# Promoter-Proximal Introns in *Arabidopsis thaliana* Are Enriched in Dispersed Signals that Elevate Gene Expression[W][OA]

**Alan B. Rose,[a,1]  Tali Elfersi,[b]  Genis Parra,[b] and Ian Korf[a,b]**

[a] Molecular and Cellular Biology, University of California, Davis, California 95616
[b] Genome Center, University of California, Davis, California 95616

**Introns that elevate mRNA accumulation have been found in a wide range of eukaryotes. However, not all introns affect gene expression, and direct testing is currently the only way to identify stimulatory introns. Our genome-wide analysis in *Arabidopsis thaliana* revealed that promoter-proximal introns as a group are compositionally distinct from distal introns and that the degree to which an individual intron matches the promoter-proximal intron profile is a strong predictor of its ability to increase expression. We found that the sequences responsible for elevating expression are dispersed throughout an enhancing intron, as is a candidate motif that is overrepresented in first introns and whose occurrence in tested introns is proportional to its effect on expression. The signals responsible for intron-mediated enhancement are apparently conserved between *Arabidopsis* and rice (*Oryza sativa*) despite the large evolutionary distance separating these plants.**

## INTRODUCTION

The positive effect of introns on gene expression, which has been termed intron-mediated enhancement (IME) (Mascarenhas et al., 1990), is often eclipsed by the better known roles of promoters, enhancers, silencers, and chromatin modifications in controlling transcription. However, IME has been observed in a wide range of eukaryotes, including vertebrates, invertebrates, fungi, and plants (Buchman and Berg, 1988; Palmiter et al., 1991; Okkema et al., 1993; Koziel et al., 1996; Simpson and Filipowicz, 1996; Duncker et al., 1997; Lugones et al., 1999; Ho et al., 2001; Xu and Gong, 2003; Juneau et al., 2006), suggesting that it reflects a fundamental feature of gene expression. In many cases, introns have a larger influence than do promoters in determining the level and pattern of expression (Virts and Raschke, 2001; Wang et al., 2002; Jeong et al., 2006).

The mechanism of IME is largely unknown. Some efficiently spliced introns boost expression more than 10-fold, while others have little or no effect, arguing against generic mechanisms related to splicing that would apply to all introns equally. Introns enhance gene expression by increasing the steady state amount of mature mRNA in the cell (Callis et al., 1987; Dean et al., 1989; Rethmeier et al., 1997; Rose and Last, 1997; Nott et al., 2003), apparently without significantly changing mRNA stability (Rethmeier et al., 1997; Nott et al., 2003). Although modest effects of introns on transcription have been noted (Rose and Last, 1997; Fong and Zhou, 2001; Furger et al., 2002), these are insufficient to entirely explain the increase in mRNA accumulation. IME is therefore not simply due to traditional transcriptional enhancers embedded in introns, although examples of these are known. The distinction from enhancer elements is underscored by the need for introns to be within transcribed sequences and in their natural orientation to increase expression and that an enhancing intron must be within ~1 kb of the transcription start to have an effect (Rose, 2004). Taken together, these observations suggest a model where signals present in introns render the transcription machinery more processive, increasing the likelihood that full-length polyadenylated mRNAs will be formed and accumulate. In the absence of these signals, the polymerase may tend to dissociate and produce short, truncated transcripts that are rapidly degraded.

The different enhancing ability of introns suggests that some contain more stimulatory signals than do others, yet prior attempts to identify the responsible sequences by deletion analysis have largely failed to identify any specific elements necessary for IME (Clancy et al., 1994; Rose and Beliakoff, 2000). For example, removing 883 nucleotides from the 1028-nucleotide first intron of the maize (*Zea mays*) *Sh1* gene has a negligible effect on its enhancing ability (Clancy and Hannah, 2002), and all sequences in *Arabidopsis thaliana TRP1* intron 1 can be individually deleted without reducing the stimulation it mediates (Rose and Beliakoff, 2000). The sequences responsible for increasing expression are therefore likely to be redundant and dispersed throughout enhancing introns.

Here, we report an experimental and computational analysis of the intron sequences responsible for elevating expression. In addition to confirming the distributed nature of enhancing sequences, we found that the signals responsible for boosting expression are most abundant in introns near the start of transcription and that the compositional differences between promoter-proximal and later introns can be used to predict the ability of an intron to stimulate expression.

## RESULTS

### The Distribution of Stimulatory Sequences

To determine if IME signals are dispersed and redundant, we created a series of hybrid introns (Figure 1) containing fragments of an enhancing intron (*UBQ10* intron 1) within a context of an otherwise nonenhancing intron (*COR15a* intron 2). Each was inserted at the same location in the *TRP1:β-glucuronidase* (*GUS*) reporter gene used previously to compare the *UBQ10* and *COR15a* introns (Rose, 2002), and steady state expression levels were determined in homozygous single-copy transgenic *Arabidopsis*. Each of the first three quarters of the *UBQ10* intron was individually sufficient for increasing mRNA accumulation (Figure 1B; see Supplemental Table 1 online). Even though the sequences at the 3′ end of the intron appear to be less active than those at the 5′ end, they are also further from the start of
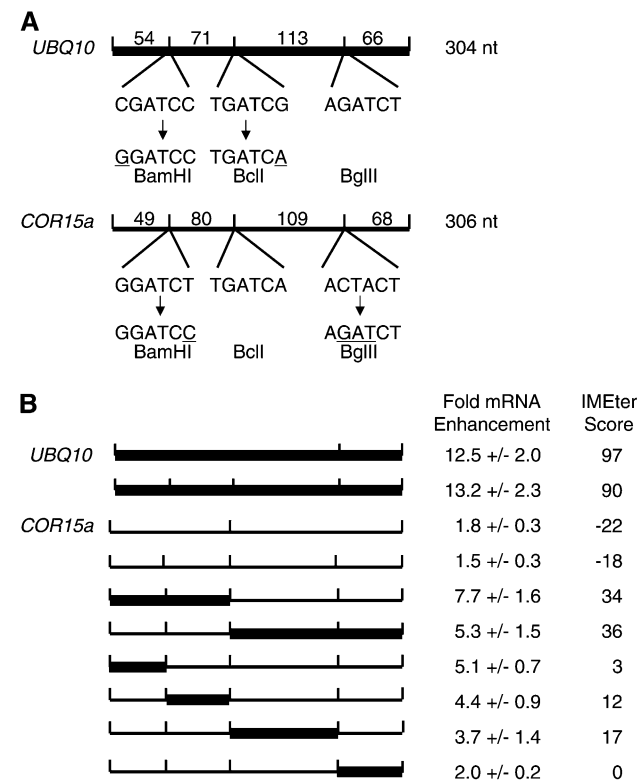
transcription in the hybrid introns, a variable known to diminish the stimulating effects of the *UBQ10* intron (Rose, 2004). Still, the ability to boost expression is clearly dispersed throughout at least the first 238 nucleotides of the 304-nucleotide *UBQ10* intron, and the enhancing effect of the entire intron is roughly the sum of the stimulation mediated by each part. This finding has significant implications for the mechanism of IME, which must be very different from more familiar regulatory elements, such as promoters, enhancers, mRNA secondary structures, or small RNAs, which depend on discrete or localized sequences.

### Computational Analysis of Introns

To better understand the mechanism of IME, we undertook a bioinformatics approach with two related goals: (1) to build a discriminator that would be able to predict the enhancing ability of an untested intron and (2) to identify sequence motifs responsible for IME. We expected that introns near the start of transcription would be particularly enriched in IME signals because most enhancing introns are first introns (Tables 1 and 2) and because of the ~1 kb positional requirement. Since introns from the middle or 3′ end of genes have not been observed to greatly affect expression, we expected these distal introns to have few IME signals. To determine if compositional properties in promoter-proximal and distal introns were correlated with IME, we built a word-based discriminator called the IMEter. The IMEter reports a log-odds score based on the frequencies of all possible words (nucleotide sequences of a given length): a positive score indicates the input sequence is similar to proximal introns, and a negative score indicates similarity to distal introns.

The IMEter was trained on a set of introns whose positions are known relative to their transcription start sites. We collected all introns in the *Arabidopsis* genome whose parent gene was represented by full-length cDNAs and formed an IMEter training set with half of these genes. Variable parameters for the training protocol include word size and the threshold positions of proximal and distal introns. For example, a particular parameter set may define a word size of 4, proximal introns as those appearing <200 bp from the transcription start, and distal introns as those appearing >800 bp from the start site. To evaluate a particular parameter set, we ran the IMEter on six experimentally tested introns (Rose, 2002) and plotted the IMEter scores against the observed level of mRNA accumulation (Figure 2A, closed symbols). These six introns represent the largest available data set in any species for which the absence of variables other than the intron, and the use of homozygous single-copy lines, permitted quantitative comparison of expression levels with minimal variation. We found that several values for word size and intron location result in accurate predictions for the enhancing ability of these six introns (see Supplemental Figure 1 online). Using the optimal parameters of word size 5, and proximal and distal introns split at 400 bp from the transcription start site, a regression line had an $R^2$ value of 0.90.

The high correlation between IMEter score and experimentally observed enhancement suggests that the IMEter can be used to predict the ability of introns to stimulate expression. To test this, we chose six previously uncharacterized *Arabidopsis* introns that have a range of IMEter scores but are similar in GC composition



**Figure 1.** Enhancing Signals Are Distributed throughout *UBQ10* Intron 1.

**(A)** Modifications to facilitate the construction of hybrid introns. Restriction enzyme sites were created by changing the underlined nucleotides. nt, nucleotides.

**(B)** Enhancement by hybrid introns. The degree to which each of the diagrammed introns stimulates steady state *TRP1:GUS* mRNA accumulation relative to the intronless control in single-copy transgenic plants is presented as mean ± SD (*n* varies from 5 to 11 as detailed in Supplemental Tables 1 and 2 online). Thick and thin lines show sequences derived from the *UBQ10* and *COR15a* introns, respectively, and vertical lines indicate the restriction sites detailed in **(A)**.

**Table 1.** Qualitative Analysis of Expression in *Arabidopsis*

| Gene | No. | mRNA | IMEter | Rice | Reference |
|---|---|---|---|---|---|
| *RHD3* (At3g13870) | 1 | + | 71 | 63 | Wang et al. (2002) |
| Histone H3 (At4g40040) | 1 | + | 82 | 74 | Chaubet-Gigot et al. (2001) |
| Histone H3 (At4g40030) | 1 | + | 63 | 65 | Chaubet-Gigot et al. (2001) |
| *EF-1α* A1 (At1g07920) | 1 | + | 100 | 72 | Curie et al. (1993) |
| *EF-1α* A3 (At1g07940) | 1 | + | 52 | 48 | Chung et al. (2006) |
| *eEF-1β* (At2g18110) | 1 | + | −4 | −10 | Gidekel et al. (1996) |
| *TWN2* (At1g14610)[a] | 1 | + | 20 | 17 | Zhang and Somerville (1997) |
| *TWN2* (At1g14610)[a] | 2 | + | 27 | −8 | Zhang and Somerville (1997) |
| *Cox5c-1* (At2g47380) | 1 | + | 60 | 68 | Curi et al. (2005) |
| *Cox5c-2* (At3g62400) | 1 | + | 46 | 35 | Curi et al. (2005) |
| *ACT1* (At2g37620) | 1 | + | 41 | 4 | Vitale et al. (2003) |
| *KC01* (At5g55630) | 1 | + | 8 | −14 | Czempinski et al. (2002) |
| *PRF1* (At2g19760) | 1 | + | 50 | 63 | Jeong et al. (2006) |
| *PRF2* (At4g29350) | 1 | + | 49 | 28 | Jeong et al. (2006) |
| *PRF3* (At5g56600) | 1 | − | 10 | 21 | Jeong et al. (2006) |
| *PRF4* (At4g29340) | 1 | − | 17 | −1 | Jeong et al. (2006) |
| *PRF5* (At2g19770) | 1 | − | 8 | −1 | Jeong et al. (2006) |
| *ADF1* (At3g46010) | 1 | + | 66 | 39 | Jeong et al. (2007) |
| *FAD2* (At3g12120) | 1 | + | 8 | −32 | Kim et al. (2006) |
| *SUVH3* (At1g73100) | 1 | + | 12 | 8 | Casas-Mollano et al. (2006) |
| *SUVH3* (At1g73100) | 2 | − | −4 | −8 | Casas-Mollano et al. (2006) |
| *ATMHX* (At2g47600) | 1 | + | 33 | 34 | David-Assael et al. (2006) |
| *UBQ3* (At5g03240) | 1 | + | 53 | 43 | Norris et al. (1993) |
| *UBQ10* (At4g05320)[b] | 1 | + | 97 | 68 | Norris et al. (1993) |
| *ATPK1* (At3g08730)[b] | 1 | + | 56 | 39 | Zhang et al. (1994) |
| *TCH3* (At2g41100)[b] | 1 | − | −7 | −15 | Sistrunk et al. (1994) |
| *COR15a* (At2g42540)[b] | 2 | − | −22 | 3 | Baker et al. (1994) |

No. indicates the intron number with respect to the 5′ end. The plus sign in the mRNA column indicates that the intron elevates expression. In the rice column, the IMEter was trained in rice.
[a] These two introns were not tested separately.
[b] These four introns are also included in Figure 2.

and length and are all found within 500 bp of the start of single-intron genes. The expression levels mediated by all six introns in single-copy *TRP1:GUS* fusions correlate with their IMEter scores (Figure 2A, open symbols), giving an $R^2$ value of 0.89 for the entire set of experimentally tested introns.

### *Arabidopsis* Introns

At least 21 *Arabidopsis* introns have been found by others to enhance expression. Quantitative comparisons between them usually cannot be made due to differences in the promoter used, the intron location, or the methods of transformation and measuring expression. Despite this limitation, these introns can be used to evaluate the IMEter because they are known to enhance expression. The IMEter assigns a positive score to all but one of these introns (Table 1), and the exception (from the *eEF-1β* gene) may affect expression by a mechanism unrelated to IME because it elevates expression when inserted upstream of the promoter in reverse orientation (Gidekel et al., 1996). Furthermore, 18 of the remaining 20 introns give a score of 20 or higher, even though <4.8% of the *Arabidopsis* introns analyzed generate scores that high. By contrast, three of the six introns shown not to affect expression have negative scores, and just one has a score over 10. Thus, the IMEter recognizes virtually all known enhanc-

ing introns from *Arabidopsis* and would have predicted their stimulating ability.

To analyze genome-wide properties of IMEter scores, we ran the IMEter on the testing set, which comprised the half of the *Arabidopsis* genome not used for training. On average, IMEter scores for introns are slightly negative. Approximately 2% of introns have scores >50, and <1% have scores of 80 or more. Average IMEter scores decline steadily with distance from the start of the gene, becoming negative at ~500 bp and leveling off ~1200 bp from the 5′ end (Figure 2C). This pattern is in striking agreement with the enhancement caused by a single intron as it is moved down a gene, which declines with distance until it is lost entirely between 550 and 1100 bp from the start of transcription (Rose, 2004). This pattern clearly indicates that introns near the start of a gene are compositionally distinct from later introns. IMEter scores of proximal introns are not uniformly high, however, but rather form a distribution centered just above 0, consistent with the inability of many first introns to elevate expression (Figure 2D).

### Motifs Associated with Enhancement

To identify the sequences most responsible for the high IMEter scores of enhancing introns, and thus candidates for involvement

**Table 2.** Qualitative Analysis of Expression in Rice

| Gene | No. | mRNA | IMEter | Reference |
|------|-----|------|--------|-----------|
| *TPI* | 1 | + | 140 | Xu et al. (1994), Snowden et al. (1996) |
| *CDPK2* | 1 | + | 37 | Morello et al. (2006) |
| *rubi3* | 1 | + | 191 | Sivamani and Qu (2006) |
| *tua1* | 1 | + | 90 | Jeon et al. (2000) |
| *tua1* | 2 | + | −2 | Jeon et al. (2000) |
| *tua1* | 3 | + | −10 | Jeon et al. (2000) |
| *tua2* | 1 | + | 174 | Fiume et al. (2004) |
| *tua3* | 1 | + | 274 | Fiume et al. (2004) |
| *tub16* | 1 | + | 158 | Morello et al. (2002) |
| *GAMyb* | 1 | + | 147 | Washio and Morikawa (2006) |
| *RPBF* | 1 | + | −118 | Washio and Morikawa (2006) |
| *ACT1* | 1 | + | 210 | McElroy et al. (1990) |
| *PLD1* | 1 | + | 63 | Ueki et al. (1999) |
| Glutelin | 1 | − | −14 | Ueki et al. (1999) |
| *Sh1* | 1 | + | 203 | Maas et al. (1991), Ueki et al. (1999) |
| *Ubi1* | 1 | + | 187 | Ueki et al. (1999) |
| *Adh1* | 1 | + | 51 | Schunmann et al. (2004) |

For these experiments, the IMEter was trained in rice. No. indicates intron number. The plus sign indicates enhancement; the minus sign indicates no enhancement. All introns are from rice genes except the last three, which are from the maize *Sh1*, *Ubi1*, and *Adh1* genes.

in the mechanism of IME, we searched for sequence motifs among the 100 highest scoring introns with NestedMICA (Down and Hubbard, 2005). Several motifs were found (see Supplemental Figure 2 online), and the one whose abundance in the 12 tested introns best correlates with its ability to elevate *TRP1:GUS* expression is shown in Figure 3A. This motif appears numerous times and is dispersed throughout enhancing introns but is nearly absent from nonenhancing introns. The overlap between the motif and IMEter score density suggests that the IMEter is detecting the motif (Figure 3B), although the IMEter may be detecting other signals as well. The IMEter scores of the *UBQ10-COR15a* hybrid introns (Figure 1B) are proportional to their effect on expression ($R^2 = 0.93$), indicating that the sequences recognized by the IMEter and those responsible for enhancement colocalize at this level of resolution.

### Rice Introns

Some dicot introns can elevate expression in a monocot (Vain et al., 2004), suggesting that IME signals are conserved across species. To test this, we evaluated the 12 *Arabidopsis* introns with measured effects on *TRP1:GUS* expression using an IMEter that had been trained with rice (*Oryza sativa*) introns. The rice IMEter scores of these *Arabidopsis* introns generally correlate with their ability to stimulate expression ($R^2 = 0.74$), indicating that these distantly related plants share some signals in common (Figure 2B). Even though IME has been observed in *Caenorhabditis elegans* and *Drosophila melanogaster*, an IMEter trained with introns from these species had no predictive value for *Arabidopsis* introns ($R^2 = 0.32$ and 0.1, respectively), suggesting that either IME signals are not universally conserved or that general compositional differences interfere with their discovery.

Although we do not have quantitative expression values for rice introns, and therefore cannot use regression to correlate

IMEter scores with expression levels, we trained the IMEter in rice and analyzed the 13 rice introns and three maize introns whose enhancing ability has been reported in the literature (Table 2). Thirteen of the enhancing introns had positive IMEter scores, nine of which were well over 100, while an intron known to have no effect on expression gave a negative score. An examination of all rice introns shows that only 2% of introns have scores of 100 or more (see Supplemental Figure 3A online). As in *Arabidopsis*, we find that promoter-proximal introns in rice have higher IMEter scores than distal introns (see Supplemental Figure 3B online). Since the IMEter apparently identifies enhancing introns in rice, we looked for motifs in the 100 highest scoring introns with NestedMICA just as we had done for *Arabidopsis* (see Supplemental Figure 4 online). The most common motif we found is very similar to the core of the *Arabidopsis* motif (Figure 3C).

### DISCUSSION

Here, we show that signals concentrated in and dispersed throughout promoter-proximal introns are responsible for IME and identify a motif that comprises part of the IME signal. The IMEter algorithm we developed provides a measure of the abundance of promoter-proximal signals in any intron, which strongly predicts the ability of that intron to stimulate gene expression.

Although the correlation between the enhancing ability of introns and their IMEter scores is very high, it is not absolute. One potential explanation for discrepancies between IMEter scores and our quantitative expression data is that the distribution of enhancing sequences within an intron could influence its ability to elevate expression. For example, stimulatory sequences near the 5′ end of an intron might have the greatest impact on enhancement, while those closer to the 3′ end could influence overall IMEter score but have much less effect on expression. Thus, the two hybrid introns with *UBQ10* sequences at the 5′ end had comparable effects on mRNA accumulation (Figure 1), even though the one with more *COR15a* sequence had a lower IMEter score. The contribution made to the total IMEter score by each part of an intron can be visualized by scanning the sequence in a sliding window (see Figure 3B), providing a high-resolution method to locate the regions that may be involved in the mechanism of enhancement. The IMEter will be refined to improve accuracy as more is learned about the sequences that boost expression.

The variables that preclude quantitative comparisons of the introns in Tables 1 and 2 are the least for different introns reported in the same publication, allowing a limited evaluation of the IMEter with these introns. In the only work to compare different introns at one location in the same reporter gene, the first intron from the rice *tua1* gene was found to stimulate expression to a much higher degree than does the second or third intron from that gene (Jeon et al., 2000), consistent with their IMEter scores (Table 2). Similarly, both the enhancing ability and IMEter score of the first intron of the *Arabidopsis SUVH3* gene are higher than those of the second intron (Table 1), although the second intron had the disadvantage of being further from the promoter because the introns were tested in their natural locations (Casas-Mollano et al., 2006). In most experiments, the promoters used were from the same gene as the intron being tested. The
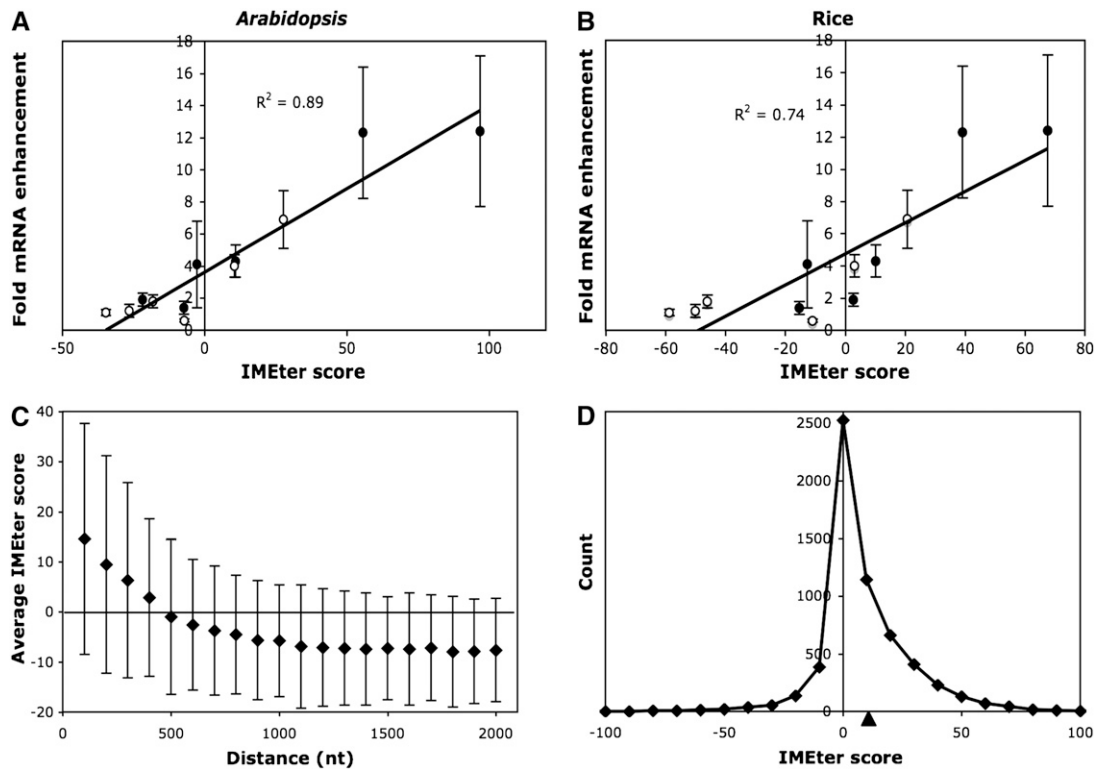
**Figure 2.** IMEter Scores in *Arabidopsis* and Rice Introns.

**(A)** Correlation between IMEter scores and the enhancing ability of *Arabidopsis* introns in *Arabidopsis*. Solid symbols indicate the original six introns tested, and open symbols represent introns chosen for their IMEter scores. Error bars indicate SD (*n* varies from 5 to 27 as detailed in Supplemental Table 2 online).

**(B)** Correlation between IMEter scores and the enhancing ability of *Arabidopsis* introns in *Arabidopsis* using an IMEter trained with the sequence of rice introns. Symbols are as in **(A)**.

**(C)** IMEter scores vary with intron location. Introns were separated into groups based on the distance between the start of transcription of the gene in which they are found and the first nucleotide of the intron. An *x* axis value of 100 includes introns up to 100 nucleotides from the start, 200 indicates introns between 101 and 200 nucleotides from the start, and so on. The average IMEter score of the introns in each group is shown, with the error bars indicating SD (*n* ranges from 943 to 1760 introns per data point).

**(D)** Distribution of IMEter scores in promoter-proximal introns. IMEter scores were calculated for introns located <400 bp from the start of transcription. The mean score (10.6) is indicated by the triangle.

potential for gene-specific interactions between promoters and introns makes this approach quite reasonable but complicates even intrastudy comparisons of different introns. With this caveat, introns whose effects on expression are comparable to their IMEter scores include those from the histone H3, *Cox5c*, *PRF*, *UBQ*, and *tua* genes (Tables 1 and 2). The largest discrepancy between IMEter score and stimulation is the very low-scoring rice *RPBF* intron, which has a larger effect than the high-scoring *GAMyb* intron (Washio and Morikawa, 2006). Determining whether the *RPBF* intron increases expression by IME or an unrelated mechanism, such as an enhancer element, alternative promoter, or some other positive regulatory element located within the intron, will require further experimentation.

Monocot and dicot introns differ in several ways. For example, the GC content of introns is 32% in *Arabidopsis* and 39% in rice, and the average length of introns in rice is more than twice that

in *Arabidopsis*. In addition, monocots are better able to splice introns from other species or introns containing stem-loop structures than are dicots (Keith and Chua, 1986; Goodall and Filipowicz, 1991), indicating that some aspects of intron recognition and splicing have diverged in these two groups. Despite these differences, the intron sequences that mediate enhancement must be sufficiently conserved for a rice-trained IMEter to generate scores for *Arabidopsis* introns that correlate with their ability to elevate expression in *Arabidopsis*. The decline in average IMEter scores with distance from the start of transcription (Figure 2C; see Supplemental Figure 3B online) and the inability of enhancing introns to affect expression from the 3′ UTR in both *Arabidopsis* and rice (Snowden et al., 1996; Rose, 2004) are further evidence that the mechanism of IME is conserved in these plants.

Why should enhancing signals be most common in proximal introns? Transcripts that originate from spurious promoters in the
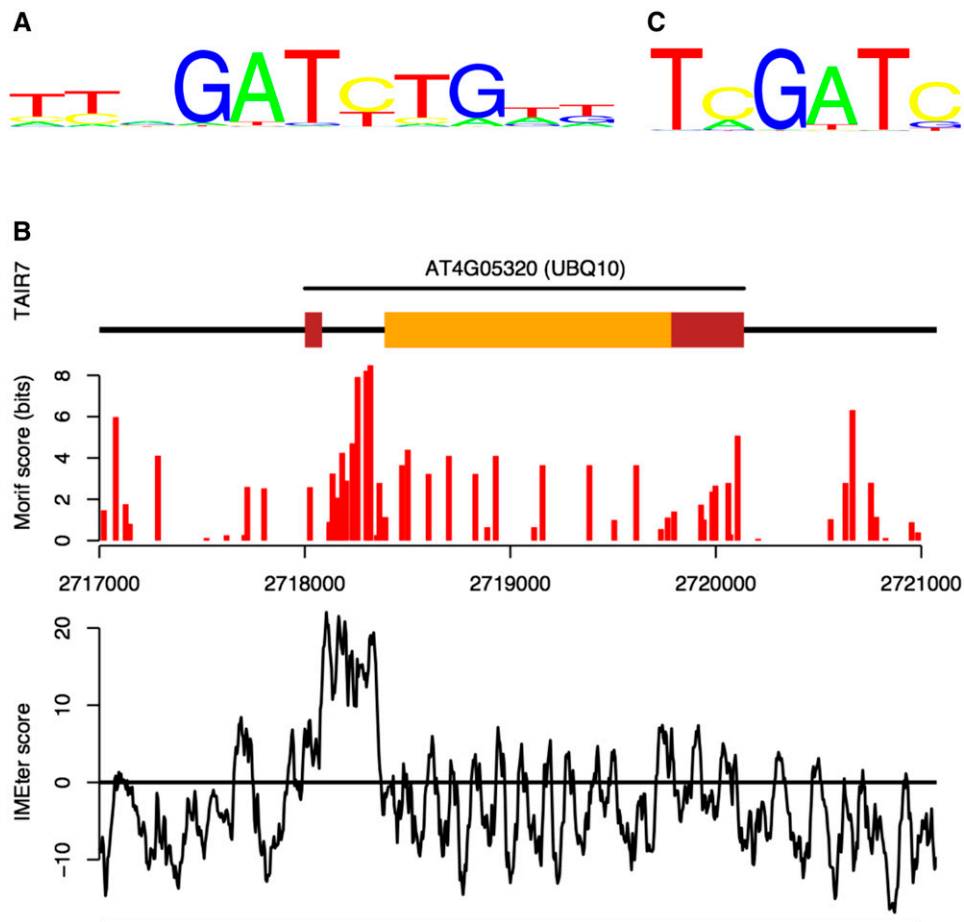
**Figure 3.** Motifs in Enhancing Introns.

**(A)** The most common motif found using NestedMICA in high-scoring *Arabidopsis* introns. The height of each letter reflects nucleotide frequency at that position.
**(B)** Distribution of IMEter scores and motifs within the *UBQ10* gene. The exon-intron structure is shown at the top with the 5′ and 3′ untranslated regions (UTRs) shaded darker. Vertical lines represent positions of the motif shown in **(A)** where the degree of matching is indicated by the height of the line. The IMEter score density in 50-bp windows is shown at the bottom.
**(C)** The most common motif found using NestedMICA in high-scoring rice introns.

body of a gene are potentially dangerous because they could encode partial proteins with dominant-negative effects, and those from intergenic regions could result in antisense transcripts that could silence genes. A close association of promoters with signals that increase processivity would help to ensure that the majority of stable RNAs from a particular locus correspond to complete mRNAs. The C-terminal domain of the largest subunit of RNA polymerase II is a likely candidate for regulation by introns because the C-terminal domain is known to bind to splicing factors and also to regulate the activity of the transcription machinery (Meinhart et al., 2005). Other mechanisms are also possible, and even those in which intron sequences operate at the DNA or chromatin level have not been ruled out. While the available biochemical data do not support intron-specific differences in transcript initiation (Dean et al., 1989; Rose and Last, 1997), splicing efficiency (Rose, 2002), or RNA stability (Rethmeier

et al., 1997), future technological improvements may reveal effects missed in earlier studies. The exon junction complex proteins deposited on mRNA during splicing promote nuclear export in *Xenopus* (Zhou et al., 2000) and increase association of the mRNA with polysomes in mammals (Wiegand et al., 2003; Nott et al., 2004). Similar mechanisms may operate in plants, as the increase in mRNA accumulation does not fully account for the boost in enzyme activity caused by introns in *Arabidopsis* (Rose, 2004).

It is unclear why some genes contain enhancing introns and some do not. Although highly expressed genes tend to have more enhancing signals than do poorly expressed ones, many abundantly expressed genes have introns with low IMEter scores. Perhaps there are several mechanisms to ensure complete transcription, and the IMEter is detecting only one kind. Alternatively, IME-like signals may be present in the 5′ UTR or coding

sequences of these genes, bypassing the need for enhancing introns. Experiments to test these possibilities are underway.

The success of many biotechnology ventures depends on optimizing the expression of a transgene from another species. To elevate expression, the transgene should be engineered to contain an intron from the host species to avoid cross-species problems in intron recognition and splicing. The IMEter will be beneficial for finding host introns with a desired degree of enhancement or for identifying IME motifs whose abundance in an intron can be modified. These motifs will also form the basis for the biochemical isolation of factors involved in the mechanism of IME.

## METHODS

### Hybrid Introns

To introduce *Bam*HI sites into the *UBQ10* and *COR15a* introns, the regions of the intron 5′ and 3′ to the target site were amplified in separate PCR reactions using primers (see Supplemental Figure 5 online) containing the desired change. The PCR products were cloned and subsequently ligated together using the *Bam*HI site, and the entire intron was verified by sequencing. The *Bgl*II or *Bcl*I sites were introduced by the same strategy into introns in which the *Bam*HI site had been created. Hybrids between the modified *UBQ10* and *COR15a* introns were made by conventional cloning using *Bam*HI, *Bcl*I, and *Bgl*II, all of which generate the same cohesive end.

### Expression Analysis

Introns were inserted as a *Pst* I restriction fragment between exons 1 and 2 of a translational fusion between the *Arabidopsis thaliana TRP1* gene and the *Escherichia coli uidA* gene (*GUS*) as described previously (Rose, 2002). Following *Agrobacterium tumefaciens*–mediated transformation of the constructs into *Arabidopsis* (ecotype Columbia) by floral dip (Clough and Bent, 1998), lines containing a single copy of the transgene were identified by individually placing 100 T2 seeds derived from each of 36 T1 plants on agar medium containing 50 µg mL$^{-1}$ kanamycin. Lines that segregated kanamycin resistance and sensitivity in a 3:1 ratio were subjected to three gel blots in which genomic DNA was digested with *Bam*HI, *Pst*I, or *Bgl*II. Homozygous lines in which all three enzymes indicated a single-copy insert were identified as those whose T3 progeny were 100% kanamycin resistant. All of the single-copy lines obtained (between one and seven for each construct) were used in expression studies. Plants were grown in Sunshine Mix #1 (Sun Gro Horticulture) at 20°C and 60% humidity under continuous light. Total RNA was isolated from leaves of homozygous 3-week-old T3 plants using the RNeasy Plant mini kit (Qiagen). Steady state mRNA levels were determined in at least two biological replicates by hybridizing gel blots of total RNA with $^{32}$P-labeled *GUS* and *TRP1* probes (Rose, 2004). Because the *TRP1* probe includes only exons 4 through 9, it hybridizes to endogenous but not *TRP1:GUS* transcripts. The filters were scanned with a STORM 860 PhosphorImager and analyzed using ImageQuant 5.2 software (Molecular Dynamics). After correcting for differences in loading using the endogenous *TRP1* mRNA signal, the fold enhancement by an intron was calculated as the amount of *TRP1:GUS* mRNA relative to that in an intronless *TRP1:GUS* control line. Because the variation in expression between different lines containing the same construct on a single blot was less than the variation of a single line in replicate blots (see Supplemental Table 1 online), all data points for a construct ($n \geq 7$) were given equal weight in determining mean expression and standard deviations. The mRNA quantification was verified using GUS enzyme assays as detailed in Supplemental Table 2 online. Intron splicing efficiency was determined by reverse transcription using random hexamers, followed by PCR with primers that flank the intron as described (Rose, 2002).

### Genome Data

The Arabidopsis genome sequence and annotation (TAIR7) were downloaded from The Institute for Genomic Research (TIGR; http://www.tigr.org/) and processed to ensure all the genes were complete, nonredundant, and of high quality. Since intron position was an important factor, and introns may occur in UTRs, genes were required to contain both 5′ and 3′ UTRs. To remove redundancy, only one isoform of any gene was kept, and any genes that were highly similar to other genes (>95% identical over 400 nucleotides) were removed. The decision of which isoform or gene to keep was based on order of precedence in the data file. Genes with unusually short coding sequences (<50 amino acids), short introns (<60 bp), or long introns (>1000 bp) were removed to ensure that all the genes were of high quality. The final set comprised 14,220 genes (53% of known genes) and 80,784 introns. The set was split randomly in half to create training and testing sets. Accession numbers of the genes in each set are listed in Supplemental Data Set 1 online.

The *Oryza sativa* genome and annotation (release 5) were also downloaded from TIGR. The processing procedure was similar to that used for *Arabidopsis* with one additional step. We found that some of the introns had abnormally high GC compositions and were possibly erroneous. Therefore, genes containing introns with >50% GC were removed. The final set consisted of 11,857 genes and 61,753 introns that were divided into training and test sets as listed in Supplemental Data Set 1 online. The *Caenorhabditis elegans* and *Drosophila melanogaster* genes were obtained from a previous collection (Korf, 2004).

### The IMEter Algorithm

For each possible subsequence of length K in a test intron (except the conserved sequences at the 5′ and 3′ ends), its observed frequency in the group of promoter-proximal introns is divided by its expected occurrence in sequence with the average nucleotide composition of all the introns in a genome. The logarithms of the observed/expected ratios are summed, and the process is repeated using the observed frequencies in distal introns. The difference between these sums is the IMEter score, which will be positive for introns that resemble promoter-proximal introns. This can be expressed as follows:

$$S = \sum_{i=1+D}^{i \leq L-K-A} \log\left(\frac{P_{wi}}{Q_{wi}}\right),$$

where $S$ is IMEter score, $i$ is the position in the sequence, $L$ is the length of the intron, $K$ is the word size, $w_i$ is a word of length $K$ starting at position $i$, $D$ is the length of the splice donor site consensus (usually 5), $A$ is the length of the splice acceptor site consensus (usually 10), $P$ is the frequency distribution of words of size $K$ in promoter-proximal introns, and $Q$ is the frequency distribution of words of size $K$ in distal introns. The IMEter program is implemented in Perl with command line options for the various parameters. The IMEter is available for public use at http://korflab.ucdavis.edu/cgi-bin/web-imeter.pl, and the software is available from the authors on request.

### Accession Numbers

The six introns chosen for testing on the basis of their IMEter scores are from the genes At1g05810, At1g62170, At1g69930, At3g04150, At3g21970, and At5g61930. The introns previously tested (Rose, 2002) are from *UBQ10* (At4g05230), *ATPK1* (At3g08730), *TCH3* (At2g41100), *COR15a* (At2g425400), and *TRP1* (formerly known as *PAT1*: At5g17990).

Accession numbers of the genes used to train and test the IMEter in *Arabidopsis* and rice are in Supplemental Data Set 1 online. Additional accession numbers can be found in Table 1.

## Supplemental Data

The following materials are available in the online version of this article.

**Supplemental Figure 1.** Optimizing IMEter Settings.

**Supplemental Figure 2.** Motifs in High-Scoring *Arabidopsis* Introns.

**Supplemental Figure 3.** IMEter Scores in Rice Introns.

**Supplemental Figure 4.** Motifs in High-Scoring Rice Introns.

**Supplemental Figure 5.** Strategy for Introducing Restriction Sites.

**Supplemental Table 1.** Expression Data.

**Supplemental Table 2.** Verification of mRNA Quantification.

**Supplemental Data Set 1.** Genes Used to Train and Test the IMEter in *Arabidopsis* and Rice.

## REFERENCES

**Baker, S.S., Wilhelm, K.S., and Thomashow, M.F.** (1994). The 5′-region of *Arabidopsis thaliana cor15a* has cis-acting elements that confer cold-, drought-, and ABA-regulated gene expression. Plant Mol. Biol. **24:** 701–713.

**Buchman, A.R., and Berg, P.** (1988). Comparison of intron-dependent and intron-independent gene expression. Mol. Cell. Biol. **8:** 4395–4405.

**Callis, J., Fromm, M., and Walbot, V.** (1987). Introns increase gene expression in cultured maize cells. Genes Dev. **1:** 1183–1200.

**Casas-Mollano, J.A., Lao, N.T., and Kavanagh, T.A.** (2006). Intron-regulated expression of *SUVH3*, an *Arabidopsis Su(var)3-9* homologue. J. Exp. Bot. **57:** 3301–3311.

**Chaubet-Gigot, N., Kapros, T., Flenet, M., Kahn, K., Gigot, C., and Waterborg, J.H.** (2001). Tissue-dependent enhancement of transgene expression by introns of replacement histone H3 genes of *Arabidopsis.* Plant Mol. Biol. **45:** 17–30.

**Chung, B.Y., Simons, C., Firth, A.E., Brown, C.M., and Hellens, R.P.** (2006). Effect of 5′UTR introns on gene expression in *Arabidopsis thaliana.* BMC Genomics **7:** 120.

**Clancy, M., and Hannah, L.C.** (2002). Splicing of the maize *Sh1* first intron is essential for enhancement of gene expression, and a T-rich motif increases expression without affecting splicing. Plant Physiol. **130:** 918–929.

**Clancy, M., Vasil, V., Hannah, L.C., and Vassil, I.K.** (1994). Maize *Shruken-1* intron and exon regions increase gene expression in maize protoplasts. Plant Sci. **98:** 151–161.

**Clough, S.J., and Bent, A.F.** (1998). Floral dip: A simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana.* Plant J. **16:** 735–743.

**Curi, G.C., Chan, R.L., and Gonzalez, D.H.** (2005). The leader intron of *Arabidopsis thaliana* genes encoding cytochrome *c* oxidase subunit 5c promotes high-level expression by increasing transcript abundance and translation efficiency. J. Exp. Bot. **56:** 2563–2571.

**Curie, C., Axelos, M., Bardet, C., Atanassova, R., Chaubet, N., and Lescure, B.** (1993). Modular organization and development activity of an *Arabidopsis thaliana* EF-1α gene promoter. Mol. Gen. Genet. **238:** 428–436.

**Czempinski, K., Frachisse, J.M., Maurel, C., Barbier-Brygoo, H., and Mueller-Roeber, B.** (2002). Vacuolar membrane localization of the *Arabidopsis* 'two-pore' K$^+$ channel KCO1. Plant J. **29:** 809–820.

**David-Assael, O., Berezin, I., Shoshani-Knaai, N., Saul, H., Mizrachy-Dagri, T., Chen, J., Brook, E., and Shaul, O.** (2006). AtMHX is an auxin and ABA-regulated transporter whose expression pattern suggests a role in metal homeostatis in tissues with photosynthetic potential. Funct. Plant Biol. **33:** 661–672.

**Dean, C., Favreau, M., Bond-Nutter, D., Bedbrook, J., and Dunsmuir, P.** (1989). Sequences downstream of translation start regulate quantitative expression of two petunia *rbcS* genes. Plant Cell **1:** 201–208.

**Down, T.A., and Hubbard, T.J.** (2005). NestedMICA: Sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res. **33:** 1445–1453.

**Duncker, B.P., Davies, P.L., and Walker, V.K.** (1997). Introns boost transgene expression in *Drosophila melanogaster.* Mol. Gen. Genet. **254:** 291–296.

**Fiume, E., Christou, P., Giani, S., and Breviario, D.** (2004). Introns are key regulatory elements of rice tubulin expression. Planta **218:** 693–703.

**Fong, Y.W., and Zhou, Q.** (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature **414:** 929–933.

**Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J.** (2002). Promoter proximal splice sites enhance transcription. Genes Dev. **16:** 2792–2799.

**Gidekel, M., Jimenez, B., and Herrera-Estrella, L.** (1996). The first intron of the *Arabidopsis thaliana* gene coding for elongation factor 1β contains an enhancer-like element. Gene **170:** 201–206.

**Goodall, G.J., and Filipowicz, W.** (1991). Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J. **10:** 2635–2644.

**Ho, S.H., So, G.M., and Chow, K.L.** (2001). Postembryonic expression of *Caenorhabditis elegans mab-21* and its requirement in sensory ray differentiation. Dev. Dyn. **221:** 422–430.

**Jeon, J.S., Lee, S., Jung, K.H., Jun, S.H., Kim, C., and An, G.** (2000). Tissue-preferential expression of a rice α-tubulin gene, *OsTubA1*, mediated by the first intron. Plant Physiol. **123:** 1005–1014.

**Jeong, Y.M., Mun, J.H., Kim, H., Lee, S.Y., and Kim, S.G.** (2007). An upstream region in the first intron of petunia actin-depolymerizing factor 1 affects tissue-specific expression in transgenic Arabidopsis (*Arabidopsis thaliana*). Plant J. **50:** 230–239.

**Jeong, Y.M., Mun, J.H., Lee, I., Woo, J.C., Hong, C.B., and Kim, S.G.** (2006). Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members. Plant Physiol. **140:** 196–209.

**Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C., and Davis, R.W.** (2006). Introns regulate RNA and protein abundance in yeast. Genetics **174:** 511–518.

**Keith, B., and Chua, N.H.** (1986). Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco. EMBO J. **5:** 2419–2425.

**Kim, M.J., Kim, H., Shin, J.S., Chung, C.H., Ohlrogge, J.B., and Suh, M.C.** (2006). Seed-specific expression of sesame microsomal oleic acid desaturase is controlled by combinatorial properties between negative *cis*-regulatory elements in the *SeFAD2* promoter

and enhancers in the 5′-UTR intron. Mol. Genet. Genomics **276:** 351–368.

**Korf, I.** (2004). Gene finding in novel genomes. BMC Bioinformatics **5:** 59.

**Koziel, M.G., Carozzi, N.B., and Desai, N.** (1996). Optimizing expression of transgenes with an emphasis on post-transcriptional events. Plant Mol. Biol. **32:** 393–405.

**Lugones, L.G., Scholtmeijer, K., Klootwijk, R., and Wessels, J.G.** (1999). Introns are necessary for mRNA accumulation in *Schizophyllum commune.* Mol. Microbiol. **32:** 681–689.

**Maas, C., Laufs, J., Grant, S., Korfhage, C., and Werr, W.** (1991). The combination of a novel stimulatory element in the first exon of the maize *Shrunken-1* gene with the following intron 1 enhances reporter gene expression up to 1000-fold. Plant Mol. Biol. **16:** 199–207.

**Mascarenhas, D., Mettler, I.J., Pierce, D.A., and Lowe, H.W.** (1990). Intron-mediated enhancement of heterologous gene expression in maize. Plant Mol. Biol. **15:** 913–920.

**McElroy, D., Zhang, W., Cao, J., and Wu, R.** (1990). Isolation of an efficient actin promoter for use in rice transformation. Plant Cell **2:** 163–171.

**Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S., and Cramer, P.** (2005). A structural perspective of CTD function. Genes Dev. **19:** 1401–1415.

**Morello, L., Bardini, M., Cricri, M., Sala, F., and Breviario, D.** (2006). Functional analysis of DNA sequences controlling the expression of the rice *OsCDPK2* gene. Planta **223:** 479–491.

**Morello, L., Bardini, M., Sala, F., and Breviario, D.** (2002). A long leader intron of the *Ostub16* rice β-tubulin gene is required for high-level gene expression and can autonomously promote transcription both *in vivo* and *in vitro.* Plant J. **29:** 33–44.

**Norris, S.R., Meyer, S.E., and Callis, J.** (1993). The intron of *Arabidopsis thaliana* polyubiquitin genes is conserved in location and is a quantitative determinant of chimeric gene expression. Plant Mol. Biol. **21:** 895–906.

**Nott, A., Le Hir, H., and Moore, M.J.** (2004). Splicing enhances translation in mammalian cells: An additional function of the exon junction complex. Genes Dev. **18:** 210–222.

**Nott, A., Meislin, S.H., and Moore, M.J.** (2003). A quantitative analysis of intron effects on mammalian gene expression. RNA **9:** 607–617.

**Okkema, P.G., Harrison, S.W., Plunger, V., Aryana, A., and Fire, A.** (1993). Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans.* Genetics **135:** 385–404.

**Palmiter, R.D., Sandgren, E.P., Avarbock, M.R., Allen, D.D., and Brinster, R.L.** (1991). Heterologous introns can enhance expression of transgenes in mice. Proc. Natl. Acad. Sci. USA **88:** 478–482.

**Rethmeier, N., Seurinck, J., Van Montagu, M., and Cornelissen, M.** (1997). Intron-mediated enhancement of transgene expression in maize is a nuclear, gene-dependent process. Plant J. **12:** 895–899.

**Rose, A.B.** (2002). Requirements for intron-mediated enhancement of gene expression in *Arabidopsis.* RNA **8:** 1444–1453.

**Rose, A.B.** (2004). The effect of intron location on intron-mediated enhancement of gene expression in *Arabidopsis.* Plant J. **40:** 744–751.

**Rose, A.B., and Beliakoff, J.A.** (2000). Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. Plant Physiol. **122:** 535–542.

**Rose, A.B., and Last, R.L.** (1997). Introns act post-transcriptionally to

increase expression of the *Arabidopsis thaliana* tryptophan pathway gene *PAT1.* Plant J. **11:** 455–464.

**Schunmann, P.H., Richardson, A.E., Smith, F.W., and Delhaize, E.** (2004). Characterization of promoter expression patterns derived from the Pht1 phosphate transporter genes of barley (*Hordeum vulgare* L.). J. Exp. Bot. **55:** 855–865.

**Simpson, G.G., and Filipowicz, W.** (1996). Splicing of precursors to mRNA in higher plants: Mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. Plant Mol. Biol. **32:** 1–41.

**Sistrunk, M.L., Antosiewicz, D.M., Purugganan, M.M., and Braam, J.** (1994). Arabidopsis *TCH3* encodes a novel Ca2+ binding protein and shows environmentally induced and tissue-specific regulation. Plant Cell **6:** 1553–1565.

**Sivamani, E., and Qu, R.** (2006). Expression enhancement of a rice polyubiquitin gene promoter. Plant Mol. Biol. **60:** 225–239.

**Snowden, K.C., Buchholz, W.G., and Hall, T.C.** (1996). Intron position affects expression from the *tpi* promoter in rice. Plant Mol. Biol. **31:** 689–692.

**Ueki, J., Ohta, S., Morioka, S., Komari, T., Kuwata, S., Kubo, T., and Imaseki, H**. (1999). The synergistic effects of two-intron insertions on heterologous gene expression and advantages of the first intron of a rice gene for phospholipase D. Plant Cell Physiol. **40:** 618–623.

**Vain, P., Finer, K.R., Engler, D.E., Pratt, R.C., and Finer, J.J.** (2004). Inron-mediated enhancement of gene expression in maize (*Zea mays* L.) and bluegrass (*Poa pratensis* L). Plant Cell Rep. **15:** 489–494.

**Virts, E.L., and Raschke, W.C.** (2001). The role of intron sequences in high level expression from CD45 cDNA constructs. J. Biol. Chem. **276:** 19913–19920.

**Vitale, A., Wu, R.J., Cheng, Z., and Meagher, R.B.** (2003). Multiple conserved 5′ elements are required for high-level pollen expression of the *Arabidopsis* reproductive actin *ACT1.* Plant Mol. Biol. **52:** 1135–1151.

**Wang, H., Lee, M.M., and Schiefelbein, J.W.** (2002). Regulation of the cell expansion gene *RHD3* during Arabidopsis development. Plant Physiol. **129:** 638–649.

**Washio, K., and Morikawa, M.** (2006). Common mechanisms regulating expression of rice aleurone genes that contribute to the primary response for gibberellin. Biochim. Biophys. Acta **1759:** 478–490.

**Wiegand, H.L., Lu, S., and Cullen, B.R.** (2003). Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. Proc. Natl. Acad. Sci. USA **100:** 11327–11332.

**Xu, J., and Gong, Z.Z.** (2003). Intron requirement for AFP gene expression in *Trichoderma viride.* Microbiology. **149:** 3093–3097.

**Xu, Y., Yu, H., and Hall, T.C.** (1994). Rice triosephosphate isomerase gene 5′ sequence directs β-glucuronidase activity in transgenic tobacco but requires an intron for expression in rice. Plant Physiol. **106:** 459–467.

**Zhang, J.Z., and Somerville, C.R.** (1997). Suspensor-derived poly-embryony caused by altered expression of valyl-tRNA synthetase in the *twn2* mutant of *Arabidopsis.* Proc. Natl. Acad. Sci. USA **94:** 7349–7355.

**Zhang, S.H., Lawton, M.A., Hunter, T., and Lamb, C.J.** (1994). *atpk1*, a novel ribosomal protein kinase gene from *Arabidopsis.* I. Isolation, characterization, and expression. J. Biol. Chem. **269:** 17586–17592.

**Zhou, Z., Luo, M.J., Straesser, K., Katahira, J., Hurt, E., and Reed, R.** (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature **407:** 401–405.

**Promoter-Proximal Introns in Arabidopsis thaliana Are Enriched in Dispersed Signals that Elevate Gene Expression**

Alan B. Rose, Tali Elfersi, Genis Parra and Ian Korf

*PLANT CELL* 2008;20;543-551; originally published online Mar 4, 2008;
DOI: 10.1105/tpc.107.057190

This information is current as of August 28, 2008