10608–10618 *Nucleic Acids Research, 2018, Vol. 46, No. 20* *Published online 9 August 2018*
doi: 10.1093/nar/gky716

Conserved sequences in the *Drosophila mod(mdg4)* intron promote poly(A)-independent transcription termination and *trans*-splicing

Maxim Tikhonov[1], Marina Utkina[1], Oksana Maksimenko[2,*] and Pavel Georgiev[1,*]

[1]Department of the Control of Genetic Processes, Institute of Gene Biology, Russian Academy of Sciences, 34/5 Vavilov St., Moscow 119334, Russia and [2]Group of Molecular Organization of Genome, Institute of Gene Biology, Russian Academy of Sciences, 34/5 Vavilov St., Moscow 119334, Russia

Received June 18, 2018; Revised July 23, 2018; Editorial Decision July 25, 2018; Accepted July 26, 2018

ABSTRACT

Alternative splicing (AS) is a regulatory mechanism of gene expression that greatly expands the coding capacities of genomes by allowing the generation of multiple mRNAs from a single gene. In *Drosophila*, the *mod(mdg4)* locus is an extreme example of AS that produces more than 30 different mRNAs via *trans*-splicing that joins together the common exons and the 3′ variable exons generated from alternative promoters. To map the regions required for *trans*-splicing, we have developed an assay for measuring *trans*-splicing events and identified a 73-bp region in the last common intron that is critical for *trans*-splicing of three pre-mRNAs synthesized from different DNA strands. We have also found that conserved sequences in the distal part of the last common intron induce polyadenylation-independent transcription termination and are enriched by paused RNA polymerase II (RNAP II). These results suggest that all *mod(mdg4)* mRNAs are formed by joining in *trans* the 5′ splice site in the last common exon with the 3′ splice site in one of the alternative exons.

INTRODUCTION

Pre-mRNA splicing is a fundamental process in eukaryotic gene expression (1). The boundaries of exons and introns are marked by the splice donor site at the 5′-end of the intron and the splice acceptor site at its 3′-end. Splicing is catalyzed by the spliceosome, a complex that is comprised of five small nuclear ribonucleoproteins (snRNPs) and a large number of non-snRNP proteins. The most common type of splicing is known as '*cis*-splicing', which involves the ligation of exons within one pre-mRNA. *Cis*-splicing ensures the correct assembly of mRNAs, whose processing is mechanistically coupled with transcription (2,3). Eukaryotic genes usually consist of many exons, and their inclusion in mature mRNA is dependent on alternative splicing (AS). Alternative splicing is a ubiquitous regulatory mechanism of gene expression that allows generation of more than one unique mRNA species from a single gene (4,5). It has been shown that AS greatly expands the coding capacities of genomes by allowing the generation of multiple mRNAs from a limited number of genes (6).

In *Drosophila,* there are several extreme examples of AS generating multiple mRNAs from a single locus (7,8). In two cases, different mRNAs are formed by *trans*-splicing of mRNAs produced by different promoters. The first is the *lola* locus that encodes a transcription factor involved in neuron and stem cell differentiation (9). The 5′-end of *lola* transcripts contain five exons that splice into 20 variants of 3′ exons to generate 20 protein isoforms (10). The second is the *mod(mdg4)* locus that produces at least 31 variants of mRNAs with the same four 5′-terminal exons and alternative 3′-terminal exons (11,12). Unexpectedly, the 3′ alternative exons in both cases are transcribed from different promoters and are sometimes located on opposite DNA strands (10,11,13–15). These facts suggest that the high diversity of mRNAs is formed by *trans*-splicing, with the splice donor site at the 5′-end of last constitutive exon joining with one of the 3′ splice sites of alternative 3′ exons transcribed from independent promoters. Such *trans*-splicing of *mod(mdg4)* is conserved in insects, including silkworms, mosquitoes, and cotton bollworms (16–19). Using paired-end deep sequencing of mRNA in *Drosophila* interspecies hybrids, McManus *et al.* (20) confirmed that *mod(mdg4)* and *lola* mRNAs can be formed from transcripts originating from different DNA strands. They also found that at least 80 other *Drosophila* genes encode alternative mRNAs that can be generated by *trans*-splicing. Thus, such *trans*-splicing is one of the AS mechanisms that ensure generation of highly diverse mRNAs from the same locus.

[*]To whom correspondence should be addressed. Tel: +7 499 135 9906; Fax: +7 499 135 4105; Email: georgiev_p@mail.ru
Correspondence may also be addressed to Oksana Maksimenko. Email: maksog@mail.ru

The mechanisms of *trans*-splicing in higher eukaryotes are not yet well understood. Current models favor the concept of long complementary sequences that bring two separate transcripts together to promote *trans*-splicing (19,21,22). However, no obvious base-paired regions have been found between the introns predicted to be involved in *trans*-splicing in the *mod(mdg4)* locus (11,12).

To make a step toward understanding a *trans*-splicing mechanism, we used the φC31-based integration system (23) to obtain an *in vivo* model for mapping DNA regions in the last common intron that was previously shown to be critical for *trans*-splicing (12,13). As a result, we identified a 73 bp sequence that proved to be critical for *trans*-splicing and the *mod(mdg4)* gene activity *in vivo*. At the same time, deletion of highly conserved 13-bp motif from the center of the 73 bp sequence had a slight effect on *trans*-splicing and *mod(mdg4)* functions. This motif was previously shown to bind U1 small nuclear RNP through strong base-pairing with U1 snRNA (12). We also found that RNAP II paused in the region of the last common intron that could form a conserved secondary structure in RNA. These sequences define a new polyadenylation-independent transcriptional terminator that prevents transcription into the cluster of 3′ exons located downstream. Thus, our results show that all alternative *mod(mdg4)* mRNAs are formed by the *trans*-splicing mechanism that joins the 5′ and 3′ ends of exons generated from different promoters.

MATERIALS AND METHODS

RNA purification and quantitative analysis

Total RNA was isolated from 2- to 3-day-old adult males using the TRI reagent (MRC) according to the manufacturer's instructions. The nuclear and cytoplasmic RNA fractions were prepared from S2 cells grown in 100-mm dishes. The amounts of specific cDNA fragments were quantified by real-time PCR. See Supplementary Materials and Methods for details.

Cell transfection

S2 cells grown in SFX medium were cotransfected with 200 ng of pAc-Fluc reference plasmid and 800 ng of experimental plasmids using Cellfectin II (Life Technologies) as recommended by the manufacturer. After transfection, the cells were grown for 48 h and then collected.

Luciferase analysis

Fly lysate was prepared from from 2- to 3-day-old adult males and assayed using the Firefly Luciferase Assay Kit (Biotium) following the manufacturer's instructions. See Supplementary Materials and Methods for details.

RACEseq and nascent RNAseq

The nuclear modified RNA from S2 cells was prepared as described in Supplementary Materials. The amplified cDNA copies were sequenced. Libraries for MiSeq (Illumina) were prepared with Nextera XT DNA Library Prep Kit according to manufacturer's recommendations. The adapter sequence was trimmed, and reads were mapped to the *Drosophila* dm3 genome using Bowtie2. The mapped data were visualized with IGB Browser. Nascent-seq data (GSM815908, GSM815909 (24)) and 3′NT-Seq data (3′end of native transcripts sequencing) (GSM1193873 (25)) were obtained from NCBI GEO. Reads were mapped to the *mod(mdg4)* locus using Bowtie2 and visualized with IGB Browser.

ChIP analysis

Chromatin preparation and immunoprecipitation were performed as described (26), with modifications. See Supplementary Materials and Methods for details.

Transgenic lines

Drosophila strains were grown at 25°C under standard culture conditions. The transgenic Drosophila lines were generated using the φC31-mediated site-specific integration system (23) or Cas9-induced (Bloomington Stock Center: 58492) homologous recombination. See Supplementary Materials and Methods for details. Details of the cloning procedure and the sequences of primers used for construct preparation are available upon request.

RESULTS

The *mod(mdg4)* transcription is terminated in the last common intron 4

The *mod(mdg4)* locus is critical for fly viability and survival of *Drosophila* cell lines. The *mod(mdg4)* mRNAs were generated by joining four common exons with one of alternative 3′ exons, suggesting that critical sequences for *trans*-splicing should be located in the last common intron (named 'intron 4') located downstream of exon 4 (Figure 1A). To map the essential sequences, we divided the intron into nine sections of unequal length, designated *a* to *i* (Figure 1B).

Using the data available from NCBI GEO (24), we analyzed the genome-wide distribution of nascent RNA molecules in S2 cells and noticed a strong decrease in RNA accumulation downstream of exon 4 (Figure 1B). To confirm this observation and detect the potential ends of RNA molecules, we performed 3′-RACE combined with high-throughput sequencing in S2 and Kc167 cells and revealed a gradual decrease in the amount of RNA (almost to zero) in the first 500 bp of intron 4 (Figure 1B and Supplementary Figure S1A). Moreover, no poly(A) tails were detected at the 3′-ends of sequenced RNAs from intron 4, in contrast to annotated polyadenylation signals at the 3′-ends of specific isoforms (Supplementary Figure S1B, S1C).

We then used RT-qPCR to compare S2 cells and adult 2-day males (line yw^{1118}) for RNA profile in intron 4 (Figure 1C). The level of RNA was measured with 16 pairs of primers covering a 2-kb region of the intron. In S2 cells, nuclear and cytoplasmic RNA fractions were isolated separately. As expected, no transcripts related to intron 4 were detected in the cytoplasmic fraction, suggesting that the donor unspliced pre-mRNA was retained in the nucleus and not exported to the cytoplasm. The nuclear S2 RNA

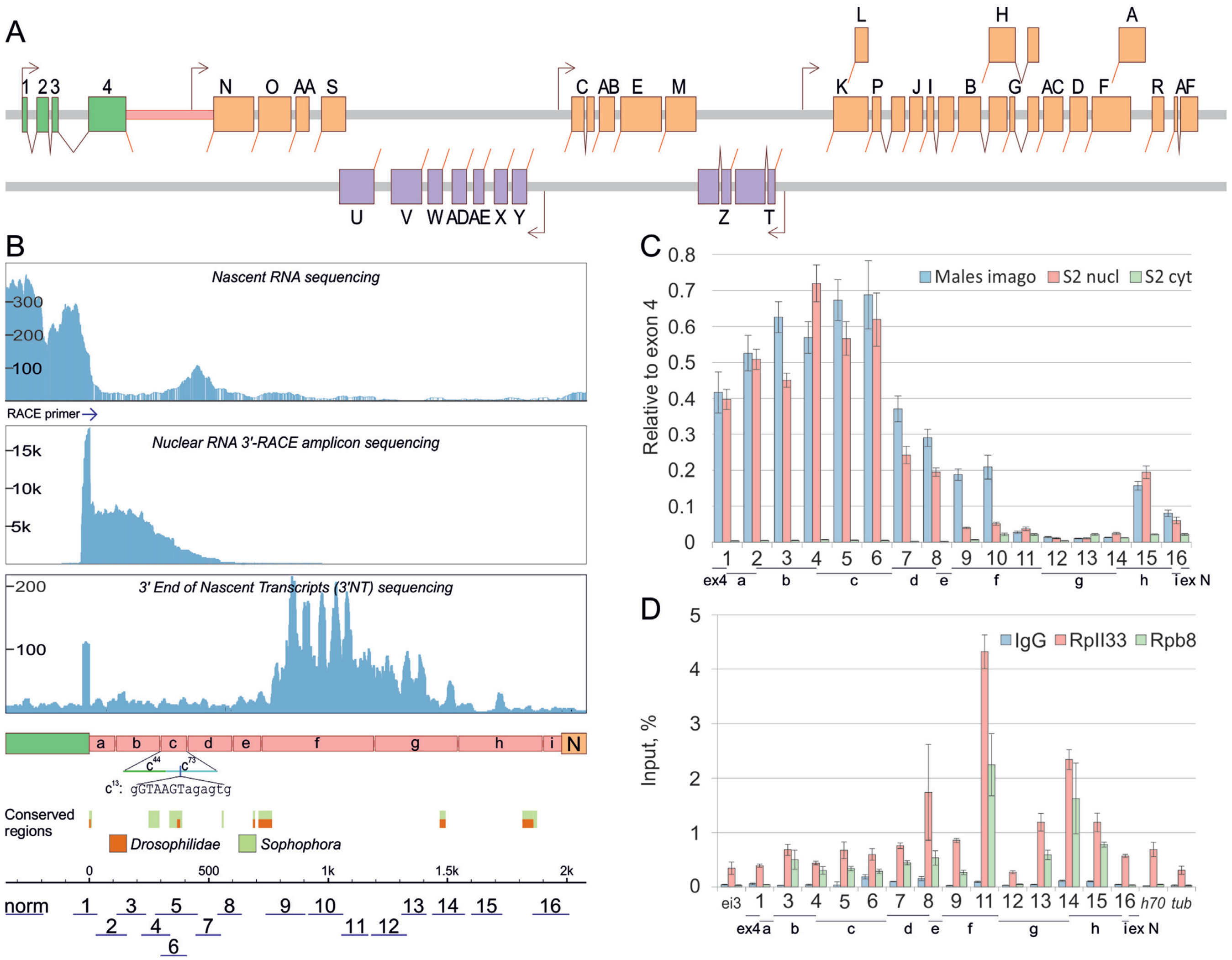

Figure 1. (**A**) Schematic organization of the *mod(mdg4)* gene locus in *Drosophila melanogaster*. The *mod(mdg4)* mRNAs are generated by joining four common exons (green boxes) with one of alternative 3′-exons placed on the same (orange boxes) or opposite (purple boxes) DNA strand. The sequence between exon 4 and the alternative exon N is defined as intron 4 (pink horizontal bar). Arrows show the TSS and directions of transcription. (**B**) Nascent RNAseq, nuclear RNA 3′-RACE amplicon sequencing, and NET-Seq profiles. Intron 4 was divided into nine regions named *a* to *i*. Region *c* was divided into 44-bp (c^{44}) and 73-bp (c^{73}) sequences. The conserved 13-bp sequence was named c^{13}. The scheme below shows conserved regions for closely related Sophophora and more common Drosophilidae groups (light green and brown boxes, respectively). The lines with numbers indicate fragments used in qPCR analysis. (**C**) RNA levels within intron 4 in adult males and in the nuclear and cytoplasmic fractions of S2 cells. Primer positions are shown in (B), at the bottom. (**D**) Enrichment of RNAP II subunits within intron 4. ChIP-qPCR analysis was carried out with the same primers as in (C) and at the following control points: exon 3–intron 3 junction (ei3), *hsp70* promoter (h70), and *tubulin-γ37C* (tub). Relative enrichment was calculated as the ratio of immunoprecipitated DNA to the input DNA. Error bars show standard deviations (n = 3).

fraction and total RNA from adult males displayed similar qPCR profiles in intron 4. We observed a gradual decrease of pre-mRNA in the *ef* region (between points 7 and 11), and no pre-mRNA was detected in the *g* region (between points 12 and 14). Taken together, these results are consistent with transcription termination in the *ef* region delimited by points 7 and 11. This region contains a 79-bp sequence highly conserved among insects (Supplementary Figure S2), which was identified previously (12).

It is also evident that transcription is initiated in the *h* region (Figure 1C). According to ModEncode data ((27); ID 5096, 5098), sequences in the *h* region are enriched with the promoter-associated H3K4Me3 histone mark (Supplementary Figure S3A). To test for promoter activity in the *h* region, we used luciferase-based assay in S2 cells (Supplementary Figure S3B). The results confirmed that the *ghi* region can drive the expression of the firefly luciferase reporter gene (*Fluc*). It appears that this promoter is responsible for the transcription of the 3′ exon encoding isoform N and, possibly, of 3′ exons located downstream that encode other isoforms (Figure 1A).

Alternative splicing and transcription termination are usually associated with pausing of the RNAP II complex (4,6). Therefore, we performed ChIP experiments with antibodies against RpII33 and Rpb8 subunits to quantify the enrichment of RNAP II at different points along the *mod(mdg4)* intron 4 in 2-day yw^{1118} males (Figure 1D). The level of RNAP II enrichment was found to be low at the control point in the exon–intron region *(ex3-int3)* of the *mod(mdg4)* locus and moderate in the *bcd* region, while an unexpectedly strong enrichment (with both antibodies) was observed in the *efg* regions, suggesting strong RNAP II pausing. Similar results were obtained by analyzing data of 3′end of native transcripts sequencing (3′NT-Seq) (25) available from NCBI GEO (GSM1193873). 3′NT-Seq detects nascent RNA transcripts, actively transcribed by RNAP II, through the capture of 3′ RNA. A high level of signal was detected at the *fg* region, linked in this case with 3′ ends formation of RNAs bound by RNAP II, which could be explained by RNAP II pausing and by active generation of RNA 3′ ends (25). Both these processes are likely to take place in the *efg* region. A high level of RNAs detected at the end of the exon 4 is due to the formation of splicing intermediates after the 5′SS cleavage. ChIP showed a high enrichment of RNAP II in the *h* region, which contains the potential promoter for the transcription of alternative downstream 3′ exons. Thus, intron 4 contains several regions in which RNAP II appears to be paused.

Mapping regions required for *trans*-splicing in intron 4

An attempt to map regions required for *trans*-splicing was made previously by using RT-PCR to test the products of *trans*-splicing between exon 4 with different lengths of intron 4 (expressed from plasmid transfected into *Drosophila* S2 cells) and the endogenous 3′ exons of *mod(mdg4)* (12). Such a model system did not take into account the role of chromatin and distance between the transcription units whose products are involved in *trans*-splicing. Therefore, we developed an alternative model system based on parts of the *mod(mdg4)* locus stably integrated into the cytogenetic region 22A (Supplementary Figure S4A).

To avoid possible RT-PCR artifacts, we developed an assay for evaluating *trans*-splicing events at the protein level (Figure 2A). One construct (named Donor) contained the endogenous region of the *mod(mdg4)* locus (including the promoter), all common exons and introns, and 3′ exon for the N isoform. The last common exon 4 was modified by insertion of an IRES sequence from the *reaper* gene (28) that could initiate translation from internal parts of the mRNA by a cap-independent mechanism. The coding region for the N isoform was separated by the *Renilla* luciferase (*Rluc*) gene sequence in order to measure, at the protein level, the amount of mRNA generated by splicing of intron 4. The second construct (named Acceptor) contained the promoter-driven 3′ exon encoding the V isoform that was separated by the firefly luciferase (*Fluc*) gene sequence. Cap-dependent translation of the *Fluc* was prevented due to premature open reading frame between the promoter and 3′ splice site. If *trans*-splicing occurred, the IRES would be fused to the *Fluc* coding region to induce translation of the firefly luciferase gene. Using the φC31-based integration system (23), both constructs with the *white* reporter were inserted into the 22A chromosome region located at approximately 10 kb from the nearest coding gene. Unfortunately, we failed to identify the *Rluc* product in our experimental system due to the problem with the too long distance between IRES and the AUG initiation codon of the *Rluc* gene.

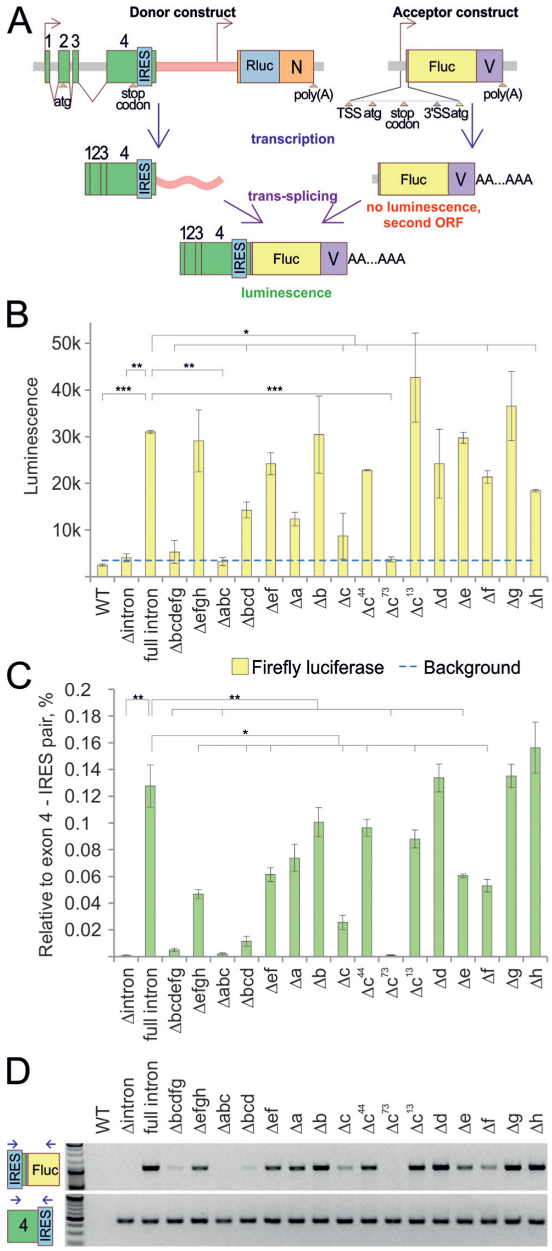

Figure 2. Mapping regions required for *trans*-splicing in intron 4. (**A**) Scheme of the model system used to map regions involved in *trans*-splicing. The Donor construct (on the left) contains the *mod(mdg4)* promoter (arrow) with the *mod(mdg4)* coding region, including four common exons (green boxes), IRES, intron 4 (pink horizontal bar), and alternative exon N (orange box) with the inserted *Rluc* gene (blue box). The Acceptor construct (on the right) contains the promoter (arrow) and alternative exon V with the *Fluc* reporter gene (yellow box). (**B**) Testing for *trans*-splicing efficiency in Donor/Acceptor heterozygotes using the *Firefly* luciferase assay. The broken blue line shows the background autoluminescence level. Each combination of *trans*-heterozygotes was analyzed in at least five replicates. (**C**) Testing for *trans*-splicing efficiency in Donor/Acceptor heterozygotes using RT-qPCR. All RT-qPCR assays were quantified against a standard sample. Levels of exon 4–IRES junction were used as reference. The histogram shows percent proportions of *trans*-spliced transcripts relative to the total amount of transcripts from the Donor construct. Error bars show standard deviations ($n = 3$). Asterisks indicate significance levels: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. (**D**) Electrophoregram of RT-PCR products from *trans*-spliced transcripts (amplified with IRES and Fluc primer pair) and from normalization target (amplified with exon 4 and IRES primer pair).

We constructed transgenic lines expressing the intact intron and derivatives with deletions of either single or several regions indicated in Figure 1B. *Trans*-splicing was tested in 2- to 3-day males heterozygous for the Donor and Acceptor constructs (Supplementary Figure S4B). A close correlation was observed between the results obtained by measuring *Fluc* luminescence (Figure 2B) and by RT-qPCR with hydrolysis probes (Figure 2C).

A single deletion of region *c* was the only one that strongly affected *trans*-splicing. Single deletions of regions *a*, *e* and *f* showed a moderate effect, with deletions of other regions (*b*, *d*, *h* or *g*) having no significant influence on *trans*-splicing (Figure 2B, C; Supplementary Table S1). Deletion of a larger fragment (*abc*) resulted in complete inactivation of *trans*-splicing, suggesting that regions *a* and *c* have partially redundant functions. Deletion of *ef* had the same effect as deletion of either *e* or *f,* which could be explained by the involvement of these two regions in the same functional module. Indeed, the most conserved structure overlaps with sequences from both regions (Supplementary Figure S2).

Region *c* contains the previously identified highly conserved 13-bp core motif (GGURAGUAGAGUG) that resembles a pseudo-5′SS sequence with the 9-base region matching to U1 snRNA (12). To test for the functional role of this motif, we made smaller deletions in the 117-bp region *c*: Δ1–44 bp (Δc^{44}), Δ45–117 bp (Δc^{73}), and Δ73–85 bp (Δc^{13}) (Supplementary Figure S2). Unexpectedly, deletion of c^{44} and c^{13} had no influence on the efficiency of *trans*-splicing, while deletion of c^{73} had an even stronger effect than deletion of the entire region *c,* resulting in almost complete inactivation of *trans*-splicing (Figure 2B–D; Supplementary Table S1).

To confirm these results, we made an Acceptor construct that contained two promoters driving divergent transcription of the 3′ exons of the T and K isoforms, with the coding regions of the T and K isoforms being substituted by *Fluc* and *mCherry* reporters (Figure 1A, Supplementary Figure S5A). The construct was integrated into the 22A chromosome region and tested for *trans*-splicing by RT-PCR in combination with the Donor constructs. The results were generally similar to those described above, but deletion of region *c* and complex deletion (*bcd*) resulted in complete inactivation of *trans*-splicing, while deletion of region *a* had only a slight effect (Supplementary Figure S5B, S5C; Table S1). Thus, the role of region c^{73} in *trans*-splicing is critical but also depends on the nature of the promoter region regulating transcription of the 3′ alternative exon.

The conserved c^{73} region is critical for *trans*-splicing in the endogenous *mod(mdg4)* locus

Our next task was to test whether the results obtained in our model system could be confirmed at the endogenous *mod(mdg4)* locus. To make deletions of the c^{73}, c^{44} and c^{13} regions we used the CRISPR/Cas9 system (29,30) (Figure 3, Supplementary Figure S6A). The sgRNAs were designed for regions *b* and *d* that, according to our test, are not essential for *trans*-splicing. The DNA template for substitution contained the region of intron 4 between the sites for the sgRNAs. The proximal site for sgRNA in the DNA template was replaced by an *Apa*I restriction site, and the *mCherry* reporter flanked by *loxP* sites was inserted into the distal site. The region of substitution was flanked by the 500-bp homology arms. We prepared DNA templates for deletions of the 13 bp core motif (Δc^{13}), the c^{44} and c^{73} regions (Δc^{44} and Δc^{73}) and for control substitution without deletion of the sequences (intact).

Seven independent events were obtained with the intact template. All of them had insertion of the *mCherry* reporter (*mod(mdg4)*^{mCh}) and the *Apa*I site, suggesting that both sgRNAs functioned effectively (Supplementary Figure S6B). Homozygous *mod(mdg4)*^{mCh} flies and *mod(mdg4)*$^{\Delta mCh}$ flies generated by *lox*P-mediated deletion of the *mCherry* reporter were fully viable. Since inactivation of the *mod(mdg4)* gene leads to embryonic lethality (31), we concluded that insertion of the reporter gene in region *d* did not interfere with the *mod(mdg4)* function.

Three independent events with the simultaneous insertion of the *mCherry* reporter and *Apa*I were obtained with the Δc^{13} template. In two cases, the 13-bp deletion was also detected and confirmed by sequencing (Supplementary Figure S6B). Flies homozygous for the *mod(mdg4)*$^{\Delta c13}$ allele were fully viable. Finally, eight and seven independent events were obtained for the deletion of the regions c^{73} and c^{44}*,* respectively. Unexpectedly, integration of the *mCherry* reporter in seven cases for Δc^{73} and five cases for Δc^{44} was not accompanied by that of the *Apa*I site. Moreover, none of integration events had deletion of either of the regions*,* including the three remaining events that contained the *Apa*I site. The absence of deletions in the integration events could be interpreted as evidence for a dominant lethal effect of the deletions or, alternatively, for a high level of recombination events during integration of the constructs.

Taking into account that deletions might induce a dominant lethal effect, we used an alternative approach with the CRISPR/Cas9 (Figure 3B, Supplementary Figure S7A). The same sgRNAs and the 500-bp homology arms were used, but the DNA template for substitution contained the *wt* region of intron 4 between the sites for the sgRNAs (*c*) and the *mCherry* reporter flanked by *lox*P and followed by the same *c* region with one of the tested deletions (*c**). It was expected that the presence of the *wt* copy of the *c* region would support *trans*-splicing at the *mod(mdg4)* locus after successful substitutions. The resulting events were tested by PCR and sequencing to select lines with correct substitutions (Supplementary Figure S7B). Next, the *wt c* region and the *mCherry* reporter were deleted by *lox*P-mediated recombination. The resulting transgenic lines contained *lox*P in the region *b* and one of the *c* variants. Using this approach, we obtained two *intact*, one Δc^{44}, one Δc^{73}, and one Δc^{13} *mod(mdg4)* fly lines. All new *mod(mdg4)* alleles were viable and could be maintained in the homozygous state. Only line homozygous for the *mod(mdg4)*$^{\Delta c73}$ allele showed reduced viability and fertility.

The next question was, whether the deletions in the intron 4 could affect *trans*-splicing of the *mod(mdg4)* isoforms. Using RT-qPCR, we measured splicing efficiency for five 3′ exons located on either the same or opposite DNA strand for lines carrying the following *mod(mdg4)* alleles: *wt*, *intact*, Δc^{44}, Δc^{73}, and Δc^{13} (Figure 3C, D). *Trans*-splicing in the *mod(mdg4)*^{intact} and *mod(mdg4)*$^{\Delta c44}$ alleles was at the wild-type level, being slightly changed in two indepen-

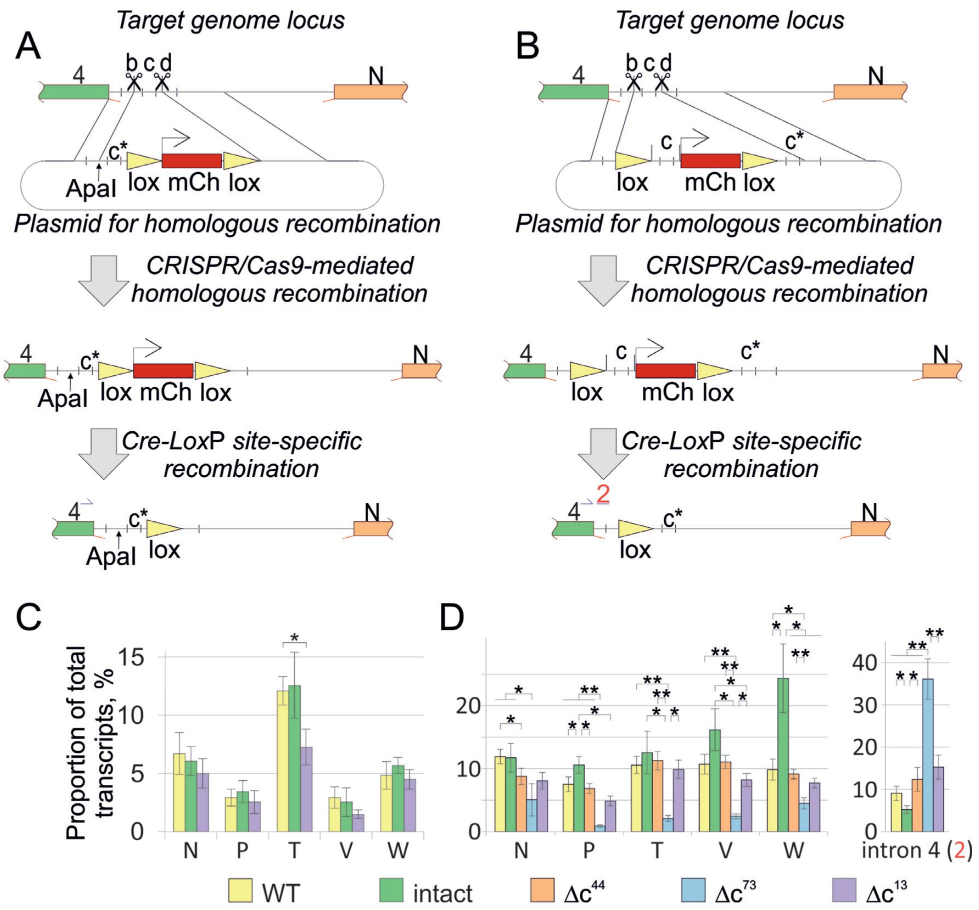

Figure 3. Functional role of the c^{44}, c^{73}, c^{13} regions from intron 4 in *trans*-splicing of the *mod(mdg4)* locus *in vivo*. (**A**) Scheme of CRISPR/Cas9 technique used to delete designed sequences from intron 4 of the *mod(mdg4)* locus. The Cas9 cleavage site in region *b* was altered into *Apa*I site. The *mCherry* reporter flanked by *loxP* sites was added into *d* cut site. (**B**) Scheme of the second variant of intron 4 editing. The template for substitution contained the unaltered region of intron 4 and the *mCherry* reporter between *loxP* sites and followed by the same region with one of the tested deletions (c*). (C, D) Comparing *trans*-splicing efficiency in $mod(mdg4)^{+}$ (WT), $mod(mdg4)^{intact}$, $mod(mdg4)^{\Delta c13}$ lines (**C**) and $mod(mdg4)^{+}$ (WT), $mod(mdg4)^{intact}$, $mod(mdg4)^{\Delta c44}$, $mod(mdg4)^{\Delta c73}$, $mod(mdg4)^{\Delta c13}$ lines (**D**) based on the results of RT-qPCR for five *mod(mdg4)* isoforms: N, P, T, V, and W (see Figure 1A) and for a point from intron 4. Lines with altered genome were obtained in accordance with the schemes shown in (A) and (B). Analysis of RNA isolated from 2-day-old males was performed in three independent replicates. The histogram shows proportions (%) of particular transcripts relative to the total amount of *mod(mdg4)* transcripts. Error bars show standard deviations ($n = 3$). Asterisks indicate significance levels: $*P < 0.05$, $**P < 0.01$.

dently obtained $mod(mdg4)^{\Delta c13}$ alleles. A significant reduction of *trans*-splicing was observed only in the case of isoform T. Thus, the core 13-bp conserved sequence proved to have only a limited impact on *trans*-splicing. At the same time, RT-qPCR showed that formation of all tested isoforms was obviously decreased in $mod(mdg4)^{\Delta c73}$flies (Figure 3D). Accumulation of unspliced RNA carrying the intron 4 sequences was also observed only in $mod(mdg4)^{\Delta c73}$ line (Figure 3D, right panel). These results are in complete agreement with the results obtained in the model transgenic system, demonstrating that the c^{73} region contains sequences that are critical for *trans*-splicing.

The conserved sequence in regions *e, f* and *g* are required for transcription termination in intron 4

According to our results (Figure 1), transcription is likely to be terminated in the *ef* region of intron 4. This region contains two cores that are invariant in *Drosophila* species and was predicted to form a stable secondary structure containing four stems and two loops (12). To directly map the regions that are essential for transcription termination, we compared RNA levels at different sites of intron 4 in males homozygous for the Donor constructs by means of RT-qPCR with either random primers or a specific primer located in the *Rluc* coding region. In both cases, deletion of *e, f* or *g* alone resulted in a severalfold increase in the amount of transcripts in downstream regions, while that of any other region had no significant effect (Figure 4A, Supplementary Figures S8, S9). Deletion of larger fragments (*ef* or *efg*) had the same effect as deletion of each individual region, suggesting that regions *e*, *f,* and *g* contain sequences that are important for the same activity.

To corroborate these results, we examined transcription in the *abcdefg* region using a heterologous transgenic model system in which this region was inserted between the *Rluc* gene driven by the *Act5C* promoter and the *Fluc* gene followed by the poly(A) signal (Figure 4B, Supplementary Figures S10, S11). The construct was inserted into the 38D chromosome region. The level of RNA in different regions

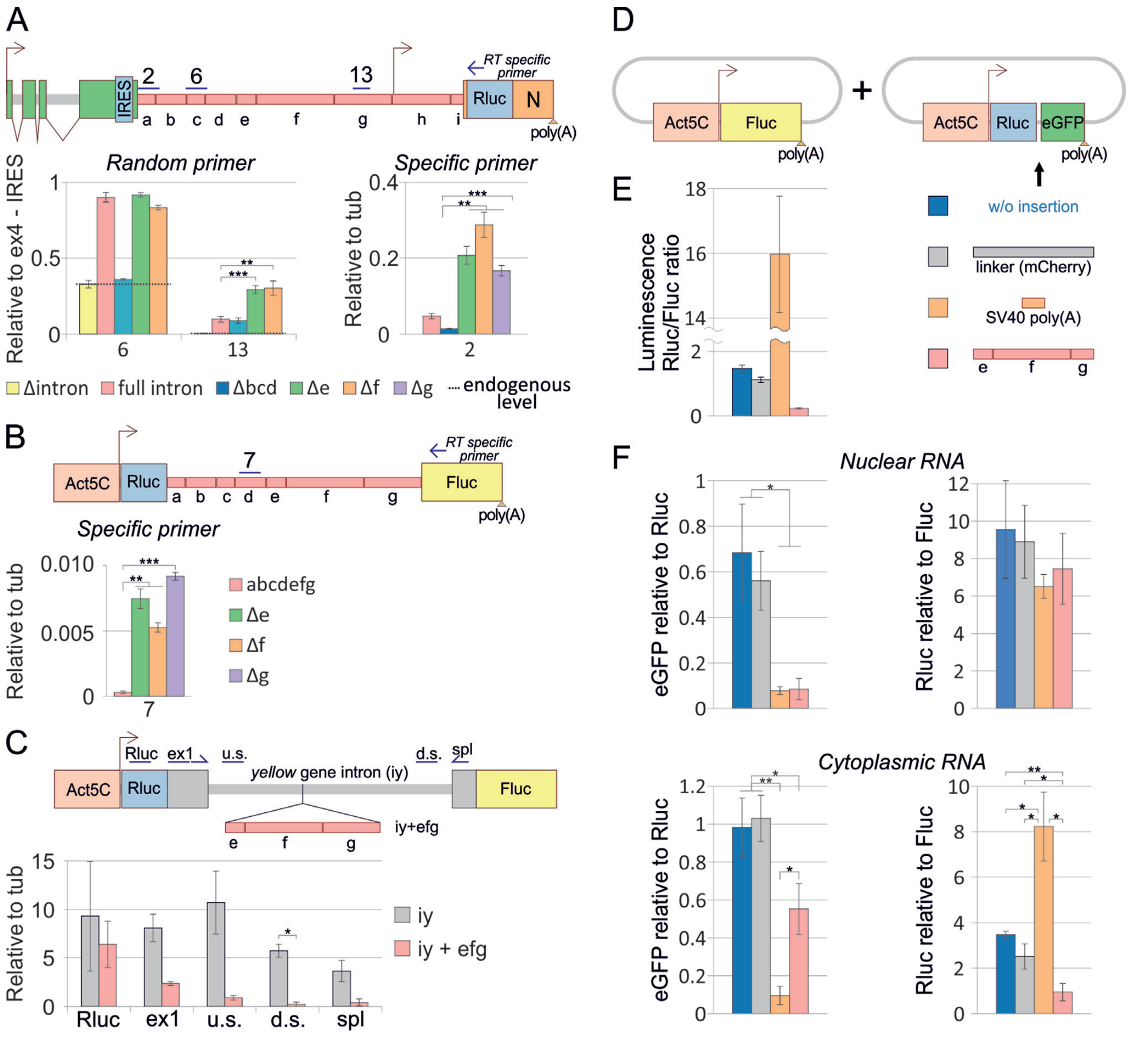

Figure 4. Transcription termination in the intron 4 of the *mod(mdg4)* locus. (**A**) Mapping the regions involved in transcription termination in intron 4 using the transgenic line homozygous for the Donor construct (see Figure 2). The histogram shows the accumulated and throughout-read transcripts within intron 4 in intact and deletion variants of the Donor construct. Dotted line shows the endogenous level of RNA corresponding to the test region. (**B**) Mapping the regions involved in transcription termination in intron 4 using the transgenic line homozygous for the construct carrying the *abcdefg* region of intron 4 inserted between the *Rluc* and *Fluc* genes under control of the *Act5C* promoter. All qRT-PCR assays were quantified against a standard sample. Levels of *ex4–IRES* junction and *tubulin-γ37C* were taken as reference. Random primers or specific primers (blue arrows) were used for RT. (**C**) Testing the ability of the *efg* region to terminate transcription in the heterologous *yellow* intron. The 2.7-kb intron with 200-bp flanking exon fragments from the *yellow* gene was inserted between the *Rluc* and *Fluc* genes driven by the *Act5C* promoter. Transgenic flies with the $y^{-}ac^{1}sc^{1}$ background were used in order to eliminate endogenous *yellow* gene influence. The histogram shows RNA levels for control points of *Rluc*, exon1 (*ex1*), and unspliced (*u.s.; d.s.*) and spliced (*spl*) products from constructs with the intact *yellow* gene intron and this intron with *efg* insertion. (**D**) Model system for testing the ability of the *efg* region to terminate transcription in the transiently transfected *S2* cells. The *efg* region, linker with similar length and polyadenylation signal from SV40 were inserted between the *Rluc* and *eGFP* genes driven by the *Act5C* promoter. *pAct5C-Fluc* reporter was cotransfected as reference. (**E**) The histogram shows Rluc-to-Fluc luminescence ratio. (**F**) The histograms show *eGFP*/*Rluc* and *Rluc*/*Fluc* ratios of RNA levels. Nuclear and cytoplasmic fractions of RNA were analyzed separately. Error bars show standard deviations ($n = 3$). Asterisks indicate significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

of the construct was tested by RT-qPCR. As previously, deletion of *e, f* or *g* alone proved to strongly increase transcription in downstream regions.

We then tested the ability of the *efg* region to affect transcription in the heterologous intron from the *yellow* gene using the construct where this region was inserted in the *yellow* intron placed between Act5C-Rluc and Fluc-poly(A) (Figure 4C). The construct was integrated into the 38D chromosome region. As a result, we observed that insertion of the *efg* region induced a strong decrease in the steady-state RNA level, with the amount of *Rluc* being only half that in the transgenic line carrying the construct without this region. The amount of RNA corresponding to the *yellow* intron and spliced RNA was decreased almost tenfold. Thus, the *efg* region inserted into the intron of unrelated gene reduces RNA level, which may be explained by induction of transcription termination and decrease in RNA stability.

To further evaluate the functional role of the *efg* region in transcription termination, we prepared a set of constructs containing the *Rluc* and *eGFP* genes under control of the *Act5C* promoter. Different sequences such as the 717-bp negative control DNA (from *mCherry*), 243-bp

SV40 poly(A) signal, and 765-bp *efg* region were inserted between the *Rluc* and *eGFP* genes. Expression of *Rluc* and *eGFP* in S2 cells was compared relative to that of *Fluc* expressed from the reference Act5C-*Fluc* construct (Figure 4D). Since insertion of the 717-bp linker between the *Rluc* and *eGFP* genes did not significantly change the expression patterns of the reporter genes, we regarded both these variants as negative controls (Figure 4E, F). The expression of reporter genes was examined separately in the nuclear and cytoplasmic RNA fractions. In the negative controls, the *eGFP* and *Rluc* genes were transcribed as a single RNA, and the ratio of *eGFP* to *Rluc* was therefore close to 1 (Figure 4F, left panels). This ratio in the nuclear RNA fraction was slightly lower because of the presence of incompletely transcribed nascent RNAs (Figure 4F, top panels). Insertion of either *efg* or *SV40* resulted in a tenfold reduction of the *eGFP/Rluc* pre-mRNA ratio in the nuclear fraction, which in both cases could be explained by transcription termination in the region between the reporter genes. The *Rluc/Fluc* ratio in the nuclear RNA fraction (Figure 4F, right panel) was almost equal for all constructs, suggesting similar levels of expression. In contrast, different RNA patterns for the *efg* and *SV40* constructs were observed in the cytoplasmic fraction (Figure 4F, bottom panels). As expected, the *Rluc/Fluc* ratio in the presence of *SV40* was four times higher than in the negative controls (Figure 4F, bottom panel), and this correlated with more than tenfold difference in the Rluc/Fluc luminescence ratio, which reflected the amount of Rluc protein (Figure 4E). These results are in agreement with effective transcription termination by *SV40* poly(A) signal and formation of the mature *Rluc* mRNA. The opposite effect was observed in the case of *efg*: compared to the negative controls, the *eGFP/Rluc* ratio was two times lower, while the *Rluc/Fluc* ratio was reduced threefold (Figure 4F, bottom panel). Thus, the cytoplasmic fraction was enriched with *Rluc*-*efg*-*eGFP* transcripts, while the nuclear fraction was enriched with *Rluc* transcripts. These results correlate with data on the Rluc/Fluc luminescence ratio for *efg,* which was decreased by factors of 6 and 60 relative to the negative controls and SV40 poly(A), respectively (Figure 4E). Taken together, these results confirm that the *efg* region accounts for the polyadenylation-independent mechanism of RNAP II transcription termination and that the RNA terminated at the *efg* region cannot be exported from the nucleus to the cytoplasm.

DISCUSSION

The *Drosophila mod(mdg4)* locus provides an extreme example of AS generating multiple mRNAs (at least 31 variants) from a single locus with the same four 5′-terminal exons and alternative 3′-terminal exons transcribed from different promoters and sometimes located on the opposite DNA strands. Thus, *trans*-splicing should account for equally efficient generation of mRNA variants with different 3′ exons. We developed an *in vivo* model for mapping DNA regions in the last common intron of *mod(mdg4)* that was previously shown to be critical for *trans*-splicing (12,13). As a result, we identified a 73-bp sequence that proved to be critical for *trans*-splicing and the *mod(mdg4)* gene activity *in vivo*. At the same time, deletion of highly conserved 13-bp motif from the center of the 73-bp sequence had no significant effect on *trans*-splicing and *mod(mdg4)* functions. This 13-bp motif was previously shown to bind U1 small nuclear RNP through strong base-pairing with U1 snRNA (12).

The results presented above are partly different from those obtained by Gao *et al.* (12), which may be explained by differences in model systems used for identification of sequences that are essential for *trans*-splicing (Figure 5A). Gao *et al.* transfected S2 cells with plasmid-borne constructs containing the *mod(mdg4)* exon 4 with different lengths of intron 4 under control of the metallothionein promoter and tested for *trans*-splicing products between the exogenous exon 4 and the endogenous 3′ exons of *mod(mdg4)* (Figure 5B). The documented *trans*-splicing events suggest that the exogenous pre-mRNAs containing only the exon 4 and part of intron 4 form stable intermediates that efficiently search for the pre-mRNAs with endogenous 3′ exons of *mod(mdg4)* located in the distant nuclear compartment. Moreover, these exogenous pre-mRNAs can efficiently compete for the endogenous 3′-exon pre-mRNAs with the endogenous 5′-exon pre-mRNAs expressed by the *mod(mdg4)* locus. In our model system, the native *mod(mdg4)* promoters generated pre-mRNAs for common 5′ exons or 3′ alternative exons (Figure 5C), and tests were performed for *trans*-splicing products between pre-mRNAs transcribed from genes located at the same site on homologous chromosomes, which recapitulated the situation in the endogenous *mod(mdg4)* locus.

On the other hand, our results are consistent with (12) in showing that the 300–400 bp region (TSA/region *c*) is most critical for *trans*-splicing (Figure 5A, Supplementary Figure S12); while deletion of the 600–800 bp region (TSB/region *ef*) has only a slight effect on this process (Supplementary Figure S13). This was confirmed in experiments with deletions generated in the *mod(mdg4)* locus using the CRISPR/Cas9 system (Figure 5A, Supplementary Figure S12). The main contradiction with (12) concerns the role of the 13-bp core TSA region that is bound by U1 snRNA. Using the S2 model system, Gao *et al.* (12) found the core TSA to be essential for *trans*-splicing, but its deletion in our model system and in the endogenous *mod(mdg4)* locus had no significant effect on *trans*-splicing. It is likely that these contradictory results are explained by high stringency of the S2 model system (Figure 5B): the exogenous pre-mRNA should search in the nucleus for the endogenous 3′-exon pre-mRNAs and compete for *trans*-splicing with the endogenous pre-mRNA encoded by the common part of the *mod(mdg4)* locus. In this case, the binding of U1 snRNA to the TSA core may improve the stability of pre-mRNA and its ability to compete with the endogenous *mod(mdg4)* pre-mRNA for *trans*-splicing.

Alternative splicing and transcription termination are usually associated with pausing of the RNAP II complex. We also found that the intron 4 contains several regions in which RNAP II appears to be paused and that the *ef* region of the last common intron contains the most conserved sequence of intron 4 with predictable secondary structure. It is possible that such an RNA structure, in combination with unknown proteins bound to region *g,* induces RNAPII pausing and conformational changes that induce

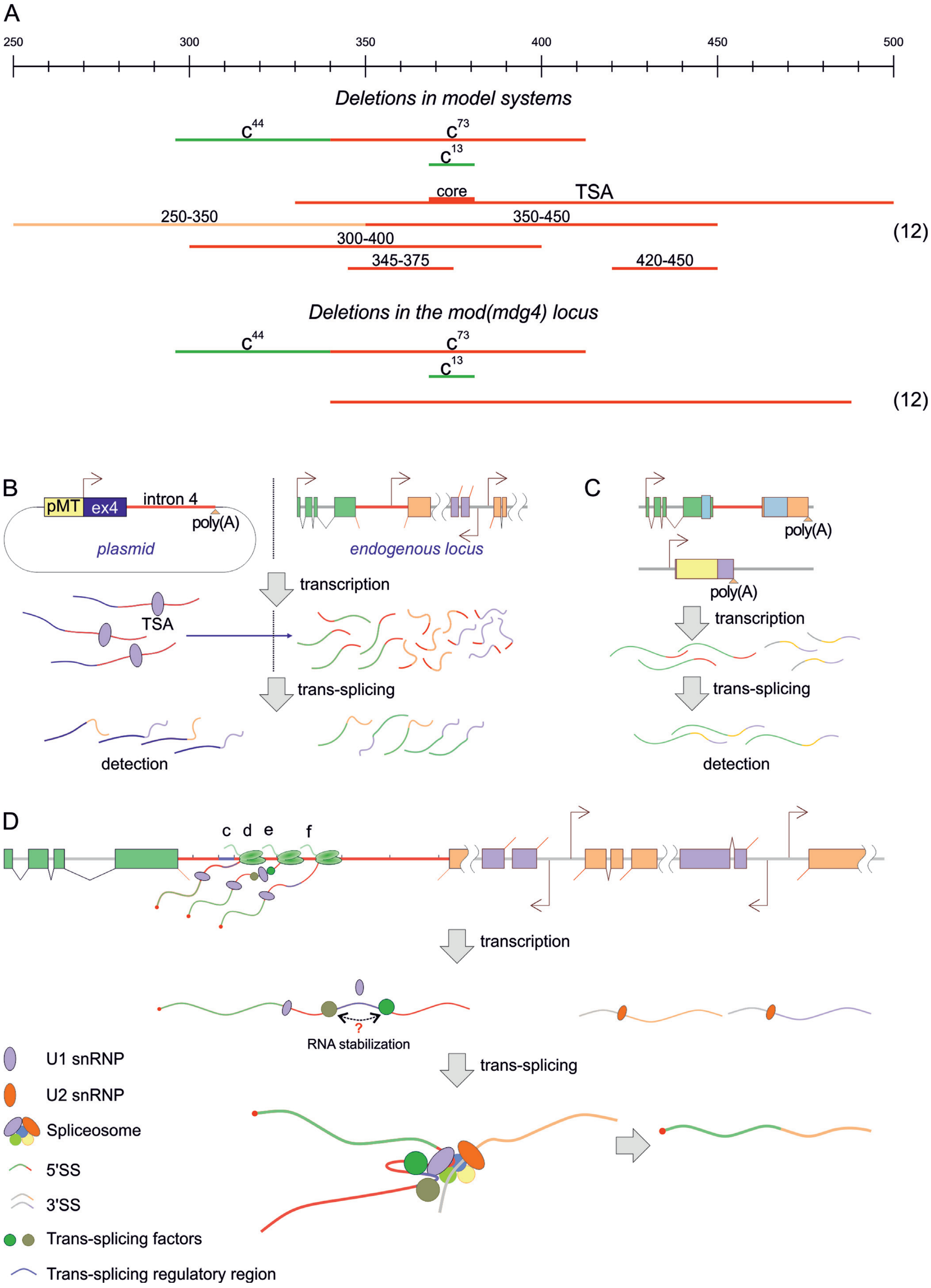

Figure 5. (**A**) Comparison of the *mod(mdg4)* intron 4 regions essential for *trans*-splicing as identified in this study (on the top) and by Gao *et al.* (12). The most critical regions are shown in red; the regions whose deletion has a moderate or no apparent effect are shown in orange and green, respectively. **(B, C)** Model systems used for mapping the regions required for trans-splicing (**B**) in Gao *et al.* (12) and (**C**) in this study. (**D**) A model of *trans*-splicing.

transcription termination without polyadenylation and prevents transcription into the cluster of 3′ exons located downstream. Similar mechanisms have been described for transcriptional termination of long noncoding microRNA genes (32,33). Such a mechanism of RNAP II termination is likely to be essential in view of previous data that U1 snRNP protects pre-mRNAs from cleavage and polyadenylation at conventional AATAAA signals in gene introns (33,34).

Transcription termination in intron 4 suggests that all *mod(mdg4)* mRNAs are formed by joining in *trans* of the 5′ splice site after the last common exon with the 3′ splice site before one of the alternative exons. As shown previously, different *mod(mdg4)* isoforms are expressed in a manner that depends on the tissue and the stage of development [(11); FlyAtlas (www.flyatlas.org)]. The *mod(mdg4)* pre-mRNAs start at multiple promoters that transcribe sequences from both DNA strands within the locus (11). Thus, the efficiency of splicing for 3′ exons should be the same, regardless of distance and orientation relative to the last common exon 4.

According to Gao *et al.* (12), all sequences required for *trans*-splicing are located in the intron 4 (Figure 5B). The c^{73}/TSA region contains the U1 snRNA binding site and additional key sites for unknown proteins involved in stimulation of *trans*-splicing. It is important to note that the deletion of c^{73} did not completely suppress *trans*-splicing in the *mod(mdg4)* locus, suggesting that additional, partially redundant sites for unknown proteins in the intron 4 are involved in this process. One of *trans*-splicing models (21) suggests that the spliceosome can mediate *trans*-splicing between the unsaturated donor and acceptor splice sites from different pre-mRNAs. In accordance with this model, regulatory proteins specifically bind to the intron 4 and, by unknown mechanism, form pre-mRNA with the 5′ unsaturated donor splice site, while transcription of alternative 3′ exons produces pre-mRNAs with unsaturated acceptor splice sites. Binding of the U1 snRNA can stabilize the unsaturated 5′ pre-mRNA complex that retains its *trans*-splicing ability while searching for a 3′ pre-mRNA with unsaturated acceptor site (Figure 5D).

The results of this study and the proposed model will hopefully contribute to our understanding of *trans*-splicing. However, a number of questions on the details of this phenomenon are still open, and further research is required to gain a deeper insight into the mechanism that generates, with equal efficiency, multiple alternative mRNAs from the same locus.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to N.A. Gorgolyuk for his help in preparing the manuscript and to F. Hasanov and A. Parshikov for fly injection. Experiments were performed using the equipment of the IGB RAS facilities supported by the Ministry of Science and Education of the Russian Federation.

FUNDING

Russian Science Foundation (RSF) [14-24-00166 to P.G.]. Funding for open access charge: RSF [14-24-00166].
Conflict of interest statement. None declared.

REFERENCES

1. Herzel,L., Ottoz,D.S.M., Alpert,T. and Neugebauer,K.M. (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.*, **18**, 637–650.
2. Oesterreich,F.C., Herzel,L., Straube,K., Hujer,K., Howard,J. and Neugebauer,K.M. (2016) Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell*, **165**, 372–381.
3. Bentley,D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.
4. Fiszbein,A. and Kornblihtt,A.R. (2017) Alternative splicing switches: Important players in cell differentiation. *BioEssays*, **39**, doi:10.1002/bies.201600157.
5. Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, **18**, 437–451.
6. Gallego-Paez,L.M., Bordone,M.C., Leote,A.C., Saraiva-Agostinho,N., Ascensao-Ferreira,M. and Barbosa-Morais,N.L. (2017) Alternative splicing: The pledge, the turn, and the prestige: The key role of alternative splicing in human biological systems. *Hum. Genet.*, **136**, 1015–1042.
7. Hemani,Y. and Soller,M. (2012) Mechanisms of Drosophila Dscam mutually exclusive splicing regulation. *Biochem. Soc. Trans.*, **40**, 804–809.
8. Horiuchi,T. and Aigaki,T. (2006) Alternative trans-splicing: A novel mode of pre-mRNA processing. *Biol. Cell* , **98**, 135–140.
9. Southall,T.D., Davidson,C.M., Miller,C., Carr,A. and Brand,A.H. (2014) Dedifferentiation of neurons precedes tumor formation in Lola mutants. *Dev. Cell*, **28**, 685–696.
10. Horiuchi,T., Giniger,E. and Aigaki,T. (2003) Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, lola. *Genes Dev.*, **17**, 2496–2501.
11. Yu,S., Waldholm,J., Bohm,S. and Visa,N. (2014) Brahma regulates a specific trans-splicing event at the mod(mdg4) locus of Drosophila melanogaster. *RNA Biol.*, **11**, 134–145.
12. Gao,J.L., Fan,Y.J., Wang,X.Y., Zhang,Y., Pu,J., Li,L., Shao,W., Zhan,S., Hao,J. and Xu,Y.Z. (2015) A conserved intronic U1 snRNP-binding sequence promotes trans-splicing in Drosophila. *Genes Dev.*, **29**, 760–771.
13. Dorn,R., Reuter,G. and Loewendorf,A. (2001) Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 9724–9729.
14. Labrador,M. and Corces,V.G. (2003) Extensive exon reshuffling over evolutionary time coupled to trans-splicing in Drosophila. *Genome Res.*, **13**, 2220–2228.
15. Mongelard,F., Labrador,M., Baxter,E.M., Gerasimova,T.I. and Corces,V.G. (2002) Trans-splicing as a novel mechanism to explain interallelic complementation in Drosophila. *Genetics*, **160**, 1481–1487.
16. Shao,W., Zhao,Q.Y., Wang,X.Y., Xu,X.Y., Tang,Q., Li,M., Li,X. and Xu,Y.Z. (2012) Alternative splicing and trans-splicing events revealed by analysis of the Bombyx mori transcriptome. *RNA*, **18**, 1395–1407.
17. Krauss,V. and Dorn,R. (2004) Evolution of the trans-splicing Drosophila locus mod(mdg4) in several species of Diptera and Lepidoptera. *Gene*, **331**, 165–176.
18. Gabler,M., Volkmar,M., Weinlich,S., Herbst,A., Dobberthien,P., Sklarss,S., Fanti,L., Pimpinelli,S., Kress,H., Reuter,G. *et al.* (2005) Trans-splicing of the mod(mdg4) complex locus is conserved between the distantly related species Drosophila melanogaster and D. virilis. *Genetics*, **169**, 723–736.
19. Kong,Y., Zhou,H., Yu,Y., Chen,L., Hao,P. and Li,X. (2015) The evolutionary landscape of intergenic trans-splicing events in insects. *Nat. Commun.*, **6**, 8734.
20. McManus,C.J., Duff,M.O., Eipper-Mains,J. and Graveley,B.R. (2010) Global analysis of trans-splicing in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12975–12979.

21. Lei,Q., Li,C., Zuo,Z., Huang,C., Cheng,H. and Zhou,R. (2016) Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol.*, **8**, 562–577.
22. Berger,A., Maire,S., Gaillard,M.C., Sahel,J.A., Hantraye,P. and Bemelmans,A.P. (2016) mRNA trans-splicing in gene therapy for genetic diseases. *Wiley Interdiscip. Rev. RNA*, **7**, 487–498.
23. Bischof,J., Maeda,R.K., Hediger,M., Karch,F. and Basler,K. (2007) An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 3312–3317.
24. Khodor,Y.L., Rodriguez,J., Abruzzi,K.C., Tang,C.H., Marr,M.T. 2nd and Rosbash,M. (2011) Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev.*, **25**, 2502–2512.
25. Weber,C.M., Ramachandran,S. and Henikoff,S. (2014) Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell*, **53**, 819–830.
26. Maksimenko,O., Bartkuhn,M., Stakhov,V., Herold,M., Zolotarev,N., Jox,T., Buxa,M.K., Kirsch,R., Bonchuk,A., Fedotova,A. *et al.* (2015) Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. *Genome Res.*, **25**, 89–99.
27. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
28. Hernandez,G., Vazquez-Pianzola,P., Sierra,J.M. and Rivera-Pomar,R. (2004) Internal ribosome entry site drives cap-independent translation of reaper and heat shock protein 70 mRNAs in Drosophila embryos. *RNA*, **10**, 1783–1797.
29. Ren,X., Sun,J., Housden,B.E., Hu,Y., Roesel,C., Lin,S., Liu,L.P., Yang,Z., Mao,D., Sun,L. *et al.* (2013) Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19012–19017.
30. Zhang,X., Koolhaas,W.H. and Schnorrer,F. (2014) A versatile two-step CRISPR- and RMCE-based strategy for efficient genome engineering in Drosophila. *G3 (Bethesda)*, **4**, 2409–2418.
31. Buchner,K., Roth,P., Schotta,G., Krauss,V., Saumweber,H., Reuter,G. and Dorn,R. (2000) Genetic and molecular complexity of the position effect variegation modifier mod(mdg4) in Drosophila. *Genetics*, **155**, 141–157.
32. Dhir,A., Dhir,S., Proudfoot,N.J. and Jopling,C.L. (2015) Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat. Struct. Mol. Biol.*, **22**, 319–327.
33. Proudfoot,N.J. (2016) Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, **352**, aad9926.
34. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.