Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky1196

Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling

Marco Catoni [1,2,*], Thomas Jonesman [1], Elisa Cerruti[1] and Jerzy Paszkowski[1,*]

[1]The Sainsbury Laboratory, University of Cambridge, Cambridge, CB2 1LR, UK and [2]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Received October 01, 2018; Revised November 08, 2018; Editorial Decision November 11, 2018; Accepted November 16, 2018

ABSTRACT

Pack-TYPE transposons are a unique class of potentially mobile non-autonomous elements that can capture, merge and relocate fragments of chromosomal DNA. It has been postulated that their activity accelerates the evolution of host genes. However, this important presumption is based only on the sequences of currently inactive Pack-TYPE transposons and the acquisition of chromosomal DNA has not been recorded in real time. Analysing the DNA copy number variation in hypomethylated Arabidopsis lines, we have now for the first time witnessed the mobilization of novel Pack-TYPE elements related to the CACTA transposon family, over several plant generations. Remarkably, these elements can insert into genes as closely spaced direct repeats and they frequently undergo incomplete excisions, resulting in the deletion of one of the end sequences. These properties suggest a mechanism of efficient acquisition of genic DNA residing between neighbouring Pack-TYPE transposons and its subsequent mobilization. Our work documents crucial steps in the formation of *in vivo* novel Pack-TYPE transposons, and thus the possible mechanism of gene shuffling mediated by this type of mobile element.

INTRODUCTION

Autonomous transposable elements (TEs) encode all factors required for transposition, whereas non-autonomous elements react to factors provided *in trans* (1). In DNA transposons, terminal inverted-repeat sequences (TIRs) of transposition-competent non-autonomous TEs are recognized by a transposase of a related autonomous element (2). The extremities of Pack-TYPE elements include terminal inverted repeats (TIRs), while their internal sequences are derived from host chromosomes (3–6). For example, Pack-MULEs are Pack-TYPE elements that populate the rice genome (7), with ∼3000 Pack-MULEs carrying segments of ∼1000 different genes (3). This heterogeneous population of transposons has had a major impact on the current organization of rice chromosomes and the evolution of rice genes (7).

Structures resembling Pack-TYPE transposons are also present in maize (3), Arabidopsis (5,8), wheat (9), soybean (10) and morning glory (4), and include TIRs deriving from both *Mutator*-like (MULE) and *En/Spm* (also known as CACTA) transposon families. However, MULE elements are characterized by relatively longer TIRs and Target Site Duplication (TSD), and therefore Pack-MULEs are better predicted in genome wide annotation studies (3,7). Although several models explaining how Pack-TYPE TEs could acquire chromosomal DNA have been proposed (3), they all remain very speculative, because the mobilization of these transposons has never been observed in real time.

Epigenetic suppression mediated by DNA methylation is one of the main reasons why TEs are immobile (11). Elimination of DNA methyltrasferase1 (MET1) activates the genome-wide transcription of TEs by erasing CpG methylation (12). Importantly, the absence of CpG methylation persists over many plant generations, even after restoration of MET1 activity (13–15). This observation prompted the construction of epigenetic recombinant inbred lines (epiRILs) from a cross between two isogenic parents: the wild-type and a *met1* mutant deficient in CpG methylation (16). F2 progeny with MET1 activity were further propagated by single-seed descent for eight generations, resulting in 68 epiRILs with mosaic methylation patterns (16). Previous studies in epiRILs reported a high-mobilization activity of the autonomous LTR transposon EVADE' and the DNA element CACTA1 (16–18); however, a genome-wide survey on mobile transposons in these plant lines is still missing.

Here, we screened for TEs mobilized in epiRILs, and we identified a new family of Pack-TYPE mobile elements with CACTA1-derived TIRs. The study of their real time mobilization suggested a new model of gene shuffling mediated

*To whom correspondence should be addressed. Tel: +44 121 41 46485; Email: m.catoni@bham.ac.uk
Correspondence may also be addressed to Jerzy Paszkowski. Tel: +48 728 195 841; Email: jurek@paszkowski.com

by this type of transposons, based on combination of previously described TE proprieties.

MATERIALS AND METHODS

Plant material

The *met1–3* derived epiRIL population was described previously (16). From this population, the ninth inbred generation was grown to produce the 67 epiRILs used in this study. Col-0 and second generation *met1–3* plants (15) were used as controls. Unless indicated otherwise, plants were grown in soil under long-day conditions (21°C, 16 h light, 8 h dark).

DNA extraction and library preparation

Arabidopsis seeds were sown on $\frac{1}{2}$ MS (Plant) medium plates and grown for 14 days before use. Approximately 10 seedlings per epiRIL were harvested (∼100 mg fresh weight) and DNA extracted using the Qiagen DNeasy Plant Mini Kit following the manufacturer's instructions (Qiagen N.V., Hilden, Germany). Before library preparation, DNA was fragmented to an average size of 350 bp using 24 cycles of 30 s with a Bioruptor Diagenode sonication device. The genomic DNA sequencing libraries were prepared using the TruSeq DNA polymerase chain reaction (PCR)-Free LT Library Prep Kit following the manufacturer's instructions (Illumina, San Diego, Calif.), starting from 1.1 μg of fragmented DNA. Libraries were validated using High Sensitivity D1000 ScreenTape on a 2200 Tapestation instrument (Agilent technologies, Santa Clara, CA) and a LightCycler 480 Instrument II using the LightCycler 480 SYBR Green I Master mix (Roche, Basel, Switzerland).

DNA sequencing, reads mapping and peak call

The DNA libraries were sequenced with 2 × 76-bp paired-end reads with a minimum of 10× coverage (24× on average) on an Illumina NextSeq 500 using the High-Output Flow Cell configuration. The raw reads were trimmed using Trimmomatic to remove adapter sequences. Reads with an averaged value of at least 15 in a 4-nt window were trimmed from both ends. After trimming, reads pairs with at least one mate shorter that 36 bp were discarded. The remaining sequences (on average 95% of raw reads) were aligned with bowtie2 (19) with the –no mixed and –non deterministic option, against the reference TAIR10 Arabidopsis genome (www.arabidopsis.org). Metrics of the sequencing analysis are displayed in Supplementary Table S1. Genome coverage was calculated for each genome position using genomecov (bedtools v2.26.0) and normalized to the Arabidopsis genome and library size. Arabidopsis genomic regions with a coverage of more than 2.5-fold or <0.2-fold of the average genome coverage in the Col-0 wild-type control (4 250 297 bp, corresponding to 3% of Arabidopsis genome) were masked to avoid calling peaks in regions with high copy number or insufficient coverage (e.g. plastids DNA, ribosomal repeats, microsatellites). Peak call was done for each epiRIL mapped bam file in comparison to the Col-0 control condition with macs2 (https://github.com/taoliu/MACS/), setting the fragment size to 75 and extended size to 400 bp, with the options –B, -SPMR and –no model and Columbia-0 as control. The Log based background subtraction was performed with macs2 bdgcmp with 'logLR' as model and a *P*-value threshold of 0.00001.

The peak lists obtained from all epiRILs were merged and peaks closer than 5 kb were joined together in a gff file using R (https://www.R-project.org/). Then, HTseq (20) was used to count reads at each peak for all conditions analysed and the raw count was normalized to peak length and library size to obtain FPKM values. Fold change of DNA coverage was calculated for each peak by the ratio between FPKM values in the tested condition and the Col-0 wild-type control. Peaks overlapping TEs (TAIR10 annotation) were further filtered to contain only regions with more than 1.8-fold change difference in at least one epiRIL.

Detection of TE insertions

New putative TE integrations were detected in each epiRIL with the Transposon Insertion Finder (TIF) tool (21), using the Arabidopsis TAIR10 genome as reference. An ad-hoc Linux bash script was designed to recursively run TIF for all sequenced epiRILs and the Col-0 control plants, using the first and last 17 bp of each TE tested as input (code reported as Supplementary Data S1). The new putative insertion loci obtained from the analysis were imported in *R* and filtered to exclude false positive calls in Col-0 sequenced samples.

PCR and cloning

To confirm new transposon insertions, PCRs were carried out using a transposon-specific primer and a primer flanking the new insertion. Amplification of the entire locus with new Pack-CACTA1a integration was performed using two primers designed on the sequences flanking the new insertion. All PCR reactions were carried out using GoTaq enzyme (Promega, Madison, WI, USA), with extension time adjusted to the expected size of the fragment amplified, following the recommended manufacturer's instructions. All PCR products were extracted from the gel using a Qiagen Gel Extraction Kit and eluted in 30 μl of water. The purified PCR products were ligated into a pGEM-T plasmid (Promega) following the manufacturer's instructions and 2.5 μl of ligated sample used to transform *Escherichia coli* DH5α cells (50 μl). Colony PCRs were performed starting from overnight grown *E. coli* colonies using the universal M13 forward and reverse primers. The amplified products were SANGER sequenced by Sigma-Aldrich (Merck, Darmstadt, Germany). All primers were designed using Geneious v9.1.2. (https://www.geneious.com), and their sequences are listed in Supplementary Table S2.

RT PCR

Total RNA was extracted from 150 mg of fresh leaf tissue or green siliques using the Trizol (Invitrogen) method according to the manufacturer's instructions and resuspended in 50 μl of water. Total RNA (10 μl) was treated with RQ1 DNase (Promega) and reverse transcribed using Superscript II (Invitrogen) following the manufacturer's instructions. RT-PCR was carried out using specific primers

designed on exons of the Pack-CACTA1a transposon and the gene target of integration (Supplementary Table S2).

BLAST search

BLAST analyses were carried out using the NCBI web interface (https://blast.ncbi.nlm.nih.gov/Blast.cgi) or the locally installed BLAST+ v2.2.25 (22), depending on the analysis. When the local version was used, blastn was run with the options –evalue 1e-5 and -max_target_seqs 100 where not specified otherwise.

Pack-CACTA analysis

Additional Pack-CACTA elements were detected in the Arabidopsis genome by blasting the 26-bp core TIR sequences (Figure 1C) of the active Pack-CACTA1a element to the Arabidopsis reference genome (TAIR10). The blastn (BLAST+ v2.2.25) was run with the options '-task blastn-short' and '-word_size 7'. The retrieved 18 matching sequences were imported in Geneious v7.1.7 (Biomatters) and manually checked to remove sequence derived from autonomous CACTA elements.

In order to identify Pack-CACTA elements in the genome of Ler-0 accession, the entire Pack-CACTA1a sequence was blasted on the PacBio Ler-0 assembly (http://www.pacb.com/). The sequences of the matching hits were retrieved from the assembly and aligned with Geneious v7.1.7 (Biomatters), using the MUSCLE algorithm and default parameters. Consensus trees were obtained from the alignment using the Temura-Nei genetic distance and Bootstrap resampling of 1000 replications. TIRs and TSDs were manually annotated.

Rice Pack-MULE analysis

The annotation of rice Pack-MULEs was obtained by Jiang *et al*. (3) and the sequence of each element was retrieved from the rice genome (http://rice.plantbiology.msu.edu/, release 6.0) using R. The sequences obtained were blasted against the same reference genome with blastn (BLAST+ v2.2.25) and the options—evalue 1e-5 and -outfmt 6. The list of 1 497 458 blast hits was filtered by alignment length >500 and evalue <1e-5, selecting only blast hits with 'query_start < 20' or 'query_end > (query_length – 20)' in order to select only results including at least one of the TIR sequences. The list was further filtered by excluding any sequence overlapping the original set of Pack-MULEs previously described (Jiang *et al*. 2011). Remaining overlapping sequences were merged to generate the final list of 2151 partial pack-MULE elements (Supplementary Data S2).

RESULTS

Pack-TYPE transposons with CACTA-derived TIRs are mobilized in epiRILs

DNAseq of PCR-free genomic libraries for the 67 epiRILs was used to search for TE mobilization, which is usually associated with increased copy number (1,23). We looked for TE-annotated loci for which read coverage was at least 1.8-fold higher than in wild-type Col-0, to select the most potentially active TEs. We then compared putative activity levels of these TEs by the average increase in sequencing coverage. According to this analysis, retrotransposon EVADE (EVD) (18) was the most active, with an average 11.8-fold coverage increase (Supplementary Table S3). Surprisingly, the second most active element, with a 6.2-fold increase, had an unusually complex structure of two short terminal DNA stretches of 451 bp (AT4TE18505) and 337 bp (AT4TE18510) (annotated as a member of the ATENSPM3 family of DNA transposons) separated by a sequence annotated as gene AT4G07526 of unknown function (Figure 1A). The putative product of this gene showed no similarity to any known transposase and the overall transposon structure, residing on chromosome 4, resembled a non-autonomous Pack-TYPE element.

AT4TE18505 and AT4TE18510 contain TIRs beginning with eight base pairs (CACTACAA) that are also a feature of the CACTA1 autonomous transposon (24), which is active in the epiRILs (16) (Supplementary Table S3). A blast search with 150 bp of both terminal sequences identified nine additional elements in the Col-0 genome with similar structures and CACTA1-like termini (Figure 1B and Supplementary Table S4). These included conserved terminal sequences of 26 bp (Figure 1C) and a 3-bp TSD (Supplementary Table S5). Their 150-bp terminal sequence identities ranged from 51.3 to 95.3% (Supplementary Table S5). Five of the elements shared related DNA sequences located between the CACTA1-like termini, while the remaining five elements carried sequences of different chromosomal origin (Figure 1B and D). Interestingly, not all of the captured sequences displayed similarities to Arabidopsis genes (Figure 1D) and DNA sequence databases did not reveal their origins. The TIR sub-terminal sequences of CACTA family transposons differ slightly between the 5′ and the 3′ ends (25). These differences were also present in the possibly active Pack-TYPE elements (Figure 1E) we discovered. Therefore, we called this family Pack-CACTA, which in the Arabidopsis Col-0 accession consists of 10 TEs divided up, according to their sequence similarities, into four groups (Figure 1B).

We found 50 new insertions of Pack-CACTA1a and 3 new insertions of Pack-CACTA2a in 8 and 3 epiRILs, respectively (Supplementary Table S6). In line epi26, Pack-CACTA1a and Pack-CACTA2a appeared mobile (Supplementary Table S6). New transposed copies were found mostly within euchromatic chromosomal regions (65%), resembling the insertion preference of the autonomous CACTA1 transposon (Supplementary Figure S1A). Both Pack-CACTA1a and PACK-CACTA2a display an increase or decrease in copy number as expected for mobile DNA transposons (Supplementary Table S7). We validated new insertions of Pack-CACTA1a by locus-specific PCR, choosing 11 random sites in epi26 and epi46 and two sites with Pack-CACTA2a insertions in epi14 and epi26 (Supplementary Figure S1B and C; Supplementary Table S8). Thus, both Pack-CACTA1a and Pack-CACTA2a were currently transposing and generating TSD of three nucleotides (Supplementary Data S3).

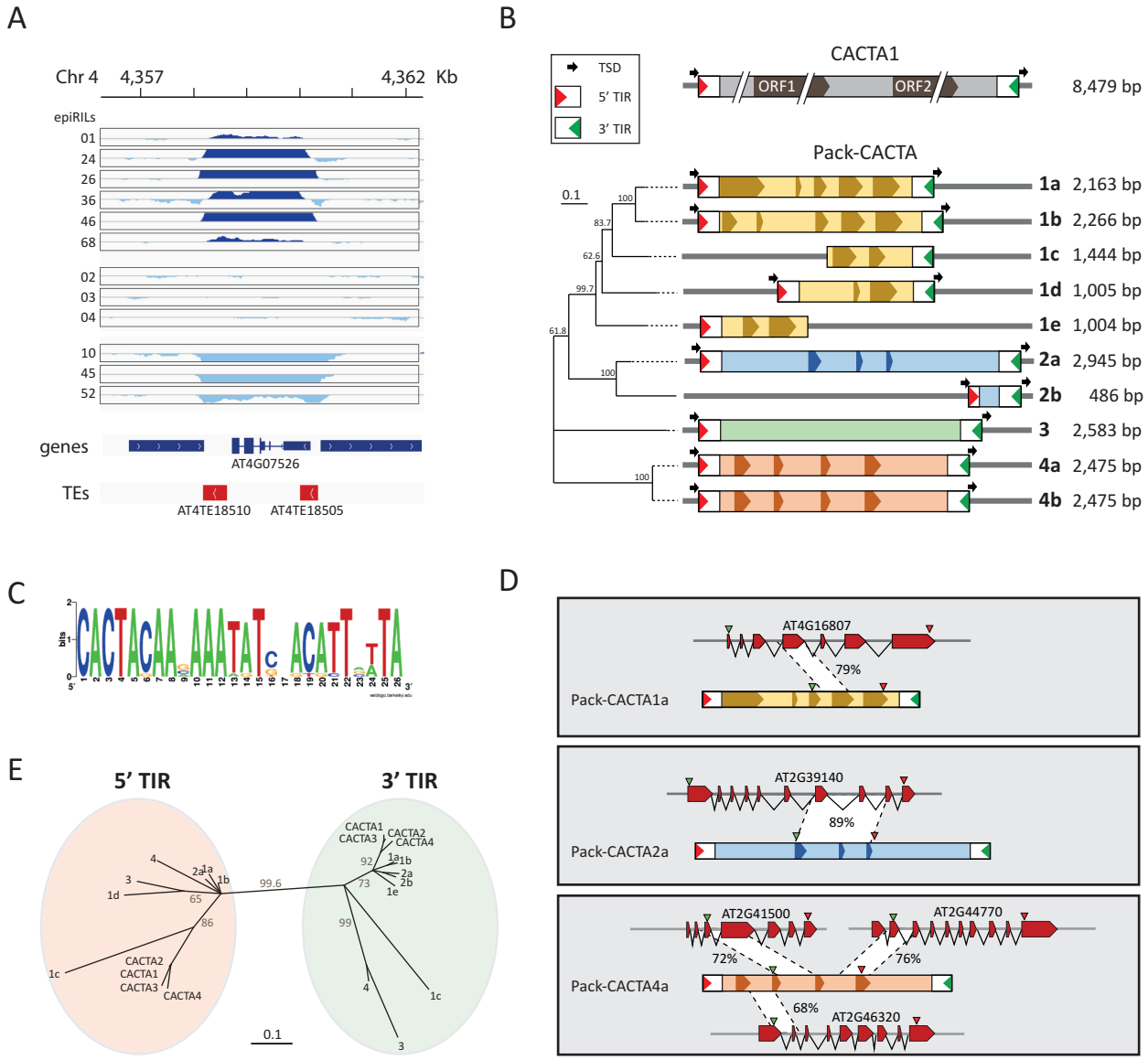

Figure 1. A novel non-autonomous pack-TYPE transposon related to CACTA becomes mobile in epiRILs. (**A**) Copy number changes in epiRILs of a Pack-TYPE element (Pack-CACTA1a) residing on chromosome 4. EpiRIL numbers are indicated left of each track. Differences to reference Col-0 control in log-transformed coverage of reads plotted in the interval of −1 to 1 are mapped to the Pack-CACTA1a for each track. TAIR10 annotation of genes (blue) and TEs (red) is displayed at the bottom. (**B**) Structure of Pack-CACTA family. Four groups of elements are marked in different colours with exons represented by a darker hue. The phylogenetic tree was obtained using alignment of full-length sequences. Numbers at each node indicate bootstrap support values of 1000 replications. The 5′ terminal inverted repeat (5′ TIR) is marked as a red triangle and 3′ TIR as a green triangle. TSDs of 3 bp (TSD) are marked as black arrows. (**C**) Sequence logo (http://weblogo.berkeley.edu/logo.cgi) obtained for alignment of TIRs of the 10 members of the Pack-CACTA family. (**D**) Chromosomal origin of Pack-CACTA sequences. The percentage identities to Arabidopsis genes are indicated for each area. With small green and red triangles are indicated respectively the predicted start (ATG) and termination of transcription. (**E**) Sequence relationships of the TIRs of Pack-CACTA groups. CACTA1 is annotated as AT2TE20205, CACTA2 as AT1TE42210, CACTA3 as AT2TE18415 and CACTA4 as AT1TE36570. Numbers at each node indicate bootstrap support values of 1000 replications.

Aberrant excisions of Pack-CACTA produce heritable DNA rearrangements and alter their transcripts

To better examine the timing and consequences of Pack-CACTA1a transposition, we characterized in more detail five new insertion sites in 10 sibling plants each of epi26 and epi46 (Supplementary Figure S2A–C). We recorded the presence or the absence of Pack-CACTA1a in individual sibling plants (Supplementary Figure S2D and E). The results are consistent with genetic segregation due to the initial hemizygosity of recent insertions and/or the frequent excision of newly inserted transposons (Figure 2A). In the case of transposon excision, we expected footprints that would be absent in the corresponding wild-type locus. In different epi26 sibling plants, we rescued sequences of both intact target loci and those containing footprints of the transposon (Figure 2B and Supplementary Data S4). Since from single plants we recovered multiple variants of transposon footprints, the results are consistent with high trans-

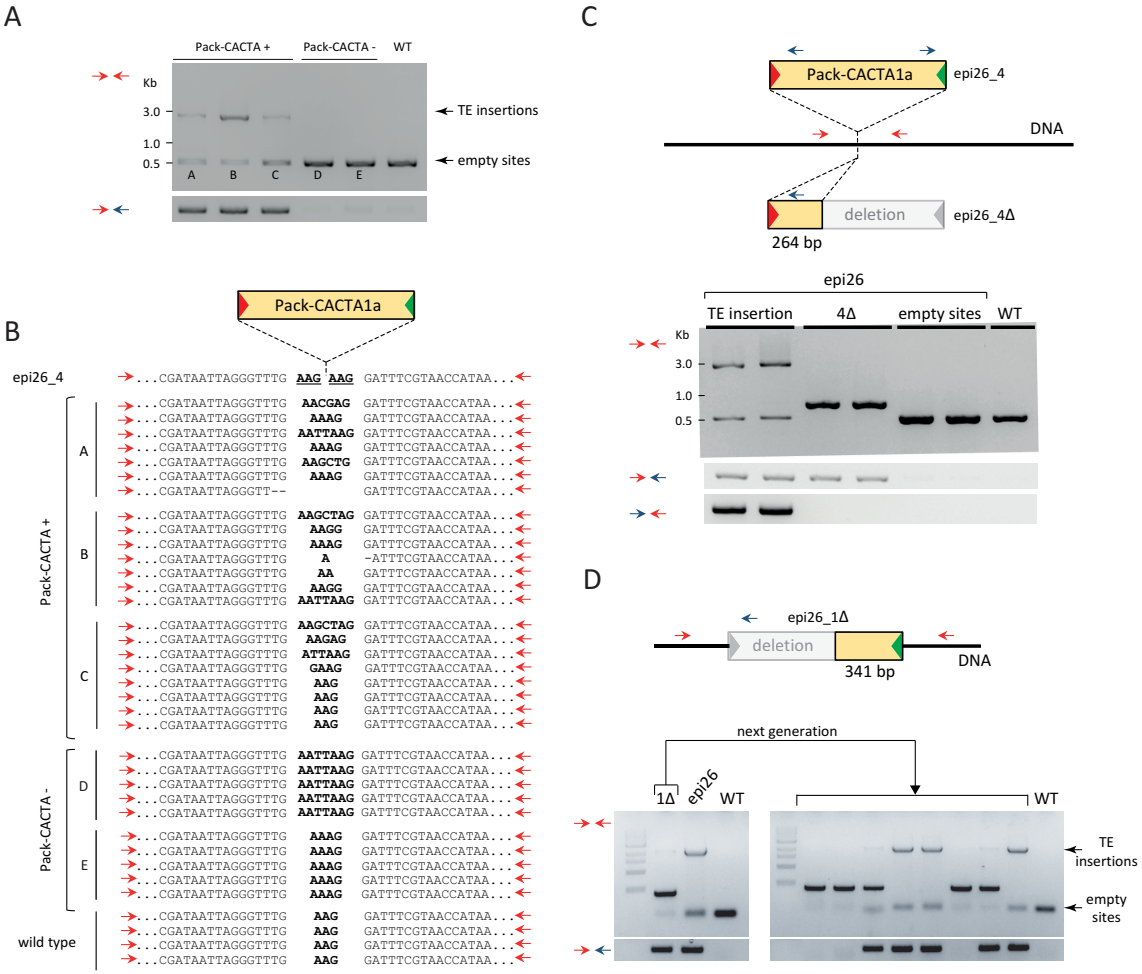

Figure 2. Hallmarks of Pack-CACTA1a excisions. (**A**) Excisions of Pack-CACTA1a from locus 4 of epiRIL26 (epi26_4) in five epi26 plants. Gel-separated PCR products were obtained with primers depicted in panel (C) as red or blue arrows. Lines marked as 'Pack-CACTA +' have a locus 4 containing a Pack-CACTA1a insert (indicated as 'TE insertions') and in plants marked as 'Pack-CACTA –'Pack-CACTA1a underwent excision, resulting in 'empty sites'. The wild-type Col-0 sample is marked as WT. (**B**) DNA sequences of PCR amplified 'empty sites' displayed in panel (A). The sequence of the initial epi26_4 containing the TE insertion is provided in the first line and TSD of 3 bp is underlined. Letters A–E left of the sequences correspond to the labelling in panel (A). (**C**) Identification of an aberrant excision of Pack-CACTA1a from epi26_4. Two lines of the gel, marked as 4Δ, represent epi26 plants that underwent aberrant excisions resulting in terminal deletion. (**D**) Identification of an additional aberrant excision event of Pack-CACTA1a (left panel) at a different locus of epi26 plants (epi26_1) and its transgenerational inheritance (right panel). All symbols and markings are as in panel (C). The line marked 1Δ represents a plant homozygote for a transposon insertion at epi26_1, in which only one Pack-CACTA1a was aberrantly excised. The remaining full-length transposon is not visible due to PCR competition. The ladder is 1 kb NEB.

positional activity of Pack-CACTA1a in epiRILs during inbreeding as well as somatic excision in tissues of individual plants.

Unexpectedly, in some plants of epi26 locus 4 (epi26_4) (Supplementary Figure S2F) we found a 264-bp fragment of the CACTA1a 5′ sequence linked to 3 bp of TSD (Figure 2C and Supplementary Data S5). Ten epi26 plants with a Pack-CACTA1a insertion at this locus contained either the entire Pack-CACTA1a or its deletion derivative (Supplementary Figure S2F and Supplementary Data S6). Therefore, it is most probable that the deletion resulted from aberrant excision of a full copy from this location. Importantly, we did not detect somatic 'reversions' to a wild type-like locus in plants homozygous for the deleted version of Pack-CACTA1a. The fragmented element was apparently immobilized at this new location (Figure 2C).

To investigate the frequency and features of aberrant excisions of Pack-CACTA1a, we examined a further 64 sibling plants each of epiRILs, epi26 and epi46, searching for additional deletion events at five loci at which complete insertions of Pack-CACTA1a were previously found. To this end, we used locus-specific primers that flanked each of the putative Pack-CACTA1a target sites (Supplementary Figure S2C). Six independent deletion events were detected in 640 PCR reactions, two in epi26 and four in epi46 (Figure 2D, Supplementary Figure S3 and Supplementary Data S6). Therefore, generation of deletion derivatives of this transposon appear to be quite common: in ∼5% of plants or in 1% of loci previously targeted by Pack-CACTA1a, and at least some of these are transgenerationally transmitted (Figure 2D and Supplementary Figure S3C). In the case of transposon insertion in an active gene, TIR deletion could

stably alter expression of the gene and/or alter the structure of its transcript. Indeed, in one case we detected a novel hybrid messenger RNA (mRNA) initiated within a promoter residing in the remaining part of Pack-CACTA1a, the second exon of the Pack-CACTA1a 'passenger gene' was fused with the exon of a downstream chromosomal gene (Figure 3 left panels and Supplementary Data S5). In a second case, the 3′ part of the transposon was inserted in a gene exon; thus, its sequence formed part of the mature gene transcript (Figure 3 right panels and Supplementary Data S7).

Specific features of real time pack-CACTA mobilization suggest a mechanism for chromosomal DNA acquisition

The sequences of remaining fragments of the transposon showed that deletions occurred at approximately equal frequencies at the 3′ (3 deletions) and 5′ terminals (4 deletions) (Figure 2C and D; Supplementary Figure S3; Supplementary Data S8). This observation suggested that if two elements with compatible 5′ and 3′ ends are inserted close to each other in the same orientation, it is possible that new sequences between can be mobilized. If such a pair of Pack-Type transposons, separated by a short stretch of chromosomal DNA, undergoes complementary deletions of each of the elements, a 'hybrid' Pack-CACTA encompassing chromosomal DNA within the respective 5′ and 3′ terminals of the former transposons would be generated (Figure 4A). To examine whether Pack-CACTA1a integrates frequently at closely linked chromosomal locations, we designed a PCR-based screen to recover chromosomal DNA stretches residing between putative neighbouring new inserts of Pack-CACTA1a (Figure 4B). Obviously, this approach can only detect insertions separated by relatively short stretches of chromosomal DNA. Therefore, it was surprising that numerous amplifications of chromosomal DNA fragments were observed in the 128 DNA samples of epi26 and epi46 used previously for the Pack-CACTA1a deletion screen (Figure 4C and Supplementary Figure S4). The sequences of 28 different PCR products from 20 plants (Supplementary Figure S5; Supplementary Data S9 and S10) all contained fragments of Arabidopsis genomic DNA of various sizes (38 to 2137 bp) between two Pack-CACTA1a copies. For three fragments, one of the two insertions of Pack-CACTA1a was also found by the NGS approach (epi26.02a, epi26.61b and epi46.19). Remarkably, 27 PCR products revealed insertions of Pack-CACTA1a as tandems in direct orientation and only one inversion (Supplementary Figure S5). The frequent formation by newly inserted copies of Pack-CACTA1a of direct repeats interspaced by short stretches of chromosomal DNA, combined with the observed formation of terminal deletion derivatives of the transposon, suggest that pairs of Pack-CACTA1a may form new 'hybrid' elements. Indeed, we recovered three deletion derivatives of Pack-CACTA1a among the 27 direct transposon repeats (Supplementary Figure S5). Importantly, most neighbouring Pack-CACTA1a insertions were separated by genic sequences (21 out of 28, Supplementary Figure S5); in one case three Pack-CACTA1a insertions were located in one gene (Figure 4D).

Obvious extrapolations from the observed properties of active Pack-CACTA1a provide crucial clues as to the mechanism by which Pack-TYPE transposons acquire and relocate stretches of chromosomal DNA (Figure 4A). However, detection of a newly formed, transpositionally active Pack-CACTA1a derivative would require the screening of several thousand plants harbouring currently active parental elements.

As an alternative, we searched for structural variants of Pack-CACTA1a elements in the newly assembled genome of the Arabidopsis Ler-0 accession and retrieved eight analogous elements (Supplementary Figure S6 and Supplementary Table S9). Of these, the sequences of three full-length elements and two deletion derivatives were closely related to Pack-CACTA1a of Col-0, with sequence identities of 73 to 90% (Supplementary Table S9). The sequences of the remaining three copies of Ler-0 Pack-CACTA elements resembled Pack-CACTA1a only in the first 992 bp and the last 325 bp with 76 and 72% identities, respectively, including the TIRs and part of the internal sequence. However, the remaining 1181 bp showed no similarity to the 'parental' element but contained a 132-bp stretch with 73% identity to the second exon of gene AT3G48620 encoding a putative protein of the outer membrane family (OMP85) (Figure 4E). This novel structure of a Ler-0 relative of Pack-CACTA1a, which consists of 5′ and 3′ regions similar to Col-0 Pack-CACTA1a and the newly acquired central part, suggests that this Pack-CACTA transposon originated by a mechanism compatible with that we recorded in real time for Pack-CACTA1a transposition in Col-0 and illustrated on Figure 4A.

In rice, previous analyses of Pack-MULE transposons considered only complete elements with both terminal repeats encompassing various fragments of rice chromosomal DNA (7). We reanalysed available rice genomic sequences for intermediates of chromosomal DNA acquisition by rice Pack-MULEs, such as transposon deletions due to aberrant excision events. We applied a blast search across the rice genome using as a query 2853 previously retrieved pack-MULE intact elements (7) and filtered results to recover incomplete elements with only one terminal sequence (see 'Materials and Methods' section). Of the 2151 entries meeting these criteria (Supplementary Data S2), 269 (13%) Pack-MULE deletion derivatives resided <10 kb from an intact Pack-MULE element (Figure 4F). Pack-MULEs consisting of terminal sequences flanked by different TSDs, making up 17% of the Pack-MULE population, were previously discarded as false positives (7). However, considering the deduced mechanism of DNA acquisition by Pack-TYPE elements, some of these putative transposons may represent cases of newly generated but not yet relocated hybrid Pack-MULEs consisting of terminal sequences from two different parental elements that underwent aberrant excisions (Figure 4G). Thus, the abundant Pack-MULEs of the rice genome, which have contributed significantly to its current organization, seem to include examples of putative transposon intermediates consistent with the scheme of chromosomal DNA acquisition during real-time Pack-CACTA mobilization in Arabidopsis (Figure 4A).

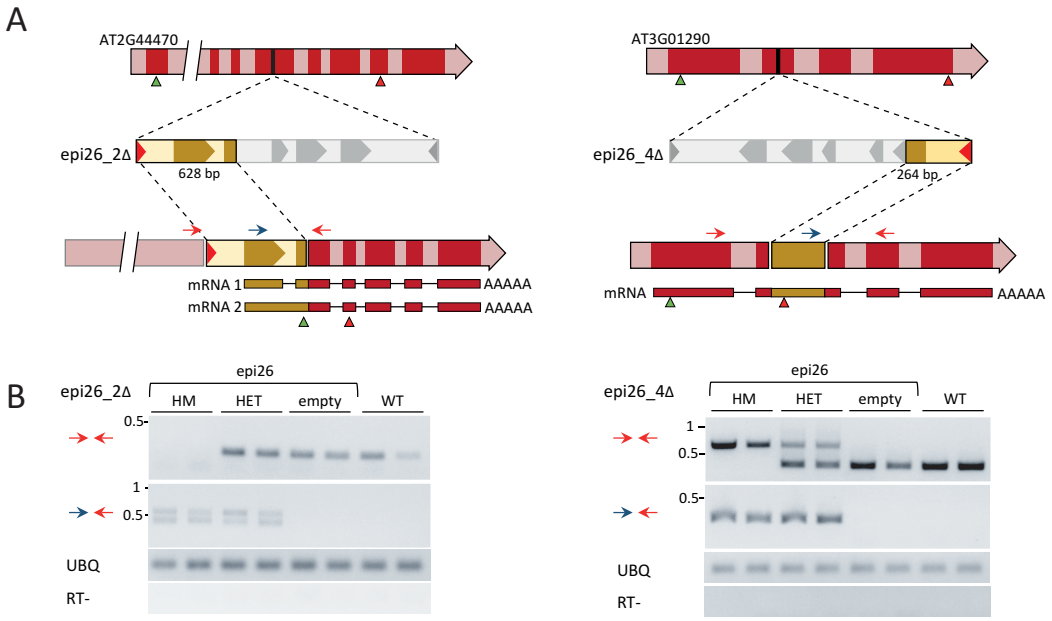

Figure 3. Deletion derivatives of Pack-CACTA1a alter gene transcripts. (**A**) Structures of new Pack-CACTA1a inserts that underwent aberrant excisions. Deleted parts of Pack-CACTA1a are in grey. Exons are in darker and introns in lighter hues. Primers used for RtPCR are shown as blue and red arrows. Deduced, novel mRNAs are displayed at the bottom, with green and red triangles representing initiation and termination codons, respectively. (**B**) Results of RT-PCR performed on single epi26 homozygous (HM), heterozygous (HET) or wild type-like (empty) plants for the corresponding Pack-CACTA1a deletion derivatives depicted in panel (A). The combination of primers used is shown on the left. Col-0 wild-type plants are shown as control (WT). Amplification of ubiquitin mRNA (UBQ) and samples without reverse transcriptase (RT-) were used as controls. Two biological replicates per genotype were used.

DISCUSSION

The effects of activation of MULE and CACTA transposons have been observed originally in maize (26,27), and then in other plants, including Arabidopsis (24,28). However, the real time mobilization of non-autonomous Pack-TYPE TEs was not reported or noticed previously. This could be due at least in part to imprecise annotation of Pack-TYPE TEs, which require specific detection approaches to be identified, as previously showed for Pack-MULEs (7). It is therefore not surprising that although copy number variation at PackCACTA1a locus was previously reported, its mobility was associated to the CACTA1 autonomous transposon (29).

Moreover, Pack-CACTA1a was found mobile only in 9 epiRILs (Supplementary Table S8), while the autonomous CACTA1 element was reported active in 18 epiRILs (18). This observation indicates that the simple expression of a potential compatible CACTA transposase is not sufficient to induce transposition of Pack-CACTA elements, suggesting that DNA methylation plays a direct role in controlling the transposition of Pack-TYPE TEs. Nonetheless, Pack-CACTA activation could be also mediated by cis-acting sequences captured between the two TIRs, or located in the chromosomal areas neighbouring a transposon insertion, similarly to the position effect already observed for MULE transposons in maize (30).

A possible contribution of Pack-TYPE transposon activity to the evolution of plant genes and genomes was deduced from analyses of rice, Brassica and soybean (3,6,31,32). It was postulated that repair of nicks and gaps created at the transposon insertions sites or arising due to the element structures might be responsible for the acquisition of chromosomal DNA (33–35). Such DNA-repair mechanisms would operate prior transposon relocation and the newly generated elements would then move to different chromosomal locations only after incorporation of chromosomal DNA. Here, studying transposition properties of Pack-CACTA1a elements *in vivo*, we propose a model where acquisition of chromosomal DNA is tightly linked to transposition. The *de novo* formation of internal deletions in newly inserted TE copies has been reported for several active DNA transposons, such as *Ac-Ds* and *Mutator* in maize (36,37), *P*-element in Drosophila (38) and *Tam3* in Antirrhinum (35). However, the model proposed for the formation of such non-autonomous TE copies, based on slip mispairing during error-prone DNA repair synthesis, does not consider aberrant excisions and thus acquisition of large stretches of DNA, as observed for Pack-TYPE transposons.

Pack-CACTA1a transposition has a tendency to produce closely spaced repeats in direct orientation. Subsequently, complementary aberrant excisions of the elements within these repeats, leaving behind one of the two TIRs, could generate a novel 'hybrid' Pack-CACTA that incorporates the chromosomal DNA that separated the two neighbouring insertions (Figure 4A).

This quite simple and efficient model for chromosomal DNA acquisition relies on the combination of known characteristics of particular DNA transposons. For example, maize *Ac*/*Ds*, *P*-element of *Drosophila, Tc1* transposon in *Caenorhabditis* and *Sleeping Beauty* in mouse tend to transpose to nearby locations, termed 'local hopping' (2,39–42). However, the orientation of elements in such integration ar-

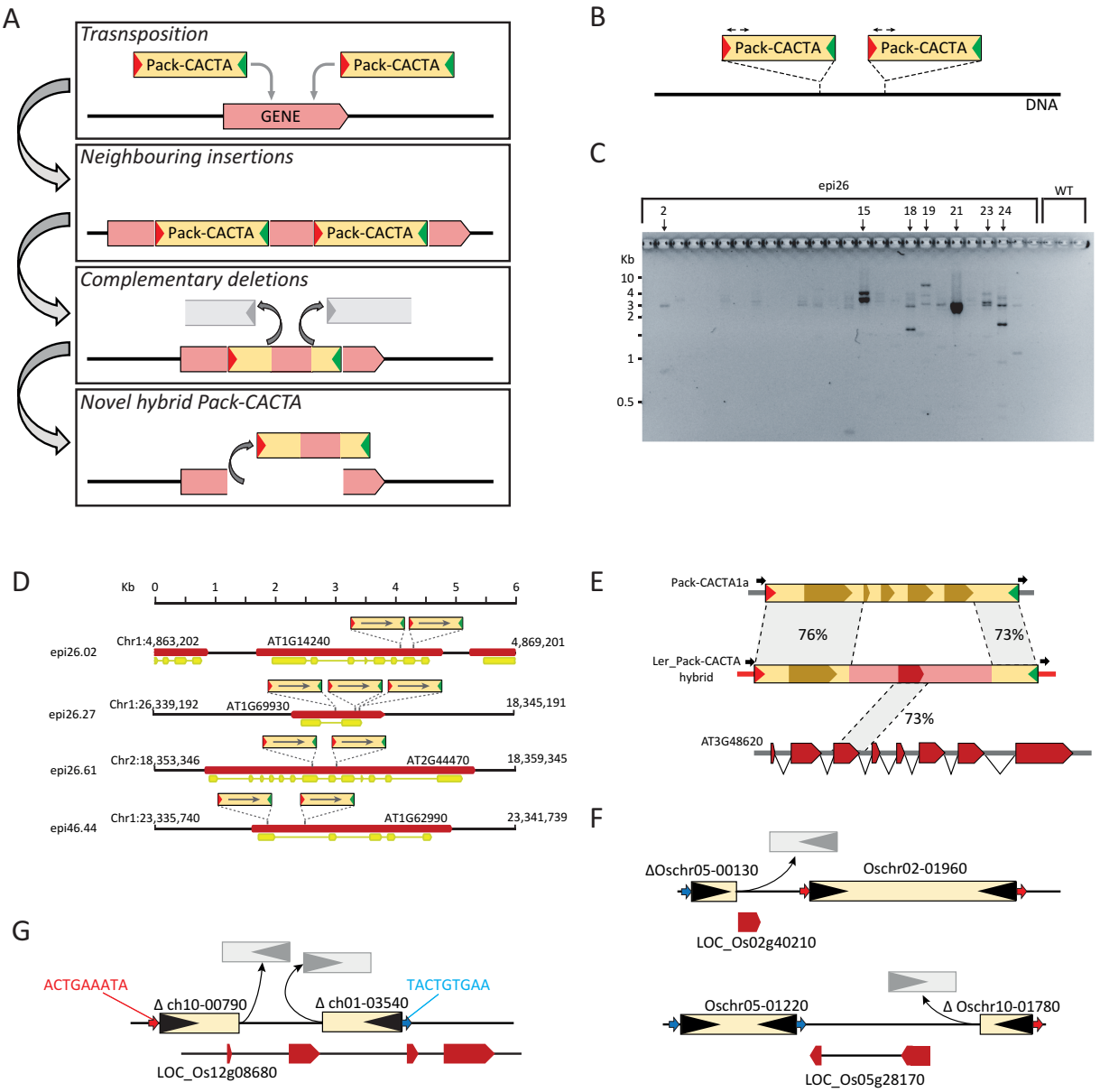

Figure 4. High frequency of Pack- CACTA1a insertion at some loci support a model for DNA capture in new pack-TYPE transposons. (**A**) Possible mechanism of genic DNA capture by Pack-CACTA1a. New transposon insertions form a closely spaced tandem array. Subsequently, neighbouring elements undergo aberrant excisions resulting in complementary, terminal deletions, which form a novel hybrid Pack-CACTA element that incorporates chromosomal DNA which separated the initial closely spaced insertions. (**B**) PCR strategy used for detection of neighbouring insertions of Pack-CACTA1a. Black arrows indicate the primers. (**C**) Results of a PCR screen for neighbouring insertions of Pack-CACTA1a in 30 epi26 plants. Cloned fragments corresponding to adjacent insertions are marked with black arrows. The full set for all tested plants is displayed in Supplementary Figure 4. (**D**) Arrangement of Pack-CACTA1a copies in four representative loci containing tandem insertions (data for other loci are presented in Supplementary Figure S5). The orientation of Pack-CACTA1a copies is indicated with an arrow (5′ to 3′ direction). Red and yellow rectangles represent genes and their transcripts, respectively. (**E**) Structure of a Col-0 Pack-CACTA1a-related element identified in the Ler accession. The percentage sequence identities with Pack-CACTA1a and the acquired gene AT3G48620 are indicated. (**F**) Examples of rice Pack-MULE terminal deletion derivatives neighbouring intact Pack-MULE elements with TIRs belonging to the same family. Black triangles represent TIRs. Blue and red arrows represent 9 bp TSD (different colours correspond to different sequences). Red rectangles represent coding regions of genes. (**G**) Example of neighbouring Pack-MULE deletion derivatives compatible with the formation of a novel hybrid Pack-MULE element. Symbols and colours as in panel (F).

rays are random and not preferentially direct, as observed for Pack-CACTA1a. Incomplete or aberrant excision is known for CACTA-like TE in soybean (43), while maize chromosomal aberrations, like breakages and fusions, were previously associated with the transposition of TIRs belonging to two different neighbouring *Ac*/*Ds* elements (44). It is interesting that Pack-CACTA1a combines both properties, i.e. efficient formation of closely linked tandem insertions in direct orientation and frequent aberrant excisions that lead to complementary deletion of one of the two TIRs of each pair. The remaining terminals of the neighbouring transposons encompass a new, single element incorporating the intervening chromosomal DNA.

The incorporation of gene fragments and their duplication by transposition to new genomic positions has been suggested for certain TE families, including MULEs, CACTAs and HELITRONs (1). This 'transduplication' process seems to be very frequent in rice and can be clearly assigned to the historical activity of Pack-MULEs (45). Importantly, transduplicated genes seem to be under purifying selection (32) or serve regulatory functions (45). This implies that Pack-TYPE transposon mobilization may have a direct influence on gene evolution, generating new gene functions or regulatory activities by shuffling parts of various coding regions across the genome. Taking into account the high rates of nearby directional insertion and aberrant excision observed here for the Pack-CACTA1a element, the 'TE-assisted' origin of genes may be strongly underestimated, especially in plants with genomes larger than Arabidopsis.

DATA AVAILABILITY

Sequencing data have been deposited in Gene Expression Omnibus under the accession number GSE120571. We have loaded on UCSC the track with all peaks in epiRILs, and the session can be accessed here https://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hub_21716_araTha1&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr1%3A6767%2D24046&hgsid=229741013_qnIdm59JYHEa9ZIGRCBFAYmqEO2h.
All the other data generated or analysed during this study are included in this published article or available upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all components of Dr. Paszkowski group for constructive discussion during the development of this project, and the three anonymous reviewers whose fair comments helped improve and clarify this manuscript.
Author contributions: M.C. and J.P. designed the experiments; T.J. prepared nucleic acid libraries; M.C., T.J. and E.C. performed validation of new transposon insertions and their arrangement, M.C. performed all other experiments and analysis of the genomic data; M.C. and J.P. wrote the paper.

FUNDING

European Research Council (EVOBREED) [322621]; Gatsby Fellowship [AT3273/GLE]. Funding for open access charge: Gatsby Fellowship [AT3273/GLE].
Conflict of interest statement. None declared.

REFERENCES

1. Lisch,D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.*, **14**, 49–61.
2. Skipper,K.A., Andersen,P.R., Sharma,N. and Mikkelsen,J.G. (2013) DNA transposon-based gene vehicles - scenes from an evolutionary drive. *J. Biomed. Sci.*, **20**, 92.
3. Jiang,N., Ferguson,A.A., Slotkin,R.K. and Lisch,D. (2011) Pack-Mutator–like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1537–1542.
4. Kawasaki,S. and Nitasaka,E. (2004) Characterization of Tpn1 family in the japanese morning glory: En/Spm-related transposable elements capturing host genes. *Plant Cell Physiol.*, **45**, 933–944.
5. Yu,Z., Wright,S.I. and Bureau,T.E. (2000) Mutator-like elements in Arabidopsis thaliana: Structure, diversity and evolution. *Genetics*, **156**, 2019–2031.
6. Zabala,G. and Vodkin,L.O. (2005) The *wp* mutation of *Glycine max* carries a Gene-Fragment-Rich transposon of the CACTA superfamily. *Plant Cell Online*, **17**, 2619–2632.
7. Jiang,N., Bao,Z., Zhang,X., Eddy,S.R. and Wessler,S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.
8. Gilly,A., Etcheverry,M., Madoui,M.-A., Guy,J., Quadrana,L., Alberti,A., Martin,A., Heitkam,T., Engelen,S., Labadie,K. *et al.* (2014) TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics*, **15**, 377.
9. Zabala,G. and Vodkin,L.O. (2005) The *wp* mutation of *Glycine max* carries a Gene-Fragment-Rich transposon of the CACTA superfamily. *Plant Cell Online*, **17**, 2619–2632.
10. Xu,M., Brar,H.K., Grosic,S., Palmer,R.G. and Bhattacharyya,M.K. (2010) Excision of an active CACTA-Like transposable element from DFR2 causes variegated flowers in Soybean [Glycine max (L.) Merr.]. *Genetics*, **184**, 53–63.
11. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated Single-Base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
12. Zilberman,D., Gehring,M., Tran,R.K., Ballinger,T. and Henikoff,S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.*, **39**, 61–69.
13. Catoni,M., Griffiths,J., Becker,C., Zabet,N.R., Bayon,C., Dapp,M., Lieberman-Lazarovich,M., Weigel,D. and Paszkowski,J. (2017) DNA sequence properties that predict susceptibility to epiallelic switching. *EMBO J.*, **36**, 617–628.
14. Kankel,M.W., Ramsey,D.E., Stokes,T.L., Flowers,S.K., Haag,J.R., Jeddeloh,J.A., Riddle,N.C., Verbsky,M.L. and Richards,E.J. (2003) Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics*, **163**, 1109–1122.
15. Saze,H., Scheid,O.M. and Paszkowski,J. (2003) Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nat. Genet.*, **34**, 65–69.
16. Reinders,J., Wulff,B.B.H., Mirouze,M., Marí-Ordóñez,A., Dapp,M., Rozhon,W., Bucher,E., Theiler,G. and Paszkowski,J. (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev.*, **23**, 939–950.
17. Debladis,E., Llauro,C., Carpentier,M.-C., Mirouze,M. and Panaud,O. (2017) Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology. *BMC Genomics*, **18**, 537.
18. Mirouze,M., Reinders,J., Bucher,E., Nishimura,T., Schneeberger,K., Ossowski,S., Cao,J., Weigel,D., Paszkowski,J. and Mathieu,O. (2009) Selective epigenetic control of retrotransposition in Arabidopsis. *Nature*, **461**, 427–430.

19. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
20. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
21. Nakagome,M., Solovieva,E., Takahashi,A., Yasue,H., Hirochika,H. and Miyao,A. (2014) Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements. *BMC Bioinformatics*, **15**, 71.
22. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
23. Chuong,E.B., Elde,N.C. and Feschotte,C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.
24. Miura,A., Yonebayashi,S., Watanabe,K., Toyama,T., Shimada,H. and Kakutani,T. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature*, **411**, 212–214.
25. Frey,M., Reinecke,J., Grant,S., Saedler,H. and Gierl,A. (1990) Excision of the En/Spm transposable element of Zea-Mays requires 2 Element-Encoded proteins. *EMBO J.*, **9**, 4037–4044.
26. McClintock,B. (1955) Controlled mutation in mais: The a1m-1-Spm system of control of gene action and mutation. Continued studies of the mode of operation of the controlling elements Ds and Ac. *Carnegie Inst. Wash. Year b.*, **54**, 245–255.
27. Robertson,D.S. (1978) Characterization of a mutator system in maize. *Mutat. Res. Mol. Mech. Mutagen.*, **51**, 21–28.
28. Singer,T., Yordan,C. and Martienssen,R.A. (2001) Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.*, **15**, 591–602.
29. Tsukahara,S., Kobayashi,A., Kawabe,A., Mathieu,O., Miura,A. and Kakutani,T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, **461**, 423–426.
30. Singh,J., Freeling,M. and Lisch,D. (2008) A position effect on the heritability of epigenetic silencing. *PLoS Genet.*, **4**, e1000216.
31. Alix,K., Joets,J., Ryder,C.D., Moore,J., Barker,G.C., Bailey,J.P., King,G.J. and (Pat) Heslop-Harrison,J.S. (2008) The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation. *Plant J.*, **56**, 1030–1044.
32. Hanada,K., Vallejo,V., Nobuta,K., Slotkin,R.K., Lisch,D., Meyers,B.C., Shiu,S.-H. and Jiang,N. (2009) The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell*, **21**, 25–38.
33. Bennetzen,J.L. and Springer,P.S. (1994) The generation of Mutator transposable element subfamilies in maize. *Theor. Appl. Genet.*, **87**, 657–667.
34. Engels,W.R., Johnson-Schlitz,D.M., Eggleston,W.B. and Sved,J. (1990) High-frequency P element loss in Drosophila is homolog dependent. *Cell*, **62**, 515–525.
35. Yamashita,S., Takano-Shimizu,T., Kitamura,K., Mikami,T. and Kishima,Y. (1999) Resistance to gap repair of the transposon Tam3 in antirrhinum majus: A role of the end regions. *Genetics*, **153**, 1899–1908.
36. Conrad,L.J., Bai,L., Ahern,K., Dusinberre,K., Kane,D.P. and Brutnell,T.P. (2007) State II dissociation element formation following activator excision in maize. *Genetics*, **177**, 737–747.
37. Hsia,A.-P. and Schnable,P.S. (1996) DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, MuDR. *Genetics*, **142**, 603–618.
38. Takasu-Ishikawa,E., Yoshihara,M. and Hotta,Y. (1992) Extra sequences found at P element excision sites in Drosophila melanogaster. *Mol. Gen. Genet.*, **232**, 17–23.
39. Bancroft,I. and Dean,C. (1993) Transposition pattern of the maize element Ds in Arabidopsis thaliana. *Genetics*, **134**, 1221–1229.
40. Carlson,C.M., Dupuy,A.J., Fritz,S., Roberg-Perez,K.J., Fletcher,C.F. and Largaespada,D.A. (2003) Transposon mutagenesis of the mouse germline. *Genetics*, **165**, 243–256.
41. Fischer,S.E.J., Wienholds,E. and Plasterk,R.H.A. (2003) Continuous exchange of sequence information between dispersed Tc1 transposons in the Caenorhabditis elegans genome. *Genetics*, **164**, 127–134.
42. Zhang,P. and Spradling,A.C. (1993) Efficient and dispersed local P element transposition from drosophila females. *Genetics*, **133**, 361–373.
43. Xu,M., Brar,H.K., Grosic,S., Palmer,R.G. and Bhattacharyya,M.K. (2010) Excision of an Active CACTA-Like Transposable Element From DFR2 Causes Variegated Flowers in Soybean [Glycine max (L.) Merr.]. *Genetics*, **184**, 53–63.
44. Ralston,E., English,J. and Dooner,H.K. (1989) Chromosome-breaking structure in maize involving a fractured Ac element. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 9451–9455.
45. Juretic,N., Hoen,D.R., Huynh,M.L., Harrison,P.M. and Bureau,T.E. (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.*, **15**, 1292–1297.