The *Brachypodium distachyon* pangenome highlights Transposable Element dynamics in the species

Somia Saidi[1,2], Mathieu Blaison[1,2], María del Pilar Rodríguez-Ordóñez[1,2], Johann Confais[1,2*] & Hadi Quesneville[1,2*]

[1] Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France.

[2] Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France.

*Corresponding Authors

Hadi Quesneville: hadi.quesneville@inrae.fr

Johann Confais: johann.confais@inrae.fr

Abstract

The role of transposable elements (TEs) in host adaptation can be explored with a pangenomic approach. Individuals of the same species undergo independent TE insertions, causing genetic variability upon which natural selection acts. This can lead to improved adaptation of individuals to their environment. The advent of third-generation sequencing has enabled use of multiple whole-genome *de novo* assemblies for a given species, avoiding bias introduced by a single reference genome. We developed a new pipeline, panREPET, for such data. It compares TE copies between each pair of individuals then identifies copies shared by a group of individuals. This gives the exact sequence and genomic context of each TE copy. We describe here TE insertions shared among 54 *Brachypodium distachyo*n genomes. We were able to date two major TE bursts corresponding to major climate events: 22 kya during the Last Glacial Maximum and 10 kya during the Holocene.

Introduction

Transposable elements (TEs) are mobile DNA elements that can invade genomes by transposition. Their insertions may be neutral to the host or deleterious, for instance, by disrupting gene expression[1,2,3]. Host defense may involve epigenetic regulation[1] or purifying selection purge deleterious TE[4]. TE insertions may also be beneficial in fostering genome evolution through introduction of functional novelties[5]. In natural populations, TE insertion polymorphism frequency is subject to drift, selection and migration[6]. Understanding how TEs can enable a species to adapt to the local environment requires a detailed study of their insertions in genomes of different individuals. At the species level, the pangenomic approach seeks to detect intra-specific diversity of TEs and to retrace their dynamics. A TE pangenome can be described by (i) TE insertions present in all individuals of the species (core-genome), (ii) insertions present only among a subset of individuals (dispensable-genome), and (iii) individual-specific insertions. Some TEs are known to be sensitive to abiotic stress and can, therefore, affect nearby genes in response to environmental conditions[5,7,8]. Using a pangenomic approach, studies have shown that recent TE insertions, at the intra-specific level, are involved in local adaptation. For example, a population genomic study of the *Arabidopsis thaliana* natural accessions revealed extensive intraspecific variation in ATCOPIA78 copies, which correlates with annual temperature variation[9]. A study of rice

suggested that TE activation might be induced by external stimuli instead of resulting from alterations of genetic factors involved in TE silencing pathways[10].

Detecting TE polymorphism is challenging because of their repeated nature. Regarding TE pangenome analysis based on the alignment of reads from different genomes of the species to an assembled reference genome, the challenge is to detect TE insertions absent from the reference. Moreover, the TE genomic environment of regions absent from the reference genome is unavailable as not assembled. The advent of third-generation sequencing now makes it possible to use several *de novo* assembled genomes of the same species. We developed a new pipeline, called panREPET, to handle them without depending on a reference genome. panREPET overcomes the reference dependency by comparing TE copies between each pair of *de novo* assembled genomes. panREPET can highlight and estimate dates of TE dynamics at the species level for all types of TE sequences.

Over 40% of the earth's surface is covered by grasslands, and the annual Mediterranean plant *Brachypodium distachyon* has emerged as a grass genomic model[11]. Since the genus is undomesticated, this model allows to avoid bias introduced by artificial selection on agricultural plants. In addition, because it grows in climate change hotspots, it empowers investigation of stress response process[11]. Several *de novo* whole-genome assemblies are available *for B. distachyon*, from 100 bp paired-end Illumina short reads (median genome coverage of 92x, mean assembled genome size of 268 Mbp, which is very close to the 272 Mbp reference genome size)[12]. The outstanding questions on *B. distachyon* TE that we address here are: what are the main transposition events in the species, which TE families transposed, and when did the transposition occur? For the oldest insertions, the question of the mechanisms of their conservation can be raised. For bursts of recent transposition on the scale of a few individuals, the question is whether external factors have contributed to their activation.

Here, we present the panREPET pipeline and its application to 54 genomes from *B. distachyon*. We describe the TE evolutionary history by estimating transposition bursts events precisely. We found correlation of the date of these events with climate changes that occurred in the past. Therefore, climate factors may, in some cases, explain TE dynamics.

Results

panTEannot

Our pangenomic approach aims to capture as much diversity as possible from all 54 genomes. Consequently, we annotated each genome for its TE content independently using the same TE reference library **(Fig. 1a)**. The TE reference library was built from the Bd21 v3.2 genome sequence using the TEdenovo pipeline from the REPET package[13] **(see Materials and Methods - Genome analysis tools)**. This TE annotation process is time-consuming and requires a fast method for serial annotation of the 54 genomes. Hence, we modified the TEannot pipeline from the REPET package[14] into a lighter version. Specifically, as we are interested in inter-individual TE variability, which is due in particular to recent TE transpositions, we focused on non-degenerate sequences, i.e. complete TE copies as a proxy. In the TEannot pipeline of REPET, the

similarity search step is done by Blaster[15], Censor[16] and RepeatMasker[17] to maximize the sensitivity. As complete TE copies are easy to detect, there is no need to maximize the sensitivity. Hence, we kept a single tool, Blaster, and removed Censor and RepeatMasker. With the same rationale, we also removed SSR search, Spurious hits removal and Fragment connection (TEannot steps 4, 5 and 7) as they have low impact on detection of complete TEs. To improve parallel computation we implemented it with the Snakemake framework[18] removing MySQL dependencies and use of a cluster job scheduler such as Sun Grid Engine or Slurm. This new implementation is called panTEannot. To benchmark its performance, we compared panTEannot to TEannot from REPET v3.0. We performed a complete TE annotation with TEannot from REPET v3.0 on the genome sequence reference Bd21 to be used as the golden standard. We also performed annotation with TEannot REPET v3.0 run without Censor and RepeatMasker, to mimic analysis by panTEannot. We compared the different TE annotations against the golden standard at the nucleotide level **(Table 1)**. We calculated sensitivity and specificity as in Baud *et al.*[19] **(see Materials and Methods - Genome analysis tools)**. TEannot from REPET 3.0 has a better sensitivity (0.9985) than panTEannot (0.8951), but panREPET has a better specificity (0.9267 against 0.8979) and is faster. As in a pangenomic context, we are more interested in detecting true positives than all potential copies including very old and thus degenerate ones, we considered this result a satisfactory compromise.

On the 54 genomes, panTEannot took 48 hours on a 32 core 64 Gb RAM virtual machine which we consider to be a good performance. The average TE sequence coverage percentage of the genome sequence is 29.9%. Unassembled genome parts correspond mostly to TEs due to their repetitive content. We observed that the percent coverage on TEs depends on the assembly quality: genomes with a TE coverage percentage below 30% are characterized by an assembled genome size after removing 'N' occurrences lower than 95% of the reference genome, corresponding to 256 Mbp **(Supplementary Table 1)**. This threshold led us to remove from our study ABR8, Bd3_1, Bd30-1, BdTR11a, BdTR12c, BdTR5i, BdTR8i, Gaz-8, Koz_3, Mur1, Tek-2 and Tek-4. Note that the BdTR11a genome is larger than 256 Mbp (about 266 Mbp) but presents only a TE sequence coverage percentage of 19%. Another useful metric for TE detection quality is the contig N50. The longer the contigs are, the better TEs are detected. Hence, we removed BdTR11a because in addition to not having a large assembled genome size, its contig N50 is low compared to others (17 kbp against 24 kbp in median).

Annotation type	TE coverage	Sensitivity (Sn)	TE intersect	Specificity (Sp)	Gene intersect	Accuracy (ACC)	Real time
Golden standard: TEannot from REPET Blaster + Censor + RepeatMasker Without step 6	0.3705	**0.9985**	1	**0.8979**	0.1550	0.9468	19:25:40
TEannot from REPET Blaster + RepeatMasker Without step 6	0.3583	0.9629	0.8799	0.9117	0.1199	0.9366	15:22:13
TEannot from REPET Blaster Without step 6	0.3389	0.9092	0.7267	0.9245	0.1001	0.9170	12:50:55
panTEannot	0.3550	0.9459	0.8994	0.9126	0.1332	0.9288	00:39:46

Blaster + RepeatMasker							
panTEannot Blaster	0.3334	**0.8951**	0.7127	**0.9267**	0.0984	0.9114	**00:20:02**

Table 1 : Comparison of performances of TEannot from REPET 3.0 with panTEannot on Bd21 genome sequence v3.2. Real times were obtained on a 32 core 64 RAM virtual machine.

panREPET pipeline

We developed the panREPET pipeline to identify shared TE insertions between genomes by comparing them pairwise. We divided the pipeline into the following steps:

Step 1: Extract TE copies according to coverage percentage in the alignment between the sequence TE copy and its reference

The coverage percentage between the TE copy and its reference sequence from the TE library is a parameter (covcons parameter), allowing extraction of either complete or partial TE copies **(Fig. 1b)**.

Step 2: Add flanking genomic sequence

In order to recognize copies that are located at the same genomic locus, we retrieve their flanking sequences by simple coordinate extension on both sides using *bedtools slop* from the Bedtools package[20] **(Fig. 1c)**.

Step 3: Bidirectional best-hit detection

Bidirectional best-hit detection first consists of aligning TE copies extended by their flanking sequences between accessions to find their best hits and vice versa **(Fig. 1d)**. Only couples of reciprocal best hits are retained. This approach was inspired by methods for orthologous gene detection[21]. Pairwise comparisons are performed by Minimap2[22], followed by Matcher[15], which merges nearby colinear alignment fragments.

Step 4: From bidirectional best hits to TE insertions

Pairs of TE copies identified in step 3 are used to build an undirected graph where the TE copies are nodes and where edges identify copies found in reciprocal best hits. In graph theory, a clique is a subset of nodes in a graph in which every node is connected to every other node in the clique. We consider a clique as a proxy of a unique TE insertion found at the same locus in several genomes. Cliques are found using NetworkX[23] version 2.4. The method *read edgelist* from NetworkX constructs the graphs that associate pairs of copies through bidirectional best hits. Then, the method *find_cliques* from NetworkX detects all maximal cliques in the constructed graphs. For each node n, a maximal clique for n is a largest complete subgraph containing n.

Finally, each TE copy of each accession is classified on core, dispensable or unique if they are found respectively in all individuals, some of them, or in only one **(Fig. 1e)**. Adding or deleting a genome will not require the pipeline to be restarted from the beginning, since pairwise genome comparisons are independent of clique building.

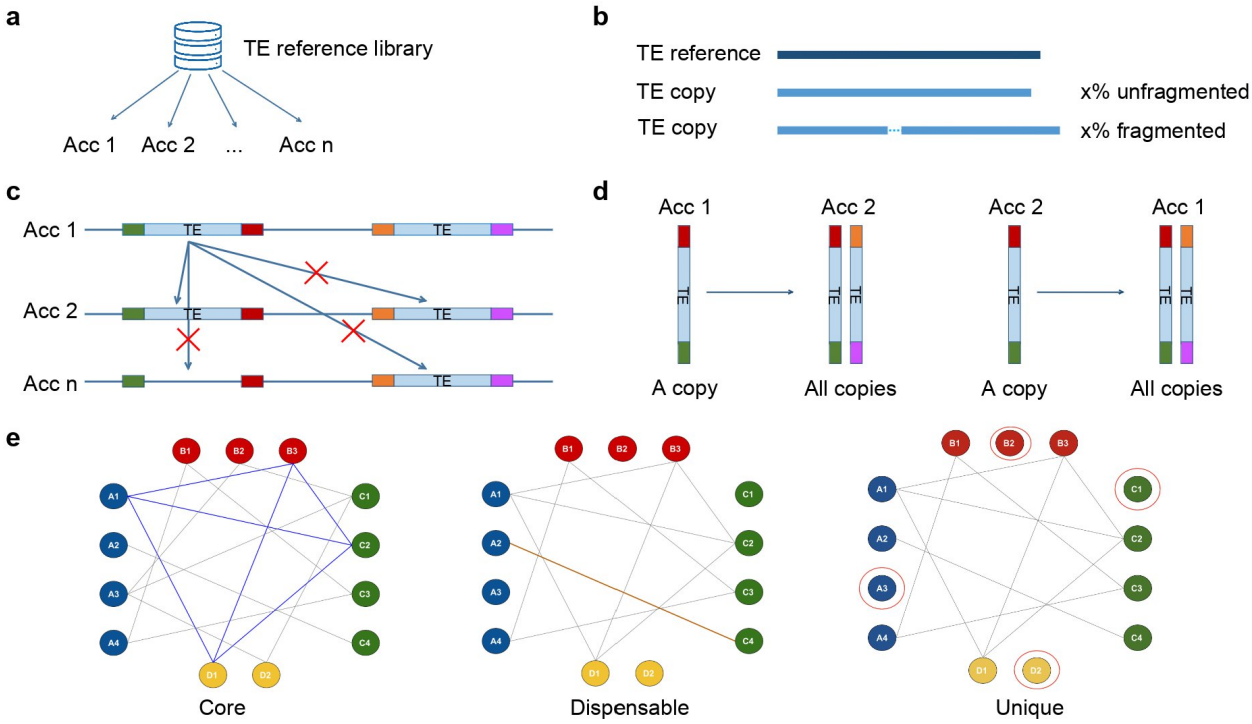

Figure 1: panREPET steps. **a** Independent TE genome annotation with panTEannot using a TE reference library. **b** Extraction of copies according to their coverage percentage in the alignment with their reference (e.g. *x* is between 95% and 105% for complete copies). **c** A TE copy from accession 1 is matched with TE copies having the same flanking regions in order to detect copies located at the same locus. **d** Bidirectional best-hit detection. **e** Clique extraction to identify TE insertion. Copy type identification: core, dispensable, or unique. A circle represents a TE copy, a color represents an accession, and a bold line represents a bidirectional best-hit detection. Acc: accession.

Retracing shared TE insertions

In order to understand how TEs have shaped the genome of the species, we identified shared TE insertions within *B. distachyon* genomes. The panREPET pipeline allowed us to detect core, dispensable and unique TE insertions among our reliable set of 42 accessions of *B. distachyon*. **Fig. 2a, 2b** represent a histogram of TE insertion number according to the number of accessions sharing the insertion. According to its shape **(Fig. 2a, 2b)**, we considered five pangenomic compartments: 1. Unique-TE insertions; 2. Cloud-TE insertions shared by 2-7 accessions (2–20%); 3. Shell-TE insertions shared by 8-38 accessions (20-95%); 4. Softcore-TE insertions shared by 39-41 accessions (95-98% of accessions) and 5. Core-TE insertions shared by all accessions. A unique-TE insertion suggests a new TE insertion in a single individual. A dispensable-TE insertion corresponds to an insert inherited from a common ancestor in some related individuals. Core-TE insertions are fixed inserts at the species level. panREPET identified 319,969 TE copies that cover their consensus over more than 95% (covcons = 95-105%). Among them, it detected 18,513 shared TE insertions including 595 core (1.2%); 1,139 soft-core (2.4%); 9,125 shell (19.4%) and 7,654 cloud (16.2%). 28,497 are unique TE insertions (60.6%). The vast majority of TE insertions appear to be unique **(Fig. 2a, 2b)**. The pipeline ran for 20 hours on a 64

core 128 Gb RAM virtual machine. By including partial copies (covcons = 75-125%), it extracted two times more copies (653,695) corresponding to 35,922 shared TE insertions and 82,347 unique insertions. We thus observed a higher proportion of core (1.4%) and unique TE insertions (69.6%) **(Fig. 2b)**. Longer TE copies tend to be unique or present in a small number of individuals **(Fig. 2c)**. The pipeline ran for 3.6 days on a 64 core 128 Gb RAM virtual machine.

We built a clustermap showing the TE copy presence or absence for each accession **(Fig. 3a, 3b)**. Gordon *et al.* estimated three genetic clusters based on >3 million SNPs[12]: (i) EDF+ with extremely delayed flowering phenotype, (ii) T+ from Eastern, predominantly Turkish, and (iii) S+ from Western, predominantly Spanish. We observed that relationships between accessions based on whole-genome SNPs is similar to those computed from shared TE insertions **(Fig. 3a, 3b)**. However, the TE insertion-based tree **(Fig. 3b)** fits even better with the gene-based tree from the Gordon *et al.* study[12] (tree not shown); we observed the same two distinct sub-clusters in the cluster T+. Bd29-1 is not clustered as expected and presents few TE insertions contrary to other accessions from its genetic cluster. We think that it does not cluster with its genetic cluster due its poor assembly quality (the lowest assembled genome size which is 253 Mbp and a low contig N50 of 13 kbp). In order to validate this hypothesis, we took a detailed look at the number of unique TE insertions according to assembly quality metrics **(SF 1, SF 2)**. We see that distribution is not uniform among accessions and does not seem to be explained by genetic clusters **(SF 1**). The number of unique insertions seems to decrease with assembly quality **(SF 2)**. Among the assembly quality metrics, the contig N50 best explains the number of unique TE insertions **(SF 2**) (linear regression with contig N50 as regressor and the number of unique TE insertions as explained variable $R^2 = 0.191$, *p*-value = 1.190e-03). We see that the best assembled genomes contain on average fewer unique TE copies than the others **(SF 1)** probably because their copies are well detected and therefore easier to put into a clique.

Unique, cloud and shell TE insertions concern mainly LTR families and the proportion of LTR families decreases as insertion becomes increasingly shared **(Fig. 2d)**. Stritt *et al.* showed that the western ancestral population represented by the S+ cluster lost retrotransposons because of several bottlenecks[24]. Supporting this hypothesis, we observe that the cluster S+ shows low intra-diversity contrary to the other genetic clusters **(Fig. 3a)**. For soft-core and core TE insertions, MITE, LINE and SINE proportions increase when insertion is more broadly shared among accessions **(Fig. 2d)**. A study on wheat has already shown high sequence conservation of MITEs[25]. Horvath *et al.* showed for 320 *B. distachyon* accessions, that retrotransposons are strongly correlated with the demographic structure whereas DNA-transposons are weakly correlated[26]. This corroborates our observations: core and soft-core TE insertions are mainly composed of DNA-transposons, particularly MITE families, while shell, cloud and unique insertions are mainly composed of LTR families and hence shape the intra-specific diversity **(Fig. 2d)**.

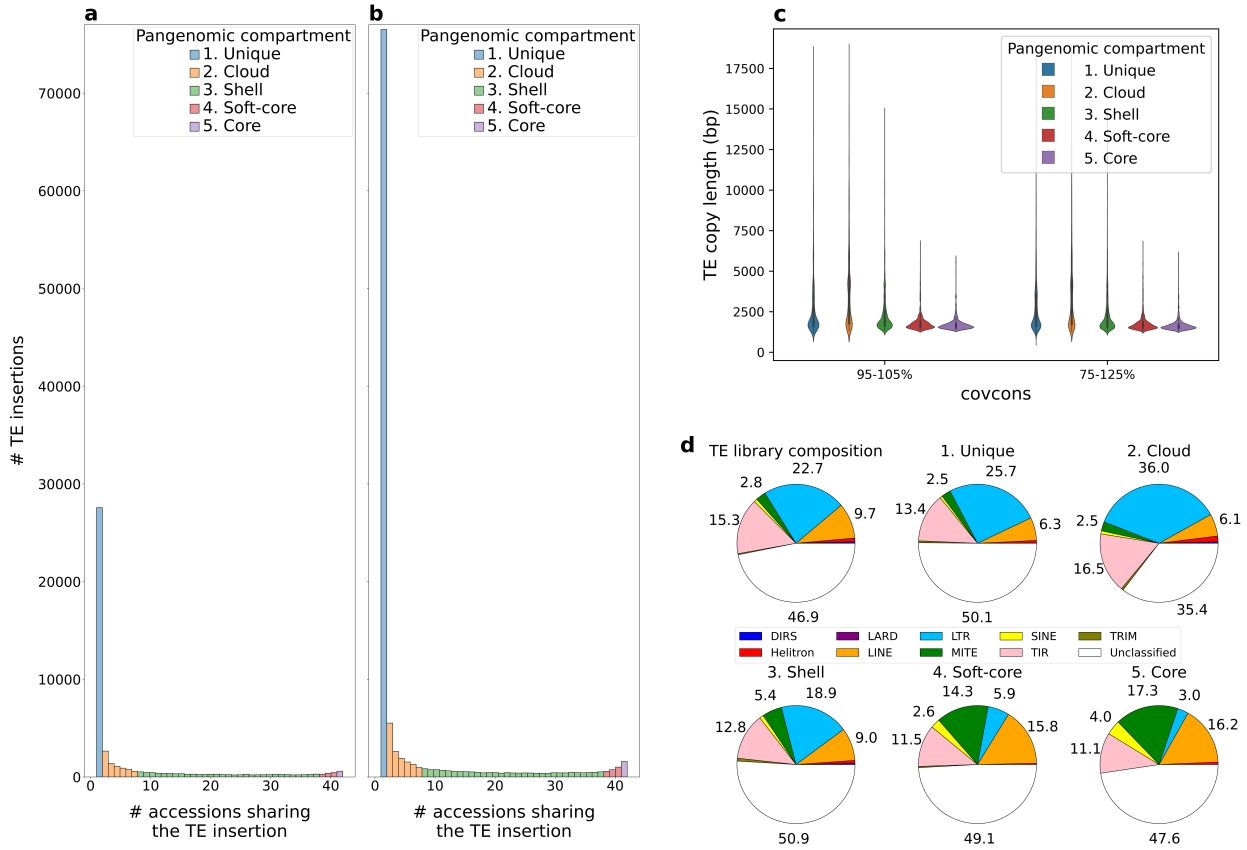

Figure 2: a, b Histograms of the number of TE insertions according to the number of accessions sharing each insertion, and the percent coverage between the TE copy and its consensus (parameter covcons). **a** covcons = 95-105% **b** covcons = 75-125% **c** Distribution of TE copy lengths according to their pangenomic compartment. **d** Composition of the TE library built *de novo* from Bd21 genome (1995 sequences) and proportions of TE insertion orders per pangenomic compartment. Only proportion labels greater than 2% are displayed.

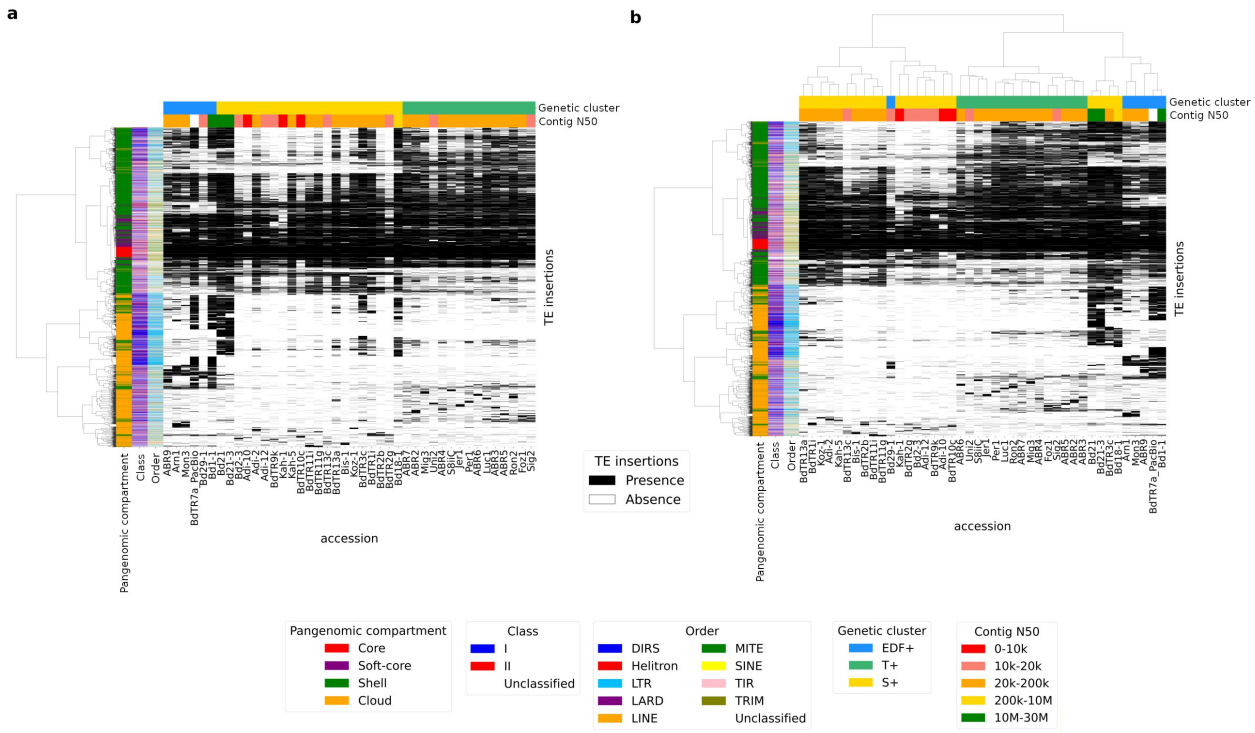

Figure 3: Clustermaps showing TE insertion presence/absence for each accession. **a** Accessions are vertically ordered according to a whole-genome SNP genetic tree (Gordon *et al.*, 2017, Supplementary Figure 4.a). **b** Accessions are vertically ordered by their shared TE insertion pattern, after clustering on presence/absence in accessions. The metric used is the Dice distance with Ward's algorithm.

Core TE insertions

Within the pangenome, about 30% of core TE insertions intersect at 80% a feature of a gene **(Table 2)**. They mainly intersect exons and introns **(Fig. 4a)**. The proportion of TE insertions intersecting a gene increases as the TE insertion is shared by more accessions **(Table 2)** suggesting a conservation of the TE by its host for functional reasons[1,27,28,29] or by fixation.

On the Bd21 genome, for all pangenomic TE compartments, the Transcription Factor Binding Site (TFBS) proportion intersecting a TE is about 4% **(Table 2)**. To test whether the predicted TFBS do not intersect the core TE sequences by chance, we shuffled the 17,589 predicted TFBS in promoter regions (-500 bp ~ +100 bp of transcription start site (TSS)) on the Bd21 genome thanks to bedtools shuffle[20] and counted the number of intersections against core TE insertions with 100% coverage on the TFBS. The null hypothesis is that there is a statistical association between core TE insertions and the shuffled TFBS. Repeating this procedure 100 times, we observed that shuffled TFBS do not intersect the core TE insertions (Chi-Square goodness of fit = 2113, *p*-value = 0.0, degrees of freedom = 99). Hence, we conclude that there is a statistically significant association between core TE insertions and TFBS in promoter regions, suggesting that these core TEs result from co-optation of a TE-TFBS[19]. Concerning the 144,761 genome-wide predicted TFBS (i.e., including TFBS not in known promoter regions), there is also no statistical association between core TE insertion and shuffled TFBS (Chi-Square goodness of fit = 179, *p*-value = 1.25e-05, degrees of freedom = 99). In the following, we will work on the TFBS predicted in the

promoters. We observe that core TEs with TFBS are mostly MITEs **(Fig. 4c)**. A study on maize has already shown domesticated MITEs carrying TFBS are involved in husk tissue-specificity[30].

Concerning the remaining core devoid of genes or TFBS (393/595 core TE insertions) **(Table 2)**, we wondered about their conservation at the species level. The hypotheses that we can test are:

TEs may be domesticated thanks to their role as essential structural components for centromeres[31,32] and for maintaining centromeric and telomeric stability and heterochromatic silencing[33]. To determine whether the remaining core TE insertions devoid of genes or TFBS have been conserved thanks to their genomic position, we plotted their positions on chromosomes **(SF 4)** and determined their distance from the centromere**,** the recombination rates of the region of insertion (cM/Mb), the TE and gene density, and their distance to the closest gene **(Fig. 4e-i).** We performed an ANOVA to compare these distributions between core TE insertions devoid of genes or TFBS and the other core TE insertions. We observe that only one core TE insertion devoid of genes or TFBS is located at the centromere, and there is a slight tendency for core TE insertions devoid of genes or TFBS to be closer to the centromere than for the other core TE insertions **(Fig. 4e, ANOVA *p*-value=0.0243)**.

TE insertions are also more likely to settle in non-recombining regions by fixation[34] as they may have been fixed by genetic drift[35,36]. Core TE insertions devoid of genes or TFBS are not more frequent in low-recombination regions than the other core TEs **(Fig. 4f, ANOVA *p*-value=0.4634)**.

Heterochromatic regions are also less subject to recombination[37], and TEs tend to accumulate there[38,39,29]. To infer the chromatin state of the genomic regions, we computed respectively TE and gene density per 50kbp intervals among chromosomes **(Figs. 4g-h)**. We observe that core TE insertions devoid of genes or TFBS are preferentially located in heterochromatin regions **(Figs. 4g-h, ANOVA *p*-values=7.673e-03 and 6.019e-09 for TE density and gene density respectively)**. In agreement, core TE insertions devoid of genes or TFBS are further from the closest gene than the other core TEs **(Fig. 4i, ANOVA *p*-value=4.840e-08)**.

Regions under positive selection and low recombination rates do not allow efficient elimination of transposable elements (the Hill-Robertson effect)[39,40]. Consequently, we observe on **Fig. 4i** that unique and cloud TE insertions devoid of genes or TFBS present a greater distance to the closest gene compared to the core and soft-core TE insertions devoid of genes or TFBS. We can then hypothesise that core TE insertions devoid of genes or TFBS are conserved by the effect of a nearby gene under positive selection.

Pangenomic compartment	Genome(s)	Genome -wide	Unique	Cloud	Shell	Soft-core	Core
Number of TE insertions in genome	Bd21	131,695	264	2,148	5,835	1,078	595
Proportion of TE insertions intersecting a gene	Bd21	3.6% (4,844)	24% (65)	25% (537)	26% (1,557)	31% (335)	30% (179)
Proportion of TE insertions intersecting a gene (median)	Pangenome (42 genomes)	3.0%	22%	25%	26%	33%	32%
Proportion of TE insertions intersecting a TFBS In promoter regions (-500 bp ~ +100 bp of TSS)	Bd21	0.63% (842)	4% (12)	4% (107)	4% (274)	5% (55)	4% (26)
Number of TE insertions intersecting a TFBS Genome-wide	Bd21	9.4% (12,481)	56% (150)	54% (1,175)	43% (2,563)	35% (387)	32% (195)

Table 2: Number of TE insertions that intersect genes (covering at least 80% of a gene feature) or TFBS (covering 100% of the TFBS) in terms of pangenomic compartments. TFBS: Transcription Factor Binding Site. TSS: Transcription Start Site.

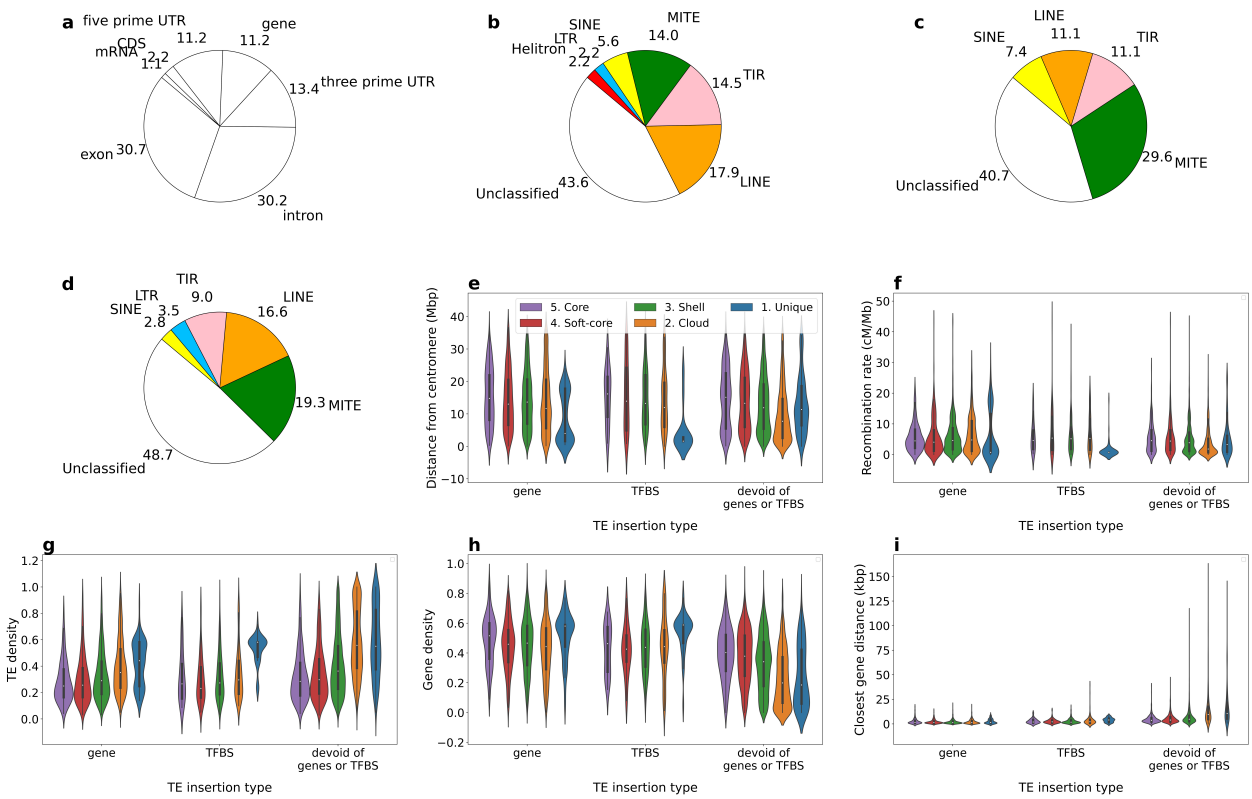

Figure 4: Analysis of core TE insertions on the Bd21 genome. **a** Gene feature that the core TE pangenomic compartment intersects by at least 80%. Only the gene feature most covered by the TE is selected. **b-d** Order of TE family proportions that the core pangenomic compartment intersects. **b** Core intersecting a gene. **c** Core intersecting a TFBS. **d** Core devoid of genes or TFBS. **e-i** Parameter distributions according to core TE insertions that intersect a gene, those that intersect a TFBS and those devoid of gene or TFBS. **e** Distance distribution from the centromere (Mbp). **f** Recombination rate (cM/Mb). **g** TE density. **h** Gene density. **i** Closest gene distance (kbp).

TE insertion dates

Age of TE insertion events can highlight their evolutionary dynamics over time. Combining genetic distances between accessions with our analysis of TEs in the pangenome allowed us to date TE insertions more accurately than current methods since we take into consideration divergence between accessions. In addition, we consider all types of TE copies and not only the LTR copies contrary to others methods[41,42,43,44]. Some other methods date TE copies using percent identity with the consensus as a proxy of age. The more the copy diverges from the consensus, the older it would be[45,46]. For all TE copies detected among accessions, we represented the distribution of TE copy-consensus identity percentage, according to the number of accessions sharing the TE insertion **(SF 3)**. As expected, core TE insertions **(SF 3)** show low identities with their consensus sequence (<65%), suggesting ancient and conserved copies.

This approach is however limited as it depends on consensus with the reference TE library. We propose a new approach for dating TE insertions based on whole-genome SNPs that takes into account species-wide evolution. TE insertions shared between accessions show that the insertions preceded the split of accession lineages from their common ancestor. TE insertion age can be then inferred as the longest whole-genome SNP distance between each pair of accessions sharing the TE insertion. We observe that the more an insertion is shared, the greater the genetic distance

based on SNPs **(Fig. 5a)**. The observed plateau is the maximum SNP distance (0.335 subs/site) among accessions, corresponding to distance between the two most distant accessions Arn1 and Ron2 (see Gordon *et al.*, 2017, Supplementary Figure 4.a). Minadakis *et al.* estimated the genetic cluster divergences based on a multispecies coalescent approach[47]: S+, T+ and EDF+ genetic clusters split at least 45 kya ago. S+ and T+ genetic clusters split 23 kya ago. For converting substitution per site into kya, we assumed that the maximum genetic distance based on whole-genome SNPs among S+ and T+ genetic clusters (0.1549 subs/site, between Bd21 and Sig2) corresponds to 23 kya.

Fig. 5a illustrates four types of TE insertion events: **(i) The ancient and not conserved TE insertions:** The insertion is shared by few to many accessions (2-36 accessions) and at least two accessions are genetically distant on the whole-genome-SNP genetic tree (i.e., with a genetic distance greater than 0.20 subs/site corresponding to 30 kya). However, the accessions that most share this type of insertion are Bd1-1, Bd18-1, Bd21, Bd21-3, BdTR3c and BdTR7a. As a reminder, they have a high quality assembly **(SF 1, Fig. 3)**. It is more likely that they share insertions because the corresponding repeated regions are better assembled than in other genomes. This type of TE insertion is numerous and Horvath *et al.* showed that TEs in *B. distachyon* are dominated by purifying selection[26], which corroborates our observation. **(ii) The ancient and conserved TE insertions**: They are shared by almost all accessions (37-42 accessions) suggesting a pre-speciation insertion. **(iii) The recent TE insertions:** They are shared by only few accessions (2-6 accessions) which are close on the SNP genetic tree (genetic distance smaller than 0.20 subs/site or 30 kya). In this case, the common ancestor underwent insertion recently. They mainly concern the couple Arn1 and Mon3 from Spain. They specifically share a mixed ancestry[47], that explains why they share so many specific TE insertions. **(iv) The TE insertions that occurred after the divergence into intra-specific genetic clusters:** The insertion is only shared by accessions from an intra-specific clade.

Hence, we dated four major transposition events in *B. distachyon*. A first major transposition event, at least 45 kya, that could correspond to pre-speciation TE insertions or TE insertions occuring during the split into the three lineages. A second major transposition event at 35 kya that involves TE mostly shared by EDF+ and T+ **(Fig. 5b)**. That could suggest that the S+ genetic cluster underwent genetic drift around 35 kya with the hypothesis that the few S+ individuals surviving the bottleneck have their copy eliminated by genetic drift or selection. A third major transposition event at 22 kya during the Last Glacial Maximum (LGM), a period conducive to thermal stress. Note in **Fig. 5a** that we observe a gap of TE insertions around 30 kya that may be explained by the abrupt decline in population sizes undergone by the species, followed by population expansion in the very recent past[24,47]. These population expansions may have triggered TE transposition around 22 kya. Finally around 10 kya, there is a burst of transposition occuring during the Holocene (11.7-7.5 kya) which is the beginning of deglaciation in Europe and the relocation of species to new environments[47]. Among the recent transpositions, we observed a burst of transposition in S+ at 11.7 kya (**Fig. 5b**). S+ may have experimented bursts of transposition during its recolonization of Europe and the Middle East.

Identifying active families that are present at very high copy number allows us to understand the impact of TEs on plant genome evolution[48]. A way to highlight families that have transposed during these major transposition events, is to represent the abundance of TE families according to their insertion type (**Fig. 5c**). We see that the oldest conserved transpositions, that occured at least

45 kya, involve in majority MITE, LINE, TIR and SINE families. LTR families are however either recent or ancient and not conserved, suggesting turnover: the two LTR families (on the left in **Fig. 5c**) correspond to Gypsy. Stritt *et al.* suggested that recent TE transpositions involve centromeric Gypsy superfamily and Copia superfamily[44].

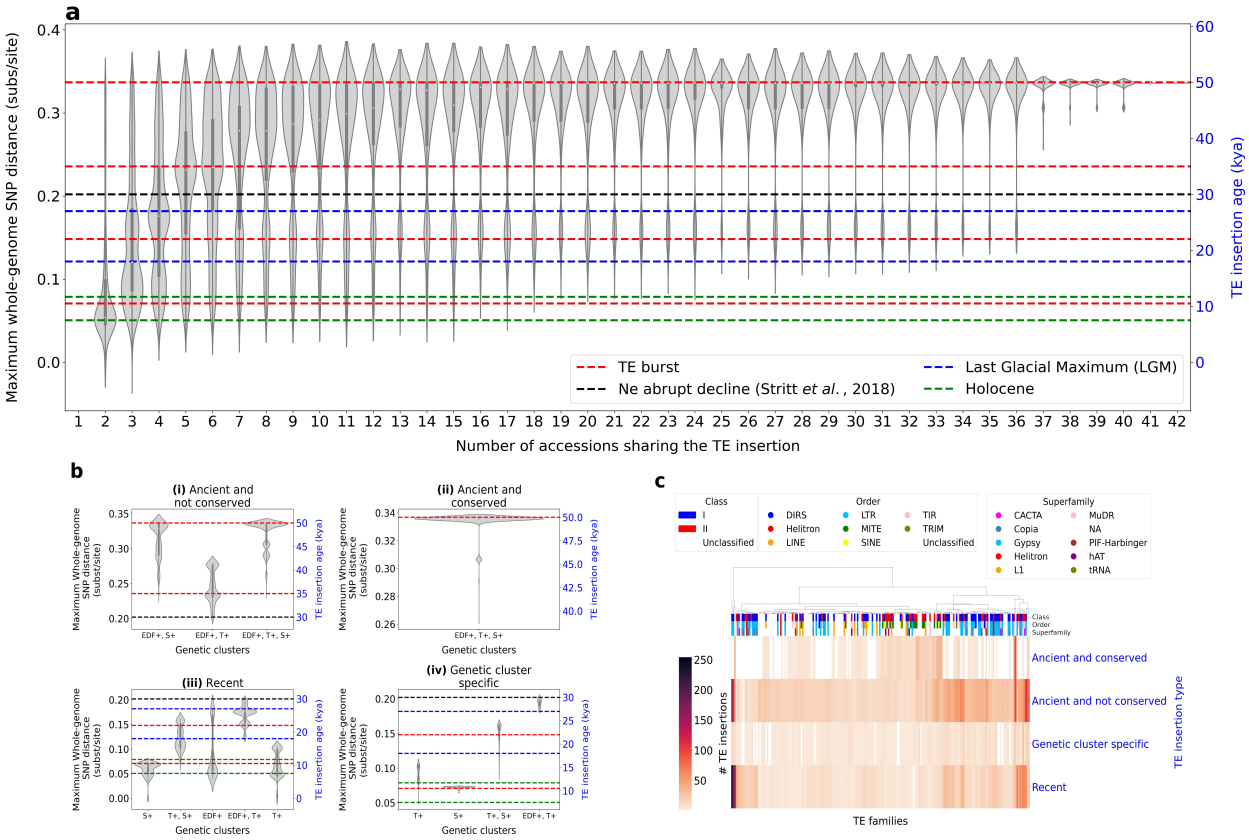

Figure 5: a Boxplots showing the age distribution of TE insertions according to the number of accessions sharing them. The age is estimated by the maximum whole-genome SNP distance between pairwise accessions within the TE clique. **b** Boxplots showing the age distribution of TE insertions according to genetic clusters to which the accessions sharing the insertion belong. We split Fig 5a. into four TE insertion types: (i) the ancient and not conserved TE insertions; (ii) The ancient and conserved TE insertions; (iii) The recent TE insertions and (iv) the TE insertions that occurred after the cluster divergence. **c** Clustermap showing the abundance of TE families according to the TE insertion types. The metric used is the Euclidean distance with ward algorithm. We selected families having more than 20 copies for readability.

Factors affecting the evolutionary dynamics of TE families

We investigated factors that may explain TE mobilization by using geo-climatic data for accessions. We are interested here in local factors which could explain recent TE transposition. In order to highlight the differential dynamics of recent TE families between accessions, we clustered the accessions according to the number of recent TE family copies **(Fig. 6)**. As a first proxy, recent TE insertion can be considered unique, i.e. occurring within only one accession. We distinguished unique TE insertions with an identity TE copy-consensus greater than 95% (95% is a cut-off already used to highlight young subfamiles[49] **(Figs. 6a, 6a')** and the other ones **(Figs. 6b, 6b').** Another way to identify recent transposition events is by selecting only TE insertions occurring after the Holocene (i.e., age < 7.5 kya). Indeed, we chose this threshold in particular because *B.*

distachyon moved southward during the last glacial period and recolonized Europe and the Middle East within the last five thousand years (Niche modeling analysis performed by Minadakis *et al.*, 2023[47]) **(Figs. 6c, 6c')**.

Similar patterns of TE family dynamics among a set of accessions may suggest common determinants for TE activity (e.g., environmental stress). To determine those common determinants, we added for each accession: its genetic cluster, the country, the climate, the elevation in meters, the flowering class, the temperature in 2016 (on the date of collection of the accession), the precipitation in 2016 and the wind in 2016 **(Figs. 6a-c)**. Unique TE family dynamics does not seem to be explained by genetic clusters but by other factors **(Figs. 6a-b)**. Accessions presenting the best contig N50 are clustered (the couples Bd1-1/BdTR7a from Turkey and Bd21/Bd21-3 from Iraq) **(Fig. 6a)**. They present fewer unique TE insertions than the other accessions **(SF 1)**, yet the same families cluster together, which is not the case in the other accessions. This may suggest that accessions with low assembly quality show many single copies which are actually noise. The couples (Bd21, Bd21-3) and (Bd1-1, BdTR7a) specifically share Copia and Gypsy LTR and MuDR TIR superfamilies **(Fig. 6a)**.

We also represented the TE family dynamics directly according to climate class for TE insertions occurring after the Holocene (age < 7.5 kya) **(Fig. 6c')** to highlight transpositions induced by abiotic conditions. We plotted the average number of TE copies per accession to avoid biases from non-homogeneous distribution of accessions in the different climate classes (Arid Bwh: 3 accessions, Humid subtropical Cfa: 3, Marine Cfb: 14, Hot-summer mediterraneen Csa: 13, Warm-summer mediterraneen Csb: 9). We observe that accessions in arid climate have activated specific TE families, and that accessions in hot-summer and warm-summer mediterranean climates have more similar TE family dynamics than accessions in humid subtropical and marine climates **(Fig. 6c')**.

Then we looked for TFBS in TEs potentially explaining a local adaptation. We looked at the intersection of the 144,761 predicted genome-wide TFBS with the supernumerary TE copies (i.e., TE families with more than 10 copies or 2 copies, see **Table 3**) and we only kept TEs intersecting by 100% a TFBS sequence **(Table 3)**. We only worked on Bd21 because the analysis required available TFBS data. Among the supernumerary recent TE insertions intersecting a TFBS, Bd21 presents in majority LTR families **(Table 3)**. We observed TFBS targeted by OsMADS56, OsMADS4, OsMADS8 and OsMADS3 aMADS-box gene family with an MIKCc type-box. Studies on *Oryza sativa* showed that they are critical for the reproductive organ[50]. OsMADS8 is required for tapetum development[51]. Its degradation by certain ambient air pollutants can lead to pollen sterility[52] in the petunia. In addition, the myb-like DNA-binding domain such as for instance in AtMYB002, is involved in drought response in *Arabidopsis thaliana*[53].

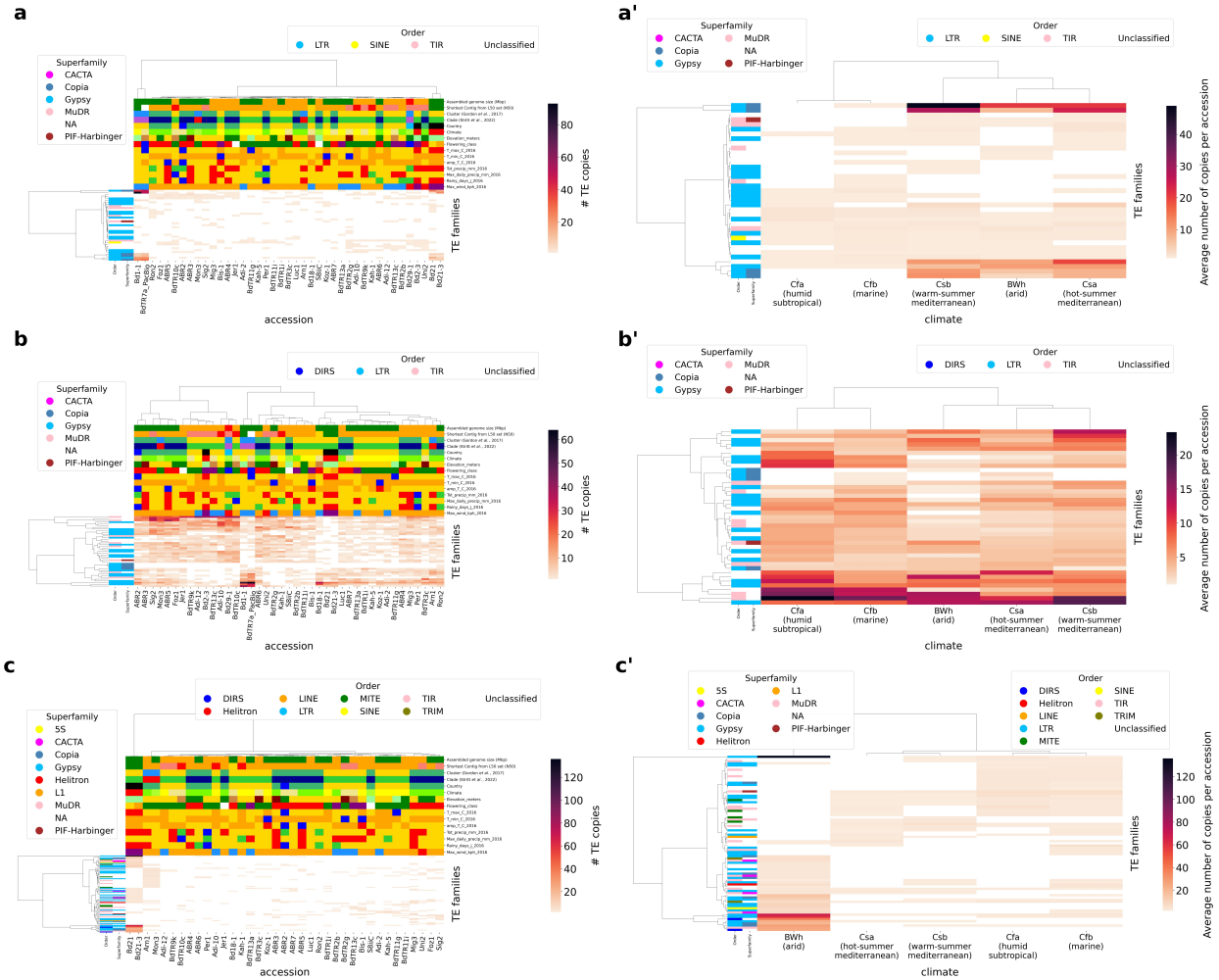

Figure 6: a-c Clustermaps showing the abundance of recent TE insertions per TE families and per accession. On the x-axis and on the y-axis, accessions and TE families are respectively clustered according to their similarity in their pattern of TE family dynamics. Color labels of columns are detailed in **Supplementary Table 2**. We used the Euclidean distance with the ward linkage method. The percent coverage between TE copy and its consensus is 95-105%. **a-b** Unique TE copies. Only TE families with at least 10 insertions for at least one individual have been selected. **a** TE copies with an identity greater than 95% with their consensus. **b** TE copies with an identity lower than 95% with their consensus. **c** TE copies that transposed after the Holocene (age < 7.5 kya). **a'-c'** Each clustermap corresponds to its counterpart, the only differences are that the climate is on the x-axis and the clustermap shows the average number of TE copies per accession in order to weight the unequal distribution of accessions in the different climate classes (Arid BWh: 3 accessions, Humid subtropical Cfa: 3, Marine Cfb: 14, Hot-summer mediterraneen Csa: 13, Warm-summer mediterraneen Csb: 9).

Figures	6a, 6a’	6b, 6b’	6c, 6c’
Recent TE insertion criteria	Unique TE insertions Copy-consensus percent indentity > 95%	Unique TE insertions Copy-consensus percent indentity <= 95%	After the Holocene (age < 7.5 kya)
Supernumeray TE families criteria	TE families with more than 10 copies	TE families with more than 10 copies	TE families with more than 2 copies
TE families intersecting a TFBS	MCL2 LTR Gypsy: 12/14 copies MCL14 LTR Copia Angela: 21/26 copies MCL42 LTR Copia: 5/6 copies	MCL2 LTR Gypsy: 18/21 copies MCL4 LTR Gypsy: 10/12 copies	MCL6 LTR Gypsy: 7/68 copies MCL2 LTR Gypsy: 29/34 copies MCL4 LTR Gypsy: 26/30 copies MCL14 LTR Copia: 11/14 copies MCL3 TIR MuDR: 13/13 copies MCL 68 LTR Copia: 10/10 copies
The main Gene Ontology (GO) inferred from rice	MCL2: ('AP2 domain containing protein, expressed', 7), ('TCP-domain protein, putative, expressed', 5), ('expressed protein', 5), ('OsMADS56 - MADS-box family gene with MIKCc type-box, expressed', 4), ('OsMADS4 - MADS-box family gene with MIKCc type-box, expressed', 4), ('tesmin/TSO1-like CXC domain containing protein, expressed', 3), ('TCP family transcription factor, putative,	MCL2: ('AP2 domain containing protein, expressed', 7), ('TCP-domain protein, putative, expressed', 5), ('OsMADS56 - MADS-box family gene with MIKCc type-box, expressed', 4), ('OsMADS4 - MADS-box family gene with MIKCc type-box, expressed', 4), ('tesmin/TSO1-like CXC domain containing protein, expressed', 3), ('TCP family transcription factor, putative, expressed', 3), ('MYB	MCL6: ('AP2 domain containing protein, expressed', 4) MCL2: ('AP2 domain containing protein, expressed', 14), ('DUF260 domain containing protein, putative, expressed', 7), ('OsMADS56 - MADS-box family gene with MIKCc type-box, expressed', 7), ('TCP family transcription factor, putative, expressed', 6), ('MYB family

	expressed', 3), ('MYB family transcription factor, putative, expressed', 3) MCL14: ('dof zinc finger domain containing protein, putative, expressed', 15), ('OsMADS8 - MADS-box family gene with MIKCc type-box, expressed', 7), ('transcription factor, putative, expressed', 6), ('GATA zinc finger domain containing protein, expressed', 4) MCL42: ('Myb-like DNA-binding domain containing protein, putative, expressed', 3), ('AP2 domain containing protein, expressed', 1), ('OsMADS3 - MADS-box family gene with MIKCc type-box, expressed', 1)	family transcription factor, putative, expressed', 3) MCL4: ('AP2 domain containing protein, expressed', 8), ('TCP family transcription factor, putative, expressed', 3), ('DUF260 domain containing protein, putative, expressed', 2), ('TCP-domain protein, putative, expressed', 2), ('MYB family transcription factor, putative, expressed', 2), ('dof zinc finger domain containing protein, putative, expressed', 1), ('OsMADS56 - MADS-box family gene with MIKCc type-box, expressed', 1)	transcription factor, putative, expressed', 4), ('tesmin/TSO1-like CXC domain containing protein, expressed', 4), ('dof zinc finger domain containing protein, putative, expressed', 3) MCL4: ('AP2 domain containing protein, expressed', 17), ('OsMADS56 - MADS-box family gene with MIKCc type-box, expressed', 4), ('TCP family transcription factor, putative, expressed', 4), ('DUF260 domain containing protein, putative, expressed', 3), ('Homeobox domain containing protein, expressed', 3) MCL14: ('dof zinc finger domain containing protein, putative, expressed', 7), ('OsMADS8 - MADS-box family gene with MIKCc type-box, expressed', 6) MCL3: ('ZOS2-02 - C2H2 zinc finger protein, expressed', 7), ('AP2 domain containing protein, expressed', 6), ('Homeobox domain

			containing protein, expressed', 6), ('no apical meristem protein, putative, expressed', 3) MCL68: ('helix-loop-helix DNA-binding domain containing protein, expressed', 10)

Table 3: Search on Bd21 genome for genome-wide predicted TFBS intersecting recent and over-represented TE insertions (covering 100% of the TFBS).

Discussion

panREPET, a new method to improve TE analysis

Several pipelines exist for detecting TEs in pangenomes. They generally analyse TEs according to a reference genome. Two approaches have tried to find neo-TE-insertions in resequenced genomes: (i) paired-end mapping (PEM) detecting discordant pairs[54,55]. However fragment insert size in paired-end sequencing is about ~400-500 bp, which is shorter than a TE[56] and (ii) the split-read method detecting truncated reads corresponding to the junction between the TE insertion and the reference sequence. Unlike the strategy of discordant reads, truncated reads can determine the TE insertion at the base level. However, the coverage is low, and this method is very sensitive to read size[10,57]. In short, this cannot distinguish between intact and truncated copies between accessions at a given locus. Quadrana *et al.* added to the split-reads approach a validation based on TSD (target site duplication) generated by most TE families upon insertion[9]. But this method cannot be used in TEs which are devoid of TSD such as Helitrons. panREPET aims to analyse precisely all types of TEs. The TEMP pipeline[58,24] combines PEM and split-read methods. In a previous study of 53 *B. distachyon* accessions, Stritt *et al*., using TEMP, identified 3,855 shared TE insertions of which 2,007 (52.0%) are present in at least one of the 53 accessions but absent from the reference genome (TE insertion polymorphisms or TIPs) and 1,848 (47.9%) shared TE insertions present in the reference genome but absent in at least one accession (TE absence polymorphisms or TAPs)[24]. In our study, using the same data on 42 accessions, panREPET found 18,513 shared TE insertions that had a coverage between the copy and reference of more than 95% and 7,514 TE insertions (40%) are absent from the reference Bd21 (i.e., the equivalent of a TIP). With panREPET, we detected many more shared TE insertions (4.8 times more) thanks to the pairwise comparison approach, illustrating the bias when depending only on a reference genome. However, we observed a smaller proportion absent from the reference (TIPs) with panREPET, because the best assembled genomes are easier to put into a clique **(SF 1)**. Note that differences in assembly quality in a pangenome may lead to underestimation of the number of accessions sharing the TE insertion.

Short reads validate position but do not validate the exact sequence composition. A TE copy is generally about 300 bp to 10 kbp long, which is often larger than the average short-read length of about 100-300 bp[56]. However, for TE regulation analysis, we need to know the exact activator and repressor sequence composition on the TE sequence[59]. Some methods as x-Transposable element analyzer (xTea) overcome the short read limitation by including long reads[60]. An additional problem is that read alignments for sequences of genomes to an assembled reference genome do not access the genetic environment of TE insertions in the genome from which the reads come. xTea tool tried to overcome this limit by using the Linked-Read technology from 10X Genomics. This utilizes microfluidics to partition and barcode DNA fragments from the same region, so that long-range information is embedded in the short-read data[60]. Plus, thanks to long-read data, xTea utilizes the fully reconstructed TE sequences to provide additional information that cannot be obtained from short-read data[60]. However, our approach is straightforward and does not involve any complex sequencing strategy as it works on several *de novo* assembled genomes. Indeed, panREPET provides direct access to the sequence of the whole TE copy, its exact insertion length, and its exact genetic environment. For instance, in our core TE insertion analyses, we were able to intersect TEs against genes and calculate the exact TE-gene coverage.

GraffiTE[61] uses several assembled genomes from the same species. It aligns each assembled genome against the reference but without aligning the others pairwise. It first detects structural variants (SVs) using genome-graphs and then annotates them in TEs. We have chosen a different paradigm: annotate the TEs in each genome independently and then compare TE copies pairwise between accessions. panREPET extracts copies that may have undergone an insertion or deletion relative to their consensus. By comparing panREPET with Minigraph[62] **(see Supplementary Results - panREPET compared with Minigraph)**, we observed that about 50-60% of TE copies detected by panREPET are retrieved by Minigraph. The remaining undetected TE copies highlight the limitation of detecting structural variants before annotating the TE copies. We see that this kind of approach can lead to the detection of partial TE copies. Indeed, a graph approach is more to detect an insertion or deletion within the TE copy as a variant, but not the entire copy **(SF 6)**. Annotating TEs first is the guarantee of detecting the complete sequence.

New biological results

A previous consensus library for *B. distachyon* was available from the TREP database and was composed of 233 sequences[63]. This library contained 74% TIR, 20% LTR (10% Copia, 10% Gypsy) and 6% Helitron. Our new consensus library is more complete with 1995 sequences including new TE families (LINE, SINE, MITE and TRIM) **(Fig. 2d)**. The TE library was built denovo from the Bd21 reference **(see Materials and methods - Genome analysis tools)**, but in the future a TE library representative of the pangenome should be considered. MITEs and SINEs are the main component of core and soft-core pangenomic compartments **(Fig. 2d)**, hence our TE library allows us to study ancient and conserved TE insertions. MITEs with TFBS are conserved at the species level. As concerns core TE insertions conserved in the species but devoid of genes or TFBS, we observed that they are preferentially located close to centromeres or in a heterochromatin region, or are conserved by the effect of a nearby gene under positive selection.

In *B. distachyon*, panREPET detected a majority of unique TE insertions **(Fig. 2a-b)**. That is expected as we extracted full TE copies (or almost, **see Results - panTEannot**) and the number

of unique TE insertions may depend on assembly quality **(SF 2)**. Interestingly, Stritt *et al.* suggested that *B. distachyon* expands populations without inter-accession breeding[24,64]; this phenomenon may lead to an excess of singletons[6].

Our approach aims to understand how individual TE lineages have evolved at the species level. A dispensable TE insertion could originate by two scenarios. First, the TE insertion could be inherited only by closely related accessions as it is recent. Second, the TE insertion could be lost by some accessions over time through excision, ectopic recombination, genetic drift or purifying selection. To distinguish these possibilities, we can estimate TE insertion age. Usually, a molecular clock of 1.3×10^{-8} substitutions/site/year is used on LTR-retrotransposons[41]. However, it does not take into account the evolutionary history of the species and it does not consider all types of TEs. Stritt *et al.* dated TE insertions among these 54 genomes of *B. distachyon* but only for LTR copies[44]. On the contrary, panREPET aims to date TE insertions for all types of TE copies thanks to the identification of shared insertions between individuals. In our study, a genetic tree based on whole-genome SNPs was available[12] as well as genetic cluster dating based on a coalescence method[47] **(see Results - TE insertion Dates)**. Hence, by fetching pairwise genetic distances among accessions, we were able to date TE insertions by considering genome-wide evolution and the specific events undergone by the species such as bottlenecks.

We dated, thanks to panREPET, the TE insertions before and after the Ice Age. We detected four ancient bursts of transpositions in *B. distachyon*: (i) at least 45 kya before or during the split into three lineages. However, we did not include the Italian genomes as they are unavailable in databases. Their addition would enable even more precise dating, as it would be the ancestral clade [64,47]; (ii) 35 kya which could correspond to the bottleneck undergone by the S+ genetic cluster; (iii) 22 kya during the Last Glacial Maximum (LGM) and occurring after the abrupt population effective size decline around 30 kya; (iv) and during the Holocene around 10 kya with the recolonization of Europe and the Middle East.

After the Holocene, the accessions Bd21 and Bd21-3 from arid climate seem to have activated specific TEs potentially induced by abiotic stresses like pollution or drought **(Fig. 6c')**. As the TFBS are only available for Bd21, we lack information on TFBS in other accessions that are not in an arid climate. Moreover, a homogeneous distribution of accessions by climate class would enable more precise inference of local adaptation, although in the end functional validations of TEs likely to be involved in adaptation will be required. panREPET aims to facilitate this type of automated query for a large dataset.

Conclusion

We developed and have presented here a new bioinformatic tool called panREPET that makes it possible to retrace and date TE insertions in a species in a more accurate way than current approaches because: (i) it provides genomic contexts and exact sequences of TE copies without depending on a reference genome (ii) it dates all types of TE copies (most current methods just allow dating for LTR families) (iii) the dating is based on shared TE insertions and hence on genetic distances between accessions and, in this study, the age given used a coalescence method which takes into account the specific events to the species such as bottlenecks. Finally, we dated

two major TE bursts corresponding to major key climate events and found specific TE families that seem to be activated in arid climate induced by pollution or drought after the Holocene.

Materials and methods

Brachypodium distachyon data

We downloaded fifty-four whole-genome *de novo* assemblies *of B. distachyon*, their gene annotations and their gene ontology from the Phytozome database (https://phytozome.jgi.doe.gov/)[12]. The genome versions are detailed in **Supplementary Table 1**.

We downloaded 17,589 predicted TFBS in promoter regions (-500 bp ~ +100 bp of transcription start site) and 144,761 predicted TFBS genome wide from the PlantRegMap database which uses the FunTFBS algorithm (https://plantregmap.gao-lab.org/download.php)[65]. Data are only available for the reference Bd21.

We fetched centromeric coordinates on Bd21 v3 from Li *et al.*, 2018, on Table 1[66]. The recombination rates are from Huo *et al.*, 2011[67], after mapping data to the version 3 reference genome assembly. We linked to the version 3 by searching for the sequences flanking the polymorphic SNPs in the genome version 3 to get the coordinates.

We extracted whole-genome SNP distances from phylogenetic tree in Newick format from Gordon et *al.*, 2017, Supplementary Figure 4.a[12]. We converted newick format to pairwise distances between accessions, using the Phylo module from Biopython v3.10.13[68].

We used the kgc R package v1.0.0.2[69] to associate each *B. distachyon* accession location with a general climate by converting the geographic coordinates into decimal degrees. We also extracted temperature, precipitation and wind from 2016, i.e. at the date of accession collection.

Genome analysis tools

We built a *de novo* TE library from the reference Bd21 v3.2 with the TEdenovo pipeline[13] from REPET v3.0 (https://urgi.versailles.inra.fr/Tools/REPET). The TE library is composed of consensus sequences which are reference sequences. We curated the annotation automatically by a second TEannot process[70] reducing the number of consensus sequences from 4475 to 1995.

We annotated the fifty-four genomes of *B. distachyon* with the TEannot pipeline from REPET v3.0 without mreps[14]. The TE classification comes from the classification of their consensus performed by the PASTEC tool from REPET v3.0[71].

The Bedtools version is 2.30.0. *bedtools intersect* is configured with option –wo writing the original A and B entries plus the number of base pairs of overlap between the two features[20].

We built clustermaps with Seaborn package from Python v3.10.13[72].

We performed ANOVA with statsmodels module from Python v3.10.13[73].

Bibliography

1.Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49-61. doi:10.1038/nrg3374

2.Bourque G, Burns KH, Gehring M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19(1):199. doi:10.1186/s13059-018-1577-z

3.Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2017;1860(1):157-165. doi:10.1016/j.bbagrm.2016.05.010

4.Langmüller AM, Nolte V, Dolezal M, Schlötterer C. The genomic distribution of transposable elements is driven by spatially variable purifying selection. Nucleic Acids Research. 2023;51(17):9203-9213. doi:10.1093/nar/gkad635

5.Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Molecular Ecology. 2019;28(6):1537-1549. doi:10.1111/mec.14794

6.Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes. 2019;10(6):419. doi:10.3390/genes10060419

7.Le TN, Schumann U, Smith NA, et al. DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. Genome Biol. 2014;15(9):458. doi:10.1186/s13059-014-0458-3

8.Makarevitch I, Waters AJ, West PT, et al. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLOS Genetics. 2015;11(1):e1004915. doi:10.1371/journal.pgen.1004915

9.Quadrana L, Bortolini Silveira A, Mayhew GF, et al. The Arabidopsis thaliana mobilome and its impact at the species level. eLife. 2016;5:e15716. doi:10.7554/eLife.15716

10.Carpentier MC, Manfroi E, Wei FJ, et al. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nat Commun. 2019;10(1):24. doi:10.1038/s41467-018-07974-5

11.Hasterok R, Catalan P, Hazen SP, et al. Brachypodium: 20 years as a grass biology model system; the way forward? Trends in Plant Science. 2022;27(10):1002-1016. doi:10.1016/j.tplants.2022.04.008

12.Gordon SP, Contreras-Moreira B, Woods DP, et al. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nat Commun. 2017;8(1):2184. doi:10.1038/s41467-017-02292-8

13.Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in De Novo Annotation Approaches. Xu Y, ed. PLoS ONE. 2011;6(1):e16526. doi:10.1371/journal.pone.0016526

14.Quesneville H, Bergman CM, Andrieu O, et al. Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLoS Comp Biol. 2005;1(2):e22. doi:10.1371/journal.pcbi.0010022

15.Quesneville H, Nouaud D, Anxolabéhère D. Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. Journal of Molecular Evolution. 2003;57(0):S50-S59. doi:10.1007/s00239-003-0007-2

16.Jurka J, Klonowski P, Dagman V, Pelton P. Censor—a program for identification and elimination of repetitive elements from DNA sequences. Computers & Chemistry. 1996;20(1):119-121. doi:10.1016/S0097-8485(96)80013-1

17.Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015. Accessed March 28, 2024. http://www.repeatmasker.org/

18.Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33. doi:10.12688/f1000research.29032.2

19.Baud A, Wan M, Nouaud D, Francillonne N, Anxolabéhère D, Quesneville H. Traces of Transposable Elements in Genome Dark Matter Co-Opted by Flowering Gene Regulation Networks. Genomics; 2019. doi:10.1101/547877

20.BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic. Accessed April 9, 2024. https://academic.oup.com/bioinformatics/article/26/6/841/244688

21.Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):157. doi:10.1186/s13059-015-0721-2

22.Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, ed. Bioinformatics. 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191

23.Hagberg A, Swart PJ, Schult DA. Exploring Network Structure, Dynamics, and Function Using NetworkX. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States); 2008. Accessed March 28, 2024. https://www.osti.gov/biblio/960616

24.Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent Activity in Expanding Populations and Purifying Selection Have Shaped Transposable Element Landscapes across Natural Accessions of the Mediterranean Grass Brachypodium distachyon. Genome Biology and Evolution. 2018;10(1):304-318. doi:10.1093/gbe/evx276

25.Yaakov B, Ben-David S, Kashkush K. Genome-Wide Analysis of Stowaway-Like MITEs in Wheat Reveals High Sequence Conservation, Gene Association, and Genomic Diversification. Plant Physiology. 2013;161(1):486-496. doi:10.1104/pp.112.204404

26.Horvath R, Minadakis N, Bourgeois Y, Roulin AC. The Evolution of Transposable Elements in Brachypodium Distachyon Is Governed by Purifying Selection, While Neutral and Adaptive Processes Play a Minor Role. Evolutionary Biology; 2023. doi:10.1101/2023.09.15.557873

27.Sinzelle L, Izsvák Z, Ivics Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. Cell Mol Life Sci. 2009;66(6):1073-1093. doi:10.1007/s00018-009-8376-3

28.Jangam D, Feschotte C, Betrán E. Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet. 2017;33(11):817-831. doi:10.1016/j.tig.2017.07.011

29.Capy P. Taming, Domestication and Exaptation: Trajectories of Transposable Elements in Genomes. Cells. 2021;10(12):3590. doi:10.3390/cells10123590

30.Fagny M, Kuijjer ML, Stam M, et al. Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. Front Genet. 2021;11:606285. doi:10.3389/fgene.2020.606285

31.Casola C, Hucks D, Feschotte C. Convergent Domestication of pogo-like Transposases into Centromere-Binding Proteins in Fission Yeast and Mammals. Molecular Biology and Evolution. 2008;25(1):29-41. doi:10.1093/molbev/msm221

32.Wong LH, Choo KHA. Evolutionary dynamics of transposable elements at the centromere.

Trends in Genetics. 2004;20(12):611-616. doi:10.1016/j.tig.2004.09.011

33.Gao D, Jiang N, Wing RA, Jiang J, Jackson SA. Transposons play an important role in the evolution and diversification of centromeres among closely related species. Front Plant Sci. 2015;6. doi:10.3389/fpls.2015.00216

34.Bartolomé C, Maside X. The lack of recombination drives the fixation of transposable elements on the fourth chromosome of Drosophila melanogaster. Genet Res. 2004;83(2):91-100. doi:10.1017/s0016672304006755

35.Jurka J, Bao W, Kojima KK. Families of transposable elements, population structure and the origin of species. Biology Direct. 2011;6(1):44. doi:10.1186/1745-6150-6-44

36.Almojil D, Bourgeois Y, Falis M, Hariyani I, Wilcox J, Boissinot S. The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes. Genes. 2021;12(6):918. doi:10.3390/genes12060918

37.Straub T. Heterochromatin Dynamics. PLOS Biology. 2003;1(1):e14. doi:10.1371/journal.pbio.0000014

38.Charlesworth B, Jarne P, Assimacopoulos S. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. III. Element abundances in heterochromatin. Genet Res. 1994;64(3):183-197. doi:10.1017/s0016672300032845

39.Bartolomé C, Maside X, Charlesworth B. On the Abundance and Distribution of Transposable Elements in the Genome of Drosophila melanogaster. Molecular Biology and Evolution. 2002;19(6):926-937. doi:10.1093/oxfordjournals.molbev.a004150

40.Charlesworth B, Jensen JD. Effects of Selection at Linked Sites on Patterns of Genetic Variability. Annu Rev Ecol Evol Syst. 2021;52:177-197. doi:10.1146/annurev-ecolsys-010621-044528

41.SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20(1):43-45. doi:10.1038/1695

42.Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics. 2007;8:218. doi:10.1186/1471-2164-8-218

43.Daron J, Glover N, Pingault L, et al. Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biology. 2014;15(12):546. doi:10.1186/s13059-014-0546-4

44.Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass Brachypodium distachyon. New Phytologist. 2020;227(6):1736-1748. doi:10.1111/nph.16308

45.Britten RJ, Baron WF, Stout DB, Davidson EH. Sources and Evolution of Human Alu Repeated Sequences. Proceedings of the National Academy of Sciences of the United States of America. 1988;85(13):4770-4774. Accessed April 18, 2024. https://www.jstor.org/stable/31869

46.Maumus F, Quesneville H. Impact and insights from ancient repetitive elements in plant genomes. Current Opinion in Plant Biology. 2016;30:41-46. doi:10.1016/j.pbi.2016.01.003

47.Minadakis N, Williams H, Horvath R, et al. The Demographic History of the Wild Crop Relative Brachypodium Distachyon Is Shaped by Distinct Past and Present Ecological Niches. Ecology; 2023. doi:10.1101/2023.06.01.543285

48.Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3(5):329-341. doi:10.1038/nrg793

49.Kapitonov VV, Jurka J. Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc Natl Acad Sci U S A. 2003;100(11):6569-6574. doi:10.1073/pnas.0732024100

50.Fornara F, Pařenicová L, Falasca G, et al. Functional Characterization of OsMADS18, a Member of the AP1/SQUA Subfamily of MADS Box Genes. Plant Physiol. 2004;135(4):2207-2219. doi:10.1104/pp.104.045039

51.Chen ZS, Liu XF, Wang DH, et al. Transcription Factor OsTGA10 Is a Target of the MADS Protein OsMADS8 and Is Required for Tapetum Development1[OPEN]. Plant Physiol. 2018;176(1):819-835. doi:10.1104/pp.17.01419

52.Rezanejad F. The response of anther and pollen development, pollen cellular material release and pollen proteins to air pollution in Petunia juss Juss. (Solanaceae). Iranian Journal of Science and Technology, Transaction A: Science. 2013;37:63-68.

53.Ambawat S, Sharma P, Yadav NR, Yadav RC. MYB transcription factor genes as regulators for plant responses: an overview. Physiol Mol Biol Plants. 2013;19(3):307-321. doi:10.1007/s12298-013-0179-1

54.Sabot F, Picault N, El-Baidouri M, et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. The Plant Journal. 2011;66(2):241-246. doi:10.1111/j.1365-313X.2011.04492.x

55.Kofler R, Gómez-Sánchez D, Schlötterer C. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. Mol Biol Evol. 2016;33(10):2759-2764. doi:10.1093/molbev/msw137

56.Groza C, Bourque G, Goubert C. A Pangenome Approach to Detect and Genotype TE Insertion Polymorphisms. In: Branco MR, De Mendoza Soler A, eds. Transposable Elements. Vol 2607. Springer US; 2023:85-94. doi:10.1007/978-1-0716-2883-6_5

57.Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011;43(3):269-276. doi:10.1038/ng.768

58.Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Research. 2014;42(11):6826-6838. doi:10.1093/nar/gku323

59.Qiu Y, Köhler C. Mobility connects: transposable elements wire new transcriptional networks by transferring transcription factor binding motifs. Biochemical Society Transactions. 2020;48(3):1005-1017. doi:10.1042/BST20190937

60.Chu C, Borges-Monroy R, Viswanadham VV, et al. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat Commun. 2021;12(1):3836. doi:10.1038/s41467-021-24041-8

61.Groza C, Chen X, Wheeler TJ, Bourque G, Goubert C. GraffiTE: a Unified Framework to Analyze Transposable Element Insertion Polymorphisms using Genome-graphs. Published online September 15, 2023. doi:10.1101/2023.09.11.557209

62.Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biology. 2020;21(1):265. doi:10.1186/s13059-020-02168-z

63.Schlagenhauf, Edith and Wicker, Thomas. The TREP platform: A curated database of

transposable elements. Available at https://trep-db.uzh.ch. Published online 2016.

64.Stritt C, Gimmi EL, Wyler M, et al. Migration without Interbreeding: Evolutionary History of a Highly Selfing Mediterranean Grass Inferred from Whole Genomes.; 2020:2020.09.03.280842. doi:10.1101/2020.09.03.280842

65.Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Research. 2020;48(D1):D1104-D1113. doi:10.1093/nar/gkz1020

66.Li Y, Zuo S, Zhang Z, et al. Centromeric DNA characterization in the model grass Brachypodium distachyon provides insights on the evolution of the genus. The Plant Journal. 2018;93(6):1088-1101. doi:10.1111/tpj.13832

67.Huo N, Garvin DF, You FM, et al. Comparison of a high-density genetic linkage map to genome features in the model grass Brachypodium distachyon. Theor Appl Genet. 2011;123(3):455-464. doi:10.1007/s00122-011-1598-4

68.Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163

69.Bryant C, Wheeler NR, Rubel F, French RH. kgc: Koeppen-Geiger Climatic Zones. Published online November 20, 2017. doi:10.32614/CRAN.package.kgc

70.Jamilloux V, Daron J, Frederic C, Quesneville H. De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue. Proceedings of the IEEE. 2016;PP:1-11. doi:10.1109/JPROC.2016.2590833

71.Hoede C, Arnoux S, Moisset M, et al. PASTEC: An Automatic Transposable Element Classification Tool. Cordaux R, ed. PLoS ONE. 2014;9(5):e91929. doi:10.1371/journal.pone.0091929

72.Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021. doi:10.21105/joss.03021

73.Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference. Published online 2010:92-96. doi:10.25080/Majora-92bf1922-011

Code availability

panTEannot source code is available at https://forgemia.inra.fr/urgi-anagen/panTEannot.
panrREPET source code is available at https://forgemia.inra.fr/urgi-anagen/panREPET.

Acknowledgement

We thank Maela Semery for providing climate classes, temperature, precipitation and wind of each accession **(see Materials and Methods - Brachypodium distachyon data)**. We thank Clémentine Vitte and Leandro Quadrana for their relevant comments. We thank Gautier Sarah for testing panTEannot.

Author information - Contributions

HQ and JC conceived the project. MB, MRO and SS developed panREPET. SS developed panTEannot and generated all manuscript results on *B. distachyon*. SS wrote the manuscript under the supervision of HQ and JC.