Conserved control signals in the transcriptome of higher plants

Khanh Tran

Thesis submitted for the degree of Doctor of Philosophy

May 2010

Discipline of Plant and Pest Science
School of Agriculture, Food, and Wine
The University of Adelaide

CHAPTER 4

FUNCTIONAL TESTING OF uORFs

CHAPTER 4 FUNCTIONAL TESTING OF uORFs

4.1 INTRODUCTION

It is well established that uORFs found in the 5′-UTR of mRNAs can down-regulate mRNA translational efficiency and/or stability (Kawaguchi and Bailey-Serres 2005; Mignone et al. 2002). Although progress has been made to elucidate the function of individual uORFs in certain species, it remains unclear what properties of a uORF determine whether they are functional or not. Recently, there have been several reports that used comparative approaches to identify putative functional uORFs (Crowe et al. 2006; Hayden and Jorgensen 2007; see Chapter 3). These approaches use sequence conservation as an indicator of a functional uORF, and are based on the rationale that essential genes are more evolutionary conserved than non-essential genes (Amsterdam et al. 2004; Jordan et al. 2002; Zhang and Li 2004). Therefore, it is expected that conserved uORFs are more likely to be functional, but the only way to know for certain is by experimental evaluation.

The comparative approach described in Chapter 3 identified conserved uORFs in higher plants. It reported that conserved uORFs are rare in the rice transcriptome, are generally short (less than 100 nt), highly conserved (50% median sequence similarity), position independent in their 5′-UTRs, and their start codon context and the usage of rare codons for translation does not appear to be important. As proof of concept on this work, two candidate rice genes, *SAMDC* and *S6K*, were chosen for functional evaluation.

The rice *SAMDC* 5′-UTR contains two overlapping uORFs: the tiny and the small uORF. The small uORF was chosen as a positive control for this study because its orthologous uORF is known to repress translation in other plant species (Hao et al. 2005; Lee et al. 1997; Mad Arif et al. 1994; Tassoni et al. 2007). For example, the Arabidopsis *SAMDC* small uORF (153 nt)

repressed translation of a GUS reporter gene by 3 to 5 fold (Hanfrey et al. 2002). Furthermore, the complete or partial removal of the uORF resulted in severe growth abnormalities, apparently due to unregulated over-expression of *SAMDC*. The other candidate chosen for further study is the long uORF (192 nt) of S6 ribosomal kinase (*S6K*), a key gene in the regulation of cell growth and energy metabolism (Wang et al. 2008). The *S6K* uORF is one of the longest conserved uORFs identified in the cereal transcriptome (Chapter 3). A long conserved uORF is less likely to have occurred by chance, and therefore may potentially affect translation.

The aim of this study was to determine if the conserved *SAMDC* small and *S6K* long uORFs from rice are functional. Mutational analysis of the rice *SAMDC* and *S6K* uORFs were performed, and a rapid quantitative *in vitro* transcription and translation system was used to evaluate their function.

4.2 MATERIALS AND METHODS

4.2.1 Plasmid construction

The 5′-UTR sequence of *SAMDC* (AK100589) and *S6K* (AK072649) were PCR amplified from KOME fl-cDNA clone vectors (Lambda FLC) using primers flanked by a *Hind*III or *Bam*HI restriction site at their 5′ ends. The primer sequences used were 5′-AATTAAGCTTTAGCCGCACCGCACGCTT-3′ (*SAMDC* forward primer, *Hind*III site underlined), 5′-ATTGGATCCTGGAGCAGGTTGGTCAGAA-3′ (SAMDC reverse primer, *Bam*HI site underlined), 5′-AATTAAGCTTGGAGTTCGCCGAGCCGAG-3′ (S6K forward primer, *Hind*III site underlined), and 5′-ATTGGATCCCTTCAGGAACTTGAGATTCAGCAG-3′ (S6K reverse primer, *Bam*HI site underlined). The PCR products (generated from 35 PCR cycles using Invitrogen AccuPrime Taq polymerase, 10 µM primers, annealing

temperature of 60°C for 30 secs, and an extension temperature of 68°C for 40 secs) were cleaved by *Hind*III and *Bam*HI restriction endonucleases (5 units of enzyme, NEBuffer 2, BSA 100 µg/ml, 16 h incubation at 37°C), and each resulting fragment was subcloned into the corresponding sites in the TNT T7 luciferase control plasmid (Promega Australia) located between the T7 promoter and the luciferase reporter gene. The resulting plasmids (SAMDC_WT and S6K_WT) created were checked for errors by DNA sequencing (at AGRF, using BDT labelling). Sequencing results were analysed by Vector NTI contig express program (Lu and Moriyama 2004).

4.2.2 Site-directed mutagenesis

Plasmids containing a point mutation (A to T) in the uORF start codon were created (SAMDC_MUT and S6K_MUT) using the QuickChange II XL Site-Directed Mutagenesis Kit (Stratagene), following the manufacturer's instructions. The SAMDC_WT and S6K_WT plasmids were used as the DNA template. The mutagenic primer sequences used were 5′-TTTAAGTGGATGTACTATTGGAGTCAAAGGGTGGC-3′ (SAMDC_MUT forward primer, point mutation underlined), 5′-GCCACCCTTTGACTCCAATAGTACATCCACTTAAA-3′ (SAMDC_MUT reverse primer), 5'-ATAGACAAGTTGGTCTCTTGAATTCAGGAAGCACATTGG-3' (S6K_MUT forward primer, point mutation underlined), and 5′-CCAATGTGCTTCCTGAATTCAAGAGACCAACTTGTCTAT-3′ (S6K_MUT reverse primer). Mutations were confirmed by DNA sequencing of the uORF region at AGRF using BDT labelling. Sequence results were analysed by Vector NTI contig express program (Lu and Moriyama 2004).

4.2.3 *In vitro* transcription and translation

All plasmids were linearised with *EcoICRI*, which cuts 58 nt downstream of the luciferase reporter gene. *EcoICRI* is an isoschizomer of *SacI*, and therefore recognises the same restriction site. However, *EcoICRI* creates blunt ends instead of 3′ overhangs, which prevents the *in vitro* synthesis of aberrant transcripts (Schenborn and Mierendorf 1985). Cleavage of plasmids was confirmed by DNA gel electrophoresis. Synthesis of capped RNA transcripts were produced from linearised plasmid DNA (the template) using the T7 RiboMAX™ Large Scale RNA Production System (Promega) following the manufacturer's instructions. RNA transcripts were quantified by using the NanoDrop ND-1000 Spectrophotometer, and equal amounts of RNA (1 µg) were analysed by gel electrophoresis (1.5% agarose gel containing 2.2 M formaldehyde, and 0.2M MOPS (3-[N-Morpholino]propanesulfonic acid) buffer) to check for correct size.

Each RNA transcript was translated in the rabbit reticulocyte lysate *in vitro* translation system (Promega) using 1 µg of RNA in 25 µl reactions. Reactions were incubated at 30°C for 1 h and kept on wet ice prior to luciferase activity measurements (Section 4.2.4).

4.2.4 Luciferase assays

Luciferase activity was measured using a luminometer (OPTIMA, BMG LabTech) programmed to perform a 2-second measurement delay followed by a 10-second measurement read. Rabbit reticulocyte lysate (2.5 µl) and 50 µl of the Luciferase Assay Reagent (Promega) was used for each read. Luciferase activity was expressed as the number of light units detected during the 10-second measurement read, and the relative luciferase activity was calculated by dividing the luciferase activity for each plasmid by the activity of the positive control plasmid (TNT T7 luciferase control plasmid). Error bars for the relative

luciferase activity was calculated based on the standard error of mean (SEM) from three independent replicates.

4.2.5 RNA folding

RNA secondary structures were predicted using the mFOLD program (Zuker 2003). Default settings were used to fold sequences: linear RNA folded at 37°C. This folding temperature is consistent with the incubating temperature used to synthesise the *in vitro* RNAs. A sequence length of 60 nt surrounding either *SAMDC* or *S6K* uORF start codon was used for RNA folding. Longer sequences were not used for the following reasons: shorter sequences are more likely to represent how RNA molecules are folded as they are transcribed, a small 15-50 nt region representing one or two stem-loops is considered more reliable for energy-based structure predictions than longer sequences that can initiate too many foldings (Pavesi et al. 2004), and the ΔG calculated from global structures becomes less representative of local structures. All mFOLD structure predictions were examined, as the most thermodynamically stable folding (lowest ΔG) is not always biologically correct.

4.3 RESULTS

4.3.1 The 5′-UTR of rice *SAMDC* and *S6K* repress *in vitro* luciferase translation

Both the rice *SAMDC* and *S6K* genes contain highly conserved uORFs (>80% amino acid sequence similarity) that are positioned centrally in their long 5′-UTRs (Figure 4.1). To determine whether the 5′-UTRs of the rice *SAMDC* and *S6K* genes were capable of repressing downstream translation of a reporter gene, the luciferase (LUC) coding sequence was fused to each 5′-UTR (Figure 4.2). These chimeric 5′-UTR-luciferase genes (SAMDC_WT and S6K_WT) were linearised and then transcribed and translated *in vitro* (Section 4.2.3).

To confirm that the SAMDC_WT and S6K_WT constructs as well as their mutant derivatives were linearised so as to prevent *in vitro* run-on transcription, gel electrophoresis was performed using vector DNA digested with *EcoICRI* (Figure 4.3). The ~5 kb bands were within the expected size (SAMDC_WT/SAMDC_MUT 4.89 kb, S6K_WT/S6K_MUT 4.71 kb). Multiple DNA bands were detected from un-linearised plasmids, as expected, which showed the dominant forms of plasmid DNA (supercoiled and relaxed).

The *in vitro* transcription/translation of linearised vector DNA showed that the *SAMDC* and *S6K* wild-type constructs repressed translation of the luciferase reporter gene (Figure 4.4). The presence of the *SAMDC* wild-type 5′-UTR repressed translation by 8.3 fold when compared to the positive control (no 5′-UTR). The *S6K* wild-type 5′-UTR repressed translation by 25 fold (Figure 4.4).

4.3.2 The uORF in 5′-UTR of rice *SAMDC* and *S6K* partly repress translation

To determine if the conserved uORFs are responsible for the translational repression of the luciferase reporter gene, site-directed mutagenesis was used to eliminate the uORFs by mutating the uORF start codon from ATG to TTG (Section 4.2.2). The single base change in the *SAMDC* uORF start codon derepressed translation of the luciferase reporter gene by 2.9 fold when compared to the *SAMDC* 5′-UTR wild-type (Figure 4.4). This result is consistent with the previous reported derepression levels of transgenic tobacco plants (Hanfrey et al. 2002). A similar result was obtained when the *S6K* long uORF was eliminated, where a 3 fold derepression was observed when compared to the *S6K* 5′-UTR wild-type. In both cases where the uORFs were eliminated the translation of luciferase was not completely restored to the levels of the positive control (no 5′-UTR, Figure 4.4).

4.3.3 Rice *SAMDC* and *S6K* uORFs may affect transcription

To confirm that the uORF repression on downstream luciferase translation was due to translational control and not transcriptional events, denaturing gel electrophoresis was performed using equal amounts of the *in vitro* RNA products (Section 4.2.3). The *SAMDC* 5′-UTR wild-type transcript containing the uORF produced the correct size band (~2.3 kb), and an additional shorter band (~1.9 kb) (Figure 4.5A, lanes D and H). However, when the *SAMDC* 5′-UTR wild-type transcript was compared with its mutant derivative (no uORF), the additional shorter band was not detected (Figure 4.5A, lanes E and I). Similarly, the *S6K* wild-type construct containing a long uORF produced two bands. One band had an expected size of ~2.1 kb, and an additional shorter band (~1.9 kb), which was not present in the *S6K* mutant (Figure 4.5A, Lanes F, G, J, and K). Interestingly, the additional shorter band of both *SAMDC* and *S6K* wild-types was similar in size (~1.9 kb; Figure 4.5A, Lanes D, F, H, and J). The additional shorter band (~1.9 kb) is unlikely the result of the full-length transcript being cleaved as the corresponding cleavage products (~500 and 370 bp, respectively) were not detected. Transcript degradation is also unlikely as no obvious smear was seen between the RNA bands. Therefore, the presence of aberrant transcription products most likely occurred from either premature transcription termination or alternative transcription initiation (Figure 4.5B).

Premature transcription termination is difficult to explain for the following reasons. An enzyme (*EcoICRI*) that produces blunt ends was used to linearise the expression vectors to avoid extraneous transcript production that is inherent with the T7 polymerase transcription system (Schenborn and Mierendorf 1985). Also, both the 3′-end of *SAMDC* and *S6K* transcripts contain the same luciferase reporter gene. Therefore if premature termination was occurring it would likely take place at the same position, and produce transcripts that are shorter than their full-length counter-parts by equal lengths. This was not observed as the difference in size between the full length and the shorter transcript of *SAMDC* and *S6K* are ~400 bp and 200 bp respectively.

The alternative explanation for the additional shorter band is that transcription in the *SAMDC* and *S6K* wild-type constructs is occurring further downstream of the T7 promoter and at the same relative position, as both the *SAMDC* and *S6K* shorter bands are similar in size (~1.9 kb) as the positive control (no 5′-UTR). This suggests that the 5′-UTRs of these constructs are not being transcribed, and that transcription initiation is occurring in the vicinity of the luciferase gene. Interestingly, mutating the uORF start codon (ATG to TTG) of both *SAMDC* and *S6K* constructs produces only full-length transcripts, suggesting that changes in potential secondary structures around the uORF start codon is effecting transcription in the 5′-UTR.

4.3.4 Predicted stem-loop structures may affect transcription but not translation

Both the *SAMDC* and *S6K* wild-type constructs produced an additional shorter band that was not observed in their corresponding mutants (uORF start codon mutated to TTG). This suggests that structural differences may have been created in the region where the uORF start codon was mutated. To investigate the possibility of secondary structure changes between the wild-type and mutant transcripts and its potential affect on transcription, a short sequence (60 nt) around each uORF was folded to predict their secondary structures (Section 4.2.5, Chapter 2). The *SAMDC* wild-type transcript is predicted to have two stem-loop motifs (ΔG -5.17) on either side of the uORF start codon (Figure 4.6A). This arrangement of stem-loops around the uORF start codon is more spaced (open configuration) in comparison to the *SAMDC* mutant (Figure 4.6B). In the *SAMDC* mutant, the mutated start codon (ATG to TTG) becomes part of the 5′ stem-loop, and as such brings the pair of stem-loops (ΔG -5.89) closer together in a closed configuration (Figure 4.6B). The ΔG of these two configurations is similar (WT ΔG -5.17 to MUT ΔG -5.89). A similar open and closed configuration was also observed with the *S6K* wild-type and mutant transcripts, respectively (Figure 4.6C and D). However, the ΔG of *S6K* wild-

type open configuration (-14) and mutant closed configuration (-5.89) were different (WT ΔG -14.00 to MUT ΔG -5.89).

The predicted secondary structure changes introduced by mutating the uORF start codon (ATG to TTG) in both *SAMDC* and *S6K* wild-type transcripts could explain the occurrence of the additional shorter transcript. One hypothesis is that mutating the uORF start codon introduces RNA conformational changes that prevent the alternative transcription initiation near the luciferase gene, and therefore full-length transcripts are produced more efficiently. Indeed, RNA folding of the region surrounding the uORF start codon of both wild-type *SAMDC* and *S6K* reveals that there are predicted secondary structures (stem-loops on either side of uORF start codon) that presumably interact with the start of the luciferase gene to form an alternative transcription site, thus promoting T7 promoter-independent extension (Zaher and Unrau 2004). However, it is unclear how these predicted secondary structures interact with the T7 polymerase. Cloning and sequencing the additional shorter band will help confirm that the shorter transcripts are truncated at the 5′-end, and may also provide further insights to uORF action at the transcriptional level. The findings reported here for two independent examples show that it is important to consider the potential roles of uORFs in all levels of gene regulation. Indeed, Hu *et al*. (2005) showed evidence for the *SAMDC* 5′-UTR in transcriptional control, but did not test if the *SAMDC* uORF is solely responsible.

Whether or not both *SAMDC* and *S6K* uORFs are implicated in transcriptional control is largely irrelevant to their effects at the translational level. The presence of aberrant transcription products, either shortened at the 5′ or 3′-end, could affect translation by compromising the 5′-UTR or inactivating the luciferase gene via C-end truncation. However, these aberrant transcription products are estimated to be only a minor contributor (<10 %) to the decrease in luciferase translation, indicating that the two uORFs are important in translational control. For example, the results in Figure 4.4 show that lanes D

and H have a 2 fold reduction in *SAMDC* full-length transcript compared to the positive control, but there is over an 8 fold reduction in luciferase translation efficiency. Similarly, the amount of *S6K* full-length transcript is similar to the positive control (Figure 4.4, lanes F and J), but there is ~25 fold reduction in luciferase translation efficiency.

4.4 DISCUSSION

4.4.1 *SAMDC* and *S6K* uORFs controls translation

This study shows, for the first time, the potential role of the *S6K* long uORF in translational control. Similar to *SAMDC* small uORF, the *S6K* long uORF shows evidence of down-regulating translation of a luciferase reporter gene (Figure 4.4). The effect of the *S6K* uORF on translational control has not been studied before in plants. However, one study reported that the *S6K* uORF is also conserved in human and mouse, and is predicted to control translation through reinitiation (Kochetov et al. 2008). The results presented here provide evidence that the *S6K* uORF is important in down-regulating translation. Eliminating the *S6K* uORF by mutating (ATG to TTG) its start codon, results in a 3 fold derepression in luciferase relative translation (Figure 4.4). This level of depression is similar to that of *SAMDC* uORF, indicating uORFs are involved in finely regulating the expression of their downstream major open reading frame. Indeed, regulation by uORFs has been suggested as a means of reducing translation and preventing over-expression of certain genes (e.g., growth factors, transcription factors or proto-oncogenes) that could be toxic to a cell (Mignone et al. 2002).

The *S6K* uORF is approximately 19% longer than *SAMDC* uORF and down-regulated luciferase translation more effectively (2.6 times), indicating that the length of a uORF could have an effect on translation. It is possible that the longer uORF of *S6K* encodes a more stable peptide, and therefore allows

prolonged inhibition of translation. Indeed, when the stability of the *SAMDC* and *S6K* uORF peptides were analysed using the ProtParam program (Wilkins et al. 1999), the *S6K* uORF was predicted to be more stable (stability index of 50.81) than the *SAMDC* uORF (stability index of 73.82). A stability index of less than 40 classifies a protein as being stable based on the occurrence of certain dipeptides that are significantly different in the unstable proteins compared with those in the stable ones, suggesting that the *SAMDC* and *S6K* uORF peptides are likely to be unstable. It has been suggested that uORF encoded peptides that have a short half-life prevents them from acting on translation of other genes (Chang et al. 2000).

Unlike the *S6K* uORF, the role of the *SAMDC* uORFs in translational control has been well studied (Hanfrey et al. 2002; Hu et al. 2005). Removing all or part of the *SAMDC* small uORF causes a 3-5 fold derepression in GUS relative translation *in planta.* This study also supports this view as the results in Figure 4.4 show that mutating the *SAMDC* uORF start codon (ATG to TTG) results in a 2.9 fold derepression in luciferase relative translation *in vitro*. This demonstrates that *in vitro* studies of uORFs are an ideal system for their primary characterisation.

4.4.2 Translational control by the *S6K* uORF

The conserved *S6K* uORF is likely to encode a bioactive peptide as selection has occurred at the peptide level (Tran et al. 2008). The *S6K* peptide is not expected to be synthesised in large amounts as the context demarcating the start codon of *S6K* uORF is weak (ctcATGa) compared to the main ORF (aagATGg). A strong sequence context is denoted by a guanine in the +4 position and a purine in the -3 position as mutations in these positions results in the greatest reduction in translation efficiency (Kozak 2005). Experimental data (Figure 4.4) confirms that the *S6K* uORF does not require an optimal sequence context to have an effect on downstream translation. In fact, an optimal sequence context alone would provide little control over the translation

of the main ORF, as initiation would predominately start at the uORF resulting in constant down regulation of the *S6K* gene, which is likely to be harmful to the plant. Rather, a weak sequence context allows for leaky scanning of the inhibitory uORF, and thus preferential initiation at the downstream main ORF. Therefore, leaky scanning of the *S6K* uORF is the likely mechanism of modulating the translation of the *S6K* gene, a mechanism that is consistent with a weak sequence context associated with a majority of reported uORFs.

4.4.3 Other features in *SAMDC* and *S6K* 5′-UTR may affect translation

The *SAMDC* and *S6K* uORFs are not solely responsible for the decrease in luciferase translational efficiency, as their mutants (no uORFs) did not restore the translational efficiency to the levels of the positive control (no 5′-UTR, Figure 4.4). One explanation for the low derepression levels of translation in the *SAMDC* and *S6K* mutants is that there are other features in the *SAMDC* and *S6K* 5′-UTR that may contribute to the translational repression of the luciferase reporter gene. This is likely given the length of the 5′-UTR of *SAMDC* (583 nt) and *S6K* (409 nt) are much longer than average (260 nt) (Kikuchi et al. 2003), and as such provide greater potential for harboring other translational repressive motifs (e.g., secondary structures). Moreover, the length of the 5′-UTR has been shown to influence ribosomal loading, with a long (>175 nt) 5′-UTR of Arabidopsis mRNA showing significantly lower than average ribosomal loading levels (Kawaguchi and Bailey-Serres 2005). The 5′-UTR length contribution to the decrease in translational efficiency could be determined in future experiments by using a positive control containing a modified wild-type 5′-UTR that substitutes the uORF with a random sequence.

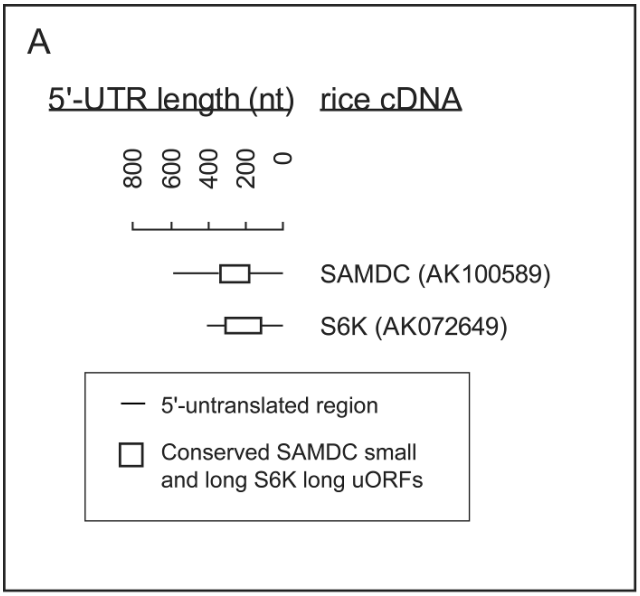
4.4.4 Evidence of uORF translational control in non-plant species

A similar LUC reporter system combined with site-directed mutagenesis of uORF start codons was used to experimentally evaluate the translational

efficiency of newly identified genes containing uORFs in *Saccharomyces cerevisiae* (Zhang and Dietrich 2005). As in this study, the uORF start codons were mutated from ATG to TTG, and the luciferase activity and mRNA levels of these mutant constructs were measured to determine the relative luciferase translational efficiency for comparison with their corresponding wild-types. They reported that some genes were translationally enhanced (~2 fold) after the uORF start codon were mutated, indicating that the uORFs have a regulatory function at the level of translation. In this study, the level of translation was slightly higher (~3 fold) when the *SAMDC* and *S6K* uORF start codons were mutated. Other studies on the *SAMDC* uORF have shown up to 5-fold translational enhancement by truncating the uORF *in vivo* (Hanfrey et al. 2002). Moreover, a 20 to 50 fold translational enhancement has been reported for the human *mdm2* oncogene as a result of downstream alternative promoter usage, which produces a shorter transcript (S-mdm2) that is devoid of two small uORFs (Landers et al. 1997). The S-mdm2 transcript was detected by an RNase protection assay that uses a probe to distinguish between the two transcript classes. These results from several independent reports on uORFs in different species show that uORFs can down-regulate the expression of genes to various extents. This is particularly evident for genes where overexpression of potent proteins is harmful.

4.5. CONCLUSION

In conclusion, this study provides evidence, for the first time, that the rice *S6K* uORF is involved in translational control, and confirms the role of the rice *SAMDC* uORF in controlling translation. It further supports the comparative approach used to identify conserved uORFs (Chapter 3). Further work is necessary to fully characterise the *SAMDC* small and *S6K* long uORFs. This may include cloning and sequencing of the additional smaller RNA transcript (Figure 4.5), and further mutagenesis work to disrupt secondary structures believed to be involved in affecting transcription.



B

```
SAMDC     AK100589
rice      MESKGGKKKSSSSRSLMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
wheat     MESKGGKK-SSSSSSLMYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
barley    MESKGGKK-SSSSSSLMYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
maize     MESKGGKK-SSSSRSMMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
sorghum   MESKGGKK-SSSSRSMMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
          ******** **** *:******************.*** ************

S6K       AK072649
rice      MNSGSTLASRVSCSSGNNGIFPPCLLPGFLGFPVVFVNFHLFKAIVRSPASSVVLGICGGRLP
wheat     MYSGRTVASRVSCSSGNNGSFPPCLCPGFLGFPVVSVNFHLFKAIVRSPASSVVLGICGRSLP
barley    MYSGRTVASRVSCSSGNNGSFPPCLRPGFLGFPVVSVNFHLFKAIVRSPASSVVLGICGRSLP
maize     MNSGRTVASSVSCPAGNSGSFPPCLLPGFLGFPVVFVNFHLFKAIVRSPASSVVLGICGGSLP
sorghum   MNSGRTVASNVSCSAGNSGSFPPCLLPGFLGFPVIFVNFHLFKAIVRSPASSVILGICGGSLP
          * ** *:** ***.:**.* ***** ********: **************************
```

Figure 4.1 Position and alignment of *SAMDC* short and *S6K* long uORFs. (A) Position of *SAMDC* and *S6K* uORFs. (B) Alignment of uORF peptides of *SAMDC* and *S6K*. The annotation line indicates fully conserved (*), strongly conserved (:) or weakly conserved amino acid residues (.).

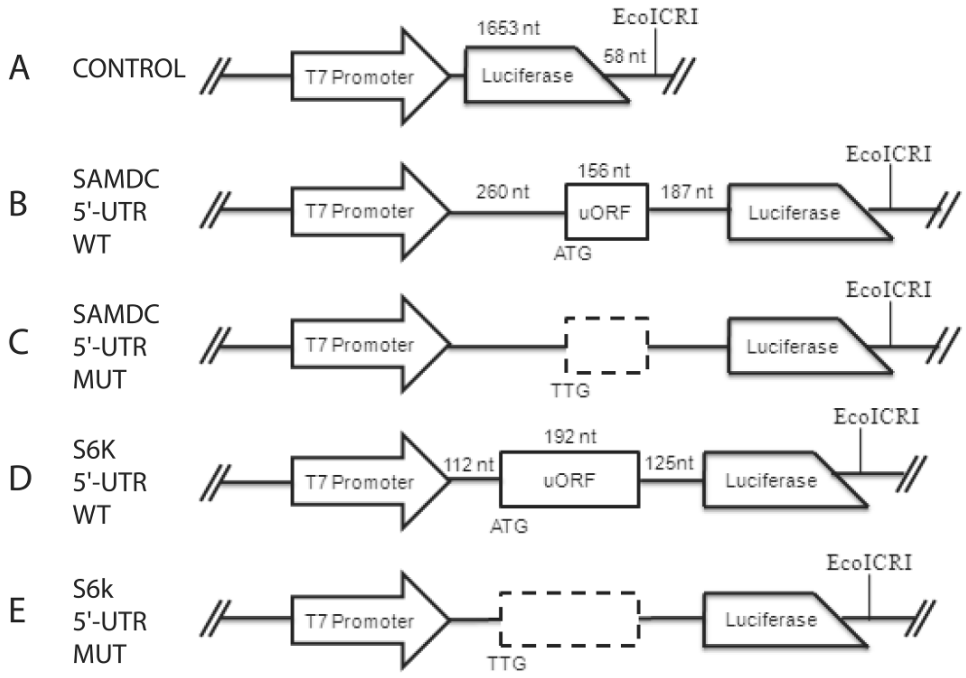


Figure 4.2 Expression vectors for *in vitro* assays. The 5′-UTR of *SAMDC* and *S6K* was placed between the T7 promoter and the firefly luciferase reporter gene of the TNT control vector (A, Promega) to construct SAMDC_WT (B) and S6K_WT (D) expression vectors. The SAMDC_MUT (C) and S6K_MUT (E) expression vectors contain a point mutation in the uORF start codon (ATG to TTG), which eliminates the uORF (indicated as a dashed box). All expression vectors were linearised with *EcoICRI* restriction endonuclease to produce blunt-ended vectors to prevent *in vitro* run-on RNA transcripts under the control of the T7 promoter (according to Promega instructions).

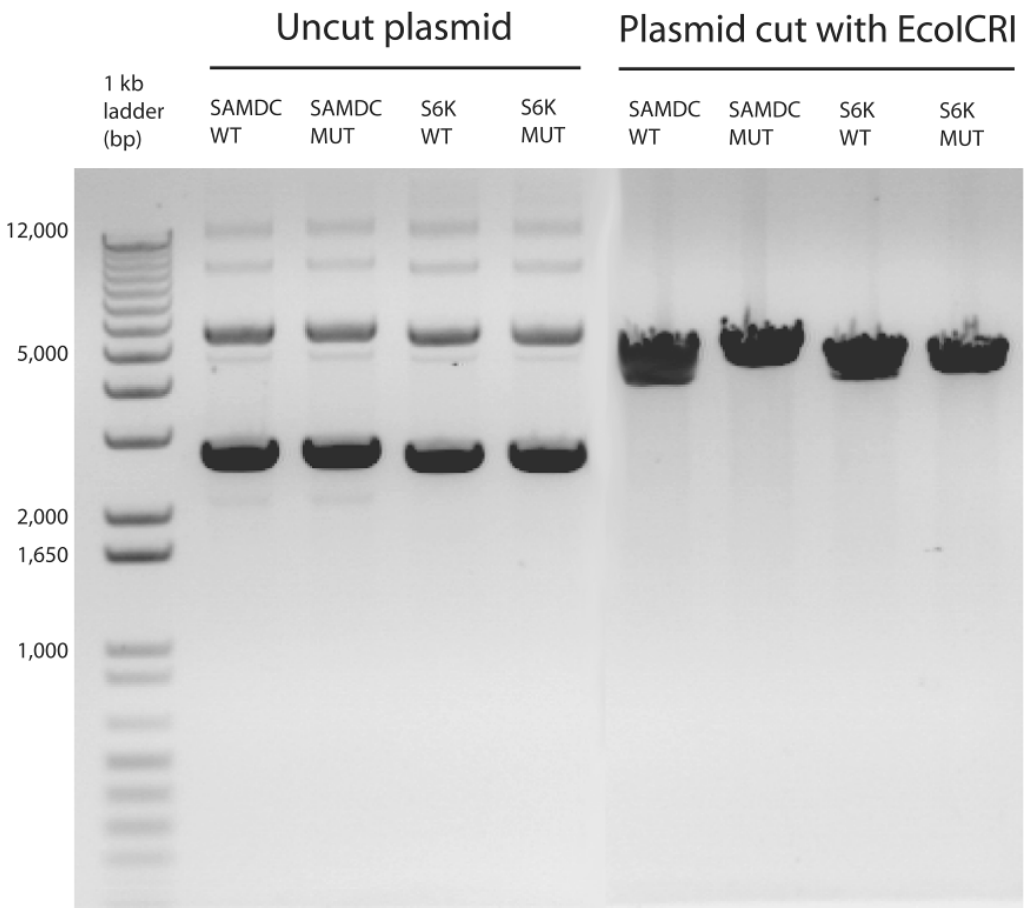


Figure 4.3 Electrophoresis of plasmid DNA (1 μg) of each expression vector. Left lane represents the 1 kb plus DNA marker (Invitrogen). The remaining lanes show uncut and cut expression vectors. The multiple bands shown for the uncut expression vectors represent the different forms of plasmid DNA, such as the supercoiled and relaxed circular DNA.

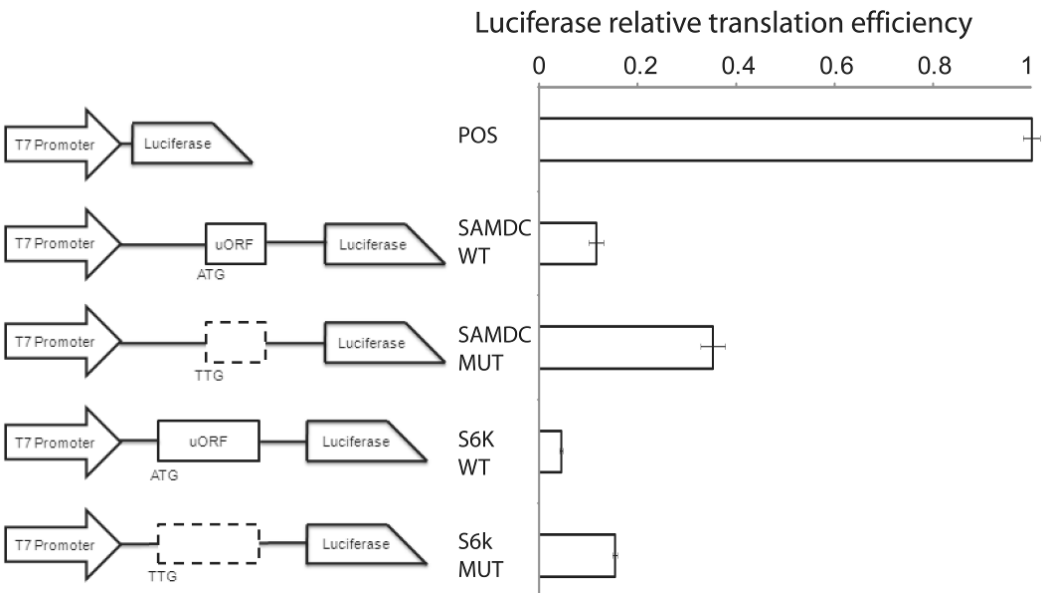
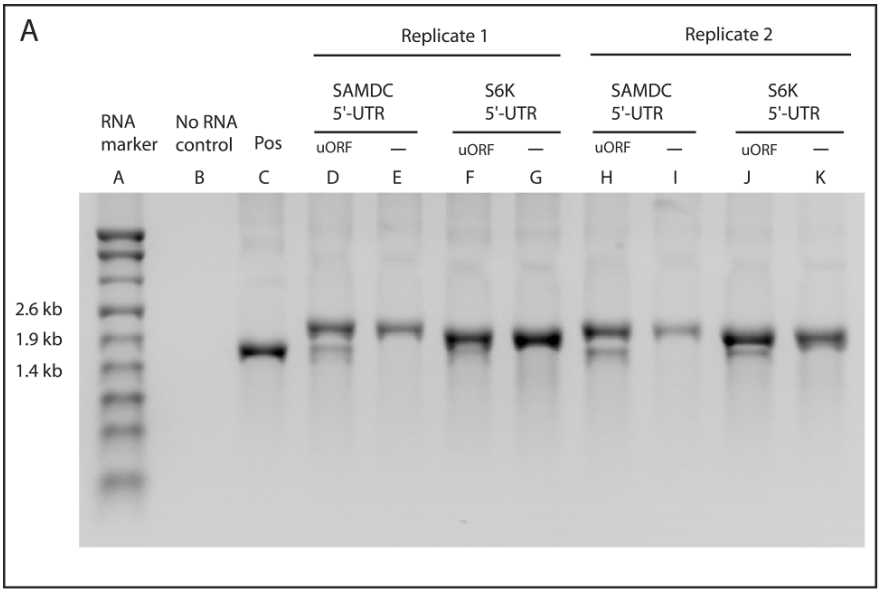
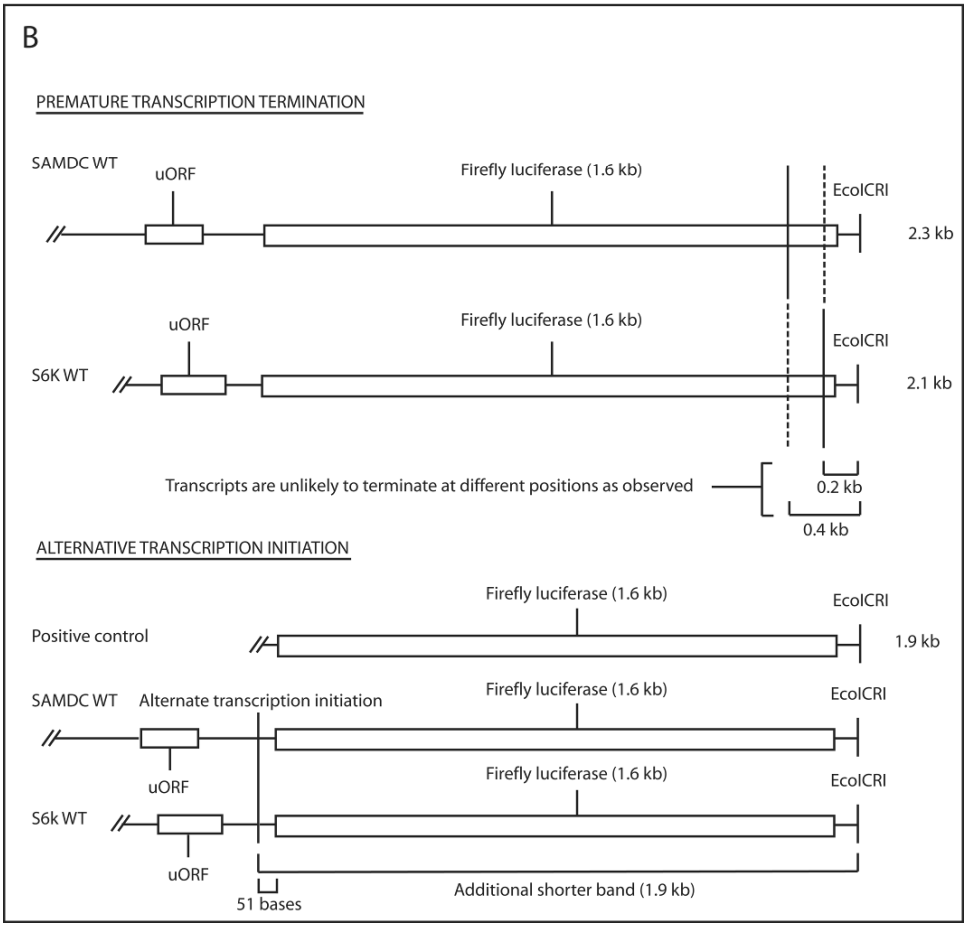


Figure 4.4 *SAMDC* and *S6K* uORFs and their effects on luciferase expression levels. RNA transcripts synthesized *in vitro* were translated in rabbit reticulocyte lysate. The luciferase translational efficiency was normalized to the positive control (POS, no 5′-UTR). Error bars for the relative luciferase activity was calculated based on the standard error of mean (SEM) from three independent replicates.

Figure 4.5 Denaturing gel electrophoresis of *in vitro SAMDC* and *S6K* RNA products. A) Electrophoresis of *in vitro* synthesized RNA through an agarose gel containing formaldehyde. Lane (A) represents the 1 kb RNA marker (Promega). Lane (B) shows the negative control where no RNA was added. The positive control (Pos), lane (C), shows the T7-luciferase RNA transcript positioned at ~1.8 kb. Expression vectors containing the 5′-UTR of SAMDC_WT (two independent clones, D and H) and S6K_WT (two independent clones, F and J) produced a major RNA transcript positioned at the expected full length ~2.3 and 2.1 kb, respectively. Mutant expression vectors that contain a point mutation in the uORF start codon (ATG to TTG) are shown in lanes (E, G, I, and K). B) Graphical representation of transcripts produced by either premature termination or alternative transcription initiation.


A
Replicate 1
Replicate 2
SAMDC 5'-UTR
S6K 5'-UTR
SAMDC 5'-UTR
S6K 5'-UTR
uORF
—
uORF
—
uORF
—
uORF
—
RNA marker
No RNA control
Pos
A
B
C
D
E
F
G
H
I
J
K
2.6 kb
1.9 kb
1.4 kb


B
PREMATURE TRANSCRIPTION TERMINATION
SAMDC WT
uORF
Firefly luciferase (1.6 kb)
EcoICRI
2.3 kb
S6K WT
uORF
Firefly luciferase (1.6 kb)
EcoICRI
2.1 kb
Transcripts are unlikely to terminate at different positions as observed
0.2 kb
0.4 kb
ALTERNATIVE TRANSCRIPTION INITIATION
Positive control
Firefly luciferase (1.6 kb)
EcoICRI
1.9 kb
SAMDC WT
Alternate transcription initiation
Firefly luciferase (1.6 kb)
EcoICRI
uORF
S6k WT
Firefly luciferase (1.6 kb)
EcoICRI
uORF
51 bases
Additional shorter band (1.9 kb)

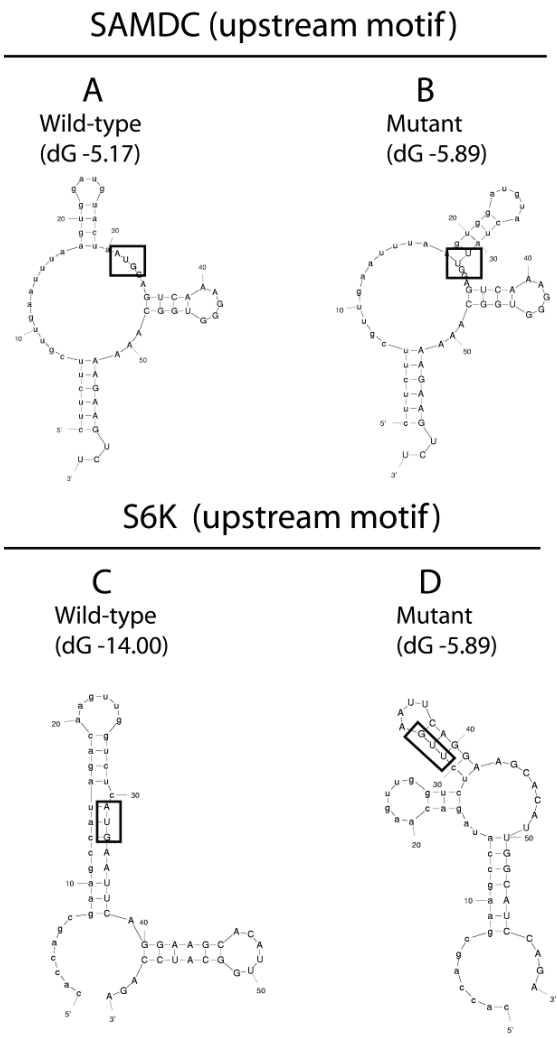


Figure 4.6 RNA folding (mFOLD) of region surrounding the start codon of the wild-type *SAMDC* and *S6K* uORFs and their mutant derivatives. The open secondary structure configuration of *SAMDC* (A) and *S6K* (C) wild-type transcripts. The closed secondary structure configuration of *SAMDC* (B) and *S6K* (D) mutant transcripts. The uORF sequence is in uppercase, and the uORF start codon is boxed.

CHAPTER 5

CONCLUSION AND FUTURE GOALS

CHAPTER 5 CONCLUSION AND FUTURE GOALS

5.1 Conserved 5′-UTR secondary structures in the cereal transcriptome

Local secondary structures in the untranslated regions of animal and plant mRNAs can regulate *in cis* translation of the main open reading frame through several different mechanisms that include reducing ribosomal loading (Kawaguchi and Bailey-Serres 2005), ribosomal attenuation and stalling (Gaba et al. 2001; Luo and Sachs 1996), and destabilising mRNAs (Ruiz-Echevarria and Peltz 2000; Vilela et al. 1999). Prior to this study, it was unclear how prevalent conserved secondary structures are, such as the stem-loop motif, in any species. It is now possible to conduct transcriptome-wide searches for conserved secondary structures due to recent advances in RNA motif prediction algorithms.

Chapter 2 describes the development of a pipeline to identify conserved 5′-UTR stem-loop motifs in four important agronomic cereal transcriptomes (i.e., the monocots rice, wheat, barley, and maize). The pipeline first used a modified reciprocal best hit method to identify putative orthologous sequences from a collection of rice full-length cDNAs and other cereal EST contigs. Finally, long 5′-UTRs (200 to 1200 nt) from putative orthologous were analysed by a comparative R-nomics program called RNAProfile to find conserved stem-loop motifs. This study allowed the identification of new genes from the cereal transcriptome that are potentially regulated by stem-loop motifs. The findings from this study concluded that 1) conserved 5′-UTR stem-loops motifs in long 5′-UTRs (200 to 1200 nt) are rare (~8%) in the cereal transcriptome, 2) are not under or over-representated in any gene class (a group of genes with related functions e.g., cell metabolism), and 3) appear to have a biological role based on higher structure (from 85% to 98%) than sequence conservation in at least three out of four cereal species.

The next major challenge is to determine if the conserved secondary structures have a functional role. To prioritise this work, a representative rice 5′-UTR stem-loop motif from each available gene class should be characterised first (approximately 20 candidates). One common procedure to measure the relative translational efficiency of a gene under the control of the 5′-UTR, is to compare native and mutagenised 5′-UTR (disrupted stem-loop base-pairing) using *in vitro* transcription and translation assays (Arnaud et al. 2007; dos Santos et al. 2008; Hulzink et al. 2002; Klinkert et al. 2006; Rogers et al. 2002; Zou et al. 2003). It is also important to determine whether changes in translational efficiency by conserved secondary structures involve their interactions with RNA-binding proteins (RBPs), as some genes are regulated coordinately by stem-loops and RBPs (Bailey-Serres et al. 2009).

Another major challenge is to determine whether some conserved stem-loop motifs have ancient biological mechanisms, and therefore it would be of interest to determine the prevalence of conserved monocot 5′-UTR stem-loops motifs in more distantly related dicot plant species (i.e., Arabidopsis) and then in lower plant species such as gymnosperms and the moss *Physcomitrella patens*. Currently, at least one example of an ancient processing mechanism has been reported for a stem-loop structure that is conserved in Arabidopsis and moss (Bologna et al. 2009), but more examples are needed. Also, it is not known what different classes of genes are being regulated by conserved 5′-UTR stem-loops motifs in more distantly related plant species and in lower plant species, and whether or not there are structural and functional differences in their stem and loop regions.

To find conserved stem-loop motifs in other species the pipeline (Chapter 2, Section 2.3.7) could be reused on more distantly related plant transcriptomes without requiring major changes. One major change envisaged is the optimisation of the rbh method for the better detection of orthologues in highly diverged species (Fulton et al. 2006; Moreno-Hagelsieb and Latimer 2008). For example, choosing appropriate BLAST options like the soft-

filtering of low information segments and using the Smith-Waterman algorithm for sequence alignment significantly improves rbh orthologue detection (Moreno-Hagelsieb and Latimer 2008). Also, an alternative algorithm called reciprocal smallest distance (Wall and Deluca 2007; Wall et al. 2003) claims to improve upon the rbh method of detecting orthologues, and as such should be tested concurrently with the rbh method. Improved detection by reciprocal smallest distance is based on the use of global sequence alignment and maximum likelihood estimation of evolutionary distances, and therefore is less likely than rbh to be misled by the presence of a paralogue with high sequence identity. Other alternative methods that apparently improve on orthologue detection should also be considered (Lee et al. 2002; Mooers 2009). The most updated datasets for plant species should be used to optimise the rbh method and to compare with the reciprocal smallest distance and other methods, so that putative orthologues identified can be used in downstream analyses (including 5′-UTR stem-loop motif discovery).

The study on conserved 5′-UTR stem-loop motifs in cereal plants provides a platform for research into local secondary structures in plant 3′-UTRs. Translation of the main open reading frame was originally thought to be less sensitive to the effect of a 3′-UTR stem-loop (Niepel et al. 1999), however more research on 3′-UTR secondary structures in both animal and plant mRNAs are required as some 3′-UTR stem-loops can down regulate translation. For example, Niepel et al. (1999) showed that a stable 3′-UTR stem-loops located near the stop codon of the main open reading frame can cause the actively translating 80S ribosome to prematurely terminate. In addition, the presence of stable 3′-UTR secondary structures at the miRNA-binding site can prevent miRNAs, especially in animals, from binding to their target and therefore stop miRNA-directed translational repression (Du and Zamore 2005). Given the role of the 3′-UTR in post-translational regulation, an open-access database, known as UTRome.org, has been developed which contains 3′-UTR structures of the worm *Caenorhabditis elegans* (Mangone et al. 2008), a model used in animal research. The pipeline that was developed for

5′-UTR stem-loop research can easily be modified to find conserved 3′-UTR stem-loops in the plant transcriptome.

5.2 Conserved uORFs in transcriptomes of higher plants

Upstream open reading frames (uORFs) can down-regulate the translation of the main open reading frame (mORF) through two broad mechanisms: ribosomal stalling and reducing reinitiation efficiency. In distantly related plants, such as rice and Arabidopsis, it was found that conserved uORFs are rare in these transcriptomes with approximately 100 loci (Hayden and Jorgensen 2007). Prior to this study, it was unclear how prevalent conserved uORFs are in closely related plants. In Chapter 3, a homology-based approach was described to identify conserved uORFs in five cereals (monocots) that could potentially regulate translation. The approach used the previously described modified reciprocal best hit method (Chapter 2, Section 2.2.2) to identify putative orthologous sequences that were then analysed by a comparative R-nomics program called uORFSCAN to find conserved uORFs. This study identified new genes that may be controlled at the level of translation by conserved uORFs. Major conclusions from this study are that the identified uORFs 1) are highly conserved (50% median amino acid sequence similarity), 2) are rare in cereal transcriptomes (<150 loci contain them), 3) are generally short (less than 100 nt), 4) are position independent in their 5′-UTRs, and 5) their start codon context and the usage of rare codons do not appear to be important for translation.

The study on conserved uORFs in the transcriptome of higher plants identified conserved uORFs that could produce bonafide bioactive peptides. To date, bioactive uORF peptides have mainly been implicated in regulating translation of the mORF and in mRNA stability, and research in both these mechanisms are ongoing (reviewed in Lovett and Rogers (1996), Meijer and Thomas (2002), and Vilela and McCarthy (2003)). Alternative roles for bioactive uORF peptides were suggested (Hayden and Jorgensen 2007; Iacono

et al. 2005) where they can act *in trans* on other biological processes, but this remains to be determined. One approach to test for putative function of bioactive uORF peptides is to screen a tilling population (Weil 2009) and/or T-DNA insertion lines (Alonso et al. 2003) for mutations in the uORF but not in the mORF, and analyse potential phenotype chances associated with these specific mutations. Another approach is to establish uORF knock-outs in species where it is possible to do homologous recombination (e.g., yeast *Saccharomyces cerevisiae*, moss *Physcomitrella patens*, and fruitfly *Drosophila melanogastor*) (Aylon and Kupiec 2004; Bi and Rong 2003; Puchta 2002). The phenotypical changes associated with the specific targeting and replacement of uORF containing genes with orthologues genes devoid of uORFs can then be analysed.

One limitation of this study is that the homology based approach used to identify conserved uORFs may be more targeted towards the identification of uORFs encoding bioactive peptides rather than uORFs that reduce reinitiation efficiency (sequence-independent uORFs). On the other hand, a homology based approach is more likely to identify bioactive peptides based on selection pressure that conserves uORFs at the amino acid level. Furthermore, one cannot rule out the possibility that the identified uORFs could also affect reinitiation. For these reasons, a homology based approach was chosen for this study. It may be possible to predict some sequence-independent uORFs that are likely to be functional based on their specific sequence organisation (i.e., uORFs that overlap the main ORF) (Kochetov et al. 2008). However, for uORFs that are positioned entirely upstream of the mORF, it will be difficult to predict their role in translational control without further biological testing.

Currently, only one example of an mRNA containing a conserved uORFs has been reported in species that diverged over long evolutionary distances (i.e., from single-celled green algae to moss to Arabidopsis), and that is the conserved uORFs in S-adenosylmethionine decarboxylase (*SAMDC*)

mRNA. Therefore, it would be of interest to determine the prevalence of conserved uORFs in more distantly related species such as *D. melanogaste*r, *Mus musculus*, and *Homo sapiens*. mRNAs containing uORFs that are conserved in both the plant and animal transcriptomes would suggest that the uORFs have an ancient origin and provide strong support for a functional role in translational control. As mentioned in Section 5.1, it is possible to re-use the uORFSCAN pipeline to find conserved uORFs in both the plant and animal transcriptomes without requiring major changes.

5.3 Primary characterisation of conserved motifs

In Chapter 4, an *in vitro* proof-of-concept study was described on two candidate rice uORFs identified *in silico* as conserved in the transcriptome of higher plants (Tran et al. 2008). The S-adenosylmethionine decarboxylase (*SAMDC*) small uORF (156 nt) was chosen as a positive control for functional analysis because its orthologous uORF is known to repress translation in other plant species (Hao et al. 2005; Lee et al. 1997; Mad Arif et al. 1994; Tassoni et al. 2007). The other candidate chosen for functional analysis was the long uORF (192 nt) of S6 ribosomal kinase (*S6K*). The *S6K* uORF is one of the longest conserved uORFs identified in the cereal transcriptome, and as such is likely to encode a bioactive peptide (Tran et al. 2008). The approach used to characterise the uORFs was based on site-directed mutagenesis of uORF start codons combined with a rapid quantitative *in vitro* transcription and translation system to evaluate translational efficiency. The findings from the functional analyses confirmed the role of the *SAMDC* uORF in controlling translation, and provided evidence, for the first time, that the *S6K* uORF can down-regulate translation of a main ORF, and therefore implicates the uORF as a *cis*-acting translational control element. Moreover, these findings underscore the validity of the comparative approach used to identify conserved uORFs as potential regulators of translation.

Two major hypotheses were drawn from the primary characterisation of both rice *SAMDC* and *S6K* uORFs. Firstly, the two uORFs appear to be implicated in transcriptional control that is, however, largely independent to their effects at the translational level (Chapter 4, Section 4.3.4). The exact mechanism of the uORF-mediated transcriptional control is unclear, but appears to involve alternative transcription initiation that is dependent on local secondary structure formation around the uORF start codon. This hypothesis is partially supported by Hu *et al*. (2005), who showed evidence for the *SAMDC* 5′-UTR in transcriptional control. More research is required to fully elucidate the mechanism of uORF-mediated transcriptional control, which may involve cloning and sequencing of the aberrant transcription product produced in both *SAMDC* and *S6K in vitro* transcription assays. In addition, it is important to determine if the aberrant *in vitro* transcription product also exists *in vivo* so as to reflect a consistent mechanism of uORF-mediated transcriptional control for both genes. The other major hypothesis is that the *S6K* uORF is likely to encode a bioactive peptide, like the known small uORF of *SAMDC*, as selection has occurred at the amino acid level. It remains to be determined if the encoded S6K peptide can affect the translational apparatus (e.g., ribosomal stalling), but it is believed that leaky scanning is the mechanism that determines whether or not the *S6K* uORF is translated. Leaky scanning is consistent with a weak sequence context demarcating the uORF start codon (*S6K* uORF sequence context is weak (ctcATGa) as denoted by the absence of a purine at -3 position and guanine at +4 position).

More in-depth analysis is required to fully characterise the rice *SAMDC* and *S6K* uORFs. For example, RNA footprinting assays (Daou-Chabo and Condon 2009; Esakova et al. 2008; Savochkina et al. 2008) can be used to detect the presence and the position of stalled ribosomes caused by binding to the proposed uORF peptide. The principle behind RNA footprinting is that proteins and/or ribosomal complexes bound to RNA will protect it from enzymatic cleavage. Therefore, by comparing the cleavage pattern of mRNAs containing the presence and absence of a uORF, it is possible to detect the

position of stalled ribosomes associated with bound uORF peptides. Another useful assay that is often performed concurrently with RNAse footprinting are RNA gel shift assays (Ebhardt and Unrau 2009; Zhu et al. 1996). RNA gel shift assays can be used to detect RNA-protein interactions based on electrophoretic separation differences between unbound RNA and protein bound RNA. Finally, RNA pull-down assays using antibodies (Russo et al. 2008) combined with mass spectrometry (Kvaratskhelia and Grice 2008) would allow for higher resolution structural analysis of protein-RNA interactions.

5.4 Future needs

ESTs represent approximately 80% of all of the public nucleotide database entries, and are therefore an important resource for bioinformatics (Benson et al. 2009). However, EST datasets are often of poor-quality and incomplete, especially EST datasets generated from single-pass reads (Picoult-Newberg et al. 1999). There is a need for better quality (full-length and error-free) and increased quantity (transcriptome-wide) of ESTs for the best possible comparative sequence analysis in orthologue detection and in RNA motif prediction. The recent availability of full-length coding sequences of the Triticeae crops wheat *(T. aestivum)* and barley (*H. vulgare*) is starting to address this need (Mochida et al. 2009). One expected advantage of using full-length mRNAs and/or completed EST datasets in orthologue detection will be the increased number of detected true orthologues, and correspondingly, the reduced detection of false positives (paralogues) (Fulton et al. 2006). The increase in detected orthologues will benefit downstream analyses such as RNA motif prediction by providing additional orthologues for comparative analysis.

RNA motif prediction has evolved from predicting structures from a single sequence to multiple related RNA sequences. Currently, the complexity of RNA motif prediction for multiple related RNA sequences can be reduced to linear time, denoted as $O(n)$, and as such is capable of finding simple

secondary structures (i.e., stem-loops) in transcriptome-wide scans using available EST sequences. A major challenge in RNA motif prediction for the future is to develop computationally efficient and reliable algorithms that are capable of finding significant (above noise) secondary structures (simple or complex) in genome-wide scans.

There are various methods for finding conserved sequence tags (CSTs) in genome comparisons including CSTminer (Castrignano et al. 2004), Genominer (Castrignano et al. 2006), and CSTgrid (Mignone et al. 2008). According to our analysis, there was no evidence of highly conserved uORFs showing significant sequence similarity at the amino acid level with coding regions based on similarity searches using 5′-UTR translations against the NCBI UniProt database (Table 3.2 and Table 3.9). Indeed, given the rarity of conserved uORFs in the plant transcriptome and their generally short length (less than 100 nt) it is unlikely that many conserved uORFs overlapping coding CSTs will be found. In the future, it is conceivable that more conserved uORFs could be found through direct genome comparisons, thus avoiding the limitations of using EST data, and as such tools for finding CSTs that could potentially overlap conserved uORFs will be more useful when that time comes.

The importance of high throughput functional assays will be paramount in the future as the number of predicted RNA motifs is expected to increase as EST datasets and RNA prediction algorithms continue to improve. High throughput assays make it easier to manage the screening of hundreds of predicted RNA motifs to identify those motifs that have a regulatory role (primary characterisation). Currently, commercial *in vitro* transcription and translation assays are used to rapidly test and compare (within hours) the translational efficiency of constructs containing either a mutagenised UTR (mutant) or endogenous UTR (wild-type) fused to a reporter gene. The rate limiting step in the biological testing of predicted RNA motifs is construct development, which can take weeks to clone and mutagenised UTRs, even

with current popular cloning (Invitrogen Gateway cloning kit) and mutagenesis (Stratagene QuickChange II XL Site-Directed Mutagenesis Kit) technologies. It may be possible to hasten (within days) the construct development step in the future as DNA synthesis technology continue to improve and be cost effective, and therefore making it possible to synthesis complete constructs (Czar et al. 2009).

A comprehensive knowledge of alternative transcript isoforms generated by alternative transcription initiation and termination and alternative splicing is required to fully elucidate mechanisms underlying post-transcriptional control. Unfortunately, using available EST data and full-length cDNAs in studying post-transcriptional control is limited as 1) it only provides a small snap-shot of the transcriptome, and 2) the combining of EST data from different sources may generate artifacts due to the loss of tissue-specific differences in transcription start and stop sites and RNA processing. The arrival of next generation sequencing technology, such as high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) (Licatalosi et al. 2008) and RNA-Seq (Wang et al. 2009), will prove useful in profiling the highly dynamic transcriptome, which should in the near future provide further insight into precise measurements of transcript levels and their isoforms, as well as RNA-protein interactions.