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

TABLE OF CONTENTS

CHAPTER 1 LITERATURE REVIEW 2

1.1 INTRODUCTION 2

1.2 EVIDENCE FOR POST-TRANSCRIPTIONAL CONTROL 3

1.2.1 Disparity between nuclear transcription rates and cytosolic mRNA levels 3

1.2.2 Cytosolic mRNA and protein levels do not always correlate 4

1.2.3 Untranslated regions important for post-transcriptional control 5

1.2.4 Control signals in untranslated regions mediate translational control 5

1.3 UPSTREAM OPEN READING FRAMES (uORFS) 6

1.3.1 Definition of uORFs 6

1.3.2 Types of uORFs 6

1.3.3 uORFs can influence mRNA stability 7

1.3.4 Start codon context of uORFs 8

1.3.5 Translation of messenger RNA containing functional uORFs 9

1.3.6 Approaches for identifying functional uORFs 10

1.3.7 Plant uORFs 11

1.3.8 Mechanisms of regulation by the plant *SAMDC* and Lc uORFs 12

1.3.9 Novel regulatory mechanisms by plant uORFs 14

1.4 MESSENGER RNA (mRNA) STRUCTURES 15

1.4.1 Introduction to RNA, structure, and stability 15

1.4.2 Types of secondary structures 17

1.4.3 Position and stability of secondary structures 18

1.4.4 Secondary structures in mRNAs from non-plant organisms 19

1.4.5 Secondary structures in mRNAs from plant organisms 20

1.4.6 Secondary structures in mRNAs from plastid genes 22

1.5 *IN SILICO* DISCOVERY OF RNA MOTIFS 23

1.5.1 RNA vs DNA motif discovery 23

1.5.2 RNA can exist in different states 24
1.5.3 Early RNA motif prediction algorithms 25
1.5.4 State-of-the-art RNA motif prediction 26

1.6 SUMMARY AND AIMS OF THE PRESENT STUDY 30

CHAPTER 2 RNA SECONDARY STRUCTURES 41

2.1 INTRODUCTION 41

2.2 MATERIALS AND METHODS 43
2.2.1 Sequence data 43
2.2.2 Orthologue searches 43
2.2.3 Algorithm for finding conserved methionine positions 44
2.2.4 RNAProfile algorithm 45

2.3 RESULTS 45
2.3.1 Selection of a suitable RNA motif prediction program 45
2.3.2 Creating a 16S rRNA dataset to test RNAProfile 46
2.3.3 RNAProfile test case 1: Archaea 16S rRNAs 47
2.3.4 RNAProfile test case 2: Cereal *Trx4* mRNAs 49
2.3.5 Features of the conserved *Trx4* 5'-UTR motif 50
2.3.6 Detection of orthologues for comparative analysis 51
2.3.7 Pipeline for discovering 5'-UTR stem-loops in cereals 52
2.3.8 RNAProfile identified conserved cereal 5'-UTR secondary structures 54
2.3.9 Rice genes with conserved 5'-UTR stem-loops have different functions 56

2.4 DISCUSSION 57
2.4.1 8% of cereal transcripts contain conserved stem-loops in long 5'-UTRs 57
2.4.2 Conserved cereal 5'-UTR stem-loop motifs may have a regulatory role 57
2.4.3 Chloroplast precursor transcripts may be translationally regulated 59
2.4.4 Some previously reported plant 5'-UTR stem-loops are not conserved 59
2.4.5 Two types of predicted conserved motifs in 5'-UTR of *Trx4* 60
2.4.6 Conserved methionine positions as indicators of translation initiation 61
2.4.7 Extending CLUSTALW annotation helps identify problematic sequences 63

2.5 CONCLUSION 64

CHAPTER 3 UPSTREAM OPEN READING FRAMES 87

3.1 INTRODUCTION 87

3.2 MATERIALS AND METHODS 89

3.2.1 Sequence data 89

3.2.2 Orthologue searches 90

3.2.3 Verification of main ORF 90

3.2.4 Statistical analysis of codon usage 91

3.3 RESULTS 92

3.3.1 The uORFSCAN pipeline for discovering uORFs 92

3.3.2 Conserved uORFs appear to be rare 93

3.3.3 Position and occupation of uORFs in 5'-UTRs 95

3.3.4 Length distribution of uORFs 96

3.3.5 Sequence conservation of uORFs 96

3.3.6 Start condon context and codon usage of uORFs 97

3.4 DISCUSSION 98

3.4.1 Conserved uORFs appear to be rare 98

3.4.2 Cereal uORFs conserved in Arabidopsis 99

3.4.3 Better quality assembled EST data is needed 101

3.4.4 Sequence dependent and independent uORFs 102

3.5 CONCLUSION 104

CHAPTER 4 FUNCTIONAL TESTING OF UORFS 128

4.1 INTRODUCTION 128

4.2 MATERIALS AND METHODS 129

4.2.1 Plasmid construction 129

4.2.2 Site-directed mutagenesis 130

4.2.3 *In vitro* transcription and translation 131

4.2.4 Luciferase assays 131
4.2.5 RNA folding 132

4.3 RESULTS 132
4.3.1 The 5'-UTR of rice *SAMDC* and *S6K* repress *in vitro* luciferase translation132
4.3.2 The uORF in 5'-UTR of rice *SAMDC* and *S6K* partly repress translation 133
4.3.3 Rice *SAMDC* and *S6K* uORFs may affect transcription 134
4.3.4 Predicted stem-loop structures may affect transcription but not translation135

4.4 DISCUSSION 137
4.4.1 *SAMDC* and *S6K* uORFs controls translation 137
4.4.2 Translational control by the *S6K* uORF 138
4.4.3 Other features in *SAMDC* and *S6K* 5′-UTR may affect translation 139
4.4.4 Evidence of uORF translational control in non-plant species 139

4.5. CONCLUSION 140

CHAPTER 5 CONCLUSION AND FUTURE GOALS 149

5.1 Conserved 5'-UTR secondary structures in the cereal transcriptome 149

5.2 Conserved uORFs in transcriptomes of higher plants 152

5.3 Primary characterisation of conserved motifs 154

5.4 Future needs 156

APPENDIX 160

1.1. uORFSCAN source code 160

1.2. Database schema used to manage the data generated by uORFSCAN 180

1.3. Overview of CMPSCAN 183

BIBLIOGRAPHY 184

ABSTRACT

Understanding the mechanisms that regulate gene expression is an important goal in bioinformatic research. There are two major levels of gene regulation: transcriptional and post-transcriptional control. Much attention has been directed to transcriptional control, but it is now clear that the untranslated regions (UTRs) of messenger RNA (mRNA) also play an important role in post-transcriptional control of gene expression. Two important control signals found in 5′-UTRs of both animal and plant mRNAs are stem-loop motifs and upstream open reading frames (uORFs).

One strategy for identifying functional uORFs in plants is to use a comparative approach (Crowe et al. 2006; Hayden and Jorgensen 2007; Pavesi et al. 2007). There are extensive EST datasets for five important cereal crops (rice, wheat, barley, maize, and sorghum). Rice is the best characterised of these cereals with a sequenced genome (Yu et al. 2002) and a cDNA database containing 32,000 clones that are enriched for 5′ full-length sequences (Kikuchi et al. 2003). In this research, comparative R-nomics was used to identify conserved stem-loop motifs and uORFs in cereals using publicly available assembled EST data.

To determine the prevalence of 5′-UTR stem-loop structures in plants a bioinformatics pipeline was developed to predict secondary structures. The pipeline used a program called RNAProfile to predict stem-loops that are conserved in both sequence and structure. The findings from this study concluded that conserved 5′-UTR stem-loops in long 5′-UTRs (200 to 1200 nt) are rare (~8%) in the cereal transcriptome, the genes themselves that contain conserved 5′-UTR stem-loop motifs are spread across different functions, and appear to have a biological role based on higher structure than sequence conservation in at least three out of four cereal species.

Another control signal that is involved in post-transcriptional control is the uORF. A recent study in distantly related plants, such as rice and

Arabidopsis, found that uORFs are rare in these transcriptomes (Hayden and Jorgensen 2007), but it is unclear how prevalent uORFs are in closely related plants. To address this question, the bioinformatics pipeline was modified to use a program called uORFSCAN to find conserved uORFs in five cereals that could potentially regulate translation. Major conclusions from this study are that the identified uORFs are highly conserved (50% median amino acid sequence similarity), are rare in cereal transcriptomes (<150 loci contain them), are generally short (less than 100 nt), position independent in their 5′-UTRs, and their start codon context and the usage of rare codons do not appear to be important for translation.

Two candidate uORFs were selected for mutational analyses, and a quantitative *in vitro* transcription and translation system was used to determine if they function in translational control. The rice *SAMDC* small and *S6K* long uORFs were shown to be capable of down-regulating translation of a luciferase reporter gene. This study has provided evidence, for the first time, that the *S6K* uORF is involved in controlling translation. In conclusion, this study has identified new genes that may be controlled at the level of translation by stem-loop motifs and conserved uORFs.

DECLARATION

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Michael Khanh Tran and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as shown below) resides with the copyright holder/s of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Chapter 3 contains material from the following publication:

Tran, M.K., C.J. Schultz, and U. Baumann. 2008. Conserved upstream open reading frames in higher plants. *BMC Genomics* **9:** 361-378.

Signature:

Date:

ACKNOWLEDGEMENTS

I would like to acknowledge with appreciation the following organisations and individuals who have been instrumental in the successful completion of my PhD.

PhD Scholarship
Australian Centre for Plant Functional Genomics (ACPFG)
Website address: www.acpfg.com.au

South Australian Partnership for Advanced Computing (SAPAC)
Website address: www.sapac.edu.au

Supervisors
Dr. Carolyn J. Schultz
Dr. Ute Baumann

Vector map construction
Dr. Andrew Jacobs

Statistical support
Dr. Andreas Schreiber

Friendship and helpful discussions
Dr. Rodney Davies

LIST OF TABLES

CHAPTER 1

Table 1.1 Comparison of programs that predict common structures in unaligned sequences.

CHAPTER 2

Table 2.1 Comparison of predicted motifs from 24 16S rRNA sequences.

Table 2.2 Effects of adding contaminating sequences.

Table 2.3 Effect of varying the region length and the number of iterations.

Table 2.4 Extended CLUSTALW annotation code.

Table 2.5 Rice translation initiation site predictions for rbh orthologue groups (4/4).

Table 2.6 Rice translation initiation site predictions for rbh orthologue groups (3/4).

Table 2.7 Conserved secondary structures predicted by RNAProfile in 4/4 cereals.

Table 2.8 Conserved secondary structures predicted by RNAProfile in 3/4 cereals.

Table 2.9 Function of rice clones with predicted conserved (4/4) 5′-UTR stem-loop motifs.

Table 2.10 Function of rice clones with predicted conserved (3/4) 5′-UTR stem-loop motifs.

CHAPTER 3

Table 3.1 The uORFs predicted by uORFSCAN in 5/5 orthologues of the 5/5 orthologue dataset.

Table 3.2 Criteria for verifying rice uORFs (uORF 5/5 result set).

Table 3.3 The uORFs predicted by uORFSCAN in 4/5 orthologues of the 5/5 orthologue dataset.

Table 3.4 The uORFs predicted by uORFSCAN in 3/5 orthologues of the 5/5 orthologue dataset.

Table 3.5 The uORFs predicted by uORFSCAN in 4/4 orthologues of the 4/5 orthologue dataset.

Table 3.6 The uORFs predicted by uORFSCAN in 3/4 orthologues of the 4/5 orthologue dataset.

Table 3.7 The uORFs predicted by uORFSCAN in 3/3 orthologues of the 3/5 orthologue dataset.

Table 3.8 Rice uORFs predicted by uORFSCAN that are conserved in Arabidopsis.

Table 3.9 Criteria for verifying rice uORFs that are conserved in Arabidopsis.

Table 3.10 Comparison of conserved cereal uORFs and their main ORF start context.

Table 3.11 ClustalW alignment of uORFs identified by uORFSCAN in 5/5 cereals and in Arabidopsis.

LIST OF FIGURES

CHAPTER 1

Figure 1.1 Six steps at which eukaryotic gene expression can be controlled.

Figure 1.2 Correlation between protein and mRNA levels for 106 genes (Δ) in yeast growing at log phase with glucose as a carbon source.

Figure 1.3 Translational efficiency of GUS in leaves of transgenic tobacco plants.

Figure 1.4 Common RNA secondary structure motifs

Figure 1.5 Two levels of translation repression mediated by the maize *Lc* (1) leader secondary structure and (2) uORF.

Figure 1.6 RNA states and folding energy profile.

Figure 1.7 Highest scoring motif occurrences output by RNAProfile on the IRE dataset with their respective energy and fitness value.

CHAPTER 2

Figure 2.1 CLUSTALW alignment of cereal thioredoxin-h4 5′-UTRs.

Figure 2.2 Highest scoring motifs predicted by RNAProfile on the *Trx4* 5′-UTR dataset with their respective energy and fitness value.

Figure 2.3 Analysis of detected rbh orthologues.

Figure 2.4 Overview of the stem-loop discovery pipeline.

Figure 2.5 5′-UTR distribution of KOME full-length cDNA clones.

Figure 2.6 CLUSTALW alignment of the nucleotide translation of the 5′ region of rice Ankyrin-2 (AK103103) and its cereal rbh orthologues TC207106 (wheat), TC148319 (barley), and TC270900 (maize).

CHAPTER 3

Figure 3.1 Overview of the uORFSCAN pipeline.

Figure 3.2 The position of uORFs conserved in four other cereals and in Arabidopsis within 5′-UTRs of rice cDNAs.

Figure 3.3 Frequency distribution of the length (nt) of rice uORFs conserved in four other cereals and in Arabidopsis.

Figure 3.4 The pattern of nucleotide sequence conservation calculated for the decanucleotide surrounding the uORF AUG triplet using WebLogo.

Figure 3.5 Relative frequencies of codons showing significant deviation (*) in codon usage between rice uORFs and rice main coding regions.

CHAPTER 4

Figure 4.1 Position and alignment of *SAMDC* short and *S6K* long uORFs.

Figure 4.2 Expression vectors for *in vitro* assays.

Figure 4.3 Electrophoresis of plasmid DNA (1 μg) of each expression vector.

Figure 4.4 *SAMDC* and *S6K* uORFs and their effects on luciferase expression levels.

Figure 4.5 Denaturing gel electrophoresis of *in vitro SAMDC* and *S6K* RNA products.

Figure 4.6 RNA folding (mFOLD) of region surrounding the start codon of the wild-type *SAMDC* and *S6K* uORFs and their mutant derivatives.

ABBREVIATIONS

aa	amino acid
ADH1	alcohol dehydrogenase-1
AGRF	Australian Genome Research Facility
ARF	auxin response factor
Avg.	average
BDT	Big Dye Terminator
BLAST	Basic Local Align Search Tool
bp	base pair
bZIP	basic region leucine zipper
CBL	calcineurin B-like
CDS	coding sequence
CNS	conserved non-coding sequence
CPA1	carbamoyl phosphate synthetase
CST	conserved sequence tag
DFCI	Dana Farber Cancer Institute
eIF4F	eukaryotic initiation factor 4F
EST	expressed sequence tag
ETT	*ETTIN*
FL	full-length
GO	gene ontology
GPRM	genetic programming for RNA motifs
GUS	β-glucuronidase
Hb	*Hordeum bulbosum*
Hv	*Hordeum vulgare*
INDEL	insertions and deletions
IRE	iron responsive element
IRES	internal ribosome entry site
IRP	iron responsive protein
KOME	Knowledge-Based Oryza Molecular Biological Encyclopedia

Lp	*Lolium perenne*
LUC	luciferase
MOPS	3-[N-Morpholino]propanesulfonic acid
mORF	main open reading frame
MP	*MONOPTEROS*
NELF	negative elongation factor
nt	nucleotide
NMD	nonsense-mediated decay
PC	*Phalaris coerulescens*
POS	positive control
rbcs	ribulose biphosphate small subunit
rbh	reciprocal best hit
RBP	RNA-binding protein
RNA	ribonucleic acid
RNApromo	RNA prediction of motifs
RNase P	ribonuclease P
S6K	S6 ribosomal kinase
SAMDC	S-adenosylmethionine decarboxylase
SAPAC	South Australian Partnership for Advanced Computing
Sc	*Secale cereale*
SCFG	stochastic context-free grammars
SECIS	selenocysteine insertion sequence
SEM	standard error of mean
SLASH	Stem-Loop Align SearcH
Taes	*Triticum aestivum*
TC	tentative contig
TIGR	The Institute for Genomic Research
TIS	translation initiation site
Trx4	*thioredoxin4*
UniProtKB	UniProt Knowledgebase
uATG	upstream start codon
uORF	upstream open reading frame

URL	universal resource locator
UTR	untranslated region