

RESEARCH

Open Access



Characterization of active transposable elements and their new insertions in tuber propagated greater yam (*Dioscorea alata*)

Sajjad Ali Panhwar^{1†}, Dandan Wang^{1†}, Fanhui Lin^{1†}, Ying Wang¹, Mengli Liu¹, Runan Chen¹, Yonglan Huang¹, Wenqiang Wu², Dongyi Huang^{1*}, Yong Xiao^{1*} and Wei Xia^{1*}

Abstract

Background Greater yam is a key staple crop grown in tropical and subtropical regions, while its asexual propagation mode had led to non-flowering mutations. How transposable elements contribute to its genetic variations is rarely analyzed. We used transcriptome and whole genome sequencing data to identify active transposable elements (TEs) and genetic variation caused by these active TEs. Our aim was to shed light on which TEs would lead to its intraspecies variation.

Results Annotation of de novo assembly transcripts indicated that 0.8 – 0.9% of transcripts were TE related, with LTR retrotransposons (LTR-RTs) accounted for 65% TE transcripts. A large portion of these transcripts were non-autonomous TEs, which had incomplete functional domains. The majority of mapped transcripts were distributed in genic deficient regions, with 9% of TEs overlapping with genic regions. Moreover, over 90% TE transcripts exhibited low expression levels and insufficient reads coverage to support full-length structure assembly. Subfamily analysis of Copia and Gypsy, the two LTR-RTs revealed that a small number of subfamilies contained a significantly larger number of members, which play a key role in generating TE transcript. Based on resequencing data, 15,002 L-RT insertion loci were detected for active LTR-RT members. The insertion loci of LTR-RTs were highly divergent among greater yam accessions.

Conclusions This study showed the ongoing transcription and transpositions of TEs in greater yam, despite low transcription levels and incomplete proteins insufficient for autonomous transposition. While our research did not directly link these TEs to specific yam traits such as tuber yield and propagation mode, it lays a crucial foundation for further research on how these TE insertion polymorphisms (TIPs) might be related to variations in greater yam traits

[†]Sajjad Ali Panhwar, Dandan Wang, Fanhui Lin contributed equally to this work

*Correspondence:

Dongyi Huang

hdongyi@hainanu.edu.cn

Yong Xiao

xiaoyong@hainanu.edu.cn

Wei Xia

saizjxiawei@hainanu.edu.cn

Full list of author information is available at the end of the article



and its tuber propagation mode. Future research may explore the potential roles of TEs in trait variations, such as tuber yield and stress resistance, in greater yam.

Keywords Active transposable elements, Truncated transposable elements, Transposable elements insertion polymorphisms, Genetic diversity

Background

Greater yam (*Dioscorea alata*) is an important staple crop grown in tropical and subtropical region, which is also an orphan crop with 90% production comes from Nigeria, Benin, Ghana, Togo, and Cote d'Ivoire in West African [1]. As a dioecious and tuberous crop, it is usually vegetatively propagated by tuber instead of seeds to maintain desirable qualities. As a result, genetic variation in greater yam is not primarily the result of sexual recombination, but rather reflects variations in its innate genetic makeup. However, the reduced ability of sexual propagation is widely observed in tuber crops because of the domestication and diversification processes [2]. Yam was also characterized by a shift from sexual to vegetative propagation, and popular landraces in a hybridization programme have never flowered or have only flowered irregularly and sparsely [3, 4]. Without sexual propagation, the genetic variations may accumulate in somatic cells, including new insertions of transposable elements (TEs). The contribution of TEs contribute to greater yam genomic variations has not been studied.

TEs comprise a substantial proportion of plant genomes, such as accounting for 88% genomic sequence in the first maize telomere-to-telomere genome version, more than 80% in coconut palm [5, 6]. TEs could shape genome evolution, re-wire and fine-tune transcriptome by increasing transcript diversity and producing regulatory siRNA via providing novel promoters, splice sites, or polyadenylation signals, thus generating transcript diversity [7]. Research showed that TEs upstream genes could keep the genetic heterozygosity in inbred population via balancing selection [8]. Most TEs were inactive in genome via epigenetic marks, in gametes the silenced status will be reinforced via epigenetic reprogramming and small RNA silencing [9–11]. However, a few TEs keep their active nature and continue to mutagenic activity [12–14], or more TEs might be reactivated under stress, specific developing stage, or genomic shocks [9, 15, 16]. Identification and characterization of active TE loci in genome would shed light on the new TE insertion tendency.

Intraspecific TE insertion polymorphic (TIP) loci were considered to be caused by active TE loci within species, which represent a major source of intraspecific variation [17]. Active TEs identification based on whole genome sequencing (WGS) data were widely conducted in human and plants. Specific somatic TE insertions in epithelial tumors were identified by using WGS data, which

influenced the expression of genes that were commonly mutated in cancer [18]. Pan-cancer analysis of ~3000 cancer genomes showed that long interspersed nuclear element caused the most frequent type of somatic structural variations (SVs) and mediated various types of SVs in cancer cells [19]. Moreover, based on WGS data, TIPs were detected for 3000 rice varieties and used for genome-wide association study which indicated that these new insertion events may be triggered by external stimuli [20]. More than 40 TIPs were found to be associated with agronomic traits in tomato based on 6,906 TIP markers [21]. These studies highlight the importance of TIP information in trait dissection and key gene validation, with important implications for breeding.

Direct evidence was also obtained by transcriptome data for active TEs. Studies conducted in human cells indicated specific TEs expressed in somatic cells were undetectable in bulk analyses, highlighting their contribution to cell heterogeneity [22, 23]. In this study, we identified and characterized TE related transcripts in greater yam based on transcriptome datasets, and explored their original genomic loci. We also featured the subfamilies to which these active long terminal repeat retrotransposons (LTR-RTs) belong. Additionally, we analyzed TIPs among different yam accessions using WGS data, some of which were validated by polymerase chain reaction (PCR). We aimed to characterize active TEs and identify their new insertion loci in greater yam, which could be used for further trait variation dissection.

Methods

Plant materials

In this study, a set of 14 greater yam accessions were utilized for the detection of TE insertion/transcription. These accessions included Da5, Da40, Da70, Da86, Da88, Da90, Da94, Da95, Da154, Da191, Da207, Da218, Da280, and Da933. Among these, ten accessions were specifically chosen for TE insertion validation, namely Da40, Da70, Da86, Da88, Da95, Da191, Da207, Da218, Da280, and Da933. Furthermore, a total of twelve accessions were selected for reverse transcription polymerase chain reaction (RT-PCR) validation, which comprised Da5, Da40, Da70, Da86, Da88, Da90, Da94, Da154, Da191, Da207, Da280, and Da933. Total RNA and DNA were extracted from the third and fourth leaves from the top of young plantlets cultured in the MS medium under the 16-hour day/8-hour night cycle. Total DNA extraction was done by the cetyltrimethylammonium bromide (CTAB)

protocol, while total RNA extraction was conducted according to Xiao et al. (2012) [24].

Expressed TEs annotation and RT-PCR validation

Transcriptome datasets of greater yam were downloaded from the National Center for Biotechnology Information website, and the detailed information was listed in Table S1a. We are using Da56 (4 transcriptomes), Da25 (4 transcriptomes), DP (1 transcriptome), DW (1 transcriptome), IB10 (1 transcriptome), T1-T4 (12 transcriptomes), and MH1 (1 transcriptome) to represent seven greater yam accessions, each with different numbers of transcriptomes. There were no samples that overlapped with the accessions used for PCR or RT-PCR validation. De novo assembly of each transcriptome was conducted by the Trinity 2.2.10 [25]. HMMER (v3.4) was used to identify the conserved domains of deduced peptides from transcripts and screen for TE related proteins. RepeatMasker (v4.1.6) was used to identify putative TEs in the greater yam's transcriptome (matched length ≥ 100 bp, score ≥ 250). The combination of these prediction results leads to the identification of transcripts containing TE related sequences. The family classification of TE related transcripts is based on the above two software annotation results.

The determination of full-length, incomplete, and non-coding TE transcripts was based on their deduced amino acid length, and the completeness of their conserved protein domain, which were obtained by the software TransDecoder [26]. The domains of full-length TE were based on the reference suggested a unified TE system [27], while the amino acids shorter than 50 was considered as noncoding.

The TE related transcript identified in the above transcriptome were used as reference sequences. Because the analysis is not including all transcripts from the transcriptomes, we used Fragments Per Kilobase Million (FPKM) values to estimated their expression levels among different samples. The FPKM values of which were estimated by RNA sequencing (RNAseq) reads mapping and assembly by hisat2 and stringtie, respectively [28–30]. The mapped reads for TE transcripts were visualized by IGV [31], and the heatmap of FPKM values were displayed by MEV4.0 [32]. The TE transcripts belonging to the Gypsy, Copia, Enhancer/Suppressor-mutator (Enspm), Mutator-Don Robertson (MULE-MuDR), PiggyBac Insertion Element from Harbinger (PIF-Harbinger), Long Interspersed Element (LINE), and hobo Activator Transposon (hAT) families were randomly selected for RT-PCR validation. Ten pairs of primers were designed for TE transcripts validation via the software Primer 5 and listed in Table S2a.

Genomic mapping of TE transcripts and LTR-RTs subclassification

The v2.0 genome sequence of greater yam was downloaded from the Phytozome v13 website (<https://phytozome-next.jgi.doe.gov/>) [1]. The identified TE transcripts were mapped to the reference genome by BLAST analysis with the parameters - identities $>95\%$ and matched length $>90\%$. The distance of TE and the nearest genes were calculated based on their genomic location, while TE overlapped with genic regions, such as the promoter region (the 2 kb upstream region from the gene start codon), 3' untranslated region (UTR, the 1 kb downstream region from the gene stop codon), intronic region, and exonic region, were determined based on the gene models of the v2.0 greater yam genome.

The LTR-RTs (Copia and Gypsy) in greater yam genome were screened by RepeatMasker (v4.1.6) and BLAST analysis by conserved reverse transcriptase peptides from PFAM database. The full-length LTR-RTs and long terminal repeat (LTR) sequences were identified by LTR_finder [33]. The overlapped site of LTR-type TEs and mapped TE transcripts were identified via comparison of genomic location. The subfamilies of Copia and Gypsy are determined by phylogenetic analysis of their reverse transcriptase sequences through CLUSTALW [34], visualized via the unrooted phylogenetic tree display in the iTOL website (v6, <https://itol.embl.de/>).

Analysis of active LTR-RT insertion sites analysis and PCR validation of TE insertions

Twenty-three LTR-RTs were randomly selected for insertion polymorphism analysis (LTR-TIPs). The LTR sequences were used as references to detection insertion loci. Twenty whole genome resequencing datasets of greater yam were downloaded from the NCBI website (Table S1b). The raw reads were analyzed via FastQC (v0.12.0) and further treated with Trimmomatic (v0.35, parameters were set as: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 HEADCROP:8 MINLEN:36 HEADCROP:15). The LTR-TIPs analysis pipeline followed the software TRACKPOSON [20], combining with BLAST analysis for precise insertion loci. The LTR-TIP loci were identified when the pair-end reads with one read matched to LTR sequence and the other matched to the greater yam reference genome. The supported reads were more than three in one sample, but could lower to two reads for other samples with the same LTR-TIP loci. Thirty-three pairs of primers were designed for LTR sequences and the corresponding flanking sequences to validate LTR-TIP loci (Table S2b).

Results

Active TEs detected in greater yam transcriptome

We analyzed a collection of 24 transcriptomes in this study, comprising eight transcriptomes produced in-house (Da56 and Da25), and 16 transcriptomes sourced from the NCBI website for tuber (2), inflorescence (1), bulbil (12), and leaf (1) tissues (Table S1a). The transcripts assembled de novo by the Trinity software for each accession - Da56 (4 transcriptomes), Da25 (4), DP (1), DW (1), IB10 (1), T1-T4 (12), and MH1 (1) underwent analysis using RepeatMasker and HMMsearch. The integration of findings from RepeatMasker and HMMsearch revealed that the TE-related transcripts in the examined transcriptome range from 604 to 1,639, constituting approximately 0.8–0.9% of the assembled transcripts in total (Table 1). Retrotransposons accounted for three to four times more transcripts than DNA transposons, with LTR retrotransposon (LTR-RTs) was the dominant type. Among LTR-RTs, the *Copia* and *Gypsy* types were the two most common main types, accounting for 43% and 22% of the detected TEs, respectively. Among DNA transposons, the EnSpm (9%) and hAT (7%) transposon family were the top two DNA transposons (Fig. 1A).

Because active TEs could cause genetic variation through their genomic insertion events, we analyzed the characteristics of these TE transcripts to determine if they contained full protein sequences necessary for transposition. Our findings revealed that a significant portion of these transcripts were non-autonomous TEs with incomplete functional domains (Fig. 1). More than 50% of TE-related transcripts were shorter than 2 kb in all TE

classes, and the majority of transcripts were shorter than 4 kb (Fig. 1B). Three TE families – *Copia*, *Gypsy*, and EnSpm - were considered to have long coding regions, but 75% to 80% of transcripts were shorter than 2 kb. We predicted the amino acid sequences for all TE-transcripts. A few full-length autonomous TEs were detected in *Copia* (1%), *Gypsy* (3%), LINE (2%), and EnSpm (5%) families, which had full length amino acid sequences longer than 1000 aa (Fig. 1C). The proportion of full-length TEs was related to their coding sequences length for hAT (21%), MULE-MuDR (24%), PIF-Harbinger (29%), and Caulimovirus (34%), with TE family of shorter sequences having more full-length TE-transcripts detected. In addition, a considerable amount of non-coding transcripts with deduced peptide length shorter than 50 aa were detected, which accounted for 40% in *Copia* (the most) and 10% in LINEs (the fewest). Therefore, the dominant proportion of these active TEs were truncated or noncoding, which would not cause severe genomic variations by their own.

Active TEs located in low gene density region

The transcripts related to TEs were aligned with the v2.0 genome sequence of greater yam [1]. We found 3,520 out of 6,878 transcripts were mapped to the genome with identities $\geq 95\%$ and coverage $\geq 90\%$. This left approximately 49% of TEs showing considerable divergence from the reference genome. Taking into account transcripts mapped to multiple loci and overlapping transcripts loci, a total of 13,796 unique TE loci were identified. This corresponds to about 2.8 active TEs per 100 kb and they are unevenly distributed across the chromosomes (Fig. 2).

Table 1 Transposable elements detected in greater yam transcriptome datasets

TE Class	Da56	Da25	DP	DW	IB10	T1-T4	MH1
DNA Transposon	260	169	165	180	256	374	162
MULE-MuDR	46	28	21	38	28	76	27
EnSpm	90	58	51	69	122	145	58
PIF-Harbinger	29	16	22	19	38	40	21
hAT	92	67	55	54	68	99	53
other	3	0	16	0	0	14	3
Retroposon	948	692	658	424	554	1265	772
LINE type	128	88	64	27	132	315	64
L1	124	82	54	26	132	309	51
RTE-BovB	4	6	10	1	0	6	13
LTR type	820	604	594	397	422	950	708
<i>Copia</i>	477	340	416	297	253	641	461
<i>Gypsy</i>	327	256	159	91	137	281	239
Caulimovirus	9	5	12	4	15	12	3
other	6	3	7	5	17	16	5
Total	1207 (1.7%)	861 (1.4%)	823 (0.8%)	604 (0.9%)	810 (0.9%)	1639 (0.2%)	934 (0.9%)

Note TEs from HMMsearch had $c\text{-value} \leq 1e-5$, $E\text{-value} \leq 1e-5$, while TEs from RepeatMasker had matched length longer than 100 bp and scores higher than 250. Transcripts from Da56 (SRR14574591 to SRR14574594) and Da25 (SRR14574587 to SRR14574590) were assembled from transcriptome datasets from leaf, stem, and tuber tissues. DP, DW, IB10, and MH1 represent the transcripts assembled from SRR1518381, SRR1518382, SRR3938623, and SRR9127739, respectively. T1-T4 represents the transcripts assembled from 12 transcriptomes - SRR7496967 to SRR7496972. The percentage indicates the ratio of TEs related transcripts to all assembled transcripts

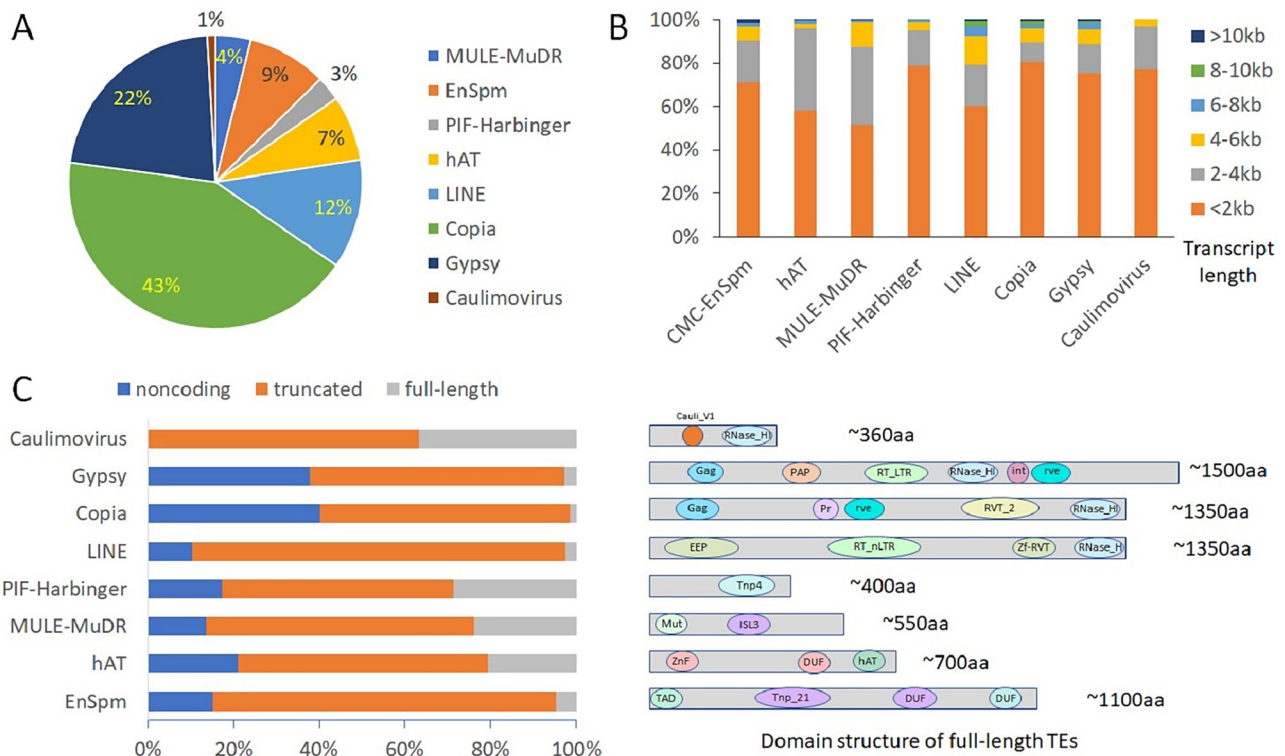


Fig. 1 Composition and characters of active TEs in transcriptome. **(A)** Pie chart of different TE families from greater yam transcriptomes; **(B)** The transcript length proportions for different TE families; **(C)** The proportion of “full-length”, “truncated”, and “noncoding” TE transcripts on the left, and the schematic diagram of full-length TEs with complete functional domains identified in this study. The truncated TE transcripts were shorter than full-length TEs and lacked complete functional domains. The noncoding transcripts had deduced amino acids less than 50 aa

The highest density of the mapped TE loci (peaks of NO. TEs per 100 kb) more frequently appeared in the low-density gene region, which was likely to be the centromeric and pericentromeric regions for all chromosomes. The average active TE density was highest in Chr6 (3.4 TEs per 100 kb), and lowest in Chr4 (2.2 TEs per 100 kb) and Chr5 (2.1 TEs per 100 kb) (Table S3 and Fig. 2).

For 13,796 mapped TE loci, the average number of mapping loci is 3.9 per TE-transcript. However, among the seven main TE families, more than 70% of TEs were single copy for LINE and MULE-MuDR, followed by PIF-Harbinger (60%), whereas only 20% of CMC-EnSpm TEs were single copy and 54% were 2 to 10 copies (Fig. 3A). DNA types TEs - hAT, CMC-EnSpm, and PIF-Harbinger had relatively higher proportion of 10 to 30 TE copies, while Gypsy and MULE-MuDR have the TE members with extremely high copies (more than 100).

Calculating the distance between TEs and the nearest protein-coding gene of the active TEs indicated that 84% of TEs were more than 2 kb away from genes, while 16% of TEs may have an influence on gene status for their close distance (less than 2 kb) from the gene or overlapping with the genic region (Fig. 3B). Out of the 9% of TEs (1,267) that overlapped with genes, 193 annotated genes were rechecked as TE-related proteins and excluded

from Table 2. For the remaining 1074 TE loci, the majority (88%, 949) were located in the intronic region and 125 TEs were overlapped with exon sequences (Table 2). Out of the 125 TEs, 98% TEs were of the “truncated” or “non-coding” type. Further analysis showed that the shared sequences between exons and TEs were short, comprising mostly about 1%~2% of the TE transcripts. Meanwhile the remaining parts of the TEs were located in the intronic region or untranslated region, with short open reading frames that were frameshift (Fig. 3C). There were 411 and 237 TEs located in the respect putative promoter and 3’UTR regions of the genes, respectively, which may influence the transcriptional status of the genes.

Most TE-transcripts have low expressions except for some conservatively expressed loci

Based on the transcriptome datasets we analyzed, the expression levels of most TEs were low. More than 50% TEs in Da25, Da56, and T1-T4 transcriptome have negligible expression levels (FPKM values < 1), while more than 90% TEs in all transcriptomes have FPKM values lower than 10 (Fig. 4A). TEs with high expression levels (FPKM ≥ 30) only account for 1–3% of TEs transcripts. Among truncated, noncoding, and full-length TE transcripts, the full-length type had a higher proportion of

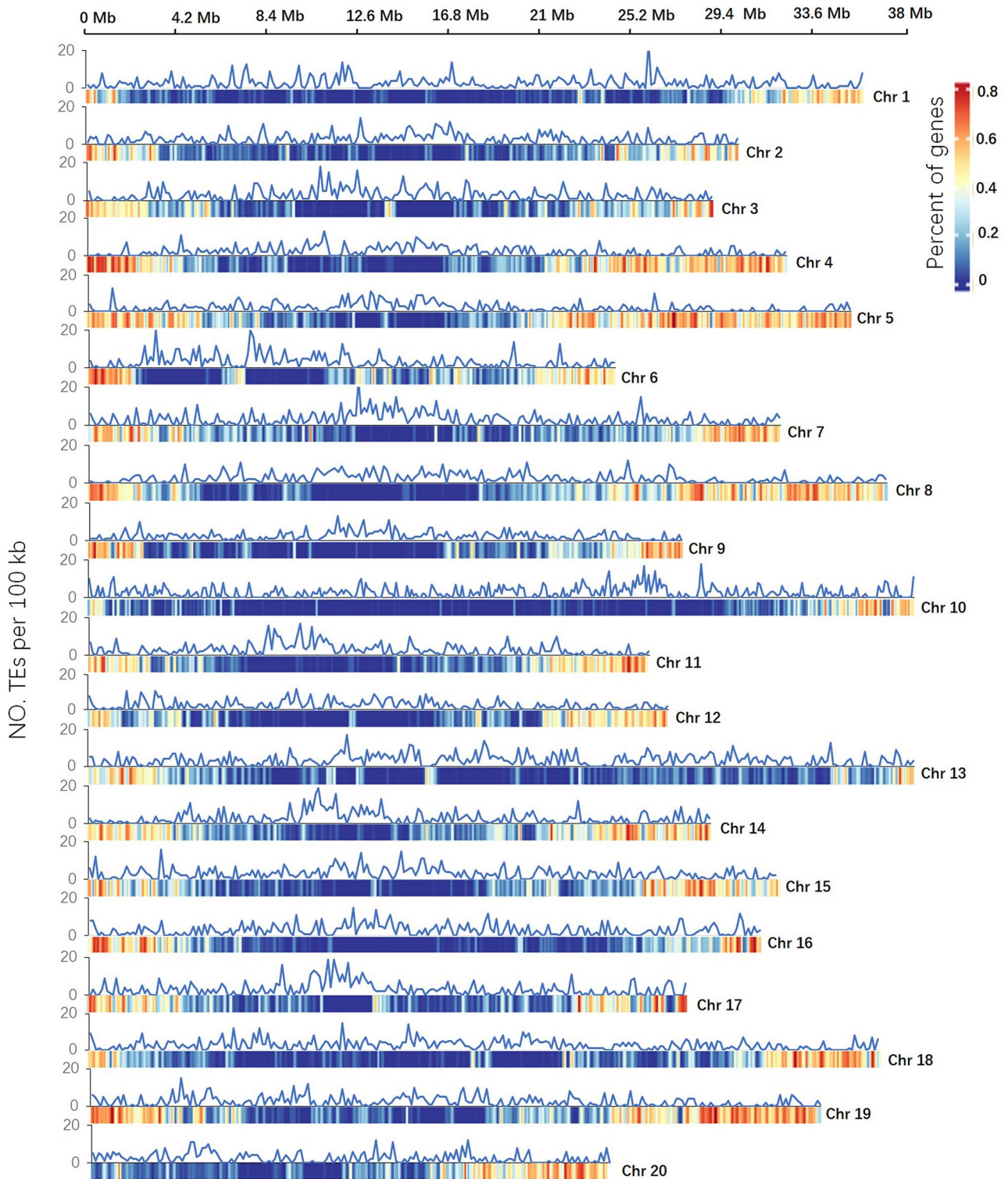


Fig. 2 The distribution of active TEs mapped to the greater yam genome. The TE density curves were set above the chromosome bars, which were TE numbers per 100 kb. The chromosome bars were color-coded to display the percentage of protein-coding genes per 10 kb window

high expression transcripts, which was consistent with the successful assembly of full-length transcripts due to high coverage RNA-seq reads (Fig. 4B). Most noncoding TE transcripts have FPKM values lower than 10, with

90% of them having a length between 200 bp and 500 bp. Additionally full-length and truncated TE transcripts exhibits a similar length distribution.

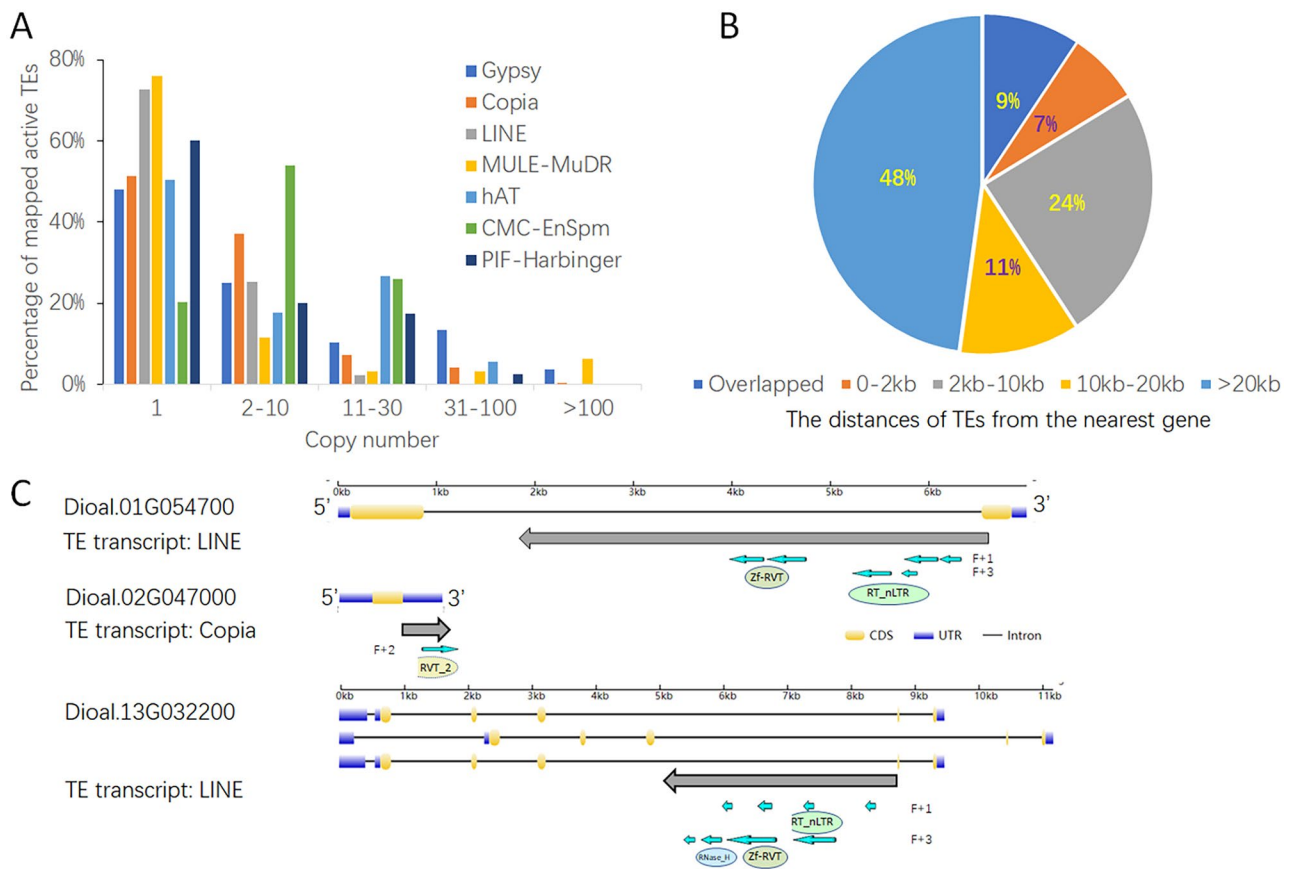


Fig. 3 The copy number and position characteristics of active TEs with adjacent genes. **(A)** Copy number of mapped expressed TE loci for different TE classes; **(B)** The proportion of different distances between active TEs and the closest genes; **(C)** Three examples of active TEs overlapped with gene exonic regions. Blue arrows under the grey TE transcripts were predicted open reading frames

TEs with high expression levels were few in the transcriptome, half of which were expressed specifically in one or two cultivars. Among the analyzed transcriptome, 69 TE transcripts had FPKM values higher than 30 in one or more transcriptomes, with only 14 having high expression in most cultivars (Fig. 4C). Further analysis of these TE transcripts indicated that some were caused by wrongly assembled transcripts which containing both TE and non-TE genes, with low coverage of RNAseq reads supporting the structure (data not shown), while others had only TE sequences that were supported by RNAseq reads (Fig. 4C).

According to the transcriptome data, a set of twelve TEs representing eight TE families were randomly selected for expression detection (Fig. 5). Nine TEs showed expression in more than one greater yam accession, except for LTR/Gypsy-1 (Fig. 5, left). Six TE loci were expressed in most of the detected greater yam accession (10/12), while the remaining loci had specific expression in certain accessions. The expression pattern was also observed in the 24 transcriptome datasets: hAT-Ac, LTR/Gypsy-2, and LTR/Copia-1 expressed in most samples, but the transcript for PIF-Harbinger and

LTR/Copia-2 had specific high expression in DP, DW, and T1-T4 for the former and Da56 for the latter (Fig. 5, right). Based on the similarity of active TE detection, sets of accessions – Da40/Da154/Da191, Da70/Da88/Da90, and Da94/Da280 – had same TE expression pattern within groups (Pearson’s Correlation Coefficient: 1.0), but showed variable expression pattern between sets of accessions (PCC<0.5).

Active LTR-RT sites contributed to the expansion of various subfamilies

LTR-RTs (*Copia* and *Gypsy*) account for a large portion of TEs distributed in genomes, as well as in TE transcripts in transcriptomes. A total of 1141 full-length reverse transcription (RT) domains for *Copia* family were extracted from the greater yam genome and used for phylogenetic analysis. The phylogenetic tree indicated that the *Copia* family could be divided into 27 subfamilies (Fig. 6A). The numbers of *Copia* loci were divergent among different subfamilies: subfamily X (130), VII (120), V (113), and IV (110) had the most TEs members. On the contrary, subfamily VIII (7), XI (10), VI (10), and VIII (12) had the least TEs members (Fig. 6A). Classification

Table 2 Numbers and location characteristics of active TEs near to the protein-coding genes

Chromosome	Number of TEs			Intron	Promoter ^a	3'UTR ^b
	Overlapped with Exon					
	full-length	truncated	noncoding			
Chr1	0	2	1	16	20	8
Chr2	0	1	1	47	17	12
Chr3	1	4	0	52	24	21
Chr4	0	2	1	52	21	6
Chr5	0	1	0	57	30	18
Chr6	0	4	5	43	13	8
Chr7	0	9	7	66	20	10
Chr8	0	0	0	54	13	16
Chr9	0	2	0	46	17	9
Chr10	0	4	3	38	16	17
Chr11	0	8	6	40	13	10
Chr12	0	4	2	29	17	8
Chr13	1	7	1	64	23	16
Chr14	0	1	1	50	32	11
Chr15	0	4	0	48	25	9
Chr16	0	0	0	50	19	12
Chr17	0	6	5	37	20	8
Chr18	0	4	7	54	16	13
Chr19	1	3	2	66	31	13
Chr20	0	9	5	40	24	12
Total	3	75	47	949	411	237

Note The head or end sequences of TEs overlapped with exons were classified as overlapped with exons. The three types of TEs - "full-length", "truncated", and "noncoding"-were determined according to their sequence length, the completeness of functional domain, as well as amino acid length. The deduced amino acid sequences shorter than 50 amino acids were considered as noncoding. ^a represented that TEs locating in the putative promoter as 2 kb upstream from the start code of the nearest genes. ^b represented that TEs locating in the 3'UTR region as 1 kb downstream from the stop code of the nearest genes

of *Copia* TE transcripts into these families showed that the numbers of transcripts were highly different between subfamilies. Subfamily X and V were the top two, each having more than 100 transcripts. Half of the numbering less than 10, including subfamilies XVIII and XX, which had zero expressed loci (Fig. 6B, top). When considering the mapped TE transcript loci, subfamilies that detected more TE transcripts, such as subfamily V and X, also exhibited a higher number of active TE loci (Fig. 6B, middle). LTR-RTs with complete and same LTR sequences were considered as recently inserted TEs, but active *Copia* loci do not show a positive correlation with the LTR feature (T-test, $p < 0.3$). For the 27 subfamilies, the portion of active TEs with LTRs was similar between active *Copia* loci and those non-active in this study, except for subfamilies VII, XII, XIX, XXIII, XXV, and XXVII (Fig. 6B, bottom).

Moreover, 985 complete RT domain sequences of the *Gypsy* family were derived from the greater yam genome and used for phylogenetic analysis. The *Gypsy* family could be divided into 18 subfamilies. Subfamily I is the most abundant subfamily with 192 members, while the other subfamilies had less than 100 members (Fig. 7). The classification of *Gypsy* transcripts showed that each subfamily had fewer than 40 detected transcripts, with

slightly more transcripts detected in subfamilies X, XI, XII, XV, and XVII (>20). The percentage of active *Gypsy* loci was lower than that of active *Copia* loci, but the portion of active *Gypsy* loci with both LTR sequences were higher than inactive loci (Fig. 7B).

LTR-RT insertion events were highly divergent among greater yam accessions

As LTR-RTs are dominant TEs in the genome and have accurate insertion borders for new insertions, 52 L sequences from the active TEs loci were selected for insertion polymorphism analysis. The phylogenetic tree constructed from these LTR sequences showed that they could be classified into 12 sub-groups: the top three sub-group I (14), II (8), and III (13) had the most members, while the remaining sub-groups had one to three members each (Fig. 8A). Based on resequencing datasets for twenty greater yam accessions, a total of 15,002 L-RT insertion loci for were identified. Among the 12 sub-groups, sub-group I had the top number of detected insertion loci (7378), followed by subgroup IV (2768), III (1656), and VIII (1422), while subgroup IX (12), V (29), and VI (78) had the least insertion loci (Table 3). The number of detected LTR-RT insertion loci was positively

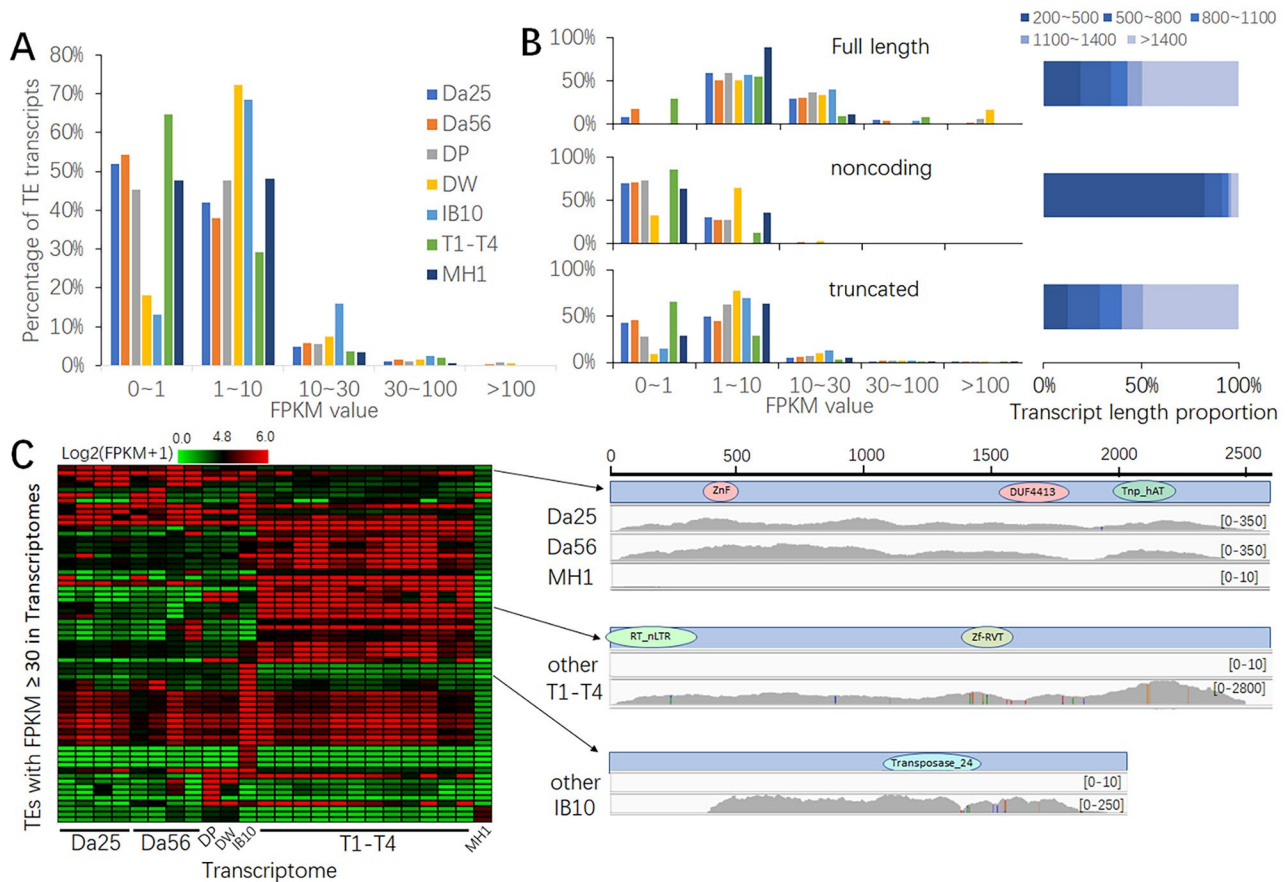


Fig. 4 The expression characteristics of active TEs. **(A)** Compositions of active TEs in different greater yam transcriptome datasets with different expression levels: 0~1 (negligible), 1~10 (low), 10~30 (median), 30~100 (high), and >100 (high). **(B)** Expression levels for three types of TE transcripts (“full length”, “noncoding”, “truncated”; detailed information in Table 1). **(C)** Heatmap of TE transcripts with FPKM ≥ 30 on the left and RNA-seq reads distribution of three TE transcripts on the right. The detailed information of the transcriptome datasets is the same as that in Table 1

correlated with the number of members in the subgroup (Pearson’s Correlation Coefficient: 0.66).

High divergence of LTR-RT insertion loci was detected among the twenty greater yam accessions (Fig. 8B and Table S4). The sample SRR13633111 had the least number of LTR-RT insertion loci (1287) detected, with approximately seven loci per million reads, followed by SRR13633103 (2994, 20 loci per million reads) and SRR13633102 (3217, 17 loci per million reads). SRR13633101 had the highest number of LTR-RT insertion loci, but the highest frequency for LTR-RT insertion detection is between 85 and 93 loci per million reads for SRR10552063, SRR10552064, and SRR10552065. The average LTR-RT loci per million reads is 52 (Table S4). The insertion loci of LTR-RTs were not correlated with the coverage of the resequencing datasets (Pearson’s Correlation Coefficient : 0.07), but they do have a positive correlation with the length of the chromosome length (Table S4).

For the detected LTR-RT insertion loci, most of them (98%) were shared in two or more than two samples,

with only 240 loci being unique to a single sample. The peak number of greater yam accessions sharing the same LTR-RT insertion loci is four samples, while two to seven greater yam accessions had more than 1000 shared LTR-RT loci (Fig. 8C). To validate these RT-LTR insertion loci, 30 insertion loci were detected via PCR, which amplify partial LTR sequence and partial nearby genomic sequence. Primers were designed for LTR and flanking sequences in ten greater yam germplasm (Fig. 8 and Figure S2). Among these insertion loci, 14 of them had expected amplicon lengths and dominant single bands, such as LTR-7069, LTR-3455, and LTR-179, which indicated polymorphic insertion variations between greater yam accessions (Fig. 9 and Figure S1). For the ten accessions used, present/absent variations for LTR-RT insertion ranged from 3/10 (LTR-3455) accessions with a present band to 10/10 present (LTR-6369 and five other loci). For the rest of the loci, two to three bands were detected, including the dominant expected amplicons, which also demonstrated more variations with divergent bands pattern for all greater yam accessions.

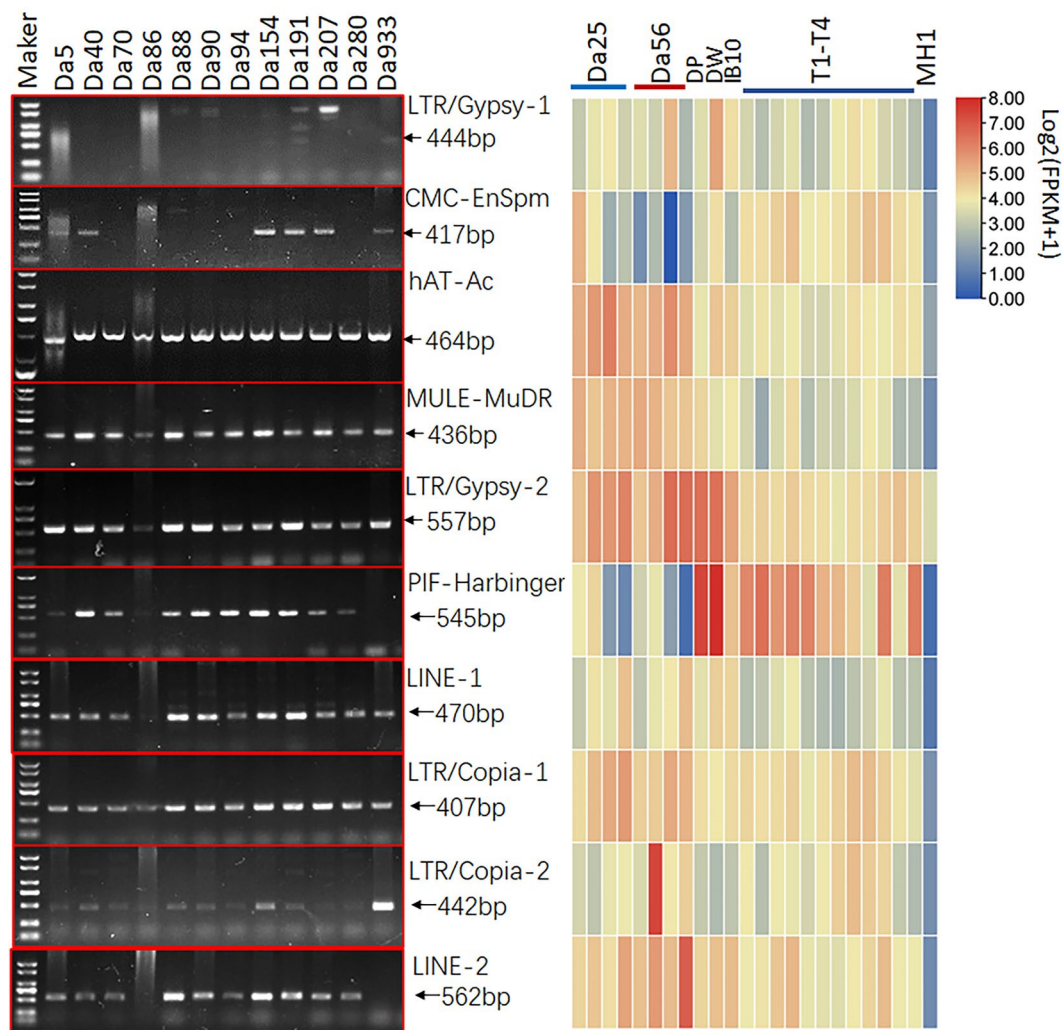


Fig. 5 RT-PCR detection (left) and an expression heatmap (right) of twelve TE transcripts. The detailed information of the transcriptome datasets used in the image is the same as in Table 1. The marker used in the image is DL2000. The original gel images were listed in Supplementary Dataset File 3

Discussion

TEs are critical force to drive genomic diversification. Many *de novo* genome sequencing and re-sequencing data has showed those genetic elements account for dramatic proportion in genome and highly variable between different individuals [17, 19, 21, 35–38]. In this study, we annotated and characterized TE-related transcripts in greater yam transcriptome via *de novo* assembly, explored their genomic location features and detected TIPs caused by active TEs. The active TEs and TIPs were compared between greater yam accessions to see their conserveness and divergence. Overall, our results shed light on the transcription features of active TEs and how they contributed to the expansion of present LTR-RT subfamilies, as well as their contribution to new TE insertions.

As we known, the vast majority of TEs in genomes are tightly surveilled and kept silenced, but still a few TEs loci or certain TE families are kept active [13, 39, 40]. In this study, the transcribed TEs accounts for 0.8–0.9% transcripts in transcriptome, while their inactive status or no extensive new insertion were partially assured by large proportion of truncated or noncoding transcript sequences: only 1–3% LTR-RTs transcripts have full-length polyprotein needed for transposition (Fig. 1C). Moreover, more than 90% TE transcripts had low expression levels (FPKM < 1), which were similar ratio as for long noncoding RNAs [41]. Therefore, only a few TE transcripts could encode autonomous TE protein and had moderate or high expression levels. The mapped TE transcripts were enriched in the gene scanty regions

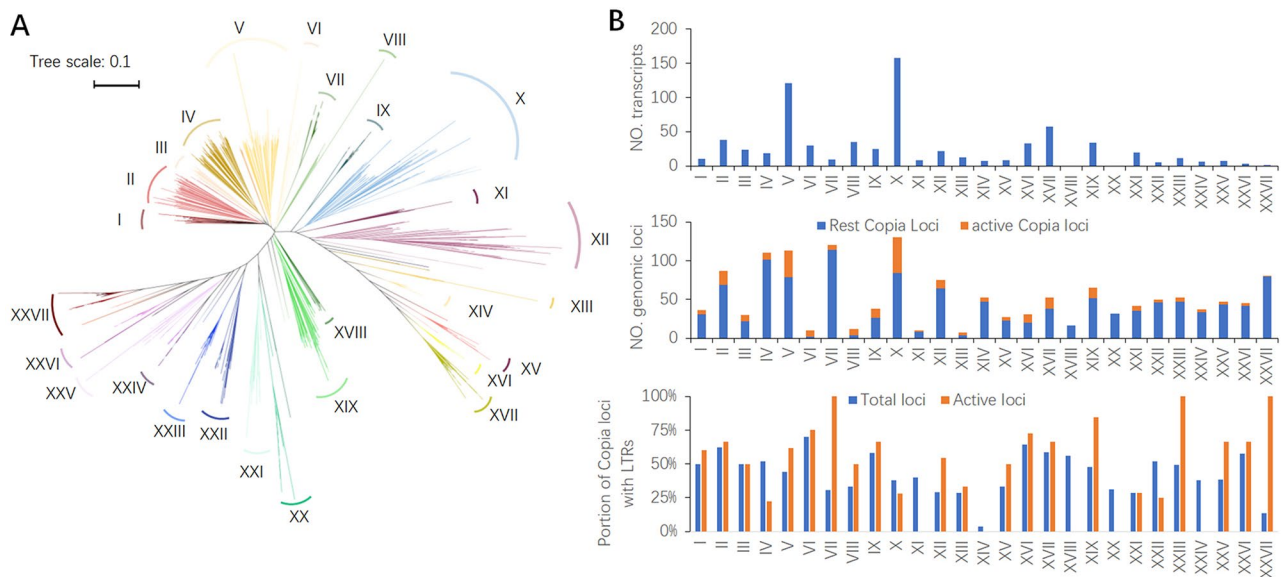


Fig. 6 Phylogenetic analysis of the *Copia* family and the classification of *Copia* transcripts, genomic loci. **(A)** The phylogenetic tree was constructed based on the RT domain peptide for *Copia* TEs; **(B)** subfamilies of *Copia* transcripts (UP), genomic loci (Mid), and the proportion of *Copia* loci with LTRs (Bottom)

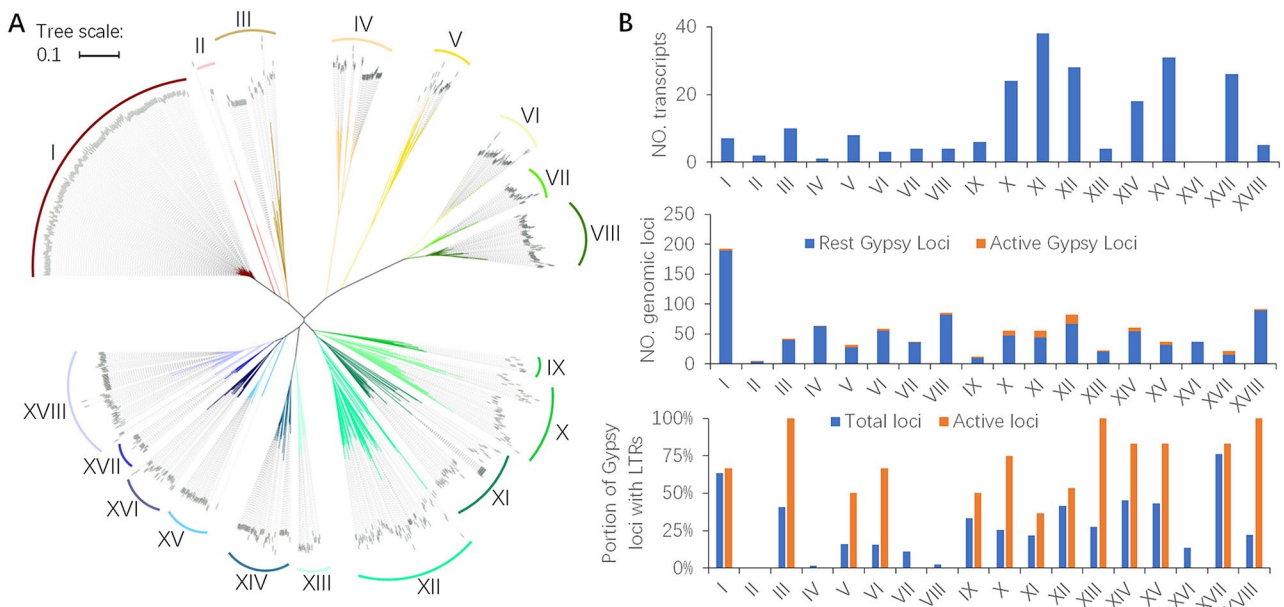


Fig. 7 Phylogenetic analysis of the *Gypsy* family and the classification of *Gypsy* transcripts, genomic loci. **(A)** Phylogenetic tree constructed based on RT domain peptide for *Gypsy* TEs; **(B)** subfamilies of *Gypsy* transcripts (UP), genomic loci (Mid), and the proportion of *Gypsy* loci with LTRs (Bottom)

- pericentromeric and centromeric regions where are usually in silencing chromatin condition, which were also reported in greater yam genome sequence that TEs are strongly anticorrelated with gene distribution [1]. All these results suggested that TEs transcription is low frequency but ongoing. Despite the robustness of our mapping strategy of TE transcript used in this study, our analysis did not attempt to detect the precise

transcription loci of these transcripts, which limits our ability to distinguish between active and inactive copies of the same retroelements. To address these limitations, future research should incorporate complementary methodologies to achieve a more comprehensive understanding of TE activity, such as long-read sequencing, to precisely map transcription loci and confirm the activity status of TEs.

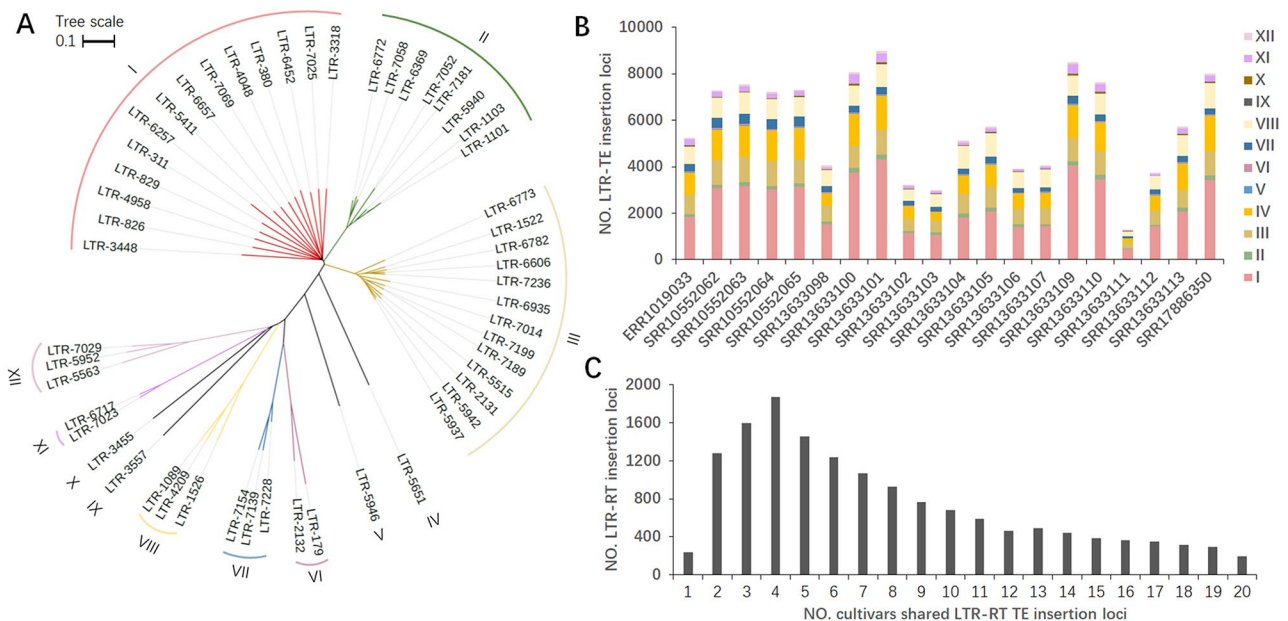


Fig. 8 Phylogenetic analysis of LTR sequences used for TIP analysis and the characteristics of TIPs identified based WGS data. **(A)** The phylogenetic tree was constructed based on LTR sequences from active TE loci; **(B)** The number and subfamily composition of TIPs identified in twenty greater yam samples with WGS data (Detailed information listed in Table S1b); **(C)** Shared TIP loci among the twenty greater yam samples

TE insertion within gene coding region or regulatory region will have sever effects on gene function, such disrupting coding sequence, provision of an alternative promoter, influencing alternative splicing events [7]. Among our mapped TE transcripts, 9% were overlapped within gene region and most were not inserted in the exon regions, with those overlapped exons only partial sequences were detected. Because we do not have the transcriptome for the greater yam that generated the reference genome, it is hard to determine the influence on the transcription status of these genes.

The annotated transcripts covered every main class of TE, while LTR-RTs (*Copia* and *Gypsy*) are the dominant types and have identical long terminal repeat sequences for new insertions. There are abundant studies showed active loci for LTR-RTs, such as in *Citrus limon*, strawberry, and sunflower [42–44]. The active TEs from *Copia* and *Gypsy* were preferentially clustered in certain subfamilies that usually had more TE members and a higher proportion of the LTR existence (Figs. 6 and 7). These results support the idea that TE transpositions are ongoing, but with a preference for active TE loci that may have more chances.

TE transposition could happen in both somatic and germ cells, leading to genome mutation. Only half of TE

transcripts were mapped to the greater yam reference genome, suggesting a high proportion of TE transcripts had restricted to certain yam accessions. This would bring ample intraspecies polymorphism. TE insertions polymorphisms (TIPs) analysis for 3,000 rice accessions found more than 50,000 TIP loci and most polymorphisms are found at very low frequency [20]. For WGS data of 602 tomato accessions, 6,906 TIPs were detected and disproportionately located within or adjacent to genes involved in environmental responses, most of which are low frequency variants [21]. When we analyzed 32 active LTR-RTs for TIPs analysis in this study, and detected 15,002 insertion loci, indicating that active TEs could contribute significantly to genome variation. The sample number with highest shared TIPs were four. This was due to the WGS dataset used containing same greater yam accession that were re-sequenced four times (Table S1b). Further PCR validation also confirmed that the results of TIP analysis are reliable and that some LTR insertion loci were conserved among different greater yam accessions. TIP is a form of structural variation, which could add genotypic information for genome wide association analysis. TIPs that overlap with genic regions would be easier to deduce their influence.

Table 3 The NO. shared transposon insertion loci between greater yam cultivars

	LTR-RT sub-groups											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1	138	2	17	48	0	1	9	14	0	0	7	4
2	842	14	50	265	1	1	13	30	0	3	29	28
3	1004	18	87	334	0	9	18	62	0	3	43	18
4	1128	37	127	389	1	7	35	60	1	15	61	10
5	861	15	127	254	1	6	50	73	0	10	48	8
6	655	21	126	223	1	4	42	91	0	13	50	14
7	502	21	148	197	3	5	36	92	1	8	46	8
8	431	22	108	160	0	5	40	104	0	9	40	10
9	314	20	79	156	2	4	46	91	0	7	40	7
10	266	16	109	123	2	2	31	84	0	5	30	10
11	232	18	75	110	0	4	26	78	0	6	33	9
12	162	13	78	93	1	3	22	58	0	6	16	8
13	179	18	59	89	1	7	24	86	0	4	19	4
14	141	21	69	78	0	2	32	78	1	1	11	10
15	145	8	60	63	3	4	22	62	0	0	10	5
16	117	13	60	65	1	3	29	57	0	3	14	4
17	116	10	54	65	4	0	13	73	2	2	6	3
18	91	8	73	34	6	1	18	66	3	0	9	7
19	46	9	78	19	2	4	24	94	4	1	7	3
20	17	4	72	3	0	6	25	69	0	0	0	2
Total	7387	308	1656	2768	29	78	555	1422	12	96	519	172

^aRepresent NO. greater yam accessions (resequencing samples) shared same LTR-RT insertion loci

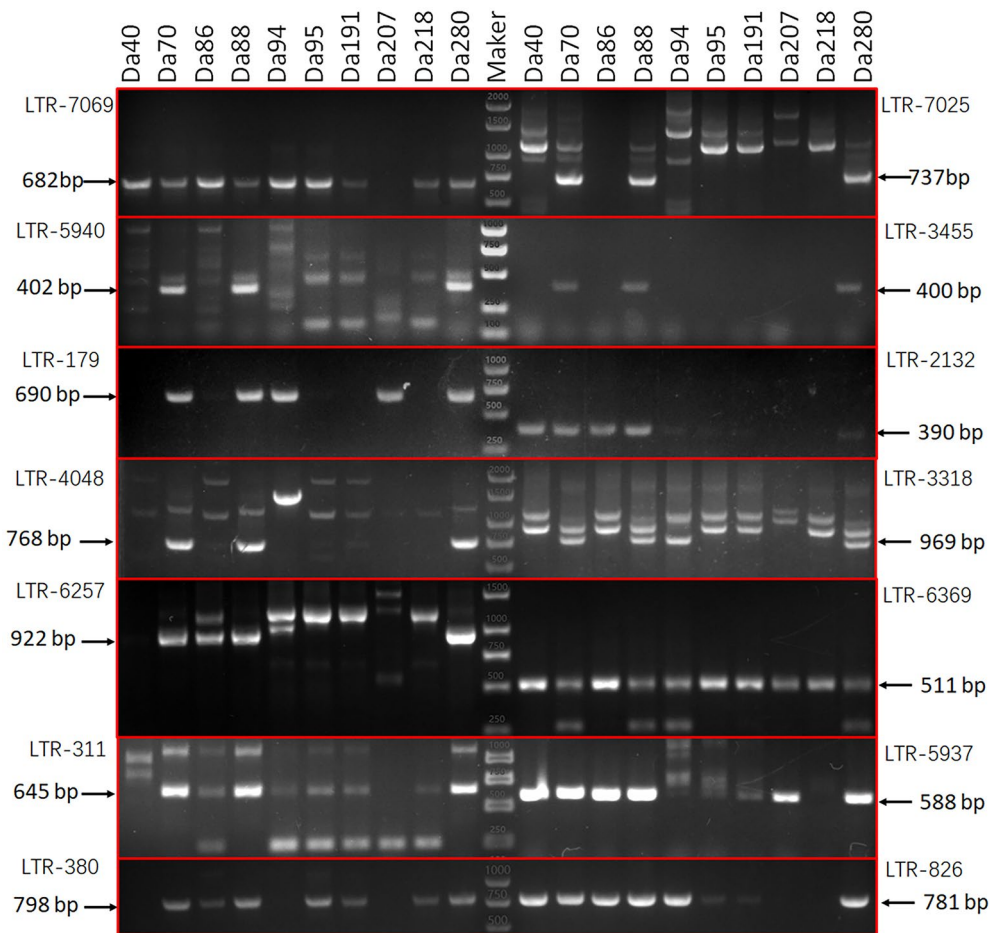


Fig. 9 The partial PCR validation results for TIPs. The pairs of primers were designed based on the LTR sequence and flanking sequence of the inserted TEs, and were listed in Table 2b. The marker used in the image is DL2000. The original gel images were listed in Supplementary Dataset File 3

Conclusions

The cultivated greater yam is usually vegetatively propagated by tubers, while its flowering ability is significantly hampered. Meanwhile, the transcription of TEs is ongoing in cells, which could lead to continuous genetic variations. Active TEs account for 0.8 – 0.9% transcripts, most of which have low expression and incomplete protein for autonomous transposition. However, the active TEs contribute to new TE insertions and cause TIPs among greater yam accessions. Although most TEs were located in intergenic regions, there are some overlapped with gene regions. potentially influencing gene functions. Ongoing TE insertions are occurring happened in greater yam, leading to abundant TIPs, which could bring continuous genetic variation to this tuber crop and may result in phenotypic changes.

Abbreviations

- TE Transposable elements
- TIPs TE insertion polymorphism
- PCR Polymerase chain reaction
- LTR-RT Long terminal repeat retrotransposon

- WGS Whole genome sequencing
- SVs Structural variations

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10779-0>.

Supplementary Material 1: Table S1a Information on transcriptome datasets used for TE related transcripts detection. Table S1b Information on wholegenomic resequencing dataset for greater yam accessions.

Supplementary Material 2: Table S2a Primers used for RT-PCR to validate active TEs. Table S2b primers used for TIP locidetection.

Supplementary Material 3: Table S3 active TEs that have been mapped to the greater yam reference genome.

Supplementary Material 4: Table S4 distribution of LTR-RTs insertion loci across greater yam chromosomes.

Supplementary Material 5: Figure S1 PCR validation results for 16 TIPs. The original gel images were listed as TIPs-1 to TIPs-13 in supplementary data.

Supplementary Material 6: Original gel images used in this study.

Acknowledgements

Thank you very much for Changmian Ji from Chinese Academy of Tropical Agricultural Sciences for the technical help for TE analysis. Thank you for the

data analysis process supported by High-Performance Computing Platform of YZBSTACC.

Author contributions

D.W., H.L., Y.H., and M.L. did the experiment. Y.W., W.W., and R.C. prepared the samples. W.X. and Y.X. wrote the manuscript. W.X., S.P. and D.W. participated in data analysis. W.X., D.H., and Y.X. participated in the design of the study and drafted the manuscript.

Funding

The work was supported by the National Natural Science Foundation of China (NO. 32160470 and NO. 31760406), as well as 2019 Hainan basic and applied basic research program (Natural Science Field) high level talent project (2019RC157).

Data availability

The transcriptome and WGS data were downloaded from the national center for biotechnology information website, whose accession numbers were listed in Table S1.

Declarations

Ethics approval and consent to participate

NOT applicable.

Consent for publication

NOT applicable.

Competing interests

The authors declare no competing interests.

Author details

¹School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication), School of Tropical Agriculture and Forestry, Hainan University, 570228 Haikou, P.R. China

²School of life sciences, Hainan University, 570228 Haikou, P.R. China

Received: 25 February 2024 / Accepted: 4 September 2024

Published online: 16 September 2024

References

- Bredeson JV, Lyons JB, Oniyinde IO, Okereke NR, Kolade O, Nnabue I, Nwadiili CO, Hřibova E, Parker M, Nwogha J, et al. Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat Commun*. 2022;13(1):2001.
- Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet*. 2013;14(12):840–52.
- Akakpo R, Scarcelli N, Chair H, Dansi A, Djedatin G, Thuillet AC, Rhone B, Francois O, Alix K, Vigouroux Y. Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. *BMC Genomics*. 2017;18(1):017–4143.
- Scarcelli N, Cubry P, Akakpo R, Thuillet A-C, Obidiegwu J, Baco MN, Otoo E, Sonke B, Dansi A, Djedatin G et al. Yam genomics supports West Africa as a major cradle of crop domestication. In: *Sci Adv*. vol. 5; 2019: eaaw1947.
- Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet*. 2023;55(7):1221–31.
- Wang S, Xiao Y, Zhou Z-W, Yuan J, Guo H, Yang Z, Yang J, Sun P, Sun L, Deng Y, et al. High-quality reference genome sequences of two coconut cultivars provide insights into evolution of monocot chromosomes and differentiation of fiber content and plant height. *Genome Biol*. 2021;22(1):304.
- Cowley M, Oakey RJ. Transposable elements re-wire and fine-tune the transcriptome. *PLOS Genet*. 2013;9(1):e1003234.
- De Kort H, Legrand S, Honnay O, Buckley J. Transposable elements maintain genome-wide heterozygosity in inbred populations. *Nat Commun*. 2022;13(1):7022.
- Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA, Martienssen RA. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*. 2009;136(3):461–72.
- Osakabe A, Jamge B, Axelsson E, Montgomery SA, Akimcheva S, Kuehn AL, Pisupati R, Lorkovic ZJ, Yelagandula R, Kakutani T, et al. The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. *Nat Cell Biol*. 2021;23(4):391–400.
- Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. *Curr Opin Plant Biol*. 2015;27:67–76.
- Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. An active DNA transposon family in rice. *Nature*. 2003;421(6919):163–7.
- Cao Y, Jiang Y, Ding M, He S, Zhang H, Lin L, Rong J. Molecular characterization of a transcriptionally active *Ty1/copia*-like retrotransposon in *Gossypium*. *Plant Cell Rep*. 2015;34(6):1037–47.
- Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet*. 2012;46(1):651–75.
- Li W, Prazak L, Chatterjee N, Gruningner S, Krug L, Theodorou D, Dubnau J. Activation of transposable elements during aging and neuronal decline in *Drosophila*. *Nat Neurosci*. 2013;16(5):529–31.
- Sun L, Jing Y, Liu X, Li Q, Xue Z, Cheng Z, Wang D, He H, Qian W. Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nat Commun*. 2020;11(1):1886.
- Cai X, Lin R, Liang J, King GJ, Wu J, Wang X. Transposable element insertion: a hidden major source of domesticated phenotypic variation in Brassica rapa. *Plant Biotechnol J*. 2022;20(7):1298–310.
- Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, et al. Landscape of somatic retrotransposition in human cancers. *Science*. 2012;337(6097):967–71.
- Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, Santamarina M, Ju YS, Temes J, Garcia-Souto D, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet*. 2020;52(3):306–19.
- Carpentier M-C, Manfroi E, Wei F-J, Wu H-P, Lasserre E, Llauro C, Debladis E, Akakpo R, Hsing Y-I, Panaud O. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat Commun*. 2019;10(1):24.
- Dominguez M, Dugas E, Benchouaia M, Leduque B, Jimenez-Gomez JM, Colot V, Quadrana L. The impact of transposable elements on tomato diversity. *Nat Commun*. 2020;11(1):4058.
- He J, Babarinde IA, Sun L, Xu S, Chen R, Shi J, Wei Y, Li Y, Ma G, Zhuang Q, et al. Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nat Commun*. 2021;12(1):1456.
- Rodriguez-Quiroz R, Valdebenito-Maturana B. SoloTE for improved analysis of transposable elements in single-cell RNA-Seq data using locus-specific expression. *Commun Biol*. 2022;5(1):022–04020.
- Xiao Y, Yang Y, Cao H, Fan H, Ma Z, Lei X, Manson AS, Xia Z, Huang X. Efficient isolation of high quality RNA from tropical palms for RNA-seq analysis. *Plant Omics*. 2012;5(6):584–9.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
- Haas B, Papanicolaou A, Yassour M, Regev A. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. 2014.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8(12):973–82.
- Perteau M, Perteau GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5.
- Perteau M, Kim D, Perteau GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;11(9):1650–67.
- Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60.
- Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics viewer (IGV). *Bioinformatics* 2022, 39(1).
- Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. MeV: MultiExperiment Viewer. In: *Biomedical Informatics for Cancer Research*. Edited by Ochs MF, Casagrande JT, Davuluri RV. Boston, MA: Springer US; 2010: 267–277.
- Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;35(suppl2):W265–8.

34. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–8.
35. Platt RN, Vandeweghe MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res*. 2018;26(1):25–43.
36. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, Barbazuk WB, Bass HW, Baruch K, Ben-Zvi G, et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nat Genet*. 2018;50(9):1282–8.
37. Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, Park PJ. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*. 2021;12(1):3836.
38. Yao Y, Maddamsetti R, Weiss A, Ha Y, Wang T, Wang S, You L. Intra- and inter-population transposition of mobile genetic elements driven by antibiotic selection. *Nat Ecol Evol*. 2022;6(5):555–64.
39. Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature*. 2018;553(7687):228–32.
40. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet*. 2017;49(10):1502–10.
41. Xia W, Dou Y, Liu R, Gong S, Huang D, Fan H, Xiao Y. Genome-wide discovery and characterization of long noncoding RNAs in African oil palm (*Elaeis guineensis* Jacq.). *PeerJ* 2020, 2(8).
42. De Felice B, Wilson RR, Argenziano C, Kafantaris I, Conicella C. A transcriptionally active *copia*-like retroelement in *Citrus limon*. *Cell Mol Biol Lett*. 2009;14(2):289–304.
43. He P, Ma Y, Zhao G, Dai H, Li H, Chang L, Zhang Z. *FaRE1*: a transcriptionally active *Ty1-copia* retrotransposon in strawberry. *J Plant Res*. 2010;123(5):707–14.
44. Qiu F, Ungerer MC. Genomic abundance and transcriptional activity of diverse *gypsy* and *copia* long terminal repeat retrotransposons in three wild sunflower species. *BMC Plant Biol*. 2018;18(1):6.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.