A new RNA-DNA interaction required for integration of group II intron retrotransposons into DNA targets

Dario Monachello, Marc Lauraine, Sandra Gillot, François Michel, Maria Costa

A new RNA-DNA interaction required for integration of group II intron retrotransposons into DNA targets

Dario Monachello[1], Marc Lauraine[1], Sandra Gillot[1], François Michel[1] and Maria Costa[1,*]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

* To whom correspondence should be addressed. Tel: +33169823215; Email: maria.costa@i2bc.paris-saclay.fr

Present Address: Dario Monachello, Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France, and, Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France. Marc Lauraine, Chronic Inflammation & Immune System, UMR 1173, Inserm, UVSQ/Université Paris Saclay, Laboratoire d'Excellence Inflamex, 2 ave de la Source de la Bièvre,78180 Montigny-le-Bretonneux, France. François Michel, Institut de Systématique, Évolution, Biodiversité, ISYEB – UMR 7205 – CNRS MNHN UPMC EPHE, Muséum National d'Histoire Naturelle, Sorbonne Universités, Paris, France.

ABSTRACT

Mobile group II introns are site-specific retrotransposable elements abundant in bacterial and organellar genomes. They are composed of a large and highly structured ribozyme and an intron-encoded reverse transcriptase that binds tightly to its intron to yield a ribonucleoprotein (RNP) particle. During the first stage of the retrotransposition pathway the intron RNA catalyses its own insertion directly into the DNA target site. Recognition of the proper target rests primarily on multiple base-pairing interactions between the intron RNA and the target DNA, while the protein makes contacts with only a few target positions by yet-unidentified mechanisms. Using a combination of comparative sequence analyses and *in vivo* mobility assays we demonstrate the existence of a new base-pairing interaction named EBS2a-IBS2a between the intron RNA and its DNA target site. This pairing adopts a Watson-Crick geometry and is essential for intron mobility, most probably by driving unwinding of the DNA duplex. Importantly, formation of EBS2a-IBS2a also requires the reverse transcriptase enzyme which stabilizes the pairing in a non-sequence-specific manner. In addition to bringing to light a new structural device that allows subgroup IIB1 and IIB2 introns to invade their targets with high efficiency and specificity our work has important implications for the biotechnological applications of group II introns in bacterial gene targeting.

INTRODUCTION

Group II introns are large self-splicing ribozymes and mobile genetic elements of bacterial origin that colonize genomes through an original, site-specific, retrotransposition pathway (1). They are widespread in bacteria and in organellar (mitochondrial and chloroplastic) genomes of plants, algae and fungi (2). Group II introns promote their own excision (self-splicing) from a precursor RNA by two consecutive transesterification reactions that result in the ligation of the flaking exons and the excision of the intron as a lariat RNA. The chemical reversibility of the transesterification reactions allows group II introns to 'reverse splice' back into the ligated-exons molecule to reconstitute the initial precursor RNA.

Although all group II introns share an evolutionarily conserved catalytic core and their overall secondary structure is always organized into six domains (I to VI; Figure 1A), important structural variations of the secondary structure allowed to define three major intron subgroups: IIA, IIB and IIC, that can be further subdivided into 10 additional subtypes (IIA1, IIA2, IIB1, IIB2, etc.) (3–5). Interestingly, introns belonging to the three major subgroups also display variations in the recognition modes of their 5' and 3' flanking exons. Thus, subgroup IIA and IIB introns use the 'EBS1' and 'EBS2' (Exon Binding Sites 1 and 2) segments to anchor the 5'-exon for splicing. These EBS1 and EBS2 sites -- which are stretches of 5-6 nucleotides lying far apart from one another in secondary structure subdomain ID -- make Watson-Crick interactions with complementary 'IBS1' and 'IBS2' (Intron Binding Sites 1 and 2) sequences at the 3' end of the 5'-exon. However, whereas in IIA introns the

EBS2 site spans the terminal loop of a hairpin structure, in subgroup IIB introns the EBS2 segment is embedded in a multi-branched internal loop (the 'EBS2 loop') as shown in figure 1A. On the other hand, IIC introns, which are smaller and typically located downstream of transcriptional terminator hairpins (6), lack the EBS2 site (and surrounding structural motifs altogether) and only retain the EBS1-IBS1 pairing (4, 7). In contrast to the 5'-exon, recognition of the 3'-exon relies essentially on a single base pairing interaction. In subgroup IIA introns, the first position of the 3'-exon (called δ') makes a Watson-Crick pairing with the nucleotide (δ) immediately upstream of the EBS1 segment (in a few cases the δ-δ' interaction can span the first two or three positions of the 3'-exon). In subgroup IIB and IIC introns, on the other hand, the first position of the 3'-exon is called IBS3 and base pairs with a nucleotide (EBS3) lying in a conserved internal loop in subdomain ID as shown in figure 1A.

In addition to their ribozyme core, mobile group II introns carry an open reading frame encoding for a multifunctional reverse transcriptase (RT) enzyme. Phylogenetic analyses of the reverse transcriptase ORFs allowed to define nine major lineages: mitochondrial-like (ML), chloroplast-like 1 and 2 (CL1 and CL2) and bacterial A-F (5, 8). Importantly, each lineage is associated with a specific RNA structural subgroup indicating an ancient association between the intron RNA and its protein, followed by coevolution of the structural features of the ribozyme with those of the encoded RT (4, 9).

The multifunctional reverse transcriptase (RT) enzyme also possesses a 'maturase' and, optionally, an endonuclease activity (8, 10). The maturase activity of the RT facilitates splicing by stabilizing the catalytically active ribozyme core (11, 12) and assisting domain VI conformational changes undergone by the intron RNA (13). After splicing completion, the excised intron lariat remains bound to its RT forming the ribonucleoprotein (RNP) particle that operates intron mobility through retrotransposition, (also referred to as retrohoming; reviewed by (14)). Retrotransposition is a highly efficient and site-specific process triggered by reverse splicing of the intron lariat directly into the top (sense) strand of the DNA target duplex. Then, the endonuclease (En) domain of the RT cleaves the bottom strand of the target generating a DNA primer that is used by the enzyme to reverse transcribe the inserted intron sequence, a mechanism known as *target DNA-primed reverse transcription* (TPRT; (15)). The very high specificity with which group II RNPs recognize their DNA targets stems essentially from the multiple base-pairing interactions (~ 12-14 bp) that the intron RNA establishes with the top strand of the target DNA (16, 17). Because these RNA-DNA pairings involve the same intron motifs (EBS and, in subgroup IIA introns, the δ position) that are put into use for binding the flanking 5' and 3' exons during splicing, group II introns invade DNA targets having the same sequence as the ligated exons with high fidelity. Importantly, this target recognition strategy allowed to turn group II introns into 'Targetrons' which are gene targeting vectors currently used for genome engineering of a wide variety of bacteria (18–20).

In addition to the RNA-DNA pairings, the protein component of the RNP also contributes to DNA target binding by recognizing a small number of specific bases flanking the EBS-IBS/δ-δ' pairings. Previous genetic and biochemical studies have suggested that protein interactions with positions lying on the 5'-distal section of the target site (upstream of IBS2) are required for unwinding the target DNA

duplex, which then enables formation of the EBS-IBS/δ-δ' pairings and intron reverse splicing (16, 17, 21, 22). On the other hand, protein interactions with nucleotides downstream of the IBS3 site, on the 3'-half of the target site, are necessary for bottom-strand cleavage by the En domain but not for the initial recognition of the DNA target site or reverse splicing (16, 17, 21, 22). Detailed investigations of the DNA binding site of the *Lactococcus lactis* Ll.LtrB group IIA intron, have identified positions T-23, G-21 and A-20 (top strand) of the target site as being directly recognized by the intron RT (17, 22). Importantly, in addition to these protein-DNA target contacts, the EBS2-IBS2 pairing between the intron RNA and the DNA target is also needed for invasion of double-stranded DNA targets. Accordingly, ectopic retrotransposition of the Ll.LtrB intron, which occurs into single-stranded DNA, was found to have relaxed target specificity as it does not require the nucleotides recognized by the protein nor the EBS2-IBS2 pairing (23, 24).

Interestingly, comparison of those DNA binding sites for subgroup IIA and IIB introns that have been defined biochemically so far (16, 21, 22, 25–27) reveal that the identity and the specific location of the nucleotides recognized by the RT vary markedly from one intron system to the other. For example, in contrast to the Ll.LtrB intron, *in vivo* mobility experiments carried out with the *Escherichia coli* EcI5 intron, a highly mobile subgroup IIB1 intron, identified positions C-18, C-17, A-15 and A-14 (top strand) on the distal 5' region of the DNA target site, as the critical bases recognized by the intron RT ((26); Figure 1B).

Although the protein-DNA target interactions are thought to initiate unwinding of the target site, their small number makes it unlikely that they are sufficient to operate DNA unwinding before at least partial establishment of the EBS-IBS/δ-δ' pairings. An alternative scenario would be that binding of the protein and pairing of the intron RNA to the DNA target occur concomitantly, with the two processes acting cooperatively to drive opening of the DNA double-helix. Interestingly, for several of the DNA target sites investigated, some of the critical bases recognized by the RT were found to lie immediately 5' to the IBS2 sequence. This is the case for the EcI5 intron for example, as all critical positions in the 5'-distal section of the intron's DNA target site are located immediately upstream of the IBS2 segment ((26); Figure 1B). This situation raises the possibility that the IBS2 site, conjointly with some of the critical distal positions on its side, could form a more extended and/or complex DNA binding site than previously thought, with the protein and the intron RNA involved in recognizing common positions.

This situation prompted us to further analyse the regions encompassing the EBS2 and IBS2 sites in a large dataset of bacterial and organellar group II introns. Here we report that these phylogenetic analyses in combination with *in vivo* mobility experiments allowed us to demonstrate, in subgroup IIB1 and IIB2 introns carrying an RT ORF, a new Watson-Crick pairing, that we designate EBS2a-IBS2a, between the intron RNA and its DNA target site. This new RNA-DNA interaction involves the nucleotide immediately upstream of the IBS2 segment in the DNA target site and an intron nucleotide lying only two positions upstream of the EBS2 sequence. Interestingly, the EBS2a-IBS2a pairing is shown to be essential for intron mobility *in vivo* but not for intron splicing. Recent structural work fully supports our findings and reveals how the RT contributes to directly stabilize the geometry of the

EBS2a-IBS2a pairing without imposing the identity of the bases. The implications of this work for future biotechnological applications of group II introns as gene targeting vectors ('Targetrons') are discussed.

MATERIALS AND METHODS

Comparative sequence analyses

Sequences of protein-encoding subgroup IIB1 and IIB2 introns were collected and arranged into the following four sets: subgroup IIB1 bacterial, mitochondrial and chloroplast introns and subgroup IIB2 cyanobacterial introns (for definition of structural subgroups IIB1 and IIB2, see (3, 28). For bacterial subgroup IIB1 introns we used the same sequences and names as the 2017 version of the Database for Bacterial Group II introns (28), except that Fa.pr.I1 was discarded, since it is identical to Eu.si.I2, and three archeal sequences were included. Sequences of mitochondrial- and chloroplast-encoded introns were collected by first using tBLASTn (29) to search the NCBI nucleotide database for sequences potentially encoding proteins similar to those associated with previously known organellar subgroup IIB1 introns. For each candidate hit, it was then manually checked that surrounding sequence stretches could be arranged into typical subgroup IIB1 secondary structure models that would include potential intron-exon junctions and EBS and IBS sites. For cyanobacterial subgroup IIB2 introns we did not use the Candales et al. (2012) database, which includes a number of introns with poor EBS2-IBS2 pairings, but generated instead our own dataset in the same way as for organellar introns.

Predicted protein sequences were separately aligned for each intron set and intron sequences with the same insertion site as a previously documented intron and a highly similar (greater than 0.9) translated ORF sequence were discarded. A phylogenetic tree was generated from alignable sections (377 codons) of 66 bacterial and 3 archaeal subgroup IIB1 ORF sequences with the help of program PhyML 3.1 (30) using the NNI tree-search strategy, the LG model of amino acid substitutions, four substitution rate categories and the approximate likelihood ratio test to estimate branch support.

Starting with appropriate alignments of ORF-containing introns, the EBS2a-IBS2a covariation is readily spotted by using the 'Mutual Information' function of the BioEdit package (31) and examining the resulting matrix. In subgroup IIB1 and IIB2 introns a stretch of three nucleotides normally separates EBS2 from the 5' branch of the ID(iv) helix. The EBS2a site is located in the middle of this segment and is flanked by an A on its 3' side, except in some mitochondrial introns (Supplementary Table S1). When fully base-paired, the EBS2-IBS2 duplex comprises five base pairs and the IBS2a site lies immediately 5' to the IBS2 sequence, usually at position -14 of the 5'-exon. In some sequences, however, both IBS2 and IBS2a are shifted towards or away from the intron-exon junction by one nucleotide. This is the case in a majority of chloroplast IIB1 introns (Supplementary Table S2),

in some cyanobacterial IIB2 introns (Supplementary Table S3) and in a few bacterial IIB1 introns: the IBS2a site is at 5'-exon position -13 in the case of introns Pol.sp.I1, I2 and I3 and at position -15 for intron Ch.lu.I1. All our alignments were generated with software BioEdit 7.2.5 (31) and are available upon request.

Plasmid constructs

The wild-type EcI5 donor and recipient plasmids were constructed according to Zhuang et al. (2009) in order to reconstitute an identical two-plasmid system for the *in vivo* mobility assays.
The final wild-type EcI5 donor plasmid was constructed sequentially (see below) using the pACD4K-C-loxP vector (TargeTron vector purchased from Sigma) as the plasmid backbone. Because the pACD4K-C-loxP vector is provided linearized at the HindIII and BsrGI sites, it was first circularized as follows: primers Fw_pACD4K_Cir and Rev_pACD4K_Cir (oligonucleotide sequences in Supplementary Table S4) were annealed and the resulting double-stranded product digested with HindIII and BsrGI before ligation with the linearized vector to generate the circular vector, 'pACD4K-C-loxP_Circ'.
In order to clone the EcI5 intron deleted of most of its ORF in domain IV and flanked by its exons, a PCR fragment (A) was first generated as follows: a DNA prep containing the *E. coli* virulence plasmid pO157 which harbours the EcI5 intron was used as a template for two distinct amplification reactions; one PCR used primers Ecl5-5' exon_New and Ecl5 P1_New and another PCR was performed with primers Ecl5-3' exon_New and Ecl5 P3_New. Then, the two synthesized PCR products were gel-purified, mixed and re-amplified with the external primers Ecl5-5' exon_New and Ecl5-3' exon_New. The resulting PCR fragment (A) was digested by HindIII and XhoI and ligated with vector pACD4K-C-loxP_Circ, previously linearized with the same restriction enzymes. In the resulting plasmid construct, 'pACD4K_EcI5-IntronΔORF', all the sequences encoding for the *Lactococcus lactis* LI.LtrB intron and its LtrA ORF have been replaced by those corresponding to intron EcI5 flanked by 30 and 95 nucleotides of its natural 5' and 3' exons, respectively. In addition, in this construct a 1551-nt long internal section of the reverse transcriptase ORF was deleted and replaced by a BamHI site, leaving only 131 and 43 nucleotides of the initial and the terminal sections of the ORF, respectively.
The complete EcI5 RT ORF with a 5'-adjacent Shine-Dalgarno (corresponding to phage T7 S10 Shine-Dalgarno sequence) was then cloned into plasmid pACD4K_EcI5-IntronΔORF downstream of the 3'-exon sequence. This cloning step was performed as follows: a PCR product was generated with primers Ecl5-5' ORF_New and Ecl5-3' ORF_New using as a template a DNA prep of the *E. coli* virulence plasmid pO157 encoding for the EcI5 intron. The resulting PCR product was digested with XmaI and XhoI and ligated with XmaI and XhoI-linearized plasmid pACD4K_EcI5-IntronΔORF in order to generate construct 'pACD4K_EcI5-IntronΔORF+3'ORF'. Finally, using this latter construct, a T7 promoter sequence immediately flanked by two BamHI sites was introduced into intron domain IV as follows: a PCR product was generated with primers EcI5_T7Fw and EcI5_T7Rev using as a template

the pACD4K_EcI5-IntronΔORF+3'ORF plasmid. The resulting PCR fragment was digested by BsaI and XmaI restriction enzymes (cleavage by BsaI generates a BamHI-digested extremity on the PCR product) and then ligated with BamHI and XmaI-digested pACD4K_EcI5-IntronΔORF+3'ORF plasmid to obtain construct 'pACD4K_EcI5-IntronΔORF+3'ORF+T7'. This plasmid construct corresponds to the final EcI5 wild-type intron 'donor plasmid' used for mobility assays.

The recipient plasmid containing the wild-type target site for the EcI5 intron was constructed sequentially (see below) using the pBR322 vector as the plasmid backbone. First, a transcriptional terminator (TΦ) for phage T7 RNA polymerase was inserted into pBR322, downstream of the tetracycline resistance gene (*tet*^{R}). This cloning step, which gave rise to construct 'pBR322_Tphi' was done as follows: a PCR fragment was generated with primers Tphi_Fw and Tphi_Rev using as DNA template the pET24b vector, after which this PCR product was digested with BsaI to generate 5' end overhangs suitable for ligation of the fragment with the pBR322 vector linearized at its unique StyI site. Then, a 629-bp DNA fragment was synthesized (Eurogentec; complete sequence in Supplementary Figure S1) encoding both the EcI5 target site (the 5' and 3' halves of the target site are 36- and 20-nt long, respectively) flanked by *E. coli* transcriptional terminators T1 and T2, and the beginning of the *tet*^{R} gene with an adjacent Shine-Dalgarno but without any upstream promoter. This 629-bp DNA fragment was cut with BstXI and NheI restriction enzymes and then ligated with plasmid pBR322_Tphi digested with AatII and NheI. This cloning generated construct 'pBR322_EcI5-TS_1XT2'. Finally, two additional copies of the transcriptional terminator T2 were inserted into this latter construct immediately downstream of the first T2 copy. Cloning was performed as follows: using plasmid pBR322_EcI5-TS_1XT2' as the PCR template, two PCR products were generated; PCR (A) was obtained with primers 2xT2_Fw and 2xT2_Rev and PCR (B) with primers 3xT2_Fw and 3xT2_Rev; PCR (A) was digested with XbaI and Acc65I and PCR (B) was cut by Acc65I and HindIII enzymes; these two PCR digests were then ligated together with plasmid pBR322_EcI5-TS_1XT2 linearized at the XbaI and HindIII sites. The resulting 'pBR322_EcI5-TS_3XT2' construct corresponds to the wild-type 'recipient plasmid' used for mobility assays.

All the mutant versions of the wild-type donor and recipient plasmids tested in this work were constructed by cloning PCR products obtained with oligonucleotides carrying the appropriated mutations.

All the constructs were verified by sequencing the entire plasmid section derived from PCR products.

Intron mobility experiments *in vivo*

E. coli HMS174(DE3) strain (Novagen) was used for all intron-mobility experiments. This strain possesses the *recA1* mutation in a K-12 background and is suitable for the over-expression of desired target genes based on the phage T7 expression system. When appropriate, antibiotics were added to LB liquid medium or LB-agar at the following final concentrations: ampicillin (Amp), 100 μg/mL; chloramphenicol (Cam), 25 μg/mL; tetracycline (Tet), 25 μg/mL.

For selection experiments aiming to test the EBS2a-IBS2a interaction, HMS174(DE3) competent cells were co-transformed by heat-shock (in 21-μL reactions, according to the manufacturer's instructions) with 40 ng of each specific donor plasmid and 40 ng of an equimolar mix containing all four recipient plasmids (aliquots of the same recipient plasmid-mixture were used for co-transformations with each one of the donor plasmids tested). After heat-shock, 80 μL of SOC media was added to the transformation reactions and the cells were incubated at 37°C for 1 hour to recover by growth. A small portion of this culture (10 μL) was serially diluted and plated onto LB-agar with Amp and Cam antibiotics in order to determine the efficiency of co-transformation for each experiment. The remainder of the culture (90 μL) was added to 25 mL of LB medium containing ampicillin, chloramphenicol and 20 mM glucose (final concentration) followed by incubation at 37°C overnight. Supplementing the media with glucose significantly reduces basal transcription in this genetic system and therefore, expression of the intron in the absence of IPTG induction. After overnight growth, the glucose was eliminated by washing the cells with 10 mL of fresh LB medium. Washed cells were resuspended in 10 mL of LB medium and a 1.8 mL portion was used for DNA plasmid purification and sequencing (described below) while a 50 μL portion of the washed cultures was inoculated into 10 mL of LB medium containing ampicillin and chloramphenicol and incubated at 37°C until early log phase was reached (OD at 590 nm = 0.2 - 0.4). At this point, a 250 μL portion of the early log phase-cultures were inoculated into 5 mL of fresh LB medium containing 100 μM IPTG (final concentration) and induced at 37°C for one hour (longer induction phases resulted in very high levels of cell death, probably because it leads to excessive expression of the *tet*^{R} gene which is known to be toxic (32)). After induction, the cultures were chilled on ice and then washed with 5 mL of ice-cold LB medium to eliminate IPTG (all centrifugations were performed at 4°C). Washed cells were resuspended in 5 mL of fresh LB medium, serially diluted and plated onto LB-agar+Amp+Tet. In order to approximatively estimate the level of intron retrotransposition in these experiments, cell plating was also performed onto LB-agar+Amp+Cam or LB-agar+Amp, for comparison with LB-agar+Amp+Tet. Each selection experiment was performed independently at least three times. Tet^{R} colonies were submitted to PCR in order to specifically analyse recipient plasmids containing an integrated intron. Colony PCR was performed with primers EcI5-13 and EcI5-14 (Supplementary Table S4) which hybridize, respectively, to the recipient plasmid backbone (upstream of the intron target site) and to intron-domain IV sequence. The resulting PCR products were purified and sequenced with primer EcI5-13 which allows to analyse the 5'-integration junction (comprising the IBS2a position) and the integrated intron-sequence up to domain IV.

To confirm that all four recipient plasmids (having distinct bases at the IBS2a position) were equally represented in the population of co-transformed cells in each selection experiment, the 1.8 mL portion of the overnight cultures mentioned above was used for plasmid purification and specific sequencing of the entire recipient plasmid-population with primers EcI5-13 and EcI5-15 (Supplementary Table S4; EcI5-15 is a reverse primer that hybridizes at the beginning of the *tet*^{R} gene). These primers allow to read the intron target site region of the recipient plasmid population from both strands (data not shown).

Mobility assays involving matched EBS2a / IBS2a combinations used 40 ng of donor and 40 ng of the corresponding recipient plasmid for co-transformation and, for the rest, were performed under identical conditions to the selection experiments. After the IPTG-induction step, cells were washed as previously described and then resuspended in 5 mL of fresh LB medium, serially diluted and plated onto different media: LB-agar+Amp+Tet, LB-agar+Amp+Cam and, optionally, LB-agar+Amp. Mobility efficiencies were determined as the ratio of (Amp^R+Tet^R)/(Amp^R+Cam^R) colonies. Identical results were obtained using the ratio of (Amp^R+Tet^R)/Amp^R colonies.

Quantification of *in vivo* intron splicing by primer extension

In order to determine if the EBS2a-IBS2a interaction has an impact on splicing efficiency, primer extension reactions were used to quantify levels of intron splicing *in vivo*. First, *E. coli* HMS174(DE3) competent cells were transformed with 40 ng of each one of the donor plasmids (in the absence of recipient plasmid) encoding for intron precursors carrying either matched or mismatched EBS2a / IBS2a combinations. Transformed cells were grown and induced under the same conditions as those used for mobility assays (described above) except that the IPTG induction step was carried out for 3.5 hours in order to allow the accumulation of intron-derived RNAs in the cell (over-expression of the intron in the absence of its recipient plasmid does not affect cell viability). Then, a 50 μL portion of each induced culture was used for extraction of total cellular RNA with the Direct-zol RNA Miniprep kit (Zymo Research). Approximately 100 ng of each total RNA extraction (or 30 ng for intron lariat and intron precursor RNA controls) were analysed by primer extension with SuperScript IV reverse transcriptase according to the manufacturer's instructions (Invitrogen) using a ^{32}P 5'-radiolabelled DNA primer. This primer (5'- TCCGGTTTCATTGTCCTGACAGC) hybridizes to a segment lying in subdomain IC of the EcI5 intron. After reverse transcription, the RNA template was degraded by adding to the 20 μL-reactions 2 μL of 3M NaOH and incubating the tubes at 85°C for 10 min followed by 40 min at 50°C. Then, the pH was brought to neutrality by adding 2 μL of 3M HCl, and the samples were ethanol-precipitated at -20°C overnight. After centrifugation, pellets were washed with 80% ethanol, dried and resuspended in 15 μL of formamide loading buffer. From these, aliquot portions were denatured at 85°C for 15 min and immediately loaded onto denaturing 5% polyacrylamide - 8 M urea gels. The gels were fixed, dried and exposed to PhosphorImager screens. After exposition, the screens were scanned using a Typhoon PhosphorImager and quantified using ImageQuant-TL software (GE Healthcare). For each experiment, intron splicing efficiency was calculated from the intensities of lariat (L) and precursor (P) bands as follows: (L / L+P) X 100.

RESULTS

Phylogenetic analyses predict a new base-pairing interaction between the intron and its flanking 5'-exon in subgroup IIB1 and IIB2 introns

We started by performing comparative sequence analysis on natural group II introns in order to search for covariation events that could be indicative of novel, previously unidentified, RNA-RNA interactions between the intron and its 5'-exon. The dataset we used for these phylogenetic analyses contained the organellar intron sequences listed in Supplementary Tables S1 and S2 and the full-length intron sequences found in the public database for bacterial group II introns (28). Note that for cyanobacterial subgroup IIB2 introns we generated a specific dataset, different from that of Candales et al. (2012) (Materials and Methods and Supplementary Table S3). Due to the high degree of structural variation of the peripheral domains of group II introns, comparative sequence analysis of these RNAs is most informative when performed at the level of each structural subgroup. In addition, it can also be potentially informative to further sub-divide subgroup-specific alignments taking into account the presence or the absence of an intron-contained RT ORF in order to try and identify RNA statistical constraints that would specifically result from the interaction of the intron ribozyme with its RT protein. Therefore, using these two criteria we manually generated and optimized specific alignments with software BioEdit 7.2.5 ((31); Materials and Methods). Careful inspection of the resulting alignments allowed us to identify two positions whose base identities vary in a highly concerted way (*i.e.*, co-vary), specifically in RT-encoding introns belonging to subgroups IIB1 and IIB2 (Figure 2). With the exception of IIC introns, IIB1 and IIB2 are the most abundant subgroups in bacterial genomes (33). The IIB2 representatives are specifically found in cyanobacteria and in chloroplastic genomes and differ from IIB1 introns by a number of base substitutions at positions lying in the central wheel and in domains V and VI and also by the presence of extra sequences, most often two stem-loops, on the 5'-side of the internal loop at the base of domain I (3, 28). Regarding the co-varying sites, and as shown in the secondary structure model of the *E. coli* EcI5 intron (Figure 1A), one of the sites is located in the 5'-exon, immediately upstream of the IBS2 sequence, at position -14, whereas the other site lies only two nucleotides upstream of the EBS2 sequence, in the multi-branched internal 'EBS2 loop' of subdomain ID. In keeping with their proximity to the established EBS2-IBS2 pairing we named these sites EBS2a (in the intron) and IBS2a (in the 5'-exon/DNA target site). Mutual information, a parameter that measures the *extent* to which the base identities at two positions are correlated with each other (34, 35), is high in all four cases (legend of Figure 2) lending strong support to this novel EBS2a-IBS2a interaction. Furthermore, the observed patterns of base distribution are consistent with the two sites interacting through Watson-Crick pairing (Figure 2). The pattern is particularly clear for bacterial IIB1 and IIB2 introns, since all four Watson-Crick combinations are represented. Interestingly, the covariation pattern observed in organellar IIB1 introns deviates somewhat from that in bacteria since there is a prevalence of the EBS2a(A)-IBS2a(U) combination over the other ones. In the case of mitochondrial IIB1 introns, however, this pattern may

result from a statistical bias, as the 27 introns analysed occupy only 10 distinct genomic locations. It should also be noted that in most chloroplast IIB1 introns, the covariant IBS2a nucleotide lies at position -13 of the 5'-exon instead of -14. Altogether, these differences suggest that the structural constraints acting on the EBS2a-IBS2a interaction might vary from bacterial to organellar IIB1 introns. Finally, it is important to mention that similar phylogenetic analyses conducted specifically on the many ORF-less IIB1 and IIB2 introns present in organellar genomes failed to provide statistical evidence for the EBS2a-IBS2a pairing (data not shown).

The EBS2a-IBS2a interaction does not seem to be required for recognition of an RNA 5'-exon by the intron ribozyme

We first sought to demonstrate experimentally the EBS2a-IBS2a interaction by testing its impact on the ability of the intron ribozyme to bind its 5'-exon RNA *in vitro*. For this purpose we used the mitochondrial Pl.LSU/2 from the brown alga *Pylaiella littoralis* since it is an RT-encoding IIB1 intron that potentially forms a EBS2a(G)-IBS2a(C) pairing with its flanking 5'-exon. Moreover, the Pl.LSU/2 intron is an ideal system for biochemical approaches due to its ability to fold and splice readily and properly (36, 37). The *in vitro* experimental system that was put into use involves purified intron lariat and an RNA oligo mimicking the 5'-exon (carrying the IBS2a, IBS2 and IBS1 sequences). When the two RNAs are combined, the 5'-exon RNA binds to the lariat intron through the EBS-IBS pairings and catalyses intron debranching, which is the reverse of the 2'-5' branch-forming reaction that occurs during the first step of splicing. In previous works we had shown that measuring the fraction of debranched molecules at different concentrations of 5'-exon makes it possible to determine the dissociation constant (Kd) of the two molecules and therefrom, the affinity of the intron for its 5'-exon ((37, 38); see also (39)). Accordingly, we perform debranching experiments using gel-purified Pl.LSU/2 intron lariat and 5'-exon RNA variants carrying different bases at the EBS2a and IBS2a sites, respectively. We expected that differences in Kd values for matched and mismatched combinations of lariat and 5'-exon RNAs would allow us to demonstrate compensatory base changes indicative of a direct interaction between the EBS2a and IBS2a sites. However, and despite multiple attempts using different experimental conditions and intron constructs (data not shown), we were unable to detect the EBS2a-IBS2a pairing by this approach. Different reasons could account for these negative results. The putative interaction may not form at all in our assay, or, it could be formed without being required for recognition of the 5'-exon RNA by the ribozyme, which would make it undetectable kinetically. Another possibility however, is that the EBS2a-IBS2a cannot come into existence because its formation requires the contribution of the Pl.LSU/2 reverse transcriptase enzyme, which was not present in our *in vitro* system. This latter possibility is consistent with the phylogenetic analyses described above since statistical support for an interaction between the EBS2a and IBS2a positions was only obtained when RT-encoding introns were analysed as a separate sub-group of sequences.

The EBS2a-IBS2a interaction between the intron ribozyme and its DNA target site is required for intron mobility.

Despite the negative results described above, evolutionary conservation of the EBS2a-IBS2a pairing spoke for an important role in some other intron function(s). One obvious possibility was its participation in the network of EBS(RNA)-IBS(DNA) interactions that the intron lariat, in complex with its RT protein, establishes with its target site during reverse-splicing into DNA.
Accordingly, we next sought to investigate the possible implication of the EBS2a-IBS2a interaction in DNA target site recognition during intron retrotransposition. As mobility of the Pl.LSU/2 intron has not been demonstrated experimentally, we decided to use as a model system for examining retrotransposition intron EcI5, a subgroup IIB1 intron encoding for a CL1-lineage RT (40) that is naturally present in the virulence plasmid pO157 of pathogenic *E. coli* strain O157:H7 (41, 42). As already mentioned, the EcI5 intron potentially forms a EBS2a(U)-IBS2a(A) pairing with its flanking 5'-exon (Figure 1A). Moreover, in a previous work this intron was shown to be highly mobile by using a quantitative 'two-plasmid' genetic assay in *E. coli* (26). In that same work, the authors identified bases C-18, C-17, A-15 and A-14 lying on the distal 5' region of the target site, immediately upstream of the canonical IBS1 and IBS2 segments, as being important for intron integration and proposed that these bases could be directly recognized by the intron RT ((26); Figure 1B). Remarkably, one of these four critical positions, A-14, which coincides precisely with our phylogenetically identified 'IBS2a' site, was found to be essential for intron integration since 100% of the target sites invaded by the wild-type EcI5 intron contained an adenine base at this position (26). We thus decided to reconstitute in our laboratory the genetic assay of Zhuang and co-workers (Materials and Methods) in order to test the possible involvement of the EBS2a-IBS2a pairing in DNA target site recognition by the EcI5 intron. The 'two-plasmid' genetic assay and the mobility experiments we carried out in *E. coli* are outlined in figure 3A. The 'donor' plasmid contains an IPTG-inducible T7lac promoter and expresses in tandem both the EcI5 intron flanked by its exons and deleted from most of its ORF (EcI5 'ribozyme') and the EcI5 RT enzyme. The EcI5 ribozyme carries a phage T7 promoter inserted in its domain IV (Figure 1A). The precise location of this T7 promoter differs from that of Zhuang and co-workers and was chosen to allow the correct folding of intron subdomain IVb. The 'recipient' plasmid contains the EcI5 target site sequence (deduced from the natural insertion site of the intron in plasmid pO157) cloned upstream of a promoter-less tetracycline resistance gene (tet^R). For mobility assays, *E. coli* HMS174(DE3), a strain expressing an IPTG-inducible T7 RNA polymerase, is co-transformed with the donor and the recipient plasmids (Materials and Methods). In this genetic context, addition of IPTG to the culture allows over-expression of the EcI5 intron ΔORF ribozyme and the RT, which results in turn in triggering intron retrotransposition. Integration events of the EcI5 ribozyme into the recipient plasmid target site are detected through emergence of tetracycline-resistant colonies due to the expression of the tet^R gene which is specifically activated by the EcI5 ribozyme-embedded T7 promoter (Figure 3A and Materials and Methods).

In order to test the EBS2a-IBS2a pairing we generated mutant versions of the EcI5 intron ΔORF and of its DNA target site that carried different bases at the EBS2a and IBS2a sites (Figure 3A). As a control for intron mobility through the canonical TPRT mechanism, a donor plasmid expressing a catalytically inactive EcI5 RT enzyme (by mutation of the catalytic motif YADD to YAHH) was also constructed. Mutant donor and recipient plasmids were then used in five independent selection experiments, conducted in parallel. In these experiments, each one of the different donor plasmids was co-transformed into *E. coli* HMS174(DE3) with an equimolar mix composed of the four different recipient plasmids (Figure 3A). Intron retrotransposition was triggered by addition of IPTG to liquid cultures grown to early log phase and integration events were detected by plating the cells on LB agar containing ampicillin and tetracycline (Materials and Methods). The 5'-integration junctions of targeted recipient plasmids in Tet^R+Amp^R colonies were then analysed by colony PCR followed by sequencing in order to determine the identity of the bases at both the EBS2a and IBS2a sites.
The results obtained show that nearly without exceptions, each EcI5 ribozyme variant only invaded recipient plasmids carrying an IBS2a nucleotide complementary to the one found at the intron EBS2a position (Figure 3B). These results clearly demonstrate that position -14 (IBS2a) of the target site is engaged in a direct Watson-Crick interaction with the EBS2a site of the EcI5 intron. The specificity of the interaction is extremely high since only three mismatched combinations were observed among a total of 178 sequenced integration events. Moreover, there was only a single G-U wobble pair selected which suggests that the EBS2a-IBS2a pairing obeys a strict Watson-Crick geometry. These results do not rule out the possibility of additional contacts between the EBS2a and/or IBS2a sites and the intron-encoded reverse transcriptase. However, if such contacts take place, they must be compatible with free exchange of all four Watson-Crick combinations at those sites, since our selection experiments demonstrate that the specificity of recognition of target site IBS2a depends only on ribozyme site EBS2a. Altogether, our results establish that EBS2a-IBS2a constitutes an additional, previously unnoticed, Watson-Crick pairing necessary for DNA target site recognition during retrotransposition of the EcI5 intron. Moreover, as indicated by our comparative sequence analyses, this interaction is most certainly at play during mobility of most, if not all, subgroup IIB1 and IIB2 introns that encode for a RT enzyme. From a structural point of view, it is interesting to note that positioning of the EBS2a site just upstream of the EBS2 segment must impose some peculiar fold-back conformation on the RNA chain that comes out of the EBS2-IBS2 helix in order to concomitantly form the Watson-Crick EBS2a-IBS2a pairing.

Impact of the EBS2a-IBS2a pairing identity on intron mobility

The above selection experiments allowed to establish the rules obeyed by the EBS2a-IBS2a pairing. However, they do not make it possible to determine accurately the ability of each one of the Watson-Crick combinations to support intron mobility. In order to gather such data, mobility assays were carried out for pairs of donor and recipient plasmids with matched EBS2a-IBS2a combinations. In

these assays intron retrotransposition was triggered by IPTG under the same experimental conditions that the ones used for selection experiments. After induction the cells were serially diluted and plated onto LB agar supplemented with different appropriate antibiotics. Mobility efficiencies were then calculated from the ratio of Amp^R+Tet^R colonies over Amp^R+Cam^R colonies as described in Materials and Methods.

The mobility efficiency obtained for the wt EBS2a(U)-IBS2a(a) combination is ~88%, which is in agreement with the efficiencies previously reported (26) and confirms that the EcI5 intron is a highly mobile retrotransposon (Figure 4). Our results show that, except for the EBS2a(A)-IBS2a(t) combination, the other Watson-Crick pairings support intron mobility with the same high efficiency (Figure 4). The relatively modest performance of the EBS2a(A)-IBS2a(t) combination (~24%) is somewhat surprising with regard to our selection experiments on the one hand, and phylogenetic data on the other, as the latter do not reveal an under-representation of this Watson-Crick combination in natural intron sequences. Accordingly, we suspect that the effect of the EBS2a(A)-IBS2a(t) combination on mobility may be due to the particular sequence context of the EBS2a and EBS2 sites in the EcI5 intron (Figure 1A) which could favour the formation of alternative and non-functional interactions engaging nucleotides EBS2a(A) or IBS2a(t). We note that the nucleotide immediately downstream of the EBS2 segment, at position 258, is an adenosine. One possibility is the formation of an alternative pairing between A258 and the thymidine at the IBS2a site which would illicitly 'extend' the EBS2-IBS2 helix by one base pair. Such an extended pairing would compete with formation of the legitimate EBS2a-IBS2a interaction and could lead to local misfolding of the multibranched 'EBS2' loop with subsequent impairment of reverse splicing of the intron into the DNA target site.

Disruption of the EBS2a-IBS2a pairing does not affect intron splicing in vivo

Having established the crucial role of the new EBS2a-IBS2a interaction in DNA target site recognition during retrotransposition, it was of importance to determine whether the interaction had a similar major impact on 5'-exon recognition by the intron during *in vivo* splicing. Recall that our initial lariat debranching experiments using the PI.LSU/2 intron were unsuccessful in detecting an interaction between the EBS2a and the IBS2a sites. However, those experiments were performed *in vitro*, in the absence of the intron-encoded RT, raising the possibility that the lack of the protein may have prevented in some way the formation of the pairing. Accordingly, we decided to investigate the impact of the EBS2a-IBS2a interaction on RT-dependent splicing by comparing *in vivo* the efficiency of splicing of RNA precursors with matched versus mismatched EBS2a-IBS2a combinations. In these RNA precursors, each EBS2a variant ribozyme is flanked by a 5'-exon containing the wild-type adenosine at the IBS2 site (Figure 5). The donor plasmids encoding these mutant constructs were transformed (in the absence of recipient plasmid) into *E. coli* HMS174(DE3) cells and over-expressed under the same experimental conditions as the ones used previously (Materials and Methods). Total

cellular RNA was then extracted from the cultures and used to quantitate splicing by the commonly used primer extension approach (43, 44) with a primer complementary to nucleotides 120-142 in subdomain IC of the EcI5 intron (Materials and Methods and Figure 1A). This assay generates a cDNA product of 142 nucleotides corresponding to spliced intron lariat ('L' band) and another predominant, longer cDNA product, derived from unspliced precursor RNA ('P' band). The results obtained show that all constructs splice with very similar efficiencies regardless of having a matched or mismatched EBS2a-IBS2a combination between the intron and its flanking 5'-exon (Figure 5). These results recapitulate those already obtained *in vitro* with the debranching assays and show that even in the presence of the RT protein, disruption of the EBS2a-IBS2a interaction does not detectably affect intron splicing. Either an all RNA EBS2a-IBS2a pairing between the intron and its flanking 5'-exon does not form at all or, if it does form, this interaction is not rate-limiting for intron splicing and therefore remains undetectable.

Regarding the 'YAHH' control experiment with a reverse transcriptase carrying a YADD to YAHH mutation in its catalytic site, it is interesting to note that this mutant RT assists intron splicing as efficiently as the wild-type enzyme. This is consistent with previous results obtained for the yeast mitochondrial introns aI1 and aI2 (45, 46) and the bacterial RmInt1 intron from *Sinorhizobium meliloti* (43).

Structural support for the EBS2a-IBS2a pairing and its stabilization by the intron reverse transcriptase

Our experimental results firmly demonstrated the existence of the EBS2a-IBS2a interaction and its crucial role in DNA target site recognition during intron mobility. Nevertheless, the potential contribution of the intron-encoded RT to the establishment of this pairing, as initially suggested by our phylogenetic analyses, remained hypothetical. Unexpectedly, however, the recent publication of the cryoelectron microscopy (cryo-EM) structures of the *Thermosynechococcus elongatus* Th.e.I3 intron (also designated as *T.el*4h) reverse splicing into DNA (13) enlightens on this question by providing structural support both for the EBS2a-IBS2a interaction and for the role of the protein in stabilizing this Watson-Crick pairing. Like the EcI5 intron, Th.e.I3 is a IIB1 intron encoding for a chloroplast-like 1 (CL1) reverse transcriptase. Moreover, these introns share extensive sequence similarities both in their ribozyme and RT components (27). Accordingly, the Th.e.I3 intron has the potential to form a EBS2a(C)-IBS2a(G) base pair with its DNA target (Figure 6A). The recent cryo-EM structures of the Th.e.I3 RNP bound to its DNA target allow to visualise two different stages of the reverse splicing process (13). Interestingly, in the structure of the pre-second step of reverse splicing (pre-2r; PDB: 6MEC), C259 (= EBS2a site) and G-14 (= IBS2a site) on the top strand of the DNA target site are found in a configuration that allows them to form an anti-parallel, canonical Watson-Crick pairing, in full agreement with our phylogenetic and experimental data (Figure 6B). Curiously, however, in the cryo-EM structure depicting the pre-first step of reverse splicing (PDB: 6ME0), C259 has a completely

opposite orientation giving rise to three clashes with donor and acceptor groups of base G-14 of the DNA target. The reasons for this discrepancy are unknown to us (the lack of awareness of the EBS2a-IBS2a interaction together with limited resolution of the map could account for it). Importantly, however, this discrepancy does not seem to have any functional significance since all the other neighbouring structures and intron-DNA target interactions (EBS1-IBS1 and EBS2-IBS2) are identical in the two maps. Anyway, the pre-2r structure now shows that the EBS2a(C)-IBS2a(G) Watson-Crick interaction forms above the canonical EBS2-IBS2 duplex and adopts a peculiar orientation since the two pairings make a ~90° angle instead of being stacked upon each other (Figure 6B). Moreover, establishment of the EBS2a(C)-IBS2a(G) pairing induces a sharp turn of the RNA chain that results in a 'loop' configuration that locks the EBS2-IBS2 helix, probably preventing it from dissociating (Figures 6B-C). Therefore, this local architecture could be decisive in driving reverse splicing of the intron into its target DNA by helping to maintain the open conformation of the target site during pairing of the intron RNA to the DNA top strand (Figure 6C). Importantly, the cryo-EM structure also shows that this intricate RNA fold is stabilized by the intron reverse transcriptase through contacts between the EBS2a(C) and IBS2a(G) nucleotides and several basic amino acid residues (R327, K332, K336) lying in a conserved segment at the boundary of protein domains RT and X(thumb) (Figure 6B and Supplementary Figure S2). The RNA groups recognized by the protein are the O2 group of base EBS2a(C) and two nonbridging phosphate oxygens belonging to the EBS2a(C) and IBS2a(G) nucleotides. Interestingly, the nature of these RNA groups enables any other Watson-Crick pair to establish the same type of protein-RNA contacts (a purine at EBS2a will have its N3 group at the same place as the pyrimidine O2). Thus, this recognition mode allows to 'sense' the precise geometry of the Watson-Crick pair without imposing the identity of the bases. This configuration rationalizes our experimental demonstration that the EBS2a-IBS2a interaction is completely reprogrammable solely through changes in the ribozyme component of the EcI5 RNP.

DISCUSSION

In this work we have demonstrated the existence in two major subclasses of group II introns of an additional Watson-Crick base pair that contributes to bind the intron RNA to its DNA target. We found that this previously unnoticed interaction, which we call EBS2a-IBS2a, plays an essential role in intron retrotransposition. Even though the EBS2a-IBS2a pairing is restricted to subgroups IIB1 and IIB2, our studies shed new light on some general aspects of DNA target site recognition by group II introns and hold promise for the development of new or pre-existing targetron systems into more extensively reprogrammable tools for gene editing in bacteria.

Reverse splicing of the intron RNA into double-stranded DNA targets constitutes a major challenge faced by the RNP particle during mobility. One of the reasons for this is that base pairing of the intron RNA to the complementary sequences on the top strand of the target site requires the RNP

not only to open the DNA double helix at the insertion site, but also prevent subsequent reassociation of the two DNA strands by branch migration at the expense of the RNA-DNA duplex. The EBS2a-IBS2a interaction represents an unanticipated structural solution to this problem as the uncommon topology of the EBS2a-IBS2a pairing, whose formation induces the RNA chain to fold backwards, should 'lock' the adjacent EBS2-IBS2 helix, preventing it from unfolding (Figures 6B-C). The fact that molecular evolution selected this solution, instead of merely extending the EBS2-IBS2 helix by one more base pair, argues indeed that the complex fold imposed by this novel interaction contributes more efficiently to the stable unwinding of double-stranded DNA target sites. Moreover, as suggested by the lack of statistical evidence for the EBS2a-IBS2a pairing in ORF-less introns and further supported by the cryo-EM structure of the Th.e.I3 intron ((13); Figure 6B), the intron-encoded RT clearly plays an essential role in stabilizing the EBS2a-IBS2a base pair, making it unlikely that the intricate fold around the EBS2-IBS2 helix could come into existence in the absence of the protein component of the RNP. Nevertheless, it is important to note that some ORF-less introns may retain the ability to form the EBS2a-IBS2a pairing as long as they remain under selective pressure for mobility (promoted by RTs synthesized *in trans*). This is well illustrated by the numerous IIB1 introns that colonize the genome of the thermophilic cyanobacterium *Thermosynechococcus elongatus* (27). All these introns share very similar ribozyme structures, but only a few of them (such as the Th.e.I3 intron) encode RT proteins; the majority are ORF-less introns which, nevertheless, can be spliced and mobilized by the RT proteins originating from a complete intron. Inspection of the 'EBS2' loop and flanking 5'-exon regions of these ORF-less introns shows that most of them conserve the potential to form the EBS2a-IBS2a pairing, as expected from the latter being necessary for their retrotransposition into double-stranded DNA target sites. Altogether, the structural and functional characteristics of the EBS2a-IBS2a pairing provide a perfect example of the co-evolution of the intron RNA structures with the RT protein. This co-evolution process is thought to be at the very basis of group II intron diversification (4, 9) and supports the 'retroelement ancestor' hypothesis for the origin of these introns.

As described in the Introduction, functional studies on mobility (23, 24) and biochemical delineation of the DNA target sites for a handful of subgroup IIA and IIB introns (16, 21, 22, 25–27) showed that the molecular drivers of duplex unwinding at the target site consist of the EBS2-IBS2 pairing and a small number (typically 1 to 4) of specific positions that are located upstream of this pairing and differ in each system. As the latter nucleotides lie outside the IBS segments known to interact with the intron RNA, they were assumed to be contacted solely by the RT protein. These studies also led to a general model for DNA target site recognition in which the RT initiates DNA melting by first binding to those critical nucleotides, followed by subsequent pairing of the intron RNA to the complementary IBS1, IBS2 and IBS3/δ' segments spanning the intron insertion site (17, 22). Identification of the EBS2a-IBS2a pairing now calls for a re-evaluation of these assumptions. First, just like the IBS2a nucleotide, some of the other critical positions of the DNA target in different intron systems might well turn out to be recognized simultaneously by the RNA and the protein moieties of the RNP. Even though we could not find evidence for possible homologs of the EBS2a-IBS2a Watson-Crick base pair in other intron subgroups, it cannot be excluded that some of the critical DNA

target positions bind directly to the intron RNA through other types of molecular interactions. Second, the requirement for the RT to stabilize the EBS2a-IBS2a pairing and the intricate fold induced by the two partners suggest that instead of taking place sequentially, binding of the RT protein and pairing of the intron RNA to the DNA target site occur concomitantly, with the two processes acting cooperatively to drive opening of the DNA double helix.

One question that remains unclear is whether the EBS2a-IBS2a base pair comes into existence during intron splicing. Disruption of this pairing did not detectably affect *in vivo* splicing in our experiments, an outcome which is in agreement with the dispensability of the EBS2-IBS2 interaction for *in vitro* self-splicing of various introns (47–49). While evolutionary conservation of the EBS2-IBS2/ EBS2a-IBS2a pairings is clearly linked to their role in intron mobility, we cannot rule out the formation of an all-RNA EBS2a-IBS2a pairing similarly stabilized by the RT protein during intron splicing. Additional structural data allowing comparison of intron RNP complexes with RNA and DNA substrates will be required to settle these questions.

Given the importance of the EBS2a-IBS2a interaction for mobility of the EcI5 intron one may wonder how introns belonging to other structural subgroups manage to invade their DNA targets. Leaving aside group IIC introns whose mobility does not involve an EBS2-IBS2 pairing but relies on recognition of DNA stem-loop structures (most often rho-independent transcription terminators; (6, 50)), it could be that representatives of other subgroups have evolved different, yet unidentified, structural devices to promote DNA strand separation. On the other hand, introns that preferentially target single-stranded DNA are certainly under much weaker selection pressure to preserve the EBS2a-IBS2a pairing. This is probably the case of the RmInt1 intron from *Sinorhizobium meliloti* (51), a bacterial IIB intron of protein subclass D that predominantly invades single-stranded DNA targets at replication forks (52, 53). Although the RmInt1 intron can form the EBS2-IBS2 interaction with its DNA target, there is no phylogenetic evidence for an EBS2a-IBS2a pairing in this intron nor in its close relatives. As for members of subgroup IIA, they differ in the organization of the distal section of subdomain ID (Supplementary Figure S3A), which includes an additional hairpin structure (helix (iii)a), while the EBS2 segment is part of a terminal, rather than internal, loop. These features result in a very different structural context around the EBS2-IBS2 helix, as can be visualized in the available cryo-EM structure of the group IIA *Lactococcus lactis* Ll.LtrB intron in complex with its RT and a 5'-exon RNA substrate ((12); Supplementary Figure S3B). The EBS2-IBS2 helix may be better stabilized within the resulting architecture, obviating the need for an additional interaction such as EBS2a-IBS2a to prevent helix unwinding. However, this scenario remains hypothetical since the cryo-EM structure displays an RNA substrate; a high-resolution structure of a group IIA intron RNP bound to its DNA target will be necessary to rigorously address this issue.

More generally, there is a need for structural data specifically pertaining to the mobility pathway(s) of group II introns. In this context, the recent cryo-EM structures of the Th.e.I3 intron RNP bound to the top strand of its DNA target (13) have been much welcome. In addition to providing structural support for our phylogenetic and experimental data on the EBS2a-IBS2a pairing, these structures also allow us now to propose an interacting partner for position -15 (top strand) of the DNA

target for bacterial IIB1 and IIB2 introns. The DNA target sites of the EcI5 and Th.e.I3 introns both contain an adenine at position -15 (Figures 1B and 6A, respectively). Moreover, in the initial studies of EcI5 mobility by Zhuang and co-workers, position A-15 was found to be the second most critical nucleotide for DNA target site recognition, just after position A-14 (now IBS2a) (26). This implies that A-15 is actively recognized in the EcI5 system and, given the relatedness between the two introns, a similar interaction should also prevail in the context of the Th.e.I3 RNP bound to the top strand of its DNA target. Furthermore, previous studies have shown that the identity of position -15 is critical as well for retrotransposition of the Th.e.I3 intron (27). We thus inspected the density maps corresponding to the two Th.e.I3 cryo-EM structures and found a portion of unassigned density just above, and contiguous to, the electron density corresponding to DNA target position G-14(IBS2a) (Supplementary Figure S4A). Suggestively, this 'empty' density can accommodate a deoxy-adenosine nucleotide (Supplementary Figure S4A), which turns out to be closely surrounded by amino acid residues lying at the end of block 2a of the 'RT' domain, in particular, glycine 173 and phenylalanine 176 (Supplementary Figures S2 and S4A). Combining these structural clues with our alignments of bacterial IIB1 and IIB2 intron sequences (Material and Methods) revealed not just a strong bias for a purine at position -15 of the DNA target, but a striking correlation between the identity of that purine and the identity of the amino acid residue at position 173 (Th.e.I3 protein numbering) of the RT enzyme (Supplementary Figure S4B). An adenine at position -15 is most often associated with a serine or alanine residue, whereas the occurrence of a guanine at that same position is strongly correlated with the presence of a glutamic acid. Such a clear-cut covariation pattern suggests the existence of specific molecular interactions between these particular nucleotides and amino acid residues. A direct interaction between DNA target G-15 and glutamic acid 173 of the RT is further supported by the phylogeny of bacterial IIB1 introns (Supplementary Figure S5), which shows that this particular nucleotide-amino acid combination arose a number of times independently during the evolution of these introns. It is also worth noting that the Th.e.I3 RT protein has a glycine at position 173 (Supplementary Figure S2) which, according to our phylogenetic data, may not be the optimal partner for DNA target position A-15. It should be interesting to explore this novel DNA-amino acid interaction experimentally and verify that engineering of residue 173 of the RT can lead to a shift in the specificity of recognition of position -15 of the DNA target by the intron RNP.

Finally, our work has direct biotechnological implications, since it should contribute to expand the use of targetrons for gene targeting in bacteria. Targetrons have been employed for genetic engineering of a wide variety of bacteria for over 20 years and are currently an established component of the toolbox of bacterial geneticists due to their ease of use, efficiency and very high specificity (18–20, 54, 55 and references therein). In addition, they constitute a tool of choice for manipulating bacteria with inefficient homology-based DNA repair processes since they operate independently from homologous recombination. This latter point is illustrated by the key contribution of the targetron technology to our current understanding of Clostridium, a genus that includes important human pathogens as well as species relevant for industrial production of solvents and chemicals, but that is refractory to conventional genetic approaches (56–59). Targetrons are largely

reprogrammable through RNA engineering of the EBS sites of the ribozyme component. However, the few (typically 2 to 6) critical positions of the target DNA that lie outside the IBS segments and whose recognition is generally thought to involve the intron-encoded RT impose limitations to the set of genomic sites that can be efficiently disrupted by a given targetron system, for information is lacking about how to engineer the RT enzyme appropriately. Nevertheless, as those nucleotides are recognized with variable, and most often, moderate stringency, it has been possible to develop computer programs (for commercially distributed targetrons) that help to identify genomic targets carrying optimal matches at these positions. Another strategy to overcome these limitations includes using a variety of group II introns so as to develop targetrons with different 'specificities' (those critical target positions outside the IBS sites), thus making it possible to choose the most suitable system according to the sequence of the genomic target to be disrupted. However, only two systems are commercially available at present. These are the subgroup IIA Ll.LtrB targetron (17, 18, 54), which was the first one to be developed, and the EcI5 targetron (26). The latter system is in principle an excellent alternative for the more widely used Ll.LtrB targetron, due to its outstanding mobility efficiency and high specificity of target recognition. Still, the use of the EcI5 targetron has remained limited so far, possibly because of the absolute requirement for an adenosine at position -14 of the DNA target (26). However, our demonstration that this requirement results from the existence of the EBS2a-IBS2a base pair should now make it possible to choose the base at position -14 at will, by merely changing the nucleotide at the EBS2a site of the ribozyme. As a consequence, the number of genomic sites accessible to this targetron will be multiplied by four, and the frequency of highly ranked potential target sites in the *E. coli* K12 genome increased from 1 per 621 nucleotide residues (26) to about 1 per 155.

Other existing or potential targetron systems based on subgroup IIB1 and IIB2 introns should benefit from our findings. One striking example is provided by the TeI3c/4c (TeI3c ribozyme and TeI4c RT protein) thermotargetron derived from two closely related IIB1 introns (forming a 'twintron') found in *Thermosynechococcus elongatus* (57). The TeI3c/4c thermotargetron allows gene editing in thermophilic bacteria and has very high mobility efficiency. Moreover, there are only two critical positions in the 5'-distal section of the TeI3c/4c DNA target site outside of the canonical IBS sites. Of those two positions, A-15 and A-14, which are recognized with high stringency, the latter is none other than the IBS2a site, so that its base content must no longer be regarded as imposed. As for A-15, our comparative phylogenetic analyses (see above and Supplementary Figures S4 and S5) strongly suggest that it not only interacts with the amino acid at position 173 (Th.e.I3 protein numbering) of the intron-encoded RT, but that it should be possible to replace it by a G, provided the amino acid at site 173 is changed to a glutamic acid. Should that turn out to be the case, the TeI3c/4c system would become the most completely reprogrammable targetron to date, with the presence of a purine, rather than a pyrimidine, at position -15 the only significant constraint when choosing a target site for this targetron. Finally, in addition to improving pre-existing systems, our work may also contribute to the future development of new targetrons derived from other IIB1, and possibly IIB2, group II introns. One reason for this is that identification of the EBS2a and IBS2a sites facilitates the precise delimitation of

the boundaries of the canonical EBS2 and IBS2 segments, which may have been improperly assigned in some cases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

The authors are grateful to Frédéric Boccard (Institute for Integrative Biology of the Cell, 91198 Gif-sur-Yvette, France) for the gift of DNA from *E. coli* strain O157:H7 containing plasmid pO157 which harbours the EcI5 intron.

FUNDING

This work was supported by the 'Comité de L'Essonne de la Ligue Nationale Contre le Cancer' [M31415 to M.C.] and by the 'BIG Lidex' program to [M.C]. Funding for open access charge: CNRS intramural funding

REFERENCES

1. Lambowitz,A.M. and Zimmerly,S. (2011) Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.*, **3**, a003616.

2. Zimmerly,S. and Semper,C. (2015) Evolution of group II introns. *Mob. DNA*, **6**, 7.

3. Michel,F., Umesono,K. and Ozeki,H. (1989) Comparative and functional anatomy of group II catalytic introns--a review. *Gene*, **82**, 5–30.

4. Toor,N., Hausner,G. and Zimmerly,S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.

5. Simon,D.M., Clarke,N.A.C., McNeil,B.A., Johnson,I., Pantuso,D., Dai,L., Chai,D. and Zimmerly,S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA N. Y. N*, **14**, 1704–1713.

6. Robart,A.R., Seo,W. and Zimmerly,S. (2007) Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 6620–6625.

7. Granlund,M., Michel,F. and Norgren,M. (2001) Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *J. Bacteriol.*, **183**, 2560–2569.

8. Zimmerly,S., Hausner,G. and Wu,Xc. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.*, **29**, 1238–1250.

9. Fontaine,J.M., Goux,D., Kloareg,B. and Loiseaux-de Goër,S. (1997) The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga Pylaiella littoralis belong to two different lineages which apparently coevolved with the group II ribosyme lineages. *J. Mol. Evol.*, **44**, 33–42.

10. Matsuura,M., Saldanha,R., Ma,H., Wank,H., Yang,J., Mohr,G., Cavanagh,S., Dunny,G.M., Belfort,M. and Lambowitz,A.M. (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev.*, **11**, 2910–2924.

11. Matsuura,M., Noah,J.W. and Lambowitz,A.M. (2001) Mechanism of maturase-promoted group II intron splicing. *EMBO J.*, **20**, 7259–7270.

12. Qu,G., Kaushal,P.S., Wang,J., Shigematsu,H., Piazza,C.L., Agrawal,R.K., Belfort,M. and Wang,H.-W. (2016) Structure of a group II intron in complex with its reverse transcriptase. *Nat. Struct. Mol. Biol.*, **23**, 549–557.

13. Haack,D.B., Yan,X., Zhang,C., Hingey,J., Lyumkis,D., Baker,T.S. and Toor,N. (2019) Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA. *Cell*, **178**, 612-623.e12.

14. Lambowitz,A.M. and Belfort,M. (2015) Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol. Spectr.*, **3**.

15. Zimmerly,S., Guo,H., Perlman,P.S. and Lambowitz,A.M. (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell*, **82**, 545–554.

16. Guo,H., Zimmerly,S., Perlman,P.S. and Lambowitz,A.M. (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J.*, **16**, 6835–6848.

17. Mohr,G., Smith,D., Belfort,M. and Lambowitz,A.M. (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev.*, **14**, 559–573.

18. Karberg,M., Guo,H., Zhong,J., Coon,R., Perutka,J. and Lambowitz,A.M. (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat. Biotechnol.*, **19**, 1162–1167.

19. Zhong,J., Karberg,M. and Lambowitz,A.M. (2003) Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic Acids Res.*, **31**, 1656–1664.

20. Enyeart,P.J., Mohr,G., Ellington,A.D. and Lambowitz,A.M. (2014) Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mob. DNA*, **5**, 2.

21. Yang,J., Mohr,G., Perlman,P.S. and Lambowitz,A.M. (1998) Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites in vitro. *J. Mol. Biol.*, **282**, 505–523.

22. Singh,N.N. and Lambowitz,A.M. (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J. Mol. Biol.*, **309**, 361–386.

23. Ichiyanagi,K., Beauregard,A., Lawrence,S., Smith,D., Cousineau,B. and Belfort,M. (2002) Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol. Microbiol.*, **46**, 1259–1272.

24. Coros,C.J., Landthaler,M., Piazza,C.L., Beauregard,A., Esposito,D., Perutka,J., Lambowitz,A.M. and Belfort,M. (2005) Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol. Microbiol.*, **56**, 509–524.

25. Jiménez-Zurdo,J.I., García-Rodríguez,F.M., Barrientos-Durán,A. and Toro,N. (2003) DNA target site requirements for homing in vivo of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. *J. Mol. Biol.*, **326**, 413–423.

26. Zhuang,F., Karberg,M., Perutka,J. and Lambowitz,A.M. (2009) EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA N. Y. N*, **15**, 432–449.

27. Mohr,G., Ghanem,E. and Lambowitz,A.M. (2010) Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol.*, **8**, e1000391.

28. Candales,M.A., Duong,A., Hood,K.S., Li,T., Neufeld,R.A.E., Sun,R., McNeil,B.A., Wu,L., Jarding,A.M. and Zimmerly,S. (2012) Database for bacterial group II introns. *Nucleic Acids Res.*, **40**, D187-190.

29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

30. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

31. Hall, T.A. (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser*., **41**, 95-98.

32. Eckert,B. and Beck,C.F. (1989) Overproduction of transposon Tn10-encoded tetracycline resistance protein results in cell death and loss of membrane potential. *J. Bacteriol.*, **171**, 3557–3559.

33. Waldern,J., Schiraldi,N.J., Belfort,M. and Novikova,O. (2020) Bacterial Group II Intron Genomic Neighborhoods Reflect Survival Strategies: Hiding and Hijacking. *Mol. Biol. Evol.*, **37**, 1942–1948.

34. Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci. CABIOS*, **7**, 347–352.

35. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.

36. Costa,M., Fontaine,J.M., Loiseaux-de Goër,S. and Michel,F. (1997) A group II self-splicing intron from the brown alga Pylaiella littoralis is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. *J. Mol. Biol.*, **274**, 353–364.

37. Costa,M., Christian,E.L. and Michel,F. (1998) Differential chemical probing of a group II self-splicing intron identifies bases involved in tertiary interactions and supports an alternative secondary structure model of domain V. *RNA*, **4**, 1055–1068.

38. Costa,M. and Michel,F. (1999) Tight binding of the 5' exon to domain I of a group II self-splicing intron requires completion of the intron active site. *EMBO J.*, **18**, 1025–1037.

39. Chin,K. and Pyle,A.M. (1995) Branch-point attack in group II introns is a highly reversible transesterification, providing a potential proofreading mechanism for 5'-splice site selection. *RNA N. Y. N*, **1**, 391–406.

40. Dai,L. and Zimmerly,S. (2002) The dispersal of five group II introns among natural populations of Escherichia coli. *RNA N. Y. N*, **8**, 1294–1307.

41. Burland,V., Shao,Y., Perna,N.T., Plunkett,G., Sofia,H.J. and Blattner,F.R. (1998) The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. *Nucleic Acids Res.*, **26**, 4196–4204.

42. Dai,L. and Zimmerly,S. (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.*, **30**, 1091–1102.

43. Muñoz-Adelantado,E., San Filippo,J., Martínez-Abarca,F., García-Rodríguez,F.M., Lambowitz,A.M. and Toro,N. (2003) Mobility of the Sinorhizobium meliloti group II intron RmInt1 occurs by reverse splicing into DNA, but requires an unknown reverse transcriptase priming mechanism. *J. Mol. Biol.*, **327**, 931–943.

44. Molina-Sánchez,M.D., Martínez-Abarca,F. and Toro,N. (2010) Structural features in the C-terminal region of the Sinorhizobium meliloti RmInt1 group II intron-encoded protein contribute to its maturase and intron DNA-insertion function. *FEBS J.*, **277**, 244–254.

45. Eskes,R., Yang,J., Lambowitz,A.M. and Perlman,P.S. (1997) Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. *Cell*, **88**, 865–874.

46. Moran,J.V., Zimmerly,S., Eskes,R., Kennell,J.C., Lambowitz,A.M., Butow,R.A. and Perlman,P.S. (1995) Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. *Mol. Cell. Biol.*, **15**, 2828–2838.

47. Jacquier,A. and Michel,F. (1987) Multiple exon-binding sites in class II self-splicing introns. *Cell*, **50**, 17–29.

48. Mullineux,S.-T., Costa,M., Bassi,G.S., Michel,F. and Hausner,G. (2010) A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA*, **16**, 1818–1831.

49. Nagy,V., Pirakitikulr,N., Zhou,K.I., Chillón,I., Luo,J. and Pyle,A.M. (2013) Predicted group II intron lineages E and F comprise catalytically active ribozymes. *RNA*, **19**, 1266–1278.

50. Mohr,G., Kang,S.Y.-S., Park,S.K., Qin,Y., Grohman,J., Yao,J., Stamos,J.L. and Lambowitz,A.M. (2018) A Highly Proliferative Group IIC Intron from Geobacillus stearothermophilus Reveals New Features of Group II Intron Mobility and Splicing. *J. Mol. Biol.*, **430**, 2760–2783.

51. Martínez-Abarca,F., Zekri,S. and Toro,N. (1998) Characterization and splicing in vivo of a Sinorhizobium meliloti group II intron associated with particular insertion sequences of the IS630-Tc1/IS3 retroposon superfamily. *Mol. Microbiol.*, **28**, 1295–1306.

52. Martínez-Abarca,F., Barrientos-Durán,A., Fernández-López,M. and Toro,N. (2004) The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res.*, **32**, 2880–2888.

53. García-Rodríguez,F.M., Neira,J.L., Marcia,M., Molina-Sánchez,M.D. and Toro,N. (2019) A group II intron-encoded protein interacts with the cellular replicative machinery through the β-sliding clamp. *Nucleic Acids Res.*, **47**, 7605–7617.

54. Perutka,J., Wang,W., Goerlitz,D. and Lambowitz,A.M. (2004) Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. *J. Mol. Biol.*, **336**, 421–439.

55. Belfort,M. and Lambowitz,A.M. (2019) Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb. Perspect. Biol.*, **11**.

56. Heap,J.T., Pennington,O.J., Cartman,S.T., Carter,G.P. and Minton,N.P. (2007) The ClosTron: a universal gene knock-out system for the genus Clostridium. *J. Microbiol. Methods*, **70**, 452–464.

57. Mohr,G., Hong,W., Zhang,J., Cui,G., Yang,Y., Cui,Q., Liu,Y. and Lambowitz,A.M. (2013) A targetron system for gene targeting in thermophiles and its application in Clostridium thermocellum. *PloS One*, **8**, e69032.

58. Kwon,S.W., Paari,K.A., Malaviya,A. and Jang,Y.-S. (2020) Synthetic Biology Tools for Genome and Transcriptome Engineering of Solventogenic Clostridium. *Front. Bioeng. Biotechnol.*, **8**, 282.

59. Wen,Z., Lu,M., Ledesma-Amaro,R., Li,Q., Jin,M. and Yang,S. (2020) TargeTron Technology Applicable in Solventogenic Clostridia: Revisiting 12 Years' Advances. *Biotechnol. J.*, **15**, e1900284.

60. DeLano, W.L. (2002) The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA.

FIGURE LEGENDS

Figure 1. Secondary structure of the *Escherichia coli* subgroup IIB1 intron EcI5 and its DNA target site. (**A**) The secondary structure of the EcI5 intron is coloured by major domains (I to VI). Tertiary base-pairing interactions between the intron and its 5' and 3' flanking exons are designated as EBS (Exon Binding Site) – IBS (Intron Binding Site) pairings whereas long-range interactions involving intronic sequences are designated by greek letters. The EBS2a site involved in the new EBS2a-IBS2a interaction demonstrated in this work is highlighted in red. For mobility assays, most of the RT ORF in domain IV (coloured in black) was deleted and replaced by a T7 RNA polymerase promoter (PT7) as indicated by the double-headed arrows. The portion of domain IV coloured in red corresponds to the sequences still present in the EcI5 'ribozyme' construct used for mobility

experiments (see Materials and Methods and Figure 3). The IVa subdomain contains the high affinity binding site for the EcI5 RT enzyme. The translation signals used for synthesis of the RT in the natural intron are circled in green. SD: Shine-Dalgarno sequence. (**B**) Natural DNA recognition site of the EcI5 intron. Coloured nucleotides on the top strand (C-18, C-17, A-15, A-14 and T+5) are those previously identified as being critical for EcI5 mobility and suggested to be directly recognized by the RT (26). The present work demonstrates that position A-14 (in red), now named IBS2a, is in fact engaged in a Watson-Crick pairing with the intron EBS2a site (see text).

Figure 2. Base covariation events suggest a new tertiary interaction between the intron and its 5'-exon and/or DNA target site. The intron sequences used to count covariation events were collected as described in Materials and Methods (see also Supplementary Tables S1 to S3 and Supplementary Figure S5). Only those sequences with at most one mismatch in the canonical, 5-bp EBS2-IBS2 pairing were used for statistics. The asterisk in chloroplastic IIB1 introns indicates that in these introns, the base that covaries with the intron EBS2a site lies at position -13, instead of -14, of the 5'-exon/DNA target site (top strand). Mutual information values calculated with BioEdit are: 1.039 for bacterial IIB1 introns; 0.408 for mitochondrial IIB1 introns; 0.958 for chloroplastic IIB1 introns and 0.635 for cyanobacterial IIB2 introns. DNA TS: DNA target site.

Figure 3. Outline of the *in vivo* selection experiments and results. (**A**) Genetic organization of the donor and recipient plasmids used for the *in vivo* mobility assays. The donor plasmid over-expresses the EcI5 'ribozyme' flanked by its 5' (5'E) and 3' (3'E) exons, from a T7lac promoter. The EcI5 'ribozyme' corresponds to the intron∆ORF construct shown in figure 1A. The T7 promoter inserted in domain IV is indicated by a green arrow. The donor plasmid also over-expresses the EcI5 reverse transcriptase from a position downstream of the intron. The recipient plasmid contains the EcI5 target site (TS) upstream from a promoterless *tet*^{R} gene. Integrated recipient plasmids result from the integration of the EcI5 ribozyme and over-express the *tet*^{R} gene from the intron-inserted T7 promoter. The brown arrows indicate the primers used for colony PCR on Tet^R colonies. In order to analyse the 5'-integration junctions (results given in (**B**)), the resulting PCR products were sequenced with a forward primer that hybridizes to a segment of the recipient plasmid backbone. *tet*^{R}: tetracycline resistance gene; *cam*^{R}: chloramphenicol resistance gene; *amp*^{R}: ampicillin resistance gene. See Materials and Methods, Supplementary Table S4 and Supplementary Figure S1 for details on plasmid constructs and experimental conditions. (**B**) Sequences of the 5'-integration junctions generated during the selection experiments with each one of the donor plasmids shown in (**A**). The 'Nb clones' ratio is the number of each sequence over the total number of tet^R clones analysed. Dots indicate identical nucleotide bases, with respect to the first sequence shown for each selection experiment.

Figure 4. Impact of the identity of the EBS2a-IBS2a pairing on mobility efficiency. The identity of the bases at the EBS2a and IBS2a sites is in upper and lower case, respectively. Appropriate donor and recipient plasmids were co-transformed into *E. coli* HMS174(DE3) and intron mobility was triggered with 100 μM IPTG at 37°C for one hour. Mobility efficiencies were calculated as the ratio of (Amp^R+Tet^R)/(Amp^R+Cam^R) colonies (see Materials and Methods). Graph bars are the mean value from at least five mobility efficiency values. Error bars represent the standard error of the mean.

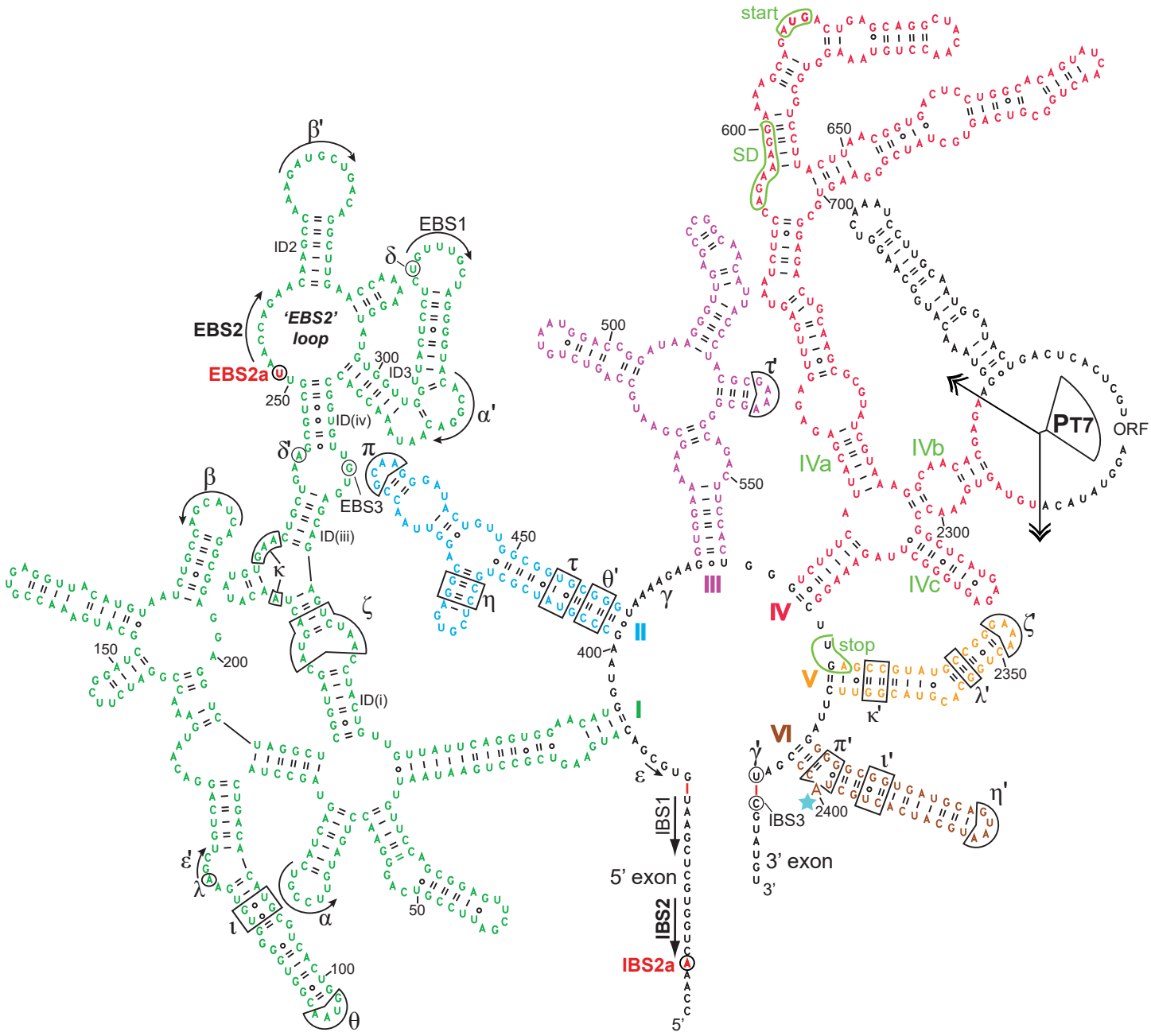
Figure 5. Impact of the EBS2a-IBS2a interaction on intron splicing *in vivo*. Primer extension reactions were performed on total cellular RNA extracted from *E. coli* HMS174(DE3) cells expressing EcI5 ribozyme (intronΔORF) precursors with the indicated matched or mismatched EBS2a-IBS2a combinations (in red; see also Materials and Methods). One representative 5% acrylamide-urea primer extension gel is shown. L(lariat) and P(precursor) are migration markers generated by primer extension of *in vitro* transcribed RNA. The identity of the lariat 'stop' band was further confirmed with additional gels (not shown) that include a sequence generated from direct sequencing of an *in vitro* transcribed wt EcI5 intronΔORF precursor. The weak band between the 'L' and 'P' cDNAs was consistently observed in all experiments but was not used for quantification because its origin is unknown.

Figure 6. Three-dimensional view of the EBS2a-IBS2a pairing and its role in DNA target site unwinding. (**A**) Secondary structure of the distal section of subdomain ID of the group II intron Th.e.I3 from *Thermosynechococcus elongatus.* The EBS2a site in this intron (C259) is highlighted in bold. The adjacent EBS2 segment involved in the canonical EBS2-IBS2 pairing is indicated with an arrow (interruption of the arrow accounts for the fact that the G in the middle of the segment is extruded from the EBS2-IBS2 helix in the cryo-EM structures (PDB: 6MEC and 6ME0)). Note that the natural Th.e.I3 intron contains a C-A mismatch at the δ-δ' interaction. The natural DNA target site of the Th.e.I3 intron is shown in magenta with the IBS2a base (G-14) on the top strand highlighted by a circle. IS: Insertion site. (**B**) Close-up views of the interactions involving the 'EBS2' loop region of the intron and the 'IBS2' section on the top strand of the DNA target site as modelled in the cryo-EM structure of the pre-2r state (PDB: 6MEC). The figure was prepared using PyMOL (60). The RNA elements are coloured according to (**A**) and the intron RT is in cyan. The positions coloured in green correspond to the strand (from U258 to A268) containing the EBS2 segment and the EBS2a site. The EBS2a-IBS2a interaction forms a Watson-Crick pairing on top of the EBS2-IBS2 helix (ladder representation) but it is not stacked upon it; rather, there is a ~90° angle between the two elements. A specific region of the RT enzyme acts as a 'lid' with basic amino acid residues (labelled) directly stabilizing the Watson-Crick geometry of the EBS2a-IBS2a pairing (see text). (**C**) Updated diagram of

the RNA-DNA base pairing interactions involved in recognition of the DNA target site by subgroup IIB1 and IIB2 introns and proposed model for the role of the new EBS2a-IBS2a interaction in promoting DNA duplex unwinding. The diagram illustrates that formation of EBS2a-IBS2a results in a ‘loop’ conformation that encloses the EBS2-IBS2 helix. This local architecture should favour reverse splicing of the intron into the DNA target site by preventing the two DNA strands from re-associating (see text).

Figure 1

A


start
SD
600
650
700
500
550
PT7
ORF
IVa
IVb
IVc
2300
2350
stop
2400
IBS3
3' exon
3'
EBS1
EBS2
'EBS2' loop
EBS2a
250
300
ID2
ID3
ID(iv)
EBS3
ID(iii)
ID(i)
150
200
450
400
50
100
IBS1
IBS2
IBS2a
5' exon
5'

B


Insertion site (IS)
IBS2a
IBS2
IBS1
IBS3
+5
dsDNA Target site
Top 5'... ATGCCAAACTGGTGCTCGAAT-CGTATGTATT...3'
Bottom 3'....TACGGTTTGACCACGAGCTTA-GCATACATAA...5'
-18 -17 -15 -14
5'-half IS
3'-half IS

Figure 2

Subgroup IIB1
(RT ORF present)

IBS2a / -14 DNA TS/ 5' exon EBS2a / Intron 'EBS2' internal loop		(24) A	(6) C	(17) G	(14) T/U
(15)	**A**	2	0	0	**13**
(17)	**C**	0	0	**17**	0
(5)	**G**	0	**4**	0	1
(24)	**U**	**22**	2	0	0

Bacteria

IBS2a / -14 DNA TS/ 5' exon EBS2a / Intron 'EBS2' internal loop		(3) A	(5) C	(0) G	(19) T/U
(22)	**A**	2	1	0	**19**
(0)	**C**	0	0	**0**	0
(4)	**G**	0	**4**	0	0
(1)	**U**	**1**	0	0	0

Mitochondria

IBS2a / -13 DNA TS*/ 5' exon EBS2a / Intron 'EBS2' internal loop		(4) A	(1) C	(4) G	(13) T/U
(12)	**A**	0	0	0	**12**
(4)	**C**	0	0	**4**	0
(1)	**G**	0	**1**	0	0
(5)	**U**	**4**	0	0	1

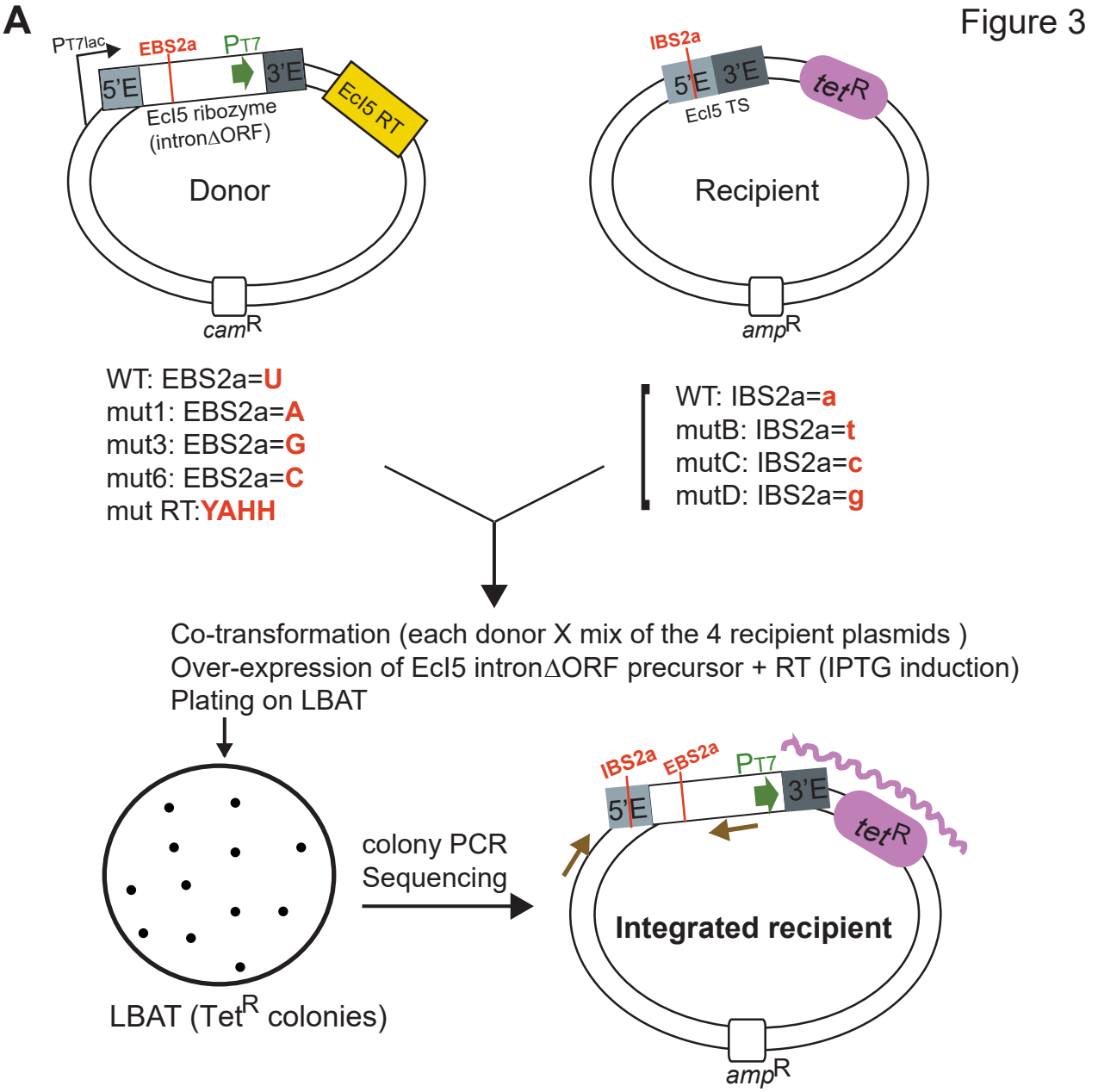
Chloroplast

Subgroup IIB2
(RT ORF present)

IBS2a / -14 DNA TS/ 5' exon EBS2a / Intron 'EBS2' internal loop		(10) A	(3) C	(1) G	(7) T/U
(12)	**A**	4	1	0	**7**
(1)	**C**	0	0	**1**	0
(2)	**G**	0	**2**	0	0
(6)	**U**	**6**	0	0	0

Bacteria

Figure 3



B

Selection exp	Nb clones	5'... IBS2a ? IBS2 IBS1 ... EBS2a EBS2 ...
Donor WT	45/45	5'...ccaaactggtgctcgaat-GTGCG.....GTTAACCAGA.....
Donor mut1	44/46	5'...ccaatctggtgctcgaat-GTGCG.....GTAAACCAGA.....
	1/46	.....a.....
	1/46	.....c.....
Donor mut3	43/44	5'...ccaactggtgctcgaat-GTGCG.....GTGAACCAGA.....
	1/44	.....t.....
Donor mut6	42/43	5'...ccaagctggtgctcgaat-GTGCG.....GTCAACCAGA.....
	1/43	.....a.....

Donor mut RT : no tetR colonies

Figure 4

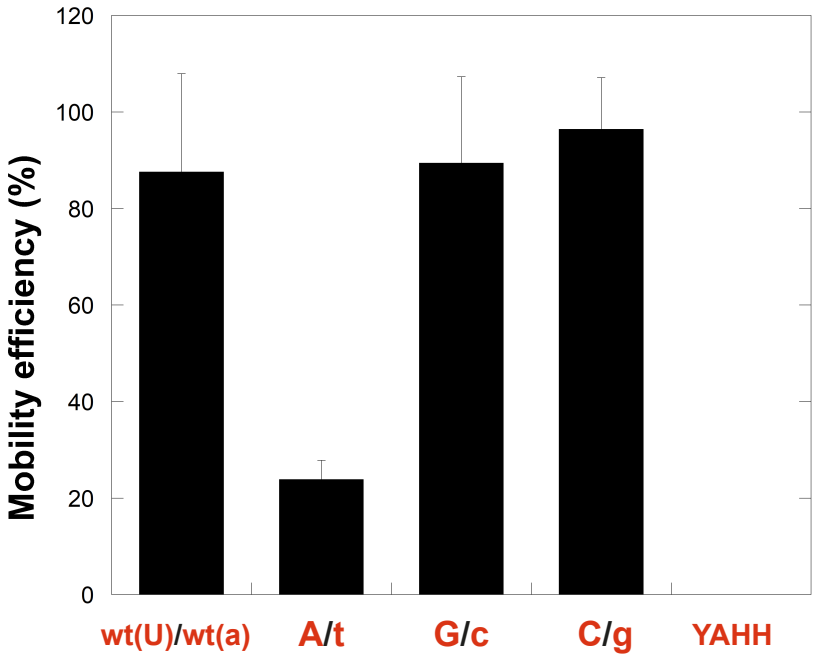
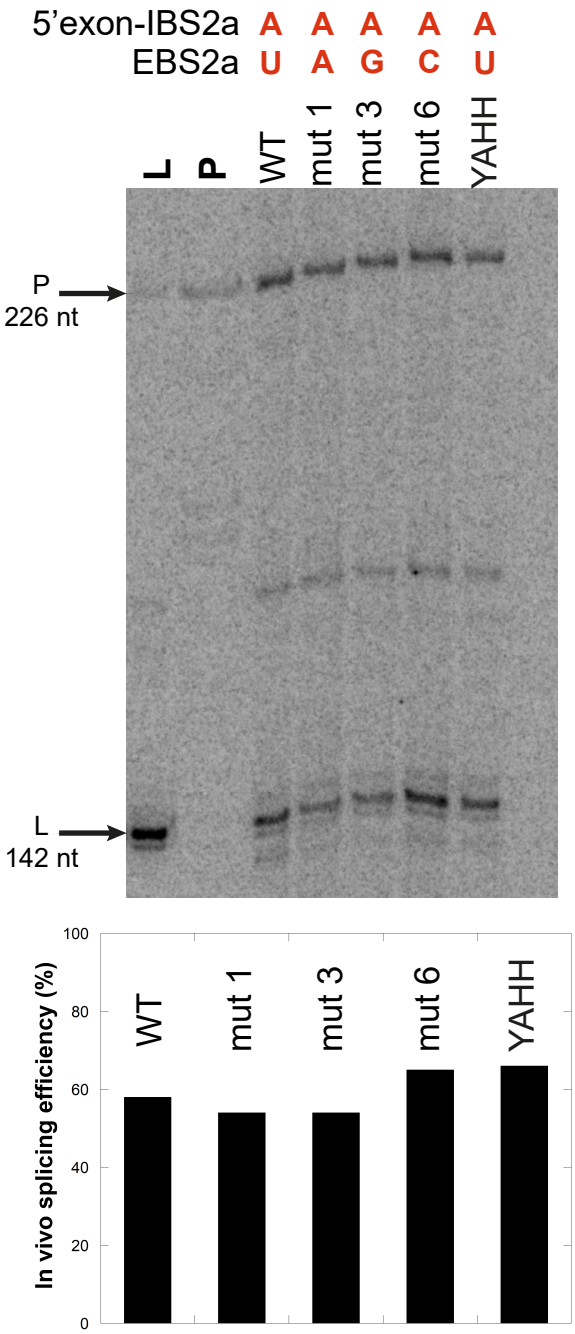
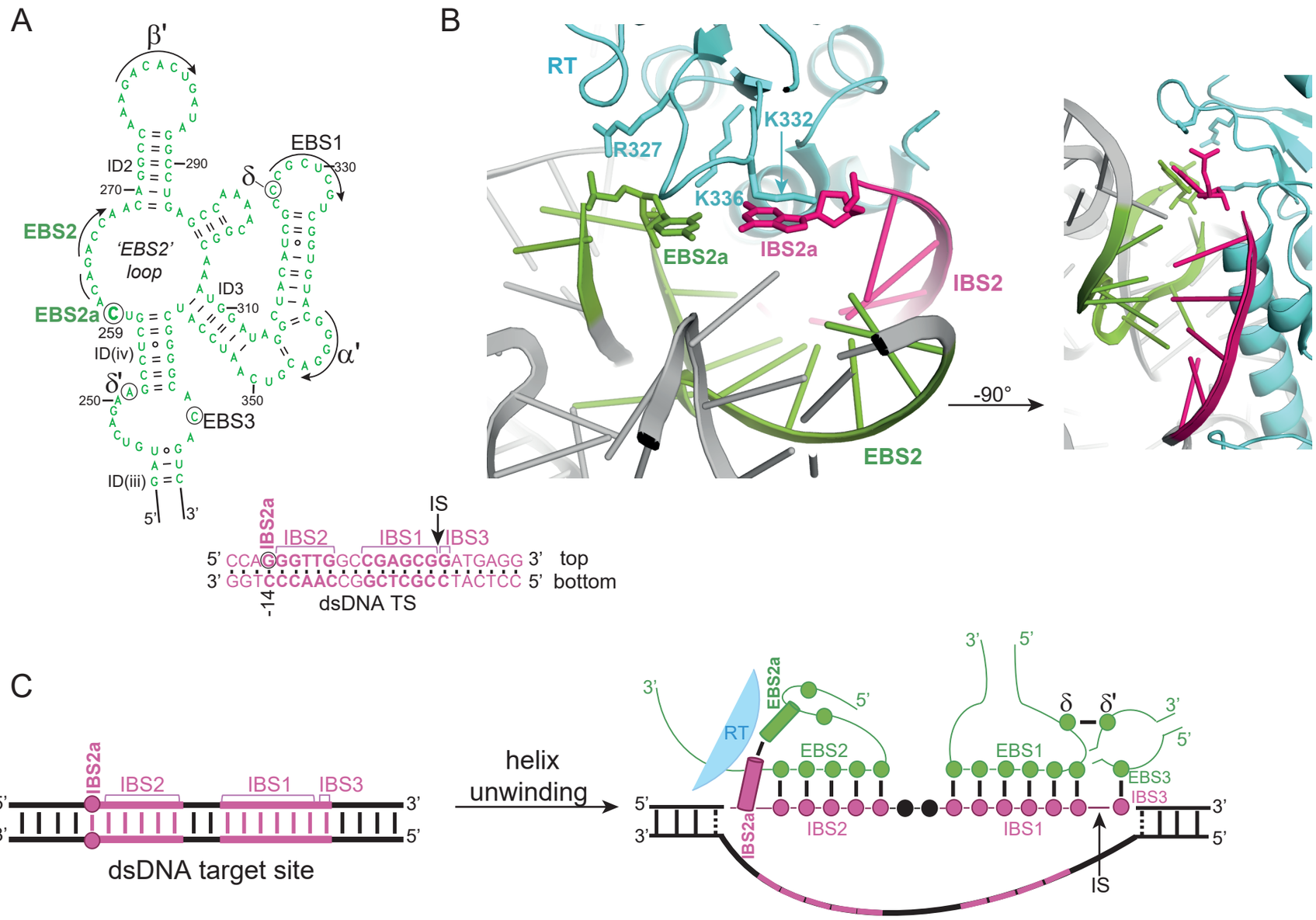


Figure 5


5'exon-IBS2a A A A A A
EBS2a U A G C U
L
P
WT
mut 1
mut 3
mut 6
YAHH
P
226 nt
L
142 nt

In vivo splicing efficiency (%)
100
80
60
40
20
0
WT
mut 1
mut 3
mut 6
YAHH

Figure 6


A
β'
EBS1
ID2
290
330
δ
270
EBS2
'EBS2'
loop
ID3
EBS2a
310
259
ID(iv)
α'
δ'
250
350
EBS3
ID(iii)
5'
3'
IBS2a
IS
IBS2
IBS1
IBS3
5' CCAGGGGTTGGCCGAGCGGATGAGG 3' top
3' GGTCCCCAACCGGCTCGCCTACTCC 5' bottom
-14
dsDNA TS
B
RT
K332
R327
K336
EBS2a
IBS2a
IBS2
-90°
EBS2
C
IBS2a
IBS2
IBS1
IBS3
5'
3'
dsDNA target site
helix
unwinding
RT
EBS2a
EBS2
EBS1
δ
δ'
EBS3
IBS3
IS