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 3

UPSTREAM OPEN READING FRAMES

CHAPTER 3 UPSTREAM OPEN READING FRAMES

This chapter contains work published in “Tran, M.K., C.J. Schultz, and U. Baumann. 2008. Conserved upstream open reading frames in higher plants. *BMC Genomics* **9:** 361”.

3.1 INTRODUCTION

RNA-omics, or more simply R-nomics, is the large-scale study of RNA structure and function (Clote 2005). One of the major challenges faced by R-nomics is to understand the regulatory mechanisms of complex signals found in the untranslated regions (UTRs) of messenger RNAs. In particular, the control signals found in the 5′-UTR of some eukaryotic mRNAs play a crucial role in translational control that can result in rapid changes to the proteome during stages of mammalian development (Le Roch et al. 2004), and in response to plant abiotic stress (Floris et al. 2009). These post-transcriptionally regulated mRNAs frequently encode important regulatory proteins (e.g., proto-oncogenes, growth factors, and transcription factors) (Mignone et al. 2002) that need to be strongly or precisely regulated for normal cellular activity. In other cases, control signals in the 5′-UTR provide continuous regulation of essential mRNAs by providing an alternative route for translation when cap-dependent translation is compromised (e.g., under stress conditions) (Holcik et al. 2000).

Translational control signals are often found in long 5′-UTRs (>100 nt) (Kozak 1987a) where they can contain either a single control signal (Raney et al. 2000) or multiple control signals that function independently (Wang and Wessler 2001) or in a coordinated fashion (Franceschetti et al. 2001; Jin et al. 2003; Yaman et al. 2003). One important translational control signal found in both prokaryotes and eukaryotes is the upstream open reading frame (uORF), a small open reading frame located upstream of the main coding region (Lovett and Rogers 1996). uORFs that are conserved are also known as conserved non-coding sequences (CNSs) (Freeling and Subramaniam 2009).

Two types of functional uORFs have been described and shown to have a demonstrated activity either *in vitro or in vivo* (Gaba et al. 2001): a) sequence-dependent uORFs encoding bioactive peptides that either cause ribosomal stalling during translation of the main open reading frame or have other biological roles (Crowe et al. 2006; Hayden and Bosco 2008; Hayden and Jorgensen 2007; Iacono et al. 2005), and b) sequence-independent uORFs that reduce reinitiation efficiency of the main open reading frame (Meijer and Thomas 2002; Vilela and McCarthy 2003). There is also evidence that both sequence-dependent and sequence-independent uORFs can influence mRNA stability (Ruiz-Echevarria and Peltz 2000; Vilela et al. 1999), however their underlying mechanisms remain unclear.

Identifying uORFs involved in regulation of gene expression is a difficult and time consuming process that is estimated to take up to 20 man-months to find a single functional uORF by random selection and testing of mRNAs (Selpi et al. 2006). To overcome this problem, computational methods have been developed to predict uORFs that are likely to be functional, and include methods such as artificial intelligence (Selpi et al. 2006), comparative approaches based on homology (Hayden and Bosco 2008; Hayden and Jorgensen 2007), and comparative approaches based on specific uORF sequence organisation (Kochetov et al. 2008).

The frequency of reported uORFs in plants is rare in comparison to mammalian systems. Early estimates on the number of characterised uORFs in plants were less than 100 (0.3%) (Tran et al. 2008), and most are described in four cereal transcriptomes. These characterised uORFs (<0.3%) in plants are much lower than the estimated number of genes that contain uORFs, which can vary from 11% (Pesole et al. 2000) to 60% (Hayden and Jorgensen 2007).

In this study, a comparative R-nomics approach was used to identify conserved uORF motifs in cereals and Arabidopsis. A bioinformatics pipeline called uORFSCAN was constructed that performs a comparative analysis on the important agronomic crops rice, wheat, barley, maize, and sorghum; and the well studied dicot plant Arabidopsis. To account for the variable quality of assembled EST data, the uORFSCAN pipeline used orthologous sequence clustering, iterative sequence analysis, and manual curation. The comparative approach of uORFSCAN is easily transferable to uORF identification in other species.

3.2 MATERIALS AND METHODS

3.2.1 Sequence data

KOME full-length rice cDNA sequences were obtained from ftp://cdna01.dna.affrc.go.jp/pub/data/CURRENT/INE_FULL_SEQUENCE_DB.zip. This file is dated Tuesday, 24 January 2006, and contains 32,127 full-length cDNA clones (originally 28,469). The Dana Farber Cancer Institute (DFCI) plant gene indices database (http://compbio.dfci.harvard.edu/tgi/) was used to obtain tentative contigs (TCs) from wheat (release 10.0, Jan 05, 580155 ESTs, 44954 TCs), barley (release 9.0, Sept 04, 370546 ESTs, 23176 TCs), maize (release 17.0, Nov 06, 695811 ESTs, 56687 TCs), and sorghum (release 8.0, Nov 05, 187282 ESTs, 20029 TCs). Data cleaning was performed on the DFCI dataset to select for sequences that are designated as tentative contigs (identifiers prefixed with “TC”), thereby excluding all singletons. All data files were imported and managed using Microsoft Access 2003. Also, the analysis was re-ran using the TIGR Plant Transcript Assemblies (last updated on October 17th, 2006) for wheat (840871 ESTs), barley (456410 ESTs), maize (1084701 ESTs), and sorghum (203575 ESTs) on the uORFSCAN pipeline, but did not find any additional conserved uORFs.

3.2.2 Orthologue searches

Similarity searches were performed at The South Australian Partnership for Advanced Computing (SAPAC) (http://www.sapac.edu.au/) using the method described in Chapter 2 (Section 2.2.2). In brief, the reciprocal best hit method (rbh) was adapted to improve the detection of putative orthologues in the presence of alternative splice forms that would otherwise give many false negatives. To account for alternative splice forms, the top hit and also similar hits in the reverse BLAST (percent identity to top hit: Δ -5%, similar length to top hit: +/- 20%) were examined for symmetry with the top hit in the forward BLAST. If there is symmetry between the forward and reverse BLASTs then we considered the reciprocal pair to be orthologous.

3.2.3 Verification of main ORF

The rice cDNA sequences containing conserved uORFs were used in a blastn search against NCBI Non-redundant database to identify uORFs predicted from ribosomal RNA genes, chloroplastic genes, and mitochondrial genes. These genes do not represent coding genes derived from the nuclear genome, and therefore have been removed from this study. Also, the main open reading frames, predicted by uORFSCAN were used to search (blastn) the coding sequence (CDS) annotations from TIGR rice pseudomolecules database (http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml). Alignments not starting from the beginning of the CDS were regarded as suspicious. As additional verification, the rice main open reading frame predictions were also compared with protein data from the UniProt Knowledgebase (UniProtKB) (www.ebi.uniprot.org/database/download.shtml). Translations of the rice cDNA sequences in the same frame as the predicted main open reading frame, starting from the 5′-untranslated region to the end of the main open reading frame, were used to search (blastp) against UniProtKB. Aligments not beginning from the start of the protein sequence were discarded if they also did not have TIGR CDS support.

3.2.4 Statistical analysis of codon usage

The *p-values* were calculated according to the following formulas:

The probability to observe the number of times each codon was present in the uORFs (n_{obs}) that was less than or equal to the expected (n_{av}) by chance alone is:

$$P = \sum_{n=0}^{n_{obs}} \begin{bmatrix} N \\ n \end{bmatrix} P^{n} \, (1\text{-}P)^{N\text{-}n} \qquad 1$$

The probability to observe the number of times each codon was present in the uORFs (n_{obs}) that was greater than or equal to the expected (n_{av}) by chance alone is:

$$P = \sum_{n=\, n_{obs}}^{N} \begin{bmatrix} N \\ n \end{bmatrix} P^{n} \, (1\text{-}P)^{N\text{-}n} \qquad 2$$

Where,

$n_{obs\,=}$ The observed number of times a codon was present in the uORFs.
$n_{av\,=}$ The average number of times a codon was present in the uORFs based on the frequency of this codon in the mORF and the sample size (the observed number of codons for the set of codons for an amino in the uORFs).

3.3 RESULTS

3.3.1 The uORFSCAN pipeline for discovering uORFs

The uORFSCAN pipeline used rice full-length cDNAs (Kikuchi et al. 2003) and wheat, barley, maize, and sorghum assembled EST data for comparative analysis (Figure 3.1). As in Chapter 2, the first step of the pipeline identified rice genes that had orthologues in wheat, barley, and maize but also in sorghum. The use of orthologous sequences allowed for better detection of the main coding region, and in turn defines the 5′-UTR that is necessary to identify conserved uORFs. For this purpose, a modified reciprocal best hit (rbh) method was used to find true orthologues by a process of eliminating paralogues (Bork et al. 1998; Tatusov et al. 1997), and was shown to perform much better than a standard one-directional BLAST. For example, in the one directional BLAST against the barley assembled EST database 19,655 sequences were identified, however this number was reduced to 5,115 (26%) sequences when the reciprocal best hit method was used (Figure 3.1, Step 1).

Only 1723 of the rice genes had conserved orthologues in the other four cereals (wheat, barley, maize, and sorghum), most likely because none of the assembled EST datasets contained the entire transcriptome. To account for missing or erroneous sequences, the orthologues were grouped into three datasets for 5′-UTR analysis (Figure 3.1, Step 2). The datasets included rice genes that had orthologues in four other cereals (5 out of 5 dataset), in three other cereals (4 out of 5 dataset), and in two other cereals (3 out of 5 dataset).

In Figure 3.1 (Step 3), the uORFSCAN program (Appendix) was developed to find conserved uORFs. uORFSCAN takes as input a FASTA file containing the rice cDNA sequence and its orthologues, and identifies for each of these sequences all the possible open reading frames (ORFs). In the first iteration, the longest conserved ORF was designated as the main coding region. However, the longest ORF is not always the main coding region when there are other ORFs of similar length. Therefore, a comparative approach was used to identify the main coding region (Figure 3.1, Step 3.1). This involved finding the longest ORF that was present in all orthologous sequences, and then iteratively reducing the number of orthologous sequences, one at a time, to determine if a longer conserved set of ORFs could be found, and finally terminating when there was no improvement. The longest ORF in at least three out of five cereals was considered the main coding region. In Figure 3.1 (Step 3.2), uORFSCAN attempts to align rice uORFs with similar length orthologous uORFs (+/- 5%) at the protein level using ClustalW (Thompson et al. 1994). Finally, uORFSCAN analysed each alignment file to determine the average conservation of the uORFs, and grouped the alignments based on the number of conserved orthologous uORFs found. For example, using the 4 out of 5 dataset generated the 4 out of 4 and the 3 out of 4 datasets (Figure 3.1). To improve the detection of functional uORFs, only uORFs from orthologous genes that shared sequence similarity were reported.

The final step (Figure 3.1, Step 4) was manual curation to verify the predicted rice main coding region of each gene by comparing it with the genome annotation and other protein data. This was necessary, as uORFSCAN is expected to be sensitive to inaccurate (e.g., frame-shifts) and/or incomplete sequence data. For example, rice full length sequences can be incomplete because of failure of the 5′ capping method (Kikuchi et al. 2003). If the coding region is truncated, this can result in an internal methionine selected as the start codon and therefore the derived 5′-UTR is actually coding sequence, which is often highly conserved and can lead to false positive predictions.

3.3.2 Conserved uORFs appear to be rare

The uORFSCAN pipeline identified nine cDNAs containing uORFs that were conserved in all five cereal orthologues (5/5 uORF dataset) (Table 3.1). Three of these cDNAs encoded multiple uORFs, one of the cDNAs being AdoMetDC, which has previously been reported to contain two uORFs

(Hanfrey et al. 2002). All nine cDNAs were manually curated and showed that they were all reliable based on the validation criteria used in this study (Table 3.2), which included the removal of the uORFs predicted from ribosomal rRNA genes (data not shown). The cDNAs included the multiple uORFs in S-adenosylmethionine decarboxylase cDNA (Hanfrey et al. 2002), alkaline phytoceramidase cDNA, calcineurin B-like (CBL)-interacting protein kinase cDNA; and a single conserved uORF in a cDNA encoding an oxidoreductase protein, ribosomal protein S6 kinase, trehalose-6-phosphate phosphatase, ubiquitin-fold protein, F9L1.29 protein, and an ankyrin-3 protein.

To account for variable quality in assembled EST data, instances where the uORFs (4/5, 3/5, 4/4, 3/4, and 3/3 result set) were conserved in only some cereal orthologues (5/5, 4/5, and 3/5 dataset) (Tables 3.3, 3.4, 3.5, 3.6, 3.7; Figure 3.1, Step 3.2) were also reported. In brief, the 4/5 result set contains 16 rice genes with a total of 20 conserved uORFs in orthologous cereal genes, the 3/5 result set contains 44 rice genes with a total of 79 conserved uORFs in orthologous cereal genes, the 4/4 result set contains 16 rice genes with a total of 23 conserved uORFs in orthologous genes, the 3/4 result set contains 113 rice genes with a total of 129 conserved uORFs in orthologous genes, and finally the 3/5 result set contains 65 genes with a total of 93 conserved uORFs in orthologous genes.

In order to identify sequence dependent uORFs, the search was extended for cereal uORFs that might also be conserved in the dicot Arabidopsis by using the rice cDNAs that contained conserved uORFs in at least two other cereal orthologues (5/5, 4/5, 3/5, 4/4, 3/4 and 3/3 result set) and the Arabidopsis Tair 7 cDNA dataset (Section 3.2.2). The uORFSCAN pipeline identified 13 rice cDNAs containing uORFs that were conserved in Arabidopsis (Table 3.8). Four of these cDNAs encoded multiple uORFs. Of the 13 cDNAs with uORFs, only 11 were verified as reliable based on manual curation (Table 3.9) that removed the uORFs predicted from a cDNA encoding a helicase. Manual curation of the helicase cDNA revealed that the genome and protein annotation for the coding region extended further upstream than predicted by uORFSCAN, highlighting the limitations of using assembled EST data where frame-shift errors was the likely reason for the false positive prediction. The reliable predictions included the multiple uORFs found in a cDNA encoding ww domain containing protein, trehalose-6-phosphate phosphatase, GAMYB-binding protein, and ankyrin-3. The latter three cDNAs contained a combination of uORFs that were conserved between the cereals (rice and at least two other cereals) and Arabidopsis, and uORFs conserved between rice and Arabidopsis (Table 3.8). uORFSCAN also identified seven rice cDNAs containing a single uORF that were conserved in Arabidopsis and in almost all cases (except cDNA encoding an auxilin-like protein) the cereals as well (Table 3.8). They included the uORFs found in a cDNA encoding phosphatase 2a protein, homeodomain containing protein, S-adenosylmethionine decarboxylase, auxilin-like protein, CBL-interacting protein kinase, protein kinase ATN1, and a hypothetical protein.

3.3.3 Position and occupation of uORFs in 5′-UTRs

Studies have shown that the position of an uORF within its 5′-UTR, which determines the pre-ORF and intercistronic distances, can have profound effects on its function (Vilela and McCarthy 2003; Vilela et al. 1999). The position of cereal uORFs within their 5′-UTRs were examined and no positional preference was found with the exception that they were not positioned too closely to the start of their individual 5′-UTR and coding region (Figure 3.2). For example, all of the uORFs conserved in five orthologous cereals (5/5 result set) and in Arabidopsis were at least positioned 20 nucleotides from the start of their 5′-UTR, which is thought to be the minimum number of nucleotides required for a functional uORF (Vilela and McCarthy 2003). The intercistronic distances for these uORFs were generally shorter than the pre-ORF distance (Figure 3.2). Also, seven uORFs were found to occupy greater than 20% of their individual 5′-UTR, and included the functional small AdoMetDC uORF (AK100589).

3.3.4 Length distribution of uORFs

Since earlier reports showed that plant uORFs can vary in length from 6 to 156 nucleotides (Franceschetti et al. 2001; Hanfrey et al. 2002; Locatelli et al. 2002; Lohmer et al. 1993; Wang and Wessler 1998; Wang and Wessler 2001), the length distribution of the cereal uORFs was examined. There are two peaks in the distribution that were found between 1 to 20 nucleotides, and 81 to 100 nucleotides (Figure 3.3). The uORFs found in the first peak are tiny with 9 (out of 14) uORFs having a length of nine nucleotides. Some of these tiny uORFs could be artefactual as a result of point mutations that insert an in-frame start and/or stop codon in the 5′-UTR. The uORF length distribution around the second peak (41 to 160 nt) tends to move towards a normal distribution. Seventy six percent of the uORFs in the length distribution are shorter than 100 nucleotides, and 48% are shorter than 40 nucleotides. The shortest conserved uORF found in four independent cDNAs was nine nucleotides, even though the cut-off length used by uORFSCAN to identify uORFs was six nucleotides (a start and a stop codon). One of the nine nucleotide uORFs was the 5′ tiny uORF found in the S-adenosylmethionine decarboxylase cDNA (Franceschetti et al. 2001), and three new uORFs, two found in a cDNA encoding alkaline phytoceramidase, and one in a cDNA encoding oxidoreductase, (Table 3.1). Two long conserved uORFs (>181 nucleotides) were found in cDNAs encoding protein kinases that included one uORF found in a cDNA encoding a CBL-interacting protein kinase and another uORF found in a cDNA encoding a ribosomal protein S6 kinase.

3.3.5 Sequence conservation of uORFs

The level of amino acid sequence conservation in cereal uORFs was generally high, as expected, based on the approach of reporting similar length orthologous uORFs that shared sequence similarity used in this study. For example, in the 5 out of 5 result set the median value is 50% sequence similarity. When the two main datasets were included (uORFs conserved in all five cereal orthologues and uORFs conserved between rice and Arabidopsis), the median value is 36% sequence similarity. The uORFs conserved between rice and at least two other cereal orthologues (5/5, 4/5, 3/5, 4/4, 3/4 and 3/3 result set) and Arabidopsis (median value of 36% sequence similarity) generally had a higher amino acid sequence similarity than those uORFs conserved between rice and Arabidopsis only (median value of 28% sequence similarity). Given that the uORFs from orthologous genes were selected to be within a given length interval for alignment purposes, the high amino acid sequence similarity may suggest that these uORFs have a functional role (e.g., ribosomal stalling) that is mediated by the encoded uORF peptide.

3.3.6 Start condon context and codon usage of uORFs

The presence of uORFs does not mean that they will be translated**.** The sequence context of some plant uORFs has been shown to be sub-optimal for efficient initiation (Joshi et al. 1997; Wang and Rothnagel 2004). Therefore the sequence context of the cereal uORF AUG codons was examined to see if there was any sequence conservation that may aid in their ribosome initiation. No informative positions in the uORF consensus sequence context were found (Figure 3.4) based on the observed number of positions that showed sequence conservation was not greater than expected by chance alone. However when the context of the AUGs demarcating the conserved uORFs were compared with the context of the AUG at the main ORF it was evident that the main ORF generally had a better sequence context denoted by a purine in the -3 position and a guanine in the +4 position (Table 3.10). There were some exceptions where the uORF sequence context was good as (Table 3.10, AK066145 and AK069526 uORF 1) or better (Table 3.10, AK060523 and AK060523) than the main ORF sequence context for ribosome initiation.

Recent work showed that ribosome stalling could occur at rare codons (Chumpolkulwong et al. 2006; Col et al. 2007; Fernandez et al. 2005; Meijer and Thomas 2003; Shu et al. 2006). Therefore the codon usage of the identified uORFs was examined to determine if they contained an increased number of rare codons. Results showed that the frequencies of some codons had a *p-value* less than <*0.05* in the rice uORF codon usage compared to the rice main coding region based on a significant deviation of observed from expected numbers of uORF codons (Section 3.2.4, Equations 1 and 2); however, the number of codons that had significant *p-values* were not greater than expected by chance (Figure 3.5).

3.4 DISCUSSION

3.4.1 Conserved uORFs appear to be rare

This study provides a method to identify conserved uORFs from large assembled EST datasets. A pipeline was developed that used a modified reciprocal best hit method to identify putative orthologous sequences that were then analysed by a comparative R-nomics program called uORFSCAN to find conserved uORFs. This pipeline was successful in identifying 29 rice uORFs that are conserved at the amino acid level (median value of 36% sequence similarity) in wheat, barley, maize, sorghum, and in some cases (33%) Arabidopsis.

The number of conserved uORFs that share sequence similarity in the transcriptome of cereals appears to be low. This is consistent with reports of conserved uORFs in distantly related plants (i.e., rice and Arabidopsis) (Hayden and Jorgensen 2007) and in *Drosophila melanogaster* (Hayden and Bosco 2008). One explanation is that genes controlled at the level of translation by uORFs have low levels of transcription (Hu et al. 2005) and therefore are under-represented in cDNA and assembled EST databases. Another explanation for the low numbers of conserved cereal uORFs is that the uORFs have evolved in both length and sequence such that they no longer share sequence similarity across minor taxonomic groups (i.e., within the cereals)

(Table 3.2 and 3.9, Table 3.11 for CLUSTAL W alignments). Furthermore, if the codon usage of cereal uORFs rather than the uORF-encoded peptide were a major controlling mechanism then amino acid sequence may not be conserved.

3.4.2 Cereal uORFs conserved in Arabidopsis

It has been shown that the amino acid sequence of uORFs in monocot and dicot plants can be similar (Hanfrey et al. 2002). Sequence similarity was observed at the amino acid level across the major taxonomic groups (e.g., Arabidopsis and rice) (Table 3.8). Eleven rice genes were identified that contained uORFs conserved in Arabidopsis, of which nine were also conserved in additional cereal orthologues (at least two others). For example, a rice cDNA encoding Ankyrin-3 contains a uORF that is conserved in the cereals and Arabidopsis, but it contains a nested uORF that appears to be conserved only in rice and Arabidopsis. Therefore, it is likely that after the split between the two major groups of angiosperms (monocots and dicots), the rice gene has gained an additional in-frame and internal start codon, that is not present in the other cereals, making a nested uORF that is shorter by 33 nucleotides. It would be of interest to determine if the nested uORF is functional.

Conservation of uORF sequence within the cereals might simply reflect a relatively recent ancestor, rather than conservation of function, therefore it is difficult to predict whether these uORFs are likely to be sequence dependent or sequence independent uORFs (Meijer and Thomas 2002; Vilela and McCarthy 2003). However, uORFs that are conserved across both monocots and dicots suggest that these uORFs have a role in a sequence dependent manner. Indeed, six rice uORFs (out of 15, excluding nested uORFs, Table 3.8) that were conserved in Arabidopsis had a biased amino acid composition that was rich in serine or arginine (at least 20%). It has been suggested that uORF peptides that are rich in serine could either promote or inhibit ribosomal stalling through their phosphorylation (Hayden and Jorgensen 2007; Wang and Proud 2007), while arginine rich motifs can be involved in RNA binding (Bayer et al. 2005).

Interestingly, of these six rice uORFs two (AK101100 and AK067412) are found in genes involved in phosphorylation, a function that appears to be over-represented in this dataset (Table 3.8). It is possible that the main protein of these genes could have dual functions, the primary function is as a *trans*-acting factor in an unknown signalling cascade, and a secondary function as a regulator of mORF expression whereby the mORF protein phosphorylates the serine-rich uORF peptides, resulting in a conformational change that allows the uORF peptides to bind and stall ribosomes (Gaba et al. 2001).

There are uORFs previously identified in Arabidopsis that were not identified in this study. For example, the Arabidopsis auxin response factor (ARF) genes (Nishimura et al. 2005) *ETTIN (ETT)* and *MONOPTEROS (MP)* contain uORFs and while orthologues of these genes were found in the rice, sorghum and wheat assembled EST datasets, the uORFs showed no sequence similarity (by ClustalW) and were of different lengths (data not shown). Similarly, uORFs found in Arabidopsis genes *AtMHX* and *AtNMT1* encoding encoding a tonoplast transporter (David-Assael et al. 2005) and a phosphoethanolamine N-methyltransferase (Tabuchi et al. 2006) respectively were not identified because the uORFs were not conserved in rice and at least two other cereals. Finally, the gene containing the uORF in Arabidopsis *sac51* encoding a bHLH-type transcription factor (Imai et al. 2006) could not be identified in the rice dataset as a clear orthologue could not be identified. Therefore, it will be of interest to monitor new rice full-length cDNAs and high quality sequences for cereals as they become available to see if more conserved uORFs can be found.

Recently, a pair-wise comparative approach was used to identify conserved uORFs within homology groups that also included paralogs and ohnologs (homologous genes arising by whole-genome duplication) using rice and Arabidopsis full-length cDNAs (Hayden and Jorgensen 2007). Compared to the 11 genes identified in this study Hayden and Jorgensen (Hayden and Jorgensen 2007) reported that 19 genes contained conserved uORFs between rice and Arabidopsis. Interestingly only four genes (S-adenosylmethionine decarboxylase, Trehalose-6-phosphate phosphatase, Auxilin-like protein, and Ankyrin-3) were in common highlighting the benefits of complementary search methods. The approach developed in this study used the modified reciprocal best hit method to find putative orthologues. It is likely that some of the homologue groups identified by Hayden and Jorgensen (Hayden and Jorgensen 2007) may not be true orthologues. For example, homologue group 12 identified by Hayden and Jorgensen (Hayden and Jorgensen 2007) were not reciprocal best hit pairs according to the analysis used in this study, and therefore are likely to be paralogues. The approach of this study is deliberately conservative, eliminating paralogues, to maximise the finding of all conserved uORFs independent of their length.

One possible criticism of the comparative approach used in this study is that uORFs as short as 9 nt were reported. However, there are two independent reports that showed that the tiny uORF of *SAMDC* is functional (Hanfrey et al. 2005; Hu et al. 2005), although there is controversy regarding the type of effect and conditions under which the tiny uORF of *SAMDC* exerts its effect on downstream translation. Therefore, there is insufficient data to conclude one way or the other, and as such a conservative approach was chosen. This has allowed us to find several genes (e.g., protein phosphatase 2a, a protein containing a ww domain, and GAMYB-binding protein) that were not found by Hayden and Jorgensen (2007), as only conserved uORFs greater than 16 codons were detected.

3.4.3 Better quality assembled EST data is needed

One unavoidable limitation of using incomplete assembled EST data for orthology determination is that orthologues could be falsely assigned in situations where sequences have multiple protein domains. This will increase the number of putative orthologues identified prior to the prediction of uORFs, which is not necessarily harmful as these predictions are manually curated.

However, to minimise this problem, a sequence coverage cutoff of at least 70% of any of the protein sequences in the alignment was used (Section 3.2.2). Also, orthologues were grouped into several datasets representing the number of orthologues that could be found for each gene. For example, the datasets included rice genes that had orthologues in four other cereals (5 out of 5 dataset), in three other cereals (4 out of 5 dataset), and in two other cereals (3 out of 5 dataset). This grouping of orthologues will also help minimise the effects of missing, incomplete, or erroneous assembled EST data.

There are reports of conserved uORFs in monocots and dicots that share high sequence similarity that were not found by the uORFSCAN pipeline, due to either lack of sequence conservation or due to limitations in the assembled ESTs currently available. For example, the uORF found in the basic region leucine zipper (bZIP)-type transcription factor *AtB2/AtbZIP11* was found to be conserved in rice and barley (Wiese et al. 2004), but not in the other cereals included in this study because the sequences are not represented in the other datasets. Current limitations include incomplete data (i.e., not all sequences are represented) and poor quality sequence data, leading to frame-shifts and incorrect prediction of uORFs. Therefore, it is possible to obtain higher numbers of conserved uORFs if the cluster size was relaxed to two out of five, but this approach would reduce the power of comparative R-nomics, and would require significant manual curation.

3.4.4 Sequence dependent and independent uORFs

The cereal uORFs identified here are likely to encode bioactive peptides as selection has occurred to conserve peptide sequence. Those cereal uORFs that showed sequence conservation at the amino acid level with Arabidopsis are likely to be classified as sequence-dependent, as the encoded uORF peptide has remained conserved across the angiosperms, suggesting the peptide is directly involved in translational control (Franceschetti et al. 2001) or has some other biological activity (Crowe et al. 2006; Hayden and Bosco 2008; Hayden and

Jorgensen 2007; Iacono et al. 2005). Some identified uORFs were conserved only within the cereals, indicating a relative recent origin or selective loss of the uORFs in Arabidopsis. The possibility that some conserved cereal uORFs could also act in a sequence-independent manner cannot be ruled out, as a recent paper reported a conserved uORF in human and mouse ribosomal protein S6 kinase genes (the same finding in this study in cereals, Table 3.1), and suggested that the uORF translational control of the main ORF was through reinitiation (Kochetov et al. 2008). Experiments are needed to confirm the biological activity of the uORF in ribosomal protein S6 kinase gene.

The sequence context surrounding a uORF (ignoring secondary structure) does not appear to play a major role in its recognition and initiation of translation of an uORF based on our analysis. It is possible that a sub-optimal uORF sequence context (compared to optimal Kozak consensus (Joshi et al. 1997) sequence for the main coding region) would allow for leaky scanning (Smith et al. 2005; Wang and Rothnagel 2004) of the uORF, and preferential initiation at the downstream main coding region. An optimal uORF sequence context would provide rigid control in the translational regulation of the main coding region, as initiation would predominantly start at the uORF resulting in reduced availability of initiation factors, such as eIF2, for re-initiation at the downstream main open reading frame.

Sequence-independent uORFs allow for low-level translation of the downstream main coding region (Child et al. 1999). Low-level translation is possible, as sequence-independent uORFs do not cause ribosomal stalling as seen in sequence-dependent uORFs. The regulatory mechanism of the sequence-independent uORF involves other factors (uORF recognition, length, stop codon environment, and the downstream intercistronic sequence) that influence reinitiation efficiency (Meijer and Thomas 2002; Vilela and McCarthy 2003), and more recently leaky scanning (Wang and Rothnagel 2004), to regulate downstream translation. The codon usage of conserved uORFs was analysed and no preferential usage of rare codons was found in the uORFs. Therefore, it is unlikely that the uORF codon usage in the examples found could contribute to low-level translation as seen for certain rare codons in *Xenopus laevis* (Meijer and Thomas 2003) and *Eschericia coli* (Chumpolkulwong et al. 2006) that can reduce translation.

3.5 CONCLUSION

This study showed that the uORFSCAN pipeline is a useful tool for identifying conserved uORFs in closely related species. This pipeline has allowed us to identify 29 conserved uORFs in the cereal transcriptome. More conserved uORFs will likely be identified once the cDNA and assembled EST datasets become more comprehensive. These conserved rice uORFs will be useful for future functional analyses that should provide some perspective into downstream translational regulation by uORFs.

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

Rice		Wheat		Barley		Maize		Sorghum		Avg. A.A. similarity (%)	Putative function[b]
Identifier	5′-UTR	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]		
AK106095	131_9_17	TC265929	113_9_16	TC148181	67_9_16	TC288369	131_9_17	TC102998	149_9_17	100	Oxidoreductase
AK103391 [c]	205_75_74	TC269775	251_75_62	TC134190	204_75_62	TC294011	215_75_75	TC103599	106_75_378	88	Trehalose-6-phosphate phosphatase
AK100589 [d,e]	240_9_334	TC264559	201_9_317	TC130707	228_9_318	TC292591	286_9_320	TC91317	260_9_329	50	S-adenosylmethionine decarboxylase (AdoMetDC)
	248_156_179		209_150_168		236_150_169		294_153_168		268_153_177	90	
	296_108_179		254_105_168		281_105_169		336_111_168		310_111_177	92	
AK073303	67_9_142	TC237149	75_9_113	TC132556	81_9_139	TC305609	127_9_69	TC102988	222_9_69	50	Alkaline phytoceramidase
	135_9_74		75_9_113		81_9_139		127_9_69		222_9_69	50	
AK072868 [e]	249_27_248	TC247418	258_27_266	TC139536	298_27_272	TC306591	444_27_564	TC102544	331_27_265	11	CBL-interacting protein kinase
	259_195_70		268_198_85		308_198_91		260_192_583		341_195_87	29	
	269_39_216		278_39_234		318_39_240		768_39_228		351_39_233	8	
	338_90_96		347_93_111		387_93_117		576_93_366		420_90_113	10	
	392_36_96		404_36_111		444_36_117		633_36_366		474_36_113	8	
AK072649	100_192_117	TC236348	79_192_117	TC133316	76_192_93	TC305793	180_192_116	TC93140	168_192_116	78	Ribosomal protein S6 kinase
AK066145	178_12_58	TC266262	149_12_73	TC134484	154_12_231	TC286452	224_12_70	TC94546	187_12_69	33	Ubiquitin-fold protein
AK064792	276_15_187	TC267323	254_15_188	TC132983	253_15_-9	TC306152	263_15_170	TC107743	230_15_150	87	F9L1.29 protein
AK060523	173_123_185	TC235416	201_126_157	TC148319	211_120_163	TC305149	255_129_195	TC103609	240_129_212	58	Ankyrin-3

[a] Pre-ORF distance_uORF length_intercistronic distance.

[b] Functional annotation based on “The UniProt Knowledgebase (UniProt)” database.

[c] One of several genes (identifiers) that are in multiple tables because different conserved uORFs were identified in the different datasets.

[d] Previously reported upstream open reading frames (see Chapter 1, Section 1.3.7).

[e] Contain one or more nested uORFs.

Ribosomal rRNA genes have been removed.

See Table 3.2 for criteria for verifying rice uORFs in 5 out of 5.

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

Accession	Full-Length cDNA[a]	Upstream & In-frame stop codon	Agreement with genome annotation[b]	Alignment of uORFSCAN identified main proteins with UniProt proteins [c]						uORF valid
				UniProt protein length (AA)	Align length (AA)	Identities (%)	Expect	Annotation	GO classication (Arabidopsis thaliana)	
AK106095	Yes	Yes	Yes	392	392	100	2.2e-217	Oxidoreductase	[go:19538] protein metabolism [go:16706] oxidoreductase activity	Yes
AK103391	Yes	Yes	Yes	371	371	100	3.4e-194	Trehalose-6-phosphate phosphatase	[go:5992] trehalose biosynthesis [go:9507] chloroplast [go:4805] trehalose-phosphatase activity	Yes
AK100589	Yes	Yes	Yes	398	398	100	1.1e-215	AdoMetDC	[go:6596] polyamine biosynthesis [go:5694] chromosome	Yes
AK073303	Yes	Yes	Yes	257	257	100	1.6e-141	Acyl-CoA independent ceramide synthase	[go:6672] ceramide metabolism [go:16020] membrane [go:3824] catalytic activity [go:16811] hydrolase activity	Yes
AK072868	Yes	Yes	Yes	443	443	100	3.6e-238	Uncharacterized protein (probable CBL-interacting serine/threonine-protein kinase 15)	[go:6468] protein phosphorylation [go:7165] signal transduction [go:5524] ATP binding [go:4672] protein kinase activity	Yes
AK072649	Yes	Yes	Yes	480	488	76	9.6e-199	Ribosomal protein S6 Kinase	[go:45946] positive translation [go:6468] protein phosphorylation [go:9507] chloroplast [go:16301] kinase activity	Yes
AK066145	No	Yes	Yes	119	119	100	1.3e-59	Membrane-anchored ubiquitin-fold protein	[go:6464] protein modification	Yes
AK064792	Yes	Yes	197[d]	109	108	57	8.4e-26	F9L1.29 protein	Not available	Yes
AK060523	Yes	Yes	Yes	166	166	100	1.9e-88	Uncharacterized protein (probable ankyrin-3)	[go:9507] chloroplast [go:5515] protein binding	Yes

[a] Used rice cDNA in blastn search against "NCBI EST_Others" database (rice) to search for longer 5' ESTs.

[b] Used rice cDNA in blastn search against "TIGR Rice Genome Annotation DB: Coding Sequences" database to verify the cDNA ORF.

[c] Translated the rice cDNA in the same frame as the main open reading frame identified by uORFSCAN (include translations upstream of predicted start Methionine). The resulting protein sequence was used in a blastp search against "The UniProt Knowledgebase (UniProt)" database.

[d] The genome annotation for the CDS is longer by the indicated number of base pairs.

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

Rice		Wheat		Barley		Maize		Sorghum		Avg. A.A.	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	similarity (%)	
AK121001	90_33_113			TC146266	76_33_71	TC279901	100_33_67	TC102588	118_33_67	60	Transcription factor
AK120494	199_21_34	TC256417	136_21_625	TC134801	394_21_404			TC97268	166_21_144	17	Hypothetical protein
AK119592	304_90_148			TC140173	311_90_110	TC297985	464_90_147	TC103116	310_90_302	72	Leucine zipper protein 16
					287_144_110		440_114_147		286_114_302	68	
					287_144_110		440_114_147		286_114_302	70	
AK104437	187_42_203	TC266855	203_42_174	TC133317	181_42_178	TC282409	251_42_164			92	RNA-binding protein cabeza
AK103391	157_123_74	TC269775	203_123_62	TC134190	156_123_62	TC294011	167_123_75			70	Trehalose
AK102376	115_24_31	TC237876	95_24_48	TC133824	87_24_47			TC101133	576_24_21	13	Zinc finger protein-like
AK102277	267_78_150	TC250018	255_78_130			TC299034	266_78_144	TC102365	258_78_137	92	AP2 domain-containing protein
	228_117_150		216_117_130				227_117_144		219_117_137	82	
	126_219_150		108_225_130				131_213_144		120_216_137	65	
AK102068	463_12_11	TC243607	181_12_14	TC136167	397_12_315			TC103028	402_12_14	33	
AK100578	249_9_10	TC241920	568_9_372			TC300179	180_9_136	TC103751	83_9_62	50	mRNA capping enzyme-like
AK099540	277_6_475			TC139607	184_6_495	TC280858	23_6_227	TC101936	513_6_465	100	Nam-like protein 2
AK073985	101_12_101	TC252583	179_12_900	TC148772	149_12_82			TC92492	123_12_329	33	RNA-binding protein FUS
AK070766	144_15_65	TC263230	129_15_44	TC134132	121_15_44	TC305003	186_15_43			50	PG4
AK065585	126_15_34	TC254095	64_15_42	TC139863	34_15_48	TC311554	192_15_22			100	Monodehydroascorbate
AK065240	112_21_17			TC132139	231_21_13	TC298549	354_21_28	TC93046	63_21_111	83	Arabinofuranohydrolase
AK065176	333_12_179	TC235016	413_12_202	TC139184	466_12_203			TC93049	357_12_192	67	Phosphatidylinositol
AK061109	35_12_98	TC263378	53_12_68			TC306071	79_12_66	TC96170	81_12_57	33	Hypothetical protein

[a] Pre-ORF distance_uORF length_intercistronic distance.
[b] Functional annotation based on "The UniProt Knowledgebase (UniProt)" database.
Identifiers may not be unique among the tables as different combinations of uORFs were conserved.
Ribosomal rRNA genes have been removed.

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

Rice		Wheat		Barley		Maize		Sorghum		Avg. A.A.	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	similarity (%)	
AK122113	73_12_635					TC283896	249_12_198	TC99912	357_12_273	33	Unknown protein
AK121846	204_39_102					TC292285	37_39_477	TC103358	136_39_228	25	K^+ efflux antiporter
AK121416	625_36_180			TC134160	116_36_41	TC282124	164_36_38			1	DNA-directed RNA polymerase II
	571_33_237				119_33_41			TC105440	34_33_140	10	
	251_126_464						102_126_10		38_132_37	2	
	248_129_464						102_126_10		38_132_37	2	
AK121398	120_27_63					TC311365	164_27_38	TC104131	111_27_82	40	Hypothetical protein
AK121128	1194_63_16					TC297930	132_63_301	TC106329	387_66_345	5	AC transposase
	1011_60_202					TC297930	132_63_301		642_57_99	1	
AK120409	257_12_1201	TC253102	218_12_752			TC313498	271_12_255			100	Cyclin T1
AK120173	386_51_59	TC252694	391_51_358					TC93304	397_51_258	1	Hypothetical protein
	33_30_433		36_30_734						183_30_493	56	
AK111899	168_21_68	TC254194	18_21_260					TC105689	513_21_42	1	Response regulator 10
AK111821	978_69_878	TC269697	396_66_78					TC96287	342_66_372	9	Transcription factor MYB86
	795_75_1055		390_72_78						336_72_372	4	
	212_69_1644		390_72_78						336_72_372	8	
	1223_63_639		396_66_78						342_66_372	10	
	1125_15_785		360_15_165						306_15_459	50	
AK111748	540_12_55			TC142603	21_12_95	TC300604	546_12_29			33	Ethylene receptor-like protein 1
AK105484	181_96_51	TC235660	198_93_56					TC93642	173_99_47	17	Homeodomain leucine zipper protein
AK103391	176_30_148	TC269775	222_30_136			TC294011	186_30_149			56	Trehalose-6-phosphate phosphatase
	130_27_197		176_27_185	TC134190	129_27_185					38	
	118_39_197		164_39_185	TC134190	117_39_185					50	
AK102966	206_9_32	TC247483	188_9_14	TC142783	160_9_138					50	Type 5 serine/threonine phosphatase
AK102080	899_9_127					TC298645	131_9_125	TC107852	461_9_24	50	Arm repeat-containing protein
	491_9_535						131_9_125		461_9_24	50	
	272_6_757						123_6_136		33_6_455	100	
AK101720	152_9_74	TC270620	187_9_264			TC289352	188_9_675			50	Probable calcium-binding
AK101319	976_9_280			TC142174	177_9_335	TC298112	153_9_176			50	Hypothetical protein
	898_72_295				446_75_0		75_72_191			12	
	544_75_646				446_75_0		75_72_191			16	
	532_87_646	TC271530	20_87_40		434_87_0					14	
	490_129_646				392_129_0		136_123_79			10	
	269_9_987				177_9_335		153_9_176			50	
AK101100	142_12_21	TC263224	132_12_14	TC132639	175_12_510					100	Protein phosphatase 2A 55 kDa
AK100780	276_21_83			TC141120	409_21_258			TC105228	86_21_152	14	Activin receptor type II precursor
AK100589	300_54_229					TC292591	343_54_218	TC91317	317_54_227	94	S-adenosylmethionine decarboxylase
AK100299	692_21_187	TC239370	264_21_11				574_21_163			1	
AK099745	136_21_245	TC269480	129_21_140	TC136177	22_21_245					17	Glutamate receptor 3.2 precursor
More …											

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

Rice		Wheat		Barley		Maize		Sorghum		Avg. A.A.	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]			Identifier	5′-UTR[a]	similarity (%)	
AK099540	145_90_523			TC13960	7 52_90_543			TC101936	381_90_513	83	Nam-like protein 2
AK072868	377_51_96	TC247418	389_51_111	TC139536	429_51_117					81	Serine/threonine kinase
AK072499	555_90_1239			TC139601	252_87_88			TC105154	157_93_526	10	Short stature homeobox protein 2
	416_9_1459	TC267242	82_9_80						431_9_336	50	
	317_69_1498		62_72_37			TC281509	24_66_263			1	
	210_33_1641				63_33_331				485_33_258	10	
	1675_66_143						24_66_263		455_63_258	9	
	1391_96_397						90_96_167		157_93_526	1	
	1258_66_560						24_66_263		455_63_258	1	
	1159_9_716		82_9_80						740_9_27	50	
	1094_9_781		82_9_80						740_9_27	50	
AK072427	7_27_136	TC258198	99_27_41			TC308361	214_27_640			13	Hypothetical protein
AK072349	376_9_36			TC137384	305_9_235	TC313267	310_9_31			100	Enhancer of polycomb-like protein
AK071762	87_12_116			TC131045	140_12_347			TC101994	100_12_109	33	Wali7 protein
AK070751	664_33_209	TC240522	226_33_83	TC142763	298_33_5					9	F7N22.3 protein
	501_60_345				208_60_68			TC106350	122_57_42	1	
	398_6_502	TC240522	264_6_72			TC294109	238_6_201			100	
AK070456	774_9_51					TC288447	587_9_472	TC97361	72_9_118	50	Molybdenum cofactor Cnx1
AK069730	770_156_22	TC246998	270_150_246	TC132118	275_159_249					15	Unknown protein
	412_153_383		270_150_246		275_159_249					47	
AK067468	3_6_164			TC138312	401_6_126	TC294470	533_6_178			100	Phosphatidylinositol 3,5-kinase-like
AK066942	259_12_32	TC253984	286_12_43	TC133589	262_12_36					67	Expressed protein
AK066073	154_75_125	TC236575	204_75_379			TC293675	596_75_169			1	Acetyl-coenzyme A synthetase
AK065538	162_24_57			TC139620	197_24_39	TC287533	93_24_70			14	Clathrin coat assembly protein
AK065237	62_9_226					TC288346	32_9_124	TC10281	0 65_9_245	50	Expressed protein
	174_9_114						152_9_4	TC102810	65_9_245	50	
AK065176	315_30_179	TC235016	395_30_202	TC139184	448_30_203					44	Phosphatidylinositol 3-and 4-kinase-like
AK065137	8_21_281	TC251833	13_21_266	TC147261	7_21_263					83	Kelch-like ECH-associated protein 1
AK064792	281_99_98	TC267323	259_99_98			TC306152	268_99_98			72	Hypothetical protein
	276_15_187		254_15_187				263_15_187			100	
AK061004	108_9_30	TC269443	96_9_25	TC151138	130_9_25					100	Peptidylprolyl isomerase
AK060780	546_6_320			TC134531	513_6_122	TC311790	53_6_34			100	Pelota (PEL1)
	440_6_426				513_6_122		53_6_34			100	
AK060523	60_27_394					TC305149	127_27_425	TC103609	133_27_421	75	Ankyrin-2
AK058965	4_186_71					TC288549	102_186_226	TC93810	51_186_67	75	Nitrilase 1
AK058513	94_24_26			TC147191	616_24_26	TC305089	645_24_26			57	Leucine aminopeptidase pre protein
	34_84_26				556_84_26		585_84_26			63	
	128_6_10				640_6_20		669_6_20			100	
	118_24_2				616_24_26		645_24_26			14	
AK058988	139_69_294	TC235910	265_69_267			TC314670	272_69_84			78	Calcium-binding protein like

[a] Pre-ORF distance_uORF length_intercistronic distance

[b] Functional annotation based on "The UniProt Knowledgebase (UniProt)" database

Identifiers may not be unique among the tables as different combinations of uORFs were conserved.

Ribosomal rRNA genes have been removed.

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

Rice		Wheat		Barley		Maize		Avg. A.A. similarity (%)	Putative function[b]
Identifer	5′-UTR[a]	Identifer	5′-UTR[a]	Identifer	5′-UTR[a]	Identifer	5′-UTR[a]		
AK121850	86_18_51	TC238796	102_18_57	TC140406	84_18_58	TC292944	94_18_72	20	Protein kinase CK2
AK104437	187_42_203	TC266855	203_42_174	TC133317	181_42_178	TC282409	251_42_164	92	RNA-binding protein cabeza
AK103140	271_36_1	TC266113	212_36_407	TC140479	182_36_1	TC281091	194_36_1	64	Protein phosphatase 2C
AK101684	158_21_12	TC253407	137_21_33	TC141318	104_21_33	TC307223	180_21_16	33	CCAAT-binding transcription factor
AK100440	246_81_195	TC235293	210_78_152	TC133630	180_78_153	TC280879	138_78_77	4	BZIP transcription factor
AK099839	147_48_82	TC237323	145_48_50	TC140250	758_48_585	TC309986	144_48_57	7	MAP3K epsilon protein kinase
AK073303	67_9_142	TC237149	75_9_113	TC132556	81_9_139	TC305609	127_9_69	50	Hypothetical protein
	135_9_74		75_9_113		81_9_139		127_9_69	50	
AK072868	392_36_96	TC247418	404_36_111	TC139536	444_36_117	TC306591	633_36_366	8	Serine/threonine kinase
	338_90_96		347_93_111		387_93_117		576_93_366	6	
	269_39_216		278_39_234		318_39_240		768_39_228	8	
	259_195_70		268_198_85		308_198_91		260_192_583	35	
	249_27_248		258_27_266		298_27_272		444_27_564	11	
AK072649	100_192_117	TC236348	79_192_117	TC133316	76_192_93	TC305793	180_192_116	81	Ribosomal protein S6 kinase homolog
AK070766	144_15_65	TC263230	129_15_44	TC134132	121_15_44	TC305003	186_15_43	50	Protein C20orf11
AK069526	737_87_60	TC265553	757_87_62	TC147034	740_87_62	TC287352	477_87_65	32	GAMYB-binding protein
	690_9_185		709_9_188		692_9_188		427_9_193	50	
	440_102_342		453_102_351		436_102_351		150_102_377	12	
AK066145	178_12_58	TC266262	149_12_73	TC134484	154_12_231	TC286452	224_12_70	33	F2E2.12
AK065585	126_15_34	TC254095	64_15_42	TC139863	34_15_48	TC311554	192_15_22	100	Monodehydroascorbate reductase
AK063875	128_78_44	TC238591	348_78_2	TC133365	519_78_2	TC312696	325_78_2	12	Prokineticin 2 precursor
AK060523	173_123_185	TC235416	201_126_157	TC148319	211_120_163	TC305149	255_129_195	60	Ankyrin-3
AK060232	38_15_10	TC273755	224_15_44	TC135479	410_15_99	TC291309	171_15_4	25	SAM-dependent methyltransferase-like

[a] Pre-ORF distance_uORF length_intercistronic distance

[b] Functional annotation based on "The UniProt Knowledgebase (UniProt)" database

Identifiers may not be unique among the tables as different combinations of uORFs were conserved.

Ribosomal rRNA genes have been removed.

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

Rice		Wheat		Barley		Maize		Avg. A.A. similarity (%)	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]		
AK121416	625_36_180			TC134160	116_36_41	TC282124	164_36_38	1	DNA-directed RNA polymerase
AK121122	743_21_34	TC266483	118_21_70	TC148472	179_21_69			14	SNF protein
AK121001	90_33_113			TC146266	76_33_71	TC279901	100_33_67	60	Transcription factor
AK120494	199_21_34	TC256417	136_21_625	TC134801	394_21_404			17	Hypothetical protein
AK120409	257_12_1201	TC253102	218_12_752			TC313498	271_12_255	100	Cyclin T1
AK119650	98_21_61	TC247011	129_21_73	TC148824	219_21_8			17	MAP kinase MAPK2
AK119592	304_90_148			TC140173	311_90_110	TC297985	464_90_147	72	Homeodomain leucine zipper protein 16
	283_111_148				287_114_110		440_114_147	68	
	280_114_148				287_114_110		440_114_147	70	
AK111887	244_21_108	TC235829	289_21_48	TC131138	232_21_48			67	Calcineurin B protein
AK111883	623_12_193	TC232511	688_12_250			TC295627	422_12_387	25	Hypothetical WD-repeat protein
	508_45_275		24_45_881				497_45_279	64	
	433_54_341		195_54_701				236_54_531	6	
	348_12_468		139_12_799				338_12_471	33	
	300_60_468		288_60_602				290_60_471	15	
	246_54_528		195_54_701				236_54_531	6	
AK111748	540_12_55			TC142603	21_12_95	TC300604	546_12_29	33	Ethylene receptor-like protein 1
AK111699	401_9_0	TC258667	378_9_360			TC307333	371_9_0	100	tRNA-dihydrouridine synthase 3
AK106310	547_72_149	TC273695	366_72_174	TC136869	1_72_168			21	Hypothetical protein
	128_21_619		471_21_120		106_21_114			17	
AK103631	376_15_48			TC139875	822_15_0	TC306875	445_15_0	25	Hypothetical protein
	328_24_87	TC253392	188_24_308		224_24_589			13	
	313_39_87		173_39_308		209_39_589			17	
	277_18_144				3_18_816	TC306875	442_18_0	20	
AK103390	277_51_25	TC236507	242_51_24	TC132330	379_51_24			69	Hypothetical protein
AK103207	167_9_20	TC269820	173_9_475			TC306369	149_9_25	50	protein kinase
AK103040	8_54_271	TC254504	13_54_233			TC293848	19_54_248	35	Single myb histone 1
	45_87_201	TC254504		TC148672	25_84_166			64	
AK102966	206_9_32	TC247483	188_9_14	TC142783	160_9_138			50	Type 5 serine/threonine phosphatase 55
AK102376	115_24_31	TC237876	95_24_48	TC133824	87_24_47			14	Zinc finger (C3HC4-type RING finger)
AK102370	127_60_11	TC255624	139_63_11	TC133336	401_57_20			30	Tubby-like protein 3
	124_63_11		139_63_11			TC312575	64_66_92	10	
AK102277	267_78_150	TC250018	255_78_130			TC299034	266_78_144	92	Unknown protein
	228_117_150		216_117_130				227_117_144	82	
	126_219_150		108_225_130				131_213_144	65	
AK102068	463_12_11	TC243607	181_12_14	TC136167	397_12_315			33	Hypothetical protein
AK101942	106_18_51	TC248321	82_18_27			TC310601	4_18_283	40	Calcium-dependent protein kinase
AK101720	152_9_74	TC270620	187_9_264			TC289352	188_9_675	50	Probable calcium-binding mitochondrial protein F19P19.26
AK101520	222_30_1281			TC136686	134_30_360	TC297598	227_30_3	11	
AK101319	976_9_280			TC142174	177_9_335	TC298112	153_9_176	50	Hypothetical protein
	898_72_295				446_75_0		75_72_191	12	
	544_75_646				446_75_0		75_72_191	16	
	532_87_646	TC271530	20_87_40		434_87_0			14	

More ...

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

Rice		Wheat		Barley		Maize		Avg. A.A.	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	similarity (%)	
	490_129_646				392_129_0		136_123_79	10	
	269_9_987				177_9_335		153_9_176	50	
AK101266	493_9_69			TC130775	526_9_432	TC285008	333_9_90	50	Thiol protease aleurain precursor
	464_9_98				526_9_432		333_9_90	50	
	457_45_69				301_45_621		270_45_117	14	
AK101100	142_12_21	TC263224	132_12_14	TC132639	175_12_510			100	Protein phosphatase 2A 55 kDa B
AK100578	249_9_10	TC241920	568_9_372			TC300179	180_9_136	50	MRNA capping enzyme-like protein
AK100539	301_45_11	TC236703	339_45_11			TC305240	183_45_11	93	Dentin sialophosphoprotein 1
AK100332	50_9_2019	TC272800	48_9_152	TC153584	275_9_342			50	Chromodomain helicase DNA binding
	262_78_1738		69_75_65		71_81_474			4	
	1769_9_300		48_9_152		275_9_342			50	
	1744_87_247		57_87_65		479_90_57			3	
	1657_9_412		48_9_152		275_9_342			50	
	1513_78_487		69_75_65		71_81_474			7	
	1505_9_564		48_9_152		275_9_342			50	
	1435_9_634		48_9_152		275_9_342			50	
	134_72_1872		72_72_65		80_72_474			1	
AK100299	692_21_187	TC239370	264_21_11			TC291351	574_21_163	1	Hypothetical protein
AK100037	449_33_85	TC234512	400_33_495	TC134276	222_33_507			90	SAC domain-containing protein
AK099852	906_9_2	TC233509	103_9_218	TC144509	159_9_177			50	Hypothetical protein
AK099745	136_21_245	TC269480	129_21_140	TC136177	22_21_245			17	Glutamate receptor 3.2
AK099676	77_18_8	TC247479	425_18_669			TC294642	38_18_8	20	ATPase
AK099625	353_9_26			TC140108	238_9_26	TC281333	102_9_25	100	Hypothetical protein
AK099540	277_6_475			TC139607	184_6_495	TC280858	23_6_227	100	Nam-like protein 2
AK074023	6_75_94			TC145114	158_78_142	TC306013	274_72_248	4	Hypothetical protein
AK073985	101_12_101	TC252583	179_12_900	TC148772	149_12_82			67	RNA-binding protein FUS
AK072868	377_51_96	TC247418	389_51_111	TC139536	429_51_117			81	Serine/threonine kinase
AK072769	272_156_77	TC265505	208_150_83			TC292123	215_153_83	20	Hypothetical protein
AK072499	317_69_1498	TC267242	62_72_37			TC281509	24_66_263	1	Short stature homeobox
AK072427	7_27_136	TC258198	99_27_41			TC308361	214_27_640	13	Hypothetical protein
AK072349	376_9_36			TC137384	305_9_235	TC313267	310_9_31	100	Enhancer of polycomb-like protein,
AK070751	664_33_209	TC240522	226_33_83	TC142763	298_33_5			9	F7N22.3 protein
	398_6_502		264_6_72			TC294109	238_6_201	100	
AK069730	770_156_22	TC246998	270_150_246	TC132118	275_159_249			15	Hypothetical protein
	412_153_383		270_150_246		275_159_249			47	
AK069726	120_78_82	TC235568	119_78_74	TC139583	107_78_72			80	CBL-interacting protein kinase 23
AK069526	214_126_544	TC265553	239_123_544	TC147034	222_123_544			80	GAMYB-binding protein
	149_246_489		174_243_489		157_243_489			63	
AK069065	133_12_97	TC266624	198_12_77	TC132959	163_12_73			33	RAD23-like protein
AK068416	254_33_43	TC239989	34_33_29			TC287928	312_33_796	20	Expressed protein
AK067468	3_6_164			TC138312	401_6_126	TC294470	533_6_178	100	Phosphatidylinositol 3,5-kinase-like
AK067412	222_84_49	TC252944	247_81_102	TC142664	123_84_118			19	Protein kinase
AK067258	246_27_25	TC247646	508_27_43	TC140304	193_27_46			38	Ankyrin-like protein
AK067123	840_72_630			TC132179	380_75_759	TC300140	66_69_62	4	Ubiquitin-specific protease 12
	579_69_894				963_66_185		66_69_62	1	
	291_72_1179				380_75_759		66_69_62	4	

More ...

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

Rice		Wheat		Barley		Maize		Avg. A.A.	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	Identifier	5′-UTR[a]	similarity (%)	
AK066952	437_57_119	TC271435	169_57_23	TC137456	357_57_231			9	Arabidopsis thaliana genomic DNA
	392_39_182		153_39_57		222_39_384			8	
AK066942	259_12_32	TC253984	286_12_43	TC133589	262_12_36			67	Expressed protein
AK066480	146_24_104	TC256019	215_24_326	TC148944	275_24_194			14	Hypothetical protein
AK066424	406_6_302			TC148993	13_6_560	TC281469	539_6_92	100	RING zinc finger protein-like
AK066073	154_75_125	TC236575	204_75_379			TC293675	596_75_169	1	Acetyl-coenzyme A synthetase
AK065998	108_72_28	TC253336	232_69_101	TC150526	760_72_308			4	Hypothetical protein
AK065863	644_45_390	TC255161	74_45_17			TC287626	306_45_23	6	Multidrug-resistance associated protein 1
AK065729	398_9_62	TC243618	239_9_191	TC134511	147_9_291			50	Hypothetical protein
	244_9_216		239_9_191		147_9_291			50	
	193_60_216		62_57_320		229_63_155			1	
AK065683	82_18_41	TC243502	103_18_600	TC153017	307_18_9			20	Cell division protein kinase 8
AK065578	470_120_117	TC249752	503_117_110			TC305318	318_117_159	1	Transformer-2-like protein
	362_108_237		339_108_283				362_108_124	77	
	325_51_331		302_51_377	TC139461	274_51_93			69	
AK065538	162_24_57			TC139620	197_24_39	TC287533	93_24_70	14	Clathrin coat assembly protein AP47
AK065240	112_21_17			TC132139	231_21_13	TC298549	354_21_28	83	Arabinoxylan arabinofuranohydrolase isoenzyme
AK065176	333_12_179	TC235016	413_12_202	TC139184	466_12_203			67	Phosphatidylinositol
	315_30_179		395_30_202		448_30_203			44	
AK065137	8_21_281	TC251833	13_21_266	TC147261	7_21_263			83	Kelch-like ECH-associated protein 1
AK065016	15_63_292			TC134594	263_63_414	TC287474	127_66_507	4	Hydroxyproline-rich
AK064864	75_15_41			TC138646	106_15_35	TC288136	159_15_32	75	Unknown protein
AK064792	281_99_98	TC267323	259_99_98			TC306152	268_99_98	72	
	276_15_187		254_15_187				263_15_187	100	
AK063846	171_12_19	TC239301	99_12_12			TC287762	164_12_12	67	protein F12M16.29
AK061109	35_12_98	TC263378	53_12_68			TC306071	79_12_66	33	Hypothetical protein
AK061004	108_9_30	TC269443	96_9_25	TC151138	130_9_25			100	Peptidylprolyl isomerase
AK060780	546_6_320			TC134531	513_6_122	TC311790	53_6_34	100	Pelota (PEL1)
	440_6_426				513_6_122		53_6_34	100	
AK059720	301_33_27	TC239546	86_33_74	TC149822	139_33_102			10	Hypothetical protein
AK059001	179_117_246	TC269581	134_120_152	TC142662	193_117_14			10	Calyx protein
	170_126_246		134_120_152		184_126_14			7	
AK058988	139_69_294	TC235910	265_69_267			TC314670	272_69_84	45	Calcium-binding protein-like
AK058880	106_51_4	TC269547	203_51_4	TC152057	83_51_420			50	Lipase class 3-like
AK058513	94_24_26			TC147191	616_24_26	TC305089	645_24_26	57	Neutral leucine aminopeptidase
	34_84_26				556_84_26		585_84_26	63	
	128_6_10				640_6_20		669_6_20	100	
	118_24_2				616_24_26		645_24_26	14	
AK058462	11_15_32			TC150825	122_15_42	TC282858	232_15_86	25	Transporter associated with antigen
AK121850	86_18_51	TC238796	102_18_57	TC140406	84_18_58			40	Kinase CK2 regulatory subunit

[a] Pre-ORF distance_uORF length_intercistronic distance
[b] Functional annotation based on "The UniProt Knowledgebase (UniProt)" database
Identifiers may not be unique among the tables as different combinations of uORFs were conserved.
Ribosomal rRNA genes have been removed.

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

Rice		Wheat		Barley		Avg. A.A. similarity (%)	Putative function[b]
Identifer	5′-UTR[a]	Identifer	5′-UTR[a]	Identifer	5′-UTR[a]		
AK122166	338_18_606	TC250897	2_18_164	TC149838	233_18_254	20	Translation initiation factor 3
AK122131	322_9_28	TC251213	242_9_30	TC147542	253_9_31	100	Chitin-inducible gibberellin-responsive
AK121850	86_18_51	TC238796	102_18_57	TC140406	84_18_58	40	Kinase CK2 regulatory subunit
AK121122	743_21_34	TC266483	118_21_70	TC148472	179_21_69	14	NF protein
AK120494	199_21_34	TC256417	136_21_625	TC134801	394_21_404	17	Hypothetical protein F17M5.140
AK119650	98_21_61	TC247011	129_21_73	TC148824	219_21_8	17	MAP kinase MAPK2
AK111887	244_21_108	TC235829	289_21_48	TC131138	232_21_48	67	Calcineurin B protein
AK106310	547_72_149	TC273695	366_72_174	TC136869	1_72_168	21	Hypothetical protein
	128_21_619		471_21_120		106_21_114	17	
AK104437	187_42_203	TC266855	203_42_174	TC133317	181_42_178	92	RNA-binding protein cabeza
AK103631	328_24_87	TC253392	188_24_308	TC139875	224_24_589	13	Hypothetical protein
	313_39_87		173_39_308		209_39_589	17	
AK103391	205_75_74	TC269775	251_75_62	TC134190	204_75_62	92	Trehalose-6-phosphate phosphatase
	157_123_74		203_123_62		156_123_62	80	
	130_27_197		176_27_185		129_27_185	38	
	118_39_197		164_39_185		117_39_185	50	
AK103390	277_51_25	TC236507	242_51_24	TC132330	379_51_24	69	Non-imprinted in Prader-Willi/Angelman syndrome region protein 2
AK103140	271_36_1	TC266113	212_36_407	TC140479	182_36_1	73	Hypothetical protein
AK103040	45_87_201	TC254504	50_84_166	TC148672	25_84_166	64	Single myb histone 1
AK102966	206_9_32	TC247483	188_9_14	TC142783	160_9_138	50	Type 5 serine/threonine phosphatase 55
AK102376	115_24_31	TC237876	95_24_48	TC133824	87_24_47	14	Zinc finger (C3HC4-type RING finger)
AK102370	127_60_11	TC255624	139_63_11	TC133336	401_57_20	30	Tubby-like protein 3
AK102068	463_12_11	TC243607	181_12_14	TC136167	397_12_315	33	Hypothetical protein
AK101684	158_21_12	TC253407	137_21_33	TC141318	104_21_33	50	CCAAT-box transcription factor
AK101539	37_12_125	TC251540	98_12_110	TC140495	36_12_129	100	CG11670-PA
AK101319	532_87_646	TC271530	20_87_40	TC142174	434_87_0	14	Hypothetical protein F14F8_120
AK101100	142_12_21	TC263224	132_12_14	TC132639	175_12_510	100	Protein phosphatase 2A 55 kDa B
AK100440	246_81_195	TC235293	210_78_152	TC133630	180_78_153	31	BZIP transcription factor, complete
AK100037	449_33_85	TC234512	400_33_495	TC134276	222_33_507	90	SAC domain-containing protein
AK099852	906_9_2	TC233509	103_9_218	TC144509	159_9_177	50	Hypothetical protein
AK099839	147_48_82	TC237323	145_48_50	TC140250	758_48_585	20	MAP3K epsilon protein kinase
AK099745	136_21_245	TC269480	129_21_140	TC136177	22_21_245	17	Glutamate receptor 3.2
AK073985	101_12_101	TC252583	179_12_900	TC148772	149_12_82	67	RNA-binding protein FUS
AK073303	67_9_142	TC237149	75_9_113	TC132556	81_9_139	100	Hypothetical protein
	135_9_74		75_9_113		81_9_139	50	
AK072868	392_36_96	TC247418	404_36_111	TC139536	444_36_117	91	Serine/threonine kinase
	377_51_96		389_51_111		429_51_117	81	
	338_90_96		347_93_111		387_93_117	53	
	269_39_216		278_39_234		318_39_240	83	
	259_195_70		268_198_85		308_198_91	65	
	249_27_248		258_27_266		298_27_272	75	
AK072649	100_192_117	TC236348	79_192_117	TC133316	76_192_93	87	Ribosomal protein S6 kinase homolog
AK072244	124_15_222	TC252797	107_15_14	TC136383	66_15_390	25	Hypothetical protein
AK072085	725_6_132	TC253625	6_6_624	TC150175	602_6_23	100	RNA polymerase II termination
	683_6_174		6_6_624		602_6_23	100	
	469_6_388		6_6_624		602_6_23	100	
	397_78_388		491_78_67		229_81_321	1	
AK070766	144_15_65	TC263230	129_15_44	TC134132	121_15_44	75	PG4
AK070751	664_33_209	TC240522	226_33_83	TC142763	298_33_5	9	F7N22.3 protein

More…

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

Rice		Wheat		Barley		Avg. A.A.	Putative function[b]
Identifer	5′-UTR[a]	Identifer	5′-UTR[a]	Identifer	5′-UTR[a]	similarity (%)	
AK069730	770_156_22	TC246998	270_150_246	TC132118	275_159_249	15	Hypothetical protein
	412_153_383		270_150_246		275_159_249	47	
AK069726	120_78_82	TC235568	119_78_74	TC139583	107_78_72	80	Hordeum vulgare mRNA for expressed sequence tag
AK069534	870_57_327	TC236981	222_57_931	TC139404	90_57_190	6	Auxilin-like protein
	603_57_594		222_57_931		90_57_190	6	
	411_93_750		982_93_135		51_96_190	6	
	1068_96_90		982_93_135		51_96_190	3	
AK069526	737_87_60	TC265553	757_87_62	TC147034	740_87_62	54	GAMYB-binding protein
	690_9_185		709_9_188		692_9_188	100	
	440_102_342		453_102_351		436_102_351	18	
	214_126_544		239_123_544		222_123_544	80	
	149_246_489		174_243_489		157_243_489	63	
AK069065	133_12_97	TC266624	198_12_77	TC132959	163_12_73	33	RAD23-like protein
AK067412	222_84_49	TC252944	247_81_102	TC142664	123_84_118	19	Protein kinase
AK067258	246_27_25	TC247646	508_27_43	TC140304	193_27_46	38	Ankyrin-like protein
AK067156	1433_6_261	TC238252	27_6_439	TC140969	143_6_148	100	Hypothetical protein
AK066952	437_57_119	TC271435	169_57_23	TC137456	357_57_231	9	Arabidopsis thaliana genomic DNA
	392_39_182		153_39_57		222_39_384	8	
AK066942	259_12_32	TC253984	286_12_43	TC133589	262_12_36	67	Expressed protein
AK066480	146_24_104	TC256019	215_24_326	TC148944	275_24_194	14	Hypothetical protein
AK066307	1325_12_14	TC264007	215_12_16	TC149400	160_12_15	100	RNA polymerase alpha subunit
	1232_6_113		109_6_128		60_6_121	100	
AK066145	178_12_58	TC266262	149_12_73	TC134484	154_12_231	67	Protein F2E2.12
AK065998	108_72_28	TC253336	232_69_101	TC150526	760_72_308	4	Hypothetical protein
AK065729	398_9_62	TC243618	239_9_191	TC134511	147_9_291	50	Hypothetical protein
	244_9_216		239_9_191		147_9_291	50	
	193_60_216		62_57_320		229_63_155	1	
AK065683	82_18_41	TC243502	103_18_600	TC153017	307_18_9	20	Cell division protein kinase 8
AK065585	126_15_34	TC254095	64_15_42	TC139863	34_15_48	100	Monodehydroascorbate reductase
AK065578	325_51_331	TC249752	302_51_377	TC139461	274_51_93	69	Transformer-2-like protein
AK065329	444_15_35	TC238280	174_15_44	TC147756	276_15_44	50	Hypothetical protein F14M19.150
AK065176	333_12_179	TC235016	413_12_202	TC139184	466_12_203	67	Phosphatidylinositol 3
	315_30_179		395_30_202		448_30_203	44	
AK065137	8_21_281	TC251833	13_21_266	TC147261	7_21_263	83	Kelch-like ECH-associated protein 1
AK063875	128_78_44	TC238591	348_78_2	TC133365	519_78_2	23	Prokineticin 2 precursor
AK061004	108_9_30	TC269443	96_9_25	TC151138	130_9_25	100	peptidylprolyl isomerase
AK060783	216_21_173	TC247469	243_21_176	TC132683	307_21_175	100	Hypothetical protein
AK060523	173_123_185	TC235416	201_126_157	TC148319	211_120_163	68	Hypothetical protein
AK060232	38_15_10	TC273755	224_15_44	TC135479	410_15_99	25	SAM-dependent methyltransferase-like
AK059720	301_33_27	TC239546	86_33_74	TC149822	139_33_102	10	Hypothetical protein
AK059394	1_9_26	TC270230	46_9_128	TC134188	182_9_266	50	Small nuclear ribonucleoprotein
AK059001	179_117_246	TC269581	134_120_152	TC142662	193_117_14	10	Calyx protein
	170_126_246		134_120_152		184_126_14	7	
AK058880	106_51_4	TC269547	203_51_4	TC152057	83_51_420	50	Lipase class 3-like

[a] Pre-ORF distance_uORF length_intercistronic distance

[b] Functional annotation based on "The UniProt Knowledgebase (UniProt)" database

Identifiers may not be unique among the tables as different combinations of uORFs were conserved.

Ribosomal rRNA genes have been removed.

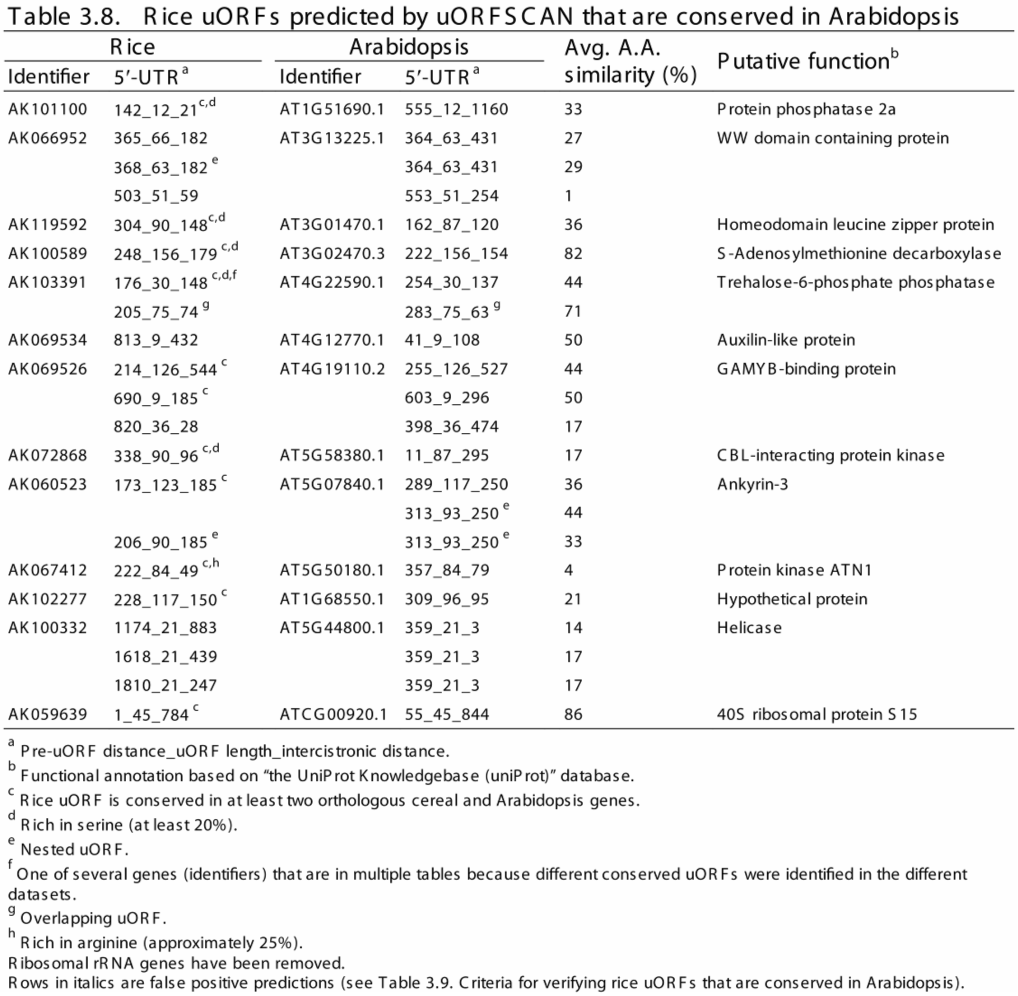


Table 3.8. Rice uORFs predicted by uORFSCAN that are conserved in Arabidopsis

Rice		Arabidopsis		Avg. A.A. similarity (%)	Putative function[b]
Identifier	5′-UTR[a]	Identifier	5′-UTR[a]		
AK101100	142_12_21[c,d]	AT1G51690.1	555_12_1160	33	Protein phosphatase 2a
AK066952	365_66_182	AT3G13225.1	364_63_431	27	WW domain containing protein
	368_63_182[e]		364_63_431	29	
	503_51_59		553_51_254	1	
AK119592	304_90_148[c,d]	AT3G01470.1	162_87_120	36	Homeodomain leucine zipper protein
AK100589	248_156_179[c,d]	AT3G02470.3	222_156_154	82	S-Adenosylmethionine decarboxylase
AK103391	176_30_148[c,d,f]	AT4G22590.1	254_30_137	44	Trehalose-6-phosphate phosphatase
	205_75_74[g]		283_75_63[g]	71	
AK069534	813_9_432	AT4G12770.1	41_9_108	50	Auxilin-like protein
AK069526	214_126_544[c]	AT4G19110.2	255_126_527	44	GAMYB-binding protein
	690_9_185[c]		603_9_296	50	
	820_36_28		398_36_474	17	
AK072868	338_90_96[c,d]	AT5G58380.1	11_87_295	17	CBL-interacting protein kinase
AK060523	173_123_185[c]	AT5G07840.1	289_117_250	36	Ankyrin-3
			313_93_250[e]	44	
	206_90_185[e]		313_93_250[e]	33	
AK067412	222_84_49[c,h]	AT5G50180.1	357_84_79	4	Protein kinase ATN1
AK102277	228_117_150[c]	AT1G68550.1	309_96_95	21	Hypothetical protein
AK100332	1174_21_883	AT5G44800.1	359_21_3	14	Helicase
	1618_21_439		359_21_3	17	
	1810_21_247		359_21_3	17	
AK059639	1_45_784[c]	ATCG00920.1	55_45_844	86	40S ribosomal protein S15

[a] Pre-uORF distance_uORF length_intercistronic distance.
[b] Functional annotation based on "the UniProt Knowledgebase (uniProt)" database.
[c] Rice uORF is conserved in at least two orthologous cereal and Arabidopsis genes.
[d] Rich in serine (at least 20%).
[e] Nested uORF.
[f] One of several genes (identifiers) that are in multiple tables because different conserved uORFs were identified in the different datasets.
[g] Overlapping uORF.
[h] Rich in arginine (approximately 25%).
Ribosomal rRNA genes have been removed.
Rows in italics are false positive predictions (see Table 3.9. Criteria for verifying rice uORFs that are conserved in Arabidopsis).

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

Accession	FL- cDNA[a]	Upstream & In-frame stop codon	Agreement with genome annotation[b]	Alignment of uORFSCAN identified main proteins with UniProt proteins [c]						uORF valid
				UniProt protein length (AA)	Align length (AA)	Identities (%)	Expect	Annotation	GO classication (Arabidopsis thaliana)	
AK101100	Yes	Yes	Yes	525	525	100	5.0e-287	Protein phosphatase 2A	[go:6470] protein dephosphorylation [go:166] nucleotide binding	Yes
AK066952	Yes	Yes	Yes	860	694	99	0	WW domain containing protein	Not available	Yes[d]
AK119592	Yes	Yes	Yes	343	343	100	6.8e-187	Homeodomain leucine zipper protein	[go:6355] regulation of transcription [go:3677] DNA binding	Yes
AK100589	Yes	Yes	Yes	398	398	100	1.1e-215	S-Adenosylmethionine decarboxylase	[go:6596] polyamine biosynthesis [go:5694] chromosome	Yes
AK103391	Yes	Yes	Yes	371	371	100	3.3e-194	Trehalose-6-phosphate phosphatase	[go:5992] trehalose biosynthesis [go:9507] chloroplast	Yes
AK069534	Yes	Yes	1066[e]	485	413	61	7.6e-117	Auxilin-like protein	Not available	Yes[f]
AK069526	Yes	Yes	Yes	483	483	83	5.8e-256	GAMYB-binding protein	[go:6468] protein phosphorylation [go:5524] ATP binding [go:16301] kinase activity	Yes
AK072868	Yes	Yes	Yes	443	443	100	3.5e-238	CBL-interacting kinase 15	[go:6468] protein phosphorylation [go:5524] ATP binding [go:16301] kinase activity	Yes
AK060523	No	Yes	Yes	166	166	99	8.2e-88	Ankyrin-3	[go:5515] protein binding	Yes
AK067412	Yes	Yes	Yes	353	353	72	1.2e-136	Protein kinase ATN1	[go:6468] protein phosphorylation [go:5524] ATP binding [go:16301] kinase activity	Yes
AK102277	Yes	Yes	Yes	338	338	99	4.9e-179	Hypothetical protein	Not available	Yes
AK100332	Yes	Yes	4092[e]	2192	872	30	5.3e-28	Helicase	[go:3676] nucleic acid binding [go:6355] regulation of transcription [go:5515] protein binding	No[g]
AK059639	No	Yes	Yes	154	154	100	2.6e-77	40S ribosomal s15 protein	[go:3735] structural part of ribosome [go:6412] protein biosynthesis	No[h]

[a] Used rice cDNA in blastn search against "NCBI EST_Others" database (rice) to search for longer 5' ESTs.

[b] Used rice cDNA in blastn search against "TIGR Rice Genome Annotation DB: Coding Sequences" database to verify the cDNA ORF.

[c] Translated the rice cDNA in the same frame as the main open reading frame identified by uORFSCAN (include translations upstream of predicted start Methionine). The resulting protein sequence was used in a blastp search against "The UniProt Knowledgebase (UniProt)" database.

[d] The protein data suggests that the main open reading frame predicted by uORFSCAN extends further upstream, but does not overlap the predicted uORFs and so the uORFs are still valid.

[e] The genome annotation for the CDS is longer by the indicated number of base pairs.

[f] A shorter protein was identified, but does not overlap the predicted uORFs and so the uORFs are still valid.

[g] A longer protein was identified indicating the main open reading frame extends further upstream, and does overlap the predicted uORFs and so the uORFs are not valid.

[h] Possibly not functional because pre-orf distance is less than 20 nucleotides that is thought to be required for translation initiation.

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

In five cereals

Identifier	uORF1	uORF2	uORF3	uORF4	uORF5	Main ORF
AK106095	131_9_17[a] CCGATGC[b]					157_1179 CCCATGG
AK103391	205_75_74 TTGATGA					354_1116 CAAATGG
AK100589	240_9_334 TGGATGT	248_156_179 CTAATGG	296_108_179[c] TTGATGT			583_1197 CCAATGG
AK073303	67_9_142 TCCATGC	135_9_74 CTCATGA				218_774 AGCATGG
AK072868	249_27_248 GGAATGC	259_195_70 AAGATGT	269_39_216 TGCATGC	338_90_96 TTCATGA	392_36_96 ACTATGG	524_1332 GTGATGG
AK072649	100_192_117 CTCATGA					409_1443 AAGATGG
AK066145	178_12_58[d] GCTATGG					248_360 GAGATGG
AK064792	276_15_187 CGGATGC					478_330 GGAATGG
AK060523	173_123_185[e] ACTATGG					481_501 CGGATGG

In rice and arabidopsis

Identifier	uORF1	uORF2	uORF3	uORF4	uORF5	Main ORF
AK101100	142_12_21 GCCATGG					175_1578 AAGATGG
AK066952	365_66_182 CCAATGA	368_63_182 ATGATGA	503_51_59 CTGATGA			613_2085 GGGATGC
AK119592	304_90_148 CCGATGA					542_1032 GCGATGG
AK100589	248_156_179 CTAATGG					583_1197 CCAATGG
AK103391	176_30_148 AACATGA	205_75_74 TTGATGA				354_1116 CAAATGG
AK069534	813_9_432 TCGATGA					1254_1602 GAGATGC
AK069526	214_126_544[d] GATATGG	690_9_185 TTGATGG	820_36_28 CATATGA			884_1455 AAAATGG
AK072868	338_90_96 TTCATGA					524_1332 GTGATGG
AK060523	173_123_185[e] ACTATGG	206_90_185 CCGATGC				481_501 CGGATGG
AK067412	222_84_49 CTGATGC					355_1059 GGGATGG
AK102277	228_117_150 TCTATGC					495_1017 GAAATGG

[a] Pre-ORF distance_uORF length_intercistronic distance.

[b] uORF or mainORF sequence context from -3 position to +4.

[c] AdoMetDC nested uORF found in this study.

[d] uORF sequence context good as main ORF.

[e] uORF sequence context better than main ORF.

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

Rice identifier	Alignment[a]	
	Upstream open reading frames conserved in 5/5 cereals but not in Arabidopsis	
AK106095 uORF1	AK106095_r_ORF_131_9_17	ML
	TC265929_w_ORF_113_9_16	ML
	TC148181_b_ORF_67_9_16	ML
	TC288369_m_ORF_131_9_17	ML
	TC102998_s_ORF_149_9_17	ML
		**
AK064792 uORF1	AK064792_r_ORF_276_15_187	MLCC
	TC267323_w_ORF_254_15_188	MLCC
	TC132983_b_ORF_253_15_-9	MLCC
	TC306152_m_ORF_263_15_170	MLCC
	TC107743_s_230_15_150	MLCC
		****
AK072868 uORF1	AK072868_r_ORF_249_27_248	MQKDVLAC-
	TC247418_w_ORF_258_27_266	MQKDVFAC-
	TC139536_b_ORF_298_27_272	MQRDVFAC-
	TC306591_m_ORF_444_27_564	MVK-IAGHL
	TC102544_s_ORF_331_27_265	MQKDVLAC-
		* : : .
AK072868 uORF2	AK072868_r_ORF_259_195_70	MCLHARELPCEGIGRVASHISPSTTLHDIGTQEYI-QRLLHVLSHYGVRRGNSTIFLDHHLGGDG
	TC247418_w_ORF_268_198_85	MCLHARELPCEGIGRVAAPVSALIDLDDTASQQHTTHLFFHVLLHNGVRRGISTIILDYHLGGDG
	TC139536_b_ORF_308_198_91	MCLHARELPCEGIGRVAAPLSALIDLDDTASQHHTAHLFFHVLLHNGVRRGISTIILDYHLGGDG
	TC306591_m_ORF_260_192_583	MWLHD-GVPCLEIGRIHKHSCTLLDLDDIGLQIYA-QQLPHAHTHTGAASCSSTIVSGFFLGGDG
	TC102544_s_ORF_341_195_87	MCLHVEELPCEGLGRVAHHIDSLPALDDLAAQEYT-HLLLLVLPHNGVRCGGSTVFLDHHLGGDG
		* ** :** :**: . *.* . * : : : . * *. **:. ...*****
AK072868 uORF3	AK072868_r_ORF_269_39_216	MLESYLVR-ELAG
	TC247418_w_ORF_278_39_234	MLESYLAR-ESAG
	TC139536_b_ORF_318_39_240	MLESYLAR-ESAG
	TC306591_m_ORF_768_39_228	-MRLWLPKPRYIL
	TC102544_s_ORF_351_39_233	MLKSYLVR-DLAG
		:. :* :
AK072868 uORF5	AK072868_r_ORF_392_36_96	-MGFDVATQPSS
	TC247418_w_ORF_404_36_111	-MGFDVASQPSS
	TC139536_b_ORF_444_36_117	-MGFDVASQPSS
	TC306591_m_ORF_633_36_366	MLRLQKALLSR-
	TC102544_s_ORF_474_36_113	-MGFDVAAQPSS
		: :: * .
AK100589 uORF1	AK100589_r_ORF_240_9_334	MY
	TC264559_w_ORF_201_9_317	MC
	TC130707_b_ORF_228_9_318	MF
	TC292591_m_ORF_286_9_320	MY
	TC91317_s_ORF_260_9_329	MY
		*
AK100589 uORF3	AK100589_r_ORF_296_108_179	-MYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
	TC264559_w_ORF_254_105_168	-MYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
	TC130707_b_ORF_281_105_169	-MYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
	TC292591_m_ORF_336_111_168	MMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
	TC91317_s_ORF_310_111_177	MMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
		******************.*** ************
AK073303 uORF1	AK073303_r_ORF_67_9_142	MP
	TC237149_w_ORF_75_9_113	MP
	TC132556_b_ORF_81_9_139	MP
	TC305609_m_ORF_127_9_69	MI
	TC102988_s_ORF_222_9_69	MI
		*
AK073303 uORF2	AK073303_r_ORF_135_9_74	MI
	TC237149_w_ORF_75_9_113	MP
	TC132556_b_ORF_81_9_139	MP
	TC305609_m_ORF_127_9_69	MI
	TC102988_s_ORF_222_9_69	MI
		*
	Upstream open reading frames conserved in 5/5 cereals and in Arabidopsis	
AK103391 uORF2	AK103391_r_ORF_205_75_74	MNCLHTCSDKKTLKKWFFIDKTVG
	TC269775_w_ORF_251_75_62	MNFLHTCSDKKTLKKWFFIDKTVG
	TC134190_b_ORF_204_75_62	MNFHHTCSDKKTLKKWFFIDKTVG
	TC294011_m_ORF_215_75_75	MNCLHTCGDKKTLKKWFFIDKTVG
	TC103599_s_ORF_106_75_378	MNCLHTCSDKKTLKKWFFIDKTVG
		** ***.****************
AK103391 uORF2	AK103391_r_ORF_205_75_74	MNCLHTCSDKKTLKKWFFIDKTVG
	AT4G22590.1_a_ORF_283_75_63	MDSSTTSSDKKTLKRWFFIDKRVG
		*:. *.*******:****** **

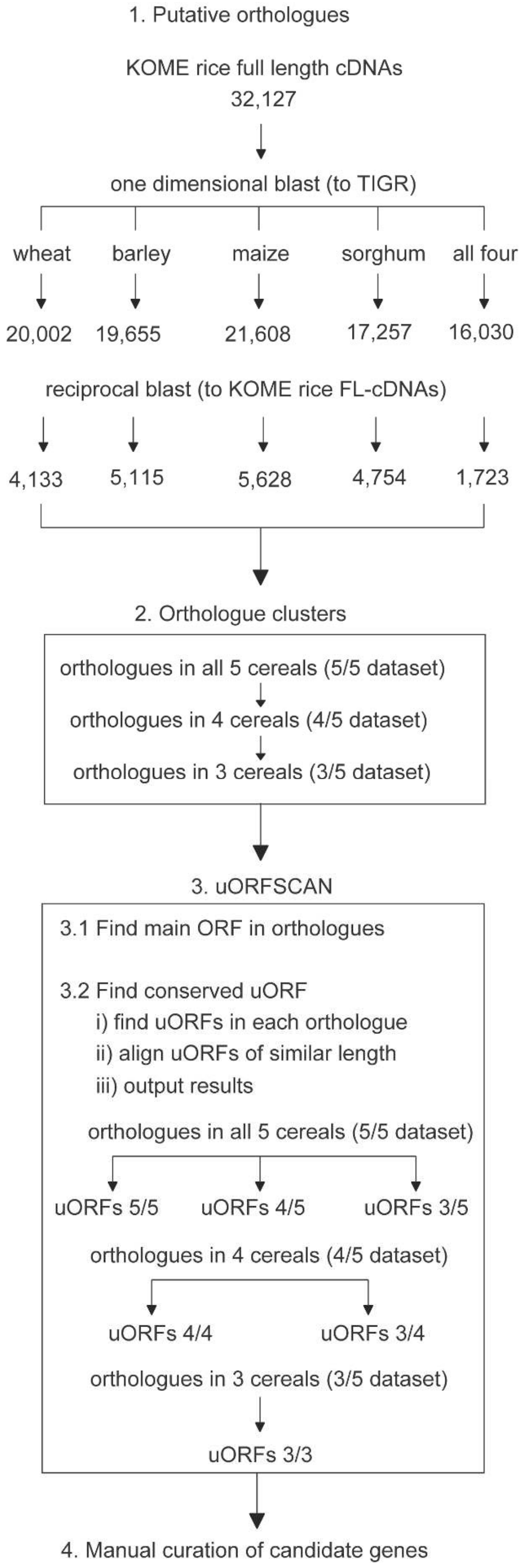
More...

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

Rice identifier	Alignment[a]	
AK100589 uORF2	AK100589_r_ORF_248_156_179	MESKGGKKKSSSSRSLMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
	TC264559_w_ORF_209_150_168	MESKGGKK-SSSSSSLMYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
	TC130707_b_ORF_236_150_169	MESKGGKK-SSSSSSLMYEAPLGYSIEDVRPAGGAKKF-SAAYSNCAKKPS
	TC292591_m_ORF_294_153_168	MESKGGKK-SSSSRSMMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
	TC91317_s_ORF_268_153_177	MESKGGKK-SSSSRSMMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
		******** **** *:*****************.*** ************
AK100589 uORF2	AK100589_r_ORF_248_156_179	MESKGGKKKSSSSRSLMYEAPLGYSIEDVRPAGGVKKFQSAAYSNCAKKPS
	AT3G02470.3_a_ORF_222_156_154	MESKGGKKKSSSSSSLFYEAPLGYSIEDVRPNGGIKKFKSSVYSNCSKRPS
		************* **:************** **:***:*:.****:*:**
AK072868 uORF4	AK072868_r_ORF_338_90_96	MT-LEHKSIYSACSMCSRTMGFDVATQPSS-
	TC247418_w_ORF_347_93_111	MTPLHSSTQHTSSSMCFCTMGFDVASQPSS-
	TC139536_b_ORF_387_93_117	MTPPRSTTQRTSSSMCFCTMGFDVASQPSS-
	TC306591_m_ORF_576_93_366	MRWESYLEKGVLPKFTMLAM-LRLQKALLSR
	TC102544_s_ORF_420_90_113	MT-LQLKSTRIFSYLCFRTMGFDVAAQPSS-
		* : :* : : *
AK072868 uORF4	AK072868_r_ORF_338_90_96	MTLEHKSIYSACSMCSRTMGFDVATQPSS-
	AT5G58380.1_a_ORF_11_87_295	MTFNF--VFISSSSSSSVFSSIFVGKPRKK
		**::. :: :.* .* .:. .. :* .
AK060523 uORF1	AK060523_r_ORF_173_123_185	MVLT-----PSPSPPPMLPKKLRALGPGLNPFAPFGMGNYYSSSR
	TC235416_w_ORF_201_126_157	MVRR-RPSSSSTSSSPMLHKNLRALGPGLNPFAPFGMGNY---SR
	TC148319_b_ORF_211_120_163	MVRR-RPSSS--SSSPMLHKNLRALGPGLNPFAPFGMGNY---SR
	TC305149_m_ORF_255_129_195	MVYAPCRSSTPPSSSPMLHKNLRALGPGLNPLAPFGMGNY---SR
	TC103609_s_ORF_240_129_212	MVYAPCRSSKPPSSSPMLHKNLRALGPGLNPFAPFGMGNY---TR
		** *..*** *:**********:******** :*
AK060523 uORF1	AK060523_r_ORF_173_123_185	-MVLTPSPSPPPMLPKKLRALGPGLNPFAPFGMGNYYSSSR
	AT5G07840.1_a_ORF_289_117_250	MLVFSSLSMTPVVIPQNLRVFGPGLNPSFPYCIANHFP---
		:*::. . .* ::*::**.:****** *: :.*::.
	Upstream open reading frames conserved in rice and in Arabidopsis	
AK103391 uORF1	AK103391_r_ORF_176_30_148	MTSSQVFLC
	AT4G22590.1_a_ORF_254_30_137	MISFQVTYF
		* * **
AK060523 uORF2	AK060523_r_ORF_206_90_185	----MLPKKLRALGPGLNPFAPFGMGNYYSSSR
	AT5G07840.1_a_ORF_313_93_250	MTPVVIPQNLRVFGPGLNPSFPYCIANHFP---
		::*::**.:****** *: :.*::.
AK101100 uORF1	AK101100_r_ORF_142_12_21	MVS
	AT1G51690.1_a_ORF_555_12_1160	MNI
		*
AK066952 uORF1	AK066952_r_ORF_365_66_182	MMKQRLILQMQVIR-LLMNVGT
	AT3G13225.1_a_ORF_364_63_431	-MSWS-ILQLQAFWGLSSGCSS
		*. ***:*.: * . .:
AK066952 uORF2	AK066952_r_ORF_368_63_182	MKQRLILQMQVIR-LLMNVGT
	AT3G13225.1_a_ORF_364_63_431	MSWS-ILQLQAFWGLSSGCSS
		*. ***:*.: * . .:
AK066952 uORF3	AK066952_r_ORF_503_51_59	-MIRSALEILLKKMLLP
	AT3G13225.1_a_ORF_553_51_254	MQYKVSHSYTFSRSYN-
		: : . :.:
AK119592 uORF1	AK119592_r_ORF_304_90_148	-------MKISTRLLWSTSFFRHKIAATIASSSSFL
	AT3G01470.1_a_ORF_162_87_120	MGFCICPLESPARLLWSTSFFRHKIMIF--------
		:: .:*************
AK069534 uORF	AK069534_r_ORF_813_9_432	MI
	AT4G12770.1_a_ORF_41_9_108	ML
		*:
AK069526 uORF1	AK069526_r_ORF_214_126_544	MEYTLYTTSSSVLHISLLEEVLGWRFSLYGDFLVISFVNCT
	AT4G19110.2_a_ORF_255_126_527	MEQVFVWPSCYHYRLFSFQEALDWRFLVRSDFLVGSFVNCT
		** .: .*. :: ::*.*.*** : .**** ******
AK069526 uORF2	AK069526_r_ORF_690_9_185	MA
	AT4G19110.2_a_ORF_603_9_296	ML
		*
AK069526 uORF3	AK069526_r_ORF_820_36_28	MSLVHNRALLE-
	AT4G19110.2_a_ORF_398_36_474	M-IFRGRCEANF
		* :.:.*. :
AK067412 uORF	AK067412_r_ORF_222_84_49	-MRAVVKRRRGGERGRCCGYWRSGASCD
	AT5G50180.1_a_ORF_357_84_79	MLAIYLSLLFSSLSCELSNLHRYKSRK-
		: :. .. . .. * :
AK102277 uORF	AK102277_r_ORF_228_117_150	MHQRLHGWNKSTSMLRDGFGVKYSGFLHIRPCGFCRGD
	AT1G68550.1_a_ORF_309_96_95	MRLRPKRTCSSVEVFG-GFHIKQQKFSFF----IVR--
		*: * : .*..:: ** :* . * .: : *

[a] Identifier_letter_ORF_pre-orf distance_orf length_intercistronic distance
letter: r = rice, w = wheat, b = barley, m = maize, s = sorghum, and a = Arabidopsis

Figure 3.1 Overview of the uORFSCAN pipeline. The pipeline consists of four steps: 1) Identifying putative orthologues using a modified reciprocal best hit (rbh) method, 2) Clustering of orthologues according to how many cereal species they are found in, 3) Using uORFSCAN program to find conserved uORFs using a comparative approach, and 4) Manual curation of predicted conserved cereal and Arabidopsis uORFs.


1. Putative orthologues
KOME rice full length cDNAs
32,127
one dimensional blast (to TIGR)
wheat
barley
maize
sorghum
all four
20,002
19,655
21,608
17,257
16,030
reciprocal blast (to KOME rice FL-cDNAs)
4,133
5,115
5,628
4,754
1,723
2. Orthologue clusters
orthologues in all 5 cereals (5/5 dataset)
orthologues in 4 cereals (4/5 dataset)
orthologues in 3 cereals (3/5 dataset)
3. uORFSCAN
3.1 Find main ORF in orthologues
3.2 Find conserved uORF
i) find uORFs in each orthologue
ii) align uORFs of similar length
iii) output results
orthologues in all 5 cereals (5/5 dataset)
uORFs 5/5
uORFs 4/5
uORFs 3/5
orthologues in 4 cereals (4/5 dataset)
uORFs 4/4
uORFs 3/4
orthologues in 3 cereals (3/5 dataset)
uORFs 3/3
4. Manual curation of candidate genes

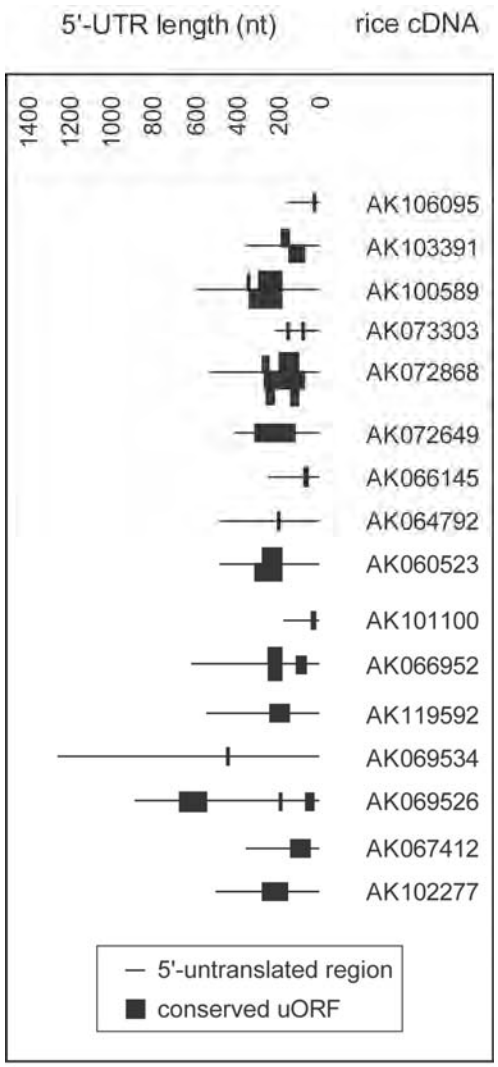


Figure 3.2 The position of uORFs conserved in four other cereals and in Arabidopsis within 5′-UTRs of rice cDNAs.

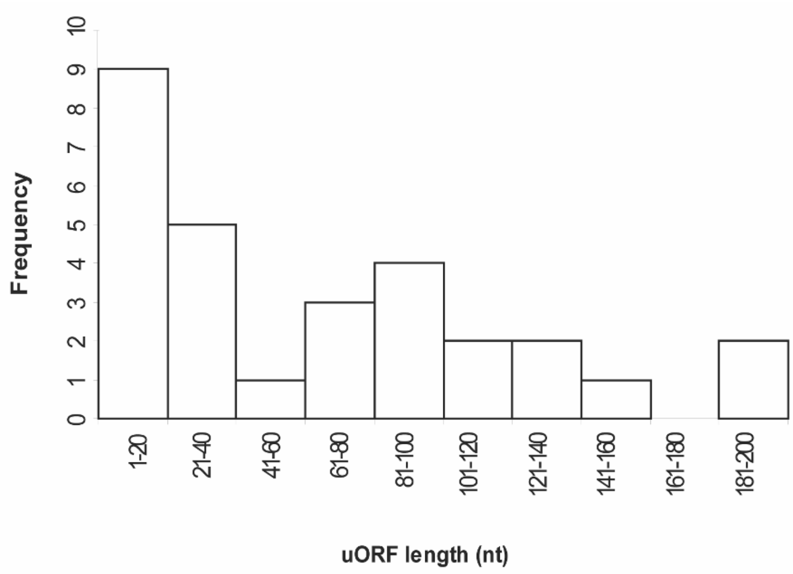


Figure 3.3 Frequency distribution of the length (nt) of rice uORFs conserved in four other cereals and in Arabidopsis.

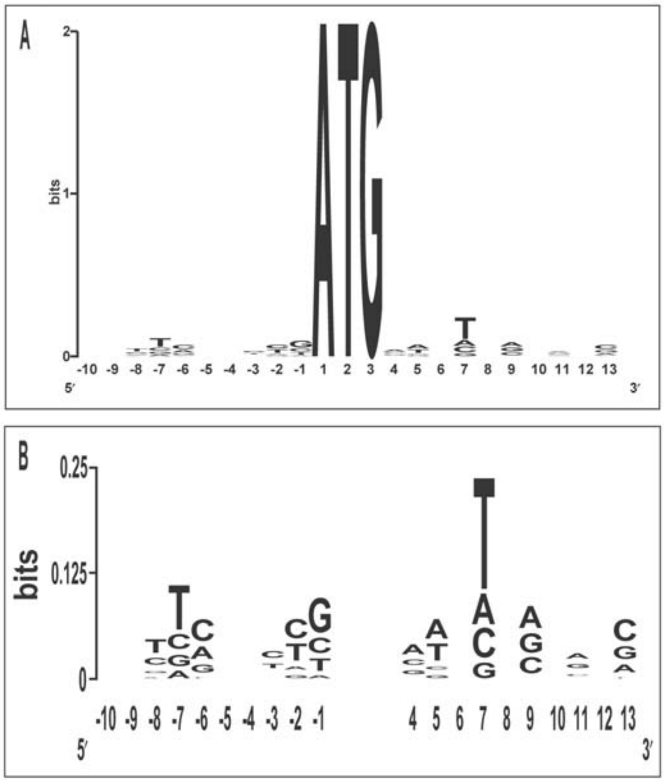


Figure 3.4 The pattern of nucleotide sequence conservation calculated for the decanucleotide surrounding the uORF AUG triplet using WebLogo (Crooks et al. 2004). The overall height of each stack indicates the nucleotide sequence conservation at that position (measured in bits), whereas the height of nucleotide symbols (A, T, G, C) within the stack reflects the relative frequency of the corresponding nucleotide at that position. (B) Positions showing detectable nucleotide sequence conservation were magnified.

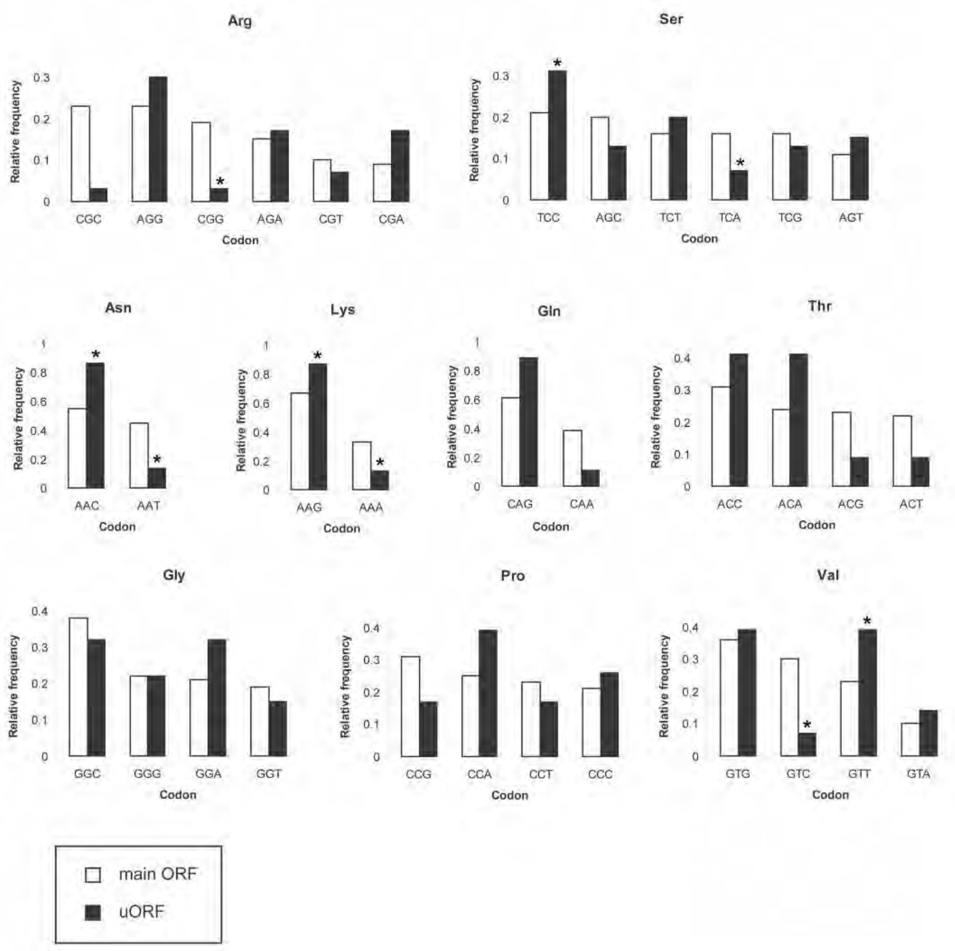


Figure 3.5 Relative frequencies of codons showing significant deviation (*) in codon usage between rice uORFs and rice main coding regions. Rice uORF codon usage calculated from the following URL: http://www.bioinformatics.vg/sms/codon_usage.html.