Analysis of the NAC Gene Family in *Triticum aestivum* Employing Novel Bioinformatics Tools

Daniel Oren

A Thesis
in
The Department
of
Biology

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science at Concordia University Montreal, Canada

March 2008

© Daniel Oren, 2008
NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
Abstract

Analysis of the NAC Gene Family in *Triticum aestivum* Employing Novel Bioinformatics Tools

Daniel Oren

A novel pipeline was created to sift through large Expressed Sequence Tag (EST) data sets of *Triticum aestivum*, and assemble only sequences identified as being of the NAC gene family. This approach bypassed the bottleneck of assembly of large EST data sets, enabling the full cycle of sequence quality trimming and assembly to take place on a Personal Computer (PC) in a matter of a few hours.

An iterative sequencing conjoined with assembly approach was employed to fully sequence full length cDNA clones. Sequences were then reduced to a unigene set using both CAP3 assembly and MegaBLAST.

A software prototype designed to find protein coding regions, particularly in EST assemblies, was built and used to automatically identify full length NAC cDNA clones for sequencing, and create NAC nucleic and amino acid sequences trimmed to the protein coding regions and conserved domains.

A total of 130 NAC gene sequences were identified, 41 of which contained the entire protein coding region. Sequences containing the entire NAC conserved domain were used for conserved domain analysis, and a phylogenetic tree was built with previously identified NAC gene families in *Oryza sativa*, and *Arabidopsis thaliana*. At least one wheat NAC gene could be found in every phylogenetic group in which a rice NAC was present. Interestingly, a number of wheat NAC clusters without any definitive rice orthologs were found.
Dedication and Acknowledgments

This thesis is dedicated to my mother and father.

I would like to thank Dr. Gulick for the opportunity to perform this research and his valuable guidance. Also, I am grateful to my committee members Dr. Titorenko and Dr. Storms for their support.

I would like to thank my colleagues Dr. Ani Dryanova, Dr. Antonio Monroy, Wucheng Liu, and Mohammed Ridha for their knowledge and assistance. I would like to give a special thanks to Zornitsa Stoyanova for her assistance in the sequencing work.
Table of Contents

LIST OF FIGURES .................................................................................................................. VIII

LIST OF TABLES ..................................................................................................................... IX

1 INTRODUCTION.................................................................................................................. 1

1.1 THE NAC TRANSCRIPTION FACTOR GENE FAMILY ....................................................... 2

1.1.1 NAC regulation..................................................................................................... 3

1.1.2 Structural analysis............................................................................................... 4

1.2 THE GENOME OF TRITICUM AESTIVUM.................................................................. 5

1.3 EST SEQUENCES .......................................................................................................... 6

1.4 EST PROCESSING PIPELINES..................................................................................... 7

1.5 OPEN READING FRAME (ORF) IDENTIFICATION SOFTWARE................................. 10

1.6 COMPLETE cDNA SEQUENCING................................................................................ 11

2 MATERIALS AND METHODS .......................................................................................... 12

2.1 SETS OF SEQUENCES USED IN ANALYSIS............................................................... 12

2.2 INITIAL DETECTION AND ASSEMBLY OF EST SEQUENCES................................ 13

2.2.1 Initial Screening of NAC ESTs............................................................................. 13

2.2.2 Trimming and assembly....................................................................................... 15

2.2.3 Further resolution of assembled sequences....................................................... 15

2.2.4 Output files.......................................................................................................... 15

2.2.5 Further analysis.................................................................................................... 16

2.3 SEQUENCE COMPLETION........................................................................................... 16
2.4 Duplicate sequence removal between sets ........................................... 19

2.5 ORF detection and translation .................................................................. 20

2.5.1 ORF detection software prototype ....................................................... 20

2.5.1.1 Determination of sequence orientation ............................................ 21
2.5.1.2 Determination of frame of translation ............................................. 21
2.5.1.3 Determination of the query and subject alignment start and end ...... 22
2.5.1.4 Method for estimating whether a sequence spans the entire ORF ...... 22
2.5.1.5 Method for locating the start and stop codons in a sequence .......... 24
2.5.1.6 Method for estimating whether the start and stop codons are correct .. 24
2.5.1.7 Frameshift and intron detection ...................................................... 27
2.5.1.8 Domain resolution ........................................................................... 28
2.5.1.9 Program output ............................................................................... 29
2.5.1.10 Methods of output analysis .......................................................... 32

2.6 Phylogenetic tree construction and alignment methods ......................... 32

3 RESULTS .................................................................................................... 34

3.1 Initial NAC detection and assembly results ........................................... 34
3.2 End tally of T. aestivum NAC genes ....................................................... 35
3.3 Effectiveness of the ORF identification methodology ............................. 38
3.4 Subdomain conservation ......................................................................... 41
3.5 Phylogenetic analysis ............................................................................. 45
3.6 Identity between Triticeum aestivum full length homeologues ............... 52

4 DISCUSSION .................................................................................................. 55

4.1 Performance of the gene family detection pipeline ................................. 55
List of Figures

Figure 1 Quality trimming software trims sequences at the edges ........................................ 8

Figure 2 A Block diagram depicting the software pipeline as it was set up to detect, assemble and confirm the presence of a NAC conserved domain in NAC ESTs in Set 1 ................................................................................................................................. 14

Figure 3 A flow chart depicting the process used to complete the sequences of full length cDNA clones ........................................................................................................................................ 17

Figure 4 Sequencing .................................................................................................................. 18

Figure 5 Assembling two tagged contig sets together to find duplicate sequences among them ........................................................................................................................................ 19

Figure 6 Coding region estimated coverage based on BLAST ............................................. 23

Figure 7 Start and stop codon determination ........................................................................... 25

Figure 8 Suspect frameshift and intron detection .................................................................... 27

Figure 9 The report file of the ORF detection software .......................................................... 31

Figure 10 NAC subdomain amino acid sequence alignments .............................................. 44

Figure 11 A phylogenetic tree of wheat, rice, Arabidopsis, and other NACs of interest. 51
List of Tables

Table 1 Initial tally of NAC contigs and singletons .................................................. 35

Table 2 Tally of NAC sequences found among sets .................................................... 37

Table 3 List of malformed wheat sequences in the analysis ....................................... 40

Table 4 The percent nucleotide sequence identity between wheat NAC homeologues that contain the entire ORF .......................................................................................... 53
1 Introduction

The dramatic rise in the amount of genomic and cDNA sequence available in database repositories has created an impetus for the development of new analysis tools that use a comparative approach for the identification and annotation of genes. For many species the greatest source of genomic information are large Expressed Sequence Tags (ESTs) collections. These collections can be systematically analyzed by comparison to the whole genome sequences of related model species. Bread wheat, *Triticum aestivum*, which is an allohexaploid species, has the largest genome among the important crops. Although the whole genome sequence is not likely to be available soon, it has one of the largest EST databases among plant species, having over 500,000 sequences. The genomes of two plant species, Arabidopsis and rice, *Oryza sativa*, have been sequenced in entirety and can provide the gene model templates to analyze the wheat EST databases and cDNA sequences.

One systematic approach to progressive genome analysis has been to characterize large gene families, since the structure of individual genes can facilitate the delineation of gene models by both intra and inter specific comparisons. The NAC transcription factor gene family in plants is a good candidate for such analysis, due to its large size and the strong conservation of functional domains within the NAC protein family.
1.1 The NAC transcription factor gene family

NAC transcription factors have only been identified in the *viridiplantae* kingdom (Olsen et al., 2005). The complete genome sequence of Arabidopsis and rice revealed that the NAC genes comprise a large gene family. The first NAC transcription factor gene, *No Apical Meristem* (NAM), was identified by its role in petunia development. Mutant NAM plants generally developed without a Shoot Apical Meristem (SAM), most died at the seedling stage, and surviving plants displayed abnormal floral development (Souer et al., 1996). Cup-Shaped Cotyledon, *CUC1* and *CUC2*, mutants identified in *Arabidopsis* plants also lack a SAM, and contain cup shaped fused cotyledons (Aida et al., 1997). A common domain was found to be shared between the NAM, *ATAF1*, *ATAF2*, *CUC1*, and *CUC2* genes and the gene family for the related transcription factors has been designated NAC.

Members of the NAC gene family have been shown to participate in wide range of processes, including shoot apical meristem development (Souer et al., 1996; Aida et al., 1997), the formation of lateral roots (Xie et al., 2000), floral organ developmental specification (Sablowski and Meyerowitz, 1998), response to viral infection (Xie et al., 1999; Ren et al., 2000; Collinge and Boller, 2001), and response to a number of abiotic stresses (Hegedus et al., 2003; Fujita et al., 2004; Tran et al., 2004).

Over-expression of the Arabidopsis *NAC1* gene caused increased lateral root formation, whereas a reduction in its expression reduced root branching. The gene was found to be induced by auxin (Xie et al., 2000), a hormone that participates in lateral root development, and was also found to regulate genes previously associated with auxin response.
NACs have also been found to participate in biotic and abiotic stress response. The Arabidopsis TIP gene has been found to act in turnip crinkle virus resistance (Ren et al., 2000). The NAC wheat GRAB1 and GRAB2 genes were found to interact with the wheat dwarf geminivirus RepA protein (Xie et al., 1999).

The Arabidopsis ATAF1 and ATAF2 have been shown to be induced in response to wounding (Collinge and Boller, 2001). Other Arabidopsis NAC genes within the same subgroup as the ATAF genes (Ooka et al., 2003) named ANAC019, ANAC055, and ANAC072 have been found to be induced in response to dehydration, high salinity and abscisic acid (ABA) (Fujita et al., 2004; Tran et al., 2004). ANAC072 in particular has been shown to participate in an ABA-dependent stress-signaling pathway and to be induced in response to reactive oxygen species. The potato (Solanum tuberosum) gene StNAC and a number of Brassica napus (rape) NAC genes have also been found to be induced by both biotic and abiotic stress responses (Collinge and Boller, 2001; Hegedus et al., 2003). NAC expression analysis has provided evidence of T. aestivum NAC participation in cold acclimation (Monroy et al., 2007).

1.1.1 NAC regulation

NACs have been shown to be regulated by a number of mechanisms. The Arabidopsis NAP gene, a member of the NAC gene family, has been shown to be directly regulated at the transcriptional level by the MAD-box transcription factors APETALA3 and PISTILLATA (Sablowski and Meyerowitz, 1998). The Maize NAC gene no-apical-meristem (NAM) related protein1 (nrp1) has been found to be silenced via gene-specific imprinting (Guo et al., 2003). NAC genes have been found to be regulated at the RNA level by miRNA; CUC1 and CUC2 were each found to have a site matching miRNA.
miR164 (Rhoades et al., 2002), and it is suggested that their mRNA is cleaved in that area (Kasschau et al., 2003). Levels of NAC1 protein have been found to be controlled through proteosomal degradation (Xie et al., 2002). *SINAT5* has been shown to act as a E3 ubiquitin protein ligase that can ubiquinate NAC1.

### 1.1.2 Structural analysis

NAC proteins are characterized by an N-terminal domain, which consists of five conserved subdomains (Souer et al., 1996; Aida et al., 1997). The structure of the conserved domain of the Arabidopsis protein ANAC19 has been determined and found to consist of a twisted β-sheet flanked on both sides with differentially structured helices (Ernst et al., 2004). NAC proteins form homo and heterodimers, and a highly conserved portion of the N-terminal domain was found to be the point of NAC protein contact for the formation of these dimers (Ernst et al., 2004). The sub-domains D and E have been found to be involved in DNA binding in the Arabidopsis NAC *AtNAM* (Duval et al., 2002). The C-terminal region has been found to be essential for the transcriptional activation functionality of a number of NAC proteins (Duval et al., 2002; Hegedus et al., 2003; Fujita et al., 2004; Tran et al., 2004).

Members of the NAC gene family have been implicated in a wide variety of gene functions, yet only small proportion of these genes have been investigated. The resolution of this gene family in *Triticum Aestivum* is sure to assist in the investigation of numerous biological pathways.
1.2 The genome of *Triticum aestivum*

*Triticum aestivum*, bread wheat, is a hexaploid species that originated from two hybridization events. The first one is estimated to have occurred 360,000 years ago between the diploid species *Triticum urartu*, which contributed the AA genome, and another diploid species, thought to be a close relative of *Aegilops speltoides* (*Sarkar and Stebbins, 1956; Dvorak and Zhang, 1990; Dvorak et al., 1993*), which contributed the BB genome to form *Triticum turgidum* ssp. *dicoccoides* (genome AABB). The second hybridization event occurred between the tetraploid *T. turgidum* and the diploid (genome DD) *T. tauschii* (*Kihara, 1944; McFadden and Sears, 1946; Dvorak et al., 1998*) to form the hexaploid *T. aestivum* (genome AABBDD). This latest hybridization event was estimated to have occurred approximately 8000 (Nesbitt and Samuel, 1996) years ago. The allohexaploid is expected to have three homeologous copies of any particular gene. However, a greater or fewer number than three copies might be present as a result of past chromosomal duplication or deletion events.

The genome of this species has not been sequenced; however a large collection of Expressed Sequence Tags (ESTs) exists. Furthermore, a related monocot species, rice, has been fully sequenced and that species can be used to estimate gene family sizes and aid in gene annotation in *T. aestivum*.

Though a number of gene families have been described in wheat (*Ridha Farajalla and Gulick, 2007*), no attempt has been made to date to describe a gene family of the projected magnitude of the NAC gene family.
1.3 EST sequences

ESTs are sequences originating from single sequencing reaction reads of cDNA clones (Nagaraj et al., 2007). Since cDNA is synthesized from mRNA through reverse transcription, cDNA clones can represent various portions of the Open Reading Frame (ORF), and the 5' and 3' Un-Translated Regions (UTRs) of the mRNA. A poly-a tail can normally be found at the 3' end of a cDNA clone. The mRNA from which the cDNA is synthesized is extracted from an organism at a particular stage of development and often from particular tissues, and so represents a snapshot of the various active genes in that tissue at a particular point in time. A gene that is more highly expressed will be more highly represented in the mRNA, and subsequently in the cDNA. Thus, clones of genes that are usually expressed at lower levels, such as transcription factors, may appear at a very low frequency in a cDNA library.

cDNA clones may not represent an entire mRNA. Since the cDNA is synthesized in vitro from the 3' end, the 5' end of the cDNA may be missing. However, the presence of a start codon in the sequence is a good indication that the entire reading frame is present. The single sequence reaction by which ESTs are produced results in about 400-800 base pairs (bp) of reliable sequence, and so generally will not encompass the entire protein coding region of a gene.

The EST GenBank database of T. aestivum contains over 500,000 sequences. Though it is a rich source of DNA sequence information, the sequences are largely un-annotated with respect to gene identity and function. They are often partial sequences of cDNA clones, in many cases of partial length cDNA clones. Further, EST sequences are not perfect, but rather contain errors within the sequence due to misreads by the automated sequencers.
Several overlapping partial length EST sequences from the same gene may be assembled into a contig sequence that represents an entire mRNA transcript. At best such resolved sequences are labeled as tentative contigs due to uncertainty introduced by poor quality sequence and cloning artifacts such as chimeric cDNA, which are clones in which two fragments from different genes were cloned into the same plasmid.

If a cDNA sequence is to be entirely resolved, a clone must be sequenced multiple times, using internal sequencing primers and the overlapping EST sequences assembled into a single complete sequence. In this thesis a number of software prototypes are presented that were built to identify, characterize, and assemble cDNA sequences. These were used to resolve the DNA sequences for NAC transcription factors from *T. aestivum*. The program was initiated from the analysis of the wheat EST data base and encompassed the additional sequencing and analysis of a large number of full length cDNA clones of NAC transcription factors.

1.4 EST processing pipelines

To process ESTs, a software pipeline is usually built that runs a number of programs in sequence, while delegating the EST sequences between them (Pertea et al., 2003; Xu et al., 2003). Initially, sequences are stored in binary files called trace files (such as ABI trace files), which represent the DNA sequence as a string of color coded peaks. DNA sequences can be extracted from these files along with a sequence quality values for each base, which reflects the probability of the nucleotide base call being correct. To this end, each trace file is put through ‘base calling software’ that assigns nucleotides and corresponding quality scores for the sequence and puts them in separate sequence and quality files. The quality scores are derived from base call error probabilities using a
formula called a phred score in which $Q = -10 \times \log_{10} (\text{probability of base call error})$ (Ewing and Green, 1998), so a base with a quality of 10 has a probability of 1 in 10 of being in error and should be correct in 90% of cases, and a base with a quality of 20 has a probability of 1 in 100 of being in error and should be correct in 99% of cases.

Bases within the first 20 nucleotides and beyond 300 nucleotides tend to contain many more errors than bases in the centre of the sequence, though with improvements in sequencing technology, reasonably high quality sequence can be obtained substantially beyond 500 nucleotides (Figure 1). Additionally, the beginning and end of the sequence may contain vector sequence derived from the plasmid into which the cDNA is inserted. Sequences are processed by vector and low quality trimming software, which trims both vector and bad quality sequence. The EST trimming software can take advantage of the fact that both low quality and vector sequence is generally found in the beginnings and ends of the EST sequences, by starting to look for bad sequence at the edges and moving towards the centre. The result is a high quality vector free sequence represented by the central region of the EST.

Figure 1 Quality trimming software trims sequences at the edges. The vector sequence and the vast majority of low quality sequence, which are trimmed by quality trimming software, reside at the edges.
Thousands of cDNA clones, many of which represent the same gene, are sequenced in many EST sequencing projects. Sequence assembly software is employed to find the ESTs that belong with each other, by looking for common overlaps in sequence. As ESTs typically represent only 400-800 nucleotides of sequence, the assembly program is also instrumental in elongating the contiguous span of sequences representing individual genes, by putting the ESTs together. The output of the assembly software is typically a combination of sequences that are a consensus of a number of ESTs, named contigs, and sequences that did not assemble with any others, named singletons.

Sequence assembly is undoubtedly the most computationally intensive part of the pipeline, as typically hundreds of thousands of sequences must be compared with each other. To reduce the intensity of this step, sequences are sometimes subdivided randomly into smaller groups, and each group is assembled with each other, after which each assembled group is assembled with each other, and so forth, until one assembled group is realized. Nevertheless, strong computational power, typically in the form of a Beowulf cluster or mainframe, and much time, typically months, are required to assemble a large EST library.

The final step in the sequence resolution pipeline is to annotate the assembled contigs and singletons. Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) is usually used to compare the unknown sequences to protein sequences of other related species, whose annotation is already known, and assigning the unknown sequences the annotation of their known homologues.

Here a software pipeline prototype is described that can look through large EST
collections and process only EST subsets of interest. By targeting a subset of a large EST collection before assembly, a gene family can be resolved on a single PC in a matter of a few hours.

1.5 **Open Reading Frame (ORF) identification software**

An ORF is an area of a gene that can be translated into a protein. The area begins with a start codon and ends with a stop codon. Identification of the protein coding region of a gene, and translation of that region into amino acids, facilitates analysis of its function, and its translated protein’s structure and conserved regions. Determining if a contig or singleton contains the 5’ end and/or 3’ of the coding region is used to determine if its corresponding cDNA is a good candidate for further sequencing.

Special considerations need to be taken when attempting to predict ORFs in EST-derived sequences. For one, ESTs tend to have errors near their ends, even when employing quality trimming software. These errors can lead to the misidentification of the start or stop codons. Additionally, EST contigs and singletons frequently do not contain enough sequence to cover both ends of the ORF, leading to erroneous identification of the start and end points of the protein coding region.

There are a number of bioinformatics software products that have been developed for the purpose of protein coding region identification in nucleic acid sequences including GRAILEXP (Xu and Uberbacher, 1997), GeneID (Guigo et al., 1992; Parra et al., 2000), GENSCAN (Burge and Karlin, 1997), and orfPredictor (Min et al., 2005). None of these was able to process the new wheat NAC sequences derived in this project with any accuracy, presumably due to them not being configured for processing plant gene sequences, or not being configured to process ESTs.
A formidable part of this thesis is devoted to the description of an ORF identification tool prototype, used to analyze the *T. aestivum* NAC sequences, that was built to automatically identify and translate protein coding regions in assembled ESTs en masse.

### 1.6 Complete cDNA Sequencing

The actual sequencing of ESTs is carried out in sequencing centers. As described earlier, the resulting EST sequences are reliable to only 500-800 bp, and so generally do not cover the gene's entire ORF, and contain errors. To obtain a high quality sequence that covers the entire ORF, ESTs of different clones that represent the same gene can be assembled with each other, and/or additional internal sequence can be obtained from the original cDNA clone through sequencing.

A common method of sequencing an entire clone is to make primers at the high quality end points of a sequence, and continue sequencing the clone from that point onwards. This is performed iteratively, until the end of the clone is reached. Both strands of a clone can be sequenced to achieve more reliable results.

An efficient approach, presented here, is to check the pool of sequences after each sequencing iteration, to see if sequences from different clones overlap with each other. This can reduce the overall number of cDNA clones that need to be resequenced, and increase the reliability of some of the remaining contigs, which can improve the speed and lower the cost of each of the following sequencing rounds.
2 Materials and Methods

2.1 Sets of sequences used in analysis

The EST sequences for the NAC transcription factor family in *T. aestivum* originate from three primary sources; each of which had specific advantages and disadvantages.

(1) EST sequences of cDNA clones from a single cultivar, Norstar, originating from the Genome Canada supported program ‘Functional Genomics of Abiotic Stress’ (FGAS) (Houde et al., 2006). This EST set was generated from a single cultivar of wheat, and would not generate contigs composed of multiple alleles from different cultivars that may contain single nucleotide polymorphisms or other variations. Since the physical clones for these accessions were available locally, they could be used for additional sequencing. There were approximately 80,000 such ESTs.

(2) Other *T. aestivum* EST sequences for which phred sequence quality files were available were obtained through the generosity of Timothy Close (University of California, Riverside) and Olin Anderson (USDA/ARS Albany, California). These include the NSF-USDA and Dupont sets of wheat ESTs. These were potentially highly useful sequences, but they were initially processed elsewhere and arrived extensively trimmed, so much that sequences from the same gene often did not overlap sufficiently during assembly to form a contig, and when contigs formed they rarely encompassed the entire coding region. There were approximately 270,000 such ESTs available.
(3) Other sequences from GenBank (Benson et al., 2007) for which no sequence quality data is available. While these sequences were numerous, they lacked quality files and originated from various libraries and cultivars, so their reliability could not be quantitatively established with accuracy except when multiple sequences for the same gene could be identified. At the time that they were retrieved from GenBank approximately 550,000 sequences were available.

The sets will be referred to as Set 1, Set 2, and Set 3 throughout the rest of the thesis.

### 2.2 Initial detection and assembly of EST sequences

A software pipeline prototype was built to facilitate fast differentiation of ESTs belonging to the same gene family from others in large data sets, trim them, and assemble them into contigs. A general block diagram depicting the pipeline as it was setup to screen the sequences in Set 1 is shown in Figure 2.

#### 2.2.1 Initial Screening of NAC ESTs

BLAST source databases were initially created for the wheat EST FASTA sets using formatdb, a utility used to create BLAST protein or nucleotide source databases. The sets were then superficially screened for NACs using TBLASTN (Altschul et al., 1990). A set of 82 rice, and 16 Arabidopsis full-length NAC protein sequences gathered from the National Center for Biotechnology Information (NCBI) Entrez protein database were used as a comparison set.

The TBLASTN output was scanned for matches, and the candidate sequences with an e value of 0.001 or less were extracted from the overall EST FASTA file, and complementary phred quality values were selected.
Figure 2 A Block diagram depicting the software pipeline as it was set up to detect, assemble and confirm the presence of a NAC conserved domain in NAC ESTs in Set 1.
2.2.2 Trimming and assembly

The sequences in Set 1 were marked for trimming using Lucy (Chou and Holmes, 2001) set to default values, and the EST FASTA and quality values were then trimmed for low quality and vector sequence. The sequence assembly program CAP3 (Huang and Madan, 1999) was run using a minimum overlap of 100 bp and a minimum 99% sequence identity over the overlap. Those sequences without quality values were given a uniform phred value of 17 as a default in order to run in CAP3.

2.2.3 Further resolution of assembled sequences

A BLAST source database was created for the candidate NAC contigs and singletons using formatdb. To confirm the identity of the candidate wheat NACs they were compared against sequences of the entire NAC domain, 2nd sub-domain, and 4th sub-domain from the rice and Arabidopsis reference protein set using TBLASTN with a maximum e value of 10. Sequences that were aligned with the entire sub domain with a score of less than 100, or had an alignment that did not start within 20 amino acids of the start of best aligned domain, or did not reach amino acid position 126 of the best aligned domain were not considered. Also, sequences that were aligned with the fourth and second subdomains with a score of less than 23, or had an alignment that did not start within seven amino acids of the start of the best aligned subdomain, or did not reach amino acid position 15 of the best aligned subdomain were not considered.

2.2.4 Output files

Sequences confirmed to belong to the gene family of interest and their associated quality values were placed in separate files, the CAP3 assembly file was trimmed to reflect only
confirmed NACs, and a report was produced noting the type of domains each contig and singleton hit, and how many ESTs were present in each contig.

### 2.2.5 Further analysis

After the pipeline was run, Set 1 contigs and singleton were further consolidated through manual inspection and by using MegaBlast with a word size of 120 and a minimum identity of 99%.

Singleton sequences of Set 3, which lack quality files, that had significant similarity to NAC transcription factors but failed to assemble under stringent conditions, were put through an additional CAP3 assembly using a threshold of 96% identity with 40 bp minimum overlap, in order to estimate the total number of novel genes that they may represent.

### 2.3 Sequence Completion

Full length clones were detected by noting if they contain the 5' end of the ORF (ATG) relative to other sequences from rice and Arabidopsis. The 5' ends were detected automatically using the sequence ORF detection software described in section 2.5.1. An iterative sequencing process was performed to complete the full length sequences of the full length NAC cDNA clones from Set 1 that were identified. This iterative process was composed of five steps (Figure 3) listed below.

1. Sequences were checked to see if they contained a poly-a tail. The presence of a poly-a tail meant that they were fully sequenced, and nothing else needed to be done.
(2) If the clone sequence was not full length, primers were designed for the clone. 5' primers were designed based on high quality existing EST sequence, and 3'
sequences were initiated from vector. Primer 3 (Rozen and Skaletsky, 2000) software was used to design the sequencing primers.

(3) The sequencing of clones was done at the Genome Quebec Innovation Centre.

(4) The resulting ABI files were then put through the phred program to generate sequence and quality files. Both sequence and quality files were trimmed of vector sequence and low quality regions using the Lucy software with default values.

(5) New sequences were assembled with the overall pool of singletons and contig candidate sequences using CAP3. A criterion of 100 bp minimum overlap and minimum 99% base pair identity over the overlap was used. This both incorporated new sequences into contigs and facilitated the bridging of any existing sequences that belonged to the same gene but that did not previously overlap the sequence (Figure 4). It also facilitated the detection of errors in sample labeling because incorrectly labeled clones would assemble with an unexpected contig.

![Diagram of sequencing](image)

**Figure 4 Sequencing.** The figure illustrates how sequences are extended from both the 5' and 3' ends. Re-assembly of extended sequences with a known pool of sequences can integrate new sequences that did not have previous overlap with the full length sequence. These in turn can provide a bridge to sequences that are extended from the 3' end vector.
2.4 Duplicate sequence removal between sets

Two contigs or singletons assembled from different EST sets were declared to be duplicates if they assembled together using an assembly identity threshold of 99% over an overlap length of 100 bp (figure 5), or were aligned together by running MegaBLAST using a word size of 120 bp and 99% identity. Removal of duplicate sequences by means of assembly was performed first and the resulting sequences run through MegaBLAST. By using these methods the sequences were reduced to a unigene set and preference was given to sequences originating from the sequencing of the full length cDNA clones from the Set 1, secondly to other sequences assembled with phred quality files from sets 1 and 2, and lastly to singletons and contigs from Set 3, for which phred quality scores were not available.

Figure 5 Assembling two tagged contig sets together to find duplicate sequences among them. In the figure above two sets, Set A and Set B, are tagged with ‘set specific’ tags and are assembled together. The sequences from Set B that assemble with sequences from Set A are duplicates.
2.5 ORF detection and translation

Using the software described below it was determined whether the NAC sequences were full length, and whether they contained the entire ORF. It was also determined whether they contained the first four subdomains, and all five subdomains. A reference set of 82 rice, and 16 Arabidopsis full-length NAC protein sequences was used for comparison. Sequences determined to contain the ORF or the subdomains were translated and trimmed to those regions.

Sequences covering the entire domain were then aligned using ClustalW (Thompson et al., 1994), and final manual trimming was performed on their edges to ensure accurate domain coverage.

2.5.1 ORF detection software prototype

The ORF detection software identifies the protein coding region of a sequence, and trims and translates the sequence accordingly. It estimates whether a sequence is full length, and whether the start or stop codons are found in a probable area.

It tackles these issues using extrinsic methods exclusively. Initially, known protein sequences are compiled into a BLAST database for comparison, and a BLASTX (Altschul et al., 1990) of the sequences to be translated is performed against them. As such, the query sequences in the BLASTX output are the sequences that are being analyzed, and the subject sequences are the known sequences that are being compared against to identify the ORF. Miscellaneous and incorrect matches can interfere with the program’s decision making. As such, all High-scoring Segment Pairs (HSPs) within a hit, in which the total aligned region of the subject is less than 20 amino acids, are ignored in all calculations.
2.5.1.1 Determination of sequence orientation

To calculate the sequence orientation, the scores of each hit in each orientation are added, and the orientation with the highest score is designated as the primary orientation. Occasionally, multiple blocks of alignment, also known as HSPs, occur within a single hit, and these can have different orientations. In these cases only the first HSP is used to calculate the score and orientation of the particular hit, since the HSPs within each hit are always presented in order of diminishing scores in the BLASTX output. So for example, if the HSP of the first hit is in the positive orientation, and the hit has a score of 545, and there are no more hits in the result, the final score for that orientation is 545. However, if there are two more hits in the result, and they are both in the negative orientation, and the score of the first HSP in the second hit is 350, and the score of the first HSP in the third hit is 250, adding to an overall orientation score of 600, the overall sequence would be designated as being in the negative orientation.

2.5.1.2 Determination of frame of translation

A similar process is undertaken to decide upon the reading frame of translation. The reading frame for each hit is taken to be that of the first HSP it contains, and the overall sequence frame is derived by adding the scores of all the hits representing each frame of translation, and taking the frame with the highest score. This method insures correct frame detection in cases where a weak, incorrect alignment occurs in the wrong frame. In case of a frame-shift, the decided frame will not necessarily be that in which the start and/or stop codon happen to be. While this strategy could result in miscalls in the determination of the existence of a start or stop codon in a sequence, sufficient warnings are provided in such cases as explained below.
2.5.1.3 Determination of the query and subject alignment start and end

The first and last aligned positions between a query and subject sequence of the best hit is used to determine if the clone is full length, and to detect the start and stop codons. Only HSPs in the decided primary orientation and reading frame are used for this determination. If the query sequence is aligned to the subject in the reverse orientation, the actual first aligned position of the query is counted from the 3' end of the sequence by taking the query length minus the last aligned position in the reverse orientation plus 1, and the query's actual last aligned position is counted from the 3' end of the sequence by subtracting the first alignment position plus 1 from the query length.

2.5.1.4 Method for estimating whether a sequence spans the entire ORF

A BLASTX alignment in which the first amino acid of the subject aligns with the query sequence normally indicates that the query is likely to be a full length cDNA. A query sequence may also be full length in cases in which it does not have high enough sequence similarity to align at the beginning of the subject sequence (Sczyrba et al., 2005). This can be the case if it has more sequence than the subject upstream of the beginning of the alignment (Figure 6).

An equation was designed to detect such sequences. In a BLASTX alignment the query positions appear in base pairs, whereas the subject positions appear in amino acids. So the likelihood that a query sequence is indeed full length can be roughly represented as (query alignment start position - frame)/(subject alignment start * 3). The frame in the equation is a number according to GFF (Generic Feature Format), which is from zero to two, rather than being from one to three.

Conversely, the likelihood that a sequence reaches the 3' end can be roughly represented
as \((\text{query length} - 3' \text{ end of the query alignment} + \text{frame} + 1)/((\text{subject length} - \text{end of the subject alignment} + 1) \times 3)\). A one is added to the query and subject to avoid division by zero.

![Diagram](image)

**Figure 6** Coding region estimated coverage based on BLAST. In the example above, the query alignment begins 50 bp downstream of the sequence start, whereas the subject alignment begins at amino acid 10. Since the 50 upstream nucleotides in the query are 20 in addition to the 30 upstream nucleotides predicted by the subject, it is likely that the query sequence encompasses the 5' of its own ORF. On the other hand, the query has 30 bp less coverage relative to the subject on the 3' end. Since the subject end is the end of its ORF, it is unlikely that the query reaches all the way to the end of its own ORF.

Values greater than one denote a higher likelihood that the sequence reaches the start or end of the coding region respectively. The likelihood decreases the more the value approaches 0. Dividing the query by the subject allows for proportional representation of the differences in distances between their alignment start and sequence start.

The more divergent the query and subject sequences are within the aligned area, the bigger the expected difference in lengths of sequence before the alignment start. That is to say, the more divergent the sequences, the more additions and deletions they should relative to one another. Therefore to determine whether a sequence reaches the 5' of the ORF, the calculation is adjusted for divergence between the query and subject. This
entails, in rough terms, taking the subject alignment start times the estimated 'divergence' between the subject and query (1 − alignment fraction identity). The result is added to the query alignment start to be divided by the subject in the formula described above. The less identity there is between the subject and query, the more leeway is given to the expectation that the two proteins will have the same length. Finally, the log base 2 of the retuned value is calculated for easier readability.

2.5.1.5 Method for locating the start and stop codons in a sequence

Since any aligned region of the query and subject produced by using BLASTX would presumably indicate a translated region of the gene, any stop codons in the query sequence before the first aligned position in the same reading frame would be an indicator of the presence of the 5' UTR, if we assume the absence of unspliced introns and sequence errors.

Sequences are translated in the chosen frame, and the start codon of the Open Reading Frame (ORF) is found by first detecting the first stop codon, if any, in the query sequence upstream of the first position of the alignment (Figure 7). If one is found, the first methionine downstream of it is taken to be the start codon; otherwise the first methionine in the sequence in frame with the alignment is taken as the start codon. The end of the coding region is taken as the first in-frame stop codon found after the start codon.

2.5.1.6 Method for estimating whether the start and stop codons are correct

In cases where the sequence is devoid of errors, the first and last codons of the coding regions found using the method outlined in the section above will be correct the vast majority of the time. However, EST sequences may be malformed in a number of ways
that could lead to incorrect first and/or last codon determinations.

Clones may contain unspliced introns and these unspliced introns frequently contain stop codons. If they are within the coding region they will cause the software to decide that a premature incorrect stop within the unspliced intron is the end of the coding region. A base call error resulting in an extra base added or deleted will create a frameshift in the sequence, resulting in incorrect sequence translation and the probable appearance of incorrect start and/or stop codons. Finally, low quality sequence may result in incorrect start or stop codon detection. To address this, the program attempts to estimate whether the start and stop codon found are within a reasonable distance from where one would expect them to be based on the alignment with the subject from the best hit (Figure 7).

![Figure 7 Start and stop codon determination](image)

*Figure 7 Start and stop codon determination.* The start codon of the ORF is found by detecting the first stop codon, if any, in the query sequence upstream of the first position of the alignment. If one is found, the first methionine downstream of it is taken to be the start codon; otherwise the first methionine in the sequence is taken. The end of the coding region is taken as the first stop codon found after the postulated start codon. The software also estimates whether the start and stop codons are found within a determined range relative to the start and stops found in the most similar homologue in rice.

The estimate for the start codon is done by deducting the subject’s first aligned position from the query’s first aligned position, giving the estimated start codon position based on
the subject. It then computes a range in which the start codon should appear by adding and deducting a leeway from the estimated start codon position. The leeway is computed, in rough terms, by adding a constant (10) plus the subject distance from the beginning of the alignment times the 'divergence' of the two sequences (1 – alignment fraction identity). A similar procedure is performed to find the estimated range in which the stop codon should appear.

The frame of translation and the fact that the subject is represented in amino acids needs to be taken into account in the calculations. The general method for doing so was discussed in section 2.5.1.4, which describes the method for estimating whether sequences reach the 5’ and 3’ ends of the coding region.
2.5.1.7 Frameshift and intron detection

Hits containing at least two HSPs with different frames but in the same orientation, where the query and subject regions of one HSP are not completely contained in the other, are taken to be suspected frameshifts, and are noted as such (Figure 8). Hits with two HSPs in the same frame and orientation, where the query and subject regions of one HSP are not contained in the other, are noted as possible unspliced introns (called 'gaps' in the report output file described in section 2.5.1.9.1).

Figure 8 Suspect frameshift and intron detection. (A) Frameshifts are detected by observing two or more HSPs in different reading frames within a single BLAST. (B) and (C) Suspect introns are detected by observing two or more HSPs of one hit containing a gap in the query relative to the subject, or a gap in the subject relative to the query. (D) One HSP contained in another is not considered to be either a frameshift or intron. (E) HSPs in different orientations are not considered to be frameshifts or introns.
2.5.1.8 Domain resolution

The conserved domains were identified by performing BLASTX against the NAC domains comprising the first four sub-domains of 82 O. sativa, and 16 Arabidopsis NACs.

The software was also designed to identify user supplied conserved domains, to estimate whether sequences in question contained a domain in its entirety, and to generate amino acid sequence and nucleotide sequence from this region. The user must initially supply a number of known domains in a FASTA file as references for a BLASTX comparison. The reference domains are compiled into a BLAST database and the conserved domains are identified in candidate sequences by performing a BLASTX comparison against them. In the resulting BLAST output file, the reference domains are the subject sequences, and the sequences whose domains are to be identified are the query sequences.

To find the domain in a query sequence, the number of base pairs between the subject alignment start position and subject domain start are subtracted from the query alignment start position and the sequence is trimmed to that position, if the query sequence is long enough on the 5' end. Similarly, the number of base pairs between the subject alignment end and the subject domain end are added to the query alignment end position, and the sequence is trimmed to that position if the query sequence is long enough on the 3' end. A user supplied 'buffer' allows for further sequence extension in either direction to ensure that the domain appears in its entirety in the trimmed sequence. Frameshifts and gaps are detected in the same manner as described in section 2.5.1.7 above.
2.5.1.9 Program output

2.5.1.9.1 Files

The software produces four FASTA files and a report file. Each gene sequence is given trimmed to the full ORF and trimmed to the NAC domain. Nucleotide and amino acid versions of the sequence are created as separate FASTA files. In cases in which the sequence could not be reliably trimmed on the 3' end and/or 5' end the sequence is left untrimmed on that side. The FASTA title is populated with indicators that note which sides were trimmed and whether there are suspected frameshifts, or unspliced introns, noted as gaps.

The reverse complement is produced for sequences in the inverse orientation, i.e., the 3' -> 5' orientation relative to the subject.

A report is created in a tab separated value file that can be readily imported into most spreadsheet programs (Figure 9). The file contains summarized information about the BLAST results of each sequence and the status of the trimming of each sequence relative to its coding region and NAC domain. More specifically, the information provided is the query name, the query description, the name of the best BLASTX hit, the frame of translation, the best hit alignment start position for the subject, the best hit alignment start position for the query, the score of the best hit, the sequence orientation, the total number of hits, whether there is a frameshift or gap, a number indicating the likelihood that the sequence covers the 5' end of the coding region, a color coded indicator noting the integrity of the 5' end trimming (see section 2.5.1.9.2 below), a number indicating the likelihood that the sequence covers the 3' end of the coding region, a color coded indicator noting the integrity of the 3' end trimming (see section 2.5.1.9.2 below), a
textual description of the trimming performed, the full translated sequence, the translated sequence trimmed to the ORF, the original nucleic acid sequence provided, the sequence trimmed to coding region in nucleic acids, any frameshifts noted in the domain, a textual description of the trimming done for the domain, the amino acid sequence of the NAC domain, and nucleic acid sequence of the NAC domain.

2.5.1.9.2 The color coding indicator in the report file

A color coding methodology was created to give a quick indication of the integrity of either end of the analyzed sequences. If the overall sequence length is estimated to cover the beginning of the translated region, and a start codon was found in the expected range, a 'green' color code is issued for the 5' end of the sequence. If a start codon was found outside of the estimated region, but the overall sequence length is estimated to cover the beginning of the coding region, a 'yellow' color code is issued. Lastly, if no start codon was found, or the sequence is not estimated to cover the start of the coding region, and a start codon was not found in the expected range, a 'red' color code is issued.

Similarly, if the sequence is estimated to cover the end of the coding region, and a stop codon was found within the expected range, a 'green' color code is issued for the 3' end. If an end codon was found outside of the expected region, but the sequence is estimated to be long enough to reach the end of the coding region, a 'yellow' color code is issued. If an end codon was not found, or the sequence is not expected to extend to the end of the coding region and a stop codon was not found inside the expected range, a 'red' color code is issued.
| Query | GB hit | Frame | Subject/Query | Strand | Hit Frame | ATG? | 5' End Shift | STOP? | 3' trim | Trim comment | Translation | Raw | ORF | NT seq | NT seq | Aroma | NT | NAC dom | NAC dom |
| SETI_JnNAC37 | gi|320746 | 2 | 1 | 158 | 456 | plus | 5 | None | 5.71 | Green | 5.54 | Green | Seq | YNLTDQEP | MSQ | GTAC | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC23 | ONAC04 | 3 | 1 | 33 | 473 | minus | 5 | None | 3.37 | Green | 7.40 | Green | Seq | EAAYPTDP | MSQ | AGGA | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC29 | ONAC04 | 1 | 1 | 76 | 484 | plus | 5 | None | 4.66 | Green | 7.70 | Green | Seq | AKKNAAAT | MSQ | AGGA | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC10 | ONAC04 | 3 | 1 | 128 | 473 | plus | 5 | None | 5.40 | Green | 7.68 | Green | Seq | SSRSREGD | MSQ | AGGA | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC14 | gi|50542 | 1 | 1 | 163 | 343 | plus | 5 | None | 5.76 | Green | 5.80 | Green | Seq | ***CVALRFT | MSQ | TTAG | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC25 | ONAC04 | 2 | 4 | 263 | 211 | plus | 5 | None | 4.82 | Green | 1.93 | Green | Seq | LSLVYTL | MRR | TCTC | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC16 | ONAC01 | 2 | 15 | 155 | 409 | plus | 5 | None | 1.69 | Green | 3.37 | Green | Seq | PHSSLTTSMS | MGG | ACCA | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC28 | ONAC05 | 1 | 4 | 310 | 205 | plus | 5 | None | 4.71 | Green | 2.19 | Green | Seq | PFFYNPFFMRMEC | CTC | ATGA | None | Both | LGPGFRF | CTGCCG |
| SETI_JnNAC32 | gi|34602 | 3 | 9 | 99 | 309 | plus | 5 | None | 2.02 | Green | 4.99 | Green | Seq | PFCYCVTF | MAV | GTC | ATGA | None | Both | LGPGFRF | CTGCCG |

**Figure 9.** The report file of the ORF detection software.
2.5.1.10 Methods of output analysis

Once the program is run, a report is produced that can be used to determine whether the sequences were trimmed correctly or not.

A 'green' indicator on both sides of a sequence usually means that the sequence covers the entire coding region, and has been trimmed properly. Manual scanning is warranted only in the cases where a suspect frameshift or gap is noted.

A yellow indicator on either side means that the start/stop codons were not detected within the expected range. If the report notes that there is a suspected frameshift or gap, then the unexpected start/stop codon is probably an artifact of the frameshift or unspliced intron. One can also look at the numbers that indicate whether the sequence is predicted to cover the either end of the coding region. If the number on the side with the yellow indicator is only weakly positive, it could be that a start/stop codons found on that side was incorrect, as the sequence does not actually encompass that side of the ORF.

Usually a 'red' indicator for either the 5' end or 3' end means that the sequence does not cover the entire coding region.

Other indicators in the report file, such as the best hit score and the subject alignment start position, can give assurances as to the reliability of the match.

2.6 Phylogenetic tree construction and alignment methods

A phylogenetic comparison was done using the region spanning the first four NAC sub-domains from \textit{T. aestivum} sequences as well as sequences compiled by Ooka et al. (2003) to construct a NAC phylogenetic tree (Ooka et al., 2003). These include NAC sequences from Arabidopsis and rice, as well as \textit{NAM} from \textit{Petunia hybrida}, \textit{SENU5} from \textit{Lycopersicon esculentum}, \textit{TERB} from \textit{Nicotiana tabacum}, and \textit{GRAB1} and \textit{GRAB2} from
*Triticum* sp. (presumed to be *Triticum monococcum*). Amino acid multiple alignments were performed using the Jalview program (Clamp et al., 2004) and utilizing MUSCLE (multiple sequence alignment with high accuracy and high throughput) (Edgar, 2004) to perform the actual alignments. The phylogenetic tree was constructed with the MEGA4 program (Tamura et al., 2007). It was constructed by employing the neighbor joining method (Saitou and Nei, 1987), using the p-distance model for estimating distances between sequences, and using a bootstrap test (Felsenstein, 1985) of 1000 replicates. The pairwise deletion option was selected so that gaps were removed through pairwise comparison. The tree was condensed using a cutoff value of 20%. 
3 Results

3.1 Initial NAC detection and assembly results

ESTs encoding NAC transcription factors from Set 1, Set 2, and Set 3 were detected and processed using the EST gene family detection software pipeline described in section 2.2. Each set was processed on a single PC (In this case a computer with a 1 GHz CPU and 1Gig of RAM) within a few hours. All the wheat ESTs that were automatically detected to be NACs by the software were confirmed to contain the NAC transcription factor domain by manual inspection.

Initially, a total of 100 NAC sequences were automatically resolved from Set 1, the set containing ESTs from the wheat cultivar Norstar (Table 1). Of these, 36 were contigs and 64 singletons. It was found that many sequences failed to assemble together under the stringent assembly criteria employed. The sequences of Set 1 were consolidated further manually resulting in a total of 67 sequences, of which 42 were contigs, and 25 were singletons.

The total number of NAC Set 2 sequences initially resolved through automatic processing amounted to 115 sequences, of which 59 were contigs and 56 were singletons. The automatic processing of Set 3, the set containing GenBank sequences, resulted in a total of 244 NAC-like sequences. Eighty one of these were contigs and 163 were singletons.
Table 1 Initial tally of NAC contigs and singletons.

<table>
<thead>
<tr>
<th>Set</th>
<th>Total</th>
<th>Contigs</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 - Norstar</td>
<td>67 (100 before manual consolidation)</td>
<td>42 (36 before manual consolidation)</td>
<td>25 (64 before manual consolidation)</td>
</tr>
<tr>
<td>Set 2 - Multiple Cultivar With quality</td>
<td>115</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>Set 3 - GenBank</td>
<td>244</td>
<td>81</td>
<td>163</td>
</tr>
</tbody>
</table>

The number represents the total NAC sequences created after the software pipeline was run with an assembly criterion of a minimum identity of 99% over at least a 100 bp sequence overlap.

3.2 End tally of T. aestivum NAC genes

After initial resolution of the wheat NAC sequences, further analysis revealed many to be too short, or were found to contain too many errors to be of use. Others were found to be duplicates between sets. As stated in the materials and methods, the sets were reduced to a unigene set, and preference was given to sequences originating from the sequencing of representative full length cDNA clones from the Set 1, secondly to other sequences assembled with phred quality files from sets 1 and 2, and thirdly to sequences and contigs from Set 3, for which phred quality scores were not available. This greatly reduced the number of sequences, especially among those in Set 3.

Of the Set 1 sequences, 38 Full length cDNA sequences were completed through iterative sequencing, and one additional full length gene sequence was assembled entirely from
existing Set 1 EST sequences (Table 2). Another full length cDNA sequence was assembled from ESTs available from GenBank using an identity threshold of 99%.

Among the three data sets, CAP3 assembly using a high sequence identity threshold (99% identity) yielded an additional 55 partial NAC sequences containing a portion of the NAC conserved domain. Of those, 24 contain all of the first four NAC subdomains, which could be used in conjunction with the full length sequences for the phylogenetic tree construction.

A CAP3 assembly using a lower 96% identity threshold over a minimal overlap of 40 bp of the GenBank sequences that did not assemble with each other at 99% yielded six unique contigs. One of those contigs contains the entire ORF, another contains the first four NAC subdomains, and four contain only the 3' portion of the conserved subdomain. Assembling the ESTs at a 96% identity threshold possibly allowed homeologues, i.e. gene copies from the three wheat genomes, to assemble into single contigs.

A total of 29 EST sequences retrieved from GenBank did not assemble with other sequences under both high and low threshold assemblies, seven of which contain all four conserved NAC sub-domains. The total number of possible NAC homeologues found amounted to 130, counting the contigs assembled at 96% identity as one each.

The sequences of Set 1 and Set 2 generally have quality scores above 30. However, even among the full length Set 1 sequences that underwent iterative sequencing, sporadic bases with a phred quality of less than 20 can be found. Sequences of Set 3 lack quality files and portions of many of the assembled contigs have a depth of only one EST. Also, some sequences are identical to others across large segments (more than 100 bp), emphasizing the possibility that a few cases of sequences representing the same gene may remain.
Table 2 Tally of NAC sequences found among sets.

<table>
<thead>
<tr>
<th></th>
<th>Full Length</th>
<th>Partial length with all first four conserved subdomains</th>
<th>Contains only 5' end of conserved domain</th>
<th>Contains only 3' end of conserved domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 – Norstar</td>
<td>39</td>
<td>8</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Set 2 - Multiple Cultivar With quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 3 - GenBank 99% identity contigs</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Set 3 - GenBank 96% identity contigs</td>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Set 3 - GenBank singletons (presumed low quality)</td>
<td></td>
<td>7</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>32</td>
<td>30</td>
<td>27</td>
</tr>
</tbody>
</table>

This is a unigene set where the same gene sequence does not appear twice. GenBank singletons are listed separately as they are presumed to be of low quality. GenBank sequences assembled at 96% minimum identity may originate from more than one gene.
3.3 Effectiveness of the ORF identification methodology

Seventy five rice NACs were previously analyzed and classified in a phylogenetic tree with Arabidopsis sequences (Ooka et al., 2003). In order to incorporate them into a phylogenetic tree with the wheat NACs, the nucleotide sequences were retrieved from the NCBI Entrez Nucleotide database and translated and trimmed using the ORF detection software. The comparison set used was composed of rice and Arabidopsis protein sequences downloaded from NCBI. The trimmed and translated sequences that had corresponding protein sequences in GenBank were compared to them, and were all found to be trimmed perfectly. The software also trimmed the sequences to the first four NAC subdomains, and the sequences were found to be trimmed perfectly in that respect too.

The ORF detection software found a few frameshifts and/or introns among the rice sequences used in Ooka et al. (2003). These sequences were manually compared to rice proteins in the GenBank Refseq database, rice cDNAs, and rice ESTs to confirm the frameshifts and unspliced introns. ONAC020 (GenBank accession AK064292) from rice was found to contain an intron with a frameshift relative to other cDNA sequences in the GenBank nr and EST databases; the frameshift was additionally confirmed with comparison with the rice genomic DNA sequence. ONAC058 (GenBank accession AK071020) was confirmed to contain an unspliced intron and frameshift. Both ONAC058 and ONAC020 were removed from the phylogenetic comparison as after unspliced intron removal and frameshift repair they were found to be identical to ONAC004 (GenBank accession AK061745) and ONAC071 (GenBank accession AK102475) respectively. ONAC075 (GenBank accession AK102902) had a frameshift which was corrected in our sequence, since the frameshift occurred within the first four sub-domains, and affected
the phylogenetic tree construction.

All the wheat NAC sequences deemed by the software to have good sequence length without error or anomalies were trimmed properly to the entire length of the coding region, and to the conserved domain. The software also helped in the identification of nine *T. aestivum* sequences with frameshift producing errors that were corrected, and three sequences with unspliced introns that were subsequently edited (Table 3). Five sequences labelled with having possible frameshifts were shown to actually have none, and four sequences with gaps could not be shown to contain introns, because there was more than one EST sequence in the est_other database in NCBI that contained the area presumed to be an intron. Two sequences with presumably missing exons, and other sequences that were found to contain multiple frameshifts and were judged to be of too low quality, were removed from the analysis. A missense mutation, not noted in the ORF detection software, in sequence SET2_BE422717, was found and corrected by comparison to two otherwise duplicate sequences, SET2_BQ805514 and SET3_9640_Contig8. It was kept rather than the other two since it was the longest sequence.
Table 3 List of malformed wheat sequences in the analysis.

<table>
<thead>
<tr>
<th>Sequence Id</th>
<th>Software warning issued</th>
<th>Malformation</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET3_9640_Contig132</td>
<td>frameshift</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25235469</td>
<td>frameshift</td>
<td>frameshift</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25264472</td>
<td>frameshift</td>
<td>frameshift</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25420264</td>
<td>frameshift</td>
<td>frameshift</td>
</tr>
<tr>
<td>SET3_Contig74</td>
<td>frameshift</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET2_BQ483881</td>
<td>frameshift</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET2_wdk3c_pk005_d17</td>
<td>frameshift</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET3_9640_Contig132</td>
<td>frameshift</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET3_Contig68</td>
<td>frameshift</td>
<td>frameshift+ intron</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET1_9_Contig6</td>
<td>frameshift</td>
<td>Intron</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET1_taNAC3</td>
<td>frameshift</td>
<td>Nucleotide Insertion was legitimate</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET3_9640_Contig70</td>
<td>frameshift</td>
<td>Not likely frameshift</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25238005</td>
<td>frameshift</td>
<td>Not likely frameshift</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25241105</td>
<td>frameshift</td>
<td>Not likely frameshift</td>
</tr>
<tr>
<td>SET2_Contig60</td>
<td>frameshift</td>
<td>Not likely frameshift</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET1_3_Contig2</td>
<td>gap</td>
<td>intron</td>
<td>Corrected</td>
</tr>
<tr>
<td>SET3_9640_Contig154</td>
<td>gap</td>
<td>intron</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET3_singleton_gi</td>
<td>25441746</td>
<td>gap</td>
<td>Intron</td>
</tr>
<tr>
<td>SET3_Contig60</td>
<td>gap</td>
<td>Intron</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET2_waw1c_pk004_h19</td>
<td>gap</td>
<td>Intron</td>
<td>Unchanged</td>
</tr>
<tr>
<td>SET3_9640_Contig144</td>
<td>Unexpected stop</td>
<td>frameshift</td>
<td>Corrected</td>
</tr>
</tbody>
</table>

The name of the sequence, the type of software warning invoked, the malformation, if any, that was discovered, and what was done with the sequence is shown in the table above.
3.4 Subdomain conservation

Among the five subdomains, subdomains A, C, and D have been shown to be the most conserved between rice and Arabidopsis NAC groups, while domains B and E were less conserved (Ooka et al., 2003). An inter-species alignment of each subdomain of two monocot species, rice and wheat, was performed using the Jalview (Clamp et al., 2004) program, ClustalW (Thompson et al., 1994), and MUSCLE (Edgar, 2004) (Figure 10). It was found that by average identity (excluding amino acid positions in which more than 75% of the sequences contained gaps) domains A and D were most conserved (74.4%, 71%) followed by domains C (66.8%), B (62%), and finally domain E (59.6%). The average amino-acid conservation, based on the AMAS method of multiple sequence alignment analysis (Livingstone and Barton, 1993), of the five domains was also investigated. Here, excluding gapped areas, domains D, E, and A had the highest conservation (6.6, 6.1, and 6) respectively, followed by domain B (5.2), and lastly domain C (4.9). It is interesting to note that domain E had the least identity between sequences, but the second highest conservation. Domains D and E have been previously found to be the DNA binding domain (DBD) of AtNAM, an Arabidopsis NAC (Duval et al., 2002). The similarity of amino acid properties in domain E may allude to similar DNA binding mechanisms among NACs.
(A)

SubDomain A alignment properties

(B)

SubDomain B alignment properties
SubDomain C alignment properties

SubDomain D alignment properties

Percent identity

Amino acid conservation
Figure 10 NAC subdomain amino acid sequence alignments. Alignment of a collection of 111 NAC sequences including 66 sequences from wheat and 55 sequences from rice. ClustalW (Thompson et al., 1994) and MUSCLE (Edgar, 2004) were used in the creation of the alignment. Aligned amino acids that did not have counterparts (contained gaps) in more than 75% of sequences were deleted. Amino acid conservation values were multiplied by ten to bring into proportion with percent identity values. The amino acid sequence depicted in the x-axis is the consensus sequence of the subdomain produced by the alignment.
3.5 Phylogenetic analysis

A phylogenetic tree of sequences containing the first four NAC subdomains was constructed (Figure 11). The analysis included 73 wheat sequences as well as a large set of sequences previously used by Ooka et al. (2003) to create a phylogenetic tree (Ooka et al., 2003). The set includes 55 NAC sequences from *O. sativa*, and 98 *A. thaliana* sequences, as well as NAM from *Petunia hybrida*, SENU5 from *Lycopersicon esculentum*, TERN from *Nicotiana tabacum*, and GRAB1 and GRAB2 from *Triticum* sp. (presumed to be *Triticum monococum*). Some Arabidopsis and rice sequences in the set composed by Ooka et al. (2003) were found to be identical to others across the four domains, and were therefore removed for clarity. These sequences are either actual duplicates, or presumed to be sequences originating from alternative splicing. *T. aestivum* clusters and their rice counterparts can be clearly discerned in most cases.

The tree was divided into 18 subgroups based on groupings previously established by Ooka et al. (2003). These were NAM, NAC1, OsNAC7, NAC2, OsNASC8, ANAC011, TIP, ONAC022, SENU, NAP, AtNAC3, ATAF, OsNAC3, TERN, ONAC001, ANAC063, ANAC001, and ONAC003. Groups OsNAC8 and NAC2 have been broken into a couple of groups each. The breakup of the groups occurred possibly as the result of adding new wheat sequences, or the removal of duplicate Arabidopsis and rice sequences. Wheat sequences are present in every group in which a rice sequence is present, and a wheat ortholog can be found for every rice sequence in groups NAC2, OsNAC8, ATAF, OsNAC3, and ONAC001. Group ANAC011 does not contain any rice sequences, but does contain a wheat sequence, *taNAC33*. A rice gene (GenBank accession ABF93685) that was not included in the Ooka et al. (2003) analysis was found to be a close ortholog.
of *taNAC33* through BLAST comparison. As it was found only much after the phylogenetic tree was constructed it was not included in it. The NAP group contains only one rice sequence, but seven wheat sequences. Five of the Set 1 wheat sequences in the NAP group are clustered together, and are not clearly associated with any Arabidopsis or rice genes. Among those only three, *taNAC7*, *taNAC3*, and *taNAC32*, were classified as homeologues (see section 3.6 below). Curiously, the other two wheat genes in the group are not associated with any rice genes either and are not classified as homeologues. The studied gene in this group, *NAP*, has been found to be a target of transcription factors needed for petal and stamen development in Arabidopsis (Sablowski and Meyerowitz, 1998). The NAM group of genes also contains two clusters of wheat NAC sequences containing more than three sequences each. The NAM group contains a number of previously studied genes, including a studied Petunia gene, NAM, and the Arabidopsis genes *CUC1*, *CUC2*, and *CUC3*, which have been found to be involved in organ development. The group also contains the wheat *GRAB2* gene, which is present in one of the two wheat sequence clusters. The gene has been found to inhibit wheat dwarf geminivirus replication in cultured wheat cells (Xie et al., 1999). It should be noted that both of the clusters in the NAM group contain sequences from Set 3, which are contigs compiled from GenBank sequences, and/or also GenBank singletons, which are not accompanied by quality files, and so strong assertions in regards to their accuracy cannot be made.
Figure 11 A phylogenetic tree of wheat, rice, Arabidopsis, and other NACs of interest. The tree is composed of the wheat NAC sequences presented in this article, and sequences gathered by Ooka et al. and previously presented in a phylogenetic tree (Ooka et al., 2003). Wheat sequences marked with a ‘SET3_singleton’ prefix are sequences that originate from Set 3 but did not assemble with other GenBank sequences and are presumed to be of low quality. Sequences marked with a ‘SET3_9640’ prefix are Set 3 sequences that assembled at a lower threshold of a minimum 96% identity over a minimum overlap of 40 bp, so they may represent more than one gene homeologue. Full length Set1 wheat sequences have a ‘taNAC’ prefix. Rice sequences have an ‘ONAC’ prefix, whereas Arabidopsis sequences have an ‘ANAC’ prefix as named by Ooka et al. (2003). Other sequences contain the name of the studied gene only. Arabidopsis sequences contain their Arabidopsis Information Resource (TAIR) gene model number, and rice sequences contain their GenBank gi number. A bootstrap test (Felsenstein, 1985) of a thousand replicates was performed, and the percentage of replicate trees in which the neighboring sequences clustered together is shown next to the branches. The tree was created with the MEGA4 program (Tamura et al., 2007). It was constructed by employing the neighbor joining method (Saitou and Nei, 1987), using the p-distance model. The pairwise deletion option was selected so that gaps were removed through pairwise comparison. The tree was condensed using a cutoff value of 20%. The wheat sequences are marked with a green diamond, the sequences by which groups were named are marked with a blue dot, and other sequences of interest are marked with a blue triangle. The names of the groups can be seen on the right hand side. The branches of groups are alternately colored pink and blue for clarity.
3.6 *Identity between Triticum aestivum full length homeologues*

The guidelines for classifying homeologue candidates were determined by observing the maximum nucleic acid identity between rice NAC sequences using ClustalW (Thompson et al., 1994). Rice NAC homologues were no more than 91% identical across the conserved subdomain and 82% identical across the whole ORF. Consequently NAC wheat sequences were searched for pairs that had a nucleic acid identity of 91% or higher between them in the region that spanned the first 4 subdomains. Among those, no pair of sequences had less than 95% identity, and so all those with a 95% identity or higher were designated as probable homeologues. Among the full length NACs, 23 genes clustered into 10 sets of probable homeologues. Three contained three members and seven contained two genes each (Table 4).

One set of homeologues consisting of genes *taNAC23*, *taNAC9*, and *taNAC10* was examined in some detail. The identity across the entire ORF was found to be 98% between *taNAC23* and *taNAC9*, 97% between *taANC23* and *taNAC10*, and 96% between *taNAC9* and *taNAC10*. A global nucleic acid alignment, using ClustalW (Thompson et al., 1994), of the three homeologue’s conserved domain, comprising all five sub-domains, revealed 9 locations where a nucleic acid differed between one of the homologues and the other two. However, an amino-acid alignment of the same region showed that all but one were synonymous mutations, alluding to the shared functionality of these proteins, and the importance of the conserved domain in their delegation. The C-terminal half had nine synonymous substitutions for 18 substitutions and a codon deletion, so more than 50% of the nucleic acid mutations caused primary structure changes in the protein.
The alignment of the 3' UTR of the three homeologues showed multiple insertions, deletions and substitutions. The alignment of the 5' UTR was done over the closest 31 nucleic acids upstream of the start codon, as taNAC23 lacked any more sequence. Over that expanse, the three homeologues had almost no identity. Interestingly, however, the 8 bp upstream of the start codon contained 100% identity among all three sequences.

Table 4 The percent nucleotide sequence identity between wheat NAC homeologues that contain the entire ORF.

<table>
<thead>
<tr>
<th>Domain identity</th>
<th>Domain identity</th>
<th>Domain identity</th>
<th>Full length identity</th>
<th>Full length identity</th>
<th>Full length identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/B</td>
<td>A/D</td>
<td>B/D</td>
<td>A/B</td>
<td>A/D</td>
<td>B/D</td>
</tr>
<tr>
<td>taNAC13 / taNAC20</td>
<td>98</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC4 / taNAC12</td>
<td>98</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC14 / taNAC31</td>
<td>98</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC32 / taNAC3 / taNAC7</td>
<td>97</td>
<td>97</td>
<td>97</td>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>taNAC29 / taNAC17</td>
<td>98</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC23 / taNAC9 / taNAC10</td>
<td>98</td>
<td>97</td>
<td>98</td>
<td>98</td>
<td>97</td>
</tr>
</tbody>
</table>
The percent identity across the first four conserved subdomains and across the entire ORF is listed. An A/B in the column heading denotes a comparison between the first two genes listed within each group. A/D and B/D in the column heading denote a comparison between the first and second gene sequence with the third gene sequence respectively.

<table>
<thead>
<tr>
<th></th>
<th>95</th>
<th></th>
<th>96</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>taNAC15/taNAC16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC19/taNAC5/taNAC25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>95</td>
<td>95</td>
<td>94</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>taNAC2/taNAC8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taNAC26/taNAC6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>98</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Discussion

4.1 Performance of the gene family detection pipeline

The software developed in this work proved to be successful in the detection and assembly of ESTs originating from genes of the same family. Differentiating between homeologues in the assembly process proved to be the largest challenge. A 99% identity threshold with a minimum overlap of 100 bp was employed in the assembly to assure that sequences from homeologous genes from hexaploid wheat did not assemble together. These stringent criteria were used because wheat homeologues have been previously found to have 97% identity within the α-tubulin gene family in wheat (Ridha Farajalla and Gulick, 2007). In Set 1 and Set 3 there were many sequences that presumably had a higher error rate than 1% across their potential overlap with other sequences, and so under the stringent criteria were not able to assemble with other ESTs that belong to the same gene.

To further consolidate sequences from Set 1 the results of the initial assembly were assembled again. Contigs that originated from the first assembly were of higher quality than the unassembled sequences from which they originated, so the error rate in some cases dropped sufficiently to enable assembly with ESTs that would not assemble in the first round of contig assembly. Additionally, MegaBLAST with a threshold identity of 99% and a word size of 120 bp was used to identify sequences of the same gene. The sequences that aligned in MegaBLAST could be assembled separately using default assembly thresholds.

These methods worked well for Set 1, however for Set 3, the set containing GenBank
sequences, the process was too time consuming. Instead, sequences were assembled using a lowered threshold identity of 96%. This lowered threshold was judged to allow homeologues to assemble together, but not paralogs, because a ClustalW comparison of known rice NAC domain sequences found the highest identity between any two paralogous genes to be 82%. Another method of increasing the number of sequences that assemble together under stringent criteria is to trim them aggressively with low sequence quality trimming software such as Lucy (Chou and Holmes, 2001). By raising the average window quality from a phred score from 15 to 20, one can decrease the number of incorrect base calls that remain in the sequence, thus significantly decreasing the chances of miss-assembly. The serious drawback of this strategy, however, is that raising the minimal average quality requirements can cause a large reduction in the size of the sequence deemed to be of good quality, resulting in smaller sequences that do not overlap with other sequences, and thus do not assemble together. That is why overlapping Set 2 sequences assembled properly even when using high assembly stringency but many sequences had insufficient overlapping regions with other sequences and remained as singletons. The effect of the shortened sequences was seen when Set 1 and Set 2 sequences were assembled together. The shorter Set 2 sequences in many instances were assembled with other sequence of the same set when they were bridged by the longer Set 1 sequences.
4.2 The ORF detection software

The software for analysis of the translated cDNA sequences was effective in detecting full length sequences and possible errors. The software’s strengths are that it trims and translates sequences, while providing the user with indicators as to the integrity of the determined ORF. It also provides data to help the user to manually determine whether the ORF found is correct.

Unlike other programs that utilize intrinsic analysis methods, such as hidden Markov models, this software uses a comparison set of sequences, and thus will be most useful for the analysis of gene families for which a large number of existing gene models are available. The growing list of whole genome sequences has dramatically increased the number of gene models available in a wide range of taxa.

The software is effective at detecting potential frameshifts, either real frameshifts or due to sequencing error, and unspliced introns in cDNA sequences. This became evident in the detection of such errors in existing cDNA clones and EST sequences. Further testing could be done with known data sets spiked with errors.

Additional functionality can be added to the software. For one, the EST quality files that denote the reliability of each base call in the sequence can be utilized. Bases of low quality can be marked within the trimmed sequence (by using upper case letters for example), and the quality of each base can be taken into account when attempting to determine if a frameshift has occurred. It could also be useful to display the sequence and the quality values of the nucleotide calls in the same view.
4.3 FGAS NACs are more common in particular phylogenetic groups

Set 1 NAC representatives, whose cDNA clones originate from the FGAS project, have higher frequency in particular phylogenetic groups within the NAC gene family. Specifically, the genes were more prevalent in the ATAF1/ATAF2, OsNAC3, and NAP groups.

Members of the ATAF group are associated with stress response, including cold stress (Nogueira et al., 2005; Ohnishi et al., 2005), and the OsNAC3 group is the group most closely related to ATAF. The NAP group is also a relatively close neighbor of ATAF and OsNAC3. The FGAS project contains a higher proportion of ESTs from cDNA libraries made from stressed tissues (Houde et al., 2006).

While frequent appearance of Set 1 sequences in the ATAF, OsNAC3, and NAP groups may be attributed to their participation in stress response pathways, it may also be due to some genes in the group being expressed in more tissue types, across a larger portion of the lifecycle, and/or at higher expression levels than other NAC genes.

Another indication that genes in the ATAF and OsNAC3 groups participate in stress response can be found in a study that used clones from the FGAS database. Monroy et al. (2007) demonstrated that six NAC clones showed significant differences in induction levels between winter and spring wheat cultivars under cold stress (Monroy et al., 2007). A sequence comparison of the six cDNA clones to the new NAC full length sequences reveal that four of the cDNA clones represent genes taNAC15 or taANC16, two homeologues in the OsNAC3 group. Another represents taNAC30 of the NAC2 group, and the last has 96% identity to taNAC6, a member of the ATAF group.
4.4 NAC ortholog groupings allude to their divergence timeline

The groups OsNAC3 and ONAC001 contain at least one wheat gene ortholog for every rice gene present, yet contain no Arabidopsis genes. Groups SENU, AtNAC3, ANAC063, and ANAC001 do not contain any genes from the monocot species analyzed. This suggests that these groups diverged after the separation of monocots and dicots.
5 Conclusion

The software presented in this thesis was well suited for the elucidation of NAC ESTs in wheat, and may be used in the study of other gene families in other species. However, further testing must be performed before its usefulness can be ascertained in a more generalized fashion. It could, potentially, save many man hours when applied to studies of a similar nature to the one presented here.

Resolving wheat sequences accurately proved to be the most trying part of the project, as the small differences between gene homeologues and the relative scarcity of NAC sequences made the task difficult. An easy way around this in future applications of the software would be to pick highly expressed gene families for study, whose sequences would be more abundant within EST libraries. Otherwise, clones can be sequenced multiple times to increase the number of available sequences. A possible strategy to more easily differentiate between sequences with very high sequence identity, such as bread wheat homeologues, would be to use genomic DNA. The introns will presumably be more divergent than the protein coding regions, and thus allow for easier differentiation. During the study enough NACs had been found that contained the first four subdomains to give representation in all phylogenetic groups previously devised by Ooka et al. (2003).

This study has shed light on the NAC transcription factor family in wheat. Despite its large size it remains relatively unknown in this species. It will undoubtedly be the attention of much future research.
6 References


Kihara H (1944) Discovery of the DD-analyser, one of the ancestors of Triticum vulgare. Agric Hortic 19: 13-14


Ohnishi T, Sugahara S, Yamada T, Kikuchi K, Yoshida Y, Hirano HY, Tsutsumi N (2005) OsNAC6, a member of the NAC gene family, is induced by various stresses in rice. Genes Genet Syst 80: 135-139

Olsen AN, Ernst HA, Leggio LL, Skriver K (2005) NAC transcription factors:
structurally distinct, functionally diverse. Trends Plant Sci 10: 79-87


Appendix I – Sequences In Nucleic Acids

7.1 Full length NAC Sequences In nucleic acids

>SET1_taNAC1
AAACTTTCCCCAATCCAATCTCCCAATCCAATCTTCCCTCTCCTGGAACCCTAAG
CCTCCATGTTACATGGACCCACCTTTTCTCTCCTCTCTCTCTCTGGCCCGCCGGCCGC
GGATGACCCGACCTCCCTTGCCCCGCTTCCGCTTCCACCCAACAGACGAGGAGCTCGT
CTCCTACTACCTCAAGCGCAAGGTGCTCGGTCGCCCGCTCAAGGTCGACGCCATCGCTG
GGATGACCTCTACAAGGTCGAGCCATGGGACCTCCCGGCAAGGTCGCGCCTCCGCAGCCG
CGATTCCCAGTGGTACTTCTTCAGCCGCCTCGACCGCAAGCACGCCAACCGCGCGCAC
CAACCGCGCCACTGCGAGGGCTACTGGAAGACCACCGGCAAGGACAGGGAGGTCCGCCA
TGGAGCTAGGGTTGCAGGAGTGAAGAAGACGCTCGTCTTCCACGCTGGCCGCGCCCAA
GGCCAGCGGCACCAACTGTTGCTACGGACGATGTACGCTGGTGGCAAAATATG
GATGCTGGATTCACTCTTACAGGAGTCTCCATATGCTGCTTCTGCTGCTTCTGGTGG
GATGGCCTAAATGTATGGCATTGCATGACATGACCTCGTGCGTCTGTTAAATTCGACA
GCATTAGCGGTACTGTTTTGCAGCCCGTGAGTTAGTGCTGTATCAATGAATTCCTCCAGT
GATATTAGTATTTTAGAACTGCTTAAAAAAAAAAAAAAAAAAAAAAA

>SET1_taNAC2
TCTCTCCCTAGTCTACACCCTCTCCCTCTCTGTCCCTCTCTGCAAGTCGCCATGAACA
TTCCAAGCAAATGCACGGCTCATGGGCGCCACTTTGCACTTGCAAAGTGAGATAGGAAGC
TGCGACCTGCCCTGCTGCATTTGCCACTGCCGTGCAAGGACGTCCCCTGTGAACTACAAG
CTACCTAGCTAGTGCGGCTATTAGTTCGTACGTACGTGCTCGAGCCACTGAGCTGAGCTA
GTTTGTGCGTGCGTGCTATTACGAGCTGTTGCCAGGTGGTCAGCGGCAGCTAGCCATGGA
TTTGGAGTTACCTCTTACAGGAGTCTCCATATGCTGCTTCTGCTGCTTCTGGTGG
GATGGCCTAAATGTATGGCATTGCATGACATGACCTCGTGCGTCTGTTAAATTCGACA
GCATTAGCGGTACTGTTTTGCAGCCCGTGAGTTAGTGCTGTATCAATGAATTCCTCCAGT
GATATTAGTATTTTAGAACTGCTTAAAAAAAAAAAAAAAAAAAAAAA

66
CATCGACGGCTCGCTCATCTTCGACCCGCCGTCGCCGGGCGCCTTCGCCCGCGACGCCGC
CGCGTTCGGGGACATGCTCACGTACCTGCAGAAGCCGTTCTGAATGAACGCGGCATCCGT
CAGACCCCTCCTCCTTACGACCTCACCACCTACATGTTGCTGACGCTGCTGTAATTGCTG
CTGCAAGCTTGGAAAACAAATTACAGATATTAGAGAAAAAAAGAAAAAAAAAAAAAAAAAA

>SET1_taNAC7
CGAGATAGATTCTGGTCTGGTCCAGTGTGTGTAAGCTTCTCCCTTCTTCTGATCGTGTGT
GTGTTAGAGATTCTGGTCTGGTCCAGTGTGTGTAAGCTTCTCCCTTCTTCTGATCGTGTGT
GTGTTAGAGAGCTTGGGTGGTTGCACGTACGAGGTACGTGCGTGTCGATCGGGTGTCGA
TCATGGCGATGGCGCAGGGGCAGGGGCAGGGGGCGGCGACGTCGCTGCCGCCGGGGTTCC
GGTTCCACCCGACGGACGAGGAGACGTCACTCTGCACCTTGCACAAACGCGCCCGCGCC
CGCCGCGCCGCTGCTCCATCATCGCCGACGTAGACATCTACAAGTTCGATCCGTGGGACC
TCCTTCTACATTACGACCACCCCATCCCATTCACATGTTGCTGACGCTGCTGTAATTGCTG
CTGCAAGCTTGGAAAACAAATTACAGATATTAGAGAAAAAAAGAAAAAAAAAAAAAAAAAA

>SET1_taNAC8
CCTCCCTACTACAATCCCCCCATCTCTCTCTCCGTAGTCTACACCCTCTCCCTCTCTCTG
TCTCTCTCTGCAAGTCGCCATGAACATTCCAAGCAAATGCACGGCTCATGGGCGCCACTT
TGCACTTGCTCAAGAGGATTCTGGTCTGGTCCAGTGTGTGTAAGCTTCTCCCTTCTTCTGATCGTGTGT
GTGTTAGAGATTCTGGTCTGGTCCAGTGTGTGTAAGCTTCTCCCTTCTTCTGATCGTGTGT
GTGTTAGAGAGCTTGGGTGGTTGCACGTACGAGGTACGTGCGTGTCGATCGGGTGTCGA
TCATGGCGATGGCGCAGGGGCAGGGGCAGGGGGCGGCGACGTCGCTGCCGCCGGGGTTCC
GGTTCCACCCGACGGACGAGGAGACGTCACTCTGCACCTTGCACAAACGCGCCCGCGCC
CGCCGCGCCGCTGCTCCATCATCGCCGACGTAGACATCTACAAGTTCGATCCGTGGGACC
TCCTTCTACATTACGACCACCCCATCCCATTCACATGTTGCTGACGCTGCTGTAATTGCTG
CTGCAAGCTTGGAAAACAAATTACAGATATTAGAGAAAAAAAGAAAAAAAAAAAAAAAAAA
>SET1_taNAC9
GCCAAAGAAGGATTTACAGCAAGGAGGTGAGGAGGAGCTGAGGCAGACTCCATGATCCATC
ACGATGCAAGCCAGGAGGAAGTACGAGGACGCTGAGGCAGACTCCATGATCCATC
CAAGCCAGGAGGATCGGATGCGGAGGAGCTGAGGCAGACTCCATGATCCATC
ATCAGGTTTCCCTTTCATTTATTTTTATTTTTTTTT

>SET1_taNAC10
AGAGCAGCAGCAGAAGAGAAGGAAAATTCAGTAGCAGTAGAACTCAGAGCAGCGGCAGCA
GAACTGCGGCGAGCTGAGGAGGAGCTGAGGCAGACTCCATGATCCATC
ACGATGCAAGCCAGGAGGAAGTACGAGGACGCTGAGGCAGACTCCATGATCCATC
CAAGCCAGGAGGATCGGATGCGGAGGAGCTGAGGCAGACTCCATGATCCATC
ATCAGGTTTCCCTTTCATTTATTTTTATTTTTTTTT

70
7.2 Other NAC gene sequences that do not contain the entire ORF
GACTCTAATGTAACTGACATTGTGTTGTATAGCTTGTGCTGTCTTCCTCGGGCCCCACAGGGCTGAGTCACTCGATGTAGAGAATGTTGGAAATATTAGCATTTTTCCGATTA
>SET1_Contig1
CCCTCTCAGAAACTTTCCTCCAAATCTCCAAATTCCAAATTCCTCCCTCTCTCTCTCTGG
ACACCTAAGCCCTCCTCAGGTACATGACCCACCCCTTCCCTCCTTCCTCCTCCTCCTCCGCGCCGC
GGCCCGCGCGGAGTACCCCGACCTCCCTTTGGCCCGGCTTTGGCTGCCCCCCTCCACCGAGAGGAG
GAAGTCTCCTCTCTACTAAGGCGAGGGTGCTCGTGCCCGCCGCTCTAAAGGGCACAGCC
CATCGTCTGAGATTGACAATTCTACAGAGTGGACATTTGGACCTCAGCTCCGCCAGGTGGCCGCT
CCGACAGCGCGACTCCAGGTGATCTCTCTCAGCCGCTAGCAGCGACGGGACGCGGACACCG
CGCCGCCAACACGCCGCACACTGAGGCTAGCTAGCTAGCTAGCTAGCTAGC

>SET1_Contig15
ACAAAAAAGCAGGCTGTACCCGGTGCCAATTCCCGGGATATCGTCGACCCACGCGTCG
CATACTCTAGCATCACACACCGCTCAGGCTACAGCAACATAGATATATGCCCTGCCTAAT
CCCTGCTACCTAGCTAGATATATACCTAGCTTCTTTCTCTCCTCTCTCGAGATCTTGTCG
ATCTCTTTCTCCTCCGTCTACCTAGCTAGCGCAATTGGTGCCCTTATAGGCAAGGCAAC
GGCATCCATCCAGATCTCTCTCCTATACCATGAAATTCAATTTCAAGCTAATGAA
tTGAGAGGAGATCGGACGTGAGCTGAGCCAGGAGGTGGTTTCAGAGAGAAGATAGTGAACAG
ATAGAGAGATAAGCATTTCCATGGGGCTGCGGGACATCGAGATGACGCTGCCGCCGGGT
TCCGCTTCTACCCGAGCGACGAGGAGCGTGACTACCTGCACGGCAAGGTGGCCA
ACCAACGCTTCGCCGGAGGAGCCACCGGGACCATGGTGGAGGTCGATCTGCACGTCCATG
AGCCATGGGAGCTGCCAGATGTGGCGAAGCTGAGCACAAACGAGTGGTACTTCTTCAGCT
TCCGCGACCGCAAGTACGCGACGGGGCTGCACAAACCGCGCCACCAGGTCCGGCTACT
CTAGATAGCTATCCCTCGAGGAACAGTTACCCGTTACCC

>SET1_Contig28
ACAAAAAAGCAGGCTGTACCCGGTGCCAATTCCCGGGATATCGTCGACCCACGCGTCG
CATACTCTAGCATCACACACCGCTCAGGCTACAGCAACATAGATATATGCCCTGCCTAAT
CCCTGCTACCTAGCTAGATATATACCTAGCTTCTTTCTCTCCTCTCTCGAGATCTTGTCG
ATCTCTTTCTCCTCCGTCTACCTAGCTAGCGCAATTGGTGCCCTTATAGGCAAGGCAAC
GGCATCCATCCAGATCTCTCTCCTATACCATGAAATTCAATTTCAAGCTAATGAA
tTGAGAGGAGATCGGACGTGAGCTGAGCCAGGAGGTGGTTTCAGAGAGAAGATAGTGAACAG
ATAGAGAGATAAGCATTTCCATGGGGCTGCGGGACATCGAGATGACGCTGCCGCCGGGT
TCCGCTTCTACCCGAGCGACGAGGAGCGTGACTACCTGCACGGCAAGGTGGCCA
ACCAACGCTTCGCCGGAGGAGCCACCGGGACCATGGTGGAGGTCGATCTGCACGTCCATG
AGCCATGGGAGCTGCCAGATGTGGCGAAGCTGAGCACAAACGAGTGGTACTTCTTCAGCT
TCCGCGACCGCAAGTACGCGACGGGGCTGCACAAACCGCGCCACCAGGTCCGGCTACT
CTTCGCCATGGCACACCTTCAAGTCCAACAACACCAACACCAACTAGAGCTCCC
TTGGGGGTAGTCCACCCCTACCCAAAGAGATGTTGACATGGTGATGGCCCATC
GGTGCTCAACCCAACCTTCACCGCGATAGCGATCGCGGAGGTGGATCTGAATAAGAACGA
CCCGTATGAGCTCCCCAAGAAGGCAAAGATGGGGGAGAAGGAGTGGTACTTCTACTGTCA
GAAGGCCACTGGCAAGGACAAGGAGATCTTCCACCCACTGCCCACTCTCATCGGCATGAA
GAAGACGCTCGTGTTCTATAAGGGCAGGGCGCCTAGGGGGGAGAAGACCAACTGGGTCAT
GCACGAGTACAGGCTTGAGATCAGCAAGCAGCCTGCATATGGCCCATCTACCGCCATCGC
CAAAGCTGCCGCCATTAATGCATCTTCCAAGAAGGAGCGTGTGGTTTGCAAAATCTTCCA
TAAGAACATTGGAGTAAGAAGGGTGACCGCCCTGTATACCGCATCTCACTCCATGCAA
CTAGAGGAGACAGCAGCAACATATTTCCCTTCAGCACTTACGCTGCCTGCCATAGGA
CTAGGGCGCATGTGTCGTCGCTGGCGCCTCCGTTGTTTCTATCTGAAGATTCTTCGTACCA
ATTGCATGCCGCAGAATCGGCTCCCGTACGAGGCCGAGCTTGTGGTGTCCCACTCATGGAA
>SET3_Cntg48
ATCTCTTCCTCTCGGAATGAAAGAGACGCTGCTCTTCTTCTACTTAAGC
CTCCCCCCCGGGGAGGGGCGTCCTTGTCGGAATGAAGAAGACGCTCGTCTTCTACCTATAGC
AAGGCCCATACCGCGCCACAGGAGATGTTGACATGGTGATGGCCCATC
GGTGCTCAACCCAACCTTCACCGCGATAGCGATCGCGGAGGTGGATCTGAATAAGAACGA
CCCGTATGAGCTCCCCAAGAAGGCAAAGATGGGGGAGAAGGAGTGGTACTTCTACTGTCA
GAAGGCCACTGGCAAGGACAAGGAGATCTTCCACCCACTGCCCACTCTCATCGGCATGAA
GAAGACGCTCGTGTTCTATAAGGGCAGGGCGCCTAGGGGGGAGAAGACCAACTGGGTCAT
GCACGAGTACAGGCTTGAGATCAGCAAGCAGCCTGCATATGGCCCATCTACCGCCATCGC
CAAAGCTGCCGCCATTAATGCATCTTCCAAGAAGGAGCGTGTGGTTTGCAAAATCTTCCA
TAAGAACATTGGAGTAAGAAGGGTGACCGCCCTGTATACCGCATCTCACTCCATGCAA
CTAGAGGAGACAGCAGCAACATATTTCCCTTCAGCACTTACGCTGCCTGCCATAGGA
CTAGGGCGCATGTGTCGTCGCTGGCGCCTCCGTTGTTTCTATCTGAAGATTCTTCGTACCA
ATTGCATGCCGCAGAATCGGCTCCCGTACGAGGCCGAGCTTGTGGTGTCCCACTCATGGAA
>SET3_Cntg50
GTCGCTGCGGCGATTGAGGATGTGAACCTCAACAAGTACGAGCCATGGGAGCTACCGGAG
AAGGCGAAGATGGGGGAAGAAGGAGATGTTGACATGGTGATGGCCCATC
GGTGCTCAACCCAACCTTCACCGCGATAGCGATCGCGGAGGTGGATCTGAATAAGAACGA
CCCGTATGAGCTCCCCAAGAAGGCAAAGATGGGGGAGAAGGAGTGGTACTTCTACTGTCA
GAAGGCCACTGGCAAGGACAAGGAGATCTTCCACCCACTGCCCACTCTCATCGGCATGAA
GAAGACGCTCGTGTTCTATAAGGGCAGGGCGCCTAGGGGGGAGAAGACCAACTGGGTCAT
GCACGAGTACAGGCTTGAGATCAGCAAGCAGCCTGCATATGGCCCATCTACCGCCATCGC
CAAAGCTGCCGCCATTAATGCATCTTCCAAGAAGGAGCGTGTGGTTTGCAAAATCTTCCA
TAAGAACATTGGAGTAAGAAGGGTGACCGCCCTGTATACCGCATCTCACTCCATGCAA
CTAGAGGAGACAGCAGCAACATATTTCCCTTCAGCACTTACGCTGCCTGCCATAGGA
CTAGGGCGCATGTGTCGTCGCTGGCGCCTCCGTTGTTTCTATCTGAAGATTCTTCGTACCA
ATTGCATGCCGCAGAATCGGCTCCCGTACGAGGCCGAGCTTGTGGTGTCCCACTCATGGAA
>SET3_Cntg53
GTCGGGGCTACGTCCCGGCGCCGGCGCGCGCGCGCTCTGCGCGCCACCACAGATCATATACAA
TGGAAACCTCTACGTCAATGGAACCTCCACCGGCTTGGCAGGCTTGGCGGAGGCTTGGCAGG
AAGGCGAAGATGGGGGAAGAAGGAGATGTTGACATGGTGATGGCCCATC
GGTGCTCAACCCAACCTTCACCGCGATAGCGATCGCGGAGGTGGATCTGAATAAGAACGA
CCCGTATGAGCTCCCCAAGAAGGCAAAGATGGGGGAGAAGGAGTGGTACTTCTACTGTCA
GAAGGCCACTGGCAAGGACAAGGAGATCTTCCACCCACTGCCCACTCTCATCGGCATGAA
GAAGACGCTCGTGTTCTATAAGGGCAGGGCGCCTAGGGGGGAGAAGACCAACTGGGTCAT
GCACGAGTACAGGCTTGAGATCAGCAAGCAGCCTGCATATGGCCCATCTACCGCCATCGC
CAAAGCTGCCGCCATTAATGCATCTTCCAAGAAGGAGCGTGTGGTTTGCAAAATCTTCCA
TAAGAACATTGGAGTAAGAAGGGTGACCGCCCTGTATACCGCATCTCACTCCATGCAA
CTAGAGGAGACAGCAGCAACATATTTCCCTTCAGCACTTACGCTGCCTGCCATAGGA
CTAGGGCGCATGTGTCGTCGCTGGCGCCTCCGTTGTTTCTATCTGAAGATTCTTCGTACCA
ATTGCATGCCGCAGAATCGGCTCCCGTACGAGGCCGAGCTTGTGGTGTCCCACTCATGGAA

95
8 Appendix II – Translated sequences

8.1 Full length sequences trimmed to the ORF

>SET1_taNAC1
MTHSSSSSAPAAPDDPTSALPGFRFHPTDEELVSYLKRKLGRPLKVDAIAEVLY
KVEWDLPIRSLRSRQSOWYSFFSLRDLKNKARTNRTATAGGGYWKITGKDREVHRGAV
VGMKKTLYVFGHAPKERTNMWNVHRELDGAAGIPQDSFFVCRIFKAKGGPGQNGAQ
YGAPFYFEVEWAEDEDDDFGMPFDRQVDVGEHEAPGAMEKYGQYMNDTQDLAGQENGVS
VLPSVDTSTSSSHESTCNSGDIIDLSGLSNFLQYIHPGQNSPMLENLSANVGD
FLNNSSEPDELKDELPIGLNDSTIWPSDGAWKTAFLDAVNGANENVPLITGQDFP
QDPELQAQLQLQDSSQILGSTMTDPLHHSTNNSVPTEDDILMVYATPFDNSMFDQFQ
TNGFLGSPATILSGIETLDGTPYDYADMDNLFNDMCSVQPSAGSSSHVFNPVLTQEV
NNPNTYSPFTQKVQPFPVAGAPSSRSLSEAGQGQLNCVVLPSQARMGSMCRFKM LDS
ISAPAPAAMEFPKSLGSVHPNTISVSTEVISHSLTVASQKWSQFKDMEMLLEFSTG
FQPDRNHRCCGGNTVTAVLGFGRFPLFLSSAIMLLLVYEVLCYIGK*

>SET1_taNAC10
MSGGQELNLPFGFRFHPTDEELVTHLCCRAGAPIAVPITEIDLYKFDPWQLPKMALY
GEKEWVFSPFDRKYPNGSRPNRAAGSYWKATGDADFVTPKPLAIKALKFLFYAGKAPK
GEKTNMIHYEMLADVDRSARKKNSLDDLWLCRIYKQCGKAPSAVDRKPAAGGGYG
GPPCAMVSSPQEKCVMGNANGGGQVTFDFAAYDPRDSMPRLHADSCECQVLSSP
DFPAGEVQSPKISEWERSFASGSGDFNVPAAGSMLEPNQGGFDPLQDLIMYWGKFP*

>SET1_taNAC11
MMSSDFAMLPPGFRFHPTDEELILHYNRRAESPCPSVLIIADVIYKFDPWALKPSASY
GERYVFSPFDRKYPNGSRPNRAAGSYWKATGDADFVTPKPLAIKALKFLFYAGKAPK
PFKPIKTNMIHYELAAADGHAYNGTRMPKFRNASMDLDDLWLCRIYKQCGKAPSAVDRKPAAGGGYG
GPPCAMVSSPQEKCVMGNANGGGQVTFDFAAYDPRDSMPRLHADSCECQVLSSP
DFPAGEVQSPKISEWERSFASGSGDFNVPAAGSMLEPNQGGFDPLQDLIMYWGKFP*

>SET1_taNAC12
MSGGQELNLPFGFRFHPTDEELVTHLCCRAGAPIAVPITEIDLYKFDPWQLPKMALY
GEKEWVFSPFDRKYPNGSRPNRAAGSYWKATGDADFVTPKPLAIKALKFLFYAGKAPK
GEKTNMIHYEMLADVDRSARKKNSLDDLWLCRIYKQCGKAPSAVDRKPAAGGGYG
GPPCAMVSSPQEKCVMGNANGGGQVTFDFAAYDPRDSMPRLHADSCECQVLSSP
DFPAGEVQSPKISEWERSFASGSGDFNVPAAGSMLEPNQGGFDPLQDLIMYWGKFP*

>SET1_taNAC13
MEHEQEQQHAMLDPGFRFHPTDEELILHYNRRAESPCPSVLIIADVIYKFDPWALKPSASY
GERYVFSPFDRKYPNGSRPNRAAGSYWKATGDADFVTPKPLAIKALKFLFYAGKAPK
PFKPIKTNMIHYELAAADGHAYNGTRMPKFRNASMDLDDLWLCRIYKQCGKAPSAVDRKPAAGGGYG
GPPCAMVSSPQEKCVMGNANGGGQVTFDFAAYDPRDSMPRLHADSCECQVLSSP
DFPAGEVQSPKISEWERSFASGSGDFNVPAAGSMLEPNQGGFDPLQDLIMYWGKFP*

>SET1_taNAC14
MTVMELKALPPGFRFHPTDEELVHLYLHIKAVDARFALAVLAVDNKCEPWLPS
LARMGEKEWVFPPCLDRKYPFTLNRATEGYWKATGGKDIFRCKGALVGMKTLVFFY
GRAPKGDKGKSWVMHYNLAAHASTGSSLSGSGRAASSKENWLVICRFKLSLVYGVVS
SAPASSACMKIGAVCECMIECCSSVADPTPLPLDLMSGCSSGASFVWPAHITCFSSNALEAG
QFNPTATAGCGAITADHGLASSYPLASFTQYGGQLHLGVLVLSLQLSSEJYHRGLAD
DMACPGNOQQQRACKGERLSASQDTGLTSDVNPHEISSFGQTFDHEPAWYG*