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

Ilya Kirov [1,2]*, Pavel Merkulov[1], Sofya Gvaramiya[1], Roman Komakhin[1], Murad Omarov[1], Maxim Dudnikov[1,2], Alina Kocheshkova[1,2], Alexander Soloviev[1], Gennady Karlov[1], Mikhail Divashuk[1,2]

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

[2]Kurchatov Genomics Center of ARRIAB, All-Russia Research Institute of Agricultural Biotechnology, Timiryazevskaya Street, 42, 127550 Moscow, Russia

Abstract

Transposable elements (TEs), which occupy significant portions of most plant genomes, 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. To expand the toolbox for TE identification in plants, we used the recently developed Cas9-targeted Nanopore sequencing (CANS) approach. Additionally, as no current bioinformatics tools automatically detect TE insertions after CANS, we developed NanoCasTE, a novel pipeline for target TE insertion discovery. We performed CANS of three copies of EVD retrotransposons in wild-type *Arabidopsis thaliana* and obtained up to 40× coverage of the targets after only a few hours of sequencing on a MinION sequencer. To estimate the ability to detect new TE insertions, we exploited the *A. thaliana ddm1* mutant, which has elevated TE activity. Using CANS, we detected 84% of these insertions in *ddm1* after generating only 4420 Nanopore reads (0.2× genome coverage), and also unambiguously identified their locations, demonstrating the method's sensitivity. CANS of pooled (~50 plants) *ddm1* plants captured >800 EVD insertions, especially in centromeric regions. CANS also identified insertions of a Ty3/Gypsy retrotransposon in the genomes of two *Aegilops tauschii* plants, a species with a large genome.

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 et al., 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 et al., 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, 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, 2020).

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., 2016; 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). 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., 2016). 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). 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 TEIs in plants with small (*A. thaliana*, 157 Mb/1C) and large (*Aegilops tauschii*, 4.3 Gb/1C) genomes. Using the *A. thaliana ddm1* mutant, which carries T-DNA and possesses elevated TE activity, we were able to detect most (84%, 21 of 25 TEIs) of the genetically inherited insertions of EVD retrotransposons in an individual plant with only 0.2× genome coverage obtained by CANS reads. This coverage is two orders of magnitude lower than what is needed for TEI identification based on whole-genome sequencing. In addition, >800 new TEIs could be detected using a pool of *ddm1* plants, resulting in a high-density EVD integration map. Finally, using *Ae. tauschii* as a model, we demonstrated that CANS is also well suited for TEI identification in plants with large, complex genomes. The application of CANS to two *Ae. tauschii* plants resulted in the capture of hundreds of insertions of a Ty3/Gypsy retrotransposon. Altogether, our results indicate that CANS and NanoCasTE are effective tools to capture TEIs in diverse plant species, 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).

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 gRNAs also recognizes the EVD1 copy. Because CANS is strand-specific, we designed gRNAs to direct Nanopore sequencing to sequence (1) EVD retrotransposons *per se* and (2) upstream and downstream sequences flanking retrotransposon insertions. The gRNAs were transcribed *in vitro* as single-guide RNAs (sgRNAs). We carried out two runs on a MinION sequencer using seven and five gRNA sets. The pilot experiment using all gRNAs 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 gRNA 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 gRNAs, designated gRNA1 to gRNA5, included one LTR gRNA (gRNA1, sgRNA mixture 1) and four gRNAs (sgRNA mixture 2) located in the inter-LTR region (Figure 1).

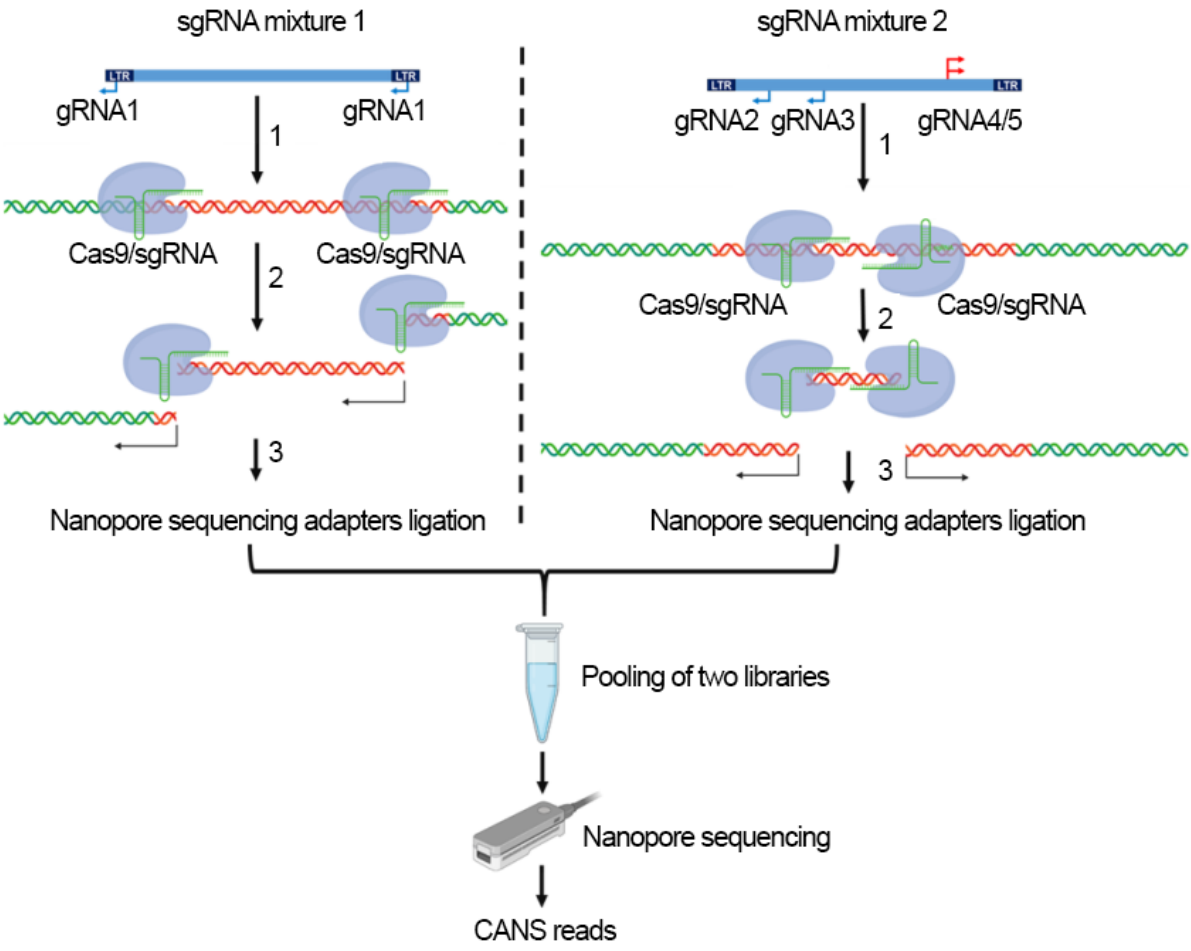


Figure 1. Schematic view of CANS of LTR retrotransposon insertions using five gRNAs divided into two mixtures. (1), (2) and (3) show the RNP binding, target DNA cleavage and ONP library preparation steps of the CANS protocol, respectively. Blue and red horizontal lines correspond to CANS reads mapped to negative and positive strands, respectively, compared to the reference genome.

The experiment using five gRNAs 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 gRNA 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, 2020).

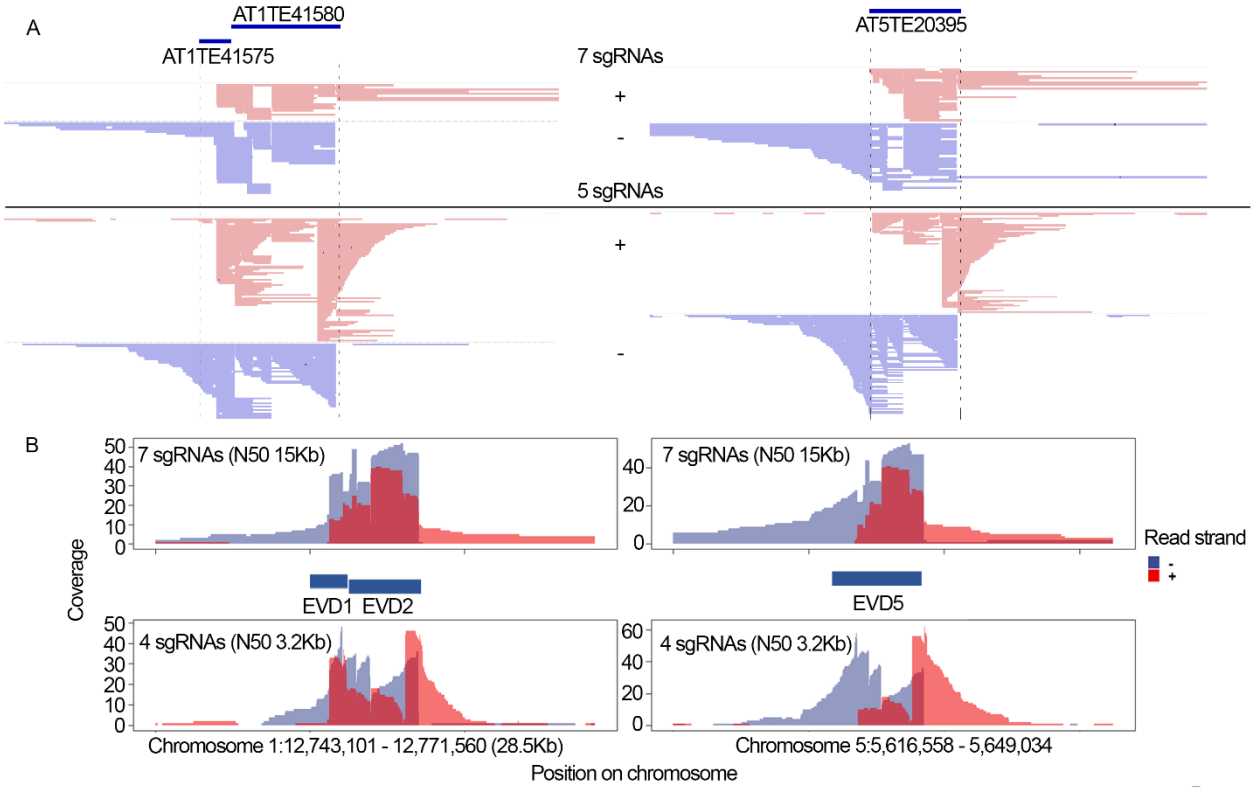


Figure 2. Targeted enrichment of EVD loci following CANS using a set of 7 or 5 gRNAs. (**A**) IGV view of Nanopore read alignment to EVD1 (AT1TE41575), EVD2 (AT1TE41580) and EVD5 (AT5TE20395) retrotransposon copies. (**B**) Coverage plot of EVD1, EVD2 and EVD5 using forward (red) and reverse (blue) reads 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 new EVD insertions in the genome using whole-genome Nanopore sequencing

To assess whether new EVD insertions could be detected using CANS, we exploited the *A. thaliana ddm1* mutant, whose EVD5 activity was previously described (Tsukahara et al., 2009). We created transgenic T_2 *ddm1* plants carrying T-DNA insertions of the EVD5 GAG open reading frame, a target for one of the five gRNAs used for CANS (G-ddm1 plants). This allowed us to simultaneously detect EVD insertions and T-DNA integration sites using a single sgRNA set and to obtain information about the sufficiency of CANS sequencing depth. To identify all EVD insertions in a G-ddm1 plant, we performed whole-genome Nanopore sequencing of one plant (plant G-ddm1-3). We collected 194,354 high-quality reads with N50 ~32Kb, corresponding to 27× coverage (the assembled genome size of *A. thaliana* is 119 Mb) (Figure 3A). We designed a pipeline called NanoWgsTE to identify EVD and T-DNA insertion sites using whole-genome Nanopore sequencing data. Using this pipeline, we identified 63 EVD and two T-DNA insertions. However, 37 TEIs (58%) were supported by only one read, which is significantly less than we expected based on Poisson distribution (*p-value* =1.85e-05 with λ= 13.5). These results may point to the occurrence of somatic EVD insertions that are shared by a subset of cells in an individual plant. Because this would affect the interpretation of the results, we performed Nanopore sequencing of two additional G-ddm1 plants (G-ddm1-1 and G-ddm1-2) that are siblings of G-ddm1-3 (progeny in the same family, Figure 3A). We obtained 9× and 10× coverage with N50 ~12Kb (Figure 3A). The NanoWgsTE pipeline revealed 51 and 62 EVD insertions in G-ddm1-1 and G-ddm1-2 plants, respectively. Of these, 34 (67%) and 45 (70%) TEIs were supported by one read, corroborating the results of G-ddm1-3 sequencing. Next, to identify truly genetically inherited TEIs, we estimated the number of EVD TEIs shared by at least two G-ddm1 plants and found 23 EVD insertions that were common to at least two sibling plants and 10 TEIs that were shared

by all three plants (Figure 3B). We then estimated the number of reads supporting these insertions (TEI reads) according to the NanoWgsTE report and found a significantly higher number of TEI reads for genetically inherited TEIs than for other TEIs (Supplementary Figure S1).

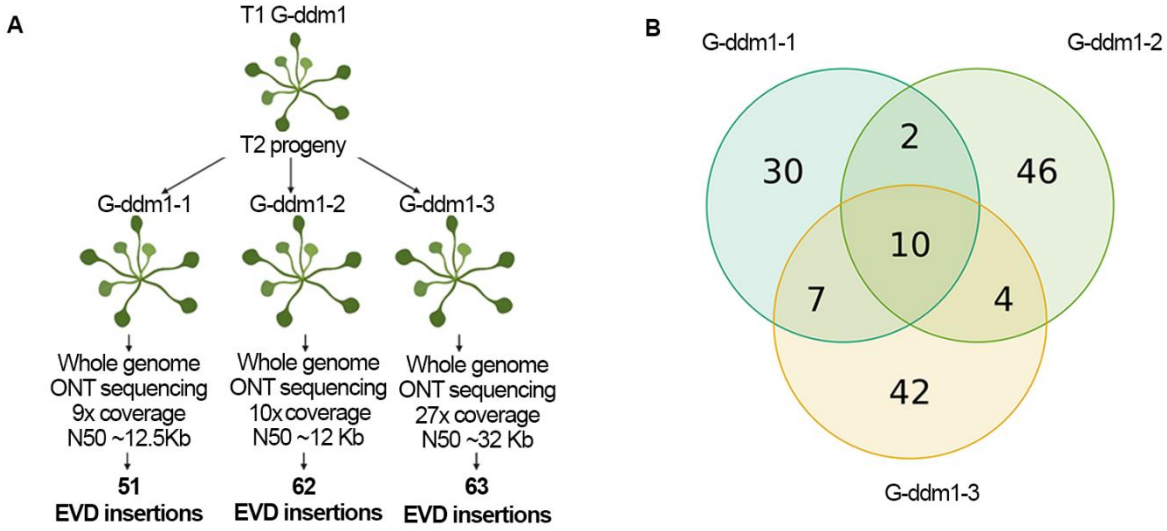


Figure 3. Detection of new EVD and T-DNA insertions in G-ddm1 plants. (**A**) Overview of whole-genome Nanopore sequencing of three T_2 sibling G-ddm1 plants and the numbers of EVD insertions detected using the NanoWgsTE pipeline. (**B**) Venn diagram showing the numbers of unique and common EVD insertions in three G-ddm1 plants identified by whole-genome Nanopore sequencing.

To further validate the suitability of CANS using these results, we selected 21 genetically inherited TEIs from a G-ddm1-3 plant. We also included four other TEIs of G-ddm1-3 that were supported by >2 TEI reads, i.e., the number of TEI reads found for >86% (18 of 21) of truly genetically inherited TEIs. Thus, using whole-genome Nanopore sequencing, we identified 25 TEIs in G-ddm1-3 plants that were genetically inherited (Supplemental Data Set 1).

CANS of *ddm1* and the NanoCasTE pipeline

We carried out CANS using five gRNAs 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 4A). The EVD sequence had up to 400× coverage by 1170 reads (25% of all CANS reads) (Supplementary Figure S2). 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 support the notion that the coverage we obtained is sufficient for EVD TEI identification, we analyzed EVD-genome junctions and found that they had ~8× coverage, which is sufficient for insertion detection (Figure 4A). As an additional control, we also checked the CANS reads covering a hemizygous T-DNA insertion site carrying a single gRNA position (gRNA2, Figure 4D) and found that even using a single gRNA, CANS generated ~4× coverage of the T-DNA integration site. These observations indicate that the number of CANS reads obtained is sufficient for TEI identification.

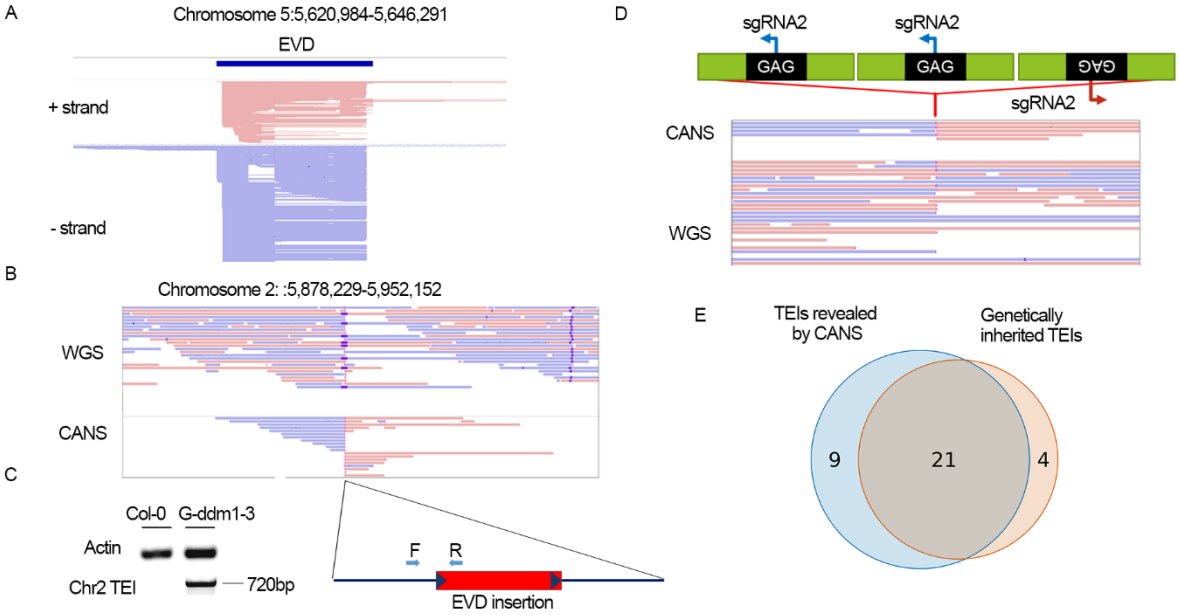


Figure 4. EVD coverage and insertion detection in a G-ddm1-3 plant using CANS. (**A**) IGV snapshot of coverage of the original copy of EVD on Chromosome 5 using forward (red) and reverse (blue) ONT reads obtained after CANS of G-ddm1-3 plant. (**B**) IGV browser snapshot of one of the 25 genetically inherited EVD insertions detected by whole-genome Nanopore sequencing and CANS. (**C**) PCR validation of the EVD insertion on Chromosome 2 (left) using primers based on the EVD and TEI flanking region (right). (**D**) Inserted T-DNA structure (upper panel), gRNA2 positions and its coverage based on CANS and whole-genome Nanopore sequencing reads. (**E**) Venn diagram of the numbers of genetically inherited and CANS-detected EVD insertions in a G-ddm1-3 plant.

To automate the detection of TE and T-DNA insertions, we developed a novel pipeline called NanoCasTE (see Methods, Figure 7). By applying NanoCasTE to the CANS dataset, we detected 30 EVD insertions. Of these, 22 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 (see above) is shown in Figure 4B. This TEI was also validated by PCR with primers designed based on EVD and insertion sites (Figure 4C). Next, we estimated how many genetically inherited TEIs were identified by CANS and found that 21 of 25 (84%) genetically inherited TEIs were captured by CANS (Figure 4E). These results suggest that even the low genome coverage (0.2× in our experiment) of CANS may be sufficient to capture most TEIs in *A. thaliana*, highlighting the good sensitivity of the method. Additionally, the newly developed NanoCasTE pipeline enables the rapid identification of novel TEIs.

Capture of EVD insertions in a *ddm1* population

Because EVD can generate new TE copies in individual *ddm1* plants, we asked whether we could detect new insertions in a population of approximately 50 *ddm1* plants. To address this issue, we isolated high molecular weight DNA from the pooled plant material and carried out CANS. After 2.5 hours of MinION sequencing, we generated 18,246 Nanopore reads with N50 ~5.5 Kb. Using the NanoCasTE pipeline, we detected 851 TEIs, including 29 high-confidence TEIs (supported by two or more reads), in the *ddm1* population. Thirteen TEIs were PCR validated using DNA from individual plants, resulting in 1 to 6 plants carrying a single TEI. Analysis of the chromosome-wide distribution of the TEIs showed that they tended to be clustered in the pericentromeric regions of all chromosomes, while the density of TEIs in chromosome arms was significantly reduced (Figure 5A).

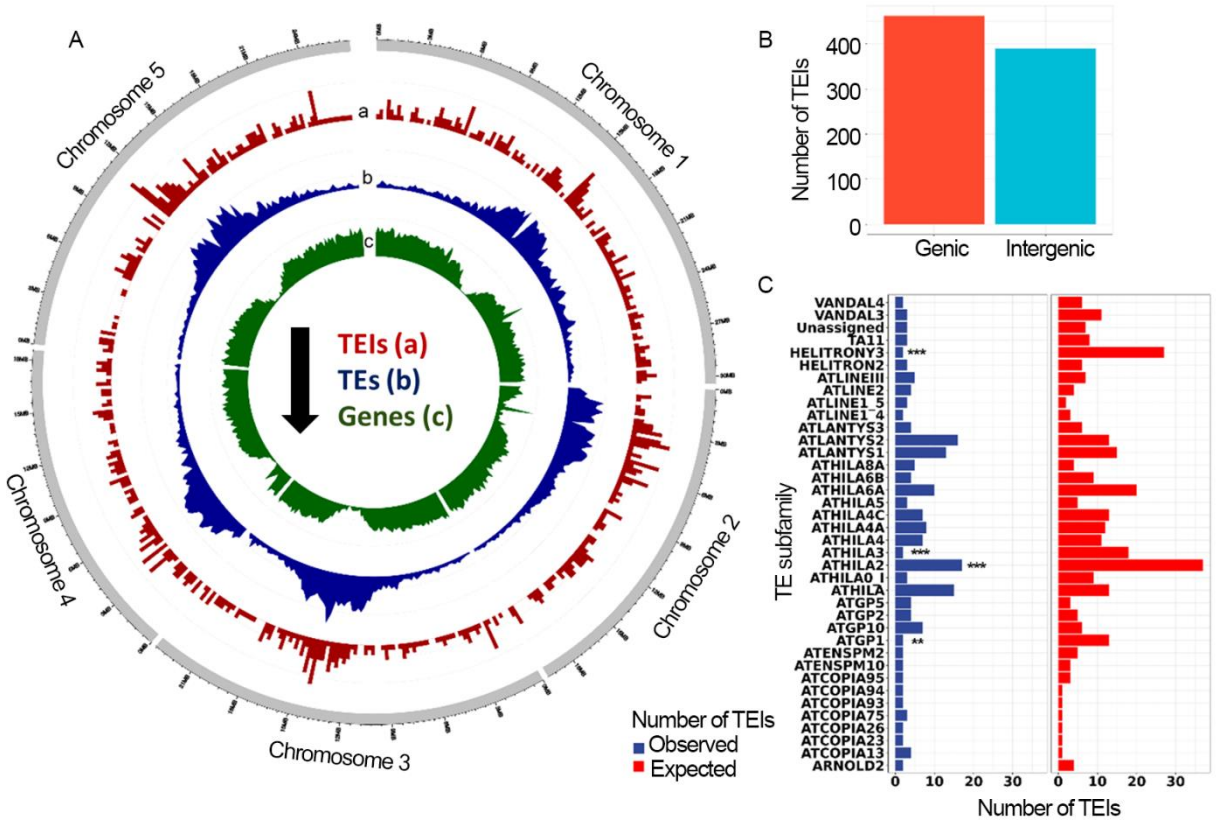


Figure 5. Genome organization of EVD insertions detected by CANS in a pooled (~50 plant) *ddm1* sample. (**A**) Chromosome distribution of TEIs detected by CANS (a), all *A. thaliana* TEs (Panda & Slotkin 2020) (b) and genes (c) along the *A. thaliana* chromosomes. (**B**) Bar plot showing the number of genic and intergenic TEIs detected by CANS. (**C**) Bar plot showing the number of EVD insertions in members of different TE subfamilies of *A. thaliana*.

Classification of these TEIs revealed that 54% (462) of them are located in genic regions, including 403 (87%) exonic and 59 (13%) intronic TEIs (Figure 5B). This value is expected by chance based on the total length of the genic regions in the assembled (119 Mb, TAIR10) *A. thaliana* genome (68.3 M, or 57% of the sequenced genome). Gene Ontology (GO) analysis revealed no enrichment of any GO categories in the set of genes with TEIs, suggesting that most EVD insertions in *ddm1* are randomly distributed among genes with different functions.

Among intergenic TEIs (389, 46%), 241 (28%) were located in the sequences of other TEs. This value is significantly higher (chi-squared test *p*-value = 0.0004) than expected by chance based on the total length of the annotated TEs in the *A. thaliana* genome (24.9 Mb, or 21% (Panda and Slotkin, 2020)). Moreover, 20 TEs had 2 or 3 different TEIs. We analyzed the number of TEIs in TEs of different subfamilies and found that members of four subfamilies were significantly underrepresented (Fisher's exact test *p*-value <0.01) in a set of TEs with detected TEIs, including HELITRONY3 (*p*-value = 6.78e-07), Athila2 (*p*-value = 0.004), Athila3 (*p*-value = 0.0002) and ATGP1 (*p*-value=0.004) (Figure 5C). This finding can be partly attributed to the differences in chromosome distribution between TEs of these families and EVD TEIs. For example, most (1104 of 1399, 79%) of the members of HELITRONY3 are located in chromosome arms, where the EVD TEI frequency is much lower than in the pericentromeric regions.

Thus, using CANS coupled with the NanoCasTE pipeline, we built a high-density physical map of EVD5 insertion sites in the *A. thaliana* genome and showed that EVD insertions are biased toward pericentromeric regions and that EVD tends to mobilize frequently into other TEs.

Targeting of LTR retrotransposons in *Ae. tauschii*

To examine whether NanoCasTE could be utilized to trace insertions of multicopy TEs in a plant with a large genome, we used *Ae. tauschii* as a model. *Ae. tauschii*, a donor of the D-genome of bread wheat (*Triticum aestivum*), has a relatively large genome (~4.3 Gb (Jia et al., 2013)). To choose a target for NanoCasTE sequencing, we aimed to identify multicopy TEs with recent copies in the genome. We selected a Ty3/Gypsy LTR retrotransposon (hereafter called Aty3_169) of the Retand lineage, with 327 BLAST hits per genome and a number of recent insertions that occurred <0.5 million years ago. This TE is >13 Kb in length and is specifically found in the bread wheat D-genome (Supplementary Figure S3), suggesting that its transposition activity occurred after the speciation of the A, B and D genome progenitors of bread wheat. We synthesized five sgRNAs including 2 gRNAs (gRNA1, gRNA2) located on LTR sequences in the both ends of the Aty_169 and three gRNAs located in the inter-LTR region (gRNA3-5). We carried out CANS using two sgRNA mixtures: sgRNA Mixture 1 (gRNA1, gRNA2) and sgRNA Mixture 2 (gRNA3-5). Genomic DNA from two *Ae. tauschii* plants (At1 and At2) was sequenced using two MinION flow cells (one flow cell per plant). In each experiment, we obtained ~70,000 reads. First, we evaluated the coverage of the reference copies of Aty3_169 based on reads. For this, only high-quality mapped reads (MQ > 40, no supplementary alignments allowed, <30 bp end clipping) were used. This analysis resulted in 228 (69%) and 247 (75%) reference copies covered by two or more ONT reads, respectively. Of these, 201 reference copies were common between the two plants (Figure 6A).

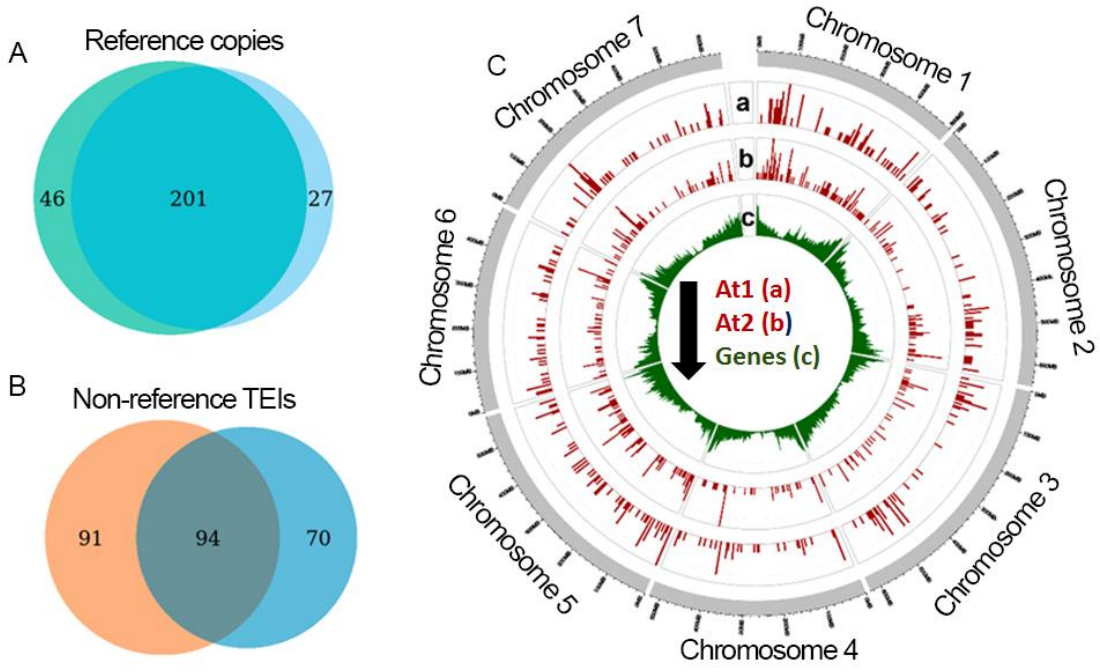


Figure 6. Insertion identification for a Ty3/Gypsy LTR retrotransposon (Aty3_169) of the Retand lineage in the genomes of two *Ae. tauschii* plants. (**A**) Numbers of reference insertions covered by two or more ONT reads. (**B**) Numbers of non-reference insertions covered by two or more Nanopore reads. (**C**) Chromosome distribution of At1 and At2 TEIs and annotated genes along the *Ae. tauschii* chromosomes.

To identify non-reference insertions, we applied the NanoCasTE pipeline to the At1 and At2 reads. We identified 185 and 164 non-reference TEIs in the At1 and At2 genomes, respectively. Of these, 94 insertions were common to both datasets (Figure 6B). We designed primers based on 10 TEIs, and 7 of them were validated by PCR.

To estimate any biases in the transposition of Aty3_169 in the *Ae. tauschii* genome, we evaluated the chromosome distribution of the TEIs. In general, TEIs were randomly distributed along the chromosomes,

although a few regions of chromosomes 1, 3, 4 and 7 were highly enriched by Aty3_169 insertions (Figure 6C). In the future, it would be interesting to assess whether these hot spots of Aty3_169 insertions could be attributed to the recent introgressions.

Discussion

Here, we demonstrated that CANS is a valuable tool for identifying TE insertions (TEIs) in both repetitive (e.g. TE rich) and genic regions of plant genomes. CANS-based detection of TEIs has several advantages over the existing method of TEI enrichment. First, the CANS protocol, including RNP complex formation, DNA cleavage and sequencing library preparation, is rapid and can be completed in 4-5 hours. Second, because a PCR amplification step is not needed, the resulting data can be used for epigenetic modification calling, although higher coverage is required for this process (Gabrieli et al., 2018; Ni et al., 2021). Third, long reads have high mappability to the genome, paving the way for TE insertion detection, even in repeat-rich regions of the genome such as centromeres and heterochromatin. Fourth, utilizing NanoCasTE allows TEIs to be identified easily using CANS data, even if hundreds of insertions are captured. Finally, CANS can be applied to small as well as large plant genomes. Despite the many advantages of CANS for TEI detection in plants, there are also several constraints to the method that need to be overcome in the future: (1) a large amount of high-molecular-weight genomic DNA (>5 µg) is required; (2) deciphering the TE sequence of the insertion is challenging and requires specific sets of overlapping sgRNAs; (3) only a small subset of sequencing pores (5-10%) are sequencing at the same time, and therefore the sequencing efficiency is lower than for whole-genome sequencing (90-98% of sequencing pores), and (4) the genotype of TEIs (homozygote or heterozygote) cannot be determined using CANS.

While whole-genome Nanopore sequencing provides excellent resolution to potentially detect all TEIs in a genome, it is expensive when the genome is relatively large. For example, in our hands and based on previous reports using different species, it is challenging to generate more than 8 Gb of data per single MinION flow cell from plant material (Lee et al., 2019; Schalamun et al., 2019; Dmitriev et al., 2021). Obtaining only 4400 CANS reads (equivalent to 0.2× coverage of the *A. thaliana* genome) on a single flow cell, we were able to detect 86% of the genetically inherited EVD insertions. Additionally, we also unraveled the positions of T-DNA containing the GAG region of the EVD retrotransposon and carrying a target site for one of the five sgRNAs. Thus, even a single gRNA can be effective for insertion detection. However, a higher number of sgRNAs and cleavage sites in genomic DNA may produce a higher proportion of target reads, as we also observed in our pilot experiment with seven sgRNAs. The use of a higher number of sgRNAs per target was also proposed to be beneficial for CANS of target genes (Gilpatrick et al., 2020). This is an important property of the CANS approach that allows more targets to be included per experiment without losing the efficiency of CANS. Based on this assumption, insertions from many TEs can be captured in one experiment using specifically designed and pooled sgRNAs. This approach was recently applied to detect TE insertions in the human genome (McDonald et al., 2021).

Here, we identified >800 TEIs in a small *ddm1* population after 2 hours of sequencing using only a part of the capacity of MinION. This experiment provided a proof of concept that CANS is a valuable tool for identifying TEIs in a population. A study on the distribution of TE insertions in the *A. thaliana* 1001 genomes population demonstrated that TEs are a major source of large-effect alleles that can be fixed in the populations, contributing to the local adaptation of different accessions (Baduel et al., 2021). Indeed, stress may induce TE activation and consequently the origination of new alleles. TEI-derived alleles have spontaneously occurred in fields of crop species and have long been exploited in plant breeding programs (Kobayashi et al., 2004; Jiang et al., 2009; Butelli et al., 2012; Domínguez et al., 2020). Although the discovery of such alleles in crop fields using whole-genome sequencing is very challenging, CANS might represent a valuable alternative to capture TEI-derived alleles in a population of cultivated plants. However, prior knowledge of active TEs capable of transposing is required for such experiments. Fortunately, this information is available for some plants. For example, we recently identified active LTR

retrotransposons in sunflower (Kirov et al., 2020a) and triticale (Kirov et al., 2020b). There are many other examples of active plant TEs that can serve as potential targets for CANS-based capture of TEIs. The development of high-throughput tools that can capture TE insertions in a population represents a milestone towards the acceleration of the plant breeding process and a more effective usage of natural variation. Further increasing the coverage of sequencing using tools such as the PromethION sequencer could greatly increase the sensitivity of this approach and might even allow the detection of TEIs that have occurred in a single plant within an entire population. Such results could then be used to design specific primers based on the insertion site in order to identify all plants carrying the target TEIs. The capture of TEIs occurring in individual plants in a population will also depend on the genome size of the species and the number of sgRNAs and target loci in the genome. These parameters must be optimized for individual species.

Estimating the integration preferences for distinct TEs is not a straightforward process, as post-integration selection pressure has a strong effect on the current TEI landscape. Therefore, artificially inducing TE activity (e.g. via stress) and exploiting mutant lines followed by TEI identification are the most direct ways to study TE integration biases and targets. CANS can be a suitable tool for identifying TEIs in such experiments. Here, we showed that EVD has preferences for integrations in the pericentromeric regions of *A. thaliana* chromosomes. It is important to note that the established EVD insertion distribution was shaped by very recent events and that post-integration selection pressure had only a minor effect on this process. While wild-type *A. thaliana* contains only a few EVD insertions, its close relative *Arabidopsis lyrata* contains multiple EVD insertions, most of them in centromeric satellite regions (Tsukahara et al., 2009). Thus, the centromere-biased distribution of EVD insertions is similar under selection pressure (*A. lyrata*) and without post-integration selection (*A. thaliana ddm1* mutant). This can be explained by the low gene density in the centromeric regions, which decreases the rate of gene insertions and relaxes the selection pressure on EVD TEIs.

Thus our pioneering work thus demonstrates that CANS is a valuable, cost-effective, rapid tool for deciphering the mobilome landscapes of plant genomes. CANS can be utilized for TEI identification in plants with small as well as large genomes and that TEs with few or multiple insertions can be detected using this method and the NanoCasTE pipeline.

Methods

Plant material and growth conditions

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

Two *Aegilops tauschii* (K-1216 and K-112) accessions were kindly provided by Dr. E.D. Badaeva (Laboratory of Genetic identification of plants, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow). The plants were grown in a greenhouse under natural conditions.

Plant transformation

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

HMW DNA isolation and size selection

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

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

gRNA design

To design gRNAs recognizing multiple copies of LTR retrotransposons we used a step-by-step strategy: (1) using FlashFry ‘discover’ tool with the ‘-positionOutput -maxMismatch 0 ‘ arguments and reference genome sequences (At 4.0 for *Ae. tauschii* and TAIR10 for *A.thaliana*) a pool of gRNAs and the corresponding genomic sites were found followed by bed file generation; (2) reference sequences of target LTR retrotransposons were blasted against the corresponding genomes followed by hit filtering (minimum hit length 100bp and minimum e-value 1e-50) and bed file consisting blast hit positions was generated; (3) the genomic location of predicted gRNA sites (step 1) were intersected with blast hit bed file (step 2) using bedtools 'intersect’ and OnOutTE ratio was calculated for each gRNAs:

OnOutTE = number of gRNA sites located on the blast hit from the corresponding TE / total number of gRNA sites

(4) We selected gRNAs with OnOutTE ratio > 0.85.

For each retrotransposon a set of gRNAs were designed including one (*A. thaliana*) or two (*Ae. tauschii*) gRNAs negative strand recognizing both LTRs and three gRNAs (one on negative and two on positive strands) located between two LTR sequences. Sequences of gRNAs are provided in Supplemental Data Set 2.

In vitro transcription of sgRNAs

SgRNAs for CANS were produced by in vitro transcription from double-stranded DNA templates containing T7 promoter that were assembled from two oligos, specific (gRNA1-5 in Supplemental Data Set 2) and universal (CRISPR_R). All oligonucleotides were ordered in Evrogen (Moscow, Russia). To synthesize double-stranded DNA templates, the following reagents were mixed per each reaction: 2µl of unique sgRNA oligo (1 µM); 2µl of CRISPR_R (1 µM); 2µl of T7 forward (5'-GGATCCTAATACGACTCACTATAG-3') and reverse (5'-AAAAAAGCACCGACTCGG-3') primers (100 µM of each); 2µl of 50x dNTP (10 µM) (Evrogen, Moscow, Russia); 10µl of 10x Encyclo buffer (Evrogen, Moscow, Russia); 1µl of Encyclo polymerase (Evrogen, Moscow, Russia); 79µl of RNAse-free water. Oligos were annealed and extended according to the PCR program: initial denaturation at 95°C (2 min); 30 cycles of denaturation at 98°C (30 sec), annealing at 60°C (30 sec), elongation at 72°C (30 sec); final elongation at 72°C (1 min). Gel electrophoresis was performed after amplification and templates were column purified.

To transcribe the templates, the Highly Efficient RNA in vitro Synthesis Kit (Biolabmix, Novosibirsk, Russia) was used. The subsequent reactions were set up and incubated at 37°C for 2h: 10µl of 5x T7 transcription buffer; 2µl of 25x DTT (250mM); 2µl of dNTP (25mM of each); 2µl of double-stranded DNA template (150-200 ng); 1µl of T7 polymerase (150 units); 33µl RNAse-free water. The sgRNAs were then purified using RNA and miRNA Extraction Kit (Biolabmix, Novosibirsk, Russia) according to the manufacturer instructions. The concentration and quality of prepared sgRNAs were estimated by Nanodrop (Thermo Scientific, USA), Qubit (Thermo Scientific, USA) and gel electrophoresis in 2% agarose gel.

Cas9/sgRNA ribonucleoprotein complexes (RNPs) assembly

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

Cas9 cleavage and library preparation

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

Library preparation for whole-genome sequencing of G-ddm1-3 was carried out using the Ligation Sequencing Kit (Oxford Nanopore Technologies, catalog no. SQK-LSK109) with 0.5 - 1 µg of input HMW DNA. The library for G-ddm1-1 and G-ddm1-2 WGS was prepared from 1 µg of both DNA using the Native Barcoding Expansion 1-12 (Oxford Nanopore Technologies, catalog no. EXP-NBD104) and the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies).

Sequencing and basecalling

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

Analysis

Base calling was done by Guppy (Version 3.2.10). The Aet_v4.0 genome sequence was downloaded from the Gramene database (ftp://ftp.gramene.org/pub/gramene/release-63/fasta/aegilops_tauschii). The TAIR10 genome was downloaded from NCBI database. The generated Nanopore reads were aligned to the reference genomes (TAIR10 for *A.thaliana* and Aet_v4.0 for *Ae.tauschii*) using minimap2 software (Li, 2018) with the following parameters: -ax map-ont -t 100. The obtained sam file was converted to bam format, sorted and indexed using SAMtools (Li et al., 2009). The obtained sorted bam files were used for NanoCasTE (CANS data) and NanoWgsTE (WGS data) pipelines.

NanoCasTE pipeline for the identification of transposon insertions following CANS

NanoCasTE is written in python and can be run as a standalone tool or via Snakefile (https://github.com/Kirovez/NanoCasTE). NanoCasTE involves several steps (Figure 7): (1) Mapping the reads to the genome using minimap2 (Li, 2018), followed by the generation of sorted bam files using SAMtools (Li et al., 2009); (2) Filtering the reads based on mapping quality (MQ > 40), read length (>750) and SA tags (no supplementary alignments are allowed); (3) Selecting mapped reads with clipped heads (reads mapped to the positive strand) or tails (reads mapped to the negative strand) with the length of the clipped parts close to the distance from the gRNA positions to the TE end; (4) Filtering the reads based on similarity of the clipped part to the target TE sequence; (5) Identifying the inserted sites and TE orientation (+ or - strand) and outputting the results.

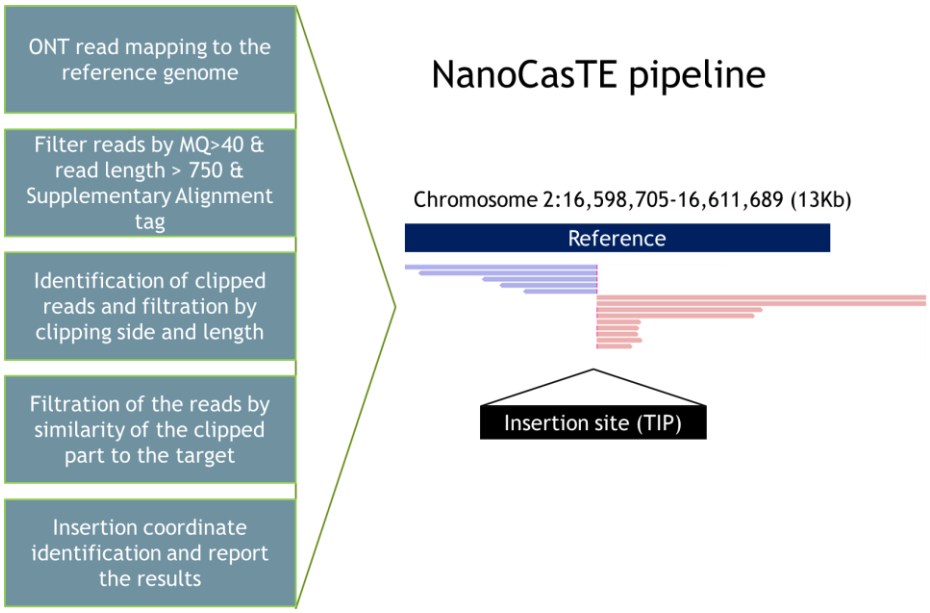


Figure 7. Schematic view of the major steps of the NanoCasTE pipeline to identify transposon insertions in the genome after CANS

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: 0.6, mapping_quality: 40, min_read_length: 750. The following lists of gRNAs were used: (1) for *A. thaliana*, 'TCTTGGTGATGAGAGTGAC, ACCCTGGATTTAAGGTGAGA, AGTTTAAGAGCTCTAGTATG, CTACAAGGTCAATCGAAAGG, TCAACACATGAAAGTCCCGA'; (2) for *Ae. tauschii*, 'CCGGGTCGTCCCTTTCTATA, GTCCCTTTCTATAGGGAGGT, CTGAGCCGTTTCGATGAGAC, TCCGGAGAATGACGTCACTC, TAAGCCGGAGATTTTTCTGT'. We used stricter filtration criteria *for Ae. tauschii* insertion selection and kept only the insertions localized on chromosomes and possessing two or more supporting reads.

NanoWgsTE pipeline

To detect TE or T-DNA insertions in the genome using Nanopore whole-genome sequencing data we designed a pipeline called NanoWgsTE (https://github.com/Kirovez/NanoWgsTE). NanoWgsTE involves several steps: (1) it converts fastq to fasta and performs BLAST search of the reads with the similarity to target sequence followed by selection of the reads into a separate file; (2) the selected reads are mapped to the genome by minimap2 (Li, 2018) followed by generation sorted bam file using SAMtools (Li et al., 2009) and BAMtools (Barnett et al., 2011); (3) the reads with clipped ends are searched in the obtained bam file assisted by pysam package (https://github.com/pysam-developers/pysam); (4) the genomic positions where clipped ends of the reads were mapped are collected and output files (bed file with putative insertion positions) and file with clipped read count data per putative insertion) are generated. We run this pipeline with default parameters for EVD5 (AT5TE20395) and T-DNA (pBI121 sequence, NCBI accession number: AF485783.1): -q 40 –mlc 500 –mbh 500. For detection of EVD5 insertions the fasta sequence was used. All putative T-DNA and EVD5 insertions were manually checked by a local instance of JBrowse (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 S2. 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 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 and final extension at 72 ◦C for 5 min. For validation of *A.thaliana* TEIs the following PCR program was used: 35 cycles of 94 ◦C for 30 s, 60 ◦C for 30 s and 72 ◦C for 2 min. Amplification products were visualized by electrophoresis in 1.5% agarose gel and observed under UV light after ethidium bromide staining. Primers used for TEI validation are listed in Supplemental Data Set 4.

Aegilops tauschii LTR retrotransposon identification and selection for CANS

Genomic sequence of *A. tauschii* (Aet_v4.0) was downloaded from the Gramene database (ftp://ftp.gramene.org/pub/gramene/release-63/fasta/aegilops_tauschii). LTR retrotransposons of chromosome 1 of *A. taushii* were predicted in the genome as previously described (Penin et al. 2021) using LTRharvest (Ellinghaus et al., 2008) and LTRdigest (Steinbiss et al., 2009). The following arguments were used for LTRdigest: -aaout yes -pptlen 10, 30 -pbsoffset 0, 3 -pdomevalcutoff 0.001. Hmm profiles of transposon domains were downloaded from the GyDB database (Llorens et al., 2010). TEsorter software (Zhang et al., 2019) was used for LTR retrotransposon classification. The LTR sequences of all predicted TEs were clustered using cd-hit with the following arguments: -T 100 -l 150 -d 100 -s 0.8 -aL 0.8 -aS 0.8 -A 80 -sc -sf -0.95 . Then we selected a single cluster containing complete TE (according to the TEsorter results) with identical LTRs pointing to its recent insertion time. This Ty3/Gypsy LTR retrotransposon (Aty3_169) of the Retand family is located on chromosome 1D:169126894..169140327

of the reference Aet_v4.0 genome. Blast search of the similarity of Aty3_169 LTR sequence to all Aet_v4 genomic sequences was carried out on the Ensembl website and resulted in >300 hits of Aty3_169 distributed along all chromosomes.

Statistics and Data Visualization

Statistical analysis was carried out in Rstudio Version 1.2.1335 (http://www.rstudio.com/) with R version 3.6.0. Visualization was carried out by ggplot2 (Wickham, 2011) and VennDiagram (Chen and Boutros, 2011) packages.

Accession Numbers

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

Acknowledgements

We thank Ekaterina Badaeva (Laboratory of Genetic identification of plants, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow) and Vincent Colot (Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France) for providing the plant material.

Author contribution statement

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

References

Baduel, P., Leduque, B., Ignace, A., Gy, I., Gil, J., Loudet, O., Colot, V., and Quadrana, L. (2021). Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. Genome Biology **22,** 138.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics **27,** 1691-1692.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., and Holmes, I.H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biology **17,** 66.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell **24,** 1242-1255.

Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., Debladis, E., Akakpo, R., Hsing, Y.-I., and Panaud, O. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nature Communications **10,** 24.

Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics **12,** 35.

Chen, J., Lu, L., Robb, S.M.C., Collin, M., Okumoto, Y., Stajich, J.E., and Wessler, S.R. (2020). Genomic diversity generated by a transposable element burst in a rice recombinant inbred population. Proceedings of the National Academy of Sciences of the United States of America **117,** 26288-26297.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics **18,** 71-86.

Clough, S.J., and Bent, A.F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J **16,** 735-743.

Dmitriev, A.A., Pushkova, E.N., Novakovskiy, R.O., Beniaminov, A.D., Rozhmina, T.A., Zhuchenko, A.A., Bolsheva, N.L., Muravenko, O.V., Povkhova, L.V., Dvorianinova, E.M., Kezimana, P., Snezhkina, A.V., Kudryavtseva, A.V., Krasnov, G.S., and Melnikova, N.V. (2021). Genome Sequencing of Fiber Flax Cultivar Atlant Using Oxford Nanopore and Illumina Platforms. Front Genet **11,** 590282-590282.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M., Colot, V., and Quadrana, L. (2020). The impact of transposable elements on tomato diversity. Nature Communications **11,** 4058.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9,** 18.

Gabrieli, T., Sharim, H., Fridman, D., Arbib, N., Michaeli, Y., and Ebenstein, Y. (2018). Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). Nucleic Acids Research **46,** e87-e87.

Gilpatrick, T., Lee, I., Graham, J.E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F.J., and Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. Nature Biotechnology **38,** 433-438.

Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet **43,** 269-276.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K.F.X., Li, D., Pan, S., Zheng, F., Hu, Q., Xia, X., Li, J., Liang, Q., Chen, J., Wicker, T., Gou, C., Kuang, H., He, G., Luo, Y., Keller, B., Xia, Q., Lu, P., Wang, J., Zou, H., Zhang, R., Xu, J., Gao, J., Middleton, C., Quan, Z., Liu, G., Wang, J., Yang, H., Liu, X., He, Z., Mao, L., Wang, J., and International Wheat Genome Sequencing, C. (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature **496,** 91-95.

Jiang, N., Gao, D., Xiao, H., and van der Knaap, E. (2009). Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. Plant J **60,** 181-193.

Jung, H., Winefield, C., Bombarely, A., Prentis, P., and Waterhouse, P. (2019). Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. Trends in Plant Science **24,** 700-724.

Kirov, I., Omarov, M., Merkulov, P., Dudnikov, M., Gvaramiya, S., Kolganova, E., Komakhin, R., Karlov, G., and Soloviev, A. (2020a). Genomic and Transcriptomic Survey Provides New Insight into the Organization and Transposition Activity of Highly Expressed LTR Retrotransposons of Sunflower (Helianthus annuus L.). Int J Mol Sci **21,** 9331.

Kirov, I., Dudnikov, M., Merkulov, P., Shingaliev, A., Omarov, M., Kolganova, E., Sigaeva, A., Karlov, G., and Soloviev, A. (2020b). Nanopore RNA Sequencing Revealed Long Non-Coding and LTR Retrotransposon-Related RNAs Expressed at Early Stages of Triticale SEED Development. Plants **9,** 1794.

Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. Science **304,** 982.

Lee, Y.G., Choi, S.C., Kang, Y., Kim, K.M., Kang, C.-S., and Kim, C. (2019). Constructing a Reference Genome in a Single Lab: The Possibility to Use Oxford Nanopore Technology. Plants **8,** 270.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34,** 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25,** 2078-2079.

Li, S., Jia, S., Hou, L., Nguyen, H., Sato, S., Holding, D., Cahoon, E., Zhang, C., Clemente, T., and Yu, B. (2019). Mapping of transgenic alleles in soybean using a nanopore-based sequencing strategy. Journal of Experimental Botany **70,** 3825-3833.

Lisch, D.J.N.R.G. (2013). How important are transposons for plant evolution? Nature Reviews Genetics **14,** 49-61.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M.,

Latorre, A., and Moya, A. (2010). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research **39,** D70-D74.

López-Girona, E., Davy, M.W., Albert, N.W., Hilario, E., Smart, M.E.M., Kirk, C., Thomson, S.J., and Chagné, D. (2020). CRISPR-Cas9 enrichment and long read sequencing for fine mapping in plants. Plant Methods **16,** 121.

Madsen, E.B., Höijer, I., Kvist, T., Ameur, A., and Mikkelsen, M.J. (2020). Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. Hum Mutat **41,** 1671-1679.

McDonald, T.L., Zhou, W., Castro, C., Mumm, C., Switzenberg, J.A., Mills, R.E., and Boyle, A.P. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing**,** 2021.2002.2010.430605.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in Arabidopsis. Nature **461,** 427-430.

Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., Bai, L., Liu, W., Xiao, C.-L., Luo, F., and Wang, J. (2021). Genome-wide Detection of Cytosine Methylations in Plant from Nanopore sequencing data using Deep Learning**,** 2021.2002.2007.430077.

Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N., and Slotkin, R.K. (2013). The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs **162,** 116-131.

Panda, K., and Slotkin, R.K. (2020). Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements **32,** 2687-2698.

Quadrana, L., Bortolini Silveira, A., Mayhew, G.F., LeBlanc, C., Martienssen, R.A., Jeddeloh, J.A., and Colot, V. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. eLife **5,** e15716.

Rabanus-Wallace, M.T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., Baez, M., Houben, A., Mayer, K.F.X., Guo, L., Poland, J., Pozniak, C.J., Walkowiak, S., Melonek, J., Praz, C.R., Schreiber, M., Budak, H., Heuberger, M., Steuernagel, B., Wulff, B., Börner, A., Byrns, B., Čížková, J., Fowler, D.B., Fritz, A., Himmelbach, A., Kaithakottil, G., Keilwagen, J., Keller, B., Konkin, D., Larsen, J., Li, Q., Myśków, B., Padmarasu, S., Rawat, N., Sesiz, U., Biyiklioglu-Kaya, S., Sharpe, A., Šimková, H., Small, I., Swarbreck, D., Toegelová, H., Tsvetkova, N., Voylokov, A.V., Vrána, J., Bauer, E., Bolibok-Bragoszewska, H., Doležel, J., Hall, A., Jia, J., Korzun, V., Laroche, A., Ma, X.-F., Ordon, F., Özkan, H., Rakoczy-Trojanowska, M., Scholz, U., Schulman, A.H., Siekmann, D., Stojałowski, S., Tiwari, V.K., Spannagl, M., and Stein, N. (2021). Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. Nature Genetics **53,** 564-573.

Schalamun, M., Nagar, R., Kainer, D., Beavan, E., Eccles, D., Rathjen, J.P., Lanfear, R., and Schwessinger, B. (2019). Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. Mol Ecol Resour **19,** 77-89.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics **8,** 272-285.

Stangl, C., de Blank, S., Renkens, I., Westera, L., Verbeek, T., Valle-Inclan, J.E., González, R.C., Henssen, A.G., van Roosmalen, M.J., Stam, R.W., Voest, E.E., Kloosterman, W.P., van Haaften, G., and Monroe, G.R. (2020). Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. Nature Communications **11,** 2861.

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research **37,** 7002-7013.

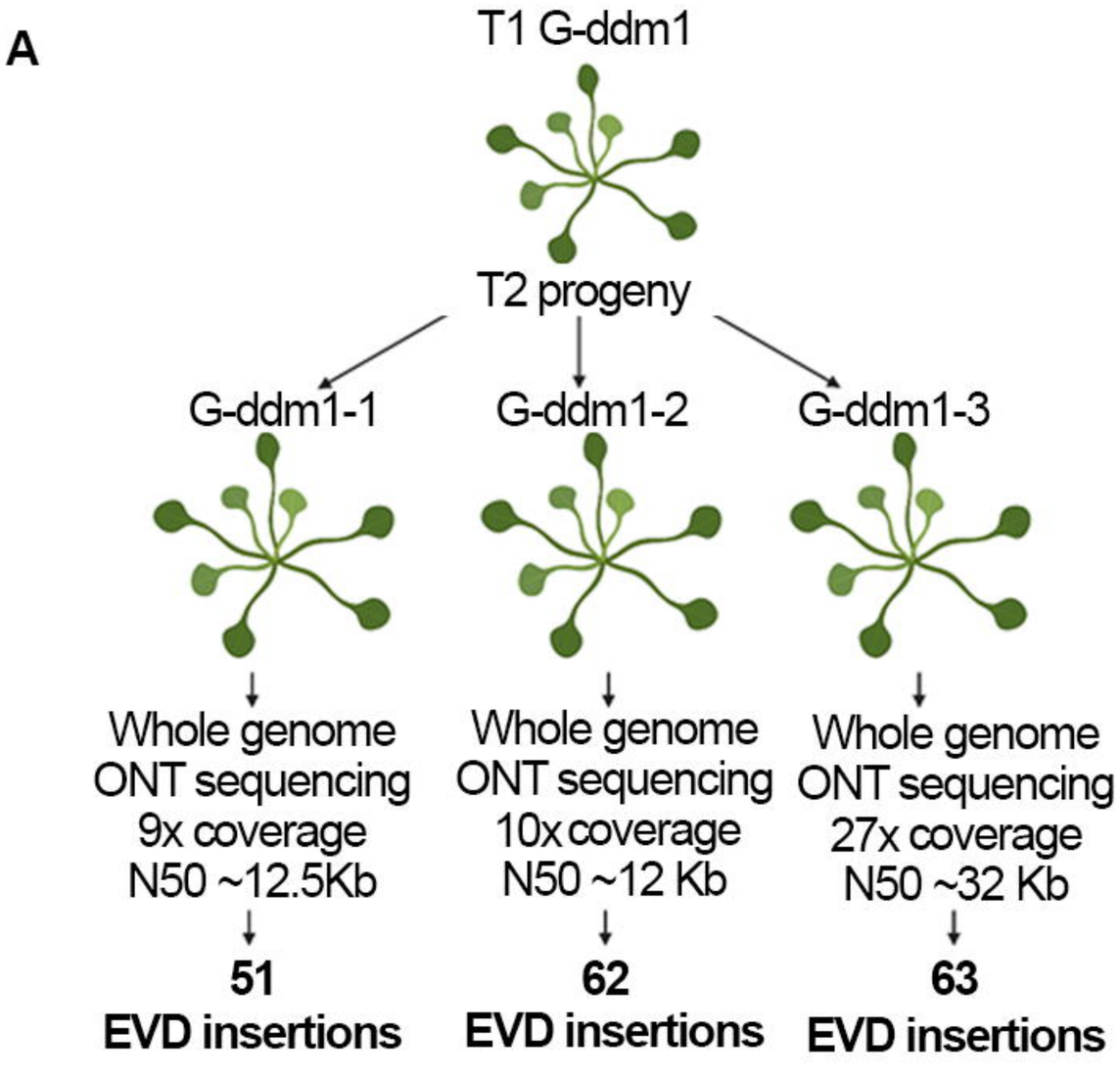
Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P.J.N.R.G. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. Nature Reviews Genetics **18,** 292-308.

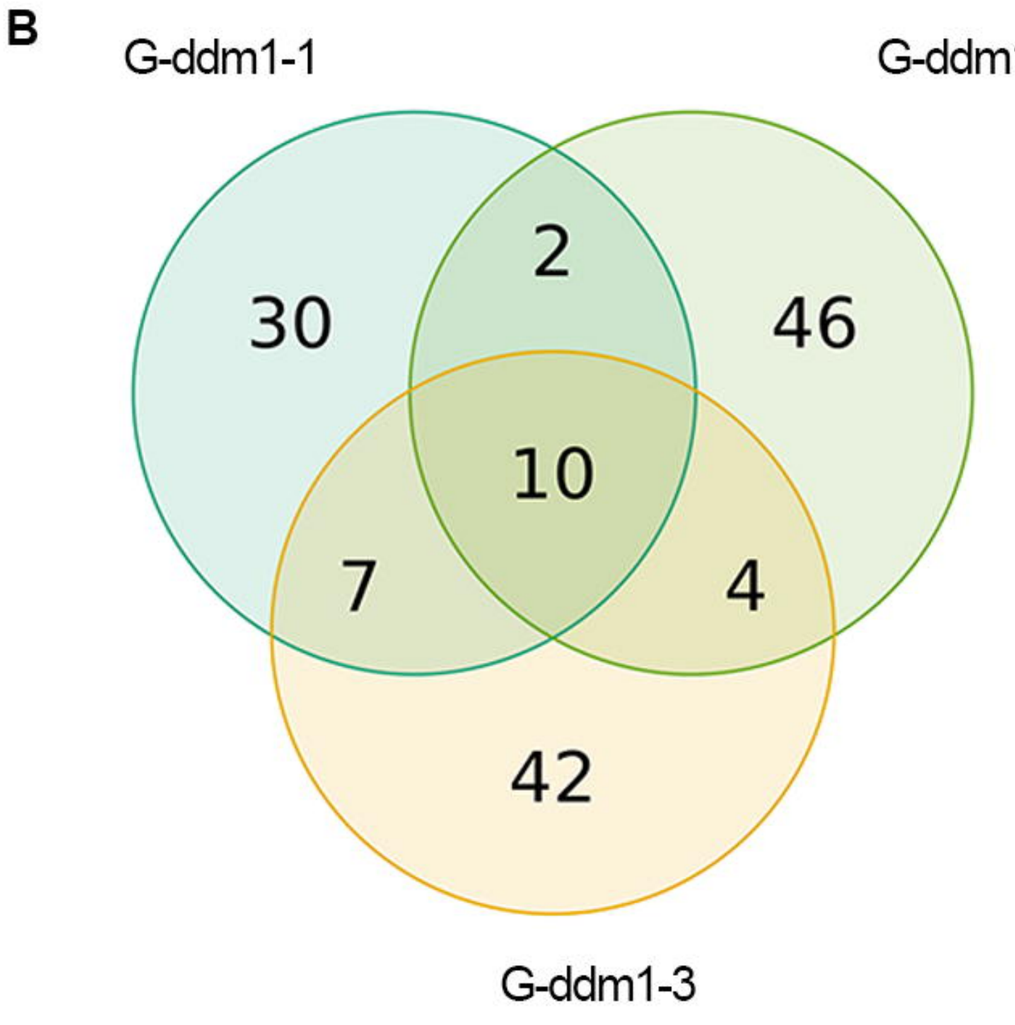
Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. Nature **461,** 423-426.

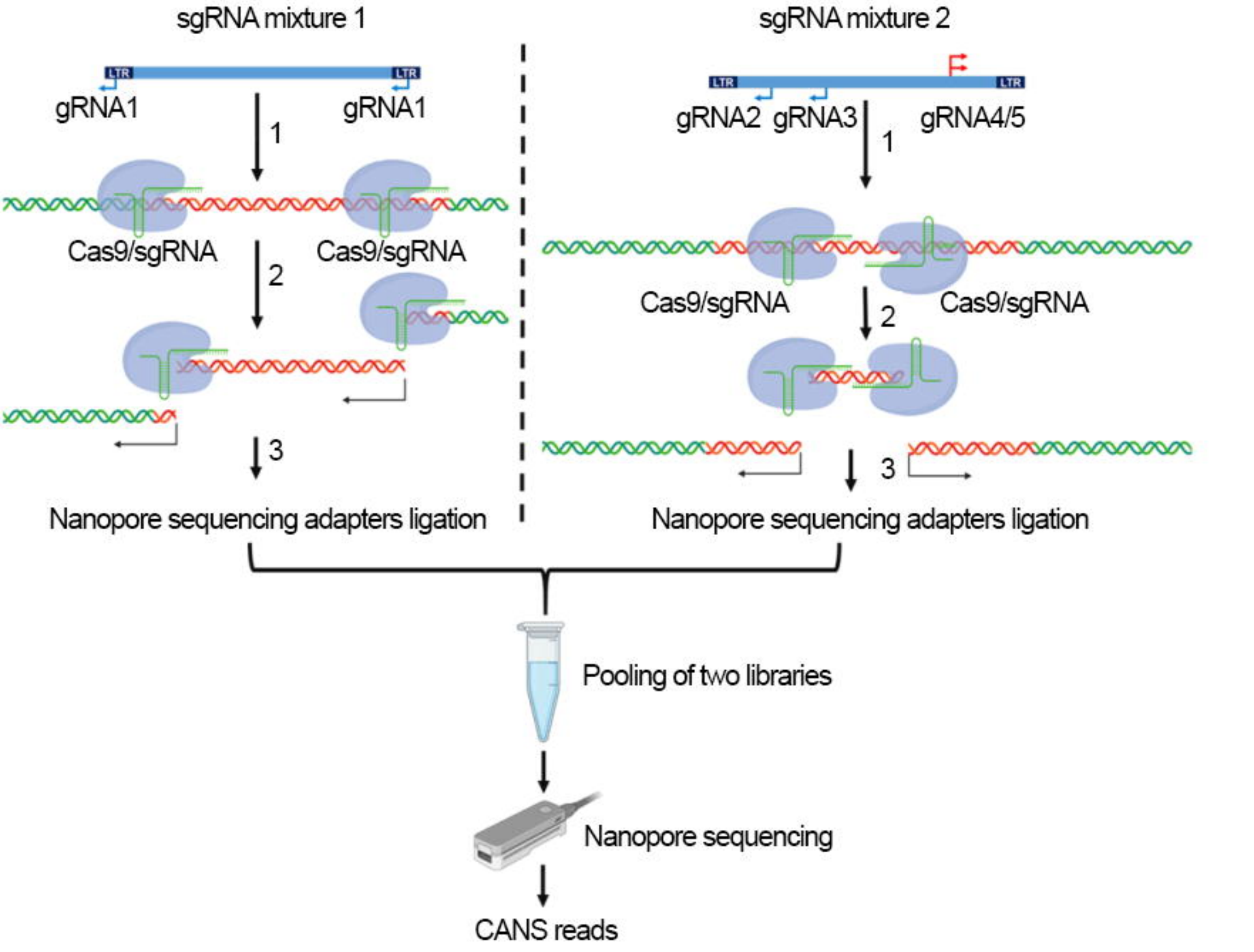
Wickham, H. (2011). ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3.2 **3,** 180-185.

Williams-Carrier, R., Stiffler, N., Belcher, S., Kroeger, T., Stern, D.B., Monde, R.-A., Coalter, R., and Barkan, A. (2010). Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. Plant J **63,** 167-177.

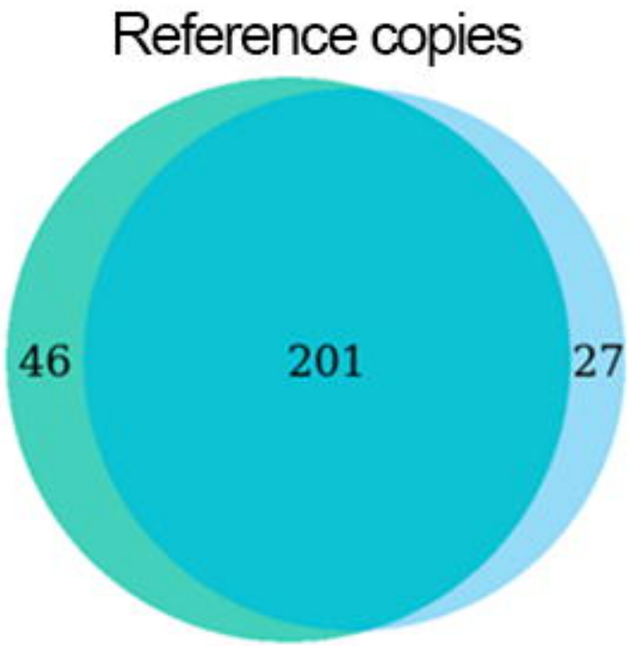
706


A
T1 G-ddm1
T2 progeny
G-ddm1-1
G-ddm1-2
G-ddm1-3
Whole genome
ONT sequencing
9x coverage
N50 ~12.5Kb
51
EVD insertions
Whole genome
ONT sequencing
10x coverage
N50 ~12 Kb
62
EVD insertions
Whole genome
ONT sequencing
27x coverage
N50 ~32 Kb
63
EVD insertions

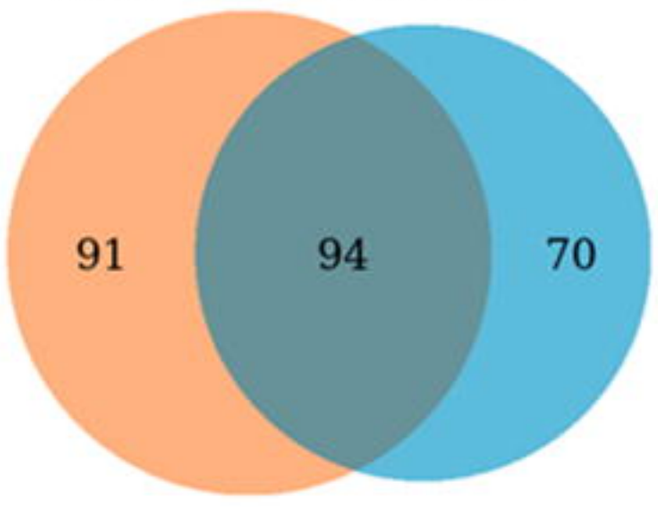

B
G-ddm1-1
G-ddm1-2
G-ddm1-3
30
2
46
10
7
4
42


sgRNA mixture 1
LTR
LTR
gRNA1
gRNA1
1
Cas9/sgRNA
Cas9/sgRNA
2
3
sgRNA mixture 2
LTR
LTR
gRNA2
gRNA3
gRNA4/5
1
Cas9/sgRNA
Cas9/sgRNA
2
3
Nanopore sequencing adapters ligation
Nanopore sequencing adapters ligation
Pooling of two libraries
Nanopore sequencing
CANS reads

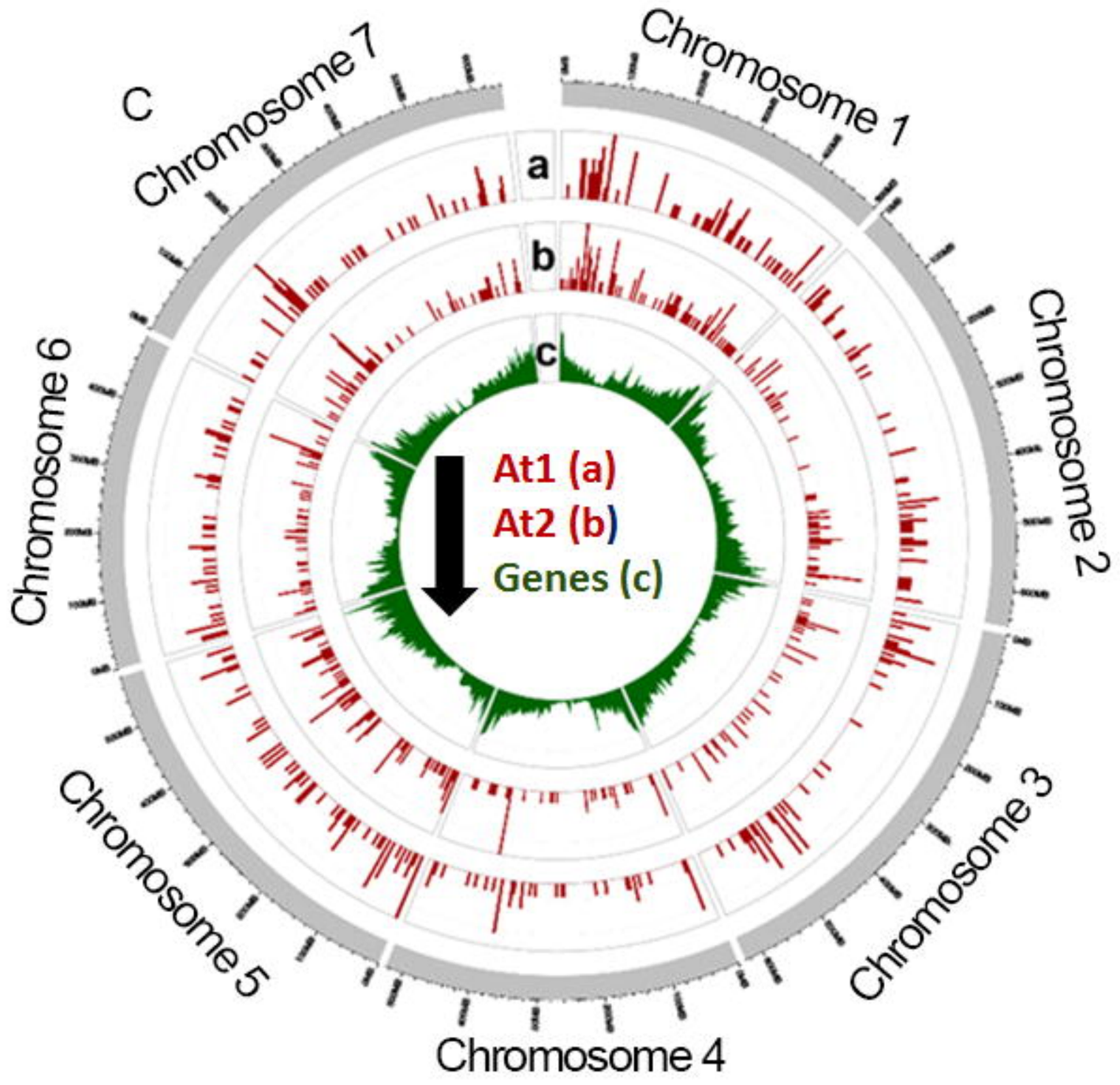
A

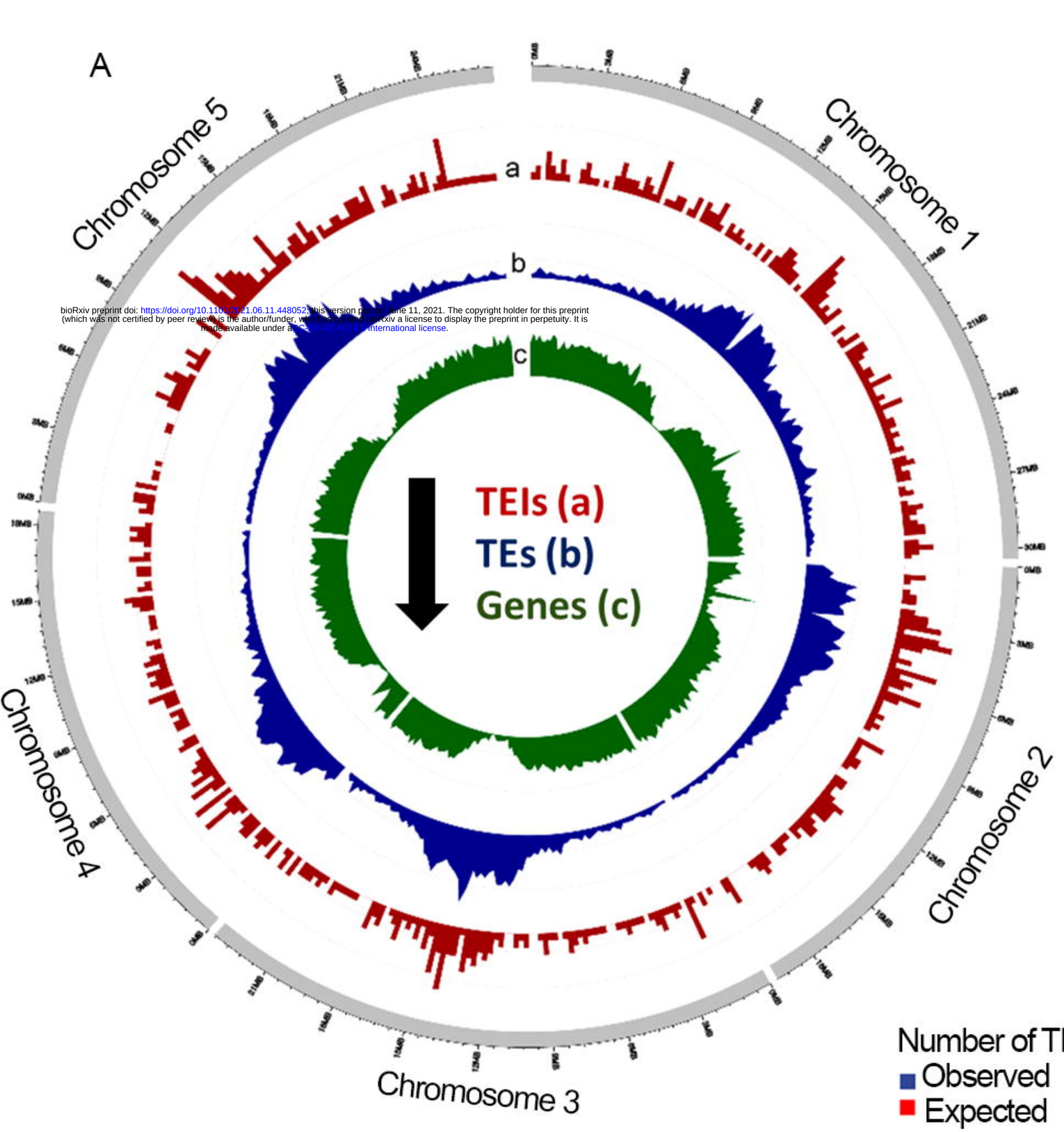


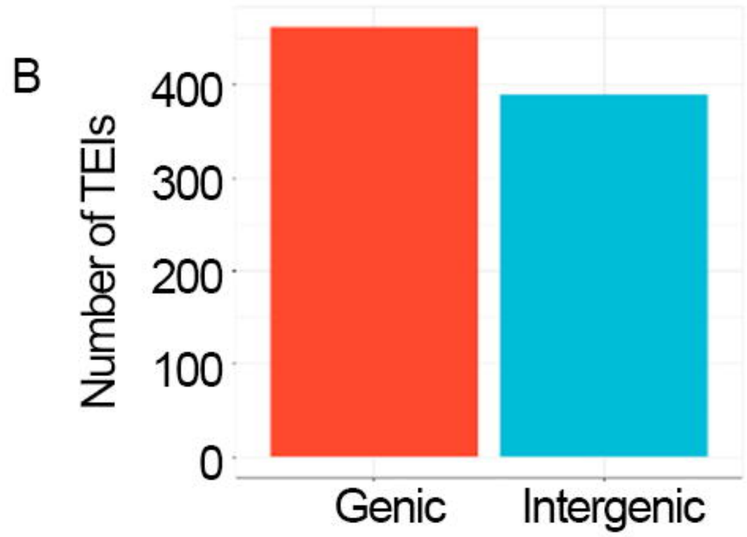
B

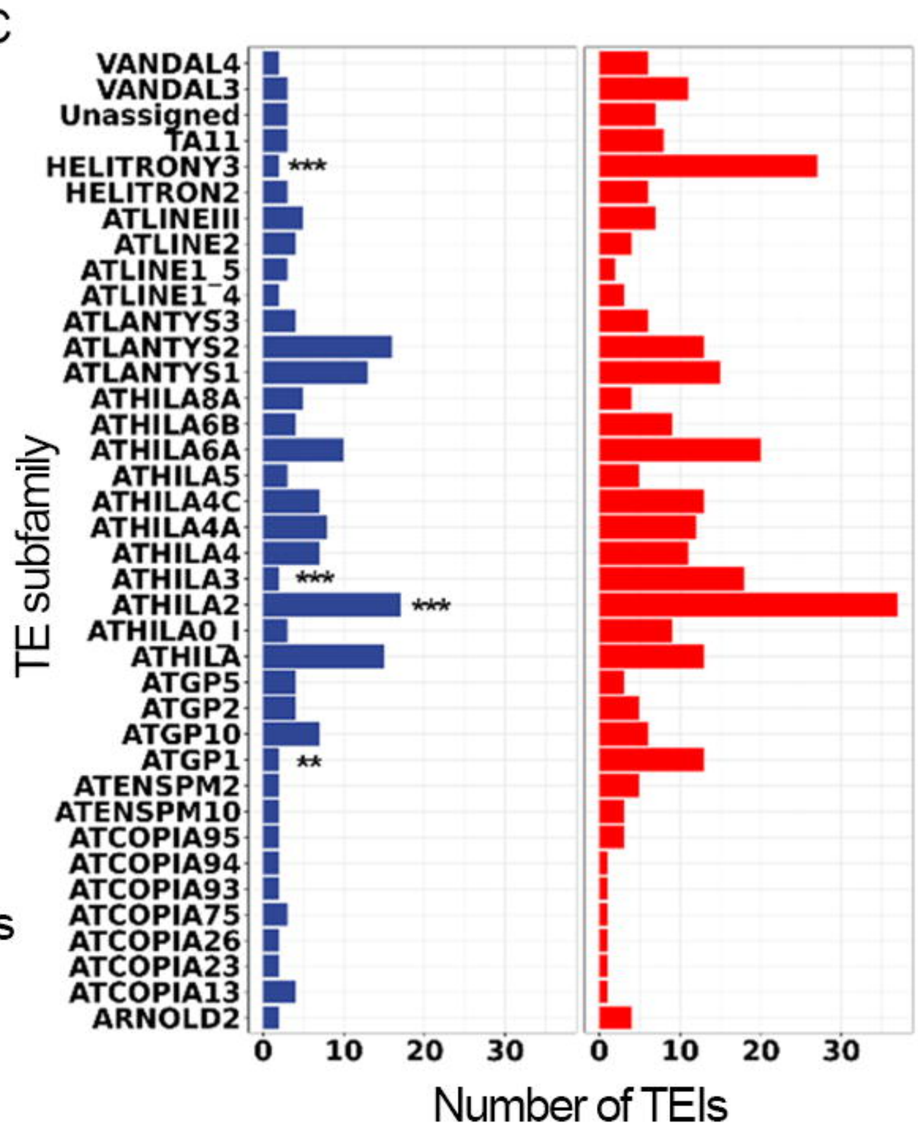


C









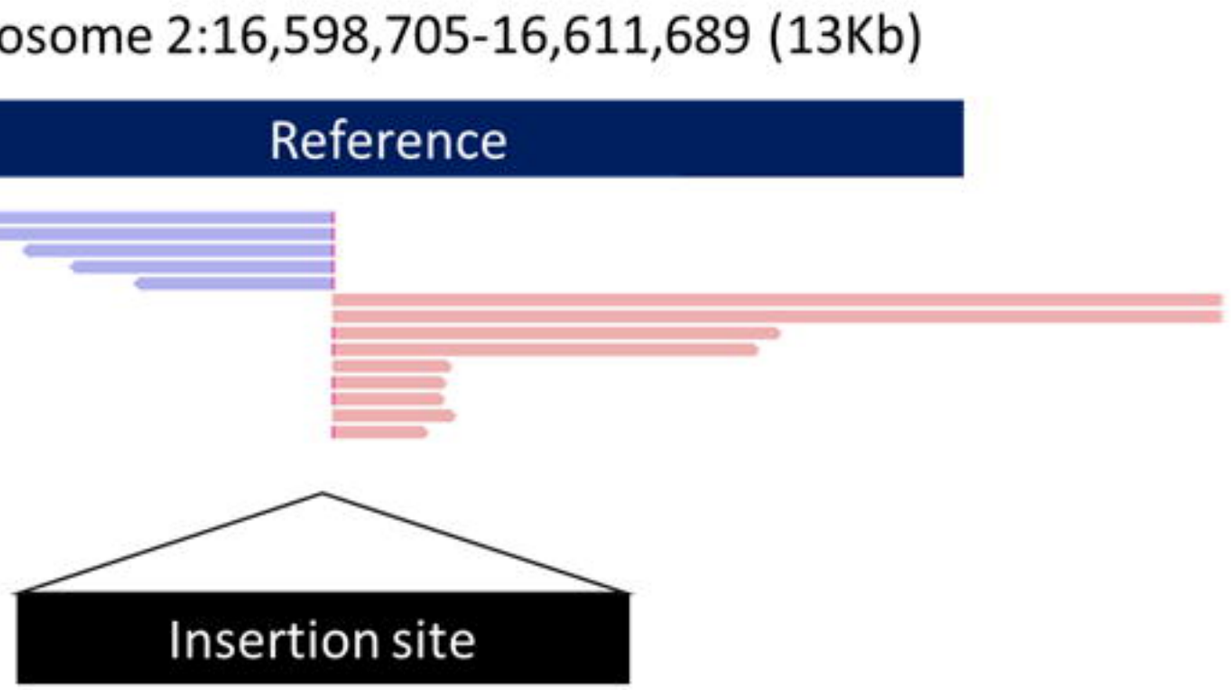
ONT read mapping to the reference genome

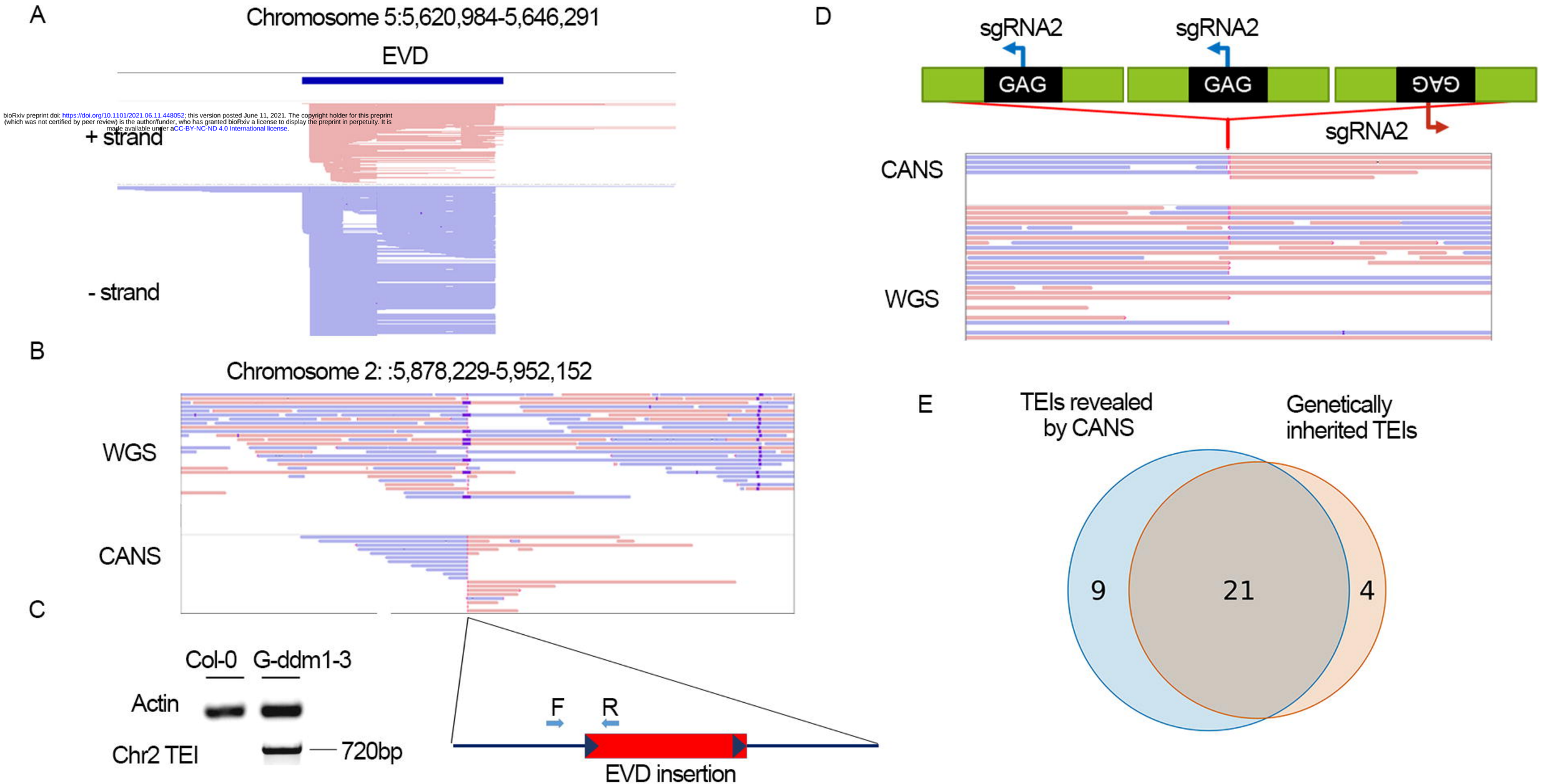
Filter reads by MQ>40 & read length > 750 & Supplementary Alignment tag

Identification of clipped reads and filtration by clipping side and length

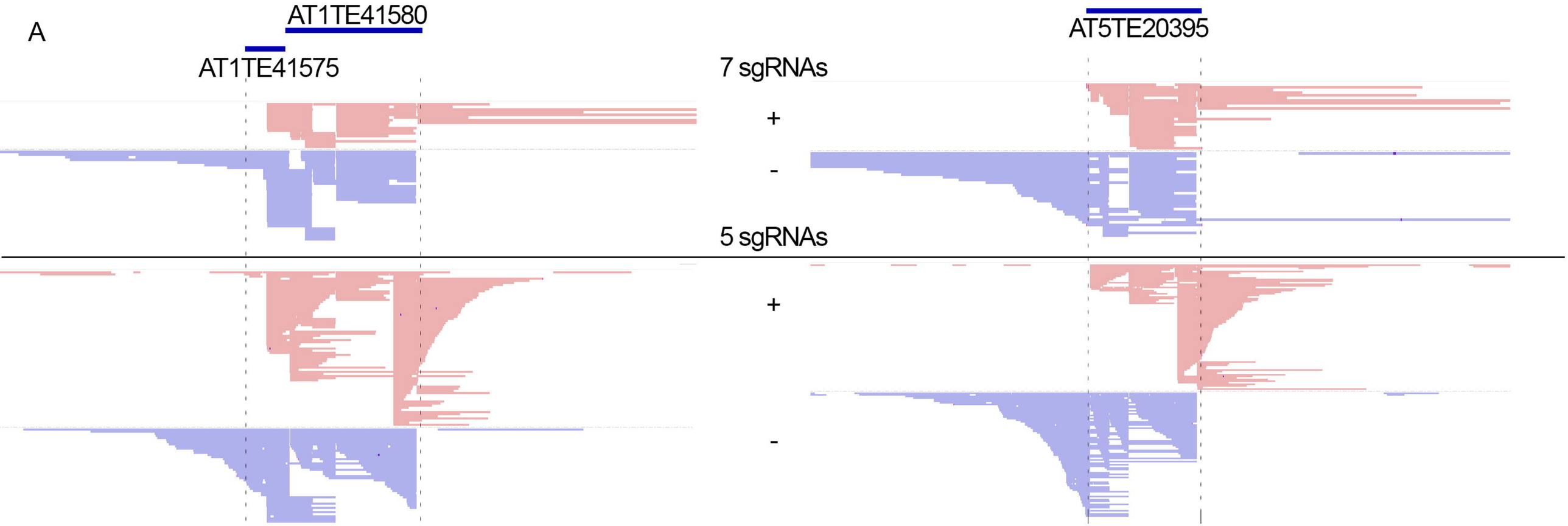
Filtration of the reads by similarity of the clipped part to the target

Insertion coordinate identification and report the results





A



B