Comprehensive analysis of both long and short read transcriptomes of a clonal and a seed-propagated model plant species reveal the prerequisites for transcriptional activation of autonomous and non-autonomous transposons in plants

Ting-Hsuan Chen
Plant and Food Research Ltd: New Zealand Institute for Plant and Food Research Ltd
https://orcid.org/0000-0002-7311-8064
Christopher Winefield (✉ **Chris.winefield@lincoln.ac.nz**)
Lincoln University https://orcid.org/0000-0002-6262-6370

Research

Keywords: Transposable element, Stress treatment, Transcription, Differential expression, Alternative splicing, Nanopore sequencing

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-859670/v1

Abstract

Background

Transposable element (TE) transcription is a precursor to its mobilisation in host genomes. However, the characteristics of expressed TE loci, the identification of self-competent transposon loci contributing to new insertions, and the genomic conditions permitting their mobilisation remain largely unknown.

Results

Using *Vitis vinifera* embryogenic callus, we explored the impact of biotic stressors on transposon transcription through the exposure of the callus to live cultures of an endemic grapevine yeast, *Hanseniaspora uvarum*. We found that only 1.7%-2.5% of total annotated TE loci were transcribed, of which 5%-10% of these were full-length, and the expressed TE loci exhibited a strong location bias towards expressed genes. These tendencies were also observed in *Arabidopsis thaliana* wild-type and *ibm2*, but not in *ddm1* mutant. Moreover, differentially expressed TE loci in the grapevine model tended to share expression patterns with co-localised differentially expressed genes. Utilising nanopore cDNA sequencing, we found a strong correlation between the inclusion of intronic TEs in gene transcripts and the presence of premature termination codons in these transcripts. Finally, we identified low levels of full-length transcripts deriving from structurally autonomous TE loci in the grapevine model.

Conclusion

Our observations in two disparate plant models representing clonally and seed propagated plant species reveal a closely connected transcriptional relationship between TEs and co-localised genes, particularly when the epigenetic silencing system is intact. We found that the stress treatment alone was insufficient to induce large-scale full-length transcription from structurally competent TE loci, a necessity for non-autonomous and autonomous mobilisation.

Background

A substantial proportion of eukaryotic genomes are comprised of Transposable elements (TEs) [1, 2]. These elements have long been considered unavoidable deleterious components of host genomes, being described variously as parasitic or junk DNA. However, there has been a recent resurgence in interest in the roles of TEs, in particular those associated with genome and organism evolution [2, 4, 9–12], as well as environmental adaptation [3, 13–15].

The first committed step in the life cycle of a transposon is the transcription of a fully competent, or autonomous, copy of a given TE. Competent de novo transcription of these TEs is an absolute requirement for TE mobilisation irrespective of TE types. Autonomous transposition of type I retrotransposons is dependent on reverse transcription of full-length transcripts derived from intact retrotransposon loci, while self-competent mobilisation of type II DNA transposons requires transcription of the element open-reading frame (ORF) encoding a functional transposase enzyme that facilitates the transposase-mediated excision of elements from competent genomic loci [10, 16]. Transposition of non-autonomous elements cannot be achieved without transcription of the associated autonomous element within the genome. Independent of TE transposition, transcription of TE loci, containing either intact or fragmented TEs, can participate in aberrant transcription or epigenetic regulation of neighbouring or host genes without generating new TE insertions [17, 18].

In eukaryotes, host genomes have evolved epigenetic regulatory networks that are targeted primarily at the suppression of TE activity through silencing of TE loci and transcripts with small interfering RNAs (siRNAs) and deposition of dense epigenetic hallmarks, including methylcytosine and suppressive histone modification (e.g. di- and tri-methylated histone H3 at lysine 9), onto TE loci [6–8]. Consequently, the complex interplay between initiation of TE transcription and epigenetic suppression sets the scene for transposon mobility and the consequential impact this mobility inevitably has on genome biology.

Despite the existence of the epigenetic silencing networks, TEs mobilise and contribute to genetic diversity naturally through evolutionary time. However, it is unclear how TEs escape from the silencing system to propagate themselves in the wild-type host genomes. Reactivation of TE transcription and mobilisation has been observed in plant mutants impaired in epigenetic silencing [19–21] and plant epigenetic recombinant inbred lines (epiRIL) [14, 22]. In wild-type backgrounds that are not impaired in epigenetic silencing, it has been widely reported that the establishment of plant tissue culture or stress treatments can result in transcriptional activation and mobilisation of TEs [23–27]. Investigation of the location of new TE insertions reveals the insertion bias of low-copy-number DNA transposons and LTR-retrotransposons avoiding heterochromatic DNA or a bias towards the gene-rich region in maise [28, 29], rice [30–32], and *Arabidopsis thaliana* [14]. Nonetheless, the location and molecular regulation of the autonomous TE loci seeding autonomous and non-autonomous mobilisation often remain elusive.

Due to the highly repetitive nature of TE sequence in genomes, most literature on TE transcriptional regulation reports this transcriptional activity at the family level [17]. Thus very little is known about the number of autonomous TE loci that contribute to the pool of TE transcripts and what proportion of measured TE transcription is contributed from the full-length transcription of these autonomous loci. This inability to effectively characterise transcriptionally active individual TE loci hampers the understanding of the determinants for their transcriptional activation and, therefore, regulation of the mobilisation of these elements and the non-autonomous partner loci throughout the genome.

To address these questions, we used *Vitis vinifera* embryogenic callus as a model system for clonally propagated crops to thoroughly interrogate the impact of exposure of these cultures to biotic stressors and to investigate the impact of such stress on the transcription of TE loci in these cultures. The biotic stressor we used was the exposure of the callus to live *Hanseniaspora uvarum*, a yeast that is commonly found in vineyards [33–35] and has previously been shown to activate TE transcription in embryogenic callus [23]. To validate our findings in grapevine, we also determined the profiles of expressed TE loci in Arabidopsis using published *A. thaliana* RNA-sequencing (RNAseq) datasets, including wild-type [36], *ibm2* [36], and *ddm1* mutants [19]. The *ddm1* mutant is impaired in epigenetic silencing of heterochromatic TEs [20] and shows high levels of TE mobilisation. The RNAseq data of the *ibm2* mutant was included as a negative control as *IBM2* is required to prevent inappropriate polyadenylation of transcripts derived from genes possessing intronic TEs [37–40], rather than directly driving transcriptional silencing of intronic TEs [36, 41].

To circumvent the commonly encountered issues with the mapping of short-read RNA-seq datasets to repetitive sequences, we utilised multiple existing tools to collectively identify 'TE expression candidates', which are determined as annotated TE loci that were mappable by unique- or multi-mapping short sequencing reads and thus potentially expressed (see Methods). Following the identification of TE expression candidates, the genome-wide landscape and properties of transcriptionally activated TE loci were investigated. We show that the transcriptional activity of TE expression candidates was highly associated with expressed genes in terms of the location distribution and time-series expression changes of TE expression candidate when the epigenetic system is not compromised. We finally employed long-read Oxford Nanopore cDNA sequencing to determine the fine structure of TE-derived transcripts in the grapevine RNA samples, with a view to detecting full-length transcription of autonomous TE loci.

Results

Most TE expression candidates are fragmented and polymorphic TE loci

Following the system established by Lizamore [23] to activate TE transcription, *V. vinifera* (Pinot noir UCD5 clone) embryogenic callus cultures were subjected to a mock treatment (hereinafter Vv_Mock; which included vigorous shaking that was expected to mount a wounding response in this tissue; see Methods) or mock treatment plus *H. uvarum* live cultures (hereinafter Vv_Yeast), and samples were collected at four time points after exposure to the yeast culture (Fig. 1a). Untreated embryogenic callus culture was included as a common time zero (Vv_T = 0) for both Vv_Mock and Vv_Yeast treatments. To comprehensively survey the transcriptional activity of TEs in this system, we generated RNA-seq data for each sample. The short sequencing reads of stranded polyadenylated transcriptome data were analysed with a computational workflow that includes multiple existing software to incorporate unique- and multi-mapping reads for identification of potentially expressed TE loci, which were denoted as 'expression candidates' (Fig. 1a; Additional file 1: Table S1; Additional file 2: Figure S1). In addition, publicly accessible *A. thaliana* RNAseq data of wild-type (Col) [36], *ibm2* [36], and *ddm1* [19] were also analysed using the same computational workflow and denoted as At_WT, At_ibm2, and At_ddm1, respectively (Fig. 1a; Additional file 1: Table S1; Additional file 2: Figure S1). This analysis workflow was consisting of three parts that differed in the software (including HTSeq [42], Bedtools [43] and TEFingerprint [44]) used to quantify the sequencing reads mapping to individual TE loci, as well as in the characteristics of the quantified reads (Fig. 1a; Additional file 2: Figure S1; see Methods for details). The three sets of TE loci passing corresponding expression threshold (see Methods) were united as a pool of expression candidates (Fig. 1a; Additional file 2: Figure S1A, B), in which those collected by htseq-count or TEFingerprint were denoted as 'trackable' expression candidates as their expression pattern was able to be tracked by virtue of unique-mapping reads, whereas those exclusively collected by Bedtools were denoted as 'un-trackable' expression candidates as the multi-mapping reads did not have suitable sequence polymorphisms to allow unambiguous mapping to individual TE-loci. Thus those expression candidates that are deemed to be trackable represent those transcribed TE loci with accumulated unique polymorphisms, such as single nucleotide variants (SNVs) or small insertions or deletions (INDELs). By contrast, un-trackable expression candidates represent a group of highly conserved TE loci, which are likely to be younger than trackable expression candidates and might include recently mobilised TE loci.

In the *V. vinifera* system, our analysis workflow identified 3,698 (1.6%), 5,524 (2.5%), and 5,531 (2.5%) of a total of 223,411 annotated TE loci as expression candidates in Vv_T = 0, Vv_Mock, and Vv_Yeast, respectively (Fig. 1b). Of the total number of loci containing TE sequence, 75%-87% were effectively excluded as having no evidence of expression and a further 11%-22% of the total annotated TE loci were excluded due to insufficient support in terms of low numbers of mapped reads and thus fell under threshold (Additional file 2: Figure S2). In each of the tested conditions (Vv_T = 0, Vv_Mock and Vv_Yeast), different sets of expression candidates were identified (Fig. 1c). The increase in the number of expression candidates uniquely found in Vv_Mock (1,139 TE loci; Fig. 1c) and Vv_Yeast (1,913 TE loci; Fig. 1c) comparing to Vv_T = 0 (329 TE loci; Fig. 1c) indicates the transcriptional activation of TE activity due to the wound and biotic stress treatments, respectively.

Similarly, in At_WT and At_ibm2, over 95% of the total 31,189 annotated *Arabidopsis* TE loci were excluded from the pool of expression candidates, leaving 1,410 and 1,342 TE loci identified as expression candidates in these two genotypes, respectively (Fig. 1d). However, in At_ddm1, the number of expression candidates was increased to 4,156 TE loci (Fig. 1d), in which 3,091 TE loci were uniquely present in the pool of expression candidates of At_ddm1 (Fig. 1e), most likely reflecting the deficiency of *ddm1* mutant in TE silencing [20].

Depending on the experimental condition in the *V. vinifera* system and genotypes in the *A. thaliana* system, 69.2%-77.4% and 87.1%-94.9% of the expression candidates, respectively, showed evidence of transcription that unique-mapping reads can represent, thus were identified as trackable expression candidates in these two plant systems (Fig. 1b, d). Only 9%-10% of the expression candidates in *V. vinifera* system and 11%-17% of the expression candidates in the *A. thaliana* system retained > 90% length integrity (hereinafter 'full-length' loci) as compared with the length of canonical elements (Fig. 1f, g). The high proportion of trackable and fragmented TE loci in the pools of expression candidates in both *V. vinifera* and *A. thaliana* systems indicate that most of the expressed TE loci are both divergent sequences and degenerated in length.

The characteristics of TE expression candidates varies by TE family

We were interested in determining whether the degree of mutation accumulation within expressed TE families varied depending on each family's mobilisation history. TE loci of more recently mobile families (i.e. younger TE families) are expected to possess a high degree of sequence conservation compared to those TE loci belonging to families that have more ancient mobilisation events (i.e. older TE families) [45, 46]. These observations suggest that younger TE families may not necessarily be transcribed to a higher level than older TE families, but younger TE families may possess a higher proportion of full-length un-trackable expression candidate loci than older TE families.

To test this assumption, we grouped grapevine expression candidates hierarchically by families, degree of polymorphism (i.e. trackable or un-trackable), and length integrity (i.e. fragmented or full-length). For Vv_T = 0, 2,565 (69%) of the total 3,698 expression candidates were trackable, whereas the remaining 1,280 expression candidates (un-trackable) remained indistinguishable (Fig. 2a). Among the total 232 TE families (corresponding to nine superfamilies) presented on the y-axis of Fig. 2a, 174 families contained expression candidates in Vv_T = 0. Among these expressed families, 102 contained fewer than ten expression candidate loci, and a further 32 families contained 10 to 23 expression candidate loci (the median abundances of expression candidates among expressed TE families; Additional file 3: Table S3), indicating that the bulk of the expressed TE families have few transcriptionally active TE loci. It was noticeable that 6 TE families had more than 100 expression candidate loci; these families were Copia-23 (211 expression candidates), Gypsy-12 (175 expression candidates), VLINE1 (245 expression candidates), VLINE4 (211 expression candidates), VLINE5 (117 expression candidates) and VLINE6 (139 expression candidates). However, of these 6 TE families, the expression candidate loci in 5 of the families were mostly fragmented and rackable TE sequences (Additional file 3: Table S3). In fact, half of the TE families containing transcribed loci (87 of the 174 expressed families) had zero un-trackable expression candidates, and a further 74 TE families had only 1 to 20 un-trackable expression candidates, most of which are fragmented. This observation demonstrates that the vast majority of the expressed TE families comprised transcriptionally active loci that are mostly fragmented and could be identified by polymorphisms (e.g. SNVs and INDELs). In contrast, Copia-23 and Copia-3 families were over-represented with expression candidates that were both un-trackable and potentially autonomous (full length). These findings were concordant with the observations in Vv_mock (Fig. 2b) and Vv_Yeast (Fig. 2c). Although more TE families were obtaining low numbers of full-length candidates in each of these two treatments than that in Vv_T = 0, most of the expressed families are still in short of un-trackable expression candidates, and Copia-3 and Copia-23 are still the two families comprising most full-length un-trackable expression candidates (Additional file 3: Table S3).

A closer look at the sequences of the canonical and 90 annotated full-length Copia-3 elements shows a condensed phylogenetic cluster mostly comprised with intact Copia-3 indicated by the presence of LTRs flanking INT domain (i.e. structurally autonomous; Additional file 2: Figure S3; also see Methods). This cluster included 26 sequences, 19 of which were structurally autonomous un-trackable candidates with over 90% read coverage of the annotated INT domain, and 4 of which were intact trackable Copia-3 with a nearly complete transcription of INT (see Methods). The majority of the remaining un-trackable expression candidates formed three groups, the largest two close to the cluster formed by intact Copia-3 loci. The opposite distal end of the tree was occupied mainly by un-expressed full-length Copia-3 loci.

The neighbour-joining tree built from the canonical and 220 full-length Copia-23 sequences revealed that the 11 un-trackable and four trackable structurally autonomous candidates with nearly complete transcription across INT were scattered in 5 broom-like clusters. These clusters were densely packed with other un-trackable candidates that had either lost intact LTRs or lacked full INT coverage (Additional file 2: Figure S4). These compact clades with short branches were distinguished from the sequences of unexpressed full-length Copia-23.

To test the contribution of the full-length Copia-3 loci to the pool of Copia-3-related transcripts, reads mapping to all Copia-3 candidates were categorised into four groups by whether they mapped to full-length/fragmented and trackable/un-trackable candidates. This analysis revealed that each category contained reads shared with one or more categories, irrespective of treatments and treatment time-point (Additional file 2: Figure S5). Nonetheless, each category obtained a unique subset of reads that only mapped to one of the four groups (pairwise combination of full-length/fragmented and trackable/un-trackable) of expression candidates, meaning none of the group was able to represent the whole collection of Copia-3 transcripts. The same analysis was applied to reads mapping to all Copia-23 expression candidates. This analysis also demonstrated reads shared across different categories and those reads unique to a single category (Additional file 2: Figure S6).

Copia-3 and Copia-23 belong to LTR-retrotransposon (LTR-TEs). This type of TEs is known for the identical long terminal repeat (LTR) at both ends upon insertion. The pair of LTRs gradually accumulates independent mutations across time; therefore, it has been believed that the more diverse the two LTR sequences, the more time that has passed since insertion [46, 47]. To test whether Copia-3 and Copia-23 were active more recently than other LTR-TEs, we analysed the divergence of each pair of LTRs, following with the estimation of the insertion time of each structurally autonomous TE loci. The insertion dates of the 87 and 177 structurally autonomous loci of Copia-3 and Copia-23, respectively, were calculated based on the divergence of individual pairs of LTRs for each element. The peak of Copia-3 and Copia-23 mobilisation was then estimated from the distribution of insertion times and found to occur approximately 0.02 and 0.017 million years ago (MYA), respectively (Additional file 2: Figure S7A, B). Peak insertion times of the other 39 LTR-TE families with at least ten intact copies were analysed in the same way (Additional file 2: Figure S7C; Additional file 4: Table S3). Most LTR-TE families experienced bursts no longer than 4.5 million years ago (MYA). Note that Copia-3 and Copia-23 were the most recently active LTR-TE families (Additional file 2: Figure S7C). Comparison of the peak insertion time of Copia-3, Copia-23 and the other 5 Copia families, which obtained trackable full-length candidates across all treatments but lacked un-trackable full-length candidates, showed that Copia-3 and Copia-23 experienced significantly more recent bursts than these five other families (Additional file 2: Figure S7D).

Together, these results suggest that LTR-TE families that experienced the most recent mobilisation burst in evolutionary time may not necessarily the TE families having the most number of expression candidate loci. Instead, they may tend to have structurally autonomous expression candidates involving in ambiguous alignment and lacking polymorphisms to facilitate the identification of the exact origin of transcription among these highly conserved loci.

TE expression candidates are not randomly distributed in the genome

A number of recent reports highlight that TE mobilisation and reintegration into the host genome often show distinct insertion bias [14, 28, 29, 30–32]. However, little is known regarding the genome-wide distribution of transcriptionally active TE loci. In order to investigate whether there is location bias between all annotated TEs and expression candidates, our analysis compartmented the annotated reference genome into genic and intergenic regions. The genic region comprised gene units, which were made of exons and introns included from the transcription start sites to the transcription stop sites of genes, and flanking regions of genes which encompassed 2kb upstream (N-flanks) and 2kb downstream (C-flanks) of corresponding translation start and stop sites (Fig. 3a). All annotated TEs intersected with specific genome compartments were categorised accordingly and hierarchically in the order of genic/intergenic regions, location within the genic region (e.g. exon, intron, flanks), and integrity (full-length or fragmented). In the grapevine system, over half of all annotated TE loci fell into intergenic regions (126,976 TEs, 56.83%), while 96,435 (43.16%) TEs co-localised with genes (Fig. 3b). About half of the annotated genic TEs of *V. vinifera* were in flanking regions, without particular preference for either side (N-flank or C-flank). As expected, intronic TEs comprised most TEs in gene units (Fig. 3b; Additional file 5: Table S4).

Expression candidates of our *V. vinifera* system were classified in the same way, with additional categories added, including the transcriptional activity of co-localised genes (i.e. TEs associated with expressed or non-expressed genes) and the presence or absence of unique-mapping reads (trackable or un-trackable). Untreated embryogenic callus (Vv_T = 0), Vv_mock, and Vv_Yeast respectively showed 71.47%, 75.69%, and 74.62% of the expression candidates were in the genic regions (Fig. 3c-e, Additional file 5: Table S4). The goodness of fit X-square test shows that the proportion of genic TEs was significantly elevated from 43% of the 'default' distribution of all annotated TEs to 71% of Vv_T = 0 expression candidates (see the comparison of 'All annotated' versus Vv_T = 0 in Fig. 3f). This distribution bias is further enhanced in Vv_mock and Vv_Yeast (Fig. 3f). Delving deeper into the insertion context, about two-thirds of these genic TE expression candidates overlapped with introns; in particular, there was a significant bias toward introns of expressed genes (Fig. 3c-e, g, h; Additional file 5: Table S4).

To examine whether the location bias of expression candidates of our *V. vinifera* system is also conserved in other plant species, the *A. thaliana* dataset was also analysed in the same way. There are 58.37% of *A. thaliana* annotated TE loci located in the genic region, while TEs in gene unit and flanking regions, respectively, comprised 16.00% and 42.37% of the total pool (Fig. 3i; Additional file 5: Table S4).

Remarkably, the majority of the TEs annotated within the 'gene unit' overlapped with exons (Fig. 3i), contributing to 12.42% of all annotated loci. This high proportion of exonic TE loci is because, based on the TE annotation established by Jin et al. [48], considerable numbers of long TEs overlap with multiple exons and introns, which were annotated based on the gene annotation file deposited in Ensembl Plants (see Methods), and therefore were preferentially categorised as exonic TEs.

In wild-type *Arabidopsis* (At_WT), TE loci located in the genic region comprised 91.28% of the 1,410 expression candidates, of which 1,093 loci co-localised with expressed genes (Fig. 3j; Additional file 5: Table S4). In addition, the proportion of gene-unit loci had skewed from 16% of total annotated TE loci (Fig. 3i) to 74.61% of the expression candidates in At_WT (Fig. 3j; Additional file 5: Table S4). The majority of these expression candidates in the gene unit were associated with expressed genes (65.67% to the total expression candidates). The expression candidates of At_ibm2 showed very similar location distribution to At_WT (Fig. 3k). However, while intergenic expression candidates comprised 8.72% of the expression candidate pool of At_WT, the intergenic proportion in At_ddm1 was significantly higher at 54.79% of expressed TE loci (Fig. 3l, m, Additional file 5: Table S4). As opposed to the high proportion of 'gene-unit' expression candidates in At_WT (74.61%), only 24.95% of the expression candidates in At_ddm1 were located in the gene unit. These results were supported by statistical tests (Fig. 3m-o) and show that the distribution of expression candidates in *Arabidopsis ibm2* mutant is highly similar to that in wild-type, whereas a striking difference was observed between *Arabidopsis* wild-type and *ddm1*.

The proposed preference for expressed TE loci to be located in genic regions may be explained by either a general increase in the proportion of genic expression candidates from most TE families or simply a reflection of the genic-enriched annotation of a few TE families that largely contribute to the pool of expression candidates. To test these two assumptions, we further delved into the location distribution of all annotated TE loci and expression candidates of individual TE families. The genic and intergenic proportions of annotated TE loci and expression candidates were plotted for families belonging to Copia, Gypsy, LINE, hAT and MULE, the five superfamilies that contributed to the majority of expression candidates (Additional file 2: Figure S8-S12). This analysis first looked at the genic and intergenic proportion of all annotated TE loci grouped by families and then categorised expression candidates, in the same way, to examine that whether the genic proportion of expression candidates is higher than that of annotated TE loci in most of the investigated TE families. For the genic and intergenic distribution of annotated TE loci belonging to Copia families, about two-thirds of the families show underrepresentation (< 50%) of genic TE loci (Additional file 2: Figure S6A). When it comes to expression candidates in Vv_T = 0, 58 of the 71 expressed families demonstrated higher genic proportions of expression candidates (Additional file 2: Figure S6B) than that of the annotated TE loci (Additional file 2: Figure S6A), meaning that the tendency of TE expression candidates to be located in the genic region is broadly presented in *V. vinifera* Copia families in Vv_T = 0. This trend in Copia families is also observed in Vv_Mock and Vv_Yeast (Additional file 2: Figure S6C-D). The analysis for *V. vinifera* TE families of Gypsy, LINE, hAT and MULE (Additional file 2: Figure S7-9) is concordant with the aforementioned findings in Copia: despite the different degrees of elevation in genic proportion, the distribution bias of expression candidates towards genic region seems to occur in most of the families broadly. The same analysis for the location preference within genic regions (exon, intron, and flanking regions) also reveals that the elevation of the intronic fraction of expression candidates is widely presented in most TE families of *V. vinifera* (Additional file 2: Figure S13-S17).

The distribution bias of TE expression candidates suggests the tolerance of transcription of TE loci within or proximal to expressed genes. We then asked whether this distribution bias only restricted to non-autonomous TE loci that are assumed to pose less stress of mutational load to host cells than the structurally autonomous TE loci. To interrogate this issue, we investigated the location distribution of structurally autonomous TE loci of Copia-3 and Copia-23, the two families estimated to mobilise most recently among all LTR-TE families in *V. vinifera*. Among the total annotated structurally autonomous TE loci of Copia-3 and Copia-23, 55 (63%) of 87 and 117 (66%) of 177 loci are within introns (Additional file 2: Figure S18A), respectively. Seventy-four and 138 of these structurally autonomous TE loci of Copia-3 and Copia-23 were identified as TE expression candidates in at least one experimental condition, respectively. Astonishingly, 59 (80%) of the 74 structurally autonomous Copia-3 expression candidate loci and 104 (75%) of the 138 structurally autonomous Copia-23 expression candidates co-localised with expressed genes, primarily within introns of expressed genes (Additional file 2: Figure S18B), indicating the distribution bias of structurally autonomous TE expression candidates toward expressed genes in the LTR-TE families that are most recently active.

The transcriptional dynamics of TEs is closely related to that of co-localised genes

The over-representation of TE expression candidates in proximal of expressed genes in the *V. vinifera* and *A. thaliana* systems (particularly At_WT and At_ibm2) indicates the tolerance of the transcriptional activity of intragenic TEs within expressed genes and suggests a' hitchhiker-like' manner of intragenic TEs in that they take advantage of the genic transcriptional permissive status for their own expression [18, 49]. Therefore these TEs might display expression dynamics that resemble a host gene's activity.

To address this possibility, differential transcriptional changes of TE expression candidates and genes of the grapevine system were detected by the computational tool DESeq2 [50]. Due to the repetitive characteristics of TEs, only a subset of expression candidates (i.e. trackable expression candidates) that were mapped by unique-mapping reads were suitable for this analysis. The differential analysis was performed on 5,869 trackable expression candidates found in at least one of the three experimental conditions of the *V. vinifera* system (Vv_T = 0, Vv_mock, and Vv_Yeast). Hierarchical clustering of differentially expressed TEs (DETEs) demonstrated various predominant expression patterns in response to different treatments (Fig. 4a, b). The mock treatment (Vv_Mock) showed that over 50% of the DETEs were transcriptionally activated in the first 3 hours (3h) of post-treatment and then returned to an expression level similar to that observed in Vv_T = 0 (Fig. 4a, c), illustrating an 'up-back' expression pattern. Interestingly, 206 of the 291 DETEs (70.79%) responded to *H. uvarum* incubation (Vv_Yeast) in an up-regulated manner (Fig. 4b, e).

In order to define the expression pattern of genes co-localised with DETEs, differential expression analysis was also applied on all expressed genes (FPKM > 1; Additional file 2: Figure S18), following by collection of differentially expressed genes (DEGs) that co-localised with DETEs (see Methods; Fig. 4d, f). Among the DEGs of Vv_Mock samples, 40 DEGs were co-localised with 45 of the 78 DETEs (Fig. 4d). In Vv_Yeast samples, 106 DEGs were co-localised with 124 of the 291 DETEs in this treatment (Fig. 4f). In an attempt to investigate the relationship of expression pattern between DEGs and the co-localised DETEs, these corresponding DETEs and DEGs were used for hierarchical clustering, in which DETEs and DEGs of similar expression pattern were grouped into the same clusters (Fig. 4g, h). Note that a small number of DETEs, especially DETEs within 2kb flanking regions of genes, might co-localise with multiple DEGs and *vice versa*. Instead of arbitrarily excluding DETEs or DEGs that fell into this scenario, the comparison of the expression pattern of co-localised DETEs and DEGs was conducted on each DETE-DEG pair. The expression pattern of the 45 DETEs co-localized with 40 DEGs in Vv_Mock was then compared with that of paired DEGs, resulting in 45 pairs of DETE-DEG comparisons summarised in Fig. 4i, where 42 (93.33%) pairs of co-localised DETE-DEG showed concordant clustering between DETEs and corresponding DEGs. The same approach was applied on co-localised DETEs and DEGs of Vv_Yeast, in which 113 (89.68%) of the total 126 co-localised DETE-DEG pairs showed the same expression pattern between paired DETEs and DEGs (Fig. 4j). These findings indicate that the dynamic expression pattern of DETEs co-localized with DEGs tended to resemble that of the paired DEGs.

Transcriptionally activated intronic TE loci are associated with intron retention and exposure of premature termination codons

Aberrant alternative splicing, such as exon skipping and intron retention, has been observed at gene loci containing epigenetically unmasked intronic TEs [51, 52]. While TE expression candidates in the *V. vinifera* and *A. thaliana* systems show a strong location bias towards expressed genes, how broad the transcriptionally active intronic TE loci associated with aberrant alternative splicing remains unanswered. To interrogate this issue on a genome-wide scale and retain intact sequence information of TE-related transcripts, Oxford Nanopore Technology (ONT) cDNA sequencing was utilised for *V. vinifera* (P. noir UCD5 clone) embryogenic callus subjected to 12 hours of continuous incubation with live *H. uvarum* culture (hereinafter Vv_Yeast12h) and the corresponding mock treatment (hereinafter Vv_Mock12h). The alignment depth of the ONT cDNA sequencing reads to the *V. vinifera* reference genome was comparable to that of the Illumina libraries (Additional file 6: Table S5). The quantification of gene and TE family expression level in the ONT dataset shows a medium to high level of correlation with that of the Illumina dataset (Spearman's correlation coefficient $\rho > 0.80$ for genes; $\rho > 0.58$ for TE families; Additional file 2: Figure S20). The FLAIR pipeline [53] was used to categorise alternative splicing events into four categories: alternative 3' splicing (Alt3), alternative 5' splicing (Alt5), intron retention (IR) and exon skipping (ES). Gene-related alternative splicing features overlapping with TEs were further collected, following by estimating the productivity (as per the definition in FLAIR pipeline, this denotes the ability of a transcript to produce protein), of gene transcripts having these alternative splicing features. First of all, among the total 21,081 alternative splicing features identified by FLAIR across the ONT libraries of Vv_Mock12h and Vv_Yeast12h, 19,526 (92.6 %) of these are related to annotated genes. Over 90% of these gene-related alternative splicing features were IR (8,806 alternative splicing features) and ES (9,378 alternative splicing features). Note that an isoform may contain multiple numbers and various types of alternative splicing features. Nonetheless, an alternative splicing feature could appear in multiple isoforms, as indicated in Fig. 5a. Notably, there are more genes than the number of associated ES features, suggesting that, for some ES events, each may involve more than one gene. Of the 19,526 gene-related alternative splicing features, only 524 (2.7%) of these overlapped with TEs (Fig. 5a). As expected, almost all TEs overlapping with Alt3, Alt5 and IR features are located within introns, while 22 of 40 ES-associated TEs overlapped with annotated exons.

To understand whether the presence of these TEs associated with the productivity of the gene transcripts, the productivity of isoforms containing these gene-and-TE-related alternative splicing was estimated using FLAIR and grouped into four types: productive (PRO), having premature termination codon (PTC, i.e. unproductive), no start codon (NGO), and having start codon but no stop codon (NST). This analysis shows that 50–68% of the isoforms having Alt3, Alt5, or ES remained productive, no matter whether the alternative splicing features overlapped with TEs (Fig. 5b). However, 80.6% of isoforms having TE-related IR were PTC, while the PTC proportion in isoforms having IR events non-overlapping with TEs was less than 45%. Looking into the estimated translation stop site of these isoforms containing TE-related IR feature, 196 of the 261 PTC isoforms exhibit premature stop codon exactly within the TE-overlapping IR feature (Additional file 7: Table S6). From the perspective of the isoform orientation, nine of the translational premature termination sites appear within TEs, two are after TEs, and the rest 186 isoforms show premature termination sites before the presence of TEs. The distance between TEs and the premature termination sites presented prior to TEs ranged from 2 bp to over 4 kb, with the first quartile, median and third quartile at 147 bp, 311 bp, and 693 bp, respectively (Additional file 7: Table S6).

Interestingly, different TE superfamilies were preferentially observed among the four types of alternative splicing features. Retrotransposon VLINE was over-represented in Alt3 and Alt5 alternative splicing events (Additional file 2: Figure S21A, B), and Harbinger, a DNA transposon, was predominantly seen in IR features (Additional file 2: Figure S21C). For ES features, MULE DNA transposon was the most predominant superfamily among all TE superfamily (Additional file 2: Figure S21D). In addition, most of these TEs were fragmented. There're only 10, 7, 34 and 1 full-length TE loci associated with Alt3, Alt5, IR, and ES features, respectively (Additional file 7: Table S6).

Identification of the potential origin of full-length transcription for autonomous mobilisation

Transcriptional activation of TEs under stress condition has been widely reported in plants and primarily investigated at the TE family level [17, 54]. Nonetheless, what proportion of these TE transcripts is derived from structurally autonomous TE loci and where are these autonomous TE loci seeding TE mobilisation in the genome remained unclear. Since our *V. vinifera* system has shown transcriptional activation of TEs under stressed conditions and demonstrated the characteristics of individual TE loci in the pool of expression candidates, we then aimed at identifying TE loci that potentially contribute full-length TE transcripts necessary for autonomous mobilisation. Due to the ability of ONT sequencing technology to sequence full-length transcripts, ONT cDNA libraries have the potential to reveal, if any, competent transcription of autonomous TE and decipher the origins of these transcripts [55]. The sequence integrity and structure of TE loci were firstly screened for structurally autonomous TE loci annotated in the genome. The structurally autonomous TE loci identified as expression candidates in the stress treatments (Vv_Mock12h and Vv_Yeast12h) were then examined for the breadth of read coverage against the TE sequence compartments whose transcription is required for autonomous mobilisation (see Methods). Intact LTR-TEs with > 90% INT coverage (Additional file 2: Figure S22 A, B), autonomous LINEs with > 0.9 breadth of coverage across whole elements (Additional file 2: Figure S23 A, B), as well as intact TIR-transposon (TIR-TE) with > 90% ORF covered by ONT reads (Additional file 2: Figure S24 A, B) were collected. This process captured 20 and 19 LTR-TE loci in Vv_Mock12h and Vv_Yeast12h libraries, respectively (Additional file 2: Figure S22 C). These include Copia-3, Copia-23, and Gypsy-V1. For LINE retrotransposon, only a single VLINE7 locus and two VLINE 8 loci were selected from Vv_Mock12h library (Additional file 2: Figure S23 C). For TIR-TE, three hAT-7 loci in Vv_Mock12h library revealed > 0.9 breadth of coverage across ORF (Additional file 2: Figure S24 C).

To further check whether the nearly full breadth of coverage resulted from contiguous full-length ONT reads across TEs, rather than a co-contribution of multiple reads, the read length of each read and the bases mapped to the potentially autonomous TEs were investigated. The ONT read should meet two criteria to prove that a full-length autonomous TE transcript was present. Firstly, depending on the type of mapped TE, it should be at least as long as the INT domain, the ORF, or the full feature of the TE locus. Secondly, this read should have its TE-mapped bases almost as much as its read length. Take Copia-23 as an example, the ONT reads mapping to the 19 structurally autonomous, seemly fully expressed, Copia-23 loci in Vv_Mock12h mainly were shorter than 3,000 bp (x-axis, Additional file 2: Figure S25 A), whereas the size of the canonical Copia-23 INT domain is 4,084 bp. The only read longer than 3 kb identified exhibited a very poor mapping to the element (y-axis, Additional file 2: Figure S25 A). In addition, the majority of these ONT reads were multi-mapping (red dots, Additional file 2: Figure S25 A). This analysis showed that most of these reads were skewed from the diagonal line, indicating the inconsistency between lengths of ONT reads and the mapped bases of these reads to TEs. To figure out the factors underlying this inconsistency, the alignment start and end sites of these ONT reads in relation to the mapped Copia-23 loci were surveyed. As illustrated in the cartoons in Additional file 2: Figure S25 B and C, the head and tail of the alignment were grouped into three categories, internal, external, and clipped. The investigation reveals that most of these ONT reads represented transcription started within the Copia-23 loci (Additional file 2: Figure S25 B). However, the tail of the reads, especially those that deviated from the diagonal line, were mostly clipped due to the sequence discrepancy between ONT reads and TEs (Additional file 2: Figure S25 C). Only a few of them extended through the annotated boundary. Overall, there is no evidence of autonomous transcription from annotated Copia-23 loci.

The situation described for the 19 autonomous Copia-23 loci was also observed in captured TE loci of Copia-3, Gypsy-V1, VLINE7, VLINE8, and hAT-7 (Fig. 6; Additional file 2: Figure S25). Only a single ONT read mapping to an autonomous Gypsy-V1 locus and eight reads of hAT-7 appeared to adequately cover the bases of the INT (Fig. 6a-c) or ORF (Fig. 6d-f) of the associated TE loci. The genome browser image of the only Gypsy-V1 locus demonstrates the full coverage of this locus by a single ONT read (Fig. 6g). The genome browser image for hAT-7 shows ONT reads covering the ORF of the hAT-7 in chromosome 14 (Fig. 6h). These suggest potential transcription of Gypsy-V1 and hAT-7 may allow limited mobilisation of these elements.

Discussion

Here we have shown an efficient strategy to capture individual loci of TE expression candidates in the *V. vinifera* and *A. thaliana* systems and detailed the characteristic profile of these expression candidates.

We demonstrated that, when the epigenetic system is not compromised, these TE expression candidates constituted less than 3% and 5% of total annotated TEs in the *V. vinifera* and *A. thaliana* systems, respectively. However, the number of TE expression candidates was significantly increased in *At_ddm1*, reflecting the epigenetic silencing function of *DDM1* [20]. Implicated by the high proportion of fragmented and trackable expression candidates, the majority of these TE expression candidates are likely relics of autonomous elements (fragmented and trackable), even in the Arabidopsis *ddm1* seedlings impaired in epigenetic silencing. It has been reported that short and fragmented TE loci (< 2kb) are often lacking CHH methylation and less likely to be targeted by the expression-dependent form of epigenetic silencing. In contrast, large and structurally intact TE loci are enriched with CHH methylation and tend to be targeted by the expression-dependent epigenetic silencing pathway [56]. In addition, with the accumulation of polymorphisms over evolutionary time, the divergence in TE sequences of individual loci may reach a point where the endogenous siRNAs system fails to recognise these elements [57], and thus results in permissive transcription of these TE loci. These mechanisms might explain the over-representation of fragmented and trackable TE loci in the expression candidate pool. However, to address this fully, the accumulation level of DNA methylation at these loci and siRNAs targeting these TEs remains to be determined.

The proportion of fragmented and trackable TE expression candidates in each expressed TE family may vary in association with the transposition history: TE families (e.g. Copia-3 and Copia-23 in grapevine) experienced the most recent mobilisation burst tended to contribute structurally autonomous and un-trackable TE loci into the expression candidate pool. It is likely due to the sequence homology of these two families that have not significantly been erased by mutations accumulated through the relatively short evolutionary time comparing with the older TE families. As a result, a sequencing read derived from the TE locus of a recently mobilised lineage of a TE family cannot be attributed to a single locus; rather, it is mappable to most of the TE loci of this lineage, and thus TE loci of the transcribed lineage were identified as expression candidates (Additional file 2: Figure S3-S4). This finding might also reflect the notion that young TE families or individual loci that mobilised recently are more likely to be active than the old elements upon specific environmental cues [45]. In Arabidopsis, the heat-sensitive LTR retrotransposon, *ONSEN*, is transcriptionally activated in wild-type plants subjected to heat-shock treatment. Still, the transposition event of *ONSEN* can only be observed in the progeny of heat-treated plant mutants, *nrpd1-3*, the lost function mutation of a gene encoding plant-specific RNA Polymerase IV, which is required for siRNA biosynthesis for RdDM pathway [45, 58]. It has been found that structurally autonomous *ONSEN* loci having identical LTRs contribute more *ONSEN* transcripts in both Arabidopsis wild-type and *nrpd1-3*, and more new insertions in the progeny of heat-treated *nrpd1-3*, than the *ONSEN* loci having non-identical pair of LTRs [45]. In the grapevine system, although there were no detectable full-length transcripts derived from structurally autonomous loci of Copia-3 and Copia-23, these two TE families may potentially contribute to new insertions upon proper environmental stimuli if the epigenetic system is compromised.

It has been reported that LTR retrotransposons, in particular from the Copia superfamily, and DNA transposons, such as MITEs, exhibit insertion bias in favour of gene-rich or transcribed gene regions [14, 28–30, 58]. Therefore, it has been proposed that TE families, particularly those with low copy numbers (less than a few hundred per genomes), prefer integrating into genetically active locations in genomes, gaining the opportunity for transcription and mobilisation [49, 59]. Nonetheless, there is no evidence to date demonstrating the location bias of transcriptionally activated TE loci. Here, for the first time, we reveal the significant location bias of TE expression candidates towards expressed genes in all experimental conditions. Compromising epigenetic silencing of TEs, as seen in the *A. thaliana ddm1* mutant, leads to a greater array of transcribed TE loci outside of gene-rich regions. In the grapevine system, this location bias of transcriptionally activated TE loci particularly favours introns of expressed genes and is observed for most TE families becoming transcriptionally active. Furthermore, most of the structurally autonomous expression candidates of Copia-3 and Copia-23 families, the two youngest LTR-TE families in grapevine, were found within intron of expressed genes. Transcriptional co-activation of co-localised TE loci and genes is not only observed in our grapevine system, but also in other plant species, including rice [60], maise [61], and *A. thaliana* [62]. Together, our observations and the findings reported in these reports suggest a mobilisation cycle that positively reinforces genic insertion that predetermines the transcriptional and thus transpositional activity.

It remains unclear whether genic TE expression candidates confer the expression activity on co-localised genes through contributing the *cis*-regulatory element in TE sequences as previously reported [4, 9, 10, 13], or whether these genic TE expression candidates acquired transcriptional activity due to the requirement of permissive chromatin or epigenetic status granted to the co-localised and transcribed genes [59]. The former indicates a positive effect of the presence of TEs on gene activity, whereas the latter suggests a 'hitchhiker-like' manner of TEs and may not necessarily benefit gene activity [59]. In the 'hitchhiker-like' scenario, epigenetically unmasked TE loci can sometimes result in the production of fused TE-gene transcripts [63, 64] or aberrant transcript isoform of the host genes [52]. Our data show that gene transcript isoforms containing intronic TEs tend to possess premature termination codon and thus are estimated to be non-productive isoforms (Fig. 5b). This finding suggests that the intronic TE loci involved in intron retention may follow the 'hitchhiker-like' scenario and influence the transcription and translation productivity of the host genes. In addition, in the analysis of the alignment start sites of ONT cDNA read relative to structurally autonomous TE loci, 35 and 73 of the total 497 ONT reads analysed were found to cross the boundary of mapped TE loci (crossing-boundary) and possess soft-clipped tail, respectively

(Fig. 6a-f; Additional file 2: Figure S19-S20). The former indicates TE transcription initiated from the upstream of the analysed TE loci, and the latter suggests a discrepancy between the reference and P. noir genomic sequences. Both (crossing-boundary and soft-clipped at the N terminals of the alignments) imply a certain amount of non-de novo TE transcription.

Although transcriptional activation of TEs under stress conditions has been widely reported, our ONT cDNA sequencing data shows that the full-length transcript derived from structurally autonomous TE loci is extremely low. This finding indicates that the so-called 'TE activation' commonly seen from short-read RNAseq may not necessarily represent transcriptional activation leading to autonomous TE mobilisation and suggests a rare possibility of a TE burst occurred in wild-type populations in a natural environment [15]. A recent study in *A. thaliana* pooled six ONT cDNA sequencing libraries, which included *A. thaliana* wild-type and four epigenetically compromised genotypes before applying an expression threshold at five ONT reads to annotate an active TE locus [55]; this generated more than 5 million reads for the model organism with a genome size of ~ 135 Mb. With a similar N50, it would require alignment depth at least 2.3 times deeper than the current Vv_mock12h or Vv_Yeast12h libraries that we have interrogated. The *A. thaliana* genotypes used in Panda and Slotkin [55] include those which were compromised in epigenetic silencing that allows the constant de-repression of TE loci, whereas the grapevine embryogenic callus in our study was initiated from wild-type *P. noir* clone UCD5 and was subjected to stress treatment. In such a fully active epigenetic silencing environment, the likelihood of high levels of transcription of autonomous elements is low. That being said, providing a stress treatment that can lead to strong transcriptional activation of autonomous TE loci in the wild-type backgrounds, more sequencing and alignment depth may be necessary to detect the full-length TE transcription than normally required to detect gene activation. For this purpose, size selection for cDNA longer than a certain threshold may lower the requirement of sequencing depth. Alternatively, using genetic backgrounds compromised in epigenetic silencing and thus predisposed to TE activation, applying chemicals that inhibit proteins of the silencing machinery [65, 66], and/or conducting stress treatment according to the stress-responsive specificity of a particular TE family/subfamily [27, 58] may raise the levels of autonomous TE transcription.

Conclusions

The expression landscape and properties of transcriptionally active TE loci have been enigmatic mainly due to the repetitive and self-proliferating characteristics of TEs. Here, using short-read and long-read sequencing technology, as well as combined analysis of unique-mapping and multi-mapping sequencing reads, we show the evidence that transcriptionally active TE loci only constitute a small portion of total annotated TEs in *V. vinifera* and *A. thaliana* when the epigenetic silencing system is not compromised. Although most TE transcripts were derived from fragmented and trackable TE loci that are incapable of autonomous mobilisation, our results indicate a strong tendency for TE expression candidates to be found within introns of expressed genes. This distribution bias is commonly found for most of the expressed TE families, and for structurally autonomous TE expression candidates of grapevine Copia-3 and Copia-23, the two LTR-TE families experienced the most recent mobilisation burst. It was also discovered that the pairs of co-localised TEs and genes shared the same differential expression patterns in response to applied stressors. We further reveal a strong association of expressed intronic TE loci that form part of the gene transcripts (i.e. TE-related intron retention) and the presence of premature termination codon within the retained intron sequences of these gene transcript isoforms. These together suggest a close relationship between the transcriptional activity of TEs and genes. Finally, despite that the transcriptional activation of TEs has been observed in response to stress in *V. vinifera* system [23], the observation that little full-length transcripts can be derived from structurally autonomous TE loci. Such conditions imply low transposition efficiency and that the proper combination of stress treatments and measurements to induce a wide scale of epigenetic relaxation (e.g. chemical inhibition [65] of critical points in the epigenetic silencing pathways) are required to trigger efficient TE mobilisation in the wild-type genetic background.

Methods

Plant material and experimental conditions

Embryogenic callus cultures were established from *V. vinifera* cv Pinor noir clone UCD5 and maintained in the C_1^P medium according to Lizamore (2013) [23]. Stress treatment was conducted essentially based on the methods established by Lizamore (2013) [23]. Before the inoculation with the biotic stressor (live *H. uvarum* cultures), the embryogenic callus cultures were subjected to vigorous shaking in the hormone-free C_1^P liquid medium (HF- C_1^P) to break the lumps of callus apart. Therefore, this procedure was considered the mock treatment, which resembles a wound treatment to the callus. For both mock and biotic stress treatments, samples were harvested at 1, 3, 6, and 12 hours as shown in Fig. 1a. A common untreated 0 hour time point (denoted as Vv_T = 0) for mock and the biotic treatment was taken prior to any treatment. All treatments and their associated time points consisted of three technical replicates. Detailed procedures of stress treatment can be found in Additional file 8: Supplementary methods.

Short-read RNA-seq and data pre-processing

Total RNA of harvested embryogenic callus was isolated according to the manufacturer's instructions of Spectrum Plant Total RNA Kit (Sigma). Each purified RNA sample was treated with DNase I following the TURBO DNA-free kit protocol (Ambion) to remove contaminating genomic DNA before being sent to New Zealand Genomics Ltd (now Otago Genomics) for library preparation and pair-end Illumina sequencing using a HiSeq 2500 sequencer.

Adapter sequences and low-quality bases were trimmed by fastq-mcf [67] before quality check using fastqc [68]. Trimmed quality reads were aligned to *V. vinifera* tRNA and rRNA sequences (Ensembl Plants database: https://plants.ensembl.org) followed by the collection of unmapped reads using samtools [69].

RNAseq data of wild-type and *ibm2 A. thaliana* were obtained from Le et al. (2015) [36] (accession codes DRA002305 and DRA002306 in DDBJ Sequence Read Archive at https://www.ddbj.nig.ac.jp/dra/), while the *ddm1* RNAseq data was collected from Oberlin et al. (2017) [19] (accession code GSE93584 in NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/). The pre-processing of these raw sequencing data follows the above-mentioned method used for *V. vinifera* dataset. *A. thaliana* tRNA/rRNA sequences were downloaded from Ensembl Plants (https://plants.ensembl.org).

Identification of TE expression candidates

This analysis is comprised of three sub-pipelines described as follows:

For the first sub-pipeline, sequencing reads unmapped to tRNA and rRNA sequences of the analysed species were aligned to the reference genome using HISAT2 [70]. The software *htseq-count* of the package Htseq [42] was then utilised to quantify unique-mapping reads aligning to annotated TE loci. TE loci having read count of more than ten were collected.

For the second sub-pipeline, reads mapping to the reference genome by the software HISAT2 were further analysed using *bedtools coverage* and *bedtools intersect* to calculate unique- and multi-mapping reads mapping to annotated TEs (a multi-mapping read would be counted multiple times) and the number of bases of a TE locus covered by reads (covered bases of a TE locus). These data were then utilised to calculate the average read depth of an individual TE's mapped region. TE loci with more than ten reads and average read depth > 5 were collected in this sub-pipeline.

The third sub-pipeline was established based on the TEFingerprint's workflow, which initially used BWA to align reads against the reference genome. TEFingerprint then processed the mapped reads to capture discordant unique-mapping reads (hereinafter dangler reads) whose paired mates were aligned to TE sequences. TE loci flanked by more than ten dangler reads on either side and had been found to have more than ten reads mapped internally were collected.

Finally, TE loci collected by each sub-pipeline were all combined to form the pool of TE expression candidates.

For the analysis of *V. vinifera* system, the 12X PN40024 grapevine reference genome was acquired from the Ensembl Plants database (https://plants.ensembl.org), and the TE annotation was established by Lizamore (2013) [23]. Grapevine's gene annotation file (version 2.1) was obtained from the Grape Genome Database at CRIBI (http://genomics.cribi.unipd.it/grape/).

For *A. thaliana* analysis, the *A. thaliana* TAIR10 reference genome and gene annotation file were downloaded from Ensembl Plants (https://plants.ensembl.org), whereas the corresponding TE annotation file was generated by Jin et al. (2015) [48] and is available from Prof. Molly Hammell's lab web page (http://hammelllab.labsites.cshl.edu/).

The downstream analysis to profile TE expression candidates' characteristics and interrogate the expression patterns across time (i.e. differential expression analysis) was conducted using R script. Details regarding the pipeline and the downstream analysis can be found in Additional file 8: Supplementary methods.

Long-read cDNA-seq and data analysis

The grapevine embryogenic callus was subjected to the mock treatment and live *H. uvarum* treatment following the method mentioned earlier. The callus of mock treatment was harvested after 12 hours of recovering on the C_1^P plate, whereas the yeast treatment involves 12 hours of continuous incubation with the yeast before harvesting.

Total RNA was extracted and separated using NORGEN Plant microRNA purification kit (Norgen Biotek) according to the manufacturer's instruction. Genomic DNA contamination was removed with the standard protocol of the TURBO DNA-free kit (Thermo Fisher). The RNA quantity was measured by Qubit RNA BR (Broad-Range) Assay Kit (Thermo Fisher), and the quality was examined using Agilent 2100 Bioanalyzer, in which the resulting RIN value of each library was above 8.

The cDNA library was prepared following the protocol of the Oxford Nanopore cDNA-PCR kit (SQK-PCS109). Briefly, 50 ng of total RNA was reverse transcribed before the strand-switching step. The resulting full-length cDNA was further enriched by PCR, which involved 12–13 amplification cycles, each with 6 min of extension step. The amplified cDNA was purified by AMPure XP beads before the ligation of the 1D sequencing adapters. Finally, the cDNA library was loaded onto a R9.4.1 MinION flow cell and then sequenced using MinKNOW (version 18.12) control software for raw data collection only.

Base-calling was carried out offline using Guppy (version 3.2.1; https://nanoporetech.com/) before removing adapter sequences using Pychopper2 [84]. The full-length reads were mapped to the 12X PN40024 *V. vinifera* reference genome using minimap2 [85] with the pre-set option *-ax splice* for long-read splice alignment. For mapping self-proliferating and highly repetitive TE sequences, the output of up to 100 secondary alignments was allowed for individual TE analysis using the flag *-N 100 -ax splice* in minimap2. For analysis based on TE family level, the ONT reads were mapped to the set of 232 canonical TE sequences by running default minimap2 using *-ax splice* before using *bedtools coverage* to quantify mapped reads at the family level. For genes, based on grapevine gene annotation (version 2.1), FLAIR [53] pipeline was then applied to obtain high fidelity isoforms and quantify the isoform expression level as transcripts per million (TPM). The TPM of isoforms derived from the same gene was then summed up at the gene level for the overall quantification of gene expression. For individual TEs, ONT reads overlapping with TEs in sense orientation were collected and quantified by *bedtools intersect* and *bedtools coverage* [43] as previously described.

To compare the correlation between ONT and Illumina platforms at the individual gene level, genes' TPM (logarithmically transformed) from ONT was plotted against FPKM (logarithmically transformed) from Illumina data, while the correlation was tested using Spearman's correlation coefficient. The same approach was applied to TE families, in which the expression level of each TE family was obtained from TEtranscripts [48] for the Illumina dataset and from alignment with the canonical sequences followed by bedtools coverage analysis for ONT data.

Individual TE loci overlapping with at least one ONT read were collected to perform an intersection with the expression candidates obtained from Illumina libraries at a time point of 12 hours. The intersected TE loci were supported by both sequencing platforms and thus considered as expressed TEs. The identification of the potential origin of full-length transcription for autonomous TE loci was conducted using Bedtools [43] and bash and R scripts. Details can be found in Additional file 8: Supplementary methods and https://github.com/ting-hsuan-chen/TE_ExpressionCandidate.

To identify expressed genes, an intersection between annotated genes with ONT TPM above one and genes with Illumina FPKM over one was performed. Genes having overlapping data from both platforms were considered transcriptionally active genes. The alternative-splicing analysis for expressed genes was conducted using FLAIR pipeline [53].

Abbreviations

Alt3	Alternative 3' splicing
Alt5	Alternative 3' splicing
Bp	Base pair(s)
cDNA	Complementary DNA
C-flank	C-terminal 2kb-flanking region of a gene
DDM1	Decrease in DNA methylation 1 (chromatin-remodelling protein)
DEG	Differentially expressed gene
DETE	Differentially expressed transposable element
DNA	Deoxyribose nucleic acid
epiRIL	Epigenetic recombinant inbred line
ES	Exon skipping
HAT	Histone acetylase
hATC	hAT C-terminal dimerization
IBM2	Increase in BONSAI methylation 2 (an RNA binding protein)
INT	Internal domain
IR	Intron retention
Kb	Kilobase pairs

LINE	Long interspersed repetitive element
LTR	Long-terminal repeat
LTR-TE	LTR transposable element
Mb	Megabase pairs
MITE	Miniature inverted-repeat transposable element
mRNA	Messenger RNA
MULE	Mutator-like element
N-flank	N-terminal 2kb-flanking region of a gene
NGO	No start codon for mRNA translation
NST	Having start codon but no stop codon for mRNA translation
nt	Nucleotide
ONT	Oxford nanopore technology
ORF	Open reading frame
PRO	Productive in terms of mRNA translation into protein
PTC	Premature termination codon for mRNA translation
RdDM	RNA dependent DNA methylation
RNA	Ribose nucleic acid
RNAseq	RNA sequencing
rRNA	Ribosomal RNA
siRNA	Small interference RNA
TE	Transposable element
TIR	Terminal inverted repeat
TIR-TE	TIR transposable element
tRNA	Transfer RNA

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All sequencing data generated in this study are available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO, https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE175475**.** Publicly available next-generation sequencing data were downloaded from NCBI GEO and DDBJ under the accession number specified in Le et al. (2015) [36] and_Oberlin et al. (2017) [19], and are listed along with general mapping statistics in Additional file 1: Table S1. The computational scripts used for this research, including a pipeline to identify TE expression candidates, is available at https://github.com/ting-hsuan-chen/TE_ExpressionCandidate.

Competing interests

The authors declare that they have no conflicts of interests.

Funding

This work was supported by the Plant and Food Research Ltd. SSIF Grapevine Genetic Improvement programme, Lincoln University internal funding, and a New Zealand International Doctoral Research Scholarship to THC.

Authors' contributions

THC and CW designed the study; THC and CW performed the experiments; THC analysed the data; THC and CW wrote the manuscript.

Acknowledgements

We would like to thank Michelle Thompson and Tirthanka Ghosh for their support in maintaining grapevine cell cultures and conducting the initial biotic stress treatment of the cultures used in this study. We would also like to thank Assoc Prof. Ross Bicknell and Dr. Darrell Lizamore for their valuable discussions and comments on this research.

Authors' information

Affiliation:

Department of Wine, Food, and Molecular Biosciences, Lincoln University, Lincoln 7647, New Zealand

Ting-Hsuan Chen, Christopher Winefield

The New Zealand Institute for Plant and Food Research Limited, Lincoln 7608, New Zealand

Ting-Hsuan Chen (present address)

Corresponding author:

Correspondence to Christopher Winefield

References

1. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115:49–63.
2. Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 2010;15:471–8.
3. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351:1083–7.
4. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9:397–405.
5. McConnell MJ, Moran JV, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science. 2017;356:eaal1641.
6. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet Lond. 2010;11:204–20.
7. Borges F, Martienssen RA. The expanding world of small RNAs in plants. Nat Rev Mol Cell Biol. 2015;16:727–41.
8. Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. Nat Plants. 2016;2:16163.

9. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol Ecol. 2013;22:1503–17.

10. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14:49–61.

11. Zhao D, Ferguson AA, Jiang N. What makes up plant genomes: The vanishing line between transposable elements and genes. Biochim Biophys Acta BBA - Gene Regul Mech. 2016;1859:366–80.

12. Yano R, Ariizumi T, Nonaka S, Kawazu Y, Zhong S, Mueller L, et al. Comparative genomics of muskmelon reveals a potential role for retrotransposons in the modification of gene expression. Commun Biol. 2020;3:1–13.

13. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.

14. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, et al. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat Commun. 2019;10:3421.

15. Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun. 2019;10:5818.

16. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8:272–85.

17. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19:199.

18. Choi JY, Lee YCG. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. PLOS Genet. 2020;16:e1008872.

19. Oberlin S, Sarazin A, Chevalier C, Voinnet O, Marí-Ordóñez A. A genome-wide transcriptome and translatome analysis of *Arabidopsis* transposons identifies a unique and conserved genome expression strategy for *Ty1/Copia* retroelements. Genome Res. 2017;27:1549–62.

20. Zemach A, Kim MY, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, Thao K, et al. The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. Cell. 2013;153:193–205.

21. Hirochika H, Okamoto H, Kakutani T. Silencing of Retrotransposons in Arabidopsis and Reactivation by the ddm1 Mutation. Plant Cell. 2000;12:357–68.

22. Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. Reconstructing de novo silencing of an active plant retrotransposon. Nat Genet. 2013;45:1029–39.

23. Lizamore DK. A study of endogenous transposon activity in grapevine (Vitis vinifera L.). [New Zealand]: Lincoln university; 2013.

24. Rakocevic A, Mondy S, Tirichine L, Cosson V, Brocard L, Iantcheva A, et al. *MERE1*, a Low-Copy-Number Copia-Type Retroelement in *Medicago truncatula* Active during Tissue Culture. Plant Physiol. 2009;151:1250–63.

25. Małolepszy A, Mun T, Sandal N, Gupta V, Dubin M, Urbański D, et al. The *LORE1* insertion mutant resource. Plant J. 2016;88:306–17.

26. Hashida S-N, Uchiyama T, Martin C, Kishima Y, Sano Y, Mikami T. The Temperature-Dependent Change in Methylation of the *Antirrhinum* Transposon Tam3 Is Controlled by the Activity of Its Transposase. Plant Cell. 2006;18:104–18.

27. Beguiristain T, Grandbastien M-A, Puigdomènech P, Casacuberta JM. Three Tnt1 Subfamilies Show Different Stress-Associated Patterns of Expression in Tobacco. Consequences for Retrotransposon Control and Evolution in Plants. Plant Physiol. 2001;127:212–21.

28. Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL. Mu1-related transposable elements of maise preferentially insert into low copy number DNA. Genetics. 1995;140:315–24.

29. Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, et al. Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome. PLOS Genet. 2009;5:e1000733.

30. Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, et al. Target Site Specificity of the *Tos17* Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome. Plant Cell. 2003;15:1771–80.

31. Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, et al. Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci. 2006;103:17620–5.

32. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461:1130–4.

33. Zhang HY, Lee SA, Bradbury JE, Warren RN, Sheth H, Hooks DO, et al. Yeasts isolated from New Zealand vineyards and wineries. Aust J Grape Wine Res. 2010;16:491–6.

34. Albertin W, Setati ME, Miot-Sertier C, Mostert TT, Colonna-Ceccaldi B, Coulon J, et al. Hanseniaspora uvarum from Winemaking Environments Show Spatial and Temporal Genetic Clustering. Front Microbiol. 2016;6:1569.

35. Drumonde-Neves J, Franco-Duarte R, Lima T, Schuller D, Pais C. Yeast Biodiversity in Vineyard Environments Is Increased by Human Intervention. PLOS ONE. 2016;11:e0160579.

36. Le TN, Miyazaki Y, Takuno S, Saze H. Epigenetic regulation of intragenic transposable elements impacts gene transcription in Arabidopsis thaliana. Nucleic Acids Res. 2015;43:3911–21.

37. Deremetz A, Roux CL, Idir Y, Brousse C, Agorio A, Gy I, et al. Antagonistic Actions of FPA and IBM2 Regulate Transcript Processing from Genes Containing Heterochromatin. Plant Physiol. 2019;180:392–403.

38. Ito H, Kim J-M, Matsunaga W, Saze H, Matsui A, Endo TA, et al. A Stress-Activated Transposon in Arabidopsis Induces Transgenerational Abscisic Acid Insensitivity. Sci Rep. 2016;6:23181.

39. Saze H, Kitayama J, Takashima K, Miura S, Harukawa Y, Ito T, et al. Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. Nat Commun. 2013;4:2301.

40. Wang X, Duan C-G, Tang K, Wang B, Zhang H, Lei M, et al. RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. Proc Natl Acad Sci. 2013;110:15467–72.

41. Saze H. Epigenetic regulation of intragenic transposable elements: a two-edged sword. J Biochem (Tokyo). 2018;164:323–8.

42. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.

43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

44. Plant and Food Research. TEFingerprint. 2019. https://github.com/PlantandFoodResearch/TEFingerprint. Accessed 21 Feb 2020.

45. Sanchez DH, Gaubert H, Drost H-G, Zabet NR, Paszkowski J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat Commun. 2017;8:1283.

46. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. 2007;17:1072–81.

47. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20:43–5.

48. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31:3593–9.

49. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol Evol. 2000;42:251–69.

50. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

51. Saint-André V, Batsché E, Rachez C, Muchardt C. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat Struct Mol Biol. 2011;18:337–44.

52. Ong-Abdullah M, Ordway JM, Jiang N, Ooi S-E, Kok S-Y, Sarpan N, et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature. 2015;525:533–7.

53. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11:1438.

54. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21:721–36.

55. Panda K, Slotkin RK. Long-read cDNA Sequencing Enables a 'Gene-Like' Transcript Annotation of Arabidopsis Transposable Elements. Plant Cell. 2020;32:2687–98.

56. Panda K, Ji L, Neumann DA, Daron J, Schmitz RJ, Slotkin RK. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA

methylation. Genome Biol. 2016;17:170.

57. Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. Curr Opin Plant Biol. 2015;27:67–76.
58. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011;472:115–9.
59. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009;19:1419–28.
60. Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, et al Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. Weigel D, editor. eLife. 2015;4:e09343.
61. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. Freeling M, editor. PLoS Genet. 2015;11:e1004915.
62. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, et al. Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci. 2012;109:E2183–91.
63. Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, et al. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. Genome Res. 2009;19:42–56.
64. Tsuchiya T, Eulgem T. An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. Proc Natl Acad Sci. 2013;110:E3535–43.
65. Thieme M, Lanciano S, Balzergue S, Daccord N, Mirouze M, Bucher E. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biol. 2017;18:134.
66. Yu C-W, Tai R, Wang S-C, Yang P, Luo M, Yang S, et al. HISTONE DEACETYLASE6 Acts in Concert with Histone Methyltransferases SUVH4, SUVH5, and SUVH6 to Regulate Transposon Silencing. Plant Cell. 2017;29:1970–83.
67. Aronesty E. Comparison of Sequencing Utility Programs. Open Bioinforma J. 2013;7:1–8.
68. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 21 Feb 2020.
69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
70. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
71. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
72. Smit AFA, Hubley R, Green P RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org. Accessed 16 Jan 2019.

73. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

74. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011;12:35.

75. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genom. 2007;8:218.

76. Jukes TH, Cantor CR. Evolution of Protein Molecules. New York: Academic Press. New York: Academic Press;; 1969.

77. Gaut BS, Morton BR, McCaig BC, Clegg MT. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci. 1996;93:10274–9.

78. Moisy C, Garrison K, Meredith CP, Pelsy F. Characterization of ten novel Ty1 copia-like retrotransposon families of the grapevine genome. BMC Genom. 2008;9:469.

79. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009.

80. Hyndman R. hdrcde: Highest Density Regions and Conditional Density Estimation. R package version 3.3. 2018. http://pkg.robjhyndman.com/hdrcde. Accessed 16 Jan 2019.

81. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2018. https://CRAN.R-project.org/package=dplyr. Accessed 21 Feb 2020.

82. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth. New York: Springer; 2002. http://www.stats.ox.ac.uk/pub/MASS4. Accessed 22 Apr 2020.

83. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. 2020. https://CRAN.R-project.org/package=gplots. Accessed 22 Apr 2020.

84. Oxford Nanopore Technologies. Pychopper: A tool to identify, orient, trim and rescue full length cDNA reads. Oxford Nanopore Technologies; 2020. https://github.com/nanoporetech/pychopper. Accessed 26 May 2020.

85. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34:3094–100.

86. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.

87. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66.

Figures

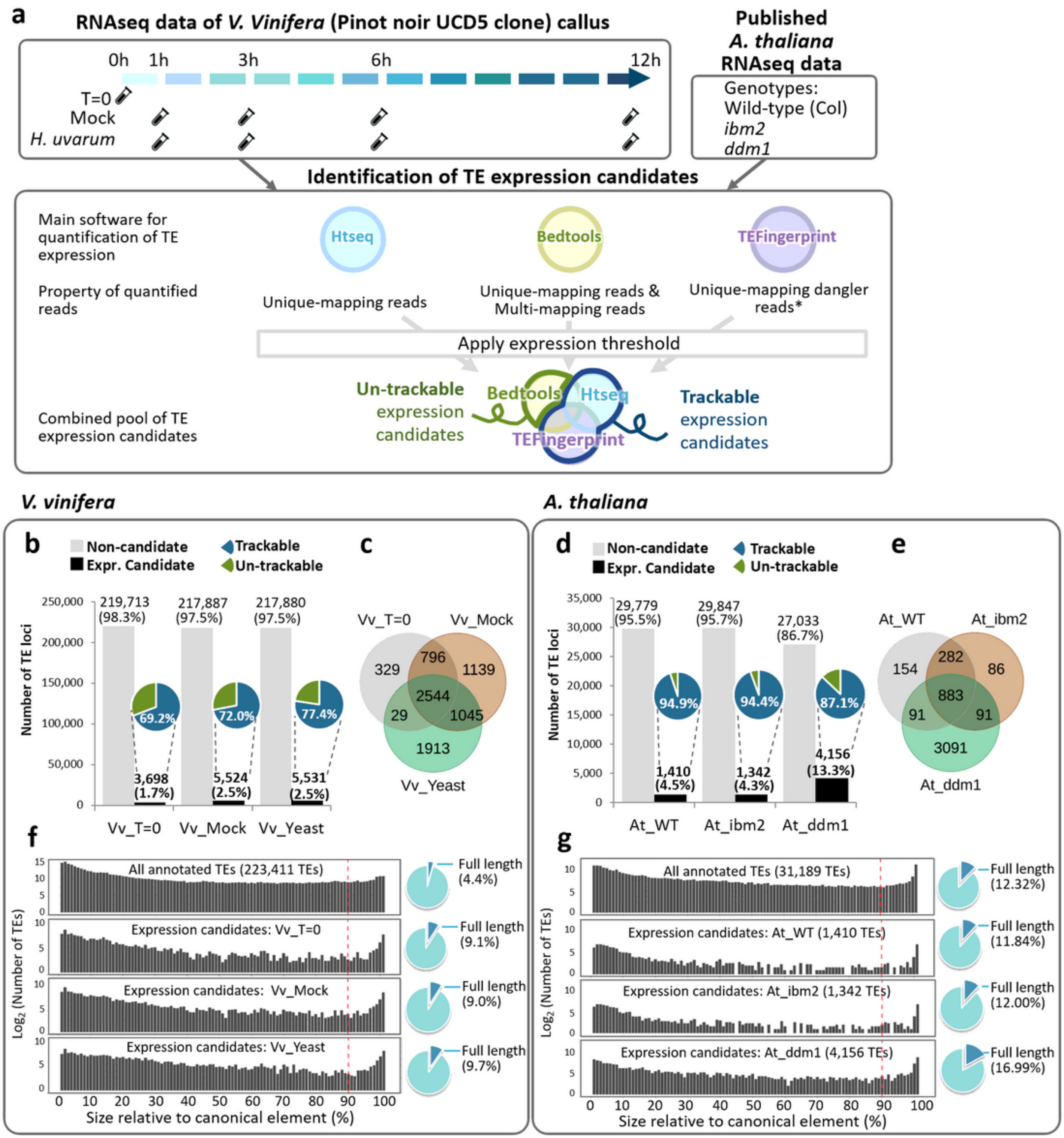


Figure 1

Identification of TE expression candidates in V. vinifera embryogenic callus and A. thaliana seedlings. a Experimental setting and simplified analysis workflow. RNAseq data from the V. vinifera embryogenic callus subjected to time-series stress treatment and published RNAseq data of Arabidopsis thaliana seedlings of wild-type [36], ibm2 [36], and ddm1[19] were analysed by the computational workflow constituted with three sub-pipelines that were mainly different in the software, and hence strategies, to quantify reads mapping to individual TE loci. The combined set of potentially expressed TE loci identified by each sub-pipeline forms the final pool of TE expression candidates. b, d Bar charts demonstrating the abundance of TE expression candidate loci identified in different experimental conditions or genotypes of the V. vinifera (b) and A. thaliana (d) systems, with pie charts showing the proportion of trackable/un-trackable loci of TE expression candidates. c, e Venn diagrams showing different sets of TE expression candidates in the V. vinifera (c) and A. thaliana (e) systems. f, g Histograms demonstrating the length integrity of all annotated TE loci and TE expression candidates, and pie charts showing the proportion of full-length TE loci among all annotated TE loci and TE expression candidates, in the V. vinifera (f) and A. thaliana (g) systems.

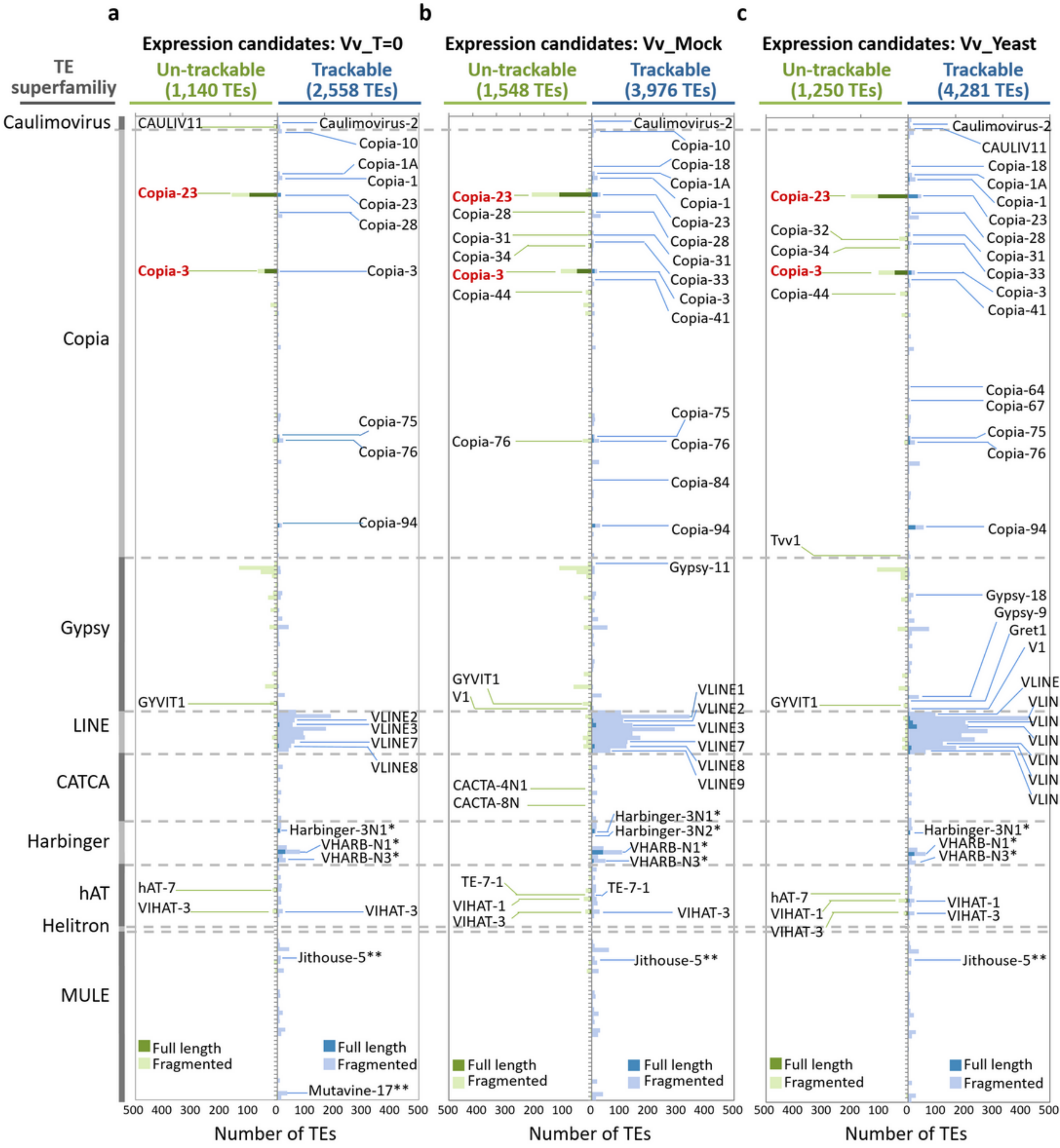


Figure 2

Transcriptionally active TE families in grapevine embryogenic callus. a−c Histograms demonstrating all expression candidates of Vv_T=0 (a), Vv_Mock (b), and Vv_Yeast (c) categorised by families, distinctiveness and integrity. Each bar represents a TE family containing expression candidates. The expression candidates were then further grouped into un-trackable (green) and trackable (blue) candidates, those of which full-length were filled with either dark green or dark blue. TE families containing at least two full-length expression candidates of either group were indicated. Note that Harbinger families missing open reading frame (ORF) encoding transposase and MULE families lack of terminal inverted repeats (TIRs) in their canonical sequences were indicated.

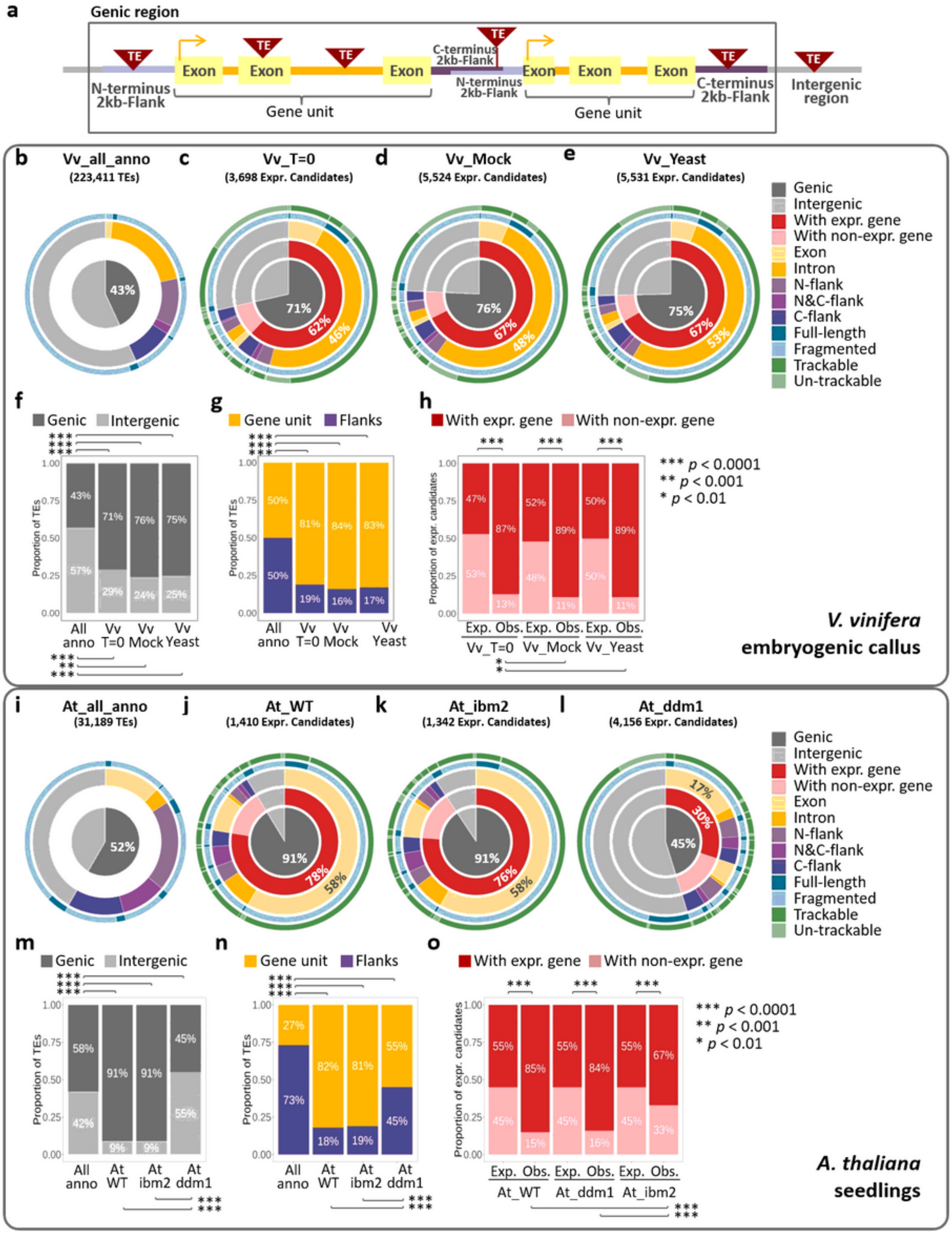


Figure 3

Location bias of expressed TE loci. a Illustration of TE insertions in the different genomic region (genic or intergenic) or location (exon, intron, or flanks) relative to genes. b-e Hierarchical classification of all annotated V. vinifera TE loci (b) and all TE expression candidates of Vv_T=0 (c), Vv_Mock (d), and Vv_Yeast (e), by location, integrity, and distinctiveness. TE loci were categorised in the order of region (centre), the transcriptional activity of co-localised genes (2nd layer), location (3rd layer), integrity (4th layer), and the presence/absence of unique-mapping reads (outer-most layer). f-h Hundred percent stacked bar charts showing the proportion of all annotated TE loci of V. vinifera and TE expression candidates (Vv_T=0, Vv_Mock, and Vv_Yeast) distributed in genic/intergenic regions (f), in gene unit (including exon and intron)/flanks (g), and with expressed/non-expressed genes (h). i-l Hierarchical classification of all annotated A. thaliana TE loci (i) and all TE expression candidates of At_WT (j), At_ibm2 (k), and At_ddm1 (l), by location, integrity, and distinctiveness. m-o Hundred percent stacked bar charts showing the proportion of all annotated A. thaliana TE loci and TE expression candidates (At_WT, At_ibm2, and At_ddm1) distributed in genic/intergenic regions (m), in gene unit/flanks (n), and with expressed/non-expressed genes (o). Colour codes for different categories of location, integrity, and distinctiveness are as indicated at the right panel. Chi-square tests with p-value < 0.01 were as indicated. Expr., expressed; Non-expr., non-expressed; N-flank, N-terminal 2kb-flanking region of gene; C-flank, C-terminal 2kb-flanking region of gene; Exp., expected proportion; Obs., observed proportion.

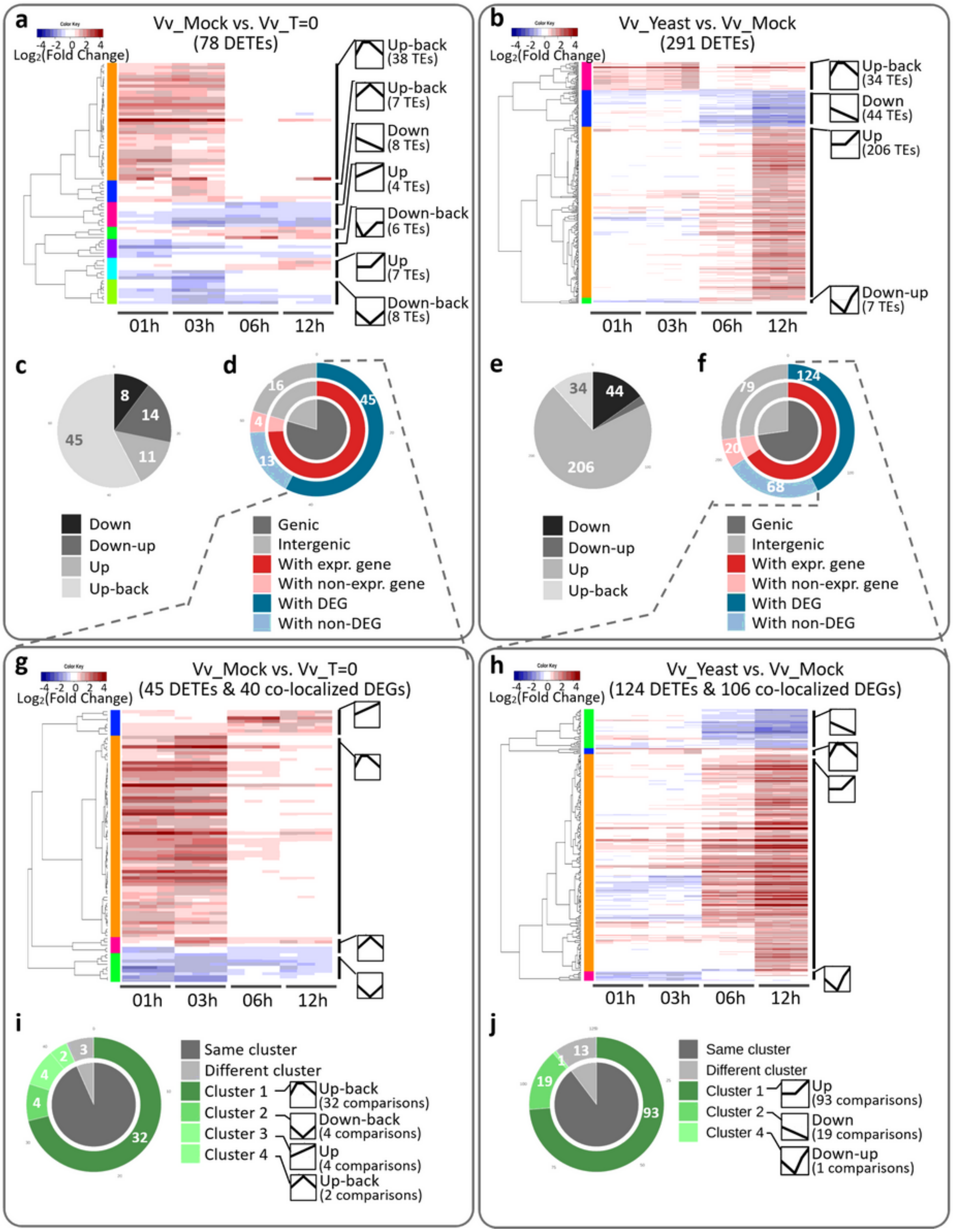
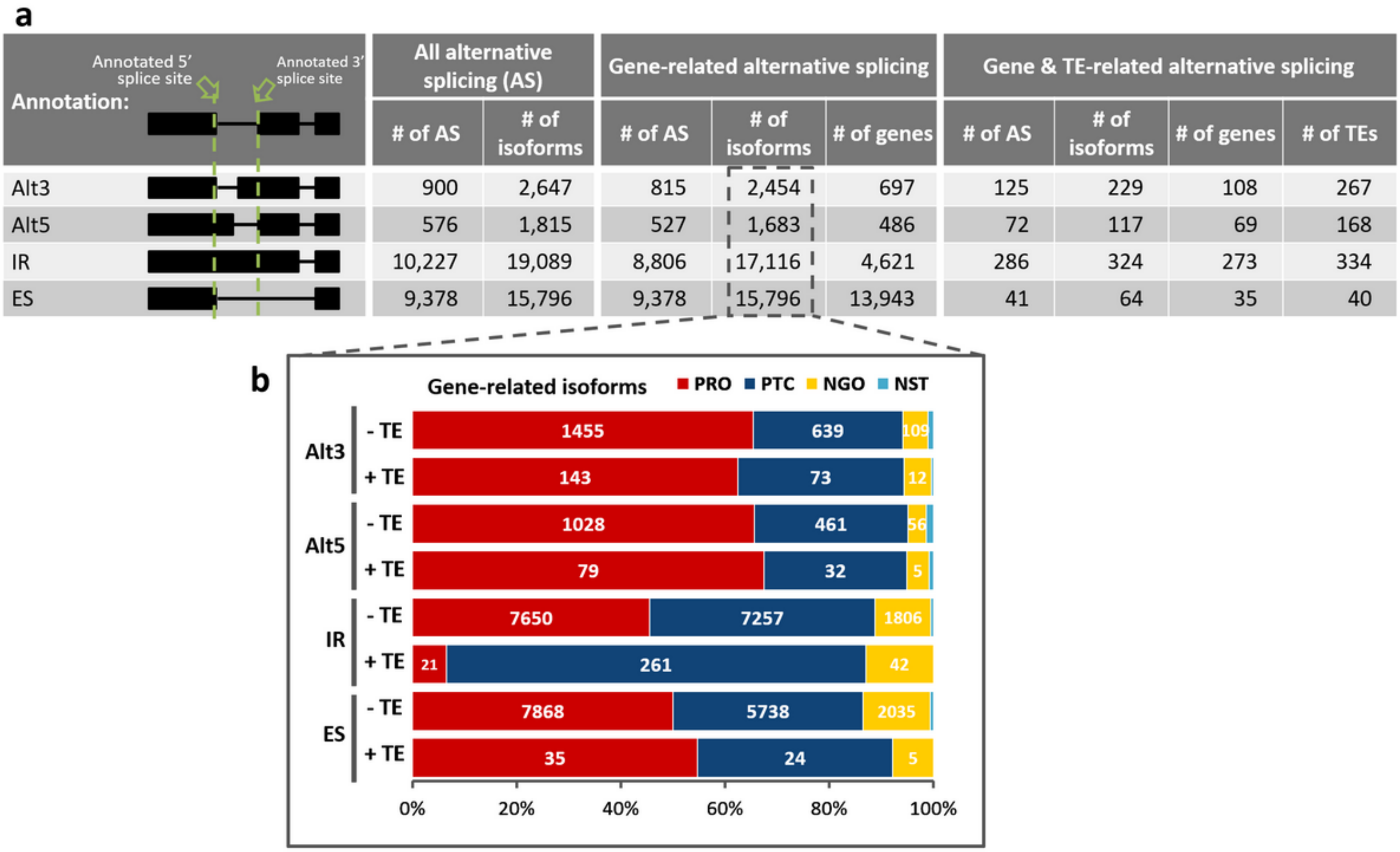


Figure 4

Expression dynamics of differentially-expressed TEs (DETEs) correlate with that of the co-localised differentially-expressed genes (DEGs). a, b Heatmaps showing hierarchical clustering of DETEs in Vv_Mock vs Vv_T=0 (a), and DETEs in Vv_Yeast vs Vv_Mock (b), based on logarithmic-transformed fold-changes. The three replicates of each time point are individually presented. The time-series expression pattern of each cluster is depicted as line graph with the number of TE loci within each cluster indicated.

c, e Pie charts to summarise the expression pattern of DETEs in Vv_Mock vs Vv_T=0 (c) and DETEs in Vv_Yeast vs Vv_Mock (e). d, f Hierarchical classification of DETEs in Vv_Mock vs Vv_T=0 (d) and DETEs in Vv_Yeast vs Vv_Mock (f) by the presence/absence of co-localised genes (genic/intergenic), the transcriptional activity of co-localised genes (with expr. gene/with non-expr. gene), and the differential test of expressed genes (with DEG/with non-DEG). g, h Heatmaps showing hierarchical clustering of co-localised DETEs and DEGs in Vv_Mock vs Vv_T=0 (g), and DETEs in Vv_Yeast vs Vv_Mock (h), based on logarithmic-transformed fold-changes. i, j Donut graphs to summarise the hierarchical clustering of co-localised DETEs and DEGs in Vv_Mock vs Vv_T=0 (i) and DETEs in Vv_Yeast vs Vv_Mock (j) by expression pattern.



Annotation:	All alternative splicing (AS)		Gene-related alternative splicing			Gene & TE-related alternative splicing			
	# of AS	# of isoforms	# of AS	# of isoforms	# of genes	# of AS	# of isoforms	# of genes	# of TEs
Alt3	900	2,647	815	2,454	697	125	229	108	267
Alt5	576	1,815	527	1,683	486	72	117	69	168
IR	10,227	19,089	8,806	17,116	4,621	286	324	273	334
ES	9,378	15,796	9,378	15,796	13,943	41	64	35	40

Figure 5

Intronic TE sequences retained in gene isoforms correlates with the presence of premature termination codon. a Illustration and summarisation of four types of alternative splicing. b A 100% bar chart to demonstrate the productivity of gene isoforms in the four types of alternative splicing, each further grouped as TE-present (+ TE) or TE-absent (- TE). AS, alternative splicing; Alt3, alternative 3' splicing; Alt5, alternative 5' splicing; IR, intron retention; ES, exon skipping; PRO, productive; PTC, premature termination codon; NGO, having no start codon; NST, having a start codon but no stop codon.

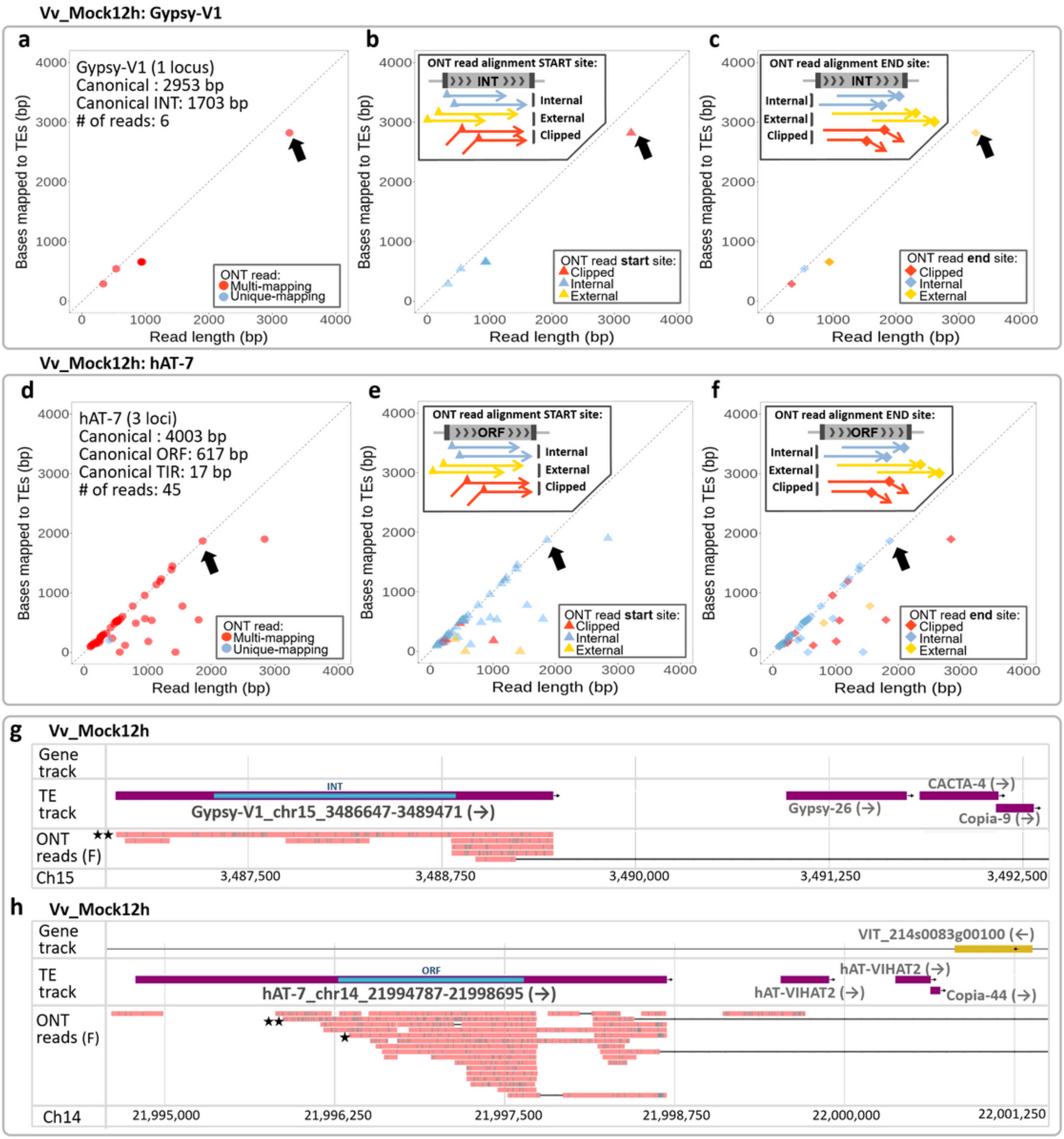


Figure 6

Identification of structurally autonomous TE loci potentially having transcripts for autonomous mobilisation. a-c The alignment properties of the ONT reads mapping to a structurally autonomous Gypsy-V1 locus whose INT domain was >90% covered by ONT reads. For each read, the read length was plotted against the number of bases overlapping with the autonomous locus and coloured by mapping specificity (multi-/unique-mapping; a) and alignment start site (b) and end site (c) relative to the TE locus.

All reads plotted in (a) were presented at the same coordinates in (b) and (c). The black arrows indicate an ONT read covering most of the structurally autonomous TE locus. Note that this ONT read is multi-mapping to other structurally autonomous Gypsy-V1 loci, but only the locus indicated here is TE expression candidate (i.e. supported by short-read data). d-f The alignment properties of the ONT reads mapping to three structurally autonomous hAT-7 loci whose ORF was >90% covered by ONT reads. (d)-(f) were plot following the approach used for (a)-(c). g-h Genome browser images of the structurally autonomous Gypsy-V1 locus (Gypsy-V1_chr15_3486647-3489471) and the representative hAT-7 locus (hAT-7_chr14_21994787-21998695), whose INT and ORF (teal blue strips) was fully covered by individual ONT reads, respectively. The two black stars label full coverage of the INT or ORF, while the single black star marks >90% coverage of the ORF. INT, internal domain; ORF, open reading frame; F, forward reads.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TableS1.xlsx
- Additionalfile2FigureS1S25.docx
- Additionalfile3TableS2.xlsx
- Additionalfile4TableS3.xlsx
- Additionalfile5TableS4.xlsx
- Additionalfile6TableS5.xlsx
- Additionalfile7TableS6.xlsx
- Additionalfile8Supplementarymethods.docx