Illuminating the plant transposon insertion landscape in real time using Cas9-targeted Nanopore sequencing and a novel pipeline

Ilya Kirov [1]*, Pavel Merkulov[1], Sofya Gvaramiya[1], Roman Komakhin[1], Murad Omarov[1], Maxim Dudnikov[1], Alina Kocheshkova[1], Zakhar Konstantinov[1], Alexander Soloviev[1], Gennady Karlov[1], Mikhail Divashuk[1]

[1]All-Russia Research Institute of Agricultural Biotechnology, Timiryazevskaya str. 42, Moscow 127550, Russia;

Abstract

Transposable elements (TEs) are a major source of genomic novelty, contributing to plant adaptation, speciation and new cultivar production. The often large, complex genomes of plants make identifying TE insertions from short reads challenging, while whole-genome sequencing remains expensive. In addition, rare somatic insertions reflecting mobilome dynamic are difficult to track using short reads. To overcome these challenges, we applied the Cas9-targeted Nanopore sequencing (CANS) approach and a novel pipeline, NanoCasTE, to trace both genetically inherited and somatic TE insertions in plants. We performed CANS of EVD retrotransposon in wild-type Arabidopsis thaliana and obtained up to 40x coverage of the targets after only a few hours of sequencing on a MinION sequencer. To estimate the ability to detect new TE and T-DNA insertions, we exploited the A. thaliana ddm1 mutant, which has elevated TE activity. Using CANS, we detect the hemizygous T-DNA insertion site and genetically inherited EVD insertions. CANS also detected >800 somatic TE insertions allowing monitoring of TE insertions in real time. By this, we built a high-density physical map of EVD insertion preferences in the A. thaliana ddm1 and found that EVD insertions are biased toward pericentromeric regions in ddm1 that differ from wild-type A. thaliana ecotypes suggesting a pivotal role of DNA methylation in the shaping of EVD insertion preferences. Using CANS, we were able to detect transposition events of ONSEN transposon family that occurred in response to heat stress and found hotspots of ONSEN insertions. Our results indicate that CANS and NanoCasTE are effective tools for the detection of TEIs in real time, opening an avenue for obtaining a deeper understanding of mobilome organization in response to stress and in different genetic backgrounds. The ability of CANS to detect genetically inherited TEIs and T-DNA integration sites make it useful for screening insertion mutants and transgenic plants.

Introduction

Transposable elements (TEs) are a diverse group of genomic elements that can transpose across the genome via RNA intermediates (Class I transposons, or retrotransposons) or via a “cut-and-paste” mechanism (Class II, DNA transposons). TEs have been identified in almost all living organisms, where they and their remnants can comprise up to approximately 93% of the genome (Rabanus-Wallace *et al.* 2021). In plants, LTR retrotransposons are a major group of TEs with hundreds of ancient as well as very recent insertions (Baduel *et al.* 2021). TE insertions (TEI) may alter gene expression, create a new transcriptional repertoire and cause genomic reorganization (Chuong, Elde and Feschotte 2017). The effects of TEs on genomic, transcriptomic and proteomic activity have made them a major force in plant species diversity and evolution (Lisch 2013). TEIs are classified as genic (exonic and intronic) or intergenic (Sultana *et al.* 2017) based on their locations in the genome. TEIs in or near genes are often more deleterious than single-nucleotide polymorphism (SNPs) and may create null alleles. However, the true impacts of TEIs on the functions of affected genes are more variable, ranging from complete inactivation to the generation of novel transcriptional programs.

Progress in pangenome sequencing and genome assembly algorithms has helped unravel the genome-wide TEI landscape in natural plant populations and germplasm (Domínguez *et al.* 2020; Baduel *et al.* 2021). These studies demonstrated that TEIs are an important source of novelty for plant adaptation (Domínguez *et al.* 2020). For example, recurrent insertions of TEs in *FLOWERING LOCUS C* (*FLC*; a key determinant of flowering time variation in plants) in different *Arabidopsis thaliana* populations have contributed to the local adaptation of this species (Baduel *et al.* 2021). Recent genome-wide association study (GWAS) analysis using 602 resequenced genomes of wild and cultivated tomatoes revealed that at least five of 17 analyzed traits were associated with TEI polymorphisms (Domínguez *et al.* 2020), including fruit color and leaf morphology. Hence, ongoing TE activity has also been widely exploited by plant breeders, as many novel and low-frequency alleles of the genes involved in agronomically important traits originated via TEIs (Kobayashi, Goto-Yamamoto and Hirochika 2004; Jiang *et al.* 2009; Butelli *et al.* 2012; Domínguez *et al.* 2020).

TE transposition activity in plant cells is under strict epigenetic control, and most TEs are silenced under non-stressed conditions (Slotkin and Martienssen 2007; Nuthikattu *et al.* 2013). To overcome this limitation for TE investigation, mutants that are defective in key genes of TE epigenetic control systems have been obtained. These plants are fertile and have elevated TE activity, making them indispensable tools for the study of TE biology (Mirouze *et al.* 2009; Tsukahara *et al.* 2009; Panda and Slotkin 2020a).

Identifying TEIs in the genome is challenging, primarily because most plant genomes possess a bulk of repetitive DNA carrying many TE insertions. This 'dark side' of the genome is difficult to explore using the short-read approach, a major tool for TEI identification in plants (Quadrana *et al.* 2016a; Carpentier *et al.* 2019; Chen *et al.* 2020; Domínguez *et al.* 2020; Baduel *et al.* 2021). Long-read sequencing using PacBio and Nanopore technologies provides a dramatic increase in the resolution of the structures of repetitive DNA sequences (Jung *et al.* 2019; Kirov *et al.* 2021a). While the whole-genome sequencing (WGS) approach theoretically allows you to identify all TEIs in an individual plant, its application in plants with large genomes is still quite expensive. In addition, WGS is even more difficult when the identification of TEIs in a plant population is needed (Handsaker *et al.* 2011). Therefore, several approaches have been proposed to capture specific genomic loci possessing TEIs without the need to sequence the rest of the genome (Li *et al.* 2019). Enriched DNA samples are usually sequenced using short reads. Therefore, identifying TEIs that have occurred in repeat-rich regions such as centromeres and in other TEs is challenging. These approaches are mostly based on oligo-probe-mediated 'fishing' of target DNA fragments (Williams-Carrier *et al.* 2010; Quadrana *et al.* 2016a). A microfluidics-based method (Xdrop) was recently developed and shown to be effective for capturing the integration sites of human papillomavirus 18 (Madsen *et al.* 2020). While these approaches are effective, they require either a PCR amplification step, which may introduce chimeric reads, or the use of specific equipment (for the Xdrop method).

The combination of TE capture with Nanopore sequencing has greatly improved integration site mapping (Li *et al.* 2019; Madsen *et al.* 2020). A novel enrichment approach combining Cas9-mediated adaptor ligation and long-read sequencing (CANS) was recently proposed (Gabrieli *et al.* 2018; Gilpatrick *et al.* 2020; Stangl *et al.* 2020). The principle of this method is based on the finding that Nanopore sequencing adapters preferentially ligate to Cas9-cleaved DNA sites. To decrease the ligation of random adapters, the DNA is initially dephosphorylated. The advantages of this approach over oligo-based enrichment include the lack of a PCR amplification step, the ability to detect DNA methylation and the shorter time needed to complete the protocol. CANS has been applied to sequence target genes in humans (Gilpatrick *et al.* 2020) and plants (López-Girona *et al.* 2020; Kirov *et al.* 2021b). The use of CANS to identify TE integration sites in human cells was recently reported (McDonald *et al.* 2021). However, the application of CANS for TE integration identification in plants has not yet been reported.

Here, we utilized CANS and developed a novel pipeline (NanoCasTE) for the identification of T-DNA and somatic as well as genetically inherited TEIs in *A. thaliana*. Using the *A. thaliana ddm1* mutant, which carries T-DNA and possesses elevated TE activity, we were able to detect all insertions of EVD retrotransposons in an individual plant with only 0.2× genome coverage obtained by CANS reads. This coverage is an order of magnitude lower than what is needed for TEI identification based on whole-genome sequencing. The comparison of distributions of hundreds of EVD insertions in ddm1 (detected by CANS) and natural *A. thaliana* populations (Baduel *et al.* 2021) revealed a key role of methylation in shaping of TEI distribution in the genome. By applying CANS to detect somatic ONSEN insertions in wild-type *A. thaliana* plants after heat-stress we showed its usefulness to monitor mobilome response to stress without need of obtaining seeds from stressed plants. Altogether, our results indicate that CANS and NanoCasTE are effective tools to capture TEIs in plants, opening an avenue for obtaining a deeper understanding of the TE integration landscapes in plant genomes and exploring the potential of TE 'mutagenesis' in plant breeding.

Results

Cas9-mediated targeted sequencing of three copies of the EVD retrotransposon in *A. thaliana*

To demonstrate that CANS is suitable for tracing individual transposon insertions in plants, we chose the EVD5 (5333 bp) LTR retrotransposon (Mirouze *et al.* 2009), which can generate new insertions in some *A. thaliana* mutants (e.g. *met1* and *ddm1* [*decrease in DNA methylation 1*]) (Tsukahara *et al.* 2009)). EVD5 belongs to the Evadé subfamily (ATCOPIA93 family) of Ty1/Copia retrotransposons, which includes other EVD elements whose activity was not detected. Of these, EVD1 (AT1TE41575, 1548 bp) and EVD2 (AT1TE41580, 5336 bp) are densely located elements with high sequence similarity (~99%) to EVD5 (Mirouze *et al.* 2009).

The CANS workflow includes several steps: (1) isolation and size selection of high amount (5 - 10µg) of high-molecular weight genomic DNA; (2) single-guide RNA design and synthesis via *in vitro* transcription; (3) Cas9-sgRNA complex assembly; (4) genomic DNA cleavage by Cas9-sgRNA complexes and dA-tailing; (4) Nanopore adapter ligation and (5) Nanopore sequencing. The detailed protocol of this procedure is available on protocols.io (https://www.protocols.io/private/DE1CFDE8C8FF11EBA7DA0A58A9FEAC02).

For CANS experiments we designed a set of seven guide RNAs located in LTRs and inter-LTR regions of EVD5 and EVD2. We predicted that one of these sgRNAs also recognizes the EVD1 copy. Because CANS is strand-specific, we designed sgRNAs to direct Nanopore sequencing to sequence (1) EVD retrotransposons *per se* and (2) upstream and downstream sequences flanking retrotransposon insertions. The sgRNAs were transcribed *in vitro* as single-guide RNAs (sgRNAs). We carried out two runs on a MinION sequencer using seven and five sgRNA sets. The pilot experiment using all sgRNAs as one mixture yielded 15,777 reads (144Mb total bases), with 231 (1.4%) on-target reads detected after 4.4 hours of sequencing. Because of the many short reads produced from the internal sgRNA cutting site, for our second experiment, we decided to use five of the most efficient sgRNAs divided into two mixtures to prevent the generation of short reads. The selected sgRNAs, designated sgRNA1 to sgRNA5, included one LTR sgRNA (sgRNA1, sgRNA mixture 1) and four sgRNAs (sgRNA mixture 2) located in the inter-LTR region (Figure 1).

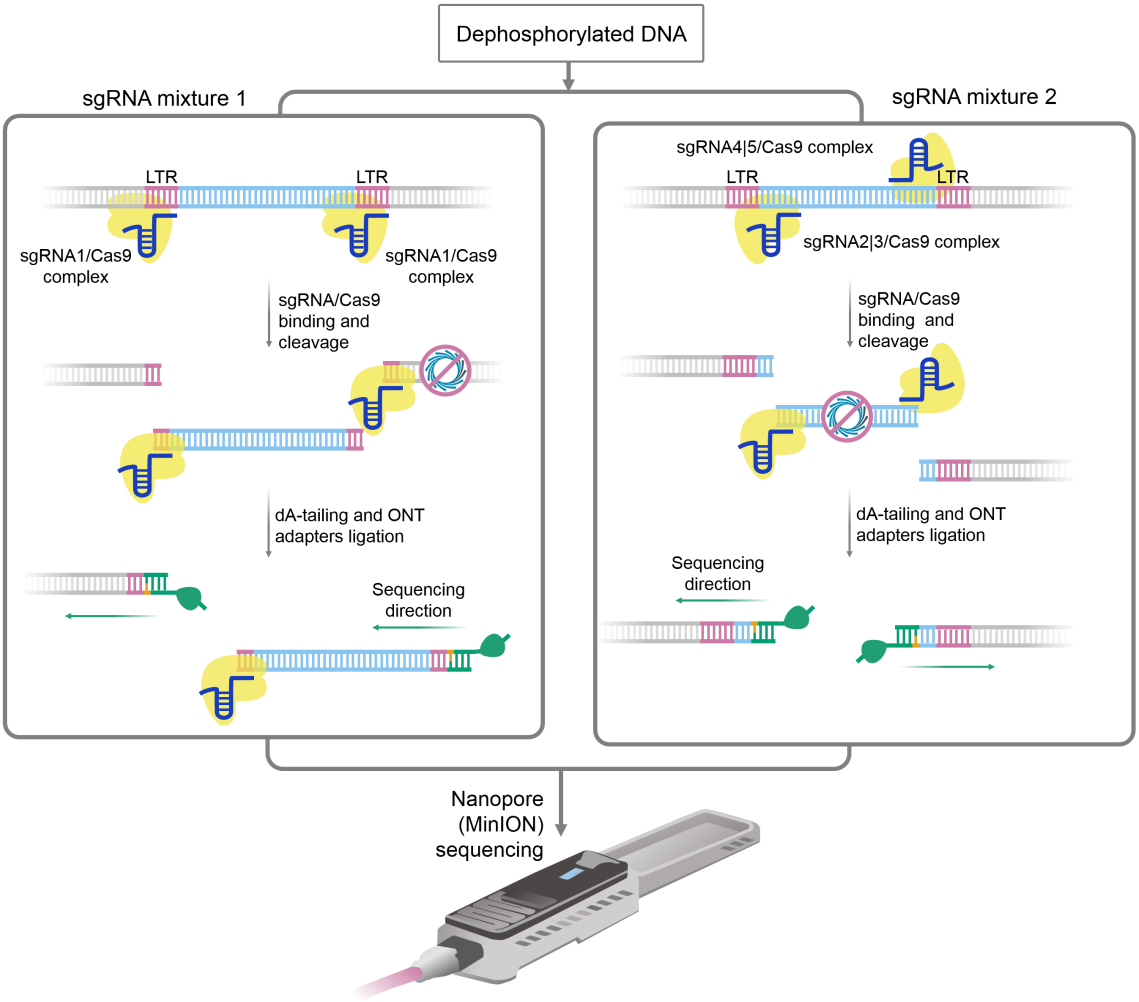


Figure 1. Schematic view of CANS of LTR retrotransposon insertions using five sgRNAs divided into two mixtures. The RNP binding, target DNA cleavage and ONT library preparation steps of the CANS protocol are shown. The DNA fragments which are not sequenced are labeled by red circle.

The experiment using five sgRNAs yielded 88,000 reads (207Mb total bases), with 259 (0.3%) on-target reads detected after 4 hours of sequencing on the MinION flow cell. Both CANS runs using Col-0 genomic DNA resulted in up to 40× coverage of the target retrotransposons EVD5 and EVD1/2 (Figure 2A, 2B). Mapped reads from the first and second runs covered 35,000 bp and 14,000 bp regions, respectively, including target EVDs and flanking regions. The broad coverage of the flanking sequences demonstrates the advantage of CANS for detecting TE insertions in the genome. Inspection of the sgRNA locations and the read distribution revealed that CANS had high strand specificity, with up to 82% of reads mapped to strands with 3'-ends unprotected by Cas9 and possessing PAM (NGG) sequences. This is another advantage of CANS, as it allows the coverage of the target region to be increased. We also noticed that the bulk of reads from internal parts of EVD2 and EVD5 were unambiguously assigned to the corresponding EVD copy, corroborating previous reports that cDNA ONP reads have high mappability (Panda and Slotkin 2020a).

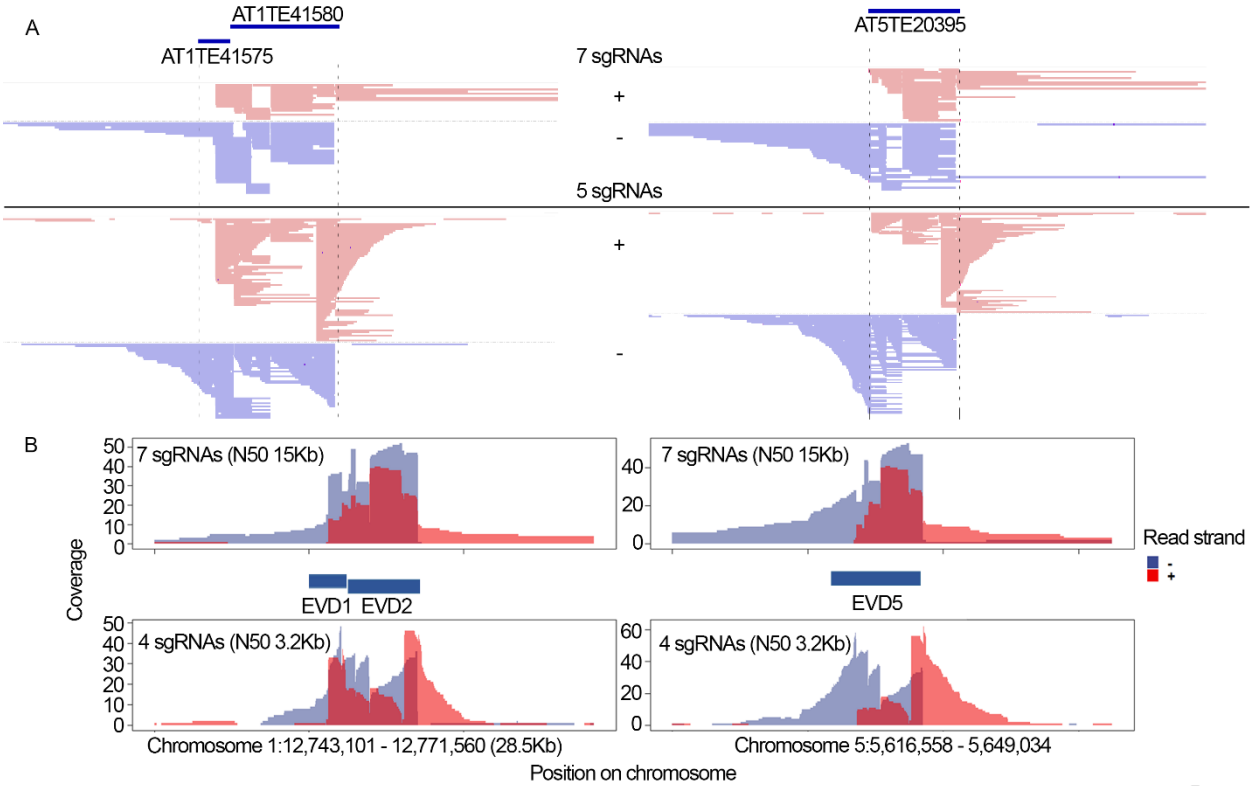


Figure 2. Targeted enrichment of EVD loci following CANS using a set of 7 or 5 sgRNAs. (**A**) IGV view of Nanopore read alignment to EVD1 (AT1TE41575), EVD2 (AT1TE41580) and EVD5 (AT5TE20395) retrotransposon copies. (**B**) Coverage plot of EVD1, EVD2 and EVD5 using forward (red) and reverse (blue) reads obtained following CANS with 7 and 5 sgRNAs.

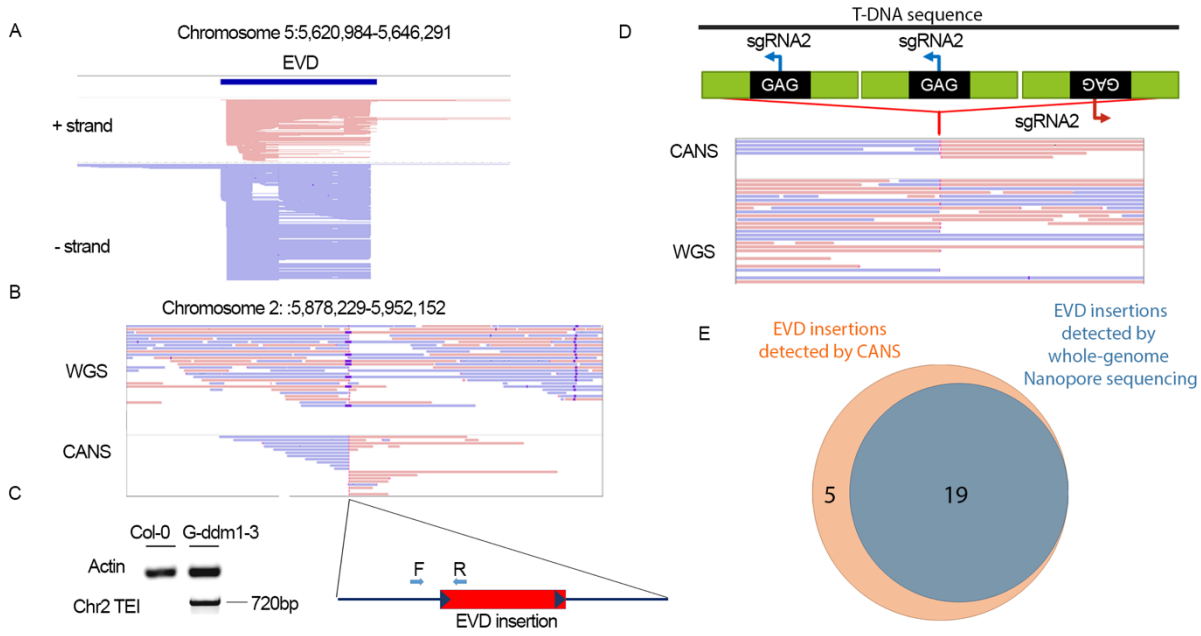
Thus, using CANS, we sequenced EVD1, EVD2 and EVD5 retrotransposon copies and their flanking sequencing with up to 40× coverage, demonstrating the advantage of this approach for detecting TE insertions.

Detection of T-DNA and new EVD insertions in the genome using CANS and a new pipeline

To assess whether non-reference EVD insertions could be detected using CANS, we exploited the *A. thaliana ddm1* mutant, whose EVD activity was previously described (Tsukahara *et al.* 2009). We created transgenic T_2 *ddm1* plants carrying T-DNA insertions which contains the EVD5 GAG open reading frame (G-ddm1-3 plants). The T-DNA is recognized by one of the five EVD sgRNAs that allowed us to simultaneously detect EVD insertions and T-DNA integration sites and to obtain information about the sufficiency of CANS sequencing depth. We carried out CANS using five sgRNAs and generated 4420 reads with N50 ~11 Kb. Following read mapping to the genome, we looked at the original EVD copy on chromosome 5 to estimate the sequence coverage (Figure 3A). The EVD sequence had up to 400× coverage by 1170 reads (25% of all CANS reads) (Supplementary Figure S1). It is worth noting that the reads spanning EVD sequences from LTR to LTR could not be assigned to the certain novel EVD insertion, as all EVD copies had identical sequences. Consequently, the observed EVD coverage primarily resulted from the alignment of Nanopore reads that originated from different EVD insertions in the genome.

To automate the detection of TE and T-DNA insertions, we developed a novel pipeline called NanoCasTE (see Material and Methods). By applying NanoCasTE to the CANS dataset, we detected 24 EVD insertions. Of these, 19 were high-confidence TEIs (TEIs supported by two or more reads). An example of one TEI located on chromosome 2 that was homozygous according to the WGS data is shown in Figure 3B. This TEI was also validated by PCR with primers designed on EVD and TEI flanking sequences (Figure 3C).

173 To support the notion that the coverage we obtained is sufficient for EVD TEI identification, we analyzed
174 EVD-genome junctions and found that they had ~8× coverage, which is sufficient for insertion detection
175 (Figure 3A).
176
177 We next checked whether the T-DNA insertions sites can be detected by CANS. The T-DNA sequence
178 (EVD5 GAG) possess the recognition site for a single sgRNA (sgRNA2, Figure 3D). Using CANS we found
179 one T-DNA insertion site which had 4x coverage (Figure 3D). The presence of this T-DNA insertion site was
180 proved by whole-genome nanopore sequencing followed by running of nanotei pipeline (Kirov *et al.*
181 2021a) that also revealed that T-DNA insertion being hemizygote. Thus, the experiments demonstrated
182 that even using a single sgRNA a hemizygote T-DNA integration site can be detected by CANS in *A. thaliana*
183 genome.
184
185 Next, we estimated the efficiency of CANS. For this, we performed whole-genome Nanopore sequencing
186 of the same G-ddm1-3 plant that was used for CANS followed by TEI calling using nanotei pipeline. We
187 found 19 EVD5 insertions supported by 2 or more WGS Nanopore reads. Comparison of TEIs identified by
188 CANS and WGS Nanopore sequencing showed that all 19 TEIs found by WGS were among the high-
189 confident TEIs captured by CANS (Figure 3E). These results suggest that even the low genome coverage
190 (0.2× in our experiment) of CANS may be sufficient to capture all T-DNA and target transposon insertions
191 in *A. thaliana*, highlighting the good sensitivity of the method. Additionally, the newly developed
192 NanoCasTE pipeline enables the rapid identification of novel TEIs from CANS data.
193
194



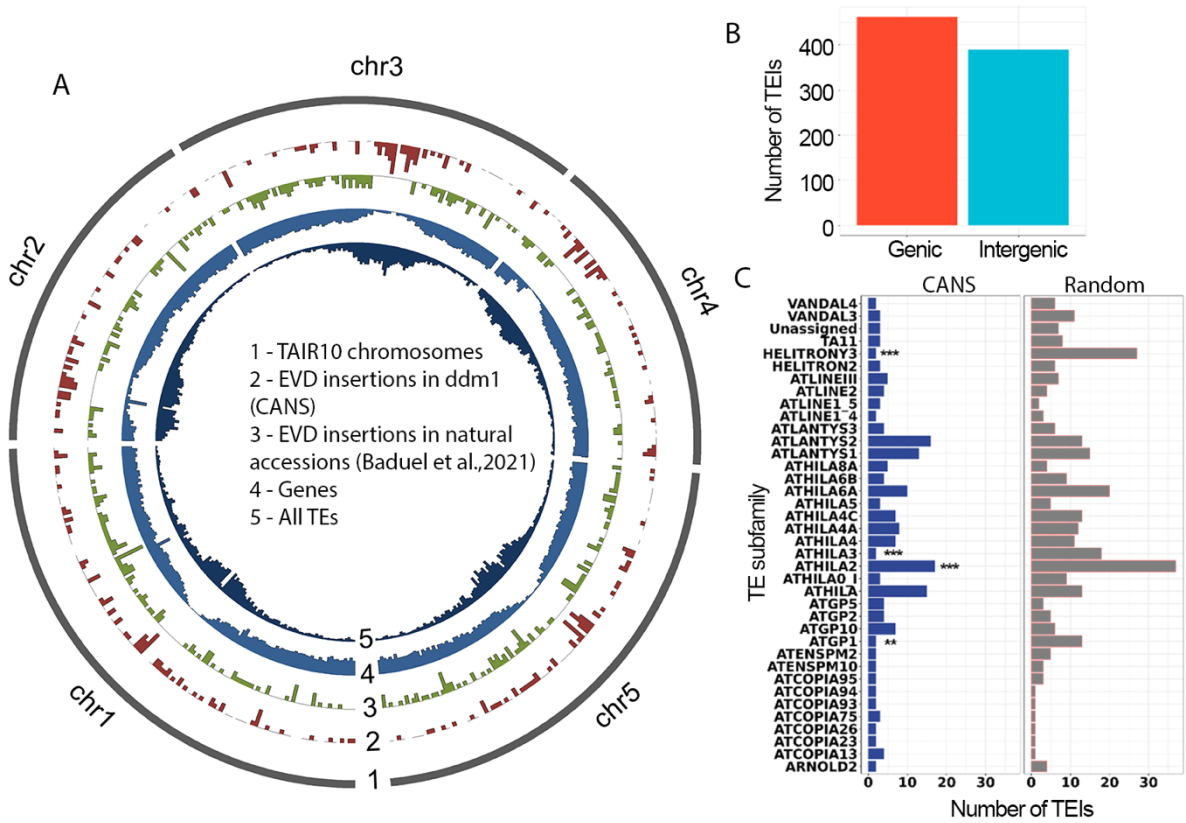
195

196 **Figure 3. EVD coverage and insertion detection in a G-ddm1-3 plant using CANS.** (**A**) IGV snapshot of
197 coverage of the original copy of EVD on Chromosome 5 using forward (red) and reverse (blue) ONT reads
198 obtained after CANS of G-ddm1-3 plant. (**B**) Inserted T-DNA structure (upper panel), sgRNA2 positions and
199 its coverage based on CANS and whole-genome Nanopore sequencing reads. (**B**) IGV browser snapshot
200 of one of the 25 genetically inherited EVD insertions detected by whole-genome Nanopore sequencing
201 and CANS. (**C**) PCR validation of the EVD insertion on Chromosome 2 (left) using primers based on the EVD
202 and TEI flanking region (right). (**E**) Venn diagram of the numbers of EVD insertions detected in G-ddm1-3
203 plant by CANS and whole-genome Nanopore sequencing followed by nanotei analysis.
204

205 Capture of somatic insertions in a *ddm1* population revealed differences in chromosome
206 distribution with wild-type accessions
207

208 It was suggested that some TEs are preferentially inserted into certain genomic regions but the current
209 picture of TEIs distribution has been shaped by purifying selection. Taking into account that CANS can
210 detect both somatic and genetically inherited TEIs we were interesting to obtain the picture of EVD
211 insertion landscape in real-time. For this we performed CANS sequencing of EVD insertions in a population
212 of approximately 50 *ddm1* plants. After 2.5 hours of MinION sequencing, we generated 18,246 Nanopore
213 reads with N50 ~5.5 Kb. Using the NanoCasTE pipeline, we detected 851 TEIs, including 29 high-confidence
214 TEIs (supported by two or more reads), in the *ddm1* population. Thirteen TEIs were PCR validated using
215 DNA from individual plants, resulting in 1 to 6 plants carrying a single TEI. Analysis of the chromosome-
216 wide distribution of the TEIs showed that they tended to be clustered in the pericentromeric regions of
217 all chromosomes, while the density of TEIs in chromosome arms was significantly reduced (Figure 4A).

218



219

220 **Figure 4. Genome organization of EVD insertions detected by CANS in a pooled (~50 plant) *ddm1***
221 **sample.** (**A**) Chromosome distribution of EVD insertions detected by CANS in ddm1 genome (2) and
222 occurred in natural *A.thaliana* accessions (Baduel *et al.* 2021) (3), TAIR10 annotated genes (4) and TEs
223 (Panda & Slotkin 2020) (5). (**B**) Bar plot showing the number of genic and intergenic TEIs detected by CANS.
224 (**C**) Bar plot showing the number of EVD insertions in members of different TE subfamilies of *A. thaliana*
225 based on CANS data (left panel) and random expectation (right panel).

226 Classification of these TEIs revealed that 54% (462) of them are located in genic regions, including 403
227 (87%) exonic and 59 (13%) intronic TEIs (Figure 5B). This value is expected by chance based on the total
228 length of the genic regions in the assembled (119 Mb, TAIR10) *A. thaliana* genome (68.3 M, or 57% of the
229 sequenced genome). Gene Ontology (GO) analysis revealed no enrichment of any GO categories in the

230 set of genes with TEIs, suggesting that most EVD insertions in *ddm1* are randomly distributed among genes
231 with different functions.

232 Among intergenic TEIs (389, 46%), 241 TEIs (28%) were located in the sequences of other TEs. This value
233 is significantly higher (chi-squared test p-value = 0.0004) than expected by chance based on the total
234 length of the annotated TEs in the *A. thaliana* genome (24.9 Mb, or 21% (Panda and Slotkin 2020a)).
235 Moreover, 20 TEs had 2 or 3 different TEIs. We analyzed the number of TEIs in TEs of different subfamilies
236 and found that members of four subfamilies were significantly underrepresented (Fisher's exact test p-
237 value <0.01) in a set of TEs with detected TEIs, including HELITRONY3 (p-value = 6.78e-07), Athila2 (p-
238 value = 0.004), Athila3 (p-value = 0.0002) and ATGP1 (p-value=0.004) (Figure 4C). This finding can be partly
239 attributed to the differences in chromosome distribution between TEs of these families and EVD TEIs. For
240 example, most (1104 of 1399, 79%) of the members of HELITRONY3 are located in chromosome arms,
241 where the EVD TEI frequency is much lower than in the pericentromeric regions. EVD insertion
242 preferences for other TEs can not be explained by their chromosome location biases. In addition, no DNA
243 motif enrichment was found in the region (TEI±500 bp) of EVD insertions. Therefore, we proposed that
244 EVD insertions may occur in the genomic regions with a certain chromatin state as it was previously shown
245 for another transposon, ONSEN (Roquis *et al.* 2021). As chromatin states are correlated with transcription,
246 we checked whether EVD more frequently insert transcriptionally poor regions. For this, we performed
247 direct Nanopore RNA sequencing of ddm1 seedlings. We found that ~90% (764) of EVD insertions are
248 located in low expressed (<2 Nanopore reads) regions in ddm1 genome that is significantly higher than
249 expected by chance (Fisher's exact test P-value 4.25e-18). If chromatin state can influence EVD insertion
250 preferences then the EVD insertion distribution may differ between genomes of ddm1 mutant and wild-
251 type *A.thaliana*. To check this assumption we compared the EVD distribution between ddm1 (851
252 insertions) and wild-type *A.thaliana* ecotypes (426 insertions of ATCOPIA93 family) (Baduel *et al.* 2021).
253 Interestingly, we detected significant differences in EVD insertion frequency in the pericentromeric
254 regions with almost no insertions found in these regions in natural accessions (Figure 4A). These results
255 provide evidence that in native genomic environment EVD tends to escape pericentromeric regions while
256 ddm1 mutation skewed the insertion distribution toward proximal chromosomal regions.

257 We then assess the differences of methylation of EVD *per se* in ddm1 genome compared to the wild-type.
258 To answer this question, we used CANS data for EVD obtained from Col-0 and ddm1 genomic DNA and
259 performed methylation calling from raw Nanopore reads. Using our custom script (Kirov *et al.* 2022) we
260 visualized methylated cytosine in 'per-read' mode. We found strong methylation of EVD5 in Col-0
261 background (Supplementary Figure S2) while EVD5 copies in ddm1 genome are barely methylated. It
262 should be noted that *ddm1* Nanopore reads aligned to EVD5 are originated from multiple copies therefore
263 the obtained methylation profile Is a snapshot of epigenetic modifications of EVD5 insertions. Taking this
264 into account and that a portion (~10%) of CANS reads from *ddm1* (Supplementary Figure S2) exhibit strong
265 methylation it could be concluded that some EVD5 insertions are de novo methylated.

266 Thus, using CANS coupled with the NanoCasTE pipeline, we built a high-density physical map of EVD
267 insertion preferences in the *A. thaliana* and showed that EVD insertions are biased toward
268 pericentromeric regions in ddm1 genome. However, the distribution of EVD insertions strongly differs
269 between ddm1 and wild-type *A. thaliana* ecotypes suggesting a pivotal role of DNA methylation in shaping
270 of EVD insertion preferences. In addition, we demonstrated that CANS data can be utilized to decipher
271 epigenetic profile of Tes.

272

Rapid detection of short-term response of target TE transposition to stress

We noticed that CANS allows us to detect both somatic and genetically inherited TEIs. This may be useful for elucidation of the short-term response of mobilome to stress. To check this, we exploited ONSEN transposon whose transposition activity in wild-type *A. thaliana* can be triggered by heat stress (Ito *et al.* 2011). To increase ONSEN activity and also induce transgenerational inheritance we applied heat-stress together with a-amanitin and zebularine treatment (Figure 5A)). We designed a set of 6 sgRNAs targeting all 8 known copies of ONSEN (ONSEN1-8) and performed CANS experiments as described above. As an additional control, we also applied EVD sgRNAs in the same run (multiplexing). After two MinION runs on the same flow cell, we collected 175,001 high quality reads with 3,924 reads aligned to the ONSEN sequences (Supplementary Figure S3).

Analysis of the data by NanoCasTE pipeline revealed 519 insertions including one insertion identified previously as a result of TAIR10 genome assembly error (Kirov *et al.* 2021a). Only one somatic EVD insertion was detected suggesting specific activation of ONSEN element but not EVD. The distribution of ONSEN insertions showed strong biases toward chromosome arms while the insertion frequency in the centromere regions was depleted (Figure 5B). Similar results were obtained by plotting of ONSEN insertions detected in natural *A. thaliana* accessions (Baduel *et al.* 2021). Most of the ONSEN insertions (83%, 430 of 519) were located in the genes corroborating with the recent studies where it was shown that >90% of ONSEN insertions were genic (Roquis *et al.* 2021). GO enrichment analysis of the genes possessing ONSEN insertions (426 genes) revealed that several GO terms were overrepresented including aspartyl esterase activity (Biological processes, GO:0045330, P-value (adjusted) = 3.5x10-2), cell wall organization or biogenesis (GO:0071554, 4.548×10-2), external encapsulating structure organization (GO:0045229, 6.873×10-3).

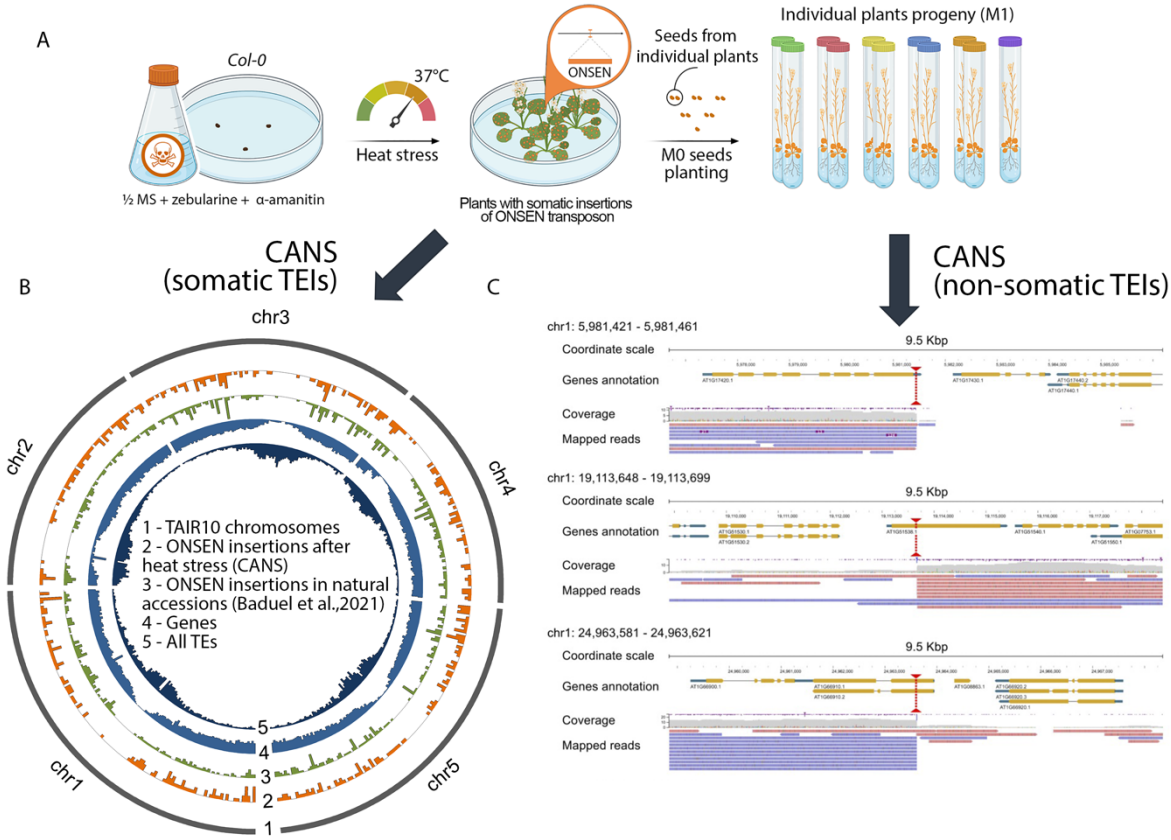


Figure 5. CANS detection of somatic and genetically inherited insertions of ONSEN after heat-stress activation. (A) Schematic overview of the experiment of ONSEN heat-activation in the presence of

299 zebularine and α-amanitin in wild-type *A. thaliana* plants. (**B**) Chromosome distribution of ONSEN
300 insertions detected by CANS in plants, 48 hours after heat stress (2), ONSEN insertions occurred in natural
301 *A. thaliana* accessions (Baduel *et al.* 2021) (3), TAIR10 annotated genes (4) and TEs (Panda and Slotkin
302 2020b) (5). **(**C) Genomic browser snapshots of three non-somatic ONSEN insertions detected by CANS in
303 M1 plants.

304

305 We found that one of the genes affected by ONSEN is FLC gene (Chr5:3,173,042..3,180,975) possessing
306 TEI in intron 1 which was previously identified as a hotspot of TE insertions (Quadrana *et al.* 2016a, 2019;
307 Baduel *et al.* 2021). Inspired by this case we analyzed all the genes that have ONSEN insertions in our and
308 (Baduel *et al.* 2021) datasets and identified 29 such genes (Supplementary Figure S3). Moreover, 3 of
309 these genes were also found among the genes with EVD insertions according to (Baduel *et al.* 2021)
310 including FLC locus, its antisense lncRNA COOLAIR (AT5G01675) and Class II formin AtFH16 gene
311 (AT5G07770).

312 We were further interested to test the ability of CANS to detect transgenerationally inherited ONSEN
313 insertions. For this, we collected seeds (M1) from five plants treated by heat-stress and grown on medium
314 containing α-amanitin and zebularine and (M0) (Figure 5A). Then, we grew 11 M1 plants (1-2 M1 plants
315 per a single M0 plant) and performed CANS sequencing of mixed DNA sample. Totally, we detected 24
316 ONSEN insertions with 16 insertions supported by 2 or more reads. Some of the insertions were located
317 in the genes (Figure 5C) providing proof of concept that CANS is useful tool to trace individual TE insertions
318 after TE activation.

319 Altogether, using CANS we were able to detect in real-time the transposition activity changes of ONSEN
320 transposon family occurred in response to heat stress. The information allowed us to determine genome
321 organization biases and hotspot of ONSEN insertions. Moreover, we showed that CANS is also useful for
322 detection of transgenerational inherited TEIs providing an opportunity for rapid identification of the
323 affected genes after TE activation in other plants including agronomically important crops.

324

325 Discussion

326 Here, we demonstrated that CANS is an easy-to-use and valuable tool for rapid identification of T-DNA
327 and TE insertions (TEIs) in both repetitive (e.g. TE rich) and genic regions of plant genomes. CANS-based
328 detection of T-DNA and TEI insertions has several advantages over the existing method of TEI enrichment
329 such as TE-capture (Quadrana *et al.* 2016b). First, the CANS protocol, including RNP complex formation,
330 DNA cleavage and sequencing library preparation, is rapid and can be completed in 4-5 hours. Second,
331 because a PCR amplification step is not needed, the resulting data can be used for epigenetic modification
332 calling to generate methylation profile of Tes or TEIs, although higher coverage is required for this process
333 (Gabrieli *et al.* 2018; Ni *et al.* 2021a). Third, long reads have high mappability to the genome, paving the
334 way for TE and T-DNA insertion detection, even in repeat-rich regions of the genome such as centromeres
335 and heterochromatin. For example, CANS revealed that many EVD insertions are located in
336 pericentromeric regions and insight of other TEs. Fourth, utilizing NanoCasTE allows TEIs to be identified
337 easily using CANS data, even if hundreds of insertions are captured. Finally, CANS is able to identify
338 somatic insertions of target TEs allowing rapid real-time monitoring of (i) mobilome response to stress as
339 we showed here using ONSEN and (ii) insertion landscape in the different genetic backgrounds as we
340 demonstrated using *ddm1*. Despite the many advantages of CANS for TEI detection in plants, there are
341 also several constraints to the method that needs to be overcome in the future: (1) a large amount of
342 high-molecular-weight genomic DNA (>2-5 µg) is required; (2) deciphering the TE sequence of the

343 insertion is challenging and requires specific sets of overlapping sgRNAs; (3) only a small subset of
344 sequencing pores (5-10%) are sequencing at the same time, and therefore the sequencing efficiency is
345 lower than for whole-genome sequencing (90-98% of sequencing pores), and (4) the genotype of TEIs
346 (homozygote or hemizygote) cannot be determined using CANS.
347
348 While whole-genome Nanopore sequencing (WGNS) provides excellent resolution to potentially detect
349 all TEIs in a genome (Kirov *et al.* 2021a; Roquis *et al.* 2021), it is expensive when the genome is relatively
350 large. For example, in our hands and based on previous reports using different species, it is challenging to
351 generate more than 8 Gb of data per single MinION flow cell from plant material (Lee *et al.* 2019;
352 Schalamun *et al.* 2019; Dmitriev *et al.* 2021). In contrast to WGNS, by only 4400 CANS reads (equivalent
353 to 0.2´ coverage of the *A. thaliana* genome) on a single flow cell, we were able to detect all genetically
354 inherited EVD insertions found by WGNS. Additionally, we also unravelled the positions of the T-DNA
355 insertion site using only one sgRNA. However, a higher number of sgRNAs and cleavage sites in genomic
356 DNA may produce a higher proportion of target reads, as we also observed in our pilot experiment with
357 seven sgRNAs. The use of a higher number of sgRNAs per target was also proposed to be beneficial for
358 CANS of target genes (Gilpatrick *et al.* 2020). This is an important property of the CANS approach that
359 allows more targets to be included per experiment without losing the efficiency of CANS. Based on this
360 assumption, insertions from many TEs can be captured in one experiment using specifically designed and
361 pooled sgRNAs. Here, we used multiplex CANS to perform simultaneous detection of EVD and ONSEN
362 insertions after heat stress. This approach was also recently applied to detect TE insertions in the human
363 genome (McDonald *et al.* 2021).
364
365 We identified >800 TEIs in a small *ddm1* population after 2 hours of sequencing using only a part of the
366 capacity of MinION. This experiment provided proof of the concept that CANS is a valuable tool for
367 identifying TEIs in a population. A study on the distribution of TE insertions in the *A. thaliana* 1001
368 genomes population demonstrated that TEs are a major source of large-effect alleles that can be fixed in
369 the populations, contributing to the local adaptation of different accessions (Baduel *et al.* 2021). Indeed,
370 stress may induce TE activation and consequently the origination of new alleles. TEI-derived alleles have
371 spontaneously occurred in fields of crop species and have long been exploited in plant breeding programs
372 (Kobayashi, Goto-Yamamoto and Hirochika 2004; Jiang *et al.* 2009; Butelli *et al.* 2012; Domínguez *et*
373 *al.* 2020). Although the discovery of such alleles in crop fields using whole-genome sequencing is very
374 challenging, CANS might represent a valuable alternative to capture TEI-derived alleles in a population of
375 cultivated plants. However, prior knowledge of active TEs capable of transposing is required for such
376 experiments. Fortunately, this information is available for some plants. For example, we recently
377 identified active LTR retrotransposons in sunflower (Kirov *et al.* 2020b) and triticale (Kirov *et al.* 2020a).
378 There are many other examples of active plant TEs that can serve as potential targets for CANS-based
379 capture of TEIs. The development of high-throughput tools that can capture TE insertions in a population
380 represents a milestone towards the acceleration of the plant breeding process and more effective usage
381 of natural variation. Further increasing the coverage of sequencing using tools such as the PromethION
382 sequencer could greatly increase the sensitivity of this approach and might even allow the detection of
383 TEIs that have occurred in a single plant within an entire population. Such results could then be used to
384 design specific primers based on the insertion site to identify all plants carrying the target TEIs. The
385 capture of TEIs occurring in individual plants in a population will also depend on the genome size of the
386 species and the number of sgRNAs and target loci in the genome. These parameters must be optimized
387 for individual species.
388
389 Estimating the integration preferences for distinct TEs is not a straightforward process, as post-integration
390 selection pressure has a strong effect on the current TEI landscape. However, CANS is sensitive enough to

391 capture somatic TEIs for which post-integration selection pressure had only a minor effect. This feature
392 of CANS provides an opportunity for the estimation of the dynamic changes of mobilome in real-time.
393 This also means that CANS can provide more TE target genes as insertions that are deleterious to be
394 inherited can be detected. Here, we showed that EVD has preferences for integrations in the
395 pericentromeric regions of *A. thaliana* chromosomes in the ddm1 genome while ONSEN transposons tend
396 to integrate into genic regions and escape pericentromeric regions in Col-0 background. It is important to
397 note that the established EVD insertion distribution was shaped by very recent events and that post-
398 integration selection pressure had only a minor effect on this process. Recently published data (Baduel *et*
399 *al.* 2021) on TEI chromosome distribution in natural accessions of *A. thaliana* provides a great resource
400 for the comparison of the genomic organization of TEIs between natural and experimentally induced
401 conditions. Such analysis performed for ONSEN transposons showed that the experimentally induced (via
402 heat stress) and naturally occurred ONSEN insertions are well correlated at the genome level. Moreover,
403 we found 29 genes that were affected by ONSEN insertions in both datasets including FLC locus implying
404 the existence of hotspots of TE insertions that results from TE integration preferences rather than from
405 the natural selection pressure. It is worth noting that our ONSEN insertions landscape is also in
406 concordance with previous results (Roquis *et al.* 2021). EVD insertion preferences in *ddm1* mutant and
407 natural accessions of *A. thaliana* (Baduel *et al.* 2021) are not consistent: in *ddm1* genome, EVD frequently
408 targets pericentromeric regions while almost no EVD insertions were found in these regions in the
409 genomes of natural accessions. Proximal chromosome distribution of EVD insertions was also previously
410 shown in the genomes epiRIL lines (progeny of ddm1 x Col-0) (Quadrana *et al.* 2019). In hypomethylated
411 mutants (e.g. ddm1 and met1), histone H2A.Z is loaded to the TEs that lack hypermethylation (e.g. EVD).
412 Quadrana et al. 2019 (Quadrana *et al.* 2019) suggested the histone H2A.Z density is shifted toward
413 pericentromeric regions in ddm1 playing a role in the preferential integration of EVD into these regions.
414 In addition, pericentromeric heterochromatin of ddm1 chromosomes is more decondensed than in wild-
415 type (Probst *et al.* 2003) which can also make these regions more attractive for EVD insertions.
416 Our pioneering work demonstrates that CANS is a valuable and cost-effective tool for rapid deciphering
417 the mobilome landscapes of plant genomes and their changes in different genetic backgrounds and stress
418 conditions in real time.
419

420 Methods

421 Plant material and growth conditions

422 Seeds of ddm1 mutants (ddm1-2, F7 generation) were kindly provided by Vincent Colot (Institut de
423 Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France). Arabidopsis thaliana Col-0 plants (wild type
424 and ddm1 mutants) were grown in a light chamber for a month under 22°C and long-day conditions (16h
425 light/8h dark). For sequencing of pooled plants, Arabidopsis seeds were surface-sterilized with 75%
426 ethanol for 2 min, then washed with 5% sodium hypochlorite for 5 min. After this, the seeds were rinsed
427 with sterile distilled water 3 times. 0,1% agarose was added to each tube and the seeds were resuspended
428 and dripped on Petri dishes with MS nutrient medium (Murashige and Skoog, 1962), supplemented with
429 3% of sucrose (PanReac AppliChem, Germany) and 1% of agar (PanReac AppliChem, Germany). Plates with
430 seeds were sealed with Parafilm (Pechiney Plastic Packaging Company, USA) and kept in the dark at 4°C
431 for 3 days for vernalization and synchronous germination. Afterwards, dishes were transferred into a light
432 chamber with 16h day/8h night photoperiods for further growth.
433

434 ONSEN activation *in vitro*

435 The principle of heat mediated activation of ONSEN has been previously described (Thieme *et al.* 2017).
436 For aseptic growth, seeds of *A. thaliana* were sterilized in 10% bleach for 10 min, rinsed three times with
437 sterile H2O, and sown on 9 cm Petri dishes containing half-strength MS media (1% sucrose, 0.6% agar

438 (Dia-M, 3440.0500), pH 5.7) supplemented with combination of sterile filtered 5 µM α-amanitin (dissolved
439 in H2O) and 40 µM zebularine (dissolved in DMSO) (Thieme *et al.* 2017). After sowing the plates were
440 placed at 4℃ in the dark for 2 days. Germinated seedlings were then grown in a growth chamber (23℃)
441 under long-day conditions (16:8h, light:dark). After 10 days, the plantlets were put at 6◦C for 24 h (lighting
442 condition unchanged). The plants were then submitted to 24 h of heat stress at 37℃ (HS). Then
443 transferred to long day conditions (23°C, 16h/8h light/dark photoperiod).

444 For seed collection from the plants after ONSEN activation the plants were grown according to the
445 previously described protocol (Pasternak, Ruperti and Palme 2020) . Briefly, plants were transferred from
446 α-amanitin/zebularine containing medium to TK1 medium in 1% agar and kept in containers with TK1 until
447 plants transit to bolting. Further, shoots were grown on Hg medium until seed formation.

448

449 ## Plant transformation

450 To create ddm1 plant carrying T-DNA od nucleotide sequence of the GAG open reading frame of EVD5
451 retrotransposon (AtCOPIA93 (Mirouze *et al.* 2009)). The GAG ORF sequence was synthesized by the Synbio
452 Technologies service (USA). The GAG fragment was cloned at the restriction sites BamH1 and Sac1 into
453 the pBI121 vector. The Agrobacterium transformation of *A. thaliana* plants was carried out as described
454 previously (Clough and Bent 1998). Primary transformants were selected based on their resistance to
455 kanamycin (50 lg/ml). At least 60 transgenic seeds of the T2 generation were selected and used for
456 segregation analysis. A plant from one T2 family (G-ddm1-3) with 3:1 segregation of the marker for
457 kanamycin resistance was then transferred to pots with soil, grown three weeks more and used for DNA
458 isolation for whole-genome and CANS sequencing.
459

460 ## HMW DNA isolation and size selection

461 High molecular weight DNA was isolated from 200-500 mg of fresh and young leaves that were
462 homogenized in liquid nitrogen. DNA isolation was carried out according to the previously published
463 protocol (https://www.protocols.io/view/plant-dna-extraction-and-preparation-for
464 -ont-seque-bcvyiw7w) with several modifications. Chloroform was used instead of dichloromethane
465 during the extraction step. Samples with added CTAB2 buffer were incubated at 65°C until the formation
466 of visible flakes. Dissolving of the precipitate in 1M NaCl was performed at 65°C for 15 minutes with
467 further cooling to the RT before adding isopropanol. Removal of contaminants residues and small-size
468 nucleotide fragments was performed by adding 2.5 volumes of 80% ethanol with further centrifugation
469 at 12000 g for 15 minutes. The obtained DNA pellet was washed with 70% ethanol and centrifuged at
470 12000 g for 5 minutes. Final DNA elution was carried out with 65 µL of nuclease-free water.

471 For whole-genome Nanopore sequencing and CANS, size-selection of the large DNA fragments was done
472 using SRE or XL Short Read Eliminator Kits (Circulomics, USA) according to the manufacture instructions.
473 The concentration and quality of isolated DNA were estimated by gel electrophoresis in 1% agarose gel,
474 NanoDrop One UV-Vis Spectrophotometer (Thermo Scientific, USA) and Quantus Fluorometer (Promega,
475 USA) using a DNA QuantiFluor ONE dsDNA System (Promega, USA). Only pure DNA (A260/A280 ~1.8 and
476 A260/A230 ~2.0 according to NanoDrop) with almost no differences in concentrations obtained by
477 Nanodrop and Quantus was used for sequencing.
478

479 ## sgRNA design

480 To design sgRNAs recognizing multiple copies of LTR retrotransposons we used a step-by-step strategy:
481 (1) using FlashFry 'discover' tool with the '-positionOutput -maxMismatch 0 ' arguments and reference
482 genome sequences (TAIR10) a pool of sgRNAs and the corresponding genomic sites were found followed
483 by bed file generation; (2) reference sequences of target LTR retrotransposons were blasted against the
484 corresponding genomes followed by hit filtering (minimum hit length 100bp and minimum e-value 1e-50)
485 and bed file consisting blast hit positions was generated; (3) the genomic location of predicted sgRNA sites
486 (step 1) were intersected with blast hit bed file (step 2) using bedtools 'intersect' and OnOutTE ratio was
487 calculated for each sgRNAs:

488 OnOutTE = number of sgRNA sites located on the blast hit from the corresponding TE / total number of
489 sgRNA sites

490 (4) We selected sgRNAs with OnOutTE ratio > 0.85.

491 For each retrotransposon a set of sgRNAs were designed including one (*A. thaliana*) sgRNAs on negative
492 strand recognizing both LTRs and three sgRNAs (one on negative and two on positive strands) located
493 between two LTR sequences. Sequences of sgRNAs are provided in Supplemental Data S 1.

494 ## In vitro transcription of sgRNAs

495 SgRNAs for CANS were produced by in vitro transcription from double-stranded DNA templates
496 containing T7 promoter that were assembled from two oligos, specific (sgRNA1-5 in Supplemental Data
497 S1) and universal (CRISPR_R). All oligonucleotides were ordered in Evrogen (Moscow, Russia). To
498 synthesize double-stranded DNA templates, the following reagents were mixed per each reaction: 2µl of
499 unique sgRNA oligo (1 µM); 2µl of CRISPR_R (1 µM); 2µl of T7 forward (5'-
500 GGATCCTAATACGACTCACTATAG-3') and reverse (5'-AAAAAAGCACCGACTCGG-3') primers (100 µM of
501 each); 2µl of 50x dNTP (10 µM) (Evrogen, Moscow, Russia); 10µl of 10x Encyclo buffer (Evrogen, Moscow,
502 Russia); 1µl of Encyclo polymerase (Evrogen, Moscow, Russia); 79µl of RNAse-free water. Oligos were
503 annealed and extended according to the PCR program: initial denaturation at 95°C (2 min); 30 cycles of
504 denaturation at 98°C (30 sec), annealing at 60°C (30 sec), elongation at 72°C (30 sec); final elongation at
505 72°C (1 min). Gel electrophoresis was performed after amplification and templates were column purified.
506 To transcribe the templates, the Highly Efficient RNA in vitro Synthesis Kit (Biolabmix, Novosibirsk,
507 Russia) was used. The subsequent reactions were set up and incubated at 37°C for 2h: 10µl of 5x T7
508 transcription buffer; 2µl of 25x DTT (250mM); 2µl of dNTP (25mM of each); 2µl of double-stranded DNA
509 template (150-200 ng); 1µl of T7 polymerase (150 units); 33µl RNAse-free water. The sgRNAs were then
510 purified using RNA and miRNA Extraction Kit (Biolabmix, Novosibirsk, Russia) according to the
511 manufacturer instructions. The concentration and quality of prepared sgRNAs were estimated by
512 Nanodrop (Thermo Scientific, USA), Qubit (Thermo Scientific, USA) and gel electrophoresis in 2% agarose
513 gel.
514

515 ## Cas9/sgRNA ribonucleoprotein complexes (RNPs) assembly

516 For CANS, 50 ng of each sgRNA were used for RNP assembly. To obtain RNPs for sgRNA Mixture 2
517 the corresponding sgRNAs were pooled together at 1:1 molar ratio in 11 µl of MQ water. Before RNP
518 assembly all sgRNA mixtures were denaturated at 95°C for 5 min, then cooled on ice. The RNPs were
519 assembled by combining 11µl of sgRNA mix with 1µl of Cas9 nuclease in 3 µl of reaction buffer (Biolabmix,
520 Novosibirsk, Russia) at a final volume of 15µl followed by incubation for 30 min at room temperature and
521 kept on ice until usage.

522

523 ## Cas9 cleavage and library preparation

524 Before CANS DNA was dephosphorylated. For this, 2-8 µg of HMW DNA was diluted in 1x CutSmart Buffer
525 and 6 µl of Quick CIP enzyme (New England Biolabs, catalog no. M0508) was added followed by incubation
526 of the reaction at 37°C for 30 minutes. The reaction was stopped by heating (80°C, 2 minutes). DNA
527 cleavage by RNPs was carried out by mixing of 40 µl of dephosphorylated DNA, 15 µl of RNPs, 1.5 µl of
528 dATP (10 mM) and 1 µl Encyclo polymerase (Evrogen, Moscow, Russia). The mixture was incubated at
529 37°C for 30 minutes followed by 5 min at 72 °C. Nanopore sequencing adapters (AMX) were ligated to the
530 Cas9 cleaved DNA by mixing the following components: 25 µl of LNB buffer (Oxford Nanopore
531 Technologies, catalog no. SQK-LSK109), 5 µl of nuclease-free water, 12.5 µl of Quick T4 DNA Ligase (New
532 England Biolabs, NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing
533 catalog no. E7180S) and Adapter Mix (SQK-LSK109). The mixture was added to the Cas9 cleaved DNA and
534 incubated for 20 min at room temperature. The samples were purified by adding equal volume of TE
535 buffer (pH 8.0) and 0.3 volume of AMPure XP Beads (Beckman Coulter, catalog no. A63881) and washed
536 twice by LFB buffer (Oxford Nanopore Technologies, catalog no. SQK-LSK109). The DNA was eluted in 15 µl
537 of elution buffer (Oxford Nanopore Technologies, catalog no. SQK-LSK109). Then sequencing library was
538 prepared according to the Ligation Sequencing Kit SQK-LSK109 protocol (Oxford Nanopore Technologies,
539 catalog no. SQK-LSK109).

540 Library preparation for whole-genome sequencing was carried out using the Ligation Sequencing Kit
541 (Oxford Nanopore Technologies, catalog no. SQK-LSK109) with 1 µg of input HMW DNA.

542 ## Sequencing and basecalling

543 Sequencing was performed by MinION equipped by R9.4.1 or R10.3 flow cells. Sequencing process was
544 operated by MinKNOW software (v.19.12.5). The detailed information on Nanopore sequencing is
545 described in Supplemental Data Set 2.

546 ## Analysis

547 Base calling was done by Guppy (Version 3.2.10). The TAIR10 genome was downloaded from NCBI
548 database. The generated Nanopore reads were aligned to the reference genomes (TAIR10 for *A.thaliana*)
549 using minimap2 software (Li 2018) with the following parameters: -ax map-ont -t 100. The obtained sam
550 file was converted to bam format, sorted and indexed using SAMtools (Li *et al.* 2009). The obtained sorted
551 bam files were used for NanoCasTE (CANS data) and nanotei (WGS data) pipelines.

552 ## NanoCasTE pipeline for the identification of transposon insertions following CANS

553 NanoCasTE is written in python and can be run as a standalone tool or via Snakefile
554 (https://github.com/Kirovez/NanoCasTE). NanoCasTE involves several steps (Figure 6): (1) Mapping the
555 reads to the genome using minimap2 (Li 2018), followed by the generation of sorted bam files using
556 SAMtools (Li *et al.* 2009); (2) Filtering the reads based on mapping quality (MQ > 40), read length (>750)
557 and SA tags (no supplementary alignments are allowed); (3) Selecting mapped reads with clipped heads
558 (reads mapped to the positive strand) or tails (reads mapped to the negative strand) with the length of
559 the clipped parts close to the distance from the sgRNA positions to the TE end; (4) Filtering the reads
560 based on similarity of the clipped part to the target TE sequence; (5) Identifying the inserted sites and TE
561 orientation (+ or - strand) and outputting the results.

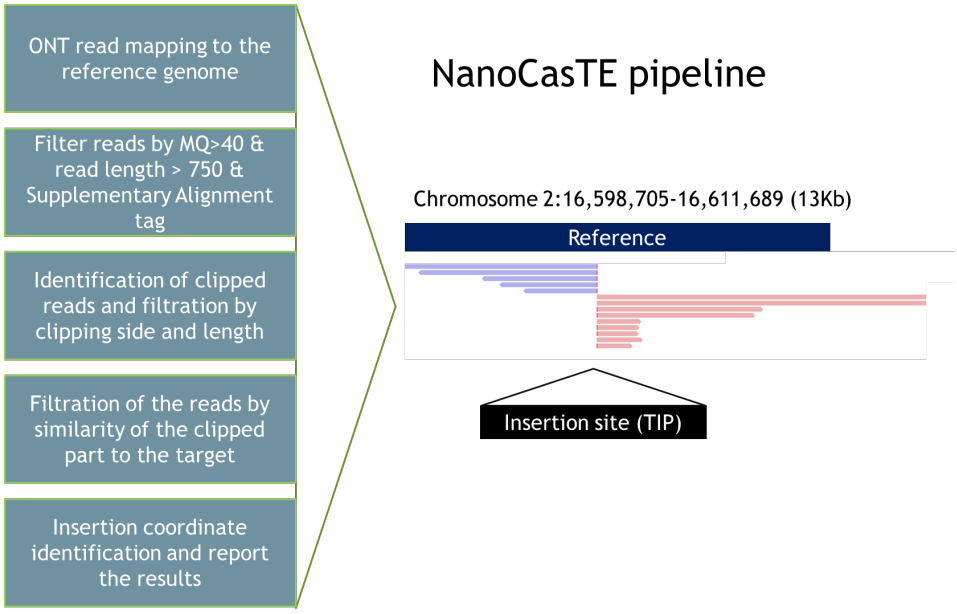


Figure 6. Schematic view of the major steps of the NanoCasTE pipeline to identify transposon insertions in the genome after CANS

NanoCasTE uses a set of stringent criteria to specifically detect TE insertions in CANS data and to distinguish them from noise signals. The pipeline reports the putative positions of the target TE insertion as well as additional information that is useful for further analysis, including the number of selected reads supporting the insertion, the total number of reads covering the insertion, the strand harboring the TE insertion and the length of the clipped parts of the selected reads.

NanoCasTE was run for all samples using the following parameters: min_len_clipped: 0.3, mapping_quality: 60, min_read_length: 700. The following lists of sgRNAs were used for *A. thaliana EVD:* 'TCTTGGTGATGAGAGTGAC, ACCCTGGATTTAAGGTGAGA, AGTTTAAGAGCTCTAGTATG, CTACAAGGTCAATCGAAAGG, TCAACACATGAAAGTCCCGA'; for *A. thaliana* ONSEN: 'GTTTAAAGCCCATGTTGAGA,TCAAGGAAGGAAATGGTGAG,ACCGTTTCTAGACGAGCAAC,TCGAGTATTCCAA GCTCTTG'.

Methylation Calling

Methylation calling from raw Nanopore data was performed using DeepSignal-plant software (Ni *et al.* 2021b) as previously described (Kirov *et al.* 2021b). For methylation visualization, the custom-made script DeepS2bam_converter (https://github.com/Kirovez/DeepS2bam_converter, accessed on 10 March 2021) was utilized. The read alignments and per-read methylation profile were visualized by JBrowse2 (Buels *et al.* 2016).

Insertion validation by PCR

For validations of TEIs we used the combination of primers including one primer located on TEI flanking regions and one primer located on TE. The sequences of primers are listed in Supplemental Data S3. The PCR was performed in a reaction volume of 25 µl using Bio-Rad T100 Thermal Cycler (Bio-Rad, USA). The reaction mixture contained 2.5 mM MgCl2, 200 mM dNTPs (Dia-M, Moscow, Russia), 5 pmols of each primer, 0.5 units of ColoredTaq polymerase (Sileks, Russia). The PCR reactions were carried out under the following conditions: Touchdown-PCR was carried out according to the following cycling program: 94 ◦C for 3 min, 94 ◦C for 30 s, 65 ◦C for 30 s and 72 ◦C for 2 min, followed by 6 cycles at decreasing annealing

592 temperatures in decrements of 1 ◦C per cycle, then 35 cycles of 30 s at 94 ◦C, 30 s at 55 ◦C, 2 min at 72 ◦C
593 and final extension at 72◦C for 5 min. For validation of *A. thaliana* TEIs the following PCR program was
594 used: 35 cycles of 94◦C for 30 s, 60◦C for 30 s and 72 ◦C for 2 min. Amplification products were visualized
595 by electrophoresis in 1.5% agarose gel and observed under UV light after ethidium bromide staining.

596

597 ## Statistics and Data Visualization

598 Statistical analysis was carried out in Rstudio Version 1.2.1335 (http://www.rstudio.com/) with R version
599 3.6.0. Visualization was carried out by ggplot2 (Wickham 2011) and VennDiagram (Chen and Boutros
600 2011) packages. The circus plots were drawn using Circos software (Krzywinski *et al.* 2009) implemented
601 in Galaxy server (Rasche and Hiltemann 2020).

602

603 ## Accession Numbers

604 All Nanopore data generated in this study are deposited in the National Center for Biotechnology
605 Information SRA database (BioProject ID PRJNA736208).
606

607 ## Acknowledgements

608 We thank Vincent Colot (Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France) for
609 providing the *ddm1* plant material and the Russian Foundation for Basic Research (grant №20-34-70032).

610

611 ## Financial support

612 This study was financially supported by Kurchatov Genomic Center of All-Russia Research Institute of
613 Agricultural Biotechnology, agreement № 075-15-2019-1667.

614

615 ## Author contribution statement

616 I.K. and M.D. designed the research; I.K., P.M., S.G., R.K., M.O., M.D., Z.K. and A.K. performed the
617 research. I.K., G.K., A.S. and M.D. analyzed data. I.K. wrote the article.

618

619 ## References

620 Baduel P, Leduque B, Ignace A *et al.* Genetic and environmental modulation of transposition
621 shapes the evolutionary potential of Arabidopsis thaliana. *Genome Biology* 2021;22:138.

622 Buels R, Yao E, Diesh CM *et al.* JBrowse: a dynamic web platform for genome visualization and
623 analysis. *Genome Biol* 2016;17:66.

624 Butelli E, Licciardello C, Zhang Y *et al.* Retrotransposons control fruit-specific, cold-dependent
625 accumulation of anthocyanins in blood oranges. *Plant Cell* 2012;24:1242–55.

626 Carpentier MC, Manfroi E, Wei FJ *et al.* Retrotranspositional landscape of Asian rice revealed by
627 3000 genomes. *Nat Commun* 2019;10:24.

628 Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and
629 Euler diagrams in R. *BMC Bioinformatics* 2011;12:35.

630 Chen J, Lu L, Robb SMC *et al.* Genomic diversity generated by a transposable element burst in a
631 rice recombinant inbred population. *Proceedings of the National Academy of Sciences of the*
632 *United States of America* 2020;117:26288–97.

633 Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts
634 to benefits. *Nature Reviews Genetics* 2017;18:71–86.

635 Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation
636 of Arabidopsis thaliana. *Plant J* 1998;16:735–43.

637 Dmitriev AA, Pushkova EN, Novakovskiy RO *et al.* Genome Sequencing of Fiber Flax Cultivar Atlant
638 Using Oxford Nanopore and Illumina Platforms. *Front Genet* 2021;11:590282–590282.

639 Domínguez M, Dugas E, Benchouaia M *et al.* The impact of transposable elements on tomato
640 diversity. *Nature Communications* 2020;11:4058.

641 Gabrieli T, Sharim H, Fridman D *et al.* Selective nanopore sequencing of human BRCA1 by Cas9-
642 assisted targeting of chromosome segments (CATCH). *Nucleic Acids Research* 2018;46:e87–
643 e87.

644 Gilpatrick T, Lee I, Graham JE *et al.* Targeted nanopore sequencing with Cas9-guided adapter
645 ligation. *Nature Biotechnology* 2020;38:433–8.

646 Handsaker RE, Korn JM, Nemesh J *et al.* Discovery and genotyping of genome structural
647 polymorphism by sequencing on a population scale. *Nat Genet* 2011;43:269–76.

648 Ito H, Gaubert H, Bucher E *et al.* An siRNA pathway prevents transgenerational retrotransposition
649 in plants subjected to stress. *Nature* 2011;472:115–9.

650 Jiang N, Gao D, Xiao H *et al.* Genome organization of the tomato sun locus and characterization
651 of the unusual retrotransposon Rider. *Plant J* 2009;60:181–93.

652 Jung H, Winefield C, Bombarely A *et al.* Tools and Strategies for Long-Read Sequencing and De
653 Novo Assembly of Plant Genomes. *Trends in Plant Science* 2019;24:700–24.

654 Kirov I, Dudnikov M, Merkulov P *et al.* Nanopore RNA Sequencing Revealed Long Non-Coding and
655 LTR Retrotransposon-Related RNAs Expressed at Early Stages of Triticale SEED Development.
656 *Plants* 2020a;9:1794.

657 Kirov I, Merkulov P, Dudnikov M *et al.* Transposons Hidden in Arabidopsis thaliana Genome
658 Assembly Gaps and Mobilization of Non-Autonomous LTR Retrotransposons Unravelled by
659 Nanotei Pipeline. *Plants* 2021a;10:2681.

660 Kirov I, Omarov M, Merkulov P *et al.* Genomic and Transcriptomic Survey Provides New Insight
661 into the Organization and Transposition Activity of Highly Expressed LTR Retrotransposons of
662 Sunflower (Helianthus annuus L.). *Int J Mol Sci* 2020b;21:9331.

663 Kirov I, Polkhovskaya E, Dudnikov M *et al.* Searching for a Needle in a Haystack: Cas9-Targeted
664 Nanopore Sequencing and DNA Methylation Profiling of Full-Length Glutenin Genes in a Big
665 Cereal Genome. *Plants* 2021b;11:5.

666 Kirov I, Polkhovskaya E, Dudnikov M *et al.* Searching for a Needle in a Haystack: Cas9-Targeted
667 Nanopore Sequencing and DNA Methylation Profiling of Full-Length Glutenin Genes in a Big
668 Cereal Genome. *Plants* 2022;11:5.

669 Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin
670 color. *Science* 2004;304:982.

671 Krzywinski M, Schein J, Birol İ *et al.* Circos: An information aesthetic for comparative genomics.
672 *Genome Res* 2009;19:1639–45.

673 Lee YG, Choi SC, Kang Y *et al.* Constructing a Reference Genome in a Single Lab: The Possibility to
674 Use Oxford Nanopore Technology. *Plants* 2019;8:270.

675 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.

676 Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools.
677 *Bioinformatics* 2009;25:2078–9.

678 Li S, Jia S, Hou L *et al.* Mapping of transgenic alleles in soybean using a nanopore-based
679 sequencing strategy. *Journal of Experimental Botany* 2019;70:3825–33.

680 Lisch D %J NRG. How important are transposons for plant evolution? *Nature Reviews Genetics*
681 2013;14:49–61.

682 López-Girona E, Davy MW, Albert NW *et al.* CRISPR-Cas9 enrichment and long read sequencing
683 for fine mapping in plants. *Plant Methods* 2020;16:121.

684 Madsen EB, Höijer I, Kvist T *et al.* Xdrop: Targeted sequencing of long DNA molecules from low
685 input samples using droplet sorting. *Hum Mutat* 2020;41:1671–9.

686 McDonald TL, Zhou W, Castro C *et al.* Cas9 targeted enrichment of mobile elements using
687 nanopore sequencing. 2021:2021.02.10.430605.

688 Mirouze M, Reinders J, Bucher E *et al.* Selective epigenetic control of retrotransposition in
689 Arabidopsis. *Nature* 2009;461:427–30.

690 Ni P, Huang N, Nie F *et al.* Genome-wide Detection of Cytosine Methylations in Plant from
691 Nanopore sequencing data using Deep Learning. 2021a:2021.02.07.430077.

692 Ni P, Huang N, Nie F *et al.* Genome-wide Detection of Cytosine Methylations in Plant from
693 Nanopore sequencing data using Deep Learning. 2021b:2021.02.07.430077.

694 Nuthikattu S, McCue AD, Panda K *et al.* The Initiation of Epigenetic Silencing of Active
695 Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs.
696 2013;162:116–31.

697 Panda K, Slotkin RK. Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of
698 Transposable Elements. *The Plant Cell* 2020a;32:2687–98.

699 Panda K, Slotkin RK. Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of
700 Transposable Elements. *The Plant Cell* 2020b;32:2687–98.

701 Pasternak T, Ruperti B, Palme K. A simple high efficiency and low cost in vitro growth system for
702 phenotypic characterization and seed propagation of Arabidopsis thaliana. *Biorxiv*
703 2020:2020.08.23.263491.

704 Probst AV, Fransz PF, Paszkowski J *et al.* Two means of transcriptional reactivation within
705 heterochromatin. *Plant J* 2003;33:743–9.

706 Quadrana L, Etcheverry M, Gilly A *et al.* Transposition favors the generation of large effect
707 mutations that may facilitate rapid adaption. *Nat Commun* 2019;10:3421.

708 Quadrana L, Silveira AB, Mayhew GF *et al.* The Arabidopsis thaliana mobilome and its impact at
709 the species level. *eLife* 2016a;5:e15716.

710 Quadrana L, Silveira AB, Mayhew GF *et al.* The Arabidopsis thaliana mobilome and its impact at
711 the species level. *eLife* 2016b;5:e15716.

712 Rabanus-Wallace MT, Hackauf B, Mascher M *et al.* Chromosome-scale genome assembly
713 provides insights into rye biology, evolution and agronomic potential. *Nature Genetics*
714 2021;53:564–73.

715 Rasche H, Hiltemann S. Galactic Circos: User-friendly Circos plots within the Galaxy platform.
716 *Gigascience* 2020;9:giaa065-.

717 Roquis D, Robertson M, Yu L *et al.* Genomic impact of stress-induced transposable element
718 mobility in Arabidopsis. *Nucleic Acids Res* 2021;49:10431–47.

719 Schalamun M, Nagar R, Kainer D *et al.* Harnessing the MinION: An example of how to establish
720 long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus
721 pauciflora. *Mol Ecol Resour* 2019;19:77–89.

722 Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome.
723 *Nature Reviews Genetics* 2007;8:272–85.

724 Stangl C, Blank S de, Renkens I *et al.* Partner independent fusion gene detection by multiplexed
725 CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nature Communications*
726 2020;11:2861.

727 Sultana T, Zamborlini A, Cristofari G *et al.* Integration site selection by retroviruses and
728 transposable elements in eukaryotes. *Nature Reviews Genetics* 2017;18:292–308.

729 Thieme M, Lanciano S, Balzergue S *et al.* Inhibition of RNA polymerase II allows controlled
730 mobilisation of retrotransposons for plant breeding. *Genome Biol* 2017;18:134.

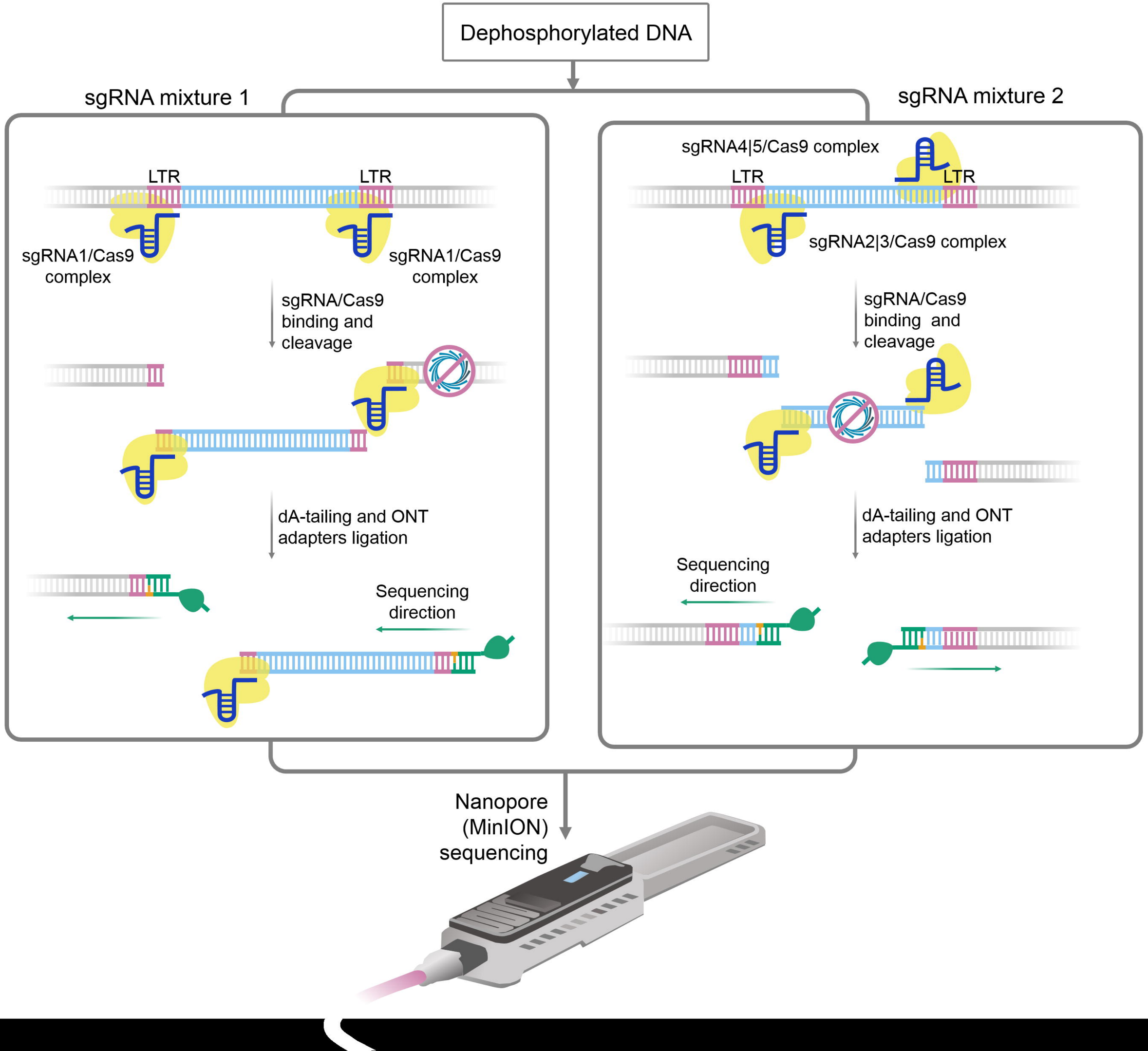
731 Tsukahara S, Kobayashi A, Kawabe A *et al.* Bursts of retrotransposition reproduced in Arabidopsis.
732 *Nature* 2009;461:423–6.

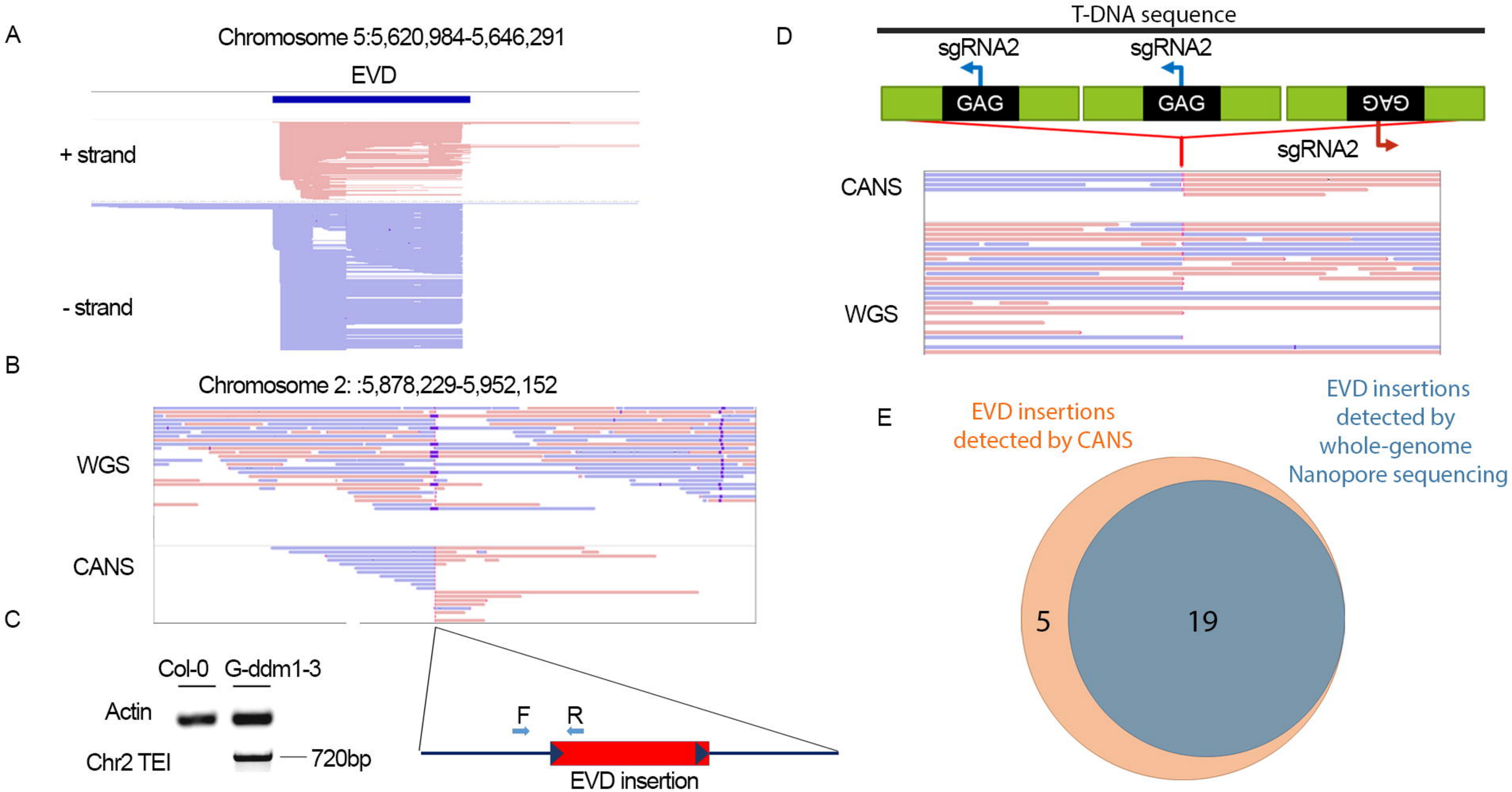
733 Wickham H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics 32* 2011;3:180–5.

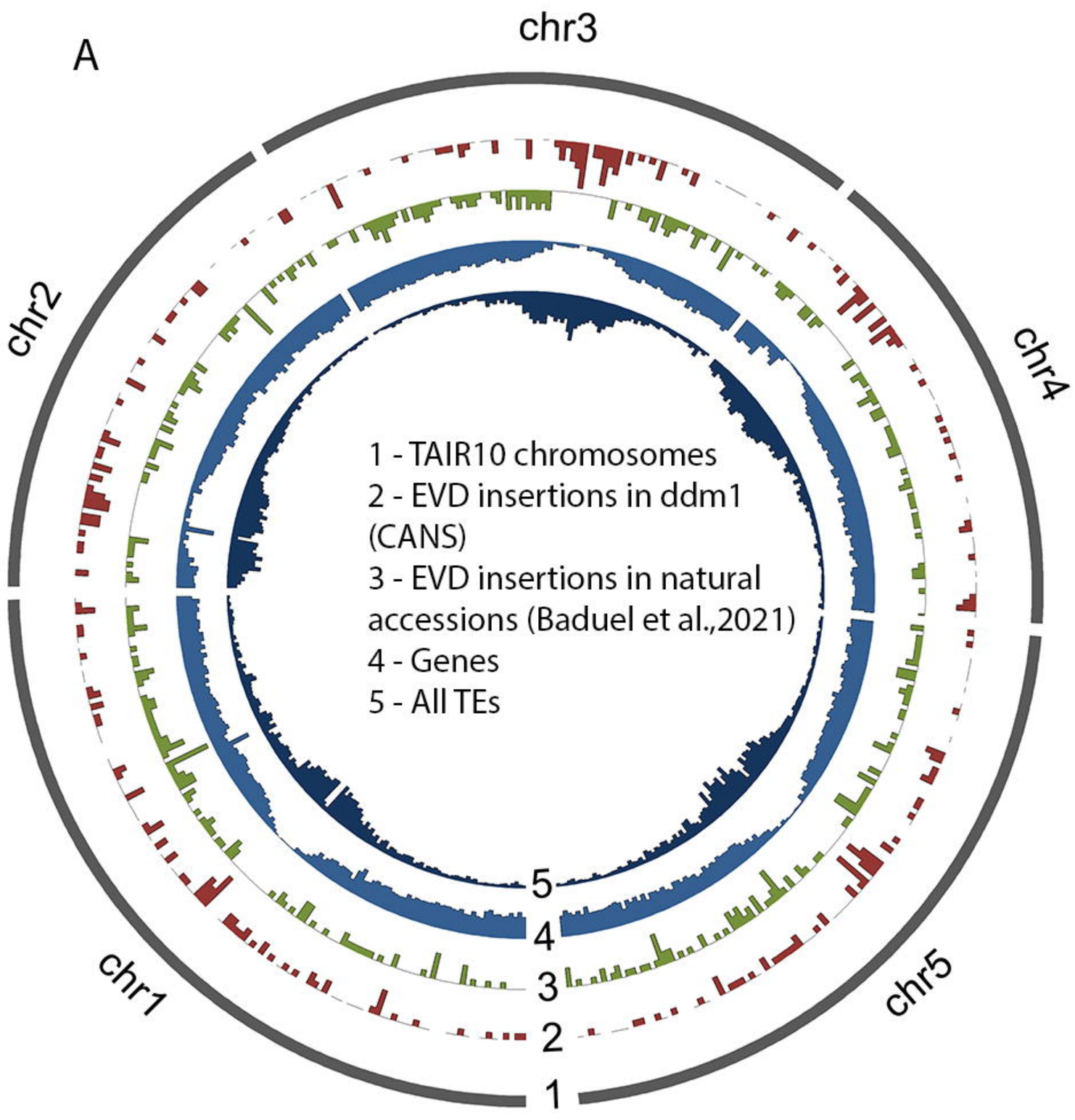
734 Williams-Carrier R, Stiffler N, Belcher S *et al.* Use of Illumina sequencing to identify transposon
735 insertions underlying mutant phenotypes in high-copy Mutator lines of maize. *Plant J*
736 2010;63:167–77.

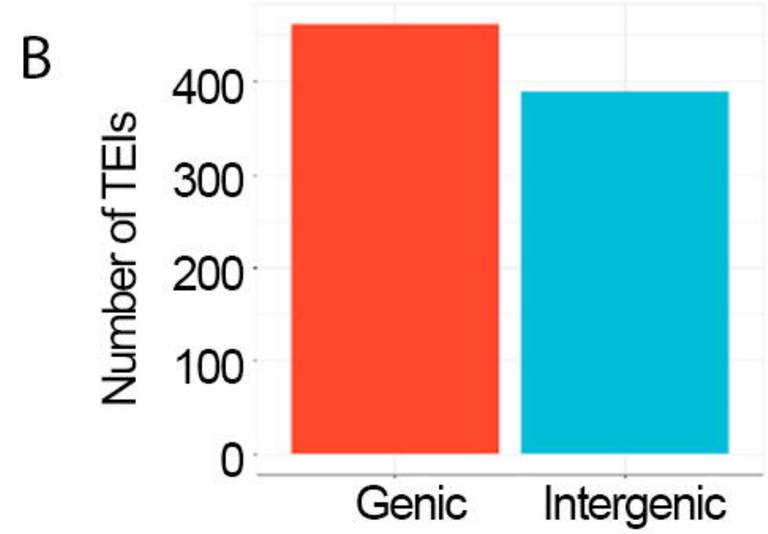
737

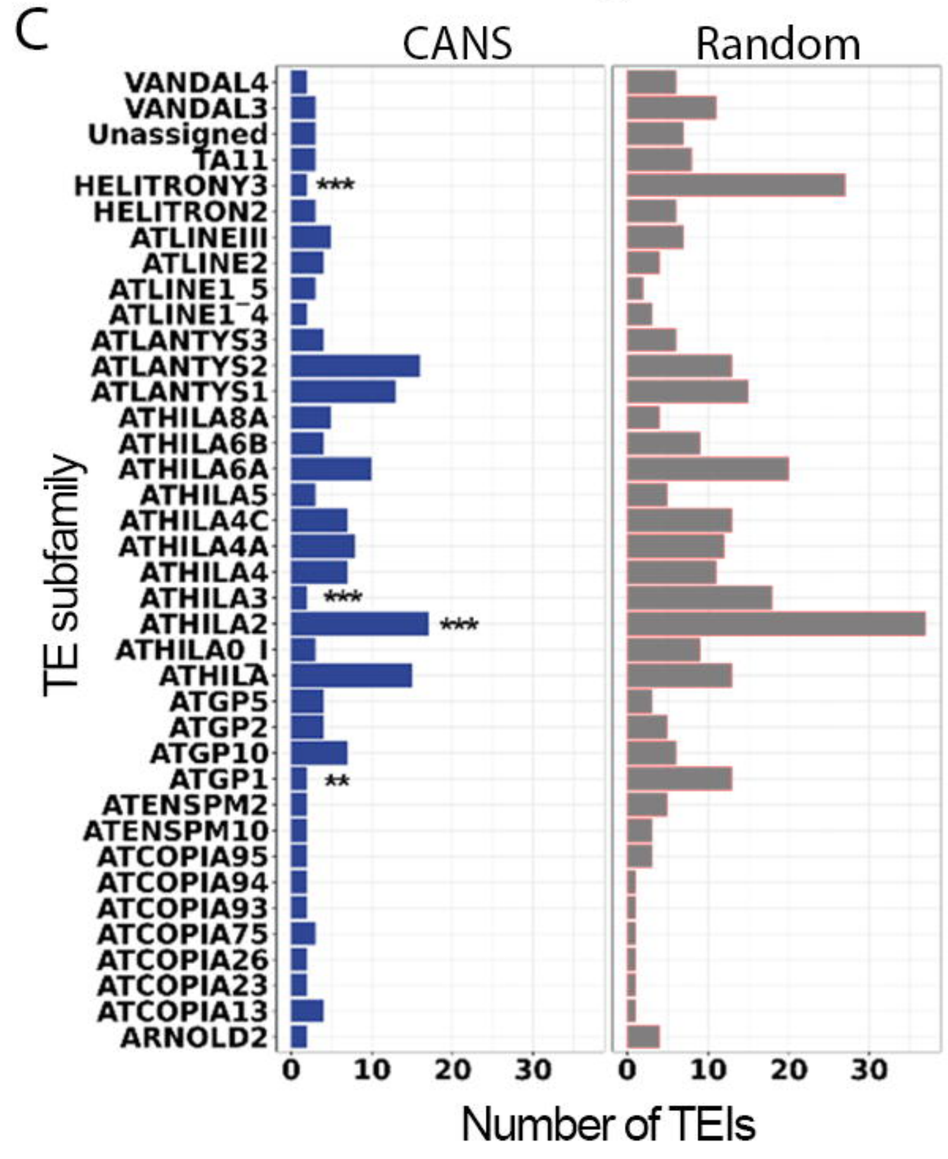
738


Dephosphorylated DNA
sgRNA mixture 1
sgRNA mixture 2
LTR
LTR
sgRNA1/Cas9 complex
sgRNA1/Cas9 complex
sgRNA/Cas9 binding and cleavage
dA-tailing and ONT adapters ligation
Sequencing direction
sgRNA4|5/Cas9 complex
LTR
LTR
sgRNA2|3/Cas9 complex
sgRNA/Cas9 binding and cleavage
dA-tailing and ONT adapters ligation
Sequencing direction
Nanopore (MinION) sequencing


A
Chromosome 5:5,620,984-5,646,291
EVD
+ strand
- strand
B
Chromosome 2: :5,878,229-5,952,152
WGS
CANS
C
Col-0
G-ddm1-3
Actin
Chr2 TEI
720bp
F
R
EVD insertion
D
T-DNA sequence
sgRNA2
sgRNA2
GAG
GAG
GAG
sgRNA2
CANS
WGS
E
EVD insertions detected by CANS
EVD insertions detected by whole-genome Nanopore sequencing
5
19


A
chr3
chr2
chr4
chr1
chr5
1 - TAIR10 chromosomes
2 - EVD insertions in ddm1 (CANS)
3 - EVD insertions in natural accessions (Baduel et al.,2021)
4 - Genes
5 - All TEs
5
4
3
2
1


B
Number of TEIs
400
300
200
100
0
Genic
Intergenic


C
CANS
Random
TE subfamily
VANDAL4
VANDAL3
Unassigned
TA11
HELITRONY3
HELITRON2
ATLINEIII
ATLINE2
ATLINE1_5
ATLINE1_4
ATLANTYS3
ATLANTYS2
ATLANTYS1
ATHILA8A
ATHILA6B
ATHILA6A
ATHILA5
ATHILA4C
ATHILA4A
ATHILA4
ATHILA3
ATHILA2
ATHILA0_I
ATHILA
ATGP5
ATGP2
ATGP10
ATGP1
ATENSPM2
ATENSPM10
ATCOPIA95
ATCOPIA94
ATCOPIA93
ATCOPIA75
ATCOPIA26
ATCOPIA23
ATCOPIA13
ARNOLD2
***
***
***
**
0
10
20
30
0
10
20
30
Number of TEIs

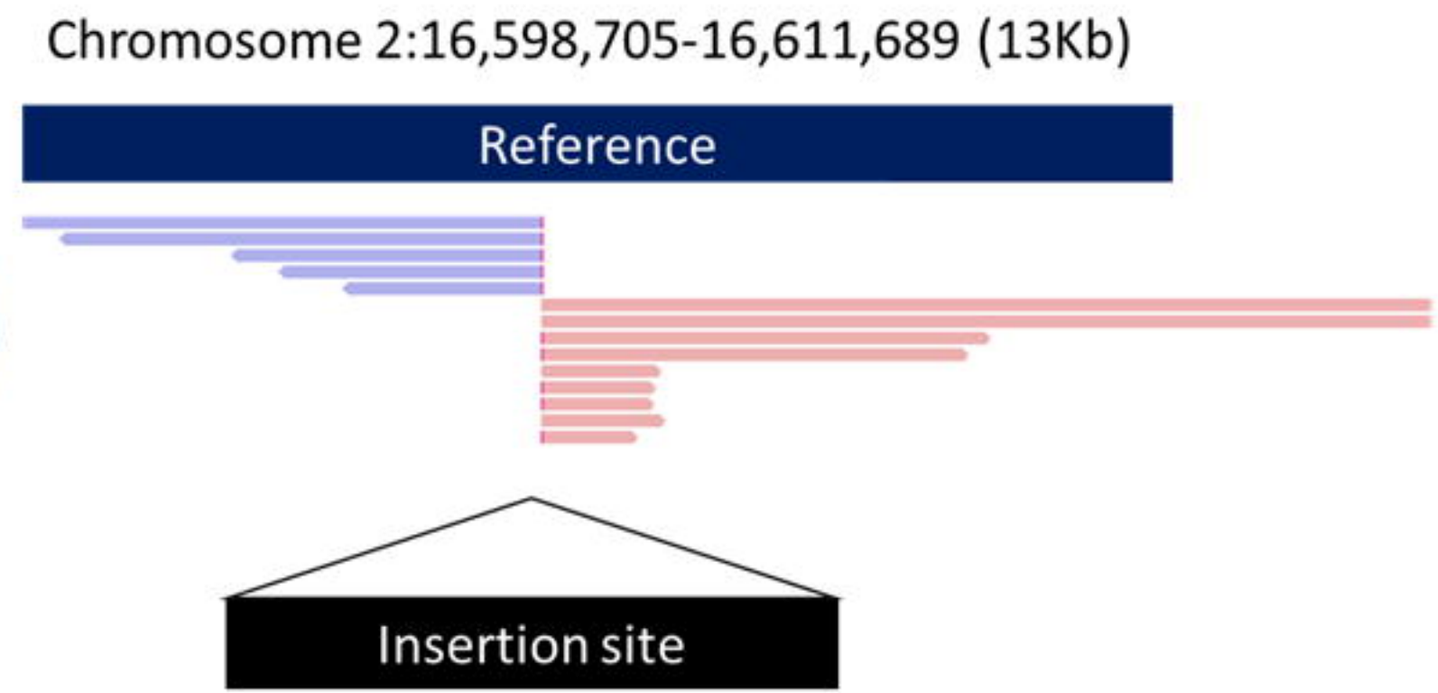
ONT read mapping to the reference genome

Filter reads by MQ>40 & read length > 750 & Supplementary Alignment tag

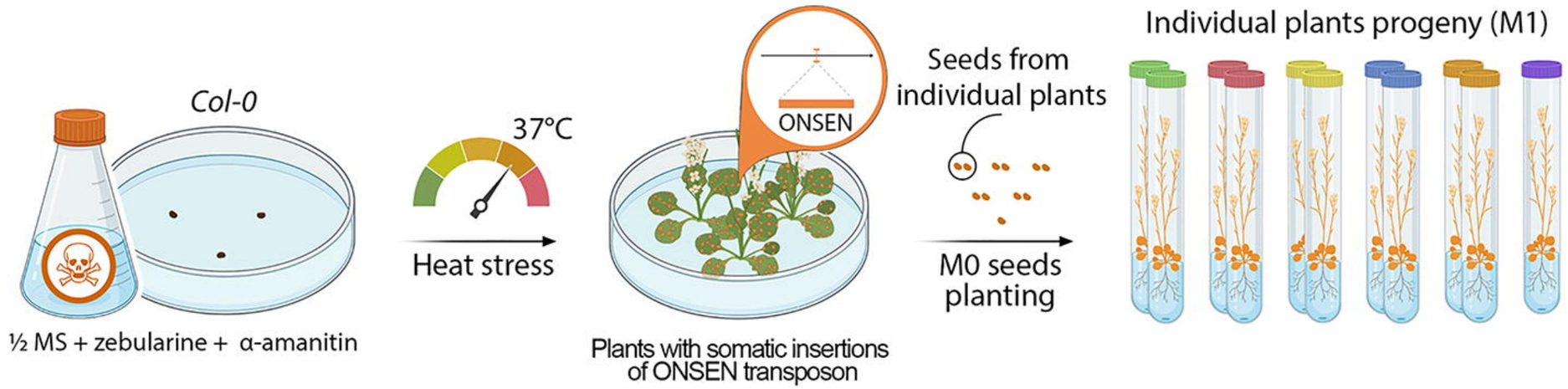
Identification of clipped reads and filtration by clipping side and length

Filtration of the reads by similarity of the clipped part to the target

Insertion coordinate identification and report the results

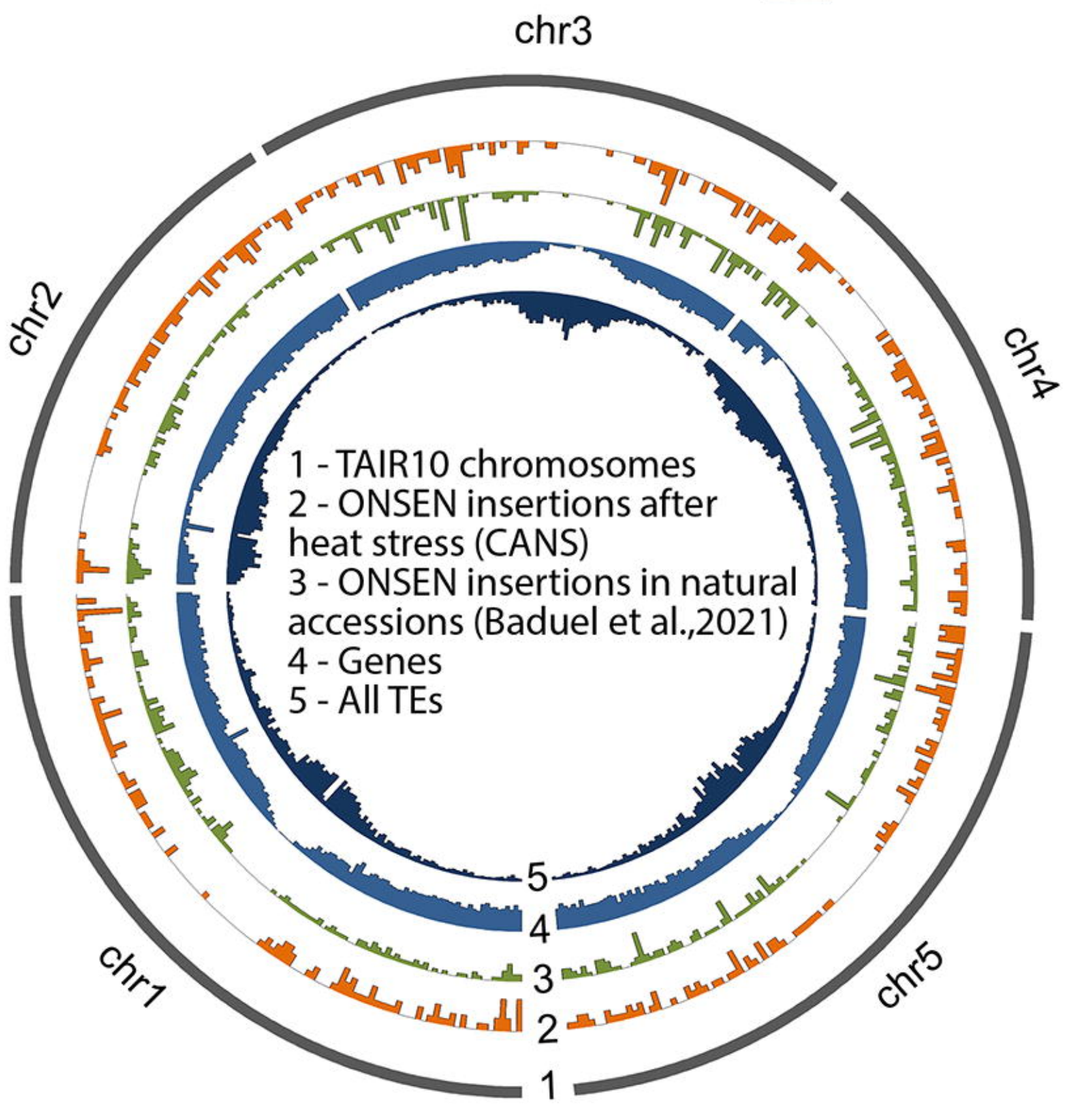


A



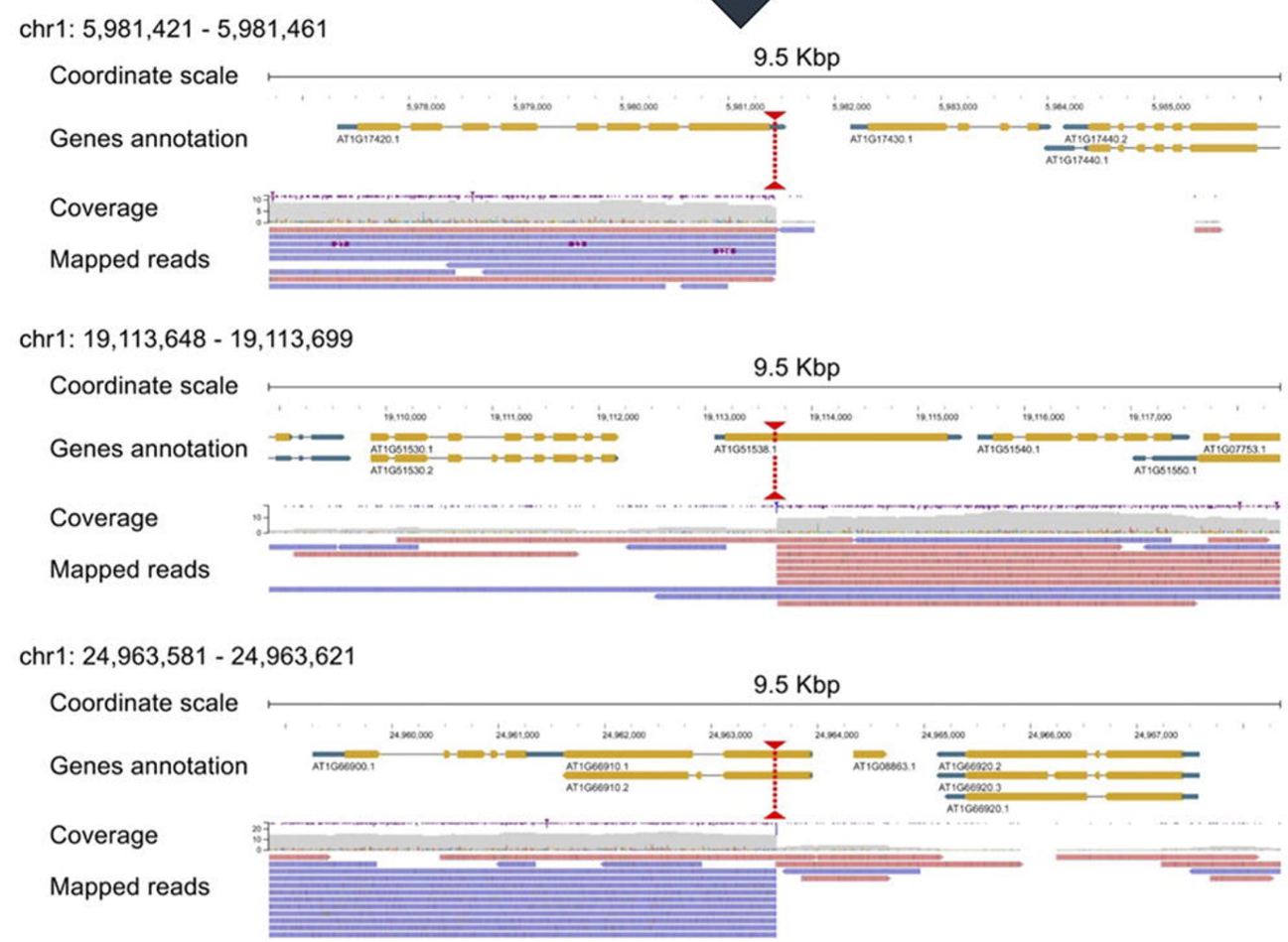
B

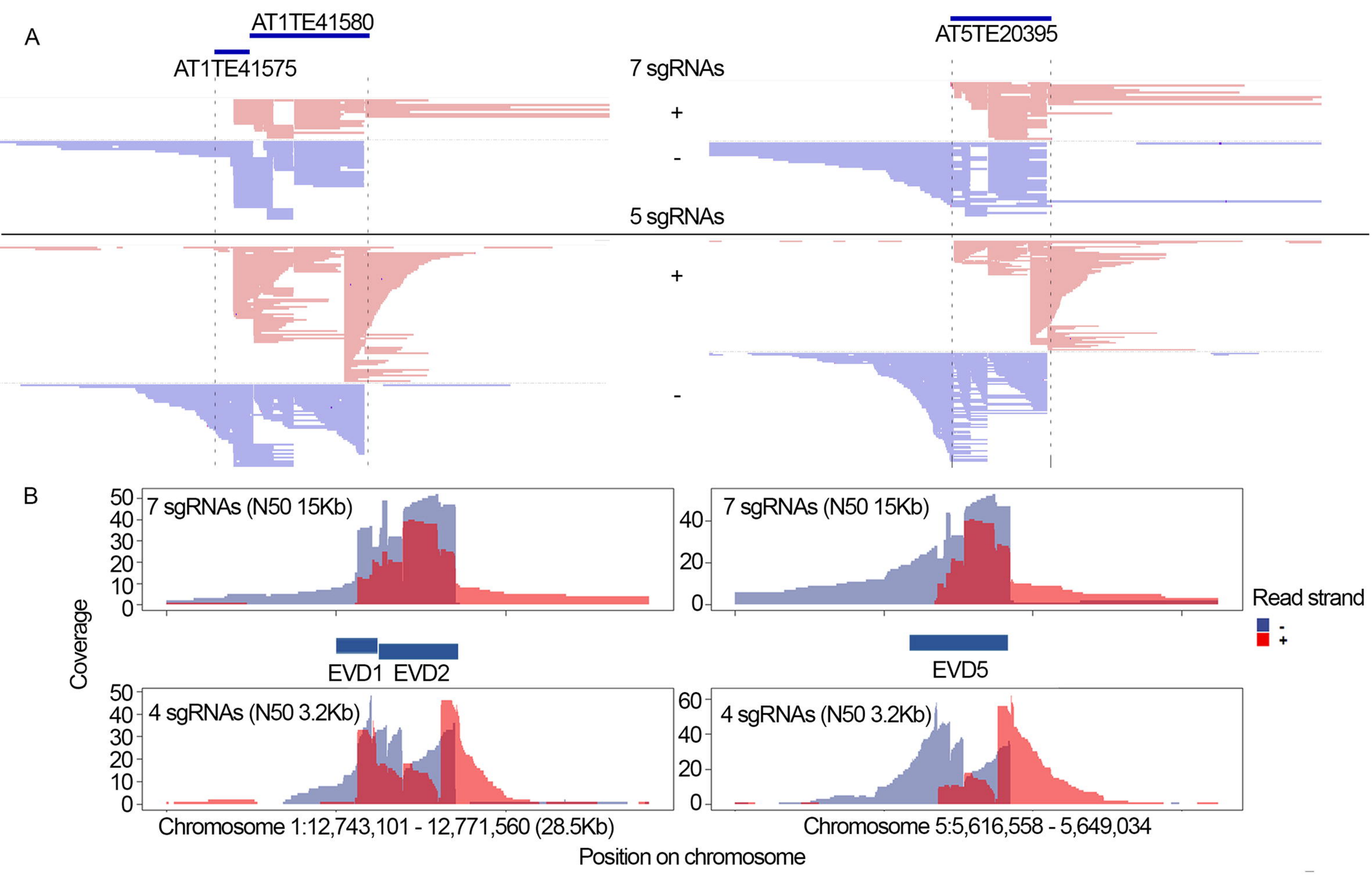
CANS (somatic TEIs)



C

CANS (non-somatic TEIs)




A
AT1TE41580
AT1TE41575
AT5TE20395
7 sgRNAs
+
-
5 sgRNAs
+
-
B
7 sgRNAs (N50 15Kb)
7 sgRNAs (N50 15Kb)
Coverage
EVD1 EVD2
EVD5
4 sgRNAs (N50 3.2Kb)
4 sgRNAs (N50 3.2Kb)
Read strand
-
+
Chromosome 1:12,743,101 - 12,771,560 (28.5Kb)
Chromosome 5:5,616,558 - 5,649,034
Position on chromosome