Sensitive detection of pre-integration intermediates of long terminal repeat retrotransposons in crop plants

Jungnam Cho1,2,3,*, Matthias Benoit1, Marco Catoni1,4, Hajk-Georg Drost1, Anna Brestovitsky1, Matthijs Oosterbeek1,5 and Jerzy Paszkowski1,6,*

Retrotransposons have played an important role in the evolution of host genomes1–3. Their impact is mainly deduced from the composition of DNA sequences that have been fixed over evolutionary time3. Such studies provide important ‘snapshots’ reflecting the historical activities of transposons but do not predict current transposition potential. We previously reported sequence-independent retrotransposon trapping (SIRT) as a method that, by identification of extrachromosomal linear DNA (eclDNA), revealed the presence of active long terminal repeat (LTR) retrotransposons in Arabidopsis1. However, SIRT cannot be applied to large and transposon-rich genomes, as found in crop plants. We have developed an alternative approach named ALE-seq (amplification of LTR of eclDNAs followed by sequencing) for such situations. ALE-seq reveals sequences of 5’ LTRs of eclDNAs after two-step amplification: in vitro transcription and subsequent reverse transcription. Using ALE-seq in rice, we detected eclDNAs for a novel Copia family LTR retrotransposon, Go-on, which is activated by heat stress. Sequencing of rice accessions revealed that Go-on has preferentially accumulated in Oryza sativa ssp. indica rice grown at higher temperatures. Furthermore, ALE-seq applied to tomato fruits identified a developmentally regulated Gypsy family of retrotransposons. A bioinformatic pipeline adapted for ALE-seq data analyses is used for the direct and reference-free annotation of new, active retroelements. This pipeline allows assessment of LTR retrotransposon activities in organisms for which genomic sequences are also predominant in LTR retrotransposons of other plants, we used the custom-made software LTRpred for de novo annotation of LTR retrotransposons in rice and tomato genomes (see Methods). Young retroelements were selected by filtering for at least 95% identity between the two LTRs and subsequently examined for their cognate tRNAs (Supplementary Fig. 1). As in Arabidopsis, around 80% of LTR retrotransposons in the tomato genome contained Met-iCAT PBS (Supplementary Fig. 1). In contrast, only 30% harboured Met-iCAT PBS in rice and Arg-CCT (Arginine tRNA-CCT anticodon) PBS was found in 60% of young LTR retrotransposons (Supplementary Fig. 1).

Results

Development of ALE-seq. Retrotransposons include a conserved sequence known as the primer binding site (PBS), where binding of the 3’ end of cognate transfer RNA initiates the reverse transcription reaction. Met-iCAT (Methionine tRNA-CAT anticodon) PBS was chosen for SIRT as it is the site present in most annotated Arabidopsis retrotransposons1. To examine whether Met-iCAT PBS sequences are also predominant in LTR retrotransposons of other plants, we used the custom-made software LTRpred for de novo annotation of LTR retrotransposons in rice and tomato genomes (see Methods). Young retroelements were selected by filtering for at least 95% identity between the two LTRs and subsequently examined for their cognate tRNAs (Supplementary Fig. 1). As in Arabidopsis, around 80% of LTR retrotransposons in the tomato genome contained Met-iCAT PBS (Supplementary Fig. 1). In contrast, only 30% harboured Met-iCAT PBS in rice and Arg-CCT (Arginine tRNA-CCT anticodon) PBS was found in 60% of young LTR retrotransposons (Supplementary Fig. 1). We used Met-iCAT PBS in our initial experiments because most retrotransposons known to be active in rice callus (for example Tos17 and Tos19) contain Met-iCAT PBS. Initially, SIRT was performed on DNA extracted from rice leaves and calli; however, we did not detect eclDNAs for Tos17 and Tos19 in rice tissues by this method (Supplementary Fig. 2). We reasoned that the short stretch of PBS used for primer design in SIRT may have impaired PCR efficiency due to the many PBS-related

1The Sainsbury Laboratory, University of Cambridge, Cambridge, UK. 2National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai, China. 3CAS-JIC Centre of Excellence for Plant and Microbial Science, Chinese Academy of Sciences, Shanghai, China. 4School of Biosciences, University of Birmingham, Birmingham, UK. 5Present address: Laboratory of Nematology, Wageningen University, Wageningen, the Netherlands. 6Present address: Radachowka 37, Kolbiel, Poland. *e-mail: jungnamcho@sippe.ac.cn; jerzy.paszkowski@slcu.cam.ac.uk
sequences present in larger genomes containing a high number of retroelements, as is the case in rice.

To counter this problem, we developed an alternative method, named ALE-seq, with improved selectivity and sensitivity of eclDNA detection. A crucial difference compared to SIRT is that ALE-seq amplification of eclDNA is separated into two reactions: in vitro transcription and reverse transcription (Fig. 1a). This decoupling of the use of the two priming sequences followed by the digestion of non-templated DNA and RNA is more selective and efficient than the single PCR amplification in SIRT.

ALE-seq starts with ligation to the ends of eclDNA of an adapter containing a T7 promoter sequence at its 5′ end and subsequent in vitro transcription with T7 RNA polymerase. The synthesized RNA is then reverse transcribed using the primer that binds the transcripts at the PBS site. The adapter and the oligonucleotides priming reverse transcription are anchored with partial Illumina adapter sequences (Supplementary Table 1), which allows the amplified products to be directly deep-sequenced in a strand-specific manner. The ALE-seq-sequences derived from retrotransposon eclDNA are predicted to contain the intact 5′ LTR up to the PBS site, flanked by Illumina paired-end sequencing adapters. We used the Illumina MiSeq platform for sequencing because its long reads of 300 base pairs (bp) from both ends cover the entire LTR lengths of most potentially active elements. The Illumina adapters were tagged to the intact LTR DNA without fragmentation of the amplicons. This, together with the long reads of MiSeq, allowed us to reconstitute the complete LTR sequences, in the absence of the reference genome sequence. The reconstituted LTRs were analysed using the alignment-based approach that complements the mapping-based approach when the reference genome is incomplete (Fig. 1b).

![Diagram](image-url)
First, we tested ALE-seq on *Arabidopsis* by examining heat-stressed Col-0 *Arabidopsis* plants, met1-1 mutant and epi12, a met1-derived epigenetic recombinant inbred line. ALE-seq cleanly and precisely recovered sequences of complete LTRs for Onsen, Copia21 and Evade in samples containing their respective eclDNA (Supplementary Fig. 3). Due to priming of the reverse transcription reaction at PBS, the reads were explicitly mapped to the 5′ LTR but not to the 3′ LTR, although the two LTRs have identical sequences. The ALE-seq reads have well-defined extremities, starting at the position marking the start of LTRs and finishing at the PBS, which is consistent with their eclDNA origin. The ends of LTRs can also be inspected for conserved sequences that would further confirm their eclDNA origin (Supplementary Fig. 4). This reduced ambiguity of read mapping in ALE-seq analysis, combined with the clear-cut detection of LTR ends, allows for explicit and precise assignment of ALE-seq results to active LTR retrotransposons.

Since SIRT failed to detect eclDNAs of rice retrotransposons known to be activated in rice callus, we examined whether ALE-seq would identify their eclDNAs. As shown in Fig. 1c–f, ALE-seq unambiguously detected eclDNAs of *Tos17* and *Tos19* in rice callus but not in leaf samples. To test whether detection of 5′ LTR sequences requires the entire ALE-seq procedure, we performed control experiments with depleted ALE-seq reactions, for example, in the absence of enzymes for either ligation, in vitro transcription or reverse transcription. All incomplete procedures failed to produce sequences containing 5′ LTRs derived from eclDNAs (Fig. 1e,f). Taken together, the data show that ALE-seq can detect eclDNAs of LTR retrotransposons in *Arabidopsis* as well as in rice with considerably greater efficiency than the SIRT method.

To examine the suitability of ALE-seq for quantitative determination of eclDNA levels, we carried out a reconstruction experiment spiking 100 ng of genomic DNA from rice callus with differing amounts of PCR-amplified *Onsen* DNA from 1 ng to 100 fg (Fig. 2a–d). The results in Fig. 2a,b show that the read-outs of ALE-seq for *Onsen* correlate well with the input amounts ($R^2 = 0.99$). The initial ALE-seq steps of ligation and in vitro transcription impinged proportionally on the input DNA, resulting in unbiased quantification of the eclDNA and minimal quantitative
distortion of the final ALE-seq data. The levels of Tos17 were similar in all the spiked samples, indicating that addition of Onsen DNA did not influence the detection sensitivity of Tos17, at least for the amounts tested (Fig. 2c,d). Thus, ALE-seq can be used to determine eclDNA levels accurately.

Most rice retrotransposons harbour Arg-CCT PBS (Supplementary Fig. 1). We tested whether the reverse transcription reaction can be multiplexed to capture both types of retrotransposons (containing Arg-CCT or Met-iCAT PBS) and whether multiplexing of the reverse transcription primers compromises the sensitivity of the procedure. ALE-seq was performed on DNA from rice callus, testing each of the reverse transcription primers separately or as a mixture of both primers in a single reaction. As shown in Fig. 2e and Supplementary Fig. 5, the levels of Tos17 recorded in the samples with both primers were similar to the Met-iCAT primer alone. We also detected the eclDNAs of the RIRE2 element containing Arg-CCT PBS (Fig. 2f), which was known to be transpositionally active in rice callus7.

Identification of Go-on retrotransposon using ALE-seq. We next used ALE-seq to search for novel active rice retrotransposons. Since many plant retrotransposons are transcriptionally activated by abiotic stresses3,10, we subjected rice plants to heat stress before subjecting them to ALE-seq. In this way we identified a Copia-type retrotransposon able to synthesize eclDNA in the heat-stressed plants (Fig. 3a–c) and named this element Go-on (the Korean word

---

**Fig. 3 | Identification of a heat-activated retrotransposon in rice.** a, b, Genome-wide plots of rice ALE-seq results as in Fig. 1. Control (a) and heat-stressed (b) rice plants were used. One-week-old seedlings were subjected to heat stress (44 °C) for 3 days. Met-iCAT PBS primer was used in reverse transcription. The levels are shown as the number of reads mapped to retroelements. Three Go-on copies are indicated in b. c, Read coverage plot for Go-on3. d, RNA-seq data showing Go-on3 and a neighbouring gene. RNA-seq data were generated using the same plant materials as in a,b. The experiment was repeated independently two times with similar results. e–g, Cumulative plots for the number of non-reference insertions of Go-on (e), Tos17 (f) and Tos19 (g) in the genomes of 388 japonica and indica rice accessions. The statistical difference was determined by iterating random selection of 200 accessions out of 388 and performing the two-tailed Wilcoxon test. **P = 2.2 × 10^{-16}.”
for 'high temperature'). The three retrotransposons with the highest eclDNA levels in heat-stress conditions all belong to the Go-on family (Fig. 3b and Supplementary Fig. 6). Although, eclDNAs were detected for all three copies, Go-on3 seems to be the youngest and possibly the most active family member, containing identical LTRs and a complete ORF (Supplementary Fig. 6). As depicted in Supplementary Fig. 6, the 5′ LTR sequences of the three Go-on copies are identical; the ALE-seq reads derived from Go-on3 LTR were also cross-mapped to other copies that could be inactive or have reduced activities. To further determine whether sequences of Go-on LTRs recovered by ALE-seq are derived from Go-on3 or also from other family members, we performed an ALE-seq experiment using reverse transcription primers located downstream of the PBS, including sequences specific for each Go-on family member (Supplementary Fig. 6). The amplified ALE-seq products revealed that the eclDNAs produced in heat-stressed rice originated only from Go-on3 (Supplementary Fig. 6). We validated the production of eclDNAs of Go-on3 by sequencing the junction of the adapter and the 5′ end of LTR (Supplementary Fig. 6) and by quantitative PCR (qPCR; Supplementary Fig. 7).

Next, we examined whether Go-on3 is transcriptionally activated in rice subjected to heat stress. RNA-seq and the qPCR with reverse transcription (RT–qPCR) data clearly showed that Go-on is strongly activated in heat-stress conditions (Fig. 3d and Supplementary Fig. 7). Similar to many other retrotransposons including ONSEN of Arabidopsis \(^{9,11,12}\), the LTR sequence of Go-on3 contains cis-acting regulatory element such as the heat shock transcription factor HSFC1-binding sequence motif (Supplementary Fig. 7), which is suggestive of its heat stress-mediated transcriptional activation (Fig. 3d). To determine whether Go-on is also activated in indica
rice, we heat-stressed plants of cv. IR64 for three days and examined Go-on RNA and DNA levels. Similar to O. sativa ssp. japonica rice, Go-on RNA and DNA accumulated markedly under heat stress (Supplementary Fig. 8), suggesting that the trigger for Go-on activation is conserved in both of these evolutionarily distant rice genotypes. Analysis of the RNA-seq data from the heat-stressed rice plants revealed a poor correlation between the messenger RNA and ecDNA levels of retrotransposons (Supplementary Fig. 9). Given that ecDNAs captured by ALE-seq in Arabidopsis and rice (Fig. 1c–f and Supplementary Fig. 3) are all known well for their transposition competence, this agrees with the notion that the ecDNA level is a better predictor of retrotransposition than the RNA level.

To relate accumulation of Go-on copies in plant populations grown in different temperatures, we analysed the historical retrotransposition of Go-on using the genome resequencing data of rice accessions from the 3,000 Rice Genome Project11. First, we retrieved the raw sequencing data for all 388 japonica rice accessions and the same number of randomly selected sequences of indica rice accessions. Using the Transposon Insertion Finder (TIF) tool12, japonica and indica sequences were analysed for the number of Go-on copies and their genome-wide distribution. Only non-reference insertions that were absent in the reference genome were scored and the cumulative number of new insertions was plotted (Fig. 3c–g). Figure 3e shows that the indica rice population grown in a warmer climate13 accumulated significantly more Go-on copies than the japonica population. As controls, we also examined the accumulation of Tos17 and Tos19, which were not activated by heat stress in our ALE-seq profile (Fig. 3a,b). Both retrotransposons showed more transposition events in japonica than in indica rice (Fig. 3f,g and Supplementary Fig. 10). Therefore, the copy number of Go-on in rice accessions correlated with their growth temperatures, which could be related to occasional Go-on activation in elevated ambient temperatures.

**Identification of FIRE retrotransposon using ALE-seq.** It was reported previously that the tomato genome (Solanum lycopersicum) experiences a significant loss of DNA methylation in fruits during their maturation, which leads to transcriptional activation of retrotransposons14. However, it was not known whether these transcriptionally activated tomato transposons synthesize ecDNA. The ALE-seq strategy might not be sensitive enough to detect ecDNA in the ~950 Mb tomato genome, which is almost three times as large as ~400 Mb genome of rice15. To address these questions, ALE-seq was carried out on DNA samples from fruits at 52 days post anthesis (DPA), when the loss of DNA methylation is most pronounced16 and from leaves as a control. We used tomato cv. M82 for these experiments, as it is commonly used for genetic studies17,18 and the sequence of the current tomato reference genome is based on cv. Heinz 1706 (ref. 19). Since retrotransposon sequences and their chromosomal distributions differ greatly between genomes of different varieties within the same plant species20–22, we could not use the standard mapping-based annotation of the ALE-seq results. As a consequence, we developed a reference-free and alignment-based approach that adopts the clustering of reads based on their sequence similarities (Fig. 1b). The reads from both samples were pooled and then clustered by sequence homology (see Methods). The consensus of each cluster was determined and used as the reference in paired-end mapping. Subsequently, the consensus sequences were used for a BLAST search against the reference genome for the closest homologues. In this way, the BLAST search was able to map the clustered ALE-seq output to reference genome annotated retrotransposons, which are most similar to the ALE-seq recovered sequences. Applying this strategy, we identified a retroelement belonging to a Gypsy family (FIRE, Fruit-Induced RetroElement) that produces significant amounts of ecDNA at 52 DPA during fruit ripening (Fig. 4a,b). We also determined the transcript levels of the FIRE element in leaves and 52 DPA fruit samples. As shown in Fig. 4c, fruit RNA levels were enhanced twofold compared to leaves, where FIRE ecDNA was barely detectable (Fig. 4a). Finally, we found that the DNA methylation status of the FIRE element was lower in fruits than leaves in all three sequence contexts (Fig. 4d,f). In contrast, the DNA methylation levels of sequences directly flanking FIRE were similar in leaves and fruits (Fig. 4e–g).

**Discussion**

Recently, a novel active retrotransposon was identified in rice by sequencing extrachromosomal circular DNA (eccDNA) produced as a by-product of retrotransposition or by nuclear recombination reactions of ecDNAs23,24. Although the method of eccDNA sequencing has advantages over SIRT, such as increased sensitivity and the discovery of sequences of the entire element, it also has limitations. For example, the method requires relatively large amounts of starting material but still shows serious limits in efficiency and indicative power for retrotransposition. The method did not detect the eccDNA of Tos19 in rice callus, where this transposon is known to move25; however, direct comparison of both methods on the same biological samples was not performed. More importantly, eccDNAs may also be the result of genomic DNA recombination26 and these background products may be misleading when extrapolating to the transpositional potential of a previously unknown element. In this respect, ALE-seq is a greatly improved tool that largely overcomes the above-mentioned limitations of previous methods and requires only 100 ng of plant DNA.

The heat-responsiveness of Go-on, the novel heat-activated Copia family retrotransposon of rice detected using ALE-seq, seems to be conferred by cis-acting DNA elements embedded in the LTR, which are similar to the heat-activated Onsen retrotransposon in Arabidopsis11,12. Although heat stress can induce production of mRNA and ecDNA of Onsen, its retrotransposition is tightly controlled by the small interfering RNA pathway27. Given that real-time transposition of rice retrotransposons has only been detected in epigenetic mutants28,29 and triggered by tissue culture conditions causing vast alterations in the epigenome30, or as a result of interspecific hybridization31, an altered epigenomic status seems to be an important prerequisite for retrotransposition. We failed to detect transposed copies of Go-on in the progeny of heat-stressed rice plants. Although Go-on produces ecDNAs after heat stress, it may be mobilized only at low frequency in wild-type rice due to epigenetic restriction of retrotransposition. On an evolutionary scale, the higher number of new insertions of Go-on in indica rice populations grown at elevated temperatures might suggest its potential mobility.

Many retrotransposons are transcriptionally reactivated during specific developmental stages or in particular cell types28,32. In tomato, fruit pericarp exhibits a reduction in DNA methylation during ripening33. This is mainly attributed to higher transcription of the DEMETER-LIKE2 DNA glycosylase gene34. Despite massive transcriptional reactivation of retrotransposons in tomato fruits, it has been difficult to determine whether further steps toward transposition also take place. Using ALE-seq, we identified ecDNA that we annotated using a reference-free and alignment-based approach to a novel FIRE element. FIRE has 164 copies in the reference tomato genome and in a conventional mapping-based approach the ALE-seq reads of FIRE cross-mapped to multiple copies, making it difficult to assign ecDNA levels to particular family members (Supplementary Fig. 11). Therefore, our strategy can be used in situations where sequence of the reference genome is unavailable or the mapping of reads is hindered by the high complexity and multiplicity of the retrotransposon population.

ALE-seq could also be applied to non-plant systems. For example, numerous studies in various eukaryotes, including mammals, found that retrotransposons are transcriptionally activated by certain diseases or at particular stages during embryo development35,36.
It was also suggested that retrotransposition might be an important component of disease progression\(^1\). Given that the direct detection of retrotransposition is challenging, it would be interesting to use ALE-seq to determine whether such temporal relaxations of epigenetic transposon silencing also result in the production of the eclDNAs, as the direct precursor of the chromosomal integration of a retrotransposon.

**Methods**

**Plant materials.** Seeds of *Oryza sativa ssp. japonica* cv. Nipponbare and *Oryza sativa ssp. indica* cv. IR64 were surface-sterilized in 20% bleach for 15 min, rinsed three times with sterile water, and germinated on time-sequential Murashige and Skoog media. Rice plants were grown in 10 l light/14 h dark at 28 °C and 26 °C, respectively. For heat-stress experiments, 1-week-old rice plants were transferred to a growth chamber at 44 °C and 28 °C in light and dark, respectively. Rice callus was induced by the method used for rice transformation as previously described\(^2\).

**Tomato plants.** (*Solanum lycopersicum* cv. MR2) were grown under standard greenhouse conditions (16 h supplemental lighting of 880 ± 150 µmol m\(^{-2}\) s\(^{-1}\) and 8 h at 15 °C). Tomato leaf tissue samples were taken from 2-month-old plants. Tomato fruit pericarp tissues were harvested at 52 DPA.

**Annotation of LTR retrotransposons.** Functional de novo annotation of LTR retrotransposons for the genomes of *Arabidopsis thaliana* (TAIR10), *Solanum tuberosum* (SL2.50) and *Oryza sativa* (MSU7; tomato, SL2.50) using Bowtie2 (ref. 43) with minor optimization. In most short-read sequencing platforms, it is often difficult to assign the unmapped reads of transposable elements (TEs) to precise genomic location. However, as MiSeq outputs relatively longer reads, we presumed that ALE-seq reads have less ambiguity than other sequencing platforms and set the parameters dealing with multimappers to default. It is only the maximum fragment length option which is set to 0 by default that was manipulated. Transplanted to 3,000 (X 3,000). The numbers of reads mapped throughout each retrotransposon were counted by the featureCounts tool of the SubRead package\(^4\) using the custom annotation file created by LTRpred. Since featureCounts recognizes multimappers by SAM file's NH tag that bowtie2 does not generate, multimapped reads are counted as one read aligned to a single genomic location, which reduces quantitation bias that often happens to multimappers. IGV was used to visualize the sequencing data. For the alignment-based approach, the forward and reverse reads were merged to yield the full-length fragment sequences and converted to fasta files using the BBTools (https://jgi.doe.gov/data-and-tools/bbtools/). The fasta files for all the samples were concatenated to get a master fasta file that is later added to CD-HIT software\(^5\) to cluster the reads by sequence similarity with the following options: -c 0.95, -p 1, -1. CD-HIT outputs a fasta file of representative reads for each cluster. The resulting fasta file was used as reference for paired-end mapping of initial fastq files. The mapped reads were counted with the featureCounts tool. Those clusters that significantly differed in the number of mapped reads in different samples were further analysed for their identities using BLAST search.

**ALE-seq library preparation.** Genomic DNA was extracted using a DNeasy Plant Mini Kit (Qiagen) following the manufacturer’s instructions. Genomic DNA (100 ng) was used for adapter ligation with 4 µl of 50µM adapter DNA. After an overnight ligation reaction at 4 °C, the adapter-ligated DNA was purified by AMPure XP beads (Beckman Coulter) at a 1:5 ratio. In vitro transcription reactions were performed using a MEGAscript RNAi kit (Thermo Fisher) with minor modifications. Briefly, the reaction was carried out for 4 h at 37 °C and the template DNA was digested before RNA purification. Purified RNA (3 µg) was subjected to reverse transcription using a Transcriptor First Strand cDNA Synthesis Kit (Roche). Transcriptor First Strand cDNA Synthesis Kit was chosen because the RTase of the kit is thermostable. This allowed the reverse transcription-inhibiting RNA secondary structure formation. The custom reverse transcription primers were added as indicated for each experiment. After the reverse transcription reaction, 1 µl of RNase A/T1 (Thermo Fisher) was added to digest non-template RNA and the reaction mixture was incubated at 37 °C for at least 30 min. Single-stranded first strand cDNA was PCR-amplified by 25 cycles using Illumina TruSeq HT dual adapter primers and the PCR product was purified by AMPure XP beads (Beckman Coulter) at a 1:1 ratio. After purification, the eluted DNA was quantified using a KAPA Library Quantification Kit (KAPA Biosystems) and run on the MiSeq v2 × 300 bp platform in the Department of Pathology of the University of Cambridge. Due to the nature of ALE-seq flow specifically amplifies ecdDNAs, some ecdDNA-free samples did not produce enough library DNAs which, although suboptimal loading, were nevertheless sequenced. It is advisable to spike in PCR-amplified retrotransposon DNA as described below.

**Preparation of full-length *Oncen* DNA.** The full-length *Oncen* copy (AT1TE12295) was amplified using Phusion High-Fidelity DNA polymerase (New England Biolabs). PCR products were run on 1% agarose gels. The full-length fragment was then purified by QIAquick Gel Extraction (Qiagen) and its concentration was measured using the Qubit Fluorometric Quantitation system (Thermo Fisher). Primers used for amplification are listed in Supplementary Table 1. RT–qPCR analyses. Samples were ground in liquid nitrogen using mortar and pestle. Total RNA was isolated using TruSeq Plant RNA Kit (Illumina) to extract total RNA following the manufacturer’s instructions. The amount of extracted RNA was estimated using the Qubit Fluorometric Quantitation system (Thermo Fisher). Complementary DNAs were synthesized using a SuperScript Vilo cDNA Synthesis Kit (Invitrogen). RT–qPCR was performed in the LightCycler 480 system (Roche) using primers listed in Supplementary Table 1. LightCycler 480 SYBR green I master premix (Roche) was used to prepare the reaction mixture in a volume of 10 µl. The results were analysed by the ΔΔCt method.

**RNA-seq library construction.** Total RNA was prepared as described above. An Illumina TrueSeq Stranded messenger RNA Library Prep kit (Illumina) was used according to the manufacturer’s instructions. The resulting library was run on an Illumina NextSeq 500 machine (Illumina) in the Sainsbury Laboratory at the University of Cambridge.

**Analysis of next-generation sequencing data.** For RNA-seq data analysis, the adapter and the low-quality sequences were removed by Trimomatic software\(^6\). The cleaned reads were mapped to the MSU7 version of the rice reference genome (http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dr/) and our custom retrotransposon annotation. Visualization of sequencing data was performed using an Integrative Genomics Viewer (IGV)\(^7\).

For ALE-seq data analysis, the adapter sequence was removed from the raw reads using Trimomatic software. For the mapping-based approach, paired-end reads were mapped to the reference genomes (*Arabidopsis thaliana*, TAIR10; rice, MSU7; tomato, SL2.30) using Bowtie2 (ref. 19) with minor optimization. In most short-read sequencing platforms, it is often difficult to assign the unmapped reads of transposable elements (TEs) to precise genomic location. However, as MiSeq outputs relatively longer reads, we presumed that ALE-seq reads have less ambiguity than other sequencing platforms and set the parameters dealing with multimappers to default. It is only the maximum fragment length option which is set to 0 by default that was manipulated. Transplanted to 3,000 (X 3,000). The numbers of reads mapped throughout each retrotransposon were counted by the featureCounts tool of the SubRead package\(^4\) using the custom annotation file created by LTRpred. Since featureCounts recognizes multimappers by SAM file’s NH tag that bowtie2 does not generate, multimapped reads are counted as one read aligned to a single genomic location, which reduces quantitation bias that often happens to multimappers. IGV was used to visualize the sequencing data. For the alignment-based approach, the forward and reverse reads were merged to yield the full-length fragment sequences and converted to fasta files using the BBTools (https://jgi.doe.gov/data-and-tools/bbtools/). The fasta files for all the samples were concatenated to get a master fasta file that is later added to CD-HIT software\(^5\) to cluster the reads by sequence similarity with the following options: -c 0.95, -p 1, -1. CD-HIT outputs a fasta file of representative reads for each cluster. The resulting fasta file was used as reference for paired-end mapping of initial fastq files. The mapped reads were counted with the featureCounts tool. Those clusters that significantly differed in the number of mapped reads in different samples were further analysed for their identities using BLAST search.

**Detection of retrotransposon insertions.** The insertions of selected retrotransposons were detected from the genome resequencing data of *japonica* and *indica* rice accessions downloaded from the 3,000 rice genome project (http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dr/). The TIF program\(^8\) was used to identify insertions using the fastq files and detect newly integrated copies. We used MSU7 (http://rice.plantbiology.msu.edu/) and ShuHu498 (http://www.mbbbase.org) for the reference of *japonica* and *indica* rice, respectively. Only non-reference insertions were considered and common insertions found in multiple accessions were counted as a single retrotransposition event.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** The custom scripts used in this study are available in http://github.com/HajkJDALE.

**Data availability.** The next-generation sequencing data that support the findings of this study are available in the Sequence Read Archive (SRA) repository with the identifier SRP159920.
Received: 27 May 2018; Accepted: 7 November 2018; Published online: 10 December 2018

References

1. Lisch, D. How important are transposons for plant evolution? Nat. Rev. Genet. 14, 49–61 (2012).
2. Chuong, E. R., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18, 71–86 (2017).
3. Griffiths, J., Catoni, M., Iwasaki, M. & Paszkowski, J. Sequence-independent identification of active LTR retrotransposons in Arabidopsis. Mol. Plant 11, 508–511 (2017).
4. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl Acad. Sci. USA 101, 12404–12410 (2004).
5. Sanchez, D. H., Gaubert, H., Drost, H., Zabet, N. R. & Paszkowski, J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat. Commun. 8, 1283 (2017).
6. Picault, N. et al. Identification of an active LTR retrotransposon in rice. Plant J. 58, 754–765 (2009).
7. Sabot, F. et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J. 66, 241–246 (2011).
8. Mironova, M. et al. Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461, 427–430 (2009).
9. Ito, H. et al. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature 472, 115–119 (2011).
10. Paszkowski, J. Controlled activation of retrotransposition for plant breeding. Curr. Opin. Biotechnol. 32, 200–206 (2015).
11. Cavrak, V. V. et al. How a retrotransposon exploits the plant’s heat stress response for its activation. PLos Genet. 10, e1004115 (2014).
12. Pietzenuk, B. et al. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. Genome Biol. 17, 209 (2016).
13. SK RGP. The 3,000 rice genomes project. GigaScience 3, 7 (2014).
14. Nakagome, M. et al. Transposon insertion finder (TIF): a novel program for detection of de novo transpositions of transposable elements. BMC Bioinform. 15, 71 (2014).
15. Xiong, Z. Y. et al. Latitudinal distribution and differentiation of rice germplasm: its implications in breeding. Crop Sci. 51, 1050–1058 (2011).
16. Zhong, S. et al. Single-base resolution methylation of tomato fruit development reveal epigeneme modifications associated with ripening. Nat. Biotechnol. 31, 154–159 (2013).
17. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641 (2012).
18. Eshed, Y. & Zamir, D. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141, 1147–1162 (1995).
19. Eshed, Y. & Zamir, D. Less-than-additive epistatic interactions of quantitative trait loci in tomato. Genetics 143, 1807–1817 (1996).
20. Quadirana, L. et al. The Arabidopsis thaliana mobilome and its impact at the species level. eLife 5, e15716 (2016).
21. Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife 5, e20776 (2016).
22. Wei, B. et al. Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. BMC Genomics 17, 536 (2016).
23. Lanciano, S. et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposition activity in plants. PLos Genet. 13, e1006630 (2017).
24. Moller, H. D. et al. Formation of extrachromosomal circular DNA from long terminal repeats of retrotransposons in Saccharomyces cerevisiae. G3 (Bethesda) 6, 453–462 (2015).
25. Moller, H. D., Parsons, L., Jorgensen, T. S., Botstein, D. & Regenberg, B. Extrachromosomal circular DNA is common in yeast. Proc. Natl Acad. Sci. USA 112, 3114–3122 (2015).
26. Cheng, C. et al. Loss of function mutations in the rice chromomethylase OsCMT3a cause a burst of transposition. Plant J. 83, 1069–1081 (2015).
27. Cui, X. et al. Control of transposon activity by a histone H3K4 demethylase in rice. Proc. Natl Acad. Sci. USA 110, 1953–1958 (2013).
28. Wang, Z. H. et al. Genomewide variation in an introgression line of rice-zizania revealed by whole-genome re-sequencing. PLos ONE 8, e74479 (2013).
29. Li, H., Freeling, M. & Lisch, D. Epigenetic reprogramming during vegetative phase change in maize. Proc. Natl Acad. Sci. USA 107, 22184–22189 (2010).
30. Slotkin, R. K. et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 136, 461–472 (2009).
31. Liu, R. et al. A DEMETER-like DNA demethylase governs tomato fruit ripening. Proc. Natl Acad. Sci. USA 112, 10804–10809 (2015).
32. Goodier, J. L. Retrotransposition in tumors and brains. Mol. DNA 5, 11 (2014).
33. Baillie, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537 (2011).
34. Mullins, C. S. & Linnebacher, M. Human endogenous retroviruses and cancer: causality and therapeutic possibilities. World J. Gastroenterol. 18, 6027–6035 (2012).
35. Cho, J. & Paszkowski, J. Regulation of rice root development by a retrotransposon acting as a microRNA sponge. eLife 6, e30038 (2017).
36. Chan, P. P. & Lowe, T. M. GtRNAdb 2. 0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 44, 184–189 (2016).
37. Daujat, M. et al. PlantRNA, a database for tRNAs of photosynthetic eukaryotes. Nucleic Acids Res. 41, 273–279 (2012).
38. Finn, R. D. et al. Pfam: the protein families database. Nucleic Acids Res. 42, 222–230 (2014).
39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
40. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
41. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 516–520 (2010).
42. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
44. Liao, Y., Smyth, G. K. & Shi, W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41, e108 (2013).
45. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006).
46. Catoni, M. et al. DNA sequence properties that predict susceptibility to epilalic switching. EMBO J. 36, 617–628 (2017).
47. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics 27, 1571–1572 (2011).
48. Catoni, M., Tsang, J. M. F., Greco, A. P. & Zabet, N. R. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. Nucleic Acids Res. 46, e114 (2018).
49. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25, 1841–1842 (2009).

Acknowledgements

This work was supported by European Research Council (EVOBREED) grant no. 326261 and a Gatsby Fellowship grant no. AT3273/GLE.

Author contributions

J.C. and J.P. conceived the research. J.C., M.B., M.C., H.-G.D., A.B. and M.O. performed experiments. J.C., M.B., M.C. and H.-G.D. analysed data. J.C. and J.P. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0202-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.C. or J.P. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2018
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
  - State explicitly what error bars represent (e.g. SD, SE, CI)

Software and code

Policy information about availability of computer code

Data collection
- SRAttoolkit, v2.8.0
- Illumina Basespace

Our web collection on statistics for biologists may be useful.
Data analysis

- Trimmomatic, v0.32
- TopHat2, v2.1.1
- Cufflinks, v2.2.1
- Bowtie2, v2.1.0
- IGV, v2.3.20
- Subread, v1.6.0
- CD-HIT, v4.6.7
- Bismark, v0.19.1
- bedtools v2.24.0
- BBtools v37.16
- R v3.3.2
- DMRCaller v3.7
- TIF v2
- LTRpred, available at http://github.com/HajkD/ALE

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data generated in this study are deposited to Sequence Read Archive (SRA) and its identification is described in the Materials and Methods section as "SRP155920" in the current version of the manuscript.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No sample-size calculation was performed. Samples size was chosen according to the standard generally accepted in the field of plant molecular biology. |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | No data were excluded from the analyses. |
| Replication | All the experimental findings were reproducible in the independent biological replications. |
| Randomization | Plants were placed randomly in the growing facility. |
| Blinding | No blinding was applied. Most of the data in this study are generated by bioinformatic analyses. Since we applied identical settings to all the samples, blinding was not essential. |

Reporting for specific materials, systems and methods
| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a                             | n/a     |
| Involved in the study           | Involved in the study |
| □ Unique biological materials  | □ ChIP-seq |
| □ Antibodies                    | □ Flow cytometry |
| □ Eukaryotic cell lines         | □ MRI-based neuroimaging |
| □ Palaeontology                 |         |
| □ Animals and other organisms   |         |
| □ Human research participants   |         |