAACTTGATGAGCG-------CTGACGAAA
REST HARROW MT ATTAACCCCAAAAG------GACTAATATTGTGGTGTTCTAA CAGCGGCAGATCAAACAAGAACTTGATGAGCGAAATCTGCTGACGAAA
PEA MT ATTAACCCCAAAAG------GACGAATATTGTGGTGTTCTAA CAGCGGCAGATCAAACAATAACTTGATGAGCGAAATCTGCTGACGAAA
BEAN MT ATTAACCCCAAAAGCAAAAGGACGAATATTGTGGTGTTCTAA CAGCGGCAGATCAAACAAGAACTTTATGAGCGAAATCTGCTGACGAAA
SOY MT ATTAACCCCAAAAG------GACGAATATTGTGGTGTTCTAA CAGCGGCAGATCAAACAATAACTTTATGAGCGAAATCTGCTGACGAAA

Figure 4.4 [A] Northern blot analysis of mitochondrial RNA from bean hybridized with ^{32}P-labelled oligoprobes specific to *rps1* regions as shown above each lane. Open arrowhead indicates the 2.9 kb transcript seen with a 3' *rps1* probe (lane 3) but not a central-located one (lane 2). Black arrowhead indicates transcript (lane 2) not seen with an oligoprobe upstream of *rps1* (lane 1). Note that lane 3 is taken from Figure 3.3 [A], lane 2. [B] Schematic of bean *rps1* showing location of oligoprobes (LB435, lane 1; LB336, lane 2; LB452, lane3; lines under numbers, which correspond to lane). Schematics of proposed *nad5* transcripts shown below, with regions containing possible 5' termini indicated as dashed lines. Note that the 3.7 kb transcript in [A] lane 2 is overrepresented because of cross-hybridization with a co-migrating transcript containing partially-duplicated *rps1* sequences (cf. Chapter 3, supporting hybridization data shown in Appendix 2).

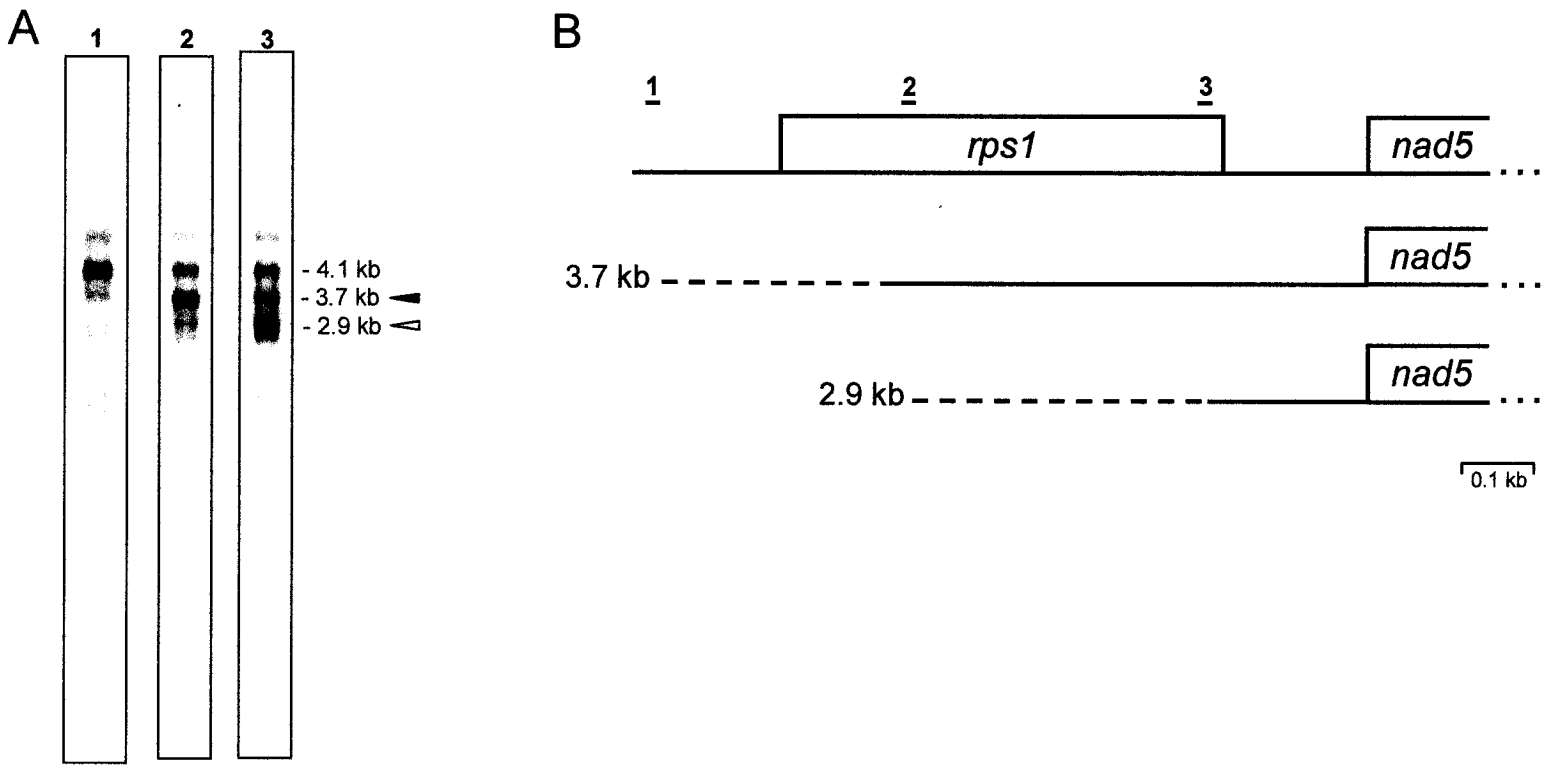
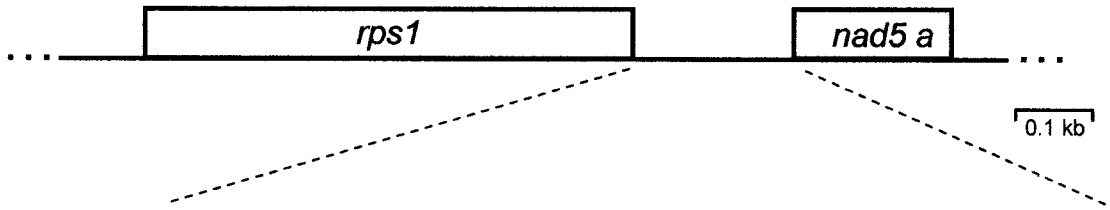

A
1
2
3
- 4.1 kb
- 3.7 kb
- 2.9 kb
B
1
2
3
rps1
nad5
nad5
3.7 kb
nad5
2.9 kb
0.1 kb

Figure 4.5 Nucleotide sequence alignment showing the *rps1-nad5ab* intergenic region. Variable sites are boxed, and variable regions are indicated by a bar above the sequence.




```
ALFALFA MT        ...TAA CAGCGGCAGATCAAACAATAACTTGATGAGCG-------CTGACGAAAAAAGTGCAGTTTTTTT---CAA
REST HARROW MT    ...TAA CAGCGGCAGATCAAACAAGAACTTGATGAGCGAAATCTGCTGACGAAAAAAGTTCAGTTTTTTT---TAA
PEA MT            ...TAA CAGCGGCAGATCAAACAATAACTTGATGAGCGAAATCTGCTGACGAAAAAAGTGCTGTTTTTTT---CAA
SOYBEAN MT        ...TAA CAGCGGCAGATCAAACAATAACTTTATGAGCGAAATCTGCTGACGAAAATCTTTCCTTTTTTTT--TCAA
BEAN MT           ...TAA CAGCGGCAGATCAAACAAGAACTTTATGAGCGAAATCTGCTGACGAAAATCTTTCCTTTTTTTTTTCAA

ALFALFA MT        TTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAATCCATTCTTTTGTTGAGGGTCTG
REST HARROW MT    TTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAATCCATTCTTTTGTTGAGGGTCTG
PEA MT            TTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAATCCATTCTTTTGTTGAGGGTCTG
SOYBEAN MT        TTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAAT--ATTCTTTTGTTGAGGGTCTG
BEAN MT           TTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAAT--ATTATTTTGTTTAGGGTCTG

ALFALFA MT        GCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTTTCTGATCACAC----TCGAAATT ATG...
REST HARROW MT    GCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTGTCTGATCACACACACTCGAAATT ATG...
PEA MT            GCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTGTCTGATCACAC----TCGAAATT ATG...
SOYBEAN MT        GCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTTGTTGTCTGATCACAC----TCGAAATT ATG...
BEAN MT           GCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTTGTTGTCTGATCACAC----TCGAAATT ATG...
```

forms, some still contain introns 1 and 2 of *nad5*. As well, stable excised introns were seen, but not in all species. Other species-specific transcript features were observed, such as a *nad5* 5' end within the *ψrps1* sequence in alfalfa, and an *rps1* monocistronic transcript in rest harrow. Through the examination of differences in transcript profiles among closely-related species we can gain insight into the nature of RNA-level regulatory signals, including those conferring stability (or conversely, degradation). Among the homologous regions within the *rps1-nad5ab* locus, nucleotide sequence identity is high among these legume species. Some lineage-specific variation can be seen (for example, in the *rps1-nad5ab* intergenic region in rest-harrow) and it may be that these create or remove specific recognition *cis*-elements that can potentially explain some of the differences in transcript patterns among species, either through a sequence recognition motif or through higher-order RNA folding. However, it cannot be ruled out that aspects of RNA-level regulation are being controlled by long-range RNA-RNA interactions involving lineage-specific sequences upstream of *rps1*. For example, Forner et al. (2005) inferred that the specificity of a processing site upstream of *cox3* that differed among three ecotypes of *Arabidopsis* was conferred by interactions between this site and ecotype-specific sequences >140 nt upstream of it.

Alternatively, the differences in transcript features among these legumes may be related to their associated nuclear-encoded machinery, and evidence points to such machinery conferring specificity in RNA-level events. For example, Bentolila et al. (2005) identified two quantitative trait loci that appear to control differences in the level of editing of a site within the *ccb206* mitochondrial gene between two ecotypes of *Arabidopsis*. Interestingly, a pentatricopeptide repeat (PPR)-containing protein has been identified near one of these loci. PPR-containing proteins have been implicated in RNA-level expression events in both mitochondria and chloroplasts and it is thought that these act as sequence-specific RNA-binding proteins that direct other machinery onto target sites on mRNAs (Andrés et al. 2007). It is as yet unclear as to how specificity is achieved by PPR proteins, but in one report, Wang et al. (2006) speculate that specificity of an mRNA by a PPR protein was lost by a single asparagine-to-serine amino acid change. Given this, it seems reasonable that these proteins (or ones providing a similar function) are capable of driving the species-specific differences in *rps1-nad5ab* RNA regulation (therefore, transcript profiles) among these legumes.

For *nad5*, 5' UTR data were known for wheat (365 nt, Pereira de Souza et al. 1991) and *Arabidopsis* (21 nt and 72 nt, Forner et al. 2007), but the absence of monocistronic *nad5* mRNAs in pea, soybean, and rest harrow is curious, and in none of the legume species did we observe *nad5* 5' ends to occur in the *rps1-nad5ab* intergenic region. This raises questions about the presence of regulatory signals within this sequence that perhaps confer transcript stability for both *rps1* and *nad5* mRNA. If so, this might relate to the absence of *ψrps1* transcripts, which appear to undergo endonucleolytic cleavage near the end of the *ψrps1* sequence and would thus no longer be associated with these signals. Alternatively, the absence of *ψrps1*-containing transcripts may be due an unknown surveillance system that degrades plant mitochondrial pseudogene transcripts, perhaps analogous to nuclear-type nonsense mediated RNA decay (reviewed in Behm-Ansmant et al. 2007). However, in rice, stable mitochondrial *ψrps11* transcripts were detected, despite the presence of two nonsense mutations (one of which is introduced by editing; Kadowaki et al. 1996). As monocistronic *rps1* mRNAs were observed in rest harrow, stability may be conferred by elements within the lineage-specific upstream sequence rather than the *rps1-nad5ab* intergenic region.

The abundance of precursor transcripts in plant mitochondria can vary even among closely-related species, as has been noted for *nad5* in wheat and maize (Pereira de Souza et al. 1991), in which the latter shows considerably more processing intermediates than the former. Interestingly, this gene appears to have complex transcript profiles in many species, as was also observed in cauliflower and chicory (Pereira de Souza et al. 1992). This raises questions as to how plant mitochondrial gene expression systems cope with the potential for translation of premature transcripts. It may be that translational machinery is part of an RNA maturation complex (cf. Choury et al. 2005) and thus translation can only occur following certain RNA maturation events. Alternatively, such premature transcripts might be translated, and the proteins are quickly detected and degraded. The development of a transformation system for plant mitochondria would greatly aid in the testing of such hypotheses. The differences seen in the transcript profiles among such closely-related species point to a continually changing composition of RNA-level regulatory elements, even when transcript sequences are highly conserved.

CHAPTER 5: COMPARATIVE ANALYSIS OF SEQUENCES PRECEDING PROTEIN-CODING MITOCHONDRIAL GENES IN FLOWERING PLANTS

Thomas Hazle and Linda Bonen

5.0 Rationale

I found that >~60 nt upstream of *rps1* in the mitochondria of legumes, the genomic context differs among alfalfa, rest harrow, pea, soybean and bean (Chapters 3 and 4). In addition, variation was observed among the homologous sequences immediately preceding *rps1*. This raised questions about the levels of conservation in sequences preceding other plant mitochondrial protein-coding genes (i.e. in regions containing signals for RNA-level regulation). The absence of the RNA-binding domain in the S1 ribosomal protein in the mitochondria of flowering plants, along with the differences at the 3' end of the SSU rRNA from the bacterial type also raise questions as to how initiation codon recognition is achieved in flowering plant mitochondria.

This chapter has been published in the journal *Molecular Biology and Evolution.* Thomas Hazle and Linda Bonen, **Comparative analysis of sequences preceding protein-coding mitochondrial genes in flowering plants.** *Mol Biol Evol* 2007 24:1101-12

5.1 Abstract

We examined the nucleotide sequences preceding 23 mitochondrial protein-coding genes held in common by maize, rice, wheat, sugar beet, tobacco, *Arabidopsis* and *Brassica* to look for features related to translation initiation and to assess the degree of conservation in mitochondrial mRNA leaders among these plants. We observed broad variation in sequence similarity as illustrated by dot plot analysis, ranging from a level rivalling that of coding sequences to complete absence of homology due to lineage-specific DNA rearrangements. Genes encoding ATP synthase subunits predominated in the latter category whereas ones encoding cytochrome c biogenesis proteins and NADH dehydrogenase subunits were primarily of the highly-conserved type. Within the region immediately preceding initiation codons, in most cases we did not observe motifs consistent with a bacterial-type Shine-Dalgarno interaction to assist in ribosome binding, nor was any other consensus sequence evident. In fact, indels in the form of tandem repeats were seen among homologues from

different plants. We did however observe a bias for high adenosine and low cytosine in the proximal ~30 nt compared to further upstream. Duplicates of some sequences in our data set were found to be associated with more than one gene within a genome. Indeed, three such families of upstream cassettes were identified and they exhibit a lineage-specific distribution among plants. Moreover, the presence of related sequences at genomic sites distant from known genes raises the possibility of future recruitment as regulatory elements. Our observations point to a dynamic nature in the make-up of the 5' leaders of plant mitochondrial mRNAs and an apparent plasticity in translational control elements.

5.2 Introduction

The mitochondrial genetic systems in various eukaryotic lineages have evolved in distinctly different directions, and some features that distinguish those of flowering plants are a low rate of nucleotide substitution, a high rate of genomic rearrangement, and very large genome sizes which range from approximately 0.2 - 2 Mbp among plants (reviewed in Bullerwell and Gray 2004). The protein-coding gene content in plant mitochondria typically consists of ~35 components which are primarily subunits of the respiratory chain and translational machinery. This set is smaller however than that of certain protists such as *Reclinomonas americana* which has 67 protein-coding genes in a mitochondrial genome of only 69 kb (Gray, Lang and Burger 2004). Non-coding sequences account for most of the expanded genome size in plants, and although certain stretches can be identified as acquired chloroplast or nuclear sequences, much is of unknown origin (reviewed in Kubo and Mikami 2007). In addition, most plant mitochondrial genomes exist in complex physical forms due to DNA recombination across repeated sequences which results in a mixed population of sub-genomic forms of the deduced master chromosome. Consequently, gene order often varies (even among closely-related plants) and coding regions can become relocated into the context of new flanking sequences, and hence acquire new regulatory signals. In exceptional cases, rearrangement events can even impact on coding sequences.

Although progress is being made in our knowledge about transcription and RNA processing events in plant mitochondria, as yet little is known about translation initiation or signals involved in the recognition of the correct start codon. This is despite a longstanding interest in this issue (cf. Dawson, Jones and Leaver 1984; Boer et al. 1985; Schuster et al.

1990). For example, Pring et al. (1992) identified three short (10-12 nt) conserved blocks within ~100 nt preceding the start codons of respiratory chain genes such as *atp6, atp4* and *cox2* in various grasses and eudicots, suggestive of a regulatory role. Based on literature reports as well as our own unpublished data, plant mitochondrial 5' untranslated regions (UTRs) often range from 100-400 nt in length and presumably contain expression elements for translational control and perhaps also for mRNA stability. Genes which are co-transcribed sometimes possess RNA cleavage sites and plant mitochondrial transcripts also typically undergo C-to-U type RNA editing, as well as splicing of group II introns in some cases. Editing, although predominantly occurring within coding sequences, has occasionally been observed in UTRs (reviewed in Shikanai 2006). For example, an editing site was identified 4 nt upstream of the *rps14* initiation codon in *Oenothera* mitochondria (Schuster et al. 1990) and in rice, there is an editing site within the 3 nt spacer separating *rpl2* and *rps19* (Kubo et al. 1996). This organization of ribosomal protein genes illustrates traces of an ancestral bacterial-type order of genes occasionally seen in plant mitochondria.

It might be anticipated that translation initiation in plant mitochondria would be similar to that of bacteria given their endosymbiotic ancestry and the highly conserved nature of core regions within their ribosomal RNAs (Gray 1992). In bacteria, initiation codon recognition is assisted by base-pairing between the pyrimidine-rich 3' end of the SSU (16S) rRNA and a purine-rich (Shine-Dalgarno) sequence within the mRNA. The latter is typically 4-5 nt in length and located 5-9 nt upstream from the start codon (reviewed in Kozak 1999, Marintchev and Wagner 2005, Nakamoto 2006). This, in conjunction with the fMet-tRNA anticodon-codon interaction is important in correct positioning of the 30S ribosomal initiation complex on the mRNA. The chloroplast translation system in flowering plants also has these features, and Shine-Dalgarno sequences have been experimentally shown to be necessary for the translation of a subset of chloroplast mRNAs (reviewed in Sugiura, Hirose and Sugita 1998). In plant mitochondria, however, the extreme 3' terminal region of the SSU (18S) rRNA is slightly shorter based on direct RNA sequencing data (cf. Schnare and Gray 1982) and it lacks a canonical anti-Shine-Dalgarno sequence. Consequently, it has been puzzling as to how the correct initiation site is recognized in plant mitochondria. Interestingly, in *Reclinomonas americana* mitochondria, which is described as the most bacteria-like of any known mitochondrial genome, candidate Shine-Dalgarno-like sequences

Table 5.1. Plant species examined in this study

Species	NCBI Accession no.	Reference
Maize (*Zea mays*)	AY506529	Clifton et al. 2004
Rice (*Oryza sativa*)	BA000029	Notsu et al. 2002
Wheat (*Triticum aestivum*)	AP008982	Ogihara et al. 2005
Sugar beet (*Beta vulgaris*)	BA000009	Kubo et al. 2000
Tobacco (*Nicotiana tabacum*)	NC_006581	Sugiyama et al. 2005
Arabidopsis thaliana	NC_001284	Unseld et al. 1997
Brassica napus	AP006444	Handa 2003

Table 5.2. Genes included in the analysis

Gene		Comments
NADH dehydrogenase	*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*	
Cytochrome b	*cob*	
Cytochrome oxidase	*cox1*, *cox2*, *cox3*	
ATP synthase	*atp1*	
	atp4 (*orf25*)	
	atp6	2 copies in wheat and *Arabidopsis*
	atp8 (*orfB*)	2 copies in wheat
	atp9	
Cytochrome c biogenesis	*ccmB*	annotated as *ccb206* in *Arabidopsis* and sugar beet[a]
	ccmFN[b]	annotated as *ccb382* in *Arabidopsis*, and *ccb577* in sugar beet[a] split into *ccmFN1* and *ccmFN2* in *Arabidopsis* and *Brassica*
	ccmFC[b]	annotated as *ccb452* in *Arabidopsis*, and *ccb438* in sugar beet[a]
Ribosomal proteins	*rps3*, *rps12*	

[a] As reported in Unseld et al. (1997) and Kubo et al. (2000).
[b] In angiosperms, *ccmF* is split into *ccmFN* and *ccmFC*, which encode the amino and carboxyl regions respectively.

have been identified upstream of most protein-coding genes (Lang et al. 1997).

In the present study, we have analysed the 100 nt stretches located immediately upstream of protein-coding genes that are present in all seven mitochondrial genomes which have been completely sequenced, namely those of maize (Clifton et al. 2004), rice (Notsu et al. 2002), wheat (Ogihara et al. 2005), sugar beet (Kubo et al. 2000), tobacco (Sugiyama et al. 2005), *Arabidopsis* (Unseld et al. 1997) and *Brassica* (Handa 2003). Such regions will include mRNA leader sequences that are expected to contain signals involved in translation initiation. Our study has revealed a wide variation in degree of sequence conservation and an apparent plasticity in translational control elements.

5.3 Methods

Nucleotide sequences upstream of plant mitochondrial protein-coding genes were obtained from the NCBI Genbank entries for the seven genomes listed in Table 5.1. Genes held in common among these plant genomes are shown in Table 5.2 with the exception that *mttB*, *matR*, *rps4*, and *ccmC* were omitted from our data set because there is uncertainty as to the position of their start codons in one or more plants. Although *ccmC* was originally designated as a pseudogene in sugar beet (Satoh et al. 2004), its status has been re-evaluated (Mower and Palmer 2006). Intraspecific comparisons were made using sequences from sugar beet normal-type cytoplasm (BA000009) and "Owen cytoplasm" (BA000024). In almost all cases, the locations of translation initiation codons were taken from the Genbank annotations. However, there were seven instances in which there is strong phylogenetic support from comparative sequence analysis for the use of a downstream in-frame AUG, that is the one which would correspond with the initiation codon annotated in other species. Moreover, the presence of frame-shifts or in-frame stop codons in homologous sequences from closely-related species usually precluded the use of the distal AUG. Those seven cases for which the proximal AUG was selected (and distance downstream of the annotated initiator) are: *Arabidopsis nad2* (33 nt), wheat *nad6* (54 nt), wheat *nad9* (291 nt), rice *cox3* (45 nt), *Arabidopsis atp9* (33 nt), maize *ccmFN* (39 nt) and wheat *ccmFN* (42 nt). In certain cases, these designations are also supported by experimental analysis (cf. wheat *nad6*, Haouazine-Takvorian et al, 1997; wheat *nad9,* Lamattina et al. 1993).

We selected a length of 100 nt preceding start codons because, although experimental data are as yet limited, plant mitochondrial 5' UTRs are typically at least this long. Sequence alignments were carried out using ClustalW with the default parameter settings (Chenna et al. 2003) and then subjected to minor correction by manual inspection. The dot matrix plot was generated using JDotter (http://athena.bioc.uvic.ca/workbench.php?tool=jdotter&db; Brodie, Roper and Upton 2004) set to default preferences with the stringency level adjusted so that even relatively short sequences of low complexity were visualized, and after identification as such they were omitted from further analysis. This was done to minimize the possibility of overlooking potentially meaningful sequence relationships.

To assess the presence of additional copies within individual plant mitochondrial genomes, we conducted BLAST (bl2seq) searches using each 100 nt sequence in our data set as query with default parameters and the filter removed (Tatusova and Madden 1999). A duplicate sequence was designated as an "upstream cassette" when at least 50 nt were found within 100 nt upstream of another gene. This included three genes not in our primary data set, namely *rps7* in maize, rice and wheat, *rpl5* in sugar beet, and *rps13* in tobacco. Copies not closely associated with a known gene were categorized based on length of sequence similarity as well as position within the query sequence. Long copies (i.e. > 40 nt) exhibited E-values ranging from 10^{-6} to 10^{-20}, with most being lower than 10^{-10}, while short copies (~20-40 nt) exceeded ~90% sequence identity. We also searched for specific short nucleotide stretches within our data set using FUZZNUC (e.g. allowing for 2 nt degeneracy within an 8 nt stretch) (http://bioweb.pasteur.fr/seqanal/interfaces/fuzznuc.html). To determine whether sequences of chloroplast origin were represented in out data set, we used BLAST (bl2seq) to query the rice (NC_001320) and *Arabidopsis* (NC_000932) chloroplast genomes.

5.4 Results

5.4.1 Variation in conservation and origin of upstream sequences

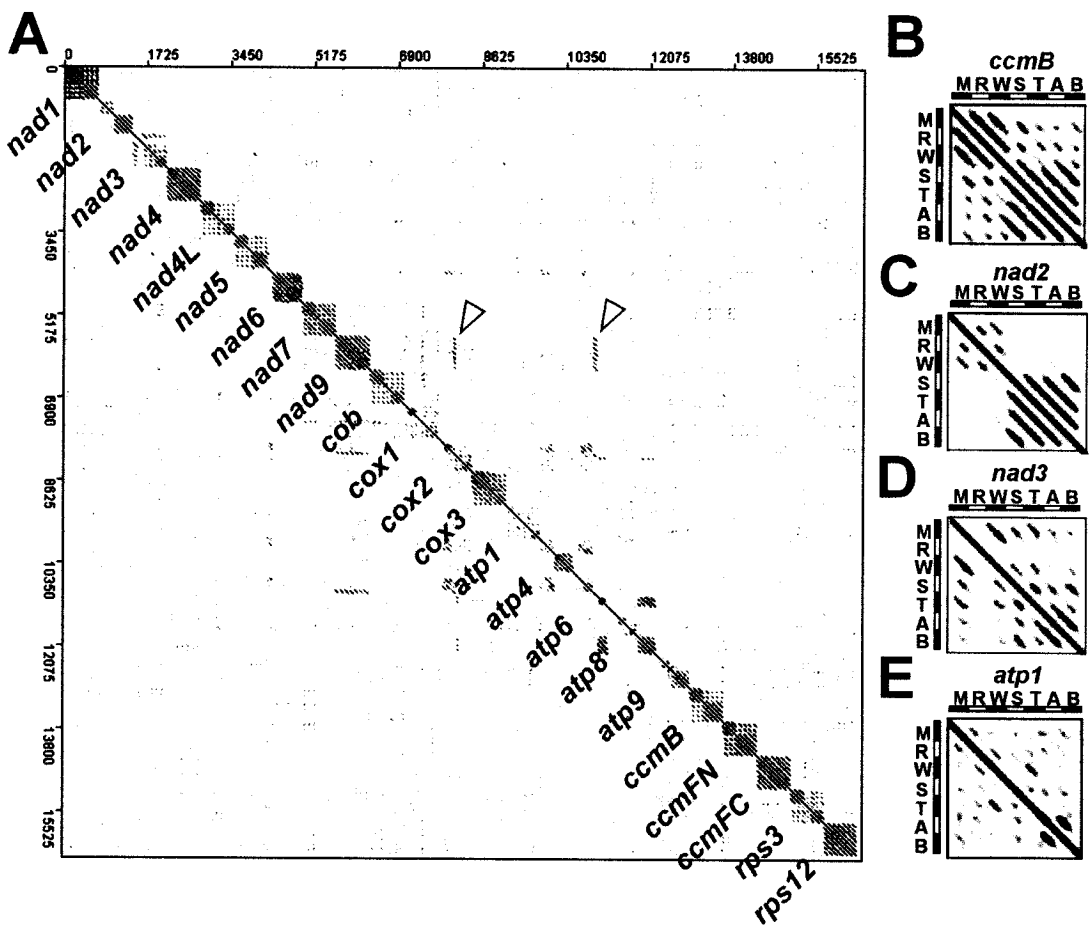
To assess the degree of conservation in sequences preceding plant mitochondrial genes and to search for potential translational regulatory signals, we compiled sequences extending 100 nt upstream of 23 protein-coding genes which are held in common among the seven flowering plant species whose mitochondrial genomes have been completely sequenced (Table 5.1). More specifically, these plants include three monocots, maize, rice,

and wheat, which diverged from a common ancestor ~50-70 Mya (Kellogg 2001) and four eudicots, sugar beet, tobacco, *Arabidopsis* and *Brassica*. The latter two species are the most closely-related, having diverged from a common ancestor ~20 Mya (Koch, Haubold and Mitchell-Olds 2000) and the monocot-eudicot split is estimated to have occurred about 150 Mya (Chaw et al. 2004). Among the 23 genes used in our analysis (Table 5.2), three are both duplicated and show variation within the 100 nt preceding the start codon, namely *atp8* and *atp6* in wheat, and *atp6* in *Arabidopsis*, so were also included in our data set. Nucleotide alignments are provided in Supplementary Figure S1. RNA editing of a genomically-encoded ACG in a few cases generates the AUG initiation codon; more specifically, there are 6 cases for *nad1*, 4 for *nad4L*, and one each for *cox1* and *atp6*. This was based on reported experimental evidence or inferred from comparative analysis.

A dot matrix plot was generated for this data set of 164 sequences strung together as a single consecutive sequence, and this is shown as a comparison with itself in Figure 5.1*A*. Signals along the diagonal (seen as a continuous line) represent identical sequences, whereas signal patterns immediately off the diagonal (squared regions) depict sequence similarity among different plants for a given gene. Signals further off the diagonal represent sequences that are duplicated and located upstream of more than one gene (Fig.5.1A, illustrated by open arrowheads, discussed below). As can be seen from Figure 1*A*, there is a broad variation in the appearance of the profiles, hence in the sequence similarity preceding different mitochondrial genes. Strikingly, some upstream regions like those associated with *nad4*, *ccmFN*, and *ccmFC* are so similar near the start codon among all seven plant species that conservation rivals that of the coding sequence. For yet others, the sequences upstream of genes such as *ccmB* (Fig.5.1*B*) show somewhat greater divergence and reflect the phylogenetic distance between the monocot and eudicot groups. In the case of *nad2* (Fig.5.1*C*), the monocot-eudicot differences are even more accentuated, with sequences appearing more divergent among the monocots, even though these plants are more closely-related than are most of the eudicot species to each other.

Among the genes exhibiting conserved upstream sequences, there were several cases in which the sequence from one plant was seen to lack homology with those in all the other species (Supplementary data, Fig. S1). One of these, namely *nad3*, is shown in Figure 1*D*, and in this case mitochondrial DNA rearrangements in the rice lineage have replaced the

Figure 5.1 Dot plot analysis of sequences preceding 23 mitochondrial protein-coding genes from seven flowering plants. [A] Upstream sequences (100 nt stretches) from the genes listed along the diagonal in the order maize (M), rice (R), wheat (W), sugar beet (S), tobacco (T), *Arabidopsis* (A) and *Brassica* (B), respectively, were strung together as a single consecutive sequence (total length of 16,400 nt) and compared against itself. Open arrowheads denote duplicated sequences located upstream of more than one gene in the sugar beet mitochondrial genome (i.e. cassette type II, see Fig.5.5a) [B-E] Enlargements of dot plots shown in panel [A] for sequences upstream of [B] *ccmB*, [C] *nad2*, [D] *nad3* and [E] *atp1*.


A
0
1725
3450
5175
8900
8625
10350
12075
13800
15525
nad1
nad2
nad3
nad4
nad4L
nad5
nad6
nad7
nad9
cob
cox1
cox2
cox3
atp1
atp4
atp6
atp8
atp9
ccmB
ccmFN
ccmFC
rps3
rps12
B
ccmB
C
nad2
D
nad3
E
atp1
M R W S T A B

ancestral-type sequence with one unrelated to anything in the databank. In the case of maize *nad6*, the recently-acquired upstream sequence was seen to be a paralogue of that preceding *atp4* (discussed below). The dot matrix plot in Figure 5.1*A* also shows that some genes lack prominent signals, and homology among plants is either uncertain or absent. For example, upstream of *atp1*, clear-cut similarity was detected only between *Arabidopsis* and *Brassica* (Fig.5.1*E*), although short stretches of similarity could be identified among various plants (Supplementary Fig.S1) as represented by the fainter signals in Figure 5.1*E*. BLAST searches using sequences in our data set as query revealed numerous matches with short duplicated segments from various genomic sites. For example, the rice *nad2* upstream sequence contains a *nad7* coding segment of 98 nt beginning approximately 30 nt upstream of the start codon. These observations are consistent with a well-known feature of plant mitochondrial genomes, namely that "bits and pieces" of genic (and flanking sequence) copies are scattered around the genome (cf. Kubo and Mikami 2007), having arisen through DNA duplication/recombination events or RT-mediated amplification of transcribed sequences with subsequent genomic integration. It is worth noting that none of the sequences in our data set exhibited similarity to sequences of chloroplast origin, which reside in plant mitochondrial genomes, although such sequences can be recruited to serve as mitochondrial expression elements (cf. Nakazono et al. 1996).

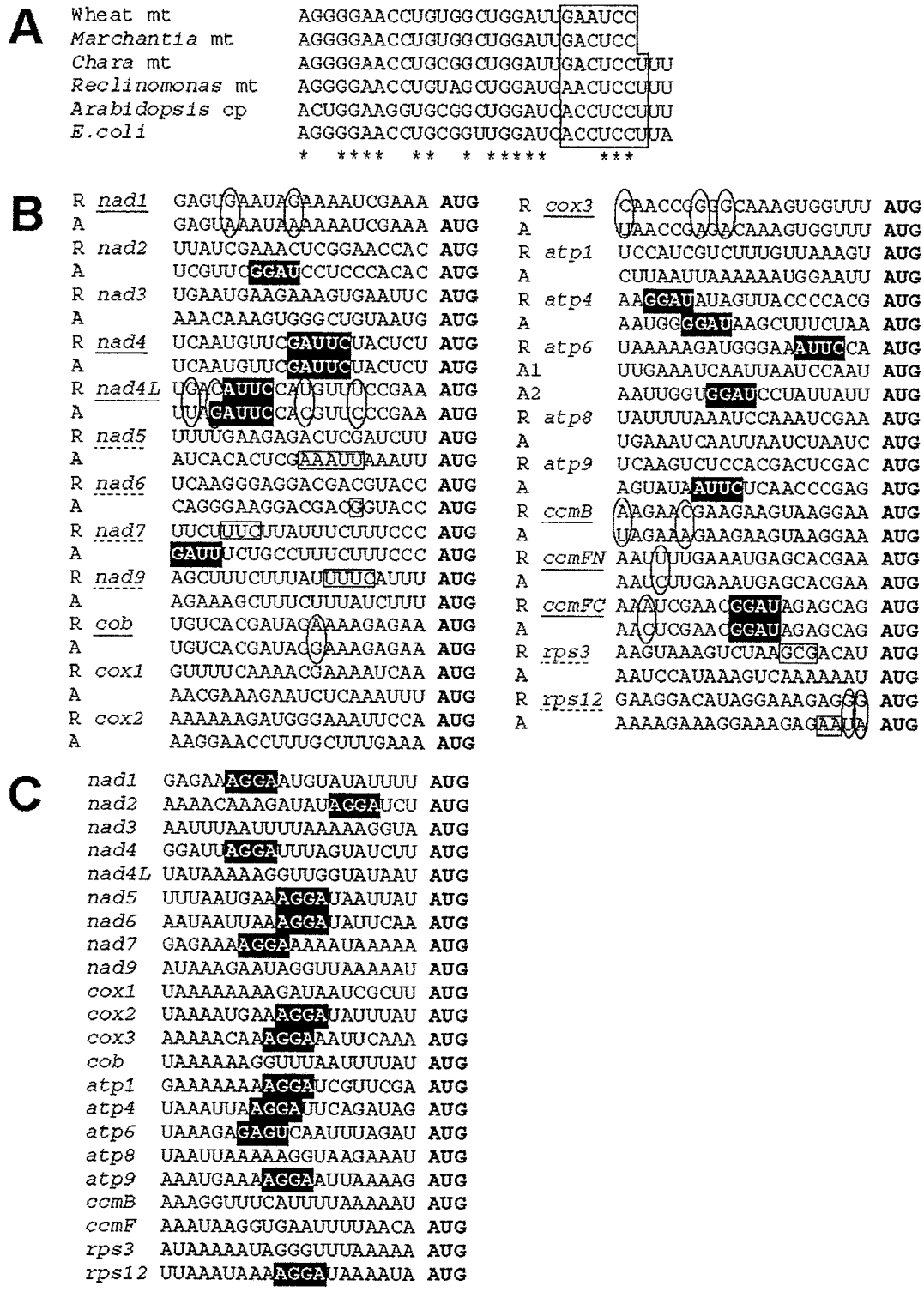
From our dot plot analysis, it can also be seen that there are several trends with respect to gene type and degree of variation among plant species. For example, sequences preceding the *atp* and *cox* genes are more volatile than those upstream of *ccm* genes, or to a lesser extent *nad* genes (Fig.5.1*A*). The *atp6* gene is particularly atypical, as it is often preceded by a fused ORF which differs among plants (reviewed in Bonen and Brown 1993) and the resulting amino-terminal region is removed by proteolytic cleavage to generate the mature ATP6 protein (cf. Krishnasamy, Grant and Makaroff 1994). In addition, *atp6* copies are located on recombinationally-active repeats in plants such as wheat (Bonen and Bird 1988) and *Arabidopsis* (Marienfeld et al. 1996) and consequently differ in their flanking upstream sequences. On the other hand, it should also be noted that some upstream sequences in our data set are constrained because of their close proximity to a co-transcribed gene. In the case of *rps12* in all seven species, the 3' end of *nad3* is only ~50 nt upstream (Supplementary Fig. S1). Similarly, *rps3* is located ~10-20 nt downstream of *rps19* in

tobacco (or a *Ψrps19* copy in the other six plant species), so that in these two cases the query included coding sequences (Supplementary Fig. S1, bold).

5.4.2 Apparent absence of a bacteria-type ribosome binding motif

To search for features consistent with a Shine-Dalgarno type base-pairing interaction, the 20 nt stretch immediately preceding each protein-coding gene in our data set was examined for sequences complementary to at least 4 consecutive nucleotides within 6 nt at the extreme 3' end of the plant mitochondrial 18S rRNA (Fig.5.2*A*). It should be noted that the 3' terminus of the mitochondrial SSU rRNA is highly conserved among plants, and this even extends to the bryophyte *Marchantia polymorpha*. However the plant mitochondrial ones differ from bacterial and chloroplast in sequence at the expected position of the anti-Shine-Dalgarno element (Fig.5.2*A*, boxed) and are slightly shorter in length. This evolutionary shift appears to have occurred after land plants diverged from the green algal lineage, in that the counterpart in *Chara vulgaris* mitochondria still is bacteria-like (Turmel, Otis and Lemieux 2003) as seen in Figure 5.2*A*. Interestingly, within the region shown in Figure 5.2*A*, only one nucleotide difference is observed between *Marchantia* and flowering plants, yet it is located within the boxed region and further reduces the pyrimidine content. At the location corresponding to the Shine-Dalgarno motif in bacterial mRNAs, only 42 of 164 plant mitochondrial sequences exhibited potential complementary to the SSU rRNA sequence. Such candidate sites for the rice and *Arabidopsis* subset are shown in Figure 5.2*B* (black boxes) and they vary in location relative to the start codon. In contrast to plant mitochondria, approximately half of the sequences upstream of the corresponding genes in *Reclinomonas* mitochondria contain potential complementary sequences (Fig.5.2*C*), which are virtually all identical and at a location more similar to that found in bacteria (cf. also Lang et al., 1997). Short stretches of relatively low complexity (e.g. homopolymeric adenosine as well as uridine) are evident in Figure 5.2*B*, but we found no obvious consensus sequence among all (or even most) sequences in the data set. Figure 5.2*B* also illustrates the degree of sequence variation observed between rice and *Arabidopsis*. More specifically, among the 23 different genes, only 14 upstream sequences appear to be homologous. In eight of those cases (Fig.5.2*B*, name underlined), differences are limited to a few nucleotide substitutions (or none in the case of *nad4*) (Fig.5.2*B*, vertical ovals), of which about 25% are C-to-U editing candidates. For the other six cases (Fig.5.2*B*, name broken underlined),

Figure 5.2 Assessment of potential Shine-Dalgarno-type base-pairing between SSU rRNA and 5'UTR mRNA sequences. [A] Alignment of 3' terminus of SSU rRNA from wheat mitochondria with mitochondrial counterparts from bryophyte (*Marchantia polymorpha*, M68929), green alga (*Chara vulgaris*, AY267353) and protist (*Reclinomonas americana*, NC_001823), as well as plant chloroplast (*Arabidopsis thaliana*, NC_000932) and bacteria (*E. coli*, AP009048). The region corresponding to the bacterial anti-Shine-Dalgarno sequence (reviewed in Marintchev and Wagner 2005) is boxed and asterisks depict identical nucleotides among the 6 sequences. The wheat mitochondrial 18S rRNA 3' terminus was experimentally determined (Schnare and Gray 1982) and is representative of other flowering plants. [B,C] Sequences of the 20 nt stretches preceding the 23 genes in our data set (see Fig.5.1) from [B] rice and *Arabidopsis*, and [C] *Reclinomonas*. Highlighted blocks (white on black) represent stretches of potential complementarity to the 3' terminal regions of their respective mitochondrial SSU rRNA. Initiation codons are shown in bold. Nucleotide substitutions between rice and *Arabidopsis* are shown by ovals (with gene names underlined) and indels are depicted by blocks (with gene names in broken-underline). Abbreviations are as in Fig. 5.1.


A
Wheat mt AGGGGAACCUGUGGCUGGAUUGAAUCC
Marchantia mt AGGGGAACCUGUGGCUGGAUUGACUCC
Chara mt AGGGGAACCUGCGGCUGGAUUGACUCCUUU
Reclinomonas mt AGGGGAACCUGUAGCUGGAUGAACUCCUUU
Arabidopsis cp ACUGGAAGGUGCGGCUGGAUCACCUCCUUU
E.coli AGGGGAACCUGCGGUUGGAUCACCUCCUUA
B
C
nad1 GAGAAAGGAAUGUAUAUUUU AUG
nad2 AAAACAAAGAUAUAGGAUCU AUG
nad3 AAUUUAAUUUUAAAAAGGUA AUG
nad4 GGAUUAGGAUUUAGUAUCUU AUG
nad4L UAUAAAAAGGUUGGUAUAAU AUG
nad5 UUUAAUGAAAGGAUAAUUAU AUG
nad6 AAUAAUUAAAGGAUAUUCAA AUG
nad7 GAGAAAAGGAAAAUAAAAA AUG
nad9 AUAAAGAAUAGGUUAAAAAU AUG
cox1 UAAAAAAAAGAUAAUCGCUU AUG
cox2 UAAAAUGAAAGGAUAUUUAU AUG
cox3 AAAAACAAAGGAAAUUCAAA AUG
cob UAAAAAAGGUUUAAUUUUAU AUG
atp1 GAAAAAAAAGGAUCGUUCGA AUG
atp4 UAAAUUAAGGAUUCAGAUAG AUG
atp6 UAAAGAGAGUCAAUUUAGAU AUG
atp8 UAAUUAAAAAGGUAAGAAAU AUG
atp9 AAAUGAAAAGGAAUUAAAAG AUG
ccmB AAAGGUUUCAUUUUAAAAAU AUG
ccmF AAAUAAGGUGAAUUUUAACA AUG
rps3 AUAAAAAAUAGGGUUUAAAAA AUG
rps12 UUAAAUAAAAGGAUAAAAUA AUG

insertion/deletions were observed between rice and *Arabidopsis* (Fig.5.2*B*, boxed). Notably, such indels occur almost immediately upstream of 14 different genes in our data set (Supplementary Fig. S1) and six of them are shown in Figure 5.3. Their direct tandem repeat nature suggests that they originated by slippage during DNA replication. Indeed, 3 of the cases shown in Figure 5.3 involve differences between the closely-related *Arabidopsis* and *Brassica*, presumably reflecting very recent evolutionary events. Such variation would be unexpected if regions near the start codons were under strong functional constraint due to the presence of a ribosome binding site that has strict requirements for its sequence and distance from AUG. Taken together, our observations suggest that classical bacteria-type ribosome binding is unlikely to be the mode of initiator recognition in plant mitochondria. When we examined the nucleotide composition of sequences immediately upstream of protein-coding genes, we observed a bias for high adenosine (36.2%) and low cytosine (16.2%) in the 30 nt preceding start codons (Fig. 5.4*A*) compared to further upstream (27.6% and 21.3% respectively). In contrast, both the G and U content are more uniformly represented in the 100 nt region preceding initiators and range from 21-28% (Fig.5.4*A*). Notably, the nucleotide content in the region extending from ~30-100 nt upstream reflects the overall A+T content (~55%) of flowering plant mitochondrial genomes (reviewed in Sugiyama et al. 2005). The asymmetry in A vs. C composition for positions -1 to -30 (Fig.5.4*A*) mimics the profile of the A vs. U content in the comparable region of yeast cytosolic 5' UTRs (Shabalina et al. 2004) and it will be of interest to learn what parameters control accessibility of the ribosome to the correct start site. We observed at least one additional AUG triplet independent of coding reading frame in about 80% of the sequences in our data set, thus disfavouring a scanning model of initiation codon recognition.

Within the coding region examined (namely, the first 100 nts), there is a bias for high uridine, which can at least partially be attributed to the third codon position U-bias observed in plant mitochondria (cf. Maier et al. 1996). An examination of the nucleotide composition close to the start codon (Fig.5.4*B*) revealed only one conspicuous feature, that is, a bias against guanosine at position -2. There was no indication of an "extended" codon-anticodon interaction, as has been suggested for translation in chloroplasts where uridine residues are typically present at position -1 (cf. Esposito et al. 2003) and the A+G content is only approximately 55% for positions -5 to -9 which would correspond to the location of the

Figure 5.3 Alignment of regions preceding initiation codons of selected plant mitochondrial genes. Short insertion/deletions in the form of direct repeats are depicted by arrows. The boxed sequence upstream of *nad6* in maize indicates non-homology (see text). Abbreviations are as in Fig 5. 1.

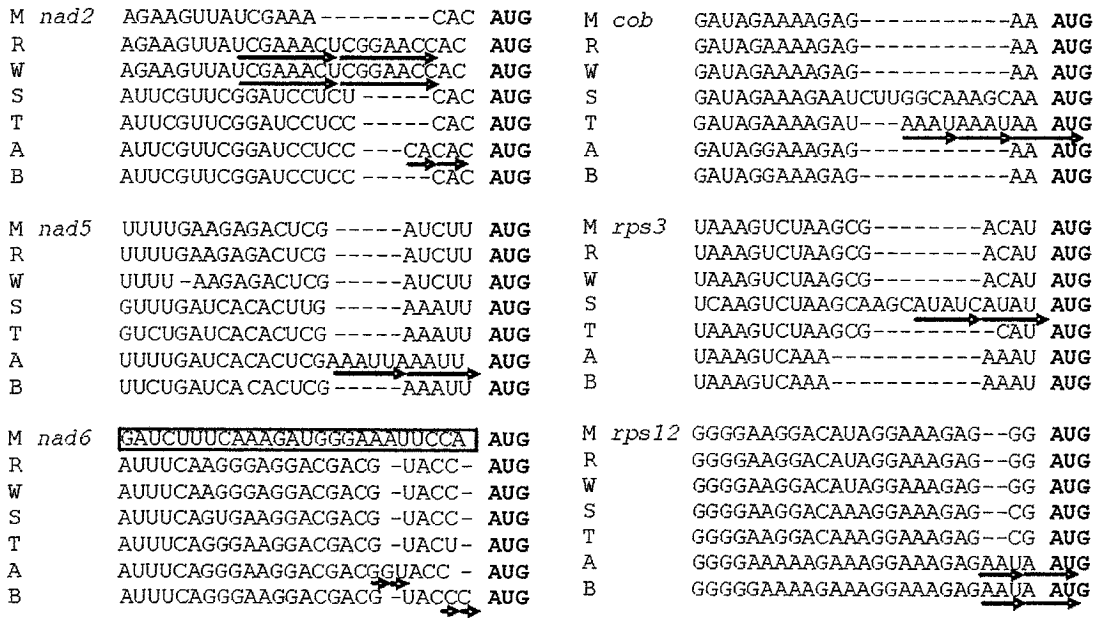
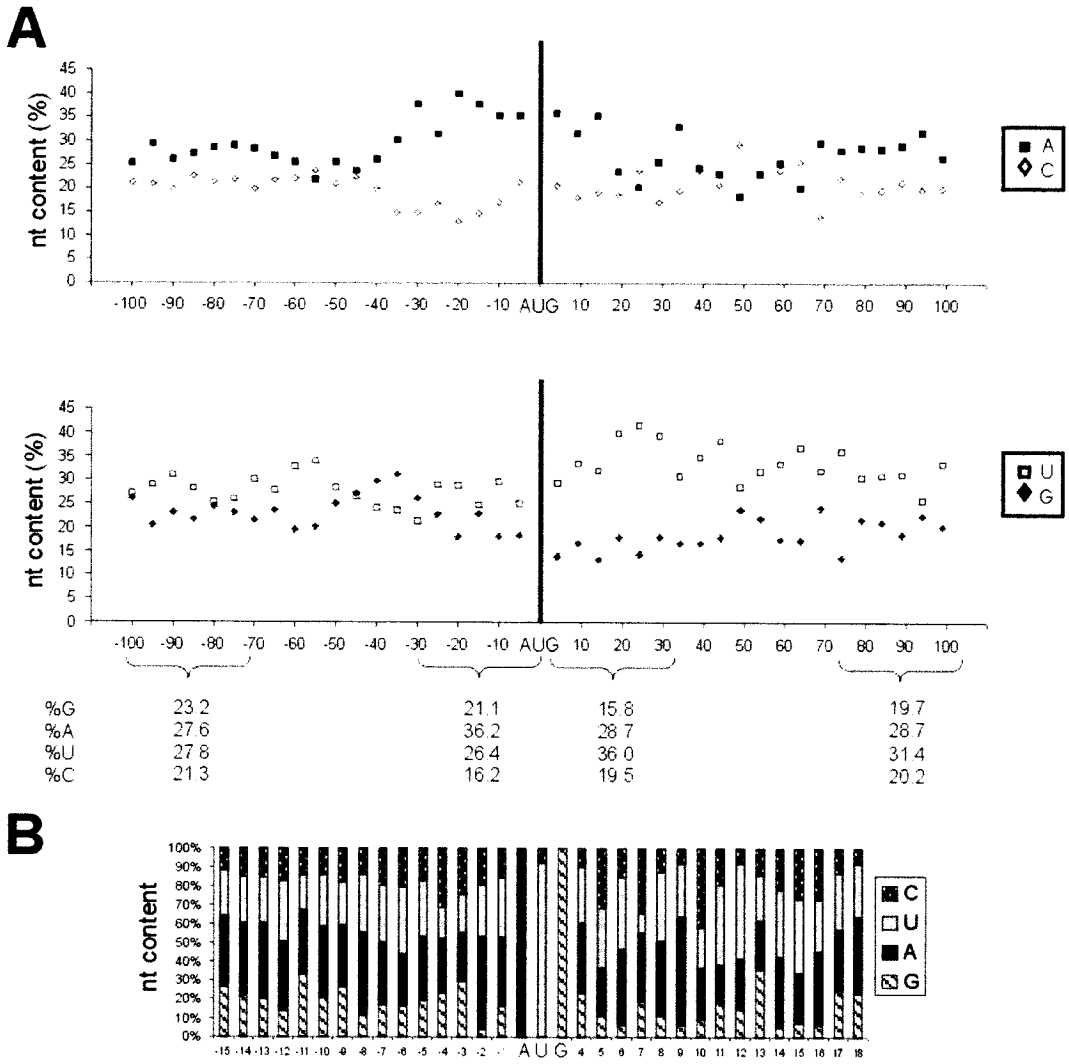

M nad2 AGAAGUUAUCGAAA--------CAC AUG
R AGAAGUUAUCGAAACUCGGAACCAC AUG
W AGAAGUUAUCGAAACUCGGAACCAC AUG
S AUUCGUUCGGAUCCUCU-----CAC AUG
T AUUCGUUCGGAUCCUCC-----CAC AUG
A AUUCGUUCGGAUCCUCC---CACAC AUG
B AUUCGUUCGGAUCCUCC-----CAC AUG
M nad5 UUUUGAAGAGACUCG-----AUCUU AUG
R UUUUGAAGAGACUCG-----AUCUU AUG
W UUUU-AAGAGACUCG-----AUCUU AUG
S GUUUGAUCACACUUG-----AAAUU AUG
T GUCUGAUCACACUCG-----AAAUU AUG
A UUUUGAUCACACUCGAAAUUAAAUU AUG
B UUCUGAUCA CACUCG-----AAAUU AUG
M nad6 GAUCUUUCAAAGAUGGGAAAUUCCA AUG
R AUUUCAAGGGAGGACGACG-UACC- AUG
W AUUUCAAGGGAGGACGACG-UACC- AUG
S AUUUCAGUGAAGGACGACG-UACC- AUG
T AUUUCAGGGAAGGACGACG-UACU- AUG
A AUUUCAGGGAAGGACGACGGUACC- AUG
B AUUUCAGGGAAGGACGACG-UACCC AUG
M cob GAUAGAAAAGAG----------AA AUG
R GAUAGAAAAGAG----------AA AUG
W GAUAGAAAAGAG----------AA AUG
S GAUAGAAAGAAUCUUGGCAAAGCAA AUG
T GAUAGAAAAGAU---AAAUAAAUAA AUG
A GAUAGGAAAGAG----------AA AUG
B GAUAGGAAAGAG----------AA AUG
M rps3 UAAAGUCUAAGCG--------ACAU AUG
R UAAAGUCUAAGCG--------ACAU AUG
W UAAAGUCUAAGCG--------ACAU AUG
S UCAAGUCUAAGCAAGCAUAUCAUAU AUG
T UAAAGUCUAAGCG---------CAU AUG
A UAAAGUCAAA----------AAAU AUG
B UAAAGUCAAA----------AAAU AUG
M rps12 GGGGAAGGACAUAGGAAAGAG--GG AUG
R GGGGAAGGACAUAGGAAAGAG--GG AUG
W GGGGAAGGACAUAGGAAAGAG--GG AUG
S GGGGAAGGACAAAGGAAAGAG--CG AUG
T GGGGAAGGACAAAGGAAAGAG--CG AUG
A GGGGAAAAAGAAAGGAAAGAGAAUA AUG
B GGGGGAAAAGAAAGGAAAGAGAAUA AUG

Figure 5.4 Nucleotide composition of regions extending 100 nt upstream and 100 nt downstream of initiation codons for the 164 plant mitochondrial sequences in our data set. [A] Upper scatterplot shows percent A (black squares) and C (open diamonds) whereas lower plot shows percent G (black diamonds) and U (open squares). Data points represent pooled values for 5 nt blocks and the position of initiation codon (+1 to +3) is indicated by a vertical line. Numerical values are given below for 30 nt stretches, namely from positions -71 to -100, -1 to -30, +4 to +33, and +74 to +103. [B] Bar graph showing nucleotide content flanking the start codons (15 nt stretches) with G [hatched], A [black], U [grey] and C [stippled]. Minor C content at +2 reflects editing sites which are converted from ACG to AUG.


A
nt content (%)
A
C
U
G
-100 -90 -80 -70 -60 -50 -40 -30 -20 -10 AUG 10 20 30 40 50 60 70 80 90 100
%G 23.2 21.1 15.8 19.7
%A 27.6 36.2 28.7 28.7
%U 27.8 26.4 36.0 31.4
%C 21.3 16.2 19.5 20.2
B
nt content
C
U
A
G
AUG

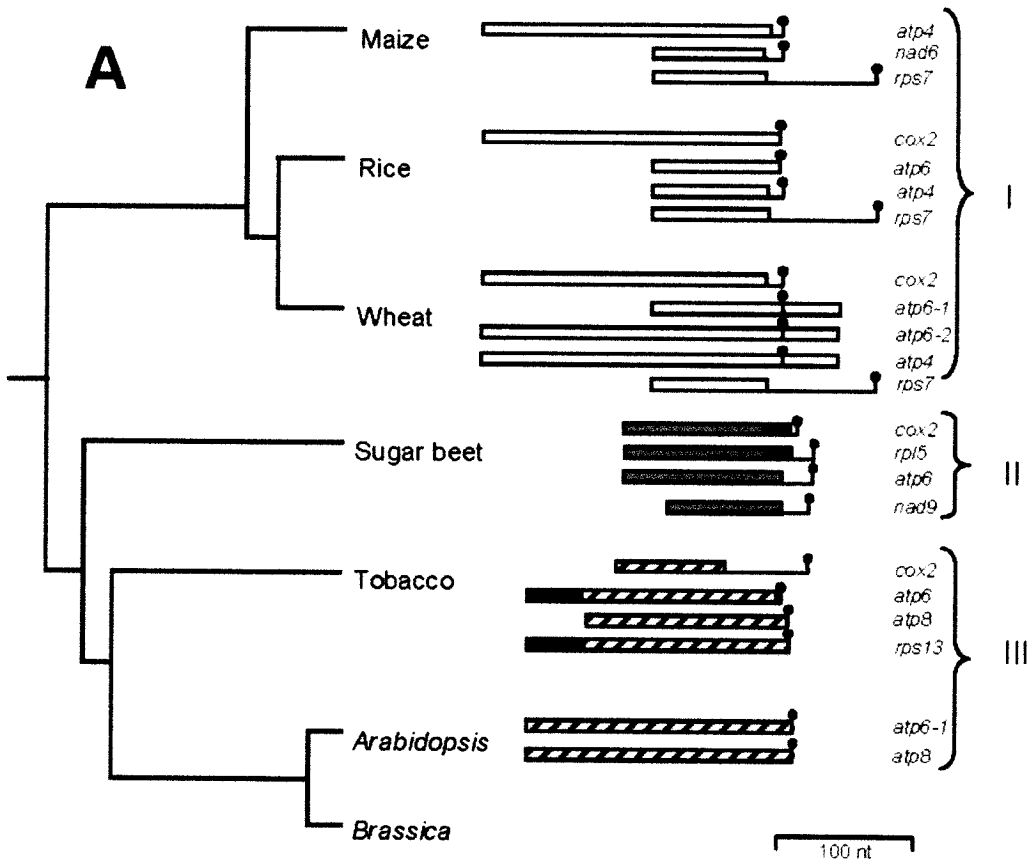
purine-rich Shine-Dalgarno motif in bacteria (Fig.5.4*B*).

5.4.3 Paralogues of "upstream sequences" preceding multiple genes

To explore sequence relationships within a given genome, we investigated dot plot signals which are off the main diagonal in Figure 5.1*A* (illustrated by open arrowheads) and thus represent sequences upstream of more than one gene in our data set. From previous work (reviewed in Bonen and Brown 1993), it has long been appreciated that stretches preceding genes are sometimes duplicated and copies are present upstream of other genes. They therefore potentially provide common regulatory signals rather than each gene having its own unique cis-elements. In our analysis, which included BLAST searches (bl2seq) querying individual genomes, a sequence was considered to be a member of an "upstream cassette" family if a copy longer than 50 nt preceding the initiation codon was present within 100 nt upstream of another known gene. Three such families, designated as types I-III, were identified (Fig. 5.5*A*, open, grey and hatched blocks, respectively) and alignments of sequences in the immediate vicinity of the initiation codon are shown in Figure 5.5*B*. Full sequence alignments are given in Supplementary Figure S2. It can be seen from Figure 5.5*A* that these "upstream cassette" families have up to five members within a genome (including incomplete copies) and they show lineage-specific distributions among the plants. Using BLAST searches, we found that some of these cassette-type sequences are also present in spacer regions (see below) or upstream of genes that had been excluded from our data set because they were not represented in all 7 plant species. The latter includes *rps7* in maize, rice and wheat (type I), *rpl5* in sugar beet (type II), and *rps13* in tobacco (type III) (Fig.5.5*A*, *B*).

Upstream cassette type I is present in all three cereals and its longest members extend ~200 nt upstream of the initiation codon and include a promoter (cf. Covello and Gray 1991). Interestingly, in the case of wheat *atp6-1*, *atp6-2* and *atp4*, homology continues ~50 nt further downstream (Fig.5.5*A*), although it should be noted that these sequences comprise part of an ORF fused to *atp6* (cf. Bonen and Brown, 1993) and do not contribute to the mature ATP6 protein. Cassette type I has a rather mosaic-like appearance in that certain stretches share greater sequence similarity than do others (Fig.5.5*B*, shaded, Supplementary Fig. S2). This was previously observed by Pring et al. (1992) who noted that one of the conserved blocks is coincident with the 5' end of the shorter members and thus consistent

Figure 5.5 Paralogous "upstream cassettes" preceding different protein-coding genes. [A] Schematic showing taxonomic relationships among the seven plants (at left) and the three families of "upstream cassettes" (designated as I-III) with names of associated genes (at right). Blocks depict cassette type I (open), type II (grey), and type III (hatched), and filled circles denote the locations of initiation codons. For type II, 3' regions of the cassette specific to *rpl5* and *cox2* are depicted by black fill. Note that *rps7*, *rpl5* and *rps13* were not in our primary data set because they are not present in all seven plant mitochondrial genomes. [B] Sequence alignments of the 3' regions of cassettes I-III. Homologous regions within a given plant are boxed, and for type I shading depicts identity among at least 9 out of 12 sequences. Dashed lines indicate gaps. An 8 nt motif common to certain members of cassette type I is shown in lowercase italics.



B

```
Type I
M atp4   UUUCUCUU---------GGAGCAGAGCUUCAUCAUAAAAGUGG AGAGUCACA AUG
M nad6   UUUCUCUUUU------GGGAGCAGAUCU----UUCAAAGAUGg gaaauucCA AUG
M rps7   UUUCUCUUUU------GGGAGCAGAGCU----CAAAAAGAUGG ACAGUAACG..50..GCUCggaaauucUCAGCUAU AUG
R cox2   UUUCU----------AGGAGCAGAGCU----AAAAAAGAUGg gaaauucCA AUG
R atp6   UUUCU----------AGGAGCAGAGCU----UAAAAAGAUGg gaaauucCA AUG
R atp4   UUUCUCUUUUGUUUU-GGGAGCAGAGCU----UAAAAGGAUAU AGUUACCCCACG AUG
R rps7   UUUCUCUUUUC-----GGGAGCAGAGCU----UAAAAAGAUGG ACAGUAACG..50..GCUCggaaauucUCAGCUAU AUG
W cox2   UUUCCUUUCUCUUUUCGGGAGCAGAGCU----CAAAAAGAUGg gaaauucCA AUG
W atp6-1 UUUC----AAAAGUGAGCGAGCAGAGCU----CAAAAAGAUGGAGUUACCUGGAG AUG AGAU..46..UAUU
W atp6-2 UUUC----AAAAGUGAGCGAGCAGAGCU----CAAAAAGAUGGAGUUACCUGGAG AUG AGAU..46..UAUU
W atp4   UUUC----AAAAGUGAGCGAGCAGAGCU----CAAAAAGAUGGAGUUACCUGGAG AUG AGAU..46..UAUU
W rps7   UUUCCUUUCUCUUUUCGGGAGCAGAGCU----CAAAAAGAUGG ACAGUAACG..50..GCUCggaaauucUCAGCUAU AUG

Type II
S cox2   UCAUUUCUCGAGAGAGCGGAGCAGUCAAAGAAUGAA CCAA AUG
S rpl5   UCAUUUCUCGAGAGAGC-----AGUCAAAGAAUGAA UAAGAAAAGCCAACUC AUG
S atp6   UCAUUUUUGAGCAAGAAGCGGAACUACA GGGAUGAAAUGAAAAGUGUUUAUU ACG
S nad9   UCAUUUUUGAGCAAGAAGCGGAACUACA AGAAAGCUUUCUUUCUCUUU AUG

Type III
T cox2   UCUCUCUUUUUUUUGGGGGGAGCAGAGCAGUCAAAGAAUGAACCAAACCAA AUG
T atp6   AUUGAGGGAAUCUUUUUAUC----UCAAUCA CA AUG
T atp8   AUUGAGGGAAUCGUUUUGUC----UCAAUCAAUCAAGA AUG
T rps13  AUUGAGGGAAUCGUUUUGUC----UCAAUCAAUCAAGA AUG
A atp6-1 CUGGAGGGAAUCAUUUUGUUGAAAUCAAUUAAUCCAAU- AUG
A atp8   CUGGAGGGAAUCAUUUUGUUGAAAUCAAUUAAUCUAAUC AUG
```

with a recombination site. Moreover certain members within a genome appear to have undergone gene conversion events in that stretches of greater similarity are seen among subsets (Fig.5.5*B*, Supplementary Fig. S2). In the case of rice *cox2* and *atp6*, the cassette extends precisely up to (but not past) the initiation codon, thus illustrating clean displacement of the pre-existing regulatory sequences and no apparent requirement for adaptation of downstream coding sequences. For *rps7* in all three cereals, approximately 90 nt of ancestral-type upstream sequence has been retained (Fig.5.5*A*, *B* and Supplementary Fig. S2), but interestingly, close to the *rps7* start codon there is an 8 nt motif (GGAAATTC) that also precedes the start codons of wheat and rice *cox2*, rice *atp6*, and maize *nad6* (Fig.5.5*B*, lowercase italics). Because of the location, it is a candidate for playing a role in start codon selection. Alternatively, it might be involved in other RNA level processing/stability or regulatory events. A search for this motif using FUZZNUC (and allowing for 2 nt degeneracy) revealed 23 other instances within our data set which are located within 20 nt of the start codon.

Cassette type II (~120 nt in length) is found only in the sugar beet mitochondrial genome (Fig.5.5*A*, supplementary Fig. S2), and the origin of its four members can be traced to *nad9*-type upstream sequences (Fig. S1). This cassette family can be divided into two sub-types because the proximal part of the *cox2* and *rpl5* copies (Fig.5.5*A*, black blocks) are non-homologous to those shared between *atp6* and *nad9* (Fig.5.5*B*). This is again suggestive of either gene conversion or difference in timing of duplication events. The eudicot cassette type III (~ 180 nt long) is well-conserved over its length (Fig.5.5*B*, Supplementary Fig. S2) although tobacco *cox2* has just a partial copy (Fig.5.5*A*, *B*) and this cassette type appears to have originated from eudicot *atp8*-type sequences (Fig.5.5*A* and supplementary Figs. S1, S2). Notably all members of cassette III (except the truncated tobacco *cox2* one) break in homology precisely at the position of the initiation codon (as was seen above for certain cassette I members). Although no duplicated cassette type III sequences were found upstream of other *Brassica* genes, two *atp8*-type copies are found in genomic spacer regions (see below). There is a fourth minor upstream repeat family in maize mitochondria (data not shown), in that *cox1* and *rps2B* are preceded by a common ~45 nt stretch plus an additional 13 nt extending into the coding region, and thus reminiscent of the wheat cassette I case described above.

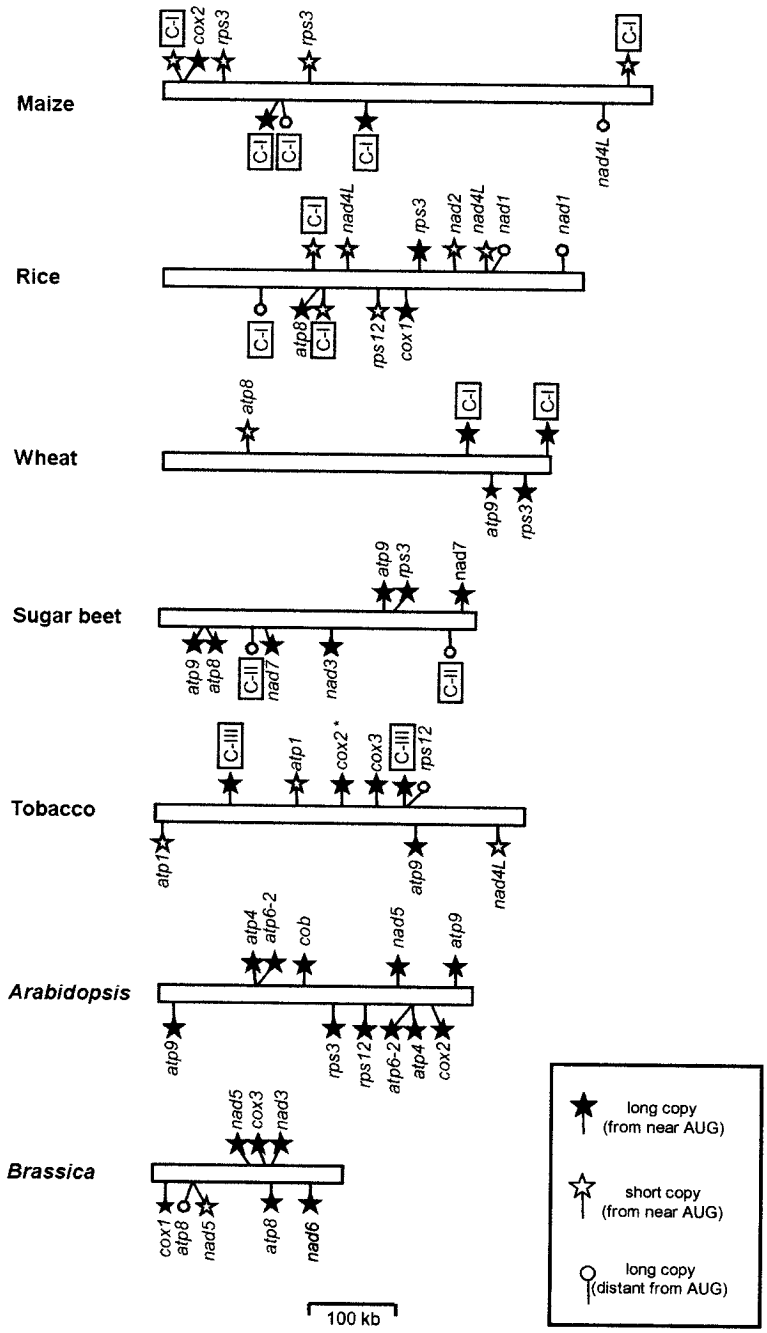
5.4.4 Additional copies of "upstream sequences" in spacer regions

We extended the BLAST (bl2seq) analysis of individual genomes to search for sequences similar to those in our data set that were more distantly located upstream (>100 nt) or downstream from a known gene and in Figure 6 their locations are schematically shown along the linearized master chromosomes for each of the seven plants. Sequences similar to the "upstream cassettes" discussed in the previous section are designated as C-I, C-II and C-III. We found 63 paralogues, which collectively represent almost every gene in our data set, with the exception of *ccm* genes which also are distinct in having among the most conservative upstream regions. Sequences were classified based on length and position relative to the start codon of the query sequence: (i) long near (Fig.5.6, black stars), (ii) short near (Fig.5.6, open stars) or (iii) long distant (Fig.5.6, open circles) as described in more detail in the Figure 5.6 legend. Approximately two-thirds are within 1 kb of a known gene and strikingly, all are in the same orientation as that gene. Thus if actively-transcribed, they might contribute regulatory cis-elements, such as for RNA processing or stability. Of the 54 in categories (i) and (ii) (Fig.5.6, black and open stars), 37 extend to or past the start codon and thus potentially include translation initiation signals. For the total set (which excludes any within 100 nt upstream of a known protein-coding gene because they were discussed above), three-quarters lie downstream of a gene and interestingly, most of those located within 100 nt (that is, 7 out of 9) are linked to genes encoding tRNAs or rRNAs.

The genomic density of these intergenic copies varies among the seven plants, with *Arabidopsis* being the highest (and representing a diversity of genes) and wheat the lowest. Moreover all of the copies in *Arabidopsis* are long (Fig.5.6, black stars) and this might reflect recent duplication without subsequent fragmentation by rearrangement. The master chromosome in *Arabidopsis* contains only two recombinational repeat elements (Unseld et al.1997) whereas wheat is reported to have 16 repeated regions (Ogihara et al. 2005). It will be of interest to learn more about the interplay between the generation/eradication vs. functional recruitment of intergenic copies. While copies were generally dispersed throughout each genome, there were cases of close linkage. For example in *Arabidopsis,* the *atp4* and *atp6-2* elements are less than 200 bp apart, and they are located on a recombinational repeat element.

Over half of these duplicated upstream sequences also include part of the coding

Figure 5.6 Locations of additional copies of "upstream sequences" in the seven plant mitochondrial genomes. Schematics of the linearized master chromosomes for maize (570 kb), rice (491 kb), wheat (453 kb), sugar beet (369 kb), tobacco (430 kb), *Arabidopsis* (367 kb) and *Brassica* (222 kb) show positions of paralogues of upstream sequences in our data set. Designation is according to the protein-coding gene with which each is associated, or as C-I to C-III (boxed) if related to "upstream cassette" types I-III (see Fig.5.5). Note that cassettes shown in Figure 5 are not included because they are located close to known genes. Black stars indicate long (>40 nt) copies that originated from regions of the element occurring near (within 15 nt of) the initiation codon. Open stars indicate short (<20 nt) copies that originated near the start codon. Open circles indicate long copies (> 40 nt) that originated further away (>15 nt) from the start codon. Copies shown above or below the bar reflect orientation in the master chromosome. The tobacco *cox2* upstream copy (asterisk) originated from part of the element not homologous to cassette III (see Fig.5.5).


Maize
C-I
cox2
rps3
rps3
C-I
C-I
C-I
C-I
nad4L
Rice
C-I
nad4L
rps3
nad2
nad4L
nad1
nad1
C-I
atp8
C-I
rps12
cox1
Wheat
atp8
C-I
C-I
atp9
rps3
Sugar beet
atp9
rps3
nad7
atp9
atp8
C-II
nad7
nad3
C-II
Tobacco
C-III
atp1
cox2
cox3
C-III
rps12
atp1
atp9
nad4L
Arabidopsis
atp4
atp6-2
cob
nad5
atp9
atp9
rps3
rps12
atp6-2
atp4
cox2
Brassica
nad5
cox3
nad3
cox1
atp6
nad5
atp8
nad6
100 kb
long copy
(from near AUG)
short copy
(from near AUG)
long copy
(distant from AUG)

sequence with which it was originally associated. In 15 cases, this has resulted in the seemingly-fortuitous creation of mosaic reading frames which have been annotated as ORFs in the mitochondrial complete genome databank entries. For example, the 5' end of *orf315* in *Arabidopsis* is composed of a duplicated *atp9* coding fragment (of ~40 codons) which is preceded by 209 nt of *atp9*-type upstream sequence. Interestingly, several of these chimeric ORFs are among ones identified to be expressed at high levels in *Arabidopsis* plants which are down-regulated for polynucleotide phosphorylase (Holec et al. 2006). This suggests that *atp9* upstream sequences may contribute to transcription and/or RNA processing/stability elements. Similarly, in sugar beet with normal-type cytoplasm, *orf246* was generated in part by the tandem duplication of a short *rps3* coding segment preceded by 154 nt of its associated upstream sequence. This *orf246* is actively transcribed, although no translation products were detected (Satoh et al. 2004).

5.5 Discussion

Plant mitochondrial genomes exhibit a striking degree of plasticity in organization, even within regions located very close to protein-coding genes, and this is illustrated in our examination of sequences immediately preceding translation initiation codons. Relatively few of the 23 genes in our data set showed strong, uninterrupted conservation in upstream sequences among all seven plants examined, and in many cases sequences were clearly non-homologous. The switching of upstream regulatory sequences via DNA rearrangements has occurred frequently during flowering plant evolution, and genes encoding ATP synthase (*atp*) subunits are particularly dynamic. In addition, certain newly-acquired upstream sequences are duplicate copies that also precede another gene (or genes) within a given genome. Our analysis focused on data from the completely-sequenced mitochondrial genomes of seven plant species, whose divergence times range from ~20 Mya to ~150 Mya, but the phenomenon of genomic fluidity impacting on upstream regulatory regions is also evident within a given species. For example, between the sugar beet "Owen CMS" and normal-type mitochondrial genomes (Satoh et al. 2004), sequences preceding *cox2* and *atp1* each show a break in homology within ~50 nt of their respective start codons, and in the case of *atp8*, upstream sequences are completely unrelated in the two cytoplasms.

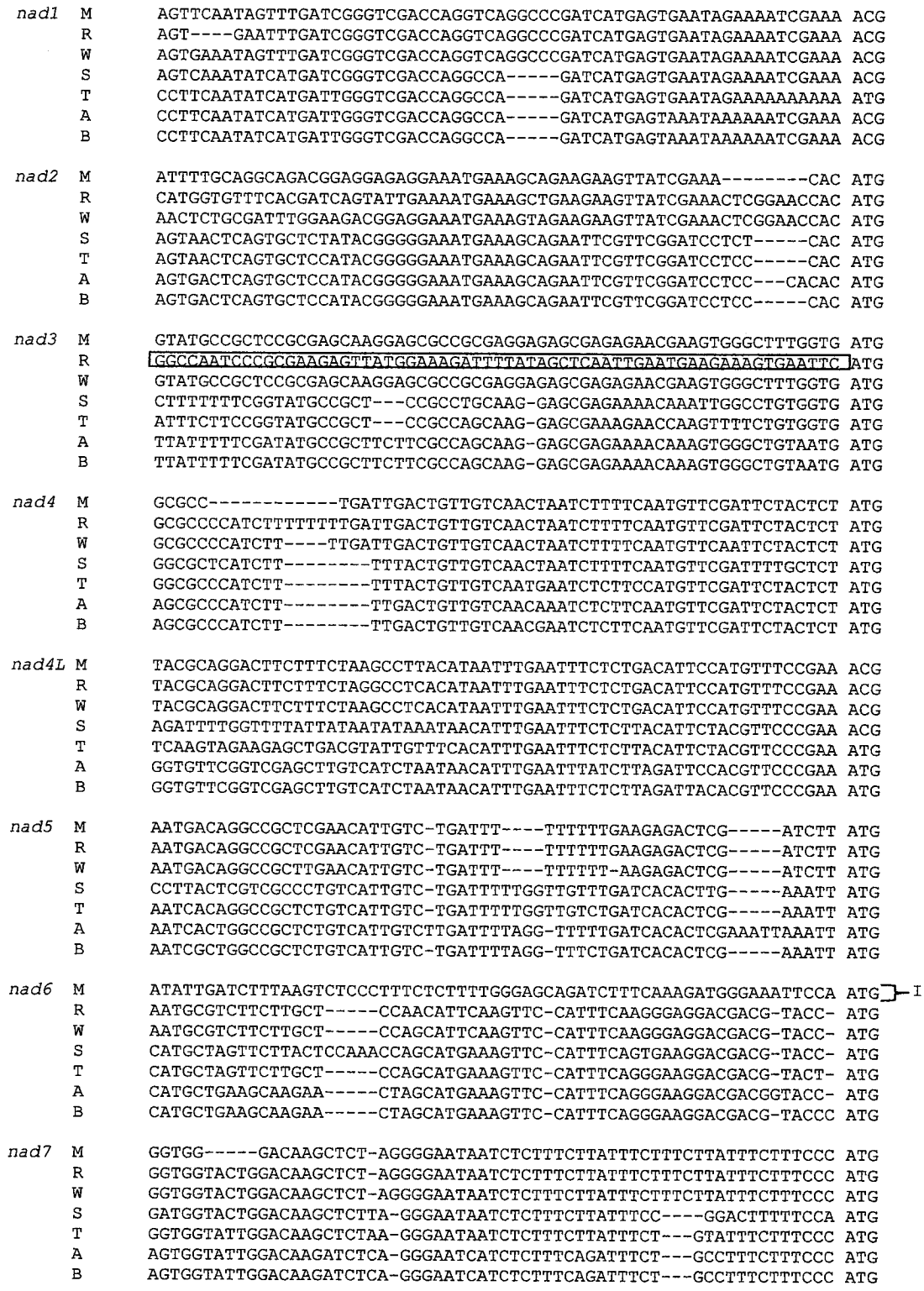
The presence of additional "upstream sequence" copies within the intergenic spacers of plant mitochondrial genomes provides an opportunity for their future recruitment as regulatory elements as well as the *de novo* creation of transcripts which might even be translated into novel polypeptides. Indeed this is a hallmark of certain types of cytoplasmic male sterility (reviewed in Hanson and Bentolila 2004), whereby detrimental proteins are generated from novel chimeric ORFs. In this regard, it is notable that *atp* sequences (both regulatory and coding) are particularly frequent contributors to the creation of such ORFs. In a recent study of CMS in rice with Boro II cytoplasm, fertility restoration was found to be mediated by pentatricopeptide repeat (PPR) type proteins which act at the mRNA turnover level to down-regulate cytotoxic *orf79* (Wang et al. 2006). This chimeric *orf79* includes a short *cox1* coding segment fused behind *cox2*-type upstream sequences (i.e. cassette type I in Figure 5.5), and that in turn is preceded by *atp6* and its associated type I upstream sequence.

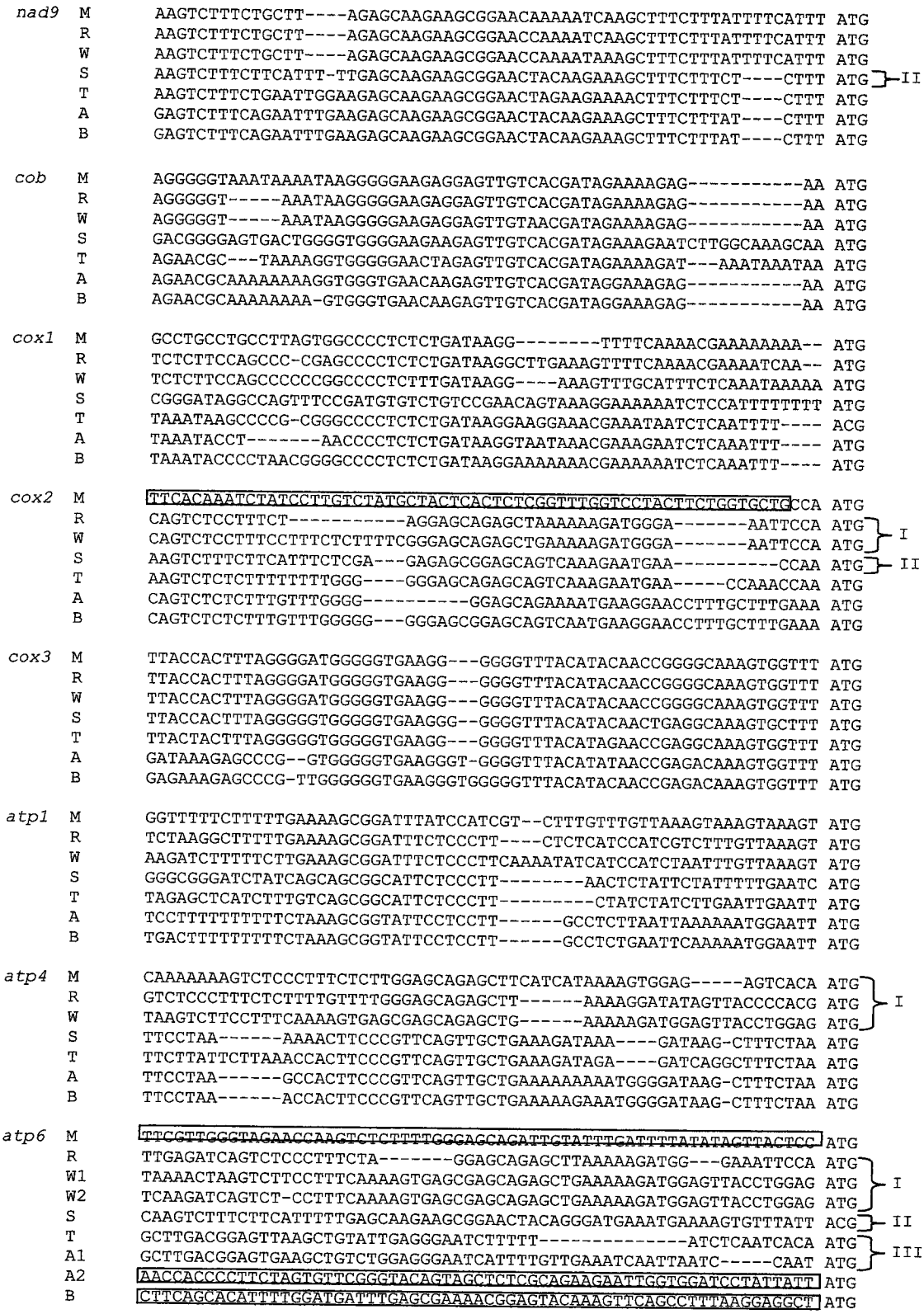
The dynamic nature of sequences preceding protein-coding genes also has implications for translational signals, and at the present time it is unclear how the correct AUG is recognized by the ribosomal machinery. Our analysis has revealed that even though the sequences preceding initiation codons show a bias for adenosine, there do not appear to be conserved purine-rich blocks complementary to the 3' terminal region of the SSU rRNA (and thus candidate elements for a bacterial-type translation initiator recognition), nor was any other consensus sequence evident. These observations in conjunction with the presence of indel sequence variation very close to initiation codons suggest that a classical Shine-Dalgarno type interaction is unlikely to play a major role in plant mitochondrial translation initiation. Moreover, the function of the S1 ribosomal protein in *E. coli* to stabilize the mRNA at the initiation codon (Sorensen, Fricke and Pedersen 1998) is excluded in that the plant mitochondrial S1 counterpart lacks the corresponding carboxy-terminal RNA binding domains (cf. Mundel and Schuster, 1996).

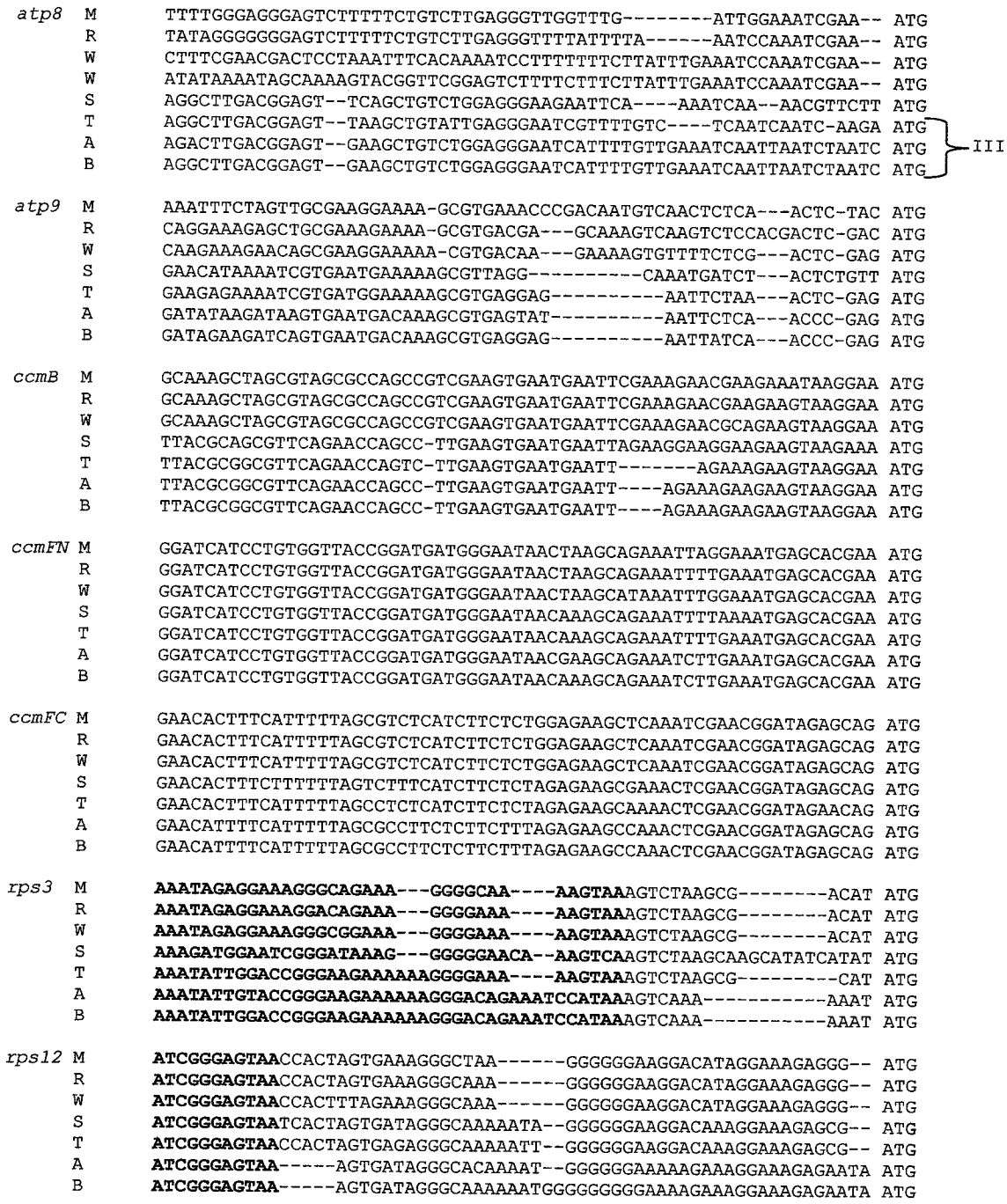
The ability of upstream sequences to be shuffled among protein-coding genes in plant mitochondria, as exemplified by the three families of lineage-specific cassettes we have described, is rather reminiscent of behaviour seen in yeast mitochondrial pseudorevertants. For example, respiratory function was seen to be restored through DNA duplication/rearrangements whereby defective *cox1* upstream sequences were replaced by a *cob* leader (Manthey and McEwen 1995). In yeast mitochondria, 5'UTRs are typically long

and they possess cis-elements which interact with gene-specific translational activators and localize protein synthesis on the mitochondrial inner membrane (reviewed in Costanzo and Fox 1990). Because of the added complexity of events such as RNA editing in plant mitochondria, it is possible that RNA binding proteins (such as PPR proteins cf. Andrés, Lurin and Small 2007) might perform multiple roles, and this could contribute to an evolutionarily-dynamic translation initiation recognition system with independent gene-specific signals. It will be important to have a greater understanding of translational signals in plant mitochondria to better appreciate the potential impact (both beneficial and detrimental) of highly rearranging genomes.

Supplementary Material S1- Clustal alignment of nucleotide sequences upstream (-1 to <-50) of plant mitochondrial genes included in the analyses. Alignments were manually corrected where appropriate with the caveat that homology was difficult to assess in some regions. Upstream sequences that are clearly unrelated to others in our data set are boxed, and upstream cassettes are indicated numerically (I-III). The *rps19* or *Ψrps19* sequence upstream of *rps3*, and the *nad3* coding sequence upstream of *rps12* are shown in bold. M=Maize, R=rice, W=wheat, S=sugar beet, T=tobacco, A=*Arabidopsis* and B=*Brassica.*


nad1 M AGTTCAATAGTTTGATCGGGTCGACCAGGTCAGGCCCGATCATGAGTGAATAGAAAATCGAAA ACG
R AGT----GAATTTGATCGGGTCGACCAGGTCAGGCCCGATCATGAGTGAATAGAAAATCGAAA ACG
W AGTGAAATAGTTTGATCGGGTCGACCAGGTCAGGCCCGATCATGAGTGAATAGAAAATCGAAA ACG
S AGTCAAATATCATGATCGGGTCGACCAGGCCA-----GATCATGAGTGAATAGAAAATCGAAA ACG
T CCTTCAATATCATGATTGGGTCGACCAGGCCA-----GATCATGAGTGAATAGAAAAAAAAAA ATG
A CCTTCAATATCATGATTGGGTCGACCAGGCCA-----GATCATGAGTAAATAAAAAATCGAAA ACG
B CCTTCAATATCATGATTGGGTCGACCAGGCCA-----GATCATGAGTAAATAAAAAATCGAAA ACG
nad2 M ATTTTGCAGGCAGACGGAGGAGAGGAAATGAAAGCAGAAGAAGTTATCGAAA--------CAC ATG
R CATGGTGTTTCACGATCAGTATTGAAAATGAAAGCTGAAGAAGTTATCGAAACTCGGAACCAC ATG
W AACTCTGCGATTTGGAAGACGGAGGAAATGAAAGTAGAAGAAGTTATCGAAACTCGGAACCAC ATG
S AGTAACTCAGTGCTCTATACGGGGGAAATGAAAGCAGAATTCGTTCGGATCCTCT-----CAC ATG
T AGTAACTCAGTGCTCCATACGGGGGAAATGAAAGCAGAATTCGTTCGGATCCTCC-----CAC ATG
A AGTGACTCAGTGCTCCATACGGGGGAAATGAAAGCAGAATTCGTTCGGATCCTCC---CACAC ATG
B AGTGACTCAGTGCTCCATACGGGGGAAATGAAAGCAGAATTCGTTCGGATCCTCC-----CAC ATG
nad3 M GTATGCCGCTCCGCGAGCAAGGAGCGCCGCGAGGAGAGCGAGAGAACGAAGTGGGCTTTGGTG ATG
R GGCCAATCCCGCGAAGAGTTATGGAAAGATTTTATAGCTCAATTGAATGAAGAAAGTGAATTC ATG
W GTATGCCGCTCCGCGAGCAAGGAGCGCCGCGAGGAGAGCGAGAGAACGAAGTGGGCTTTGGTG ATG
S CTTTTTTTCGGTATGCCGCT---CCGCCTGCAAG-GAGCGAGAAAACAAATTGGCCTGTGGTG ATG
T ATTTCTTCCGGTATGCCGCT---CCGCCAGCAAG-GAGCGAAAGAACCAAGTTTTCTGTGGTG ATG
A TTATTTTTCGATATGCCGCTTCTTCGCCAGCAAG-GAGCGAGAAAACAAAGTGGGCTGTAATG ATG
B TTATTTTTCGATATGCCGCTTCTTCGCCAGCAAG-GAGCGAGAAAACAAAGTGGGCTGTAATG ATG
nad4 M GCGCC------------TGATTGACTGTTGTCAACTAATCTTTTCAATGTTCGATTCTACTCT ATG
R GCGCCCCATCTTTTTTTTGATTGACTGTTGTCAACTAATCTTTTCAATGTTCGATTCTACTCT ATG
W GCGCCCCATCTT----TTGATTGACTGTTGTCAACTAATCTTTTCAATGTTCAATTCTACTCT ATG
S GGCGCTCATCTT--------TTTACTGTTGTCAACTAATCTTTTCAATGTTCGATTTTGCTCT ATG
T GGCGCCCATCTT--------TTTACTGTTGTCAATGAATCTCTTCCATGTTCGATTCTACTCT ATG
A AGCGCCCATCTT--------TTGACTGTTGTCAACAAATCTCTTCAATGTTCGATTCTACTCT ATG
B AGCGCCCATCTT--------TTGACTGTTGTCAACGAATCTCTTCAATGTTCGATTCTACTCT ATG
nad4L M TACGCAGGACTTCTTTCTAAGCCTTACATAATTTGAATTTCTCTGACATTCCATGTTTCCGAA ACG
R TACGCAGGACTTCTTTCTAGGCCTCACATAATTTGAATTTCTCTGACATTCCATGTTTCCGAA ACG
W TACGCAGGACTTCTTTCTAAGCCTCACATAATTTGAATTTCTCTGACATTCCATGTTTCCGAA ACG
S AGATTTTGGTTTTATTATAATATAAATAACATTTGAATTTCTCTTACATTCTACGTTCCCGAA ACG
T TCAAGTAGAAGAGCTGACGTATTGTTTCACATTTGAATTTCTCTTACATTCTACGTTCCCGAA ATG
A GGTGTTCGGTCGAGCTTGTCATCTAATAACATTTGAATTTATCTTAGATTCCACGTTCCCGAA ATG
B GGTGTTCGGTCGAGCTTGTCATCTAATAACATTTGAATTTCTCTTAGATTACACGTTCCCGAA ATG
nad5 M AATGACAGGCCGCTCGAACATTGTC-TGATTT----TTTTTTGAAGAGACTCG-----ATCTT ATG
R AATGACAGGCCGCTCGAACATTGTC-TGATTT----TTTTTTGAAGAGACTCG-----ATCTT ATG
W AATGACAGGCCGCTTGAACATTGTC-TGATTT----TTTTTT-AAGAGACTCG-----ATCTT ATG
S CCTTACTCGTCGCCCTGTCATTGTC-TGATTTTTGGTTGTTTGATCACACTTG-----AAATT ATG
T AATCACAGGCCGCTCTGTCATTGTC-TGATTTTTGGTTGTCTGATCACACTCG-----AAATT ATG
A AATCACTGGCCGCTCTGTCATTGTCTTGATTTTAGG-TTTTTGATCACACTCGAAATTAAATT ATG
B AATCGCTGGCCGCTCTGTCATTGTC-TGATTTTAGG-TTTCTGATCACACTCG-----AAATT ATG
nad6 M ATATTGATCTTTAAGTCTCCCTTTCTCTTTTGGGAGCAGATCTTTCAAAGATGGGAAATTCCA ATG I
R AATGCGTCTTCTTGCT-----CCAACATTCAAGTTC-CATTTCAAGGGAGGACGACG-TACC- ATG
W AATGCGTCTTCTTGCT-----CCAGCATTCAAGTTC-CATTTCAAGGGAGGACGACG-TACC- ATG
S CATGCTAGTTCTTACTCCAAACCAGCATGAAAGTTC-CATTTCAGTGAAGGACGACG-TACC- ATG
T CATGCTAGTTCTTGCT-----CCAGCATGAAAGTTC-CATTTCAGGGAAGGACGACG-TACT- ATG
A CATGCTGAAGCAAGAA-----CTAGCATGAAAGTTC-CATTTCAGGGAAGGACGACGGTACC- ATG
B CATGCTGAAGCAAGAA-----CTAGCATGAAAGTTC-CATTTCAGGGAAGGACGACG-TACCC ATG
nad7 M GGTGG-----GACAAGCTCT-AGGGGAATAATCTCTTTCTTATTTCTTTCTTATTTCTTTCCC ATG
R GGTGGTACTGGACAAGCTCT-AGGGGAATAATCTCTTTCTTATTTCTTTCTTATTTCTTTCCC ATG
W GGTGGTACTGGACAAGCTCT-AGGGGAATAATCTCTTTCTTATTTCTTTCTTATTTCTTTCCC ATG
S GATGGTACTGGACAAGCTCTTA-GGGAATAATCTCTTTCTTATTTCC----GGACTTTTTCCA ATG
T GGTGGTATTGGACAAGCTCTAA-GGGAATAATCTCTTTCTTATTTCT---GTATTTCTTTCCC ATG
A AGTGGTATTGGACAAGATCTCA-GGGAATCATCTCTTTCAGATTTCT---GCCTTTCTTTCCC ATG
B AGTGGTATTGGACAAGATCTCA-GGGAATCATCTCTTTCAGATTTCT---GCCTTTCTTTCCC ATG


nad9 M AAGTCTTTCTGCTT----AGAGCAAGAAGCGGAACAAAAATCAAGCTTTCTTTATTTTCATTT ATG
R AAGTCTTTCTGCTT----AGAGCAAGAAGCGGAACCAAAATCAAGCTTTCTTTATTTTCATTT ATG
W AAGTCTTTCTGCTT----AGAGCAAGAAGCGGAACCAAAATAAAGCTTTCTTTATTTTCATTT ATG
S AAGTCTTTCTTCATTT-TTGAGCAAGAAGCGGAACTACAAGAAAGCTTTCTTTCT----CTTT ATG II
T AAGTCTTTCTGAATTGGAAGAGCAAGAAGCGGAACTAGAAGAAAACTTTCTTTCT----CTTT ATG
A GAGTCTTTCAGAATTTGAAGAGCAAGAAGCGGAACTACAAGAAAGCTTTCTTTAT----CTTT ATG
B GAGTCTTTCAGAATTTGAAGAGCAAGAAGCGGAACTACAAGAAAGCTTTCTTTAT----CTTT ATG
cob M AGGGGGTAAATAAAATAAGGGGGAAGAGGAGTTGTCACGATAGAAAAGAG----------AA ATG
R AGGGGGT-----AAATAAGGGGGAAGAGGAGTTGTCACGATAGAAAAGAG----------AA ATG
W AGGGGGT-----AAATAAGGGGGAAGAGGAGTTGTAACGATAGAAAAGAG----------AA ATG
S GACGGGGAGTGACTGGGGTGGGGAAGAAGAGTTGTCACGATAGAAAGAATCTTGGCAAAGCAA ATG
T AGAACGC---TAAAAGGTGGGGGAACTAGAGTTGTCACGATAGAAAAGAT---AAATAAATAA ATG
A AGAACGCAAAAAAAAGGTGGGTGAACAAGAGTTGTCACGATAGGAAAGAG----------AA ATG
B AGAACGCAAAAAAAA-GTGGGTGAACAAGAGTTGTCACGATAGGAAAGAG----------AA ATG
cox1 M GCCTGCCTGCCTTAGTGGCCCCTCTCTGATAAGG--------TTTTCAAAACGAAAAAAAA-- ATG
R TCTCTTCCAGCCC-CGAGCCCCTCTCTGATAAGGCTTGAAAGTTTTCAAAACGAAAATCAA-- ATG
W TCTCTTCCAGCCCCCCGGCCCCTCTTTGATAAGG----AAAGTTTGCATTTCTCAAATAAAAA ATG
S CGGGATAGGCCAGTTTCCGATGTGTCTGTCCGAACAGTAAAGGAAAAAATCTCCATTTTTTTT ATG
T TAAATAAGCCCCG-CGGGCCCCTCTCTGATAAGGAAGGAAACGAAATAATCTCAATTTT---- ACG
A TAAATACCT--------AACCCCTCTCTGATAAGGTAATAAACGAAAGAATCTCAAATTT---- ATG
B TAAATACCCCTAACGGGGCCCCTCTCTGATAAGGAAAAAAACGAAAAAATCTCAAATTT---- ATG
cox2 M TTCACAAATCTATCCTTGTCTATGCTACTCACTCTCGGTTTGGTCCTACTTCTGGTGCTGCCA ATG
R CAGTCTCCTTTCT----------AGGAGCAGAGCTAAAAAAGATGGGA-------AATTCCA ATG I
W CAGTCTCCTTTCCTTTCTCTTTTCGGGAGCAGAGCTGAAAAAGATGGGA-------AATTCCA ATG I
S AAGTCTTTCTTCATTTCTCGA---GAGAGCGGAGCAGTCAAAGAATGAA----------CCAA ATG II
T AAGTCTCTCTTTTTTTTGGG----GGGAGCAGAGCAGTCAAAGAATGAA-----CCAAACCAA ATG
A CAGTCTCTCTTTGTTTGGGG----------GGAGCAGAAAATGAAGGAACCTTTGCTTTGAAA ATG
B CAGTCTCTCTTTGTTTGGGGG---GGGAGCGGAGCAGTCAATGAAGGAACCTTTGCTTTGAAA ATG
cox3 M TTACCACTTTAGGGGATGGGGGTGAAGG---GGGGTTTACATACAACCGGGGCAAAGTGGTTT ATG
R TTACCACTTTAGGGGATGGGGGTGAAGG---GGGGTTTACATACAACCGGGGCAAAGTGGTTT ATG
W TTACCACTTTAGGGGATGGGGGTGAAGG---GGGGTTTACATACAACCGGGGCAAAGTGGTTT ATG
S TTACCACTTTAGGGGGTGGGGGTGAAGGG--GGGGTTTACATACAACTGAGGCAAAGTGCTTT ATG
T TTACTACTTTAGGGGGTGGGGGTGAAGG---GGGGTTTACATAGAACCGAGGCAAAGTGGTTT ATG
A GATAAAGAGCCCG--GTGGGGGTGAAGGGT-GGGGTTTACATATAACCGAGACAAAGTGGTTT ATG
B GAGAAAGAGCCCG-TTGGGGGGTGAAGGGTGGGGGTTTACATACAACCGAGACAAAGTGGTTT ATG
atp1 M GGTTTTTCTTTTTGAAAAGCGGATTTATCCATCGT--CTTTGTTTGTTAAAGTAAAGTAAAGT ATG
R TCTAAGGCTTTTTGAAAAGCGGATTTCTCCCTT----CTCTCATCCATCGTCTTTGTTAAAGT ATG
W AAGATCTTTTTCTTGAAAGCGGATTTCTCCCTTCAAAATATCATCCATCTAATTTGTTAAAGT ATG
S GGGCGGGATCTATCAGCAGCGGCATTCTCCCTT--------AACTCTATTCTATTTTTGAATC ATG
T TAGAGCTCATCTTTGTCAGCGGCATTCTCCCTT---------CTATCTATCTTGAATTGAATT ATG
A TCCTTTTTTTTTTTCTAAAGCGGTATTCCTCCTT------GCCTCTTAATTAAAAAATGGAATT ATG
B TGACTTTTTTTTTTCTAAAGCGGTATTCCTCCTT------GCCTCTGAATTCAAAAATGGAATT ATG
atp4 M CAAAAAAAGTCTCCCTTTCTCTTGGAGCAGAGCTTCATCATAAAAGTGGAG-----AGTCACA ATG I
R GTCTCCCTTTCTCTTTTGTTTTGGGAGCAGAGCTT------AAAAGGATATAGTTACCCCACG ATG I
W TAAGTCTTCCTTTCAAAAGTGAGCGAGCAGAGCTG------AAAAAGATGGAGTTACCTGGAG ATG I
S TTCCTAA------AAAACTTCCCGTTCAGTTGCTGAAAGATAAA----GATAAG-CTTTCTAA ATG
T TTCTTATTCTTAAACCACTTCCCGTTCAGTTGCTGAAAGATAGA----GATCAGGCTTTCTAA ATG
A TTCCTAA------GCCACTTCCCGTTCAGTTGCTGAAAAAAAAATGGGGATAAG-CTTTCTAA ATG
B TTCCTAA------ACCACTTCCCGTTCAGTTGCTGAAAAAGAAATGGGGATAAG-CTTTCTAA ATG
atp6 M TTCGTTGGGTAGAACCAAGTCTCTTTTGGGAGCAGATTGTATTTGATTTTATATAGTTACTCC ATG
R TTGAGATCAGTCTCCCTTTCTA--------GGAGCAGAGCTTAAAAAGATGG---GAAATTCCA ATG I
W1 TAAAACTAAGTCTTCCTTTCAAAAGTGAGCGAGCAGAGCTGAAAAAGATGGAGTTACCTGGAG ATG I
W2 TCAAGATCAGTCT-CCTTTCAAAAGTGAGCGAGCAGAGCTGAAAAAGATGGAGTTACCTGGAG ATG I
S CAAGTCTTTCTTCATTTTTGAGCAAGAAGCGGAACTACAGGGATGAAATGAAAAGTGTTTATT ACG II
T GCTTGACGGAGTTAAGCTGTATTGAGGGAATCTTTTT----------ATCTCAATCACA ATG III
A1 GCTTGACGGAGTGAAGCTGTCTGGAGGGAATCATTTTGTTGAAATCAATTAATC-----CAAT ATG III
A2 AACCACCCCTTCTAGTGTTCGGGTACAGTAGCTCTCGCAGAAGAATTGGTGGATCCTATTATT ATG
B CTTCAGCACATTTTGGATGATTTGAGCGAAAACGGAGTACAAAGTTCAGCCTTTAAGGAGGCT ATG


atp8 M TTTTGGGAGGGAGTCTTTTTCTGTCTTGAGGGTTGGTTTG--------ATTGGAAATCGAA-- ATG
R TATAGGGGGGGAGTCTTTTTCTGTCTTGAGGGTTTTATTTTA------AATCCAAATCGAA-- ATG
W CTTTCGAACGACTCCTAAATTTCACAAAATCCTTTTTTTTCTTATTTGAAATCCAAATCGAA-- ATG
W ATATAAAATAGCAAAAGTACGGTTCGGAGTCTTTTCTTTCTTATTTGAAATCCAAATCGAA-- ATG
S AGGCTTGACGGAGT--TCAGCTGTCTGGAGGGAAGAATTCA----AAATCAA--AACGTTCTT ATG
T AGGCTTGACGGAGT--TAAGCTGTATTGAGGGAATCGTTTTGTC----TCAATCAATC-AAGA ATG
A AGACTTGACGGAGT--GAAGCTGTCTGGAGGGAATCATTTTGTTGAAATCAATTAATCTAATC ATG
B AGGCTTGACGGAGT--GAAGCTGTCTGGAGGGAATCATTTTGTTGAAATCAATTAATCTAATC ATG
III
atp9 M AAATTTCTAGTTGCGAAGGAAAA-GCGTGAAACCCGACAATGTCAACTCTCA---ACTC-TAC ATG
R CAGGAAAGAGCTGCGAAAGAAAA-GCGTGACGA---GCAAAGTCAAGTCTCCACGACTC-GAC ATG
W CAAGAAAGAACAGCGAAGGAAAAA-CGTGACAA---GAAAAGTGTTTTCTCG---ACTC-GAG ATG
S GAACATAAAATCGTGAATGAAAAAGCGTTAGG----------CAAATGATCT---ACTCTGTT ATG
T GAAGAGAAAATCGTGATGGAAAAAGCGTGAGGAG----------AATTCTAA---ACTC-GAG ATG
A GATATAAGATAAGTGAATGACAAAGCGTGAGTAT----------AATTCTCA---ACCC-GAG ATG
B GATAGAAGATCAGTGAATGACAAAGCGTGAGGAG----------AATTATCA---ACCC-GAG ATG
ccmB M GCAAAGCTAGCGTAGCGCCAGCCGTCGAAGTGAATGAATTCGAAAGAACGAAGAAATAAGGAA ATG
R GCAAAGCTAGCGTAGCGCCAGCCGTCGAAGTGAATGAATTCGAAAGAACGAAGAAGTAAGGAA ATG
W GCAAAGCTAGCGTAGCGCCAGCCGTCGAAGTGAATGAATTCGAAAGAACGCAGAAGTAAGGAA ATG
S TTACGCAGCGTTCAGAACCAGCC-TTGAAGTGAATGAATTAGAAGGAAGGAAGAAGTAAGAAA ATG
T TTACGCGGCGTTCAGAACCAGTC-TTGAAGTGAATGAATT-------AGAAAGAAGTAAGGAA ATG
A TTACGCGGCGTTCAGAACCAGCC-TTGAAGTGAATGAATT----AGAAAGAAGAAGTAAGGAA ATG
B TTACGCGGCGTTCAGAACCAGCC-TTGAAGTGAATGAATT----AGAAAGAAGAAGTAAGGAA ATG
ccmFN M GGATCATCCTGTGGTTACCGGATGATGGGAATAACTAAGCAGAAATTAGGAAATGAGCACGAA ATG
R GGATCATCCTGTGGTTACCGGATGATGGGAATAACTAAGCAGAAATTTTGAAATGAGCACGAA ATG
W GGATCATCCTGTGGTTACCGGATGATGGGAATAACTAAGCATAAATTTGGAAATGAGCACGAA ATG
S GGATCATCCTGTGGTTACCGGATGATGGGAATAACAAAGCAGAAATTTTAAAATGAGCACGAA ATG
T GGATCATCCTGTGGTTACCGGATGATGGGAATAACAAAGCAGAAATTTTGAAATGAGCACGAA ATG
A GGATCATCCTGTGGTTACCGGATGATGGGAATAACGAAGCAGAAATCTTGAAATGAGCACGAA ATG
B GGATCATCCTGTGGTTACCGGATGATGGGAATAACAAAGCAGAAATCTTGAAATGAGCACGAA ATG
ccmFC M GAACACTTTCATTTTTAGCGTCTCATCTTCTCTGGAGAAGCTCAAATCGAACGGATAGAGCAG ATG
R GAACACTTTCATTTTTAGCGTCTCATCTTCTCTGGAGAAGCTCAAATCGAACGGATAGAGCAG ATG
W GAACACTTTCATTTTTAGCGTCTCATCTTCTCTGGAGAAGCTCAAATCGAACGGATAGAGCAG ATG
S GAACACTTTCTTTTTTAGTCTTTCATCTTCTCTAGAGAAGCGAAACTCGAACGGATAGAGCAG ATG
T GAACACTTTCATTTTTAGCCTCTCATCTTCTCTAGAGAAGCAAAACTCGAACGGATAGAACAG ATG
A GAACATTTTCATTTTTAGCGCCTTCTCTTCTTTAGAGAAGCCAAACTCGAACGGATAGAGCAG ATG
B GAACATTTTCATTTTTAGCGCCTTCTCTTCTTTAGAGAAGCCAAACTCGAACGGATAGAGCAG ATG
rps3 M AAATAGAGGAAAGGGCAGAAA---GGGGCAA----AAGTAAAGTCTAAGCG--------ACAT ATG
R AAATAGAGGAAAGGACAGAAA---GGGGAAA----AAGTAAAGTCTAAGCG--------ACAT ATG
W AAATAGAGGAAAGGGCGGAAA---GGGGAAA----AAGTAAAGTCTAAGCG--------ACAT ATG
S AAAGATGGAATCGGGATAAAG---GGGGGAACA--AAGTCAAGTCTAAGCAAGCATATCATAT ATG
T AAATATTGGACCGGGAAGAAAAAAGGGGAAA----AAGTAAAGTCTAAGCG---------CAT ATG
A AAATATTGTACCGGGAAGAAAAAAGGGACAGAAATCCATAAAGTCAAA-----------AAAT ATG
B AAATATTGGACCGGGAAGAAAAAAGGGACAGAAATCCATAAAGTCAAA-----------AAAT ATG
rps12 M ATCGGGAGTAACCACTAGTGAAAGGGCTAA------GGGGGGAAGGACATAGGAAAGAGGG-- ATG
R ATCGGGAGTAACCACTAGTGAAAGGGCAAA------GGGGGGAAGGACATAGGAAAGAGGG-- ATG
W ATCGGGAGTAACCACTTTAGAAAGGGCAAA------GGGGGGAAGGACATAGGAAAGAGGG-- ATG
S ATCGGGAGTAATCACTAGTGATAGGGCAAAAATA--GGGGGGAAGGACAAAGGAAAGAGCG-- ATG
T ATCGGGAGTAACCACTAGTGAGAGGGCAAAAATT--GGGGGGAAGGACAAAGGAAAGAGCG-- ATG
A ATCGGGAGTAA-----AGTGATAGGGCACAAAAT--GGGGGGAAAAAGAAAGGAAAGAGAATA ATG
B ATCGGGAGTAA-----AGTGATAGGGCAAAAAATGGGGGGGGGAAAAGAAAGGAAAGAGAATA ATG

Supplementary Material S2 – Clustal nucleotide alignment of upstream cassette types I-III. Alignments were manually corrected with the caveat that homology was difficult to assess in some regions. Cassettes are shown in upper case, while flanking non-homologous sequences are shown in lower case. Positions within cassette type I that are identical among at least 9 of 12 sequences are indicated by asterisks. Dashes depict gaps, while dots denote nucleotide sequences not shown. Start codons are in bold. M=Maize, R=rice, W=wheat, S=sugar beet, T=tobacco and A=*Arabidopsis.*

TYPE I

```
M atp4   ...gtcaaGTTGCTCCTCAGAAAATTAGTAGAATAGTTTCATTGGCCTTCGTCAATGGGACAAACGCTCCAGTGTATGCGTTACA--GTTACTGGGAAGCTAGCATTTTTGCTTGTCATGGAATCAAGTCTAT
R cox2   ...atcatGTTGCTCCTCGGAAAACGCGTATAGTAGTCTCATTGGCCTTCGTCGATGGGACAAACGCTCCAGTGTATGCGTTACAAGGCAACTAGCATTTTGGATTAGAAGTTCGTGAAAGTATGTTCTCA--
W cox2   ...ccgttGTTGCTCTTCAGAAAACGCGTATAGTA-----AGTAGTCTTCGTCGATGGGAC-----CTCCAGTGTATGCGTTACGAGGCAACTAGCATTTTGTCATTAAAGTTCGTGAAAGAATG--------
W atp6-2 ...agaatGTTGCTCTTCAGAAAACGCGTATAGTG---------GCCTTCGTCGATGGGACAAACGCTCCAGTGTATGCGTTACAAGGCAACTAGCATTTTGTCATGGAAGTTCGTGAAAGAATG--------
W atp4              ...ctactcGAAAACGCGTATAGTA-----AGTAGTCTTCGTCGATGGGAC-----CTCCAGTGTATGCGTTACGAGGCAACTAGCATTTTGTCATTAAAGTTCGTGAAAGAATG--------
```

```
M atp4   TTGTTCGAATGTTCTTTTTTTGGAAAAACCAACCAC----------AAAAAAAGTCTCCCTTTCTCTT---------GGAGCAGAGCTTCATCATAAAAGTGGagagtcacaatg
M nad6         ...tatctTTTGTTTGGAAAAACCAAACCC-ACATATTGATCTTTAAGTCTCCCTTTCTCTTTT------GGGAGCAGATCT----TTCAAAGATGGgaaattccaatg
M rps7         ...tttcgTTTCGTTGGAAAAACCTACGCCCACATATTGATCTTTAAGTCTCCTTT-CTCTTTT------GGGAGCAGAGCT----GAAAAAGATGGacagtaacg...50...gctcggaaattctcagctatatg
R cox2   ------------TTTTTCGTTGGAAAAACCAACGCC---CGACGTCAAGATCAGTCTCCTTT-CT----------AGGAGCAGAGCT----AAAAAAGATGGGAAATTCCAATG
R atp6         ...tagagTTTCCTTGGAAAAACCGAGGCC------AGTTGAGATCAGTCTCCCTTTCT----------AGGAGCAGAGCT----TAAAAAGATGGGAAATTCCAATG
R atp4         ...attgcTTTCCTCGGAAAAACCAAGGCC------AGTTGAGATC-GTCTCCCTTTCTCTTTTGTTTT-GGGAGCAGAGCT----TAAAAGGATatagttaccccacgatg
R rps7         ...cttcgTTTCGTTGGAAAAACCGACGCC----AACGTTAAGATCAGTCTCCTTT-CTCTTTTC-----GGGAGCAGAGCT----TAAAAAGATGGacagtaacg...44...gctcggaaattctcagctatatg
W cox2   --------------TTTCGTTGGAAAAACCAACGCC----GACGTCAAGATCAGTCTCCTTT-CCTTTCTCTTTTCGGGAGCAGAGCT----GAAAAAGATGGgaaattccaatg
W atp6-1       ...attcgTTTCGTTGGAAAAACCAACGCC----GATCTAAAACTAAGTCTTCCTTTCAAAAGTGA----GCGAGCAGAGCT----GAAAAAGATGGAGTTACCTGGAGATG
W atp6-2 --------------TTTCGTTGGAAAAACCAACGCC----GACGTCAAGATCAGTCT-CCTTTCAAAAGTGA----GCGAGCAGAGCT----GAAAAAGATGGAGTTACCTGGAGATG
W atp4   --------------TTTCGTTGGAAAAACCAACGCC----GATCTAAAACTAAGTCTTCCTTTCAAAAGTGA----GCGAGCAGAGCT----GAAAAAGATGGAGTTACCTGGAGATG
W rps7         ...caggcTTTCCTTGGAAAAACCAAGGAC----AACCCCTATCTCAGTCTCCTTTCCTTTCTCTTTTC-GGGAGCAGAGCT----GAAAAAGATGGacagtaacg...50...gctcggaaattctcagctatatg
                     **** ************* ***               ******* **             ************     **********
```

TYPE II

```
S cox2   ...aaatcATAAGAGAAGAAAGCAATGCCCAAAGACTCCCATGCCTTTCTTGGTTGGA-AAACCCAACCGGTGATTTCTGACAAGTCTTTCTTCATTTCTCGAGAGAGCGGAGCAGTCAAAGAATGAAccaaatg
S rpl5    ...cgtatATAA-ATAGAAAAGCAATGCCCAAGGACTCCCATGCCTTTCTTGGTTGGAAAAAACCAACTATAAATATCCTACAAGTCTTTTTTCATTTCTCGAGAGAGC-----AGTCAAAGAATGAAtaagaaaagccaactcatg
S atp6    ...attgaATAAGAGAAGAAAGCAATGCCCAAAGACTCCCATGCCTTTCTTGGTCGGACCAACCCAACCGGTGATTTCCGACAAGTCTTTCTTCATTTTTGAGCAAGAAGCGGAACTACAgggatgaaatgaaaagtgtttattacg
S nad9                              ...gtgcaTCCCATGCCTTTCCTGGTTGGA-AAACCCAACCGGTGATTTCCGACAAGTCTTTCTTCATTTTTGAGCAAGAAGCGGAACTACAagaaagctttctttctctttatg
```

TYPE III

```
T cox2                                                                ...tcttcCTTTTTTTCTCCCATGCTTTCCGTTGGTCAACAACCAACTAAAGTGCTCTATACTTCTTCACTACTCGTAC
T atp6       ...ctcttTTTATATATGAAATTACTTCGTCCTTT--TTTTTAGCCCTTTTTCGTTTGTCCATCTTTTTTTCTCCCATGCTTTCCGTTGGTCAACAACCAACTAAAGTGCTCTATACTTCTTCACTACTCGTAC
T atp8                                    ...ggttcTTATTAGCCCTTTTTCGTTTGTGCATCTTT----CTCCCATGCTTTCCGTTGGTCAACAACCAACCAAAGTGCTCTATACTTCTTCACTACTCGTAC
T rps13      ...atatcTTTATATACGAAATTACTTCGTCCTTTTTTTTTTAGCCCTTTTTCGTTTGTCCATCTTTTTTTCTCCCATGCTTTCCGTTGGTCAACAACCAACCAAAGTGCTCTATACTTCTTCACTACTCGTAC
A atp8     ...cccgcGATCAAACTATCAATCTCATAAGAGAAGAAATCTCTATGCCCCCTTTTTCTTGG--------TTTTCTCCCATGCTTT-TGTTGGTCAACAACCAACCACAAACTTTCTATAGTTCTTCACTACTCCTAG
A atp6-1   ...tacggGATCAAACTATCAATCTCATAAGAGAAGAAATCTCTATGCCCCCTTTTTCTTGG--------TTTTCTCCCATGCTTT-TGTTGGTCAACAACCAACCACAACTTTCTATAGTTCTTCACTACTCCTAG
```

```
T cox2   AGGGAAGGTGGAagaaatgaagtctctctttttttttgggggggagcagagcagtcaaagaatgaaccaaaccaaatg
T atp6   AGGCTTGACGGAGTTAAGCTGTATTGAGGGAATCTTTTTATCTCAATCAcaatg
T atp8   AGGCTTGACGGAGTTAAGCTGTATTGAGGGAATCGTTTTGTCTCAATCAATCAAGAATG
T rps13  AGGCTTGACGGAGTTAAGCTGTATTGAGGGAATCGTTTTGTCTCAATCAATCAAGAATG
A atp6-1 AGACTTGACGGAGTGAAGCTGTCTGGAGGGAATCATTTTGTTGAAATCAATTAATCCAATATG
A atp8   AGGCTTGACGGAGTGAAGCTGTCTGGAGGGAATCATTTTGTTGAAATCAATTAATCTAATCATG
```

CHAPTER 6: GENERAL DISCUSSION

6.1 Transfer of *rps1* to the nucleus in legumes: possible model

The gene encoding the S1 ribosomal protein (*rps1*) is nuclear-located in alfalfa and its close relatives barrel-medic, sweet clover and fenugreek, yet is mitochondrial-encoded in more distantly related species within the Phaseoleae subfamily (rest harrow, pea, soybean and bean). From this, I inferred a recent transfer of *rps1* to the nucleus, and as such the nuclear-located *rps1* sequences in alfalfa and its close relatives appear to be at a stage following transfer that allows for a rare glimpse of some its initial adaptive changes. Comparative analysis of the *rps1* upstream sequence in the nucleus with its ancestral (mitochondrial) counterparts that can still be seen upstream of *rps1* in rest harrow allows for the proposal of a model of the steps involved in this transfer.

6.1.1 Escape of *rps1* from the mitochondria and migration to the nucleus

The functional *rps1* gene underwent several DNA rearrangements in its upstream genomic environment during recent legume evolution. Following the split of the pea lineage from the rest of these legumes, a 20 nt stretch of sequence was inserted into an otherwise well-conserved region upstream of mitochondrial-located *rps1* (and this can still be seen in rest harrow), and this stretch contained a motif that would be used as a nuclear intron 3' splice site. A copy of *rps1* escaped from the mitochondria and migrated to the nucleus, likely as cDNA as all the editing sites seen in other legumes (except the silent-edited site) appear as the edited form in the present-day nuclear copy. However, it cannot be excluded that this event was DNA-mediated, as the editing sites in present-day legumes may have contained genomically-encoded uracil residues at the time of transfer.

6.1.2 Integration into the nuclear genome: the 'transition stage'

One scenario is that the *rps1* sequence along with its flanking regions became integrated at its currently observed nuclear genomic location, which includes a promoter and the 5' splice site of an intron. This region of integration may have contained remnants of a duplicated (or partially-duplicated) gene. The translocated *rps1* sequence might even have integrated into the intron of a functional (but redundant) gene, in which case the motif in the residual sequence upstream of *rps1* could have taken over the role of the 3' splice site, and the remainder of the gene downstream would appear to have been lost. Alternatively, the promoter and 5' intron splice site (also signals for transcription termination) could have been

acquired following integration through rearrangement events in the nuclear genome, although once in the nucleus, a translocated sequence likely has a very limited time to become active (and under functional constraint) before it accumulates deleterious mutations.

Translation initiation of *rps1* might have been possible albeit inefficient due to the presence of an AUG triplet ~45 nt upstream that would have likely interfered with ribosome scanning (see Figure 3.5). As well, it appears that the mitochondrial-type initiator codon would have been in an unfavourable context for recognition in the cytosol. Despite this, "leaky scanning" (cf. Marintchev and Wagner 2005) in the 5' UTR could have allowed for low-level translation of the *rps1* mRNA, and translation initiation at both the mitochondrial-type initiation codon and the AUG triplet adjacent to it (which appears to be in a favourable context, i.e. adenosine at position -3) would likely have synthesized a functional S1 ribosomal protein. However, the efficiency of synthesis of the nuclear-encoded protein would possibly have been comprimised due to mitochondrial-suited codon usage (Chapter 3). The sequence of the mitochondrial-encoded S1 ribosomal protein appears to have inherent targeting peptide-like features and this may have permitted mitochondrial-targeting at the time of activation, and protease cleavage of an N-terminal extension would not have been required. Under these circumstances, even if it were not able to fully adopt the role provided by the mitochondrial copy, expression of *rps1* in the nucleus might have allowed for a reduction in the level of expression of the mitochondrial copy. Thus, both copies would be under functional constraint until one gained (or regained) sufficient expression to become the lone copy. Expression of the mitochondrial copy could have been reduced (although not lost altogether), either through changes in the make-up of regulatory elements embedded within mitochondrial sequences, or though changes in (for example) specificity of nuclear-encoded machinery involved in RNA-level regulation. Such an alteration of the composition of elements regulating expression of mitochondrial genes can occur in short periods of evolutionary time (cf. Chapters 4, 5). Also, certain evolutionary changes affecting the mitochondrial *rps1* sequence might have been tolerated due to a somewhat relaxed constraint. Such changes might include loss of editing (which only occurs at a few sites in *rps1*), indel mutations (not causing a frame-shift), and missense mutations. The nuclear-located copy would then have been under functional constraint and selective pressure could maintain its presence as it acquired adaptive mutations. It is interesting to speculate that if

such a 'transition stage' (i.e. there is a functional copy of a gene in both cellular compartments) occurs during a period of rapid divergence of species within a flowering plant lineage, then depending on how frequently the nuclear copy 'wins out' over the mitochondrial copy, then even among very close relatives there could be differences in the status of a given mitochondrial gene similar to what was observed for *rps19* in cereals (Fallahi et al. 2005).

6.1.3 Adaptation of *rps1* to its new location

At the time of transfer the sequence of the translocated *rps1* sequence would be identical to that of its mitochondrial-located counterpart (and thus apparently suitable to encode a functional S1 ribosomal protein), however some adaptation would almost certainly have been required to optimize expression in the nucleus. Such adaptive changes that occurred in nuclear-located *rps1* appear to have included the loss of the ATG located in the 5' UTR, as well as the ATG used as the initiation codon in the mitochondria, leaving the adjacent ATG (apparently becoming the sole candidate for translation initiation) in a favourable context (i.e. adenosine at position -3). Substitutions in the coding sequence could have improved the mitochondrial-targeting features of the S1 ribosomal protein, and as this would have required changes in the primary amino acid sequence, it may be that co-evolutionary changes would have had to occur in genes that encode, for example, other ribosomal proteins in close association within the ribosome. Other substitutions appear to have shifted the codon usage to suit translation in the cytosol. Changes affecting acquired regulatory features (e.g. promoter strength) might also have been necessarily. Following adaptation, the nuclear-located *rps1* alone might have been sufficient to provide the S1 ribosomal protein and the mitochondrial copy would have been redundant.

6.1.4 Inactivation of the mitochondrial copy

Once it became redundant, the mitochondrial copy was inactivated apparently by acquiring an insertion containing an in-frame stop codon (although additional inactivating-type mutations were seen among the pseudogenes in alfalfa, sweet clover and fenugreek). Alternatively, mitochondrial-located *rps1* might have been inactivated through the loss of abundant *rps1* transcripts (Chapter 3) as was seen for mitochondrial *cox2* in soybean (Adams et al. 1999). Perhaps this loss was a result of the introduction of a processing site in the *rps1* sequence (and possibly resulting instability of the remaining *rps1*-containing transcript),

which may be related to the genomic environment upstream of *rps1* (which is common to the legume species having a mitochondrial *rps1* pseudogene) in a manner similar to what was seen with *cox3* among ecotypes of *Arabidopsis*, in which long-range RNA interactions created a lineage-specific processing site (Forner et al. 2005; see Chapter 4, discussion). Northern blot analysis and end-mapping experiments are being done to determine if the loss of abundant *ψrps1* transcripts and the proposed processing site are common to sweet clover and fenugreek, as well. If so, transcript-loss would be a candidate for the method of common inactivation. While the accumulation of missense mutations can render a gene inactive, this was not apparent in the *ψrps1* sequences. This might not be surprising considering the low rate of nucleotide substitution in plant mitochondria, and the high rate of DNA rearrangements might explain why flowering plant pseudogenes appear to be lost from mitochondrial genomes prior to their deterioration from point mutations.

Under this model of the transfer of *rps1* to the nucleus, there would be a period of time in which a potentially aberrant S1 ribosomal protein would be encoded, either through translation initiation at an upstream-located AUG triplet in the nucleus (which would have extended the protein by 15 amino acids), or through translation of the *rps1* pseudogene in the mitochondria, which would have resulted in a protein that was truncated by 58 amino acids. These abnormal S1 proteins would have had the potential to integrate into the ribosome and thus exclude their functional counterparts. If so, this may not have been detrimental because 1) as discussed above, the loss of abundant steady-state mitochondrial *rps1* transcripts may have predated the nonsense mutation, 2) the levels of the aberrant proteins might have been low, either because of the above, or because the upstream AUG in the nuclear-located *rps1* gene was in an unfavourable context for initiation codon recognition, or 3) plant mitochondria may have a surveillance mechanism to detect and remove such proteins. Alternatively, the possible aberrant S1 ribosomal proteins might have conferred an impaired mitochondrial function that would only have been apparent during periods of high mitochondrial demand. It would be interesting to know if cases such as this could lead to, for example, cytoplasmic male sterility, in a similar manner to that which is caused by (aberrant) chimeric mitochondrial proteins resulting from mitochondrial genome rearrangements at a locus containing a duplicated piece of a respiratory chain gene (reviewed in Hanson and Bentolila 2004).

6.1.5 Nuclear-located *rps1* in sunflower and cotton

To gain insight into other *rps1* transfer events, I examined EST sequences for sunflower and cotton (see Appendix 1). The EST sequences along with the absence of *rps1* sequences in their mitochondria (based on Southern data; Adams et al. 2002b) indicate that S1 is nuclear-encoded in these plants from independent transfer events. Comparison of these sequences with alfalfa S1 and mitochondrial-encoded S1 sequences revealed similarities among these three nuclear-encoded copies. As in alfalfa, the sunflower S1 ribosomal protein contains no N-terminal extension (based on an upstream in-frame stop codon), and the presence of a sequence encoding a targeting peptide is unknown for cotton due to insufficient sequence data. In addition, 11 sites in the S1 sequence appear to contain nuclear-sequence specific amino acid residues in two of the three species (Appendix 1). The same initiation codon used in alfalfa appears to be used in sunflower (and also cotton if there is no acquired presequence), and the mitochondrial-type initiator codon has been lost in both cases. In cotton, like alfalfa, this site is AAG, and the corresponding site in sunflower (ACA) also provides an adenosine at position -3. The similarity of these nuclear-located *rps1* features seems to contrast with the diverse ways in which mitochondrial *rps10* (also thought to frequently undergo transfer) has become functional in the nucleus in various flowering plants. For example, cases have been documented where translocated *rps10* has acquired a presequence (encoding a mitochondrial targeting peptide) and others where one has not. In the case of lettuce, the *rps10* sequence has integrated into a non-mitochondrial gene, and the acquired N-terminal region appears not to be cleaved after mitochondrial import (summarized in Table 1.2).

6.2 Expression of the *rps1* gene in flowering plants and translation initiation

The examination of sequences preceding *rps1* in the mitochondria of legumes revealed variation among even closely-related species, and this variation includes indels, substitutions and even editing status (Chapter 3). In addition, sequences further upstream (yet still within the *rps1* 5' UTR region) were found to be non-homologous, and therefore lineage-specific. This lack of conservation, along with the deviation of the plant mitochondrial S1 ribosomal protein from the bacterial-type raised questions about how translation initiation occurs in flowering plant mitochondria, and in a more general sense,

about the nature of sequences in regions expected to contain RNA-level regulatory signals. Thus, an analysis was included in this study that broadened the data set by exploiting the complete mitochondrial genome information available in the databank (Chapter 5).

In bacteria, the S1 ribosomal protein consists of a ribosome binding domain and an RNA binding domain (Subramanian 1983), and the latter feature is not found in mitochondrial lineages. As this may be related to the evolutionary loss of Shine-Dalgarno type ribosome binding in translation initiation, it is curious that many mRNAs in the mitochondrion of the protist *Reclinomonas americana* have potential Shine-Dalgarno sequences. It would be interesting to learn if this reflects an 'intermediate stage' in which the function of the RNA binding domain in assisting Shine-Dalgarno ribosome binding has been replaced in the protist lineage, and perhaps a derivative of this mechanism has overtaken the role formerly provided by this type of ribosome binding in the mitochondria of plants. While it is possible that plant mitochondria use mRNA-specific machinery in initiator codon recognition, somewhat analogous to the translational activators seen in yeast mitochondria (Costanzo and Fox 1990), RNA specificity would have to accommodate the frequent lack of conservation immediately preceding initiator codons (including non-homology due to replacement; Chapter 5). Interestingly, one of the yeast translational activators is a PPR protein, and in plant mitochondria, the role of such proteins is thought to confer sequence specificity for other machinery involved in RNA-level regulatory events (many of which, at least for the *rps1* locus in legumes, have a lineage-specific nature to them) (Andrés et al. 2007). Perhaps machinery for initiation codon recognition is part of a complex including RNA maturation machinery (cf. Choury et al. 2005), and therefore translation initiation is tightly coupled with other RNA-level events such as editing and intron removal. Another contributing factor in start codon recognition may be the accessibility of such machinery (including the ribosome) to the region of the initiator, as a conspicuous feature of sequences immediately preceding plant mitochondrial start codons is a nucleotide bias (for adenosine and against cytosine; Chapter 5). It is unclear as to how this bias in the 5' UTR might influence accessibility, however, it may be related to the formation of secondary structures, which has been shown to affect translation efficiency of bacterial and eukaryotic mRNAs (reviewed in Marintchev and Wagner 2005).

The different levels of conservation among regions upstream of plant mitochondrial genes (especially lack thereof), along with the apparent lineage-specific features of plant mitochondrial transcripts of conserved regions of a locus (i.e. *rps1-nad5ab*) point to an expression system that uses something other than conserved recognition elements (see Chapters 4, 5). Continuous recruitment of regulatory signals seems inefficacious, but perhaps is made feasible by 1) the availability of copies of (potential) regulatory sequences located in spacer regions; 2) through co-transcription, the ability to exploit regulatory sequences already in use by another gene (as seen for *rps1* in some legume species); 3) the apparent flexibility in composition of regulatory elements to collectively provide adequate expression, as well as seemingly flexible requirements for an individual element (perhaps aided by dynamic nuclear-encoded machinery); and 4) the latitude in the relative location of at least certain signals, as evinced by UTRs of variable lengths (cf. Forner et al. 2007). The last point may be greatly facilitated by the apparent absence of plant mitochondrial transcription terminators (Hoffman et al. 2001).

6.3 The S1 ribosomal protein gene in flowering plants

Among flowering plant mitochondrial genes, *rps1* is inferred to be among the more dynamic with respect to gene transfer as it is thought to have been lost from the mitochondria independently 33 times during angiosperm evolution (Adams et al. 2002b). The sporadic nature of these events may reflect that *rps1* is exceptionally amenable to transfer. In addition to its apparent targeting peptide-like features, S1 ribosomal protein is short and hydrophilic relative to respiratory chain proteins, and these features are thought to facilitate transfer. Perhaps the relatively high number of times *rps1* has relocated to the nucleus is related to the recombinationally active region preceding it. As rest harrow-specific sequences upstream of *rps1* contributed seemingly vital elements for nuclear expression, then the success of this particular transfer event appears to have been dependent upon the exploitation of derived mitochondrial sequence features specific to that lineage. While it is unclear as to how frequently residual mitochondrial sequences contribute to nuclear gene activation, the fluid nature of plant mitochondrial genomes, even near genes raises questions about the role of mitochondrial DNA rearrangements in providing (translocatable) sequences containing various gene and flanking sequence combinations, some of which may fortuitously

contribute signals for nuclear expression along with the coding sequence. Therefore, the placement of a suitable genomic context (or conversely, its subsequent loss) in the mitochondria from lineage to lineage of flowering plants may in part explain the sporadic nature of mitochondrial gene loss (relocation) (cf. Adams et al. 2002b). It would be interesting to learn if regions upstream of other mitochondrial genes that have been implicated in multiple independent transfers have also undergone frequent rearrangements. For example, *rps12*, which has the lowest number of inferred losses (6) from flowering plant mitochondria (Adams et al. 2002b), is closely-linked to upstream *nad3* in all seven completely-sequenced flowering plant mitochondrial genomes. However, it cannot be ruled out that many of the *rps12* sequences detected via Southern hybridization by Adams et al. 2002b represent pseudogenes.

A mitochondrial-type *rps1* homologue could not be detected in either of the completely-sequenced nuclear or mitochondrial genomes of *Arabidopsis*. The apparent loss or functional replacement of mitochondrial S1 in this lineage (perhaps reminiscent of its occasional loss in bacterial lineages as reviewed in Lecompte et al. 2002) may be indicative of flexibility, perhaps in the structural features the S1 ribosomal protein. In *E. coli*, this protein interacts directly with two other ribosomal proteins (S11 and S18) and is part of a cluster also containing S6 and S21 (Sengupta et al. 2001). Despite this, only the N-terminal region of the S1 ribosomal protein is involved in interactions with other proteins in the ribosome and this is reflected in their weak association, as S1 dissociates easily during physio-chemical experimentation (Subramanian 1983). This may explain why variability in the C-terminal region among mitochondrial lineages (as seen in *Marchantia* and *Physcomitrella*, see Figure 1.5 B) appears to be tolerated, and indicates that the C-terminal binding domain in the chloroplast-type S1 ribosomal protein (Hirose and Sugiura 2004, Merendino et al. 2003) may not preclude its ability to substitute for the mitochondrial-type, as may have happened in the *Arabidopsis* lineage. It will be of interest to know if losses of *rps1* in the mitochondria of other flowering plant lineages reflect replacement of the mitochondrial-type S1 ribosomal protein. This idea of flexibility of this protein might also be underscored by the absence of editing of legume *rps1* transcripts at sites predicted to be so, particularly by the editing at corresponding sites in wheat (Gonzalez et al. 1993). Additional

mitochondrial *rps1* data for flowering plant species is needed to learn more about how this gene is evolving in the mitochondria.

The S1 ribosomal protein, like many other ribosomal proteins in *E. coli* has extraribosomal functions (cf. Wool 1996). One in particular that has undergone recent examination is that in *trans*-translation. This is a quality control system by which bacterial mRNAs lacking a stop codon translate an additional degradation tag with the aid of a transfer/messenger RNA (tmRNA) containing a tRNA-like domain along with a surrogate reading frame (reviewed in Dulebohn et al. 2007). The RNA binding domain of the S1 ribosomal protein in *E. coli* has been shown to be important in this function. *Trans*-translation has not been demonstrated in the mitochondria of plants, and if this function was lost, it may have relaxed to a degree the constraint on the RNA binding domain, and thus facilitated its loss or functional replacement. Interestingly, the tRNA-like domain in the tmRNAs of *E. coli* are reminiscent of the t-element-like structural feature found at the 3' end of the *nad6* and *ccmC* mRNA, which is internal to the coding sequence in the mitochondria of *Arabidopsis* (Forner et al. 2007). It would be interesting to learn if these are related to a *trans*-translation type mechanism.

6.4 Future directions

By investigating a mitochondrial gene that had recently been transferred to the nucleus, I was able to examine some changes that I inferred to be adaptive. A similar analysis on mitochondrial genes in a transition state (i.e. a functional copy in both compartments) and thus ideally at a stage prior to the accumulation of such adaptive changes might allow for the assessment of the functionality of a gene in the very early stages of transfer. It would also be interesting though mutational analyses to attempt to recreate the nuclear *rps1* sequence (and its upstream region) as it would have been at the time of transfer, and see if a (potentially) functional S1 protein was localized in the mitochondria.

By examining mitochondrial transcript features among closely-related legume species I was able to infer a very plastic composition of RNA-level regulatory elements involved in the expression of *rps1* and its neighbouring genes. It would be worthwhile to investigate in more detail the nature of these signals (for example, promoter vs. processing site generating the *nad5* 5' end in alfalfa), and also to examine if such differences in RNA-level regulation

are seen at different developmental stages as was seen for mitochondrial genes in wheat by Li-Pook-Than et al. (2004). As virtually nothing is known about the nature of the nuclear-encoded machinery involved in plant mitochondrial RNA-level expression, additional work in this area is necessary, particularly examining PPR proteins, as these show promise in providing a pivotal role in governing much of mitochondrial RNA-level events.

6.5 Concluding remarks

The transfer of plant mitochondrial genes to the nucleus is contingent upon the success of several steps of this process, and the seemingly remote chances of success contrasts with the inferred high frequency of these events in flowering plants. The examination of the recently transferred *rps1* gene in legumes has allowed for the proposal of a model that explains how this gene became activated in the nucleus, and how it adapted to replace its mitochondrial-located counterpart. In addition, the recruitment (and re-recruitment) of regulatory signals both mitochondrial- and (presumably) nuclear-encoded results in different mitochondrial transcript features, even among very close relatives, and illustrates the lineage-specific nature of plant mitochondrial RNA-level regulation. Comparative studies such as these not only provide insight into how plant mitochondrial genomes have evolved (and continue to evolve), but can also aid in the development of fields such as genetic modification of plants (by examining how foreign genes can become suited for plant nuclear expression), and cytoplasmic male sterility, which is an important feature of some crop plants.

REFERENCES

Adams KL, Palmer JD 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29:380-95

Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD 2000. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 16:354-7

Adams KL, Daley DO, Whelan J, Palmer JD 2002a. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. Plant Cell 14:931-43

Adams KL, Qiu YL, Stoutemyer M, Palmer JD 2002b. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A 99:9905-12

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD 1999. Intracellular gene transfer in action: dual transcription and multiple silencings of nuclear and mitochondrial cox2 genes in legumes. Proc Natl Acad Sci U S A 96:13863-8

Andrés C, Lurin C, Small ID. 2007. The multifarious roles of PPR proteins in plant mitochondrial gene expression. Physiol. Plantarum 129:14-22

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796-815

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K 1990. Current protocols in molecular biology. Current Protocols, Boston

Behm-Ansmant I, Kashima I, Rehwinkel J, Saulière J, Wittkopp N, Izaurralde E 2007. mRNA quality control: an ancient machinery recognizes and degrades mRNAs with nonsense codons. FEBS Lett 581:2845-53

Bentolila S, Elliott LE, Hanson MR 2007. Genetic architecture of mitochondrial editing in Arabidopsis thaliana. Genetics. [Epub ahead of print]

Bentolila S, Chateigner-Boutin AL, Hanson MR 2005. Ecotype allelic variation in C-to-U editing extent of a mitochondrial transcript identifies RNA-editing quantitative trait loci in Arabidopsis. Plant Physiol 139:2006-16

Binder S, Brennicke A 2003. Gene expression in plant mitochondria: transcriptional and post-transcriptional control. Philos Trans R Soc Lond B Biol Sci 358:181-8

Blanchard JL, Lynch M 2000. Organellar genes: why do they end up in the nucleus? Trends Genet 16:315-20

Boer PH, McIntosh JE, Gray MW, Bonen L 1985. The wheat mitochondrial gene for apocytochrome b: absence of a prokaryotic ribosome binding site. Nucleic Acids Res 13:2281-2292

Bonen L 2006. Mitochondrial genes leave home. New Phytol 172:379-81.

Bonen L, Bird S 1988. Sequence analysis of the wheat mitochondrial atp6 gene reveals a fused upstream reading frame and markedly divergent N-termini among plant ATP6 proteins. Gene 73:47-56

Bonen L, Brown GG 1993. Genetic plasticity and its consequences: perspectives on gene organization and expression in plant mitochondria. Can J Bot 71:645-660

Bonen L, Calixte S 2006. Comparative analysis of bacterial-origin genes for plant mitochondrial ribosomal proteins. Mol Biol Evol 23:701-12

Brennicke A, Grohmann L, Hiesel R, Knoop V, Schuster W 1993. The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Lett 325:140-5

Brodie R, Roper RL, Upton C. 2004. JDotter: a Java interface to multiple dotplots generated by dotter. Bioinformatics 20:279-281

Bullerwell CE, Gray MW 2004. Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. Curr Opin Microbiol 7:528-534

Chanut FA, Grabau EA, Gesteland RF 1993. Complex organization of the soybean mitochondrial genome: recombination repeats and multiple transcripts at the atpA loci. Curr Genet 23:234-47

Chaw SM, Chang CC, Chen HL, Li WH 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424-41

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD 2003. Multiple sequence alignment with the Clustal series of programs. Nucl Acids Res 31:3497-3500

Choi C, Liu Z, Adams KL 2006. Evolutionary transfers of mitochondrial genes to the nucleus in the Populus lineage and coexpression of nuclear and mitochondrial Sdh4 genes. New Phytol 172:429-39

Choury D, Farré JC, Jordana X, Araya A 2005. Gene expression studies in isolated mitochondria: Solanum tuberosum rps10 is recognized by cognate potato but not by the transcription, splicing and editing machinery of wheat mitochondria. Nucleic Acids Res 33:7058-65

Claros MG, Vincens P 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 241:779-86

Clifton SW, Minx P, Fauron CMR et al. (13 co-authors) 2004. Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol 136:3486-3503

Costanzo MC, Fox TD 1990. Control of mitochondrial gene expression in Saccharomyces cerevisiae. Annu Rev Genet 24:91-113

Covello PS, Gray MW 1991. Sequence analysis of wheat mitochondrial transcripts capped in vitro: definitive identification of transcription initiation sites. Curr Genet 20:245-251

Dawson AJ, Jones VP, Leaver CJ 1984. The apocytochrome b gene in maize mitochondria does not contain introns and is preceded by a potential ribosome binding site. EMBO J 3:2107-2113

Dulebohn D, Choy J, Sundermeier T, Okan N, Karzai AW 2007. Trans-translation: the tmRNA-mediated surveillance mechanism for ribosome rescue, directed protein degradation, and nonstop mRNA decay. Biochemistry 46:4681-93

Emanuelsson O, Brunak S, von Heijne G, Nielsen H 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953-71

Esposito D, Fey JP, Eberhard S, Hicks AJ, Stern DB 2003. In vivo evidence for the prokaryotic model of extended codon-anticodon interaction in translation initiation. EMBO J 22:651-656

Fallahi M, Crosthwait J, Calixte S, Bonen L 2005. Fate of mitochondrially located S19 ribosomal protein genes after transfer of a functional copy to the nucleus in cereals. Mol Genet Genomics 273:76-83

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 35:3676-92

Forner J, Weber B, Wiethölter C, Meyer RC, Binder S 2005. Distant sequences determine 5' end formation of cox3 transcripts in Arabidopsis thaliana ecotype C24. Nucleic Acids Res 33:4673-82

Franzetti B, Zhou DX, Mache R 1992. Structure and expression of the nuclear gene coding for the plastid CS1 ribosomal protein from spinach. Nucleic Acids Res 20:4153-7

Gan X, Kitakawa M, Yoshino K, Oshiro N, Yonezawa K, Isono K 2002. Tag-mediated isolation of yeast mitochondrial ribosome and mass spectrometric identification of its new components. Eur J Biochem 269:5203-14

Gaut BS 2002. Evolutionary dynamics of grass genomes. New Phytol 154:15–28

Giegé P, Brennicke A 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci U S A 96:15324-9

Gonzalez DH, Bonnard G, Grienenberger JM 1993. A gene involved in the biogenesis of c-type cytochromes is co-transcribed with a ribosomal protein gene in wheat mitochondria. Curr Genet 24:248-55

Groth-Malonek M, Pruchner D, Grewe F, Knoop V 2005. Ancestors of trans-splicing mitochondrial introns support serial sister group relationships of hornworts and mosses with vascular plants. Mol Biol Evol 22:117-25

Gray MW 1992. The endosymbiont hypothesis revisited. Int Rev Cytol 141:233-357

Gray MW 1988. Organelle origins and ribosomal RNA. Biochem Cell Biol 66:325-48

Gray MW, Lang BF, Burger G 2004. Mitochondria of protists. Annu Rev Genet 38:477-524

Handa H 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis. Nucl Acids Res 31:5907-5916

Hanson MR, Bentolila S 2004. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell 16:S154-S169

Haouazine-Takvorian N, Takvorian A, Jubier MF, Lejeune B 1997. Genes encoding subunit 6 of NADH dehydrogenase and subunit 6 of ATP synthase are co-transcribed in maize mitochondria. Curr Genet 31:63-69

Hazle T, Bonen L 2007. Comparative analysis of sequences preceding protein-coding mitochondrial genes in flowering plants. Mol Biol Evol 24:1101-12

Hirose T, Sugiura M 2004. Functional Shine-Dalgarno-like sequences for translational initiation of chloroplast mRNAs. Plant Cell Physiol 45:114-7

Hoffman M, Kuhn J, Däschner K, Binder S 2001. The RNA world of plant mitochondria. Prog Nucleic Acid Res Mol Biol 70:119-154

Holec S, Lange H, Kuhn K, Alioua M, Borner T, Gagliardi D 2006. Relaxed transcription in Arabidopsis mitochondria is counterbalanced by RNA stability control mediated by polyadenylation and polynucleotide phosphorylase. Mol Cell Biol 26:2869-2876

Kadowaki K, Kubo N, Ozawa K, Hirai A 1996. Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. EMBO J 15:6652-61

Kellogg EA 2001. Evolutionary history of the grasses. Plant Physiol 125:1198-1205

Knoop V 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet 46:123-39

Knoop V, Altwasser M, Brennicke A 1997. A tripartite group II intron in mitochondria of an angiosperm plant. Mol Gen Genet 255:269-76

Koch MA, Haubold B, Mitchell-Olds T 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol 17:1483-1498

Komarova AV, Tchufistova LS, Dreyfus M, Boni IV 2005. AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in Escherichia coli. J Bacteriol 187:1344-9

Kozak M.1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187-208

Krishnasamy S, Grant RA, Makaroff CA 1994. Subunit 6 of the Fo-ATP synthase complex from cytoplasmic male-sterile radish: RNA editing and NH2-terminal protein sequencing. Plant Mol Biol 24:129-141

Kubo N, Ozawa K, Hino T, Kadowaki K 1996. A ribosomal protein L2 gene is transcribed, spliced and edited at one site in rice mitochondria. Plant Mol Biol 31:853-862

Kubo T, Mikami T 2007. Organization and variation of angiosperm mitochondrial genome. Physiol Plantarum 129:6-13

Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T 2000. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNA-Cys (GCA). Nucl Acids Res 28:2571-2576

Kühn K, Bohne AV, Liere K, Weihe A, Börner T 2007. Arabidopsis phage-type RNA polymerases: accurate in vitro transcription of organellar genes. Plant Cell 19:959-71

Lamattina L, Gonzalez D, Gualberto J, Grienenberger JM 1993. Higher plant mitochondria encode an homologue of the nuclear-encoded 30-kDa subunit of bovine mitochondrial complex I. Eur J Biochem. 217:831-838

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-497

Lavin M, Herendeen PS, Wojciechowski MF 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst Biol 54:575-94

Lecompte O, Ripp R, Thierry JC, Moras D, Poch O 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 30:5382-90

Li-Pook-Than J, Carrillo C, Bonen L 2004. Variation in mitochondrial transcript profiles of protein-coding genes during early germination and seedling development in wheat. Curr Genet 46:374-80

Lorković ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W 2000. Pre-mRNA splicing in higher plants. Trends Plant Sci 5:160-7

Lynch M, Koskella B, Schaack S 2006. Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727-30

Mackenzie SA 2005. Plant organellar protein targeting: a traffic plan still under construction. Trends Cell Biol 15:548-54

Maier RM, Zeltz P, Kossel H, Bonnard G, Gualberto JM, Grienenberger JM 1996. RNA editing in plant mitochondria and chloroplasts. Plant Mol Biol 32:343-365

Manthey GM, McEwen JE 1995. The product of the nuclear gene PET309 is required for translation of mature mRNA and stability or production of intron-containing RNAs derived from the mitochondrial COX1 locus of Saccharomyces cerevisiae. EMBO J 14:4031-4043

Marienfeld J, Unseld M, Brandt P, Brennicke A 1996. Genomic recombination of the mitochondrial atp6 gene in Arabidopsis thaliana at the protein processing site creates two difference presequences. DNA Res. 3:287-290

Marintchev A, Wagner G 2005. Translation initiation: structures, mechanisms and evolution. Q Rev Biophys 3:197-284

Merendino L, Falciatore A, Rochaix JD 2003. Expression and RNA binding properties of the chloroplast ribosomal protein S1 from Chlamydomonas reinhardtii. Plant Mol Biol 53:371-82

Mower JP, Palmer JD 2006. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Mol Gen Genomics 276:285-293

Mundel C, Schuster W 1996. Loss of RNA editing of rps1 sequences in Oenothera mitochondria. Curr Genet 30:455-460

Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA 2001. Excision of group II introns as circles. Mol Cell 8:201-11

Nakamoto T 2006. A unified view of the initiation of protein synthesis. Biochem Biophys Res Comm 341:675-678

Nakamura Y, Gojobori T, Ikemura T 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res 28:292

Nakazono M, Nishiwaki S, Tsutsumi N, Hirai A 1996. A chloroplast-derived sequence is utilized as a source of promoter sequences for the gene for subunit 9 of NADH dehydrogenase (nad9) in rice mitochondria. Mol Gen Genet 252:371-378

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K 2002. 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

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, et al. 1992. Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol 223:1-7

Ogihara Y, Yamazaki Y, Murai K et al. (14 authors) 2005. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucl Acids Res 33:6235-6250

Ong HC, Palmer JD 2006. Pervasive survival of expressed mitochondrial rps14 pseudogenes in grasses and their relatives for 80 million years following three functional transfers to the nucleus. BMC Evol Biol 6:55-71

Peeters N, Small I 2001. Dual targeting to mitochondria and chloroplasts. Biochim Biophys Acta 1541:54-63

Pereira de Souza A, Jubier MF, Lejeune B 1992. The higher plant nad5 mitochondrial gene: a conserved discontinuous transcription pattern. Curr Genet 22:75-82

Pereira de Souza A, Jubier MF, Delcher E, Lancelin D, Lejeune B 1991. A trans-splicing model for the expression of the tripartite nad5 gene in wheat and maize mitochondria. Plant Cell 3:1363-78

Pring DR, Mullen JA, Kempken F 1992. Conserved sequence blocks 5' to start codons of plant mitochondrial genes. Plant Mol Biol 19:313-317

Sambrook J, Fritsch EF, Maniatis T 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sandoval P, León G, Gómez I, Carmona R, Figueroa P, Holuigue L, Araya A, Jordana X 2004. Transfer of RPS14 and RPL5 from the mitochondrion to the nucleus in grasses. Gene 324:139-47

Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T 2004. The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol Genet Genomics 272:247-256

Schnare MN, Gray MW 1982. 3'-terminal sequence of wheat mitochondrial 18S ribosomal RNA: further evidence of a eubacterial evolutionary origin. Nucl Acids Res 10:3921-3932

Schuster W, Unseld M, Wissinger B, Brennicke A 1990. Ribosomal protein S14 transcripts are edited in Oenothera mitochondria. Nucl Acids Res 18:229-233

Sengupta J, Agrawal RK, Frank J 2001. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. Proc Natl Acad Sci U S A 98:11991-6

Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ 2004. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucl Acids Res 32:1774-1782

Shikanai T 2006. RNA editing in plant organelles: machinery, physiological function and evolution. Cell Mol Life Sci 63:698-708

Small I, Peeters N, Legeai F, Lurin C 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4:1581-90

Sorensen MA, Fricke J, Pedersen S 1998. Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. J Mol Biol 280:561-569

Subramanian AR 1983. Structure and functions of ribosomal protein S1. Prog Nucleic Acid Res Mol Biol 28:101-42

Subramanian S, Fallahi M, Bonen L 2001. Truncated and dispersed rpl2 and rps19 pseudogenes are co-transcribed with neighbouring downstream genes in wheat mitochondria. Curr Genet 39:264-272

Sugiura M, Hirose T, Sugita M 1998. Evolution and mechanism of translation in chloroplasts. Annu Rev Genet 32:437-459

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272:603-15

Tatusova TA, Madden TL 1999. Blast 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247-250

Terasawa K, Odahara M, Kabeya Y, Kikugawa T, Sekine Y, Fujiwara M, Sato N 2007. The mitochondrial genome of the moss Physcomitrella patens sheds new light on mitochondrial evolution in land plants. Mol Biol Evol 24:699-709

Timmis JN, Ayliffe MA, Huang CY, Martin W 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-35

Turmel M, Otis C, Lemieux C 2003. The mitochondrial genome of Chara vulgaris: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. Plant Cell 15:1888-1903

Unseld M, Marienfeld JR, Brandt P, Brennicke A 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat Genet 15:57-61

Wang Z, Zou Y, Li X et al. (13 authors). 2006 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

Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS. 2002 Intron size and genome size in plants. Mol Biol Evol 19:2346-52

Wikström N, Savolainen V, Chase MW 2001. Evolution of the angiosperms: calibrating the family tree. Proc Biol Sci 268:2211-20

Wolfe KH, Li WH, Sharp PM 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84:9054-8

Wool IG 1996. Extraribosomal functions of ribosomal proteins. Trends Biochem Sci 21:164-5

APPENDIX

Appendix 1A

Alignment of deduced S1 sequences from EST data for *Helianthus* (sunflower)and *Gossypium* (cotton), nuclear-encoded S1 in alfalfa, and mitochondrial-encoded S1 in rest harrow and tobacco, showing nuclear-specific amino acid substitutions (green)

```
CLUSTAL W (1.83) multiple sequence alignment


HELIANTHUS        -MTVY[ ]SRLFR[ ]SNSSF[ ]LSSGKALKSQVSRFRDDMLLVDAG[ ]GAP[ ]LTMPDELR--DIP 57
COTTON            -MSTY[ ]SRLFP[ ]SNSSM[ ]LCSGQAIKAEVLRLREQTLLVDAG[ ]GSPRICTQDETT-LTRS 58
ALFALFA_NUC       -MSIYLSRLFPRCNSSSFLCSGKALQSEVLRLGKETFLVDAGPGTP[ ]NCTRDELTRVPIN 59
RESTHARROW        MMSIYLSRSFPRCNSSLFLCSGKALQSEVLRLGEEMFLVDAGPGTPRNCMQDELTGVPIN 60
TOBACCO           MMSIYLSRLFPRSNSSFFLCSGNALQSEVLRLREEMFLVDAGLGTPRICMQDELTGVPIN 60
                   *: *:** * :.*** :*.**:*::::* *: .: :***** *:*:    **

HELIANTHUS        DVTNFENKVGFMEATAG--ETLIKTQLLERMFMDLVAGE[ ]D[ ]KERAIARFNESVG--STE 113
COTTON            AINRFEDKVGFFDKGVGGGESLVRKRFLERFFVDLVSGESI[ ]KERAAVRFNDLLGSRSLD 118
ALFALFA_NUC       PATRFENKVGFLDRAAG--ETHIRKKNLERLFIDLVAGE[ ]LIKERAAARFNDMAG--STD 115
RESTHARROW        RATRFENKVGFLDLVAG--ESLIKKKILERLFIDLVAGESLIKERAAARFNDLVG--STD 116
TOBACCO           RATRFENKVGFLDLVAG--ESLIKEQILERFFIDLVAGESLIKERAAARFNDLVG--STD 116
                    ..**:****::  .*  *: :: : ***:*:***:**.  **** .***:  *  * :

HELIANTHUS        TVAGEPAIVLSRRYTQNQALIELKKKWQSHKKVKGFPLEKVKGGYSVAIAG[ ]IAFLPFR- 172
COTTON            VVAGEPTLLNPRRFRQTLVWQELNKHWRNNSKVKGFFIDKV[ ]GGYSVAIAG[ ]VAFLPYSR 178
ALFALFA_NUC       VVADEPLLLLPRRFRQDRAWMKLNKIWRTNTKVKGFIIDKV[ ]GGYSVAIAGFIAYLPFRS 175
RESTHARROW        VVAGEPLLLLPRRFRQNRAWMKLNKIWRTNTKVKGFIIDKVKGGYSVAIAGFITFLPFRS 176
TOBACCO           VVAGEPLLLLPRRFRQNRAWMELNKIWRTNTKVKGFIIEKVKGGYSVAIAGFITFLPFR- 175
                  .**.** :: .**: *  .  :*:* *:.:.***** ::**:*********::::**:

HELIANTHUS        -ASLARHKLMNDKFTIETFNPKRKKLTVFYSFLFSSKHYYYYFAAVLFFST 222
COTTON            FGRKMVRDSDKDRFTIERIDPKKR[ ]LVVF---------------------- 207
ALFALFA_NUC       HSKRQRRRISNDQFTIESINPKNK[ ]IVVF---------------------- 204
RESTHARROW        HNKRKKRRISNDQFTIESINPKRTNIVVF---------------------- 205
TOBACCO           ---RRRKRISNDRFTIENINPKKTNIVVF---------------------- 201
                        :   :*:**** ::**. .:.**
```

Appendix 1B

Nucleotide alignment of *rps1* EST sequences for *Helianthus* (sunflower). Proposed *rps1* initiation codon shown in green, in-frame upstream stop codon shown in red.

```
DY919152.1            ATCAG[ ]CAGACA[ ]ACGGTGTACATGAGCCGATTGTTCCGGAAATCCAATTCAAGCT 74
DY919027.1            ------[ ]CAGACA[ ]ACGGTGTACATGAGCCGATTGTTCCGGAAATCCAATTCAAGCT 54
DH0AMM29ZF01ZZM1      ATCAG[ ]CAGACA[ ]ACGGTCTACATGAGCCGATTGTTCCGGAAATCCAATTCAAGCT 62
EL469775.1            ----------------ACGGTCTACATGAGCCGATTGTTCCGGAAATC-AATTCGAGCT 42
DY954844.1            ACCAG[ ]CAGACA[ ]ACGGTCTACATGAGCCGATTGTTCCGGAAATCCAATTCAAGCT 117
EL491471.1            ATCAG[ ]CAGACA[ ]ACGGTCTACATGGGCCGATTGTTCCGGAAATCCAATTCAAGCT 120
                                      ***** ****** ****************** ***** ****
```

Appendix 1C

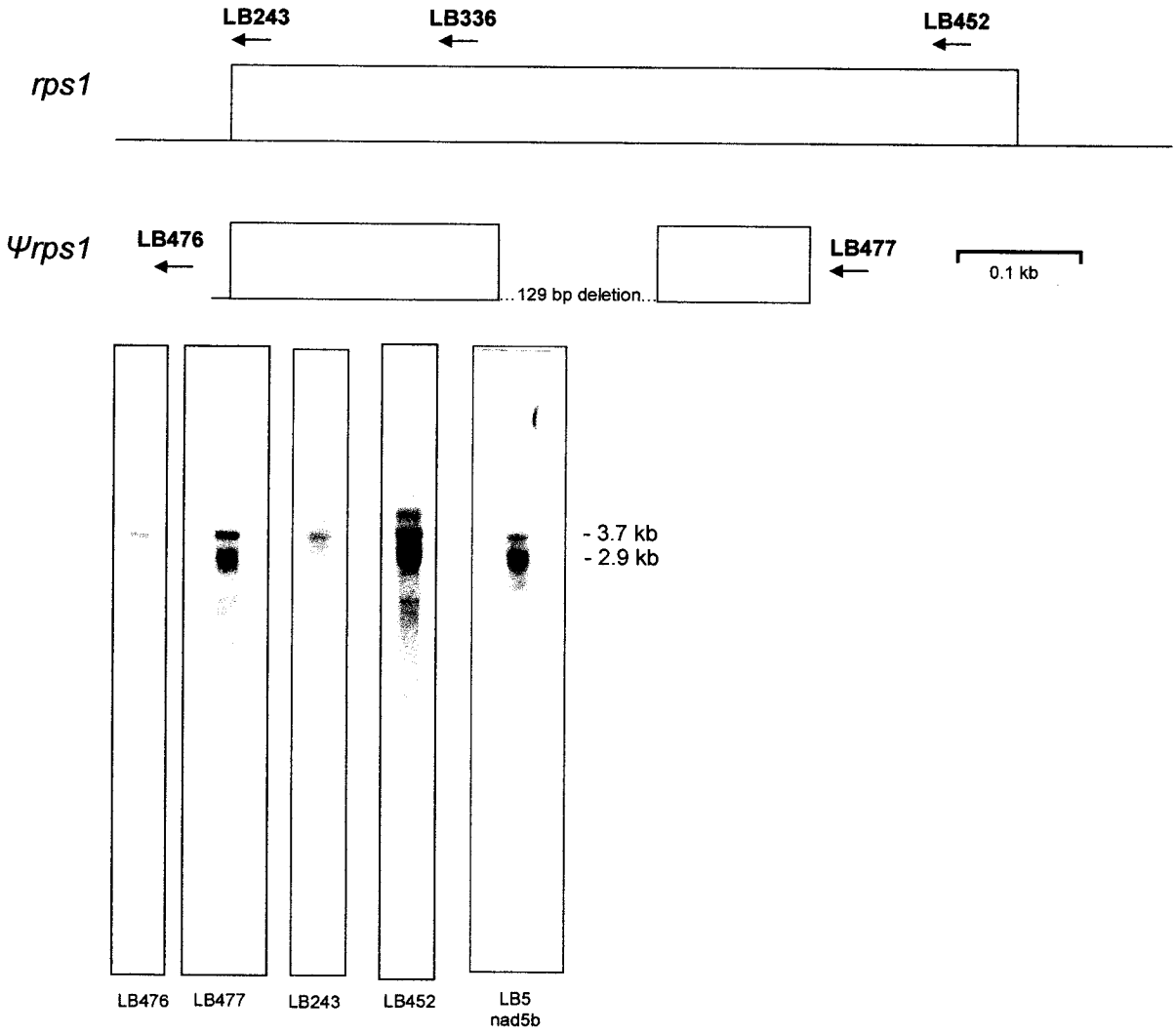
Nucleotide alignment of *rps1* EST sequences for *Gossypium* (cotton). AUG corresponding to alfalfa nuclear-located *rps1* initiation codon shown in green.

```
GossypiumBQ408259RC     ------------------------------------------------------------
GossypiumBQ414414RC     ------------------------------------------------------------
GossypiumDW519806.1     ------------------------------------------------------------
GossypiumDW226586.1     ------------------------------------------------------------
GossypiumBG447460       CGGCACCAGCCTGAGCAGCCTTTATTTTTCGCTCAGCCTACCTACCTCCGGCCTTGGCAC 60
GossypiumBQ405749       ----------------------------------GTAAATTCAAAAAAAAAAACAGAAA 25
GossypiumDR455986.1     --------------------------------TGCTTGGAAATCAAAAAAGAAAA-AAGAA 28

GossypiumBQ408259RC     ------------------------------------------------------------
GossypiumBQ414414RC     ------------------------------------------------------------
GossypiumDW519806.1     ------------------------------------------------------------
GossypiumDW226586.1     ------------------------------------------------------------
GossypiumBG447460       TGGCGAGCCTCCGCTGCGAGCACGTATATGAGTCGCTTGTTTACAAAATCTAATTCC 120
GossypiumBQ405749       AGAAAAGAACCCTCAAAGAGCACGTATATGAGTCGCTTGTTTCCAAAATCTAATTCC 85
GossypiumDR455986.1     AGAAAAGAACC-TCAAAGAGCACGTATATGAGTCGATTGTTTCCAAAATCTAATTCC 87
```

Appendix 2

Northern blot analysis of bean mtRNA showing cross-hybridization of a 3.7 kb transcript to *ψrps1*. This is evident in Figure 4.4A, lane 2 (for which the 32 P-labeled probe used was LB336). Shaded regions indicate homology. LB476 5'TTATGAAATTGCGCAGGAGG, LB477 5'GGAATCAACCAGATCTTCAC



Appendix 3

Raw sequence data for mitochondrial- and nuclear-located *rps1* sequences, including flanking regions.

>BEAN MT RPS1 RAW SEQUENCES
GTCATGCCTTTTCCTCTTCTTGCTTTCTATTTTCTTCCACTACTTTTTTCCTNNNNNNAGGTATACAAGTGCGTCTCTTTTGCTTCTATCTAA
CAAGGCCCAAGTTACAGCGGGCTCTTTCTATTATGAAATGGTTCATCTNNNTCTGTTTACTACTAGTCTCTCTATATATCATTTTCGATTTGG
TTGGGTTAAATCCCGACCTATTTTTCTTGAAACGTTATTTTATCACGGGTTACGTGGTCTTTTGAAAATCTTCGGTTTCTATATTCCAGGACT
TCTTATTGTGCAAATCGCTTCTCTTTTGTGTTACACCAGCTTGCCCATGGAGGATCCCGCAGGAGGTAACATGGATTCCTCTCTTTTGTGTTA
CACCAGCTTGCCCATGGAGGATCCCGCAGGAGGTAACATGGATTCCTCGGGGGGCTCCCACTCAATCTCAGTTGGAGGTGAGAGTACCATGGG
AGGGGGTAGTAGTAGGGGATCTGGTTGGACTTCATTTGACCTCGACGTCTTAGCGGAGCCCACCGCTAATGAAAATGAAAGAGGCGAGGAAGT
AGCCCAGCCGAATCCCCCAAACGCCCTGGATAATCCGGTCGCTTACCCGGGGGAGGCTCAGGAGGCGCTGCCGCAAGCGCCTGCACAGGCTGA
AGGAGCCCAGCCGCCCCCTACCCCTATAAAAAAAAATGAAATAATAGGCGGGGATAGTGTGGAATCCATTGAGCGGAGGCTTTTGGGGAGATTC
TCCTTCCGCCCATGAGATCACAATGGCCCGCCCGGATTGAGGCCGAAGACCTCTTCGAGGTCAAGGTGGAGATCATCCAGCAGATGGCAGTCC
TGATTGAGGCCGAAGACCTCTTCGAGGTCAAGGTGGAGATCATCCAGCAGATGGCAGTCCTTGATCCAAGTGGCGATTGGACGGGACGGGGAG
CGCGGGCCCTGGACAATCCCCGTGCGGCCTCAGGGGAACCTTCCTTAATAGAGCTCTATCGCCTTAGGGACGAGCTCCATGAGGGAGGAGTAC
AATCCGACCGATGCCTTTGAAAGGCTAAAGGAAAATCTTTTCACCAGGGTAGAGAATCTGGATGAGAACTCCACTACTTAATCTACTGGCGTG
GTGCTACTGGATGCTTTTTATCAATAGAACGGGAAGAGTAGGAAAAAAACATATTTTATGATGAGCATCTATTTGAGTCGATCATTTCCAAGA
TCTAATTCCAGTTTATTCTTATGTAGTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTCTTGGTGGATGCAGGACCT
GGGACCCCCAGAATTTGTATGCAAGATGAGCTTACAGGAGTGCCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTA
GTGGCCGGTGAATCACTGATCAAAAAGAAGATTTTGGAGAGATTATTCATCGATCTAGTGGCCGGCGAATCACTGATCAAAGAGCGAGCAGCC
GCCAGGTTTAATGATTTGGTGGGATCCACAGATGTAGTGGCTGGTGAACCGCTTCTTCTTCCACGAAGATTCAGACAAAACCTAGCTTGGATG
AAACTGAACAAGATTTGGCGAACGAATACAAAGGTAAAAGGCTTTATTCTTGATAAAGTAAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTC
ATTACTTTTCTTCCATTCCGTTCTCACAACAAAAGAAGAAGGAAAAAGATATCGAATGATCGATTCACCATTGAGAGCATTAACCCCAAAAGC
AAAAGGACGAATATTGTGGTGTTCTAACAGCGGCAGATCAAACAAGAACTTTATGAGCGAAATCTGCTGACGAAAATCTTTCCTTTTTTTTTT
TCAATTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAATATTATTTTGTTTAGGGTCTGGCACTTATTACAGGCCGCTCT
GTCATTGTCTGATTTTTTGTTGTCTGATCACACTCGAAATTATGTATCTACTTATCGTATTTTTGCCCCTGATCGGTAGTTCCTTTGCAGGTT
TTTTTGGACGTTTTTTAGGATCAGAAGGAAGCGCTATAATGACCACTACGTGCGTTTCATTCTCTTCGATCTTATCTTTGATTGCTTTTTATG
AAGTCGCACTGGGAGCTAGTGCTTGCTATCTAAGAATTGCTCCATGGATCTCATCGGAAATGTTTGATGCTTCTTGGGGCTTCTT

>FENUGREEK NUCLEAR RPS1 INTRON
AATTAGTAGTTTTGAAGTGAATACTGGTATGCTTAGGAAAATTGAAGAGCCACTAGATCCGTGGCTCAAAGACATGCATGCAGTCTTGACACA
TTAAGATAAAAGTAATTAGTAGTAGAAATTCAAAGAAAGTGGACAAATAATGAATCTTCAAAACCTCTATTTTCGTCCATTCCATAGAGGTAA
GATTTTACAATTTCATCTTTGTATTGTGTTTGTTCCTTTGGTGCTGTTTTGTTTCTCCGTCATTATATGATGTGTGCGCTGTCCCCATTTCCT
ACTTTCTCTGGCAATGTGTTTTAGGTTATGTGTGTTTTCTTGTTTTTTTTTTtttATAAAAAAGGAACTGATGTCATCTTATAAAAAAAGGGT
GATCGCCCCTGACCTTTTTTtATTCATTTTTTAATCAATAGTTCTATTACACTGATTAG

>ALFALFA NUCLEAR-LOCATED RPS1 INTRON
...ATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTCTGGTGGGGTGATT
GAGATAAAAGCAATCAGTAGTTTTGAAGTGAATGTTGGTGACTGGTATGCTTAGGGAAATTGAAGAGCCACTAGATCCTTGGCTCTCGAAGAT
ATGCCTGCAGTCGTGACACATTGCTCCCGTTAGTTTTTTGGATGGAATTTTAGATTTAGTCCACTCCATAAAGATAAGATTTTAAAATTTCAT
GTTTTTGTTGTGTTTGTTTCTCCATCAGTATATGATGTGTGTGCTGACCACATTTGTGTTATCTTACGTTTGTTTCTCTGGCAATGTGTTTTA
GGTTATGTGTGTATTCTTGTTCTTCATTTTATGGCTTTCTTATTAATTTTTTTTTTTTTAAAAGAAACTGATCTCATATATAAAAGAAGGGTG
ACCGTCCCTTTGGCCTATTTTTATTCATTTCATCATATTATTCCTTTTTTAATTGATAGTATTATTACATTGATTAG(EXON2)

>SWEET CLOVER NUCLEAR RPS1 INTRON
AATTGGTAGTTTTGAAGTGAATATTGGTGTGCTTAGGAAAATTGAAGAGCCACTAGATCCATGGCTCAAAGACATGCATGCAGTCTTGACACA
TTAAGATAAAAGTAATTGGTAGTAGAGATTCAAAGAAAGTGGACAAATAATGAATCTTCAAAACCTCTATTTTCGTCCATTCCATAGAGGTAA
GATTTTACAACTCATCTTTGTATTGTGTTTGTTCCTTTGGTGCTGTTTTGTTTCTCCGTCATTATATGATGTGTGCGCTGTCCCCATTTCCTA
CTTTCTCTGGCAATGTGTTTTAGGTTGCGTGTGTTTTCTTGTTCTTTTTTTTTTTTATGGCTTTCTTATTCATTTWAAAAAAAAAAAaGGAACT
GATCTCATCTTATAATAAAAGGGTGATCGCCCTTGACCTTTTTTTTTAATCAATAGTTCTATTACACTGATTAG

>ALFALFA MITOCHONDRIAL-LOCATED ΨRPS1
AGTTGACGACAATGCTAAACCTGTCACGACCGTAGTGGGTCAAATAGAATGATTCCTATATTACCTTGCATA
AAGAAAGCAGCTCGGTCTATGAGTCCAATATATCCGTACCTGTCAAGTGGAAGAGTAAGAGTGGCAGAGTCTTGGAGAGCGCAGAGTCTAATC
TGGTAAATATCTCCTATAATAAGGCGGAGAACGGATAATAACTCGGCAAATGTCCGATTTCGGAGATAAGTGTGCCTTTCTTTTGATCAATAG
AACAGGAAGAGGAGGAAAAGAAAAGCTTATGATGAGCATCTATTTGAGTCGATAATTTCCAAGATGTAATTCAATTTTATTCTTATGTAGTGG
AAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTATTGGTGGATGCAGGACCTGGGACCCCCATAAATTGTATGCAAGATGA
GCCTACAGGAGTGCCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGTGAATCACTGATCAAAAAGAA
GATTTTGGAGATATTATTCATCGATCTAGTGGCCGGCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCAC
AGATGTAGTGGCTGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGAGCTTGGATGAAACGGAACAAGATTTGGCGAACGAATACAAAGGT
AAGGTAAAAAAAAGGCTTTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTATTCCATTCCGTTCTCA
CAACAAAAGGAATAAAAGGAGGATATCGAATGATCGATTCACCATTGAGAGCATTAACCCCAAAAGGACGAATATTGTGGTGTTCTAACAGCG
GCAGATCAAACAATAATCTTGATGAGCGCTGACGAAAAAAGTGCAGTTTTTTTTCAATTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCT
ATCACCTTAATCCATTCTTTTGTTGAGGGTCTGGCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTTTCTGATCACACTCGAAA
TTATGTATCTACTTATCGTATTTTTGCCCCTGCTCGGTAGTTCCGTAGCAGGTTTTTTCGGACGTTTTCTAGGATCAGAAGGAACCGCTATAA
TGACCCTACGTGCGTTTATTTTCTTCGATCTTATCTTTGATTGCTTTTTATGAAGTCGCACCGGGAGCTAGTGCTTGCTATCTAAGAATTGCT
CCATGGATCTCATCGGAAATGTTTGATGCTTCTTGGGGCTTTTTTGGCGACCGTGAAGTCACCGGATGAATTGCCGAGTAGATAGATCAGAAC
CCGACGCGGCTGTTGCTCCGCGGCGATACGGACTTGACCCGCTCTACCCCCGGGGGCACATAGCATGTCGGGAAGAAKGGGGGGAYATAYTCG
RYGWACYACTCCYTTTGGGGCTGTGCGCYCTGCYTTTYGRATCGMTAA

>PEA MT RAW SEQUENCES
CAAGACATTGCAACCCACCTTCATCACGATTCCGAGTCTTTGGAGCAACTCTTGACCATTTTGAAGAATCTTACTGAATTAGGTCTACAAAGT
CAAGAGTTTCAACAAATTATCCTCTTTCTCTCGCAGTGAAGTCCCGATGGCAAACTATATATAACATTTGAATTTCTCTTACATTCCACGTTT
CCGAAACGGATCCTATCAAATATTTTACATTTTCTATGATCATCTCTATTTTAGGTATTCGGGGAATCCTCCTTAATAGACGAAATATTCCTA
TTATGTCAATGCCAATTGAATCAATGTTATTAGCTGTTAATTCGAACTTTTTGGTATTTTCCGTTTCTTCGGATGATATGATGGGTCAATCAT
TTGCTTCATTGGTTTCAACGGTGGCAGCTGCGGAATCAGCTATTGGGTTAGCCATTTTCGTTATAACTTTCCGAGTCCGAGGGACTATTGCTG
TAGAATTTATTAATAGCATTCAAGGTTAAACATGACTCCTAGAGAATTACCAAAATACGAAGTTATCTTTTCTCCTTTCGTTCTCTTCTTTAT
TTATTTTTTATTTTTACTTGGTTGGCAGGGCAGGGTCGGGGCCTTTATCGCTGGGCGAGCGCATCCGATTTATTCTAAAGTCCTTTCCTAAAC
CACTTCCCGTTTAGTTGCTGATAGATAGAACCTTTCTAAATGAGATTGAGTTCCACGCATATGCAAGCTAGAAAGATGCTATTTGCTGCTATT
CCATCTATTTGTGCATTANGTTCGAATAAGATATCNAATCTATAATGAAGAAATGATAGTAGCTCGTTGGTTTTATAGGCTTCATCATATTCA
GTCGGNAAAAGTTTAGGTAATACNTTTCAANGTAACTCTCGACGGGAGAATCCAGGCTATTCAGGAAGAATCGCAGCAATTCCCCCAATCCTA
ACGAAGTANTTNCCCCNGGAATCTAANGANCAACACCGAATTACTTAGGATCNGTTTGGCNAANTTGNGGGCNCCGTANTAAAANCCNTNNCC
ANGGGNCCCCCTGGGNCNCCTAAGNGCNA--(RESTRICTION MAPPING SUGGESTS THAT THESE SEQUENCES OVERLAP, OR
THERE IS A SMALL GAP)--
TGGAAGCCGAAAAAATNGGAATTCCTTCGGGAGNGCTTGGGGGTTNTTAAGAAATGGTTCNGGGGGGGGGGGNTTTTTTTNAGAATNAATAAGN
CGAATNGNATTTAAATNCANGTTTATGCTAACNGAAGGAGCGGATCCAATNCAAGGACAACTTTTTTTTCCATGAAGCAGCTTGCAAGAAAAG
GAGTTAGTTCCCTCCGGGGGTTCTGATAGCCAATGCTTCNTACGCTCCAGCCCNTAGCCCTGAACTCCTTAGCACAAGCACTAAGTAAGCCCG
TGAAGGAAAGGCCTTCAATATCGCACCGCACCTACTAGGACTGAAATTTAGTTTTCGAATAGGACTACTTCTTTCTTTTCTGTACGCTAGGGC
CGCCCGCACGCTTTAAGGGTTGTTGAACGTATTGCACTAAGAAAGACTCCTGCCGTTGCGGTAAGATTTTCTTTGAAAGCCGCTCAAGCAATT
TCTTTAAAGAAGAGCAAGCAAGCAAGACAAGATAAGGCTCGAAAGCTGCGAAGAAAGTCGAGCTGCTCTAATCGAGCTGTGAAATCGGTGTGG
TGAGGTATTGGAAGTGTTCAGCGAATATAGATAAAGAAACTCAAAAAGAGCACCGTGCTATACTGGCAGAGCTTCAGCAAAACTAGGTAGGTA
GCTTTCGCTAATGAGAGATGGATTTGGGGATTATATTAGGTCTCCTAATTAATGGAATTAGTTTTCTCTTGATTAGTTTCTTTTTAAAGAAGC
ATTCTCCCTTTACTCGATCTTTAATTGAATTATCGATAGAACGTTGATCAATATTAAGGAAGAGATCGAGGATCACAGCAGGATCCTATCTGA
TCTTTGTCAACACAGGACTCGAAGCCCCCTGCTCAAACAAAAGGGTAATCCTTCGGAATTGAAAATTCCTTTCGGATTTCTTATATTAATTAT
AGTTGAGGGGAGAAGACTCATTTAAAATTCTATCTATATAGGGGGGGGGAGCCATTGTGAAGCCTAGAGAGGCGGAAGCGGCAAATCGCTTCT
ACCGAGTTCCACAACACTTAGCTTAGTAGCTTAACTCTTTCTACTGGCGTGGCGCTGCTGGATGCTTCTGATCAATAGAACAGGAAGAGGAGG
AAAAAAAAGCTTATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATGTAATTCAAGTTTATTCTTATGTAGTGGAAAGGCCTTACAATCT
GAAGTTTTACGCTTAGGGGAAGAAATGTTCTTGGTGGATGCAGGACCTGGGACCCCCAGAAATTGTATGCAAGATGAGCCTACAGGAGTGCCA
ATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGTGAATCACTGATCAAAAAGAAGATTTTGGAGAGATTA
TTCATCGATCTAGTGGCCGGCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCACAGATGTAGTGGCCGGT
GAACCGCTTCTTCTTCTTCCACGAAGATTCAGACAAAACCGAGCTTGGATGAAACTGAACAAGATTTGGCGAACGAATACAAAGGTAAAAGGC
TTTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTCTTCCATTCCGTTCTCACAACAAAAGGAAAAAA
AGGAGGATATCGAATGATCGATTCACCATTGAGAGCATTAACCCCAAAAGGACGAATATTGTGGTGTTCTAACAGCGGCAGATCAAACAATAA
CTTGATGAGCGAAATCTGCTGACGAAAAAAGTGCTGTTTTTTTCAATTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAA
TCCATTCTTTTGTTGAGGGTCTGGCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTGTCTGATCACACTCGAAATTATGTATCT
ACTTATCGTATTTTTGCCCCTGCTCGGTAGTTCCGTAGCTGGTTTTTTCGGACGTTTTCTAGGATCAGAAGGAACCGCTANTATGACCACTAC
GTGCGTtTCATTCTCTTCGATCTTATCTTTGATTGCTTTTTATGAAGTCGCACTGGGAGCTAGTGCTTGCTATCTAARAATTGCTCCATGGAT
CTCAGaGRAAATGTNTGATGCTTCTTGGGGCTTCTT--(intron 1)--
GTTCGATAGCCTGACCGTAGTGATgTtAATTGTGGTwATCATAAGTAGCTTGGTCAATCGAATTCCCAGGCGCCATGGCGGCHGGGAGCATGC
GACGTCGGGCCCAATTCGCCCTATAGTGAGTCKTA...ATTCNGTTAGCGTCTTTCACTTAATGAATCACGCNTTTTTCAAAGCATTACTATT
CCTGAGTGCANNTTCNNTNATTCATGCCATNTCNNATGAGCAANATATGCGGAAGATGGGGGGGCTTGCCTCCTCGTTCCTTTTTACCTATGC
CATGATGCTCATGNNCAGCTTATCTCTAATTNNATTTCCTTTTCTAACTNNATTTTATTCCAAAGATGTGATATTAGAGCTGGCTTACACTAA
GTATACCATAAGTGGTAATTTTGCTTTCTGGTTGGGAAGTGTAGAGCTGTCTTACACTAAGTATACCATAAGTGGTAATTTTGCTTTCTGGTT
GGGAAGTGTNTCTGTCCTTTTCACTTCTTATTACTCTTNTCGTTCACTTTTTCTAACATTTCTAGTACCAANTAATTCAGTCGGGCGAGACNT
CTTACGATGTCATGATGCGCCCATTCCTATGGCCATTCNTTTAATACNTCTGGCTGTCGGGAGTNTCTTTGTAGGANACTTGGCCAAAGTGTG
ACCCGTTAGCCTNTAAGTAAGTANTGTGACGAAGCGGNTGTTGCTCACCCGANACGNTCGTACGANGTNACAANTCACCCAANACGNTCATCC
GGNGTGNACAGNAATTGGGGNTNGGANGCGGGCGANANNCCCcGCCAATGGNTGAGATGNTCAGTCGACTNCCNNCCCCNTNNNNGGGGKCCG
GACNCCTNACGAGTGAGCAGRAAAGGGAGGNGGANAGAGGNCNTGGNGANCNGNGAN

>REST HARROW MT RAW SEQUENCES
GTAGGAAAGGGAGCCAATGATAGATGCTTTGAAATAGCGTAATAGAAAGAAAGATTGTCTCAATATCAAGGCAAGCTTTAGAGAATAGGGGCG
AGTAATTGAATAAATCCCTTCCATGGATTTTCCGGAAGAGTCACTCGCCAGAACAGGGGAAGGGAATTCGGAACTCAATCACCAAGTGGACCT
TCCCATTGTCTATGCCTGCAACAAAGGAAGAGGTGAGCTACCAACTCGTCACCTCCCCTATCGTTCCACTTCTGTCTGTTCTTTCATGGCTAT
TCTTTCATTGTCAAGTGTCGGACTCTGGTCTTTTTGGAACATTAAAGTGTTCCGTGAGTTAGATTTCCCCTCTCTTCTTTAATCTTTTCTTCC
CAAAGCTCACTAATTTAGTGATAGAGGCTCTAATGCTGCTAGCGACGGGTAGGTACTAGGAAGAGTTTACGGTATAAAATTTCTCGACCTATA
GCGAAAAAAGAAAAAGATAAGAAAGCTGCTAGCAGAGGGTGGAGAAGTCGTCTGGTTATCAAAATTATGTATAGAAAGAAGATTCGTCTTGCC
GGTAAATGCTGTAGGTAGACTGAACCTTGTTCTCTATCCGCTAGCACGCAGACAAAAACTAGGATAAATTTATGGCAAAGTCAGTGTGCATCG
AAGGTCAAGAAACTCGAGTCATAACATTTCTACAAGTGCTTAGGTGGGCTCATTCCGAACCTAATCGTATCGTTTGTGAATCTTTCTCCCATG
CATTTCTGGATCAACCAACCAGCGATTTACAACAAACAAGTCTTTCCTCTTGTTGGGGGAGCAGTGAAAAAATGAACCAAAGCAAATGATATT
CCAATTTTTTGATTGGGTGAAACGTAATTGTTGCTGTTGCTTGCGCGTAGAAAAGCCGACTGGTTCAGCTTCAGCAGTCCTGCCGACTGCTGA
ACCAGTCCTGCCTACTAATGTTTCGGCGGGTAAAGCAGAATTTGAGAAGGGCCCCAGTTTTAGTGAATGGAAAGTAAATATGAAGCCACCAGT
AGGGGAGGGCCCTAGCGCATTACCCCTCGACTCATCTGCCGTACCCGGCCCTCTTGAACAATTTGAGATAATCCCTTTTCTTCCTATGAAGAT
AGGTGACTTGTATTTCTCATTCACAAATCCCTCTTTGTTTATGCTGCTCACTCTCAGTTTGGTCTTACTGCTGGTTCATTTTGTTACTAAAAA
GGGAGGAGGAAAGTTAGGATGAGAACTCCAGTACTTAATCTACTGGGGTGGTGATTAGGTCCATTCCGTGGCGCTGCTGGATGCTTTTGATCA
ATAGAACAGGAAGAGTAGGAAAAAAAAAGCTTATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATGTAATTCAAGTTTATTCTTATGTA
GTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTCTTGGTGGATGCAGGACCTGGGACCCCCAGAAATTGTATGCAAG
ATGAGCTTACAGGAGTGCCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGTGAATCACTGATCAAAA
AGAAGATTTTGGAGAGATTATTCATCGATCTAGTGGCCGGCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGAT
CTACAGATGTAGTGGCTGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGACAAAACCGAGCTTGGATGAAACTGAACAAGATTTGGCGAA
CGAATACAAAGGTAAAAGGCTTTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTCTTCCATTCCGTT

CTCACAACAAAAGGAAAAAAAGGAGGATATCAAATGATCAATTCACCATTGAGAGCATTAACCCCAAAAGGACTAATATTGTGGTGTTCTAAC
AGCGGCAGATCAAACAAGAACTTGATGAGCGAAATCTGCTGACGAAAAAAGTTCAGTTTTTTTTAATTCCGAAGCGAAACCTAGCGTCTCCTG
ATAACACTCTATCACCTTAATCCATTCTTTTGTTGAGGGTCTGGCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTGGTTGTCTGATC
ACACACACTCGAAATTATGTATCTACTTATCGTATTTTTGCCCCTGCTCGGTAGTTCCGTAGCAGGTTTTTTCGGACGTTTTCTAGGATCAGA
AGGAACCGCTATAATGACCACTACGTGCGTTTCATTCTCTTCGATCTTATCTTTGATTGCTTTTTATGAAGTCGCACTGGGAGCTAGTGCTTG
CTATCTAAGAATTGCTCCATGGATCTCATCGGAAATGTTTGATGCTTCTTGGGGCTTCTTGTTCGATAGCCTGACCGTAGTGATGTTAATTGT
GGTTA

>SOYBEAN RPS4 RAW SEQUENCE
CATAGTCCCGNGGCATTCCCCNTGACCTCGAGATAAAATGAAAAAACGGACGATAAACAGTTTGATTGTTAAGGGTACCATTATTTCATTAGT
AAGGGAATTTTTATTGGAACGGGCTTCGATTCCCTTTATACGCGATGTGGCGAATCAGACTGATTTAACGACGACGAGATATGCGATTTAAAA
CTTGTCGTCTACTTTCAGGAAATGTTCGGAACAGAGAACTTACAATAATACAACGCCGCGTTCTCCGAAGATTGAGGAACAAGAAGAGATCTA
TTAAGAGAAAGATTTCTCCGAGAAAAAATATGAACAGTTACATCCAATCACAAACTACACGAAAGTTGCCCCTTTTTCATGCAGATTTACCCA
TCACAGAGATGCACAGAGGAACAGAACGAACTTCATATATCCCTTTTCTACTCAATCCAGAAACAAGATCGGACGTTATTCCGGTTCGTCTCC
ATTTTCGTGAAACTATTCCTCAAGCAAGGCAGCCGATAAGTCATCGAAGGGTTTGTGTGAATAATCGAATTGTAAGCATTACTCGTTTGAAAG
TTTCCCACGGTGATCTAATATCTTTTCAAGAAAATGACGCGAGAATCCGCGGTGAAGAAATAAGGAGATCTTTCTATATCGAAATATCAGTTG
ATAAAATCATAGGCAAATTCCTGGATCACCCGGTAAGAATGTGGAGAAGAACCAAAACAGAATGGTTCCACCTACTCAAAACTAAGCGGGGAT
GCCGCCTACTACTAAAATCCCGGTTTTTGCAACAACAGTTGCGTTATTCTATGCAAGAAGAATACTTAGAAAGAACAAAGAAGTTTG

>SOYBEAN RPS1 RAW SEQUENCE
GATCCGAAAAAGTATGCTTAGGCAGTTCCTTCGCTGAGCACAACAGAATGAAGAGGAATTTGTATCATTTCAAGTCCCTATTCTTATCGAAGA
GAAGAAACGAGAAAAACCAAGATCTTCCTACTCGAACAAGAAGTCCTATAGTTTACAACTCTTCTTTATATAGTAATTCGACATATTGCTCCT
CATCCCCCCATCCGTTTCCTAGGAAGAGAAGAATCAAAAGGATCGAACTACCTACTCATTATTCGGAGGTGAATCATAGAACACCAAAAGCGG
TGGTATTTTATGGACCTAACATAGGTCACATCCCTCACGACATAAGATTGAAAGATCCAAACCTTCCTCTTCGGAGCGGAAACGGACGTGGCC
AAAACATATAAAGATCGGCGTAGTCGCTCATAGGGACATATCGATCCGGATAGAGGATAGTCTAGATCGATTCATAGATAGATTTCTATCCAT
ATAGAAAAGAAATGAATTGATCGCTAATAGAATATTAAACAACAACGGAGCAGAGTATACTATTGAAACTCTCCAAAAGAACCTTGAAAAATT
AACAACCGAAGGCCAAAACTCTTTCTTTTTTAAAAAGCTTATAAAAGAGAGACCTTCGAATTAAACGGCCGATTTGATTAACAAGCTGCTCCT
ATGCGCAAGAAGCATTCCAAGATTTTTTGAATGAGAAACGCTGTTCCCAAAAGTCCCATGTCTTTTTTGGCCAACAACCAACCACAATCAAAA
CACTAAACACTAGTCCTACTCCATCAGCTACAGGCTTTTCGGAGTGTCTGGAGGGCAGAGAAGCTTAGACCGCTGAAGGAACTTTCTTATAGA
GTAGGGAAAACAGTAAACTTAATCAACTCCGTATACGAATAGAAACAAAAGAGAGACAAAACAGTCAATGTTGAGTTGAGCCATCGACAACTC
TCGTGCAAGCACAAGGAGGAGTAGCCCGATGCTTGTGAGTGACACGGATAAGTGCACTTATTATTTATAAAATAATAGAGAAAAATAAAAAGT
TCGGCGAATAGACCGATTTGAATCGAACACTCAACACCGGCGTGGCGTTTCCGTATGTTTTTGATCAATAGAACAGGAAGAGGAGGAAAAAAA
AGCTTATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATCTAATTCCAGTTTATTCTTATGTAGTGGAAAGGCCTTACAATCTGAAGTTT
TACGCTTAGGGGAAGAAATGTTCTTGGTGGATGCAGGACCTGGGACCCCCAGAATTTGTATGCAAGATGAGCCTACAGGAGTGCCAATCAACC
GAGCCACCAGGTTTGAGAATAAGGTGGGATCCCTGGATCTAGTGGCCGGTGAATCACTGATCAAAAAGAAGATTTTGGAGAGATTCTTCATCG
ATCTAGTGGCCGGCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCACAGATGTAGTGGCTGGTGAACCGC
TTCTTCTTCTTCCACGAAGATTCAGACAAAACCTAGCTTGGATGGAACTGAACAAGATTTGGCGAACGAATACAAAGGTAAAAGGCTTTATTA
TTGATAAAGTAAAAAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTCTTCCATTCCGTTCTCACAACAAAAGAAGAAGGAAAA
AGATATCGAATGATCGATTCACCATTGAGAGCATTAACCCCAAAAGGACGAATATTGTGGTGTTCTAACAGCGGCAGATCAAACAATAACTTT
ATGAGCGAAATCTGCTGACGAAAATCTTTCCTTTTTTTTTTCAATTCCGAAGCGAAACCTAGCGTCTCCTGATAACACTCTATCACCTTAATAT
TCTTTTGTTGAGGGTCTGGCACTTATTACAGGCCGCTCTGTCATTGTCTGATTTTTTGTTGTCTGATCACACTCGAAATTATGTATCTACTTA
TCGTATTTTTGCCCCTTATCGGTAGTTCCTTTGCAGGTTTTT

Appendix 4

Nucleotide alignment of mitochondrial *ψrps1* sequences in alfalfa, sweet clover and fenugreek, and pea mitochondrial *rps1* sequences showing out-of-frame regions (shaded) and in-frame stop codons (bold). Note that the barrel-medic sequence (AC145156) is identical to that of alfalfa.

```
ALFALFA          ATGATGAGCATCTATTTGAGTCGATAATTTCCAAGATGTAATTCAATTTTATTCTTATGT
SWEETCLOVER      ATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATGTAATTCAAGTTTATTCTTATGT
FENUGREEK        ATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATGTAATTCAAGTTTATTCTTATGT
PEA              ATGATGAGCATCTATTTGAGTCGATCATTTCCAAGATGTAATTCAAGTTTATTCTTATGT

ALFALFA          AGTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTATTGGTGGAT
SWEETCLOVER      AGTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTATTGGTGGAT
FENUGREEK        AGTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTATTGGTGGAT
PEA              AGTGGAAAGGCCTTACAATCTGAAGTTTTACGCTTAGGGGAAGAAATGTTCTTGGTGGAT

ALFALFA          GCAGGACCTGGGACCCCC---------ATAAATTGTATGCAAGATGAGCCTACAGGAGTG
SWEETCLOVER      GCAGGACCTGGGACCCCC-----ATAAATTAATTGTATGCAAGATGAGCCTACAGGAGTG
FENUGREEK        GCAGGACCTGGGACCCCCACCCCAAAAATTAATTGTATGCAAGATGAGCCTACAGGAGTG
PEA              GCAGGACCTGGGACCCCC---------AGAAATTGTATGCAAGATGAGCTTACAGGAGTG

ALFALFA          CCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGT
SWEETCLOVER      CCAATCAACCGAGCCACCAGGTTTGAGAATAAGGCGGGATTCCTGGATCTAGTGGCCGGT
FENUGREEK        CCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGT
PEA              CCAATCAACCGAGCCACCAGGTTTGAGAATAAGGTGGGATTCCTGGATCTAGTGGCCGGT

ALFALFA          GAATCACTGATCAAAAAGAAGATTTTGGAGATATTA-----TTCATCGATCTAGTGGCCG
SWEETCLOVER      GAATCACTGATCAAAAAGAAGATTTTGGAGATATTATATTATTCATCGATCTAGTGGCCG
FENUGREEK        GAATCACTGATCAAAAAGAAGATTTTGGAGATATTA-----TTCATCGATCTAGTGGCCG
PEA              GAATCACTGATCAAAAAGAAGATTTTGGAGAGATTA-----TTCATCGATCTAGTGGCCG

ALFALFA          GCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCACAG
SWEETCLOVER      GCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCCCAG
FENUGREEK        GCGAATCACTGATCAAAGAG--------CGCCAGGTTTAATGATTTGGTGGGATCCACAG
PEA              GCGAATCACTGATCAAAGAGCGAGCAGCCGCCAGGTTTAATGATTTGGTGGGATCCACAG

ALFALFA          ATGTAGTGGCTGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGA---------GCTT
SWEETCLOVER      ATGTAGTGGCTGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGA---------GCTT
FENUGREEK        ATGTAGTGGCTGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGA---------GCTT
PEA              ATGTAGTGGCCGGTGAACCGCTTCTTCTTCTTCCACGAAGATTCAGACAAAACCGAGCTT

ALFALFA          GGATGAAACGGAACAAGATTTGGCGAACGAATACAAAGGTAAGGTAA---AAAAAAGGCT
SWEETCLOVER      GGATGAAACTGAACAAGATTTGGCGAACGAATACAAAGGTAAGGAAGGTAAAAAAAGGCT
FENUGREEK        GGATGAAACTGAACAAGATTTGGCGAACGAATACAAAGGTAAGGTAA---AAAAAAGGCT
PEA              GGATGAAACTGAACAAGATTTGGCGAACGAATACAAAGGTAA------------AAGGCT

ALFALFA          TTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTA
SWEETCLOVER      TTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTC
FENUGREEK        TTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTC
PEA              TTATTATTGATAAAGTCAAAGGAGGTTATTCAGTAGCCATCGCGGGTTTCATTACTTTTC

ALFALFA          TTCCATTCCGTTCTCACAACAAAAGGAATAAAAGGAGGATATCGAATGATCGATTCACCA
SWEETCLOVER      TTCCATTCCGTTCTCACAACAAAAGGAAAAAAAGGAGGATATCGAATGATCGATTCACCG
FENUGREEK        TTCCATTCCGTTCTCACAACAAAAGGAAAAAAAGGAGGATATCGAATGATCGATTCACCG
PEA              TTCCATTCCGTTCTCACAACAAAAGGAAAAAAAGGAGGATATCGAATGATCGATTCACCA

ALFALFA          TTGAGAGCATTAACCCCAAAAGGACGAATATTGTGGTGTTCTAA
SWEETCLOVER      TTGAGAGCATTAACCCCAAAAGGACTAATATTGTGGTGTTCTAA
FENUGREEK        TTGAGAGCATTAACCCCAAAAGGACTAATATTGTGGTGTTCTAA
PEA              TTGAGAGCATTAACCCCAAAAGGACGAATATTGTGGTGTTCTAA
```