Traces of transposable element in genome dark matter co-opted by flowering gene regulation networks.

Agnès Baud[1], Mariène Wan[1], Danielle Nouaud[2], Dominique Anxolabéhère[2-3], 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*
[3]*Université Paris-Sorbonne, Paris, France*
* : Corresponding author

1 Abstract

Transposable elements (TEs) are mobile, repetitive DNA sequences that are the primary contributors to the genome bulk. They contribute 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, based on *k-mers*, able to find degenerated TE sequences. With this new algorithm, we detect up to 10% of the *A. thaliana* genome deriving from TEs not yet identified, bringing to 50% the part of this genome deriving from TE. A significant fraction of them overlap conserved non-coding sequences identified in crucifers and rosids, as well as transcription factor binding sites. They can be over-represented in some gene regulation networks such as the flowering gene network. This suggests a functional role of these sequences being conserved over 100 million years, since the Cretaceous when flowering plants spread.

2 Introduction

Transposable elements (TEs) are mobile, repetitive DNA sequences are the primary contributors to the genome bulk in many organisms. 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, they are also controlled by their host through multiple pathways involving RNAi machinery. Hence, TEs remain quiet in the genome for long period of time. until reactivated after events such as horizontal transfers or genomic shocks. Consequently, they invade genomes reccurently by wave of transposition bursts that ceased rapidly when repressed by host defense mechanisms subsequently triggered. In parallel, TE sequences accumulate mutations. This process results in an inactivation of the sequence that becomes with time too degenerated to be functional, 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 TE sequences over long periods of times. To explore this question, we developed 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]. We generated a library of consensus repeat sequences that we appended to the *A. thaliana* TE reference library in order to compile a "*Brassicaceae*" library. This compiled TE library was used to annotate the *A. thaliana* Col-0 genome to explore 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, achieves highly sensitive detection as it founds one third more TEs than the current official annotation [8]. The fact that many *A. thaliana* TE copies can be detected by consensus sequences built in apparented species can be seen as an evidence supporting the origin of these *A. thaliana* repeats in their common ancestors.

But, our ability to recognize the part of the dark matter deriving from TEs, is still limited by the sensitivity of current alignment algorithms. 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

500 pb of genes. Their epigenetic status, their nucleotide composition, and their long-term conservation in orthologous positions attest an old TE origin. Moreover, they overlap with experimentaly identified Transcription Factor Binding Sites (TFBS) suggesting that they have been co-opted for new functional roles. Hence interestingly, we found their over-representation in 5' of flowering genes. A significant fraction of these TEs overlap TFBSs able to bind TFs known to be involved in flowering. Their overlaps with Conserved Non-coding Sequences (CNS) suggest a long-term impact of TEs on flowering since the Cretaceous period when flowering plant successfully spread over the earth.

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 [6]. Briefly, 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 ecotype, *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/). Orginal coordinates have been projected 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.

TFBS and CNS data

We use TFBS data compiled by Heyndrickx *et al.* [19] which are identified from ChIP-seq experiments, downloaded on http://bioinformatics.psb.ugent.be/cig_data/RegNet/.

CNSs data were retrieved from Van de Velde *et al.* work [20] available at http://bioinformatics.psb.ugent.be/cig_data/Ath_CNS/Ath_CNS.php, and Haudry *et al.*[21] at http://mustang.biol.mcgill.ca:8885/download/A.thaliana/AT_CNS.bed

Binding motifs analysis

Binding motifs have been search with the MEME suite server [22] at http://meme-suite.org.

Orthologous genes analysis

OrthoMCL[23] 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 method.

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 [24] and *revseq* from Emboss 6.1.0 [25] 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. A repeat family can stop to multiply in one species, but continue in a close relative. 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 the other 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.

As shown in our previous study [6], 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 nucleotide 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.

Supplementary file 1 presents performance assessment of duster. It shows that Duster outperforms the standard tools in term of speed. Sensitivity is higher and Specificity is lower for Duster, but coverage is higher, suggesting that our tool detects many more potential TEs not previously known. The false positive rate remains under 0.01.

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 TEs, Brassicaceae, or *A. thaliana* REPET annotation (see Material and Methods). Hence, we got 53341 that we called Duster-specific copies. We did the same with TAIR10 and Brassicaceae annotations to obtain respectively 320 TAIR10-specific and 39489 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 $<10^{-293}$, $9.79\ 10^{-14}$, $<10^{-293}$). Figure 1B shows the distance to the closest 5', or 3', TE copies for each annotated gene. It shows that duster-

specific copies are found more abundant close to genes than other copies (Chi-square *p-value* $<10^{-293}$, compared with *Brassicaceae*-specific, TAIR10-specific, and TAIR10 copies). Similarly, *Brassicaceae*-specific are more abundant than TAIR10-specific copies. They are found more often in up-stream relative to genes (Figure 1B, Chi-square p-value $7.84\ 10^{-61}$), as for *Brassicaceae*-specific, and TAIR10 TE copies (Chi-square p-value respectively $2.37\ 10^{-27}$, $1.78\ 10^{-44}$). On the contrary, there is no significative difference for TAIR10-specific TE copies (Chi-square p-value respectively 0.92) indicating no insertional bias. Figure 1C shows TE copies distribution on the chromosomes. It shows that Duster, and with a lesser trend *Brassicaceae* TE copies, follows the gene chromosomal distribution (see right panel of figure C), where as TAIR10 TEs shows an opposite one. Duster and *Brassicaceae* TEs have unusual chromosomal distribution compared to annotated TEs from TAIR10.

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 Luo *et al.*[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. However, it is worth to note here that H3K27me3 is the most preeminent mark for the method-specific TEs, also known as predominantly present in euchromatine preferentially associated with genes that are expressed at low levels or in a tissue-specific manner [29–32].

TE conservation in flowering plants

The Duster and Brassicaceae TE sequences appearing old, we investigated the conservation of the TE copies searching for overlaps with known conserved non-coding sequences (CNSs) identified in previous studies. We compared with CNSs identified in Crucifers [21] and in rosids [20]. For both datasets, a significant fraction of TE copies overlaps with these CNSs (spanning from 5.32 to 20.4%, see table 1). Some TE sequences overlap CNSs conserved in 12 rosids species: *Arabidopsis thaliana, Carica papaya, Glycine max, Malus domestica, Populustrichocarpa, Fragariavesca, Medicago truncatula, Lotus japonicus, Theobroma cacao, Ricinus communis, Manihot esculenta,* and *Vitis vinifera*. As Fossil rosids are known from the Cretaceous period, being estimated with molecular clock between 125 and 99.6 million years ago [33–35], we show here a remarkable conservation of 1521 and 1213 TE insertions identified by respectively Duster or Brassicaceae methods, over more than 100 million years, the double of what is detected with traditional annotation approach, as avalailable in TAIR10 TE annotation. We show also here that the Duster approach is able to detect more TEs overlapping CNSs than the Brassicaceae method.

Table 1: Number of CNS and percentage of TE annotations that overlaps.

	CrucifersCNSs	Rosids CNSs	RosidsCNSs – score 11
All duster	14774 *14.1%*	20758 *19.8%*	1521 *1.45%*
All brassicaceae	7798 *9.39%*	13702 *16.5%*	1213 *1.46%*
All TAIR10 TEs	1583 *5.32%*	4934 *16.6%*	527 *1.77%*
Duster-specific	8599 *16.1%*	7423 *13.9%*	488 *0.91%*
Brassicaceae-specific	5526 *14.0%*	7607 *19.3%*	605 *1.53%*
TAIR10-specific	44 *13.7%*	53 *16.6%*	3 *0.94%*
5' Duster-specific	907 *14.7%*	1179 *19.0%*	77 *1.25%*
5' Brassicaceae-specific	309 *11.7%*	539 *20.4%*	26 *0.98%*
5' TAIR10-specific	3 *6.38%*	7 *14.9%*	0 *0.0%*

Looking at the the 77 rosids CNSs present in the 12 rosids species associated with duster-specific copies in the 500 bp upstream of genes, we found 67 duster-specific sequences. These sequences analysed with MEME-ChIP [22] found a significant T/G rich motif of 15 nucleotids (E-value 1.1 10^{-6}) present on 43 sequences. This motif matches with MA1281.1 (AT5G02460), MA1274.1 (OBP3), MA1268.1 (AT1G69570), MA1267.1 (AT5G66940), MA1272.1 (AT2G28810), MA1279.1 (COG1), MA1158.1 (MGP) with E-values < 1.85 10^{-2}, refered by the JASPAR database [36] as C2H2 zinc finger factors of Dof-type. Some of these motifs relate to: (i) AT3G55370 (OBP3), known to encode a nuclear localized DOF domain containing TF expressed primarly in root, responsive to salicylic acid in leaves and petals, (ii) AT1G69570 (CDF5) is invoved in flower development and photo-period, (iii) AT1G29160 (COG1) acts as a negative regulator in the phytochrome-mediated light responses, (iv) AT1G03840(MGP) regulates tissue boundaries and asymmetric cell division and (v) AT5G02460.1 is probably involved in early processes for vascular development.

We then looked in details 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 upstream 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 that the Duster-specific set contain older 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. It 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. This result still illustrates the interest of our cross-species TE annotation approach and the efficiency of Duster over the REPET annotation procedure.

TE contribution to architecture of gene regulatory networks.

We investigated the functional role of these TE sequences that can be supposed to be co-opted for some regulatory property. We chose two Gene Regulatory Networks (GRNs) for which TEs can be suspected to

play a role. In particular genes controlling flower development in *Arabidopsis thaliana* are a good candidates as some alleles have been reported to be controlled by a TE sequence in *A. thaliana*: the FLOWERING WAGENINGEN (FWA) locus [36], and FLOWERING LOCUS C (FLC) [37,38].

We considered genes reported in the Chen et al. 2018 [39] paper which describe the architecture of GRN controlling flower development in *Arabidopsis thaliana*. We searched in the 500bp upstream regions of these genes the presence of Duster-specific, Brassicaceae-specific, and TAIR10-specific TEs.

Duster-specific regions are enriched in 5' of flowering genes, 36.3% of presence to be compared to 22.5% considering all genes (chi-square p-value=3.49 10^{-5}, table 2). Brassicaceae-specific and Buisine-specific regions look not enriched with respectively 8.28% compared to 9.63% (chi-square p-value=0.56), and 0% *vs* 0.17% (chi-square p-value=0.60).

Table 2: Method-specific TE overlap with the 500 bp gene upstream

	Duster-specific	Brassicaceae-specific	TAIR10-specific
All genes	22.5%	9.63%	0.17%
All flowering genes	36.3%	8.28%	0.0%
Stress genes	21.8%	10.6%	0.0%

To futher explore over-representation of Duster-specific and Brassicaceae-specific TEs in GRNs, we selected stress GRNs genes that are also suspected to be linked to TEs as reported by several studies suggesting that they can be triggered during plant stress responses including salt [40], wounding [41], bacteria [42], and viruses [43]. We took genes expressed in various stress conditions reported in Barah et al. [44]. We searched in the 500bp upstream regions of these genes the presence of Duster-specific, Brassicaceae-specific, and TAIR10-specific TEs (table 2), and found that Duster-specific regions do not appear enriched in the 500bp upstream of Stress genes (chi-square p-value=0.82), nor Brassicaceae-specific TE (chi-square p-value=0.67) or TAIR10-specific TEregions (chi-square p-value=0.59).

TEs and transcription factors binding sites

The conservation and the over-representation shown previously, suggest that these TEs may play a functional role. We then investigated if they can regulate gene expression by looking if they can bind TFs. We looked at the co-occurrence of transcription factors binding sites (TFBSs) identified with 27 TFs ChIP-seq experiments from Heyndrick et al. [19] and all studied TE annotations.

We show that Duster regions are enriched in TFBSs compared to Brassicaceae and TAIR10 with respectively 29.0% 19.5%, and 14.6% (Table 3). Method-specific regions amplified this observation with 33%, 27% and 24% of overlap with TFBSs (respectively). This trend is increased again when looking at regions in the 500bp upstream of genes (respectively 45% and 29%, TAIR10-specific being not really testable due to low counts). Interestingly, those regions contain 840 and 283 for respectively Duster-specific and Brassicaceae-specific TFBS regions linking more than 7 TFs called hereafter *hot TFBS*s. These regions suggest the presence of a hub target genes that are attributed the important function of providing crosstalk between different processes [45].

Table 3: Method-specific TE overlaps with TFBS identified by ChIPseq

	Hot TFBSs	TFBSs
All duster	1.10%	29.0%
All brassicaceae	0.51%	19.5%
All TAIR10 TEs	0.40%	14.6%
Duster-specific	1.57%	33.2%
Brassicaceae-specific	0.72%	26.7%
TAIR10-specific	0.94%	24.1%
5’ Duster-specific	2.15%	45.0%
5’ Brassicaceae-specific	0.79%	29.9%
5’ TAIR10-specific	2.13%	17.0%

Interestingly, looking in the 5' of flowering genes, we found ten key TFs known to be involved in flowering that can be found bound to Duster-specific and Brassicaceae-specific regions (AGL-15, AP1, PI, AP3, FLM, SEP3, AP2, SOC1, SVP, LFY). Few are bound to circadian rhythm and light response TFs (PIF3, PRR5, PRR7) and development (GL3). Most of them are found in Duster-specific regions, very few with Brassicaceae-specific, and none with TAIR10-specific. Interestingly, some Duster-specific regions are associated with several TFBSs.

We found 3109 and 2854 Duster-specific regions that overlap respectively Crucifers and Rosids CNS, and a TFBS. Among them, 188 are highly conserved as they overlap CNSs present in the 12 rosids species used for their identification, suggesting a presence in the common ancestors of the rosids for more than 100 Myr ago. In addition, we counted 24 of these highly conserved Duster-specifc regions overlapping a *hot TFBS*s suggesting the presence of a highly conserved hub target genes providing crosstalk between different processes. The top five highly conserved TFBSs from Duster-specific regions are AGL-15, PRR5, AP1, PI, and SEP3 (respectively 739, 570, 468, 429, 420 occurences), all involved directly in flowering process, except PRR5 which is more related to circadian rhythm and light response.

From the CNSs present in the 12 rosids species associated with duster-specific copies in the 500bp upstream of genes, we found 67 duster-specific sequences. Among them, 42 are target genes of floral regulators according to Chen *et al.* [39] and 18 of these display a colocalization of Duster-specific regions with a highly conserved CNSs and a TFBS (Table 4).

Table 4 : Eighteen floral regulators genes from Chen et al. [39] with a Duster-specific region in the upstream 500bp which colocalize with highly conserved CNSs and TFBSs

Gene	TFBS	Profil from MEME-suite
AT1G03970	FHY3	SOC1
AT1G20930	AGL-15	SOC1
AT1G25380	AP3; PI	SOC1
AT1G49760	AGL-15	
AT1G75110	GL1	
AT2G05920	AGL-15; PIF4	SOC1
AT2G30400	AGL-15	
AT2G30970	AGL-15; AP1; SEP3	
AT2G33750	SEP3	SOC1
AT2G41370	AGL-15; AP1; AP3; PI	
AT3G02040	AP3; AGL-15; PI	SOC1
AT3G14172	AGL-15	
AT3G19170	AGL-15; PIF4; PIF3; PI; AP3	SOC1
AT4G00810	PIF	SOC1
AT4G32680	FY3	SOC1
AT4G37260	SEP3	SOC1
AT5G03680	AP1; SEP3; FLM; PIF4	
AT5G13990	PIF4	

Using the *MEME Suite* we identified 10 of them with a sequence corresponding to a binding motif of the SOC1 TF. Thirteen of the 67 downstream genes are controled by floral regulation motifs of one or several type II TF-MADSs (AG, AP1, AP2, AP3, BLR, ETT, FLM, JAG, LFY, PI, RGA, SEP3, SVP).

5 Discussions

A need for new dedicated repeat detection algorithm

RepeatMasker [46], Censor [47,48], 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 statistical 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 in real sequences. Consequently, the model rejects some alignment differently according to the sequence lengths, and with a model that seems discutable. In Duster, we keep all regions that match with two *k-mers*, and the parameters we empirically 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.

The presented work highlights 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 impact of TE copies

TEs are important source of variation on which selection can operate to evolve species. In plants many examples show TE copies providing new phenotypes [49]. We call this phenomenon « domestication » when TE sequences are retrain in new genes, or « co-option » when TE insertions affects existing ones. TE sequences that became functional for the host are maintained conserved by selective pressure and can be recognized as such over long period of time. Other TE copies devoided of host function are gradually removed from the genome by accumulating point mutations and deletions. Note however that, because they are functional, gene regulatory regions accumulate point mutations and deletions at a lesser rate than other regions. Consequently, TE insertions in these regions, even if neutral, could be difficult to remove once installed.

Binding on TE sequences, TF control TEs transpositional activity. Consequently, their TFBS are spread out across the genome when multiplied throughout the genome. Sometimes, the same TE family is inserted close to several genes. Hence, nearby genes may become regulated by the same TF which can make them evolve

into gene regulation networks. Genome-wide assessment revealed that hundreds of TEs have been co-opted into regulatory regions of mammalian genes [50,51]. TEs have also been involved in both the creation of new regulatory networks [52,53] and in the rewiring of preexisting ones [54]. Such networks are observed, for example, with the Dayspleeper gene in *A. thaliana* [55]. Interestingly here, this gene share caracteristics with hAT transposases, suggesting a domestication as a new TF. Another interesting example is the retrotransposon ONSEN in *Arabidopsis* [56]. Thieme *et al.* [57] showed that following heat stress-dependent mobilization of ONSEN, the progenies of the treated plants contain up to 75 new ONSEN insertions. Progenies with additional ONSEN copies show a broad panel of environment-dependent phenotypic diversity. This suggests that a part of the new TE insertions affect some gene expression in a temperature-dependant manner. It suggests that TEs sequences might have played a role in individual local adaptation for the mutations they induce during burst of transposition. Some of these transposition burst might be the result of activation from environmental stresses, that in turn promote environment-sensitive phenotypes.

However, little is known about this long-term impact of TEs. In this study, CNS data suggest that some identified TEs have been inserted more than 100 Myr ago, an age corresponding to the Cretaceous period. The most important evolutionary event during this period was, perhaps, the spread of flowering plants (Angiosperms) which colonize all the earth. Flowering plants were particularly successful as they colonized and replace the old competing flora. As key genome evolution player, TEs have certainly played an important role. We may have detected few TE insertions that have played such a role. Indeed, Duster-specific copies appears to be old, degenerated, short, and surprisingly close to genes in 5', at a distance corresponding to the gene regulatory regions. They would be specifically maintained in these regions because they bring function for the host, probably a regulatory module acting on the neigbour gene.

The Duster-specific TEs we identified might have play an importante role far in the past to build new pathways for adaptating flowering plants to their environnements. Indeed, we found Duster-specific copies over-represented in 5' gene regions of the flowering GRN. A significant fraction of these copies overlaps TFBSs that bind TF known to be involved in flowering. Moreover, the histone H3K27me3 mark is identified predominantly for the method-specific TEs (see figure 2). This histone mark has been described associate with low level expressed or tissue-specific genes [32] as expected for example for flower development.

Our results suggest a possible link between the success of flowering plants during the Cretaceous and a role of TEs co-option in flowering GRN. This is very tempting, but further analyses are needed to establish this causal role. This study is a first step toward this analysis, providing candidates yet unknown.

Flowering have been extensively studied providing whealth of available data. Consequently, the data we used are clearly biased towards flowering. Impacts affecting other GRNs might be still discovered with our ancient TE annotation as new data will be made available.

Interestingly, our results suggest that identifying the very old TE copies could help to identify TE regulatory modules selected long time ago. This could support the TFBS detection by ChipSeq experiment, but also suggest a TE origin to many TFBS.

6 Conclusions

In this study, we investigate the TE contribution to Arabidopsis genome bulk at timescale still inaccessible for concurrent approaches, thanks to a new tool we developed called Duster. Duster implements a new efficient algorithm that allows to identify a new 10% of nucleotids annotated as TE. 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 result helping to understand plant evolution and plant adaptation, by providing clues that identifies ancient TEs remnants in genes regulatory regions, providing potential regulation modules. Some TE copies identified here may have been selected far in the past for adaption to changing environments.

7 Acknowledgment

We are gratefull to KlaasVandepoele for reading a first version of this manuscript and to point out interesting TFBSs and CNSs data, that help us to improve the manuscript. We also 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 Author contributions

HQ: conceive the study, develop Duster, run the software. AB identified method-specific sets and describe their general features. MW did conservation analysis among *Brassicaceae*. DA, DN, HQ, did CNS, TFBS and profile analysis. DA and HQ draft the manuscript. All authors read and commented the manuscript.

9 Materials & Correspondence.

All material and information request should be addressed to Hadi Quesneville at hadi.quesneville@inra.fr

10 Code availability

Duster code is available on github distributed as part of the TEfinder package at https://github.com/urgi-anagen/TE_finder

11 Data availability

Brassicaceae TE copies from *Arabidopsis thaliana* Col-0 ecotype, *Arabidopsis lyrata*, *Capsella rubella*, *Arabis alpina*, *Brassica rapa* and *Schrenkiella parvula*, used for Duster annotation, and TAIR10 TE, Duster, Brassicaceae annotations can be downloaded from https://urgi.versailles.inra.fr/files/sequence/PubAthaDarkmatterData/PubAthaDarkmatterData.zip

12 Figure legends

Figure 1: **Structural characteristics of Duster-specific, *Brassicaceae*-specific, TAIR10-specific, and TAIR10TE copies.** (A) TE length distribution, (B) distance to the closest 5', or 3', TE copies for each annotated gene, (C) TE copies distribution on the chromosomes. Left panel: density plot in 100kb windows overlapping by 10kb. Right panel: density differences in each 100kb bins

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**. They are 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*.

14 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. [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. Heyndrickx KS, de Velde JV, Wang C, Weigel D, Vandepoele K. A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana. The Plant Cell.

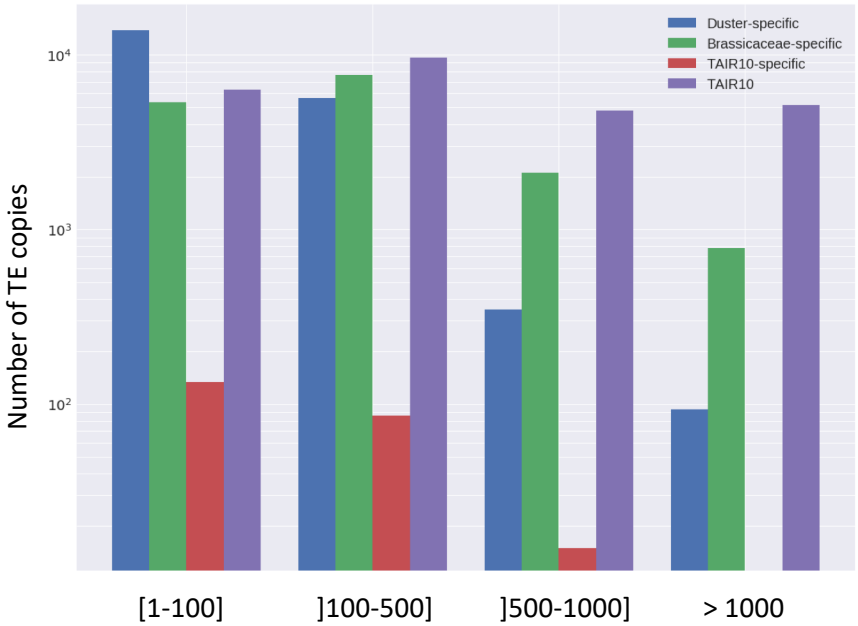
2014;26: 3894–3910. doi:10.1105/tpc.114.130591
20. de Velde JV, Heyndrickx KS, Vandepoele K. Inference of Transcriptional Networks in *Arabidopsis* through Conserved Noncoding Sequence Analysis. The Plant Cell. 2014;26: 2729–2745. doi:10.1105/tpc.114.127001
21. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genetics. 2013;45: 891–898. doi:10.1038/ng.2684
22. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–W208. doi:10.1093/nar/gkp335
23. 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
24. Eddy SR. SQUID - library of functions for biological sequence analysis Copyright (C) 1992-2002 Washington University School of Medicine. Available: http://eddylab.org/software/squid/squid-1.9g/
25. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.
26. 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.
27. 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
28. 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
29. Turck F, Roudier F, Farrona S, Martin-Magniette M-L, Guillaume E, Buisine N, et al. Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. PLOS Genetics. 2007;3: e86. doi:10.1371/journal.pgen.0030086
30. Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, et al. Whole-Genome Analysis of Histone H3 Lysine 27 Trimethylation in Arabidopsis. PLoS Biol. 2007;5. doi:10.1371/journal.pbio.0050129
31. Oh S, Park S, Nocker S van. Genic and Global Functions for Paf1C in Chromatin Modification and Gene Expression in Arabidopsis. PLOS Genetics. 2008;4: e1000077. doi:10.1371/journal.pgen.1000077
32. Jacob Y, Stroud H, Leblanc C, Feng S, Zhuo L, Caro E, et al. Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature. 2010;466: 987–991. doi:10.1038/nature09290
33. Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proceedings of the National Academy of Sciences. 2004;101: 1904–1909. doi:10.1073/pnas.0308127100
34. Magallón S, Castillo A. Angiosperm diversification through time. American Journal of Botany. 2009;96: 349–365.
35. Hohmann N, Wolf EM, Lysak MA, Koch MA. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. The Plant Cell. 2015; tpc.15.00482. doi:10.1105/tpc.15.00482
36. 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
37. Liu J, He Y, Amasino R, Chen X. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. Genes Dev. 2004;18: 2873–2878. doi:10.1101/gad.1217304

38. Strange A, Li P, Lister C, Anderson J, Warthmann N, Shindo C, et al. Major-Effect Alleles at Relatively Few Loci Underlie Distinct Vernalization and Flowering Variation in Arabidopsis Accessions. PLOS ONE. 2011;6: e19949. doi:10.1371/journal.pone.0019949
39. Chen D, Yan W, Fu L-Y, Kaufmann K. Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. Nature Communications. 2018;9. doi:10.1038/s41467-018-06772-3
40. 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
41. 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
42. 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
43. 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
44. Barah P, B N MN, Jayavelu ND, Sowdhamini R, Shameer K, Bones AM. Transcriptional regulatory networks in *Arabidopsis thaliana* during single and combined stresses. Nucleic Acids Research. 2016;44: 3147–3164. doi:10.1093/nar/gkv1463
45. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5: 101–113. doi:10.1038/nrg1272
46. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015.
47. 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.
48. 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
49. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14: 49–61. doi:10.1038/nrg3374
50. 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
51. 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
52. 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
53. 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
54. 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
55. Bundock P, Hooykaas P. An Arabidopsis hAT-like transposase is essential for plant development. Nature. 2005;436: 282–284. doi:10.1038/nature03667
56. 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
57. Thieme M, Lanciano S, Balzergue S, Daccord N, Mirouze M, Bucher E. Inhibition of RNA

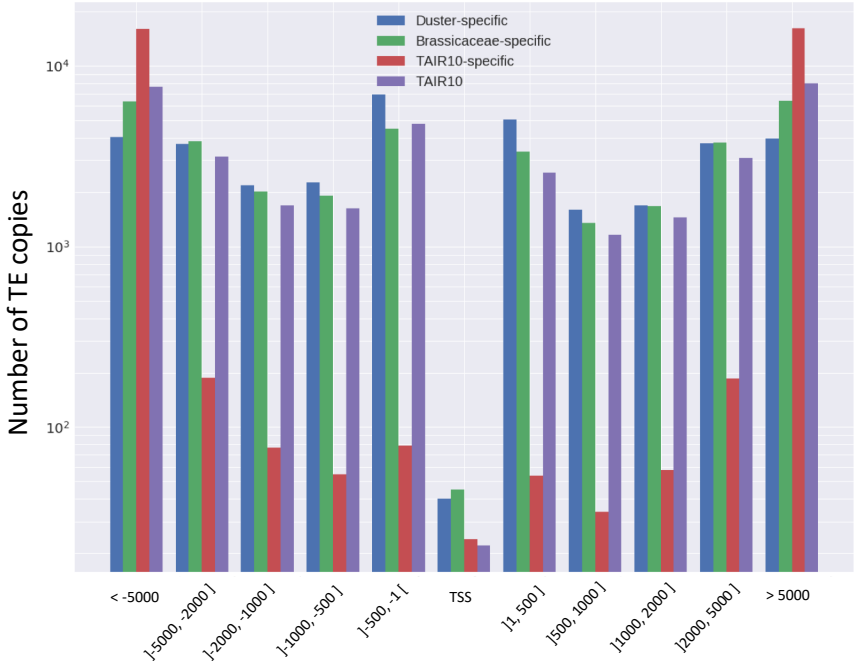
polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biology. 2017;18. doi:10.1186/s13059-017-1265-4

Fig.1

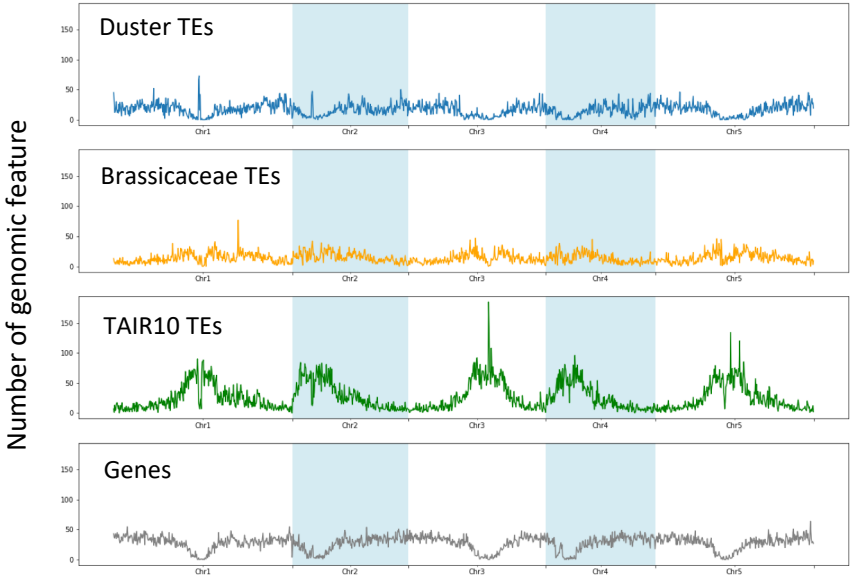
A)


Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
Number of TE copies
10^4
10^3
10^2
[1-100]
]100-500]
]500-1000]
> 1000

B)


Duster-specific
Brassicaceae-specific
TAIR10-specific
TAIR10
Number of TE copies
10^4
10^3
10^2
< -5000
]-5000, -2000]
]-2000, -1000]
]-1000, -500]
]-500, -1 [
TSS
]1, 500]
]500, 1000]
]1000, 2000]
]2000, 5000]
> 5000

C)


Number of genomic feature
Duster TEs
Brassicaceae TEs
TAIR10 TEs
Genes
150
100
50
0
Chr1
Chr2
Chr3
Chr4
Chr5

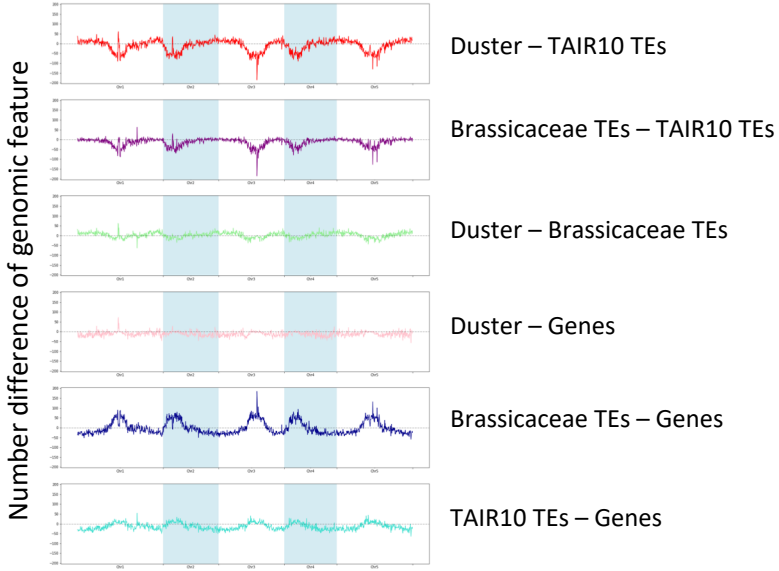
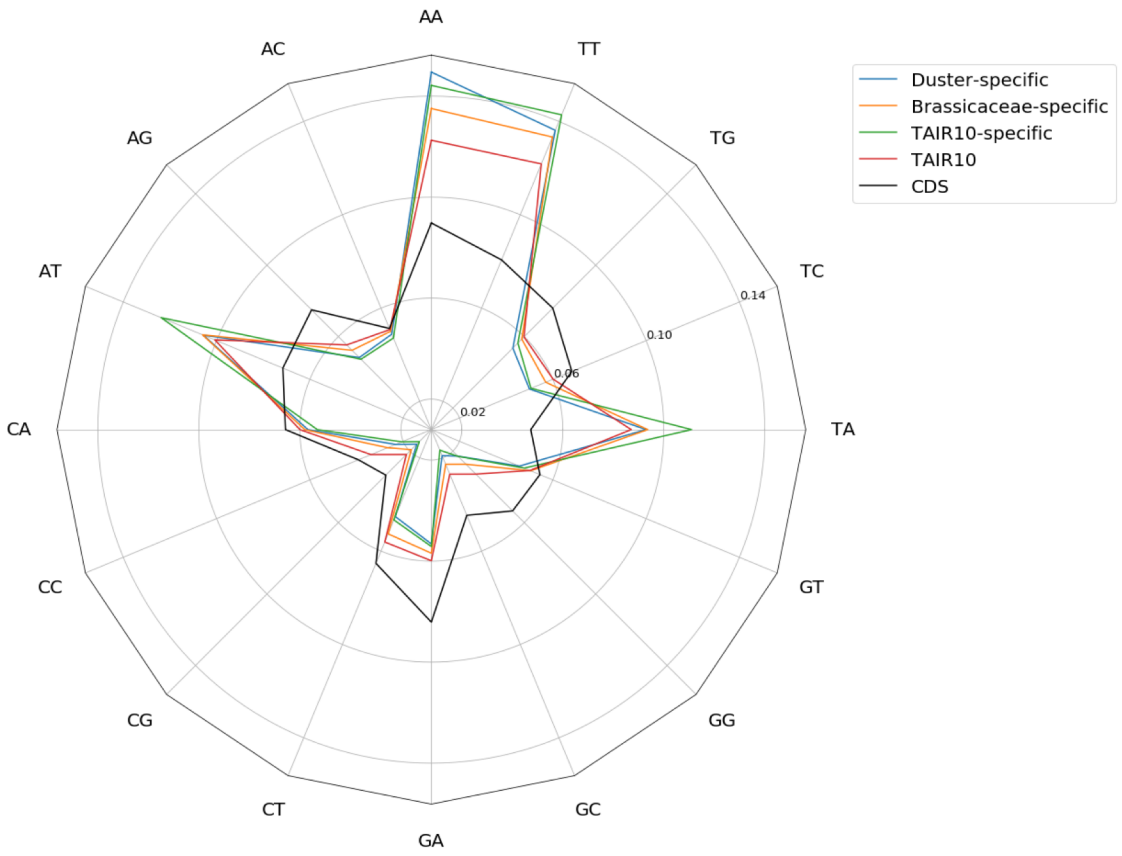

Number difference of genomic feature
Duster – TAIR10 TEs
Brassicaceae TEs – TAIR10 TEs
Duster – Brassicaceae TEs
Duster – Genes
Brassicaceae TEs – Genes
TAIR10 TEs – Genes

Fig.2

A)


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

B)

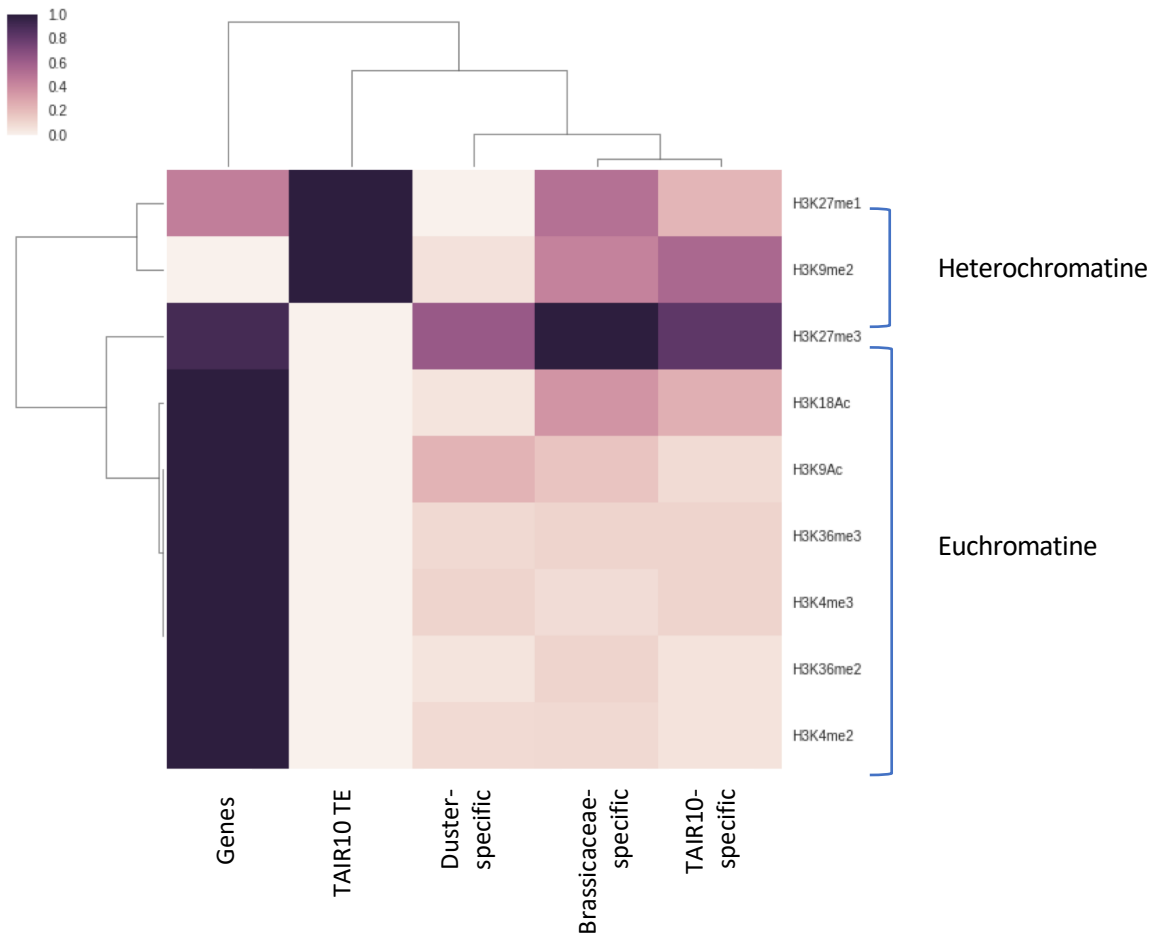
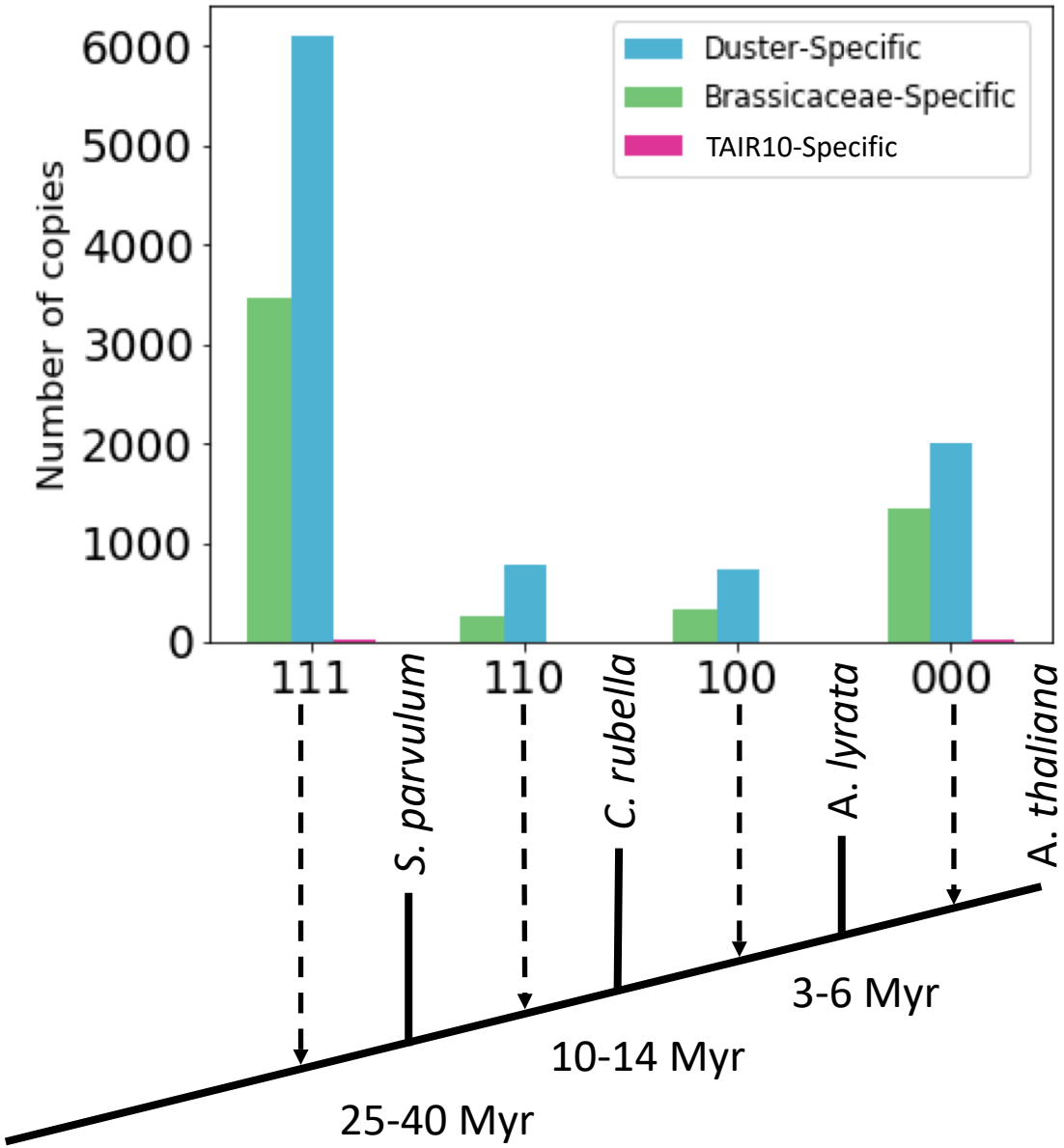

1.0
0.8
0.6
0.4
0.2
0.0
H3K27me1
H3K9me2
H3K27me3
H3K18Ac
H3K9Ac
H3K36me3
H3K4me3
H3K36me2
H3K4me2
Heterochromatine
Euchromatine
Genes
TAIR10 TE
Duster-specific
Brassicaceae-specific
TAIR10-specific

Fig.3


Duster-Specific
Brassicaceae-Specific
TAIR10-Specific
Number of copies
6000
5000
4000
3000
2000
1000
0
111
110
100
000
S. parvulum
C. rubella
A. lyrata
A. thaliana
3-6 Myr
10-14 Myr
25-40 Myr