Traces of past transposable element presence in *Brassicaceae* genome dark matter

Agnès Baud[1], Mariène Wan[1], Danielle Nouaud[2], Dominique Anxolabéhère[2], Hadi Quesneville[1]

[1]*URGI, INRA, Université Paris-Saclay, 78026, Versailles, France*
[2]*IJM, Institut Jacques Monod, CNRS UMR 7592, Université Paris-Diderot, Paris, France*

1 Abstract

Transposable elements (TEs) are mobile, repetitive DNA sequences that have been found in every branch of life. In many organisms TEs are the primary contributors to the genome bulk. They invade genomes recurrently by wave of transposition bursts that ceased rapidly as repressed by host defense mechanisms subsequently triggered. The sequences become immobile and start to degrade, fading away in the genome sequence so that it cannot be recognized as such. It contributes then to the so-called “dark matter of the genome”, this part of the genome where nothing can be recognized as biologically functional in first instance.

We developed a new method able to find these old and degenerated TE sequences. With the new algorithm we implemented, we detect up to 10% of the *A. thaliana* genome deriving from TEs not yet identified. Altogether we bring to 50% the part of the genome deriving from TE in this species. Interestingly these sequences are generally very short, about 500bp, and found in the upstream 500pb of genes. Their epigenetic status and their nucleotide composition suggest an old TE origin.

2 Introduction

Transposable elements (TEs) are mobile, repetitive DNA sequences that have been found in practically every branch of life. In many organisms TEs are the primary contributors to the genome bulk. They are one of the main causes for genome size along with polyploidy. Hence, they can represent up to 85% of some genomes such as the wheat and maize[1–5].

TEs, through their ability to amplify, invade genomes. But mobilization of TEs is usually deleterious, and hosts developped epigenetic defence mechanisms to limit their harmful effects. TEs are controlled both transcriptionally and post-transcriptionally through multiple pathways involving RNAi machinery.

Because of these epigenetic controls, TEs remain quiet in the genome for long period of time. They are thought to be potentially reactivated after events such as: a horizontal transfer, genomic shocks such as hybridization or loss of epigenetic control for example following a heat shock. Hence, they invade genomes reccurently by wave of transposition bursts that ceased rapidly when repressed by host defense mechanisms subsequently triggered. TE sequences are then maintained immobile, and their sequence start to accumulate mutations. This process results in an inactivation of the sequence that becomes with time too degenerated to be functional. At this stage, defense mechanisms are no more needed, and the sequence continues to degrade, fading away in the genome sequence so that it cannot be recognized as such. It contributes then to the so-called “dark matter of the genome”, this part of the genome where nothing can be recognized as biologically functional in first instance.

Little is known about this evolution and of the nature and impact of repeated sequences over long periods of times. To explore this question, we developed recently an innovative repeat annotation approach - that we name *cross-species TE annotation* because it uses closely related species to enhance detection sensitivity of very old and degenerated repeated sequences[6]. We analyzed the genome of several *A. thaliana* relatives that diverged approx. 5-40 Myr [7]and generated a library of consensus repeat sequences that we appended to the *A. thaliana* TE reference library in order to compile a “*Brassicaceae*” library that was used to annotate the Col-0 genome and to explore deeper, long term TE presence on genome evolution. Our *Brassicaceae* TE annotation, excluding annotations overlaping CDS, covers over 31.8Mb (26.7%) of the *A. thaliana* genome, achieving highly sensitive detection as it founds one third more than the current official TE annotation [8]. The fact that many *A. thaliana* TE copies can be detected by consensus sequences built in foreign species is presumably most parsimoniously explainable by differential selective bursts among the *Brassicaceae* lineages, and can be seen as an evidence supporting the ancient origin of the *A. thaliana* repeats, and an

explaination to the chromosome-level distribution of old versus young copies in this species.

In this study, we present a new tool that we developed to better exploit this strategy. Our new algorithm is able to find older and more degenerated TE sequences. Indeed, with the tool we implemented, we detect up to 10% more of the *A. thaliana* genome deriving from TEs not yet identified. Combining several strategies and tools, we bring to 50% the part of the genome deriving from TE in this species. Interestingly the new sequences that we detect are generally very short and found in the upstream 500pb of genes. Their epigenetic status, their nucleotide composition, and their long-term conservation in orthologous positions suggest an old TE origin.

3 Material and Methods

Genome sequences

Genome sequences were obtained from the following sources: *A. thaliana* ecotype Col-0 (TAIR10 release) (http://www.phytozome.com/arabidopsis.php); *A. lyrata* (v1.0, http://www.phytozome.com/alyrata.php); *C. rubella* (initial release, http://www.phytozome.com/capsella.php); *A. alpina* (preliminary release, courtesy of Eva-Maria Willing, George Coupland, and Korbinian Schneeberger); *Schrenkiella parvulum* (formerly *Thellungiella parvula;* v2.0, http://thellungiella.org/data/); *B. rapa* (v1.2, http://www.phytozome.com/napacabbage.php).

Genome annotation

TAIR10 gene and TE annotations, where retrieved from URGI web site (https://urgi.versailles.inra.fr/gb2/gbrowse/tairv10_pub_TEs/).

The "*Brassicaceae"* TE annotation was obtained in a previous published study (Maumus et al, 2014). Briefly they were obtained as follows. For all the genomes from five *Arabidopsis thaliana* ecotypes that have been assembled (Col-0, Ler-1, Kro-0, Bur-0 and C24) and *Arabidopsis lyrata*, *Capsella rubella*, *Arabis alpina*, *Brassica rapa*, *Thellungiella salsuginea*, and *Schrenkiella parvula*, the TEdenovo pipeline from the REPET package (v2.0) [9–11]was used with default parameters and with combining both similarity and structural branches. Consensus sequences derived from the structural branches which use LTR Harvest, were retained only when they presented pfam domains typical of LTR retrotransposons. Classification of the consensus sequences was performed by REPET looking for characteristic structural features and similarities to known TEs from Repbase (17.01) [12], and by scanning against the Pfam library (26.0) [13] with HMMER3 [14]. All library of repeat sequences consensus that have been generated, were compiled in a "*Brassicaceae*" library, that was used to annotate the Col-0 genome with TEannot from the REPET package with default settings.

Brassicaceae TE copies

For all the genomes from *Arabidopsis thaliana* Col-0 ecotypes, *Arabidopsis lyrata*, *Capsella rubella*, *Arabis alpina*, *Brassica rapa* and *Schrenkiella parvula*, we used the REPET package v2.5 with its two pipelines TEdenovo and TEannot. On each genome, the similarity branch of TEdenovo was used with default parameters, followed by a TEannot with defaults parameters (sensitivity 2). From this first annotation, we selected consensus sequences that have at least one full length copy (*i.e.* aligned over more than 95% of the consensus length) to run a second TEannot pass. This procedure was demonstrated to improve the quality of annotation [15]. Copies from consensus annotated as 'PotentialHostGene' were removed.

Prediction accuracy

We will denote as true positives (TP), the number of predicted TE nucleotids that truly belong to a TE copy. False positives (FP) are the number of predicted TE nucleotids that do not belong to a TE copy. True negative (TN) are the number of nucleotids truly not predicted as belonging to a TE copy (correct rejection), and false negative (FN) are the missed TE copy nucleotids by the TE prediction.

Sensitivity, also called true positive rate, given by the formula TP/(TP+FP), is obtained by calculating the nucleotid fraction of the predicted TE that overlap with the TE reference annotation.

Specificity, also refered as true negative rate is a bit tedious to calculate. Indeed, it is given by TN/(TN+FP),

but TN and FP are difficult to determine in the case of TE, as we must be certain to know all TE copies of a genome, and that appears to be not really possible. However, we could consider (as a first approximation) that genes are not TEs, nor derive from them, and use this information to better determine TN and FP. In this context, FP are predicted TE nucleotides that overlap a gene annotation, and TN are gene regions not predicted as TE.

The accuracy given by ACC=(TP+TN)/(TP+TN+FN+FP) correspond to the rate of good predictions.

Epigenetic data

We used small RNA map from Lister et al. (2008) [16] corresponding to dataset GSM277608 from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) with lift up to TAIR10 assembly. The occurrences of multiply mapping reads were distributed evenly among genomic copies. This small RNA datasets derive from inflorescences of plants grown at 23°C with 16 hours light period.

We used the 10 chromatin marks maps (H3K18ac, H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K9ac, H3K9me2 et H3) from Luo et al. [17].

Reads that overlap an annotation were counted with the CompareOverlapping.py script (option –O) from S-Mart package [18].

We normalized counts by calculating the ratio between the mean number of reads that overlap an annotation over the number of overlapping reads from the input.

Hierarchical clustering of epigenetic marks is computed from the normalized ratio using the *seaborn* python library with the correlation metric and a standard-scale normalization for each mark.

Orthologous genes analysis

OrthoMCL [19] version 2.0 was used to identify orthologous genes between *A. thaliana, A .lyrata, C. rubella*, and *S. parvulum*. From the 21689 groups obtained, we retained only 6921 with 4 genes all originated from a different species to limit the paralogs false positives of this methods.

Statistical analysis

We used python libraries *pynum*, *scipy for* statistics, *matlibplot* for graphics and *panda* for data manipulation. *Jupyter notebooks* were used to monitor the analysis.

Sequence and coordinate manipulation.

We obtained random sequence using *shuffle* from the SQUID 1.9g package [20] and *revseq* from Emboss 6.1.0 [21] packages.

Genome coordinates were manipulated with the S-Mart package [18]. In particular we used `modifyGenomicCoordinates` (version 1.0.1) and `CompareOverlapping` (version 1.0.4) to respectively extend coordinates in 5' of genes, and find overlaps.

4 Results

Duster: a new approach for analyzing old degenerated transposable elements

After separation from a common ancestor, repeat families have different destinies in different genomes. For example, a specific repeat family can multiply again in one species and not in another one. The burst of an autonomous repeat family is a highly selective process: only the copies that have accumulated limited mutational drift are functional and capable to burst. Such a selective burst allows the multiplication of the best conserved copies, *i.e.* the ones that are closest to the ancestral sequence. Therefore, the TE families that maintain activity in some genomes should longer preserve the ancestral sequence as compared to a decaying pool of relatives in another genome. As a consequence, a repeat copy from one species is most likely to be relatively old if it is most similar to a sequence established from a foreign species.

Consequently, identifying TEs in a species with reference sequences found in the studied species but also in closely related species, will detect older TE copies than those that are found only with the reference sequence

from the studied species. Indeed, we unravel old TE sequences which would not have been recognizes otherwise.

We developed a software called *Duster* able to compare a genome sequence, here considered as a query sequence, to a large amount of TE sequences, *i.e.* a sequence library. Its algorithm used *k-mers* to search for similar sequences without performing nucleotid aligments. Hashed *k-mers* values allows to speed-up the search. Sensitivity is obtained allowing one mismatches in *k-mers* every *n* consecutive nucleotids. Details of the algorithm are given in Supplementary file 1, but we can summarize the algorithm as comparing *k-mers* between the genome and each sequence from the library and reporting matches when at least two *k-mers* are found on the same alignment diagonal (i.e. the difference between the coordinates on the query and the sequence library are identical) with a maximal distance of *d k-mers*. The region bounded by the two-extreme *k-mer* position are reported as matching. Two matching regions on the genome separated by less than *x k-mers* are merged. At the end of this first pass, the region identified on the genome can be used as a new sequence library for a new search (the *-n* parameter). This procedure is repeated until genome coverage increased by less than 1% if *-n* is set to 0.

Duster performance assessment

To assess Duster performances, we compute its prediction accuracy. The accuracy (ACC) is obtained by calculating sensitivity (Sn) and specificity (Sp) of the predicted TE annotation by comparing the prediction with a reference annotation at the nucleotid level (see Material and Methods). We used here as reference, the official annotation for *A. thaliana* from TAIR. ACC considers both Sn and Sp and conveniently propose an aggregated value. Consequently, we decided to maximize this value in our benchmarks. For this test, the sequence library is the TE sequences from TAIR annotation.

We empirically choose for Duster a parameter set that appear to give the best result in our hands, by optimizing the annotation accuracy using TE copy sequences from other Brassicaceae species (data not shown). With this parameter, we compared Duster performances, benchmarking it with tools that implement other algorithm that could be used for similar analysis. We choose for this comparison BLAST [22] and MegaBLAST [23] two widely used sequence comparison algorithms. As they are not designed to be run on a long genomic sequence, we run them through Blaster [24] which pre- and post-process respectively input sequence and output results to facilitate their usage.

Table 1 shows the results obtained with Duster, BLAST and MegaBLAST. Two runs of Duster were performed varying the distance required between two *k-mers* (*–d* parameter) with 0 or 5, and the position shift on the genomic sequence (*-S* parameter) with 15 (size of the *k-mer*) or 7 (overlapping *k-mer* by 7 bp). The parameter *(–d 0; –S 15*) is less sensitive by construction than (*-d 5; -S 7*). Note that parameter *–f* is the minimum sequence size to be considered in the library of sequences (*L* in supplementary file 1), and *–n* is the number of iterations. Here the two sets of parameters are set to only consider sequence longer than 100 bp with only 1 iteration. We also chose a *k-mer* length of 15 (parameter *–w*) and a potential nucleotide mismatch every 4 nucleotides (parameter *–k*).

Table 1 shows that Duster outperforms the other tools in term of speed: five to seven minutes versus 38 minutes at best with MegaBLAST parallelized on 4 threads. Sensitivity is higher for Duster and BLAST with 0.99. Specificity is lower for Duster, but coverage is higher, suggesting that our tool detects many more potential TEs not previously known. To assess false positive rate, differently, we run Duster on a shuffled genome sequence respecting dinucleotides composition and a reversed but not complemented sequence. Coverage on the shuffled sequence remains under 0.001 and 0.01 on the reversed one.

According to the way we compute false positives, this suggests that many genes have regions deriving from old TEs not detected with other tools. As this is one purpose of our new tools, we considered that it is a good result, in particular as it is not biologically inconsistent.

Table 1 : Comparing tool performances using TE sequences from official TAIR A. thaliana *TE annotation. *computed on a Linux workstation with Intel Xeon® CPU E3-1270 v3 @ 3.50 GHz x 8 and 15.5 Go of RAM.*

TOOLS	PARAMETERS	COV.	SN	SP	ACC	TIME*
Duster	-w 15; -k 4; -d 0; -f 100; -S 15; -n 1	0,27 (shuff.: 0% rev.:0.1%)	0,99	0,91	0,93	5.0 m
Duster	-w 15; -k 4; -d 5; -f 100; -S 7; -n 1	0,31 (shuff.: 0.1% rev.:1%)	0,99	0,87	0,89	7.0 m
Blaster/MegaBLAST	-S 2 ; -L 200	0,20	0,96	0,99	0,98	1,18h
Blaster/MegaBLAST-4Threads	-S 2 ; -L 200	0,20	0,96	0,99	0,98	38m
Blaster/BLAST	-S 2	0,23	0,99	0,98	0,98	17h
Blaster/BLAST-2Threads	-S 2 ; -L 200	0,22	0,97	0,98	0,98	8,8h
Blaster/BLAST-4Threads	-S 2 ; -L 200	0,22	0,97	0,98	0,98	6,15h

Up to fifty percent of the A. thaliana genome derives from transposables elements

Considering that Duster may detect interesting new TE sequence in *A. thaliana* genome, we run an analysis with all *Brassicaceae* TE copies we previously annotate on *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Capsella rubella*, *Schrenkiella parvulum*, *Arabis alpina*, and *Brassica rapa* (see Material and Methods). We used the previous parameter setting with *–d* 5 and *–S* 7, but change *–n* to 0 for leaving the algorithm to iterate until it reaches a genome coverage difference between two successive iterations less than 1%.

Overall, gathering TAIR10, *Brassicaceae* and Duster TE annotations we cover 49.75% of the genome sequence. It corresponds to a net increase of +29,72% when compared to the reference TAIR10 TE annotation which cover 20.03%, and by +10.60% if compared to the *Brassicaceae* 39.15% TE coverage.

Structural properties of Duster-specific copies

To characterize the new repeated compartiment found by Duster, we extract from its annotations, copies that did not overlap any Gene, TAIR10, *Brassicaceae*, or *A. thaliana* REPET annotation (see Material and Methods). Hence, we got what we called the Duster-specific copies. We did the same with TAIR10 and *Brassicaceae* annotations to obtain respectively TAIR10-specific and *Brassicaceae*-specific copies by removing any copies that overlap any other annotation.

We characterized these copies by comparing their length, chromosome distribution, and position relative to genes (figure 1). Duster-specific copies appear significantly shorter than *Brassicaceae*-specific, TAIR10-specific, and TAIR10 copies (Figure 1A, Chi-square p-value respectively 0.0, 9.79E-14, 0.0). They are found more often in up-stream relative to genes (figure 1B, Chi-square p-value 7.84E-61), as for *Brassicaceae*-specific, and TAIR10 copies (Chi-square p-value respectively 2.37E-27, 1.78E-44). On the contrary, there is no significative difference for TAIR10-specific copies (Chi-square p-value respectively 0.92). Figure 1C shows the distance to the closest 5', or 3', TE copies for each annotated gene. It shows that duster-specific copies are found closer than other copies (Chi-square p-value being 0.0, for *Brassicaceae*-specific, TAIR10-specific, and TAIR10 copies). Interestingly *Brassicaceae*-specific, and TAIR10 copies follows the same general trend, but TAIR10-specific copies exihibit an opposite behavior. Figure 1D shows TE copies distribution on the chromosomes. It shows that Duster-specific, and with a lesser trend *Brassicaceae*-specific

copies, follows the gene chromosomal distribution, where as TAIR10 shows an opposite one.

Finally, we examined the nucleotid composition of the sequences counting the dinucleotids. Figure 2A presents the counts as a radar plot. It shows similar profiles for all TE copies (Duster-specific, *Brassicaceae*-specific, TAIR10-specific, and TAIR10 TEs). Interestingly, TAIR10-specific first, followed by Duster-specific have the strongest bias in TT and AA dinucleotids. TAIR10-specific alone is characterized by the strongest bias in AT and TA. These biases, also shared by other TE copies but in a lesser extent, were supposed to be the consequence of the deamination process of methylated cytosine. The fact that TAIR10-specific and Duster-specific have the highest "A-T" richness may indicate that they have undergone a longer mutational process and they are consequently older.

Epigenetic profiles

We investigate the epigenetic status of identified TE copies considering small RNAs, and chromatin marks. Small RNA where taken from Lister et al. [16] experiment, available as mapped data on the genome. Intersection between this dataset and our annotations shows 4.30%, 30.89%, 17.87%, 60.44% of matching TE copies from respectively Duster-specific, Brassicaceae-specific, TAIR10-specific, and TAIR10 TE datasets. We analyzed 9 epigenetic marks from [17] also available as mapped data. The hierarchical clustering algorithm identify clear distinct profiles for genes and for TAIR10 TEs (figure 2B). TAIR10 TEs are enriched with heterochromatin marks (H3K27me1 and H3K9me2), and genes with euchromatin marks (H3K18ac, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, and H3K9ac). Duster-specific, *Brassicaceae*-specific, and TAIR10-specific are associated to the TAIR10 TE profile by the clustering algorithm, indicating that their profiles are closer to a TE profile than a gene one. *Brassicaceae*-specific and TAIR10-specific have a quite similar profile which appears more similar to TAIR10 TEs than Duster-specific marks, in particular because of the higher density of heterochromatic marks. Duster-specific copies appear here to have very few heterochromatic and euchromatic chromatin marks.

TE conservation in the Brassicaceae

We looked at the conservation of Duster-specific, Brassicaceae-specific, TAIR10-specific copies among the *Brassicaceae*. We considered only regions close to orthologous genes found with OrthoMCL (see Material and Methods). We chose *A. thaliana*, *A .lyrata*, *C. rubella*, and *S. parvulum* as they span divergence time from 5 to 40 Myr. Orthologous genes with 5 kb 5' flanking regions were aligned on the *A. thaliana* region containing both the orthologous gene and a TE copy from a method-specific set. We end-up with 9610, 5385, and 34 TEs respectively for Duster-specific, Brassicaceae-specific, and TAIR10-specific that can be analysed. We considered a TE copy as present if more than 50% of the *A. thaliana* annotated TE copy nucleotids are found identical in the pairwise alignment. Figure 3 shows sequence conservation between species represented with 3 digits where position 1, 2 and 3 stand for presence, denoted with a 1, or absence with 0, on respectively *A .lyrata*, *C. rubella*, and *S. parvulum*.

We show that the Duster-specific set contain more old TEs, followed by Brassicaceae-specific as 111 column of the histogram which correspond to the presence of a TE on orthologous position in the 4 species is higher. Interestingly 000 column is quite high also. This corresponds to TE found only in *A. thaliana*, but as they belong to method-specific sets, they escaped the TE detection from REPET simple *de novo* procedure when limited to *A. thaliana*. Consequently, this indicates that they can only be detected with TEs found in other species. These copies can then result either from horizontal transfer from these other species, or simply identified in others genomes because better conserved there. All these results illustrate the interest of our cross-species TE annotation approach and the efficiency of Duster over the REPET annotation procedure.

5 Discussions

A need for new dedicated repeat detection algorithm

RepeatMasker [25], Censor [26,27], and Blaster [19] are the most used tools to annotate TE sequences in genomes. All these tools encapsulate BLAST calls with pre- and post-processing allowing to analyse a genomic sequence. Hence, they all have the intrinsic limits of BLAST, relying in particular on seeds to find alignements. These seeds in BLAST are *k-mers* of size 11 by default. BLAST requires two k-mers on the same diagnonal (i.e. alignment without gaps) to process further the alignment in order to check if it is

relevant. Alignement scores are used to check relevance with alignment scores threshold determined through a probabilistic model. According to this, two reasons may explain the poor sensitivity compared to Duster.

First, two k-mers are needed to start an alignment. With default parameters, this requires at least an exact match of 22 nucleotides between two sequences. This can be reduced as the seed length is a parameter of BLAST, and reduced to 14 with some implementation (with WU-BLAST, seed size can be 7), but still needs an exact match. For Duster, we allow mismatches in the *k-mers*, and the two k-mers may overlap. With the setting used in this analysis we required a match with 21 nucleotides but where some mismatches may occur.

The second reason is relative to the relevance test based on an alignment score threshold. Indeed, even if the exact match of 22 nucleotides is found, an ungaped alignment is produced to test its score to the probabilistic model. The result depends on sequences length and on a model that even if mathematically sophisticated, seems biologically too simple as it considers independence between successive nucleotides, and their equiprobability. We know today that these two assumptions are false. Consequently, the model rejects some alignment differently according to the sequences length, and with a model that seems discutable. In Duster, we keep all regions that matches with two k-mers, and the parameters we choose show very few false positive (0.001).

We see here that BLAST is not the proper algorithm to find small degenerated TEs. In fact, it has been developed for another purpose, finding the best matches in databanks, to a sequence given as a query. The usage made for finding TEs corresponds to an important deviation from its first purpose for which it has been shown performant.

Duster has been designed in particular for finding old and degenerated TE copies. In addition to a different strategy for k-mers, it could be considered as an alignment free algorithm. BLAST does many alignments before reporting a match. In our case, we do not really need an alignment, juste boundary coordinates. This explain the gain in term of speed of this algorithm. We may consider that boundaries are not precise as we relie on k-mers and consequently have a precision limit linked to the k-mer size and its coordinate shift on the genomic sequence. With the parameters used in this analysis, the precision is about 7 nucleotides. We think that it is enough for our purpose to identify regions, and even not reasonable to have a better one with very old and degenerated TE copies.

This demonstrate the interest of having specifically developed tools for some hard-biological questions. It advocates the needs for a new generation of sequence finding tools, designed for the biological question asked, perhaps replacing BLAST sometimes by more adapted algorithms.

Long term evolution of TE copies

Duster-specific copies appears to be old, degenerated, and surprisingly close to genes, in 5' at a distance corresponding to the gene regulatory regions. A first explanation would be that they are specifically maintained in these regions because they have a functional role for the host, probably a regulatory module acting on the neigbour gene. The other TE copies with no function would have been gradually removed from the genome by accumulating point mutations and deletions. With time, only those that have acquired a function remains in the genomes. In this case, identifying the very old TE copies might mean to identify regulatory modules selected long time ago. They would be involved in the creation of important pathways as they are still present.

Alternatively, we can argue that the gene regulatory regions accumulate point mutations and deletions at a lesser rate than other regions because they are functional. Indeed, deletions or point mutations may affect the gene regulation and consequently are counter-selected. In such a case, a TE insertion in these regions would be difficult to remove once installed.

These two scenarios are not incompatible. TE copies may be present in the 5' regulatory region because of either one or the other. The challenge would be to determine which copies follows the first or the second.

Evolutionary impact of Duster-specific copies

TEs are important source of variation on which selection can operate to evolve species. In plants many examples are well documented[28]. A small amount of TE copies may acquire by themselves a functional role during evolution. We call this phenomenon « domestication » when the TE encoded function remains the same for the host, or « exaptation » otherwise. These functional sequences are then maintained by selective pressure and can be recognized as such because they are conserved. In some celebrated cases TEs have been co-opted to play key organismal functions such as the generation of antibody diversity in the

vertebrate immune system [29] and the maintenance of telomeres in *Drosophila* [30]. A striking example in plants is the Mustang protein family essential for flower development and fitness which derives from a TE family [31].

TE genome invasion may also increase the number of TE transcription-factor-binding sites, linking nearby genes into transcriptional networks. Such networks are observed, for example, with the Dayspleeper gene in *A. thaliana*, where the transposase have been exapted as transcription factors [32]. Genome-wide assessment revealed that hundreds of TEs have been co-opted into regulatory regions of mammalian genes [33,34] . TEs have also been involved in both the creation of new regulatory networks [35,36] and in the rewiring of preexisting ones [37]. In plants, several examples of temperature sensitive gene expression have been reported [38–40]. One striking example is the the cold-responsive expression of the transcription factor Ruby, responsible for the red colour of blood-orange [41]. Many other studies suggest that plants can also be highly sensitive to various other stress including salt [38], wounding [42], bacteria [43], and viruses [44].

Genes located nearby TEs could be also affected by their epigenetic control and then become epigenetically regulated [45]. The best studied example in plants is the FWA locus in *A. thaliana* which is epigenetically silenced by a TE insertion. Ectopic expression result in late flowering phenotype [46,47]. We analyzed genome-wide DNA methylation maps obtained at single-nucleotide resolution in *Arabidopsis* [48] and showed that although the majority of TE sequences are methylated, ~25% are not. Moreover, a significant fraction of TE sequences densely methylated at CG, CHG, and CHH sites (where H=A, T or C) had no or few matching siRNAs and were therefore unlikely to be targeted by the RNA-directed DNA methylation (RdDM) machinery. We provided evidence that these TE sequences acquire DNA methylation through spreading from adjacent siRNA-targeted regions. Further, we showed that methylated and unmethylated TE sequences tend to be more abundant close to genes in euchromatin. However, this trend is less pronounced for methylated TE sequences located 5’ to genes. Based on these findings, we proposed that DNA methylation spread has a negative impact on neighboring gene expression through promoter methylation.

TEs can also mediate chromosomal rearrangements [49], a phenomenon well documented in maize [50]. They also mediate some gene movements [51–53]. This common feature of plant genomes, bring genes into new genomic contexts which could affect their regulation. These events may also increase the number of gene copies reducing selective pressure operating on it, allowing then gene copies to acquire new pattern of expression or new functions. Consequently, duplicated genes are more subject to changes in their function, but also their regulations. An illustration of this phenomenon is shown at the Sun locus in tomato, where TE-mediated gene variations are responsible for differences in fruit shape [54]. TEs can even operate simultaneously by moving genes to change their regulation context, duplicating them to decrease selection pressure, and bring new regulation modules, potentially sensitive to the environnement. A good example is the R locus in *A. thaliana* involved in the resistance to *Hyaloperonospora parasica* [55] . Interestingly, genes involved in resistance to bioagressors are often found duplicated, forming clusters of transposed sequences, suggesting an important role of TEs for plant resitances.

Altogether, these results strongly suggest that TE mediated gene duplication or regulation may build gene networks sensitive to environmental condition. These networks would be of particular interest for adapting the hosts to their environment. The Duster-specific TEs we identified, are old and located close to genes in regions known to contain gene regulatory regions. They might have play an importante role in the past to build new pathways for adaptating *Brassicaceae* to their environnements. Futher analysis are needed to locate among them, those that played a role. This study is a first step toward this analysis providing candidates yet unkown until now.

6 Conclusions

TEs represent quantitatively important components of genome sequences, and as shown by several examples described in the literature, there is no doubt today that modern genomic DNA has evolved in close association with TEs. Through their amplification, TEs participate to the DNA turnover in genome sequence by duplicating DNA and bringing new sequences, hence forming the raw material for genetic innovations.

In this study, we investigate the TE contribution to Arabidopsis genome bulk, thanks to a new tool we developped called Duster, at timescale still inaccessible for concurrent approaches. Duster implements a new efficient algorithm that allows to identify a new 10% of nucleotids which are annotated as TE sequence in the genome of *A. thaliana*. Hence, we dig deeper into the dark matter than previously done, recognizing old and degenerated TEs sequences, undetectable with other methodologies.

We delivered a key knowledge helping to understand plant evolution and plant adaptation, by providing clues that identifies TEs in genes regulatory regions, providing potential regulation modules. Some TE copies identified here may have been selected in the past for adaption to changing environments.

7 Acknowledgment

We thank Michaël Alaux and Johann Confais for their comment on the manuscript. This work was performed using the facilities of the URGI platform (https://urgi.versailles.inra.fr/)

8 Cited literature

1. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115: 49–63. doi:10.1023/A:1016072014259
2. Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP. Cross species selection scans identify components of C4 photosynthesis in the grasses. J Exp Bot. 2017;68: 127–135. doi:10.1093/jxb/erw256
3. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326: 1112–1115. doi:10.1126/science.1178534
4. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics. 2002;3: 329–341. doi:10.1038/nrg793
5. Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, et al. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biology. 2018;19: 103. doi:10.1186/s13059-018-1479-0
6. Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat Commun. 2014;5: 4104. doi:10.1038/ncomms5104
7. Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, et al. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. Mol Biol Evol. 2016;33: 394–412. doi:10.1093/molbev/msv226
8. Buisine N, Quesneville H, Colot V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics. 2008;91: 467–475. doi:10.1016/j.ygeno.2008.01.005
9. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1: 166–175. doi:10.1371/journal.pcbi.0010022
10. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS ONE. 2011;6: e16526. doi:10.1371/journal.pone.0016526
11. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: an automatic transposable element classification tool. PLoS ONE. 2014;9: e91929. doi:10.1371/journal.pone.0091929
12. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6: 11. doi:10.1186/s13100-015-0041-9
13. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2018; doi:10.1093/nar/gky995
14. Accelerated Profile HMM Searches [Internet]. [cited 25 Oct 2018]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195
15. Jamilloux V, Daron J, Choulet F, Quesneville H. De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue. Proceedings of the IEEE. 2017;105: 474–481. doi:10.1109/JPROC.2016.2590833
16. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133: 523–536. doi:10.1016/j.cell.2008.03.029
17. Luo C, Durgin BG, Watanabe N, Lam E. Defining the functional network of epigenetic regulators in Arabidopsis thaliana. Mol Plant. 2009;2: 661–674. doi:10.1093/mp/ssp017
18. Zytnicki M, Quesneville H. S-MART, a software toolbox to aid RNA-Seq data analysis. PLoS ONE.

2011;6: e25988. doi:10.1371/journal.pone.0025988
19. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–2189. doi:10.1101/gr.1224503
20. Eddy SR. SQUID - library of functions for biological sequence analysis Copyright (C) 1992-2002 Washington University School of Medicine [Internet]. Available: http://eddylab.org/software/squid/squid-1.9g/
21. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.
22. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402.
23. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24: 1757–1764. doi:10.1093/bioinformatics/btn322
24. Quesneville H, Nouaud D, Anxolabéhère D. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol. 2003;57 Suppl 1: S50-59. doi:10.1007/s00239-003-0007-2
25. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015.
26. Jurka J, Klonowski P, Dagman V, Pelton P. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem. 1996;20: 119–121.
27. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 2006;7: 474. doi:10.1186/1471-2105-7-474
28. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14: 49–61. doi:10.1038/nrg3374
29. Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 2005;3: e181. doi:10.1371/journal.pbio.0030181
30. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F. Transposons in place of telomeric repeats at a Drosophila telomere. Cell. 1993;75: 1083–1093. doi:10.1016/0092-8674(93)90318-K
31. Joly-Lopez Z, Forczek E, Hoen DR, Juretic N, Bureau TE. A Gene Family Derived from Transposable Elements during Early Angiosperm Evolution Has Reproductive Fitness Benefits in Arabidopsis thaliana. PLOS Genetics. 2012;8: e1002931. doi:10.1371/journal.pgen.1002931
32. Bundock P, Hooykaas P. An Arabidopsis hAT-like transposase is essential for plant development. Nature. 2005;436: 282–284. doi:10.1038/nature03667
33. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441: 87–90. doi:10.1038/nature04696
34. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA. 2007;104: 8005–8010. doi:10.1073/pnas.0611223104
35. Wang J, Keightley PD, Halligan DL. Effect of divergence time and recombination rate on molecular evolution of Drosophila INE-1 transposable elements and other candidates for neutrally evolving sites. J Mol Evol. 2007;65: 627–639. doi:10.1007/s00239-007-9028-6
36. Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA. 2006;103: 8101–8106. doi:10.1073/pnas.0601161103
37. Ackerman H, Udalova I, Hull J, Kwiatkowski D. Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol. 2002;19: 884–890. doi:10.1093/oxfordjournals.molbev.a004145
38. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461: 1130–1134. doi:10.1038/nature08479
39. Ivashuta S, Naumkina M, Gau M, Uchiyama K, Isobe S, Mizukami Y, et al. Genotype-dependent transcriptional activation of novel repetitive elements during cold acclimation of alfalfa (Medicago sativa). Plant J. 2002;31: 615–627.
40. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011;472: 115–119. doi:10.1038/nature09861
41. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-

specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24: 1242–1255. doi:10.1105/tpc.111.095232
42. Mhiri C, Morel J-B, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien M-A. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Mol Biol. 1997;33: 257–266. doi:10.1023/A:1005727132202
43. Grandbastien M-A, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa A-PP, et al. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res. 2005;110: 229–241. doi:10.1159/000084957
44. Buchmann RC, Asad S, Wolf JN, Mohannath G, Bisaro DM. Geminivirus AL2 and L2 proteins suppress transcriptional gene silencing and cause genome-wide reductions in cytosine methylation. J Virol. 2009;83: 5005–5013. doi:10.1128/JVI.01771-08
45. Makarevitch I, Stupar RM, Iniguez AL, Haun WJ, Barbazuk WB, Kaeppler SM, et al. Natural Variation for Alleles Under Epigenetic Control by the Maize Chromomethylase Zmet2. Genetics. 2007;177: 749–760. doi:10.1534/genetics.107.072702
46. Kinoshita Y, Saze H, Kinoshita T, Miura A, Soppe WJJ, Koornneef M, et al. Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. Plant J. 2007;49: 38–45. doi:10.1111/j.1365-313X.2006.02936.x
47. Fujimoto R, Kinoshita Y, Kawabe A, Kinoshita T, Takashima K, Nordborg M, et al. Evolution and control of imprinted FWA genes in the genus Arabidopsis. PLoS Genet. 2008;4: e1000048. doi:10.1371/journal.pgen.1000048
48. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res. 2011;39: 6919–6931. doi:10.1093/nar/gkr324
49. Fiston-Lavier A-S, Anxolabehere D, Quesneville H. A model of segmental duplication formation in Drosophila melanogaster. Genome Res. 2007;17: 1458–1470. doi:10.1101/gr.6208307
50. Yu C, Zhang J, Peterson T. Genome Rearrangements in Maize Induced by Alternative Transposition of Reversed Ac/Ds Termini. Genetics. 2011;188: 59–67. doi:10.1534/genetics.111.126847
51. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37: 997–1002. doi:10.1038/ng1615
52. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431: 569–573. doi:10.1038/nature02953
53. Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci USA. 2011;108: 16116–16121. doi:10.1073/pnas.1109273108
54. van der Knaap E, Sanyal A, Jackson SA, Tanksley SD. High-resolution fine mapping and fluorescence in situ hybridization analysis of sun, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements. Genetics. 2004;168: 2127–2140. doi:10.1534/genetics.104.031013
55. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res. 2008;18: 1924–1937. doi:10.1101/gr.081026.108

9 Figure legends

Figure 1: Structural characteristics of Duster-specific, *Brassicaceae*-specific, and TAIR10-specific copies. (A) TE length distribution, (B) TE 5' or 3' position relative to genes, (C) distance to the closest 5', or 3', TE copies for each annotated gene, (D) TE copies distribution on the chromosomes.

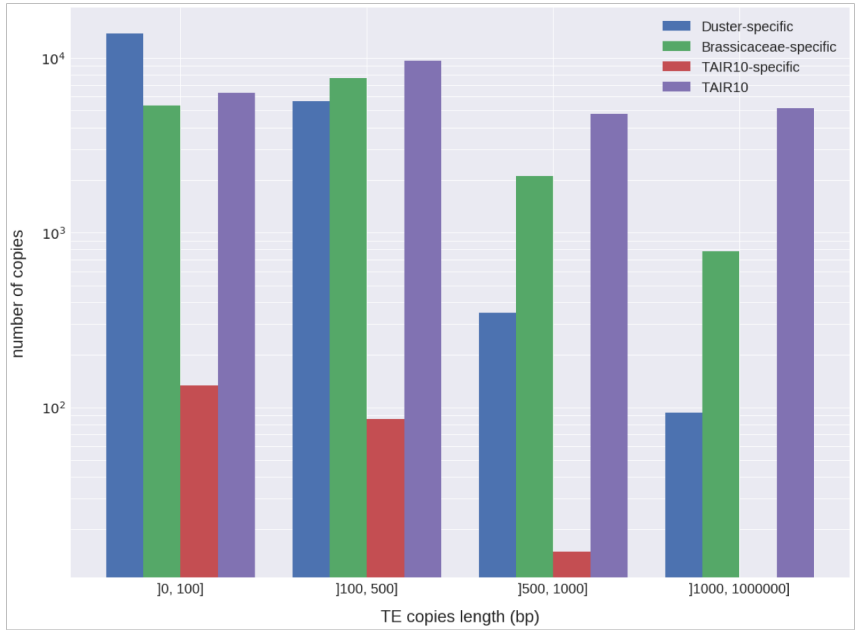
Figure 2: Composition of Duster-specific, *Brassicaceae*-specific, and TAIR10-specific copies. (A) radar plot of di-nucleotid composition of the sequences. (B) The hierarchical clustering of TEs and genes with respect to heterochromatin marks (H3K27me1 and H3K9me2) and euchromatin marks (H3K18ac, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, and H3K9ac).

Figure 3: Effectifs of sequence conservation in orthologous position between species, represented with 3 digits where position 1, 2 and 3 stand for presence, denoted with a 1, or absence with 0, on respectively *A*

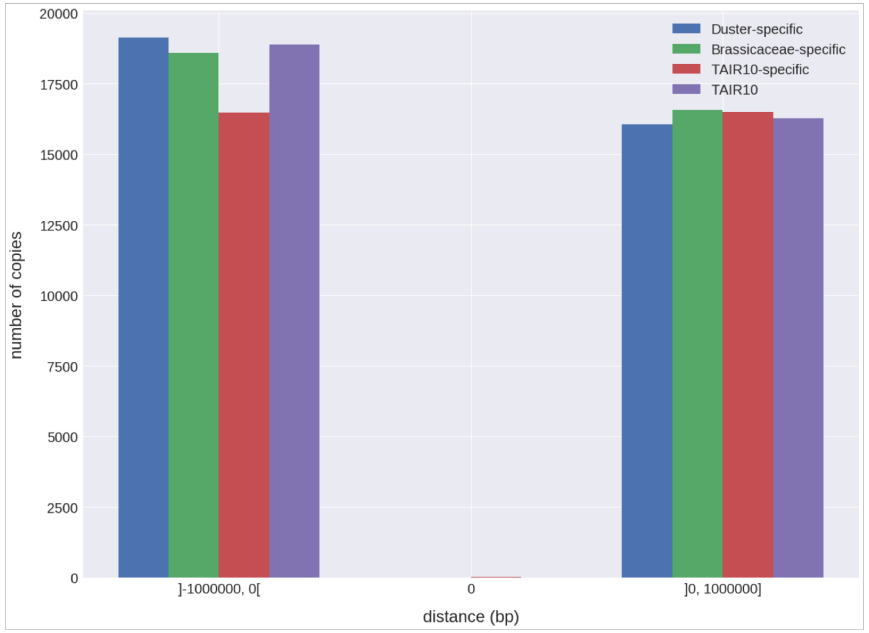
.*lyrata*, *C. rubella*, and *S. parvulum*.

Fig.1

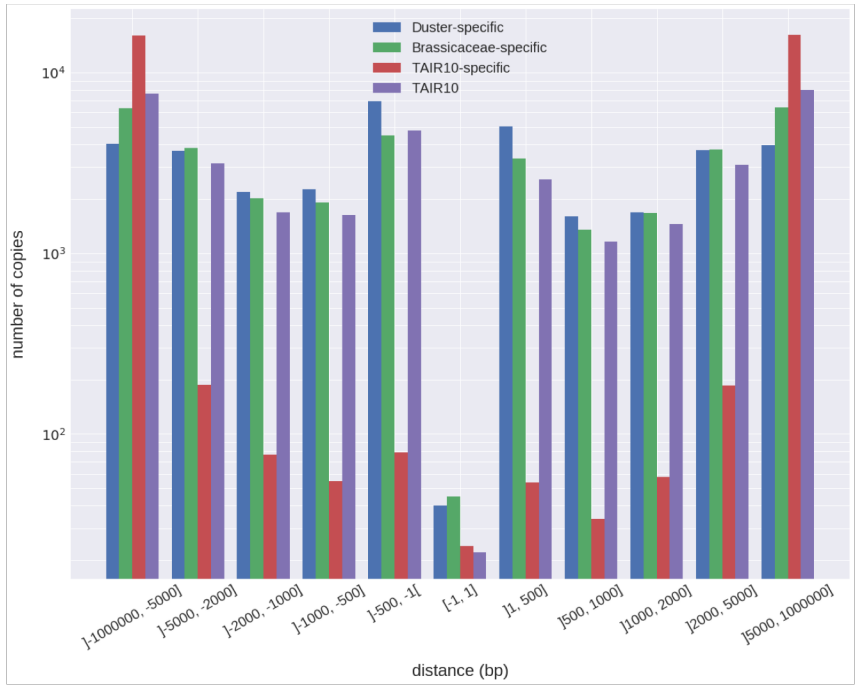
A)


Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
number of copies
10^4
10^3
10^2
]0, 100]
]100, 500]
]500, 1000]
]1000, 1000000]
TE copies length (bp)

B)


Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
number of copies
20000
17500
15000
12500
10000
7500
5000
2500
0
]-1000000, 0[
0
]0, 1000000]
distance (bp)

C)


Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
number of copies
10^4
10^3
10^2
]-1000000, -5000]
]-5000, -2000]
]-2000, -1000]
]-1000, -500]
]-500, -1[
[-1, 1]
]1, 500]
]500, 1000]
]1000, 2000]
]2000, 5000]
]5000, 1000000]
distance (bp)

D)

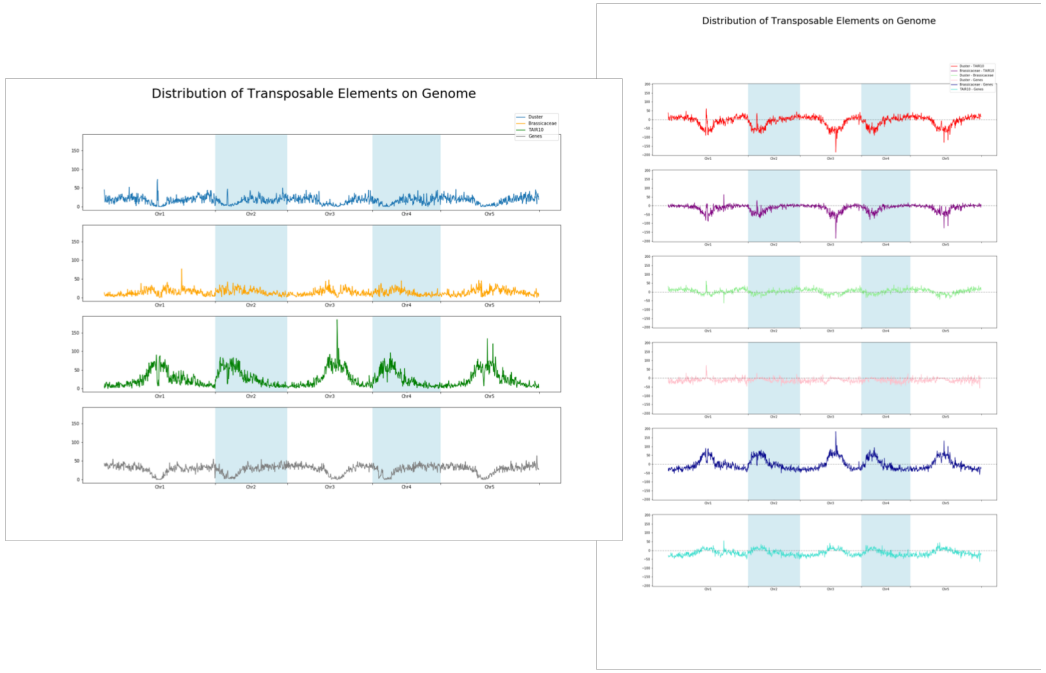
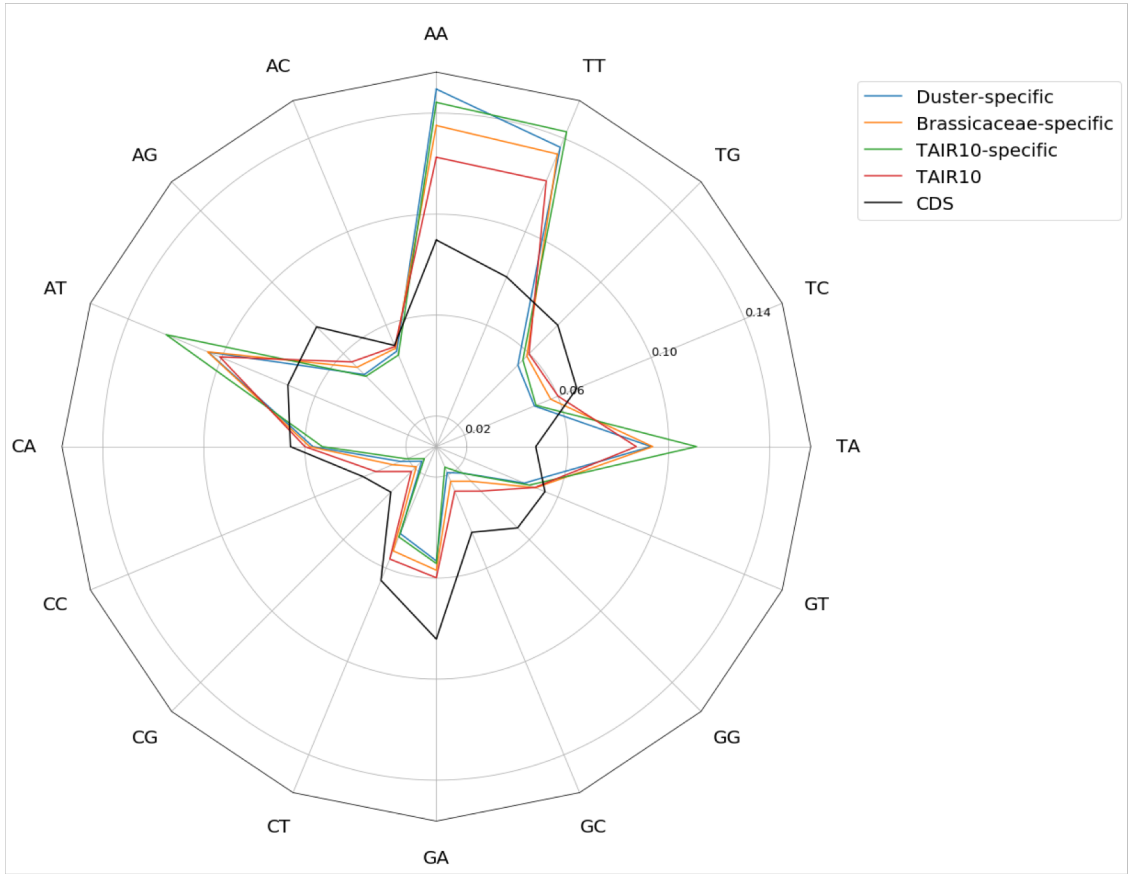

Distribution of Transposable Elements on Genome
Distribution of Transposable Elements on Genome

Fig.2

A)


AA
AC
TT
AG
TG
AT
TC
CA
TA
CC
GT
CG
GG
CT
GC
GA
0.02
0.06
0.10
0.14
Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
CDS

B)

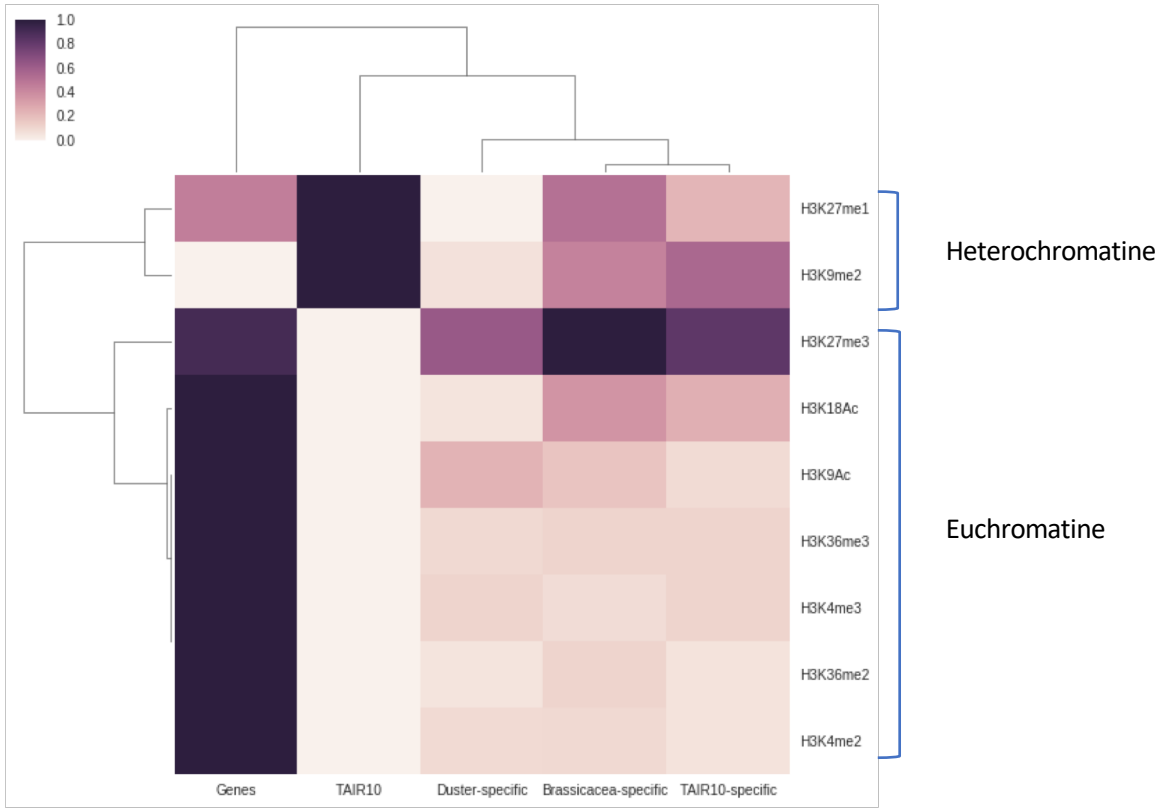
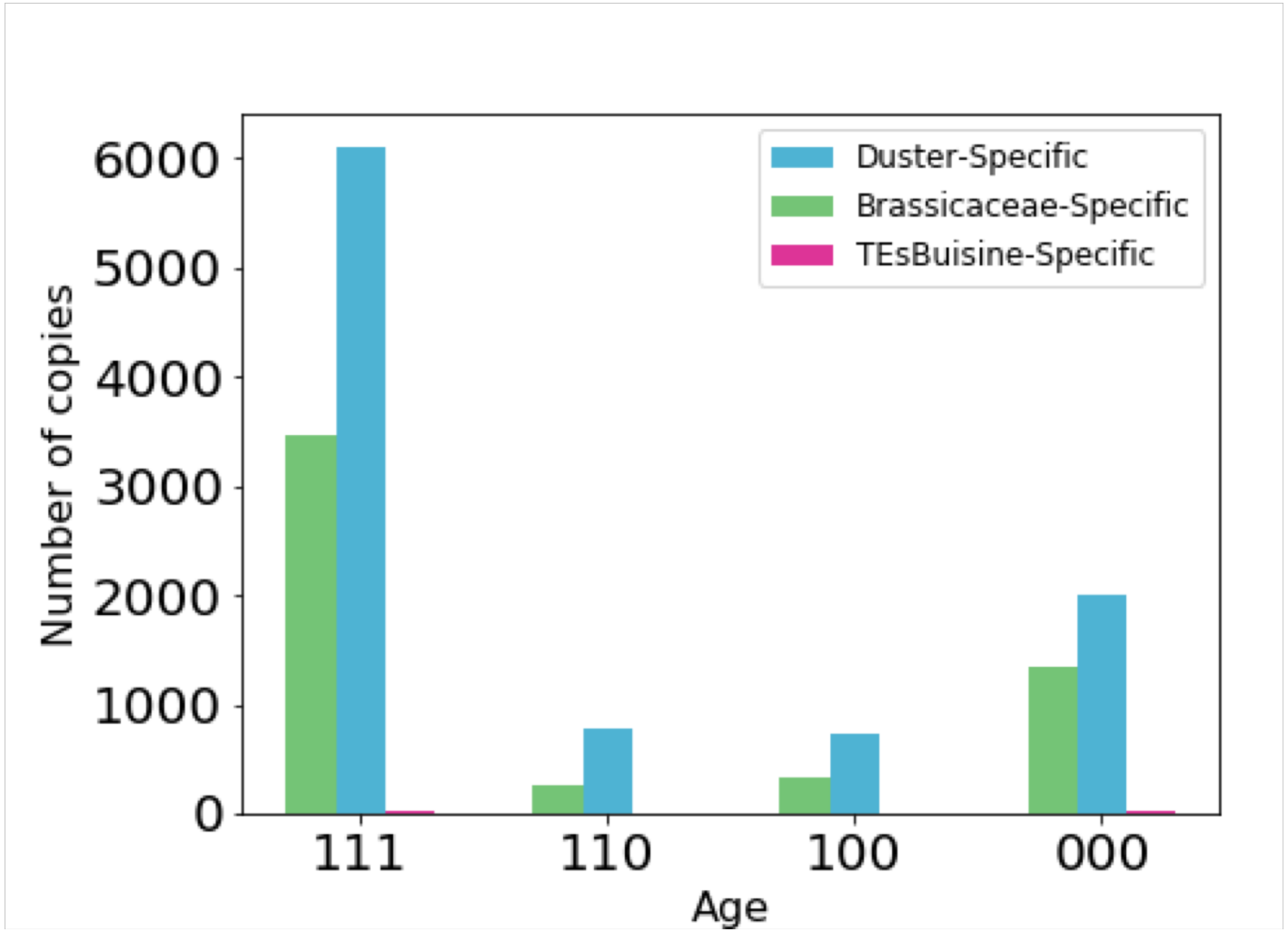

1.0
0.8
0.6
0.4
0.2
0.0
H3K27me1
H3K9me2
H3K27me3
H3K18Ac
H3K9Ac
H3K36me3
H3K4me3
H3K36me2
H3K4me2
Genes
TAIR10
Duster-specific
Brassicacea-specific
TAIR10-specific
Heterochromatine
Euchromatine

Fig.3


Duster-Specific
Brassicaceae-Specific
TEsBuisine-Specific
Number of copies
6000
5000
4000
3000
2000
1000
0
111
110
100
000
Age