5238–5246 *Nucleic Acids Research, 2006, Vol. 34, No. 18* *Published online 26 September 2006*
doi:10.1093/nar/gkl688

The proteins encoded by the *pogo*-like *Lemi*1 element bind the TIRs and subterminal repeated motifs of the *Arabidopsis Emigrant* MITE: consequences for the transposition mechanism of MITEs

Céline Loot, Néstor Santiago, Alicia Sanz and Josep M. Casacuberta*

Departament de Genètica Molecular, Laboratori de Genètica Molecular Vegetal CSIC-IRTA, Jordi Girona 18, 08034 Barcelona, Spain

Received May 22, 2006; Revised August 10, 2006; Accepted September 7, 2006

ABSTRACT

MITEs (miniature inverted-repeated transposable elements) are a particular class of defective DNA transposons usually present within genomes as high copy number populations of highly homogeneous elements. Although an active MITE, the *mPing* element, has recently been characterized in rice, the transposition mechanism of MITEs remains unknown. It has been proposed that transposases of related transposons could mobilize MITEs in *trans*. Moreover, it has also been proposed that the presence of conserved terminal inverted-repeated (TIR) sequences could be the only requirement of MITEs for mobilization, allowing divergent or unrelated elements to be mobilized by a particular transposase. We present here evidence for a recent mobility of the *Arabidopsis Emigrant* MITE and we report on the capacity of the proteins encoded by the related *Lemi*1 transposon, a *pogo*-related element, to specifically bind *Emigrant* elements. This suggests that *Lemi*1 could mobilize *Emigrant* elements and makes the *Lemi*1/*Emigrant* couple an ideal system to study the transposition mechanism of MITEs. Our results show that *Lemi*1 proteins bind *Emigrant* TIRs but also bind cooperatively to subterminal repeated motifs. The requirement of internal sequences for the formation of proper DNA/protein structure could affect the capacity of divergent MITEs to be mobilized by distantly related transposases.

INTRODUCTION

Transposable elements (TE) can be divided into two classes according to their structure and transposition mechanism. Class 1 elements transpose by a replicative mechanism involving an RNA molecule that is reverse transcribed before integration, while class 2 elements are mobilized by a cleavage and strand-transfer mechanism usually known as 'cut and paste'. Although the enzymes required for transposition can be encoded by the mobile element itself, non-autonomous defective elements that can be mobilized in *trans* exist for both classes of elements. Autonomous DNA elements encode transposases that are able to bind to the terminal sequences of the element and that catalyse the cleavage and strand-transfer reactions. As most non-autonomous DNA elements are mutation derivatives of their autonomous counterparts, they usually show sequence similarity to them and can often be mobilized by the related transposases.

MITEs (miniature inverted-repeated transposable elements) are usually classified as non-autonomous DNA transposons because they share structural characteristics with these elements. MITEs contain terminal inverted-repeated (TIR) sequences and do not have any coding capacity. The existence of putative transposons sharing extensive sequence similarities to some MITE families and potentially coding for transposases has led to propose that MITEs could be deletion derivatives of DNA transposons that are mobilized in *trans* by these elements (1,2).

Most MITEs have been characterized by computer-assisted searches and, until very recently, the proposal of a precise mechanism of transposition of these elements has been prevented by the lack of an actively transposing MITE. The characterization of a rice MITE named *mPing*, whose transposition is induced in anther-derived cell cultures (3) leaving excision footprints behind (4), confirmed that MITEs can transpose by a 'cut and paste' mechanism typical of DNA transposons. Nevertheless, although *mPing* has extensive sequence similarities to a DNA transposon potentially coding for a transposase of the *mariner* superfamily, the *Ping* element, the rice varieties where *mPing* was mobilized do not contain any complete *Ping* element that could account for the mobilization of these elements. For this reason it was proposed that a different DNA transposon, *Pong*, which is

*To whom correspondence should be addressed. Tel: +34 93 4006142; Fax: +34 93 2045904; Email: jcsgmp@ibmb.csic.es

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

only distantly related to *Ping*/*mPing* elements, could be the source of transposase (5). Different phylogenetic analysis indeed suggested that transposases not directly related to a particular MITE family could be responsible for the mobilization of these elements (6,7). Moreover, although it has not yet been proved that the capacity of any DNA transposase (either of the same or of a different family) can mobilize a MITE copy, a recent report shows that rice *mariner*-like transposases can bind *in vitro* and in yeast assays to the TIRs of non-related *Stowaway* MITEs (8). The presence of similar TIR sequences could thus be the only requirement for interaction with transposases non-directly related to the MITE, in line with the hypothesis that MITEs could be mobilized by distantly related transposases (9).

The *Emigrant* MITE (10) is present in some 500 copies in the genome of *Arabidopsis* and other *Brassicae*. A phylogenetic analysis has shown that *Arabidopsis* contains different subfamilies of the *Emigrant* elements consistent with the amplification of one or a few 'master' copies at different times during the evolution of these species (11). The *Columbia* ecotype of *Arabidopsis* contains a single-copy *pogo*-related element called *Lemi*1 that has extensive sequence homology with *Emigrant* elements from which the latter was probably derived by internal deletion (12).

Here we show that some of the *Emigrant* insertions are polymorphic among different *Arabidopsis* ecotypes or even among different individuals of the same ecotype, suggesting that they have transposed in a recent past and that *Arabidopsis* probably contains an enzymatic activity capable of mobilizing these elements. We also show that the proteins encoded by *Lemi*1 specifically bind to *Emigrant* TIRs, suggesting that *Lemi*1 could provide the transposase mobilizing *Emigrant* elements. Moreover, we show that *Lemi*1 proteins cooperatively bind to subterminal repeated motifs, suggesting that the internal sequences of MITEs, and not only the TIR sequences, could also be important for transposase interaction and mobilization of some of these elements.

MATERIALS AND METHODS

Amplification and cloning of *Emigrant* and *Lemi*1 sequences

For the analysis of *Emigrant* polymorphisms, the *loci* containing *Emigrant* insertions were amplified by PCR with primers corresponding to flanking genomic sequences. For *Emi*158, the primers were e158-5′ (5′-CCATATTCACAATTTTAC-3′) and e158-3′ (5′-GCTTAAATAAATAGAAAGAG-3′).

*Lemi*1 sequences of different *Arabidopsis ecotypes were amplified by PCR using the 372 (5′-CTCTGTCTTTGATCCACA-3′) and 1616 (5′-GGTCCTATTAGTTCATCTG-3′) primers. PCR products were cloned into the pCRII-TOPO vector (Invitrogen) and sequenced. The sequences were aligned using the Genedoc program.*

PCR products obtained by amplification of *Arabidopsis Columbia*-0 genomic DNA with the primers 372–1616 and 372–1972 (1972, 5′-CCATTTTATATCAGGATAGTTATA-3′) were cloned into the pTZ57R vector (Fermentas) to generate the pLemi1 and pLemi3 plasmids (containing the *orf*1 and the region comprising the two *orf*s, respectively). The codon stop interrupting *orf*1 was removed from both constructs by PCR site-directed mutagenesis using the 497W (5′-CGATTTAAAAGTA**TGG**CTTGAAG-3′) and 520W (5′-CTTCAAG**CCA**TACTTTTAAATCG-3′) primers. PCR products were cloned into the pTZ57R vector to produce the pLemi0 and pLemi4 plasmids. The intron of *orf1-2* was removed by PCR from pLemi4 using partially complementary primers (WI5′: 5′-CTGATCCAGGCTATCGCAATAGAATGG-3′ and WI3′: 5′-CGCTATCGGACCTAGTCTAACAGGG-3′) that span the donor and acceptor splicing sites. The PCR product obtained was cloned into the pTZ57R vector to generate the pLemi5 plasmid. During the PCR amplification of the *Lemi*1 sequences, a mutation was introduced leading to the formation of a stop codon and therefore generating a truncated protein of 264 amino acids corresponding to the entire DNA-binding domain and a truncated catalytic domain. The PCR product was cloned into the pTZ57R vector to obtain the pLemi2 plasmid.

Protein production

For gel retardation assays, the *Lemi*1 proteins were produced as fusion proteins linked to the glutathione *S*-transferase (GST) using the pGEX-KG system.

The *Lemi*1 sequences were amplified by PCR from pLemi0, pLemi1, pLemi2 with the EcoRI-390 and 1684-SacI primers (5′-CCGAATTCCAACGATGGCGTCTC-3′ and 5′-GGCTCGAGAGATAACGCTATCGG-3′, respectively) and from pLemi5 using the EcoRI-390 and 1949-SacI primers (5′-GGCTCGAGTTTATATCAGGATAGTTATAG-3′). The PCR products were digested by EcoRI and SacI and ligated into the pGEX-KG plasmid, resulting in pGEX0, pGEX1, pGEX2 and pGEX5. pGEXCo-1 was generated by replacing an AccI–AccI fragment of pGEX5 by the fragment containing the 372–1616 *Lemi*1 sequence obtained by PCR from *Arabidopsis Coimbra*-1 genomic DNA.

Electrophoretic mobility shift assays (EMSAs)

The region of *Arabidopsis Columbia*-0 genomic DNA containing the *Emi*126 element was amplified with the primers e126-5′ (5′-CAGCGATAACATTTGATCTTC-3′) and e126-3′ (5′-GCATGTATCTTAAACCATTG-3′) and cloned in the pTZ57R vector (Fermentas) to generate the pEmi1 plasmid. This plasmid was used to obtain the *Emi*126 TIR1 and *Emi*126 TIR2 probes, as well as a non-specific (NS) competitor DNA fragment. The *Emi*126 TIR1 probe was produced by PCR amplification using the 5′-TIR1C1 and 3′-TIR1C1 primers (5′-ATTAAATCATTGACGAAGAAC-3′ and 5′-CCAAATTAGGAAAATTTCTC-3′, respectively), and the *Emi*126 TIR2 probe was obtained using the e126 5′-TIR2 and e126 3′-TIR2 primers (5′-GAAGAATTTATTAATTTATAGAGG-3′ and 5′-CAAAAATATCATGACAATCCG-3′, respectively). The non-specific competitor DNA was generated by PCR on the same pEmi1 plasmid with primers flanking the *Emi*126 insertion: e126c and e126-3′ (5′-TTAAATGATATTATCGATATTC-3′ and 5′-GCATGTATCTTAAACCATTG-3′, respectively). The PCR products were cloned into the pTZ57R vector (Fermentas) to generate, respectively, the pEmi6, pEmi3 and pEmi4 plasmids. The *Emi*126 TIR1 probe (99 bp of the *Emi*126 3′-terminal sequence plus 69 bp of flanking genomic sequence), the Emi126 TIR2 probe (64 bp of *Emi*126 5′-terminal sequence

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

plus 105 bp of flanking genomic sequence) and the NS probe (169 bp of genomic sequences flanking Emi126) were generated by SacI–BamHI digestion of the pEmi6, pEmi3 and pEmi4 plasmids, respectively, and radioactively labelled, when necessary, with [α-^{32}P]dCTP using Klenow polymerase (Roche) by standard procedures.

The *Emi*126-L probe was produced by EcoRI–HindIII digestion of the pEmi3 plasmid to give a fragment of 209 bp containing 64 bp of Emi126 and 146 bp of flanking sequences.

The *Emi*158 locus was amplified by PCR on *Arabidopsis Columbia*-0 genomic DNA using the e158-5′ (5′-CCATATT-CACAATTTTAC-3′) and e158-3′ (5′-GCTTAAATAAATA-GAAAGAG-3′) primers, and cloned in pTZ57R to generate the pEmi158 plasmid. This plasmid was used to produce the *Emi*158 TIR1 and *Emi*158 TIR2 probes by PCR using, respectively, the e158-5′TIR1 and e158-3′TIR1 primers (5′-TTTGAAAAGTTCTTTATTTATAT-3′, 5′-TTA-TAAATTGATAAATATTAATTT-3′) and the e158-5′TIR2 and e158-3′TIR2 primers (5′-GAAGAATTTATTAATTTAT-AAAAG-3′, 5′-ATGCTTAAATAAATAGAAAGAG-3′). The PCR products were cloned into the pCRII-TOPO vector (Invitrogen) to obtain, respectively, the pEmi8 and pEmi9 plasmids. The TIR1 and TIR2 *Lemi*1 terminal sequences and flanking regions were amplified by PCR on *Arabidopsis Columbia*-0 genomic DNA using the *Lemi*1-5′TIR1 and *Lemi*1-3′TIR1 primers (5′-CCATACATCAAACATAGCTT-ATAC-3′ and 5′-CTTTCAAAAATCTGAAAACCAAAAA-TTC-3′) and the *Lemi*1-5′TIR2 and *Lemi*1-3′TIR2 primers (5′-GAAGAATTTATTAATTTAGAGAGG-3′ and 5′-GTC-AATTTGTCGAAAAAATTTACAATC-3′), respectively. The PCR products were cloned into the pCRII-TOPO vector (Invitrogen) to generate the pEmi10 and pEmi11 plasmids. The probes *Emi*158 TIR1 (99 bp of *Emi*158 sequence and 88 bp of flanking genomic DNA), *Emi*158 TIR2 (64 bp of *Emi*158 sequence and 123 bp of flanking DNA), *Lemi*1 TIR1 (106 bp of *Lemi*1 sequence and 93 bp of flanking genomic DNA) and *Lemi*1 TIR2 (64 bp of *Lemi*1 sequence and 134 bp of flanking DNA) were obtained by EcoRV–BamHI digestion of pEmi8, pEmi9, pEmi10 and pEmi11 plasmids and were radioactively labelled, when necessary, with [α-^{32}P]dCTP using Klenow polymerase (Roche) by standard procedures.

EMSAs were performed by incubating 30, 60 or 120 ng proteins with 1 μg poly(dI–dC), 1 μg BSA in binding buffer (25 mM HEPES, pH 7.6, 40 mM KCl, 2 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT and 10% glycerol) for 10 min on ice. Radioactively labelled DNA probe (1 ng, 20 000 c.p.m.) was added to the mixture and the incubation on ice continued for another 20 min. In competition assays, the reaction was pre-incubated with a 5- or 50-fold molar excess of ice-cold competitor DNA. The assays were resolved in a 4% polyacrylamide gel.

DNase I footprinting assays

Samples of the EMSA reactions were digested by 0.05 U of DNase I (Roche) for 1 min at room temperature. The enzyme was diluted in dilution buffer (16 mM $MgCl_2$ and 8 mM $CaCl_2$). Reactions were stopped using STOP buffer (50 mM Tris–HCl, pH 8, 0.1 M NaCl and 0.5% SDS). DNA was purified by phenol–chloroform extraction and ethanol precipitation. The cleavage pattern was analysed by electrophoresis on a 6% polyacrylamide sequencing gel. DMS/piperidin reactions were performed following standard procedures to reveal G positions and were used to localize the DNAse I protected regions.

Bioinformatic analysis

The sequences flanking *Emigrant* elements in *Arabidopsis Columbia*-0 ecotype were used to search the database of *Landsberg erecta* sequences delivered by Monsanto (13) and available on the TAIR website (*arabidopsis*.org/Cereon/index.jsp). The comparison of the *Columbia*-0 and *Landsberg erecta* sequences allowed the detection of *Emigrant* insertion/deletion polymorphisms.

RESULTS AND DISCUSSION

Some *Emigrant* elements transposed recently in *Arabidopsis*

A genome-wide analysis of *Arabidopsis* (*Columbia* ecotype) showed that this genome contains different subfamilies of the *Emigrant* family of MITEs generated by the burst owing to amplification that occurred at different times during the evolution of this genome (11). The *Emi*A subfamily groups young *Emigrant* elements while the *Emi*0 subfamily probably contains the oldest ones (11). In order to look for evidence of recent mobility, we amplified 10 regions containing five *Emi*A and five *Emi*0 insertions in the *Columbia* ecotype from DNA obtained from 14 *Arabidopsis* ecotypes by using PCR. Two out of five *Emi*A insertions (*Emi*126 and *Emi*158) were found to be polymorphic whereas none of the *Emi*0 showed insertion polymorphism. The amplification of the *Emi*158 region gave two bands in the *Coimbra*-4 ecotype (Figure 1A, lane 6), suggesting that individuals of this ecotype are polymorphic for the *Emi*158 insertion. The PCR analysis of 10 different individuals indeed revealed that some *Coimbra*-4 individuals contain the insertion, some do not contain the insertion and some are heterozygous for the insertion (Figure 1B).

As an important fraction of the genome of the *Landsberg erecta Arabidopsis* ecotype has been sequenced, we looked for possible insertion polymorphisms of the whole *Emigrant* population in this ecotype. Of the 158 genomic regions searched, we found the corresponding *Landsberg erecta* sequence of 78 (55%). The analysis of these regions revealed that eight *Emigrant* insertions (10.25%) are polymorphic when comparing *Columbia* and *Landsberg erecta* genomes (data not shown). Seven of these polymorphic *Emigrant* insertions belong to the *Emi*A subfamily confirming that this subfamily groups the youngest *Emigrant* elements (data not shown).

The results presented here suggest that *Emigrant* elements were mobilized recently during *Arabidopsis* evolution and that some of their insertions have not had time to become fixed in this genome.

The sequencing of the empty sites revealed in two cases the presence of extra nucleotides coinciding with that of *Emigrant* TIRs and/or short deletions of sequences

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

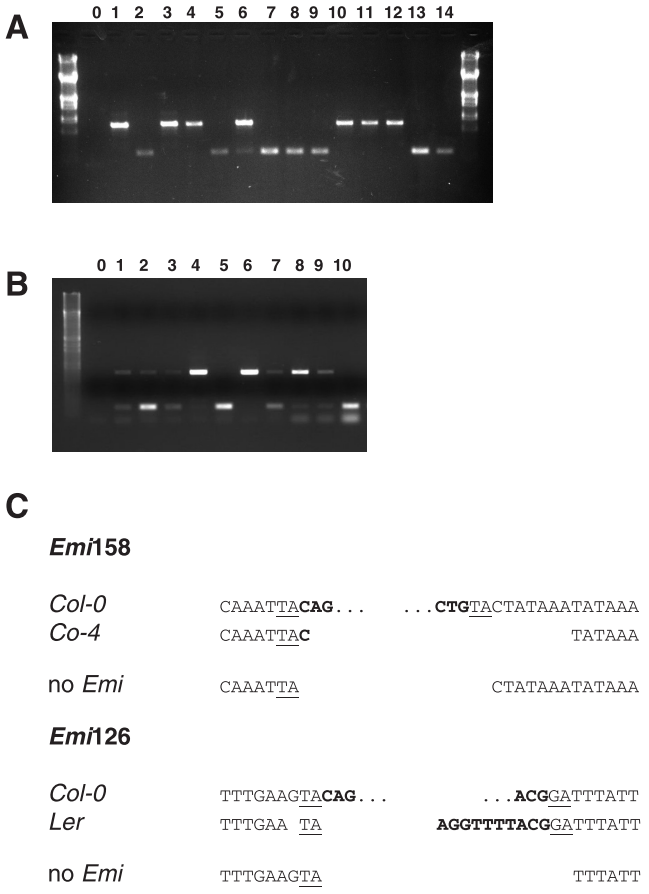

Figure 1. *Emigrant* insertion and excision polymorphisms. (**A**) PCR amplification with oligonucleotides flanking the *Emi*158 insertion from water (0) or DNA from the following *Arabidopsis* ecotypes: 1, *Columbia*; 2, *Landsberg*; 3, *RLD*; 4, *Wassilewskija*; 5, *Canterbury*-1; 6, *Coimbra*-4; 7, *Dijon*-G; 8, *Estland*; 9, *Geneva*-0; 10, *Kashmir*-1; 11, *Moscow*; 12, *Niederzenz*-1; 13, *Tsu*-0; 14, *Nossen*. (**B**) PCR using DNA from 10 different *Coimbra*-4 individuals. (**C**) Comparison of the sequences *loci* corresponding to two polymorphic *Emigrant* insertions. The name of the polymorphic elements, as well as the name of the ecotypes compared, is shown on the left. The sequence of the theoretical empty site is shown below the sequences for comparison.

flanking *Emigrant* insertions (Figure 1C). These sequence differences to the expected empty site probably represent excision footprints and suggest that although most of the polymorphisms detected are probably due to differential *Emigrant* insertions, a few of them were probably generated by *Emigrant* excisions. Our results thus suggest that in the recent past the genome of *Arabidopsis* contained an enzymatic activity that was able to excise and reinsert *Emigrant* MITEs.

Analysis of the *Emigrant*-related *pogo*-like transposon *Lemi*1 in different *Arabidopsis* ecotypes

It has been proposed that *Emigrant* elements originated by a severe deletion of a putative *pogo*-like transposon called *Lemi*1 that shows extensive sequence homology with *Emigrant* elements (12). *Lemi*1 is present only in one copy of the *Arabidopsis Columbia* ecotype, and displays an *orf* potentially coding for a *pogo*-like transposase interrupted by a STOP codon in position 39 of the protein and a frameshift in position 385 of the protein [(12) and Figure 2A]. It has been proposed that the splicing of a putative intron could allow overcoming the frameshift (12). The sequence of the putative donor and acceptor splicing sites (data not shown and Figure 2B, respectively) perfectly fit the consensus for plant introns (14) and a consensus branch point sequence is found at the correct distance from the acceptor AG (data not shown). In order to get insight on the original structure of the *Lemi*1 element we have analysed the *Lemi*1 sequences of different *Arabidopsis* ecotypes. *Lemi*1 is present as a single-copy element in all the *Arabidopsis* ecotypes that we have analysed, and we have not been able to obtain evidences of mobility of *Lemi*1 in those genomes (data not shown). Using primers complementary to internal *Lemi*1 sequences we have amplified, sequenced and compared *Lemi*1 sequences obtained from seven different *Arabidopsis* ecotypes. In spite of the high similarity found (from 94 to 100% identical over 1245 bp), the *Lemi*1 sequences are polymorphic at particularly important positions (Figure 2B). The *Lemi*1 sequence does not contain the STOP codon found in *Columbia* in four different ecotypes, *Ms-0*, *RLD* and *Dijon*-G, and is probably present in two different alleles (only one of them containing the STOP) in the *Tsu*-0 ecotype, as we have obtained two types of sequences when amplifying this region from this ecotype. On the other hand, the *Lemi*1 sequence has an insertion of 56 bp in the region of the frameshift that restores the coding capacity in a single *orf* in *Coimbra*-1 and *Coimbra*-4 (Figure 2B). The presence of this insertion in *Lemi*1-related sequences of *Gossypium hirsutum*, *Solanum demisum* and *Medicago truncatula* (data not shown) suggests that the original *Lemi*1 element had a single *orf* of 1559 bp. Nevertheless, *Coimbra*-1 and *Coimbra*-4 also have a difference with respect to the *Lemi*1 sequences found in all other ecotypes at one of the two invariable nucleotides of the putative splicing acceptor site (Figure 2B). The presence of an acceptor splicing consensus site exclusively in the *Lemi*1 sequences where the *orf* is interrupted by a frameshift (Figure 2B) could also indicate the functionality of a spliced protein.

The proteins encoded by *Lemi*1 specifically and cooperatively bind *Lemi*1 and *Emigrant* sequences

Transposase binding to the terminal regions of the mobile element is the first step of the transposition process. In order to test the capacity of the proteins encoded by the original *Lemi*1 transposon to bind *Lemi*1 and *Emigrant* sequences, we reconstructed the consensus *Lemi*1 coding sequence by replacing the STOP codon found in the *Columbia Lemi*1 sequence with the tryptophan coding triplet found in *Ms-0*, *RLD*, *Dijon-D* and *Tsu*-0 ecotypes by site-directed mutagenesis. We expressed in *Escherichia coli* the two proteins encoded by the modified *Lemi*1 element as GST fusions: the protein encoded by the first *orf* (GST–Orf1) and the protein that would be produced by splicing the predicted *Lemi*1 intron and consisting in a fusion of most of Orf1 and Orf2 (GST–Orf1-2) (Figure 3). On the other hand, we generated a construct containing the insertion found in *Coimbra*-1 and *Coimbra*-4 (GST–Orf1-2+) by replacing an AccI–AccI fragment (see Figure 2A) of the *Columbia* sequence by that

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

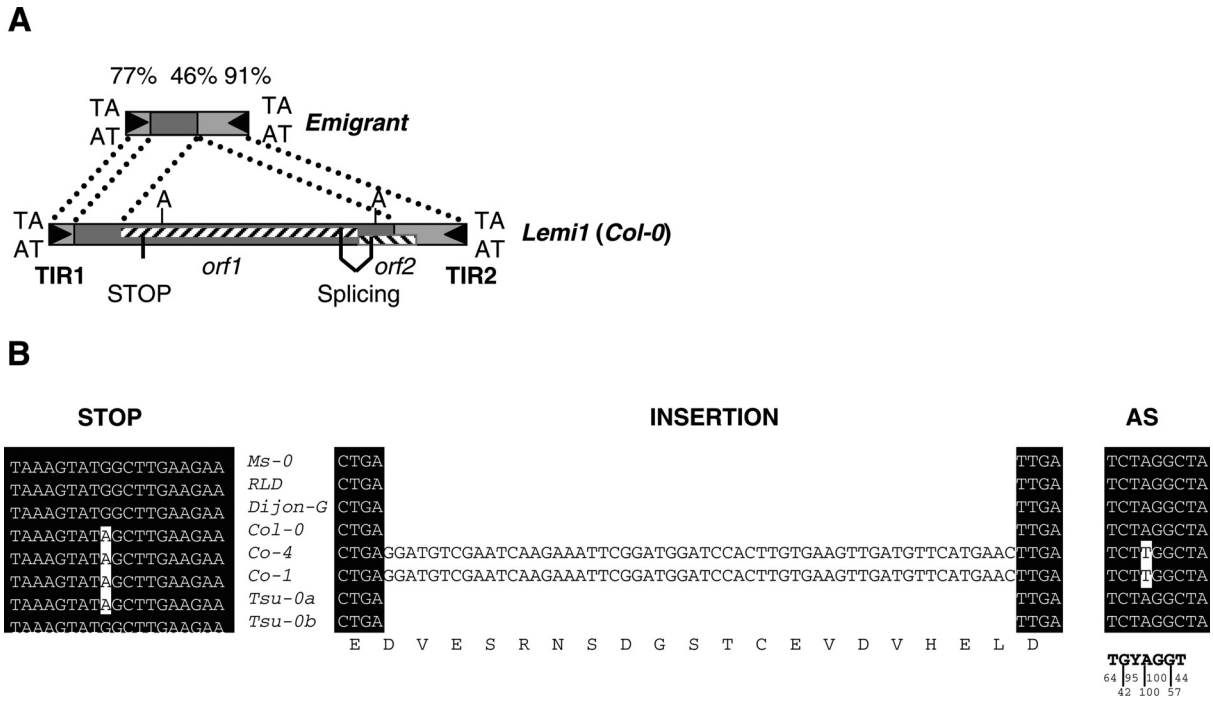

Figure 2. Structure and sequences of the *Lemi*1 elements from different *Arabidopsis* ecotypes. (**A**) The structure of the *Columbia Lemi*1 element is shown compared with that of a consensus *Emigrant* element. The percentage of identity between both sequences in the common regions is shown on the top of the *Emigrant* scheme. The two *orfs* of *Lemi*1, as well as the position of the stop codon interrupting the *orf*, and of the putative intron are shown. The position of the two AccI (A) used for cloning (see Materials and Methods) is shown. (**B**) Comparison of the DNA sequences of three polymorphic regions (the region containing the stop codon in Columbia, STOP; the region containing an insertion in Co-1 and Co-4 ecotypes, INSERTION; and the region of the putative acceptor splicing site, AS) of *Lemi*1 from seven different *Arabidopsis* ecotypes: *Moscow* (*Ms*-0), *RLD*, *Dijon*-G, *Columbia* (*Col-0*), *Coimbra*-4 (*Co*-4), *Coimbra*-1 (*Co*-1) and two alleles of *Tsu*-0. The accession numbers of the sequences are DQ888704, DQ888705, DQ888706, At2g06660, DQ888708, DQ888709, DQ888710 and DQ888711, respectively. The polypeptide encoded by the insertion sequence is shown on the bottom of the corresponding sequence. The plant consensus sequence of the splicing acceptor site sequence (numbers below sequences represent the percentage occurrence of indicated bases) is shown underneath the corresponding sequences.

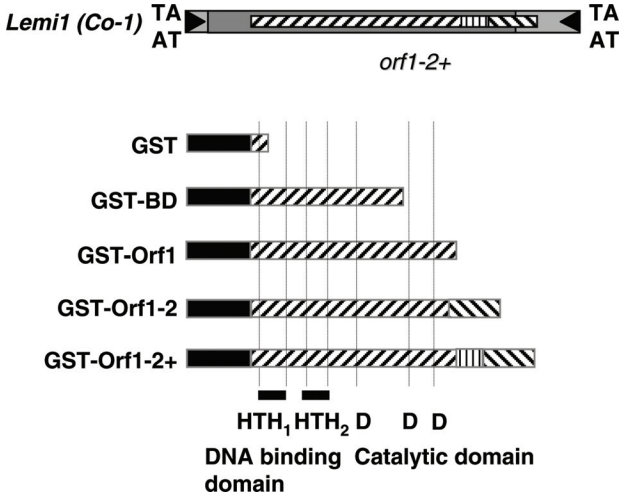

Figure 3. *Lemi*1 protein constructs. The schemes representing the different fusion proteins used are shown compared with the structure of the *Lemi*1 element found in the *Coimbra*-1 ecotype. The position of the two HTH motifs of the putative DNA-binding domain and the three conserved Aspartic residues of the putative catalytic domain are indicated.

of the *Coimbra*-1. As control proteins, we expressed in *E.coli* GST fusions with truncated *Lemi*1 proteins: the GST construct, in which the STOP codon of the *Columbia Lemi*1 sequence was maintained and that encodes a GST protein fused to a short polypeptide of 38 amino acids, and the GST–BD construct, in which a STOP codon was introduced at position 264 of the protein and that encodes for a GST protein fused with a truncated Orf1 protein that contains the whole DNA-binding domain and a truncated catalytic domain. These proteins were used to perform EMSA with radioactively labelled probes corresponding to the terminal sequences of *Lemi*1 and *Emigrant* elements. We first tested the ability of *Lemi*1 proteins to bind its own terminal sequences. EMSA analysis showed that the GST–Orf1 protein binds *Lemi*1 TIR1 and TIR2 probes giving one major retarded band, B1 (TIR1), or three retarded bands, B1–B3 (TIR2) (Figure 4), suggesting that the reconstructed protein has retained the ability to specifically bind the transposon TIR sequences. In order to test if *Lemi*1 proteins could also specifically bind the TIRs of *Emigrant* elements we performed EMSA analysis with the TIRs of the *Emi*126 element, a polymorphic *Emigrant* element belonging to the young *Emi*A family (11). These analyses showed that *Lemi*1 proteins specifically bind *Emigrant* TIRs (Figure 5). Although the control GST protein does not bind to *Emi*126 probes, the GST–Orf1, GST–Orf1-2 and GST–Orf1-2+ proteins specifically bind to both TIR1 and TIR2 *Emi*126 probes. Binding to TIR1 gave two retarded bands (B1 and B2, Figure 5A), while binding to TIR2 gave three retarded bands (B1, B2 and B3, Figure 5B). The three proteins tested seem to bind *Emigrant* probes in a similar way, suggesting that the

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

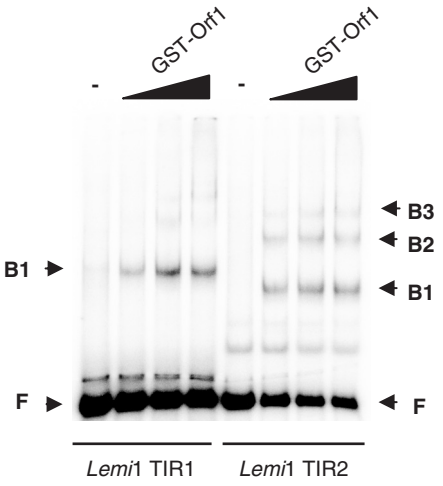

Figure 4. *Lemi*1 proteins binding to *Lemi*1 TIRs. Increasing concentrations of the GST–Orf1 protein were incubated with radioactively labelled probes corresponding to the 5′- and 3′-terminal sequences of *Lemi*1 (*Lemi*1 TIR1: 106 bp of *Lemi*1 sequence and 93 bp of flanking genomic DNA; *Lemi*1 TIR2: 64 bp of *Lemi*1 sequence and 134 bp of flanking genomic DNA) and were analysed by EMSA. The migrating position of the free probes (F) and the different retarded bands (B1–B3) is shown on both sides of the panel.

C-terminal part of the catalytic domain, which coincides with the polypeptide encoded by *orf2*, does not participate in the binding of *Lemi*1 protein(s) to DNA. The GST–BD polypeptide binds *Emigrant* TIR1 and TIR2 sequences but, although its binding of TIR1 is very similar to that of GST–Orf1, GST–Orf1-2 and GST–Orf1-2+ (Figure 5A), its binding to TIR2 appears to be very different. While GST–Orf1, GST–Orf1-2 and GST–Orf1-2+ proteins gave three retarded bands with TIR2, GST–BD gave only one retarded band even at high-protein concentrations (Figure 5B, lanes 5–7). Therefore, although the DNA-binding domain of *Lemi*1 protein(s) is sufficient for specific binding, a region of the catalytic domain absent from the GST–BD protein is needed for multiple binding to TIR2. A suggestive possibility is that this region, which is absent from the GST–BD polypeptide and is located within the catalytic domain, could mediate protein–protein interactions allowing multiple binding to DNA. Transposase dimerization domains can be located within the DNA-binding domain, e.g. in the *mariner*-like *Sleeping Beauty transposase* (15), in the C-terminal part of the protein, in many *hAT* transposases such as *Hermes* (16), or within both the DNA-binding domain and the catalytic domain similar to the one in the case of the *mariner*-like *Mos*1 transposase (17–19).

*Lemi*1 proteins bind to the *Emigrant* TIR and to subterminal repeated motifs

A key step of transposition is the formation of a precise DNA/protein structure that requires transposase dimerization and allows DNA cleavage and strand-transfer reactions (20). In most cases this structure consists of a synaptic complex in which the transposase catalyses the reaction in *trans* to the transposon end at which it is bound, but transposases can also catalyse the reaction in *cis*, as it has been shown for the mariner *Himar*1 element (21). Transposase dimerization is essential to form active complexes, and transposase dimerization can take place in *cis*, with a single transposase unit bound to the DNA (21), or between transposases bound to the DNA side-by-side, or in *trans* to form paired end complexes (17). In order to determine whether the different retarded bands obtained with *Emigrant* TIR1 and TIR2 probes are the result of multiple protein units bound to a single DNA molecule or of complexes containing multiple DNA molecules, we have performed EMSA analysis in the presence of competitor molecules of different sizes. The competition of GST–Orf1 binding to *Emi*126 TIR2 probe with two different unlabelled *Emi*126 TIR2 fragments of different sizes did not reveal any difference in the mobility of retarded bands that could be an indication of the presence of multiple DNA fragments in protein/DNA complexes (Figure 6). Thus, the retarded bands obtained in EMSA experiments are probably the result of the binding of multiple *Lemi*1 proteins to a single DNA molecule.

We performed DNase I footprinting analysis to determine the *Lemi*1-binding sites in *Emi*126. These experiments showed that GST–Orf1-2 binds *Emigrant* TIRs but also other internal sequences (Figure 7). In the case of TIR2 the DNase I protection covers a continuous region of 56 bp including the TIR and two repeated motifs that coincide with the 3′-half of the TIR sequence (Figure 7B and see Figure 5B for sequence details). The TIR2 footprint is flanked by a DNase I hypersensitive band indicating that binding of *Lemi1* protein(s) induce(s) a distortion of the target DNA. Protein binding often affects DNA structure and in particular transposases often distort DNA upon binding (22). The DNase I footprinting analysis of TIR1 shows a protection that covers two regions: 23 bp of the TIR itself and a 29 bp region consisting of two repeats of a sequence coinciding with the 3′-half of the TIR in opposite orientation and separated from the TIR by 22 bp (Figure 7A and see Figure 5A for sequence details).

These experiments show that multiple binding to *Emi*126 TIR1 and TIR2 sequences revealed by EMSA experiments is the result of recognition of the TIR itself and the subterminal repeated sequences by *Lemi*1 proteins. Binding to one, two or three of these sequences could explain the different retarded bands obtained in EMSA. Nevertheless, although *Lemi*1 proteins gave three retarded bands in EMSA with the TIR2 probe, they gave only two with the TIR1 probe (Figure 5). This suggests that GST–Orf1, GST–Orf1-2 and GST–Orf1-2+ proteins can simultaneously bind the three binding motifs present in TIR2 while they bind, but not simultaneously, the three binding motifs found in TIR1. As there are no major differences in the binding motifs found in both TIRs, the different binding should be explained by the different arrangement of the binding motifs in both TIRs. Indeed, the TIR and the two subterminal motifs are contiguous and in the same orientation in TIR2 while in TIR1 the subterminal repeats are found in reverse orientation and separated from the TIR sequence (Figure 5). The need for an internal region of the catalytic domain of *Lemi*1 proteins for multiple binding to TIR2 suggests that protein–protein interactions play an important role in *Lemi*1 binding and different arrays of DNA-binding motifs would probably modify the protein–protein interactions that can take place.

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

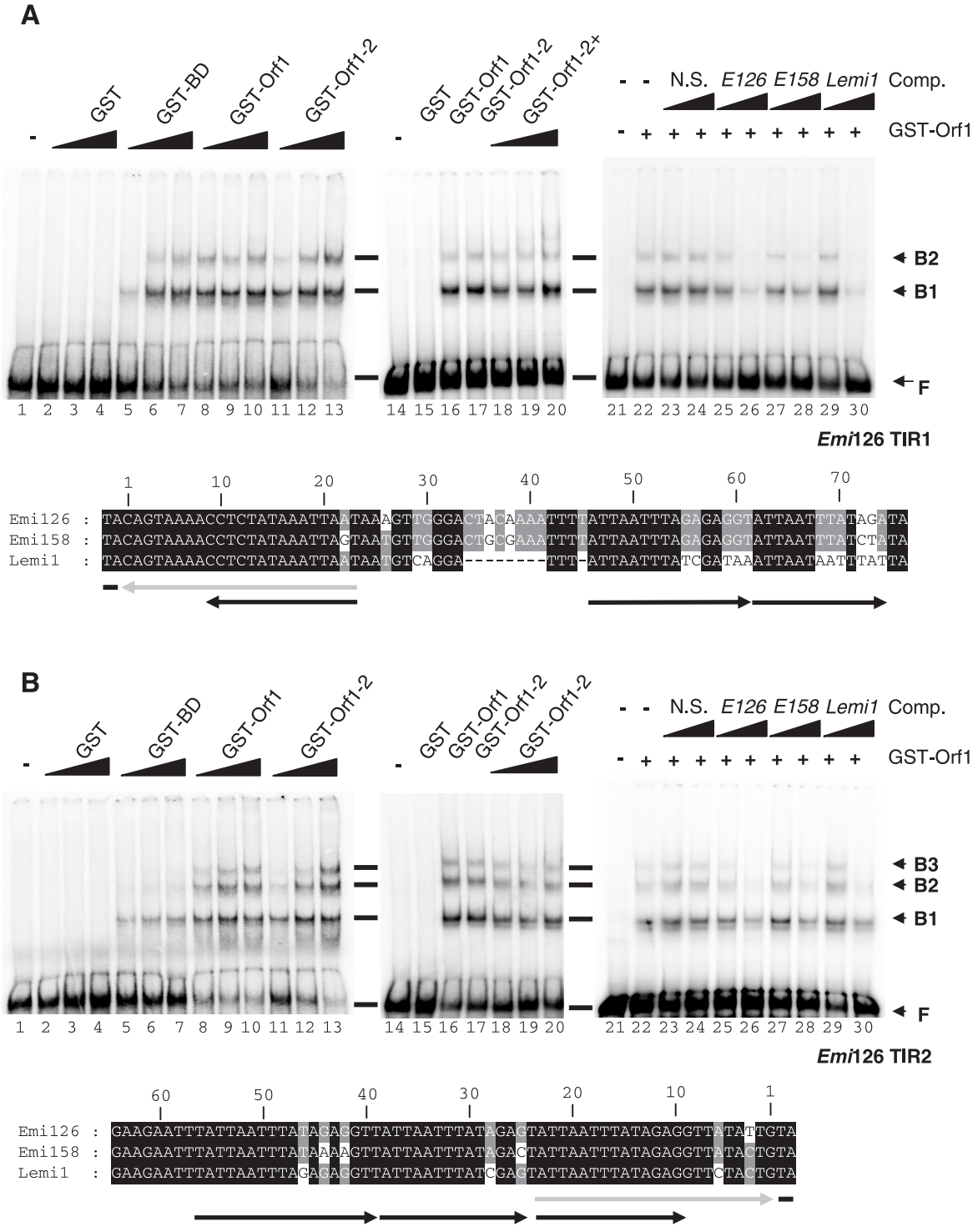

Figure 5. DNA-binding analysis of the different *Lemi*1 proteins to the *Emi*126 TIRs. Increasing concentrations of the indicated proteins (left and middle panels) or a fixed concentration of GST–Orf1 protein (right panel) where incubated with radioactively labelled probes corresponding to *Emi*126 TIR1 (99 bp of the 5′-terminal sequence of *Emi*126 element plus 69 bp of the flanking genomic sequence) (**A**) or to *Emi*126 TIR2 (64 bp of the 3′-terminal of *Emi*126 plus 105 bp of the flanking genomic sequence) (**B**), and were analysed by EMSA. Control reactions with no recombinant protein (−) were also included as controls. Competition experiments (right panel) were performed by including in the reaction mixture increasing concentrations (ratios: 1/5 and 1/50) of ice-cold non-specific competitor (NS, 169 bp of genomic sequence flanking *Emi*126 insertion), *Emi*126 TIR1(A) or TIR2 (B) fragments, *Emi*158 TIR1 (99 bp of the 3′-terminal of *Emi*158 plus 88 bp of the flanking genomic sequence) (A) or *Emi*158 TIR2 (64 bp of the 3′-terminal of *Emi*158 plus 123 bp of the flanking genomic sequence) (B), *Lemi*1 TIR1 (106 bp of the 3′-terminal of *Lemi*1 plus 93 bp of the flanking genomic sequence) (A) or *Lemi*1 TIR2 (64 bp of the 3′-terminal of *Lemi*1 plus 134 bp of the flanking genomic sequence) (B). The migrating position of the free probes (F) and the different retarded bands (B1–B3) is shown on the right. An alignment of the region comprising the TIR and subterminal sequences of the *Emi*126, *Emi*158 and *Lemi*1 is shown on the bottom. The position of the TIR and subterminal sequences is indicated by grey and black arrows. The position of the target site duplication is shown by a solid line. Nucleotides are numbered from the first nucleotide of the TIR.

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

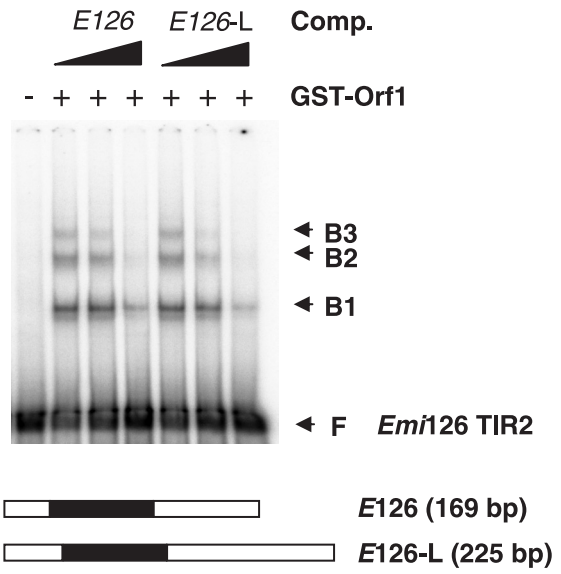

Figure 6. Competition experiments with *Emi*126 TIR2 fragments of different sizes. GST–Orf1 protein was incubated with Emi126 TIR2 probe in the presence of increasing concentrations (ratios: 1/1, 1/10 and 1/30) of ice-cold *Emi*126 TIR2 (*E*126) or a DNA fragment containing the same *Emi*126 region flanked by longer genomic and plasmid DNA sequence (*E*126-L). The migrating position of the free probe (F) and the different retarded bands (B1–B3) is shown on the right. A scheme representing *Emi*126 TIR2 (*E*126) and *E*126-L used as probe and/or competitors is shown on the bottom. *Emi*126 sequences are shown as a closed boxes whereas genomic and plasmid DNA sequences are shown as open boxes.

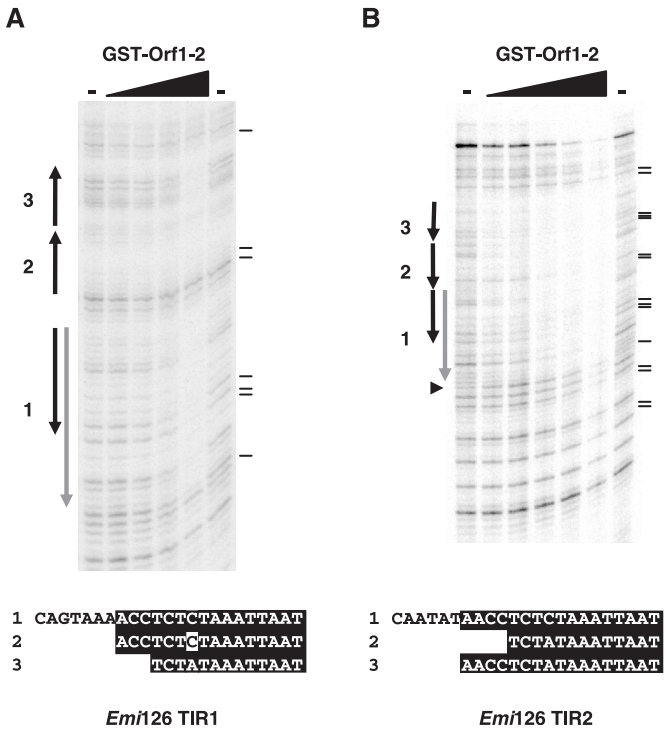

Figure 7. Footprinting analysis of the *Lemi*1 proteins binding to *Emi*126. Increasing concentrations of the GST–Orf1-2 were incubated with radioactively labelled *Emi*126 TIR1 (**A**) and TIR2 (**B**) probes. The position of the TIR sequence is shown by a grey arrow and the position of the TIR and subterminal repeated motifs are indicated by solid arrows. The position of G nucleotides of the upper strand (TIR2) or the bottom strand (corresponding to C nucleotides in the upper strand) (TIR1), revealed by DMS reactions, are shown on the right of each gel. G (or C) positions, numbered from the first nucleotide of the TIR, as in Figure 5, are 1, 9, 10, 12, 34, 37 and 80 (TIR1), and −23, −22, −13, −12, −6, 1, 9, 10, 12, 25, 27, 41, 42, 44, 61 and 64 (TIR2). A closed triangle indicates the DNase I hypersensible position flanking the protected region in TIR2. A sequence comparison of the TIRs and the repeated motifs is shown at the bottom. Identical nucleotides appear as white letters on black boxes.

The existence of subterminal repeats with a capacity to bind the transposase and required for proper activity has also been recently reported for some rice *mariner*-like transposases (8) as well as other *mariner*/Tc1 elements such as *Sleeping Beauty* (15) or *pogo* (23), and for transposons of other families, such as the CACTA and *hAT* families [reviewed in (24)]. The cooperative binding of *Lemi*1 protein(s) to *Emigrant* subterminal motifs indicates that this can also be the case for MITEs, suggesting that the TIRs could not always be the only requirement for transposase binding to MITE sequences and transmobilization.

*Lemi*1 proteins bind differently to *Lemi*1 and *Emigrant* sequences

In order to test the capacity of *Lemi*1 proteins to bind different *Emigrant*-related sequences and analyse the importance of subterminal repeated motifs for binding, we performed binding and competition experiments with different sequences. In addition to *Emi*126, we analysed the binding of *Lemi*1 proteins to *Emi*158, another polymorphic *Emigrant* element (see Figure 1) that has well-conserved TIR and subterminal repeated motifs and *Lemi*1, which has consensus TIR sequences but presents important differences in the subterminal regions of TIR1 (Figure 5). The competition experiments with ice-cold *Emi*126, *Emi*158 and *Lemi*1 probes show that both *Emi*158 and *Lemi*1 bind *Lemi*1 proteins with an efficiency similar to that of *Emi*126 (Figure 5A and B, right panels). This suggests that TIR itself is the primary determinant for Lemi1 binding. Thus, *Emigrant*-related elements that have diverged in their subterminal regions can efficiently bind *Lemi*1 proteins, similarly to what has been recently found for the interaction of *Stowaway* MITEs with related but distinct transposon families (8). However, EMSA analysis using *Lemi*1 TIRs as probes showed that, although elements that have diverged in their subterminal repeats can bind *Lemi*1 proteins, their binding is different. Indeed, although Lemi1 binding to the well-conserved *Lemi*1 TIR2 is similar to that of *Emi*126, the binding to *Lemi*1 TIR1 is different. Indeed, Lemi1 binding to *Lemi*1 TIR1 produces only one major retarded band (Figure 4) instead of the two obtained with *Emi*126 TIR1 (Figure 5). This suggests that *Lemi1* proteins only bind *Lemi*1 TIR motif itself and not the divergent subterminal motifs found in this sequence. These results thus show that multiple binding to *Emigrant*-related elements depends on the conservation of the sequence and the relative arrangement of the subterminal repeated motifs. Thus, although *Lemi*1 proteins could efficiently bind the TIRs of divergent *Emigrant* elements that have conserved the TIRs, their final DNA–protein structure will greatly depend on the number and the relative arrangement of the subterminal repeated motifs. It has been shown that the formation of a proper nucleoprotein complex, which can depend on the

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020

binding of the transposase to subterminal repeated motifs, is a key regulatory step in transposition [reviewed in (25)]. Moreover, in some cases, such as for the *Sleeping Beauty* transposon, transposase binding to a subterminal repeat greatly enhances transposition (15). Interestingly, the subterminal repeat of *Sleeping Beauty* consists of the 3′-half of the TIR, similarly to what we describe here for the *Lemi*1/*Emigrant* elements. We thus propose that the conservation of the subterminal repeated motifs could modify the potentiality of a particular *Emigrant*-related element to be mobilized by the *Lemi*1-encoded transposase.

CONCLUSIONS

The results presented here show that the *Arabidopsis Emigrant* MITE has transposed in the recent past and that the protein(s) encoded by the *pogo*-like *Lemi*1 element specifically and cooperatively bind *Emigrant* elements. This suggests that *Lemi*1 could mobilize *Emigrant* elements and makes the *Lemi*1/*Emigrant* couple an ideal system to study the transposition mechanism of MITEs. Our results show that the sequence of the TIR itself is the primary determinant for binding. But our results also show that, once the binding to the TIR is accomplished, *Lemi*1 proteins can also bind subterminal repeated motifs. These results thus show that, at least for the *Emigrant*/*Lemi*1 system, the final protein/DNA structure depends on transposase binding to subterminal repeated motifs, and suggests that the conservation of internal sequences could influence the ability of a particular MITE to be mobilized by a related transposase.

ACKNOWLEDGEMENTS

The authors wish to thank Michael Joulie for his help in the analysis of *Lemi*1 copy number. We are grateful to Elena Casacuberta, Lluïsa Espinás, Inmaculada Hernández-Pinzón, Paloma Mas and Soraya Pelaz for their critical reading of the manuscript. This work was funded by the Spanish Ministerio de Ciencia y Tecnología (Grant BIO2003-01211) and the Centre de Referència en Biotecnologia (CeRBa) from the Generalitat de Catalunya. C.L. is recipient of an I3P Postdoctoral contract from the Spanish Ministerio de Ciencia y Tecnología. Funding to pay the Open Access publication charges for this article was provided by Ministerio de Ciencia y Tecnología (Grant BIO2003-01211).

Conflict of interest statement. None declared.

REFERENCES

1. Feschotte,C., Zhang,X. and Wessler,S.R. (2002) Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II*. American Society for Microbiology Press, Washington DC, pp. 1147–1158.
2. Casacuberta,J.M. and Santiago,N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, **311**, 1–11.
3. Kikuchi,K., Terauchi,K., Wada,M. and Hirano,H.Y. (2003) The plant MITE mPing is mobilized in anther culture. *Nature*, **421**, 167–170.
4. Nakazaki,T., Okumoto,Y., Horibata,A., Yamahira,S., Teraishi,M., Nishida,H., Inoue,H. and Tanisaka,T. (2003) Mobilization of a transposon in the rice genome. *Nature*, **421**, 170–172.
5. Jiang,N., Bao,Z., Zhang,X., Hirochika,H., Eddy,S.R., McCouch,S.R. and Wessler,S.R. (2003) An active DNA transposon family in rice. *Nature*, **421**, 163–167.
6. Feschotte,C., Swamy,L. and Wessler,S.R. (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics*, **163**, 747–758.
7. Jiang,N., Feschotte,C., Zhang,X. and Wessler,S.R. (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.*, **7**, 115–119.
8. Feschotte,C., Osterlund,M.T., Peeler,R. and Wessler,S.R. (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. *Nucleic Acids Res.*, **33**, 2153–2165.
9. Zhang,X., Jiang,N., Feschotte,C. and Wessler,S.R. (2004) PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics*, **166**, 971–986.
10. Casacuberta,E., Casacuberta,J.M., Puigdomenech,P. and Monfort,A. (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterization of the Emigrant family of elements. *Plant J.*, **16**, 79–85.
11. Santiago,N., Herraiz,C., Goni,J.R., Messeguer,X. and Casacuberta,J.M. (2002) Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **19**, 2285–2293.
12. Feschotte,C. and Mouches,C. (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.*, **17**, 730–737.
13. Jander,G., Norris,S.R., Rounsley,S.D., Bush,D.F., Levin,I.M. and Last,R.L. (2002) *Arabidopsis* map-based cloning in the post-genomic era. *Plant Physiol.*, **129**, 440–450.
14. Lorkovic,Z.J., Wieczorek Kirk,D.A., Lambermon,M.H. and Filipowicz,W. (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci.*, **5**, 160–167.
15. Izsvak,Z., Khare,D., Behlke,J., Heinemann,U., Plasterk,R.H. and Ivics,Z. (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *J. Biol. Chem.*, **277**, 34581–34588.
16. Hickman,A.B., Perez,Z.N., Zhou,L., Musingarimi,P., Ghirlando,R., Hinshaw,J.E., Craig,N.L. and Dyda,F. (2005) Molecular architecture of a eukaryotic DNA transposase. *Nature Struct. Mol. Biol.*, **12**, 715–721.
17. Auge-Gouillou,C., Brillet,B., Germon,S., Hamelin,M.H. and Bigot,Y. (2005) Mariner Mos1 transposase dimerizes prior to ITR binding. *J. Mol. Biol.*, **351**, 117–130.
18. Richardson,J.M., Dawsonm,A., O'Hagan,N., Taylor,P., Finnegan,D.J. and Walkinshaw,M.D. (2006) Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J.*, **25**, 1324–1334.
19. Zhang,L., Dawson,A. and Finnegan,D.J. (2001) DNA-binding activity and subunit interaction of the mariner transposase. *Nucleic Acids Res.*, **29**, 3566–3575.
20. Lohe,A.R., Sullivan,D.T. and Hartl,D.L. (1996) Subunit interactions in the mariner transposase. *Genetics*, **144**, 1087–1095.
21. Lipkow,K., Buisine,N., Lampe,D.J. and Chalmers,R. (2004) Early intermediates of mariner transposition: catalysis without synapsis of the transposon ends suggests a novel architecture of the synaptic complex. *Mol. Cell. Biol.*, **24**, 8301–8311.
22. Watkins,S., van Pouderoyen,G. and Sixma,T.K. (2004) Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res.*, **32**, 4306–4312.
23. Wang,H., Hartswood,E. and Finnegan,D.J. (1999) Pogo transposase contains a putative helix–turn–helix DNA binding domain that recognizes a 12 bp sequence within the terminal inverted repeats. *Nucleic Acids Res.*, **27**, 455–461.
24. Kunze,R. and Weil,C.F. (2002) The hAT and CACTA superfamilies of plant transposons. In Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II*. American Society for Microbiology Press, Washington DC, pp. 565–610.
25. Gueguen,E., Rousseau,P., Duval-Valentin,G. and Chandler,M. (2005) The transpososome: control of transposition at the level of catalysis. *Trends Microbiol.*, **13**, 543–549.

Downloaded from https://academic.oup.com/nar/article-abstract/34/18/5238/3112142 by Universitat Autonoma de Barcelona user on 17 January 2020