Detection of Transposable Element Insertions in Arabidopsis Using Sequence Capture

Leandro Quadrana, Amanda Bortolini Silveira, Erwann Caillieux, Vincent Colot

Detection of transposable element insertions in Arabidopsis using sequence capture

Leandro Quadrana[1]* Amanda Bortolini Silveira[1], Erwann Caillieux[1], Vincent Colot[1*]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Ecole Normale Supérieure, PSL Research University, 75005 Paris, France.

* correspondence to leandro.quadrana@ens.psl.eu and vincent.colot@ens.psl.eu

ABSTRACT

Transposable elements (TEs) are repetitive DNA sequences that have the ability to mobilize in the genome and create major effect mutations. Despite the importance of transposition as a source of genetic novelty, we still know little about the rate, landscape and consequences of transposition. This situation stems in large part from the repetitive nature of TEs, which complicates their analysis. Moreover, TE mobilization is typically rare and therefore new TE (i.e. non-reference) insertions tend to be missed in small-scale population studies. This chapter describes a TE-sequence capture approach designed to identify transposition events for most of the TE families that are potentially active in *Arabidopsis thaliana*. We show that our TE-sequence capture design provides an efficient means to detect with high sensitivity and specificity insertions that are present at a frequency as low as 1/1000 within a DNA sample.

INTRODUCTION

Transposable elements (TEs) are ubiquitous parasitic DNA sequences capable of self-propagating across genomes. Although designated by their discoverer Barbara McClintock in the 1940s as controlling elements because of their effect on genes, TEs have long been considered as mere selfish DNA. However, this situation has radically changed with the advent of genomics and TEs are now universally recognized as powerful engines of genome evolution. Nonetheless, despite the importance of TE mobilization in the creation of genetic novelty, the precise extent to which TE insertions can contribute to heritable phenotypic variation remains unknown. This lack of knowledge is largely due to the repetitive nature of TE sequences, which makes them 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. To circumvent the difficulties to detect TE insertions *de novo* using short-read sequencing technologies, we have designed a TE-sequence capture method that concentrates the sequencing power on DNA sequences that immediately flank TE insertions (Quadrana et al., 2016). Here, we provide a detailed description of our TE-sequence capture design, which is aimed at enabling a comprehensive and cost-effective assessment of TE mobilization in *A. thaliana* (Figure 1).

2 Materials

2.1. Design of TE-sequence capture probes

1. Reference genome
2. High quality TE annotation
3. List of target TE families (i.e. potentially mobile TEs)

2.2 DNA Extraction and library preparation

1. High quality DNA extracted using CTAB method
2. 1.5 mL plastic tubes.
3. KAPA HTP/LTP library preparation kit (Roche, Ref. 07961871001)
4. Freshly prepared Ethanol 80%

5. Elution buffer (10mM TrisHCl, pH 8.0). Prepare 100X Elution buffer stock by dissolving 121.14 g of Tris in 800 mL H2O. Adjust the pH of the solution to 8.0 using HCl. Make up the volume to 1 L with H2O, autoclave the solution and keep it at room temperature for up to 1 year. To obtain Elution buffer 1X dilute 1:100 the Elution Buffer 100X with PCR-grade water.
6. Agencourt AMPure XP (Beckman Coulter, Ref. A63880)
7. TruSeq DNA Single Indexes Set A (Illumina, Ref. 20015960)
8. NanoDrop Spectrophotometer (ThermoFisher, Ref. ND2000)
9. Qubit Fluorometer(ThermoFisher Ref. Q33238)
10. Qubit dsDNA HS Assay Kit (ThermoFisher Ref. Q32851)

2.3 TE-sequence capture, library amplification and sequencing

1. SeqCap EZ Prime Developer Probes (Roche Ref. 08247595001). This kit contains the SC Wash Buffers (tubes 1, 2 and 3), Stringent Wash Buffer (tube 4), 2x SC Hybridisation Buffer (tube 5), Hybridisation Component A (tube 6) and the Bead Wash Buffer (tube 7).
2. Plant Capture Enhancer (Roche)
3. 1.5 mL plastic tubes.
4. DNA LoBind Tubes 1,5 mL (Eppendorf, Ref. 0030108051)
5. 18 gauge x 1 1/2" Thin Wall Needle (Terumo, Ref. NN-1838R)
6. QIAquick PCR Purification kit (Qiagen, Ref. 28104)
7. Dynabeads M-270 Streptavidin (Invitrogen Ref. 65305)
8. Magnetic rack for 1.5 mL tubes.
9. KAPA HTP/LTP library preparation kit (Roche, Ref. 07961871001)
10. Agencourt AMPure XP (Beckman Coulter, Ref. A63880)
11. 2100 Bioanalyzer Instrument (Agilent, Ref. G2939BA). Other automated electrophoresis systems can also be used.
12. Bioanalyzer DNA1000 Chip (Agilent, Ref. 5067-1504)
13. SpeedVac (ThermoFisher, Ref. DNA130)

14. Oligos (see Table 1)

Oligo	**Sequence (5'-3')**
TS-PCR-1	AATGATACGGCGACCACCGAGA
TS-PCR-2	CAAGCAGAAGACGGCATACGAG
HE-Universal-oligo	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
HE-Index-2-oligos	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 1. List of oligos for TE-sequence capture. Note that the sequence for only one HE-Index-2-oligos is provided. HE-Index-oligos for other indexes can be obtained by modifying the underlined sequence.

2.4. Bioinformatic analysis

1. Multifasta file with complete reference genome sequence
2. Desktop Linux workstation with at least 4Gb of memory and 4 CPUs.
3. Fasta file with a collection of TE sequences targeted by TE sequence capture
4. Bowtie2 v2.3.2 (available at https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.5.1/)
5. Samtools v1.2.1 (available at http://www.htslib.org/download/)
6. bedtools v2.29.3 (available at https://bedtools.readthedocs.io/en/latest/index.html)
7. Picardtools (available at https://broadinstitute.github.io/picard/)

2.5. Estimating sensitivity

1. High quality DNA extracted from samples containing known TE insertions
2. List of known TE insertions
3. Same as 2.3 and 2.4

3 Methods

3.1. Design of TE-sequence capture probes

This section describes the procedure used to select TE sequences to design capture probes. This protocol is based on the SeqCap EZ Prime Developer Probes application (Roche) and could be implemented for any species for which a reference genome sequence with well annotated TEs is available. The reference genome sequence of *A.thaliana* (TAIR10) contains ~32,000, mostly degenerate TE copies, which belong to 326 distinct families, split in equal parts into class I retrotransposons, which are mobilized using a "copy and paste" mechanism and class II DNA transposons, which transpose through a "cut and paste" mechanism (Ahmed et al., 2011; The Arabidopsis Genome Initiative, 2000). So far, transposition activity has been documented experimentally for eleven TE families, mainly based on studies carried out in the reference accession Col-0 (Ito & Kakutani, 2014; Tsay et al., 1993). To determine the set of TE sequences mobile or potentially mobile in *A. thaliana* we combined the results of two complementary approaches:

- Bioinformatic analysis of consensus TE sequences

A comprehensive repository of TE consensus sequences from diverse eukaryotic organisms is available through the Repbase Update database (Jurka, 2000; Jurka et al., 2005). These consensus sequences are routinely used in alignmentment-based algorithms (such as RepeatMasker) to detect and mask TEs from genome sequences. Out of 326 consensus sequences identified in *A. thaliana*, 83 contain intact open reading frames, suggesting that they are still capable of mobilization. Representatives of all 83 TE families, 31 of which were also detected as potentially mobile in the population genomics approach (see below), were included in the TE-sequence capture design.

- Population genomics of TE insertion polymorphisms

We used short-read resequencing data publically available for 211 natural accessions (Schmitz et al., 2013; Schneeberger et al., 2011) to detect TE insertions polymorphisms in nature, which may be indicative of recent transposition. To this end we deployed our SPLITREADER pipeline (Quadrana et al., 2016), which searches among the set of Illumina short reads that cannot be completely mapped on the reference TAIR10

genome for so-called “split-reads” that contain TE extremities. Moreover, because most TE families generate short target site duplications (TSDs) of fixed size upon insertion, we looked for split-reads covering TE junctions that are absent from the reference genome and that produce, when mapped to the insertion site, a sequence overlap of the size of TSDs. An improved version of this bioinformatic approach can be found in Baduel et al. 2020. Overall, we identified non-reference TE insertions with TSDs for 131 TE families, of which 14 were not included in the TE-sequence capture design because of their very high copy numbers (>500 copies) in the reference *A. thaliana* genome, which would likely negatively impact the efficiency and evenness of the TE- sequence capture.

Based on these two complementary approaches, we considered a total of 318 reference TE copies belonging to 167 distinct families for TE-sequence capture. This list includes 75 families of LTR-retrotransposons, eight families of non-LTR retrotransposons, 65 families of DNA transposons and 19 families of helitrons (Supplementary Table 1). Probes covering the first and last 400bp of each of the 318 reference TE copies were chosen and produced by Roche using proprietary technologies. . Ultimately, these SeqCap Capture Oligos defined 634 capture targets, ranging from 53-322bp. Given the high sequence similarity between closely related TE families, TEs belonging to 181 TE families could in fact be captured with our design.

3.2 DNA Extraction and library preparation

This section describes the procedure to obtain high-quality genomic DNA and library preparation based on the CTAB method and KAPA HTP/LTP library preparation kit (Roche), respectively.

1. Extract DNA from plant material (1–2 g fresh weight, we use aerial parts of 3-week-old plants grown under long day conditions) with CTAB protocol (Allen et al., 2006). 1.2 µg of DNA is needed for this protocol (includes sonication test).

2. Quantify DNA and place 1.1 µg in a final volume of 55 µL (complete with PCR-grade water if necessary). Sonicate to obtain fragment sizes around 300bp. We use a Covaris sonicator with Peak Incident Power 140, Duty Factor 10%, Cycles per burst 200 and time 100''.
3. Run 100ng of sonicated and non-sonicated samples side by side in 1.5 % TAE gel. A smear should be visible between 200 and 600 bp.
4. Prepare WGS Library for illumina sequencing without performing dual size selection. This protocol was optimized for KAPA HTP/LTP library preparation kit (Roche), however other library kits (such as Illumina TruSeq or NEB Nebnext) should also perform well. To prepare one library, mix 8µl PCR-grade water, 7µl 10X KAPA End Repair Buffer and 5µl KAPA End Repair Enzyme. Add the 50µl of the fragmented DNA (from step 3.2.3). Mix and incubate at 20ºC for 30 minutes.
5. Add 120µl Agencourt AMPure XP and mix by pipetting up and down several times. Incubate the tube at room temperature for 10 minutes.
6. Place the tube on a magnet until the liquid is clear and discard the supernatant. Wash the beads with 200µl ethanol 80% twice. Allow the beads to dry at room temperature and continue immediately to the next step.
7. Remove the tubes from the magnet and add 42µl of water, 5µl 10X KAPA A-Tailing Buffer and 3µl KAPA A-tailing Enzyme. Mix by pipetting up and down. Incubate at 30ºC for 30 minutes.
8. Add 90µl PEG/NaCL SPRI solution and mix thoroughly by pipetting up and down until getting an homogenous solution. Incubate at room temperature for 10 minutes.
9. Repeat step 6.
10. Remove the tubes from the magnet and add 30µl of water, 10µl 5X KAPA Ligation Buffer, 5µl KAPA T4 DNA Ligase and 5µl of 15mM of TruSeq DNA Single Indexes adapter. Mix by pipetting up and down. Incubate at 20ºC for 15 minutes.
11. Add 50µl PEG/NaCL SPRI solution and mix thoroughly by pipetting up and down until getting an homogenous solution. Incubate at room temperature for 10 minutes.
12. Repeat step 6.

13. Remove the tubes from the magnet, resuspend the beads in 30µl of Elution Buffer (10mM Tris-Hcl pH 8.0) and incubate at room temperature for 2 minutes. Add 50µl PEG/NaCL SPRI solution and mix thoroughly by pipetting up and down until getting an homogenous solution. Incubate at room temperature for 10 minutes.
14. Repeat step 6.
15. Remove the tubes from the magnet, resuspend the beads in 25µl of Elution Buffer (10mM Tris-Hcl pH 8.0) and incubate at room temperature for 2 minutes. Place the tubes in the magnet and incubate until the solution is clear. Transfer the supernatant to a new tube and proceed with the pre-capture library amplification.
16. Mix 25µl of KAPA HiFi HotStart Ready Mix, 5µl of 100µM TS-PCR Oligo 1/2 (100 µM) and 20µl of Sample Library (from step 20) in a PCR tube. Amplify in thermocycler using the following program.

    Step 1: 45 seconds at 98ºC

    Step 2: 15 seconds at 98ºC

    Step 3: 30 seconds at 60ºC

    Step 4: 35 seconds at 72ºC

    Step 5: Repeat steps 2-4 six times.

    Step 6: 1 minute at 72ºC
17. Clean up amplified Library using a PCR purification kit, such as QIAquick PCR Purification kit (QIAGEN), and elute the sample in 50 µl of PCR grade water.
18. Quantify the concentration of the pre-captured amplified library using Nanodrop. If more than one sample will be pooled for multiplexing, quantify using Qubit and Qubit dsDNA HS Assay Kit according to the manufacturer indications.

3.3 TE-sequence capture, library amplification and sequencing

This section describes the procedure to capture TE sequences from a standard Illumina library prepared with the Kappa HTP/LTP kit (Figure 2). The TE sequence capture protocol presented here is based on a customized dual capture SeqCap EZ Library (Roche NimbleGen).

1. In case of multiplexing, mix together equal mass of each amplified DNA library to obtain a total combined mass of 1.25 µg. Up to 24 samples can be multiplexed.
2. Mix together equal amounts of Hybridization Enhancer (HE) oligos so the resulting Multiplexing HE oligo pool contains 1,000 pmol of HE-index-oligos (1 µM). Add 1,000 pmol of HE-Universal-oligo (1 µM) and vortex for two seconds.
3. Add 10 µl of Plant Capture Enhancer (Roche) in a new 1.5ml tube.
4. Add 1 µg of DNA sample library mix from step 1 to the 1.5 ml tube containing Plant Capture Enhancer. Add the 2,000 pmol of Multiplex HE oligo pool prepared in 2. Close the tube's lid and make a hole in the tube's cap with a 18 gauge needle.
5. Dry the mix in a vacuum concentrator (SpeedVac) on high heat.
6. Once the sample is dry, cover the hole with a sticker. Add 7.5 µl 2X SC Hybridization Buffer and 3 µl SC Hybridization Component A (SeqCap EZ Prime Developer Probes). Vortex the sample for 10 seconds and centrifuge at maximum speed for 10 seconds.
7. Incubate the tube at 95ºC for 10 minutes. Centrifuge the samples at maximum speed for 10 seconds.
8. Mix the sample with an aliquot of SeqCap Capture Oligos (SeqCap EZ Prime Developer Probes) in a PCR tube. Add 2.25 µl of PCR-grade water. Vortex for three seconds and spin down.
9. Incubate in a thermocycler at 47ºC (with heated lid turned on at 57ºC) for 72 hours.
10. Prepare the wash buffer solutions as indicated below:

    1X SWB. Mix 40 µl Stringent Wash Buffer (SWB) and 360 µl PCR-grade water. Incubate at 47ºC

    1X WBI 47C. Mix 10 µl of Wash Buffer (WB) I and 90 µl of PCR-grade water. Incubate at 47ºC

    1X WBI RT. Mix 20 µl of Wash Buffer (WB) I and 180 µl of PCR-grade water.

    1X WBII. Mix 20 µl of Wash Buffer (WB) II and 180 µl of PCR-grade water.

    1X WBIII. Mix 20 µl of Wash Buffer (WB) III and 180 µl of PCR-grade water.

1X BWB. Mix 200 µl Bead Wash Buffer (BWB) and 300 µl of PCR-grade water.

11. Allow the Dynabeads M-270 Streptavidin to warm at room temperature for 30 minutes. Vortex the beads for 15 seconds and aliquot 100 µl in a 1.5 ml DNA LoBind tube.
12. Place the tube in a magnet and once clear, remove the supernatant. Add 200 µl of 1X BWB. Remove the tube from the magnet and vortex for 10 seconds.
13. Repeat step 12.
14. Place the tube in a magnet rack and once clear, remove the supernatant. Remove the tube from the magnet and resuspend the beads in 100 µl of 1X BWB. Transfer the solution to a new PCR tube.
15. Place the tube in a magnet and once clear, remove the supernatant.
16. Transfer the hybridized samples (from step 9) to the tube containing the capture beads.
17. Mix by pipetting up and down. Incubate the tube at 47ºC for 45 minutes. Vortex every 15 minutes to ensure the beads remain in suspension.
18. After 45 minutes of incubation, add 100 µl 1X WBI 47ºC. Mix by vortexing 10 seconds and transfer all to a new 1.5 ml tube. Place the tube in a magnet and once clear remove the supernatant.
19. Remove the tube from the magnet and add 200 µl of 1X SWB. Mix by pipetting up and down ten times. Incubate at 47ºC for five minutes.
20. Place the tube in a magnet and once clear, remove the supernatant. Repeat step 19.
21. Place the tube in a magnet and once clear, remove the supernatant. Remove the tube from the magnet and add 200 µl of 1X WBI RT. Mix by vortexing for two minutes.
22. Place the tube in a magnet and once clear, remove the supernatant. Remove the tube from the magnet and add 200 µl of 1X WBII. Mix by vortexing for one minute.
23. Place the tube in a magnet and once clear, remove the supernatant. Remove the tube from the magnet and add 200 µl of 1X WBIII. Mix by vortexing for 30 seconds.
24. Place the tube in a magnet and once clear, remove the supernatant. Remove the tube from the magnet and add 50 µl of PCR-grade and proceed with the middle-captured library amplification

25. Mix 50µl of KAPA HiFi HotStart Ready Mix, 10µl of 100µM TS-PCR Oligo 1/2 (100 µM) and 50µl of Sample Library (from step 29) into two PCR tubes. Amplify in thermocycler using the following program.

    Step 1: 45 seconds at 98ºC

    Step 2: 15 seconds at 98ºC

    Step 3: 30 seconds at 60ºC

    Step 4: 35 seconds at 72ºC

    Step 5: Repeat steps 2-4 four times.

    Step 6: 1 minute at 72ºC

26. Combine the two PCR reactions and clean up using a PCR purification kit, such as QIAquick PCR Purification kit (QIAGEN), and elute the sample in 50 µl of PCR grade water.
27. Repeat capture hybridization (steps 2-8)
28. Incubate in a thermocycler at 47ºC (with heated lid turned on at 57C) for 12 hours.
29. Repeat capture washing (steps 10-24).
30. Remove the tube from the magnet and add 50 µl of PCR-grade water and proceed with the oost-captured library amplification
31. Mix 50µl of KAPA HiFi HotStart Ready Mix, 10µl of 100µM TS-PCR Oligo 1/2 (100 µM) and 50µl of Sample Library (from step 29) into two PCR tubes. Amplify in thermocycler using the following program.

    Step 1: 45 seconds at 98ºC

    Step 2: 15 seconds at 98ºC

    Step 3: 30 seconds at 60ºC

    Step 4: 35 seconds at 72ºC

    Step 5: Repeat steps 2-4 13 times.

    Step 6: 1 minute at 72ºC

32. Combine the two PCR reactions and clean up using a PCR purification kit, such as QIAquick PCR Purification kit (QIAGEN), and elute the sample in 50 µl of PCR grade water.
33. Quantify the library using NanoDrop. The library yield should be >500ng.
34. Run 1µl post capture library on a Bioanalyzer DNA 1000 chip. A successfully constructed library should have an average fragment size between 150 and 500bp.
35. Evaluate capture efficiency by qPCR pre-capture (step 3.2.23) and post-capture (3.3.37) libraries using primers spanning capture targets as well as non-captured sequences. Capture efficiency (CE) may be calculated using the following formula:

$$CE = \frac{eff^{\Delta Ct}_{target}}{\sqrt[f]{\prod_0^f eff^{\Delta Ct, non-target_i}_{non-target_i}}}$$

where CE indicates the Capture efficiency, *eff* is the amplicon efficiency, *ΔCt* is the Ct difference between the post-capture and pre-capture samples, *target* and *non-target* indicate the capture and non-capture amplicons, respectively, and f represents the number of non-capture amplicons analyzed. A successful captured library has >500-fold enrichment.

Sequence 200pM of pooled library with 20% PhiX on Illumina sequencer, >76 cycles, paired-end reads. Note that 20 million paired-end reads per library is sufficient for most applications.

3.4. Bioinformatic analysis

The first step to identify non-reference TE insertions based on short-read sequencing data obtained by TE-sequence capture is to detect informative reads partially mapping on the TE sequences covered by the probes. A multifasta file containing the complete set of captured TEs must be generated using the coordinates of full-length TE sequences annotated in the reference genome and the FastaFromBed command from Bedtools. This fasta file is used to builds a Bowtie2 index as detailed below:

```
FastaFromBed -i captured_TE_sequences.bed -g TAIR10.fa  > captured_TE_sequences.fa
Bowtie2-build captured_TE_sequences.fa captured_TE_sequences
```

Detection of reads mapping partially over TE sequences.

After mapping short-reads on the collection of TE sequences, collect those reads mapping partially (with at least 20bp soft-clipped at 5' or 3' read's end).

```
###mapping pair-ends using bowtie2 in local mode (--local) and reporting as much as four equally possible mapping positions (-k 4).
bowtie2 -x captured_TE_sequences -1 in_1.fastq -2 in_2.fastq -S in.sam --local --very-sensitive -k 4
samtools view -H in.sam > in-TE.sam
###collecting pair ends of reads with at least 20nt softclipped at 5' or 3' read's end.
samtools view -F 4 -S $in.sam | awk '$6~/^[2-8][0-9]S/ || $6~/[2-8][0-9]S$/ || $6~/^1[0-9][0-9]S$/ || $6~/1[0-9][0-9]S$/ {print $0}' >> in-TE.sam
### collecting unmapped reads with the paired read mapped
samtools view -f 4 -F 8 -S in.sam | awk '{print $0}' >> in-TE.sam
samtools view -f 8 -F 4 -S in.sam | awk '{print $0}' >> in-TE.sam
### convert sam to bam
samtools view -Sb in-TE.sam > in-TE.bam
```

Identification of potential novel TE insertion sites

```
## extracting split-clipped mapping on specific TEs (i.e. ATCOPIA93)
samtools view -F 4 in-TE.bam | grep ATCOPIA93 | awk '$6~/^[2-8][0-9]S/ || $6~/[2-8][0-9]S$/ || $6~/^1[0-9][0-9]S/ || $6~/1[0-9][0-9]S$/ {print $1"\t"$10"\t"$11}' | sort -k1,1 -k2,2 -u | awk '{print "@"$1"\n"$2"\n+\n"$3}' > in-ATCOPIA93-split.fastq

## extracting discordant-pair reads mapping on specific TEs (i.e. ATCOPIA93)
samtools view -F 8 in-TE.bam | grep ATCOPIA93 | awk '{print $1}' | sort -u >> reads.name.disc

samtools view -f 64 -u in-TE.bam > in-TE-disc-first.bam=
samtools view -f 128 -u in-TE.bam > in-TE-disc-second.bam

java -Xmx10g -XX:+UseSerialGC -Djava.io.tmpdir=./ -cp /usr/share/java/picard.jar net.sf.picard.sam.FilterSamReads INPUT=in-TE-disc-first.bam FILTER=includeReadList READ_LIST_FILE=reads.name.disc OUTPUT=in-ATCOPIA93-selected-disc-first.sam TMP_DIR=./
```

```
java -Xmx10g -XX:+UseSerialGC -Djava.io.tmpdir=./ -cp /usr/share/java/picard.jar net.sf.picard.sam.FilterSamReads INPUT=in-TE-disc-second.bam FILTER=includeReadList READ_LIST_FILE=reads.name.disc OUTPUT=in-ATCOPIA93-selected-disc-second.sam TMP_DIR=./

cat in-ATCOPIA93-selected-disc-first.sam | awk '$1!~/^@/ {print $1"\t"$10"\t"$11}' | sort -u -k1,1 -k2,2 | awk '{print "@"$1"|1\n"$2"\n+\n"$3}' > in-ATCOPIA93-disc.fastq

cat in-ATCOPIA93-selected-disc-second.sam | awk '$1!~/^@/ {print $1"\t"$10"\t"$11}' | sort -u -k1,1 -k2,2 | awk '{print "@"$1"|2\n"$2"\n+\n"$3}' >> in-ATCOPIA93-disc.fastq

##Mapping split- and discordant-reads on reference genome
bowtie2 -x TAIR10_index -U in-ATCOPIA93-split.fastq, in-ATCOPIA93-disc.fastq -S in-ATCOPIA93-local.sam --local --very-sensitive
samtools view in-ATCOPIA93-local.sam > in-ATCOPIA93-local.bam

##Identification of putative insertion sites
BamToBed -i in-ATCOPIA93-local.bam -split > ATCOPIA93_bamtobed.bed
mergeBed -i ATCOPIA93_bamtobed.bed -c 4 -o count_distinct -d 500 | awk '$4>1 && $3-$2<3000' > ATCOPIA93_putative_insertions.bed
```

3.5. Estimating sensitivity

This section describes how we calculated the sensitivity of our TE sequence capture approach using DNA extracted from plants with validated novel TE insertions.

In order to estimate empirically the sensitivity of our TE-sequence capture approach, we took advantage of a population of Arabidopsis TE accumulation (TEA) lines previously characterized using Illumina mate-pair libraries (Quadrana et al., 2019). The large physical distance (~5 kb) between mate-pair reads was used to determine the complete sequence of the insertions with high accuracy. We selected eight TEA lines carrying non-reference TE insertions for the LTR-retroelement *ATCOPIA93* and we extracted DNA from four weeks-old plants as described in 3.2. Each DNA sample was serially diluted (from 1:50 to 1:6400) and

then pooled together, leading to a differential representation of each DNA sample. The resulting pooled DNA was processed as indicated in 3.2 and 3.3. Data analysis was performed as in 3.4. The set of non-reference *ATCOPIA93* insertions detected by TE-sequence capture was compared to the validated set of insertions previously identified using mate-pair WGS (Quadrana et al., 2019). TE-sequence capture identified 95% of *ATCOPIA93* non-reference insertions for dilutions above 1:800, although the sensitivity of our TE-sequence capture drops <30% for dilutions below 1:800. Moreover, the number of reads supporting each *ATCOPIA93* insertion strongly correlates with the dilution factor (Figure 2). Thus, he number of supporting reads is a reliable estimator of the TE insertion representation and can therefore be used to identify rare transposition events and to determine with reasonable accuracy the allele frequency of polymorphic TE insertions within populations of up to 1000 individuals.

4 Conclusions

We have described a comprehensive protocol for the detection of TE insertions in Arabidopsis using a custom-designed TE-sequence capture approach. This protocol uniquely combines all necessary steps from choosing target TE sequences, designing probes, performing sequence capture and analysing the short-read sequencing data. Thanks to the high sensitivity of TE-sequence capture, it offers a cost-effective option to investigate transposition as well as TE insertion polymorphisms (TIPs). Indeed, we have previously applied this protocol for the study of TIPs in a large panel of Arabidopsis natural accessions (Quadrana et al., 2016) as well as of transposition in Arabidopsis TEA lines (Quadrana et al., 2019). As the sensitivity of TE-sequence capture is determined by the number of target sequences and sequencing depth, this approach can readily be implemented for the study of transposition in species with large genomes, such as maize and wheat. Finally, because of the increasing interest in understanding the role of TEs in the generation of large effect mutations, we predict that TE-sequence capture will be a key component of genomic tool-kits in the future.

Notes

1- To prevent damage of oligos due to multiple freeze/thaw cycles, after resuspending the oligos in PCR-grade water they should be aliquoted into small volumes (20µl) and stored at -20°C.

Acknowledgements

We thank members of the Colot group for discussions, specially P. Baduel for critical reading of the manuscript. Support was from the Agence National de la Recherche (ANR-09-BLAN-0237, the Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, ANR-11-IDEX-0001-02 PSL* Research University to V.C) and the Centre National de la Recherche Scientifique (MOMENTUM program, to L.Q.).

REFERENCES

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., & Quesneville, H. (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. In *Nucleic Acids Research* (Vol. 39, Issue 16, pp. 6919–6931). https://doi.org/10.1093/nar/gkr324

Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S., & Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature Protocols*, *1*(5), 2320–2325.

Baduel, P. Quadrana, L. Colot, V. (2020) Efficient detection of transposable element insertion polymorphisms between genomes using short-read sequencing data. *Methods in Molecular Biology*

Ito, H., & Kakutani, T. (2014). Control of transposable elements in Arabidopsis thaliana. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *22*(2), 217–223.

Jurka, J. (2000). Repbase Update: a database and an electronic journal of repetitive elements. In *Trends in Genetics* (Vol. 16, Issue 9, pp. 418–420). https://doi.org/10.1016/s0168-9525(00)02093-x

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. In *Cytogenetic and Genome Research* (Vol.

110, Issues 1-4, pp. 462–467). https://doi.org/10.1159/000084979

Quadrana, L., Bortolini Silveira, A., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., & Colot, V. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. *eLife*, *5*. https://doi.org/10.7554/eLife.15716

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini Silveira, A., Engelen, S., Baillet, V., Wincker, P., Aury, J.-M., & Colot, V. (2019). Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nature Communications*, *10*(1), 3421.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., Schork, N. J., & Ecker, J. R. (2013). Patterns of population epigenomic diversity. *Nature*, *495*(7440), 193–198.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., Smith, L. M., Cao, J., Fitz, J., Warthmann, N., Henz, S. R., Huson, D. H., & Weigel, D. (2011). Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10249–10254.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.

Tsay, Y. F., Frank, M. J., Page, T., Dean, C., & Crawford, N. M. (1993). Identification of a mobile endogenous transposon in Arabidopsis thaliana. *Science*, *260*(5106), 342–344.

FIGURES

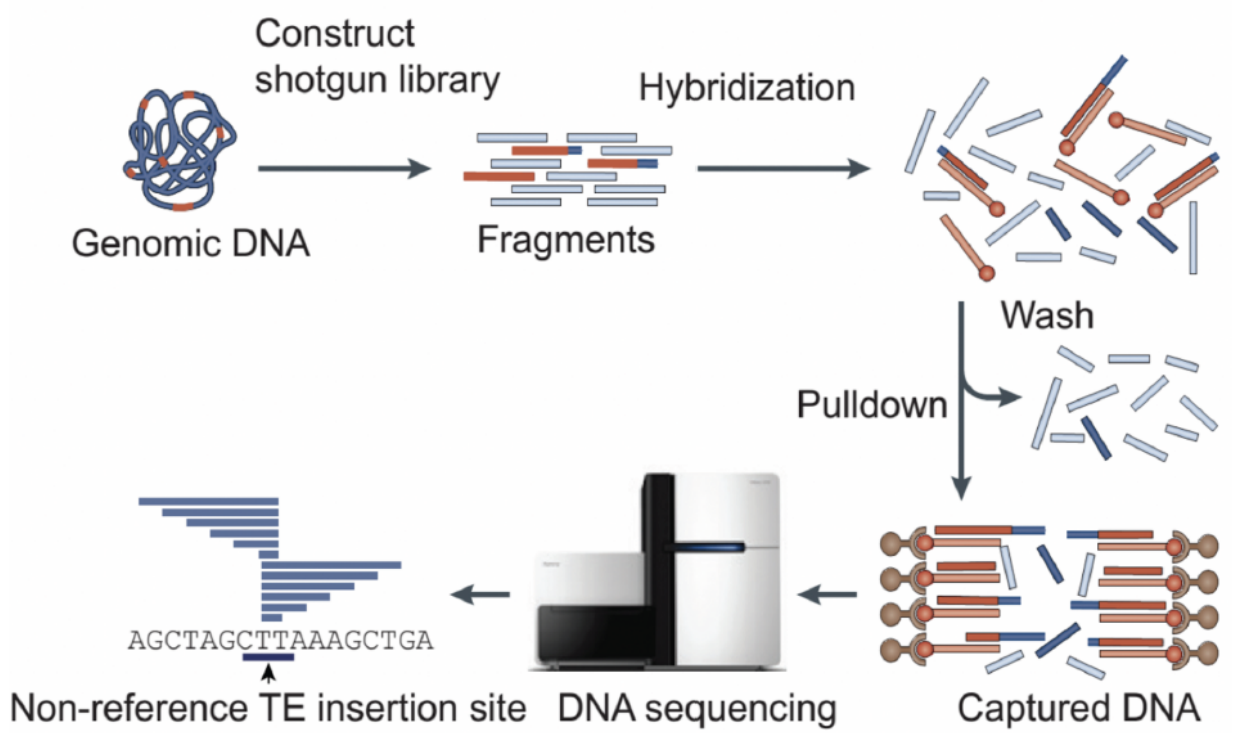

Figure 1. Description of the TE-capture workflow. Target TE sequences are indicated in red. After sequencing, reads are analyzed to detect clusters of reads indicating the presence of non-reference TE insertions.

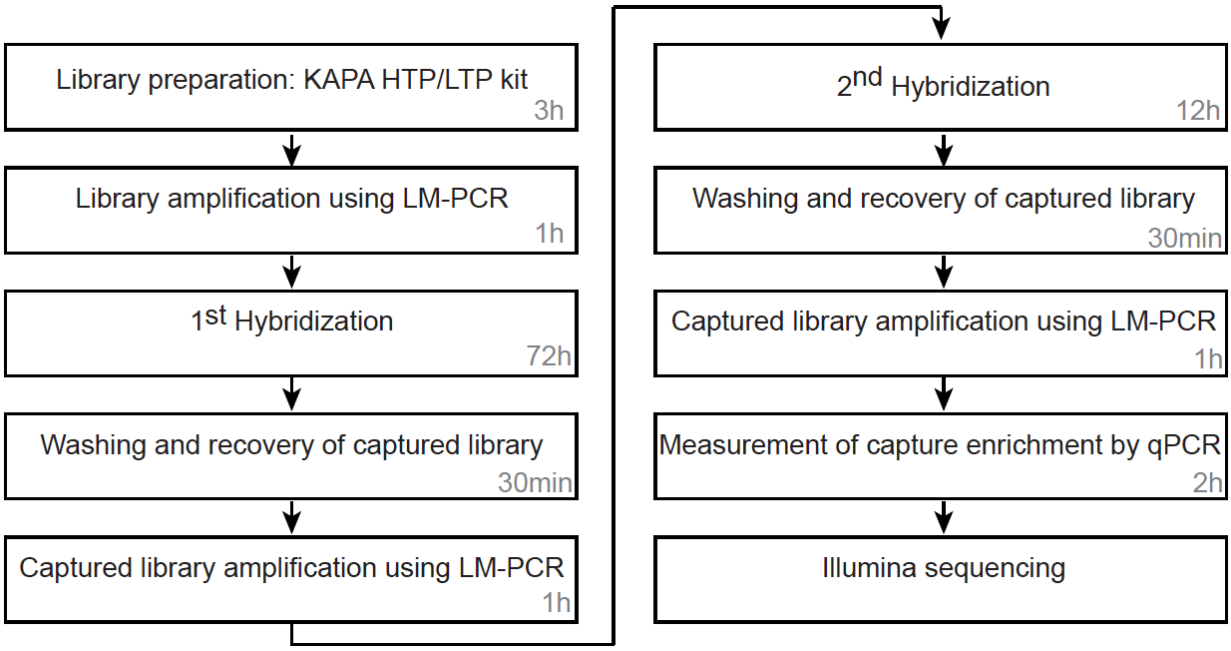

Figure 2. TE Sequence capture workflow.

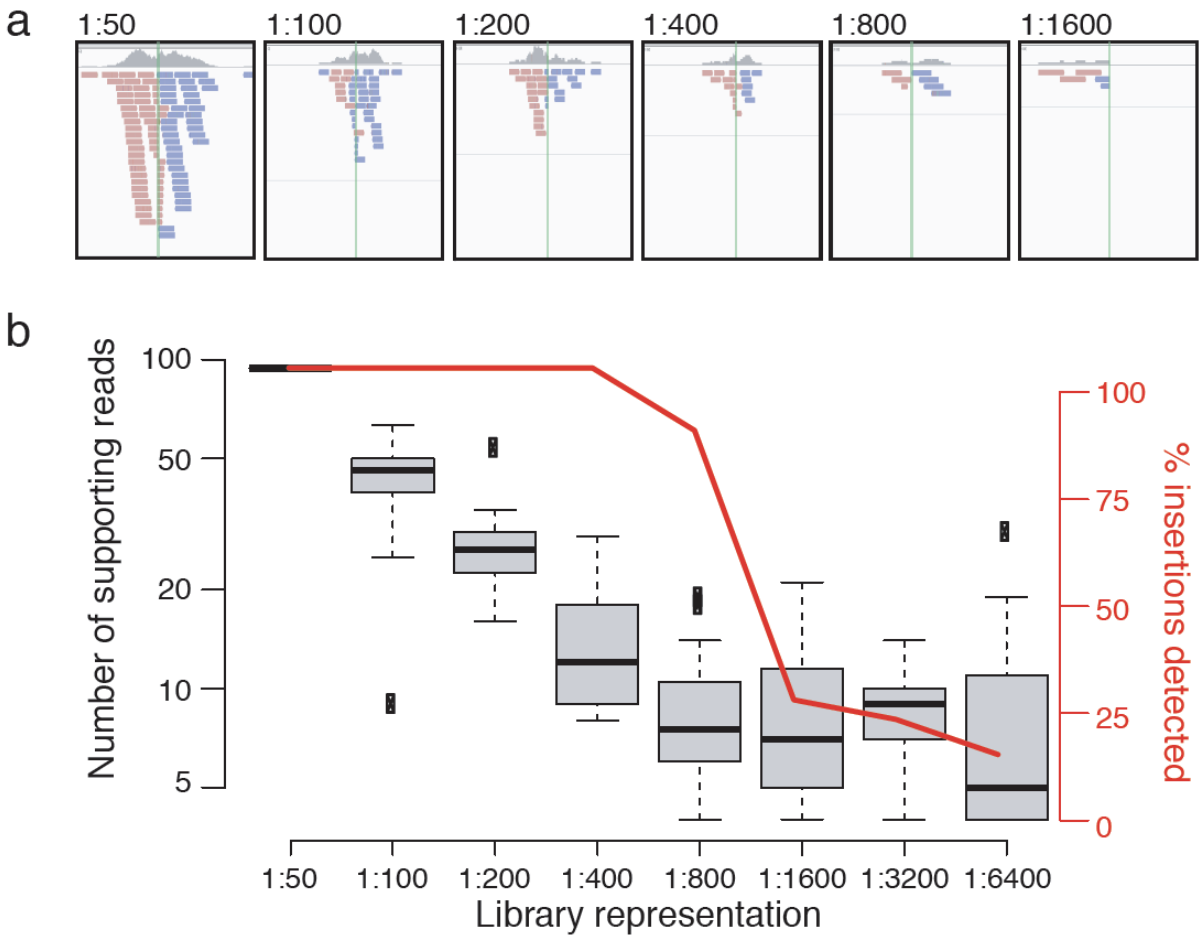

Figure 3. Sensitivity of TE sequence capture. a) Examples of TE insertions detected at different dilution factors. **b**) Number of supporting reads per insertion site as well as percentage of TE insertions detected.