Supplementary Data for:

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,*]

Content:

Supplementary Table S1

List of 27 subgroup IIB1 fungal and algal mitochondrial introns that potentially encode a reverse transcriptase

Organism	gene	Accession ID	Intron coordinates	EBS2a and EBS2 (1)
Candida zemplinina	LSU	AY445918	2516-324	UA CACGUA
Juglanconis oblonga	cox1	KY575054	71345-74396	CAGCUAUAA
Monilinia fructicola	cox1	MK163638	99802-102662	CAGCUAUAA
Termitomyces sp. Mi166	cox1	MH725795	12788-15624	CAGCUACAA
Paracoccidioides brasiliensis	cox1	AY955840	41071-43890	UAGCAAUAC
Erysiphe pisi ctg17027	nad2	CACM01017027	3202-458	CAACUAUUC
Coleochaete scutata	cox1	NC_045180	117025-114193	CAGCUAUAG
Halamphora sp. CSP-2018a	cox1	MF997424	274-3197	CAGCGAUAC
Psammoneis japonica	cox1	NC_037989	9425-12339	CAGCGAUAC
Candida frijolesensis	cox1	NC_015814	13095-16036	CAGCGAUAG
Candida parapsilosis	cox1	NC_005253	12690-15605	CAGCGAUAC
Juglanconis oblonga	cox1	KY575054	64525-67280	UACUAGCUA
Tremella fuciformis TF11	LSU	MF422654	34351-37182	UAAUACCAA
Bracteacoccus aerius	LSU	NC_024755	18355-15720	UGAAAACUC
Pilayella littoralis	LSU	Z48620	4052-6489	UGAAAACAA
Ulva pertusa	LSU	NC_035722	29773-32136	UGAAGACUG
Pyropia yezoensis	LSU	NC_017837	1504-3833	AGAAAACAG
Ophiocordyceps sinensis	LSU	NC_034659	680-3620	UAAAAGGUG
Pyronema omphalodes	LSU	NC_029745	163098-166120	UAAAAGGUG
Termitomyces sp. T13	LSU	MH725796	60465-63157	UAAAAGGUG
Pyrrhoderma noxium	LSU	NBII01000013	24138-26853	UAAAAGGUG
Nitella hyalina	LSU	NC_017598	6816-9271	CAAAAGGUG
Pilayella littoralis	LSU	Z48620	543-2952	UAAAAGGUG
Botryococcus braunii	SSU	LT545992	106815-109395	UAAACCUAG
Ulva sp. UNA00071828	LSU	KP720617	33326-35774	AAAAAUUUG
Microbotryum cf. violaceum	LSU	KC285587	60321-57841	UUACAUAGA
Tremella fuciformis TF05	LSU	MF422649	39594-42448	UAAAUACAG

(1) Nucleotides that participate in canonical base pairing are underscored.

[maximal pairwise similarity of alignable sections : 71.2 % of encoded amino acids for Halamphora vs Psammoneis]

Supplementary Table S2

List of 22 subgroup IIB1 algal chloroplast introns that either potentially encode a reverse transcriptase or are closely related to introns that potentially encode a reverse transcriptase.

Organism	gene	Accession ID	Intron coordinates	EBS2a and EBS2 (1)
Gloeotilopsis planctonica	petA	KX306824	76416-77570	UAAUAGUAA(3)
Gloeotilopsis planctonica	psaC	KX306824	191001-193484	AAACGGACA(3)
Gloeotilopsis planctonica	petA	KX306824	77799-80311	UAAGCUUCA
Gloeotilopsis sarcinoidea	ycf3	KX306821	137378-134868	UAAGAGGUA
Gloeotilopsis sarcinoidea	psbH	KX306821	64452-66900	UAAACCUAA
Pedinomonas tuberculata	petA	NC_025530	77423-79780	UCACUCCGA
Gloeotilopsis sarcinoidea	atpF	KX306821	248719-246056	UAACUCGUA
Watanabea reniformis	petB	NC_025526	123866-126876	AAAAUCGUA
Gloeotilopsis planctonica	psbF	KX306824	124715-122798	UAAGGCUAA(3)
Gloeotilopsis sarcinoidea	psaJ	KX306821	210039-207490	UAAGCUGUA(3)
Gloeotilopsis planctonica	ycf12	KX306824	120626-118964	UAACACUAA(3)
Gloeotilopsis planctonica	clpP	KX306824	21722-23158	UUAAGCCAA(3)
Gloeotilopsis planctonica	psbH	KX306824	45724-48746	UCAAAUUAA
Gloeotilopsis planctonica	atpH	KX306824	207894-205242	UAAAGUUAA
Pseudocharacium americanum	cemA	NC_034711	150783-154348	UCAAACGAA
Pseudoneochloris marina	psbB	NC_034710	26551-28940	UUAAGCCAA
Gloeotilopsis sarcinoidea	psbB	KX306821	55589-58170	UUACGCCAA
Capsosiphon fulvescens	psaB	NC_039920	103507-101263	UUAGGAAAA
Rhodochaete parvula	petB	NC_031180	105841-103409	ACAAGCGAA(2)
Hydrodictyon reticulatum	psbA	NC_034655	26465-29026	UGACGAGCU
Pseudobryopsis hainanensis	rbcL	MH591092	11499-14008	UAAUGAUGA
Chlamydomonas subcaudata	psbA	GQ868651	210-4064	AUACCCGGA

(1) The IBS2a site is located at 5’ exon position -13, except for R. parvula, H. reticulatum, P. hainensis and C. subcaudata introns, where it lies at position -14. Nucleotides that participate in canonical base pairing are underscored.
(2) An U was assumed to bulge out in the middle of IBS2
(3) The intron-contained ORF is vestigial, whereas the rest of the sequence is highly similar to that of a protein-encoding intron in the same genome.

[maximal pairwise similarity of alignable sections : 85.8 % of encoded amino acids for G. planctonica petA 77799 vs. G. sarcinoidea ycf3]

Supplementary Table S3

List of 21 cyanobacterial subgroup IIB2 introns that potentially encode a reverse transcriptase

Organism	Accession ID	Intron coordinates	EBS2a and EBS2(1)
Cyanothece sp. PCC 7424	CP001291	1076135-1073508	AAAUAUUUG
Cyanothece sp. PCC 7424	CP001291	3732792-3735419	AAAUAUUUG
Cyanothece sp. PCC 7424	CP001291	4808449-4811196	AAACAUCGA
Cyanothece sp. PCC 7822	CP002198	3119850-3117088	AAAAGUGAU
Cyanothece sp. ATCC 51142	CP000806	415659-418281	UGAUAUGGA(2)
Cyanothece sp. PCC 8801	CP001287	1463534-1466142	UUACGCCGA(2)
Crocosphaera chwakensis CCY0110	AAXW01000034	38411-41184	AAAAACCGA
Lyngbya majuscula	AY652953	7499-4974	UUAGUCCAA
Microcystis aeruginosa NIES-843	AP009552	4471193-4468823	UCACUGAAA
Acaryochloris marina MBIC11017 plasmid pREB6	CP000843	6057-3604	UAAUCAUUA
Nostoc sp. PCC 7120 plasmid pCC7120delta (N.sp.I4)	AP003604	45422-47908	UUAUGCAAA
Calothrix sp. (Cx.sp.I1)	X71404	446–2898	UGAUAAAAC
Anabaena variabilis ATCC 29413 plasmid A	CP000119	23162-25652	UUACGUCGG
Anabaena variabilis ATCC 29413 plasmid C	CP000121	50858-53348	UUACGUCGG
Cyanothece sp. ATCC 51142	CP000806	4828919-4826465	AUACAUAAA
Microcystis aeruginosa PCC 7806	CP020771	1355174-1357793	UAAUUAGCU
Microcystis aeruginosa PCC 7806	CP020771	1654278-1656897	UAAUUAGCU(3)
Microcystis aeruginosa NIES-843	AP009552	510767-513382	UAAACAUUG
Microcystis aeruginosa NIES-843	AP009552	5746618-5749233	UAAACAUUG
Microcystis aeruginosa NIES-843	AP009552	4725373-4722761	UAAAUAUUG(3)
Microcystis aeruginosa NIES-843	AP009552	3033405-3029653	UAAACAUUG

(1) Nucleotides that participate in canonical base pairing are underscored.
(2) The IBS2a site is located at 5’ exon position -13
(3) The IBS2a site is located at 5’ exon position -15

Supplementary Table S4

DNA oligonucleotides used for constructing the wild-type EcI5 donor and recipient plasmids and for analyses of the Tet^R colonies as described in Material and Methods. Sequences in bold indicate the restriction sites used for cloning.

Fw_pACD4K_Cir	5'_GACTAT**AAGCTT**CGGCACGGCCTGATGGAGGCCGCATGTGAGAGGGC GGCCGC**TGTACA**TCACGT
Rev_pACD4K_Cir	5'_ACGTGA**TGTACA**GCGGCCGCCCTCTCACATGCGGCCTCCATCAGGCCG TGCCG**AAGCTT**ATAGTC
Ecl5-5' exon_New	5'_AACCAA**AAGCTT**CCCCTCTAATAGAATTCCCATGCCAAACTGGTGCTCG AAT
Ecl5-3' exon_New	5'_AATATA**CTCGAG**GGTACC**CCCGGG**CTGCAGTCATTTGTGGCGCGGCGA TA
Ecl5 P1_New	5'_**GGATCC**TCTCGTGTTGCCTTTACGATACGC
Ecl5 P3_New	5'_CGTAAAGGCAACACGAGA**GGATCC**TGTAGTGAAACCGGC
Ecl5-5' ORF_New	5'_AATATA**CCCGGG**CCCCGTTTAAACCCTAAGAAGGAGATATACCTATGAC TGAGCAGGCTACAACCTGTAAA
Ecl5-3' ORF_New	5'_AAACCT**CTCGAG**CCAACCTGGAGAACACCGACTAGTCAAGCCTTTCTAA GCCCACTCTCATG
EcI5_T7Fw	5'_ACGGCT**GGTCTC**GGATCCAGTAATACGACTCACTATAGGAGGATCCTG TAGTGAAAC
EcI5_T7Rev	5'_TTAGTT**CCCGGG**CTGCAGTCATTTGTG
Tphi_Fw	5'_AATATA**GGTCTC**ACTTGTGATCCGGCTGCTAACAAAGCCCGA
Tphi_Rev	5'_ATAATT**GGTCTC**ACAAGTATCCGGATATAGTTCCTCCTTTCA
2xT2_Fw	5'_AATATA**TCTAGA**TAAGTAGGTGAGGGTGGCGG
2xT2_Rev	5'_ATAATT**GGTACC**GTTTGTAGAAACGCAAAAAGGCC
3xT2_Fw	5'_AATATA**GGTACC**TAAGTAGGTGAGGGTGGCGG
3xT2_Rev	5'_ATAATT**AAGCTT**GTTTGTAGAAACGCAAAAAGGCC
EcI5_13	5'_CTGATCGATAGCTGAAACGC
EcI5_14	5'_TCGTGTTGCCTTTACGATACG
EcI5_15	5'_AAGCCTATGCCTACAGCATCC

Supplementary Figure S1. Synthetic DNA fragment used for cloning the recipient plasmid.

The synthesized 629-bp DNA carries the *E. coli rrnB* transcription terminators T1 and T2 (delimited by arrowheads and underlined) flanking the natural EcI5 target site (underlined in yellow). T1 also terminates phage T7 RNA polymerase but T2 does not. The IBS sequences at the target site are highlighted: IBS1 and IBS2 in green, IBS3 in gray and the new IBS2a site investigated in this work, in red. The beginning of the *tet*^{R} gene is underlined in blue and the upstream SD sequence (underlined in black) corresponds to phage T7 S10 Shine-Dalgarno. The *tet*^{R} gene is devoid of transcription promoter. The restriction sites BstXI and NheI that were used for cloning this DNA fragment are in bold (see Materials and Methods).

AAACCAACCAAACGTATGGCTGATCGATAGCTGAAACGCCGTAGCGCCGATGGTAGTG
BstXI

TGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTC
<

AGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTA
T1 terminator **>**

GGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGACGTCCTAG

GAGAGGAGAATTCCCATGCCAAACTGGTGCTCGAATCGTATGTATTTTTCTGGTGACTC
2a **2** **1** **3**
< IBS motifs >
GAGAATTCTAAGTAGGTGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCAT

CAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTAGA
< **T2 terminator** **>** XbaI

CCATTCTTGGTACCCCATTCTTAAGCTTAATAATTTTGTTTAACTTTAAGAAGGAGATATA
Acc65I HindIII **SD**

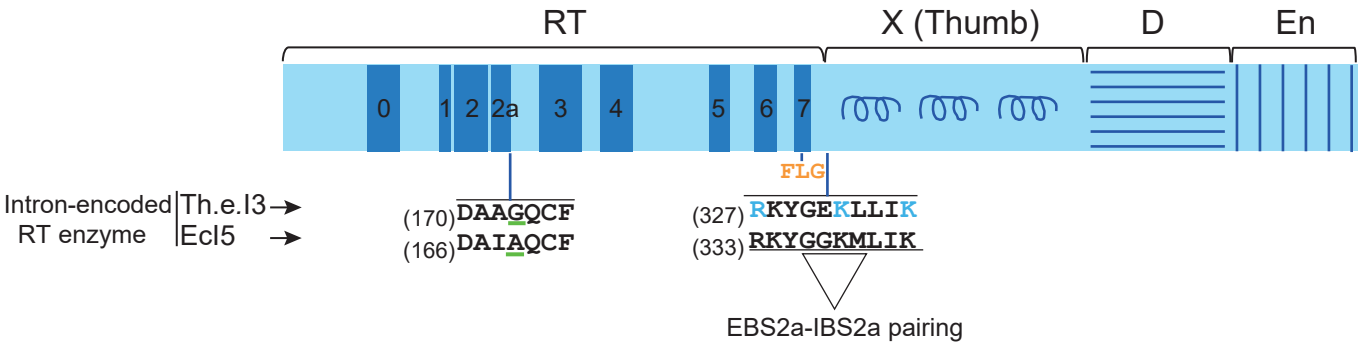
CATATGAAATCTAACAATGCGCTCATCGTCATCCTCGGCACCGTCACCCTGGATGCTGT
< ***tet*^{R}**

AGGCATAGGCTTGGTTATGCCGGTACTGCCGGGCCTCTTGCGGGATATCGTCCATTCC

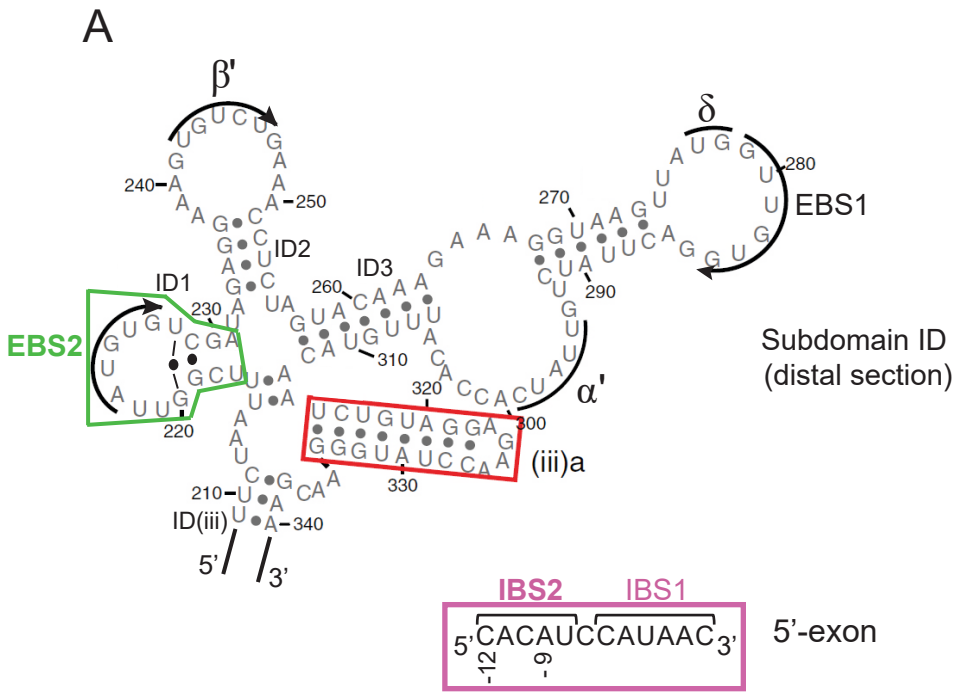
GACAGCATCGCCAGTCACTATGGCGTGCTGCTAGCATATCCC
< NheI >

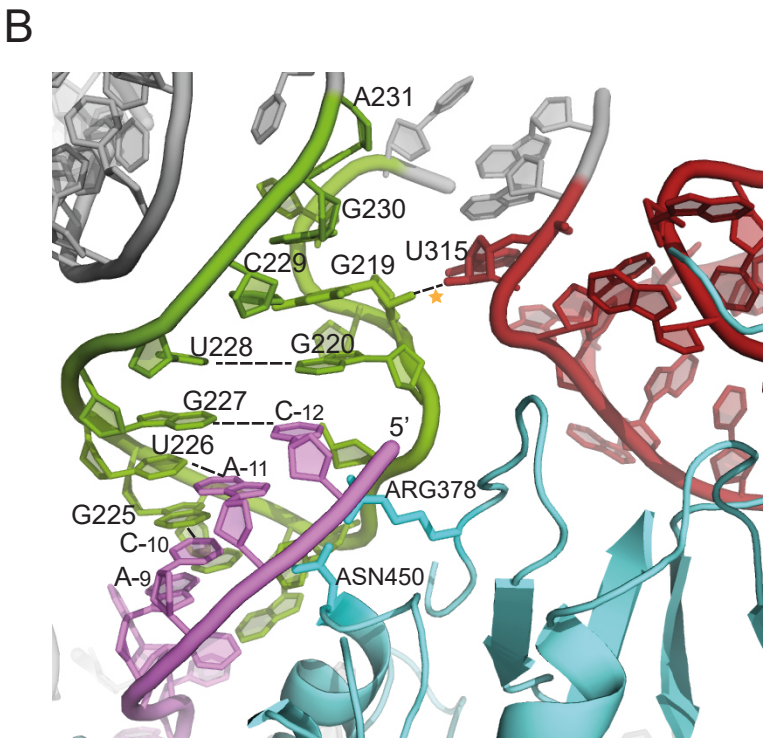
Supplementary Figure S2. General organization of group II intron reverse transcriptases and protein motifs involved in recognition of the DNA target site by subgroup IIB1 and IIB2 introns.

The functional protein domains are, 'RT': reverse transcriptase domain with the different conserved blocks (blue dark sections) of residues typical of group II intron reverse transcriptases; 'X/thumb': maturase domain containing the three alpha helices (dark blue loops) typical of this domain; 'D': DNA-binding domain; 'En': Endonuclease domain. The protein segment involved in stabilization of the EBS2a-IBS2a interaction lies at the boundary between 'RT' and 'X/thumb' domains and is only three amino acids downstream the universally conserved FLG motif (in orange) found in block 7. The numbers between brackets correspond to the numbering of the first residue of the motif shown in the complete sequence of the Th.e.I3 and EcI5 reverse transcriptase enzymes. The amino acids in blue are those directly contacting the EBS2a-IBS2a pairing and labelled in figure 6B. This motif is well conserved in all bacterial IIB1 and IIB2 introns analysed in figure 2. However, IIB2 introns contain extra residues (~16) inserted between the conserved arginine (R) and the first conserved lysine (K) of the motif. RT enzymes encoded by the organellar IIB1 introns analysed in figure 2 have more divergent forms of the motif and carry insertions at the same position as bacterial IIB2 RT enzymes. The motif lying at the end of block 2a harbours the amino acid residue (underlined in green) potentially involved in recognition of position -15 of the DNA target site (see text and Supplementary Figures S4 and S5).



Supplementary Figure S3

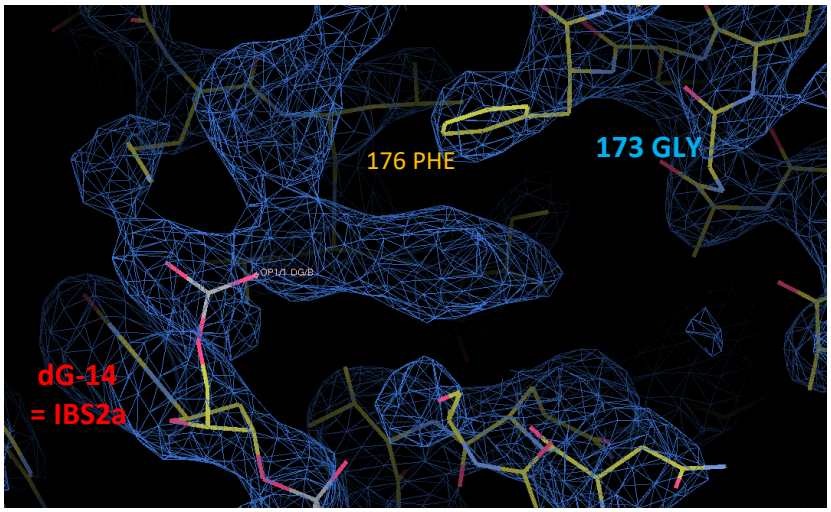




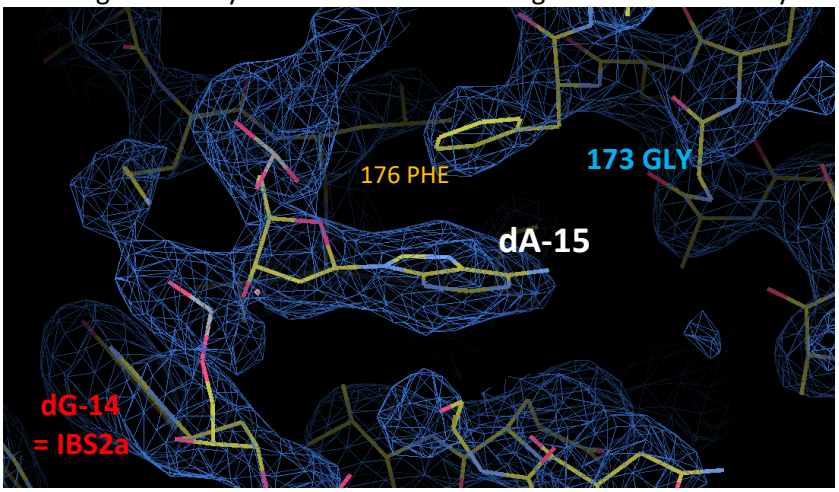
Supplementary

Figure S4

A



Docking of a deoxy-Adenosine in the unassigned electron density



B

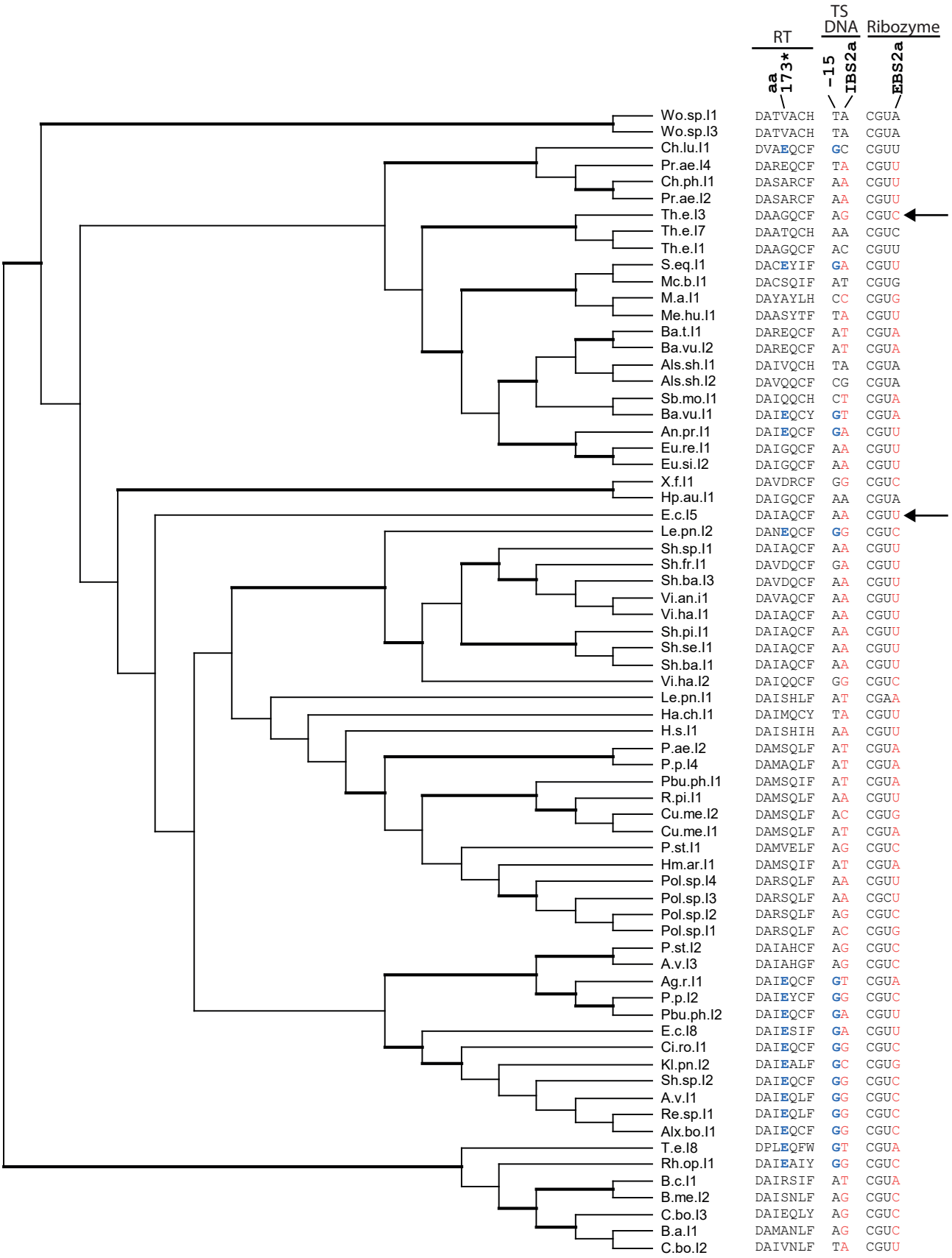
Bacterial subgroup IIB1 introns carrying a RT ORF (69 sequences)

aa at 173* / nt at -15 of DNA TS		15 S	14 A	5 G	21 E	3 D	3 Q	5 V	3 other(T,M,R)
39	**A**	14	13	5	3	1	–	1	2
20	**G**	–	–	–	17	2	1	–	–
7	**T**	1	–	–	1	–	–	4	1
3	**C**	–	1	–	–	–	2	–	–

Bacterial subgroup IIB2 introns carrying a RT ORF (21 sequences)

aa at 173* / nt at -15 of DNA TS		8 S	3 G	8 E	2 other(N,K)
12	**A**	8	3	–	1
5	**G**	–	–	5	–
1	**T**	–	–	1	–
3	**C**	–	–	2	1

Supplementary Figure S5



Supplementary Figure legends S3 to S5

Supplementary Figure S3. Structure and stabilization mode of the all-RNA EBS2-IBS2 interaction between subgroup IIA introns and their 5'-exons. (**A**) Secondary structure of the distal section of subdomain ID of Ll.LtrB intron from *Lactococcus lactis*, a subgroup IIA representative. Two distinctive structural features of subgroup IIA introns are the location of the EBS2 segment, embedded in the terminal loop of ID1 helix (encircled in green), and the presence of helix (iii)a (encircled in red). The IBS-containing section of the 5'-exon of the Ll.LtrB intron is shown encircled in magenta. (**B**) Three-dimensional structure of the EBS2-IBS2 helix formed between the Ll.LtrB intron and its 5'-exon RNA as revealed by the cryo-EM structure of the Ll.LtrB intron in complex with its RT enzyme, LtrA (PDB: 5G2X). The RNA elements are coloured according to the colours of the boxes in (**A**) and the LtrA protein is in cyan. The all-RNA EBS2-IBS2 duplex is stabilized by different types of interactions as follows: base-stacking of pairs (U228-G220 and C229-G219) upon the first EBS2-IBS2 pair (G227-C-12); 2'-OH-mediated interaction between ID1 and (iii)a helixes (indicated by an orange star) and EBS2-IBS2 helix-LtrA protein contacts involving residues ARG378 (at the linker region between RT and X/thumb domains) and ASN450 (at the X/thumb domain). Figure of the cryo-EM structure was prepared with PyMOL (1).

Supplementary Figure S4. Potential for a direct interaction between position -15 of the DNA target site and the intron-encoded reverse transcriptase in bacterial IIB1 and IIB2 introns. (**A**) The top panel shows a close-up view of the cryo-EM map (EMD-9106) of the pre-2r structure (PDB: 6MEC) around DNA target position G-14 (IBS2a). Unassigned density (blue mesh) can be seen above the electron density corresponding to the IBS2a site. This extra density is contiguous to IBS2a density and is surrounded by amino acid residues lying in block 2a of the 'RT' domain, in particular, phenylalanine 176 and glycine 173. The bottom panel shows that a deoxy-Adenosine can be satisfactorily docked into the unassigned density (manual docking using COOT (2)), suggesting that it could correspond to the Th.e.I3 DNA target position A-15. (**B**) The local protein environment of A-15 revealed by the cryo-EM structure drove analyses of DNA target site and intron-encoded RT sequences of bacterial IIB1 and IIB2 introns. The tables show the correlations found between the identity of the base at position -15 of the DNA target and that of the amino acid residue at position 173 of the RT (the asterisk indicates that residue '173' corresponds to the Th.e.I3 protein numbering). The covariation patterns support the existence of a specific interaction between a guanosine at position -15 of the DNA target and a glutamic acid (E) residue at RT position 173*.

Supplementary Figure S5. Phylogeny of bacterial IIB1 introns further supports an interaction between position -15 of the DNA target site and the intron-encoded reverse transcriptase. The phylogenetic tree displayed was generated by PhyML 3.1 (3) from translated ORF sequences of 66

bacterial and 3 archaeal IIB1 introns (see Materials and Methods). Thickened lines correspond to approximate likelihood branch support values equal to, or greater than 0.95. TS DNA: target site DNA. Bases engaged in a Watson-Crick interaction at the EBS2a and IBS2a sites are coloured in red. Co-occurrences of a guanosine (G) at position -15 of the DNA target site and a glutamic acid (E) residue at position 173 of the RT (Th.e.I3 protein numbering) are highlighted in blue. The phylogenetic tree reveals that this specific nucleotide (G)-amino acid (E) combination arose independently several times during evolution of bacterial IIB1 introns, providing further support to a direct interaction between the two elements.

Supplementary References

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

2. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, **66**, 486–501.

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