Etude des conséquences d'une perte de méthylation de l'ADN sur la mobilisation des éléments transposables

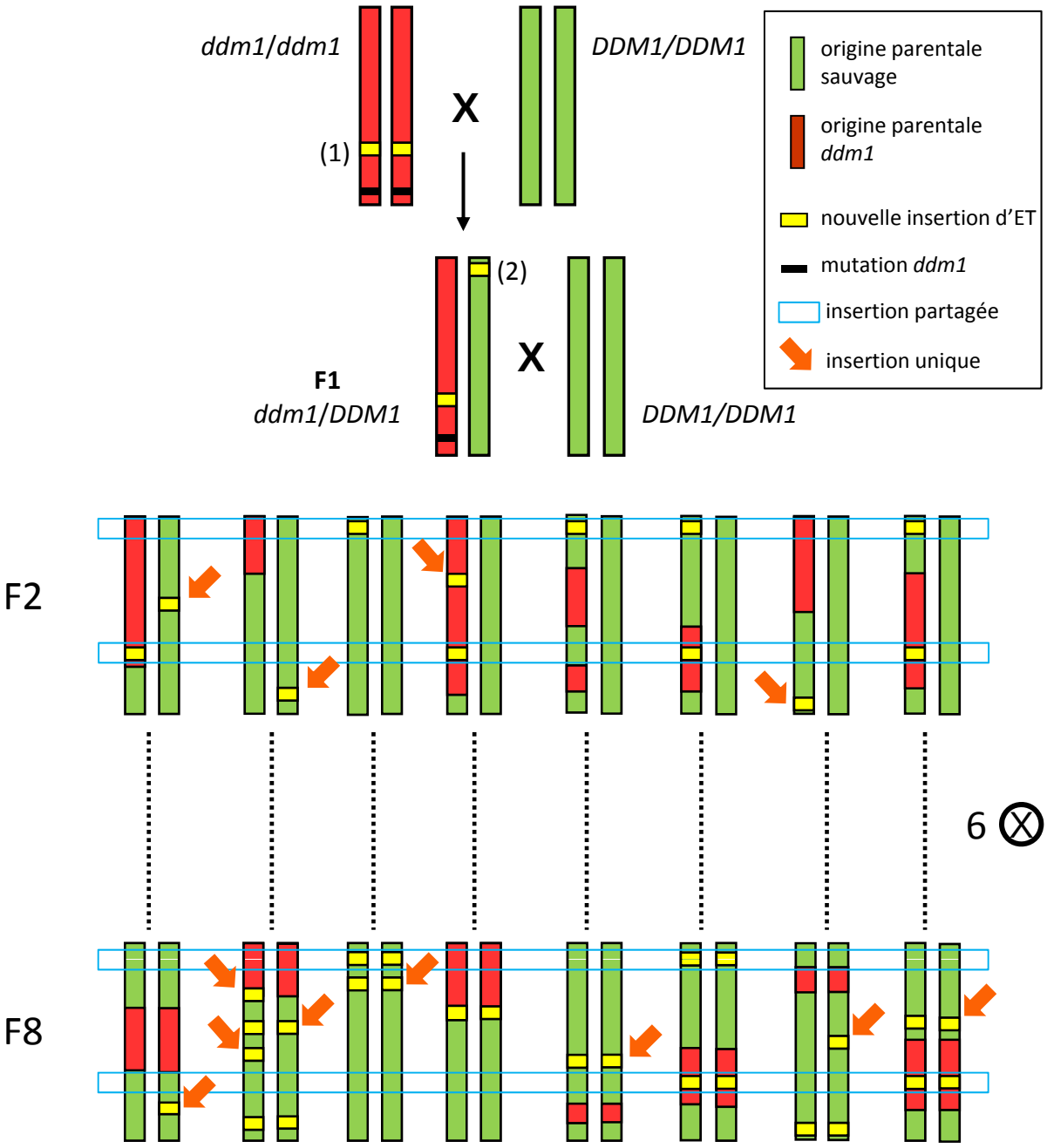


Figure 2.1 : Evènements de mobilisation potentiellement détectables dans les epiRIL.
La perte de méthylation induite dans le mutant *ddm1* cause la mobilisation de certains ET (1). Cette hypométhylation étant transmise à la descendance (du moins à certains locus), on peut s'attendre à de nouveaux évènements de transposition dans celle-ci. Alors que les nouvelles insertions produites dans le parent *ddm1* ou la F1 (1 et 2) vont ségréger dans la population d'epiRIL et vont donc être présentes dans plusieurs lignées (insertions partagées), celles produites à partir de la F2 seront propres à chaque lignée (insertions uniques). En F8, chaque lignée présente donc un jeu de nouvelles insertions, composé d'insertions partagées et d'insertions uniques.

Comme décrit dans l'introduction, une perte drastique de la méthylation de l'ADN, telle qu'induite par une mutation dans le gène *DDM1*, engendre une réactivation transcriptionnelle massive des ET. Cela dit, il n'est pas établi dans quelle mesure cette réactivation transcriptionnelle des ET se traduit par leur mobilisation. Afin de répondre à cette question ainsi que pour déterminer le profil d'insertion des ET, nous avons procédé au séquençage du génome de 53 epiRIL ainsi que des deux lignées parentales.

2.1 Identification des évènements de transposition ayant eu lieu lors de la génération des epiRIL

La population d'epiRIL est issue d'un croisement entre un parent sauvage et le mutant *ddm1-2* suivi d'un rétrocroisement avec le parent sauvage à la suite duquel seuls les individus homozygotes pour l'allèle sauvage du gène *DDM1* ont été sélectionnés et autofécondés sur 6 générations (fig. 2.1). Compte tenu du schéma de croisement, chaque point du génome est en moyenne d'origine sauvage dans 75% des lignées et hérité du parent *ddm1* dans 25% les lignées, sauf bien sûr à proximité du locus *DDM1*, systématiquement d'origine sauvage. Etant donné que l'hypométhylation induite par la mutation *ddm1* est transmise de façon stable pour de nombreux locus, les epiRIL présentent des profils de méthylation contrastés (fig. 2.1).

Afin d'identifier les évènements de transposition ayant eu lieu dans le mutant *ddm1* ou lors de la production des epiRIL (fig. 2.1) nous avons, en partenariat avec le Génoscope, réalisé le séquençage Illumina « paired-end » de banques « mate-pair » du génome de plus d'une cinquantaine de ces lignées ainsi que de deux individus sauvages et d'un mutant *ddm1* cousins des parents utilisés pour générer la population d'epiRIL.

Le séquençage paired-end, contrairement au séquençage de lectures uniques permet d'identifier des variations structurales en se basant sur la détection de paires dites discordantes, à savoir de paires dont les deux lectures ne sont pas positionnées à la bonne distance l'une de l'autre ou mal orientées par rapport au génome de référence. Le choix de banques mate-pair permet d'obtenir des lectures appariées situées physiquement à une

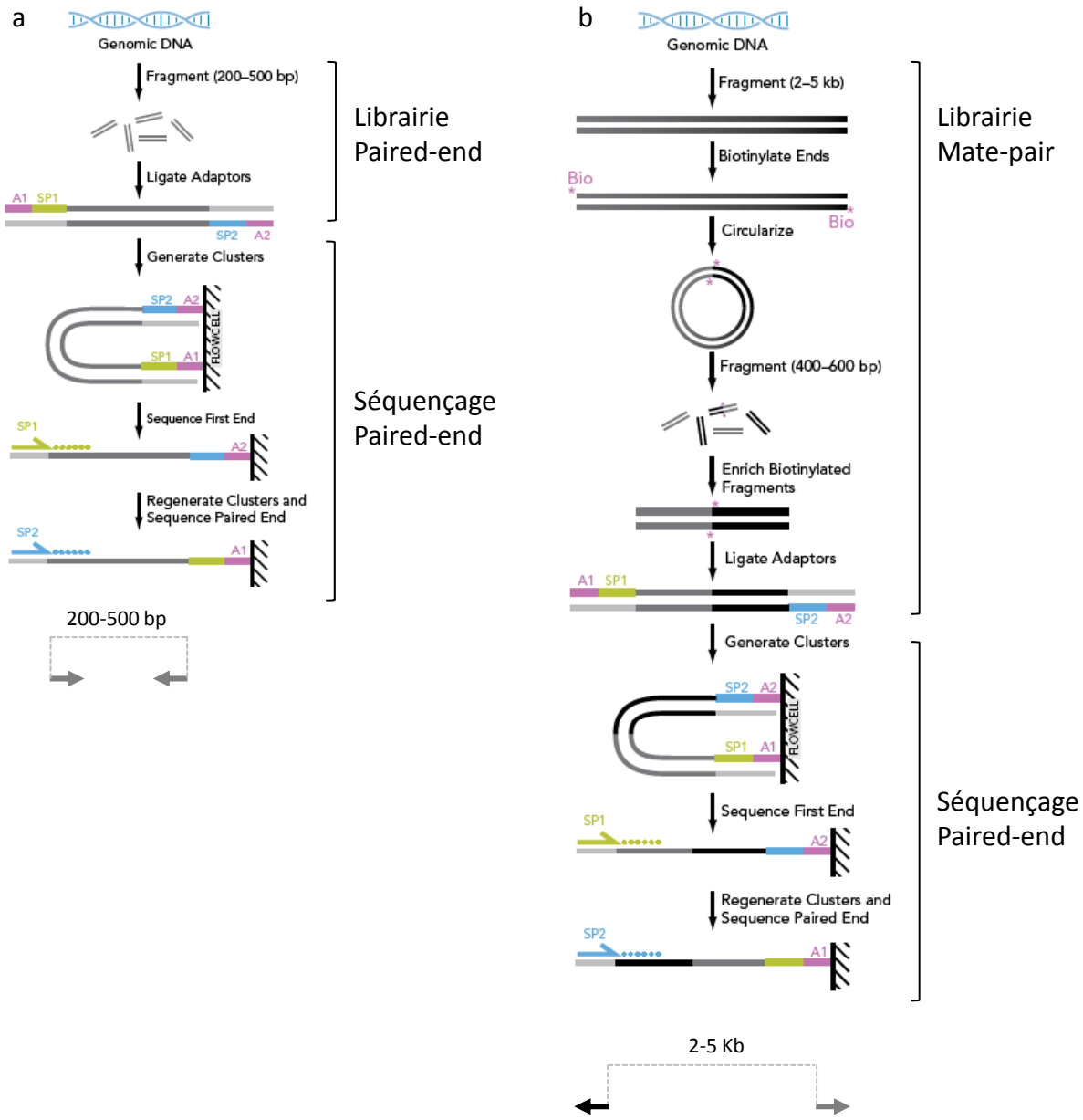


Figure 2.2 : Protocole de séquençage paired-end de banques paired-end (a) vs banques mate-pair (b).

grande distance l'une de l'autre (jusqu'à plusieurs kb), contrairement aux banques paired-end classiques qui couvrent de petites distances (500 bp maximum) (fig. 2.2). Le choix de banques mate-pair présente deux avantages majeurs : (i) elle permet d'augmenter le nombre de lectures qui soutiennent une variation structurale donnée et donc l'exhaustivité et la robustesse de la détection de ces évènements ; (ii) elle produit une plus grande couverture horizontale des variations structurales ce qui permet d'avoir des informations sur la quasi-totalité de la séquence insérée et notamment sa partie interne. Cette propriété est particulièrement importante dans le cadre de la détection de nouvelles insertions d'ET car l'une des difficultés majeures est l'identification précise du locus donneur. Avec l'approche que nous avons choisie, même un très faible niveau de polymorphisme, où qu'il soit localisé dans la séquence de l'ET, est suffisant pour discriminer entre plusieurs donneurs.

Afin d'optimiser la détection des nouvelles insertions d'ET basée sur des données de séquençage de banques mate-pair, j'ai participé à l'élaboration de TE-Tracker, un programme spécifiquement dédié à cette problématique. Ce travail est présenté dans un manuscrit en cours de finalisation, reproduit ci-après.

TE-Tracker: systematic identification of transposition events through whole-genome resequencing

Arthur Gilly[1,*], Mathilde Etcheverry[4,5,6,*], Mohammed-Amin Madoui[1,*], Julie Guy[1], Antoine Martin[4,5,6], Tony Heitkam[4], Karine Labadie[1], Jérémie Le Pen[4,5,6,†], Patrick Wincker[1,2,3], Vincent Colot[4,5,6,‡] and Jean-Marc Aury[1,‡]

[1]Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France
[2]Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France
[3]Université d'Evry, UMR 8030, CP5706, Evry, France
[4]Institut de Biologie de l'Ecole Normale Supérieure, F-75230 Paris Cedex 05, France
[5]Centre National de la Recherche Scientifique (CNRS), UMR 8197, F-75230 Paris Cedex 05, France
[6]Institut national de la santé et de la recherche médicale (INSERM) U1024, F-75230 Paris Cedex 05, France

[*]These authors contributed equally
[†]Current address: Gurdon Institute and Department of Biochemistry, University of Cambridge, The Henry Wellcome Building of Cancer and Developmental Biology, Tennis Court Rd, Cambridge CB2 1QN, UK.
[‡]Corresponding authors

Abstract

The recent explosion in genome sequencing has confirmed that transposable elements (TEs) and their relics constitute the major fraction of genomic DNA in most known eukaryotes, including mammals and many plants (Lopez-Flores and Garrido-Ramos, 2010). Furthermore, resequencing experiments have provided compelling evidence that while TE mobilization is strongly repressed in most instances, this repression is not absolute and can be allievated in both pathological and normal situations (Huang et al., 2012).In particular, mobilization of specific TEs has been linked to diseases, including cancer, as well as to the normal development of the brain in humans (Lee et al., 2012; Hancks and Kazazian, 2012; Muotri et al., 2010). Until now, the identification of TE mobilization events using next generation sequencing (NGS) technologies has often relied on the use of general-purpose structural variant (SV) detection algorithms, which produce very large numbers of putative breakpoints and leave the calling step to the biologist. Here, we present a computational method for accurately detecting both the identity and destination of newly mobilized TEs in re-sequenced genomes. This method, called TE-Tracker, accepts as input the widely used BAM format, and does not rely on prior TE annotation but solely on the extraction of specific association patterns within paired-end alignments. We show that noise levels are orders of magnitude lower when using TE-Tracker instead of generic SV detection algorithms. We test TE-Tracker on a synthetic human genome and use it to provide a comprehensive view of transposition events induced by loss of DNA methylation in Arabidopsis.

Introduction

TEs and their abundant relics have been found in nearly all organisms and have been classified into several families based on sequence features and transposition mechanisms (Lopez-Flores and Garrido-Ramos, 2010). So-called DNA-transposons generally exhibit cut-and-paste transposition, while retrotransposons use an ARN intermediate and thus transpose using a copy-and-paste mechanism. Retro-elements are further divided into two subclasses, depending on the presence or absence of Long Terminal Repeats (LTR). The biological role of TEs has been the subject of great controversy, and although they had been assimilated to "selfish" or "junk" DNA for some time (Doolittle and Sapienza, 1980), they are now recognized as important factors in the evolution of genome structure and function (Hurst and Werren, 2001; Rebollo et al., 2012).

Indeed, it has been estimated that mobilization of LTR-retrotransposons is responsible for up to one tenth of spontaneous germline mutations (Maksakova et al., 2006) in laboratory mice. Similarly, mobilization of the human LINE1 (L1) non-LTR retrotransposon was found to account for 19% of the structural variation between individual genomes (Kidd et al., 2010), and has been linked to over a hundred human diseases (Hancks and Kazazian, 2012). TEs can also be mobilized in somatic cells and although this has been associated with pathological situations, such as cancer in humans (Lee et al., 2012; Iskow et al., 2010), there is now growing evidence that transposition does occur during normal development and cellular differentiation, notably in the brain (Thomas et al., 2012). Despite these observations, TE mobilization

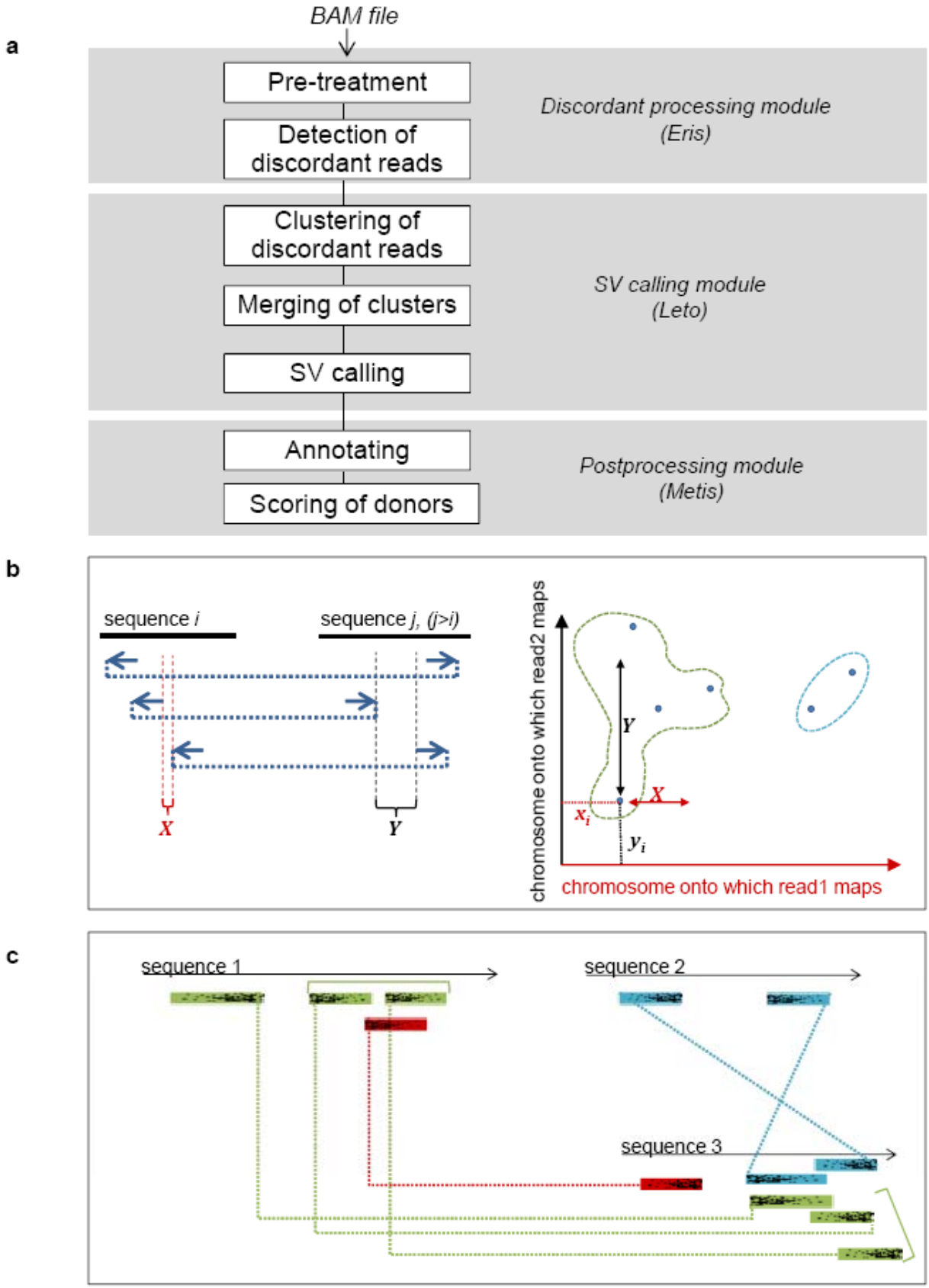


Figure 1 : TE-Tracker overview and main algorithms.

a. Main steps of the TE-Tracker pipeline. **b**. Principle of the single linkage clustering. Left: genome view of three anomalously spaced read pairs (blue arrows and lines). Discordant reads are ordered by increasing order on position chromosome and onset. Read pairs that are distant by no more than X on their ordered side and by no more than Y on their unordered one are clustered together. Right: graphical representation of the Single-Linkage Clustering (SLC) algorithm implemented in TE-Tracker. **c**. Cluster merging and calling. The bracketed green clusters overlap on sequence 3 and are close to each other on sequence 1 and will therefore be merged. However, if the red cluster is also present all green clusters will be rejected. Considering the set of blue and green clusters independently leads to multiple donors or acceptors.

remains a rare event, notably because of mechanisms, such as DNA methylation, that keep TEs in a transcriptionnally inactive form most of the time (Slotkin and Martienssen, 2007; Zamudio and Bourc'his, 2010). Additional control mechanisms must also exist, as transcriptional reactivation of TEs does not invariably lead to their mobilization. Yet, the extent to which TE mobilization is controlled post-transcriptionally is largely unknown.

With the advent of NGS technologies, it is now conceivable to resequence whole genomes in order to characterize TE mobilization systematically. However, this task is complicated by the inherently repeated, most often defective nature of TE sequences and by their frequent clustering. Perhaps as a result, few bioinformatic tools have been developed specifically for the detection of newly mobilized TEs in resequenced genomes and these tools often rely on prior annotation (Nellaker et al., 2012), or are restricted to the analysis of specific TE sequences (Lerat, 2010). Nonetheless, several studies have attempted to identify new TE insertions de novo using structural variant (SV) detection tools. Four broad types of such methods have been described over the past few years. They are based on the analysis of either (i) depth of coverage, (ii) split reads, or (iii) discordant paired reads, or else (iv) on de novo assembly (Alkan et al., 2011). Type (i) methods give a quantitative measure of the number of extra TE copies but do not provide information about their location. Type (ii) and (iii) methods identify one-sided events in the form of clusters of anomalously mapped reads, but they do not combine these one-sided events to produce bona fide TE insertions. In addition, the heavy computational burden of type (iv) methods as well as their poor performance with repeated sequences preclude their use for large-scale detection of new TE insertions (Li et al., 2011).
More recently, several programs have attempted to adopt an integrative approach by combining results from several methods (Jiang et al., 2012; Rausch et al., 2012; Sindi et al., 2012), but their signal-to-noise ratio is still typically low when considering specific types of structural variation (See Methods).Finally, none of these general-purpose tools can identify the donor TE or provide the internal sequence of transposed copies.

Here, we present TE-tracker, a new method dedicated to the systematic and robust identification of newly mobilized TEs based on whole-genome resequencing (Figure 1.a). Since TE-Tracker performs pairwise association of read mate groups that are anomalously spaced or oriented, discordant reads can be anchored on both sides of the recipient sequence until their mates overlap over the donor. Not only does this allow to detect new TE insertions de novo and without prior annotation, it also makes it possible to determine both the length and orientation of the newly inserted sequences, as well as the identity of the corresponding donor TEs.
We have used TE-Tracker to analyse the resequenced genome of Arabidopsis in the progeny of a mutant plant compromised for DNA methylation. Our findings indicate that despite extensive transcriptional reactivation of numerous TEs in the parental plant, only few of

these are mobilized, thus providing strong evidence of the importance of post-transcriptional mechanisms in preventing TE mobilization.

Results

The TE-Tracker pipeline

As with type (iii) SV detection methods (see Introduction), TE-Tracker starts with the analysis of the orientation and spacing of clustered (Figure 1.b) discordant paired reads to detect new insertions (Medvedev et al., 2009). Unlike these methods though, TE-Tracker aims to cover the entire length of the new TE insertion by searching at the new location for left and right-side clusters that overlap with each other (Figure 1.c). As a result, TE-Tracker can detect insertions and determine the donor copy for TEs that are up to 2L in length, where L is the mean size of DNA fragments used for sequencing. For example, in order to fully characterize the transposition landscape of Alu elements in the human genome ($\approx$ 300 bp), TE-Tracker would require a short fragment paired-end library of 150 bp mean length, whereas longer, recircularized fragments (such as mate-pairs) would have to be used for larger elements.

When more than two clusters satisfy the overlap condition at the same location, TE-Tracker attempts to consider them together, within some limits (Figure 1.c and see Methods), which allows orienting insertion events with respect to the donor TE sequence. The clustering algorithm has been optimized for speed and TE-Tracker was found to run between one and four order of magnitudes faster than three common pipelines (Zeitouni et al., 2010; Chen et al., 2009; Quinlan et al., 2010) that accept long fragment sequencing data as input (Supplementary Figure 1).

Typically, TE families contain mostly defective copies that are unable to be mobilized because of truncations or other mutations in their coding or regulatory sequences. Nonetheless, their potentially mobile copies are difficult to predict on the basis of sequence integrity alone, and there are no programs to date that attempt to identify those that transpose among potential candidates. Given that TE families may contain several mobile copies that differ from each other by a few sequence polymorphisms, we have included in TE-Tracker a donor-scoring feature, which selects within clusters those reads that contain discriminating polymorphisms (See Methods).

All of the features described above are implemented into a pipeline comprising three main programs (Figure 1.a), which allow pre- as well as post-processing steps such as filtering of the input alignment or annotation of the output result to be also performed (see Methods).

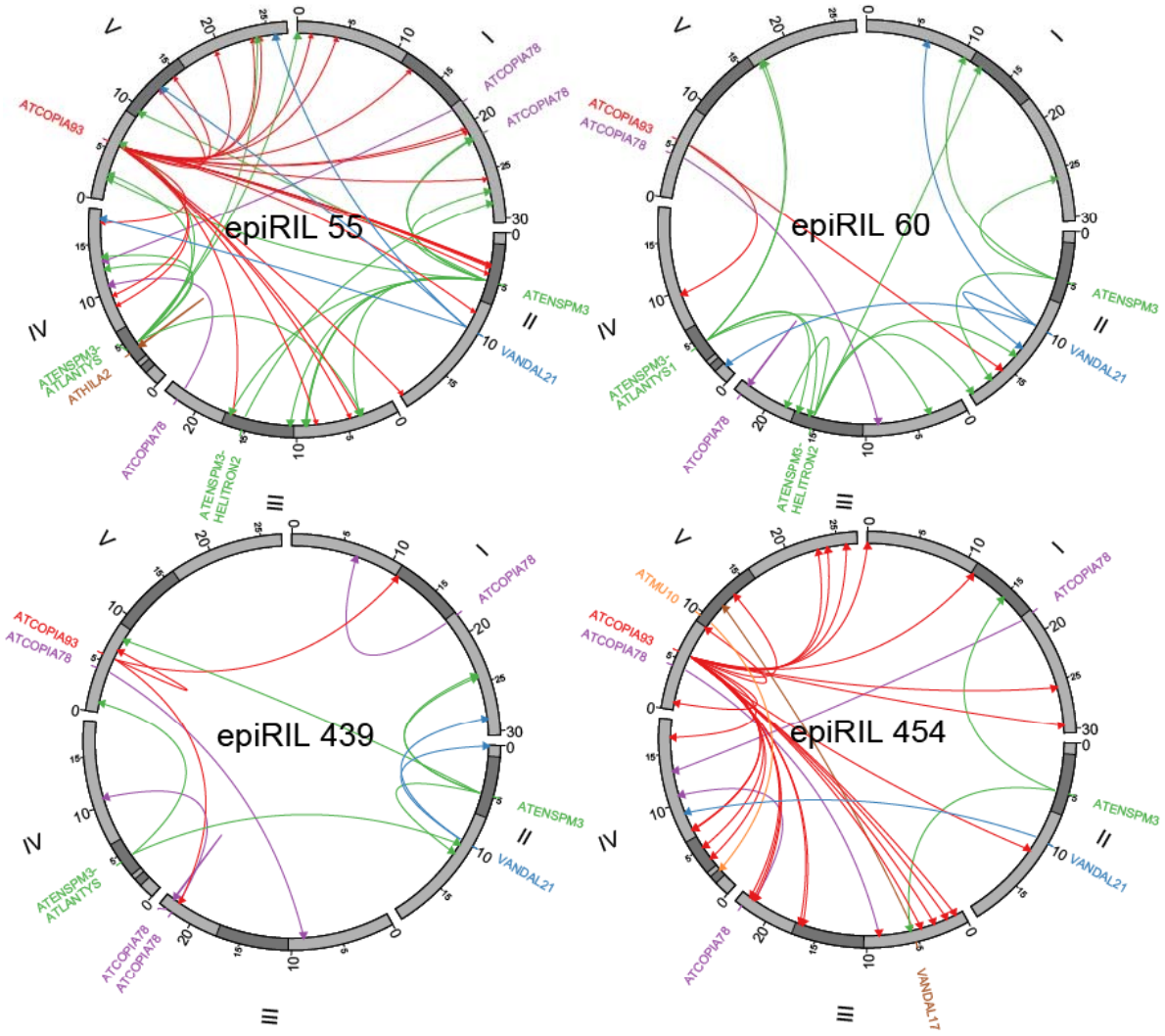


Figure 2 : Circos representation of new TE insertion events detected in four epiRILs . Exterior circle represents the five chromosomes of Arabidopsis with pericentromeric regions and heterochromatic knob on chromosome 4 in dark grey. Arrows link donor TEs with the new insertion sites. Only events mapped with no ambiguity (no multiple acceptor sites and no similarity with events detected in wt) are represented.

Assessment of TE-Tracker performance on simulated Arabidopsis and human chromosomes

We first tested TE-Tracker by simulating a mate-pair sequencing strategy for the complete Arabidopsis reference genome using an in-house paired read simulator (see Methods). Importantly, although the simulator took into account realistic rates and types of sequencing errors as well as of chimeric reads, this did not lead to any transposition event being detected in this un-rearranged version of the Arabidopsis genome, which suggests a null type-II error rate. When 80 random insertions involving two copies of the retro-element COPIA93 were simulated between a pair of Arabidopsis chromosomes and sequenced using the same simulator, 75 were detected, which gives a sensitivity of 95% (Supplementary Table 1). The remaining five insertions were located either in tandem or other nearby ($<$ 50 kb) repeat elements, or in sequence gaps.

We also performed similar tests on two human chromosomes, for which we simulated the mobilization of two, 6 kb-long L1-type elements that differ by 124 nucleotides and that have been described as active in the human brain (Baillie et al., 2011). Of the 20 random insertions generated (with random donor), 17 were detected (Supplementary Table 2), and the three remaining ones were inserted in sequence gaps. Furthermore, only one L1 donor was misattributed in this set, despite there being about one hundred distinct, potentially mobile full length L1 on these two chromosomes.

Application of TE-tracker to the exploration of the transposition landscape in Arabidopsis

We then applied TE-Tracker to the identification of novel TE insertions in a set of four Arabidopsis epigenetic recombinant inbred lines (Johannes et al., 2009) derived from a cross between a wild type (*wt*) plant and a plant mutated for the *DECREASE IN DNA METHYLATION 1* (*DDM1*) gene. *ddm1* mutant plants in which DNA methylation as well as transcriptional silencing of TEs is severely compromised (Vongs et al., 1993), inducing their transcriptional reactivation and thus potentially leading to their re-mobilization (Lippman et al., 2004; Singer et al., 2001; Miura et al., 2001; Tsukahara et al., 2009). The four epiRILs together with two *wt* lines were sequenced using Illumina mate-pair libraries (around 5.5 kb mean length), which should therefore enable the detection of new insertions for almost all of the TEs that are potentially active in the genome, as over 90% of all full-length annotated Arabidopsis TEs are less than 11 kb long (Buisine et al., 2008; Ahmed et al., 2011). Effective sequencing coverage (after alignment) ranged from 17X to 37X (Supplementary Table 3). Results are illustrated in Figure 2 and summarized in Supplementary Tables 4-9. TE-Tracker reported a total of 220 distinct insertions among the four epiRILs, of which 162 matched annotated TE sequences (Supplementary Tables 4-9, column "Donor annotation"). The vast

majority (136) of these insertions were not detected in the two *wt* parental lines, as expected if most transposition events occurred in the *ddm1* parental line or during the propagation of the epiRILs. To validate these results, a random set of 69 potentially novel insertions as well as three insertions also shared with the two *wt* parents were tested by PCR. In all 72 cases, the presence of the insertion could be confirmed (Supplementary Table 10), which provides further evidence of the high specificity of TE-Tracker. Furthermore, sequencing of 26 PCR products corresponding to new insertions was used to evaluate the performance of TE-tracker in identifying donor TEs. In all but one case, the donor scoring module was able to identify the correct TE donor sequence. Also, sequencing of both ends of 12 new insertions confirmed the presence of a target site duplication in each case, as expected for true transposition events (Supplementary Figure 2, Supplementary Table 10). Among these, we validated several insertions involving composite sequences that were not previously annotated as full-length TE units, namely fragments of ATENSPM3 and ATLANTYS1 and fragments of ATENSPM3 and HELITRON2 (Supplementary Figure 3).

Of the 136 distinct novel TE insertions identified, six were shared among the four epiRILs (Supplementary Tables 4-9). This proportion is lower than that expected (exact one-sided binomial test, p-value= 2.137e-11) if all insertions had occurred in the *ddm1* parental line used to establish the epiRIL population (Johannes et al., 2009)and suggests therefore that TE mobilization could also occur in subsequent generations in at least some cases. Furthermore, transposition in the epiRILs concerns only a small number of TE families (Supplementary Tables 4-9), which is consistent with a previous report of TE mobilization in *ddm1* (Tsukahara et al., 2009). These findings, together with the fact that most TE sequences are transcriptionally reactivated in *ddm1* (Lippman et al., 2004), suggest therefore an important role of postranscriptional mechanisms in preventing TE mobilization in Arabidopsis. Our analysis indicates in addition that mobilization, when it occurs, often concerns only one of the potentially mobile TE members of a given family. For instance, despite there being two highly similar copies of the LTR retroelement family COPIA93, only one appears to be mobile in the genome of the Columbia accession, as was previously reported (Mirouze et al., 2009). However, there are exceptions to this rule, as exemplified by the fact that several members of the LTR retroelement family COPIA78, which is closely related to COPIA93, have been mobilized. As many of these new COPIA78 insertions are shared among at least two of the epiRILs, transposition is likely to have taken place in most cases in the parental *ddm1* line or in the F1, which contradicts a previous claim that COPIA78 cannot transpose in this mutant background (Ito et al., 2011). Furthermore, in the case of COPIA78 insertions, the donor scoring feature often yielded two potential donors with similar high scores. Detailed analysis of one such COPIA78 insertion showed the existence of distinct sequential blocks corresponding to either donor. This is in agreement with previous reports indicating that similarly to what is seen in viruses (Goodrich and Duesberg, 1990), two RNA intermediates

matching distinct LTR-TE family members could be encapsidated together. As a result, TE sequences could undergo recombination by template switching during cDNA synthesis (Jordan and McDonald, 1998), thus leading to the insertion of a composite sequence presenting block-wise similarity to both of the parent elements (Supplementary Figure 4). However, we cannot rule out in the case of the COPIA78 element that the new insertion corresponds in fact to to the mobilization of still non-assembled COPIA78 donor elements. Resequencing also revealed that the DNA transposon VANDAL21 inserts preferentially close to the transcription start site of genes and in the same orientation as these (9 /12 instances and 8 /9 instances, respectively), likely indicating a role for transcription initiation of the target locus in the insertion of this particular TE family. Finally, we note that overall, new TE insertions are spread across the entire genome (Figure 2), which contrasts with the pericentromeric localization of most TE sequences present in the Arabidopsis genome. This provide therefore strong evidence that purifying selection plays an important role in eliminating insertions that occur within the gene-rich regions of Arabidopsis chromosomes.

Discussion

We have presented here a program, TE-Tracker, that accurately detects both the source and destination of novel transposition events in resequenced genomes. Since TE-Tracker only relies on the detection and clustering of discordant paired reads and not on TE annotation, it is generic and enables to track any mobilized TE, irrespective of its identity. With appropriate parameterization, TE-Tracker is also able to detect all potential donor sequences for a given insertion, and by discriminating reads that map better to a particular donor, it can attribute the correct one among them if they differ by at least one nucleotide. Furthermore, TE-Tracker is faster and produces significantly less noise than common SV detection programs (see Methods), therefore allowing the researcher to focus exclusively and exhaustively on TE mobilization events in a re-sequenced genome.

Clustering algorithms such as TE-Tracker are generally memory-intensive when run over a large number of points; in particular, it is known that the optimal performance of the single-linkage algorithm used in TE-Tracker is $\mathcal{O}(n^2)$, where n the number of points (Sibson, 1973). In an omics context, this will result in increased computational load proportional to the number of discordant reads, either because of larger genomes or higher sequencing depth. For TE-Tracker, we choose to favor speed at the expense of memory use. For performance optimization, we developed a seed-type heuristic that reduces the amount of pairs in memory to a fraction of the total number (see Methods). Furthermore, at any given time, read mate mappings that belong to different pairs of chromosomes and are mapped in a specific orientation are considered independently and sequentially, which implies that performance of TE-Tracker will not depend on overall genome size/sequencing depth but on the average

sequence size/sequencing depth for individual chromosomes. Hence, discordant reads are subdivided in up to $4 \times \left(\!\!\binom{k}{2}\!\!\right)$ chunks where k is the number of chromosomes. This is why performance evaluation for a pair of two human chromosomes can be considered to reflect performance over the human genome as a whole. For an organism with very large chromosomes, it would be possible to reduce concerns for memory load by adding a further optimization step, for example one that would perform two-dimensional Fast Fournier Transform for spatial density estimation of read pair mappings. This would further scale down the size of the individual proximity matrices considered by the program, however during our tests, we never encountered a case where such a procedure was necessary.

Acceptor sites are generally defined by TE-Tracker within a few kb, depending on the size of the mobilized TE but in no case can be less than a few dozens of bp, because of split reads at the precise breakpoint, which are often left unmapped by a general-purpose aligner such as BWA. These "blind spots" can nevertheless be overcome by performing an additional split-aligning step within the insertion interval defined by TE-Tracker (e.g. using BWA's Smith-Waterman implementation) and searching for appropriate motifs in read sequences. This search algorithm was not included in the standard distribution of TE-Tracker due to its low (combinatorial) performance, but could easily be added so as to match the base-pair precision provided by integrative structural variant detection tools such as PRISM.

To conclude, we show that TE-tracker, when combined with a relevant sequencing protocol, provides an efficient and reliable method for detecting new transposition events in re-sequenced genomes. Indeed TE-Tracker stands in marked contrast to general-purpose structural variant detection tools, which typically generated lists of thousands of clusters and/or breakpoints, which the biologist needs to sieve through. Furthermore, none of these tools provide information on donor TE sequences. We believe that the development of biologically constrained programs (BCPs) such as TE-Tracker, which aim at providing robust answers to specific questions, will greatly help overcome the noise issues that typically plague the analysis of short read sequencing data.

Methods

Sequencing

DNA was extracted from seedlings grown under long-day conditions, using DNeasy Qiagen kits. About 10 microgram of genomic DNA were sonicated separately to a 4-6 Kb size range using the E210 covaris instrument (Covaris, Inc., USA). Libraries were prepared following Illumina's protocol (Illumina Mate Pair library kit). Briefly, fragments were end-repaired and biotin labeled. A size selection of fragments with length of interest (around 5Kb) was performed. DNA were then circularized and linear, non-circularized DNA were eliminated by digestion. Circularized DNA were fragmented to 300-700-bp size range using covaris E210. Biotinylated DNA were purified, end-repaired, then 3'-adenylated, and Illumina adapters were added. DNA fragments were PCR-amplified using Illumina adapter-specific primers. Finally, the PCR amplified libraries (350-650 bp) were size-selected. Libraries were then quantified using a Qubit Fluorometer (Life technologies) and libraries profiles were evaluated using an Agilent 2100 bioanalyzer (Agilent Technologies, USA). Each library was sequenced using 100 base-length read chemistry in a paired-end flow cell on the Illumina GAIIx (2 lanes) or HiSeq2000 (1 lane) (Illumina, USA).

Read Mapping

Reads were then mapped with BWA v. 0.6.1 (Li and Durbin, 2009), using the parameters `-R 10000 -l 35 -O 11`, onto the TAIR10 reference sequence (Lamesch et al., 2012). Reads hanging over chromosome ends were removed using `picard CleanSam`, duplicate pairs were removed using `picard MarkDuplicates` (http://picard.sourceforge.net). Supplementary Table 3 displays mean vertical coverage, as well as the portion of the genome covered by less than 10 reads, in percent, for each of the lines.

TEtracker algorithm

The `TE-Tracker` algorithm comprises three Perl modules. The three modules are named `eris`, `leto` and `metis` and deal respectively with preprocessing, clustering of discordant pairs and postprocessing (annotation and scoring of results). The core of the pipeline lies in the `leto` module, which is itself but a wrapper around the `slclust` program, written in C++ using the `boost` library (Siek et al., 2000) for better performance.
The pipeline's design was user-driven and centered around modularity. It is therefore possible to replace the modules with custom ones, provided the command-line argument set is consistent with other elements of the pipeline. Conversely, and under the same conditions, it is possible to use another clustering program in lieu of `slclust`, all it takes is to change one single path in the general configuration file.

TE-Tracker will use that file (`SV.conf`) to determine all parameters to be used during execution, as well as the number of times that each step needs to be performed.
In the following sections we will describe the workings of the pipeline's most important steps.

Preprocessing

Each alignment file was preprocessed using the `eris` module. Parameters used were `-treat_bam input:0-1`, which removed all mappings whose best match contained more than 1 mismatch.
The use of this parameter was driven by our concern for noise reduction and was justified by the average high coverage depth of all our lines. Testing showed that what would later be verified as being noise was significantly reduced using this procedure.

Discordant pair detection

Let a mapped read $r(c, o, l)$ be defined by its chromosome c, its orientation o ($+$ or $-$) and its localisation l on c, and a read pair $p(r_a, r_b)$, the couple of two reads with $l_a < l_b$ and $d = l_b - l_a$ the distance between two reads of the same pair if $c_a = c_b$. Let $P_i = p_{i_1}, .., p_{i_n}$ the set of the n different mapping possibilities for the pair i. From the P_i we calculate the median M , the median absolute deviation (MAD) and upper and lower limits of the distances from the d_i , respectively d_{inf} , d_{sup} with $d_{inf} = M - 3.MAD$ and $d_{sup} = M + 3.MAD$. For a large insert library, a pair mapping $p_i(r_a, r_b) \in Pi$ is a proper mapping if $c_a = c_b$, $(o_a, o_b) = (-, +)$ and $d_{inf} < l_b - l_a < d_{sup}$. If such mapping do not exist in P_i at least once, the pair is considered as discordant and its mapping possibilities are classified following their mapping signatures. For each mapping possibility of one read pair $p_i \in P_i$, we have:

$$\text{if } c_a = c_b \begin{cases} \text{if } (o_a, o_b) = (-, +) & : \begin{cases} p_i \in Del \text{ if } d_i > d_{sup} \\ p_i \in Ins \text{ if } d_i < d_{inf} \end{cases} \\ \text{if } (o_a, o_b) = (+, -) & : p_i \in Dup \\ \text{if } (o_a, o_b) \in \{(-, -); (+, +)\} & : p_i \in Inv \end{cases}$$

$$\text{if } c_a \neq c_b : p_i \in Trans$$

Del, Ins, Dup, Inv, $Trans$, being sets of discordant read pairs having respectively a deletion, insertion, duplication, inversion and translocation signatures. For an organism with m chromosomes, we split the sets Del, Ins, Dup and Inv into n subsets (i.e. Del_1, ... ,Del_m, Ins_1, ... ,Ins_m, ...) and the $Trans$ set into $4 \times A_n^2$ subsets (i.e. $Trans_{(1,+);(2,+)}$, $Trans_{(1,+);(2,-)}$, $Trans_{(1,-);(2,+)}$, $Trans_{(1,-);(2,-)}$, ..., $Trans_{(m-1,-);(m,-)}$).

Single-linkage clustering and merging

We also calculated upper and lower limits of the depth of coverage c_{inf} , c_{sup} using M and MAD. Pairs whose reads both map in a genomic region with a very high coverage depth (typically $> 1000\times$ and containing repeated elements and low complexity sequences) are discarded from the discordant pair set. Discordant reads are sorted by l_a and clustered using single linkage clustering. For each subset, we built $G = (V, E)$, an undirected graph where nodes V are pairs $(p_i, p_j) \in E$ if:

$$|l_{i_a} - l_{j_a}| < \frac{4s}{c_{inf}} \tag{1}$$

and

$$|l_{i_b} - l_{j_b}| < d_{sup} - d_{inf} \tag{2}$$

with l_{i_a} and l_{j_a} respectively the position of the reads a of the pair i and the reads a of the pair j, and s the read size. Two pairs are linked if the distance between the two reads r_{j_a}, r_{j_a} is smaller than expected relative to the coverage depth variation (1) and the fragment size variation (2).
The single linkage processus starts from a single read pair, the seed. It tries to link it to the next available pairs; when linking is not possible anymore the last linked pair is used as seed, which is helpful in terms of CPU and memory usage. Nearby clusters with identical signatures are merged, this cluster extension allows no penalization due to low covered regions that may interrupt the linking process. After merging, the clusters are filtered by their size, rejecting those larger than d_{sup}.

Calling

Intra and interchromosomal translocations are called by searching overlapping clusters of different orientation(Figure 1.c) at the donor location that allows detection of translocation up to $2 \times M + 6 \times MAD$ bp. A deletion pattern cluster overlapping a duplication pattern cluster is needed to call an intrachromosomal translocation in the sense donor orientation. If a deletion pattern cluster does not overlap any duplication pattern cluster and if its size is over d_{inf} , this cluster supports a deletion. An inversion pattern cluster overlapping a inversion pattern cluster of the opposite orientation is needed to call an antisense intrachromosomal translocation.

Donor scoring

When several donors are available for one putative insertion site, the `leto` module outputs them all and groups them by an unique "acceptor ID". It is possible to assign a donor score to each using the `metis` module. Discordant reads anchoring at the acceptor site on one side,

and at every successive potential donor on the other, are extracted from the discordant BAM file generated by the `eris` module with parameters `-discordant=sorted`. For each donor, the program goes through the multiple mappings of each read, and looks for mates on the donor-side that map significantly better on a particular donor than on all the others. The count of these mates constitutes the score. The idea behind this is that multiple copies of a single TE in the genome, although they share most of their sequence, diverge by at least some SNPs. If this is true, then for each such insertion there will be discordant reads spanning these discriminating positions. With lenient mapping parameters such as ours, they will map on other copies of the original donor element, however their score will be lower there because of the SNPs they contain. Thus, a common score of 0 means that there is complete sequence identity between all possible donors.

TE-Tracker Parameterization

Most of TE-Tracker's parameters are self-explanatory and concern only the pipelining features of the program. The only two parameters that significantly influence output are X and Y, involved in the central Single-Linkage clustering algorithm.

There are two ways to deal with these parameters. By default, TE-Tracker produces a point estimate of the optimal parameters through statistical inference; however it is also possible to specify them manually.

Optimal parameters are evaluated as follows: since X is dependent on the coverage distribution, and Y on the insert size distribution, they have to be estimated in quite a different way. Estimation of Y is straightforward: it corresponds to a sufficiently large quantile of the empirical insert size distribution.

Similar distribution-based estimation for X lead to systematic underestimation. Thus another method was implemented, based on Let (C_n) be the random variable that gives coverage depth for each position n in the genome.

We model (C_n) as a two-state Markov chain, where one state corresponds to low-coverage (vertical coverage 0 or 1) regions and the second to normally covered ones. Under this model, the probability of a gap of size k is given by the probability of a k-sized null path: $p_{l_k} = \pi_{h \to l} \prod_{i=1}^{k} p_{ll}$ where $\pi_{h \to l}$ is the transition probability from the normal state to the low-coverage state and p_{ll} is the probability of staying in the low-coverage state. Both of these values can be empirically estimated by traversing a "pileup" file built from the alignment file.

We then calculate $p_l(k)$ for gaps of increasing size until a k_0 is encountered with probability $p_l(k_0)$ below a certain threshold. We then assign $X = k_0$, since k_0 will correspond to the maximum gap size that can be expected to be found under this model with a probability below the specified threshold. Typically, we found that satisfactory values of X were obtained with a very stringent treshold, for example when $p_l(k)$ was so small that it became indiscernible from 0 under double floating point precision.

Running TE-Tracker

In order to maximize sensitivity, we ran the `leto` module on a regular grid of increasing X and Y parameters (see below) and pooled the results by traversing the grid. A step size of 50 was chosen for X, while a step size of 300 was chosen for Y. X ranged from 50 to 2000, Y from 300 to 6000, which amounts to a total of 800 runs of the `leto` module. For this, we needed only one run of the prefiltering `eris` module, which allowed us to complete each grid in about a day's time.

The high number of points on the grid allowed us to draw a three-dimensional surface onto which the number of events outputted for each run was shown (Supplementary Figure 5). We noticed a clear maximum in all of the lines, and in every occasion, the point estimation of the optimal (X, Y) parameters was found to be quite close to it.

For our traversing program, we first looked for the said maximum, then we built a dictionary of donors from the insertions found in it. Then, we went through every run's output file and added data to the dictionary: we performed the cartesian product of the dictionary and the output file and added events that did not overlap either on donor or acceptor site. This allows to build a comprehensive landscape of all insertions that appear at least once on the grid. When an event was found in several points of the grid, the one supported by the most reads was kept, which ensures that the optimal parameters for each insertion were used for each line of the final output file. Looking for the maximum first allows to reduce the computational impact of such a combinatorial algorithm.

Performance evaluation

Comparison with other software

In order to assess its performance, we compared the events produced by TE-Tracker to those outputted by other software. However, this procedure proved difficult because all existing programs display only one edge of a mobilized transposon, with no or wrongful event size information. In addition, several others such as SHORE(Ossowski et al., 2008) do not assemble *INV*, *INS*, *DEL* and *DUP* events into putative two-sided events, which makes it difficult to quantify how much of an event the program has really detected.

Additionally, few existing programs were documented to handle large insert size libraries such as those produced by our mate-pair sequencing approach. We therefore performed our tests on two sequencing runs of the same epi-RIL line, the first one using a mate-pair preparation (median insert size 5kb), the second a more usual paired-end one with median insert size of 500bp.

Events involving COPIA93 and ATENSMP3, that had been PCR-validated and for which the donor has been clearly identified by biological means, were used in this benchmark. Seventeen COPIA93-events and 6 ATENSPM3-events in total had been validated on this particular line. Results are presented in Supplementary Tables 11 and 12. For programs other than TE-Tracker, a positive find was called whenever the program's output contained a cluster that overlapped both event's extremities.

Paired-end tests

For the paired-end run, we tested the popular SV-detection programs BreakDancer Max(Chen et al., 2009), SHORE(Ossowski et al., 2008), PRISM(Jiang et al., 2012) and SVDetect(Zeitouni et al., 2010). However, SVDetect was dismissed further on due to its long run-time in a single-core setting, as well as SHORE due to its ambiguous results and lack of support.

Mate-pair tests

Only BreakDancer(Chen et al., 2009), Hydra(Quinlan et al., 2010) and SVDetect (Zeitouni et al., 2010) claimed to be able to detect structural events based on large-fragment sequencing. For all these programs, we also assessed computational performance, for which we compared both CPU time and maximum memory use. These results are presented in Supplementary Figure 1.

Simulation

As mentioned in the main text, we also performed tests on simulated mate-pair sequencing runs to assess theoretical sensitivity and specificity. For this, we used the custom-built `Sphinx` script that generates artificial transpositions in a reference sequence. It is able to generate sense and antisense transpositions, as well as cut-paste, copy-paste and deletion events.

This tampered reference sequence is then supplied as an argument to the `SimSeqG.pl` script, which simulates mate-pair sequencing according to the features found in an existing alignment. It first scans an existing BAM file for aligned pairs and considers their fragment size and quality.

Using this data, it then builds a constant-frequency bin histogram for both the quality distribution of each base and the fragment size distribution for paired-end oriented mates and mate-pair oriented ones.

`SimSeqG.pl` has been designed according to the BCP principle and aims to mirror the sequencing process as closely as possible:

- A fragment size is sampled from the empirical distribution;
- a location is randomly selected in the genome and a sequence corresponding to the sampled fragment length is extracted starting at this position;
- the fragment is circularized and a random splice length is chosen from the pair-end fragment length distribution;
- a splice start is randomly chosen around the circularization point and the sequence is extracted from the circularized fragment;
- both ends of the subfragment are extracted and sequenced: for each base, the program will draw a quality corresponding to its position from the empirical quality distribution. Then, it will produce a sequencing error at that position with probability given by the base quality.
- once in a while, at a rate defined from the BAM learning set, a configuration leading to the production of a parasitic paired-end fragment is produced and the result is sequenced in a similar way;
- reads and qualities are then written in FASTQ format.

PCR Validation

A list of primers used for the validation of detected insertions is provided in Supplementary Table 13.

Set operations

Every set operation on shared insertions were performed using overlapping acceptor regions as the main criterion. The program used for intersecting regions was `bedops`(Neph et al., 2012).

Data Access

TE-Tracker is free for academic noncommercial use and can be downloaded from http://www.genoscope.cns.fr/
All additional programs as well as sequenced and simulated FASTQ files are also available at

this address.

Acknowledgments

This work was supported by the French National Agency for Research (ANR; Project Acronym EPIMOBILE).

Disclosure Declaration

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Figure Legends

Figure 1 TE-Tracker overview and main algorithms – **a.** Main steps of the TE-Tracker pipeline. **b.** Principle of the single linkage clustering. Left: genome view of three anomalously spaced read pairs (blue arrows and lines). Discordant reads are ordered by increasing order on position chromosome and offset. Read pairs that are distant by no more than X on their ordered side and by no more than Y on their unordered one are clustered together. Right: graphical representation of the Single-Linkage Clustering (SLC) algorithm implemented in TE-Tracker. **c.** Cluster merging and calling. The bracketed green clusters overlap on sequence 3 and are close to each other on sequence 1 and will therefore be merged. However, if the red cluster is also present all green clusters will be rejected. Considering the set of blue and green clusters independently leads to multiple donors or acceptors.

Figure 2 Circos representation of new TE insertion events detected in four epiRILs – Exterior circle represents the five chromosomes of Arabidopsis with pericentromeric regions and heterochromatic knob on chromosome 4 in dark grey. Arrows link donor TEs with the new insertion sites. Only events mapped with no ambiguity (no multiple acceptor sites and no similarity with events detected in wt) are represented.

References

Ahmed I, Sarazin A, Bowler C, Colot V, and Quesneville H. 2011. Genome-wide evidence for local dna methylation spreading from small rna-targeted sequences in arabidopsis. *Nucleic Acids Res* **39**: 6919–31. Ahmed, Ikhlak Sarazin, Alexis Bowler, Chris Colot, Vincent Quesneville, Hadi England Nucleic Acids Res. 2011 Sep 1;39(16):6919-31. doi: 10.1093/nar/gkr324. Epub 2011 May 17.

Alkan C, Coe BP, and Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–76. Alkan, Can Coe, Bradley P Eichler, Evan E Canadian Institutes of Health Research/Canada Howard Hughes Medical Institute/ England Nat Rev Genet. 2011 May;12(5):363-76. doi: 10.1038/nrg2958. Epub 2011 Mar 1.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al.. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–7. Baillie, J Kenneth Barnett, Mark W Upton, Kyle R Gerhardt, Daniel J Richmond, Todd A De Sapio, Fioravante Brennan, Paul M Rizzu, Patrizia Smith, Sarah Fell, Mark Talbot, Richard T Gustincich, Stefano Freeman, Thomas C Mattick, John S Hume, David A Heutink, Peter Carninci, Piero Jeddeloh, Jeffrey A Faulkner, Geoffrey J 090385/Wellcome Trust/United Kingdom 090385/Z/09/Z/Wellcome Trust/United Kingdom BB/H005935/1/Biotechnology and Biological Sciences Research Council/United Kingdom England Nature. 2011 Oct 30;479(7374):534-7. doi: 10.1038/nature10531.

Buisine N, Quesneville H, and Colot V. 2008. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**: 467–75. Buisine, Nicolas Quesneville, Hadi Colot, Vincent Genomics. 2008 May;91(5):467-75. doi: 10.1016/j.ygeno.2008.01.005. Epub 2008 Mar 14.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al.. 2009. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677–81. Chen, Ken Wallis, John W McLellan, Michael D Larson, David E Kalicki, Joelle M Pohl, Craig S McGrath, Sean D Wendl, Michael C Zhang, Qunyuan Locke, Devin P Shi, Xiaoqi Fulton, Robert S Ley, Timothy J Wilson, Richard K Ding, Li Mardis, Elaine R HG003079/HG/NHGRI NIH HHS/ P01 CA101937/CA/NCI NIH HHS/ Nat Methods. 2009 Sep;6(9):677-81. doi: 10.1038/nmeth.1363. Epub 2009 Aug 9.

Doolittle WF and Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–3. Doolittle, W F Sapienza, C ENGLAND Nature. 1980 Apr 17;284(5757):601-3.

Goodrich DW and Duesberg PH. 1990. Retroviral recombination during reverse transcription. *Proc Natl Acad Sci U S A* **87**: 2052–6. Goodrich, D W Duesberg, P H 5-R35-CA39915-04/CA/NCI NIH HHS/ Proc Natl Acad Sci U S A. 1990 Mar;87(6):2052-6.

Hancks DC and Kazazian H H J. 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**: 191–203. Hancks, Dustin C Kazazian, Haig H Jr Research Support, N.I.H., Extramural Review England Current opinion in genetics & development

Curr Opin Genet Dev. 2012 Jun;22(3):191-203. doi: 10.1016/j.gde.2012.02.006. Epub 2012 Mar 8.

Huang CR, Burns KH, and Boeke JD. 2012. Active transposition in genomes. *Annual Review of Genetics* **46**: 651–75. Huang, Cheng Ran Lisa Burns, Kathleen H Boeke, Jef D P30 CA006973/CA/NCI NIH HHS/ Annu Rev Genet. 2012;46:651-75. doi: 10.1146/annurev-genet-110711-155616.

Hurst GD and Werren JH. 2001. The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet* **2**: 597–606. Hurst, G D Werren, J H England Nat Rev Genet. 2001 Aug;2(8):597-606.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, and Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253–61. Iskow, Rebecca C McCabe, Michael T Mills, Ryan E Torene, Spencer Pittard, W Stephen Neuwald, Andrew F Van Meir, Erwin G Vertino, Paula M Devine, Scott E 1R01CA132065/CA/NCI NIH HHS/ 5P01CA116676/CA/NCI NIH HHS/ F32HG004207/HG/NHGRI NIH HHS/ R01 CA077337-10/CA/NCI NIH HHS/ R01 CA132065-01A2/CA/NCI NIH HHS/ R01 CA132065-02/CA/NCI NIH HHS/ R01 CA132065-03/CA/NCI NIH HHS/ R01 HG002898/HG/NHGRI NIH HHS/ R01 HG002898-06/HG/NHGRI NIH HHS/ R01CA086335/CA/NCI NIH HHS/ R01CA116804/CA/NCI NIH HHS/ R01GM078541/GM/NIGMS NIH HHS/ R01HG002898/HG/NHGRI NIH HHS/ Cell. 2010 Jun 25;141(7):1253-61. doi: 10.1016/j.cell.2010.05.020.

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, and Paszkowski J. 2011. An sirna pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**: 115–9. Ito, Hidetaka Gaubert, Herve Bucher, Etienne Mirouze, Marie Vaillant, Isabelle Paszkowski, Jerzy England Nature. 2011 Apr 7;472(7341):115-9. doi: 10.1038/nature09861. Epub 2011 Mar 13.

Jiang Y, Wang Y, and Brudno M. 2012. Prism: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**: 2576–83. Jiang, Yue Wang, Yadong Brudno, Michael Canadian Institutes of Health Research/Canada England Oxford, England Bioinformatics. 2012 Oct 15;28(20):2576-83. doi: 10.1093/bioinformatics/bts484. Epub 2012 Jul 31.

Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, et al.. 2009. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS genetics* **5**: e1000530. Johannes, Frank Porcher, Emmanuelle Teixeira, Felipe K Saliba-Colombani, Vera Simon, Matthieu Agier,

Nicolas Bulski, Agnes Albuisson, Juliette Heredia, Fabiana Audigier, Pascal Bouchez, David Dillmann, Christine Guerche, Philippe Hospital, Frederic Colot, Vincent PLoS Genet. 2009 Jun;5(6):e1000530. doi: 10.1371/journal.pgen.1000530. Epub 2009 Jun 26.

Jordan IK and McDonald JF. 1998. Evidence for the role of recombination in the regulatory evolution of saccharomyces cerevisiae ty elements. *J Mol Evol* **47**: 14–20. Jordan, I K McDonald, J F J Mol Evol. 1998 Jul;47(1):14-20.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, and Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**: 837–47. Kidd, Jeffrey M Graves, Tina Newman, Tera L Fulton, Robert Hayden, Hillary S Malig, Maika Kallicki, Joelle Kaul, Rajinder Wilson, Richard K Eichler, Evan E HG004120/HG/NHGRI NIH HHS/United States P01 HG004120-03/HG/NHGRI NIH HHS/United States Howard Hughes Medical Institute/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States Cell Cell. 2010 Nov 24;143(5):837-47. doi: 10.1016/j.cell.2010.10.027.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al.. 2012. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic Acids Res* **40**: D1202–10. Lamesch, Philippe Berardini, Tanya Z Li, Donghui Swarbreck, David Wilks, Christopher Sasidharan, Rajkumar Muller, Robert Dreher, Kate Alexander, Debbie L Garcia-Hernandez, Margarita Karthikeyan, Athikkattuvalasu S Lee, Cynthia H Nelson, William D Ploetz, Larry Singh, Shanker Wensel, April Huala, Eva 5P41HG002273-09/HG/NHGRI NIH HHS/ England Nucleic Acids Res. 2012 Jan;40(Database issue):D1202-10. doi: 10.1093/nar/gkr1090. Epub 2011 Dec 2.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette L J r, Lohr JG, Harris CC, Ding L, Wilson RK, et al.. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**: 967–71. Lee, Eunjung Iskow, Rebecca Yang, Lixing Gokcumen, Omer Haseley, Psalm Luquette, Lovelace J 3rd Lohr, Jens G Harris, Christopher C Ding, Li Wilson, Richard K Wheeler, David A Gibbs, Richard A Kucherlapati, Raju Lee, Charles Kharchenko, Peter V Park, Peter J Cancer Genome Atlas Research Network F32 AG039979/AG/NIA NIH HHS/United States F32AG039979/AG/NIA NIH HHS/United States K25AG037596/AG/NIA NIH HHS/United States R01 GM082798/GM/NIGMS NIH HHS/United States R01GM082798/GM/NIGMS NIH HHS/United States RC1HG005482/HG/NHGRI NIH HHS/United States U01 HG005209/HG/NHGRI NIH HHS/United States U01 HG005725/HG/NHGRI NIH HHS/United States U01HG005209/HG/NHGRI NIH HHS/United States U01HG005725/HG/NHGRI NIH HHS/United States U24 CA144025/CA/NCI NIH HHS/United States U24CA144025/CA/NCI NIH HHS/United

States Research Support, N.I.H., Extramural United States Science (New York, N.Y.) Science. 2012 Aug 24;337(6097):967-71. doi: 10.1126/science.1222077. Epub 2012 Jun 28.

Lerat E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**: 520–33. Lerat, E England Heredity (Edinb). 2010 Jun;104(6):520-33. doi: 10.1038/hdy.2009.165. Epub 2009 Nov 25.

Li H and Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**: 1754–60. Li, Heng Durbin, Richard 077192/Z/05/Z/Wellcome Trust/United Kingdom England Oxford, England Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18.

Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H, et al.. 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat Biotechnol* **29**: 723–30. Li, Yingrui Zheng, Hancheng Luo, Ruibang Wu, Honglong Zhu, Hongmei Li, Ruiqiang Cao, Hongzhi Wu, Boxin Huang, Shujia Shao, Haojing Ma, Hanzhou Zhang, Fan Feng, Shuijian Zhang, Wei Du, Hongli Tian, Geng Li, Jingxiang Zhang, Xiuqing Li, Songgang Bolund, Lars Kristiansen, Karsten de Smith, Adam J Blakemore, Alexandra I F Coin, Lachlan J M Yang, Huanming Wang, Jian Wang, Jun Nat Biotechnol. 2011 Jul 24;29(8):723-30. doi: 10.1038/nbt.1904.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al.. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–6. Lippman, Zachary Gendrel, Anne-Valerie Black, Michael Vaughn, Matthew W Dedhia, Neilay McCombie, W Richard Lavine, Kimberly Mittal, Vivek May, Bruce Kasschau, Kristin D Carrington, James C Doerge, Rebecca W Colot, Vincent Martienssen, Rob Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. England Nature Nature. 2004 Jul 22;430(6998):471-6.

Lopez-Flores I and Garrido-Ramos MA. 2010. The repetitive dna content of eukaryotic genomes. *Genome dynamics* **7**: 1–28. Lopez-Flores, I Garrido-Ramos, M A Switzerland Genome Dyn. 2010;7:1-28. doi: 10.1159/000337118. Epub 2012 Jun 25.

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, and Mager DL. 2006. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet* **2**: e2. Maksakova, Irina A Romanish, Mark T Gagnier, Liane Dunn, Catherine A van de Lagemaat, Louie N Mager, Dixie L PLoS Genet. 2006 Jan;2(1):e2.

Medvedev P, Stanciu M, and Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**: S13–20. Medvedev,

Paul Stanciu, Monica Brudno, Michael Nat Methods. 2009 Nov;6(11 Suppl):S13-20. doi: 10.1038/nmeth.1374.

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**: 427–30. Mirouze, Marie Reinders, Jon Bucher, Etienne Nishimura, Taisuke Schneeberger, Korbinian Ossowski, Stephan Cao, Jun Weigel, Detlef Paszkowski, Jerzy Mathieu, Olivier England Nature. 2009 Sep 17;461(7262):427-30. doi: 10.1038/nature08328. Epub 2009 Sep 6.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, and Kakutani T. 2001. Mobilization of transposons by a mutation abolishing full dna methylation in arabidopsis. *Nature* **411**: 212–4. Miura, A Yonebayashi, S Watanabe, K Toyama, T Shimada, H Kakutani, T England Nature. 2001 May 10;411(6834):212-4.

Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, and Gage FH. 2010. L1 retrotransposition in neurons is modulated by mecp2. *Nature* **468**: 443–6. Muotri, Alysson R Marchetto, Maria C N Coufal, Nicole G Oefner, Ruth Yeo, Gene Nakashima, Kinichi Gage, Fred H 1-DP2-OD006495-01/OD/NIH HHS/United States DP2 OD006495/OD/NIH HHS/United States DP2 OD006495-01/OD/NIH HHS/United States R01 MH088485-03/MH/NIMH NIH HHS/United States R01MH088485/MH/NIMH NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Nature Nature. 2010 Nov 18;468(7322):443-6. doi: 10.1038/nature09544.

Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, and Ponting CP. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol* **13**: R45. Nellaker, Christoffer Keane, Thomas M Yalcin, Binnaz Wong, Kim Agam, Avigail Belgard, T Grant Flint, Jonathan Adams, David J Frankel, Wayne N Ponting, Chris P R01 NS031348/NS/NINDS NIH HHS/ Medical Research Council/United Kingdom Wellcome Trust/United Kingdom Cancer Research UK/United Kingdom England Genome Biol. 2012 Jun 15;13(6):R45. doi: 10.1186/gb-2012-13-6-r45.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al.. 2012. Bedops: high-performance genomic feature operations. *Bioinformatics* **28**: 1919–20. Neph, Shane Kuehn, M Scott Reynolds, Alex P Haugen, Eric Thurman, Robert E Johnson, Audra K Rynes, Eric Maurano, Matthew T Vierstra, Jeff Thomas, Sean Sandstrom, Richard Humbert, Richard Stamatoyannopoulos, John A 1U54HG004592/HG/NHGRI NIH HHS/ 5U01ES017156/ES/NIEHS NIH HHS/ England Oxford, England Bioinformatics. 2012 Jul 15;28(14):1919-20. doi: 10.1093/bioinformatics/bts277. Epub 2012 May 9.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, and Weigel D. 2008. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome Res* **18**: 2024–33. Ossowski, Stephan Schneeberger, Korbinian Clark, Richard M Lanz, Christa Warthmann, Norman Weigel, Detlef Genome Res. 2008 Dec;18(12):2024-33. doi: 10.1101/gr.080200.108. Epub 2008 Sep 25.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, and Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**: 623–35. Quinlan, Aaron R Clark, Royden A Sokolova, Svetlana Leibowitz, Mitchell L Zhang, Yujun Hurles, Matthew E Mell, Joshua C Hall, Ira M 1F32HG005197-01/HG/NHGRI NIH HHS/ DP2OD006493-01/OD/NIH HHS/ Genome Res. 2010 May;20(5):623-35. doi: 10.1101/gr.102970.109. Epub 2010 Mar 22.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, and Korbel JO. 2012. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339. Rausch, Tobias Zichner, Thomas Schlattl, Andreas Stutz, Adrian M Benes, Vladimir Korbel, Jan O England Oxford, England Bioinformatics. 2012 Sep 15;28(18):i333-i339. doi: 10.1093/bioinformatics/bts378.

Rebollo R, Romanish MT, and Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**: 21–42. Rebollo, Rita Romanish, Mark T Mager, Dixie L Canadian Institutes of Health Research/-Canada Annu Rev Genet. 2012;46:21-42. doi: 10.1146/annurev-genet-110711-155621. Epub 2012 Aug 16.

Sibson R. 1973. Slink - optimally efficient algorithm for single-link cluster method. *Computer Journal* **16**: 30–34. O9837 Times Cited:152 Cited References Count:10.

Siek J, Lee L, and Lumsdaine A. 2000. Boost graph library. http://www.boost.org/libs/graph/.

Sindi SS, Onal S, Peng LC, Wu HT, and Raphael BJ. 2012. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol* **13**: R22. Sindi, Suzanne S Onal, Selim Peng, Luke C Wu, Hsin-Ta Raphael, Benjamin J R01 HG005690/HG/NHGRI NIH HHS/ R01 HG5690/HG/NHGRI NIH HHS/ England Genome Biol. 2012;13(3):R22. doi: 10.1186/gb-2012-13-3-r22.

Singer T, Yordan C, and Martienssen RA. 2001. Robertson's mutator transposons in a. thaliana are regulated by the chromatin-remodeling gene decrease in dna methylation (ddm1). *Genes & development* **15**: 591–602. Singer, T Yordan, C Martienssen, R A Genes Dev. 2001 Mar 1;15(5):591-602.

Slotkin RK and Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272–85. Slotkin, R Keith Martienssen, Robert R01 GM067014-01A1/GM/NIGMS NIH HHS/United States Research Support, N.I.H., Extramural Review England Nature reviews. Genetics Nat Rev Genet. 2007 Apr;8(4):272-85.

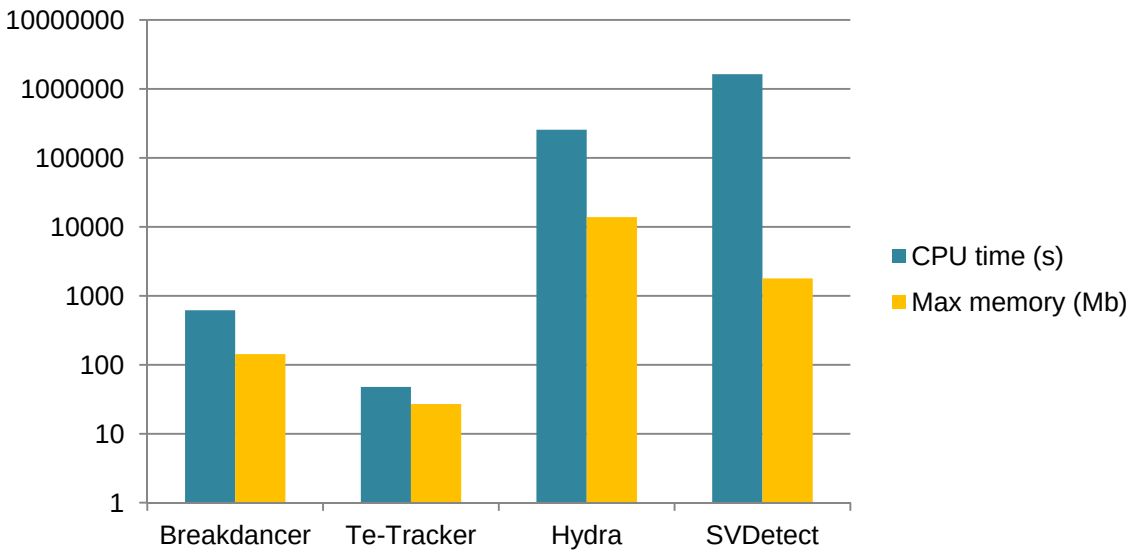
Thomas CA, Paquola AC, and Muotri AR. 2012. Line-1 retrotransposition in the nervous system. *Annual review of cell and developmental biology* **28**: 555–73. Thomas, Charles A Paquola, Apua C M Muotri, Alysson R 1-DP2-OD006495-01/OD/NIH HHS/ 1F31NS076198-01A1/NS/NINDS NIH HHS/ R01 MH094753-01/MH/NIMH NIH HHS/ Annu Rev Cell Dev Biol. 2012;28:555-73. doi: 10.1146/annurev-cellbio-101011-155822.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, and Kakutani T. 2009. Bursts of retrotransposition reproduced in arabidopsis. *Nature* **461**: 423–6. Tsukahara, Sayuri Kobayashi, Akie Kawabe, Akira Mathieu, Olivier Miura, Asuka Kakutani, Tetsuji England Nature. 2009 Sep 17;461(7262):423-6. doi: 10.1038/nature08351. Epub 2009 Sep 6.

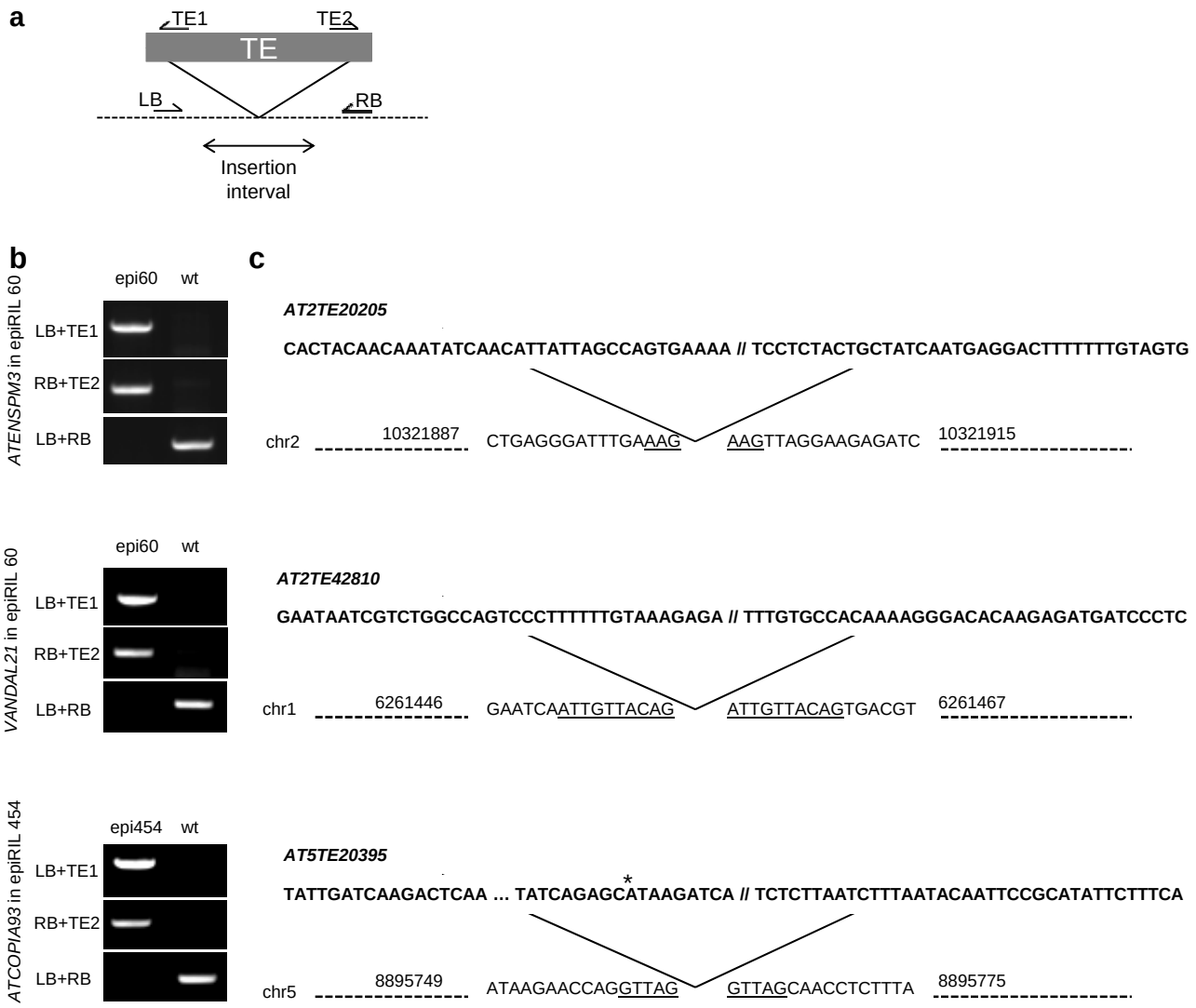
Vongs A, Kakutani T, Martienssen RA, and Richards EJ. 1993. Arabidopsis thaliana dna methylation mutants. *Science* **260**: 1926–8. Vongs, A Kakutani, T Martienssen, R A Richards, E J New York, N.Y. Science. 1993 Jun 25;260(5116):1926-8.

Zamudio N and Bourc'his D. 2010. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity* **105**: 92–104. Zamudio, N Bourc'his, D England Heredity (Edinb). 2010 Jul;105(1):92-104. doi: 10.1038/hdy.2010.53. Epub 2010 May 5.

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, and Barillot E. 2010. Svdetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* **26**: 1895–6. Zeitouni, Bruno Boeva, Valentina Janoueix-Lerosey, Isabelle Loeillet, Sophie Legoix-ne, Patricia Nicolas, Alain Delattre, Olivier Barillot, Emmanuel England Oxford, England Bioinformatics. 2010 Aug 1;26(15):1895-6. doi: 10.1093/bioinformatics/btq293.



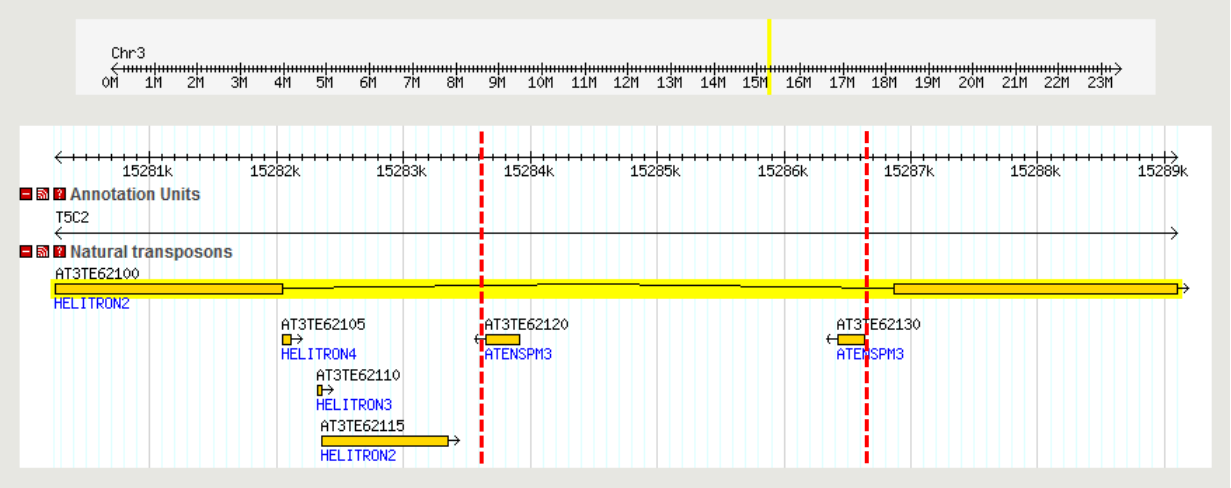
Supplementary figure 1 : Comparative evaluation of TE-Tracker time and memory usage. Both variables were evaluated using Platform Computing® Load Sharing Facility. Time is displayed in seconds, peak memory usage in megabytes, both displayed in log-scale.



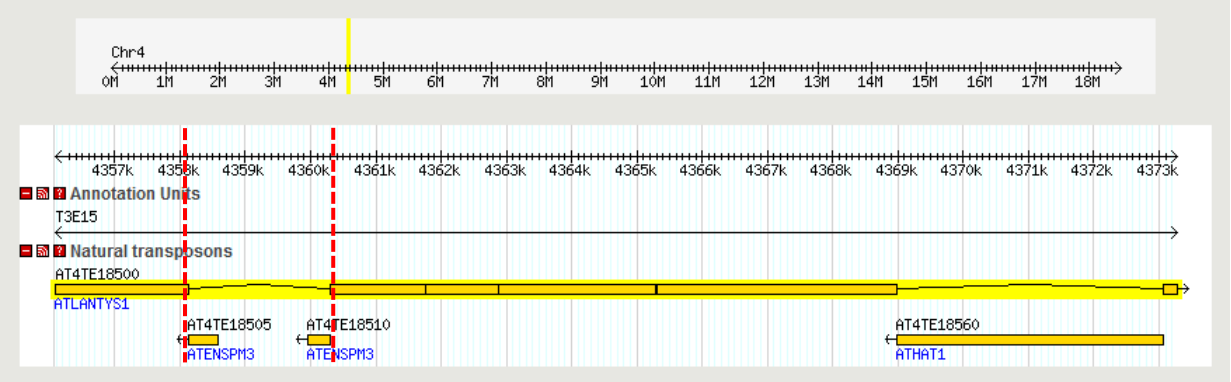
Supplementary figure 2 : Validation of new TE insertions.

a. Scheme of the PCR validation procedure, arrows represent primers. **b.** PCR results for 16 three new insertions identied in two epiRILs and involving three distinct donors : *ATENSPM3* (*AT2TE20205*), *VANDAL21* (*AT2TE42810*) and *ATCOPIA93* (*AT5TE20395*). **c.** Sequencing results of the PCR products that correspond to the borders of new insertions. TE sequence is represented in bold and target site duplications generated by TE insertion are underlined. The star highlights the single polymorphism located in the extremity of the TE and that discriminates between the two ATCOPIA93 potential donors.

a *ATENSPM3,ATENSPM3,HELITRON2*



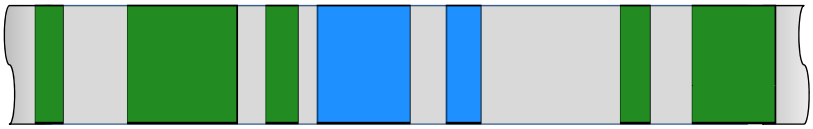
b *ATENSPM3,ATLANTYS1,ATENSPM3*



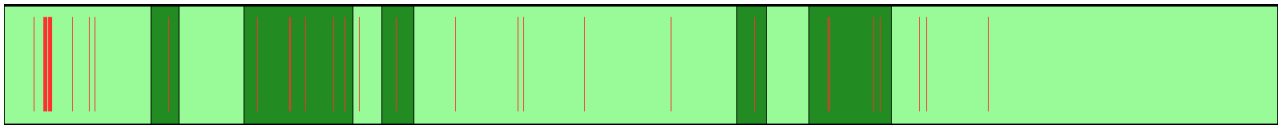
Supplementary figure 3 : **TAIR (Lamesch et al., 2012) Gbrowse display of composite elements detected by TE-Tracker.**

a. Element composed of two sequences annotated as ATENSPM3 and one annotated as HELITRON2. **b.** Element composed of two sequences annotated as ATENSPM3 and one annotated as ATLANTYS1. Red dotted lines indicate the boundaries of the mobile sequence as detected by TE-Tracker.

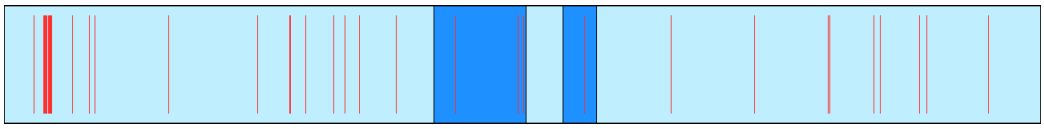
Putative composite sequence



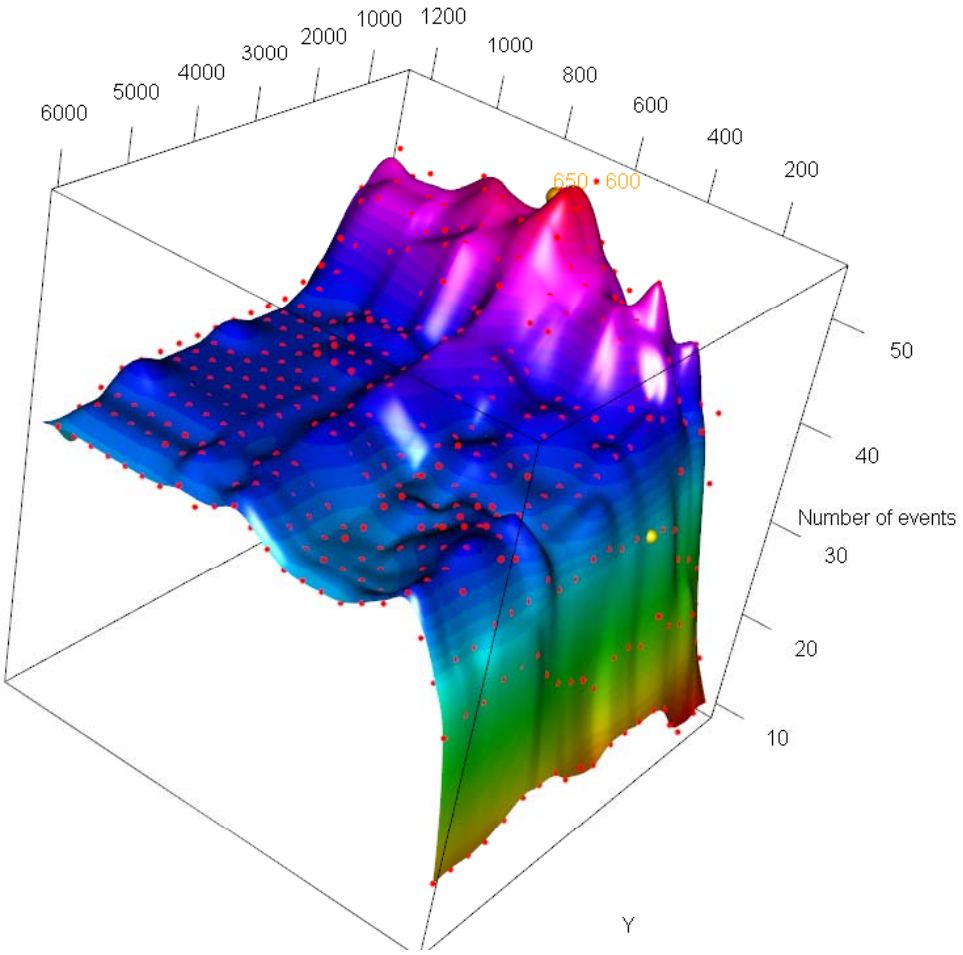
AT5TE15240



AT3TE89830



Supplementary Figure 4 : **Multiple-origin unique regions for donor suggest a recombination-like event during transposition.**
Clusters of reads mapping to acceptor region chr3:8763824..8765272 of epiRIL 454 have their mates mapping on several genomic regions corresponding to ATCOPIA78 sequences, which indicates multiple potential donors. The donor scoring feature extracted mates that mapped significantly better on one copy than on all others; this yielded not one, but two candidates. Piling up the said reads led to the configuration above. In light blue and green are the two copies, aligned unto each other. Red ticks indicate variants (SNP and indels) between the two copies. In dark blue and green are the regions delimited by specific reads for each copy, respectively. On top is the putative inserted sequence with blocks coming from either copy. Ticks not included in darker regions represent non-unique variations that are therefore also found when aligning to other *ATCOPIA78* sequences.



Supplementary figure 5 : **Number of events on two-dimensional grid for epiRIL 439.** Example of surface generated by a 800-point grid in X-Y space-Surface was loess-interpolated, coordinates of the maximum are displayed in orange, Markov-chain point estimate is displayed by the orange sphere near the peak, distribution estimate is displayed by a yellow sphere.

	Acceptor stire		Donor			
ID	chr.	position	chr.	start	stop	detected
1	chr1	248294	chr5	5630011	5634975	x
2	chr1	623731	chr5	5630011	5634975	x
3	chr1	770974	chr5	5630011	5634975	x
4	chr1	1071330	chr5	5630011	5634975	x
5	chr1	2037961	chr5	5630011	5634975	x
6	chr1	2257882	chr5	5630011	5634975	x
7	chr1	2481930	chr5	5630011	5634975	x
8	chr1	2871763	chr5	5630011	5634975	x
9	chr1	3340610	chr5	5630011	5634975	x
10	chr1	5016892	chr5	5630011	5634975	x
11	chr1	5046091	chr5	5630011	5634975	x
12	chr1	5250696	chr5	5630011	5634975	x
13	chr1	6416036	chr5	5630011	5634975	x
14	chr1	6715956	chr5	5630011	5634975	x
15	chr1	6912218	chr5	5630011	5634975	x
16	chr1	7982407	chr5	5630011	5634975	x
17	chr1	8603037	chr5	5630011	5634975	x
18	chr1	9735777	chr5	5630011	5634975	x
19	chr1	10284309	chr5	5630011	5634975	
20	chr1	10671044	chr5	5630011	5634975	x
21	chr1	12351929	chr5	5630011	5634975	x
22	chr1	13861366	chr5	5630011	5634975	x
23	chr1	14152163	chr5	5630011	5634975	x
24	chr1	16793854	chr5	5630011	5634975	x
25	chr1	17841838	chr5	5630011	5634975	x
26	chr1	19477905	chr5	5630011	5634975	x
27	chr1	20198447	chr5	5630011	5634975	x
28	chr1	20795441	chr5	5630011	5634975	x
29	chr1	20917772	chr5	5630011	5634975	x
30	chr1	22738695	chr5	5630011	5634975	x
31	chr1	22788970	chr5	5630011	5634975	x
32	chr1	23436613	chr5	5630011	5634975	x
33	chr1	24714782	chr5	5630011	5634975	x
34	chr1	25429340	chr5	5630011	5634975	x
35	chr1	25902534	chr5	5630011	5634975	x
36	chr1	27314695	chr5	5630011	5634975	x
37	chr1	27552632	chr5	5630011	5634975	x
38	chr1	27828210	chr5	5630011	5634975	x
39	chr1	29071608	chr5	5630011	5634975	x
40	chr5	1505621	chr5	5630011	5634975	x

	Acceptor stire		Donor			
ID	chr.	position	chr.	start	stop	detected
41	chr5	1674163	chr5	5630011	5634975	x
42	chr5	2174400	chr5	5630011	5634975	x
43	chr5	3304484	chr5	5630011	5634975	x
44	chr5	3910297	chr5	5630011	5634975	x
45	chr5	4230314	chr5	5630011	5634975	x
46	chr5	6104389	chr5	5630011	5634975	x
47	chr5	6854645	chr5	5630011	5634975	x
48	chr5	6906526	chr5	5630011	5634975	x
49	chr5	7812336	chr5	5630011	5634975	x
50	chr5	9091588	chr5	5630011	5634975	x
51	chr5	9698725	chr5	5630011	5634975	x
52	chr5	10365579	chr5	5630011	5634975	x
53	chr5	10661178	chr5	5630011	5634975	x
54	chr5	11826789	chr5	5630011	5634975	
55	chr5	12309261	chr5	5630011	5634975	x
56	chr5	13315486	chr5	5630011	5634975	x
57	chr5	13727742	chr5	5630011	5634975	x
58	chr5	14112116	chr5	5630011	5634975	x
59	chr5	14356539	chr5	5630011	5634975	
60	chr5	14432804	chr5	5630011	5634975	
61	chr5	15059811	chr5	5630011	5634975	x
62	chr5	15108704	chr5	5630011	5634975	x
63	chr5	16366871	chr5	5630011	5634975	x
64	chr5	17754143	chr5	5630011	5634975	x
65	chr5	18929477	chr5	5630011	5634975	x
66	chr5	19550283	chr5	5630011	5634975	x
67	chr5	19717309	chr5	5630011	5634975	x
68	chr5	21710387	chr5	5630011	5634975	x
69	chr5	22466087	chr5	5630011	5634975	x
70	chr5	22514806	chr5	5630011	5634975	x
71	chr5	23367420	chr5	5630011	5634975	x
72	chr5	23406966	chr5	5630011	5634975	x
73	chr5	23418427	chr5	5630011	5634975	x
74	chr5	23445199	chr5	5630011	5634975	x
75	chr5	24842981	chr5	5630011	5634975	x
76	chr5	24946033	chr5	5630011	5634975	x
77	chr5	25260678	chr5	5630011	5634975	x
78	chr5	25589670	chr5	5630011	5634975	x
79	chr5	26965212	chr5	5630011	5634975	x

Supplementary Table 1: Synthetic TE insertions in Arabidopsis

acceptor site		donor		
chromosome	position	chromosome	start	stop
chr19	3943363	chr12	101539821	101545842
chr19	8369980	chr12	75268648	75274681
chr19	9250040	chr12	75268648	75274681
chr19	18299455	chr12	101539821	101545842
chr19	31836999	chr12	101539821	101545842
chr19	32550986	chr12	75268648	75274681
chr19	37320442	chr12	101539821	101545842
chr19	38820257	chr12	101539821	101545842
chr19	40646550	chr12	75268648	75274681
chr19	42456336	chr12	75268648	75274681
chr19	42984248	chr12	101539821	101545842
chr19	45046186	chr12	101539821	101545842
chr19	52203029	chr12	101539821	101545842
chr19	52246544	chr12	75268648	75274681
chr19	54702746	chr12	101539821	101545842
chr19	55041491	chr12	75268648	75274681
chr19	58771303	chr12	75268648	75274681

Supplementary Table 2: Synthetic TE insertions in Human

line	coverage	*p(c)* < 10 (%)	sequencing
55	24.64	2.61	GAIIx
60	27.55	1.70	GAIIx
454	17.68	11.39	HiSeq2000
439	32.71	0.62	HiSeq2000
wt1	29.54	0.84	GAIIx
wt2	37.27	0.48	HiSeq2000

Supplementary Table 3: Sequencing and alignment properties

Supplementary Table 4-9 :
Ces tables sont présentées dans l'annexe 1 du manuscrit de thèse

	TE	validation of the presence	validation of the donnor
present in epiRILs only	*ATENSPM3*	13/13	9/9
	VANDAL21	5/5	3/3
	ATCOPIA93	42/42	14/14
	ATCOPIA13	1/1	-
	ATCOPIA78	4/4	2/3
	ATENSPM3;ATLANTYS1;ATENSPM3	2/2	ND
	ATENSPM3;ATENSPM3;HELITRON	3/3	-
present in wt	*ATGP2*	1/1	ND
	ATHILA6A	1/1	ND
	ATCOPIA78	1/1	ND
TOTAL		72/72	25/26

Supplementary Table 10: Summary of PCR validation of new insertions

	Paired-End		Mate-Pair		
	True Positives(/17)	Total Positives	True Positives	Total Positives	Correctly Scored
TE-Tracker	-	-	100%	352	17
BreakDancerMax***	17%	1600	0%	2859	-
PRISM**	47%	9641	-	-	-
SHORE			-	-	-
SVDetect*	?	?	94%	125929	-
Hydra	-	-	82%	42521	-

Supplementary Table 11: Comparative sensitivity evaluation for common structural variation detection tools (for *ATCOPIA93* insertions)

	Paired-End		Mate-Pair		
	True Positives(/6)	Total Positives	True Positives	Total Positives	Correctly Scored
TE-Tracker	-	-	100%	176	6
BreakDancerMax***	100%	1600	0%	2859	-
PRISM**	100%	9641	-	-	-
SHORE			-	-	-
SVDetect*	?	?	100%	125929	-
Hydra	-	-	100%	42521	-

Supplementary Table 12: Comparative sensitivity evaluation for common structural variation detection tools (for *ATENSPM3* insertions)

*We did not manage to get SVDetect to produce results on paired-end data (program crashed or run did not finish).

** Results filtered for events supported by more than 10 reads

*** For BreakDancer, we allowed for an extra 500 base pairs left or right.

A blank cell means the test was not performed, a question mark means the program did not terminate, a hyphen indicates an unsupported function

primers flanking the acceptor sites				
primer name	primer sequence	acceptor site		
		chr	start	stop
55_COPIA13_i1_F	CCCACATCCAAAATCCCTTT	chr3	17341732	17341847
55_COPIA13_i1_R	TGGTGCAGAATATGGAATGG			
55_COPIA78_i1_F	ATCATCACCGTGGTCCTTGT	chr4	10739659	10739988
55_COPIA78_i1_R	GACCACTGAAATCGGTCCAC			
55_COPIA78_i2_F	TGATATGTGTGGCTCTTGTTCA	chr4	13252525	13253326
55_COPIA78_i2_R	CCTCGAAAATTTAGGGCAGA			
55_COPIA93_i1_F	GTAAGTGCATTTGATGGATTCG	chr1	3990167	3990279
55_COPIA93_i1_R	CCATGAATCGGATTTTAAACTGT			
55_COPIA93_i2_F	ACCATCACCAAACTCAACTGG	chr1	12169447	12169578
55_COPIA93_i2_R	TGAATCTCCGGATCGCTTAC			
55_COPIA93_i3_F	ATTGATTTCTTTTGGCAGTGACT	chr1	20679800	20679956
55_COPIA93_i3_R	CTACATGAATCCCCACCACA			
55_COPIA93_i4_F	GCTGATCTGCAGCTTATTTCCT	chr1	25815758	25815886
55_COPIA93_i4_R	CTGGACATGCAGCTTAACAAAG			
55_COPIA93_i5_F	GTAGTGAGTCTTGTGAGCAGAGGT	chr2	3215890	3216071
55_COPIA93_i5_R	CCATAAAGATACTGACGCCACA			
55_COPIA93_i6_F	CCGGTAGCTCTACTAAGCGAAG	chr2	3275722	3275965
55_COPIA93_i6_R	CTATGTGAGGTGGGGAAAGGT			
55_COPIA93_i7_F	ATGGTGCCTTGATCTTCAGC	chr2	3505788	3506083
55_COPIA93_i7_R	AGAGAGTCCTCACCTAACAAGCA			
55_COPIA93_i8_F	AAGCTCTGAACCATCCATGC	chr2	4171056	4171205
55_COPIA93_i8_R	AAAGATCCCCTTATCCTTGGAA			
55_COPIA93_i9_F	TCTTCAGCAGTAACCACACGA	chr2	8272168	8272305
55_COPIA93_i9_R	TTGGATGGGGGAGATAAGAAC			
55_COPIA93_10_F	AGGACCATCTGATGACTTTATTCT	chr2	19475806	19475903
55_COPIA93_10_R	TACGCCGGATCCTCCTAGT			
55_COPIA93_i11_F	AAACCCCTAAAACCCTTCTTCT	chr3	4589949	4590097
55_COPIA93_i11_R	GTGAAATGGCTCATGGTTAAAG			
55_COPIA93_i12_F	GAGGAGATACGGGAAATGTGA	chr3	8036011	8036108
55_COPIA93_i12_R	GGATCACACATGCTTGAAAAGA			
55_COPIA93_i13_F	GGACTTGAGAGGGCAAATAGTG	chr4	8561572	8561742
55_COPIA93_i13_R	CTCTAAATTCAACCTCACCGAAC			
55_COPIA93_i14_F	ACCCGTAAGACTTTGCGGTA	chr5	12603975	12604164
55_COPIA93_i14_R	CCATCGAGTTAACGGCATTC			
55_COPIA93_i15_F	TCCGACTTGCTGGGTTTACT	chr5	14550492	14550576
55_COPIA93_i15_R	ATGAGATGGTCTTTCGCTACG			
55_COPIA93_i16_F	AGAAGGCCATGTCTTGATCCT	chr5	19519546	19519749
55_COPIA93_i16_R	TGGTATCCACGTTGGAAGAAG			
55_SPM3_i1_F	CACACCTTCACCCATGTGAT	chr1	21486789	21486913
55_SPM3_i1_R	CAACCCTTTACCAGGGTACCTA			
55_SPM3_i2_F	TGTGTTCAAGGGACAAAAACAG	chr1	27101182	27103667
55_SPM3_i2_R	TACGTCCCGTTTCTTTGGTC			
55_SPM3_i3_F	GATCTTATGTTACGTCAAAGCCAAC	chr3	9043438	9043531
55_SPM3_i3_R	TGGAGTCCTGCCTTTATTCTTG			
55_SPM3_i4_F	CCAAACCTATGTGTTGACGAAG	chr3	10647578	10647846
55_SPM3_i4_R	GGGAAAGTCGACAGCGATAG			
55_SPM3_i5_F	GGAAAGCCAAGATTTTGATGAC	chr3	16618765	16619206
55_SPM3_i5_R	CAACACCATTCATGACAAAACC			
55_SPM3_i6_F	TTTCTCAGGCAGGAAGAGACA	chr5	24028648	24028773
55_SPM3_i6_R	GCGATCAAGCTCTCATCAAAC			
55_COPIA78_i1_F	TGCAGATCGTCAACGTTTGT	chr3	8764557	8764584
55_COPIA78_i1_R	TTCAGCACGCCACTACATTC			
55_COPIA78_i2_F	TTTTCCCAATCCAAAAATCA	chr3	22615358	22615562
55_COPIA78_i2_R	TTAAATGATGACGCGGAAAG			
60_COPIA93_i1_F	ATGCTGCAACCTCTGAACCT	chr2	15677093	15677345
60_COPIA93_i1_R	TCTCATGTGCTCTTGGTTGG			
60_SPM3_i1_F	ACTAGATTTCGCTGTCGTTCC	chr1	25956040	25956309
60_SPM3_i1_R	AGCGAAAATCAATAGCGTGA			
60_SPM3_i2_F	CAATTTCGACACCTTTCTCG	chr2	10321820	10321904
60_SPM3_i2_R	AAGCAGAAGAAACCCATCATC			
60_SPM3_i3_F	CAGAAACTCCCCTTTTCAGG	chr2	17097260	17097372
60_SPM3_i3_R	AATTTGCATTTGTCGTCAGG			
60_SPM3HELI_i1_F	CCAAACAACAAACCAAGTAATCG	chr1	12178270	12181765
60_SPM3HELI_i1_R	TGGAATCTTAGCTTGTTTTGG			
60_SPM3HELI_i2_F	TGTCTGGCATTACAACATTTAAAA	chr3	3464127	3467214
60_SPM3HELI_i2_R	CACTGATGATTCCGGGAAGA			
60_SPM3HELI_i3_F	ACGGGCTCTTGGTGTATCTG	chr3	18274281	18277593
60_SPM3HELI_i3_R	GGATCCTTAATGTAAGCTTTTCTGG			
60_SPM3ATLA_i1_F	CGAATATCGATGGACTAACAATACC	chr3	3464127	3467214
60_SPM3ATLA_i1_R	TCGTGTTTACGGAGAACGTG			

Supplementary Table 13 (part1) : primer sequences for PCR validation of new TE insertions

primers flanking the acceptor sites				
primer name	primer sequenC e	acceptor site		
		chr	start	stop
60_SPM3ATLA_i2_F	CGTCGAGGCTTGAGAGAAGT	chr3	15773104	15778796
60_SPM3ATLA_i2_R	TGTCCCACCCATCTCTCTTT			
60_VANDAL21_i1_F	CTAAGACACCAATCCGGTGA	chr1	6261367	6261460
60_VANDAL21_i1_R	CGGCTGAGTTAGTTGTGGAG			
60_VANDAL21_i2_F	TTTGGAGTTGACGGAAGATG	chr2	12660632	12660713
60_VANDAL21_i2_R	GCCGTGTGACGTTAGTAGGA			
60_VANDAL21_i3_F	AATCAATCCAGCGAAGAAGAC	chr4	1151516	1151621
60_VANDAL21_i3_R	TTCGTCGGTAAATGAACCAG			
439_COPIA93_i1_F	ATCAACAAATAGATATCTTTGCATTTT	chr1	11211722	11211794
439_COPIA93_i1_R	TTCTTTTCGGATTCTGTGTGG			
439_COPIA93_i2_F	ATGGTTCGATTGTGGAGGAA	chr3	21970792	21970868
439_COPIA93_i2_R	CAACATTCTAGGGTTTTACCTTGC			
439_COPIA93_i3_F	TCTCCCACCCAAAATCTCAG	chr5	6591834	6592026
439_COPIA93_i3_R	GCGTGTTGTTTTGCAGTGAT			
439_SPM3_i1_F	AAAAGATCAACATGGTCACTCG	chr1	24091117	24091194
439_SPM3_i2_F	TGGACTTCGACGTTAGATCTTG	chr1	24334475	24334558
439_SPM3_i3_F	GCTGAGAATGATGTTTGGTTG	chr2	11505759	11505852
439_SPM3_i4_F	AACGTCAAATACATTCGCCA	chr5	7815404	7815488
439_VANDAL21_i1_F	CGATTTGACCCCAAAACAAA	chr1	28631078	28631250
439_VANDAL21_i2_F	TCCAGAGCTTCCCGTCAGTTT	chr2	79850	80054
454_COPIA93_i1_F	CTACTGGCGTGGATGCTTTT	chr1	3115	3249
454_COPIA93_i2_F	TATGACTGGGAAGCACGATG	chr1	11191545	11191795
454_COPIA93_i3_F	TTCTTCCAAAATTTGATTACTTGC	chr1	25653718	25654122
454_COPIA93_i4_F	CCCAAGTGTAGTCCAAAACG	chr1	29616254	29616578
454_COPIA93_i5_F	TGTACAAAAGCTAGGTAGGAGGATG	chr2	11496506	11496664
454_COPIA93_i6_F	TTTTGTGATATTTCTGCGTAGATTG	chr3	807190	807542
454_COPIA93_i7_F	TTGGAATTCGATTCAACTAGAGG	chr3	1737581	1737875
454_COPIA93_i8_F	TGTGCGTAGCTTTTCCTTGA	chr3	3083438	3083594
454_COPIA93_i9_F	GCCTTAATGAATGGCCCAATA	chr3	4492935	4493027
454_COPIA93_10_F	GACCCCAACCAATCAATCAC	chr3	16580047	16580428
454_COPIA93_i11_F	AAAATGTCACTTATCATTGCAAGC	chr3	16894460	16894609
454_COPIA93_i12_F	CGCTTGCTAGGCAACCTGTA	chr3	21980027	21980291
454_COPIA93_i13_F	GAAGGGGAAACCCTATGGAA	chr4	3651904	3652219
454_COPIA93_i14_F	TTATCTTTAATGTGACGATTGCTTT	chr4	5006317	5006790
454_COPIA93_i15_F	ACGCATCATCAGGCTTCTCT	chr4	7003389	7003463
454_COPIA93_i16_F	CCAACGTGGGTTAGATCCAT	chr4	7083314	7083461
454_COPIA93_i17_F	CTTGAGGGGCTGTCTCTGAT	chr4	16682304	16682584
454_COPIA93_i18_F	GGATTCTGCTTTTAGACCATCAA	chr5	729472	729675
454_COPIA93_i19_F	CGGTGTTTGTAGTGGGAACC	chr5	8895666	8895869
454_COPIA93_i19_R	GCAAAAACATTCGTACATTTCG			
454_COPIA93_i20_F	TCTCTTGTCTTGTGCGACTCA	chr5	22921922	22922268
454_COPIA93_i21_F	TGACAAGAAAACCACAAGACAAA	chr5	23913821	23913992
454_COPIA93_i22_F	GCTACGCTGACTGGGCTTCT	chr5	25834575	25834843
wt_COPIA78_i1_F	TGTGTCGTTGAAACACATCC	chr4	9320384	9320617
wt_COPIA78_i1_R	ACAATCCGTACCAAGCGAAC			
wt_ATGP2_i1_F	TAAACCCGTAAGGCGCAAG	chr3	14200805	14201497
wt_ATHILA6A_i1_R	ATTCAAAAACCGGAGAAGCA	chr3	5285507	5285698

primers within donor TEs	
VANDAL21_LB_R	CTTGCAGGAGGAAAACG
VANDAL21_RB_F	TAGGCGGCATGACTGATTTC
COPIA93_LB_R	CAAGCACAAACGGACTGATG
COPIA93_RB_F	TTGCATAAGTCTTGCGCTTG
SPM3_LB_R	TAAGTGTTGGCGCTGAAGTG
SPM3_RB_F	AAAAGTGTGGAGGCCATCAG
ATHILA6A_LB_R	GCAAAGTGTGCAAGAATGATG
ATGP2_LB_R	GATGCGGACTGCCTAAAGTC
ATGP2_RB_F	GGGAAAAGCAAGGGATTTGT
COPIA13_LB_R	GTGCGAATCCAACCCACTAT
COPIA13_RB_F	CCTGGTCAAGGCATTTTGTT
SPM3HELI_LB_R	GGGCATAACATAGCGTTTATGA
SPM3HELI_RB_F	TCGAAGTATTGGCGGAATAAGT
SPM3ATLA_LB_R	CCATAACAGGGGCATAAGACA
SPM3ATLA_RB_L	TTTTAGCTTCCCGCTTAGTGA
COPIA78_LB_R	TGAGAGGGGGAGGAGGTATT
COPIA78_RB_F	AATACCTCCTCCCCCTCTCA
COPIA78(1kb)_LB_R	CTTGTAATGACCCAAGAAGTTGCTCTATTG
COPIA78(1kb)_RB_F	GGCTTACATTATTCAACTACTAGTGATTACAAGC

Supplementary Table 13 (part2) : primer sequences for PCR validation of new TE insertions

2.2 Résultats du séquençage du génome des epiRIL : identité des ET mobilisés, fréquence et dynamique de mobilisation

A ce jour, les génomes de 53 epiRIL ainsi que d'un individu *ddm1* et de deux individus sauvages ont été séquencés et analysés avec TE-Tracker. Les séquençages ont été effectués sur des ADN extraits de lots de plantes de la génération F9 (au moins 12 plantes par lot), ce qui doit permettre l'identification des insertions communes à toutes les plantes au sein du lot, mais pas celles propres à chacune d'entre elles. Ainsi donc, en séquençant un lot de plantes F9, ce sont les insertions présentes dans la plante mère unique F8 que l'on détecte.

Les listes de nouvelles insertions pour chacune des epiRIL ainsi que pour le mutant *ddm1* que j'ai utilisées pour l'ensemble des analyses rapportées ci-après sont placées dans l'annexe 2. Elles correspondent aux listes obtenues à partir de TE-tracker auxquelles ont été soustraites les insertions également détectées dans les plantes sauvages ainsi que les évènements ambigus (voir Matériels et Méthodes).

2.2.1 Analyses globales de la mobilisation des ET dans les epiRIL et le mutant *ddm1*

Le séquençage du génome du mutant *ddm1* n'a révélé que quatre évènements de transposition faisant intervenir quatre familles d'ET (fig. 2.3), la famille de transposons à ADN de type EN/SPM *ATENSPM3* et trois familles de rétroéléments à LTR, *ATCOPIA21*, *ATCOPIA57* et *ATCOPIA78*. Les familles *ATENSPM3* et *ATCOPIA21* avaient précédemment été décrites comme mobiles dans le mutant *ddm1* (Miura et al. 2001, Tsukahara et al. 2009). En revanche, *ATCOPIA57* n'a jamais été décrit auparavant comme étant mobile. La famille *ATCOPIA78,* qui regroupe les ET également connus sous le nom d'*ONSEN,* a fait l'objet de plusieurs études récentes qui ont montré la mobilité des *ONSEN* chez les mutants affectés dans la voie du RdDM soumis à des conditions de stress chaleur (Ito et al. 2011). Cependant, jamais leur mobilité n'avait pu être mise en évidence dans le mutant *ddm1*. Le faible nombre d'insertions identifiées ici mais également l'absence de nouvelles insertions d'ET connus pour être fortement mobiles dans ce fond génétique mutant (notamment le rétroélément *ATCOPIA93* et le transposon à ADN *VANDAL21*) peuvent paraître contradictoires au regard

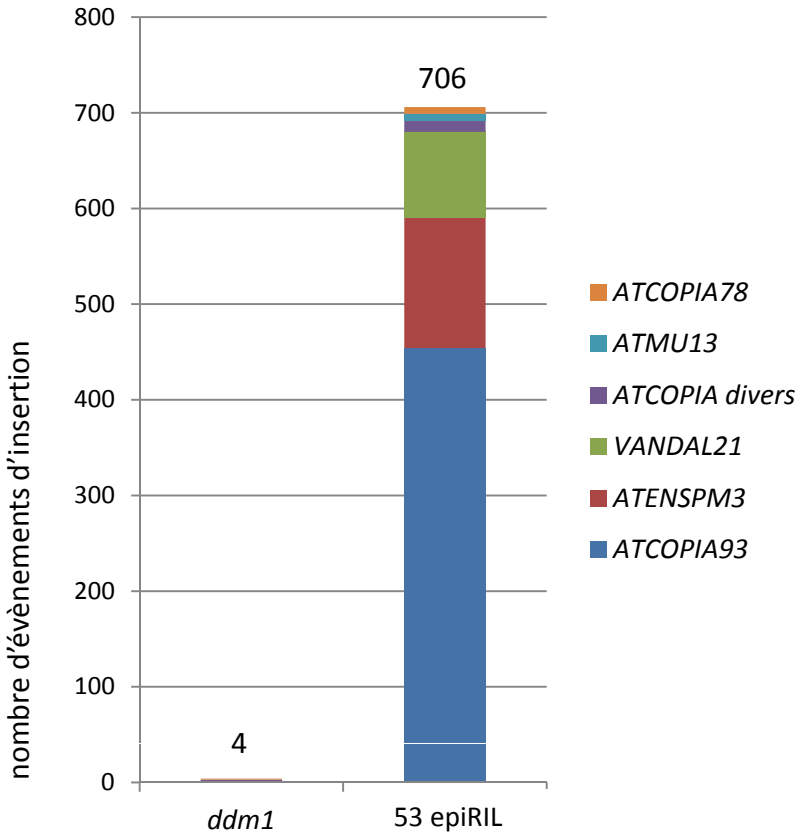


Figure 2.3 : Nombre d'évènements d'insertion par famille d'ET mobilisés dans le mutant *ddm1* et les epiRIL.

de travaux antérieurs (Tsukahara et al. 2009). Cependant, les études précédentes portaient sur des plantes *ddm1* issues de huit autofécondations successives contre quatre seulement dans notre cas, ce qui pourrait expliquer le plus faible nombre d'insertions détecté par nous.

Le séquençage du génome des 53 epiRIL a permis de mettre en évidence un total de 706 évènements d'insertion d'ET (fig. 2.3). Le contraste important entre ce nombre et celui des nouvelles insertions détectées dans le mutant *ddm1* est un premier indicateur de l'activité des ET dans les epiRIL, malgré la restauration de la fonction DDM1 dès la F1 (*ddm1* étant une mutation récessive, Vongs et al. 1993). Cependant, seules 13 familles d'ET sont concernées : neuf familles de rétroéléments à LTR de la superfamille COPIA (*ATCOPIA13*, ATCOPIA20, *ATCOPIA21, ATCOPIA31, ATCOPIA51, ATCOPIA63, ATCOPIA78, ATCOPIA93 et ATRE1*), deux familles de transposons à ADN, l'une de type Mu (*VANDAL21* et *ATMU*) et l'autre de type EN/SPM (*ATENSPM3*). Le nombre de familles mobilisées est donc faible au regard du nombre total de familles que compte le génome d'Arabidopsis (>300) (Buisine et al. 2008, Ahmed et al. 2011) et surtout en comparaison avec la quantité de famille réactivées transcriptionnellement suite à la perte de méthylation induite par la mutation *ddm1* (Lippman et al. 2004, Zemach et al. 2013). Plusieurs hypothèses permettent d'expliquer cet apparent paradoxe. La plus triviale est que, bien qu'exprimée, une importante proportion des ET est constituée d'éléments dégénérés, présentant des ORF mutées, tronquées voire absentes et sont, par conséquent, incapables de transposer. Au laboratoire, nous avons ainsi déterminé, sur la base de plusieurs critères de séquence, que plus de la moitié des familles d'ET du génome d'Arabidopsis était composée de copies sans capacité de transposition (données non montrées). La deuxième hypothèse est l'existence de mécanismes contrôlant l'activité des ET à un niveau post-transcriptionnel. En effet la répression post-transcriptionnelle voire post- traductionnelle a été mise en évidence pour plusieurs ET. On peut citer notamment l'effet dose décrit pour les transposons à ADN *Ac* du maïs qui arrêtent de transposer malgré la production de leur transposase (Scofield et al. 1993, Kunze et al. 1993) ou encore la régulation des rétroéléments à LTR de type COPIA Ty1 chez *S. cerevisae* par des protéines codées par le génome hôte qui seraient impliquées dans la dégradation de l'ADNc (Lesage and Todeschini 2005). Enfin, un des aspects cruciaux est la régulation tissu-spécifique de l'expression des ET. De par notre schéma expérimental nous ne nous intéressons qu'aux évènements d'insertion qui ont été transmis au travers des générations,

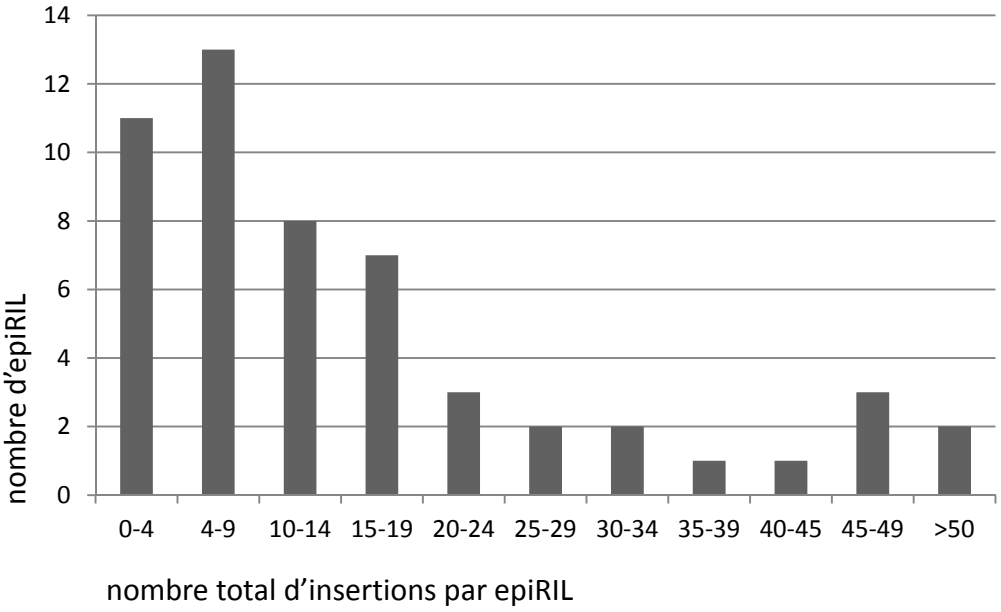


Figure 2.4 : Distribution du nombre de nouvelles insertions dans les epiRIL.

ce qui implique que les insertions ont eu lieu dans des cellules qui « contribuent » à la génération suivante. Or, les analyses transcriptionnelles dont nous disposons ont été réalisées sur plantules entières majoritairement composées de cellules somatiques. Il est possible que certains ET transcriptionnellement actifs dans les tissus somatiques soient réprimés passivement (absence de facteurs de transcription) ou activement (via le TGS ou le PTGS par exemple) dans certains tissus, notamment dans les organes reproducteurs. A ce sujet, il faut noter que, comme décrit dans l'introduction, la régulation des séquences répétées semble différer dans les gamètes et les tissus somatiques.

Il est notable que la majorité des familles présentant des éléments mobiles appartiennent à la superfamille de rétroéléments à LTR COPIA qui s'est amplifiée et diversifiée récemment dans le génome d'Arabidopsis (Peterson-Burch et al. 2004).

Nos résultats montrent également que le nombre d'insertions diffère dramatiquement entre les différentes familles pour lesquelles nous avons mis en évidence une mobilisation. De fait, la très vaste majorité (>95%) des évènements de transposition identifiés est due à la mobilisation d'ET appartenant à trois familles seulement (*ATCOPIA93, ATENSPM3* et *VANDAL21*). Cette disparité est particulièrement grande pour les 9 familles de LTR-COPIA mobiles puisque seul *ATCOPIA93* présente une fréquence de mobilisation élevée, contribuant à 65% de l'ensemble des évènements de transposition détectés dans les epiRIL. Deux raisons peuvent être évoquées ici : (i) le nombre de copies mobiles par famille est très variable, (ii) certaines familles présentent des copies extrêmement mobiles. L'identification des locus donneurs permettra de trancher entre ces hypothèses (voir plus loin).

Enfin, le nombre de nouvelles insertions par lignée est hautement variable entre les epiRIL allant de 0 à 87 (fig. 2.4). L'existence de lignées présentant un grand nombre de nouvelles insertions suggère des phénomènes de « burst » qui concernent principalement *ATCOPIA93* (cf tables en annexe 2). On notera cependant que la lignée 122 qui possède le plus de nouvelles insertions (87) le doit presqu'autant à *ATENSPM3* (32 insertions) qu'à *ATCOPIA93* (46 insertions).

Rétroéléments			
famille	donneurs	RT	LTR
ATCOPIA93	AT5TE20395 (évadé)	oui	oui
	AT1TE41580 (attrapé)	oui	oui
ATCOPIA78	AT1TE59755	oui	oui
	AT3TE89830	oui	oui
	AT5TE15240	oui	oui
	AT3TE92525	oui	oui
ATCOPIA21	AT5TE65370	oui	oui
ATCOPIA20	AT2TE13385	oui	oui
ATCOPIA31	AT1TE38210	oui	oui
ATCOPIA63	AT1TE57025	non	oui
	AT5TE33540	non	oui
ATCOPIA13	AT2TE23850+AT2TE23855	oui	oui
ATRE1	AT1TE72060	oui	oui
ATCOPIA51	AT1TE36030+AT1TE36035	oui	oui
Transposons à ADN			
famille	donneurs	Transposase	TIR
ATENSPM3	AT2TE20205	?	oui
	AT3TE62120+AT3TE62130+ AT3TE62100 (composite1)	non	oui
	AT4TE18510+AT4TE18500+ AT4TE18505 (composite2)	non	oui
VANDAL21	AT2TE42810	oui	oui
ATMU13	AT4TE23205+AT4TE23185 +AT4TE23200+AT4TE23195 +AT4TE23190	oui	oui

Figure 2.5 : Copies donneuses identifiées par TE-tracker.

2.2.2 Identification, caractérisation et propriétés des locus donneurs

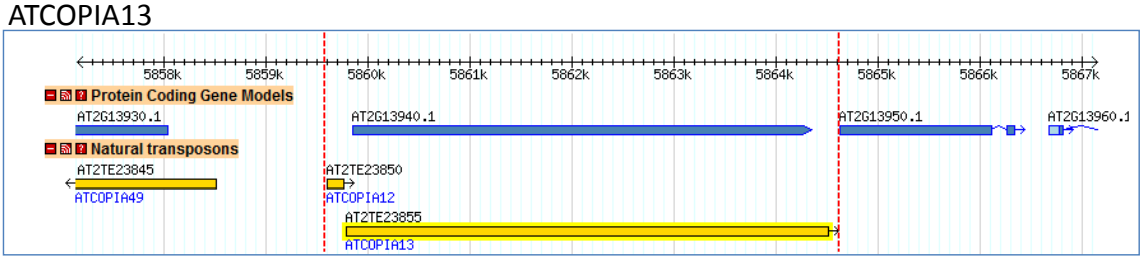
L'approche par séquençage de banques mate-pair associée au module d'attribution de score de TE-tracker permet de déterminer précisément le locus donneur pour la majorité des insertions détectées. Le nombre et l'identité des donneurs mobilisés dans les epiRIL sont résumés dans la table fig. 2.5.

Au total, nous avons identifié 19 locus donneurs avec une bonne confiance dans les epiRIL (score élevé et /ou validation par PCR-reséquençage) et leur nombre varie de 1 à 4 selon les familles. On peut donc conclure ici que les différences drastiques de fréquence de mobilisation observées entre familles ne sont pas dues à des variations majeures du nombre de copies mobilisées.

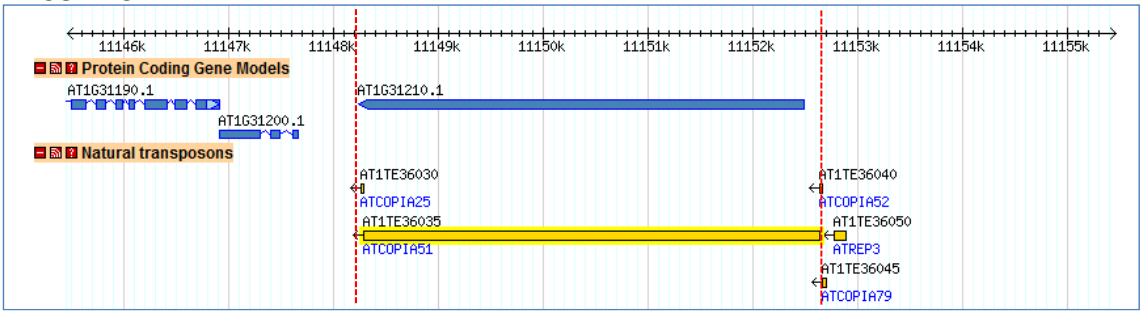
Parmi les différents locus donneurs identifiés, certains couvrent plusieurs annotations d'ET (d'après l'annotation de Buisine et al. 2008) et/ou sont plus étendus que celles-ci. Ces unités « composites » détectées comme mobiles et validées par PCR sont représentées fig. 2.6. J'ai pu identifier des structures de LTR et TIR pour les deux composites de rétroéléments à LTR et les trois composites de transposons à ADN respectivement en réalisant des BLAST des extrémités. Enfin, les deux composites *ATENSPM3* contiennent chacun un gène (*AT3G43340* et *AT4G07526)* qui n'est pas annoté comme un gène d'ET et des analyses de BLASTp et PSI-BLAST n'ont pas permis de mettre en évidence un quelconque lien entre les protéines putatives codées par ces gènes et des protéines caractéristiques d'ET connus. Cependant, très peu d'informations sur ces gènes sont disponibles et si la protéine putative codée par *AT3G43340* présente un domaine pseudouridine synthase, celle codée par *AT4G07526* ne présente aucun domaine ni fonction connus.

Afin de déterminer l'autonomie potentielle des différents locus donneurs, je me suis appuyée sur les résultats d'une étude réalisée au laboratoire basée sur la détection de RT et LTR (pour les rétroéléments) ou de transposase et TIR (pour les transposons à ADN) chez tous les ET annotés du génome d'Arabidopsis (fig. 2.5). Parmi les 14 LTR-rétroéléments identifiés, tous présentent des LTR de séquences quasi-identiques et 12 présentent une ORF codant une RT à priori intacte. Les deux autres copies donneuses appartiennent à la famille *ATCOPIA63*, et leur mobilisation résulte vraisemblablement de l'utilisation d'une RT codée soit par un autre élément de la même famille (au moins un *ATCOPIA63* code une RT intacte

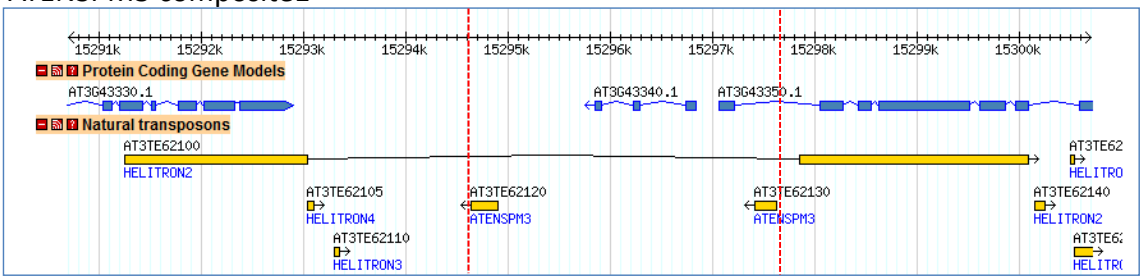
ATCOPIA13



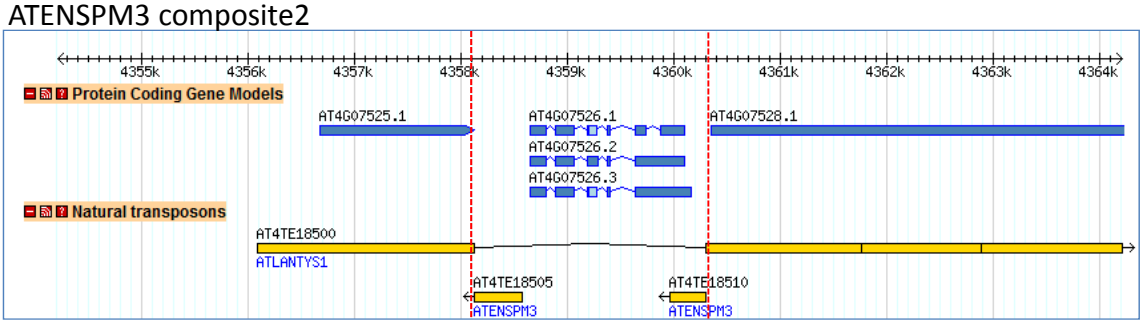
ATCOPIA51



ATENSPM3 composite1



ATENSPM3 composite2



ATMU13

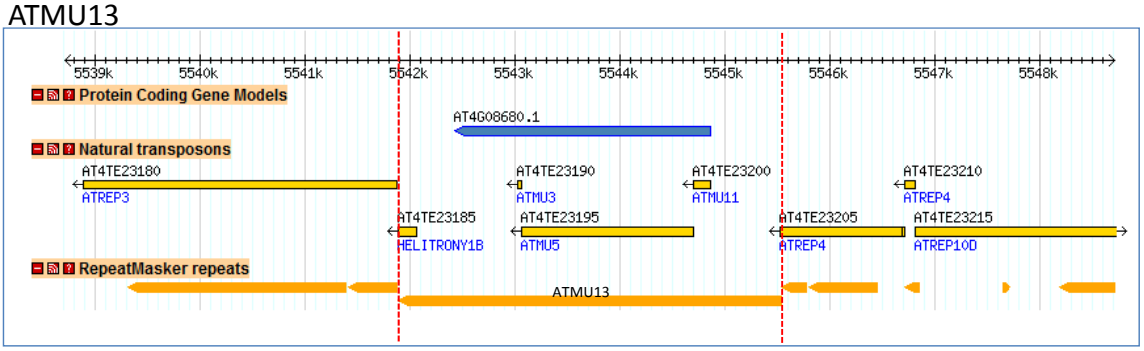


Figure 2.6 : Vues Gbrowse des cinq locus donneurs « composites ». Les régions encadrées par des pointillés rouges correspondent aux unités mobiles.

Remarque : L'annotation établie par « RepeatMasker » a été rajoutée pour le dernier composite car elle correspond exactement à la séquence mobilisée. Pour les autres composites, cette annotation n'apportait pas plus d'information que « Natural transposon ».

selon nos analyses), soit par celle d'un ET appartenant à une autre famille LTR-COPIA. Pour les transposons à ADN, tous les donneurs présentent des TIR mais seuls les locus donneurs de *VANDAL21* et *ATMU13* présentent une ORF continue. Le donneur *ATENSPM3* (locus *AT2TE20205)* présente lui une ORF interrompue ce qui suggère la présence d'introns. Cette observation est en accord avec les travaux effectués sur les éléments Spm du maïs qui avaient mis en évidence un épissage alternatif donnant lieu à plusieurs protéines (Masson et al. 1989). Enfin, il est probable que les deux composites *ATENSPM3* soient mobilisés en trans.

D'autre part, comme cela est décrit dans le manuscrit présenté ci-dessus, certaines nouvelles insertions de LTR-COPIA sont attribuées à deux donneurs avec un score élevé. Une analyse fine des lectures correspondant à certaines de ces insertions a permis de mettre en évidence leur structure chimérique (cf. figure supplémentaire 4 du manuscrit Gilly et al.). Ce résultat n'est pas inattendu car nous savons que lors de la rétrotransposition des rétroéléments à LTR deux ARN correspondant à des locus distincts peuvent être encapsidés ensemble et la RT peut changer de matrice lors de la transcription inverse.

Dans un premier temps, j'ai comparé la liste des locus donneurs dans les epiRIL avec la liste des copies potentiellement mobiles pour chacune des familles mobilisées en me focalisant sur les copies potentiellement autonomes (d'après les analyses réalisées au laboratoire non montrées). Pour les rétroéléments qui présentaient des erreurs d'annotation (*ATCOPIA13* et *ATCOPIA51*) j'ai étendu la recherche de LTR en dehors de l'annotation pour la détermination des locus potentiellement mobiles (fig. 2.7). Pour la majorité des familles, il existe plus d'éléments potentiellement mobiles autonomes que ceux qui transposent dans les epiRIL et les raisons de cette différence restent énigmatiques. Le cas le plus frappant est celui d'*ATCOPIA93*, qui présente trois copies potentiellement mobiles dont deux d'entre elles sont identiques à plus de 99% et sont connues sous les noms d'*EVADE* et *ATTRAPE*. Bien que ces deux copies soient mobiles dans les epiRIL, *ATTRAPE* n'a été attribué qu'à un seul évènement de transposition contre plusieurs centaines pour *EVADE*. C'est d'ailleurs la première fois que la mobilisation d'*ATTRAPE* a été détectée car l'étude de la mobilisation des éléments *ATCOPIA93* dans divers mutants n'avait mis en évidence que la mobilisation d'*EVADE* (Mirouze et al. 2009).

Rétroéléments		
famille	Copies mobilisées autonomes (LTR+RT)	Copies potentiellement mobiles et autonomes (RT+LTR)
ATCOPIA93	2	3
ATCOPIA78	4	8
ATCOPIA21	1	1
ATCOPIA20	1	1
ATCOPIA31	1	2
ATCOPIA63	0	1
ATCOPIA13	1	4
ATRE1	1	2
ATCOPIA51	1	4
Transposons à ADN		
famille	Copies mobilisées autonomes (transposase+TIR)	Copies potentiellement mobiles et autonomes (transposase+TIR)
VANDAL21	1	6
ATENSPM3	?	?
ATMU13	1	?

Figure 2.7 : Nombre de copies mobiles autonomes et de copies apparentées potentiellement mobiles et autonomes.

	FAMILLE	TOTAL	UNIQUES	PARTAGEES
Rétroélements	ATCOPIA93	454	452	2
	ATCOPIA78	7	0	7
	ATCOPIA63	4	4	0
	ATCOPIA21	2	2	0
	ATCOPIA13	2	2	0
	ATRE1	1	1	0
	ATCOPIA20	1	0	1
	ATCOPIA31	1	1	0
	ATCOPIA51	1	1	0
Transposons à ADN	ATENSPM3	136	125	11
	VANDAL21	90	89	1
	ATMU13	8	8	0
TOTAL		707	685	22

Nombre faible d'insertions, toutes partagées
Nombre élevé d'insertions, la plupart uniques
Nombre faible d'insertions, toutes uniques

Figure 2.8 : Dynamique de mobilisation des différentes familles d'ET dans les epiRIL.

2.2.3 Dynamique de transposition des ET dans les epiRIL

Afin d'estimer plus finement la dynamique de mobilisation pour les différentes familles d'ET, j'ai déterminé pour chacune d'entre elles le nombre d'insertions partagées (qui ont donc eu lieu soit dans le mutant *ddm1* soit dans la plante F1 unique à partir de laquelle a été réalisé le croisement en retour à l'origine de chacune des epiRIL), ainsi que le nombre d'insertions uniques (fig. 2.8). Comme attendu au vu du faible nombre d'insertions détectées dans le mutant *ddm1*, la très grande majorité des insertions identifiées dans les epiRIL est spécifique à chacune des lignées. Ceci permet de conclure définitivement que les ET sont toujours actifs dans les epiRIL après ségrégation de la mutation *ddm1*.

Néanmoins, la dynamique de mobilisation diffère selon les familles. On peut distinguer : (1) les familles d'ET présentant un petit nombre d'insertions toutes partagées (*ATCOPIA78* et *ATCOPIA20*) indiquant une mobilisation exclusivement dans *ddm1* ou la F1, (2) les familles présentant un nombre élevé d'insertions dont une majorité sont non partagées entre epiRIL (*ATCOPIA93*, *ATENSPM3* et *VANDAL21*) ce qui témoigne d'une mobilisation dans *ddm1* et /ou la F1 et dans les épiRIL, (3) les familles d'ET ayant un faible nombre d'insertions toutes non partagées, révélant une mobilisation exclusivement dans les epiRIL. Parmi elles, certaines présentent des nouvelles insertions dans plusieurs epiRIL (*ATMU13*, *ATCOPIA21* et *ATCOPIA12*) et d'autres dans une seule (*ATCOPIA63*, *ATCOPIA51*, *ATCOPIA31* et *ATRE1*).

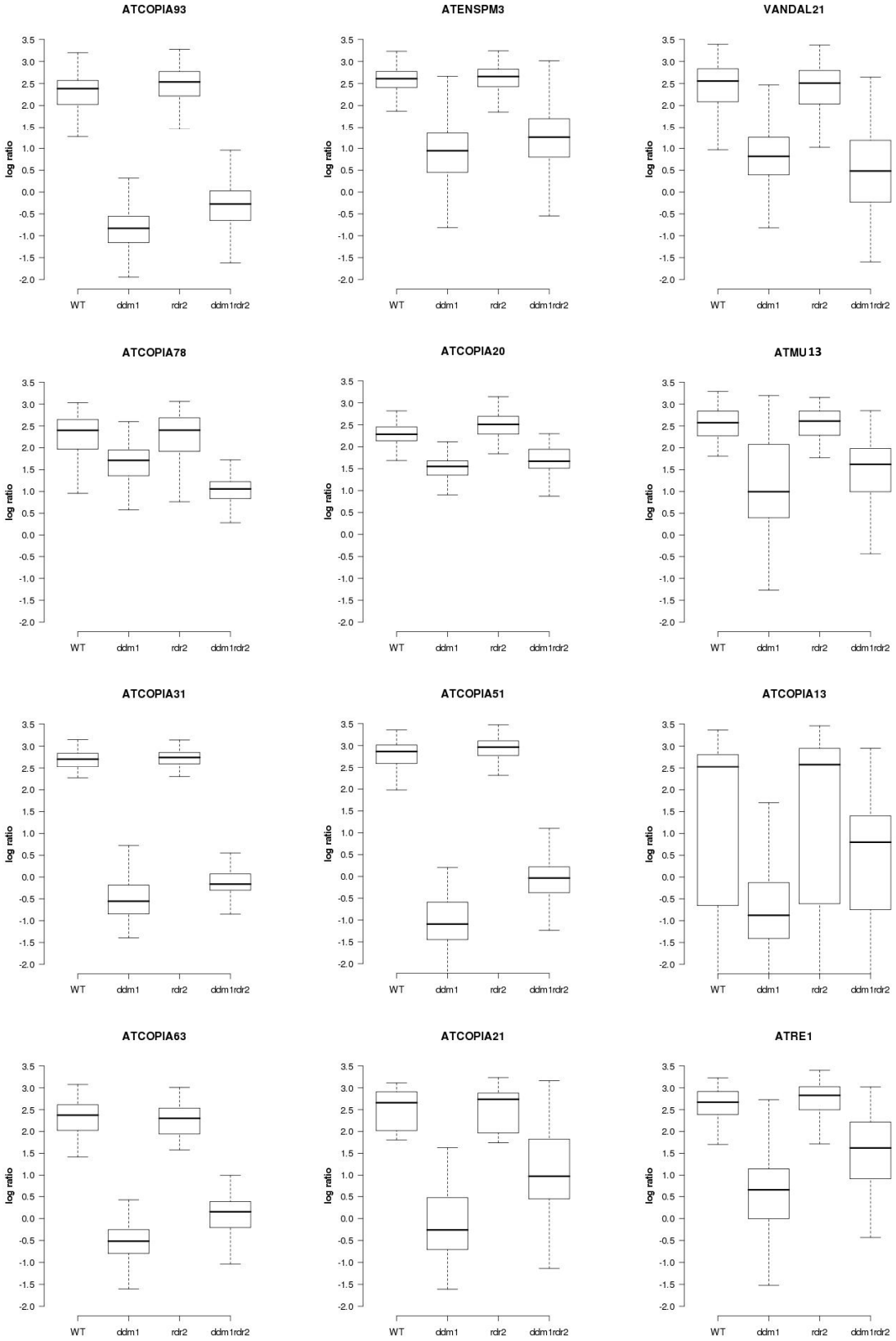


Figure 2.9 : Distribution chez le sauvage et les mutants *ddm1*, *rdr2* et *ddm1rdr2,* du niveau de méthylation (MeDIP-chip) rapporté par les sondes couvrant les séquences des différentes familles d'ET pour lesquelles de nouvelles insertions ont été détectées.

2.2.4 Etude comparative des différentes familles comptant des ET mobiles dans les epiRIL

Afin de mieux comprendre les différences de fréquence et de dynamique de mobilisation des ET appartenant aux différentes familles dans les epiRIL, j'ai étudié l'implication des différentes voies de la méthylation de l'ADN (dépendante de DDM1 et RdDM) dans le contrôle de ces ET. Pour ce faire, j'ai réalisé des expériences d'immunoprécipitation des cytosines méthylées suivies d'hybridation sur puce à ADN (MeDIP-chip) chez le sauvage et le mutant *ddm1* ainsi que chez le mutant *rdr2* et le double mutant *ddm1rdr2*. J'ai également tiré avantage de données de séquençage profond de petits ARN dans ces mêmes fonds génétiques ainsi que du transcriptome de *ddm1* (ARN-chip) disponibles au laboratoire.

Les analyses présentées ci-dessous ont été effectuées à l'échelle de la famille d'ET car les techniques de MeDIP-chip et d'ARN-chip ne permettent pas, dans la plupart des cas, d'obtenir des données spécifiques pour chaque copie au sein d'une famille à cause des trop grandes similarités de séquence.

Comme attendu, toutes les familles comptant des ET mobiles dans les epiRIL présentent une perte de méthylation dans le mutant *ddm1* en comparaison avec le sauvage (fig. 2.9). Cependant, l'intensité de cette perte varie entre les familles. Notamment, les familles *ATCOPIA78* et *ATCOPIA20* qui n'ont engendré que des insertions partagées présentent une perte de méthylation moindre dans le mutant *ddm1* que toutes les autres familles mobilisées. Cette observation indique donc l'existence de mécanismes capables d'assurer le maintien de la méthylation en l'absence de l'action de DDM1 pour ces familles. Il faut néanmoins considérer ici que ce type d'analyses ne permet pas de déterminer si une perte partielle de méthylation observée à l'échelle d'une famille d'ET est due à la perte complète au niveau de certains membres de la famille alors que les autres membres gardent un niveau élevé de méthylation ou si cela traduit une perte partielle concertée au niveau de tous les membres.

Les résultats de l'analyse de méthylation dans le mutant *rdr2* montrent que la compromission du RdDM seul n'engendre pas de perte de méthylation globale pour aucune des familles. Il faut noter ici que le MeDIP-chip ne permet pas de mettre en évidence une perte subtile de méthylation. Or, la voie du RdDM est responsable du maintien de la méthylation en contexte CHH qui est de faible intensité (40% maximum des cytosines sont

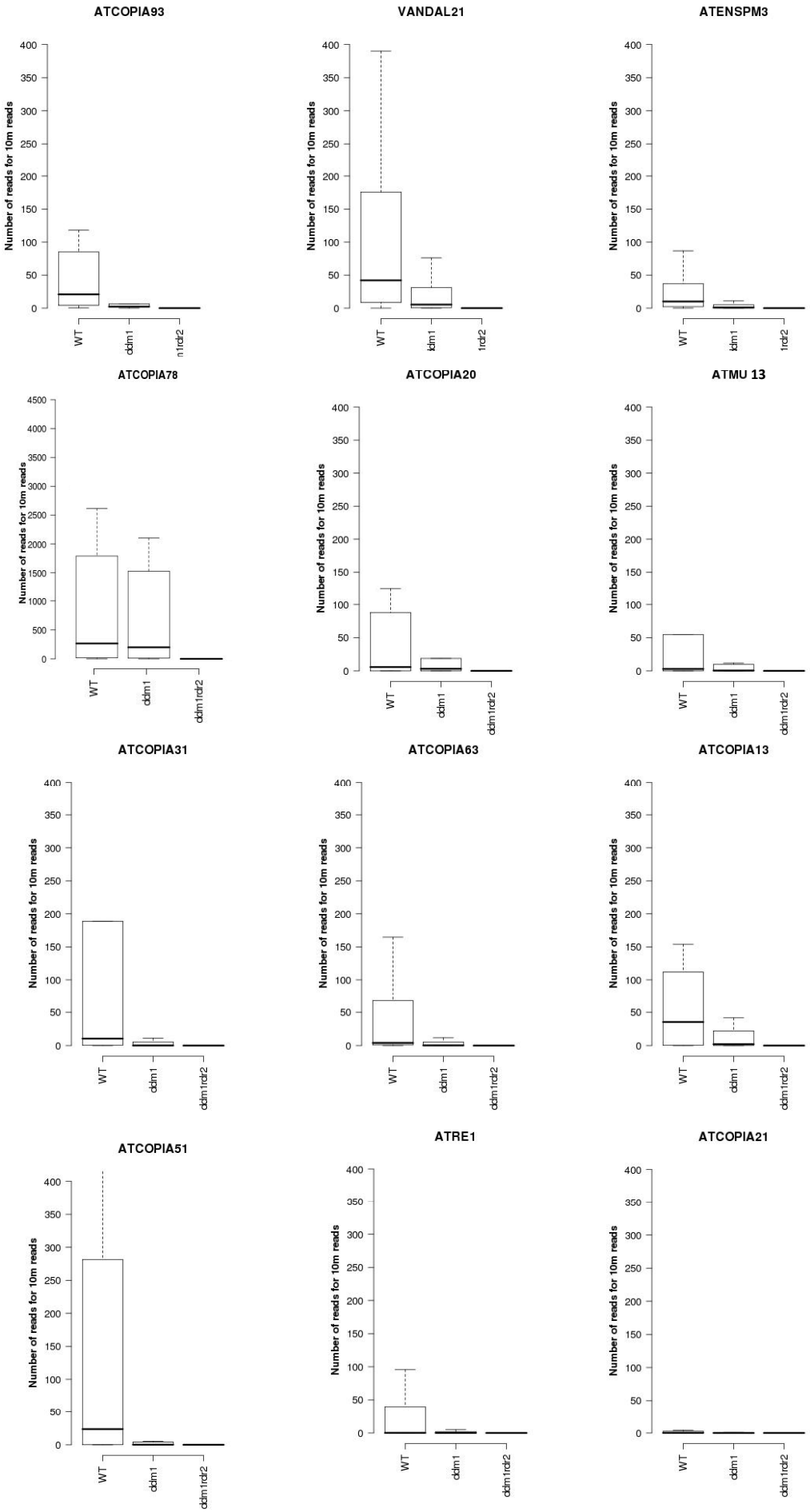


Figure 2.10 : Abondance chez le sauvage et les mutants *ddm1* et *ddm1rdr2* des siARN de 24nt ciblant les différentes familles d'ET pour lesquelles de nouvelles insertions ont été détectées. A noter que l'échelle des ordonnées est différente pour *ATCOPIA78.*

méthylées à une position donnée). Cependant, l'importance du RdDM dans le contrôle de la méthylation de la famille *ATCOPIA78* est révélée dans le double mutant *ddm1rdr2* qui présente, pour ces séquences, une perte de méthylation plus importante que dans le simple mutant *ddm1*. Pour cette famille, il est donc clair que c'est l'action du RdDM qui lui permet de garder un certain niveau de méthylation dans le mutant *ddm1*. En revanche, *ATCOPIA20* ne présente pas une perte de méthylation plus importante dans le double mutant *ddm1rdr2*, ce qui suggère que le maintien d'un niveau de méthylation relativement élevé dans *ddm1* pour les membres de cette famille n'est pas dû à l'action du RdDM. Enfin, pour la majorité des familles, le niveau de méthylation dans le double mutant *ddm1rdr2* bien que très faible, apparait légèrement supérieur à celui mesuré dans le simple mutant *ddm1*. Ceci est probablement dû au fait que le double mutant *ddm1rdr2* n'a subi que deux cycles d'autofécondation successives alors que le simple mutant *ddm1* en a subi quatre, or l'hypométhylation induite par la mutation *ddm1* a tendance à s'aggraver au cours des générations.

Les données issues du séquençage profond des petits ARN sont en accord avec celles concernant la méthylation de l'ADN (fig 2.10). *ATCOPIA78* est dix fois plus ciblée en moyenne par les siARN de 24nt que les autres familles considérées ici. De plus, elle est la seule famille à conserver quasi totalement ses siARN dans le mutant *ddm1* alors que les autres familles les perdent drastiquement voire totalement. En revanche, et comme attendu, on observe une perte totale des siARN de 24nt ciblant les *ATCOPIA78* dans le double mutant *ddm1rdr2*. Ces observations expliquent donc pourquoi *ATCOPIA78* est la seule famille à présenter une perte de méthylation plus importante dans *ddm1rdr2* que dans *ddm1* et pourquoi la perte de méthylation dans *ddm1* est aussi forte pour les autres familles.

Enfin, l'analyse différentielle d'expression par ARN-chip dans le mutant *ddm1* révèle des comportements contrastés entre les différentes familles comptant des d'ET mobilisés (fig. 2.11). Aucune corrélation ne peut être mise en évidence entre le niveau d'activation transcriptionnelle de ces familles dans *ddm1* et les fréquences et dynamiques de mobilisation. Une fois encore, il faut prendre en considération que les expériences d'ARN-chip sont effectuées sur plantules entières, or c'est l'expression dans les tissus qui donneront lieu à la génération suivante qui importe. Si par exemple les évènements de

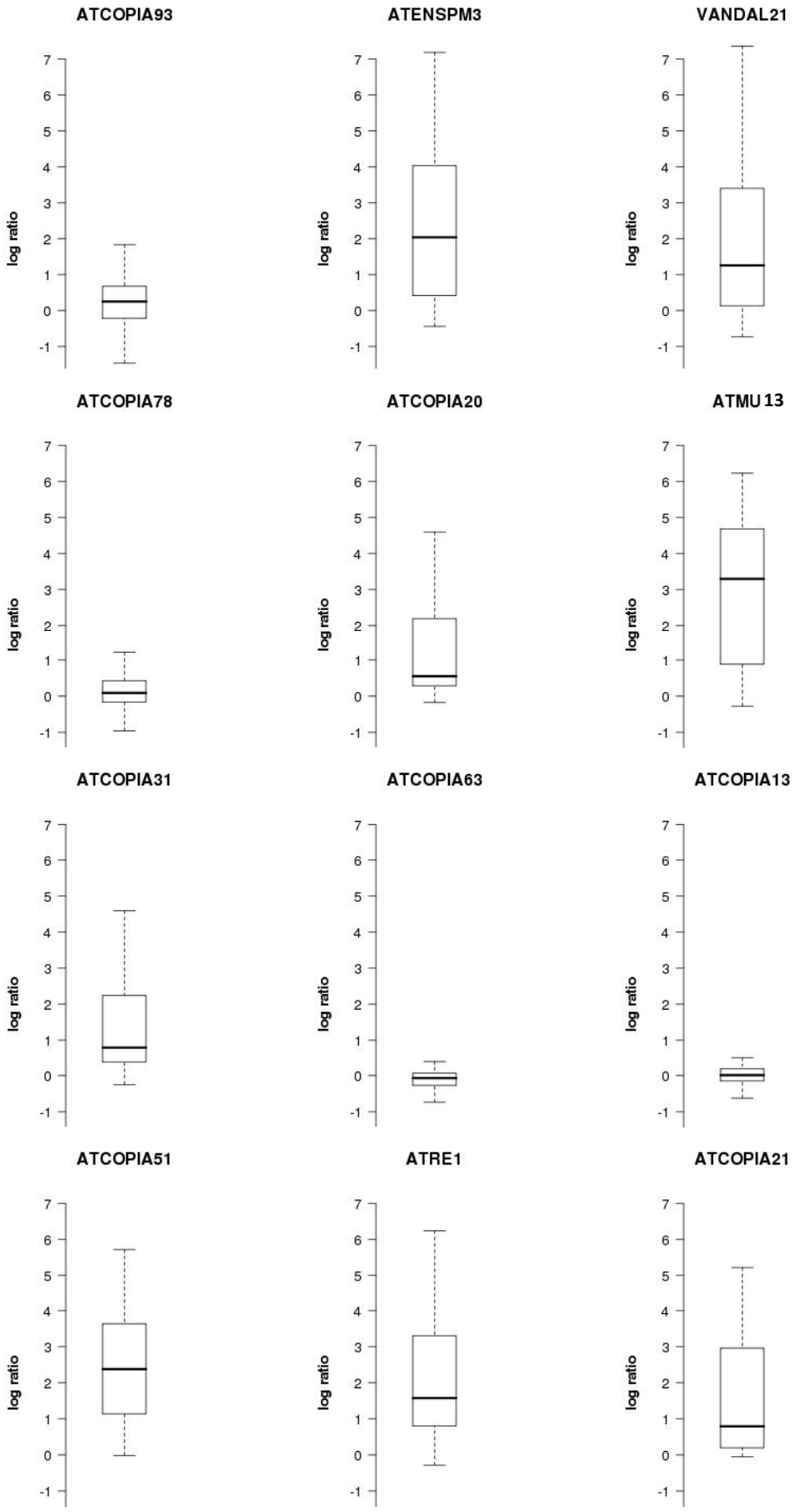


Figure 2.11 : Distribution du niveau d'expression différentielle (ARN-chip) rapportée par les sondes couvrant les séquences des différentes familles d'ET pour lesquelles des nouvelles insertions ont été détectées entre *ddm1* et sauvage.

mobilisation que nous avons détectés sont dus à une activation spécifique dans les gamètes ou les inflorescences, elle ne sera pas détectée ici.

Dans l'ensemble, les analyses de la méthylation de l'ADN, de l'expression et de l'abondance en siARN de 24nt dans les différents fonds mutants ne permettent pas d'expliquer les différences drastiques de fréquence et dynamique de mobilisation observées entre les différentes familles d'ET. Néanmoins, les familles responsables d'insertions toutes partagées ne présentent qu'une perte modérée de méthylation dans le mutant *ddm1* en comparaison avec les autres familles. Pour le cas d'*ATCOPIA78*, cela s'explique clairement par un contrôle de la méthylation par la voie du RdDM. Cette action forte du RdDM sur cette famille pourrait également expliquer le fait que la mobilisation ne soit possible que dans *ddm1* ou dans la F1. De fait, un ciblage fort par la machinerie du RdDM chez le sauvage et qui perdure chez le mutant *ddm1* est la signature caractéristique du phénomène de reméthylation décrit dans l'introduction. Il est donc probable que les copies résidentes d'*ATCOPIA78* soient efficacement reméthylées donc réinactivées après restauration de la fonction DDM1, mais également que les nouvelles insertions soient ciblées par ce mécanisme et donc rapidement réprimées également. Ce point sera développé plus en détail dans la partie 3.

Enfin, s'il est vrai que les familles d'ET qui comptent les copies les plus mobiles montrent une forte perte de méthylation dans *ddm1*, cela est vrai aussi pour la majorité de celles présentant un faible nombre d'insertions toutes non partagées. Concernant ces dernières, une mobilisation restreinte à un très faible nombre de lignées pourrait être due à un défaut occasionnel de reméthylation dans ces lignées spécifiquement. De fait, le mécanisme de reméthylation présente une pénétrance variable selon les séquences. Afin de tester cette hypothèse, j'ai évalué l'état de méthylation d'un de ces éléments, *AT2TE23855 (ATCOPIA13),* dans les epiRIL à partir des données de MeDIP-chip disponibles au laboratoire. J'ai dû me restreindre à cet élément car c'est le seul qui présente des sondes suffisamment spécifiques sur la puce à ADN utilisée pour le MeDIP-chip. Bien que n'ayant transposé que dans deux epiRIL, *AT2TE23855* est dans un état hypométhylé dans plusieurs autres lignées, et cela dans les proportions attendues dans le cas de la ségrégation mendélienne de locus stablement hypométhylés (dans 25% des epiRIL). L'absence de transposition dans ces dernières lignées reste donc énigmatique mais pourrait être liée à la ségrégation de facteurs agissant en trans.

	ATCOPIA93 (AT5TE20395)			VANDAL21 (AT2TE42810)			ATENSPM3 (AT2TE20205)		
epiRIL	ORIGINE PARENTALE	INS. PARTAGEE	INS. UNIQUE	ORIGINE PARENTALE	INS. PARTAGEE	INS. UNIQUE	ORIGINE PARENTALE	INS. PARTAGEE	INS. UNIQUE
8	ND		2	WT		3	WT	2	
14	WT			DDM1		1	DDM1		
24	WT			WT			WT		
36	WT		4	DDM1		5	WT		
46	WT			WT			WT	1	
52	WT			WT	1	1	WT	1	1
53	WT			WT			WT		
55	DDM1	1	24	WT		5	DDM1		11
60	WT	1	1	DDM1	1	4	WT		3
69	WT			WT		4	WT		3
70	ND		3	WT		2	WT		
92	DDM1	1	5	WT	1	5	DDM1		
95	DDM1	1	22	WT	1		WT		8
98	DDM1	1	8	DDM1		5	DDM1	2	
99	WT			WT			WT		
108	WT		2	WT	1		WT	1	
118	WT			WT			WT		
122	DDM1	1	46	ND	1	1	DDM1		32
137	WT			ND			WT		3
144	WT			WT		3	WT		2
150	WT		1	WT		1	WT		
159	WT		22	WT		2	WT		1
166	WT			WT			WT		
193	DDM1	1	23	ND		5	DDM1		
218	WT			DDM1	1		DDM1		
222	WT			WT			WT	1	
252	WT			WT		2	WT		1
260	ND			WT		8	WT		
276	WT	1	3	ND		5	WT	1	
277	WT			WT			WT		4
340	WT			DDM1		2	WT	1	
344	WT			WT			WT		
356	DDM1		40	WT			WT	3	
362	ND	2	37	ND		2	ND		
363	ND		12	WT			WT	1	
366	DDM1		15	WT			WT	1	5
368	WT			WT			WT		
371	WT	1	1	WT	1	7	WT	1	
375	WT	1	9	WT			WT		
377	ND		41	ND	1	2	ND		
393	DDM1		10	WT			WT		
408	DDM1		7	WT			WT		8
421	DDM1		18	DDM1		3	DDM1		
425	DDM1		42	WT		3	WT	1	
437	DDM1		10	WT			WT		
439	ND		3	DDM1		2	WT		4
454	DDM1		24	WT	1	1	WT	2	
458	WT	1	5	DDM1		2	WT		
466	WT			WT			DDM1		
467	WT	1	1	WT			WT		
480	WT			WT			WT		
503	WT	1	11	WT		2	WT		3
559	WT			WT		1	WT		

Figure 2.12 : Tableau récapitulatif des insertions d'*AT5TE20395*, *AT2TE42810* et *AT2TE20205* et origine parentale du locus donneur dans les epiRIL.

2.2.5 Mobilisation des ET en fonction de l'origine parentale du locus donneur et de la présence d'insertions partagées.

Pour les trois ET potentiellement autonomes les plus mobiles *AT5TE20395* (*ATCOPIA93*), *AT2TE20205* (*ATENSPM3*) et *AT2TE42810* (*VANDAL21*) j'ai cherché à déterminer leur « potentiel de mobilisation », autrement dit à établir dans quelle mesure la présence d'une copie potentiellement active dans les epiRILs (locus donneur hérité du parent *ddm1* et /ou insertions partagées héritées du parent *ddm1* ou de la plante F1) se traduit par la mobilisation de ces ET. Pour ce faire, j'ai mis en relation de façon systématique l'origine parentale du locus donneur et la présence d'insertions partagées avec la présence d'insertions uniques pour chacune des epiRIL. Pour établir l'origine parentale des locus donneurs je me suis appuyée sur les cartes de recombinaison des epiRILs établies au laboratoire (annexe 3 Colome-Tatche et al. 2012). Les résultats de cette analyse sont regroupés figure 2.12.

- *ATCOPIA93*

Il apparaît très clairement que dans le cas d'*ATCOPIA93*, la présence du locus donneur dans un intervalle d'origine *ddm1* (donc théoriquement hypométhylé) et/ou la présence de l'insertion partagée est systématiquement associée à la présence d'insertions uniques. Cependant, des insertions uniques sont également détectées dans des epiRIL ne présentant pas ces caractéristiques. Cette dernière observation pourrait refléter soit une mobilisation dans les premières générations des epiRIL avant fixation à l'état homozygote de la copie *AT5TE20395* d'origine sauvage (fig. 2.13 (1)), soit l'héritage à l'état hémizygote de la copie partagée pendant quelques générations avant élimination par ségrégation (fig. 2.13 (2)). Dans les deux cas, la présence de copies actives a pu engendrer l'apparition de nouvelles insertions uniques. Ces hypothèses pourront être testées au moyen d'études verticales au sein de chaque lignée en y analysant à chaque génération l'état de méthylation du locus donneur et la présence de l'insertion partagée.

- *VANDAL21*

Dans le cas de *VANDAL21* la tendance concernant l'origine parentale du donneur est la même que celle observée pour *ATCOPIA93*, à savoir la présence quasi-systématique (8 cas

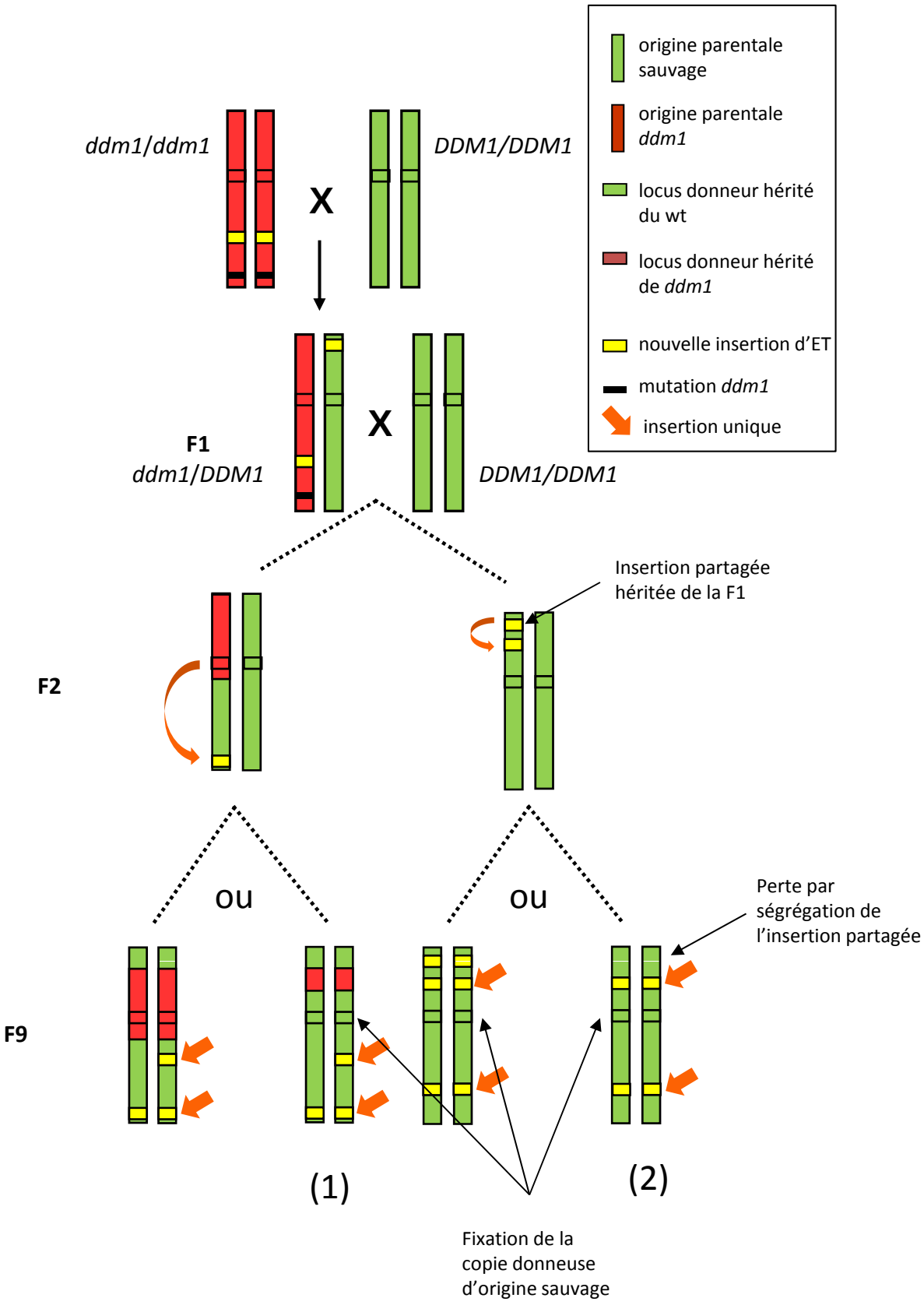


Figure 2.13 : Origines possibles des insertions uniques dans les epiRIL possédant un locus donneur hérité du parent sauvage et aucune insertion partagée (cas 1 et 2).

sur 9) de nouvelles insertions uniques quand le donneur est hérité de *ddm1*. En revanche, il n'en va pas de même pour le cas de la présence de l'insertion partagée qui n'est pas systématiquement associée à celle d'insertions uniques.

- *ATENSPM3*

Enfin, dans le cas d'*ATENSPM3*, la tendance est complètement inversée car l'origine *ddm1* du locus donneur est corrélée négativement avec la présence de nouvelles insertions uniques (seulement deux cas sur neuf). L'hypothèse la plus triviale pouvant expliquer cette observation est directement liée au mode de transposition d'*ATENSPM3* par un mécanisme de couper/coller. De fait, si la mobilisation d'un transposon à ADN est suivie de la réparation du site donneur par « non-homologous end joining » et que la nouvelle insertion n'est pas fixée dans la descendance, alors on a disparition de l'ET dans la lignée. Afin de tester cette hypothèse j'ai réalisé des PCR pour la détection de la présence d'*AT2TE20205* au locus donneur dans les epiRIL (fig. 2.14). *AT2TE20205* est systématiquement absent au locus donneur lorsqu'il a été hérité du parent *ddm1,* ce qui traduit une excision vraisemblablement ancestrale et héritée du parent *ddm1*. Il est également remarquable que la présence d'une ou plusieurs insertions partagées est négativement corrélée avec la présence de nouvelles insertions uniques dans le cas d'*ATENSPM3*. S'il est vrai que la mobilisation des SPM se traduit le plus souvent par un site donneur vide après excision, la mobilisation des insertions partagées dans chaque lignée qui produit des insertions uniques doit souvent être associée avec la « perte » de ces insertions partagées. Cette hypothèse peut être testée en recherchant dans chaque lignée présentant des insertions uniques d'*ATENSPM3* la présence d'empreintes attestant de la présence antérieure des insertions partagées. De fait, j'ai déjà mis en évidence un premier cas dans l'epiRIL 439 qui présente la signature d'excision d'une des insertions partagées (fig. 2.15). Quoiqu'il en soit, cela n'explique pas la raison pour laquelle tant pour *VANDAL21* qu'*ATENSPM3* les insertions partagées sont mobilisées dans certaines lignées mais pas dans d'autres.

En résumé il apparaît donc que la présence d'une ou plusieurs copies potentiellement mobiles d'*AT5TE20395* est systématiquement associée avec sa prolifération dans chacune des epiRIL bien qu'à des fréquences très variées (le nombre d'insertions uniques est très variable selon les lignées). En revanche, ce n'est pas le cas pour *AT2TE20205* et *AT2TE42810*

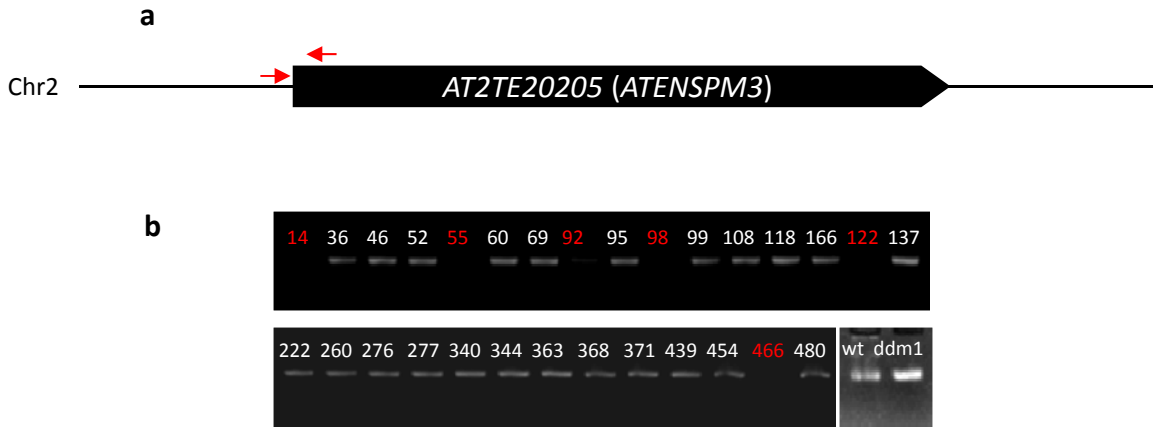
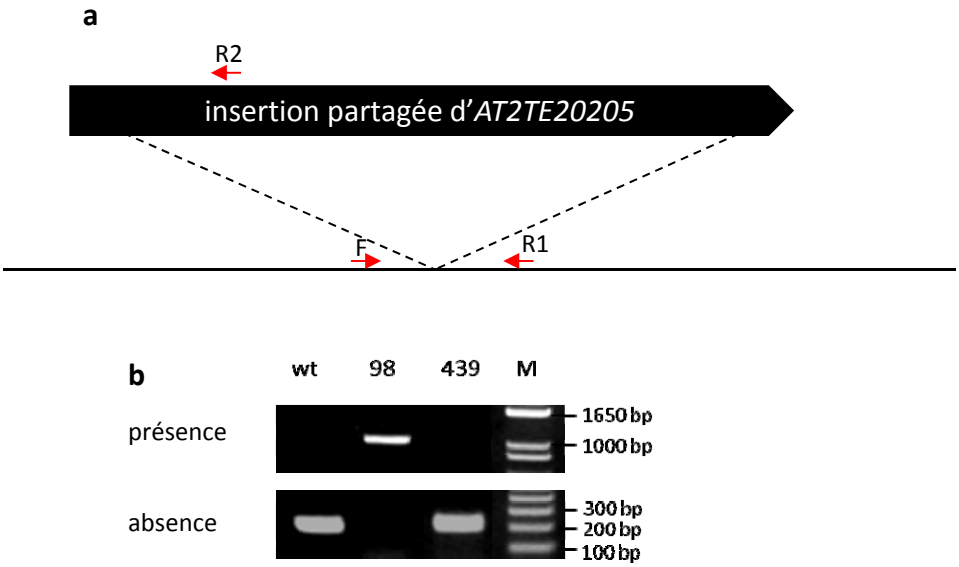


Figure 2.14 : Test de la présence d'*AT2TE20205* (*ATENSPM3*) au locus donneur.
(**a**) Représentation schématique du positionnement des amorces de PCR utilisées pour tester la présence d'*AT2TE20205* au locus donneur (l'échelle n'est pas respectée). (**b**) PCR montrant la présence/absence d'*AT2TE20205* au locus donneur. Les lignées en rouge ont le locus donneur dans un intervalle hérité du parent *ddm1*.



c chr4 : 16482535-16482686

```
ref  CTAGTGCGAATCCAGTCTGTCAGACAATGAGTAAACGAGTGACATGTCATCTCACGCATA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
439  CTAGTGCGAATCCAGTCTGTCAGACAATGAGTAAACGAGTGACATGTCATCTCACGCATA

ref  CTTATTCCCTGAGAAGA----TAATTATTAACATTAACATCGAATGAATTTAAAAATAAG
     |||||||||||||||||    ||||||||||||||||||||||||||||||||||||||
439  CTTATTCCCTGAGAAGATAATTAATTATTAACATTAACATCGAATGAATTTAAAAATAAG

ref  TATTACAAGTACCAAATAGTATTAAGCAGAATTTAT
     |||||||||||||||||||||||||||||||||||
439  TATTACAAGTACCAAATAGTATTAAGCAGAATTTAT
```

Figure 2.15 : Excision d'une insertion partagée d'*AT2TE20205* (*ATENSPM3*) dans l'epiRIL 439.
(**a**) Représentation schématique du positionnement des amorces PCR utilisées pour montrer l'absence (F+R1) ou la présence (F+R2) d'une insertion partagée d'*AT2TE20205* (l'échelle n'est pas respectée). (**b**) PCR montrant la présence *vs* absence de l'insertion partagée dans le sauvage, l'epiRIL 439 et l'epiRIL 98 qui seule présente l'insertion partagée. (**c**) Résultat du séquençage du produit de PCR correspondant à l'absence de l'insertion dans l'epiRIL 439 aligné sur le génome de référence. Les nucléotides en bleu correspondent à l'empreinte laissée par l'excision d'*AT2TE20205*.

car certaines lignées présentant des insertions partagées de ces ET ne possèdent aucune insertion unique. Cela peut indiquer que ces insertions partagées ne sont plus actives dans ces lignées.

2.2.6 Etude de l'activation en trans des transposons à ADN *ATENSPM3* et *VANDAL21*

L'absence de la mobilisation du locus donneur *AT2TE20205* lorsqu'il est hérité du sauvage alors que de nouvelles insertions, potentiellement actives, sont présentes peut sembler contradictoire avec des travaux effectués sur les éléments Spm du maïs qui mettaient en évidence la réactivation de copies silencieuses (alors appelées cryptiques) lorsque mises en présence de copies actives (Fedoroff 1989). Cette réactivation passait notamment par une étape de déméthylation de la copie cryptique grâce à l'action de la protéine TNPA codée par la copie active. Il est possible que les éléments *ATENSPM3* d'Arabidopsis ne présentent pas cette propriété, ce qui expliquerait l'absence de mobilisation de la copie donneuse héritée du parent sauvage. Afin de tester cette hypothèse, j'ai mesuré par McrBC-qPCR le niveau de méthylation de la copie au site donneur lorsqu'il est hérité du parent sauvage dans des epiRIL présentant ou non des nouvelles insertions d'*AT2TE20205* ainsi que chez la plante sauvage et le mutant *ddm1* (fig. 2.16). Plusieurs epiRIL, bien qu'ayant hérité le locus donneur du parent sauvage, présentent une hypométhylation drastique de son extrémité 5'. Cette hypométhylation n'est observée que pour des epiRIL présentant des nouvelles insertions d'*At2TE20205,* ce qui suggère qu'elle est bien causée par ces nouvelles copies. Cependant, toutes les epiRIL présentant des nouvelles insertions ne montrent pas l'hypométhylation de l'extrémité 5', ce qui peut signifier soit que ce phénomène n'a pas une pénétrance complète soit que, dans ces lignées, les nouvelles insertions ne sont pas/plus actives (voir partie 3). Notons qu'il n'y a pas de lien apparent entre l'hypométhylation du locus donneur et le nombre ou l'ancienneté (unique vs partagée) des nouvelles insertions (fig. 2.12). Quoiqu'il en soit, il semble que cette hypométhylation induite en trans ne se traduise pas par la mobilisation car aucune excision du locus donneur *AT2TE20205* hérité du sauvage n'a été observée.

J'ai étendu cette analyse à *VANDAL21* bien que ce phénomène d'hypométhylation en trans n'est jamais été mis en évidence pour les éléments de type Mu. Les résultats de cette

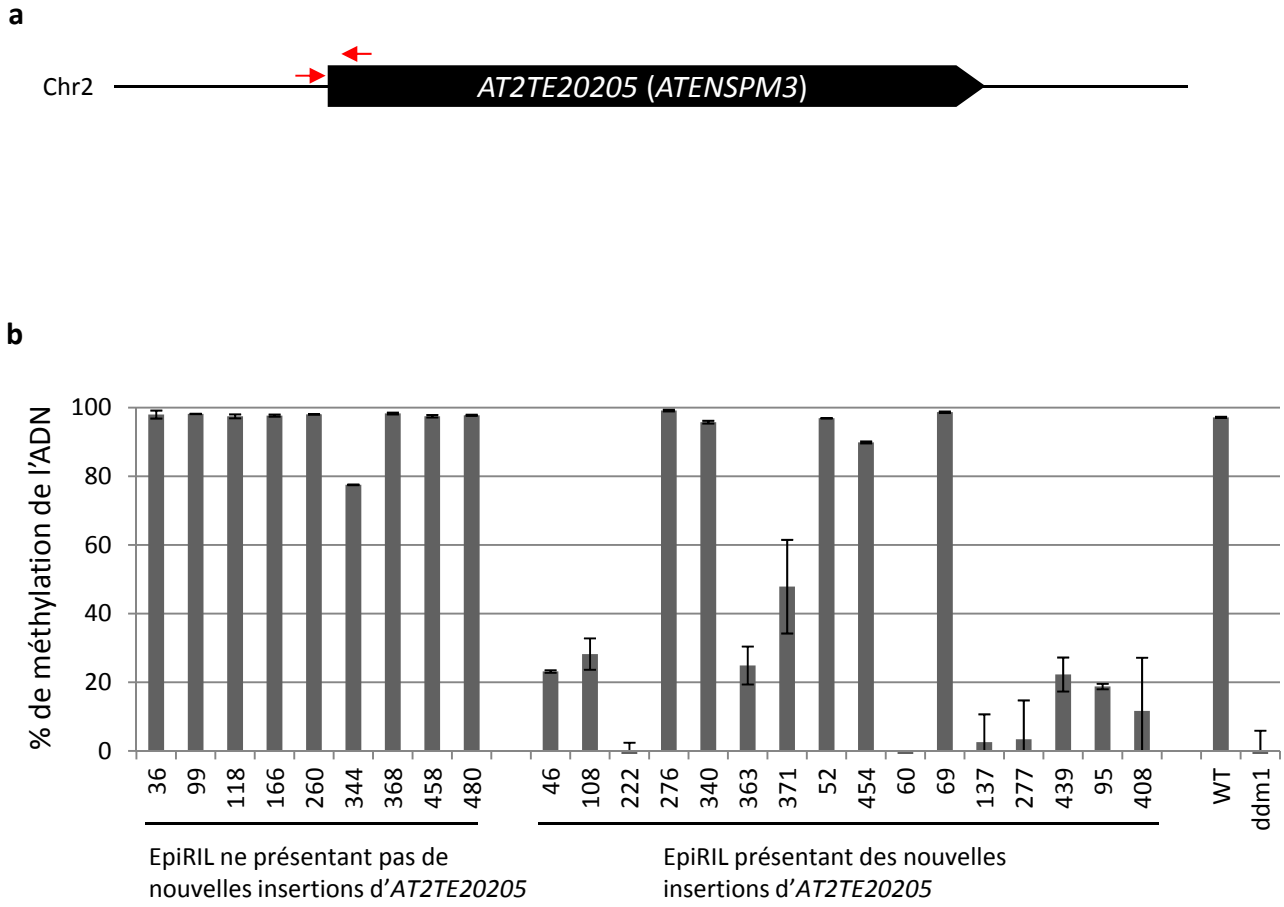


Figure 2.16 : Les nouvelles insertions d'*AT2TE20205* (*ATENSPM3*) induisent l'hypométhylation du locus donneur.

(**a**) Représentation schématique du positionnement des amorces de qPCR utilisées pour mesurer le niveau de méthylation en 5' du locus donneur *AT2TE20205* (les échelles ne sont pas respectées). (**b**) % de méthylation (McrBC-qPCR) de l'extrémité 5' du locus donneur en fonction de la présence ou non de nouvelles insertions. Les epiRIL présentant des nouvelles insertions sont ordonnées par ordre croissant de celles-ci. Les barres d'erreur représentent l'écart type entre deux répliques techniques.

analyse montrent clairement l'hypométhylation systématique du locus donneur hérité du parent sauvage lorsque de nouvelles insertions sont présentes (fig. 2.17). Ces analyses ont été valorisées dans une publication réalisée dans le cadre d'une collaboration avec l'équipe du Dr. T Kakutani (voir en annexe 4 : Fu et al, 2013) qui a notamment montré que cette hypométhylation est induite en trans par une protéine codée par *VANDAL21* et qu'elle est transitoire car le locus reméthyle après ségrégation de la copie active. Ce phénomène engendre la réactivation transcriptionnelle de la copie cryptique et est systématiquement associé à son excision dans les tissus somatiques (mais probablement dans une petite fraction de cellules car elle n'a pu être mise en évidence que par « nested PCR ») (fig. 2.17). En revanche la transmission germinale de l'excision du locus donneur n'a été observée que dans deux epiRIL (fig. 2.17). Notons ici que l'aspect systématique du phénomène d'hypométhylation en trans du locus donneur hérité du sauvage lorsque de nouvelles copies sont présentes suggère qu'au moins une des nouvelles insertions est active dans ces lignées.

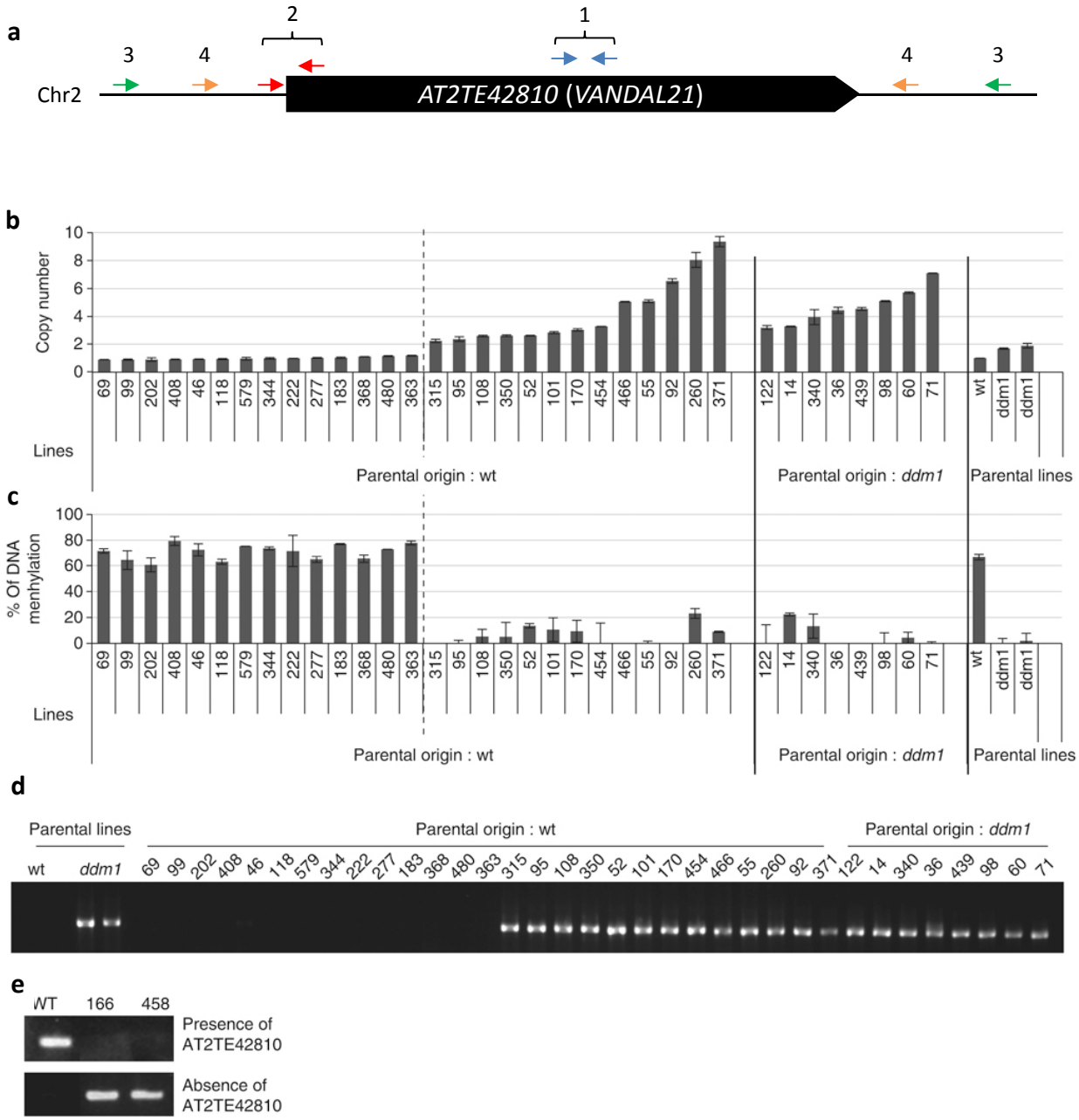


Figure 2.17 : Les nouvelles insertions d'*AT2TE42810* (*VANDAL21*) induisent l'hypométhylation du locus donneur ainsi que son excision.

(**a**) Représentation schématique du positionnement des amorces de PCR et qPCR sur le locus donneur *AT2TE42810* (les échelles ne sont pas respectées). (**b**) Estimation du nombre de copies d'*AT2TE42810* dans chacune des epiRIL estimé par qPCR en utilisant les amorces 1. (**c**) % de méthylation à l'extrémité 5' du locus donneur estimé par McrBC-qPCR en utilisant les amorces 2. Les barres d'erreur représentent l'écart type entre deux répliques techniques. (**d**) Analyse de l'excision somatique d'*AT2TE42810* par « nested » PCR en utilisant les amorces 3 puis 4. (**e**) Transmission germinale de l'allèle excisé dans deux epiRIL détectée par PCR classique en utilisant les amorces 2 pour la présence et 3 pour l'absence. L'origine parentale du locus est sauvage pour l'epiRIL 166 et *ddm1* pour 458.

2.3 Distribution des nouvelles insertions d'ET

Chez Arabidopsis, les ET sont principalement localisés au niveau des régions péricentromériques, ce qui pourrait résulter soit d'un ciblage préférentiel soit de leur élimination rapide après insertion le long des bras chromosomiques.

Diverses études portant sur le ciblage de plusieurs ET ont révélé qu'il était essentiellement dicté par des propriétés de la transposase/intégrase et qu'il peut se faire à plusieurs niveaux. Tout d'abord, comme bon nombre d'endonucléases, ces enzymes peuvent reconnaitre spécifiquement certaines séquences d'ADN. Ensuite, certaines transposases/intégrases contiennent un domaine d'interaction avec des modifications d'histones. Cela a été montré pour certains chromovirus (rétroéléments à LTR GIPSY) qui présentent un chromodomaine reconnaissant la marque H3K9me2 engendrant leur insertion dans l'hétérochromatine (Gao et al. 2008). Enfin, la transposase/intégrase peut interagir avec des protéines associées spécifiquement à l'ADN ou la chromatine comme cela a été très bien décrit pour les rétroéléménts à LTR Ty chez *Saccharomyces cerevisiae*. De fait, le ciblage des éléments Ty3 (rétroélément à LTR GYPSY) au niveau des gènes d'ARNt se ferait d'une manière analogue au recrutement de POL III à ces locus, par une interaction de l'intégrase avec TFIIIC (Aye et al. 2001, Lesage and Todeschini 2005). De même, le ciblage de Ty5 (rétroélément à LTR COPIA) dans l'hétérochromatine serait dû à l'interaction de l'intégrase avec la protéine hétérochromatique SIR4 (Xie et al. 2001). Il est important de noter que ces spécificités ne sont pas caractéristiques des différentes superfamilles d'ET car des ciblages différents impliquant des mécanismes divers ont été identifiés pour des ET d'une même superfamille et ce, au sein d'une même espèce ou entre deux espèces apparentées. C'est notamment le cas des éléments appartenant aux familles de rétroéléments à LTR COPIA Ty1, Ty2, Ty4, et Ty5 chez la levure *S. cerevisiae* (Lesage and Todeschini 2005) ou encore des deux rétroéléments à LTR COPIA apparentés *Tal1* et *EVADE* (*ATCOPIA93*) respectivement présents chez *Arabidopsis lyrata* et *Arabidopsis thaliana* (Tsukahara et al. 2012).

D'autres paramètres peuvent également affecter la localisation des nouvelles insertions d'ET. On peut mentionner notamment les cas du transposon à ADN *Ac* du maïs et de l'élément P de la drosophile qui tendent à s'insérer à proximité du locus donneur par des

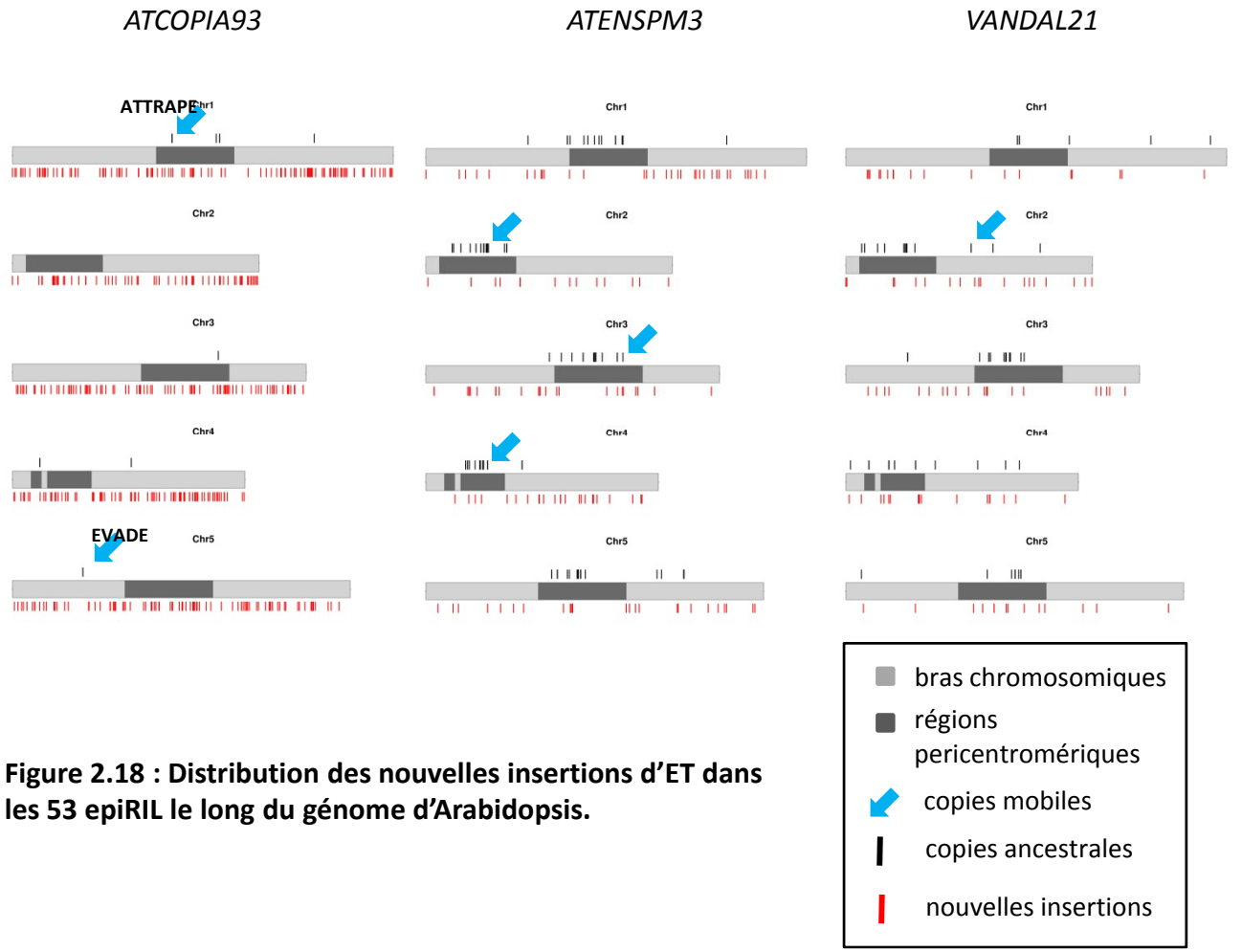


Figure 2.18 : Distribution des nouvelles insertions d'ET dans les 53 epiRIL le long du génome d'Arabidopsis.

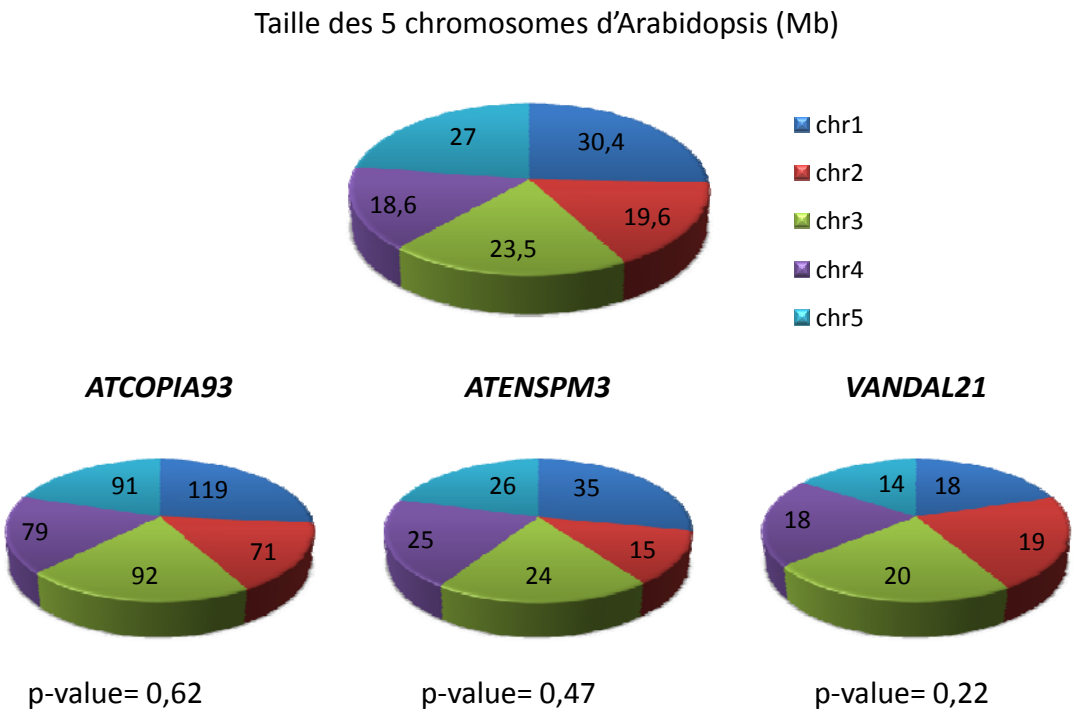


Figure 2.19 : Distribution des nouvelles insertions d'ET sur les chromosomes d'Arabidopsis.
Les p-value ont été obtenues par un test de χ^2.

mécanismes encore méconnus mais basés sur le couplage de la transposition de ses éléments avec la réplication (Greenblatt 1984, Dooner and Belachew 1989).

L'analyse des sites d'insertion des nouvelles copies d'ET dans les epiRIL doit permettre de discriminer entre les deux hypothèses concernant la distribution des ET dans le génome d'Arabidopsis car les évènements d'insertions sont identifiés juste après qu'ils ont eu lieu et donc bien avant que la sélection naturelle ou la dérive génétique ne puissent conduire à leur élimination. De fait, seuls les évènements de transposition causant la létalité car invalidant des gènes essentiels à la survie échappent à notre analyse.

2.3.1 Distribution des nouvelles insertions d'ET le long du génome d'Arabidopsis

J'ai analysé les distributions des nouvelles insertions d'ET le long des chromosomes d'Arabidopsis pour les trois familles les plus mobiles dans les epiRIL en ne prenant en compte que les insertions uniques afin de pouvoir déterminer si elles ont lieu préférentiellement dans des intervalles d'origine *ddm1* ou sauvage (fig. 2.18).

Dans les trois cas, les nouvelles insertions sont réparties sur les 5 chromosomes sans préférence significative pour celui ou ceux sur lesquels se trouvent les locus des donneurs (fig. 2.18 et 2.19). De plus, la distribution des nouvelles insertions n'est pas significativement différente de celle attendue théoriquement sous hypothèse de hasard entre les régions péricentromériques et les bras chromosomiques (fig. 2.20) (d'après la définition des régions péricentromériques établie dans Bernatavichute et al. 2008). Cependant, bien que non statistiquement significatives, des tendances se dessinent pour *VANDAL21* et *ATENSPM3* qui présentent une proportion plus importante de nouvelles insertions respectivement dans les régions péricentromériques et les bras chromosomiques. Enfin, il apparaît que la distribution des nouvelles insertions contraste dramatiquement de celle des copies résidentes de chacune des familles, qui sont majoritairement trouvées dans les régions péricentromériques (bien que, pour *ATCOPIA93*, le trop faible nombre de copies résidentes ne permettent pas de réaliser de test statistique) (fig. 2.20).

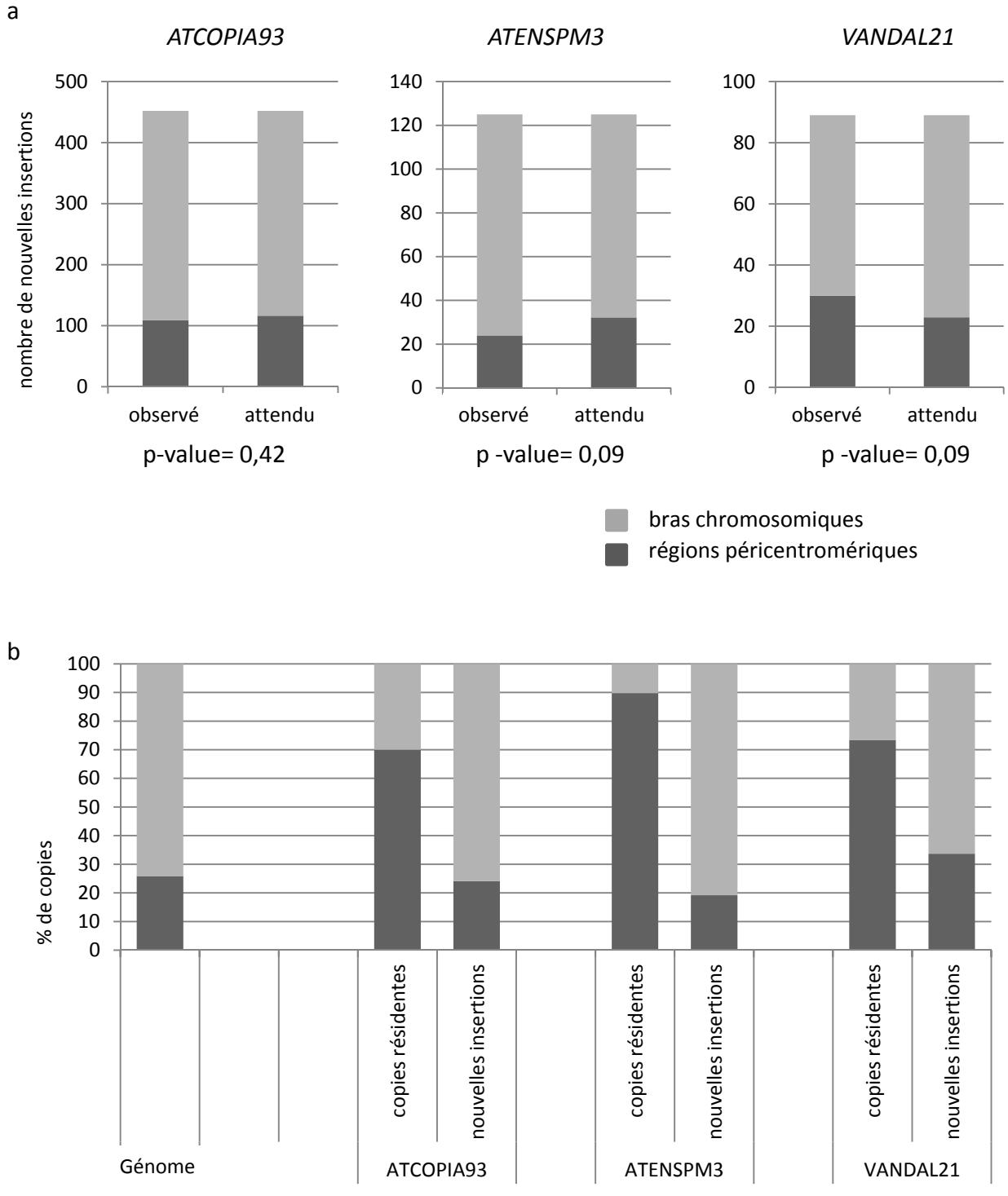


Figure 2.20 : Distribution des nouvelles insertions d'ET entre régions péricentromériques et bras chromosomiques par rapport à une distribution attendue sous hypothèse de hasard (a) et par rapport à la distribution des copies résidentes (b).
Les p-value ont été obtenues par un test de χ^2.

2.3.2 Distribution des nouvelles insertions d'ET en fonction des états chromatiniens

Comme décrit dans l'introduction, quatre états chromatiniens distincts ont été décrits chez Arabidopsis. Ces états sont caractérisés par des combinaisons différentes de marques chromatiniennes et sont associés à des activités transcriptionnelles contrastées. J'ai donc cherché à déterminer si les nouvelles insertions d'ET présentaient une préférence pour certains états. Il est essentiel de noter ici que la carte génomique des états chromatiniens établie au laboratoire est basée sur des expériences réalisées sur des plantules entières qui sont composées quasi-exclusivement de cellules somatiques différenciées. Bien que la dynamique chromatinienne au cours du développement soit encore peu connue, il est probable que la distribution des états chromatiniens le long du génome d'Arabidopsis soit quelque peu différente dans les cellules contribuant à la génération suivante (zygote, cellules souches du méristème apical, gamètes…). De fait, nous savons déjà que des variations plus ou moins importantes de la méthylation de l'ADN se produisent lors de la reproduction (voir introduction).

Comme décrit dans l'introduction, les états chromatiniens forment des domaines de petite taille dans la majorité des cas. J'ai donc restreint mon analyse aux insertions cartographiées dans des intervalles accepteurs d'une taille maximale de 1kb afin de limiter les situations de recouvrement de plusieurs domaines. Pour les trois familles d'ET analysées, la distribution des nouvelles insertions en fonction des états chromatiniens est significativement différente du hasard. De plus, trois profils de distribution très distincts voire contrastés se dessinent pour les trois familles (fig. 2.21).

Dans le cas d'*ATCOPIA93*, les insertions dans les états chromatiniens CS2 et CS4 sont surreprésentées, tandis que celles dans le CS1 sont sous représentées. Les insertions dans le CS3 sont quant à elles observées dans la proportion attendue théoriquement sous une hypothèse de hasard. Cependant, dans les epiRIL, les régions héritées du parent *ddm1* présentent des altérations majeures du CS3 à savoir une perte des marques répressives 5meC et H3K9me2 ainsi qu'un gain de marques associées à la transcription comme H3K4me2. Afin d'affiner mon analyse, j'ai donc déterminé l'origine parentale des intervalles accepteurs localisés dans des domaines définis comme CS3 chez la plante sauvage. Le résultat de cette analyse révèle un biais très marqué pour les intervalles chromosomiques

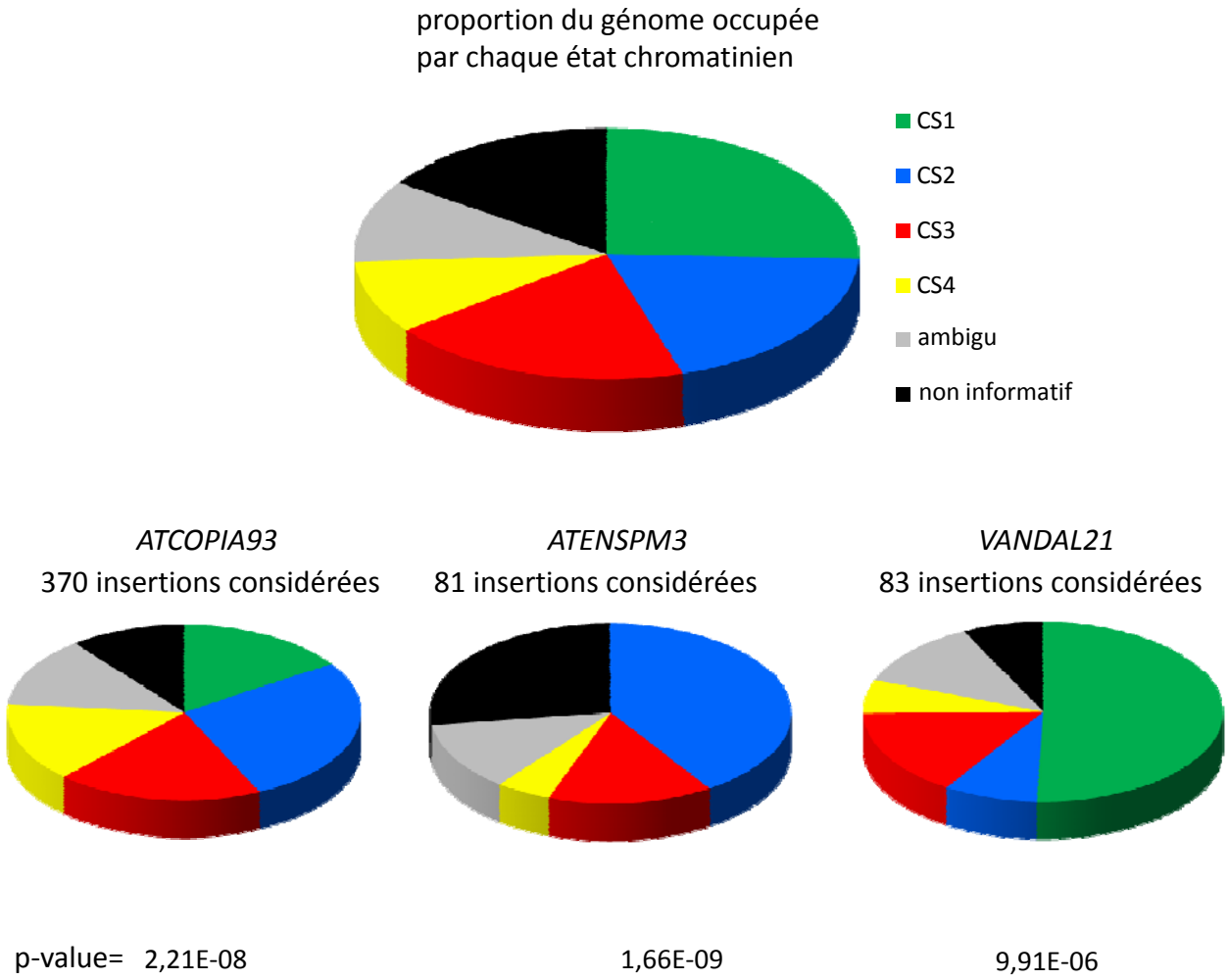
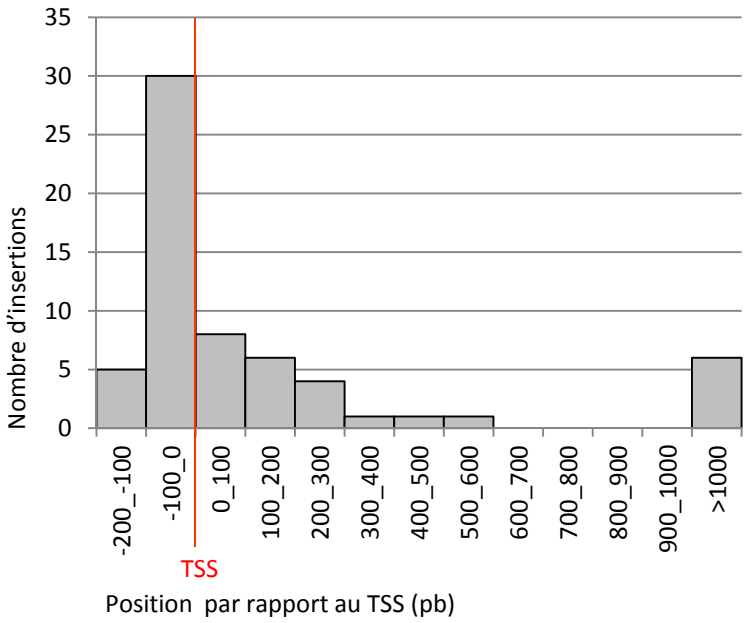


Figure 2.21 : Proportion des nouvelles insertions d'ET dans chacun des état chromatiniens.
En plus des quatre états chromatiniens définis par Roudier et al, 2011 décrits dans l'introduction, j'ai également représenté sur cette figure l'état qualifié d'ambigu (en gris) qui correspond aux sondes de la puce à ADN associées à des marques caractéristiques de plusieurs états distincts ainsi que les régions du génome qui correspondent aux sondes ne rapportant aucun signal pour les marques testées (en noir). Les p-value ont été obtenues par un test de χ^2.

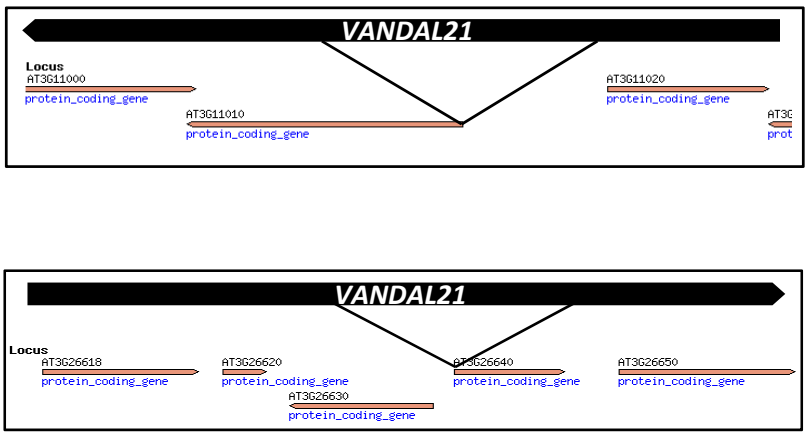
d'origine *ddm1* (47/56 insertions considérées contre 14/56 sous hypothèse de hasard p-value = 2,3E-24 avec un test de χ2). Pour résumer, *ATCOPIA93* montre donc une distribution préférentielle des nouvelles insertions dans le CS2, le CS4 et le CS3 dérivé de *ddm1*. La caractéristique principale partagée par ces trois états est l'absence de méthylation de l'ADN. Cela coïncide avec le déficit en insertion dans le CS1 car la méthylation de l'ADN est aussi une des marques assez fréquemment associée avec cet état. Afin de confirmer cette hypothèse j'ai analysé directement l'état de méthylation (à partir de données de MeDIP-chip) des sites accepteurs des nouvelles insertions d'*ATCOPIA93* en prenant en compte leur origine parentale. Trois cas de figure sont à prendre en considération pour les intervalles accepteurs qui peuvent être : (i) des régions qui ne sont pas ou peu méthylées dans le sauvage et dans le mutant *ddm1*, (ii) des régions qui sont méthylées dans le sauvage mais qui perdent leur méthylation dans le mutant *ddm1* et (iii) des régions qui sont méthylées fortement dans le sauvage et dans le mutant *ddm1*. Il apparait que 85% des insertions d'*ATCOPIA93* ont lieu dans des intervalles de catégorie (i) 15% dans des intervalles de type (ii) et 0% dans des intervalles de types (iii). Cependant, lorsque les insertions sont localisées dans des régions de type (ii), ces régions sont très majoritairement hérités de *ddm1* (22/25 insertions considérées contre 6,25/25 sous hypothèse de hasard, p-value= 3,5E-13 avec un test de χ2). Considérés ensemble, ces résultats confirment l'insertion préférentielle d'*ATCOPIA93* dans les régions non méthylées. On peut alors envisager que, tout comme certaines enzymes de restriction, l'intégrase des *ATCOPIA93* soit sensible à la méthylation de l'ADN. D'autres hypothèses sont bien sûr possibles, notamment une préférence pour d'autres marques chromatiniennes dont la distribution est corrélée négativement avec celle de la méthylation de l'ADN, comme par exemple le variant d'histone H2AZ (Zilberman et al. 2008).

Les nouvelles insertions d'*ATENSPM3* présentent une distribution encore plus contrastée. Les insertions dans le CS1 sont absentes tandis que celles dans le CS2 sont fortement surreprésentées. Le trop faible effectif d'insertions dans le CS3 ne permet pas de déterminer s'il existe un biais dans l'origine parentale des sites accepteurs. A ce stade de l'analyse, il est difficile d'expliquer ce profil de distribution en ne considérant qu'un seul paramètre. Le biais en faveur du CS2 pourrait être dû à un ciblage spécifique des insertions vers ce type chromatinien par exemple grâce à l'interaction de la transposase avec une marque

a



b



c

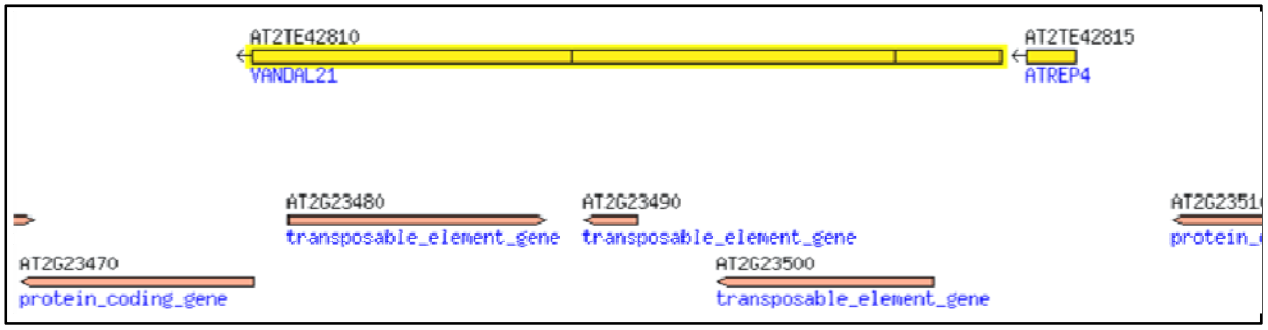


Figure 2.22 : Distribution préférentielle des nouvelles insertions de *VANDAL21* à proximité du TSS des gènes.
(**a**) Distribution des nouvelles insertions de *VANDAL21* par rapport au TSS des gènes. (**b**) Deux exemples de nouvelles insertions de *VANDAL21* à proximité du TSS des gènes et dans la même orientation que le gène. (**c**) Vue Gbrowse du locus donneur *AT2TE42810*.

chromatinienne caractéristique de cet état : H3K27me3. Cependant, cette hypothèse n'explique pas l'absence totale d'insertions dans le CS1 uniquement. Cette absence suggère quant à elle des insertions préférentielles en dehors de la chromatine transcriptionnellement active. Cela n'explique pas alors le biais vers CS2 plutôt que vers CS3 ou CS4 qui sont également très peu actifs transcriptionnellement.

Enfin, la distribution des nouvelles insertions de *VANDAL21* montre une préférence très nette pour le CS1 au détriment des autres catégories qui sont toutes également sous-représentées. Si le faible nombre d'insertions dans le CS3 ne permet pas ici non plus de tester statistiquement la présence d'un biais dans l'origine parentale des sites accepteurs, on notera cependant que 9/12 sont d'origine *ddm1* contre ¼ attendu. Ces deux observations suggèrent un biais d'insertion préférentiel de *VANDAL21* dans les régions chromatiniennes transcriptionnellement actives. Ce résultat est en contradiction avec l'observation précédente qui montrait, pour les nouvelles insertions de *VANDAL21*, une tendance (non significative) vers les régions péricentromériques, théoriquement riches en domaines CS3. On pourrait alors faire l'hypothèse d'une régulation à deux niveaux : une attraction au niveau des régions péricentromériques au sein desquelles s'effectuerait un ciblage spécifique dans les régions transcriptionnellement actives. En effet, bien que composées majoritairement d'ET, les régions péricentromériques contiennent également des gènes.

2.3.3 Distribution des nouvelles insertions d'ET par rapport aux gènes

L'analyse de la localisation des nouvelles insertions d'ET par rapport aux gènes est importante non seulement pour ajouter à la compréhension du ciblage éventuel des insertions mais également pour explorer l'impact de ces nouvelles insertions sur la régulation de l'expression des gènes au sein desquels ou à proximité desquels elles ont eu lieu.

En me basant sur le même jeu d'insertions que pour la distribution en fonction des états chromatiniens (insertions cartographiées dans des intervalles < 1Kb), j'ai recherché combien d'entre elles sont situées dans un gène ou à une distance maximale de 200pb d'un gène. Les proportions ainsi obtenues sont respectivement de 70%, 55% et 77% pour *ATCOPIA93*,

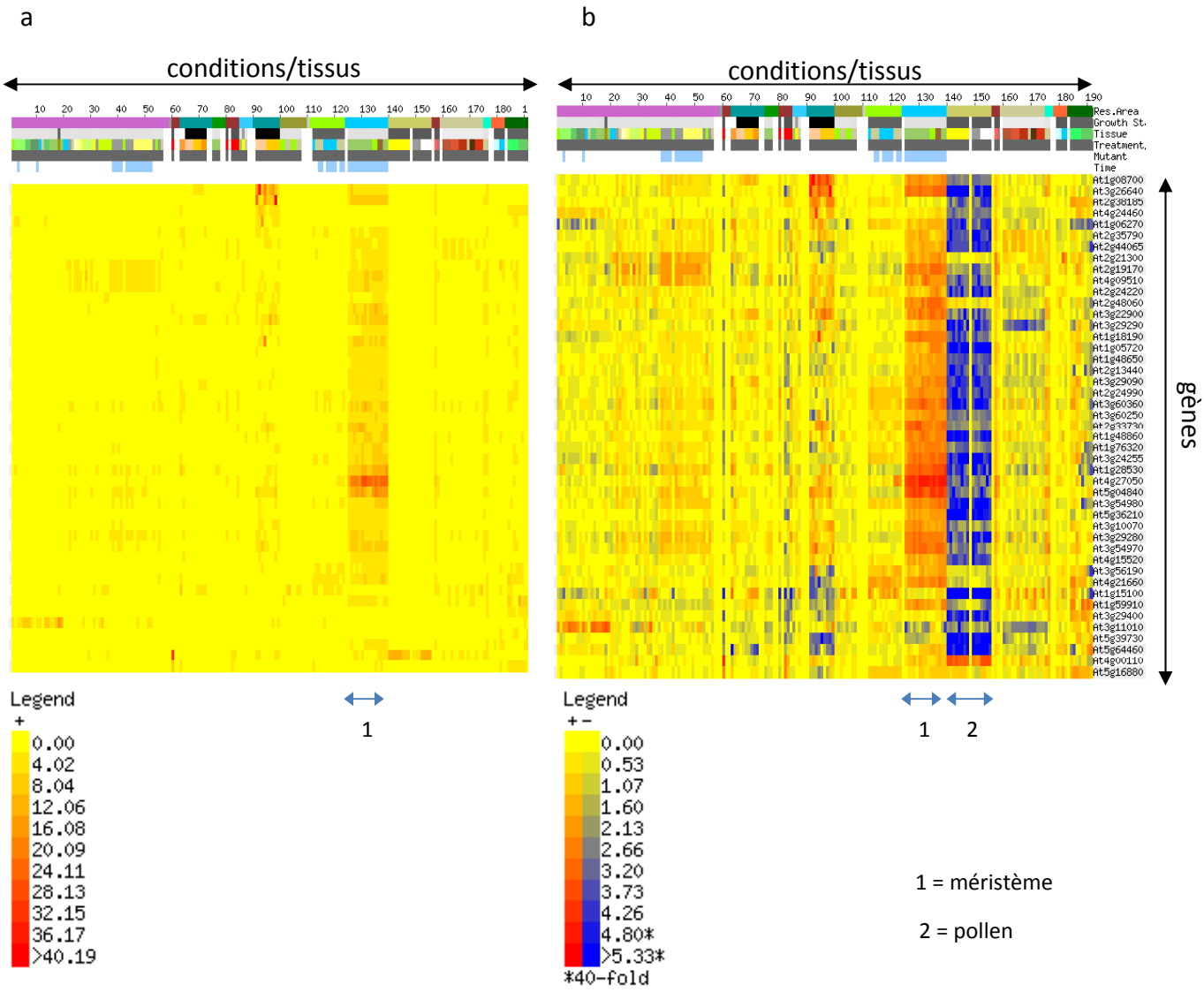


Figure 2.23 : e-northern des gènes ciblés par les nouvelles insertions de *VANDAL21*. Figure établie à partir de données de transcriptomes publiques (http://bar.utoronto.ca). (**a**) Représentation de l'expression en « fold change » par rapport aux conditions contrôles. (**b**) Représentation de l'expression transformée en «log(fold change) ».
Remarque : l'absence d'expression dans le pollen n'est pas spécifique de ces gènes car des tirages aléatoires de gènes présentaient des profils similaires (données non montrées). Ce n'est pas le cas en revanche pour l'expression méristématique.

ATENSPM3 et *VANDAL21* (les résultats de cette analyse sont listés dans l'annexe 5). Afin de déterminer si ces proportions sont significativement différentes du hasard nous avons réalisé des simulations aléatoires (bootstrap) de listes d'intervalles (1000 listes pour chaque ET). Il apparaît que le nombre d'insertions de *VANDAL21* au sein ou à proximité d'au moins un gène est significativement supérieur au hasard (p-value = 0,03) et non significativement différentes du hasard pour les insertions d'*ATCOPIA93* et *ATENSPM3* (p-value = 0,16 et 0,84 respectivement).

Afin d'affiner mes analyses quant à la position de ces insertions par rapport aux gènes, je me suis ensuite focalisée sur les insertions dont la localisation à proximité des gènes peut être déterminée encore plus précisément, dans un intervalle de 200pb. Ainsi, j'ai concentré mon analyse sur les insertions située à <400pb d'un gène (taille de l'intervalle accepteur ajoutée à la distance maximum par rapport au gène). J'ai cherché à déterminer si ces insertions présentaient un biais pour les extrémités 5' ou 3' des gènes. Les insertions de *VANDAL21* sont majoritairement localisées dans la région 5' des gènes en amont du TSS (fig. 2.22a). De plus, ces insertions proches du TSS sont en grande majorité orientées dans le sens du gène ce qui suggère un lien entre les mécanismes d'insertion et de transcription (fig. 2.22b). Il est notable que le locus donneur *VANDAL21* (*AT2TE42810*) montre également ces caractéristiques (fig. 2.22c). Ces mêmes observation ont été faites par nos collaborateurs (Fu et al. 2013).

J'ai ensuite cherché à déterminer si ces gènes présentaient des profils d'expression particuliers en me basant sur des données publiques d'expression. Pour les insertions d'*ATENSPM3* et *ATCOPIA93* aucun profil d'expression significatif n'a pu être mis en évidence (données non montrées). En revanche, les gènes présentant des nouvelles insertions de *VANDAL21* à proximité du TSS tendent à montrer une expression méristématique (fig. 2.23). Ce profil constitue un indicateur du tissu dans lequel *VANDAL21* a transposé dans les epiRIL. En effet, si, comme le suggèrent nos données, *VANDAL21* s'insert préférentiellement près des gènes transcriptionnellement actifs et que les gènes ciblés sont, pour la plupart actifs principalement dans le méristème apical, cela suggère que *VANDAL21* est mobilisé majoritairement dans les cellules souches du méristème apical plutôt que dans les gamétophytes ou l'embryon précoce. Bien sûr, cette analyse ne nous renseigne pas sur la

possibilité d'évènements de transposition dans d'autres types cellulaires ou organes ne contribuant pas à la génération suivante.

2.4 Discussion

2.4.1 Remarques générales sur l'approche expérimentale

Bien que les conséquences de la perte du contrôle épigénétique sur la mobilisation des ET ont déjà été étudiées à plusieurs reprises, l'approche décrite ici est originale par de nombreux aspects. C'est en effet la première fois que la mobilisation est déterminée à l'échelle du génome pour tous les ET, sans apriori d'annotation et ce à l'échelle d'une population. Le séquençage paired-end de librairies mate-pair associé à la méthode d'analyse que nous avons développée permet l'identification robuste des évènements d'insertion et, dans la majorité des cas, de la copie donneuse. Cependant, notre méthode ne permet pas à ce jour d'identifier les évènements d'insertion d'une taille supérieure à 11kb (cf. manuscrit Gilly et al.). Bien que peu nombreux au sein du génome d'Arabidopsis, certains ET >11 kb ont été identifiés comme potentiellement mobiles sur la base de critères de séquence. L'une des améliorations prévue pour notre méthode d'analyse est l'ajout d'un module pouvant identifier de tels évènements. La détection des insertions sans *a priori* sur les annotations a permis également de mettre en évidence la mobilisation de structures complexes composites de plusieurs annotations d'ET. Si à ce stade de l'analyse je ne me suis intéressée qu'aux évènements d'insertion dont le donneur recouvrait au moins partiellement une annotation d'ET, il est important de souligner que des évènements d'insertions n'impliquant aucun ET annoté ont aussi été détectés. S'il s'agit souvent de séquences centromériques imbriquées les unes dans les autres et qui constituent probablement des faux positifs, une analyse approfondie des séquences sans annotation pourrait néanmoins permettre l'identification éventuelle de nouveaux ET.

2.4.2. Eléments mobilisés suite à la perte du contrôle épigénétique

Nous avons identifié un petit nombre de familles présentant des ET mobiles dans le mutant *ddm1* et dans les epiRIL, ce qui est en accord avec des études antérieures réalisées sur ce mutant (Tsukahara et al. 2009). Cependant, nos travaux ont pu mettre en évidence la mobilisation d'éléments appartenant à des familles jamais identifiées comme mobiles chez Arabidopsis ou alors, comme dans le cas des *ATCOPIA78*, mobiles dans des conditions différentes des nôtres. Si on ajoute à cette observation que peu de copies seulement au sein de chacune de ces familles sont mobiles, il apparaît que très peu d'ET sont mobilisés suite à la perturbation du contrôle épigénétique induite par la mutation *ddm1*. Ce nombre paraît d'autant plus faible si on le compare à la réactivation transcriptionnelle massive de très nombreux ET dans le mutant *ddm1 (Lippman et al. 2004, Zemach et al. 2013)*. Plusieurs aspects dont certains ont été évoqués précédemment sont à prendre en compte ici pour tenter d'expliquer ces apparentes contradictions.

Le premier aspect concerne l'activation transcriptionnelle de copies défectueuses, incapables de transposer. Plus de la moitié des familles d'ET d'Arabidopsis serait composée de telles copies. Cependant, il n'est pas aisé de déterminer quelles copies d'ET sont potentiellement mobiles ou non en s'appuyant uniquement sur des critères de séquence. Ceux choisis au laboratoire (notamment présence LTR + RT pour les rétroéléments et de TIR + transposase pour les transposons à ADN) ne sont pas très strictes. On ne peut donc ne pas exclure que le nombre de familles sans aucun membre mobilisable soit encore plus élevé.

L'autre aspect essentiel à considérer est que la mobilisation au travers des générations nécessite l'expression dans des cellules qui participent à la génération suivante comme les cellules souches du méristème apical, les gamètes ou le zygote. L'activation transcriptionnelle des ET dans le mutant *ddm1* a été évaluée sur des plantules entières qui sont majoritairement composées de cellules somatiques et il est envisageable, voire probable que les profils d'expression des ET soient différents dans les cellules contribuant à la génération suivante. Cette hypothèse est d'ailleurs en accord avec les observations qui suggèrent que le contrôle des ET est renforcé au moment de la reproduction grâce à des mécanismes faisant intervenir différentes voies de l'ARNi. Afin de tester cette hypothèse, il

faudrait comparer les profils d'expression (par hybridation *in situ* par exemple) de plusieurs ET mobiles ou seulement potentiellement mobiles.

Cependant, ces différentes hypothèses ne permettent pas d'expliquer le cas très particulier des ET de la famille *ATCOPIA93 EVADE* et *ATTRAPE*. Comme décrit précédemment, ces deux ET présentent une très grande similarité de séquence (>99%), des LTR intègres et identiques ainsi que des ORF continues en apparence intactes. Cependant seul *EVADE* est fortement mobilisé suite à une perte de méthylation de l'ADN (Mirouze et al. 2009 et nos travaux). En outre, des travaux réalisés en collaboration avec l'équipe d'Olivier Voinnet (présentés dans la partie 3) indiquent que seul *EVADE* est exprimé dans le mutant *met1*, et on peut supposer qu'il en est de même dans le mutant *ddm1* et les epiRIL. Or, des travaux récents indiquent que certains ET sont marqués par H3K27me3 (marque caractéristique de la répression par le complexe PRC2) dans le mutant *met1* (Deleris et al. 2012) et que *EVADE* et *ATTRAPE* en font partie. Cependant, les analyses ne permettent pas de discriminer entre ces deux locus. L'absence de réactivation transcriptionnelle d'*ATTRAPE* dans *met1* (et probablement dans *ddm1*) pourrait donc résulter de son ciblage spécifique en fond mutant par le complexe PRC2. Des expériences de CHIP-qPCR utilisant des amorces spécifiques d'*ATTRAPE* ou d'*EVADE* permettraient de tester cette hypothèse.

Quoiqu'il en soit, le décalage qui existe entre le nombre d'ET réactivés transcriptionnellement dans le mutant *ddm1* et le nombre d'ET mobiles n'était pas inattendu. De fait, les mutants *ddm1* sont viables et fertiles et peuvent être propagés sur au moins huit générations d'autofécondation, ce qui est manifestement incompatible avec une mobilisation intempestive de plusieurs centaines d'ET.

2.4.3. Distribution des nouvelles insertions d'ET le long du génome d'Arabidopsis

L'étude de la distribution des ET le long du génome d'Arabidopsis a permis de mettre en évidence des préférences d'insertions spécifiques de chacune des familles étudiées.

J'ai montré tout d'abord que les ET appartenant aux familles *ATCOPIA93*, *VANDAL21* et *ATENSPM3* ne présentent pas de biais d'insertion à l'échelle du génome entre les régions

péricentromériques et bras chromosomiques, en contradiction avec la distribution des copies résidentes.

Si des travaux antérieurs avaient montré une absence apparente de biais concernant les insertions d'*ATCOPIA93* et *ATENSPM3* le long du génome d'Arabidopsis (Miura et al. 2001, Tsukahara et al. 2012) l'étude de la distribution des nouvelles insertions en fonction des états chromatiniens a permis de mettre en évidence certains biais. Dans le cas d'*ATCOPIA93* qui semble s'insérer préférentiellement dans les régions non méthylées du génome, il serait intéressant de tester si l'intégrase présente une affinité différente pour l'ADN méthylé et non méthylé.

Le ciblage préférentiel de *VANDAL21* dans le CS1 et à proximité du TSS des gènes est en accord avec les études réalisées sur d'autres éléments de la superfamille Mu chez le riz et le maïs (Naito et al. 2009, Liu et al. 2009). Il semble donc que les éléments Mu, bien que très diversifiés (Wicker et al. 2007) présentent les mêmes spécificités de site d'insertion chez les différents organismes étudiés. Cependant, c'est la première fois à ma connaissance qu'une préférence d'orientation dans le sens du gène a été mise en évidence pour ce type d'éléments (voir également en annexe 4 l'article Fu et al, 2013).

Considérées ensemble, ces différentes observations suggèrent que la suraccumulation des ET (du moins de ceux considérés ici) dans les régions péricentromériques du génome d'*Arabidopsis thaliana* ne résulte pas d'un biais préférentiel d'insertion mais plutôt d'une élimination des insertions situées le long les bras chromosomiques. La première raison évoquée est la contre sélection des insertions ayant un effet délétère sur les gènes au sein ou à proximité desquels elles ont lieu. Le génome d'Arabidopsis étant très compact, une insertion sur un bras chromosomique a statistiquement de grandes chances de tomber dans ou à proximité immédiate d'un gène. L'autre force pouvant entrainer l'élimination des ET dans les bras chromosomiques est la recombinaison homologue. Chez Arabidopsis, il a été montré que la recombinaison homologue est très fortement inhibée dans les régions péricentromériques (Giraut et al. 2011). Il est donc possible que des insertions, n'ayant potentiellement aucun effet délétère, soient cependant plus fréquemment éliminées si elles sont localisées dans les bras chromosomiques, plus recombinogènes, que celles localisées dans les régions péricentromériques par dérive génétique.