



HAL
open science

Sensitive detection of pre-integration intermediates of LTR retrotransposons in crop plants

Jungnam Cho, Matthias Benoit, Marco Catoni, Hajk-Georg Drost, Anna Brestovitsky, Matthijs Oosterbeek, Jerzy Paszkowski

► **To cite this version:**

Jungnam Cho, Matthias Benoit, Marco Catoni, Hajk-Georg Drost, Anna Brestovitsky, et al.. Sensitive detection of pre-integration intermediates of LTR retrotransposons in crop plants. *Nature Plants*, 2019, 5 (1), pp.26-33. 10.1038/s41477-018-0320-9 . hal-04171191

HAL Id: hal-04171191

<https://hal.science/hal-04171191>

Submitted on 22 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Title**

2 Sensitive detection of pre-integration intermediates of LTR retrotransposons in crop plants

3

4 **Authors**

5 Jungnam Cho*, Matthias Benoit, Marco Catoni, Hajk-Georg Drost, Anna Brestovitsky,

6 Matthijs Oosterbeek and Jerzy Paszkowski*

7 The Sainsbury Laboratory, University of Cambridge, Cambridge CB2 1LR, UK

8 *Corresponding author

9

10 **Correspondence:**

11 Jungnam Cho jungnam.cho@slcu.cam.ac.uk and Jerzy Paszkowski

12 jerzy.paszkowski@slcu.cam.ac.uk

13

14

15 **Abstract**

16 **Retrotransposons have played an important role in the evolution of host**
17 **genomes^{1,2}. Their impact on host chromosomes is mainly deduced from the composition**
18 **of DNA sequences, which have been fixed over evolutionary time. These studies provide**
19 **important “snapshots” reflecting historical activities of transposons but do not predict**
20 **current transposition potential. We previously reported Sequence-Independent**
21 **Retrotransposon Trapping (SIRT) as a methodology that, by identification of**
22 **extrachromosomal linear DNA (eclDNA), revealed the presence of active LTR**
23 **retrotransposons in *Arabidopsis*⁹. Unfortunately, SIRT cannot be applied to large and**
24 **transposon-rich genomes of crop plants. We have since developed an alternative**
25 **approach named ALE-seq (amplification of LTR of eclDNAs followed by sequencing). ALE-**
26 **seq reveals sequences of 5' LTRs of eclDNAs after two-step amplification: *in vitro***
27 **transcription and subsequent reverse transcription. Using ALE-seq in rice, we detected**
28 **eclDNAs for a novel *Copia* family LTR retrotransposon, *Go-on*, which is activated by heat**
29 **stress. Sequencing of rice accessions revealed that *Go-on* has preferentially accumulated**
30 **in *indica* rice grown at higher temperatures. Furthermore, ALE-seq applied to tomato**
31 **fruits identified a developmentally regulated *Gypsy* family of retrotransposons.**
32 **Importantly, a bioinformatic pipeline adapted for ALE-seq data analyses allows the direct**
33 **and reference-free annotation of new active retroelements. This pipeline allows**
34 **assessment of LTR retrotransposon activities in organisms for which genomic sequences**
35 **and/or reference genomes are unavailable or are of low quality.**

36

37 Chromosomal copies of activated retrotransposons containing long terminal repeats
38 (LTRs) are transcribed by RNA polymerase II, followed by reverse transcription of transcripts
39 to extrachromosomal linear DNAs (eclDNA); these integrate back into host chromosomes.
40 Because of the two obligatory template switches during reverse transcription, the newly
41 synthesized eclDNA is flanked by LTRs of identical sequence. Their subsequent divergence
42 due to the accumulation of mutations correlates well with length of time since the last
43 transposition, and thus transposon age³. However, the age of LTR retrotransposons cannot
44 be used to predict their current transpositional potential. Moreover, predictions are further

45 complicated by recombination events that occur with high frequency between young and
46 old members of a retrotransposon family⁴; thus old family members also contribute to the
47 formation of novel recombinant elements that insert into new chromosomal positions.
48 Although, retrotransposon activities can be easily measured at the transcriptional level^{5,6},
49 the presence of transcripts is a poor predictor of transpositional potential due to
50 posttranscriptional control of this process^{7,8}. In addition, direct detection of transposition by
51 genome-wide sequencing to identify new insertions is too expensive and time-consuming to
52 be applied as a screening method. Clearly, the development of an expeditious approach to
53 identify active retrotransposons that predict their transposition potential would be
54 welcomed. We previously described the SIRT strategy for *Arabidopsis* that led to the
55 identification of eclDNA of a novel retroelement and subsequent detection of new
56 insertions⁹. Thus, the presence of eclDNAs, the last pre-integration intermediate, was shown
57 to be a good predictor of retrotransposition potential.

58 Retrotransposons include a conserved sequence known as the primer binding site
59 (PBS), where binding of the 3' end of cognate tRNA initiates the reverse transcription
60 reaction⁹. Met-iCAT PBS was chosen for SIRT as it is the site involved in the majority of
61 annotated *Arabidopsis* retrotransposons⁹. To examine whether Met-iCAT PBS sequences are
62 also predominant in LTR retrotransposons of other plants, we used the custom-made
63 software *LTRpred* for *de novo* annotation of LTR retrotransposons in rice and tomato
64 genomes (see Materials and methods). Young retroelements were selected by filtering for at
65 least 95% identity between the two LTRs and subsequently examined for their cognate
66 tRNAs (Supplementary Figure 1a). As in *Arabidopsis*, around 80% of LTR retrotransposons in
67 the tomato genome contained Met-iCAT PBS (Supplementary Figure 1a). In contrast, only 30%
68 harboured Met-iCAT PBS in rice, and Arg-CCT PBS was involved in 60% of young LTR
69 retrotransposons (Supplementary Figure 1a). Nonetheless, we used Met-iCAT PBS in our
70 initial experiments because most retrotransposons known to be active in rice callus (e.g.
71 *Tos17* and *Tos19*) contain Met-iCAT PBS. Initially, SIRT was performed on DNA extracted
72 from rice leaves and callus; however, we did not detect eclDNAs for *Tos17* and *Tos19* in rice
73 tissues by this method (Supplementary Figure 2a and b). We reasoned that the short stretch
74 of PBS used for primer design in SIRT may have impaired PCR efficiency due to the many

75 PBS-related sequences present in larger genomes containing a high number of
76 retroelements, as is the case in rice.

77 To counter this problem, we developed an alternative method, named amplification
78 of LTR of eclDNAs followed by sequencing (ALE-seq), with significantly improved selectivity
79 and sensitivity of eclDNA detection. A crucial difference to SIRT is that ALE-seq amplification
80 of eclDNA is separated into two reactions: *in vitro* transcription and reverse transcription
81 (Figure 1a). This decoupling of the use of the two priming sequences by production of an
82 RNA intermediate is significantly more selective and efficient than the single PCR
83 amplification in SIRT.

84 ALE-seq starts with ligation to the ends of eclDNA of an adapter containing a T7
85 promoter sequence at its 5' end and subsequent *in vitro* transcription with T7 RNA
86 polymerase. The synthesized RNA is then reverse transcribed using the primer that binds
87 the transcripts at the PBS site. The adapter and the oligonucleotides priming reverse
88 transcription are anchored with partial Illumina adapter sequences, which allows the
89 amplified products to be directly deep-sequenced in a strand-specific manner. The ALE-seq-
90 sequences derived retrotransposon eclDNAs are predicted to contain the intact 5' LTR up to
91 the PBS site, flanked by Illumina paired-end sequencing adapters. We used the Illumina Mi-
92 seq platform for sequencing because its long reads of 300 bp from both ends cover the
93 entire LTR lengths of most potentially active elements. It is worth noting that the Illumina
94 adapters were tagged to the intact LTR DNA without fragmentation of the amplicons. This
95 together with the long reads of Mi-seq allowed us to reconstitute the complete LTR
96 sequences, even in the absence of the reference genome sequence. The reconstituted LTRs
97 were analysed using the alignment-based approach that complements the mapping-based
98 approach when the reference genome is incomplete (Figure 1b).

99 First, we tested ALE-seq on *Arabidopsis* by examining heat-stressed Col-0
100 *Arabidopsis* plants¹¹, *met1-1* mutant⁹ and epi12¹², a *met1*-derived epigenetic recombinant
101 inbred line. ALE-seq cleanly and precisely recovered sequences of complete LTRs for *Onsen*,
102 *Copia21* and *Evade* in samples containing their respective eclDNA (Supplementary Figure 3a
103 to g)^{8,9,11}. Due to priming of the reverse transcription reaction at PBS, the reads were
104 explicitly mapped to the 5' but not to the 3' LTR, although the two LTRs have identical

105 sequences. The ALE-seq reads have well-defined extremities, starting at the position
106 marking the start of LTRs and finishing at the PBS, which is consistent with their ecdDNA
107 origin. The ends of LTRs can also be inspected for conserved sequences that would further
108 confirm their ecdDNA origin (Supplementary Figure 1b). This reduced ambiguity of read
109 mapping in ALE-seq analysis, combined with the clear-cut detection of LTR ends, allows for
110 explicit and precise assignment of ALE-seq results to active LTR retrotransposons.

111 Since SIRT failed to detect ecdDNAs of rice retrotransposons known to be activated in
112 rice callus, we examined whether ALE-seq would identify their ecdDNAs. As shown in Figure
113 1c to f, ALE-seq unambiguously detected ecdDNAs of *Tos17* and *Tos19* in rice callus, but not
114 in leaf samples. To test whether detection of 5' LTR sequences requires the entire ALE-seq
115 procedure, we performed control experiments with depleted ALE-seq reactions, for example,
116 in the absence of enzymes for either ligation, *in vitro* transcription, or reverse transcription.
117 All incomplete procedures failed to produce sequences containing 5' LTRs derived from
118 ecdDNAs (Figure 1e and f). Taken together, the data show that ALE-seq can detect ecdDNAs
119 of LTR retrotransposons in *Arabidopsis* as well as in rice with considerably greater efficiency
120 than the SIRT method.

121 To examine the suitability of ALE-seq for quantitative determination of ecdDNA levels,
122 we carried out a reconstruction experiment spiking 100 ng of genomic DNA from rice callus
123 with differing amounts of PCR-amplified full-length *Onsen* DNA from 1 ng to 100 fg (Figure
124 2a to d). The results in Figure 2a and b show that the readouts of ALE-seq for *Onsen*
125 correlate well with the input amounts ($R^2=0.99$). The initial ALE-seq steps of ligation and *in*
126 *vitro* transcription impinged proportionally on the input DNA, resulting in unbiased
127 quantification of the ecdDNA and minimal quantitative distortion of the final ALE-seq data.
128 Noticeably, the levels of *Tos17* were similar in all the spiked samples, indicating that
129 addition of *Onsen* DNA did not influence the detection sensitivity of *Tos17*, at least for the
130 amounts tested (Figure 2c and d). Thus, ALE-seq can be used to accurately determine
131 ecdDNA levels.

132 Most rice retrotransposons harbour Arg-CCT PBS (Supplementary Figure 1a). We
133 tested whether the reverse transcription reaction can be multiplexed to capture both types
134 of retrotransposons (containing Arg-CCT or Met-iCAT PBS) and whether multiplexing of the

135 reverse transcription primers compromises the sensitivity of the procedure. ALE-seq was
136 performed on DNA from rice callus, testing each of the reverse transcription primers
137 separately or as a mixture of both primers in a single reaction. As shown in Figure 2e, the
138 levels of *Tos17* recorded in the samples with both primers were similar to the Met-iCAT
139 primer alone. Importantly, we also detected the ecdNAs of the *RIRE2* element containing
140 Arg-CCT PBS (Figure 2f), which was known to be transpositionally active in rice callus⁷.

141 We next used ALE-seq to search for novel active rice retrotransposons. Since many
142 plant retrotransposons are transcriptionally activated by abiotic stresses^{11,13}, we subjected
143 rice plants to heat stress before subjecting them to ALE-seq. In this way we identified a
144 *Copia*-type retrotransposon able to synthesize ecdDNA in the heat-stressed plants (Figure 3a
145 to c) and named this element *Go-on* (the Korean for 'high temperature'). The three
146 retrotransposons with the highest ecdDNA levels in heat-stress conditions all belong to the
147 *Go-on* family (Figure 3b and Supplementary Figure 4a). Although, ecdDNAs were detected for
148 all three copies, *Go-on3* seems to be the youngest and, thus, possibly the most active family
149 member, containing identical LTRs and a complete ORF (Supplementary Figure 4a). As
150 depicted in Supplementary Figure 4a, the 5' LTR sequences of the three *Go-on* copies are
151 identical; thus the ALE-seq reads derived from *Go-on3* LTR were also cross-mapped to other
152 copies that are possibly inactive or have reduced activities. To further determine whether
153 sequences of *Go-on* LTRs recovered by ALE-seq are indeed derived from *Go-on3* or also from
154 other family members, we performed an ALE-seq experiment using RT primers located
155 further downstream of the PBS, including sequences specific for each *Go-on* family member
156 (Supplementary Figure 4a). The amplified ALE-seq products revealed that the ecdDNAs
157 produced in heat-stressed rice originated only from *Go-on3* (Supplementary Figure 4b). We
158 validated the production of ecdDNAs of *Go-on3* by sequencing the junction of the adapter
159 and the 5' end of LTR (Supplementary Figure 4c) and by qPCR (Supplementary Figure 5a).

160 Next, we examined whether *Go-on3* is transcriptionally activated in rice subjected to
161 heat stress. RNA-seq and the RT-qPCR data clearly showed that *Go-on* is strongly activated
162 in heat-stress conditions (Figure 3d and Supplementary Figure 5b). The LTR sequence of *Go-*
163 *on3* contains the heat-responsive sequence motifs (Supplementary Figure 5c), which is
164 consistent with its heat stress-mediated transcriptional activation (Figure 3d). To determine
165 whether *Go-on* is also activated in *indica* rice, we heat-stressed plants of *IR64* for three days

166 and examined *Go-on* RNA and DNA levels. Similar to *japonica* rice, *Go-on* RNA and DNA
167 accumulated markedly under heat stress (Supplementary Figure 6a and b), suggesting that
168 the trigger for *Go-on* activation is conserved in both of these evolutionarily distant rice
169 genotypes. Analysis of the RNA-seq data from the heat-stressed rice plants revealed a poor
170 correlation between the mRNA and ecdDNA levels of retrotransposons (Supplementary
171 Figure 7a and b). This agrees with the notion that the ecdDNA level is a better predictor of
172 retrotransposition than the RNA level.

173 To determine whether *Go-on* proliferation increases in rice grown at elevated
174 temperatures, we analysed the historical retrotransposition of *Go-on* using the genome
175 resequencing data of rice accessions from the 3,000 Rice Genome Project¹⁴. First, we
176 retrieved the raw sequencing data for all 388 *japonica* rice accessions and the same number
177 of randomly selected sequences of *indica* rice accessions. Using the Transposon Insertion
178 Finder (TIF) tool¹⁵, *japonica* and *indica* sequences were analysed for the number of *Go-on*
179 copies and their genome-wide distribution. Only unique insertions that were absent in the
180 reference genome were scored and the cumulative number of new insertions was plotted
181 (Figure 3e to g). Figure 3e shows that the *indica* rice population grown in a warmer climate
182 accumulated significantly more *Go-on* copies than the *japonica* population. As controls, we
183 also examined the accumulation of *Tos17* and *Tos19*, which were not known to be activated
184 by heat stress. Both retrotransposons showed more transposition events in *japonica* than in
185 *indica* rice (Figure 3f and g). Therefore, *Go-on* as a heat-activated retroelement has
186 undergone specific accumulation in *indica* rice subjected to a warmer climate.

187 It was reported previously that the tomato genome experiences a significant loss of
188 DNA methylation in fruits during their maturation, which leads to transcriptional activation
189 of retrotransposons^{16,17}. However, it was not known whether these transcriptionally
190 activated tomato transposons synthesise ecdDNA. It was questionable whether the ALE-seq
191 strategy is sensitive enough to detect ecdDNA in the tomato genome, which is three times
192 larger than that of rice¹⁸. To address these questions, ALE-seq was carried out on DNA
193 samples from fruits at 52 days post anthesis (DPA), when the loss of DNA methylation is
194 most pronounced¹⁶, and from leaves as a control. It is important to note that we used
195 tomato cultivar (cv.) M82 for these experiments, as it is commonly used for genetic
196 studies^{19,20}, and that the sequence of the current tomato reference genome is based on cv.

197 Heinz 1706¹⁸. Since retrotransposon sequences and their chromosomal distributions differ
198 largely between genomes of different varieties within the same plant species²¹⁻²⁴, we could
199 not use the standard mapping-based annotation of the ALE-seq results. As a consequence,
200 we developed a reference-free and alignment-based approach that adopts the clustering of
201 reads based on their sequence similarities (Figure 1b). Briefly, the reads from both samples
202 were pooled and then clustered by sequence homology (See Materials and methods). The
203 consensus of each cluster was determined and used as the reference in paired-end mapping.
204 Subsequently, the consensus sequences were used for a BLAST search against the reference
205 genome for the closest homologues. In this way, the BLAST search was able to map the
206 clustered ALE-seq output to reference genome annotated retrotransposons, which are most
207 similar to the ALE-seq recovered sequences. Applying this strategy, we identified a
208 retroelement belonging to a *Gypsy* family (*FIRE*, *Fruit-Induced RetroElement*) that produces
209 significant amounts of ecDNA at 52 DPA during fruit ripening (Figure 4a). We also
210 determined the transcript levels of the *FIRE* element in leaves and 52 DPA fruit samples. As
211 shown in Figure 4b, fruit RNA levels were enhanced twofold compared to leaves, where *FIRE*
212 ecDNA was barely detectable (Figure 4a). Finally, we found that the DNA methylation status
213 of the *FIRE* element was lower in fruits than leaves in all three sequence contexts (Figure 4c
214 and e). In contrast, the DNA methylation levels of sequences directly flanking *FIRE* were
215 similar in leaves and fruits (Figure 4d to f).

216 Recently, a novel active retrotransposon was identified in rice by sequencing
217 extrachromosomal circular DNA (eccDNA) produced as a by-product of retrotransposition or
218 by nuclear recombination reactions of ecDNAs²⁵⁻²⁷. Although the method of eccDNA
219 sequencing has certain advantages over SIRT, such as increased sensitivity and the recovery
220 of sequences of the entire element, it also has certain limitations. For example, the method
221 requires relatively large amounts of starting material but still shows serious limits in
222 sensitivity and indicative power for retrotransposition. The method did not detect the
223 eccDNA of *Tos19* in rice callus, where this transposon is known to move²⁵. Most importantly,
224 eccDNAs may also be the result of genomic DNA recombination²⁷ and these background
225 products may be misleading when extrapolating to the transpositional potential of a
226 previously unknown element. In this respect, ALE-seq is a significantly improved tool that

227 largely overcomes the above-mentioned limitations of previous methods, and requires only
228 100 ng of plant DNA.

229 The heat-responsiveness of *Go-on*, the novel heat-activated *Copia* family
230 retrotransposon of rice detected using ALE-seq, seems to be conferred by *cis*-acting DNA
231 elements embedded in the LTR, which are similar to the heat-activated *Onsen*
232 retrotransposon in *Arabidopsis*¹¹. Although heat stress can induce production of mRNA and
233 ecdDNA of *Onsen*, its retrotransposition is tightly controlled by the small interfering RNA
234 pathway¹¹. Given that real-time transposition of rice retrotransposons has only been
235 detected in epigenetic mutants^{28,29} and triggered by tissue culture conditions causing vast
236 alterations in the epigenome^{7,30}, or as a result of interspecific hybridization³¹, an altered
237 epigenomic status seems to be an important prerequisite for retrotransposition. In fact, we
238 failed to detect transposed copies of *Go-on* in the progeny of heat-stressed rice plants. Thus,
239 although *Go-on* produces ecdDNAs after heat stress, it may be mobilized only at low
240 frequency in wild type rice due to epigenetic restriction of retrotransposition. Nevertheless,
241 on an evolutionary scale, the higher copy number of *Go-on* in *indica* rice populations grown
242 at elevated temperatures is compatible with potential mobility.

243 Many retrotransposons are transcriptionally reactivated during specific
244 developmental stages or in particular cell types³²⁻³⁴. In tomato, fruit pericarp exhibits a
245 reduction in DNA methylation during ripening. This is largely attributed to higher
246 transcription of the *DEMETER-LIKE2* DNA glycosylase gene^{17,35-37}. Despite massive
247 transcriptional reactivation of retrotransposons in tomato fruits, it has been difficult to
248 determine whether further steps toward transposition also take place. Using ALE-seq, we
249 identified ecdDNA that we annotated using a reference-free and alignment-based approach
250 to a novel *FIRE* element. *FIRE* has 164 copies in the reference tomato genome and in a
251 conventional mapping-based approach the ALE-seq reads of *FIRE* cross-mapped to multiple
252 copies, making it difficult to assign ecdDNA levels to particular family members. Therefore,
253 our annotation strategy can be used in situations where sequence of the reference genome
254 is unavailable or the mapping of reads is hindered by the high complexity and multiplicity of
255 the retrotransposon population.

256 ALE-seq could also be applied to non-plant systems. For example, numerous studies
257 in various eukaryotes, including mammals, found that retrotransposons are transcriptionally
258 activated by certain diseases or at particular stages during embryo development³⁸⁻⁴⁰. It was
259 also suggested that retrotransposition might be an important component of disease
260 progression⁴¹. Given that the direct detection of retrotransposition is challenging, it would
261 be interesting to use ALE-seq to determine whether such temporal relaxations of epigenetic
262 transposon silencing also result in the production of the ecdDNAs, as the direct precursor of
263 the chromosomal integration of a retrotransposon.

264 **Materials and methods**

265 Plant materials

266 Seeds of *Oryza sativa ssp. japonica cv. Nipponbare* and *Oryza sativa ssp. indica cv. IR64* were
267 surface-sterilized in 20% bleach for 15 min, rinsed three times with sterile water and
268 germinated on ½-MS media. Rice plants were grown in 10 h light / 14 h dark at 28°C and
269 26°C, respectively. For heat-stress experiments, 1-week-old rice plants were transferred to a
270 growth chamber at 44°C and 28°C in light and dark, respectively. Rice callus was induced by
271 the method used for rice transformation as previously described⁴².

272 Tomato plants (*Solanum lycopersicum cv. M82*) were grown under standard greenhouse
273 conditions (16 h supplemental lighting of 88 w/m² at 25°C and 8 h at 15°C). Tomato leaf
274 tissue samples were taken from 2-month-old plants. Tomato fruit pericarp tissues were
275 harvested at 52 days post anthesis (DPA).

276

277 Annotation of LTR retrotransposons

278 Functional *de novo* annotation of LTR retrotransposons for the genomes of TAIR10
279 (*Arabidopsis*), MSU7 (rice) and SL2.50 (tomato) was achieved by the *LTRpred* pipeline
280 (<https://github.com/HajkD/LTRpred>) using the parameter configuration: minlenltr = 100,
281 maxlenltr = 5000, mindistltr = 4000, maxdisltr = 30000, mintsd = 3, maxtsd = 20, vic = 80,
282 overlaps = "no", xdrop = 7, motifmis = 1, pbsradius = 60, pbsalilen = c(8,40), pbsoffset =
283 c(0,10), quality.filter = TRUE, n.orf = 0. The plant-specific tRNAs used to screen for primer
284 binding sites (PBS) were retrieved from GtRNAb⁴³ and plantRNA⁴⁴ and combined in a
285 custom fasta file. The hidden Markov model files for gag and pol protein conservation
286 screening were retrieved from Pfam⁴⁵ using the protein domains RdRP_1 (PF00680), RdRP_2
287 (PF00978), RdRP_3 (PF00998), RdRP_4 (PF02123), RVT_1 (PF00078), RVT_2 (PF07727),
288 Integrase DNA binding domain (PF00552), Integrase zinc binding domain (PF02022),
289 Retrotrans_gag (PF03732), RNase H (PF00075) and Integrase core domain (PF00665).
290 Computationally reproducible scripts for generating annotations can be found at
291 <http://github.com/HajkD/ALE>.

292

293 ALE-seq library preparation

294 Genomic DNA was extracted using a DNeasy Plant Mini Kit (Qiagen) following the
295 manufacturer's instruction. Genomic DNA (100 ng) was used for adapter ligation with 4 μ l of
296 50 μ M adapter DNA. After an overnight ligation reaction at 4°C, the adapter-ligated DNA
297 was purified by AMPure XP beads (Beckman Coulter) at a 1:0.5 ratio. *In vitro* transcription
298 reactions were performed using an MEGAscript RNAi kit (Thermofisher) with minor
299 modifications. Briefly, the reaction was carried out for 4 h at 37°C and RNase was omitted
300 from the digestion step prior to RNA purification. Purified RNA (3 μ g) was subjected to
301 reverse transcription (RT) using a Transcriptor First Strand cDNA Synthesis Kit (Roche). The
302 custom RT primers were added as indicated for each experiment. After the RT reaction, 1 μ l
303 of RNase A/T1 (Thermofisher) was added and the reaction mixture was incubated at 37°C
304 for at least 30 min. Single-stranded first strand cDNA was purified by AMPure XP beads
305 (Beckman Coulter) at a 1:1 ratio and PCR-amplified by 25 cycles using Illumina TruSeq HT
306 dual adapter primers. After purification, the eluted DNA was quantified using a KAPA Library
307 Quantification Kit (KAPA Biosystems) and run on the MiSeq v3 2 X 300 bp platform in the
308 Department of Pathology of the University of Cambridge. The oligonucleotide sequences are
309 provided in Supplementary Table 1.

310

311 Preparation of full-length *Onsen* DNA

312 The full-length *Onsen* copy (AT1TE12295) was amplified using Phusion High-Fidelity DNA
313 polymerase (New England Biolabs). PCR products were run on 1% agarose gels. The full-
314 length fragment was then purified by QIAquick Gel Extraction (Qiagen) and its concentration
315 measured using the Qubit Fluorometric Quantitation system (Thermo Fisher). Primers used
316 for amplification are listed in Supplementary Table 1.

317

318 RT-qPCR analyses

319 Samples were ground in liquid nitrogen using mortar and pestle. An RNeasy Plant Mini Kit
320 (Qiagen) was used to extract total RNA following the manufacturer's instructions. The
321 amount of extracted RNA was estimated using the Qubit Fluorometric Quantitation system
322 (Thermo Fisher). cDNAs were synthesized using a SuperScript VILO cDNA Synthesis Kit
323 (Invitrogen). Real-time quantitative PCR was performed in the LightCycler 480 system
324 (Roche) using primers listed in Supplementary Table 1. LightCycler 480 SYBR green I master
325 premix (Roche) was used to prepare the reaction mixture in a volume of 10 μ l. The results
326 were analysed by the $\Delta\Delta$ Ct method.

327

328 RNA-seq library construction

329 Total RNA was prepared as described above. An Illumina TruSeq Stranded mRNA Library
330 Prep kit (Illumina) was used according to the manufacturer's instructions. The resulting
331 library was run on an Illumina NextSeq 500 machine (Illumina) in the Sainsbury Laboratory
332 at the University of Cambridge.

333

334 Analysis of next-generation sequencing data

335 For RNA-seq data analysis, the adapter and the low-quality sequences were removed by
336 Trimmomatic software. The cleaned reads were mapped to the MSU7 version of the rice
337 reference genome using TopHat2. The resulting mapping files were processed to the
338 Cufflinks/Cuffquant/Cuffnorm pipeline for quantitation and visualized in an Integrative
339 Genomics Viewer (IGV).

340 For ALE-seq data analysis, the adapter sequence was removed from the raw reads using
341 Trimmomatic software. For the mapping-based approach, paired-end reads were mapped to
342 the reference genomes (Arabidopsis, TAIR10; rice, MSU7; tomato, SL2.50) using BOWTIE2
343 with minor optimization (-X 3000). The numbers of reads for each retrotransposon were
344 counted by the featureCounts tool of the SubRead package. IGV was used to visualize the
345 sequencing data. For the alignment-based approach, the forward and reverse reads were
346 merged and converted to fasta files. The fasta files created for all the samples were

347 concatenated and clustered by CD-HIT software with the following options: -c 0.95, -ap 1, -g
348 1. The resulting fasta file for the representative reads was used as the reference for paired-
349 end mapping. The mapped reads were counted with the featureCounts tool.

350 For Bisulfite sequencing analysis, raw sequenced reads derived from tomato fruits (52 DPA)
351 and leaves were downloaded from the public repository (SRP008329)¹⁶ and re-analysed as
352 previously described⁴⁶, with minor modifications. Briefly, high-quality sequenced reads were
353 mapped with Bismark⁴⁷ on the cv. Heinz 1706 reference genome (<https://solgenomics.net>),
354 including a chloroplast sequence obtained from GenBank database (NC_007898.3) to
355 estimate the conversion rate. After methylation call and correction for unconverted
356 cytosines, the methylation proportions at each cytosine position with a coverage of at least
357 3 reads were used to generate a bedGraph file for each cytosine context, using the R
358 Bioconductor packages DMRCaller
359 (<https://www.bioconductor.org/packages/release/bioc/html/DMRCaller.html>) and
360 Rtracklayer⁴⁸. The IGV browser was used to visualize the methylation profiles.

361

362 Detection of retrotransposon insertions

363 The insertions of selected retrotransposons were detected from the genome resequencing
364 data of *japonica* and *indica* rice accessions downloaded from the 3,000 rice genome project
365 (PRJEB6180). The Transposon Insertion Finder (TIF) program¹⁵ was used to identify the split
366 reads in the fastq files and detect newly integrated copies. We used MSU7
367 (<http://rice.plantbiology.msu.edu>) and ShuHui498 (<http://www.mbkbase.org>) for *japonica*
368 and *indica* rice, respectively. Only unique non-redundant insertions were considered.

369

370 Data accessibility

371 The next generation sequencing data generated in this study are deposited in the GEO
372 repository under accession numbers GSEXXXXX.

373

374 **References**

- 375 1. Lisch, D. How important are transposons for plant evolution? *Nat Rev Genet* **14**, 49–
376 61 (2012).
- 377 2. Chuong, E. B., Elde, N. C. & Feschotte, C. elements²: from conflicts to benefits. *Nat.*
378 *Publ. Gr.* **18**, 71–86 (2017).
- 379 3. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes.
380 *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12404–10 (2004).
- 381 4. Sanchez, D. H., Gaubert, H., Drost, H., Zabet, N. R. & Paszkowski, J. High-frequency
382 recombination between members of an LTR retrotransposon family during
383 transposition bursts. *Nat. Commun.* **8**, 1–6 (2017).
- 384 5. Picault, N. *et al.* Identification of an active LTR retrotransposon in rice. *Plant J.* **58**,
385 754–765 (2009).
- 386 6. Cavrak, V. V. *et al.* How a Retrotransposon Exploits the Plant’s Heat Stress Response
387 for Its Activation. *PLoS Genet.* **10**, (2014).
- 388 7. Sabot, F. *et al.* Transpositional landscape of the rice genome revealed by paired-end
389 mapping of high-throughput re-sequencing data. *Plant J.* **66**, 241–246 (2011).
- 390 8. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in Arabidopsis.
391 *Nature* **461**, 1–5 (2009).
- 392 9. Griffiths, J., Catoni, M., Iwasaki, M. & Paszkowski, J. Sequence-Independent
393 Identification of Active LTR Retrotransposons in Arabidopsis. *Mol. Plant* 1–4 (2017).
394 doi:10.1016/j.molp.2017.10.012
- 395 10. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering
396 plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
- 397 11. Ito, H. *et al.* An siRNA pathway prevents transgenerational retrotransposition in
398 plants subjected to stress. *Nature* **472**, 115–9 (2011).
- 399 12. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in Arabidopsis.

- 400 *Nature* **461**, 427–430 (2009).
- 401 13. Paszkowski, J. Controlled activation of retrotransposition for plant breeding. *Curr.*
402 *Opin. Biotechnol.* **32**, 200–206 (2015).
- 403 14. The 3, 000 rice genomes project. The 3 , 000 rice genomes project. *Gigascience* **3**, 1–6
404 (2014).
- 405 15. Nakagome, M. *et al.* Transposon Insertion Finder (TIF): a novel program for detection
406 of de novo transpositions of transposable elements. *BMC Bioinformatics* **15**, 71
407 (2014).
- 408 16. Zhong, S. *et al.* Single-base resolution methylomes of tomato fruit development
409 reveal epigenome modifications associated with ripening. *Nat. Biotechnol.* **31**, 154–9
410 (2013).
- 411 17. Cristina, R. *et al.* Diversity , distribution and dynamics of full-length Copia and Gypsy
412 LTR retroelements in *Solanum lycopersicum*. *Genetica* **145**, 417–430 (2017).
- 413 18. Consortium, T. tomato genome. The tomato genome sequence provides insights into
414 fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
- 415 19. Eshed, Y. & Zamir, D. An Introgression Line Population of *Lycopersicon pennellii* in the
416 Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated
417 QTL. *Genetics* **141**, 1147–1162 (1995).
- 418 20. Eshed, Y. & Zamir, D. Less-Than-Additive Epistatic Interactions of Quantitative Trait
419 Loci in Tomato. *Genetics* **143**, 1807–1817 (1996).
- 420 21. Quadrana, L. *et al.* The *Arabidopsis thaliana* mobilome and its impact at the species
421 level. *Elife* **5**, 1–25 (2016).
- 422 22. Stuart, T. *et al.* Population scale mapping of transposable element diversity reveals
423 links to gene regulation and epigenomic variation. *Elife* **5**, 1–27 (2016).
- 424 23. Wei, B. *et al.* Genome-wide characterization of non-reference transposons in crops
425 suggests non-random insertion. *BMC Genomics* **17**, 1–13 (2016).

- 426 24. Zhang, Q. & Gao, L. Rapid and Recent Evolution of LTR Retrotransposons Drives Rice
427 Genome Evolution During the Speciation of AA-Genome *Oryza* Species. *G3 (Bethesda)*.
428 **7**, 1875–1885 (2017).
- 429 25. Lanciano, S. *et al.* Sequencing the extrachromosomal circular mobilome reveals
430 retrotransposon activity in plants. *PLOS Genetics* **13**, (2017).
- 431 26. Møller, H. D. *et al.* Formation of Extrachromosomal Circular DNA from Long Terminal
432 Repeats of Retrotransposons in *Saccharomyces cerevisiae*. *G3 (Bethesda)*. **6**, 453–62
433 (2015).
- 434 27. Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D. & Regenberg, B.
435 Extrachromosomal circular DNA is common in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
436 E3114–22 (2015).
- 437 28. Cheng, C. *et al.* Loss of function mutations in the rice chromomethylase OsCMT3a
438 cause a burst of transposition. *Plant J.* **83**, 1069–1081 (2015).
- 439 29. Cui, X. *et al.* Control of transposon activity by a histone H3K4 demethylase in rice.
440 *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1953–8 (2013).
- 441 30. Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. & Kanda, M. Retrotransposons of
442 rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. U. S. A.* **93**,
443 7783–7788 (1996).
- 444 31. Wang, Z. H. *et al.* Genomewide Variation in an Introgression Line of Rice-Zizania
445 Revealed by Whole-Genome re-Sequencing. *PLoS One* **8**, 1–12 (2013).
- 446 32. Li, H., Freeling, M. & Lisch, D. Epigenetic reprogramming during vegetative phase
447 change in maize. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 22184–22189 (2010).
- 448 33. Slotkin, R. K. *et al.* Epigenetic Reprogramming and Small RNA Silencing of
449 Transposable Elements in Pollen. *Cell* **136**, 461–472 (2009).
- 450 34. Hsieh, T. F. *et al.* Genome-wide demethylation of *Arabidopsis* endosperm. *Science*
451 *(80-)*. **324**, 1451–1454 (2009).
- 452 35. Bernacchia, E. T. G., How, S. M. A. & Gallusci, L. S. D. R. P. Tissue dependent variations

- 453 of DNA methylation and endoreduplication levels during tomato fruit development
454 and ripening. *Planta* **228**, 391–399 (2008).
- 455 36. Cao, D. *et al.* Genome-wide identification of cytosine-5 DNA methyltransferases and
456 demethylases in *Solanum lycopersicum*. *Gene* **550**, 230–237 (2014).
- 457 37. Liu, R. *et al.* A DEMETER-like DNA demethylase governs tomato fruit ripening. *Proc.*
458 *Natl. Acad. Sci.* **112**, 10804–10809 (2015).
- 459 38. Goodier, J. L. Retrotransposition in tumors and brains. *Mob. DNA* **5**, 11 (2014).
- 460 39. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the
461 human brain. *Nature* **479**, 534–7 (2011).
- 462 40. Xie, Y., Rosser, J. M., Thompson, T. L., Boeke, J. D. & An, W. Characterization of L1
463 retrotransposition with high-throughput dual-luciferase assays. *Nucleic Acids Res.* **39**,
464 1–11 (2011).
- 465 41. Mullins, C. S. & Linnebacher, M. Human endogenous retroviruses and cancer:
466 Causality and therapeutic possibilities. *World J. Gastroenterol.* **18**, 6027–6035 (2012).
- 467 42. Cho, J. & Paszkowski, J. Regulation of rice root development by a retrotransposon
468 acting as a microRNA sponge. *Elife* 1–21 (2017).
- 469 43. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA
470 genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, 184–189
471 (2016).
- 472 44. Daujat, M. *et al.* PlantRNA, a database for tRNAs of photosynthetic eukaryotes.
473 *Nucleic Acids Res.* **41**, 273–279 (2012).
- 474 45. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, 222–230
475 (2014).
- 476 46. Catoni, M. *et al.* DNA sequence properties that predict susceptibility to epiallelic
477 switching. *EMBO* **36**, 617–628 (2017).
- 478 47. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for

- 479 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
- 480 48. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer[®]: an R package for interfacing
481 with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

482 **Figure legends**

483 Figure 1. Detection of ecDNA by ALE-seq

484 **a**, The workflow of ALE-seq. The colour code is indicated in a box. **b**, Analysis pipeline of ALE-
485 seq results. The sequenced reads can be mapped to the reference genome or aligned to
486 each other to obtain a cluster consensus. **c** and **d**, Genome-wide plots of rice ALE-seq results
487 from leaf (**c**) and callus (**d**). The levels are shown as number of reads mapped to each
488 retrotransposon. Dots represent annotated retrotransposons; those corresponding to *Tos17*
489 and *Tos19* are indicated. **e** and **f**, Read plots mapped to *Tos17* (**e**) and *Tos19* (**f**). The black
490 bars represent retrotransposons and white arrowheads indicate LTRs.

491

492 Figure 2. Sensitivity and specificity of ecDNA detection by ALE-seq

493 **a-d**, ALE-seq reconstruction experiment with varying amounts of PCR-amplified *Onsen* DNA
494 added to rice callus DNA. Genome browser image with the read coverage (**a** and **c**) and
495 quantitated read counts (**b** and **d**) for *Onsen* (**a** and **b**) and *Tos17* (**c** and **d**) loci. The amounts
496 of *Onsen* DNA added were 1 ng, 100 pg, 10 pg, 1 pg or 100 fg; 100 ng of rice callus DNA was
497 used. Note that read coverage values are Log10-converted in **a**. For **b** and **d**, values are
498 shown as Log10-converted counts per million reads. **e** and **f**, Read coverage plots for the
499 ALE-seq of rice callus using different RT primers. *Tos17* and *RIRE2* transposons depicted
500 below the plots as in Figure 1.

501

502 Figure 3. Identification of a novel heat-activated retrotransposon in rice

503 **a** and **b**, Genome-wide plots of rice ALE-seq results as in Figure 1. Control (**a**) and heat-
504 stressed (**b**) rice plants were used. One-week-old seedlings were subjected to heat stress
505 (44°C) for 3 days. The levels are shown as the number of reads mapped to the
506 retroelements. Three *Go-on* copies are indicated in **b**. **c**, Read coverage plot for *Go-on3*. **d**,
507 RNA-seq data showing *Go-on3* and a neighbouring gene. RNA-seq data were generated
508 using the same plant materials as in **a** and **b**. **e-g**, Cumulative plots for the number of unique

509 insertions of *Go-on* (**e**), *Tos17* (**f**), and *Tos19* (**g**) in the genomes of 388 *japonica* and *indica*
510 rice accessions.

511

512 Figure 4. Identification of a tomato retrotransposon activated in fruit pericarp

513 **a**, Read coverage plot for the *FIRE* retrotransposon identified in tomato fruit pericarp. **b**, The
514 RNA levels of *FIRE* in leaves and fruits determined by RT-qPCR. The levels are means \pm
515 standard deviation (sd) of two biological replicates performed with three technical repeats
516 each. Normalization was done against *SICAC* (Solyc08g006960). **c**, Genome browser image
517 for the DNA methylation levels at *FIRE* elements in leaves and fruits of tomato. **d-f**,
518 Quantitation of DNA methylation levels. The levels are the averages of percent DNA
519 methylation in the indicated regions. The upstream and downstream regions are immediate
520 flanking sequences with the same length as *FIRE*.

521

522 Supplementary Figure 1. PBS and LTR terminal sequences of LTR retrotransposons in
523 *Arabidopsis*, rice and tomato

524 **a**, The frequency of tRNAs used for targeting PBS. LTR retrotransposons were annotated by
525 *LTRpred* and selected for young elements by filtering LTR similarities higher than 95%. The
526 total numbers of retrotransposons analysed in each species are shown below. **b**, The
527 conserved sequences of 5' and 3' ends of LTR. The first and last five nucleotides of LTRs are
528 displayed. The images were generated by the WebLogo tool
529 (<http://weblogo.berkeley.edu/logo.cgi>).

530

531 Supplementary Figure 2. SIRT results from leaves and calli of rice

532 **a**, Genome-wide plots for SIRT performed in leaves and **(b)** in calli of rice.

533

534 Supplementary Figure 3. Ale-seq detection of eclDNAs of *Arabidopsis* retrotransposons.

535 Genome-wide plots (**a**, **b**, **d** and **f**) and read coverage plots (**c**, **e** and **g**) for ALE-seq profiles of
536 *Arabidopsis* Col-0 wt (**a**), heat-stressed Col-0 (**b** and **c**), *met1-1* (**d** and **e**), and epi12 (**f** and **g**).

537

538 Supplementary Figure 4. *Go-on* retrotransposon family

539 **a**, Schematic structure of *Go-on* retrotransposons. The genomic coordinates and LTR
540 similarities of each copy are shown at the left and right, respectively. Red boxes, ORFs; blue
541 boxes, PBS; white arrowheads, LTRs. Note that the sequences of the upstream LTRs through
542 the PBS are identical in all three copies. The sequence variation specific for each element is
543 indicated. Primers used for sequencing and qPCR analyses are shown as arrows. **b**, Multiple
544 sequence alignment of the genomic sequences of three *Go-on* copies and the sequenced
545 ALE clones. ALE-seq was performed using the RT primer specific to *Go-on3* indicated as “a”
546 in **a**. The resulting single-stranded first strand cDNA was PCR-amplified, cloned to the pGEM
547 T-easy vector, and sequenced. Multiple sequence alignment was performed by ClustalW
548 (<http://www.genome.jp/tools-bin/clustalw>) and visualized by boxshade tools
549 (https://www.ch.embnet.org/software/BOX_form.html). **c**, Sequencing of the ALE-seq
550 product of *Go-on3* showing the junction region of the adapter and LTR. Sequences in red
551 and black are the adapter and *Go-on* LTR, respectively.

552

553 Supplementary Figure 5. Heat stress-triggered transcriptional activation of *Go-on*

554 **a** and **b**, The relative levels of DNA (**a**) and RNA (**b**) of *Go-on3* determined by qPCR. Heat
555 treatment (44°C) was applied to 1-week-old rice seedlings for the periods indicated; +3r
556 means 3 days of recovery in normal growth conditions after heat stress. The levels are
557 means \pm sd of three biological replicates performed in three technical repeats. For DNA
558 analysis, Day 0 levels are set to 3, reflecting three genomic copies of *Go-on* in *japonica* rice.
559 Normalization was done against *eEF1 α* . Asterisks represent significant statistical difference
560 as determined by Student’s t-test. ** $P < 0.005$; * $P < 0.05$; n.s., not significant. **c**, The
561 sequence of the left LTR and PBS of *Go-on3*. The sequences in red are the heat-related cis-
562 acting sequence motifs predicted by PlantPan 2.0 tool
563 (<http://plantpan2.itps.ncku.edu.tw/index.html>). The PBS is shown in blue.

564

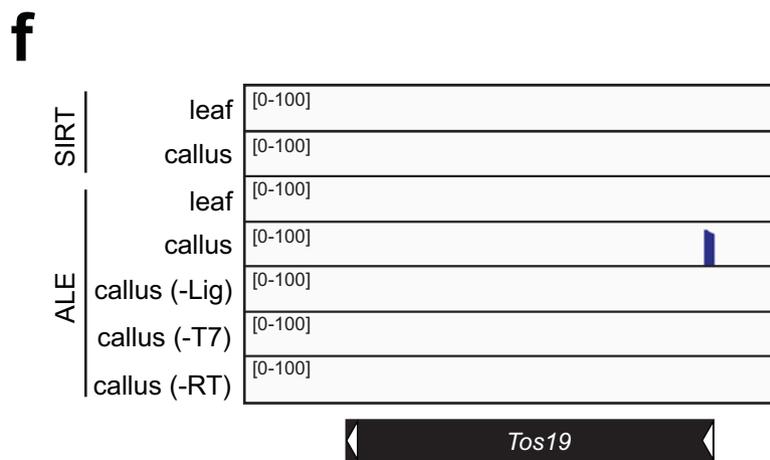
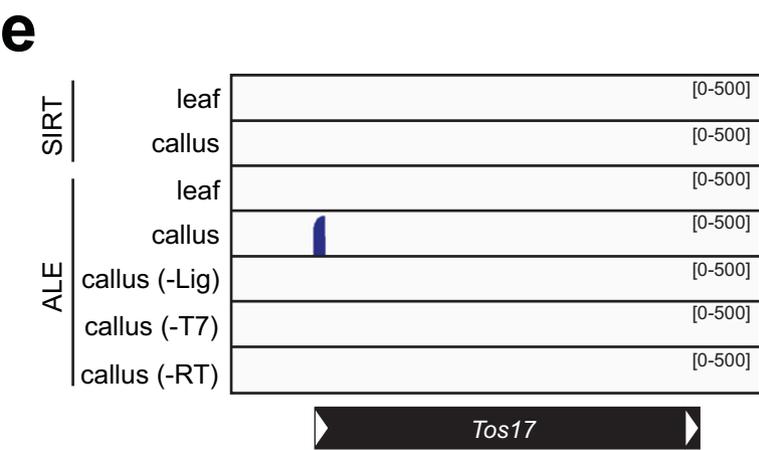
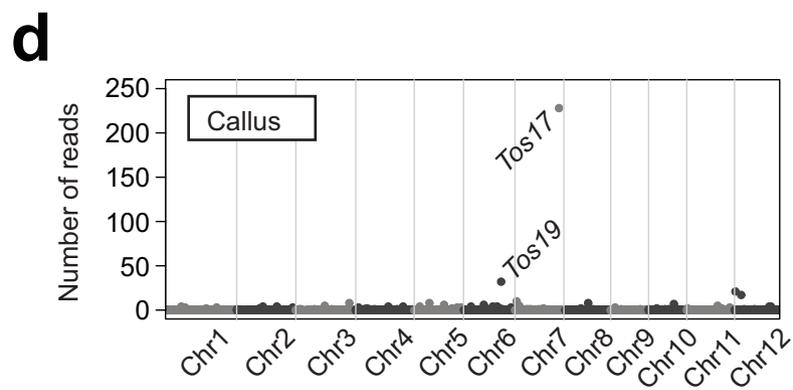
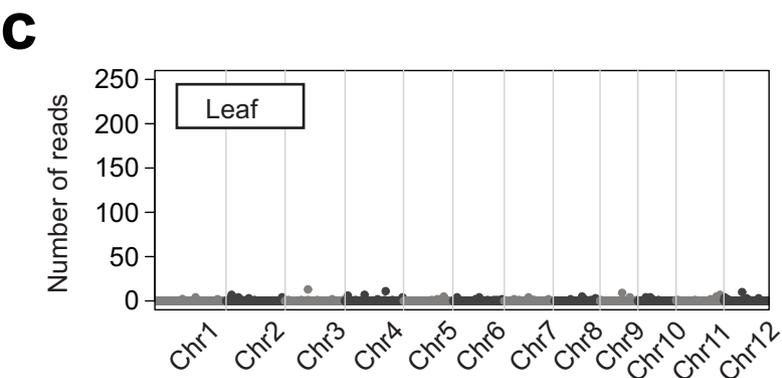
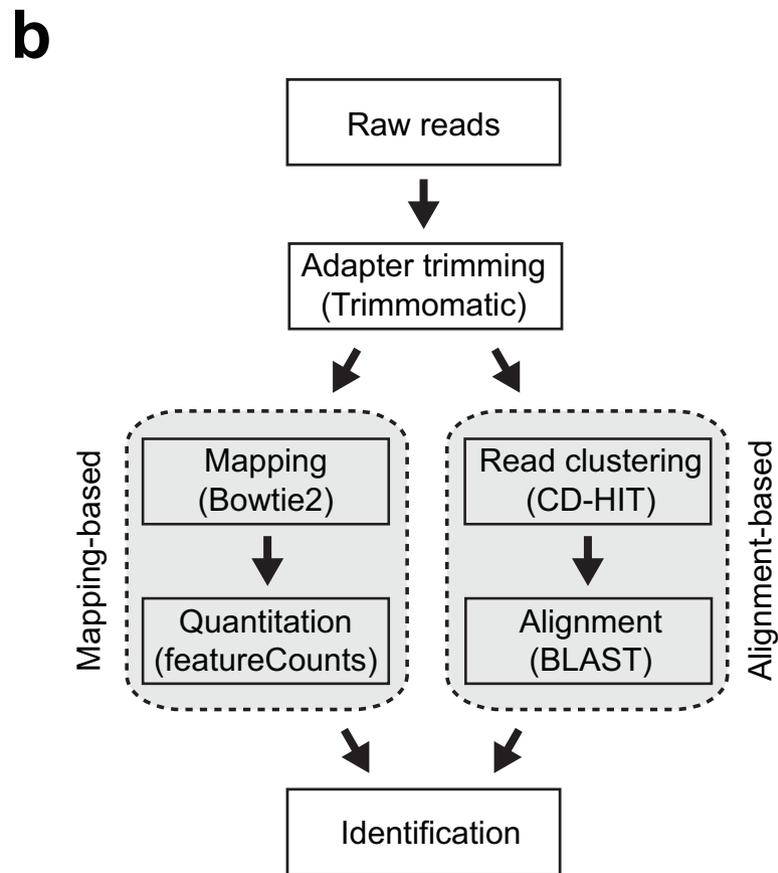
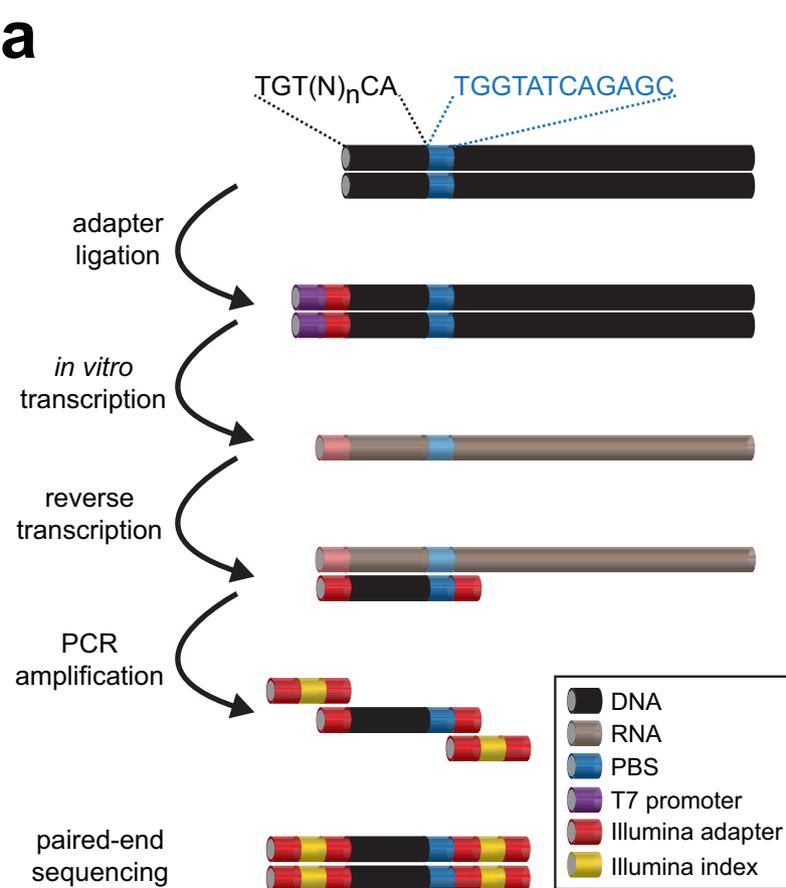
565 Supplementary Figure 6. Heat stress-triggered activation of *Go-on* in *indica* rice

566 **a** and **b**, qPCR analyses for DNA (**a**) and RNA (**b**) levels of *Go-on* in *indica* rice. The levels are
567 means \pm sd of three technical repeats. The levels of replicate 1 of control sample are set to 2
568 (**a**) reflecting 2 genomic copies of *Go-on* in *indica* rice. Asterisks represent significant
569 statistical difference as determined by Student's t-test. $**P < 0.005$.

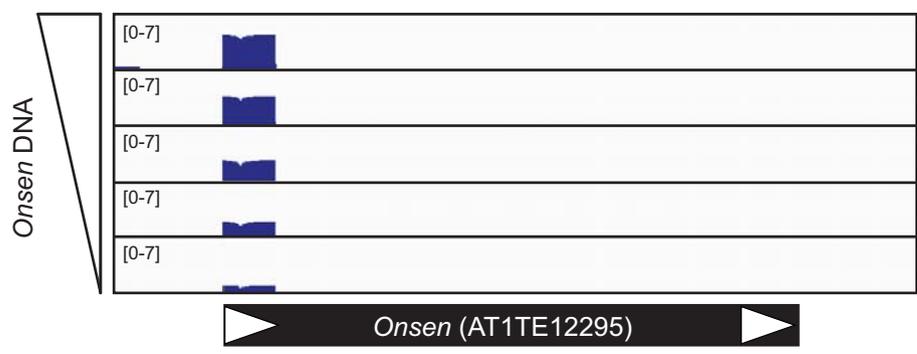
570

571 Supplementary Figure 7. Comparison of mRNA and ecDNA levels

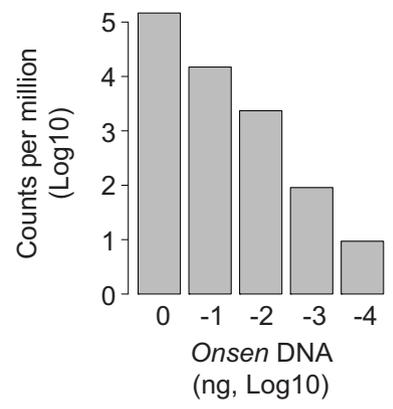
572 **a**, Scatter plot for fold changes in RNA-seq and ALE-seq profiles in the control and
573 heat-stressed rice plants. Each dot represents an individual retroelement. Three *Go-on*
574 retroelements are circled in red. **b**, Read coverage plot for a selected retrotransposon
575 showing evidence of transcriptional activation upon heat stress not followed by synthesis of
576 ecDNAs.

Figure 1.

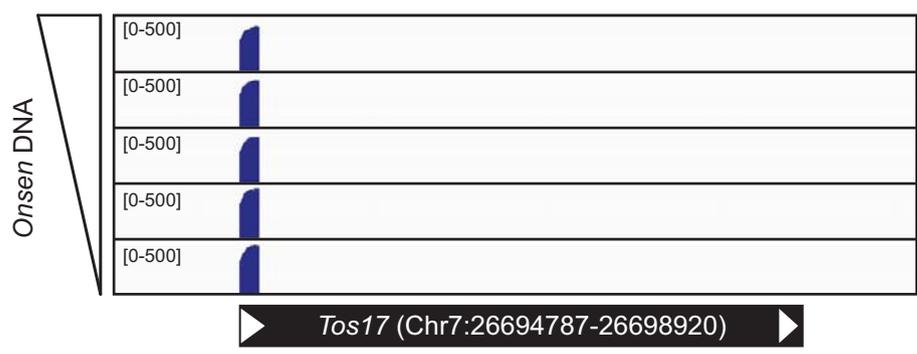
a



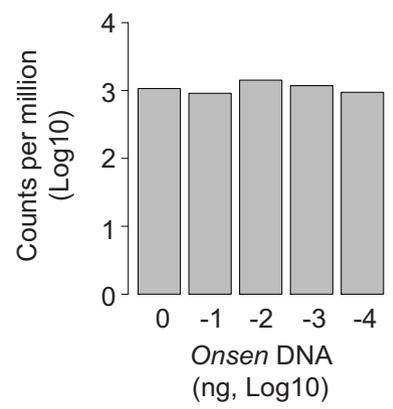
b



c



d



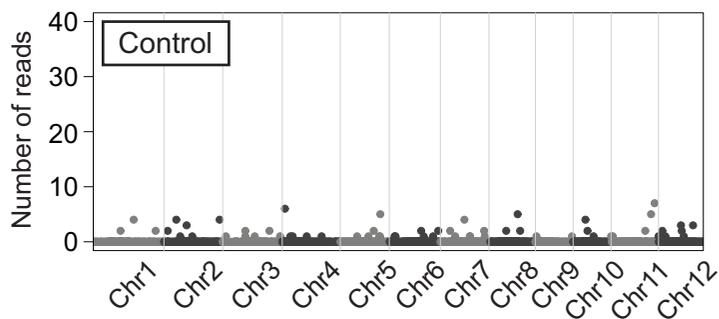
e



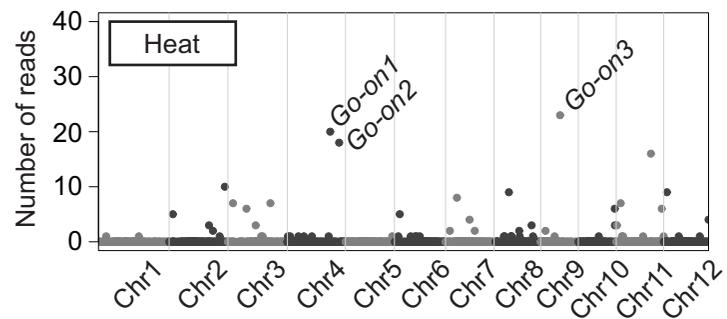
f



a



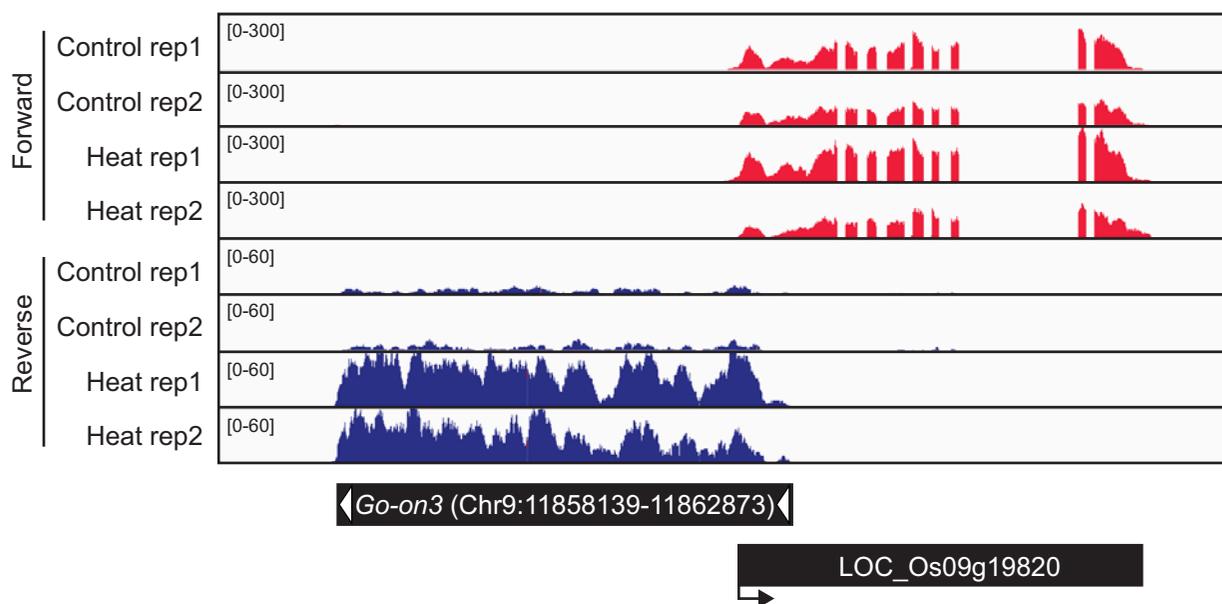
b



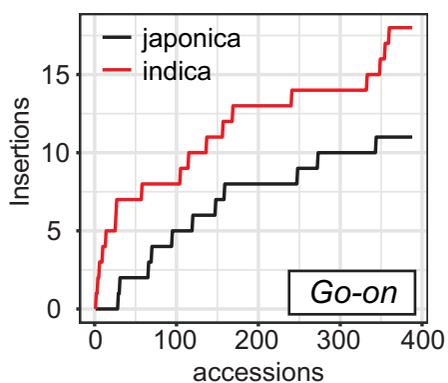
c



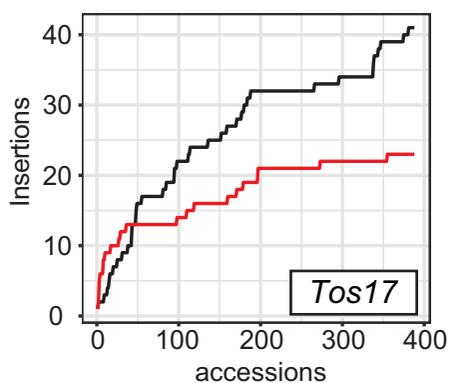
d



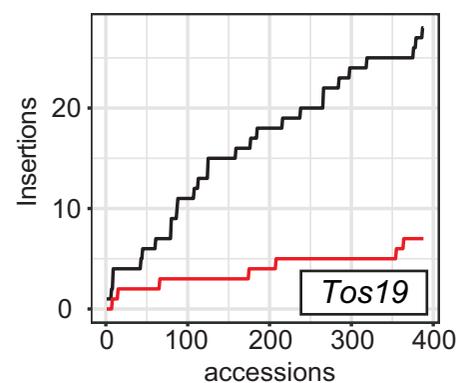
e



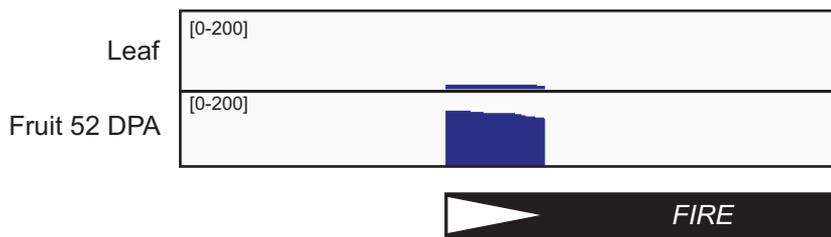
f



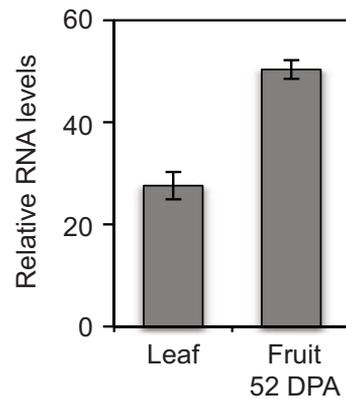
g



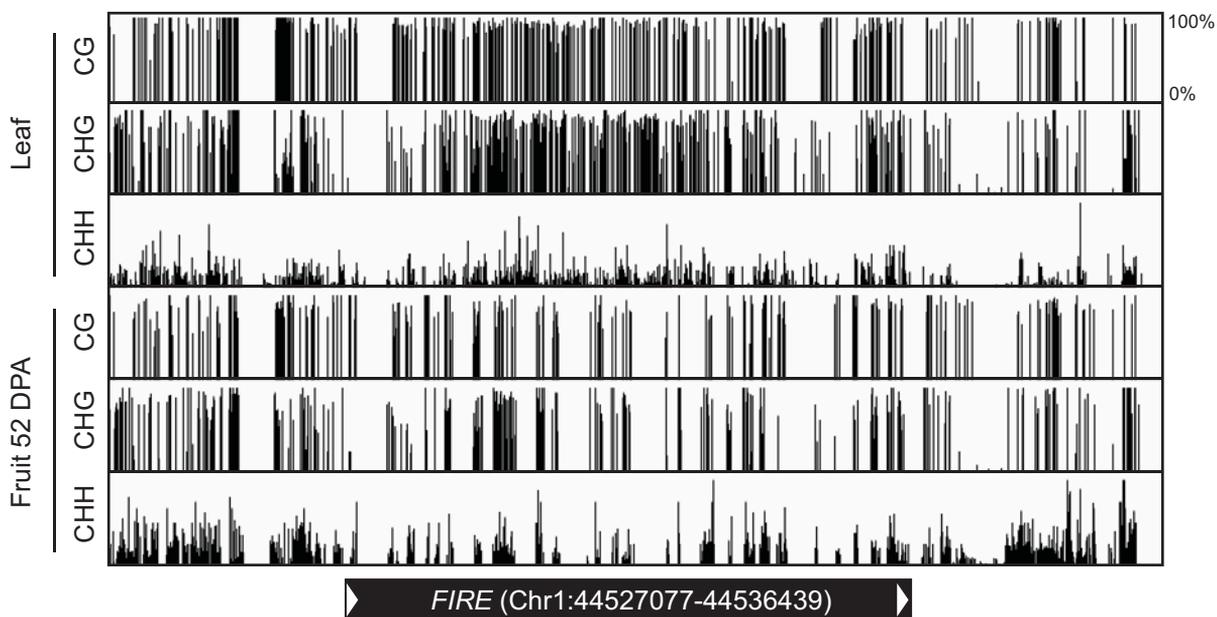
a



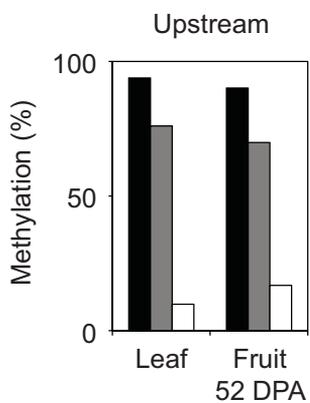
b



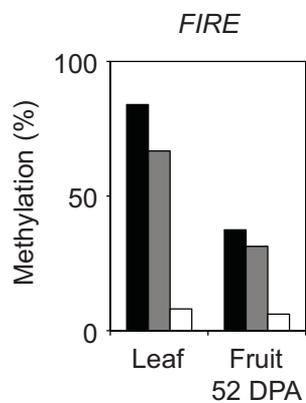
c



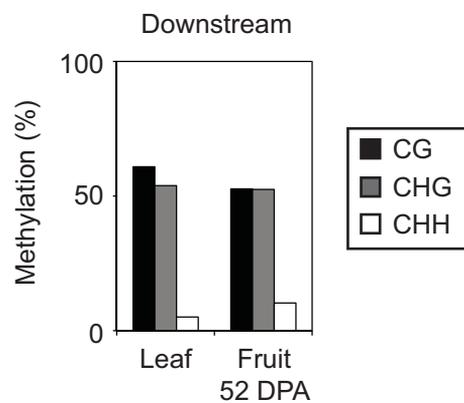
d



e

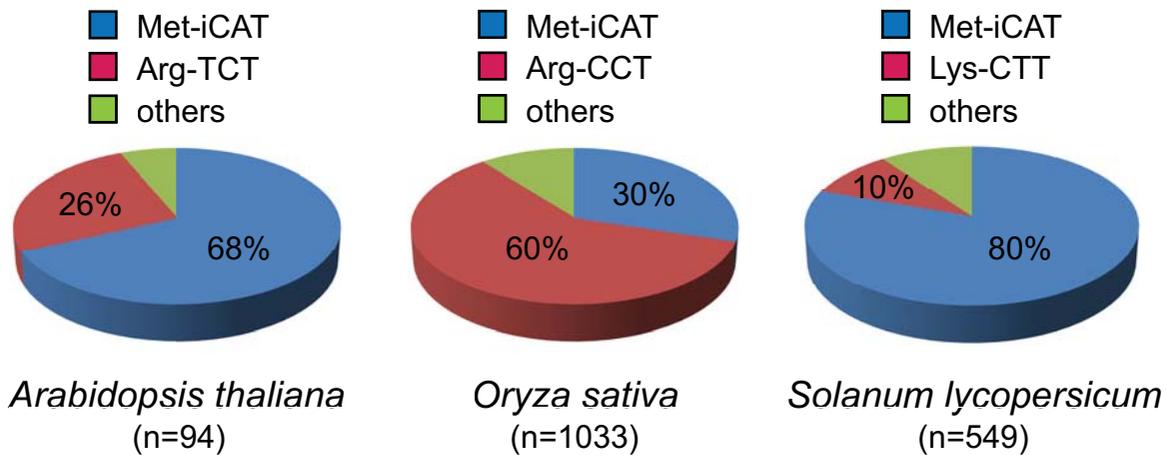


f

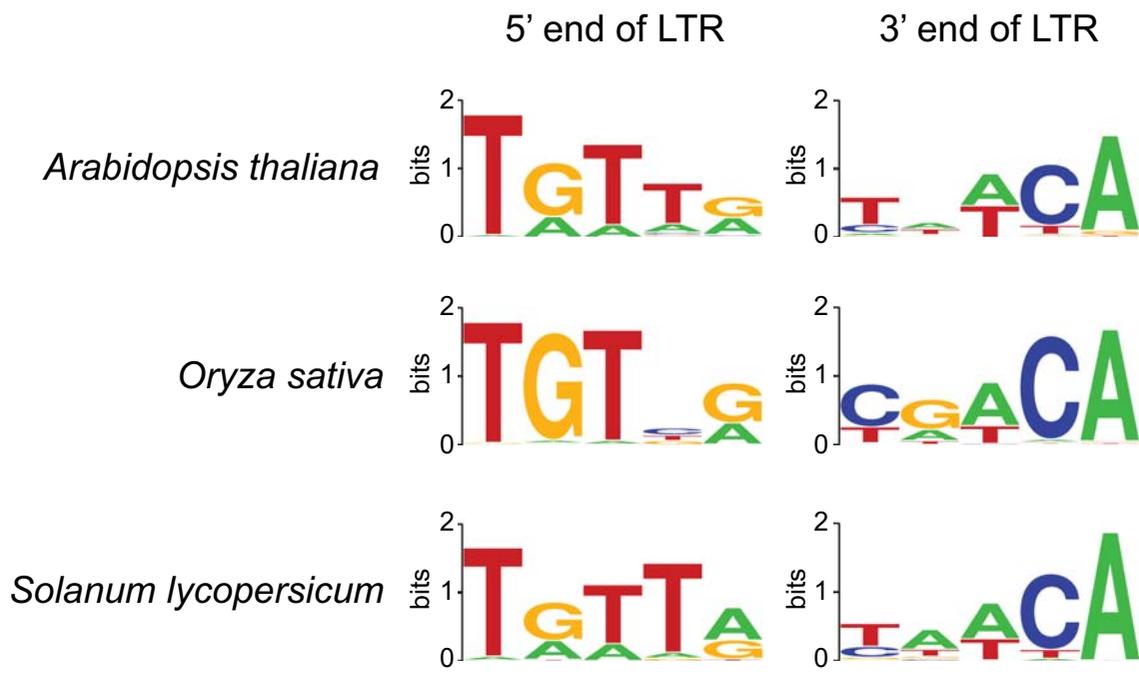


Supplementary Figure 1.

a

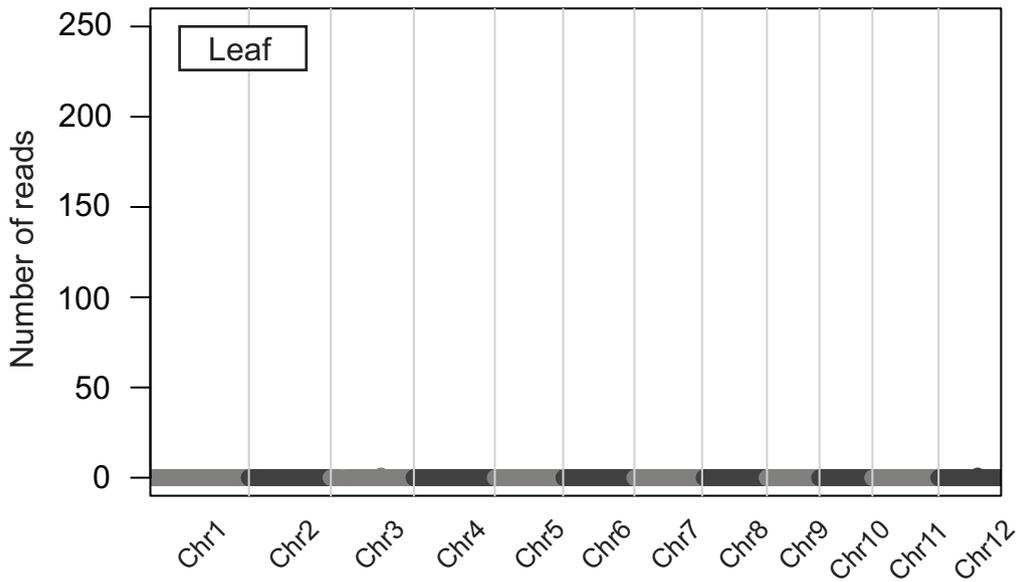


b

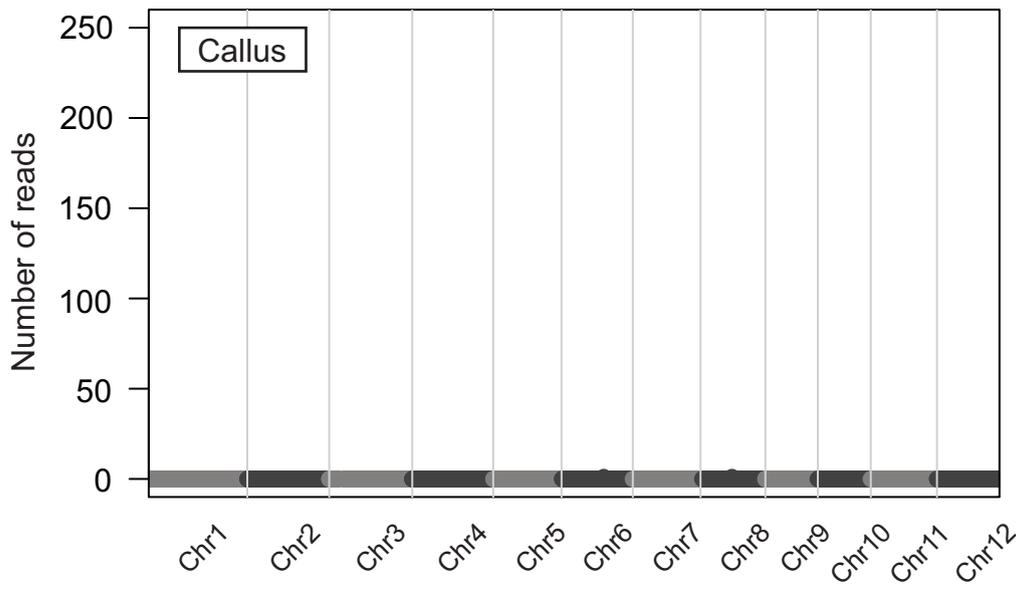


Supplementary Figure 2.

a

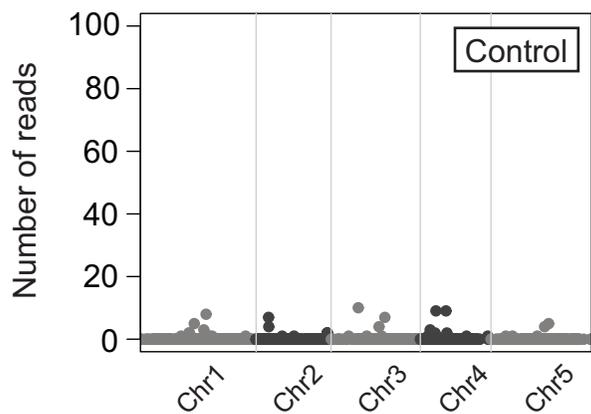


b

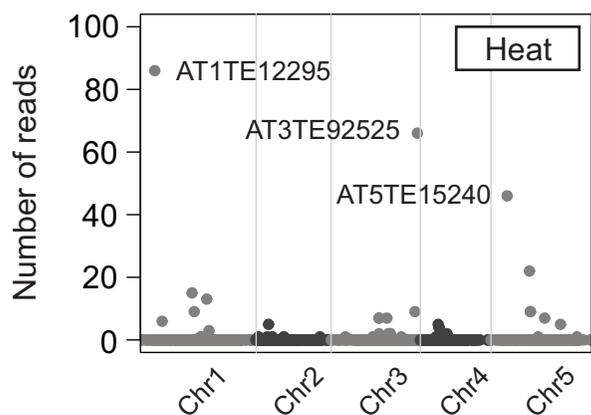


Supplementary Figure 3.

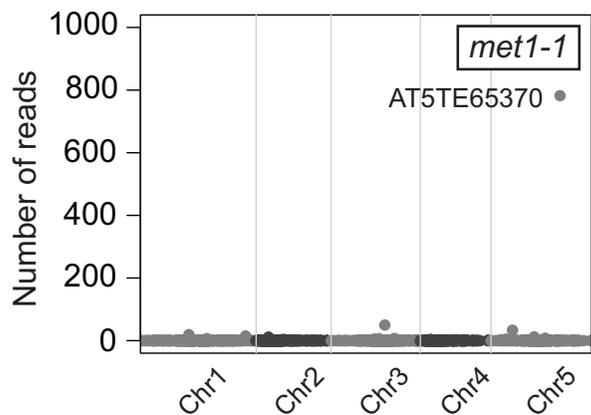
a



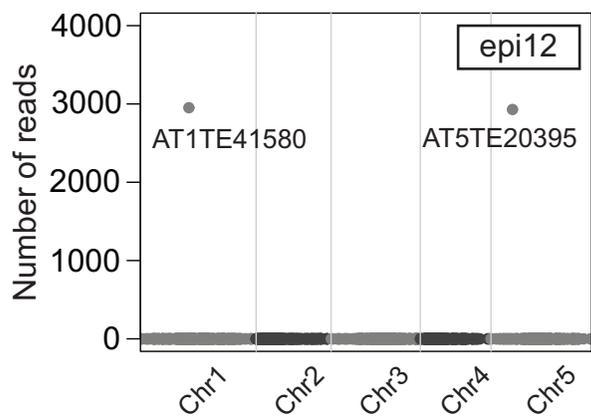
b



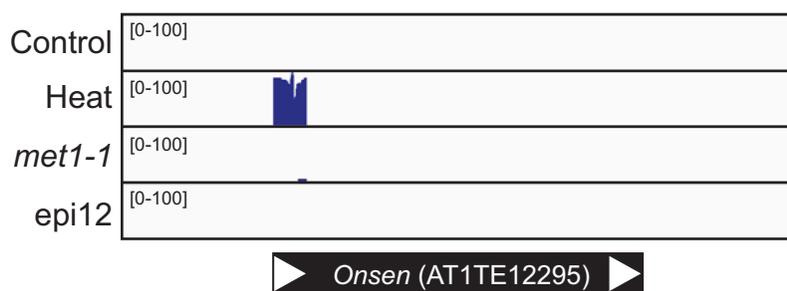
d



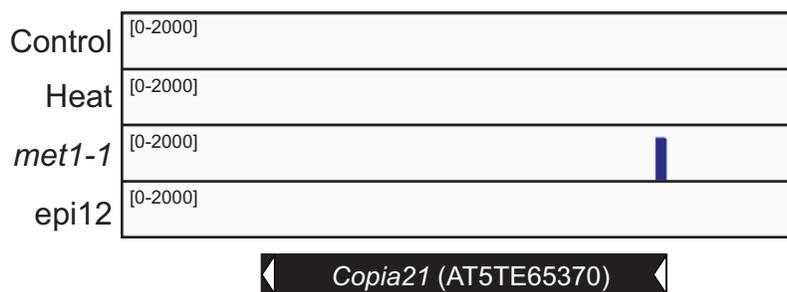
f



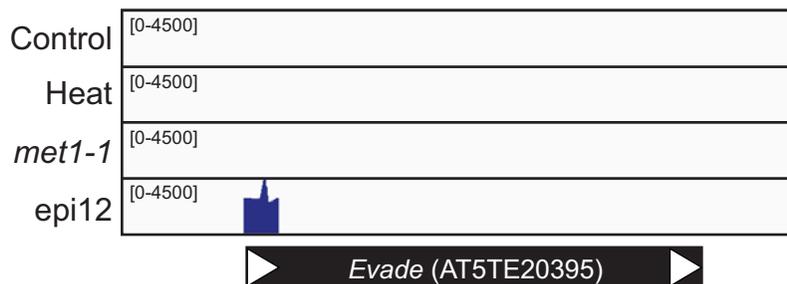
c



e

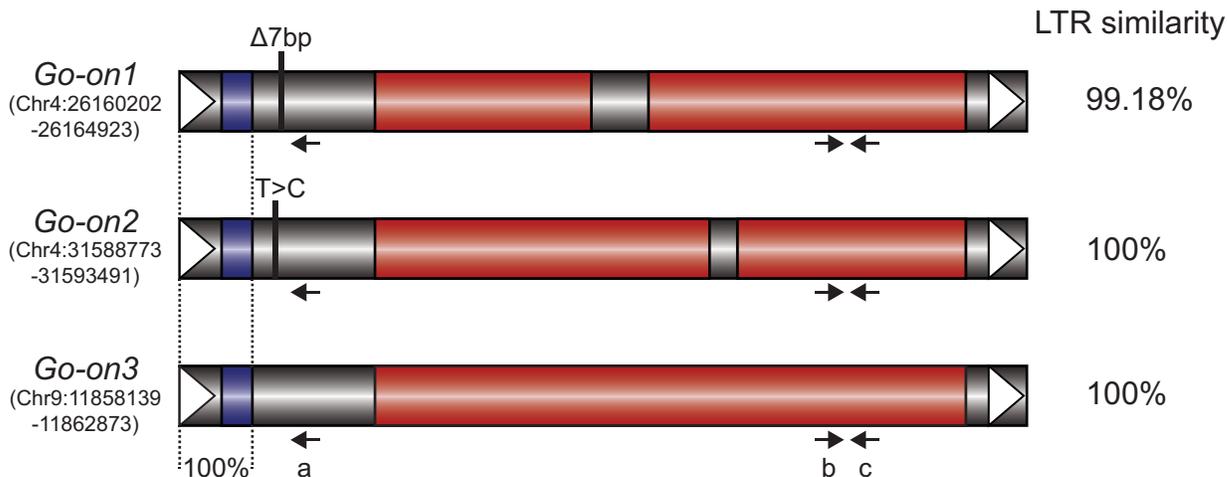


g



Supplementary Figure 4.

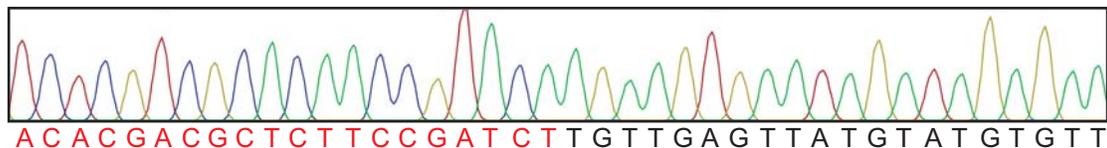
a



b

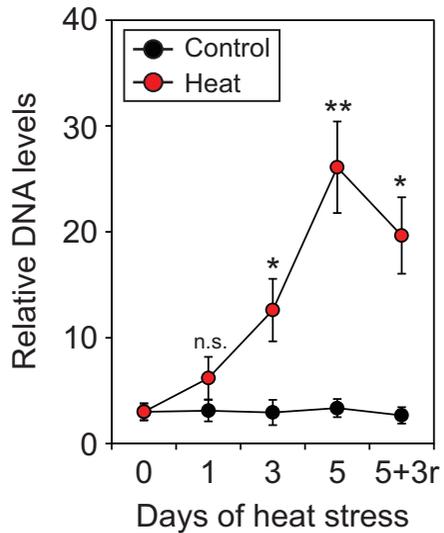
<i>Go-on1</i>	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GG	-----	CT AAACCCT AGCCT CGCCGGAG
<i>Go-on2</i>	123	TGGT AT CAGAGCCAAT CGGCT GGC GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
<i>Go-on3</i>	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
clone1	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
clone2	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
clone3	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
clone4	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		
clone5	123	TGGT AT CAGAGCCAAT CGGCT GGT GGCT GGC GAC GGGC GAC GGCT AAACCCT AGCCT CGCCGGAG		

c

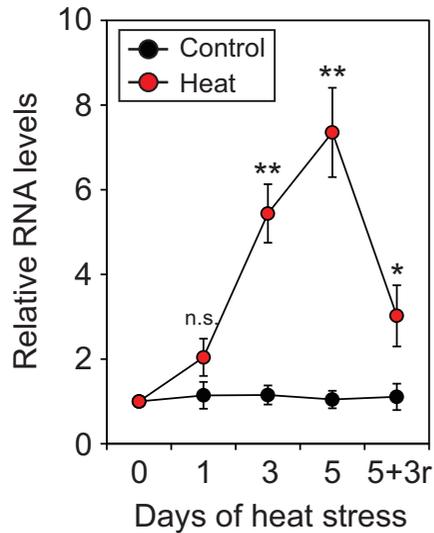


Supplementary Figure 5.

a



b



c

5' -TGTTGAGTTATGTATGTGTTGG**CCCAT**GAGG**CCCAT**TATACTAC

1 10 20 30 40

TCATATGTACATGTATATAGCAGAGTTAGAGAAATGAAAAAAGTAG

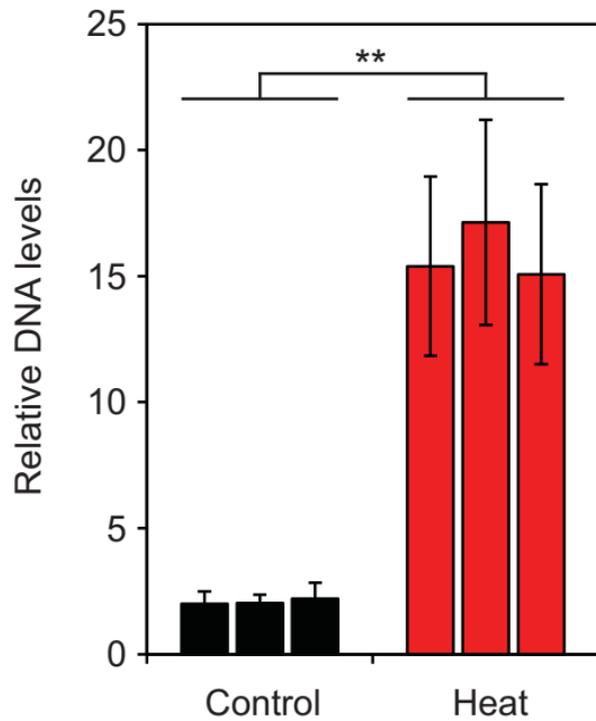
50 60 70 80

TGAAGCTTCTAGAGAAAAATTCCCAA**ACTTCA****TGGTATCAGAGC**-3'

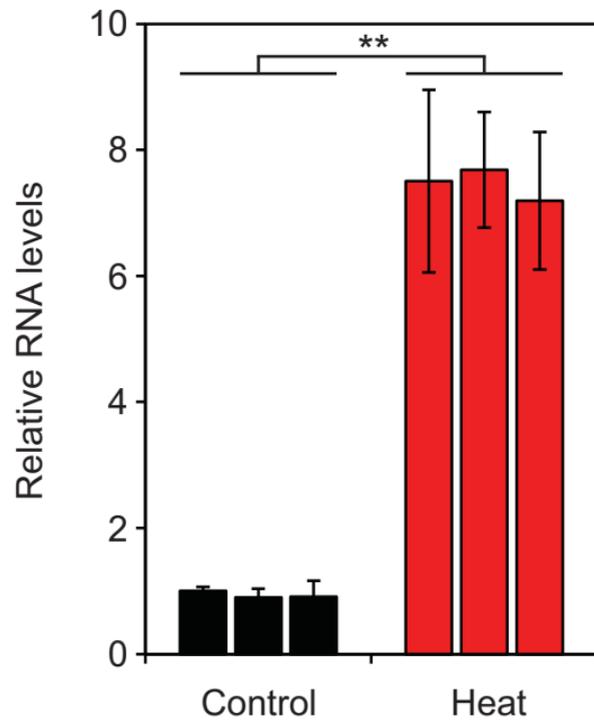
90 100 110 120 130 134

Supplementary Figure 6.

a

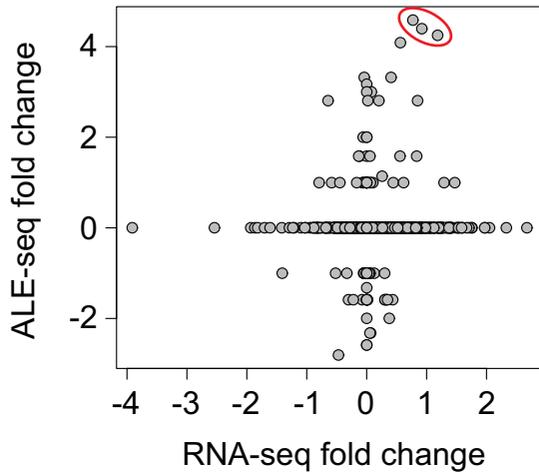


b



Supplementary Figure 7.

a



b

