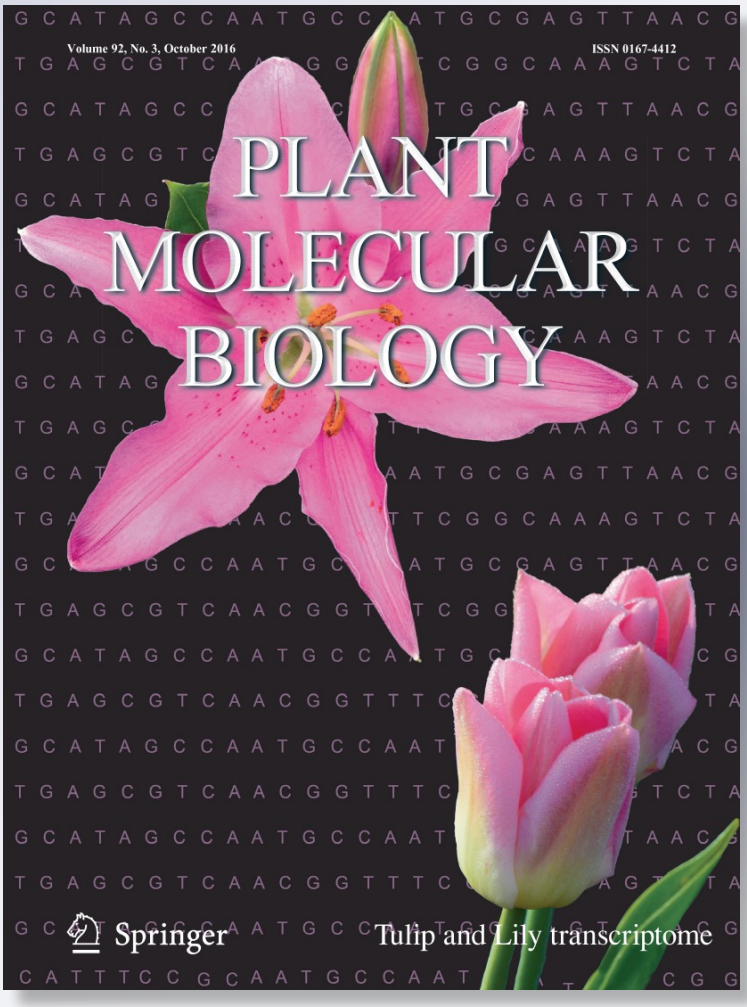
Intron sequences that stimulate gene expression in Arabidopsis

Alan B. Rose, Amanda Carter, Ian Korf & Noah Kojima

Plant Molecular Biology
An International Journal on Molecular Biology, Molecular Genetics and Biochemistry

ISSN 0167-4412
Volume 92
Number 3

Plant Mol Biol (2016) 92:337-346
DOI 10.1007/s11103-016-0516-1



Springer

Plant Mol Biol (2016) 92:337–346
DOI 10.1007/s11103-016-0516-1

Intron sequences that stimulate gene expression in *Arabidopsis*

Alan B. Rose[1] · Amanda Carter[1] · Ian Korf[1] · Noah Kojima[1,2]

Received: 4 March 2016 / Accepted: 14 July 2016 / Published online: 5 August 2016

Abstract
Key message **Related motifs strongly increase gene expression when added to an intron located in coding sequences.**
Abstract Many introns greatly increase gene expression through a mechanism that remains elusive. An obstacle to understanding intron-mediated enhancement (IME) has been the difficulty of locating the specific intron sequences responsible for boosting expression because they are redundant, dispersed, and degenerate. Previously we used the IMEter algorithm in two independent ways to identify two motifs (CGATT and TTNGATYTG) that are candidates for involvement in IME in *Arabidopsis*. Here we show that both motifs are sufficient to increase expression. An intron that has little influence on expression was converted into one that increased mRNA accumulation 24-fold and reporter enzyme activity 40-fold relative to the intronless control by introducing 11 copies of the more active TTNGATYTG motif. This degree of stimulation is twice as large as that of the strongest of 15 natural introns previously tested in the same reporter gene. Even though the CGATT and TTNGATYTG motifs each increased expression, and CGATT matches the NGATY core of the longer motif, combining the motifs to make TTCGATTTG reduced the stimulating ability of the TTNGATYTG motif. Additional substitutions were used to test the contribution to IME of other residues in the TTNGATYTG motif. The verification that these motifs are active in IME will improve our ability to predict the stimulating ability of introns, to engineer any intron to increase expression to a desired level, and to explore the mechanism of IME by seeking factors that might interact with these sequences.

Keywords Intron · Gene expression · *Cis*-element · Motif · *Arabidopsis*

Introduction

Shortly after introns were discovered, it was found that removing the introns from a gene can significantly reduce or completely eliminate its expression (Gruss et al. 1979). Similarly, adding an intron from the host organism to an intronless gene such as a bacterial reporter was often shown to increase its expression as a transgene (Buchman and Berg 1988; Callis et al. 1987; Palmiter et al. 1991). The broad diversity of organisms in which introns can increase expression (Duncker et al. 1997; Juneau et al. 2006; Lu and Cullen 2003; Okkema et al. 1993) suggests that gene regulation by introns is very widely conserved and therefore likely ancient. Investigations into the means through which introns boost expression have revealed several different mechanisms. For example, many introns contain transcriptional enhancer elements (Bianchi et al. 2009; Deyholos and Sieburth 2000), and the deposition of the exon junction complex proteins onto the mRNA during splicing can increase the efficiency with which an mRNA is exported from the nucleus and translated

Electronic supplementary material The online version of this article (doi:10.1007/s11103-016-0516-1) contains supplementary material, which is available to authorized users.

✉ Alan B. Rose
abrose@ucdavis.edu

[1] Department of Molecular and Cellular Biology, University of California, 1 Shields Avenue, Davis, CA 95616, USA

[2] Present address: David Geffen School of Medicine at the University of California, 10833 Le Conte Avenue, Los Angeles, CA 90095, USA

(Le Hir et al. 2003; Lu and Cullen 2003; Matsumoto et al. 1998; Nott et al. 2004; Wiegand et al. 2003). In addition to these well-known methods for increasing expression, many introns have a less understood effect that has been termed intron-mediated enhancement (IME). In its broadest sense, IME can include any case where an intron-containing construct is more highly expressed than an intronless counterpart, regardless of the basis for the difference. Here, we use the term IME more narrowly to mean an increase in mRNA accumulation caused by some introns but not others. Key characteristics that differentiate between IME and enhancer elements are that for IME, the intron must be transcribed (Callis et al. 1987; Jeon et al. 2000; Snowden et al. 1996) and within approximately 1 Kb of the promoter to increase mRNA accumulation (Rose 2004).

The observation that efficiently spliced introns vary widely in their effect on expression in plants (Rose 2002) suggests the existence of specific stimulatory sequences that are more abundant in some introns that in others. However, traditional approaches to identify sequences involved in IME have largely failed. The number of introns with measured effects on expression is insufficient to permit meaningful computational searches for sequences that are shared by stimulatory introns and are absent or reduced in non-stimulatory introns. Furthermore, deletion analysis has often shown that the stimulating ability of introns is remarkably unaffected by removing large portions (up to 90% of the intron) as long as the intron can be spliced (Chung et al. 2006; Clancy et al. 1994; Jeon et al. 2000; Luehrsen and Walbot 1994). Hybrid *Arabidopsis* introns constructed using the stimulatory *UBQ10* (At4g05320) intron and the non-stimulatory *COR15a* (At2g42540) intron revealed that the sequences responsible for increasing expression are dispersed throughout the *UBQ10* intron (Rose et al. 2008). The inability to identify the sequences involved in IME has hampered investigations into the mechanism of IME, and has limited our ability to predict which introns will affect expression and which will not.

Due to these limitations, very few candidate sequences for involvement in IME have been proposed. One is a 35 nt region from the 1028 nt first intron of the maize *Sh1* gene. The full length *Sh1* intron stimulates expression 20- to 50-fold, and a 145 nt derivative in which 86% of the intron is deleted has a similar effect (Clancy and Hannah 2002). Removing an additional 35 nt from the 145 nt deletion derivative causes the stimulation to drop by two-thirds (Clancy and Hannah 2002). The authors concluded that T-richness was more important for IME than precise sequence because an unrelated 35 nt T-rich sequence from another part of the wild-type *Sh1* intron is equally effective at promoting expression when substituted for the original 35 nt sequence. Derivatives of the *Arabidopsis TRP1* (At5g17990) intron in which the T-richness of the intron was raised or lowered also led to the conclusion that IME is caused by T-rich sequences, and that clusters of Ts are more important than isolated T residues for intron recognition and IME (Rose 2002). In contrast to these T-rich sequences, the only other candidate to have been implicated in IME in any organism is GTGCCGCG. This GC-rich octamer was found by comparing the sequences of three stimulating introns from the maize *GapA1, Sh1*, and *Adh1* genes. A *GapA1:GUS* fusion, which is not expressed at all without an intron, is expressed at a low level when a 92 bp fragment containing four copies of this motif is used to replace the first *GapA1* intron (Donath et al. 1995). Deletion and hybrid intron analysis identified a 118 nt region of the *Arabidopsis AtMHX* leader intron that increases expression, although this fragment has a much larger effect on translation than mRNA accumulation (Akua and Shaul 2013). The vague and inconsistent natures of the sequences previously proposed have left the sequences responsible for increasing mRNA levels largely undefined.

Alternative approaches to search for the sequences involved in IME were made possible by the development of the IMEter algorithm (Rose et al. 2008). The IMEter is based on the reasoning that because IME requires introns to be near the promoter to increase expression, the composition of promoter-proximal introns might differ from other introns. Any differences in average composition may be due to the subset of promoter-proximal introns that increase expression by an IME mechanism. The IMEter software separates all the introns in a genome into two groups, promoter-proximal and distal, and then calculates the frequency of occurrence of all possible k-mers of a given length (such as pentamers) in each group. The intron under study (or any sequence) is compared to each profile, generating an IMEter score that reflects the degree to which that intron or sequence resembles promoter-proximal introns. The strong correlation between the IMEter score of an intron and its ability to increase mRNA accumulation supports the idea that the IMEter detects the sequences responsible for IME that are enriched in promoter-proximal introns (Akua and Shaul 2013; Rose et al. 2008).

Even though the IMEter does not directly reveal which sequences are involved in IME, it has been used in two ways to identify candidate sequences. One was to search for sequences that are over-represented in introns with high IMEter scores using motif-finding programs such as NestedMICA or MEME (Bailey et al. 2006; Down and Hubbard 2005). This led to the finding that the sequence TTNGATYTG (N = any nucleotide, Y = C or T) is more abundant in introns with high IMEter scores than in low-scoring introns (Rose et al. 2008). In the set of natural *Arabidopsis* introns whose effect on expression has been measured in single-copy transgenic lines, the number of matches to this motif correlates with the ability of each intron to stimulate

mRNA accumulation (Rose et al. 2008). The other approach was to find the regions within a stimulating intron that make the largest contribution to its overall IMEter score, and presumably have the largest effect on expression, by scanning the sequence with a sliding window. Of the eleven highest peaks in IMEter score in the strongly stimulating 304 nt *UBQ10* intron, six are due to the sequence CGATT, and three result from one instance each of the similar sequences CGAAT, CGATC, and AGATC. Rearranging the sequences responsible for all eleven peaks through a total of 46 nt changes reduces the stimulating ability of this intron by 47% (from 13 times more steady-state mRNA than the intronless control for the wild-type intron to 7 times more mRNA for the modified intron) (Parra et al. 2011). The residual activity of this intron indicates that other sequences are also involved in IME. Furthermore, the non-stimulating *COR15a* intron is converted into one that stimulates mRNA accumulation more than sixfold by adding eleven copies of CGATT (Parra et al. 2011). These results support the idea that CGATT and related sequences contribute to IME.

Here we report additional testing of the CGATT and TTNGATYTG motifs for their involvement in IME by exploring their capacity to increase the stimulating ability of the *COR15a* intron. We found that rearranging nucleotides within the *COR15a* intron to make six or eleven perfect matches to either motif changed it into a stimulating intron, in one case far exceeding the effects of strongly stimulating natural introns. The importance of specific nucleotides within the TTNGATYTG motif was also explored by mutating individual and combinations of residues.

Results

Comparing the CGATT and TTNGATYTG motifs

To test the relevance of the CGATT and TTNGATYTG motifs in IME, derivatives of the *COR15a* intron were designed in which six or eleven copies of either motif were introduced (Fig. 1). To maintain the length and nucleotide composition of the intron, motifs were made by identifying stretches of contiguous nucleotides with the necessary composition and then rearranging the order of nucleotides to match the motif (Fig. 1b, c). To make CGATT, regions of five to seven nucleotides composed of at least two Ts and one each of A, C, and G were rearranged. When the starting sequence was longer than 5 nucleotides, the sequence created was CGATT followed by the extra nucleotide(s). Similarly, all but one of the TTNGATYTG motifs were made by rearranging 9 nt sequences composed of at least one A, two Gs, and either five Ts or four Ts and one C, with the remaining nucleotide used as the N at position 3 of the motif. The other copy of TTNGATYTG was generated from an 11 nt starting sequence (Fig. 1c). The introns with eleven instances of either motif contained the same changes as the introns with six copies plus an additional five copies (Fig. 1a). The resulting introns that contain 6 or 11 instances of the short (CGATT) or long (TTNGATYTG) motif were named the *COR15a*6S, *COR15a*11S, *COR15a*6L, and *COR15a*11L introns. The *COR15a*11S intron was reported previously (Parra et al. 2011). Each intron was inserted at the same location between exons 1 and 2 of a *TRP1:GUS* reporter gene (Fig. 1d), which was introduced into *Arabidopsis* plants by *Agrobacterium*-mediated transformation. Steady-state levels of *GUS* mRNA and enzyme activity were measured in leaves of 3-week-old homozygous single-copy transgenic plants as described (Rose et al. 2008). The effect of an intron on expression was calculated as the fold increase in *GUS* mRNA or enzyme activity derived from the intron-containing *TRP1:GUS* fusion relative to that in a control line containing an otherwise identical intronless *TRP1:GUS* fusion. Complete expression results are presented in Table 1 and Supplementary Table 1, and a sample RNA gel blot is shown in Fig. 2.

The *COR15a*6S and *COR15a*11S introns increased mRNA accumulation 5.7-fold and 7.0-fold respectively relative to the intronless control (Fig. 2; Table 1), although the difference between these introns is not statistically significant. The value for the *COR15a*11S intron differed slightly from that previously published [6.6-fold (Parra et al. 2011)] because the average now includes ten new data points in addition to the original 15. The observed values are very similar to the predicted stimulation of these introns of 5.3- and 6.5-fold respectively based on their IMEter 2.0 scores and the best-fit line of IMEter 2.0 scores plotted versus mRNA accumulation of all wild-type introns previously tested in this reporter gene (Fig. 3). Thus, the effect of the sequence CGATT on mRNA accumulation was consistent with its influence on IMEter score. As was observed previously (Rose 2004), the effect of the intron measured at the level of GUS enzyme activity was approximately twice that seen at the level of mRNA accumulation (Table 1). The exon junction complex proteins that are deposited on spliced mRNAs might increase the yield of protein produced per unit of mRNA by facilitating mRNA export and its association with ribosomes (Matsumoto et al. 1998; Nott et al. 2004; Wiegand et al. 2003).

The TTNGATYTG motif had a substantially greater ability to convert the *COR15a* intron into one that significantly increased gene expression. The *COR15a*6L intron increased *GUS* mRNA accumulation 14-fold, similar to the effect of the strongly stimulating *UBQ10* intron (Table 1). The *COR15a*11L intron stimulated mRNA accumulation a remarkable 24-fold and GUS activity 40-fold, nearly twice as much as the strongest of the 15 natural introns ever tested in the *TRP1:GUS* reporter gene (Fig. 3). The effects on

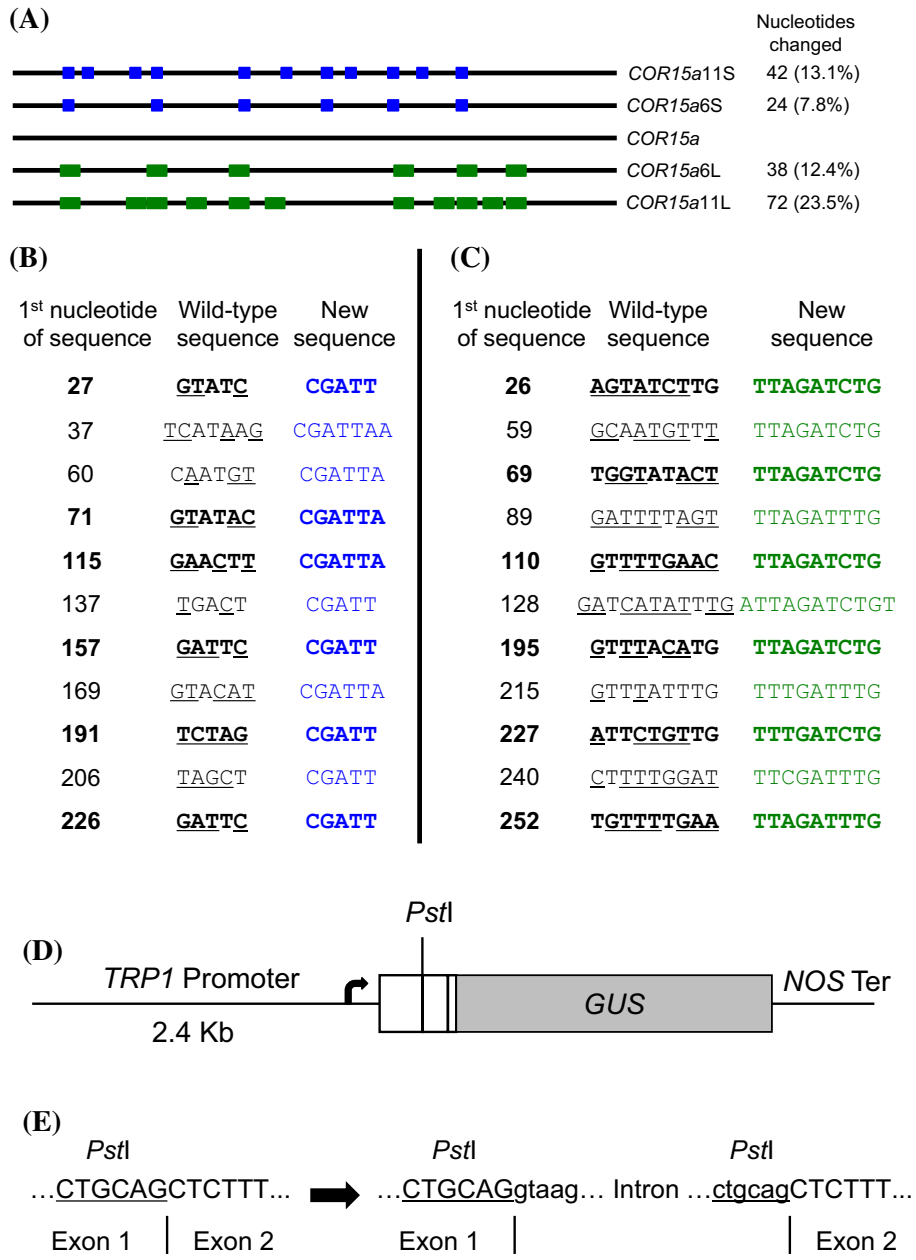


Fig. 1 Details of the sequence changes made to introduce motifs and the *TRP1:GUS* reporter gene. **a** Diagram of the locations of the CGATT (*blue* rectangles) and TTNGATYTG (*green* rectangles) motifs created in the *COR15a* intron. **b, c** The sequences rearranged to match the CGATT (**b**) and TTNGATYTG (**c**) motifs. The *rows not in bold* indicate changes present only in introns with 11 copies of the motif. The nucleotides altered are underlined in the wild-type sequences. **d** The *TRP1:GUS* reporter gene fusion. The *arrow* denotes the start of transcription, *white rectangles* indicate *TRP1* protein coding exons, the *grey box* represents the *GUS* gene, and *NOS* Ter indicates the transcriptional terminator from the *nopaline synthase* gene. **e** Sequence details of intron insertion. Introns, immediately preceded by a *Pst*I site and in which the last six nucleotides form a *Pst*I site, were inserted into the *Pst*I site at the 3′ end of *TRP1* exon 1. Splicing at the normal splice sites generates a mature mRNA that is identical in sequence to that from the intronless control. *Upper case and lower case letters* indicate exon and intron sequences respectively

mRNA accumulation of the *COR15a*6L and *COR15a*11L introns were much greater than would be expected based on their IMEter 2.0 scores (Fig. 3). This suggests that our ability to predict the effect of an intron on mRNA production could be improved by considering both its IMEter score and the number of TTNGATYTG motifs it contains.

Combining the CGATT and TTNGATYTG motifs

The CGATT and TTNGATYTG motifs might represent different manifestations of the same stimulating element because the CGATT motif matches the central NGATY portion of the longer motif. If the sequences are functionally related, combining motifs to make the sequence TTCGATTTG might have a larger effect on expression than the TTNGATYTG motif. In the *COR15a*6L intron, the N at position 3 was an A in five of the six motifs, and the Y at position 7 was a C in five of the copies (Fig. 1c). To test the effects of a combined motif, a derivative of the *COR15a*6L intron was made in which all six copies of the motif had the N at position 3 converted to a C and the Y at position 7 converted to a T. The resulting intron, named *COR15a*6L(C3T7), is the same length as the *COR15a* intron but has a slightly different composition due to the 11 nucleotide changes necessary to create the desired sequence. The *COR15a*6L(C3T7) intron increased mRNA accumulation 6.7-fold (Table 1), which is roughly half the effect of the *COR15a*6L intron (14-fold). A derivative of the *COR15a*6L intron, *COR15a*6L(C3), in

Table 1 The effect on expression of *COR15a* intron 1 modified to contain motifs

Intron name	Motif sequence	Number added	IMEter 2.0 score	Increase in mRNA[a]	Increase in GUS activity[a]
COR15a			7.9	1.9 ± 0.4	2.3 ± 0.7
*COR15a*6S	CGATT	6	21.8	5.7 ± 1.2	9.0 ± 1.7
*COR15a*11S	CGATT	11	27.7	7.0 ± 1.0	12.7 ± 2.1
*COR15a*6L	TTNGATYTG	6	14.2	14.0 ± 2.5	25.9 ± 5.1
*COR15a*11L	TTNGATYTG	11	25.4	24.1 ± 3.2	39.9 ± 16.1
*COR15a*6L(C3)	TT**C**GATYTG	6	26.9	10.1 ± 1.9	15.9 ± 2.6
*COR15a*6L(C3T7)	TT**C**GAT**T**TG	6	28.2	6.7 ± 0.8	11.2 ± 2.1
*COR15a*6L(T5A6)	TTNG**TA**YTG	6	4.7	3.0 ± 0.4	3.5 ± 0.5
*COR15a*6L(A7)	TTNGAT**A**TG	6	5.3	3.9 ± 0.6	5.7 ± 1.5
*COR15a*6L(TACTG)	TT**TAC**T**G**TG	6	4.7	3.6 ± 0.7	5.2 ± 1.0
UBQ10			49.0	14.6 ± 2.9	25.1 ± 5.3

Nucleotides changed from the starting motif are bold and underlined

[a]Relative to the intronless control. Numbers are mean ± standard deviation

which only the N position of each motif was converted to C, but the Y at position 7 was the same as in the *COR15a*6L intron (usually a C), had an intermediate effect, increasing mRNA accumulation 10-fold (Table 1). Therefore, changing the nucleotide at positions 3 to a C, and the residue at position 7 to a T, had similarly deleterious effects on expression. Thus, even though the CGATT motif increased mRNA accumulation as separate sequences within an intron, it reduced the activity of the TTNGATYTG motif.

Testing nucleotides within the TTNGATYTG motif

The greater ability of the long motif than the short motif to increase expression suggests that the dinucleotides at each end of the TTNGATYTG motif might influence motif activity. The nucleotides flanking the six CGATT motifs in the *COR15a*6S intron are TA…TT, TG…AT, TT…AA, AA…AT, TT…TT, and AT…TG, which include some partial matches to TT…TG but are generally more A-rich. To test the importance of the nucleotides at each end of the TTNGATYTG motif, a derivative of the *COR15a*6L intron was made in which the central five nucleotides of each motif were converted to a presumed non-stimulating sequence, so that any residual motif effect must be due to the dinucleotides at each end. The non-stimulating sequence chosen was TACTG, which has a very low IMEter score and reduces the stimulating ability of the *UBQ10* intron when substituted for the sequence CGATT (Parra et al. 2011). The resulting intron, *COR15a*6L(TACTG), increased mRNA accumulation 3.6-fold, which is twice the effect of the *COR15a* intron. This suggests that TT at the beginning of the motif and the TG at the end had a small effect on expression.

Two additional derivatives of the TTNGATYTG motif were constructed to test the functional significance of conserved nucleotides. In the *COR15a*6L(A7) derivative of the *COR15a*6L intron, all six copies of the motif contained the purine A rather than a pyrimidine at position 7. In the other, the GAT triplet at the center of each motif was converted to GTA, creating the *COR15a*6L(T5A6) derivative. Both modified motifs had significantly less effect on expression than did the TTNGATYTG motif. The *COR15a*6L(A7) and *COR15a*6L(T5A6) introns increased mRNA accumulation 3.9-fold and 3.0-fold, respectively (Table 1). In both cases, most of the remaining activity was probably contributed by the terminal TT and TG dinucleotides, as the effect on mRNA accumulation of the *COR15a*6L(A7) and *COR15a*6L(T5A6) introns was similar to that of the *COR15a*6L(TACTG) intron. Because swapping the order of the nucleotides at positions 5 and 6, or changing position 7 from a pyrimidine to an A, caused as great a reduction in motif activity as changing all five central nucleotides, these results suggest that the nucleotides at position 5, 6, and 7 of the TTNGATYTG motif are very important for IME function and are sensitive to change.

Line to line variation

The difference in expression between all independent single-copy lines containing the same construct was very small. In 44 (80 %) of the 55 instances in which more than one line containing the same construct was tested on the same blot (Supplementary Table 1), the amount of *TRP1:GUS* mRNA in the most highly expressed line differed from that in the lowest expressing line by less than 20 %, and in only three cases (5.5 %) did the ratio between the highest and lowest expressing lines exceed 1.4 (1.41, 1.56, and 1.68). The consistency between single-copy lines suggests that the results are reliable for the two constructs (containing introns *COR15a*6L(C3T7) and *COR15a*6L(A7)) for which only one single-copy line was identified, although additional lines might yield slightly different results.

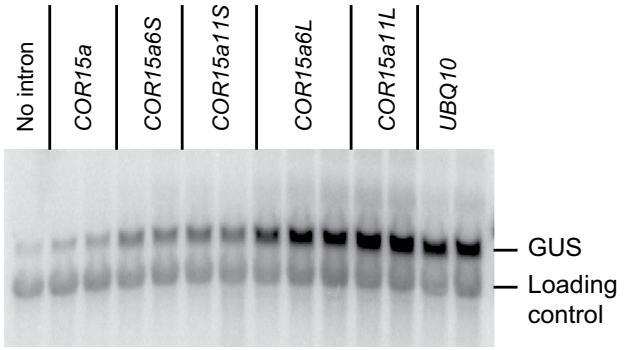


Fig. 2 RNA gel blot of lines containing modified introns. Each *lane* contains 10 µg of total RNA from an independent homozygous single-copy transgenic line, in which the *TRP1:GUS* transgene contains the indicated intron. The filter was hybridized with *GUS* and loading control probes

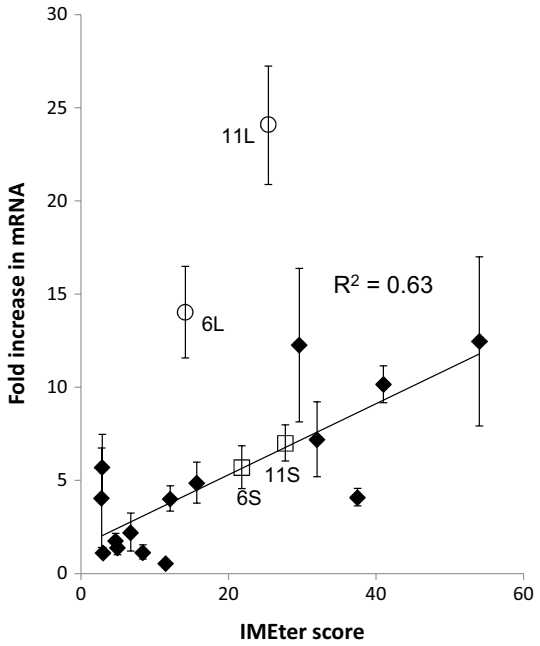


Fig. 3 IMEter scores and the stimulating ability of introns. The fold increase in *TRP1:GUS* mRNA accumulation caused by all 15 natural *Arabidopsis* introns ever tested in this reporter gene (*filled symbols*), and derivatives of the *COR15a* intron containing 6 or 11 copies of the CGATT (*squares*, marked 6S and 11S) or TTNGATYTG (*circles*, marked 6L and 11L) motifs, are plotted against their IMEter 2.0 scores. The genes from which the introns were isolated are At1g05810, At1g62170, At1g66080, At1g69930, At2g36960, At2g41100, At2g42540, At3g04150, At3g08730, At3g21970, At4g05320, At5g17990 (introns 1 and 6), At5g36230, and At5g61930. A similar plot using IMEter 1.0 scores of 12 of these introns was previously published (Rose et al. 2008). The trend line and R^2 value apply only to the 15 natural introns. *Error bars* indicate standard deviations

Splicing efficiency

An apparent decrease in expression caused by nucleotide changes within a motif might actually be caused by a reduction in intron splicing efficiency or accuracy. Any un- or mis-spliced transcripts would reduce the pool of mRNA capable of producing active GUS enzyme. Improper splicing also could affect mRNA levels if a stop codon created early in a transcript by retained intron sequences, or by a frame-shift resulting from use of an alternative splice site, activated nonsense-mediated mRNA decay. Therefore, the splicing of the introns was examined in several ways. In all RNA gel blots, the *TRP1:GUS* mRNA derived from the intron-containing constructs co-migrated with that from the intronless control (2.5 Kb), and no bands migrating in the position expected for unspliced transcripts (2.8 Kb) were observed for any construct (Fig. 2, for example). Splicing was also evaluated using two types of RT-PCR analysis. In the first, cDNA was amplified using primers that flank the site of the intron in the *TRP1:GUS* fusions. The only products that were detected co-migrated with the product derived from the intronless control, or were faint non-specific products that were also present in the intronless control lane (Fig. 4b). The absence of other products, especially those that co-migrate with products from genomic DNA amplifications that mark the size of unspliced transcripts, indicate that splicing was efficient and the use of alternative splice sites was not detected. To avoid differences in amplification efficiency related to the lengths of the products, cDNA and genomic DNA controls were amplified with two additional pairs of primers. The reverse primer downstream of the intron was the same in both pairs, and both pairs amplified products of nearly the same size (536 vs. 561 bp). In one pair, the forward primer was in the intron to detect unspliced mRNA, while in the other it was upstream of the intron to detect spliced mRNA (Fig. 4a). The amount of product derived from unspliced or aberrantly spliced mRNA was 5 % or less than that from normally spliced mRNA (Fig. 4c), indicating that all of the introns with modified motifs were efficiently and accurately spliced.

Discussion

The related motifs CGATT and TTNGATYTG, which were identified using the IMEter algorithm in two different ways, strongly increase expression from within an intron. Some evidence supporting a role for the sequence CGATT in IME was presented previously (Parra et al. 2011). Here we extended those findings by varying the number of motifs in an intron, by testing the activity of the TTNGATYTG motif for the first time, by comparing and combining the CGATT and TTNGATYTG motifs, and by evaluating the importance of specific nucleotides within the longer motif. Relatively small differences in expression were made detectable through the use of single copy transgenic lines because line to line variation in expression was almost entirely avoided

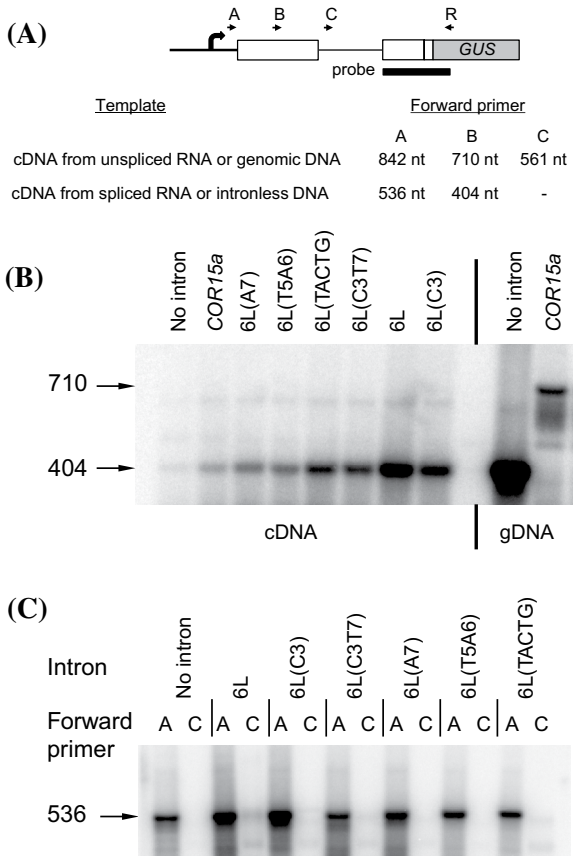


Fig. 4 RT-PCR analysis of intron splicing efficiency. **a** Detail of the *TRP1:GUS* reporter gene showing the location of primer annealing sites, probe, and the expected sizes of products using forward primers A, B, and C with reverse primer R. Primer C is in a region of the *COR15a* intron (*thin line*) unchanged by the addition of motifs. **b** Products derived from cDNA or genomic DNA (gDNA) templates from transgenic lines containing *TRP1:GUS* fusions with the indicated introns, amplified by PCR using primers B and R. The '*COR15a*' portion of the names of motif-containing introns was omitted for clarity. **c** Products derived from cDNA as in **b**, amplified by PCR using forward primer A or C as indicated and reverse primer R. Control amplifications using gDNA as template (not shown) verified that that primers C and R could generate the expected products

by eliminating differences in transgene copy number, as has been observed by others (Nagaya et al. 2005; Schubert et al. 2004).

The TTNGATYTG motif was remarkably effective at increasing expression. It was approximately four times more active than CGATT, as six copies of the long motif provided roughly twice the stimulating effect of 11 copies of the shorter sequence. Every copy of the long motif contributed an additional twofold increase in mRNA accumulation and a nearly fourfold increase in GUS enzyme activity. The upper limit of the increase in expression from adding additional copies of this motif remains to be determined.

The additive nature of the stimulation by the TTNGATYTG and CGATT motifs is consistent with the observation that the degree to which deletion derivatives of the *UBQ10* intron and *UBQ10:COR15a* hybrid introns increase *TRP1:GUS* mRNA accumulation appears linearly related to the amount of *UBQ10* intron sequence present in each intron (Supplementary Fig. 1). However, for some other introns, deleting more than 80% of the intron does not noticeably diminish its stimulating ability (Clancy and Hannah 2002), suggesting that the size of their influence on expression is determined by something other than the cumulative effects of stimulatory sequences, perhaps the act of splicing. Introns that increase expression either through redundant sequences or as a consequence of splicing could still have an effect with some parts deleted because neither mechanism relies on a single discrete sequence such as an enhancer element. Therefore, differentiating between the many possible mechanisms through which an intron stimulates expression requires thorough analysis, even if no individual intron sequences are necessary.

Even though the combined motif TTCGATTTG is a perfect match to the TTNGATYTG consensus, it was only half as effective at increasing expression as when the motif was predominantly TTAGATCTG. The difference in stimulating ability of two sequences that are both perfect fits to the motif illustrates that even though a computational approach successfully identified TTNGATYTG as being involved in IME, the observed nucleotide frequency at each position of the motif in high-scoring introns does not necessarily reflect importance for biological activity.

The sequence requirements for maximizing expression are still being explored. What we showed here is that the motif was slightly more active when the nucleotide at position 3 was A rather than C, that converting the core of the motif from GAT to GTA almost entirely eliminated its effect on expression, and the activity profile at position 7 was C>T>>A. The TT at beginning and TG at the end apparently also made minor contributions to motif function.

The T-rich nature of the TTNGATYTG motif is consistent with previous suggestions that IME is caused by T-rich sequences, and especially clusters of Ts (Clancy and Hannah 2002; Rose 2002). However, neither of the two 35 nt T-rich portions of the maize *Sh1* intron previously implicated in IME (Clancy and Hannah 2002) contains a perfect match to either the CGATT or the TTNGATYTG motif. Similarly, the T residues added to the *Arabidopsis TRP1* first intron that slightly increase its effect on mRNA accumulation (Rose 2002) do not create either motif. Clearly there are sequences in addition to CGATT and TTNGATYTG that are involved in IME.

The identification of sequences that stimulate expression will facilitate investigations into the mechanism of IME, which remains mysterious (Gallegos and Rose 2015). The isolation of a factor that binds to either motif would be a major step forwards. Even though introns exist in both the DNA being transcribed and in the RNA that is produced,

and could potentially affect expression from either location, there is some evidence that IME is caused by the intron sequences in the DNA. Stimulating intron sequences increase mRNA accumulation equally well regardless of the orientation of the fragment (Rose et al. 2011). This supports a DNA-based mechanism because inverting the fragment drastically alters the sequence of the RNA into the reverse complement, while the changes to the double stranded DNA are relatively minor, only affecting orientation and the sequence at the new junctions.

If IME is caused by a transcription factor that binds to the TTNGATYTG motif and related sequences, this would be an unusual transcription factor for a gene transcribed by RNA polymerase II because it operates from locations far downstream of the transcription start site. Downstream regulatory sequences are common in genes transcribed by RNA polymerase III but these are usually located 50–80 bp from the start of transcription (Dieci et al. 2007). The *UBQ10* intron stimulates expression when up to 550 bp from the promoter but not 1100 bp or more (Rose 2004). While enhancer elements can be located downstream of the start site in genes transcribed by RNA polymerase II, enhancers operate over much larger distances than do introns. Furthermore, enhancers can function from untranscribed locations while IME operates only when the intron is within transcribed sequences (Callis et al. 1987; Jeong et al. 2006; Snowden et al. 1996). Neither the CGATT nor the TTNGATYTG motif is currently listed in the AtcisDB, Plant CARE, or PLACE databases of known cis regulatory elements in plants (Higo et al. 1999; Lescot et al. 2002; Yilmaz et al. 2011).

While IME could be caused by a factor that binds the motif in a sequence-specific manner, there are numerous alternative ways in which introns could be acting to increase mRNA production. The motif could promote sequence non-specific interactions with proteins, such as histones or histone modifying enzymes, to affect nucleosome positioning, occupancy, or modification that creates a chromatin state favoring transcription initiation (Chodavarapu et al. 2010; Schwartz et al. 2009; Segal et al. 2006; Tilgner et al. 2009). It is even possible that IME is caused by the physical properties of the intronic DNA itself, such as the ease with which the strands are separated to form the transcription bubble, its ability to dissipate the torsional strain generated by transcription (Niu and Yang 2011), or its propensity for three-dimensional looping that facilitates re-initiation by bringing both ends of a gene into close proximity (Moabbi et al. 2012). Further analysis of the sequence requirements for motif function should help differentiate between these possibilities.

There are several practical ways in which the TTNGATYTG motif may prove to be useful. The number of introns known to boost expression is still relatively small, and many are subject to patent protection. Now, any intron can be converted into one that increases mRNA accumulation by creating the appropriate number of motifs for the desired level of expression, and introns apparently can be made that have a larger effect than do naturally occurring introns. These introns are expected to activate the expression of most genes in virtually all tissues, based on the broad stimulating ability of the *UBQ10* intron (Emami et al. 2013). However, differences between promoters in their response to introns containing this motif are possible. The computational approach that was used to identify the TTNGATYTG motif in *Arabidopsis* detected a related motif (TCGATC) in rice (Rose et al. 2008), suggesting that the TTNGATYTG motif might affect expression in both monocots and dicots. Alternatively, for any of the 34 species for which online IMEters are available (Gallegos and Rose 2015), the same approach could be used to identify IME-associated motifs in the plant of interest. The ability to easily generate a variety of new stimulating introns could be particularly advantageous in cases where a high level of expression of several genes at once is needed because this could reduce potential silencing problems that can arise from having large regions of homologous sequences in different transgenes (Eamens et al. 2008). Adapting the IMEter software to recognize specific stimulating sequences will improve our ability to recognize naturally occurring stimulating introns and to predict their effect on expression.

Ultimately, the main benefit of identifying the sequences responsible for IME and further investigations into their function will be an increased understanding of how gene expression is controlled in eukaryotes.

Materials and methods

Creating modified introns

The sequence of the *COR15a* intron was visually scanned for anagrams of either CGATT or TTNGATYTG, which were then rearranged to match the motif. In a few cases, the motifs were created from sequences that were one or two nucleotides longer than the motif, with the extra nucleotides placed at either end. The changes to test the importance of individual residues were made in all six copies of the motif in the *COR15a*6L intron. The resulting introns were synthesized by Epoch Biolabs, Sugar Land, Texas, or by Biomatik, Cambridge, Ontario, verified by sequencing, and inserted as *Pst*I restriction fragments between exons 1 and 2 of a *TRP1:GUS* reporter gene (Fig. 1d, e) as previously described (Rose et al. 2008).

Measuring gene expression

The intron-containing *TRP1:GUS* reporter genes were introduced into *Arabidopsis thaliana* ecotype Columbia by

Agrobacterium-mediated transformation using the floral dip method (Clough and Bent 1998), and transformants were identified by plating T_1 seeds on media containing kanamycin. One hundred T_2 seeds from each individual T_1 plant were tested for resistance to kanamycin, and those that gave a resistant:sensitive ratio of 3:1 indicative of a single locus of insertion were screened by genomic DNA gel blots to identify those that contained a single copy *GUS* transgene. From a total for all constructs of 361 starting T_2 lines, 271 were tested by DNA blot, of which 31 were verified to be single-copy when digested with each of three enzymes (*Pst*I, *Bam*HI, or *Bgl*II). The T_3 seeds from several T_2 plants for each line were tested for kanamycin resistance to identify those derived from homozygous T_2 individuals. Expression experiments were performed using leaf tissue of 3-week-old homozygous T_3 plants grown in Professional Growing Mix (Sun Gro Horticulture, Agawam, MA) at a density of 500 plants per 170 cm^2 jumbo square pot. All single-copy lines obtained were analyzed and given equal weight in calculations. For the two constructs where only one single-copy line was found, homozygous T_3 seeds derived from two different T_2 individuals were tested separately. Enzyme assays for GUS activity were performed using MUG as the substrate (Jefferson 1987), and the concentration of total protein in the extract was determined by dye binding assay using Pierce Coomassie Plus reagent with BSA as the standard. The RNA gel blots were performed and band intensity was quantified by PhosphoImager as described previously (Rose et al. 2008). Briefly, 10 µg of total RNA was loaded per lane in a denaturing gel, electrophoresed for 100 min at 80 volts, and transferred to Brite Star membranes, all using Northern Max kit reagents (Ambion). The filters were probed with *GUS* as well as exons 4–9 of a *TRP1* cDNA as a loading control. The *TRP1* probe hybridizes to transcripts from the endogenous *TRP1* gene but not the *TRP1:GUS* transgene, which contains *TRP1* sequences only upstream of the start of exon 3. The above-background signal in the *TRP1:GUS* band, corrected for differences in loading using the *TRP1* band, was calculated relative to the amount of *TRP1:GUS* mRNA from the intronless control. Log gene expression data was analyzed for differences between constructs, using a mixed model that adjusted for blot to blot differences, and included random effects for the line and the date of the biological replicate. Residual normality was checked using a Wilk Shapiro test and homoscedasticity using a Levene ANOVA. Post hoc comparisons among the levels of categorical predictors were based on least squares means, using a protected least significant difference.

Splicing efficiency

Total RNA was digested with RQ1 RNAse-free DNAse (Promega) and reverse-transcribed using random primers. The resulting cDNA, as well as control genomic DNA, was amplified by 25 cycles of PCR using the primers indicated in Fig. 4. Primer A is OAR212 (5′-CTCTACTGTGTTGGTTGAGCAATCG), primer B is OAR28 (5′-GAAGAAGCAACTTGACCGGAG), primer C is OAR213 (5′-CCTCATAAGTAAGGATCTTAG CAGGC), and primer R is OAR37 (5′-TAACGCGCTTT CCCACCAACG). Equal volumes of product were subjected to electrophoresis in 3% NuSieve 3:1 (Lonza) agarose gels for 2 h at 100 volts, transferred to Gene Screen Plus (Perkin Elmer) membranes using 0.4 N NaOH, and hybridized with a ^{32}P-labeled probe (see Fig. 4a). The intensity of bands in PhosphoImager scans was measured using ImageQuant software (Molecular Dynamics).

Acknowledgments We thank Dr. Neil Willits for statistical analysis, and Dr. Lesilee Rose and Jenna Gallegos for helpful comments on the manuscript. This work was supported in part by the United States Department of Agriculture, Grant Number 2006-35301-17072.

Author contributions A.R designed and carried out most of the experiments, and wrote the manuscript. A.C. and N.K. assisted in the experiments. I.K. provided bioinformatic guidance in choosing motif mutations to test.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akua T, Shaul O (2013) The *Arabidopsis thaliana MHX* gene includes an intronic element that boosts translation when localized in a 5′ UTR intron. J Exp Bot 64:4255–4270

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369–W373

Bianchi M, Crinelli R, Giacomini E, Carloni E, Magnani M (2009) A potent enhancer element in the 5′-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene. Gene 448:88–101

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

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

Chodavarapu RK et al (2010) Relationship between nucleosome positioning and DNA methylation. Nature 466:388–392

Chung BY, Simons C, Firth AE, Brown CM, Hellens RP (2006) Effect of 5′UTR introns on gene expression in *Arabidopsis thaliana*. BMC Genomics 7:120

Clancy M, Hannah LC (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 LC, Vasil IK (1994) Maize *Shrunken-1* intron and exon regions increase gene expression in maize protoplasts. Plant Sci 98:151–161

Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. Plant J 16:735–743

Deyholos MK, Sieburth LE (2000) Separable whorl-specific expression and negative regulation by enhancer elements within the *AGAMOUS* second intron. Plant Cell 12:1799–1810

Dieci G, Fiorino G, Castelnuovo M, Teichmann M, Pagano A (2007) The expanding RNA polymerase III transcriptome. Trends Genet 23:614–622

Donath M, Mendel R, Cerff R, Martin W (1995) Intron-dependent transient expression of the maize *GapA1* gene. Plant Mol Biol 28:667–676

Down TA, Hubbard TJP (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res 33:1445–1453

Duncker BP, Davies PL, Walker VK (1997) Introns boost transgene expression in *Drosophila melanogaster*. Mol Gen Genet 254:291–296

Eamens A, Wang MB, Smith NA, Waterhouse PM (2008) RNA silencing in plants: yesterday, today, and tomorrow. Plant Physiol 147:456–468

Emami S, Arumainayagam D, Korf I, Rose AB (2013) The effects of a stimulating intron on the expression of heterologous genes in *Arabidopsis thaliana*. Plant Biotechnol J 11:555–563

Gallegos JE, Rose AB (2015) The enduring mystery of intron-mediated enhancement. Plant Sci 237:8–15

Gruss P, Lai CJ, Dhar R, Khoury G (1979) Splicing as a Requirement for Biogenesis of Functional 16 S Messenger-RNA of Simian Virus-40. Proc Natl Acad Sci USA 76:4317–4321

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27:297–300

Jefferson RA (1987) Assaying chimeric genes in plants: the GUS gene fusion system. Plant Mol Biol Rep 5:387–405

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

Jeong YM, Mun JH, Lee I, Woo JC, Hong CB, Kim SG (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 ME, Nislow C, Davis RW (2006) Introns regulate RNA and protein abundance in yeast. Genetics 174:511–518

Le Hir H, Nott A, Moore MJ (2003) How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci 28:215–220

Lescot M et al (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30:325–327

Lu S, Cullen BR (2003) Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. RNA 9:618–630

Luehrsen KR, Walbot V (1994) Addition of A- and U-rich sequence increases the splicing efficiency of a deleted form of a maize intron. Plant Mol Biol 24:449–463

Matsumoto K, Wassarman KM, Wolffe AP (1998) Nuclear history of a pre-mRNA determines the translational activity of cytoplasmic mRNA. EMBO J 17:2107–2121

Moabbi AM, Agarwal N, El Kaderi B, Ansari A (2012) Role for gene looping in intron-mediated enhancement of transcription. Proc Natl Acad Sci U S A 109:8505–8510

Nagaya S, Kato K, Ninomiya Y, Horie R, Sekine M, Yoshida K, Shinmyo A (2005) Expression of randomly integrated single complete copy transgenes does not vary in *Arabidopsis thaliana*. Plant Cell Physiol 46:438–444

Niu DK, Yang YF (2011) Why eukaryotic cells use introns to enhance gene expression: splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. Biol Direct 6:24

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

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

Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL (1991) Heterologous introns can enhance expression of transgenes in mice. Proc Natl Acad Sci USA 88:478–482

Parra G, Bradnam K, Rose AB, Korf I (2011) Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Res 39:5328–5337

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

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

Rose AB, Elfersi T, Parra G, Korf I (2008) Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. Plant Cell 20:543–551

Rose AB, Emami S, Bradnam K, Korf I (2011) Evidence for a DNA-based mechanism of intron-mediated enhancement front. Plant Sci 2:98

Schubert D, Lechtenberg B, Forsbach A, Gils M, Bahadur S, Schmidt R (2004) Silencing in Arabidopsis T-DNA transformants: the predominant role of a gene-specific RNA sensing mechanism versus position effects. Plant Cell 16:2561–2572

Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon–intron structure. Nat Struct Mol Biol 16:990–995

Segal E et al (2006) A genomic code for nucleosome positioning. Nature 442:772–778

Snowden KC, Buchholz WG, Hall TC (1996) Intron position affects expression from the *tpi* promoter in rice. Plant Mol Biol 31:689–692

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R (2009) Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol 16:996–1001

Wiegand HL, Lu S, Cullen BR (2003) Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. Proc Natl Acad Sci USA 100:11327–11332

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E (2011) AGRIS: the Arabidopsis gene regulatory information server, an update. Nucleic Acids Res 39:D1118–D1122