Analysis of the cDNAs of Hypothetical Genes on Arabidopsis Chromosome 2 Reveals Numerous Transcript Variants

Yong-Li Xiao*, Shannon R. Smith, Nadeeza Ishmael, Julia C. Redman, Nihkil Kumar, Erin L. Monaghan, Mulu Ayele, Brian J. Haas, Hank C. Wu, and Christopher D. Town

The Institute for Genomic Research, Rockville, Maryland 20850

In the fully sequenced Arabidopsis (Arabidopsis thaliana) genome, many gene models are annotated as “hypothetical protein,” whose gene structures are predicted solely by computer algorithms with no support from either expressed sequence matches from Arabidopsis, or nucleic acid or protein homologs from other species. In order to confirm their existence and predicted gene structures, a high-throughput method of rapid amplification of cDNA ends (RACE) was used to obtain their cDNA sequences from 11 cDNA populations. Primers from all of the 797 hypothetical genes on chromosome 2 were designed, and, through 5’ and 3’ RACE, clones from 506 genes were sequenced and cDNA sequences from 399 target genes were recovered. The cDNA sequences were obtained by assembling their 5’ and 3’ RACE polymerase chain reaction products. These sequences revealed that (1) the structures of 151 hypothetical genes were different from their predictions; (2) 116 hypothetical genes had alternatively spliced transcripts and 187 genes displayed polyadenylation sites; and (3) there were transcripts arising from both strands, from the strand opposite to that of the prediction and possible dicistronic transcripts. Promoters from five randomly chosen hypothetical genes (At2g02540, At2g31270, At2g33640, At2g35550, and At2g36340) were cloned into report constructs, and their expressions are tissue or development stage specific. Our results indicate at least 50% of hypothetical genes on chromosome 2 are expressed in the cDNA populations with about 38% of the gene structures differing from their predictions. Thus, by using this targeted approach, high-throughput RACE, we revealed numerous transcripts including many uncharacterized variants from these hypothetical genes.

The first plant genome project, the whole-genome sequencing of Arabidopsis (Arabidopsis thaliana), was completed at the end of 2000 (Arabidopsis Genome Initiative, 2000). However, the goal of the project is not only the collection of the complete sequence of the Arabidopsis genome but also understanding the function of each gene. This is the major goal of the National Science Foundation 2010 projects, and this challenge has been taken up by many Arabidopsis researchers (Chory et al., 2000). Identifying each gene in the fully sequenced Arabidopsis genome and uncovering their function will provide crucial information for biologists to understand plant physiology, genetics, development, and evolution. One approach to achieve this goal, which has been taken by several groups (Ceres, Stanford Genome Center, Salk Institute, Plant Gene Expression Center, RIKEN, and Institut National de la Recherche Agronomique/Genoplante), is to obtain full-length cDNA sequence for every gene in the genome by large-scale sequencing.

The group of genes/gene products with the most obscure functions, from the annotation of the Arabidopsis genome produced by The Institute for Genomic Research (TIGR), is called “hypothetical proteins.” These genes and their products are predicted solely by computer algorithms such as Genscan (Burke and Karlin, 1997), Genemark.hmm (Lukashin and Borodovsky, 1998), and various splice-site prediction programs (Uberbacher and Mural, 1991; Hebbsgaard et al., 1996; Brendel and Kelle, 1998). At the time of genome completion, 5,690 of the annotated genes were designated as “hypothetical.” However, with increased database content and improved annotation methods, this number decreased over time so that the final TIGR annotation release (Version 5, February 2004) contained only 2,248 hypothetical proteins among the 26,207 annotated open reading frames (ORFs).

The lack of expressed sequence tag (EST)/cDNA matches from Arabidopsis or protein homologs from other species indicates that there is no experimental evidence for the expression, structure, or function of these genes and also suggests that if they are expressed, their expression is likely localized, transient, at low levels, or under specific conditions. Thus, large-scale, undirected cDNA cloning and sequencing projects may have less chance of recovering them. In contrast, we applied a more specific and sensitive PCR-based targeted...
approach, high-throughput RACE (Frohman et al., 1988), to investigate the expression of all hypothetical genes on chromosome 2 under a variety of growth conditions to obtain cDNA sequences of expressed genes. Although several reports based on amplicon or oligonucleotide array hybridization and massively parallel signature sequencing (MPSS) showed that many genes annotated as “hypothetical proteins” were expressed (Kim et al., 2003; Yamada et al., 2003; Meyers et al., 2004; Redman et al., 2004), their cDNA sequences could not be recovered by these methods. In contrast, the cDNA sequences of expressed hypothetical genes can be obtained by our approach and are of great value for validating the predicted gene structures and laying a foundation for the analysis of their functions in vivo. In a previous pilot study, 138 out of 169 hypothetical genes tested were expressed in our cDNA populations with full-length cDNA sequences obtained from 16 hypothetical genes analyzed (Xiao et al., 2002). In this study, we developed a high-throughput RACE method and examined all the hypothetical genes on Arabidopsis chromosome 2 (797 when the work was initiated). We expanded our cDNA populations from six to 11 to increase the likelihood of detecting expression of these hypothetical genes. As a result, we recovered numerous transcripts from annotated hypothetical genes including many examples of alternate splicing and multiple polyadenylation sites, transcripts arising from both strands, transcripts arising only from the strand opposite to that of the prediction, and possible dicistronic transcripts. These results provide a statistically more significant data set for comparing genes structures predicted ab initio by computer programs with those supported by experimental evidence. To begin an investigation of the function of some of these hypothetical genes, we generated green fluorescent protein (GFP) and β-glucuronidase (GUS) gene reporter constructs of five randomly chosen genes (At2g02540, At2g31270, At2g33640, At2g35550, and At2g36340). The expression patterns of these five promoter-reporter constructs display highly localized tissue and/or developmental stage specificity, providing an explanation for their low abundance in all EST collections sequenced to date. To our knowledge, this is the first large-scale study focusing on annotated hypothetical genes by high-throughput RACE.

RESULTS

Generation, Sequencing, and Analysis of RACE Products

In this study, we tested all of the 797 genes from chromosome 2 that were annotated as hypothetical (Supplemental Table I). To increase the RACE success rate, the cDNA populations were expanded from the six tissues/treatments in our previous study (Xiao et al., 2002) to a total of 11 populations including tissue exposed to cold; heat; 2,4-dichlorophenoxyacetic acid (2,4-D); hydrogen peroxide (H₂O₂); UV; indole-3-acetic acid (IAA); salt and pathogens (Xanthomonas campestris pv campestris and Pseudomonas syringae); and tissue from callus, roots, and young seedlings, as well as sampling at different time points for some treatments (see “Materials and Methods”). To assess the transcript diversity of this set of samples, each RNA was labeled and hybridized to the whole-genome Affymetrix chip designed by TIGR (Redman et al., 2004). Collectively, 17,501 genes were assigned a “present” call in one or more tissues and 4,880 genes were not detected in any RNA samples by Microarray Analysis Suite 5.0 (J.C. Redman and C.D. Town, unpublished data).

RACE reactions were performed on pooled cDNA populations. Nested RACE reactions were performed after first-round RACE to increase specificity, and their products were cloned. For each targeted gene, 12 colonies from each end (5’ or 3’ RACE) were screened by PCR using gene-specific primers, and positive clones were sequenced. Among the 797 hypothetical genes on chromosome 2, insert-positive clones were obtained from 506 genes (Supplemental Table I). These were sequenced from both ends and produced a total of 11,237 good reads. The sequences clustered and assembled using the Program to Assemble Spliced Alignments (PASA; Haas et al., 2003), which first aligns each sequence read to the finished Arabidopsis genome and then assembles each cluster of overlapping reads into a minimal number of transcript isoforms consistent with the collection of sequences. The entire flow scheme and outcome of the RACE process is shown in Figure 1. Only 399 targets yielded clones whose sequences matched the initial target gene and can thus unambiguously be called “expressed” in our new cDNA populations (Fig. 1; Supplemental Table I). For the remaining 107 RACE-positive target genes, either no sequence was generated or the sequences did not match the correct target. Among the 4,880 genes not detected on the Affymetrix chip, 327 genes were tested in our RACE experiments and 160 were found to be expressed. As expected, our PCR-based method is more sensitive than Affymetrix hybridization.

Analysis of Gene Structures

Assemblies Aligning to Individual Target Loci

As shown in Figure 1, there are 589 PASA assemblies that align to 366 single-target genes with 40 assemblies matching regions that span more than one locus (see below). Among 366 single-target matches, 215 (58.7%) genes have at least one transcript whose structure is identical to its prediction over the alignment region. Of these, the structures of 181 genes with ORFs are the same as predictions, adding only 5’ and 3’ untranslated regions (UTRs) to the original prediction (Fig. 2A). For the other 34 genes the RACE assemblies do not cover the entire predicted ORF, suggesting that they may not represent the complete transcript. Of the 215 matched genes, 108 consist of single exons.
The expressed sequences from the other 151 genes (41.3%) reveal gene structures that are different from their predictions, with 11 genes having predicted start codons within regions that are actually introns (Fig. 2B) and 26 genes with predicted stop codons in regions that are actually introns (Fig. 2C). Sixty-seven genes display intron/exon boundary or boundaries that are different from their predictions (Fig. 2D). Thirty-seven gene predictions fail to identify existing exon/exons (Fig. 2E), and three genes are predicted with additional exons (Fig. 2F). The predicted introns from 71 genes are not supported by the cDNA sequences obtained (Fig. 2G), and 12 genes display introns within their predicted exons (Fig. 2H). For a particular gene whose structure is different from its prediction, there are usually discrepancies in more than one of the above categories. Specific examples of each type of discrepancy are shown in Figure 2.

Alternate Splicing and Polyadenylation. In our analysis, if any one of a gene’s splice isoforms is consistent with the prediction, the gene is flagged only as “alternatively spliced transcript,” not “different from prediction.” Among our target genes, alternatively spliced transcripts were found from 116 target regions. The alternative transcripts can be classified as showing an alternative donor site (31 cases), an alternative
acceptor site (37 cases), unspliced intron (71 cases), and exon skipping (six cases). For a particular gene, the alternatively spliced transcripts may fall into two or more categories. For example, there are three different assemblies from gene At2g29930 that represent alternatively spliced transcripts (Fig. 3). The differences among them are in the structure of the first intron. One assembly (assembly 1) has an intron in the 5’ UTR and an in-frame ATG 21 bp upstream of the predicted initiation codon (ATG site). Therefore, the ORF encoded by assembly 1 is seven amino acids larger than predicted. However, assembly 2 has an unspliced first intron that does not cause a change in the encoded protein, while assembly 3 displays a different intron acceptor site creating a larger intron that results in a shorter protein. The longest ORF in each of the three transcripts is different from the predicted ORF, but all are in the same reading frame. Overall, in our study, about 29% (116/399) of the hypothetical genes have alternatively spliced transcripts.

In addition, polyadenylation sites were identified in 187 genes with multiple polyadenylation sites from 83 genes. The distance between different polyadenylation sites was from 1 to 543 bp. Only one poly(A) was found from 104 genes; two poly(A)s were found from 41 genes; three poly(A)s were found from 42 genes.

**Figure 2.** Examples of comparison of experimentally derived cDNA sequences to the predictions, A to H. Different types of relationships between the predicted gene structure (identified by AGI identifier: At2g#####) and that inferred from RACE sequence (Assembly). The continuous upper line indicates the uninterrupted genomic sequence, the second line the spliced alignment of the predicted hypothetical gene model to the genomic sequence, and subsequent lines the spliced alignments of RACE-derived sequence. Vertical dotted lines highlight differences in splice site locations. All the predicted gene structures are annotation Version 1. The GenBank accession numbers are At2g02320 assembly, AY168990; At2g33400 assembly, AY144106; At2g10850 assembly, AY501348; At2g01240 assembly, AY168991; At2g03020 assembly, AY219083; At2g06800 assembly, AY464642; At2g01960 assembly, AY168991; and At2g04850 assembly, DQ069802.
27 genes; four poly(A)s were found from 11 genes; and only four genes were found with five poly(A) sites. All the poly(A) sites, even the two from At2g16140 that are 543 bp apart, occur after the stop codon so that none of these alternative polyadenylation events affects the protein sequences. All the genes identified with poly(A) site(s), the poly(A) location, and the distance of each poly(A) from the stop codon are shown in Supplemental Table II.

Transcripts from the Opposite Strand. From 20 targeted regions, transcripts were obtained from the strands opposite from their predictions. For example, in the region of hypothetical gene At2g39975, a poly(A) tail is present in each of four assemblies formed. However, the alignment of assemblies with genomic and predicted At2g39975 coding sequence (CDS) reveals poly(T) tails aligning to the 5' end of the At2g39975 prediction, which clearly suggests that these transcripts arise from the opposite strand (Fig. 4A). Another example is in the region of At2g24480. Not only do all assemblies display poly(T) sequences that align to the 5' end of the predicted gene (At2g24480), but one assembly (no. 3) also contains two introns, whose splice-site orientations (GT-AG) demonstrate that this transcript is from the opposite strand (Fig. 4B). The other cases are similar to these two and were judged by the same criteria.

Transcripts from Both Strands. At three targeted regions (At2g32890, At2g03460, and At2g27160), transcripts from both strands were found. At the At2g32890 locus (Fig. 5A), six assemblies align to the predicted gene and genomic sequences. Among them, two contain poly(A) tails at the 3' end of the At2g32890 prediction, suggesting that their direction of transcription is the same as predicted for the At2g32890 gene. However, the other four assemblies display poly(T) tails that align to the 5' end of the prediction, indicating that these transcripts are derived from the other strand (Fig. 5A). Two different assemblies were obtained from the predicted At2g03460 gene region (Fig. 5B). For assembly 1, the orientation of intron splice sites (GT-AG) and the poly(A) tail demonstrate that the direction of transcription is the same as predicted for At2g03460, although the structure is different because the acceptor site of intron 2 is 37 bp downstream from its prediction and the acceptor site of intron 1 is also different. By contrast, the orientation of the splice sites shown by the alignment of assembly 2 to the genome indicates that this transcript is derived from the other strand (Fig. 5B). A similar situation occurs in the region of hypothetical gene At2g27160 (Fig. 5C). One assembly (no. 1) displays an intron whose boundaries are different from those in the prediction but whose transcription is in the same direction as its prediction (Fig. 5C). All the other assemblies (nos. 2, 3, and 4) have poly(T) tracts aligning to the predicted 5' end of At2g27160, and one of them (no. 4) predicts four introns, indicating that these three assemblies are all transcribed from the opposite strand (Fig. 5C). Thus, although the transcripts obtained from these three regions may not be full-length cDNAs, the poly(A) tails and the orientations of the intron splice sites clearly demonstrate that the transcripts recovered from these regions are from both strands.

Assemblies Aligning to More Than One Locus

Forty assemblies at 28 regions align to more than one locus, encompassing a total of 57 loci, 36 of which are from our target list. At seven of these regions, the single-aligned assembly simply overlaps the neighboring gene on the other strand, an observation that is not uncommon in either Arabidopsis or other organisms. The other 33 assemblies that align to two or more genes at 21 regions may indicate either the existence of
dicistronic transcripts or a necessity for merging of existing gene models.

**Dicistronic Transcripts.** At four target gene regions, dicistronic transcripts were recovered. The assembly from At2g18200 actually covers the nearby At2g18210 region (Fig. 6A). Two ORFs found in this transcript are identical to the ORFs predicted for At2g18210 and At2g18200, and the distance between them is 819 bp (Fig. 6A). Similarly, the transcript obtained from At2g07708 is extended to the At2g07707 region and the two ORFs within this transcript, separated by 582 bp, are identical to At2g07708 and At2g07707, respectively (Fig. 6B). The orientation of both transcripts was confirmed by their poly(A) tails. Another dicistronic transcript was obtained from assembling sequences from the At2g10060 RACE products and contains two ORFs of 208 and 164 amino acids separated by 279 bp. Both gene structures and the encoded ORFs differ from those of the predictions (Fig. 6C). The transcript from the At2g14350 locus covers only the predicted region without any extension to other loci. However, the At2g14350 prediction has four introns, while the assembly contains only three introns, skipping the smallest predicted intron (9 bp) that contains an in-frame stop codon. Consequently, two ORFs are encoded by this single transcript (Fig. 6D). The first ORF is 210 amino acids and starts 9 bp (three amino acids) upstream of the predicted At2g14350 start site because there is a 1-bp difference between our recovered cDNA sequence and the genomic sequence. This difference was supported by two different clones and confirmed by manually checking their chromatograms. The second ORF contains 205 amino acids and starts at 45 bp downstream from the stop site of the first ORF and ends at the stop site predicted for At2g14350 (Fig. 6D). Both ORFs are in the same reading frame as the original prediction.

**Gene Merging.** At three target regions, assemblies aligning to more than one gene suggest that the existing gene models should be merged because each of the transcripts obtained contain only one uninterrupted ORF (Fig. 7). The assemblies from At2g40316 cover the nearby gene At2g40313 and alternatively spliced transcripts were detected in this region (Fig. 7A). The three different transcript forms all have continuous ORFs with 276, 214, or 195 amino acids separated by 279 bp. Both gene structures and the encoded ORFs differ from those of the predictions (Fig. 6C). The transcript from the At2g14350 locus covers only the predicted region without any extension to other loci. However, the At2g14350 prediction has four introns, while the assembly contains only three introns, skipping the smallest predicted intron (9 bp) that contains an in-frame stop codon. Consequently, two ORFs are encoded by this single transcript (Fig. 6D). The first ORF is 210 amino acids and starts 9 bp (three amino acids) upstream of the predicted At2g14350 start site because there is a 1-bp difference between our recovered cDNA sequence and the genomic sequence. This difference was supported by two different clones and confirmed by manually checking their chromatograms. The second ORF contains 205 amino acids and starts at 45 bp downstream from the stop site of the first ORF and ends at the stop site predicted for At2g14350 (Fig. 6D). Both ORFs are in the same reading frame as the original prediction.

**Figure 4.** Examples of transcripts from strands opposite to the original prediction. Solid lines are genomic sequences in genomic strands or exon sequences in prediction and assemblies. Dotted lines show the alignment of intron/exon boundaries. Splice site, GT-AG or CT-AC, is shown in genomic strands. Poly(T) tails are shown as TTTTT in assemblies. Start and stop codons are shown in predicted CDS. The predicted gene structures are annotation Version 1. None of the opposite-strand transcripts shown here has a good ORF. The GenBank accession numbers are transcript on the opposite strand of At2g39975 (assembly1), DQ069850; transcript on the opposite strand of At2g24480 (assembly2), DQ069848; and transcript on the opposite strand of At2g24480 (assembly3), DQ069849.
completely covers the nearby gene At2g47730, annotated as glutathione S-transferase (Fig. 7C). However, the longest ORF from this transcript is the same as At2g47730. The portion of the transcript covering the predicted At2g47720 region contains an unpredicted exon, which results in a small potential ORF (64 amino acids) at this part of the transcript that is different from the predicted At2g47720 ORF (72 amino acids; Fig. 7C). Thus, this portion of the transcript might be just the UTR region. The remaining 14 assemblies span more than one gene but extend only partially into nearby genes (data not shown), thus not permitting clear conclusions.

Promoter-Reporter Fusion Expression Patterns from Five Hypothetical Genes

Five genes (At2g02540, At2g31270, At2g33640, At2g35550, and At2g36340) were chosen at random for a pilot study of expression patterns, in which their presumptive promoter regions, approximately 1.2-kb upstream sequence from ATG site, were cloned into GUS or GFP reporter constructs (pYXT1 and pYXT2; Fig. 8A) and transformed into Arabidopsis. Three independent transgenic lines were obtained from At2g02540 and At2g36340 and two independent lines from each of At2g31270, At2g33640, and At2g35550. Their expression...
patterns were examined (Fig. 8, B–F). The At2g02540 promoter-reporter construct is expressed in vascular tissues of leaf and different parts of the flower, and in siliques, but is not expressed in roots and old stems (Fig. 8B). Reporter constructs from At2g31270, At2g33640, At2g35550, and At2g36340 are all expressed in flowers, but in different parts and at different developmental stages (Fig. 8, C–F). The At2g31270 construct is expressed in carpels and petals (Fig. 8C). The At2g33640 construct is expressed exclusively in pollen (Fig. 8D). The At2g35550 construct is expressed in carpels, stamens (especially in the anther), and petals (Fig. 8E). The At2g36340 construct is expressed in young anthers and some young seeds in siliques (Fig. 8F), but not in mature pollen and in some but not all young flowers (Fig. 8F, arrows). Thus, the promoter-reporter fusion constructs from all five hypothetical genes tested exhibit tissue-specific and/or developmental stage-specific expression patterns.

DISCUSSION

In our study, we examined the structure and expression of all 797 genes on Arabidopsis chromosome...
that were originally annotated as hypothetical (Version 1 data). To do this, we developed a high-throughput method with increased specificity that included automated primer design, the use of nested primers, more restricted screening for insert-positive colonies, and a set of scripts for semiautomated sequence assembly and analysis. As a result, a great deal of useful information was obtained from this study.

Expression Analysis

In an attempt to recover more expressed hypothetical genes, we expanded the cDNA populations and simplified the RACE process by pooling them together while at the same time increasing the input of each component cDNA so as not to compromise the concentration of each transcript in this complex pool.
Without the specific expression-testing step (Xiao et al., 2002), expression is determined by the generation of RACE product(s) that match the intended target. The overall expression frequency observed (50%; 399/797) is lower than in our previous report (82% [137/169]; Xiao et al., 2002). This is most likely due to the smaller number of clones screened as well as to false priming during RACE of the complex pool as compared to the tissue-by-tissue assay. The numbers of hypothetical genes that are reported to be expressed by hybridization methods range from 37% for the whole-genome tiling array (Yamada et al., 2003) to 81% for the chromosome 2 amplicon array (Kim et al., 2003), with the Affymetrix ATH1 array reporting an intermediate value (60%; Redman et al., 2004). Despite some limitations of our high-throughput RACE protocol, this PCR-based method is still more sensitive than hybridization-based assays for the detection of gene expression. Among the hypothetical genes that are reported to be expressed by hybridization methods range from 37% for the whole-genome tiling array (Yamada et al., 2003) to 81% for the chromosome 2 amplicon array (Kim et al., 2003), with the Affymetrix ATH1 array reporting an intermediate value (60%; Redman et al., 2004). Despite some limitations of our high-throughput RACE protocol, this PCR-based method is still more sensitive than hybridization-based assays for the detection of gene expression. Among the hypothetical genes represented on, but not detected by, the ATH1 Affymetrix chip, 327 genes were included in our RACE experiments and 160 were found to be expressed. With the whole-genome tiling array (Yamada et al., 2003), only 190 of the 399 genes found expressed in our study were called present. Comparing our results to transcription data generated by MPSS from callus, inflorescence, leaves, root, and silique (Meyers et al., 2004), 170 out of the 399 genes for which transcripts were detected in this study have MPSS expression support.

**Gene Structure Analysis**

Insert-positive clones from RACE reactions were obtained for 506 out of the 799 genes attempted. However, the sequences generated matched only 399 target gene regions. For the remaining 107 gene targets, either the sequencing reactions failed or the sequences generated did not match the right target genes. Clearly some mispriming and amplification of nontarget regions coupled with false-positive clone identification occurred during our experiments. Among the 366 individual hypothetical genes matched by our sequences with full-length or partial cDNAs, about 38% of the predicted gene structures have some degree of inaccuracy. This is similar to an earlier study in which incorporation of 5,000 full-length cDNA sequences into the Arabidopsis genome annotation revealed that 33% of the gene models needed modifications (Haas et al., 2002). In our study, the number of incorrect gene models may be underestimated, since approximately half of the cDNA sequences are likely incomplete due either to failed RACE reactions or short sequence reads.

**Alternate Splicing**

Previous reports based on full-length cDNA and EST alignments have suggested that 10% or fewer Arabidopsis genes are alternately spliced (Haas et al., 2002, 2005; Zhu et al., 2003). However, in this study, 116 out of
399 expressed genes (approximately 29.0%) display alternatively spliced transcripts, a number consistent with but statistically more reliable than that reported previously (four out of 16; Xiao et al., 2002) and not much less than the value of 38% reported for the human genome (Brett et al., 2000; Kan et al., 2001; Modrek et al., 2001). The hypothetical genes that we examined lacked EST or cDNA support most likely because of low expression levels, localized or transient expression, or expression only under certain biological conditions. However, when a specific targeted approach was used in our experiments, a high level of sequence coverage was generated at many of the target regions. Although not all sequences matched their target regions, there were 8,757 sequences aligning to 399 target regions, corresponding to an average coverage of 22 reads per region, which facilitated the identification of multiple splice isoforms at particular regions. Use of a diverse pool of cDNA populations could also contribute to the higher percentage of alternative-splicing isoforms observed in our study, since it is well accepted that different tissues or biological conditions may produce different transcript variants. Lazar and Goodman (2000) showed that one alternative-splicing isoform of SRI in Arabidopsis could play a role in cellular adaptation to a high-temperature environment. In pumpkin (Cucurbita maxima), light can regulate the alternative splicing of the hydroxyypyruvate reductase gene (Mano et al., 1999, 2000). Recently, genome-wide analysis showed that environmental stress conditions significantly affected alternative-splicing profiles in Arabidopsis (Iida et al., 2004). Although many reports about alternative splicing of individual genes demonstrated that alternatively spliced transcripts have important biological functions in different plants (Zhou et al., 1998; Mano et al., 1999; Hartung and Puchta, 2000; Lazar and Goodman, 2000; Asakura et al., 2002; Jasinski et al., 2002; Quesada et al., 2003; Savaldi-Goldstein et al., 2003), it is unknown at this point which alternative-splicing transcripts in our study have real biological function and which are just misspliced products, especially in the 71 cases of intron retention that has been reported as a major phenomenon in alternative splicing in Arabidopsis (Ner-Gaon et al., 2004). Therefore, we believe that the high percentage of alternatively spliced transcripts found in our study are not specific to hypothetical genes but are observed because of the facets of our experimental approach discussed above.

Multiple Poly(A) Sites

Among the 187 genes for which polyadenylated transcripts were recovered, 83 genes (about 44%) display more than one poly(A) site. Multiple polyadenylation sites have been reported previously (Dean et al., 1986; Graber et al., 1999; Lupold et al., 1999; Hartung and Puchta, 2000; Magnotta and Gogarten, 2002). Alternative splicing of FCA gene transcripts leads to alternative polyadenylation, and interactions between the products of these splice isoforms control Arabidopsis flowering time (Quesada et al., 2003; Simpson et al., 2003). However, none of the alternate polyadenylation events observed in this study affect the coded-protein sequence. Searching of the sequences 300 bp upstream of each poly(A) site revealed that only about 27% (89/331) of the sequences contained the consensus AATAAA poly(A) signal, indicating that this signal (AAUAAA) is not consistently utilized in plants (Graber et al., 1999). Further investigation will be required to determine whether any of these alternate polyadenylation events has functional significance.

Unusual Transcripts

In addition to expected transcripts from the targeted hypothetical genes, some unusual transcripts were found, including transcripts from the strand opposite to the targeted prediction, transcripts from both strands at a particular locus, and transcripts covering more than one locus.

Transcripts from the Opposite Strand. From 20 regions, the transcripts recovered are not from the strands on which the predicted genes are located but from the opposite strand, as confirmed by the presence of poly(A) tails or the orientation of splice sites (Fig. 4; see Supplemental Fig. 1 for a detailed explanation). In every case, the ORFs in these opposite-strand transcripts are shorter than 100 amino acids. In contrast, the predicted ORFs on the other strand are always longer, explaining why they would be favored by the gene prediction programs. These opposite-strand transcripts may be noncoding transcripts (MacIntosh et al., 2001; Jones, 2002), or they may simply encode small proteins (Wen et al., 2004; for review, see Xia, 2004). Furthermore, the existence of opposite-strand transcripts does not preclude the possibility that there is also a transcript from the predicted strand that simply was not detected in our experiment, an idea supported by the fact that four out of these 20 gene regions do show both sense and antisense transcription in the MPSS data (Meyers et al., 2004). Among the remaining 16 genes, 14 do not show any expression by MPSS and the other two show only sense expression. At the whole-genome level, Yamada et al. (2003) detected 1,846 annotated genes with only antisense expression, including eight of the 20 regions found in our study.

Transcriptions from Both Strands. At three target regions (At2g32890, At2g03460, and At2g27160), transcripts were obtained from both strands (Fig. 5). As with the transcripts from opposite strands only, the transcripts from the strands opposite from predicted genes have short or nonexistent ORFs. The overlap regions of the pair of transcripts are 95 bp from the At2g32890 region, 695 bp from the At2g03460, and 505 bp from the At2g27160 region. Although antisense transcripts have been found in plants previously, their occurrences and functions are still unclear (Schmitz
and Theres, 1992; Dolfini et al., 1993; Cock et al., 1997; Quesada et al., 1999; for review, see Terryn and Rouze, 2000). One report demonstrated that activation of a retrotransposon could change the expression of adjacent genes and result in transcripts from both strands (Zijlstra and Hohn, 1992; Dolfini et al., 1993; Cock et al., 1997; Quesada et al., 1999; for review, see Terryn and Rouze, 1992; Dolfini et al., 1993; Cock et al., 1997; Quesada et al., 1999; for review, see Terryn and Rouze, 2000). One report demonstrated that activation of a retrotransposon could change the expression of adjacent genes and result in transcripts from both strands (Zijlstra and Hohn, 1992; Dolfini et al., 1993; Cock et al., 1997; Quesada et al., 1999; for review, see Terryn and Rouze, 1992; Dolfini et al., 1993; Cock et al., 1997; Quesada et al., 1999; for review, see Terryn and Rouze, 2000).

Dicistronic Transcripts. The transcripts found from four targeted hypothetical gene regions (At2g18200, At2g07708, At2g10060, and At2g14350) show clear dicistronic characteristics (Fig. 6) as defined by (1) two distinct and nonoverlapping coding regions contained within a single processed transcript; and (2) each CDS is more than 100 amino acids. There are likely more dicistronic transcripts in the Arabidopsis genome because, even in our study, there are 33 assemblies matching two loci that extend partially to nearby loci. Additionally, there are at least 20 examples of transcripts corresponding to two adjacent genes found in TIGR’s latest annotation process (Haas et al., 2005; ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/DATA_RELEASE_SUPPLEMENT/polycistronicTranscripts.txt.gz). The mechanism by which the downstream ORF could be translated is still unclear. In the four cases described here, the presence of ATG codons (out of frame) in the intercistronic regions argues against the ribosome scanning mechanism, which requires the absence of any ATG codons in the intercistronic region (Levine et al., 1991). Another mechanism of dicistronic transcript translation is that the presence of an internal ribosome entry site (IRES; Peabody and Berg, 1986) could initiate the translation of the second CDS, which may be more applicable to our cases. IRESs have been identified from different viruses, and they vary in structural, sequence, length, and functional requirements (Belsham and Sonenberg, 2000; Pestova et al., 2001; Dorokhov et al., 2002). To date, except for some reports about the expression of dicistronic constructs made by virus sequences (Zijlstra and Hohn, 1992; Urwin et al., 2000; Toth et al., 2001; Jaag et al., 2003), there is only one report about functional dicistronic transcripts in plants (Garcia-Rios et al., 1997), in which tomPROM locus encodes γ-glutamyl kinase and γ-glutamyl phosphate reductase separated by only 5 bp (Garcia-Rios et al., 1997). Coordinate regulation of this pair of genes as a single operon is consistent with their functions in the same biological pathway and suggests that the pairs of hypothetical genes on the dicistronic transcripts reported here may also have related functions. Therefore, functional study of the ORFs in each dicistronic transcript is needed. The intercistronic regions from all four cases in our study are from 45 to 819 bp and do not have obvious consensus sequences among them except for a 10-bp consensus sequence ATCACATGGT obtained from Multiple Em for Motif Elicitation (Bailey and Elkan, 1994). However, except for the shortest 45-bp intercistronic sequence, the other three sequences contain numerous palindrome sequences that might form the hairpin structures that are typical for IRES. More experiments such as protein expression analysis are needed to confirm this possibility. In addition to two proteins being coded by dicistronic transcripts, there is a possibility that a cis-acting selenocysteine insertion sequence (SECIS) located at the 3’ UTR region translates stop codon (UGA) into selenocysteine and results in a fusion protein (Berry et al., 1991; Rother et al., 2001; Howard et al., 2005; for review, see Low and Berry, 1996). An algorithm is now available for identification of SECIS elements (Zhang and Gladyshev, 2005). However, SECISs were not found by searching the sequences of three dicistronic transcripts with TGA stop codon at first ORF (http://genome.unl.edu/SECISearch.html).

Expression Patterns of Hypothetical Genes Tested

Expression patterns from five randomly chosen hypothetical genes (At2g02540, At2g31270, At2g33640, At2g35550, and At2g36340) were examined. The reporter construct driven by the promoter of At2g02540 is expressed in young vascular tissue of above-ground parts of Arabidopsis (Fig. 8B). Several genes that are expressed in vascular tissues belong to the homeodomain-Leu zipper (HD-ZIP) class of transcription factors (Mattsson et al., 2003; for review, see Ye et al., 2002) or have promoters containing a domain recognized by Leu zipper (HD-ZIP) factors (Ayre et al., 2003). However, At2g02540 apparently does not have an HD-ZIP domain but has a domain homology to a group of plant transcription factors named as ZF-HD for zinc-finger homeodomain proteins that might be involved in the mesophyll-specific expression of the C4 phosphoenolpyruvate carboxylase gene in C4 species of the genus Flaveria (Windhovel et al., 2001).

The other four promoter-reporter constructs tested from At2g31270, At2g33640, At2g35550, and At2g36340 are all expressed in different parts of the floral organ at different development stages (Fig. 8). According to the Arabidopsis ABC model of flower development (Coen and Meyerowitz, 1991), At2g31270, which gave strong expression in carpel (whorl 4), should belong to group C. The protein encoded by At2g31270 is similar to the CDT1 protein that promotes DNA replication in yeast (Saccharomyces cerevisiae; Nishitani et al., 2000) and was recently identified as an Arabidopsis CDT1 protein
affecting leaf cell division and cell proliferation (Castellano Mdél et al., 2004). Detection of At2g31270 expression in young flowers is consistent with this function. However, Castellano Mdél et al. (2004) also detected its expression in leaf and root. The differences in the length of promoter used and the fusion reporter protein may account for the variation. The promoter-reporter construct of At2g33640 is expressed in pollen, a part of the stamen (whorl 3), and should belong to the group C or B. At2g33640 has a DHHC zinc-finger domain that was first isolated from Drosophila (Mesilaty-Gross et al., 1999), which might be involved in protein-protein or protein-DNA interactions (Putilina et al., 1999). The promoter-reporter construct of At2g35550 is expressed in petals (whorl 2) and stamens (whorl 3) but not in carpels (whorl 4), which should be in gene group B. The protein sequence encoded by At2g35550 has a domain that matches a protein from soybean (Glycine max), GBP, which binds to GAGA element dinucleotide repeat DNA (Sangwan and O’Brien, 2002) and is likely involved in DNA binding. The Gbp gene is expressed in soybean leaves and is induced in symbiotic root nodules (Sangwan and O’Brien, 2002). Interestingly, the promoter-reporter construct of At2g36340 shows expression in young anthers and some young seeds in the silique, but not in the pollen or some young flowers even in the same inflorescence (Fig. 8E, arrows). At2g36340 was recently reported as a member of a new gene family in Arabidopsis that encodes a nuclear protein with DNA-binding activity that is regulated by KNAT7 (Curaba et al., 2003). However, the expression pattern of At2g36340 detected here is different from the expression pattern of GeBP gene, the typical member of this gene family. GeBP is expressed in the apical meristem and young leaf primordial, but not in flowers (Curaba et al., 2003). This might be due to the fact that At2g36340 was identified as a GeBP family member only by amino acid sequence similarity (Curaba et al., 2003). Although none of the five genes tested are still annotated as “hypothetical protein” in the current TIGR annotation, their biological functions in Arabidopsis are still unclear. To understand these genes, including in which cell types, at which developmental stages they are expressed, and which genes they regulate, more detailed experiments, such as sections, in situ hybridization, knock-out, and overexpression are definitely needed. Overall, the promoter-reporter constructs from all five hypothetical genes tested have very localized and/or transient expression patterns providing a likely explanation for their absence from EST or cDNA collections.

Development of a High-Throughput Method

Based on our previous results, a high-throughput pipeline was developed to clone and analyze cDNA sequences of hypothetical genes. The first advance was the development of a Perl script for automatic primer design that combines Primer3 (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_ www.cgi) and BLASTn programs. The primer design script functions as follows: (1) accepts multifasta format input sequence; (2) uses the Primer3 program to design outside primer pair; (3) blasts designed primers against TIGR ATH1 cDNA database to check their uniqueness; and (4) if the outside primers are accepted, Primer3 is used to design a nested primer pair omitting the uniqueness check. This script outputs outer and nested primers for each gene and product sizes based on predicted hypothetical gene structures. The criteria for primer design are that they should be 18- to 35-nucleotide-long, having 20% to 80% GC content with melting temperature ≥70°C, which enables touchdown PCR. The primers were designed to give a 200- to 500-bp overlap between the 5’ and 3’ RACE products so that the 5’ RACE sequences and 3’ RACE sequences could be assembled together. The second protocol adjustment is the use of 96-well format for all PCR reactions, cloning reactions, transformations into Escherichia coli, and colony screening. For each target gene, 12 clones at 5’ RACE and 3’ RACE cloning reaction were screened by PCR for insert, and all insert-positive clones were sequenced from both ends. A sequence analysis script automatically (1) retrieves sequences for each gene from the sequence database and assembles them using TIGR Assembler (Sutton et al., 1995); (2) retrieves predicted CDS sequence and genomic sequence for each gene from the annotation database; and (3) aligns genomic, predicted CDS, and assembled experimental cDNA sequences together using the dds/gap2 alignment program (Huang et al., 1997) for manual inspection/validation. The entire set of sequences was subsequently mapped to the Arabidopsis genome by PASA for final analysis. This high-throughput method has greatly increased the speed of cloning and analysis of cDNA of hypothetical genes. The primer design script and sequence analysis scripts are available upon request.

Plant Material

Arabidopsis (Arabidopsis thaliana) ecotype Columbia-0 seeds/plants were subjected to a variety of treatments as described below. Callus, roots, and young plants and the treatments of cold and heat were obtained as described in the previous study (Xiao et al., 2002). Pathogen-treated plants in this study were a mixture of Xanthomonas-treated and Pseudomonas-treated plants. Before infection, the plants were grown at 25°C and 8-h photoperiod for 21 d. The leaves were inoculated with fresh cultures, and aerial plant parts were harvested 24 h later for Xanthomonas infection and 12 h later for Pseudomonas infection. For UV treatment, soil-grown plants as above were exposed to 400,000 µj/cm² or 800,000 µj/cm² UV light. Aerial plant parts were harvested at 2, 6, 16, and 30 h following treatment. The treatments with salt, 2,4-D, IAA, and H₂O₂ all were completed with liquid-cultured plants. Sterile seeds were put at 4°C for 48 h and then inoculated into 200 mL of sterile 0.5 × Murashige and Skoog salts (Murashige and Skoog 1962), 10 g/L sucrose, 0.5 × vitamins (vitamins added after autoclaving), 0.5 × Glycine-Glycine mix, pH adjusted to 5.8 with KOH. Plants were grown under a 24-h photoperiod with shaking at 100 rpm. Treatments were carried out by adding the chemical challenge to the liquid at its designated final concentration at approximately 14-d postgermination. All tissue (whole plant) was harvested at indicated intervals posttreatment. Treatment concentrations and collection time points are as follows: 250 mM NaCl harvested at 1, 3, 6, and 24 h; 50 µM 2,4-D harvested at 1, 3, 6, and 24 h; 10 µM IAA harvested at 1 and 3 h; and 5 mM and 25 mM H₂O₂ harvested at 1, 3, 8, and 24 h. For all treatments with multiple time points, equal masses of tissue were combined for total RNA isolation with the exception of IAA treatments, for which RNA was isolated for each individual time point and then equal masses of RNA mixed for mRNA isolation.

Vectors Construction, Transformation, and Plant Growth

To create the Gateway-compatible Gus reporter construct pYXT1 from binary vector pBI121, pBI121 was first digested with HindIII and the ends blunt-ended using Klenow. Then, Smal was used to release the 35S promoter fragment from pBI121. The large fragment lacking the 35S promoter (pBI121 backbone) was then recovered from an agarose gel and dephosphorylated with calf intestinal alkaline phosphatase. Finally, the Gateway reading frame

MATERIALS AND METHODS

Selection of hypothetical genes, experimental concept, methods of construction of cDNA populations, and RACE reactions are the same as our previous study (Xiao et al., 2002) except that nested PCR was used in the RACE reactions according to the instructions of the Marathon cDNA amplification kit (BD Biosciences Clontech).
A fragment (Invitrogen) was ligated into the pBl21 backbone and transformed into DB3.1 competent cells (Invitrogen). The plasmid with the correct orientation of the reading frame A insertion is pYXT1. The Gateway-compatible GFP reporter construct, pYXT2, was made from the binary vector pET-15GAL4-VP16UasmGFPF5SER (Bougourd et al., 2000) as follows. First, pET-15GAL4-VP16UasmGFPF5SER was digested with SacII. The small fragment (about 5.3 kb) was recovered and cloned into the SacII site in pGEMZf1 (+; Promega) to form an intermediate plasmid. The large fragment was recovered and reserved for regulation in the future. The intermediate plasmid was digested with BamHI, dephosphorylated by calf intestinal alkaline phosphatase, then ligated to the Gateway reading frame B (RfB) fragment (Invitrogen) and transformed into DB3.1 competent cells (Invitrogen). Colonies with the correct orientation of RfB insertion were digested with SacII and again the 5.3-kb fragment containing RfB was recovered. This 5.3-kb fragment was ligated into the large fragment obtained from the SacII digestion of pET-15GAL4-VP16UasmGFPF5SER. The plasmid with the correct orientation of the SacII insertion containing RfB is pYXT2. All cloning junctions in the pYXT1 and pYXT2 constructs were confirmed by sequencing.

Amplification of the promoter fragments with attB sites, as well as βP and LR reactions, were completed according to the protocols in the Gateway Cloning Technology booklet (Invitrogen). Arabidopsis (Columbia ecotype) genomic DNA was used to amplify promoter fragments of five hypothetical genes. The primers used include gene-specific sequence as well as the attB site sequence for βP cloning and are as follows: At2g13300, upstream primer, AAGGCTGGAATCAATGAAATCCATCACCA; downstream primer, AT2g06750, upstream primer, ACCAAGCAGGCTGCATCGTACGATACGATACG; downstream primer, AT2g35550, upstream primer, AAAAAGCAGGCTTTTGTTTTGGGTGAATATGAAAATCTT, downstream primer, At2g31270, upstream primer, AAAAAGCAGGCTGATCTAGATCAGATTCTAAGAGACG; downstream primer, At3g09700, upstream primer, GTGAAATGTGTGAGCTGGAAGTTGCCA; downstream primer, At4g39070, upstream primer, AGAAAGCTGGGTAGTCCTTGTTTGCGTTTGTAGCAGA; and downstream primer, At2g06770, upstream primer, AGAAAGCTGGCTGCCCACCTCCACTATTACCATAACTA; downstream primer, AAAAAGCAGGCTACTATAGCAACCTGTTCAAGAGACG, downstream primer, AGAAAGCTGGGTGAGCTGGAAGTTGCCA; and downstream primer, AGAAAGCTGGGTAATCCATCACCA. The amplified promoter fragments with attB sites were introduced into the pDONR207 entry vector (Invitrogen) using a βP cloning reaction and then transferred into pYXT1 or pYXT2 destination vectors with an LR cloning reaction. Resulting constructs were transformed into Arabidopsis using Agrobacterium tumefaciens (strain GV3101) and transfomants were selected on kanamycin, according to standard methods (Clough and Bent, 1998). Three independent transformations were done for each construct.

Sequence data from this article can be found in the GenBank/EMBL data libraries, and accession numbers are shown in Supplemental Table III.

ACKNOWLEDGMENTS

We thank all members of the plant group at TIGR for their help, especially Bin Feng for his computational support and Robin Buell, Francoise Thibaud-Nissen, and Beverly Underwood for critical reading of the manuscript.

Received March 29, 2005; revised June 1, 2005; accepted August 3, 2005; published October 21, 2005.

LITERATURE CITED


Cooper GE, Hunkapiller M, Vachon G (2003) GeBP, the first member of a new gene family in Arabidopsis, encodes a nuclear protein with DNA-binding activity and is regulated by KNAT1. Plant J 34: 305–317


Copyright © 2005 American Society of Plant Biologists. All rights reserved.


Hypothetical Genes on Arabidopsis Chromosome 2

Plant Physiol. Vol. 139, 2005