



142
722
THS

This is to certify that the
thesis entitled

Experimental verification of translational potential of
computationally predicted small protein-coding regions in

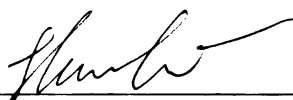
Arabidopsis thaliana

presented by

Shan Yin

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Plant Biology



Major Professor's Signature

8/23/10

Date

MSU is an Affirmative Action/Equal Opportunity Employer

LIBRARY
Michigan State
University

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

| DATE DUE | DATE DUE | DATE DUE |
|----------|----------|----------|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

**Experimental verification of translational potential of
computationally predicted small protein-coding regions
in *Arabidopsis thaliana***

By

Shan Yin

A THESIS

Submitted to
Michigan State University
In partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Plant Biology

2010

ABSTRACT

Experimental verification of translational potential of computationally predicted small protein-coding regions in *Arabidopsis thaliana*

By

Shan Yin

Protein coding genes with small open reading frames (sORFs, < 100 amino acids) have been found to be involved in diverse biological processes. However, sORFs tend to be missed by both genetic screening and annotation efforts. Based on a simple measure of nucleotide composition bias, ~1000 “intergenic” sORFs in *Arabidopsis thaliana* are predicted to be protein-coding. However, there is little direct experimental evidence to validate the translation of predicted protein-coding sORFs. Here, I report the experimental verification of sORF translational potential. We identified 577 sORFs that are expressed and potentially translated in 7-day old seedlings and an additional 441 sORFs that are possibly small (<300 base pairs) non-coding RNA genes based on RNA-seq evidence. Rapid Amplification of cDNA End (RACE) of 38 selected sORFs was performed to obtain the 5' untranslated region (UTR) and translation of 6 out of 15 sORF candidates was verified using a tobacco transient expression system. Importantly, some sORFs that overlap with annotated RNA genes turn out to be protein-coding, indicating the effectiveness of computational tools in coding region prediction. This experimental study demonstrates the translational potential of computationally predicted small protein coding regions, and provides information that will help improving current annotation and gene prediction algorithms.

TABLE OF CONTENTS

| | |
|--|----|
| LIST OF TABLES..... | iv |
| LIST OF FIGURES..... | v |
| INTRODUCTION | 1 |
| 1. Small open reading frames (sORFs) in diverse species..... | 1 |
| 2. sORFs may play important roles in plant growth and development..... | 2 |
| 3. sORFs are missed by gene finders..... | 4 |
| 4. Major goal of this study..... | 6 |
| OBJECTIVES..... | 7 |
| 1. Verification of transcription and translation of sORF candidates in <i>Arabidopsis thaliana</i> seedlings..... | 7 |
| 1.1 Methods..... | 7 |
| 1.1.1 Polysome associated mRNA isolation..... | 7 |
| 1.1.2 RNA-seq..... | 8 |
| 1.1.3 sORF selection and Rapid Amplification of cDNA Ends (RACE)..... | 9 |
| 1.2 Results and discussion..... | 10 |
| 1.2.1 Transcriptional and translational evidence obtained by polysome associated mRNA isolation and RNA-seq..... | 11 |
| 1.2.2 Other transcriptional and translational evidence of sORF candidates and the results of RACE..... | 12 |
| 2. Verification of translational potential of sORF candidates..... | 13 |
| 2.1 Methods..... | 14 |
| 2.1.1 Translational fusion construction..... | 14 |
| 2.1.2 Translation of sORFs in tobacco..... | 16 |
| 2.1.3 Significance of different transcriptional/translational evidence in predicting sORF translational potential..... | 17 |
| 2.2 Results and discussion..... | 18 |
| 3. Preliminary functional study of sORF candidates..... | 20 |
| 3.1 Methods..... | 20 |
| 3.11 Selection of T-DNA insertion lines and phenotype observation..... | 20 |
| 3.12 Homozygote identification by PCR..... | 21 |
| 3.2 Results and discussion..... | 21 |
| CONCLUSIONS..... | 22 |
| REFERENCES..... | 44 |

LIST OF TABLES

| | |
|--|-----------|
| Table 1. RNA-seq evidence of sORF candidates..... | 36 |
| Table 2. RPKM values of sORF candidates..... | 36 |
| Table 3. Properties of sORF candidates that are tested by RACE..... | 37 |
| Table 4. Properties of sORF candidates with 5'UTRs that were tested in tobacco transient expression system..... | 42 |
| <u>Table 5. Fisher's exact test ($\alpha = 0.05$).....</u> | <u>43</u> |

LIST OF FIGURES

Images in this thesis are presented in color.

| | |
|--|----|
| Figure 1. Example of confirmation of sORF candidates on TAIR..... | 25 |
| Figure 2. RACE results of sORF candidates..... | 27 |
| Figure 3. Fusion construction to verify translation of sORF candidates..... | 30 |
| Figure 4. Confocal microscopy images of live leaf epidermal cells of tobacco showing transient expression of sORF candidates..... | 32 |
| Figure 5. List of SALK lines studied..... | 34 |

Introduction

1. Small Open Reading Frames (sORFs) in diverse species.

Small open reading frames (sORFs) are defined as sequences typically less than 300 base pairs and can theoretically be translated into small proteins (Basrai et al, 1997). sORFs have been identified in many species including *Saccharomyces cerevisiae* (Kastenmayer et al, 2006), *Hydra* (Bosch & Fujisawa, 2001), mouse (Frith et al, 2006), human (Oyama et al, 2007), and *Arabidopsis thaliana* (Hashimoto et al, 2008). In addition, they were shown to be involved in diverse biological processes (Hashimoto et al, 2008; Matsubayashi et al, 2006; Kastenmayer et al, 2006).

In *Saccharomyces cerevisiae*, 299 sORFs were identified by either experimental approaches or computational analysis (Kastenmayer et al, 2006). Phenotypes of gene-deletion mutants for 247 annotated sORFs have been analyzed and 22 sORFs were found to be important for survival of *S. cerevisiae* in diverse conditions such as at high temperature and in the presence of DNA damage and replication-arrest agents. Moreover, many yeast sORFs are conserved among eukaryotes (Kastenmayer et al, 2006), suggesting that sORFs are functional among diverse species.

Small protein-coding genes have also been found in animals. More than 800 small peptides were found in *Hydra*, and were predicted to be involved in muscle contraction and neuronal differentiation (Bosch & Fujisawa, 2001). In

mouse, genetic studies have led to the discovery of ~3,700 small protein-coding sORFs (Frith et al, 2006). Several small protein-coding genes, such as *polished rice* (*prn*) (Galindo et al, 2007; Kondo et al, 2007) and Bearded (Leviten et al, 1997), have also been found in *Drosophila*. *Polished rice* was first assumed to be a non-coding RNA required for the formation of epithelial architectures. Further study revealed that it is actually a polycistronic small peptide gene encoding at least four independent and highly related small peptides (Galindo et al, 2007; Kondo et al, 2007). Bearded encodes an 81 amino acid protein and is involved in lateral inhibition (Leviten et al, 1997; Chanet et al, 2009). Aside from sORFs in animals and fungi, they are found in plants which are detailed in the next section.

In addition to sORFs that appear to be protein coding genes, some sORFs are located in the 5' UTR of protein-coding genes. They are named as upstream ORFs (uORFs) and are involved in translational regulation (Lovett et al, 1996; Vilela et al, 2003). Proteomic analysis in human K562 cells has confirmed translation of uORFs in human (Oyama et al, 2007). Several uORFs have also been found in plant species such as *A. thaliana* and rice and were shown to regulate the translation of main ORFs and are highly conserved among plant species (Locatelli et al, 2001; Rahmani et al, 2009). The confirmation of uORFs in human and plants suggests that uORFs are commonly distributed in many species, and may be a necessity for translational regulation.

2 sORFs may play important roles in plant growth and

development.

Similar to other eukaryotes, many sORFs have been identified and confirmed to play important roles in diverse biological processes including regulation of gene translation (Lovett et al, 1996; Vilela et al, 2003), hormonal homeostasis, and plant growth and development (Hashimoto et al, 2008; Matsubayashi et al, 2006; Oyama et al, 2007). For example in *A. thaliana*, the *POLARIS* (*PLS*) gene was first identified in a mutation line of *A. thaliana* (Topping & Lindsey, 1997). Its transcribed mRNA consists of three small ORFs encoding nine, eight, and 36 amino acid small peptides, respectively. *PLS* is expressed strongly in seedling roots and *p/s* mutant plants have short roots and reduced leaf vascularization. Overexpression of the 36 amino acid peptide complements the root mutant phenotype. Thus, the 36 amino acid peptide is considered functional (Casson et al, 2002). Further studies indicated that this 36 amino acid peptide is involved in hormonal homeostasis (Chilley et al, 2006). Another example is *A. thaliana* *ROTUNDIFOLIA4* (*ROT4*) encoding 53 amino acids and was also identified by forward genetics (Narita et al, 2004). Over-expression of this gene results in round leaves and short floral organs. Although *ROT4* is a member of the novel gene family *ROT FOUR LIKE* (*RTFL*), it can cause unique phenotypes (Narita et al, 2004). sORF genes are also found in plant species other than *A. thaliana*. For example, *Enod40* is a small protein-coding gene (36 aa) in soybean that plays important roles in nodule formation in legumes (Charon et al, 1997), and is widely conserved among plant species (Kouchi et al, 1999).

Although sORFs exist in diverse species and are involved in diverse biological processes, the numbers of sORFs that have been identified through forward genetics, particularly in plants, remain small. Therefore, one major question is whether there are more sORFs in different organisms that remain unidentified.

3 sORFs are missed by gene finders.

Due to their small size (< 100 amino acids), genic regions with sORFs tend to be missed by either genetic screening or computational prediction methods (Basrai et al, 1997). Genetic screening is relatively more challenging for sORF genes compared to longer protein coding genes because the probability of insertions of the transfer DNA (T-DNA) into sORF regions is relatively small. Because there are many random small ORFs in the genome, it is difficult to distinguish functional sORFs from random ones with computational predictions (Basrai et al, 1997). In addition, because the size of sORFs are quite small, the contribution of each codon to coding potential measure is larger than that of larger genes which may skew any measure of protein-coding potential, such as codon adaptation index (CAI) (Basrai et al, 1997).

Given the difficulty in finding protein-coding sORFs by experimental and computational tools, the Coding Index (CI) measure was developed to predict the coding potential of sORFs in *A. thaliana* (Hanada et al, 2007). CI is the average posterior probability that a sORF is protein-coding. This algorithm is based on the

nucleotide composition bias of protein-coding genes widely used in gene finders (Hanada et al, 2007). More than 7000 sORFs in *A. thaliana* had a CI value above the threshold, and were predicted to be small protein-coding genes. Several lines of evidence were used to demonstrate the transcription and/or translation of predicted sORFs including Expression Sequence Tag (EST) and Rapid Amplification of cDNA End (RACE) data, tiling array data (Yamada et al, 2003; Stolc et al, 2005; Hanada et al, 2007), stress tiling array data (Matsui et al, 2008), and peptide evidence (Castellana et al, 2008; Baerenfaller et al, 2008). Also, conservation of predicted sORFs among plant species has been demonstrated (Hanada et al, 2007).

Currently, most known protein-coding sORFs were first annotated through expression-based analyses. Serial Analysis of Gene Expression (SAGE) has been utilized to identify and compare global gene expression patterns in *S. cerevisiae* (Velculescu et al, 1995; Velculescu et al, 1997; Basrai and Hieter 2002). RT-PCR, Northern blotting, microarray, gene-trap and proteomics analysis have also verified transcription and translation of several sORFs in many species including yeast, *A. thaliana*, and human (Olivas et al, 1997; Oshiro et al, 2002; Kumar et al. 2002; Hanada et al, 2007; Oyama et al, 2007). These studies provide supporting evidence for the expression of a number of sORFs. However, most sORFs computationally predicted from yeast (Kastenmayer et al, 2006) and *A. thaliana* (Hanada et al, 2007) are not annotated and do not have experimental evidence for their expression in transcriptomic or proteomic studies. Therefore,

this study is focused on providing experimental evidence for the translation of computationally predicted sORFs.

4 Major goal of this study

Given the involvement of sORFs in multiple biological processes, it is important to obtain experimental evidence of translation of sORFs in *A. thaliana* to facilitate downstream functional studies. Another important reason for assessing the translation of predicted sORFs is that many sORFs are found to overlap with newly annotated RNA genes. For example, one assumed non-coding RNA (ncRNA) gene, *polished rice*, was shown to be a small protein-coding gene (Inagaki et al, 2005; Tupy et al, 2005). Similarly, we cannot rule out the possibility that some sORFs may not code for proteins. sORFs that only have transcriptional evidence instead of translational evidence could be non-coding RNA or just transcriptional noise. The major purpose of this study was to evaluate these possible scenarios by verifying the translation of predicted sORFs as mRNAs that are associated with ribosome are likely translated. Based on this idea, I have assessed the genome-wide patterns of translation in *A. thaliana*. In addition, I have chosen sORFs that have transcriptional and translational evidence and confirmed their translation. Finally, I have conducted a preliminary study of sORF functions using T-DNA insertion lines.

Objectives

The overall goal of this work was to experimentally validate transcription and translation of computationally predicted sORFs and determine expression levels and patterns of sORFs for functional studies. To achieve this goal, the first objective of this study was to experimentally verify the global transcription and translation of computationally predicted sORFs in 7-day old *A. thaliana* seedlings. The second objective was to further validate translation of selected sORF candidates with a transient expression system in tobacco, and determine the expression patterns of sORFs. The third objective was to study the functions of selected sORFs shown to be translated using T-DNA insertion lines.

Objective 1 Verification of global transcription and translation of sORF candidates in *Arabidopsis thaliana* seedlings

Over 7,000 potential sORFs have been predicted (Hanada et al, 2007), and diverse transcriptional and translational evidence has been obtained for those sORF candidates. However, there is no experimental evidence for the translation of most sORF candidates. The first objective is to obtain transcriptional and translational evidence of sORF candidates using RNA-sequencing (RNA-seq).

1.1 Methods

1.1.1 Polysome associated mRNA isolation

Polyribosomal complexes from 7-day old seedlings of *A. thaliana* were immunoprecipitated in Julia Bailey-Serres's lab at the University of California-Riverside. Transgenic *A. thaliana* that overexpressed ribosomal protein (RP) with His6-FLAG dual epitope tag were made (Zanetti, 2005). Anti-FLAG agarose-conjugated beads were used for affinity purification of ribosomal complexes by immunoprecipitation (Zanetti, 2005). Both immunoprecipitated (IP) RNA and total RNA of 7-day old seedlings of *A. thaliana* was isolated.

1.1.2 RNA-seq

RNA-seq is a newly developed transcriptome profiling approach using high-throughput sequencing technology (Wang et al, 2009). RNA-seq was performed in the Research Technology Support Facility (RTSF) at MSU. Poly-A containing mRNA molecules from immunoprecipitated RNA (IP RNA) and total RNA samples of 7-day old seedlings of *A. thaliana* were obtained by using poly-T oligo-attached magnetic beads. Purified mRNA was then fragmented using divalent cations and reverse transcribed into first strand cDNA with random hexamer primers. Second strand cDNA was synthesized by DNA polymerase I and RNaseH. Finally, Illumina Genome Analyzer was utilized to obtain sequence information of the IP RNAs and total RNAs. 36-basepair reads were generated. Reads were mapped to the TAIR 9 version of the *A. thaliana* genome using Bowtie (Langmead et al, 2009). Alignments that have less than 2 mismatches in the 5'-most 5 bases of the read were reported. No more than 10 alignments for a read were allowed to be reported. The Reads per Kilobase of Exon Mapped Per

Million Mapped Reads (RPKM) were calculated using Tophat (Trapnell et al, 2009). The minimum defined intron size was set to be 5 bp, and the maximum intron size was 3000 bp. All the other settings were by default. A sORF was defined as having evidence of translation if it had at least one sequence tag from the total sample and an RPKM of >2.25 in the IP sample.

1.1.3 sORF selection and Rapid Amplification of cDNA Ends (RACE)

sORF candidates were selected for verification of translation based on their transcriptional and translational evidence. First, the top 100 sORFs with the highest numbers of RNA-seq reads from IP RNA sample were selected. Second, transcription and translational evidence, including peptide and EST evidence, for these top 100 sORF candidates were obtained by checking the annotation information from The *Arabidopsis* Information Resource (TAIR) (e.g. Figure 1). Finally, surrounding areas of those sORF candidates in the genome were also examined to determine if a sORF overlaps with newly annotated genes. sORF candidates that do not overlap with an annotated protein-coding gene and have transcriptional (EST evidence, RACE evidence, tiling array data, RNA-seq data) and translational (peptide evidence, RNA-seq data) evidence were analyzed further.

Primers were designed for selected sORF candidates by Primer Premier 5.0 (Premier Biosoft, California). FirstChoice RLM-RACE (Ambion, Texas) was

processed to obtain the 5' full length UTR of sORF candidates. A 5' RACE library was constructed using extracted total RNA from 7-day old seedlings of *A. thaliana*. Calf Intestine Alkaline Phosphatase (CIP) treatment was used to get removed 5'-phosphates from degraded RNA or DNA. Tobacco acid pyrophosphatase (TAP) treatment was used to remove the 5' cap from full-length mRNA. 5' RACE Adapter was added to decap full-length mRNA. Reverse transcription and nested PCR was used to obtain the full-length cDNA. DNA obtained by nested PCR was purified by Qiagen Gel Extraction Kit, and ligated to the pGEM T Easy vector (Promega, Wisconsin). Recombinant plasmids were transformed into *E.coli* (pGEM T EASY cloning protocol). Transformed *E.coli* cells were spread onto plates with ampicillin (100 µg/ml), and incubated overnight at 37°C. Single colonies of transformed *E. coli* were selected and incubated overnight at 37°C. Plasmids were extracted by the alkaline lysis method (Birnboim, 1979). Extracted plasmids were mixed with M13 forward and reverse primers (5'-CGCCAGGGTTTTCCCAGTCACGAC-3', 5'-AGCGGATAACAATTTTCACACAGGA-3', pGEM T EASY cloning protocol) and sterilized water (dH₂O), and were sequenced with an ABI PRISM® 3730 Genetic Analyzer at the MSU RTSF. The sequence of the RACE products was searched against the *A. thaliana* genome sequence using Basic Local Alignment Search Tool (BLAST, Altschul et al, 1990) at National Center for Biotechnology Information (NCBI) website to obtain the 5' UTR of sORF candidates.

1.2 Results and Discussion

1.2.1 Transcriptional and translational evidence obtained by polysome associated mRNA isolation and RNA-seq.

Short reads from the library of total RNA (~7 million) and Polysome Immuno-precipitated (IP) RNA (~6 million) nucleotide sequences were generated with Illumina high-throughput sequencing. RNA-seq data from total RNA provides information on genes that are transcribed in 7-day old seedlings. On the other hand, polysome IP data shows genes that are transcribed and theoretically translated. Among 7,408 predicted sORF candidates, 7.8% (577) have RNA-seq reads in both total RNA and IP RNA (Table 1), indicating their transcription and potential translation in 7-day old seedlings. Among those, 13 sORF candidates have IP RNA-seq RPKM values higher than 400. 80 sORF candidates have RNA-seq RPKM values ranging from 10 to 400. 484 sORF candidates have RNA-seq RPKM values ranging from 0.04 to 10 (Table 2). Given that the calculated mean number of RNA-seq reads for protein-coding genes is 30.1 in IP RNA and 27.1 in total RNA, the expression levels of most sORF candidates are relatively low, consistent with earlier observation based on tiling array studies (Hanada et al. 2007).

441 predicted sORFs have total RNA-seq reads but no IP RNA-seq reads (Table 1), which suggests that they are not ribosome-associated and likely not translated. Alternatively, it is possible that their transcripts did not have polyA tail and thus were not reverse-transcribed under the conditions used in this study. The rest of the predicted sORFs that lack either IP RNA-seq reads or total RNA-

seq reads (Table 1) may be false positives. Other potential explanations could be that they are not expressed in 7-day old seedlings. On the other hand, searching the sequences of sORFs with more than 5,000 RNA-seq RPKM values against *A. thaliana* genome sequences has confirmed that sequences of those sORFs are repetitive in the *A. thaliana* genome. Therefore, the RPKM values for those sORFs may not represent their true expression levels. Excluding these sORFs with >5000 RPKM, the expression levels of most sORFs are low (average 0.517).

1.2.2 Additional transcriptional and translational evidence for 100 sORF candidates and the results of RACE.

Transcriptional and translational evidence for 100 predicted sORFs was obtained from TAIR. Among the top 100 sORF candidates with the most RNA-seq reads ranging from 10.0 to 15848.1 RPKM, there is peptide evidence for 15 of these top 100 sORFs. However, among these 15, nine sORFs have a different direction of translation from the peptide (Castellana et al, 2008; Baerenfaller et al, 2008). All of these 100 sORF candidates are expressed in 7-day old seedlings based on tiling array and RNA-seq data. 37 of these 100 sORFs do not have EST evidence. Eight of those 100 sORFs have ESTs overlapping with another predicted or annotated protein-coding gene or RNA gene. 45 of those 100 sORFs have sequence overlapping with another predicted or annotated protein-coding gene or RNA gene.

By searching for transcriptional and translational evidence of sORF

candidates based on TAIR v.9 annotation, I found that sequences of many sORF candidates overlap with protein-coding genes or RNA genes that were annotated after the prediction of sORFs (based on v.7 annotation). There are three different types of overlaps. The first type of sORF candidates has the same reading frame as annotated protein-coding genes. The second type has a different reading frame from annotated protein-coding genes. The third type overlaps with annotated RNA genes. From these three types, 38 sORF candidates were selected to obtain their full length cDNA (Table 3). Properties of all the tested sORF candidates, including sORF names, transcriptional evidence (RNA-seq, EST) and translational evidence (RNA-seq, peptide), and surrounding areas of location of sORFs in the *A. thaliana* genome, are listed in Table 3. 5' UTRs was obtained for the 14 sORF candidates in the list, (Table 3). In addition, 5' UTRs of eight additional sORF candidates are available from published resources (Moskal et al, 2007). For failed RACE reactions, it is possible the primers designed are not specific. Another possibility is that the expression levels of sORF candidates in 7-day old seedlings are too low to be detected by RACE. Or possibly the transcripts of those sORFs lack polyA tail, and thus were not reverse-transcribed when the RACE library was made.

Objective 2 Verification of translational potential of sORF candidates

RNA-seq data indicates that hundreds of predicted sORFs could be transcribed and translated in 7-day old seedlings of *A. thaliana*. My second

objective was to further experimentally validate translation of sORF candidates in a tobacco transient expression system.

2.1 Methods

2.1.1 Translational fusion construct

Modified plasmid pMDC83 (Invitrogen), in which the Green Fluorescent Protein (GFP) gene has been replaced by the Yellow Fluorescent Protein (YFP) gene, was utilized for the fusion construct. The 5' UTR and predicted coding sequence (CDS) of each sORF candidate was PCR amplified, fused with an YFP reporter gene (with start codon deleted), and TOPO TA Cloning® (Invitrogen, California) was performed. The transcription of the sORF candidate and YFP fusion was driven by a cauliflower mosaic virus (CaMV) 35S promoter (Figure 3A). Therefore, if the sORF candidate can be translated in frame, the YFP reporter gene can also be translated.

To test the validity of the fusion expression system, positive control and negative controls were included. As a positive control, AT5G45420, an ER membrane protein (Slabaugh, Held, Brandizzi, unpublished), was fused with the YFP reporter gene. Two negative controls were used including a known non-coding RNA gene (AT1G12013) with and without a start codon in the construct (Figure 3B). The negative controls allow us to see if a RNA gene or a fragment of random DNA allows in frame translation of the YFP reporter gene or not.

After the 5' UTR of sORF candidates were obtained, primers were

designed with Primer Premier 5.0 (Premier Biosoft), and PCR was performed to obtain the 5' UTR and CDS of sORF candidates. DNA obtained by PCR was purified by the Qiagen Gel Extraction Kit. Insertion of 5' UTR and CDS of sORF candidates (without stop codon) into TOPO vector was done by the TOPO cloning kit to insert the 5'UTR and CDS of sORF candidate into Invitrogen pTOPO plasmid (Figure 3A). Recombinant plasmid was transformed into *E. coli* DH5a (TOPO cloning protocol). Transformed *E. coli* cells carrying recombinant plasmids were grown and recovered at 37°C for one hour and were spread on LB plate with spectinomycin (100 µg/ml) and grew overnight.

Single colonies of transformed *E. coli* were picked and grown overnight at 37°C. PCR using the M13 forward primer (TOPO cloning protocol) and sORF specific reverse primers was performed to confirm the insertion. Transformed *E. coli* colonies verified by PCR were selected for sequencing. Plasmid DNA was isolated using the alkaline lysis method to extract recombinant plasmids from transformed *E. coli*. Extracted plasmids were sequenced using the GW1/GW2 primers 5'-GTTGCAACAAATTGATGAGCAATGC-3', 5'-GTTGCAACAAATTGATGAGCAATTA-3' (TOPO cloning protocol) at RTSF.

Sequences of the recombinant plasmids were searched against the *A. thaliana* genome sequence available on TAIR and compared with the TOPO plasmid sequence. Constructed TOPO plasmid with the expected direction of sORF insertion was selected. Modified pMDC183, in which the YFP reporter gene has replaced GFP reporter gene, was the destination vector. Constructed

TOPO plasmids were mixed with the destination vector and LR recombinant reaction was done by the Invitrogen gateway cloning kit. Destination vectors carrying 5'UTR and CDS of sORF candidates were transformed into *E. coli* with the same transformation method mentioned above. Transformed *E. coli* cells were spread on the LB plate with kanamycin (50 µg/ml), and grew overnight.

Single colonies of transformed *E. coli* were picked and cultivated overnight at 37°C. PCR with sORF specific primers was done to test the insertion of 5'UTR and CDS of sORF candidates. Transformed *E. coli* colonies that were verified by PCR were selected for sequencing. Isolation of plasmids was done by the alkaline lysis method to extract the constructed plasmids. Extracted plasmids were sequenced using GW1/GW2 primers (Invitrogen) at the MSU RTSF.

Sequencing data was searched against *A. thaliana* genome sequence available in TAIR and was compared with the TOPO plasmid and destination vector sequences (Invitrogen). Constructed destination vectors with correctly inserted sORF sequence were selected for downstream analyses.

2.1.2 Translation of sORFs in tobacco

Constructed destination vectors carrying the 5'UTR and CDS of sORF candidates (Table 4) were transformed into *Agrobacterium tumefaciens* strain GV3101 using the Freeze-Thaw Method (Weigel, 2002). Transformed *Agrobacterium* broth was spread on plates with gentamycin (15 µg/ml),

kanamycin (50 µg/ml) and rifampicin (30 µg/ml), and was incubated at room temperature overnight. Single colonies of transformed *Agrobacterium* were picked up and cultivated overnight at 28°C. PCR with sORF specific primers was done to test the efficiency of transformation of constructed plasmids.

Transient transformation was performed to express sORF-YFP in tobacco (*Nicotiana tabacum*) cells (Sparkes et al, 2006). Agro broth was cultivated overnight, and 200ul of the culture (Optical Density or OD ~1-2, A600) was taken and microcentrifuged at 8000g for one minute. Supernatant was removed, and the pellet was resuspended with 1ml sterilized water. OD of the suspended Agro cells was measured at A600. Dilution or further growth of the Agro culture was done to make the OD about 0.1. Agro cells were infiltrated into tobacco leaves, and the infiltrated tobacco was kept under light for 72 hours. After that, the transformed areas of tobacco leaves were detached, and observed under the inverted laser scanning confocal microscope (Olympus Spectral FV 1000). YFP signals were detected with a 514nm excitation line of argon laser and 500ms exposure time.

2.1.3 Significance of different transcriptional/translational evidence in predicting sORF translational potential

There are diverse transcriptional and translational evidence available for sORF candidates. To determine which one(s) play significant roles in predicting protein-coding potentials of sORF candidates and thus optimize the selection criteria of sORF candidates for further analysis, Fisher's exact test was

performed on the results of sORF translation in tobacco and available transcriptional and translational evidence of tested sORFs (Table 4).

Fisher's exact test is a statistical significance test which is used in the analysis of contingency tables where sample sizes are small (Fisher, 1922). The significance of deviation from the null hypothesis can be calculated exactly. This test is used to check the significance of association of two classifications. In this study, Fisher's exact test was performed to determine the significance of relationships between tobacco transient expression and transcriptional/translational evidences of sORF candidates.

2.2 Results and Discussion

Fusion constructs and tobacco infiltration for 15 sORF candidates were carried out. YFP translation was detected for the positive control and not for the negative controls (Figure 4 A-D). Six fusion proteins had positive YFP signals comparing to controls (Figure 4 G-R). Nine sORF-YFP fusion proteins were not detected (Table 4). Among the 15 sORF candidates tested, seven had RNA-seq evidence, three had peptide evidence, nine were predicted to be experiencing purifying selection, and 11 had EST evidence (Table 4).

To determine whether there is a relationship between the translational potential of tested sORF candidates and a specific type of transcriptional/translational evidence. Fisher's exact tests were performed. Table 5 lists all the tested pair-wise categorical variables, with two levels each (Yes/No).

The null hypothesis is that relative proportions of one variable are independent from those of the other variable. For example, if we want to determine if one sORF candidate with EST evidence has a higher probability to be translated, the null hypothesis is that the proportion of sORF candidates with EST evidence are not significantly different between those that are translated and those that are not. The results of Fisher's exact test (Table 5), indicate that the proportions of sORF candidates with EST evidence/peptide evidence/purifying selection evidence are not significantly different between those that are translated and those that are not. By contrast, the proportions of sORF candidates that have RNA-seq evidence are not the same between those that are translated and those that are not.

My findings suggest that if sORF candidates have RNA-seq reads, and RACE products can be obtained, it is highly probable that those sORF candidates are translated in tobacco. The sORF candidates that have positive translation results in tobacco almost have the same ratio of EST evidence or conservation evidence as those that show negative results. Thus, it appears that RNA-seq and RACE are two important elements for selection of sORF candidates. However, the sORFs that were not translated in tobacco still could possibly be protein-coding genes. The reasons why they were not translated in tobacco could be that they are not expressed in 7-day old seedlings, or they can be expressed in *A. thaliana* but not in tobacco. Also, it is possible that they are expressed late in tobacco, and more than 72 hours will be needed to observe the YFP signal. Or possibly they need other unknown elements to initiate their

transcription. Moreover, maybe the sORFs function in tissue specific regulation.

3 Preliminary functional studies of sORF candidates

3.1 Method

3.11 Selection of T-DNA insertion lines and phenotype observation

The protein coding potential of six sORFs was validated by RNA-seq and translation in tobacco but their functions remain elusive. Therefore, it is of interest to further study the possible functions of sORF candidates. sORFs were selected for functional characterization based on several criteria: First, sORF candidates that were expressed in tobacco and showed positive results were preferentially selected. Second, sORF candidates that have RNA-seq reads and are conserved between *A. thaliana* and *Arabidopsis lyrata* were chosen as a priority. Third, sORF candidates that have related sequences in the genome and have a large number of RNA-seq reads were excluded.

T-DNA insertion lines for selected sORF candidates were identified from TAIR. SALK lines (Alonso et al, 2003) that have T-DNA inserted in the coding sequence (CDS) of sORF candidates were selected (Figure 5). Eight seeds from each SALK line and wild type *A. thaliana* were planted. Seeds were sterilized (Chatfield, 2005), stratified in 4°C for two days to break dormancy, and were planted into soil. Wild type *A. thaliana* were planted close to different SALK lines to provide a similar growth environment between SALK lines and wild type

controls.

Morphological characteristics of SALK lines, including the number, shape, color, size and appearance of leaves, the height of plants, and the number, shape, color, size and appearance of flowers were closely monitored until seeds were mature and collected.

3.12 Homozygote identification by PCR

PCR using the T-DNA specific primer and the target gene specific primer as well as PCR using a pair of target gene specific primers was done to check if the SALK lines were homozygous or not. If a tested SALK line was homozygous, PCR using the T-DNA specific primer and the target gene specific primer could amplify DNA fragments, but PCR using a pair of target gene specific primers might not be able to amplify DNA fragments. If a tested SALK line was heterozygous, PCR using the T-DNA specific primer and the target gene specific primer could amplify DNA fragments, and PCR using a pair of target gene specific primers could amplify DNA fragments.

3.2 Results and Discussion

Seven SALK lines (Figure 5) of six sORF candidates were selected out of 50 sORFs that have RNA-seq reads and are conserved between *A. thaliana* and *A. lyrata*. T-DNA was inserted in the CDS regions of sORFs in all 6 lines. Two plants from one SALK line for one sORF candidate appeared shorter and smaller

than the other 6 plants of the same SALK line and wild type. However, more replicates would be necessary to establish the statistical significance of the difference. No other dramatic difference in leaves, flowers and height of plants were observed. PCR results show that some plants of this SALK line are homozygous, but they did not show a short phenotype. Therefore, it is not clear if the phenotype results from the insertion of T-DNA in the sORF CDS region.

The absence of dramatic phenotypes of sORF SALK lines may be because of four reasons: First, the selected sORF candidates do not play significant roles in influencing physical appearance of *A. thaliana* under normal growing conditions. Second, the difference in phenotype may be subtle and could not be detected with the phenotyping performed in this study. Third, it is possible that the selected sORF candidates belong to gene families or are functionally redundant with other genes. Therefore, knockout of those sORFs may not result in obvious phenotypes. Fourth, the sORFs could still be expressed despite the T-DNA insertions due to read through transcription.

Conclusion

My study shows that RNA-seq is effective in providing transcriptional and translational evidence of sORF candidates. It not only uncovered hundreds of sORF candidates that are possibly translated, but also provided information on sORF candidates that are potential RNA genes. Based on RNA-seq data, expression levels of sORF candidates is substantially lower than longer protein

coding genes. It is possible that only a small amount of sORFs are needed for their functions. In addition, these sORFs may express in a tissue or cell type specific manner and the significantly lower level of expression can be a consequence of the fact that we only look at the whole seedling. Future experiments are needed to resolve these possibilities.

Many sORF candidates overlapped with newly annotated protein-coding genes or RNA genes. Some sORFs that overlapped with annotated RNA genes turned out to be protein-coding. This indicates the effectiveness of our computational tool in predicting protein-coding regions. In addition to protein coding genes misannotated as RNA genes, another possible explanation for the sequence overlap between sORFs and newly annotated genes could be that they are nested genes. Nested genes have been confirmed in many organisms, including human, yeast and *Drosophila* (Kumar, 2009). It is also possible explanation could be that neither gene model is correct. Further studies are necessary to examine these alternative explanations.

According to the results of Fisher's exact test, RNA-seq data is the most important evidence in predicting coding potential of sORF candidates. If a sORF candidate has RNA-seq reads and RACE products can be obtained for this sORF, it is highly possible that the sORF will be translated in the transient expression test in tobacco. This indicates the effectiveness of RNA-seq in validating transcription and translation of sORFs. What we may do next could be to select

more sORF candidates primarily based on RNA-seq data and try RACE to obtain their 5' UTR. By doing this, more sORF candidates could be tested by transient expression in tobacco, and more sORF candidates could be available for functional study.

No obvious phenotypes were observed for T-DNA mutation lines of five sORF candidates. The lack of dramatic phenotypes may be due to subtle changes that have not been detected by us or possible functional overlaps with other genes. Or maybe the sORFs still could be expressed despite the T-DNA insertions. Also, the sORFs may function only under specific environments instead of condition tested. Therefore, it will be necessary to determine if sORF belongs to a gene family. The expression of sORFs in SALK lines should be examined using RT-PCR. Finally, the SALK lines should be re-examined under alternative growth conditions such as abiotic stress.

In summary, this study has validated transcription and translation of hundreds of computationally predicted sORFs in *A. thaliana* for further functional study in *A. thaliana* and other species. Moreover, expression levels and expression patterns of several sORFs have been studied, which will also be helpful for functional study of these sORFs. This study can potentially provide background knowledge and help determine what these sORFs really do in diverse species and how to better identify them.

Figure 1. Example of confirmation of sORF candidates on TAIR

Transcriptional and translational evidence of sORF candidates were checked against TAIR annotation v.9. The top track shows where in the *A. thaliana* genome the sORF candidate is located. Locus, Protein Coding Gene Models, and non-coding RNAs lines indicate v.9 annotated genes. Hanada et al, 2007 Gene Models show the model of sORF candidates. *A. thaliana* cDNAs are transcriptional evidence for sORF candidates. AtProteome and AtPeptide show the available peptides as translational evidence. T-DNAs/transposon line refers to the available insertion lines which are useful information for further functional study of sORF candidates.

Figure 1. Example of confirmation of sORF candidates on TAIR

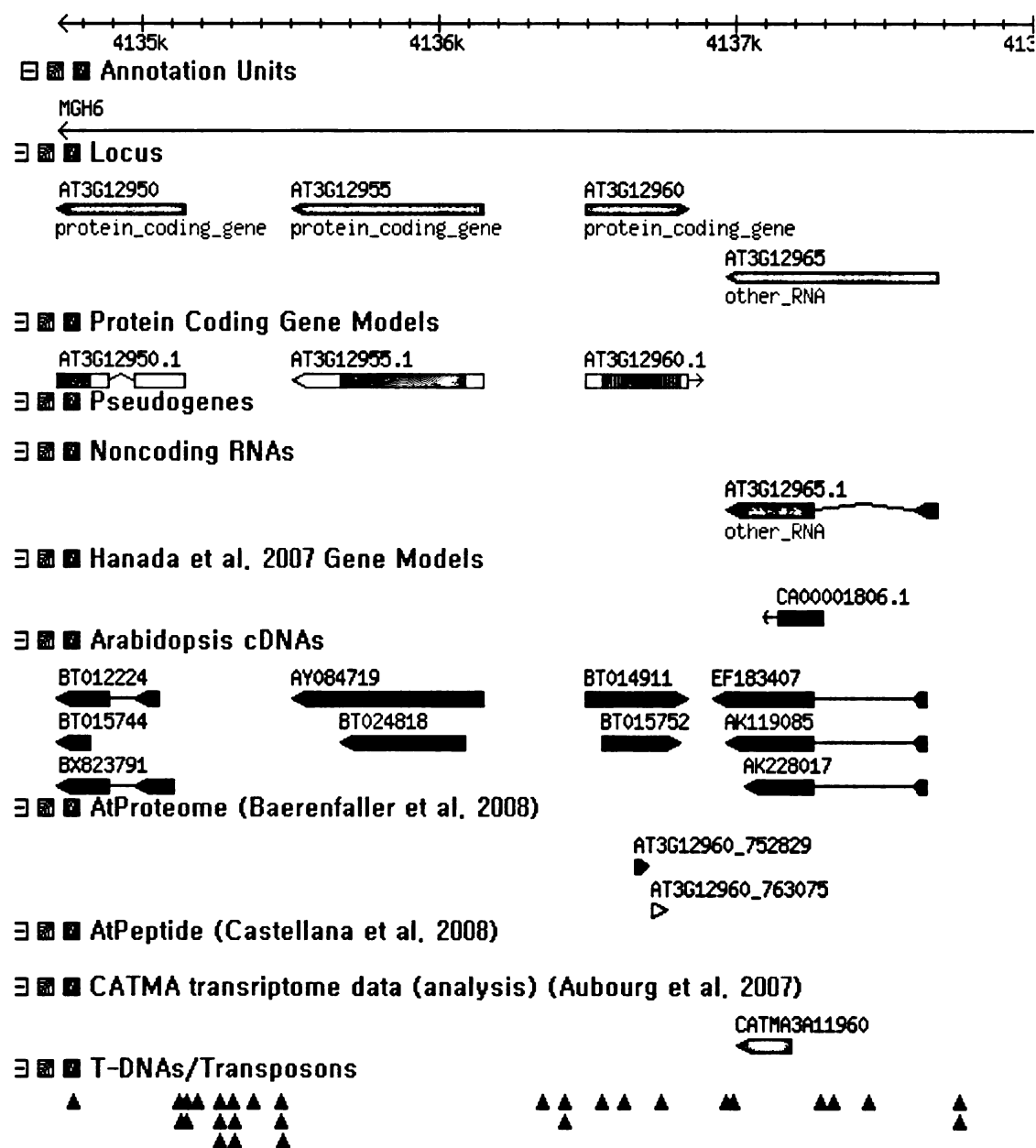


Figure 2. RACE results of sORF candidates.

Outer and nested PCR products for each sORF candidate were loaded in two neighboring lanes. For each sORF, the left lane was loaded with PCR product that was produced by 5' RACE outer primer, and the right lane was loaded with PCR product which was produced by 5' RACE inner primer. If the nested PCR product contains more than one DNA fragment with different sizes, all the DNA fragments obtained were purified, cloned, and sequenced to see which one matches the target sORF candidate. If more than one DNA fragment matches the same sORF candidate, the longer DNA fragment will be chosen for further cloning.

Figure 2. RACE results of sORF candidates.

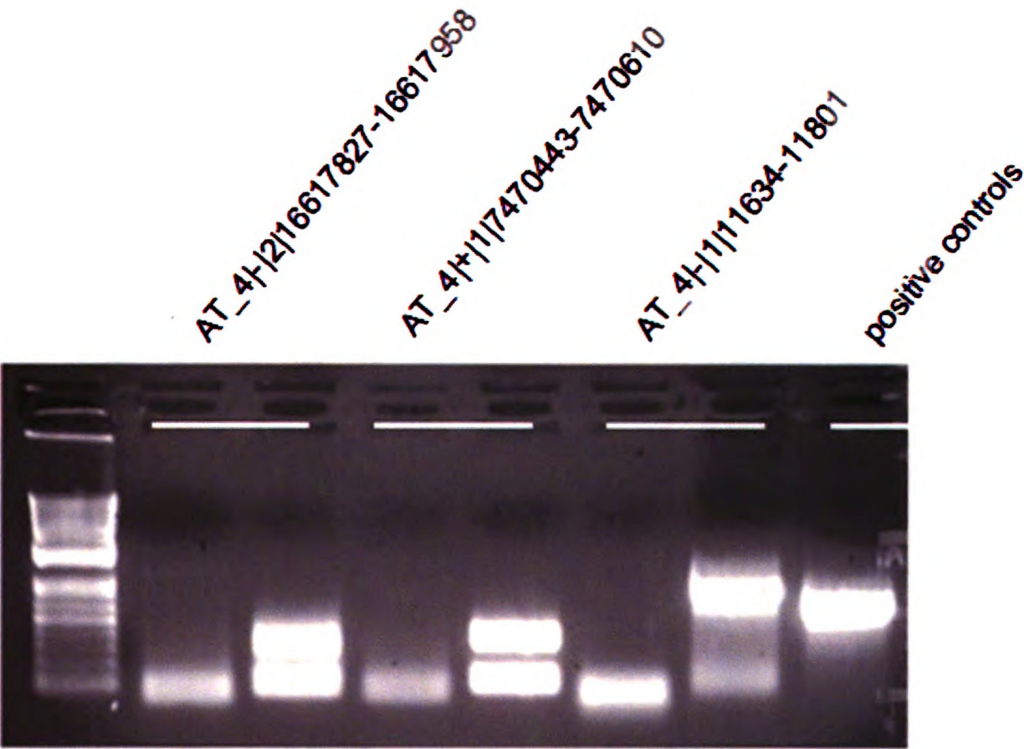
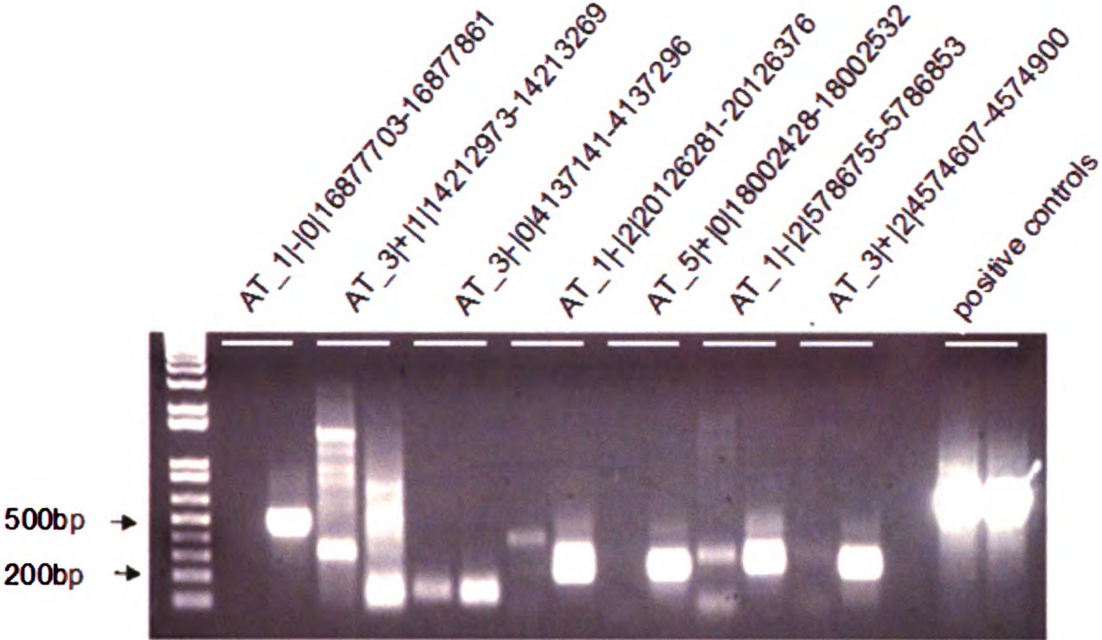


Figure 2 continued.

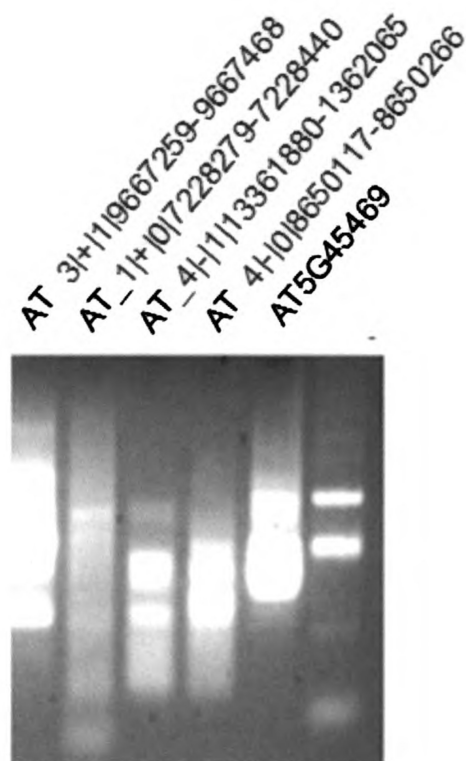
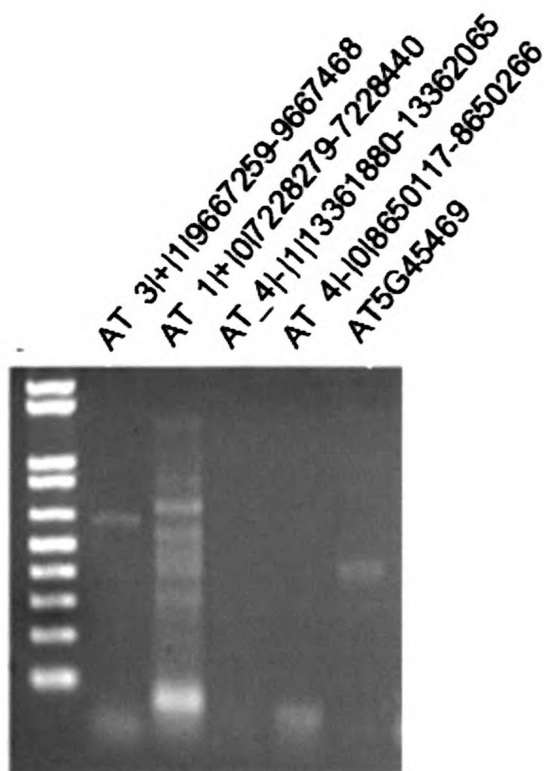
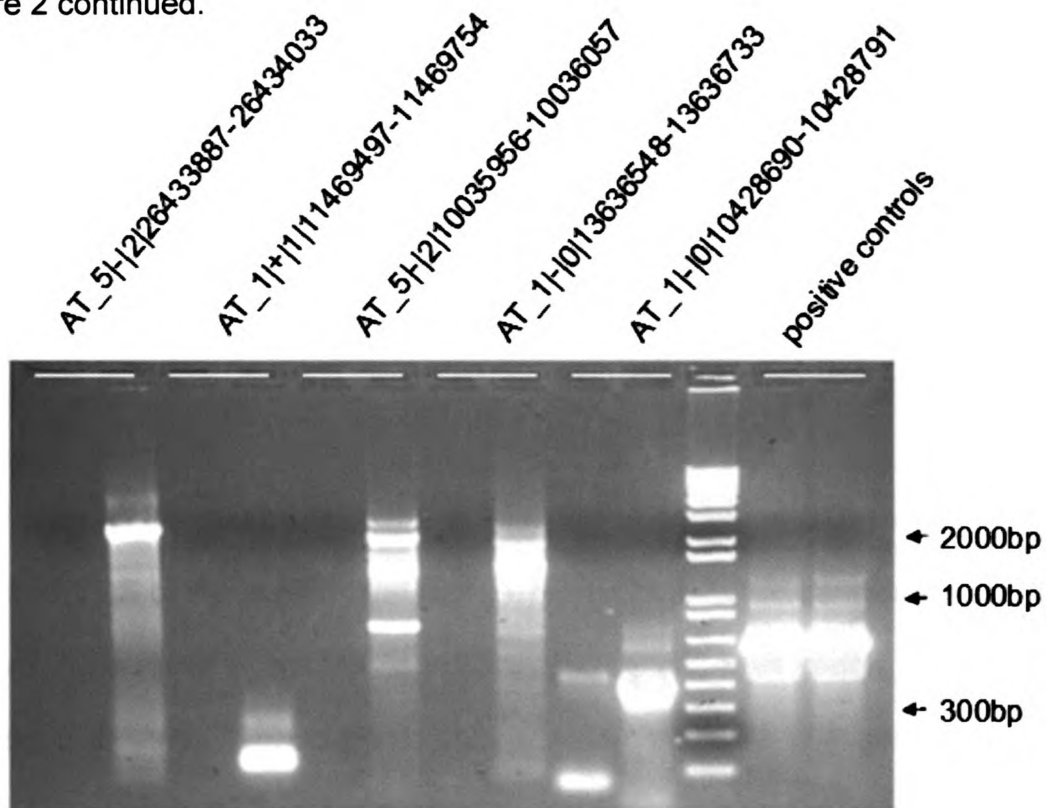
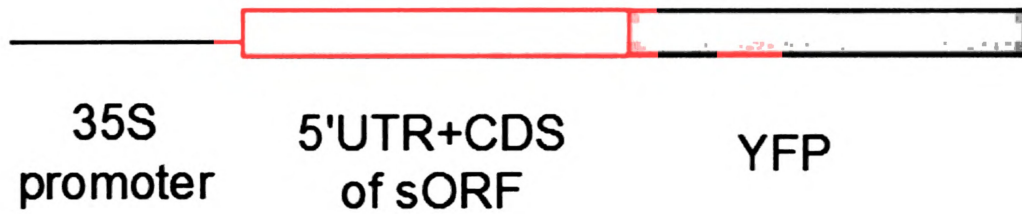


Figure 3. Fusion construction to verify translation of sORF candidates.

A. 5' UTR and CDS of the sORF candidate (stop codon of sORF candidate has been deleted) was fused with CDS of YFP reporter gene (start codon of YFP gene has been deleted). Transcription of sORF candidate and YFP gene was driven by 35S promoter. **B.** The first negative control of this construct was to use sequence of well-known RNA gene to fuse CDS of YFP reporter gene (start codon of YFP gene has been deleted). Transcription of RNA gene and YFP gene was driven by 35S promoter. The second negative control contains ATG upstream of the sequence of RNA gene. This is to see if a random DNA sequence (ORF without stop codon) could be transcribed and translated too.

Figure 3. Fusion construction to verify translation of sORF candidates.

A.



B.

Negative controls:

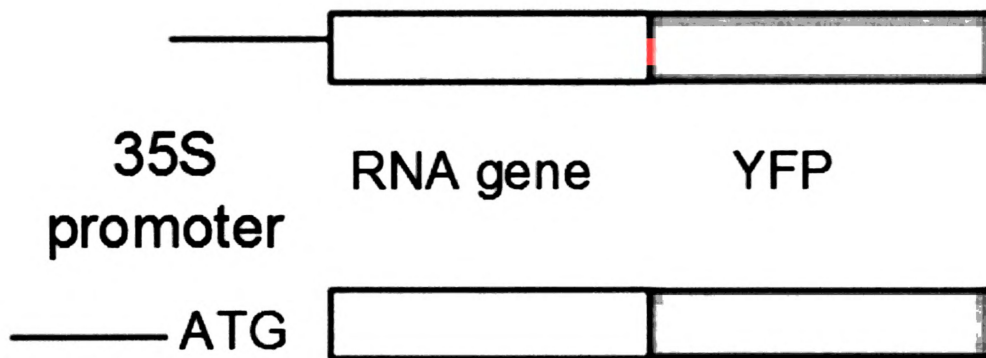


Figure 4. Confocal microscopy images of live leaf epidermal cells of tobacco showing transient expression of sORF candidates.

Transformed tobacco leaves were cut off 72 hours after transformation and observed at a confocal microscope to detect YFP signal. The chloroplast were pseudocolored red. **A.** Transient expression of a YFP construct (35S + YFP with start codon) without sORF insertions, a positive control, in tobacco epidermal cells. Scale bar, 20µm. **B.** Merge of **A** and the transmission light microscope image. The transmission image shows the outline of the epidermal cells. Scale bar, 20µm. **C.** Transient expression of ORF in a snoRNA (ORF in snoRNA +YFP) in tobacco epidermal cells. Scale bar, 20µm. **D.** Merge between **C** and the transmission light microscope image. Scale bar, 20µm. **E.** Epidermal cells of tobacco infiltrated with untransformed *Agrobacterium* (negative control). Scale bar, 20µm. **F.** Merge between **E** and the transmission light microscope image. Scale bar, 20µm. **G, I, K, M, O, Q.** Transient expression of six sORFs (sORF + YFP) in tobacco epidermal cells. **H, J, L, N, P.** Merge between **G, I, K, M, O, Q** and their respective transmission light microscope images. Scale bars in **G-H, I-J, K-L, O-R** are 10µm; scale bars in **M-N** are 20µm.

Figure 4. Confocal microscopy images of live leaf epidermal cells of tobacco showing transient expression of sORF candidates.

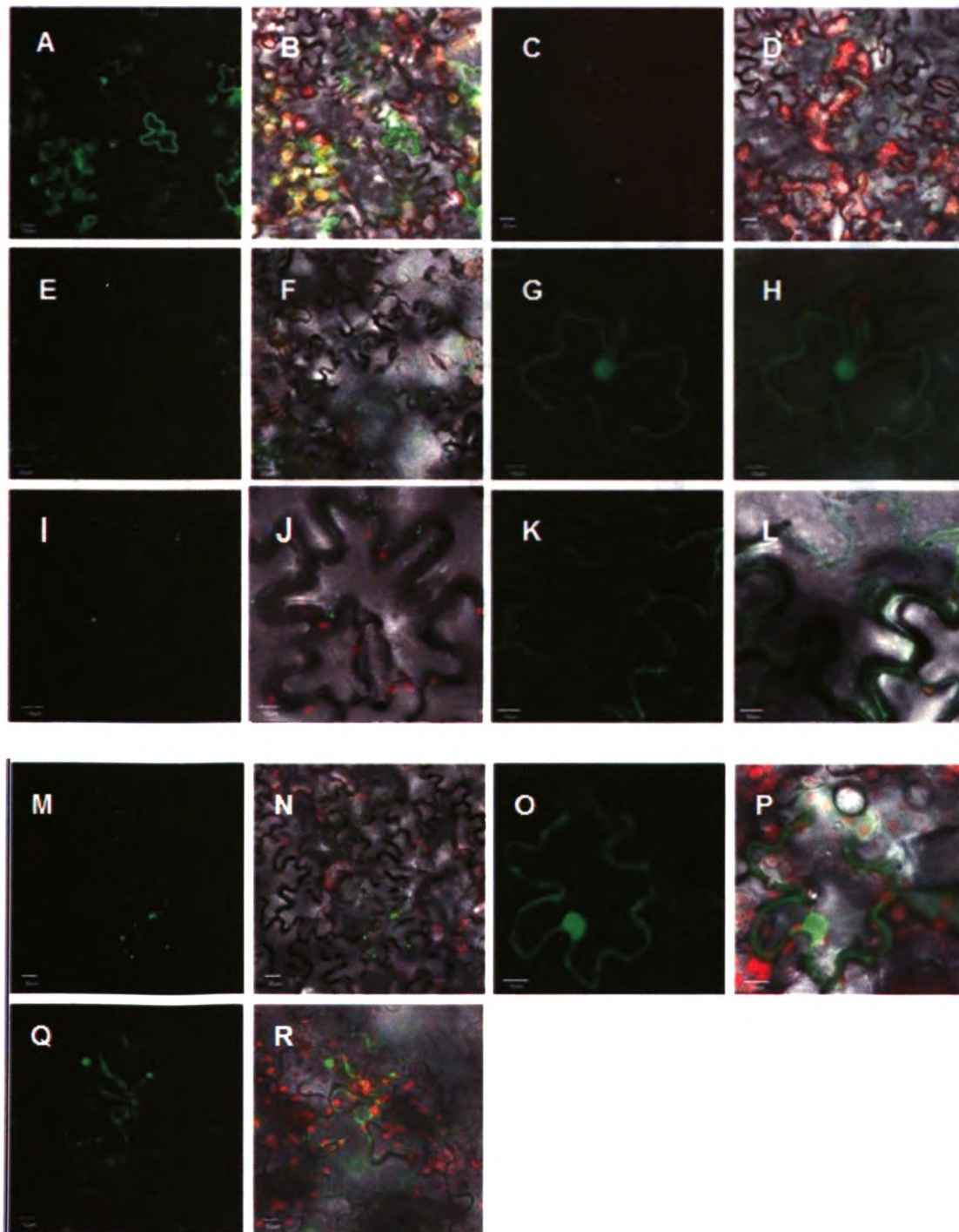


Figure 5. List of SALK lines studied.

| sORF candidates | SALK lines | Location of insertion |
|-------------------------------|-----------------------------|-----------------------|
| AT_4 0 8650117- 8650266 | SALK_021946 | |
| AT_1 0 7228279- 7228440 | SALK_101109 | |
| AT_3 +1 9667259- 9667468 | SALK_114025C SALK_114346 | |
| AT_4 +1 13361880- 13362065 | SALK_057047 | |

Figure 5 continued.

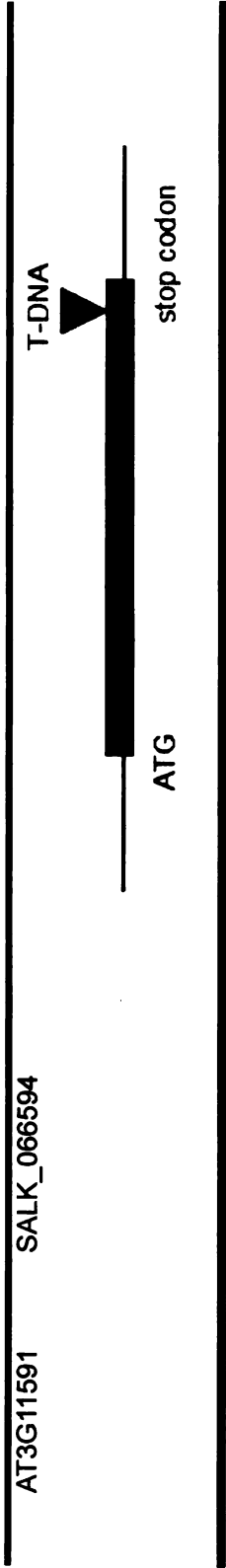


Table 1. RNA-seq evidence of sORF candidates.

| sORFs | Number of sORFs | IP RNA tags | Total RNA tags |
|-----------------------------|--------------------|-------------|-------------------|
| With RNA-seq evidence | 577 | Y | Y |
| | 441 | N | Y |
| Without RNA-seq evidence | 6,390 | N | N |

Table 2. RPKM values of sORF candidates.

| Expression level | Number of sORFs | RPKM values |
|------------------|-----------------|-------------|
| High | 13 | > 400 |
| Median | 80 | 10-400 |
| Low | 484 | 0.04-10 |

Table 3. Properties of sORF candidates that are tested by RACE

| sORFs | RNA-seq | Peptide | EST | 5'UTR | Surrounding Area in genome |
|----------------------------|----------------|----------------|------------|--------------|--|
| AT_4 - 0 17338876-17339082 | Y | N | Y | N | |
| AT_1 - 0 2443968-2444120 | Y | N | N | N | 300bp from a protein-coding gene,different direction,1300bp from a protein-coding gene,different direction |
| AT_3 + 1 7831136-7831378 | Y | N | Y | N | |
| AT_2 - 1 10458645-10458863 | Y | N | N | N | |
| AT_4 + 0 7470406-7470531 | Y | N | Y | N | |
| AT_2 - 1 11298405-11298521 | Y | N | Y | N | |
| AT_2 + 0 10458532-10458681 | Y | N | N | N | |
| AT_3 - 1 19168805-19168915 | Y | N | N | N | |
| AT_4 - 0 7233445-7233648 | Y | Y | Y | N | Sequence overlapping with a protein-coding gene, different direction |
| AT_1 + 2 3545268-3545375 | Y | N | Y | N | Sequence overlapping with an "other RNA", different direction |

Table 3 continued.

| | | | | | |
|----------------------------|---|---|---|---|--|
| AT_3 - 1 1849172-1849297 | N | N | Y | N | Sequence overlapping with an "other RNA", same direction |
| AT_1 - 1 3745307-3745540 | Y | N | Y | N | Sequence overlapping with an "other RNA", different direction |
| AT_5 - 1 6222522-6222680 | N | N | Y | N | |
| AT_1 - 0 16877703-16877861 | Y | Y | Y | N | Sequence overlapping with a protein-coding gene, same direction. |
| AT_3 - 0 4137141-4137296 | Y | N | Y | N | EST overlapping with a "other RNA", same direction.300bp from a protein-coding gene,different direction |
| AT_4 + 1 7470443-7470610 | Y | N | Y | N | |
| AT_4 - 1 11634-11801 | Y | Y | Y | N | sequence overlapping with a protein-coding gene, same direction, same reading frame. EST correspond to protein-coding gene |
| AT_5 - 2 26433887-26434033 | Y | Y | Y | N | EST overlapping with a protein-coding gene, different direction |

Table 3 continued.

| | | | | | |
|--------------------------------|---|---|---|---|--|
| AT_5 - 2 10035956- 10036057 | Y | Y | N | N | |
| AT_1 - 0 13636548- 13636733 | Y | N | N | N | |
| AT_1 - 0 10428690- 10428791 | Y | Y | Y | N | Sequence overlapping with an "other RNA" |
| AT_1 + 0 7228279- 7228440 | Y | N | N | N | |
| AT_4 - 1 13361880- 13362065 | Y | N | Y | N | |
| AT_4 - 0 8650117- 8650266 | Y | N | Y | N | EST overlapping with a protein-coding gene, different direction |
| AT_2 + 0 6671842- 6672111 | N | N | Y | Y | |
| AT_3 + 1 14212973 -14213269 | Y | N | Y | Y | |
| AT_1 - 2 20126281- 20126376 | Y | N | Y | Y | |
| AT_5 + 0 18002428 -18002532 | Y | N | Y | Y | Sequence overlapping with a protein-coding gene (longer, but still sORF), same direction, same reading frame |
| AT_1 - 2 5786755- 5786853 | Y | N | N | Y | Sequence overlapping with a protein-coding gene (longer, but still |

Table 3 continued.

| | | | | | |
|--------------------------------|---|---|---|---|--|
| AT_1 - 2 5786755- 5786853 | Y | N | N | Y | sORF), same direction, same reading frame, RACE of the protein-coding gene has been got |
| AT_3 + 2 4574607- 4574900 | Y | Y | Y | Y | Sequence overlapping with a protein-coding gene, same direction, same reading frame |
| AT_4 - 2 16617827- 16617958 | Y | N | Y | Y | Sequence overlapping with a protein-coding gene (very short, sORF), different direction, RACE for the protein-coding gene have been obtained. |
| AT_1 + 1 11469497 -11469754 | Y | Y | Y | Y | Sequence overlapping with an "other RNA", same direction |
| AT_1 - 1 8544986- 8545123 | Y | N | Y | Y | Sequence overlapping with an "other RNA", different direction |
| AT_5 - 2 6212882- 6212980 | Y | N | Y | Y | |
| AT_2 + 2 10458498 -10458590 | Y | N | N | Y | |
| AT_1 - 2 2407648- 2407782 | Y | N | Y | Y | |

Table 3 continued.

| | | | | |
|------------------------------|---|---|---|---|
| AT_1 - 0 2443968- 2444120 | Y | N | Y | Y |
| AT_3 + 1 9667259- 9667468 | Y | N | Y | Y |

Table 4. Properties of sORF candidates with 5'UTRs that were tested in tobacco transient expression system

| sORF candidates | RNA- seq | Peptide | Conservation | EST | Translated in tobacco | CI |
|--------------------------------|-------------|---------|--------------|-----|--------------------------|-------|
| AT2G15318 | N | N | Y | Y | N | 0.469 |
| AT3G03341 | Y | N | Y | Y | N | 0.627 |
| AT1G47265 | N | Y | Y | Y | N | 0.512 |
| AT2G13547 | N | N | Y | Y | N | 0.758 |
| AT5G45469 | N | N | Y | Y | N | 0.403 |
| AT_5 + 0 18002428- 18002532 | Y | N | N | Y | N | 0.904 |
| AT_3 + 1 9667259- 9667468 | Y | N | N | Y | N | 0.602 |
| AT2G07820 | N | N | Y | Y | N | 0.349 |
| AT3G28899 | N | N | Y | Y | N | 0.580 |
| AT_3 + 1 14212973- 14213269 | Y | N | N | Y | Y | 0.308 |
| AT_3 + 2 4574607- 4574900 | Y | Y | Y | Y | Y | 0.573 |
| AT_1 + 1 11469497- 11469754 | Y | Y | Y | Y | Y | 0.569 |
| AT_1 - 2 5786755- 5786853 | Y | N | Y | Y | Y | 0.478 |
| AT_1 - 2 20126281- 20126376 | Y | N | N | Y | Y | 0.986 |
| AT3G11591 | Y | N | Y | Y | Y | 0.478 |

Table 5. Fisher's exact test ($\alpha = 0.05$)

| | EST evidence | No EST evidence | RNA-seq | No RNA- seq | peptide evidence | No peptide evidence | purifying selectio n | No purifying selection |
|----------------------|---|--------------------|-------------------------------------|----------------|---|---------------------------|---|------------------------------|
| sORF translated | 5 | 1 | 6 | 0 | 2 | 4 | 3 | 3 |
| sORF untranslated | 9 | 0 | 3 | 6 | 1 | 8 | 2 | 7 |
| P-value | 0.400 | | 0.017 | | 0.525 | | 0.329 | |
| conclusion | The null hypothesis is not rejected. | | The null hypothesis is rejected. | | The null hypothesis is not rejected. | | The null hypothesis is not rejected. | |

Reference

- Alonso J. M., Stepanova A. N., Leisse T. J., Kim C. J., Chen H., Shinn P., Stevenson D. K., Zimmerman J., Barajas P., Cheuk R., Gadrinab C., Heller C., Jeske A., Koesema E., Meyers C. C., Parker H., Prednis L., Ansari Y., Choy N., Deen H., Geralt M., Hazari N., Hom E., Karnes M., Mulholland C., Ndubaku R., Schmidt I., Guzman P., Aguilar-Henonin L., Schmid M., Weigel D., Carter D. E., Marchand T., Risseuw E., Brogden D., Zeko A., Crosby W. L., Berry C. C., Ecker J. R. 2003. Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. *Science*. **310**: 653-657
- Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., Baginsky, S. 2008. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*. **320**:938-941.
- Basrai, M. A., Hieter, P. 2002. Transcriptome analysis of *Saccharomyces cerevisiae* using serial analysis of gene expression. *Methods Enzymol.* **350**: 414–444.
- Basrai, M. A., Hieter, P., Boeke, J. D. 1997. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Res.* **7**: 768-771.
- Bosch, T. C., Fujisawa, T. 2001. Polyps, peptides and patterning. *Bioessays*. **23**: 420–427.
- Casson, S. A., Chille, P. M., Topping, J. F., Evans, I. M., Souter, M. A., Lindsey, K. 2002. The POLARIS gene of *Arabidopsis* encodes a predicted peptide required for correct root growth and leaf vascular patterning. *Plant Cell*. **14**: 1705–1721.
- Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., Briggs, S. P. 2008. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl Acad. Sci. USA*. **105**: 21034-21038.
- Chanet, S., Vodovar, N., Mayau, V., Schweisguth, F. 2009. Genome Engineering-Based Analysis of Bearded Family Genes Reveals Both Functional Redundancy and a Nonessential Function in Lateral Inhibition in *Drosophila*. *Genetics*. **182**: 1101–1108
- Charon, C., Johansson, C., Kondorosi, E., Kondorosi, A., Crespi, M. 1997. *enod40* induces dedifferentiation and division of root cortical cells in legumes. *Proc. Natl Acad. Sci. USA*. **94**: 8901–8906.
- Chille, P. M., Casson, S. A., Tarkowski, P., Hawkins, N., Wang, K. L., Hussey,

- P. J., Beale, M., Ecker, J. R., Sandberg, G. K., Lindsey, K. 2006.** The POLARIS peptide of *Arabidopsis* regulates auxin transport and root growth via effects on ethylene signaling. *Plant Cell*. **18**: 3058–3072.
- Fisher, R. A. 1922.** On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*. **85**: 87–94.
- Frank, M. J., Smith, L. G. 2002.** A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. *Curr. Biol*. **12**: 849–853.
- Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., Grimmond, S. M. 2006.** The abundance of short proteins in the mammalian proteome. *PLoS Genet*. **2**: e52.
- Frohman, M. A., Dush, M. K., Martin, G. R. 1988.** Rapid production of full-length cDNAs from rare transcripts by amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA*. **85**: 8998-9002.
- Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H., Shiu, S. H. 2007.** A large number of novel coding small open reading frames in the intergenic regions of the *A. thaliana* genome are transcribed and/or under purifying selection. *Genome Res*. **17**: 632-640
- Hashimoto, Y., Kondo, T., Kageyama, Y. 2008.** Lilliputians get into the limelight: Novel class of small peptide genes in morphogenesis. *Develop. Growth Differ*. **50**: S269–S276.
- Inagaki, S., Numata, K., Kondo, T., Tomita, M., Yasuda, K., Kanai, A., Kageyama, Y. 2005.** Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*. *Genes Cells*. **10**: 1163–1173.
- Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W. C., Yang, H., Carter, C. D., Wheeler, D., Davis, R. W., Boeke, J. D., Snyder, M. A., Basrai, M. A. 2006.** Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res*. **16**:365–373.

- Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., Kageyama, Y. 2007.** Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology*. **9**: 660 - 665
- Kouchi, H., Takane, K., So, R. B., Ladha, J. K., Reddy, P. M. 1999.** Rice ENOD40: Isolation and expression analysis in rice and transgenic soybean root nodules. *Plant J*. **18**: 121–129.
- Kumar, A. 2009.** An Overview of Nested Genes in Eukaryotic Genomes. *Eukaryotic Cell*. **8**: 1321–1329.
- Kumar, A., Harrison, P. M., Cheung, K. H., Lan, N., Echols, N., Bertone, P., Miller, P., Gerstein, M. B., Snyder, M. 2002.** An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol*. **20**: 58–63.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. 2009.** Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. **10**:R25
- Lee, P. M. 2004.** Bayesian Statistics, an introduction (3rd ed. *Wiley*).
- Leviten, M. W., Lai, R. C., Posakony, J. W. 1997.** The *Drosophila* gene Bearded encodes a novel small protein and shares 3' UTR sequence motifs with multiple Enhancer of split Complex genes. *Development*. **124**: 4039-4051
- Lovett, P. S., Rogers, E. J. 1996.** Ribosome regulation by the nascent peptide. *Microbiol. Rev*. **60**: 366–385.
- Matsubayashi, Y., Sakagami, Y. 2006.** Peptide Hormones in Plants. *Annu. Rev. Plant Biol*. **57**:649–74.
- Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T. A., Okamoto, M., Nambara, E., Nakajima, M., Kawashima, M., Satou, M., Kim, J. M., Kobayashi, N., Toyoda, T., Shinozaki, K., Seki, M. 2008.** *Arabidopsis* Transcriptome Analysis under Drought, Cold, High-Salinity and ABA Treatment Conditions using a Tiling Array. *Plant and Cell Physiology*. **49**:1135-1149.

- Moskal Jr, W. A., Wu, H. C., Underwood, B. A., Wang, W., Town, C. D., Xiao, Y. 2007.** Experimental validation of novel genes predicted in the un-annotated regions of the *Arabidopsis* genome. *BMC Genomics*. 8:18.
- Narita, N. N., Moore, S., Horiguchi, G., Kubo, M., Demura, T., Fukuda, H., Goodrich, J., Tsukaya, H. 2004.** Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in *Arabidopsis thaliana*. *Plant J*. 38: 699-713.
- Olivas, W. M., Muhlrads, D., Parker, R. 1997.** Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res*. 25: 4619–4625.
- Oshiro, G., Wodicka, L. M., Washburn, M. P., Yates III, J. R., Lockhart, D. J., Winzeler, E. A. 2002.** Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res*. 12: 1210–1220.
- Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T., Sugano, S. 2007.** Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome. *Molecular & Cellular Proteomics*. 6:1000–1006.
- Sparkes, I. A., Runions, J., Kearns, A. Hawes, C. 2006.** Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nature Protocols*. 1:2019-2025.
- Stolc, V., Samanta, M. P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D. C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S., Ulrich, E. L., Zhao, Q., Wrobel, R. L., Newman, C. S., Fox, B. G., Phillips, G. N., Markley, J. L., Sussman, M. R. 2005.** Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl Acad. Sci. USA*. 102: 4453–4458.
- Topping, J. F., Lindsey, K. 1997.** Promoter trap markers differentiate structural and positional components of polar development in *Arabidopsis*. *Plant Cell*. 9:1713–1725.

- Trapnell, C., Pachter, L., Salzberg, S. L. 2009.** TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. **25**: 1105-1111
- Tupy, J. L., Bailey, A. M., Dailey, G., Evans-Holm, M., Siebel, C. W., Misra, S., Celniker, S. E., Rubin, G. M. 2005.** Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*. **102**: 5495–5500.
- Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. 1995.** Serial analysis of gene expression. *Science*. **270**: 484–487.
- Velculescu, V., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett Jr., D. E., Hieter, P., Vogelstein, B., Kinzler, K. 1997.** Characterization of the yeast transcriptome. *Cell*. **88**: 243–251.
- Vilela, C., Mccarthy, J. E. 2003.** Regulation of fungal gene expression via short open reading frames in the mRNA 5' untranslated region. *Mol. Microbiol.* **49**: 859–867.
- Wang, Z., Gerstein, M., Snyder, M. 2009.** RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. **10**: 57-63.
- Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., et al. 2003.** Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*. **302**: 842–846.
- Zanetti, M. E., Chang, I. F., Gong, F., Galbraith, D. W., Bailey-Serres, J. 2005.** Immunopurification of Polyribosomal Complexes of *Arabidopsis* for Global Analysis of Gene Expression. *Plant Physiology*. **138**: 624–635.

