Genome Biology

RESEARCH

Open Access

Ultra-sensitive detection of transposon insertions across multiple families by transposable element display sequencing

Pol Vendrell-Mir[1†], Basile Leduque[1†] and Leandro Quadrana[1*]

†Pol Vendrell-Mir and Basile Leduque contributed equally to this work.

*Correspondence: leandro.quadrana@cnrs.fr

[1] Institute of Plant Sciences Paris-Saclay (IPS2), Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Université Evry, Université Paris-Saclay, Gif Sur Yvette 91190, France

Abstract

Background: Mobilization of transposable elements (TEs) can generate large effect mutations. However, due to the difficulty of detecting new TE insertions in genomes and the typically rare occurrence of transposition, the actual rate, distribution, and population dynamics of new insertions remain largely unexplored.

Results: We present a TE display sequencing approach that leverages target amplification of TE extremities to detect non-reference TE insertions with high specificity and sensitivity, enabling the detection of insertions at frequencies as low as 1 in 250,000 within a DNA sample. Moreover, this method allows the simultaneous detection of insertions for distinct TE families, including both retrotransposons and DNA transposons, enhancing its versatility and cost-effectiveness for investigating complex "mobilomes." When combined with nanopore sequencing, this approach enables the identification of insertions using long-read information and achieves a turnaround time from DNA extraction to insertion identification of less than 24 h, significantly reducing the time-to-answer. By analyzing a population of *Arabidopsis thaliana* plants undergoing a transposition burst, we demonstrate the power of the multiplex TE display sequencing to analyze "evolve and resequence" experiments. Notably, we find that 3–4% of de novo TE insertions exhibit recurrent allele frequency changes indicative of either positive or negative selection.

Conclusions: TE display sequencing is an ultra-sensitive, specific, simple, and cost-effective approach for investigating the rate and landscape of new TE insertions across multiple families in large-scale population experiments. We provide a step-by-step experimental protocol and ready-to-use bioinformatic pipelines to facilitate its straightforward implementation.

Keywords: Transposable elements, Transposition, Evolve and resequencing, TE enrichment, Target sequencing, TE dynamics

Background

Transposable elements (TEs) are sequences that can move and replicate around the genome. TEs are categorized in two broad classes based on their mobilization mechanisms. Class I retrotransposons move via an RNA intermediate, while class II DNA transposons use a "cut and paste" mechanism instead. TEs are further divided into superfamilies and families based on distinctive sequence features, such as the presence of specific terminal repeats or conserved protein domains [1].

Although TEs have long been regarded as mere genomic parasites [2], this view has dramatically shifted and it is now widely appreciated that TEs are key players in the functional as well as the structural evolution of genomes. Notwithstanding, the extent to which TE insertions contribute to heritable phenotypic variation remains unknown in most eukaryotes. This lack of knowledge is largely due to their repetitive nature, which makes TEs less amenable than single copy DNA to genome-wide analyses, particularly using short-read sequencing technologies. Moreover, TE mobilization is typically rare and therefore new TE insertions tend to be missed in small-scale population studies. Hence, being capable of detecting new TE insertions in large populations of individuals is essential to study transposition activity and its impact on genome function and evolution.

Several methods have been developed to detect new TE insertions using short-read [3–12] or long-read [13, 14] whole genome resequencing data. Due to the difficulties to detect non-reference TE insertions, the sensitivity and specificity of these methods are strongly penalized by limited sequencing depth [15], rendering the detection of low-frequency TE insertions virtually impossible. Alternatively, several methods relying on the selective throughput sequencing of TE extremities, either using capture approaches [16–18], Cas9-mediated target DNA cleavage followed by nanopore sequencing [19–21], or specific amplification [22–25], have been developed. Capture approaches rely on the use of biotinylated oligos to selectively pull-down DNA fragments with complementary sequences [16–18]. Although these methods provide good sensitivity, designing specific oligos is challenging and their production costly, limiting their wide implementation. Approaches combining Cas9-mediated cleavage and adapter ligation followed by long-read sequencing have been recently proposed as an interesting alternative, as it allows to sequence native DNA without additional steps [19, 20]. However, the enrichment efficiency of such methods is limited, precluding its use for population level studies. On the other hand, approaches based on target amplification of specific sequences theoretically offer the highest sensitivity. These methods have primarily been used to detect somatic insertions [22–25], and their applicability in investigating germline insertions remains underexplored. Importantly, because PCR amplification is prone to the generation of chimeric or non-specific products, these methods typically require case-by-case modifications of experimental and bioinformatic steps to limit the number of false positives. As a result, methods based on target amplification have been tailored to study specific TEs, mostly mammalian L1s or ERVs, which raises questions about their versatility and applicability to analyze diverse types of TEs, including both retrotransposons and DNA transposons. In addition, amplification approaches often target single TEs, which brings into question their ability to assess transposition of multiple TE families

simultaneously, a crucial factor for investigating transposition activity in species with a large number of distinct mobile TE families.

Here, we describe a cost-effective (less than 12 EUR per sample's library), simple, and ultra-sensitive assay to detect non-reference TE insertions in genomes. The method is based on classical low-throughput transposon display assays [22, 26], but includes critical modifications that make it readily compatible with high-throughput short- and long-read sequencing techniques. We demonstrate its utility to investigate germline insertions using the model plant *Arabidopsis thaliana*, and we show that the method provides ~ 250,000-fold enrichment of target TEs, enabling the detection of low-frequency insertions in large populations. The obtained libraries can be analyzed using nanopore sequencing technologies, leading to longer reads and, importantly, a turnaround time from DNA extraction to insertion identification of less than 24 h, significantly reducing the time-to-answer. As a proof-of-concept, we investigated the short-term evolutionary dynamics of new TE insertions in an evolve and resequencing (E&R) experiment using large populations of *A. thaliana* plants experiencing ongoing TE invasions. Last, we provide a detailed hands-on protocol, including critical steps and quality controls, as well as off-the-shelf bioinformatic pipelines to analyze the obtained sequencing results.

Results

TE display sequencing

Classic transposon display assays rely on genomic DNA digestion using restriction enzymes followed by ligation of an end-compatible custom adapter, and PCR amplification using a primer specific to the target TE sequence and another complementary to the adapter [22, 26]. To adapt this assay for high-throughput detection of TE insertions using next generation sequencing (NGS), we replaced enzymatic digestion with random fragmentation by sonication, producing fragments of approximately 200–500 bp (Fig. 1a). Similar to standard NGS library preparations, the fragmented DNA is end-repaired and A-tailed to enable the ligation of custom adapters, which are compatible with Illumina sequencing. We designed the adapters to avoid amplification by adapter primers only, which would otherwise lead to the generation of random whole genomic libraries. Specifically, the priming sites on the adapter are not present ab initio, but generated through a first round of extension from the TE primer, analogous to the strategy used in vectorette/splinkerette PCR strategies [27, 28]. Unlike these strategies, which utilize hairpins or partially mismatched "bubbled" adapters, the present method relies on asymmetric Illumina forked-adapters. These adapters are much shorter and readily compatible with NGS libraries, providing advantages compared to vectorette/splinkerette PCR adapters. To block extension from the asymmetric forked-adapter, its short arm includes a 3′ dideoxy nucleotide (Fig. 1b). Even in the unlikely situation where an adapter is extended first, producing templates with full adapter sequences at both ends, these fragment ends will preferentially anneal to form an intramolecular "panhandle" structure, which cannot be amplified further. The "suppression PCR" effect occurs because the complementarity of the adapter's primer is shorter than that of the adapter itself, favoring intramolecular pairing. In combination, the universal adapter's design and primer sequences ensure that only

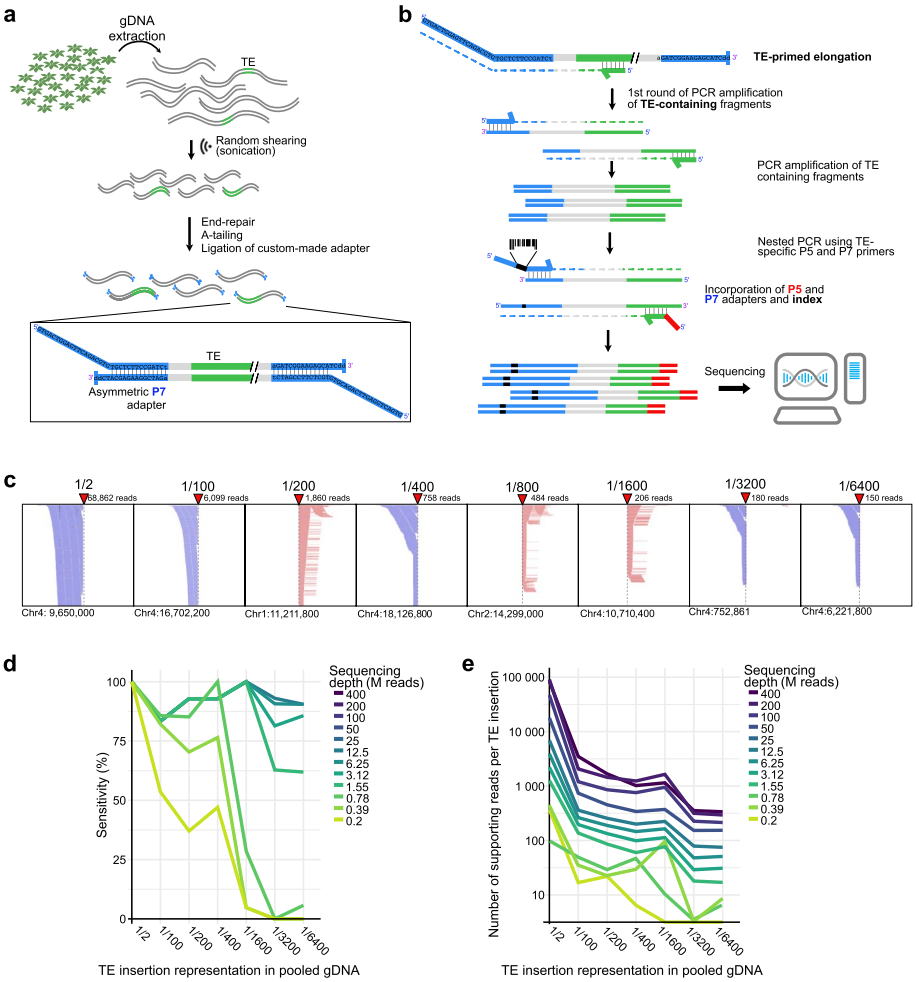

Fig. 1 Transposable display sequencing method. **a** Schematic representation of the initial steps to prepare adapter-containing whole genome libraries. **b** Schematic representation of specific amplification of TE-containing adapted fragments and incorporation of P5 and P7 adapters compatible with Illumina sequencing. **c** Examples of TE insertions detected at different dilution factors. **d** Sensitivity of TEd-seq to detect TE insertions at different dilution factors and in relation to sequencing variable depth. **e** Number of reads supporting each TE insertion present at different dilution factors in relation to variable sequencing depth

bona fide PCR templates containing both the target TE site and adapter sequence are generated during the first rounds of extension.

To increase further the sensitivity and specificity of the assay, we included a nested PCR step using a TE-specific primer closer to the TE edge together with an adapter's primer (Fig. 1b), respectively containing the P5 and P7 sequences required for cluster recognition on Illumina sequencing machines. The adapter's primers also include sequence barcodes, enabling the multiplexing of up to hundreds of samples in the same sequencing run. Critically, to improve sequencing quality due to the low diversity of the amplicon library, nested TE-specific primers contain different numbers of spacer nucleotides before the P5 adapter. Such straightforward modifications significantly improve library diversity and overall sequencing quality (Additional file 1: Fig. S1).

To analyze the data generated by this assay, we developed a ready-to-use bioinformatic tool that takes as input the sequenced reads, targeted TE extremity, and reference genome sequence to detect TE insertions. Specifically, following demultiplexing, reads are processed bioinformatically to eliminate PCR duplicates and select paired-reads containing the target TE sequence end, which is trimmed to retrieved informative insertion sites sequences. Alignment of reads on the reference genome and their clustering allows the detection of reference and non-reference insertion sites (Fig. 1c).

Highly specific and ultra-sensitive detection of new TE insertions

To assess the sensitivity and specificity of the transposable element display sequencing (TEd-seq) approach to detect non-reference TE insertions, we took advantage of a population of TE accumulation (TEA) lines in *A. thaliana* that have underwent extensive transposition following the epigenetic activation of TEs [12]. The identity and genomic location of heritable TE insertions in each of these lines were previously characterized using Illumina mate-pair libraries [18, 29] and TE sequence capture [18], providing a gold standard data set to test the sensitivity and specificity of TEd-seq. We selected eight lines carrying distinct repertoires of non-reference TE insertions of the LTR retrotransposon *ATCOPIA93* (also known as *Évadé*). We extracted genomic DNA (gDNA) from each of these TEA lines and serially diluted them in DNA extracted from the reference wild-type Col-0 plants, leading to different concentrations of TEA line's DNAs (from 1/2 to 1/6400). Pooling together these differentially diluted gDNA thus leads to a mixed gDNA with 212 non-reference TE insertions present at different dilutions (Additional file 1: Table S1).

The resulting pooled DNA was subjected to TEd-seq using primers specific to the mobile *ATCOPIA93* copy (*AT5TE20395*), sequenced to a depth of 400 M pair-end reads, and analyzed bioinformatically to detect non-reference TE insertion sites. Comparison between the set of non-reference *ATCOPIA93* insertions detected by TEd-seq to the gold standard set of previously identified insertions [29] showed that 87–100% of TE insertions were detected for all the dilutions (Fig. 1c, d). Importantly, the number of reads supporting each *ATCOPIA93* insertion correlated well with the dilution factor (Fig. 1c, e), indicating that the number of supporting reads is a reliable estimator of the TE insertion representation.

To evaluate the influence of sequencing depth on the sensitivity of TEd-seq, we subsampled decreasing numbers of row paired-end reads obtained in the TEd-seq experiment and detected non-reference TE insertions. As expected, sensitivity decreased in relation to sequencing depth (Fig. 1d). Specifically, 80% of the TE insertions present in the 1/6400 dilution were detected with as little as 6 M pair-end reads. On the basis of the empirical sensitivity obtained from the serially diluted DNA, we applied a subsampling approach to estimate the theoretical limit of detection for TEd-seq (Additional file 1: Fig. S2). This analysis indicated that 400 M reads are sufficient to detect TE insertions present at dilutions as low as 1/250,000.

Sensitive detection of TE insertions using ONT sequencing

Recent developments in sequencing technologies based on Oxford Nanopore Technologies (ONT) are revolutionizing the output and speed of genomic analysis. We reasoned

that in addition to allowing the sequencing of long DNA fragments, which can facilitate the identification of insertion sites within highly repetitive or low complexity regions, its portability and real-time generation of sequencing data can significantly reduce the time-to-answer. We thus tested the possibility of sequencing a TEd-seq library using ONT (Fig. 2b). We optimized the fragmentation and size-selection steps to obtain longer DNA fragments, which are more suitable for ONT sequencing.

We performed a TEd-seq library of the retrotransposon *ATCOPIA93* using gDNA extracted from a single TEA line (TEA500) diluted 1/200 in Col-0 gDNA. As expected, the average size of ONT reads is longer than the fragments obtained by short-read sequencing (814.27 vs 350 nt, respectively), with some reads reaching more than 3000 nt (Fig. 2b and c). We next detected non-reference TE insertions using a modified bioinformatic pipeline compatible with ONT reads (see Methods). Comparison between the set of non-reference *ATCOPIA93* insertions detected by TEd-seq to the gold standard set of previously identified insertions [29] in this TEA line showed 100% correspondence, demonstrating that combination of TEd-seq with long-read ONT sequencing is highly specific. To test the sensitivity of ONT sequencing to detect non-reference TE insertions, we generated and sequenced the TEd-seq library containing the mixed pool of serially diluted gDNA from the different TEA lines. With as little as 1 M ONT reads, we identified more than 70% of insertions present in 1/400 dilution (Fig. 2d). This sensitivity is close to the one obtained by ~ 1 M short-reads, indicating that the major factor

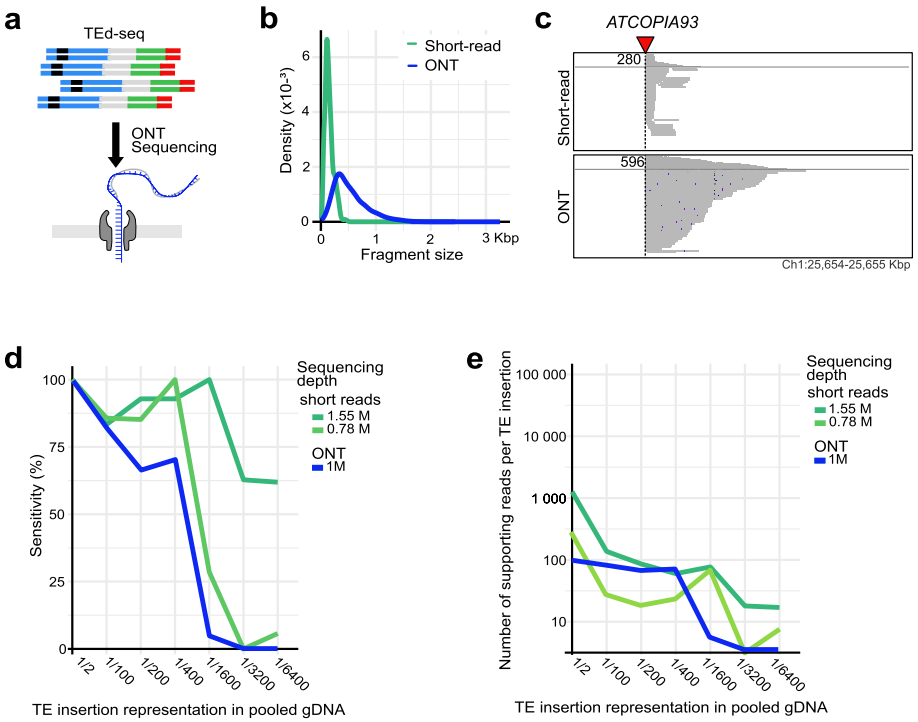

Fig. 2 Transposable elements display long-read sequencing. **a** Schematic representation of TEd-seq sequencing using ONT. **b** Fragment size distribution of short-read and ONT-based TEd-seq. **c** Example of a TE insertion detected by short-read and ONT-based TEd-seq. **d** Sensitivity of ONT-based TEd-seq to detect TE insertions at different dilution factors. **e** Number of reads supporting each TE insertion present at different dilution factors for both short-read and ONT-based TEd-seq

influencing the sensitivity of TEd-seq is the depth of coverage. In combination, these results establish that TEd-seq libraries are compatible with ONT sequencing, and that the results are comparable to the ones obtained with short-read sequencing. Notably, the portability and short hand-off time of ONT sequencing offer a turnaround time of less than 24 h from DNA extraction to insertion identification, significantly expediting the investigation of TE mobilization in routine experiments.

Simultaneous detection of insertions from multiple TE families using multiplex TEd-seq

Eukaryotic genomes typically contain several active TE families. To effectively investigate transposition activity, it is desirable to simultaneously detect non-reference insertions for multiple TE families. To explore this possibility, we designed TEd-seq primers specific for the *CACTA* and *MuDR* DNA transposons *ATENSPM3* and *VANDAL21*, which are highly mobile in the population of TEA lines [29]. Using gDNA from TEA line 454 (dilution 1/10) and 500 (dilution 1/200), we generated, for each TEA line, independent TEd-seq libraries for *ATCOPIA93*, *VANDAL21*, and *ATENSPM3*, together with multiplexed TEd-seq libraries using an equimolar mix of the three TE-specific primers during the amplification steps (Fig. 3a). Based on 1 M ONT reads, we detected insertions for the three TE families (Fig. 3b). Remarkably, between 97.77 and 100% of the homozygous non-reference insertions previously identified in the TEA lines [29] were

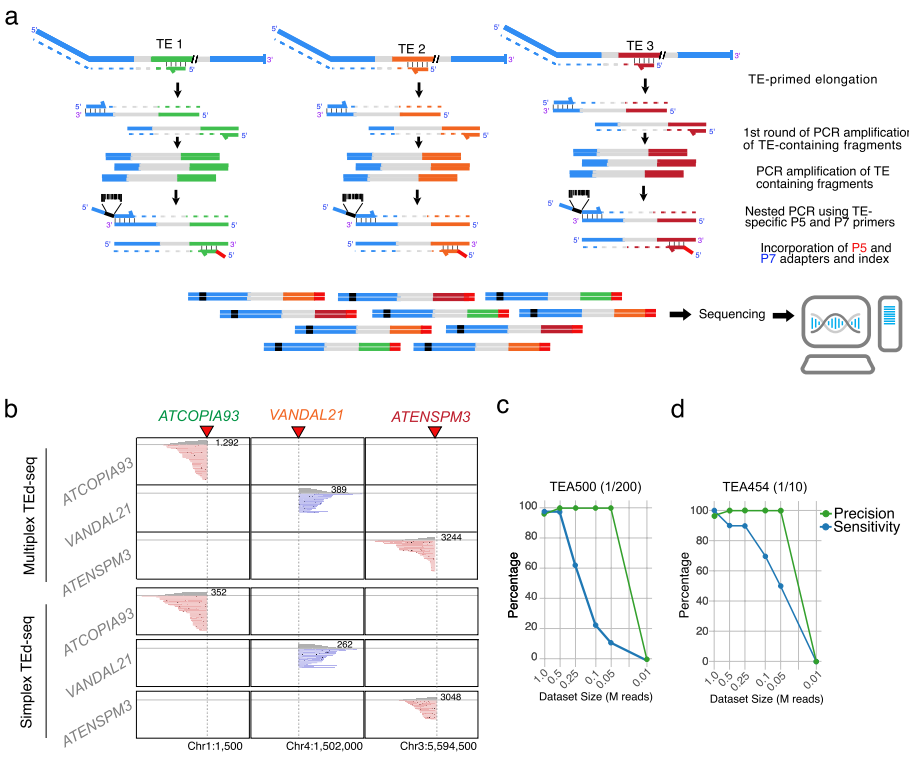

Fig. 3 Simultaneous detection of insertions for different families by multiplexed transposable element display sequencing. **a** Schematic representation of specific multiplexed amplification of TE-containing adapted fragments and incorporation of P5 and P7 adapters compatible with Illumina sequencing. **b** Examples of TE insertions from different TE families detected by multiplex and simplex TEd-seq. **c** and **d** Precision and sensitivity of multiplex TEd-seq to detect TE insertions present in different TEA lines

detected in our multiplexed TEd-seq experiments (Fig. 3c, d), providing a global sensitivity of ~ 98% and enabling the possibility to study several TE families simultaneously. In addition, out of the 96 non-reference TE insertions detected in the multiplexed libraries, only three appear to be false positives, leading to an overall precision rate of 96.8% (Fig. 3c, d). Altogether, these results established multiplex TEd-seq as a remarkably cost-effective, highly sensitive, and specific method to simultaneously detect non-reference insertions of different TE families, enabling the investigation of rich mobilomes.

Multiplex TEd-seq empowers evolve and resequencing experiments

Although most TE-induced mutations are likely deleterious, empirical and theoretical evidence suggest that a continuum of selective effects must exist [30]. However, direct estimations of the evolutionary dynamics of new TE insertions remain largely underexplored. Whole genome sequencing (WGS) of pools of individuals from experimentally evolved populations, such as in common garden competition experiments or experimental evolution setups, has enabled the identification of genomic responses to selection. The so-called evolve and resequencing (E&R) approaches use short-read WGS to identify segregating polymorphisms and characterize their changes in allele frequency across generations [31, 32]. Detection of complex DNA sequence variations, such as TE insertion polymorphisms, as well as of rare variants remains very challenging using WGS, limiting the wide implementation of E&R approaches to investigate the evolutionary dynamics of new TE insertions. The ultra-high sensitivity of TEd-seq combined with its ability to simultaneously detect non-reference TE insertions for multiple TE families in a cost-effective manner may therefore empower E&R experiments. As a proof-of-concept for the utility of TEd-seq in quantifying the presence and frequency of non-reference TE insertions in these experimental experiments, we generated a largely isogenic wild-type population of *Arabidopsis* plants undergoing a burst of transposition for the endogenous TE families *ATCOPIA93*, *VANDAL21*, and *ATENSPM3*. Specifically, we crossed wild-type Col-0 plants with distinct TEA lines (TEA 523, 54, and 393) containing active copies for these three TEs families. F1 plants were selfed to obtain thousands of F2 seeds. An equal number of F2 seeds (3500) from each cross, together with naïve Col-0 seeds (500), were mixed to obtain a so-called TE invaded population (G0) (Fig. 4a). We performed two parallel E&R experiments by growing 2500 plants from the TE invaded population in environmental chambers reproducing realistic climates, including temperature, humidity, and light intensity and quality (Additional file 1: Fig. S3) recorded at an experimental station near Paris (see Methods). All produced seeds at the end of the life cycle were collected from each population. This was repeated for another generation to obtain G2 offspring.

To assess the ability of TEd-seq to accurately estimate the allele frequency of non-reference TE insertions, we performed multiplex TEd-seq (*ATCOPIA93*, *VANDAL21*, and *ATENSPM3)* on two random samples of 1000 G1 offsprings from population 1. In total, we identified 340 and 321 non-reference insertions in each sample, most of which belonged to the LTR retrotransposon *ATCOPIA93* and were shared between samples (Fig. 4b and c). We next estimated the allele frequency of each TE insertion by comparing the numbers of reads supporting new (i.e., non-reference) TE insertions in relation to that for fixed (i.e., reference) insertions. This analysis revealed that allele

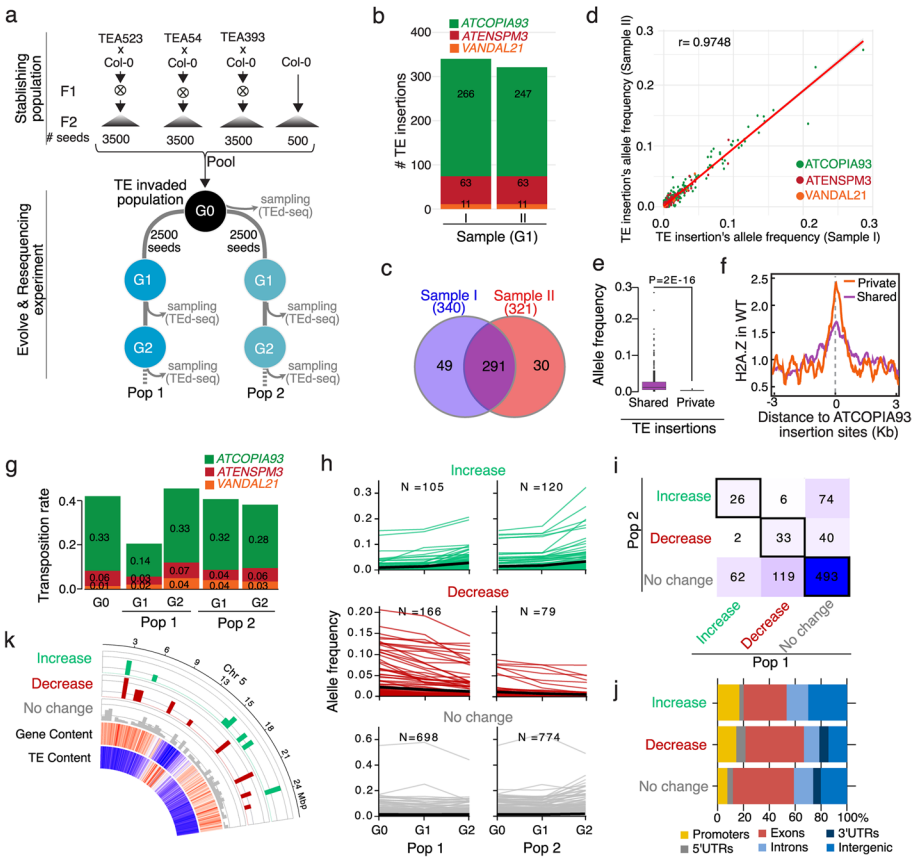

Fig. 4 Investigation of the evolutionary dynamics of new TE insertions using TEd-seq. **a** Schematic representation of the evolve and resequencing (E&R) experiment developed to study the evolutionary dynamics of new TE insertions. Two independent populations (Pop. 1 and Pop. 2) were used to run two E&R replicated experiments. **b** Number of TE insertions identified by multiplex TEd-seq in two independent samples of 1000 offspring of G1 Pop. 1 for the three TE families (*ATCOPIA93*, *VANDAL21*, and *ATENSPM3*) transposing actively in the experimental population. **c** Number of private and shared insertions detected in the two samples. **d** Comparison of allele frequencies of all TE insertions estimated in the two populations. **e** Allele frequencies of shared and private TE insertions. Statistically significant differences are indicated (MWU test). **f** Meta-analysis of levels of H2A.Z levels around shared and private *ATCOPIA93* insertion sites. **g** Estimated transposition rate across generations in both replicated populations. Three independent samples were analyzed in each case. **h** Allele frequencies change across generations in both populations. Insertions showing increase, decrease, or no change allele frequencies were selected by comparison with allele frequencies trajectories calculated under random genetic drift. **i** Comparison of TE insertions detected as increase, decrease, or no change among populations. **j** Fraction of TE insertions identified as recurrent increase, decrease, or no change within genic features and intergenic regions. **k** Circos representation of TE insertions identified as recurrent increase, decrease, or no change in chromosome 5 of *Arabidopsis*. Density of annotated genes and TEs are indicated as heatmaps

frequency estimations were highly consistent between samples (Fig. 4d), demonstrating that TEd-seq offers reliable quantification of the allele frequency of TE insertions.

Shared insertions among offspring samples should correspond to segregating alleles in the G1, and as such they are expected to be present at higher allele frequencies than private insertions. To test this expectation, we compared the allele frequency of shared and private non-reference TE insertions. As predicted, shared insertions were covered by a much higher number of reads than private ones (Additional file 1: Fig. S4), and we estimated that segregate at frequencies higher than 0.025, compared

to 0.004 for the latter (Fig. 4e). Thus, private insertions represent de novo heritable transposition events.

It was previously shown that *ATCOPIA93* integrates preferentially within nucleosomal DNA containing the histone variant H2A.Z [29]. Analysis of the integration sites of *ATCOPIA93* in relation to the presence of H2A.Z confirmed this preference and revealed a higher occupancy of H2A.Z at de novo compared to shared TE insertion sites (Fig. 4f). This result are consistent with counterselection of TE insertions within H2A.Z containing regions.

Next, we set out to investigate the transgenerational dynamics of active transposition on the E&R experiment. We performed multiplex TEd-seq on DNA extracted from independent samples of 1000 offspring of G0, G1, and G2 generations for both replicated populations (Fig. 4a). In total, we identified 5799 insertions, 48% of which were private to individual samples. Using the number of private insertions as a proxy for de novo TE mobilization events, we estimated the transposition rate across replicates and generations. Estimated transposition rates were very similar between populations and generations (Fig. 4g), consistent with lineal, rather than exponential [29], accumulation of new TE copies across generations.

To characterize the evolutionary dynamics of segregating TE insertions, we estimated the allele frequency across generations of insertions shared among samples. To detect insertions potentially subjected to selection, we compared their allele frequencies with neutral expectations based on random genetic drift and considering less than 5% of outcrossing rate [33]. Segregation dynamics of most (~63%) TE insertions were indistinguishable from the expected random genetic drift. Conversely, 19% and 14% of the TE insertions show a statistically significant increase or decrease in allele frequencies in at least one population, respectively. In addition, 26 (3%) and 33 (3.9%) of these insertions show paralleled allele frequency changes in both populations (Fig. 4i). Such parallel changes are unlikely to be produced by chance (hypergeometric test p value $< 6.3E-6$ and $6.9E-8$ for increases and decreases, respectively), suggesting recurrent positive or negative selection of these TE insertions across replicated populations. In addition, 493 TE insertions show no changes in allele frequency across populations, suggesting that they correspond to neutral, or nearly neutral, mutations (Fig. 4i). The distribution of TE insertions with reproducible increase, decrease, or no changes in allele frequencies shows no chromosome-wide bias (Fig. 4k), suggesting that the observed changes in allele frequencies are not due to selection on underlying haplotypes (i.e., background selection), but rather on individual TE insertions. Moreover, TE insertions with recurrent decrease in allele frequency show overrepresentation over exonic features compared to neutral TE insertion mutations (Fig. 4j), consistent with exonic insertions being more likely associated with deleterious effects.

Among the 18 genes with insertions in their gene body or promoter that are increasing in population frequency, we identified 14 GO terms. These terms are primarily associated with transmembrane transporters, suggesting a role for these TE insertions in cell-environment interaction and transport processes. We identified 56 GO terms associated with the 28 genes carrying counterselected TE insertions. These terms are involved in various binding activities (e.g., nucleotides and nucleosides), cell wall structure and organization, which are essential for maintaining cell integrity and ensuring proper

cellular interactions, as well as responses to environmental changes. For instance, we found an *ATENSPM3* insertion with consistent decrease in allele frequency across populations within the gene encoding *KNUCKLES* (*KNU*), which is involved in an essential regulation of flower determinacy [34] (Additional file 1: Fig. S5). In combination, these results indicate that the vast majority of new TE insertions are neutral or nearly neutral and that as much as 4% of new TE insertions are affected by positive or negative selection in our E&R experiment.

Discussion

Here, we adapted the classical transposable element display assay to be readily compatible with modern sequencing technologies [26]. Several methods have been developed to enrich libraries for TE insertion sequences, employing different adapter ligation strategies. These include tagmentation-based approaches [9, 10], as well as protocols involving sonication followed by end-repair, A-tailing, and ligation of modified adapters [8, 11, 22, 23, 25]. Additionally, some of these methods have been adapted to the use of long-read technologies to improve the sensitivity over repetitive regions [14, 35] or to analyze the methylation status of non-reference TE copies [36]. When compared to these methods (Additional file 2), TEd-seq demonstrates several advantages, including higher sensitivity, compatibility with both short- and long-read sequencing platforms, and precise estimation of TE insertion allele frequencies in large-scale populations. Furthermore, we have optimized TEd-seq to be able to detect insertions for several TE families simultaneously, including the LTR-retrotransposon *ATCOPIA93* and the DNA transposons *VANDAL21* and *ATENSPM3*. Such a multiplex TEd-seq approach offers a powerful approach to study complex mobilomes, such as the one of the model plant *Arabidopsis* [37] or *Drosophila melanogaster* [38] with a much lower cost per sample compared to probe-based TE enrichment methods [18]. Indeed, the estimated per-sample cost is about 12 EUR, offering a cost-effective approach. Its simplicity and versatility significantly facilitate its implementation to study diverse TE families, in *Arabidopsis* as well as in other species [39]. Furthermore, its remarkable sensitivity makes it an ideal approach for mapping high-resolution TE insertion landscapes, as recently demonstrated by the investigation of somatic integration preferences for chimeric LTR retrotransposons between *ATCOPIA93* and *Tal1* (transposon of *Arabidopsis lyrata* 1) [40].

Our results demonstrate that TEd-seq is compatible with both short- as well as long-read sequencing technologies. Compared to the short-read sequencing, TEd-seq with ONT produces reads that range between 0.5 and 3 kbp, potentially unlocking the detection of non-reference TE insertions within highly repetitive sequences, such as satellite repeats, telomeres, and centromeres. Furthermore, we show that the sensitivity and specificity of TE insertion identification is comparable between both sequencing technologies. Importantly, the portability and little hand-off time of ONT sequencing offers a turnaround time from DNA extraction to insertion identification of less than 24 h, expediting the investigation of TE mobilization for routine experiments. Its simplicity and low upfront cost, combined with the steady deline in ONT sequencing prices, will greatly facilitate the study of transposition across genomes.

The combination of ultra-high sensitivity, accuracy, and cost-effectiveness are particularly suitable for large-scale experiments. As a proof-of-concept, we implemented

multiplex TEd-seq to quantify transposition dynamics across generations in populations of *Arabidopsis* plants undergoing a transposition burst. Our results show that the global transposition rate is relatively constant across generations. This observation is at odds with the notion that once initiated, TE invasion follows a chain reaction as every new copy may serve as a source for new transposition events. Unlike previous experimental populations, which are akin to mutation accumulation lines, our E&R approach involved thousands of individuals growing in a realistic natural environment. Under such circumstances, moderately deleterious mutations could be purged by natural selection, potentially limiting runaway accumulation of new TE insertions. Indeed, ~4% of segregating TE insertions show consistent decrease in allele frequency in replicated populations, pointing to significant counter selection. However, given that our proof-of-concept E&R has only two generations, estimation of counter selected TE insertions is likely underpowered. Extending these experiments to additional generations, as well their replication across distinct environments, would offer more precise estimations of fitness effect for new TE insertion mutations.

Conclusions

TEd-seq enables the simultaneous study of heritable transposition for several TEs and in a large number of individuals. Implementing TEd-seq on E&R experiments will empower the investigation of the evolutionary dynamics of new TE insertions in real time. In addition, the ultra-high sensitivity offered by this approach would be valuable to investigate somatic TE insertions, which are typically present in only one or few nuclei and therefore appear at very low frequencies in sequencing data. We provide a step-by-step protocol and bioinformatic tool to readily implement TEd-seq for studying transposition activity across virtually any TE type in any organism.

Methods

Plant material

The *A. thaliana* reference accession Col-0 and the TEA lines (known as epiRILs) used in this work were described before [18, 29]. To test the method's sensitivity, gDNA from different TEA lines were diluted into Col-0 gDNA before preparing the TEd-seq libraries (Additional file 1: Table S1). To generate the TE invaded populations, three TEA lines (epiRIL54, epiRIL523, and epiRIL393, referred here as TEA54, TEA523, and TEA393) were crossed with Col-0 and an equal number of seeds from independent F2 segregating populations were collected and mixed together to generate a founder population. One thousand seeds of this mix were sowed and plants were grown for the whole life cycle in a growth chamber (Percival) under environmental conditions simulating the climate of February 2022 to June 2022 (30′ resolution) nearby Paris (48°16′55.1″N 2°40′25.3″E) (Additional file 1: Fig. S3). All seeds were collected and pooled. Samples of 1000 seeds were germinated on 1/2MS media and grown under standard conditions. Ten days old seedlings were collected in bulk, flash frozen in liquid nitrogen, and stored at − 80 °C until use. Plant material was grinded using pestle and mortar and used for DNA extraction with the CTAB method [41].

TEd-seq protocol for short-reads

Fifty microliters of 20 ng/µl gDNA was sonicated in a Covaris M220 sonicator (peak incident power 75 W, duty factor (18.6%), 200 cycles per burst, for 175 s) to obtain fragments of 200–600 bp. Fragmented DNA was subjected to end repair and A-tailing using the NEBNext Ultra II End Prep Enzyme Mix (NEB, Ref:E7645) in a total volume of 60 µl, followed by ligation using TEd-seq forked-adapters (final concentration of 0.8 µM), NEBNext Ultra II Ligation Master Mix, and NEBNext Ligation Enhancer. Adapter-ligated DNA was size selected using AMPure XP Beads (Beckman Coulter, ref:A63880) and used as template for selective amplification of TE containing fragments with NEBNext Ultra II Q5 Master Mix PCR and primers TEDseq_outer_for and TE_outer_rev (Additional file 1: Table S2). The following PCR conditions were used: 98 °C for 30 s, followed by 98 °C for 10 s, 61 °C for 75 s for 20 cycles with a last cycle of 5′ at 61 °C. PCR products were purified and size selected using AMPure XP Beads (0.9X). The DNA purified product was subjected to a second round of PCR with nested primers containing custom Illumina P5 and barcoded P7 adapters (P7_Primer_index#, P5_TE_primer, Additional file 1: Table S2). The PCR conditions were 98 °C for 30 s; 6 cycles of 98 °C for 10 s and 61 °C for 75 s; 10 cycles of 98 °C for 10 s and 72 °C for 75 s; and a final step at 72 °C for 5 min. Finally, the DNA library was purified and size selected using AMPure XP Beads (0.9X) and resuspended in 30 µl of TrisHcl 10 mM. Short-read sequencing was performed by BGI Tech Solutions (Hong Kong). Multiplexed TED-seq was conducted using equimolar concentrations of TE_outer_rev specific to each TE family and P5_TE_primers (see Additional file 1: Table S2). The dataset of previously validated non-reference TE in the TEA lines [18] can be found in Additional file 3.

TEd-seq protocol for long-reads

Fifty microliters of 20 ng/µl gDNA was sonicated using a Covaris sonicator (50 W, 2% duty factor, 200 cycles per burst, 3 min) to obtain 750–1500 bp fragments. Fragmented DNA was processed with the TEd-seq protocols described above, adjusting the AMPure XP Beads molarity rate (0.7X) to select fragments larger than 700 bp.

One microgram of pooled TEd-seq libraries was subjected to end repair and A-tailing using the NEBNext® FFPE DNA Library Prep Kit (NEB:E6650S). Nanopore adapter ligation and library construction was conducted using the Nanopore Ligation Sequencing Kit (ONT, SQK-LSK110) following manufacturer's instructions. Libraries were loaded on a R9.4.1 flow cell (ONT) according to the manufacturer's instructions. A standard 72-h sequencing was performed on a PromethION 2 Solo (ONT). TEd-seq ONT library was also conducted using the Nanopore Ligation Sequencing Kit (ONT, SQK-LSK114) and sequenced on a R10.4.1 flow cell (ONT). MinKNOW software (version 23.11.7) was used for sequence calling.

TEd-seq short-reads bioinformatic pipeline

Short-read libraries were demultiplexed by TE family, and TE sequences were trimmed from both ends using cutadapt (V2.6) [42] with options: -e 0.1, -O 25. PCR duplicates were removed using bbmap clumpify (V38.18) with options: -subs = 2, dedupe, -Xmx32g. Reads were mapped to the reference genome using bowtie2 (V2.3.5.1) [43] with parameters: –local, –very-sensitive. TE insertions were called using a custom

pipeline accessible in GitHub. Briefly, read coverage was estimated at each genome position; positions with more than 30 reads for *ATCOPIA93*, 10 reads for *ATENSPM3*, or 3 reads for VANDAL21 were considered as possible new insertions. Insertions overlapping reference TEs of the same family and in plastid genomes were excluded from the analysis. TE insertions detected in different samples and located less than 500 bp were considered to be shared.

TEd-seq long-reads bioinformatic pipeline

Raw reads were basecalled using guppy basecaller (V6.4.6 + ae70e8f) for R9 libraries and Dorado (V0.6.0) for R10 libraries. Reads were then trimmed and demultiplexed with cutadapt (V2.6) [42] using the barcoded custom P7 sequence and the P5 TE sequence with options: -e 0.2, -O 25. Reads containing multiple adapters were discarded. Trimmed and demultiplexed reads were mapped to the reference genome using minimap2 (V2.17-r941) [44]. TE insertions were called using a custom pipeline accessible in GitHub.

Allele frequency and transposition rate estimations

Three independent samples of 1000 seeds from each population and from generations 1 and 2 were grown for 10 days in 1/2MS medium. DNA was then extracted in bulk from all the seedlings of each sample. These three independent samples served as biological replicates to estimate allele frequency. For generation 0, due to a limited number of seeds, only 1000 seeds were sown. Multiplexed short-read TEd-seq was performed on each individual sample as described above. For generation 0, we created two independent short-read TEd-seq libraries from the same DNA extraction and treated them as technical replicates.

To estimate allele frequencies for shared and non-shared TE insertions, the number of reads whose pairs (second-in-pair) map at each insertion site were counted per sample and normalized by the average number of second-in-pair reads across reference TE insertions. For *ATCOPIA93*, we included four reference copies (genomic coordinates: Chr1:12,755,096–12760399, Chr3:8,674,540–8675098, Chr3:10,181,732–10182633, and Chr5:5,629,978–5,635,310). In the case of *ATENSPM3* and *VANDAL21*, we used three and one reference TE copies, respectively (genomic coordinates: Chr1:12,942,825–12,947,010, Chr2:4,385,856–4,394,311, and Chr2:4,900,803–4909281 for *ATENSPM3* and Chr2:10,001,255–10001765 for VANDAL21). All putative insertions detected in a 5-kbp window from the reference insertion sides and insertions with average allele frequencies lower than 2‰ in G0 and G1 were discarded. Using the observed allele frequencies from generation 0 (p), the expected frequencies were estimated assuming a mixing mating system with a selfing rate (s) of 0.95. Changes in allele frequency due to genetic drift were identified by comparing predicted to observed frequencies in each generation using a Z-test. Insertions with an adjusted p value lower than 0.05, following the same trend across generations (increasing from G0 to G1 and G1 to G2, or decreasing in the same manner), were categorized as changing in frequency. All shared transposable element insertions identified in the E&R experiment, along with their allele frequency estimates for each sample, are provided in Additional file 4.

To estimate the transposition rate, the reads for each TE family from the E&R were demultiplexed and downsampled to the same number of reads using Seqtk subseq

(V1.3-r106). TED-seq germline short-reads pipeline was then run. Windows of 1000 bp were opened from each TE insertion position, and each insertion was categorized as private or shared based on whether it was detected in a single sample or across multiple samples, respectively. Private insertions from each sample were considered for estimating the transposition rate. The pipeline used to estimate allele frequencies and to estimate transposition rates is accessible in GitHub https://github.com/LeanQ/TEd-seq-EandR-analyses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03512-x.

Additional file 1. Figures S1–S5, Tables S1 and S2.

Additional file 2. Comparison of different TE enrichment methods.

Additional file 3. Catalog of reference and validated non-reference TE insertions used in this study

Additional file 4. List of TE insertions identified in the E&R populations, along with their allele frequency estimates for each sample.

Acknowledgements
We thank members of the Quadrana group, Sayuri Tsukahara, and Tetsuji Kakutani for discussions and advice.

Peer review information
Robert Kofler and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
LQ conceived the project and wrote the manuscript. LQ and BL established the TEd-seq protocol and bioinformatic analysis, with additional input from PV. PV developed the multiplex and long-read TEd-seq, and performed population experiments and analyses. All authors interpreted the data, read, and approved the paper.

Funding
This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 948674 to LQ).

Data availability
Availability of data and materials: Original scripts are available on GitHub; https://github.com/LeanQ/TEd-seq-short-read, https://github.com/LeanQ/TEd-seq-long-read and https://github.com/LeanQ/TEd-seq-EandR-analyses and archived on Zenodo under the doi: https://doi.org/10.5281/zenodo.14645345 [45]. Sequencing data are available at the European Nucleotide Archive (ENA) under project PRJEB79758 (https://www.ebi.ac.uk/ena/browser/view/PRJEB79758) [46]. A detailed TEd-seq method is publicly available at protocols.io (dx.doi.org/https://doi.org/10.17504/protocols.io.dm6gp378pvzp/v1) [47].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Leandro Quadrana is a Guest Editor for Genome Biology but was not involved in the editorial process of this manuscript. The other authors declare that they have no competing interests.

Received: 11 September 2024 Accepted: 24 February 2025
Published online: 06 March 2025

References

1. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
2. Orgel LE, Crick FH, Sapienza C. Selfish DNA. Nature. 1980;288:645–6.
3. Kofler R, Gómez-Sánchez D, Schlötterer C. PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. Mol Biol Evol. 2016;33:2759–64.

4. Chen J, Wrightsman TR, Wessler SR, Stajich JE. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. PeerJ. 2017;5: e2942.
5. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27:1916–29.
6. Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. 2014;42:6826–38.
7. Baduel P, Quadrana L, Colot V. Efficient detection of transposable element insertion polymorphisms between genomes using short-read sequencing data. Methods Mol Biol. 2021;2250:157–69.
8. Gohl DM, Freifeld L, Silies M, Hwa JJ, Horowitz M, Clandinin TR. Large-scale mapping of transposable element insertion sites using digital encoding of sample identity. Genetics. 2014;196:615–23.
9. Stern DL. Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites. bioRxiv. 2017 [cited 2025 Jan 14]. p. 037762. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/037762v5.abstract
10. Ryu J, Chan W, Wettengel JM, Hanna CB, Burwitz BJ, Hennebold JD, et al. Rapid, accurate mapping of transgene integration in viable rhesus macaque embryos using enhanced-specificity tagmentation-assisted PCR. Mol Ther Methods Clin Dev. 2022;24:241–54.
11. Veeranagouda Y, Didier M. Transposon insertion site sequencing (TIS-Seq): an efficient and high-throughput method for determining transposon insertion site(s) and their relative abundances in a PiggyBac transposon mutant pool by next-generation sequencing. Curr Protoc Mol Biol. 2017;120:21.35.1–21.35.11.
12. Johannes F, Porcher E, Teixeira FK, Saliba-colombani V, Albuisson J, Heredia F, et al. Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. 2009;5: e1000530.
13. Mohamed M, Sabot F, Varoqui M, Mugat B, Audouin K, Pélisson A, et al. TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. Genome Biol. 2023;24:63.
14. Ivančić D, Mir-Pedrol J, Jaraba-Wallace J, Rafel N, Sanchez-Mejias A, Güell M. INSERT-seq enables high-resolution mapping of genomically integrated DNA using Nanopore sequencing. Genome Biol. 2022;23:1–13.
15. Vendrell-Mir P, Barteri F, Merenciano M, González J, Casacuberta JM, Castanera R. A benchmark of transposon insertion detection tools using real data. Mob DNA. 2019;10:53.
16. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond T a., De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011:479:534–7.
17. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. Elife. 2016;5. Available from: https://doi.org/10.7554/eLife.15716
18. Quadrana L, Silveira AB, Caillieux E, Colot V. Detection of transposable element insertions in Arabidopsis using sequence capture. Methods Mol Biol. 2021;2250:141–55.
19. Merkulov P, Gvaramiya S, Dudnikov M, Komakhin R, Omarov M, Kocheshkova A, et al. Cas9-targeted nanopore sequencing rapidly elucidates the transposition preferences and DNA methylation profiles of mobile elements in plants. J Integr Plant Biol. 2023;65:2242–61.
20. McDonald TL, Zhou W, Castro CP, Mumm C, Switzenberg JA, Mills RE, et al. Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nat Commun. 2021;12:3586.
21. Artesi M, Hahaut V, Cole B, Lambrechts L, Ashrafi F, Marçais A, et al. PCIP-seq: simultaneous sequencing of integrated viral genomes and their insertion sites with long reads. Genome Biol. 2021;22:97.
22. Uren AG, Mikkers H, Kool J, van der Weyden L, Lund AH, Wilson CH, et al. A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. Nat Protoc. 2009;4:789–98.
23. Philippe C, Vargas-Landin DB, Doucet AJ, Van Essen D, Vera-Otarola J, Kuciak M, et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife. 2016;5:1–30.
24. Dawes JC, Webster P, Iadarola B, Garcia-Diaz C, Dore M, Bolt BJ, et al. LUMI-PCR: an Illumina platform ligation-mediated PCR protocol for integration site cloning, provides molecular quantitation of integration sites. Mob DNA. 2020;11:7.
25. Steranka JP, Tang Z, Grivainis M, Huang CRL, Payer LM, Rego FOR, et al. Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome. Mob DNA. 2019;10:8.
26. Yephremov A, Saedler H. Technical advance: display and isolation of transposon-flanking sequences starting from genomic DNA or RNA. Plant J. 2000;21:495–505.
27. Arnold C, Hodgson IJ. Vectorette PCR: a novel approach to genomic walking. PCR Methods Appl. 1991;1:39–42.
28. Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA. An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. 1995;23:1087–8.
29. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, et al. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat Commun. 2019;10:3421.
30. Fiston-Lavier A-S, Carrigan M, Petrov D a., González J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. 2010:39:e36.
31. Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity. 2016;116:248.
32. Kofler R, Schlötterer C. A guide for the design of evolve and resequencing studies. Mol Biol Evol. 2014;31:474–83.
33. Abbott RJ, Gomes MF. Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. Heredity (Edinb). 1989:62:411–8.
34. Payne T, Johnson SD, Koltunow AM. KNUCKLES (KNU) encodes a C2H2 zinc-finger protein that regulates development of basal pattern elements of the Arabidopsis gynoecium. Development. 2004;131:3737–49.
35. Yasir M, Turner AK, Lott M, Rudder S, Baker D, Bastkowski S, et al. Long-read sequencing for identification of insertion sites in large transposon mutant libraries. Sci Rep. 2022;12:1–9.
36. Philippe C, Cristofari G. Genome-wide young L1 methylation profiling by bs-ATLAS-seq. Methods Mol Biol. 2023;2607:127–50.
37. Baduel P, Leduque B, Ignace A, Gy I, Gil J Jr, Loudet O, et al. Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. Genome Biol. 2021;22:138.

38. Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res. 2018;46:9524–36.
39. Huang H-Y, Zhang S, Choucha FA, Verdenaud M, Tan F-Q, Pichot C, et al. Harbinger transposon insertion in ethylene signaling gene leads to emergence of new sexual forms in cucurbits. Nat Commun. 2024;15:4877.
40. Tsukahara S, Bousios A, Perez-Roman E, Yamaguchi S, Leduque B, Nakano A, et al. Centrophilic retrotransposon integration via CENH3 chromatin in Arabidopsis. Nature. 2025; Available from: https://doi.org/10.1038/s41586-024-08319-7
41. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc. 2006;1:2320–5.
42. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011:17:10–2.
43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
44. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
45. Vendrell-Mir P, Leduque B, Quadrana L. Transposable element display sequencing (TEd-Seq) bioinformatic analysis. Zenodo. https://doi.org/10.5281/zenodo.14645345 (2025)
46. Vendrell-Mir P, Leduque B, Quadrana L. Transponsable element display sequencing. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB79758 (2025). Access Feb 20.
47. Vendrell-Mir P, Leduque B, Quadrana L. Protocol TE display sequencing (TED-Seq). protocols.io. https://doi.org/10.17504/protocols.io.dm6gp378pvzp/v1 (2024).

Publisher's Note