Ventimiglia Maria (Orcid ID: 0000-0003-2718-7517)
Usai Gabriele (Orcid ID: 0000-0001-9982-4883)
Zuccolo Andrea (Orcid ID: 0000-0001-7574-0714)
Mascagni Flavia (Orcid ID: 0000-0001-9747-8040)

Genome-wide identification and characterisation of exapted transposable elements in the large genome of sunflower (*Helianthus annuus* L.)

Maria Ventimiglia[1]*, Giovanni Marturano[2]*, Alberto Vangelisti[1], Gabriele Usai[1], Samuel Simoni[1], Andrea Cavallini[1], Tommaso Giordani[1], Lucia Natali[1], Andrea Zuccolo[2,3] and Flavia Mascagni[1]

[1]Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, 56124 Pisa, Italy; [2]Crop Science Research Center, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà 33, 56127 Pisa, Italy; [3]Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

Authors for correspondence:
Flavia Mascagni
Email: flavia.mascagni@unipi.it
Andrea Zuccolo
Email: andrea.zuccolo@kaust.edu.sa
*These authors contributed equally to this work.

Summary

Transposable elements (TEs) are an important source of genome variability, playing many roles in the evolution of eukaryotic species. Besides well-known phenomena, TEs may undergo the exaptation process and generate the so-called exapted transposable element genes (ETEs). Here we present a genome-wide survey of ETEs in the large genome of sunflower (*Helianthus annuus* L.), in which the massive amount of TEs, provides a significant source for exaptation.

A library of sunflower TEs was used to build TE-specific Hidden Markov Model profiles, to search for all available sunflower gene products. In doing so, 20,016 putative ETEs were identified and further investigated for the characteristics that distinguish TEs from genes, leading to the validation of 3,530 ETEs.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.16078

The analysis of ETEs transcription patterns under different stress conditions showed a differential regulation triggered by treatments mimicking biotic and abiotic stress; furthermore, the distribution of functional domains of differentially regulated ETEs revealed a relevant presence of domains involved in many aspects of cellular functions. A comparative genomic investigation was performed including species representative of Asterids and appropriate outgroups: the bulk of ETEs resulted specific to the sunflower, while few ETEs presented orthologues in the genome of all analysed species, making the hypothesis of a conserved function.

This study highlights the crucial role played by exaptation, actively contributing to species evolution.

Significance statement: The molecular domestication of transposable elements leads to the formation of novel genes through exaptation. By setting up a protocol for systematically discovering exapted transposable elements (ETEs) in large genome species, we identified and validated 3,530 ETEs in sunflower genome. The identification of ETEs in *Helianthus annuus* can be considered further proof of the fundamental contribution that TEs had in the rising of genetic novelties, probably influencing different biological processes during the evolution of the sunflower.

Introduction

Transposable elements (TEs) are genomic sequences that can transfer themselves from their original chromosomal site to different ones; they constitute a substantial component of most eukaryotic genomes and play a key role in defining genome structure, function, and evolution (Rebollo *et al.*, 2010; Fedoroff, 2012; Bourque *et al.*, 2018; Dubin *et al.*, 2018).

TEs are divided into two classes. Class I TEs transpose via a replicative mechanism, which requires the formation of an RNA intermediate, and Class II TEs transpose through a 'cut & paste' conservative model. TIR elements are Class II TEs characterised by the presence of terminal inverted repeats (TIRs) and a transposase (TPase) as their only coding domain. Long terminal repeat retrotransposons (LTR-REs) are Class I TEs, characterised by the presence of long terminal repeats (LTRs) that flank the GAG-POL internal coding region, constituted by group-specific antigen (GAG), protease (PROT), reverse transcriptase (RT), ribonuclease H (RH), and integrase (INT). Elements belonging to this LTR-RE order are found in plant genomes divided into two superfamilies, *Copia* and *Gypsy*, which present the same coding domains that are displayed in a different order.

Due to their peculiar capability of moving from one locus to another, TEs may cause a wide range of conspicuous modifications in gene expression and function, such as the modelling of new regulatory networks and the creation of new genes (Feschotte, 2008; Cosby *et al.*, 2021). In addition to the well-known exon shuffling activity (Moran *et al.*, 1999; Feschotte and Wessler, 2001; Morgante *et al.*, 2005), another way TEs can originate novel genes is through the mechanism of exaptation (Joly-Lopez and Bureau, 2018). This process mediates the formation of new genes or sequences with new functions called Exapted Transposable Elements (ETEs) (Hoen and Bureau, 2015). Although transposition can be detrimental to the host genome stability, under particular circumstances these events may also produce positive effects, i.e., when the activity of a TE provides a selective advantage to the host, the TE may stop behaving as a self-replicative separate entity and begin to be kept in vertical inheritance through phenotypic selection. In some cases, TE-exaptation does not lead to the production of new complete genes (full molecular domestication): transduplication, for instance, is a mechanism by which a TE incorporates and mobilises a gene or gene fragments into a new genomic position leading often to the production of pseudogenes (Lisch, 2013). Another mechanism by which TEs can participate in molecular exaptation is called exonization, which occurs when a TE inserts into a gene or near a gene and it can be incorporated as a novel cassette exon (Hoen and Bureau, 2012). Other exaptation events affect gene regulation, leading to the production of new transcription factors or cis-regulatory sequences, that contribute to the establishment of novel patterns of gene expression (Wray, 2007).

Joly-Lopez and Bureau (2018) proposed a phased model to describe the TE exaptation mechanism. According to this model, a TE may eventually provide a phenotype beneficial to the host, and because of positive selection is then fixed in the genome over a short evolutionary time and finally undergoes persistence and diversification. A TE that, throughout evolution, has overcome all phases could be considered exapted.

In this context, ETEs can play important roles in host physiology. In plants, they are involved in flowering (Cowan *et al.*, 2005; Joly-Lopez *et al.*, 2012), development (Bundock and Hooykaas, 2005; Knip *et al.*, 2012; Knip *et al.*, 2013), light signaling (Hudson *et al.*, 2003; Lin *et al.*, 2007), and stress response (McClintock, 1984; Chénais *et al.*, 2012; Wheeler, 2013; Makarevitch *et al.*, 2015). For instance, Arabidopsis FAR1(far-red impaired response 1), the first ETE discovered in plants, derives from an ancient Mutator-like transposase, and modulates phyA-signaling homeostasis (Lin *et al.*, 2007). DAYSLEEPER, another ETE identified in *A. thaliana* (Bundock and Hooykaas, 2005), encodes a protein including conserved domains from

the hAT DNA transposon family (hobo, Ac, and Tam3) which acts as a transcription factor regulating several genes.

Class II TEs are particularly prone to be exapted (Feschotte and Pritham, 2007). This is because the transposase enzyme encoded by Class II elements contains DNA binding domains (DBDs) that bind to regulatory regions of the TE itself, thus modifying the expression of neighboring genes and giving rise to new regulatory elements and pathways (Feschotte, 2008). The exaptation of transposases has indeed occurred numerous times during the evolution of fungi, plants, and animals (Cowan *et al.*, 2005; Quesneville *et al.*, 2005; Babu *et al.*, 2006; Casola *et al.*, 2008).

One of the first systematic approaches aimed at identifying ETEs was developed by Hoen and Bureau (2015) for *A. thaliana*; their method exploits the different genetic attributes between TEs and ETEs, both from a structural and functional point of view. Some features remain common between a TE and its derived ETE; for example, highly conserved regions of the TE coding domains are also conserved in the derived ETE. The level of similarity that an ETE shares with its ancestral TE is sufficient to allow its identification but not to maintain the function of the TE. Moreover, ETEs are generally present in the genome with a lower number of copies than TEs, and most TEs are not expressed but silenced, commonly by small interfering RNAs (siRNAs), while ETEs are stably expressed and tend not to be a target for siRNAs (Cowan *et al.*, 2005; Jiao and Deng, 2007). However, no single genomic feature alone would allow a robust identification of ETEs. Because of this, the validation of putative ETEs should consider several parameters.

Despite the great importance for the evolution of genomes, the majority of known ETEs have been discovered mostly by chance, as single events (Hoen and Bureau, 2015), except for the work of Cowan *et al.* (2005) on *Oryza sativa* and that of Hoen and Bureau (2015) on *A. thaliana*, whereas the availability of an entirely sequenced high-quality genome allows this type of study to be conducted on a genome-wide level.

To address this paucity of genome-scale surveys of ETEs, in this work, we followed the Hoen and Bureau (2015) method, to devise an *ad hoc* bioinformatic pipeline, to investigate the exaptation events involving Class I TE LTR-REs, and Class II TE TIR elements of the cultivated sunflower (*Helianthus annuus* L.) whose genome size (3.6 Gbp, Badouin *et al.*, 2017) is by far larger than those of Arabidopsis or rice. Although a reference genome for sunflower was only recently completed (Badouin *et al.*, 2017), it has been known for many years that the sunflower genome (3.6 Gbp) is composed of a large percentage of repeated sequences, mostly represented by TEs, estimated from a minimum of 62% (Cavallini *et al.*, 2010) up to a

maximum of 81% (Mascagni *et al.*, 2015). LTR-REs constitute a very large fraction of the genome of sunflower (Giordani *et al.*, 2014; Mascagni *et al.*, 2017a), thus making these elements more subject, on a quantitative basis, to the phenomenon of exaptation.

The goals of this work were to (1) test a bioinformatic method to identify TE exaptation events leading to full molecular domestication in a species with a large genome, the sunflower; (2) localise the identified fully domesticated ETEs along the genome and investigate their expression pattern to gain insight into their potential function; (3) infer the evolutionary history of the identified ETEs.

Results

Genome-wide discovery of sunflower full-length TEs

The identification of ETEs was carried out by creating HMM profiles for LTR-REs and TIR elements coding domains starting from a comprehensive collection of 6,163 full-length LTR-REs and 7,472 full-length TIR elements. LTR-REs were divided into two main superfamilies, *Copia* and *Gypsy*, which in turn were arranged into nine lineages and five lineages, respectively (Wicker *et al.*, 2007; Neumann *et al.*, 2019; Mascagni *et al.*, 2020; Table 1). The annotation of TIR elements refers to the superfamily level, classifying them into five main superfamilies: *CACTA*, *hAT*, *Mutator*, *PIF-Harbinger*, and *Tc1-Mariner* (Wicker *et al.*, 2007; Table 1).

The estimation of the insertion time of sunflower LTR-RE *Copia* and *Gypsy* lineages (Supporting Information Fig. S1), shows that all lineages had a proliferation burst about one million years ago, except for the *Copia* lineage *Bianca*, whose elements appear to still be active and mobile.

Identification of ETEs in the sunflower genome

A complete autonomous element was selected for each lineage of LTR-RE and each superfamily of the TIR order (Supporting Information Table S2). For each of these prototypical elements, the translated sequences of their coding domains were retrieved. Altogether thus obtaining 14 GAG-POL sequences for the LTR-RE order, one for each lineage, and 5 TPase sequences for the TIR elements, one for each superfamily. It is to be noted that LTR-REs and TEs are highly variable in sequence, and that variations occur even within LTR-RE lineages and TIR superfamilies. However, the transposon coding portions are quite conserved, in fact they are used to efficiently assign an element to one lineage or another (Mascagni *et al.*, 2017b; Ventimiglia *et al.*, 2019). In this sense, the use of a prototypical sequence for each LTR-RE

lineage or TIR superfamily should ensure to cover most of the intra-lineage and intra-superfamily variations, respectively.

To build the HMM profiles, the GAG-POL and TPase aminoacidic sequences were split into fragments of 100 amino acids each, as shown in Figure 1. Each fragment was used as a query in similarity searches against the collection of TEs belonging to the same lineage/superfamily. All the positive matches were retrieved and aligned separately for each lineage/superfamily. Altogether 188 multiple sequence alignments were created and used to build the corresponding HMM profiles. The amino acid sequences of 75,390 sunflower nuclear gene products corresponding to 61,327 nuclear genes were then searched using the 188 HMM profiles.

The scan provided 27,178 matches in 21,841 sunflower nuclear genes, therefore considered as putative ETEs. 18,107 ETEs were derived from precursors belonging to the LTR-RE order and 1,909 from the TIR order. For 1,825 of these ETEs, it was not possible to infer the origin because they showed similarity with HMM profiles built on both TE orders, due to a possible artifact. This inconsistent subset was excluded from the further analyses, thus taking into consideration 20,016 genes as putative ETEs.

The 20,016 putative ETEs were evaluated based on their i) repetitiveness, ii) similarity with already known TEs, iii) siRNA coverage, and iv) expression, assigning scores ranging from a minimum of -2 to a maximum of +1, as reported in Supporting Information Table S1. In Figure 2, scores assigned to putative ETEs for each feature considered are reported.

All attribute data were summarised into a single total score per putative ETE. Only 3,530 ETEs (i.e., those scoring at least +3) were considered validated, corresponding to 5.8% of the 61,327 sunflower nuclear genes (Supporting Information Table S3). The highest possible score of +4 was reached by 954 of the validated ETEs (1.56% of all nuclear genes; Table 2). ETEs appeared to have been generated from any of the TE superfamilies considered in this study. For 409 (11.59%) of the validated ETEs, it was not possible to infer their superfamily of derivation; these not classifiable (NC) ETEs showed similarity with more than one LTR-RE superfamily (395 ETEs classified as LTR-RE_NC) or TIR superfamily (14 elements classified as TIR_NC). Figure 3 reports the distribution of scores and the proportion of the superfamilies from which the ETEs most likely originated.

Distribution of ETEs in sunflower chromosomes

In Figure 4, an overview of the distribution of validated ETEs across the 17 sunflower chromosomes is reported. The density of ETEs in 3 Mbp intervals, spanning the *H. annuus*

HanXRQ genomes (Badouin *et al.* 2017), was compared with the chromosomal localisation of sunflower repeats.

ETEs displayed distinct chromosomal distribution profiles compared to sunflower repeats. The difference resulted as significant by Pearson correlation (R squared = 0.3896, pvalue= 0.0001; Supporting Information Table S4 and Figure S2).

Functional characterisation of validated ETEs

The 3,530 validated ETEs encode 3,906 protein products of sunflower, 2,347 of which are assigned to at least one GO term (Conesa *et al.*, 2005; Usai *et al.*, 2017). The GO annotations were inspected separately for each principal class, namely Molecular Function, Biological Process, and Cellular Component (respectively MF, BP, and CC). Most of the MF terms were *binding* (GO:0005488) and *catalytic activity* (GO:0003824), whereas BP terms were not as skewed to few terms, with the most abundant being *cellular process* (GO:0009987) and *metabolic process* (GO:0008152). Regarding CC terms, most of ETEs are annotated as *cell* (GO:0005623) and *cell part* (GO:0044464) (Figure 5a). By means of enrichment analysis, the GO terms of the ETEs were compared to those of the entire transcriptome of *H. annuus*. Overall, nine GO terms were significantly overrepresented in sunflower ETEs (Figure 5b).

Possible regulation of ETEs in response to stress mimicking biotic and abiotic stimuli in sunflower roots was evaluated (Figure 6). A total of 1,499 ETE genes were detected as differentially regulated during the biotic/abiotic stress on sunflower roots (Supporting Information Table S5). Interestingly, especially IAA, ABA, and MeJA affect over and under expression of ETEs with respectively 898, 536, and 249 differentially regulated genes. On the contrary, some treatments showed no effect on ETEs regulation such as BRA, GA3, and STRI.

Finally, concerning the 1,499 differentially expressed ETEs, the distribution of functional protein domains was analysed using the PFAM database. Overall, we detected a major occurrence of "pentatricopeptide repeat (PPR) repeat" (87 ETEs), "PPR repeat family" (81 ETEs), and "protein kinase domain" (51 ETEs) as reported in Table 3.

Reconstructing the evolutionary patterns of ETEs

To infer the evolutionary history of sunflower ETEs we performed a comparative genomic investigation using lettuce and artichoke as representatives of Asterids II, coffee for Asterids I, and grapevine and Arabidopsis as outgroup species. From this synteny analysis, the bulk of ETEs resulted specific to sunflower and closely related species (i.e., lettuce and artichoke), indicating that ETEs dynamics are related to speciation events (Figure 7a).

Considering the ETE orthologs distribution across the genomes, 1,833 ETEs resulted specific to the sunflower, and closer species showed more orthologous ETEs, such as sunflower and lettuce (398 ETE orthologs) and sunflower and artichoke (265 ETE orthologs). Few ETEs orthologs were discovered in other situations. For instance, 24 ETE orthologs were found in each of the six analysed species (Figure 7b), an observation which might suggest they originated ancestrally and display conserved functions.

Discussion

Out of the many roles played by TEs in the evolution of eukaryotic genomes, one of the less investigated is their capability to provide raw material to originate new genes through the process of exaptation (Hoen and Bureau, 2015). In this work, we identified 20,016 putative ETEs, of which 3,530 were validated as ETEs, integrated into the sunflower genome with specific cellular functions. The few previous studies that aimed at the identification of ETEs on a genome-wide scale were conducted on *A. thaliana* and *O. sativa*, which have small/medium-sized genomes, and identified low numbers of ETEs (Cowan *et al.*, 2005; Hoen and Bureau, 2015). The high number of ETEs identified in this study probably reflects the large size of the sunflower genome and its abundance of TEs (Badouin *et al.*, 2017, Mascagni *et al.*, 2015, Natali *et al.*, 2012)

Compared to the genomic composition of TEs in sunflower, ETEs displayed a distinct relative abundance. For instance, the highly abundant *Mutator* superfamily gave rise to a low number of ETEs; in contrast, the *Copia* superfamily, moderately abundant, seemed to have given rise to many ETEs. This data raised the question of whether some TE domains were more inclined than others to undergo exaptation events and/or to be retained. Indeed, we observed that within our set of ETEs, the domains of *Copia* LTR-RE were the most prone to being exapted, showing an inverse correlation to the abundance of these superfamilies in the sunflower genome. This result was quite unexpected, as we know that most known ETEs are derived from DNA TEs (Hoen and Bureau, 2015), but it could be related to the huge abundance of retrotransposons in sunflower (Mascagni *et al.*, 2015).

The distribution of ETEs was assessed along sunflower chromosomes, and compared with the density of TEs, revealing how ETEs are enriched in euchromatic regions where the frequency of TEs is lower, whereas a high presence of repeats is related to a lower abundance of ETEs. It may be hypothesised that ETEs derived from TEs inserted in a euchromatic site are more prone to be evolutionary retained than those derived from elements inserted in the heterochromatin.

The functional characterisation of ETEs performed through GO enrichment analysis allowed us to highlight significant enrichment in the functional *binding* category and disparate other cellular functions. This heterogeneity suggests that, during the evolution of the sunflower genome, ETE genes have helped to fix several important biological processes in this species. It is known that ETEs played a crucial role in various biological systems, such as the control of development in plants (Bundock and Hooykaas, 2005; Knip *et al.*, 2012; Knip *et al.*, 2013). In eukaryotes, ETEs can also be involved in chromosome segregation, centromere binding, heterochromatin formation, meiotic recombination, TE silencing, chromosome stability, programmed chromosomal rearrangements, and translational regulation (Feschotte and Pritham, 2007; Sinzelle *et al.*, 2009). This underlines the importance of ETEs in genome evolution and function.

By probing ETEs expression in cDNA libraries from sunflower roots under different conditions, we were able to highlight the association of ETEs expression with stress response. We report a total number of 1,499 differentially expressed ETEs, that were investigated at the level of functional domains: the search of similarity to Pfam families in differentially expressed ETEs revealed that most of them are related to PPR domains, which are sequence-specific RNA-binding proteins present in large numbers in plant genomes (O'Toole *et al.*, 2008) and are involved in many aspects of RNA editing (Delannoy *et al.*, 2007; Nakamura and Kobayashi, 2012). Other highly represented categories are leucine-rich repeat (LRR) domain-containing proteins, found in a large gene family involved in plant defense (McHale *et al.*, 2006), and pentatricopeptide-repeat containing proteins, a very heterogeneous class of proteins, involved in RNA editing, with pleiotropic effect related to plant development and environmental adaptation (Barkan and Small, 2014). The fact that both these highly represented categories are involved in plant response to environmental stimuli is in line with the hypothesis that ETE formation and fixation is adaptive. Furthermore, also ETEs showing functional domain belonging to F-box protein (PF00646) could possibly be involved in plant defense since these motifs are capable to bind and cooperate with LRR (Kiperos and Pagano, 2000), nevertheless possible functions related to this functional domain could also be implied in signal transduction and regulation of cell cycle (Craig and Tyers, 1999). Finally, we retrieved specific domains involved in cellular response to stimuli, especially signal transduction and transcription factor activity. Several ETEs coding for protein tyrosine and serine/threonine kinase (PF07714) were found amongst differentially expressed genes, these elements are probably involved in triggering molecular signaling cascade deriving from growth (especially hormone stimulation) and physiological variations in response to stress as observed in *Arabidopsis thaliana*, *Dacus*

carota, and *Pisum sativum* (Hardie, 1999; Ghelis, 2011). In addition, Pfams related to transcription factor families, different from those retrieved in *A. thaliana* (Hoen and Bureau, 2015) and *O. sativa* (Cowan *et al.*, 2005) were retrieved: in particular, PF00249, representing a Myb-like DNA-binding domain conserved in plants (Kranz *et al.*, 2001). This is consistent with data reported for other species, i.e., about half of eukaryotic ETEs with putative or known functions are transcription factors (Feschotte and Pritham, 2007).

Finally, we attempted to reconstruct the evolutionary patterns of sunflower ETEs in related species representatives of Asterids. The bulk of ETEs resulted specific to the sunflower, suggesting that most sunflower putative ETEs originated after *Heliantheae* separation. As a matter of fact, the presence of orthologous ETEs correlates with the taxonomic distances of the species investigated, as lettuce and artichoke display a high number of shared orthologs which are not found in coffee and the outgroups. Moreover, there were a few ETEs shared among all species, likely representing ETEs formed ancestrally and vertically inherited in the descendants.

In conclusion, the identification of sunflower ETEs can be considered further evidence of the fundamental contribution that TEs had to the rise of genetic novelties, probably influencing different biological processes during the evolution of sunflower, which seems to have resulted from the functional analysis of the identified genes. The results obtained could be used in future studies to screen candidate targets, also assessing the phenotypical effects of specific ETE disruption.

Experimental procedures

Collection of TEs

A collection of LTR-RE and TIR elements was isolated from the HanXRQr2.0-SUNRISE version of the *H. annuus* genome, deposited in NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCA_002127325.2), using software that are based on the identification of TE structural characteristics, as following: LTR-REs were first retrieved by LTRharvest (included in GenomeTools v1.5.9 (Ellinghaus *et al.*, 2008) with the following parameter settings: '-minlenltr 100 -maxlenltr 10000 -mindistltr 1500 -maxdistltr 25000 -mintsd 5 -maxtsd 5 -motif tgca -vic 10' and subsequently characterised at lineage level by searching coding domains within the sequences, using the tool Domain based ANnotation of Transposable Elements (DANTE) on the Galaxy platform (http://galaxyproject.org). The collected information was then parsed using an in-house built script to detect and filter out nested elements. Only LTR-REs complete of all coding domains were included in the library.

TIR elements belonging to the five major plant TIR superfamilies were identified using the software TIR-Learner, included in the EDTA package, with default options (Ou *et al.*, 2019).

The abundance of each LTR-RE lineage and each TIR superfamily was estimated by mapping Illumina reads of sunflower (SRR5004633, available at https://www.ncbi.nlm.nih.gov/sra/SRR5004633) onto the reference TE library. Mapping was performed using CLC Genomics Workbench v9.5.3 (CLC-BIO, Aarhus, Denmark), with the following parameters: mismatch cost = 1, deletion cost = 1, insertion cost = 1, similarity = 0.9, and length fraction = 0.9. Abundance values were reported as the percentage of mapped reads per total reads.

The insertion time of LTR-REs was estimated by comparing the two LTRs of each element (SanMiguel *et al.*, 1998). The two LTRs were first aligned using the Stretcher program (EMBOSS package v6.6.0.0; Rice *et al.*, 2000), and then the nucleotide distances between the LTRs were measured using the Kimura two-parameter method (K2P; Kimura, 1980), implemented in the Distmat program (EMBOSS package; Rice *et al.*, 2000), applying a synonymous substitution rate that is twice that calculated for sunflower genes *i.e.*, $2*10^{-8}$ (Mascagni *et al.*, 2017a), according to SanMiguel *et al.* (1998).

Prediction of ETEs

To identify and retrieve the sequences of GAG-POL proteins encoded by LTR-REs and TPase proteins encoded by TIR elements, we used the Pfam search tool v33.1 (Bateman *et al.*, 2004); examining the TEs belonging to our library, the Pfam protein domain prediction tool was used to choose 19 prototypical sequence elements: 14 (one for each LTR-RE lineage) presenting a GAG-POL region complete of all the protein domains, and 5 (one for each TIR superfamily) with an intact TPase domain.

The prototypical sequences were then split into segments of 100 amino acids in length then used to perform a tblastn search with the BLAST tool v2.6.0+ (Altschul *et al.*, 1997) on the pool of elements of the same lineage/superfamily (threshold e-value 1e-5).

The significant matches were retrieved and aligned using MUSCLE v3.8.31 (Edgar, 2004). The multiple sequence alignments were then processed using the tool *hmmbuild* (with default options), which is part of HMMER v3.3 software (Finn *et al.*, 2011), generating a Hidden Markov Model (HMM) profile for each multi-alignment. The HMM profiles were used to search all the gene products of sunflower (retrieved from https://www.ncbi.nlm.nih.gov/genome/?term=txid4232[orgn]) using *hmmscan* (HMMER), thus identifying the sunflower genes

showing correspondence with the HMM profiles. Of these, only the nuclear genes which showed similarity to a single order of elements (LTR-REs or TIR elements), were retained as putative ETEs.

Validation of ETEs

To figure out the reliability of ETE predictions, a score was computed evaluating four attributes for each putative ETE: repetitiveness, similarity with already known TEs, siRNA coverage, and expression (Supporting Information Table S1). The scores of each attribute were added up for every ETE giving a cumulative score, then compared to a significance threshold to distinguish the validated ETEs.

Repetitiveness: Genomic reads of sunflower (SRR5004633, https://www.ncbi.nlm.nih.gov/sra/SRR5004633) were mapped with CLC Genomics Workbench v9.5.3 (mismatch cost 1, insertion cost 1, deletion cost 1, length fraction 0.9, similarity fraction 0.9) onto the DNA sequences of the putative ETEs. As a comparison, our TE library was analysed to obtain the *average coverage* values used as an estimate of the repetitiveness (Natali *et al.*, 2013). The thresholds were chosen as follows: as *A* and *B* correspond to the first and third quartiles of the *average coverage* distribution resulting from the mapping that occurred on TEs only, ETEs with *average coverage* < *A* were assigned a score of +1, those between *A* and *B* were assigned a value of 0, and those resulting > *B* were assigned a score of -1. For putative ETEs with higher *average coverage* we also evaluated the repetitiveness of the specific region, which showed similarity with an HMM profile (and therefore with a TE), also called a 'matching region.' The HMM matching regions of each ETE were then retrieved and searched through tblastn in the sunflower genome. Matching regions with significant alignments (identity percentage 70.00, length 70, e-value 1e-5) were assigned a further penalty (-2).

Similarity with known TEs: Three BLAST searches (thresholds: identity 70%, length 70%, e-value 1e-5) were performed: i) blastn of putative ETEs against the TE library to identify which of the putative ETEs had similarities with known TEs over the majority of its length (putative ETEs not showing similarity with TEs were assigned score +1); ii) blastn using as queries the exons of putative ETEs providing significant hits in i). Putative ETEs whose exons provided at least one significant match were considered to possess similarity along the coding portions (score -1), otherwise the similarity was limited to the intronic regions (score 0). iii) tblastn of the HMM matching regions of putative ETEs providing significant hits in ii) against the TE library. ETEs showing high similarity with TEs within the exon portion corresponding to the matching region were assigned a score of -2.

siRNA coverage: 482 siRNAs identified by Badouin *et al.* (2017) were mapped onto all sunflower ETEs using BWA v0.7.17-r1188 (aln/samse algorithm with the -n parameter set to 500) (Li and Durbin, 2009). Putative ETEs not targeted by siRNAs were assigned a score of +1. For the others, it was determined whether mapping occurred in the intronic regions (score 0) or the exonic regions repeating the BWA mapping using only exon regions. We chose to penalise a putative ETE that mapped a siRNA in the exonic regions (score -2).

Expression: Putative ETE expression was evaluated based on the average RPKM (reads per kilobase per million mapped reads) value calculated using three cDNA libraries of sunflower leaves (SRR4996792, SRR4996801, SRR4996807) that are publicly available at NCBI SRA (accession number SRP092742, available at https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP092742). The cDNA reads were mapped on a reference constituted by all sunflower genes and the TE library using the CLC Genomics workbench v. 9.5.3 with the following settings: mismatch cost 1, insertion cost 1, deletion cost 1, length fraction 0.9, similarity fraction 0.9). ETEs with RPKM > 1 were assigned a score of +1; ETEs with RPKM between 0 and 1 were assigned a score of 0; ETEs with RPKM = 0 were assigned a score of -1.

Finally, the TE domains that were more prone to be exapted were identified by reporting the percentage of the domains of *Copia, Gypsy,* and TIR elements involved in exaptation processes. ETEs derived by *Copia* LTR-REs, *Gypsy* LTR-REs, and TIR elements were investigated separately.

Genomic localisation of ETEs

To analyse the genomic localisation of ETEs along sunflower chromosomes, ETE distribution was correlated with the abundance of repeated sequences. First, the genome assembly was split into 3-Mbp bins using BEDTools v2.27.1 (Quinlan and Hall, 2010). Then the genome was masked using RepeatMasker v4.0.3 (Smit *et al.*, 2013-2015) fed with the obtained sunflower repeat libraries (Mascagni *et al.*, 2015). The process was run with default parameters. Finally, ETE locations were intersected with the masking results using BEDTools. Pearson correlation on ETE and masking data was performed using GraphPad Prism v9.0.0 (GraphPad Software, Inc., La Jolla, CA, USA). A graphical representation of the data was produced using the "ggplot2" R package (Wickham, 2016).

Functional characterisation of ETEs

Gene Ontology (GO) term annotation of all sunflower gene products was obtained by submitting the sequences to InterProScan v5.45-80.0 (Jones *et al.*, 2014) against the Pfam database (Sonnhammer *et al.*, 1997). The distribution of GO terms associated with validated ETE was visualised with WEGO v2.0 (Ye *et al.*, 2018). In addition, Fisher's exact test was performed comparing the GO terms of validated ETEs against the complete GO set of sunflower genes using Blast2GO (Conesa *et al.*, 2005), hence, in order to reduce the complexity arising from the GO terms, we performed REVIGO (Supek *et al.*, 2011) using the 'tiny similarity' parameter and selecting only GO terms that had a distribution higher than 10%.

Expression of ETEs was analysed in available Illumina cDNA libraries from roots exposed to treatments mimicking stress, available under project accession SRP092742 of SRA. In particular, the libraries concerning the treatments of sunflower roots with abscisic acid (ABA), ethylene (ACC), brassinosteroids (BRA), gibberellic acid (GA3), auxin (IAA), methyl jasmonate (MeJA), sodium chloride (NaCl), polyethylene glycol (PEG), salicylic acid (SA), strigolactones (STRI) and kinetin (KIN), along with 6 control libraries were chosen. Reads from cDNA libraries were trimmed using Trimmomatic v0.38 (Bolger *et al.*, 2014), removing low-quality reads and adapters. Then, high-quality reads were mapped on the sunflower transcriptome using the CLC Genomics Workbench with the following parameters: mismatch penalties = 2, gap open penalties = 3, length fraction = 0.9, and similarity fraction = 0.9. Expression values for reads mapped for each gene were normalised with reads per kilobase of exon per million mapped reads (RPKM) (Mortazavi *et al.*, 2008).

Differentially expressed genes (DEGs) were retrieved by comparing treatments versus respective control libraries. DEGs were obtained using EdgeR (Robinson *et al.*, 2010) with a quasi-likelihood statistical test. Genes with an absolute fold change value over 2 and false discovery rate (FDR)-corrected p-value under 0.05 (Benjamini *et al.*, 1995) were considered differentially expressed. Transcripts that showed an RPKM value < 1 in each library were not considered for differential expression analysis.

Finally, we examined the distribution of the Pfam families of differentially expressed ETEs.

Phylogenetic inference of orthology

A comparative analysis was performed between sunflower and three species representatives of Asterids: *Lactuca sativa* (Reyes-Chin-Wo *et al.*, 2017), *Cynara cardunculus* var. *scolymus* (Scaglione *et al.* 2016), and *Coffea arabica* (Denoeud *et al.*, 2014). Furthermore, *Vitis vinifera* (Jaillon *et al.*, 2007) and *Arabidopsis thaliana* (Pucker *et al.*, 2016) were used as

outgroups. The ETEs' orthology relationships were investigated using OrthoFinder version 2.5.4 (Emms and Kelly, 2019). Only sequences displaying shared synteny, as computed with MCScanX (Wang *et al*., 2012), were considered orthologs.

Acknowledgements

The authors acknowledge funding from the Department of Agriculture, Food, and Environment of the University of Pisa, Italy, Project "Plantomics"

Author Contributions

Conceptualisation, AC, F.M. and A.Z.; data curation and methodology, M.V., G.M. and F.M.; investigation, M.V., G.M., A.V., G.U. and S.S.; M.V., G.M., A.V., G.U., S.S., A.C., T.G., L.N., A.Z. and F.M. discussed the data, wrote the manuscript, and contributed to its final form.

Conflict of interest

The authors declare that they have no competing interests.

Data availability

The data that support the findings of this study are available in NCBI at https://www.ncbi.nlm.nih.gov/assembly/GCA_002127325.2, https://www.ncbi.nlm.nih.gov/genome/?term=txid4232[orgn], https://www.ncbi.nlm.nih.gov/sra/SRR5004633, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP092742. ETE discovery pipeline is available at: https://github.com/ETEdiscovery2022/ETE_discovery

Short legends for Supporting Information

Figure S1 Insertion time estimated for *Copia* (a) and *Gypsy* (b) LTR-RE lineages in *H. annuus* genome **Figure S2** *H. annuus* ETEs' distinct chromosomal distribution profiles compared to sunflower repeats. Pearson correlation between the percentage of TE masked bases in the genome, and the density of ETEs in 3 Mbp intervals.

Table S1 Genetic attributes and scoring system. (i): evidence associated with the intron; (e): evidence associated with the exon; (!): evidence associated with the region sharing similarity with the HMMs profile.

Table S2 Prototypical sequence elements chosen for each lineage of the LTR-RE order and

each superfamily of the TIR order in *H. annuus*.

Table S3 List of 3,530 validated ETEs discovered in *H. annuus* genome.

Table S4 Pearson correlation between the density of ETEs in 3 Mbp intervals of the *H. annuus* genome, and the chromosomal localisation of sunflower repeats. The difference between the distribution profiles resulted significant (R squared = 0.3896, pvalue= 0.0001).

Table S5 Differentially expressed ETEs in response to biotic/abiotic stresses in *H. annnus* roots.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Babu, M.M., Iyer, L.M., Balaji, S. and Aravind, L. (2006) The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res.* **34**, 6505-6520.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B. *et al.* (2017) The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, **546**, 148-152.

Barkan, A. and Small, I. (2014) Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* **65**, 415-442.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L. *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res.* 32, D138-D141.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289-300.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S. *et al.* (2018) Ten things you should know about transposable elements. *Genome Biol.* **19**, 1-12.

Bundock, P. and Hooykaas, P. (2005) An *Arabidopsis hAT*-like transposase is essential for plant development. *Nature*, **436**, 282-284.

Casola, C., Hucks, D. and Feschotte, C. (2008) Convergent domestication of *pogo*-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.* **25**, 29-41.

Cavallini, A., Natali, L., Zuccolo, A., Giordani, T., Jurman, I., Ferrillo, V., Vitacolonna, N., Sarri, V., Cattonaro, F., Ceccarelli, M. *et al.* (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. *Theor. Appl. Genet.* **120**, 491-508.

Chénais, B., Caruso, A., Hiard, S. and Casse, N. (2012) The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7-15.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676.

Cosby, R.L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E.J. and Feschotte, C. (2021) Recurrent evolution of vertebrate transcription factors by transposase capture. *Science*, **371**, eabc6405.

Cowan, R.K., Hoen, D.R., Schoen, D.J. and Bureau, T.E. (2005) *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. *Mol. Biol. Evol.* **22**, 2084-2089.

Craig, K.L., Tyers, M 1999. The F-box: a new motif for ubiquitin dependent proteolysis in cell cycle regulation and signal transduction. *Progress in Biophysics and Molecular Biology*, **72**: 299-328.

Delannoy, E., Stanley, W.A., Bond, C.S. and Small, I.D. (2007) Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochem. Soc. Trans.* **35**, 1643-1647.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G. *et al.* (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, **345**:1181-1184.

Dubin, M.J., Scheid, O.M. and Becker, C. (2018) Transposons: a blessing curse. *Curr. Opin. Plant Biol.* **42**, 23-29.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids* Res. **32**, 1792-1797.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 1-14.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1-14.

Fedoroff, N.V. (2012) Transposable elements, epigenetics, and genome evolution. *Science,* **338**, 758-767.

Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397-405.

Feschotte, C. and Pritham, E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331-368.

Feschotte, C. and Wessler, S.R. (2001) Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **98**, 8923-8924.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-W37.

Ghelis, T. (2011) Signal processing by protein tyrosine phosphorylation in plants. *Plant Signal. Behav*. **6**, 942-951.

Giordani, T., Cavallini, A. and Natali, L. (2014) The repetitive component of the sunflower genome. *Curr. Plant Biol.* **1**, 45-54.

Hardie, D.G. (1999) Plant protein serine/threonine kinases: classification and functions. *Annu. Rev. Plant Biol.* **50**, 97-131.

Hoen, D.R. and Bureau, T.E. (2012) Transposable element exaptation in plants. In: *Plant transposable elements*. Springer, Berlin, Heidelberg, pp. 219-251.

Hoen, D.R. and Bureau, T.E. (2015) Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Mol. Biol.Evol.* **32**, 1487-1506.

Hudson, M.E., Lisch, D.R. and Quail, P.H. (2003) The *FHY3* and *FAR1* genes encode transposaserelated proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *Plant J.* **34**, 453-471.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. *et al.* (2007) French-Italian Public Consortium for Grapevine Genome Characterization The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463-467.

Jiao, Y. and Deng, X.W. (2007) A genome-wide transcriptional activity survey of rice transposable element-related genes. *Genome Biol.* **8**, 1-19.

Joly-Lopez, Z., Forczek, E., Hoen, D.R., Juretic, N. and Bureau, T.E. (2012) A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in Arabidopsis thaliana. *PLoS Genet.* **8**, e1002931.

Joly-Lopez, Z. and Bureau, T.E. (2018) Exaptation of transposable element coding sequences. *Curr. Opin. Genet. Dev.* **49**, 34-42.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. *et al.* (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics,* **30**, 1236-1240.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.

Kipreos, E.T. and Pagano, M. (2000) The F-box protein family. *Genome Biol.* **1**, 1-7.

Knip, M., de Pater, S. and Hooykaas, P.J. (2012) The *SLEEPER*genes: a transposase-derived angiosperm specific gene family. *BMC Plant Biol.* **12**, 1-15.

Knip, M., Hiemstra, S., Sietsma, A., Castelein, M., de Pater, S. and Hooykaas, P. (2013) DAYSLEEPER: a nuclear and vesicular-localized protein that is expressed in proliferating tissues. *BMC Plant Biol.* **13**, 1-11.

Kranz, H., Scholz, K. and Weisshaar, B. (2000) c-MYB oncogene-like genes encoding three MYB repeats occur in all major plant lineages. *Plant J.* **21**, 231-235.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics,* **25**, 1754-1760.

Lin, R., Ding, L., Casola, C., Ripoll, D.R., Feschotte, C. and Wang, H. (2007) Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science*, **318**,1302-1305.

Lisch, D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49-61.

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., RossElbarra, J. and Springer, N.M. (2015). Transposable elements contribute to activation of maize genes in response to abiotic stress. *PloS Genet.* **11**, e1004915.

Mascagni, F., Barghini, E., Giordani, T., Rieseberg, L.H., Cavallini, A. and Natali, L. (2015) Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. *Genome Biol. Evol.* **7**, 3368-3382.

Mascagni, F., Giordani, T., Ceccarelli, M., Cavallini, A. and Natali, L. (2017a) Genome-wide analysis of LTR retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). *BMC Genom.* **18,** 1-16.

Mascagni, F., Cavallini, A., Giordani, T. and Natali, L. (2017b) Different histories of two highly variable LTR retrotransposons in sunflower species. *Gene*. **634**, 5-14.

Mascagni, F., Vangelisti, A., Usai, G., Giordani, T., Cavallini, A. and Natali, L. (2020) A computational genome-wide analysis of long terminal repeats retrotransposon expression in sunflower roots (*Helianthus annuus* L.). *Genetica*, **148**, 13-23.

McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792-801.

McHale, L., Tan, X., Koehl, P. and Michelmore, R.W. (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7**, 1-11.

Moran, J.V., DeBerardinis, R.J. and Kazazian, Jr H.H. (19999 Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530-1534.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997-1002.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628.

Nakamura, T., Yagi, Y. and Kobayashi, K. (2012) Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. *Plant Cell Physiol.* **53**, 1171-1179.

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg, L. *et al.* (2013) The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genom.* **14**, 1-14.

Neumann, P., Novák, P., Hoštáková, N. and Macas, J. (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA,* **10**, 1-17.

O'Toole, N., Hattori, M., Andres, C., Iida, K., Lurin, C., Schmitz-Linneweber, C., Sugita, M., Small, I. (2008) On the Expansion of the Pentatricopeptide Repeat Gene Family in Plants. *Mol. Biol. Evol.* **25**, 1120-1128.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T. *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1-18.

Pucker, B., Holtgräwe, D., Rosleff Sörensen, T., Stracke, R., Viehöver, P. and Weisshaar, B. (2016) A de novo genome sequence assembly of the *Arabidopsis thaliana* accession Nie-derzenz-1 displays presence/absence variation and strong synteny. *PLoS One*, **11**, e0164321.

Quesneville, H., Nouaud, D., and Anxolabehere, D. (2005) Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the *P*-transposable element. *Mol. Biol. Evol.* **22**, 741-746.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

Rebollo, R., Horard, B., Hubert, B. and Vieira, C. (2010) Jumping genes and epigenetics: towards new species. *Gene*, **454**, 1-7.

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D.O., Truco, M.J. *et al.* (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* **8,** 1-11.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277.

Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2010) EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998)**.** The paleontology of intergene retrotransposons of maize. *Nature Genet.* **20**, 43-45.

Scaglione, D., Reyes-Chin-Wo, S., Acquadro, A., Froenicke, L., Portis, E., Beitel, C., Tirone, M., Mauro, R., Lo Monaco, A., Mauromicale, G. *et al.* (2016) The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Sci. Rep.* **6**, 1-17.

Sinzelle, L., ZIzsvak, Z. and Ivics, Z. (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* **66**, 1073-1093.

Smit, A.F.A., Hubley, R. and Green, P. (2013-2015) RepeatMasker Open-4.0.

Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405-420.

Supek, F., Bošnja, M., Škunca, N. and Šmuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.

Usai, G., Mascagni, F., Natali, L., Giordani, T. and Cavallini, A. (2017) Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. *Tree Genet. Genomes*, **13**, 1-12.

Ventimiglia, M., Pugliesi, C., Vangelisti, A., Usai, G., Giordani, T., Natali, L., Cavallini, A. and Mascagni, F. (2020). On the trail of *Tetu1*: genome-wide discovery of CACTA transposable elements in sunflower genome. *Int. J. Mol. Sci*, **21(6)**, 2021.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids* Res. **40,** e49-e49.

Wheeler, B.S. (2013) Small RNAs, big impact: small RNA pathways in transposon control and their effect on the host stress response. *Chromosome Res.* **21**, 587-600.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982.

Wickham, H. (2016) Data analysis. In: *ggplot2.* Springer, Cham: 189-201.

Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206-216.

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A. *et al.* (2018) WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Res.* **46**, W71-W75.

Tables

Table 1 Number and genomic abundance of LTR-RE and TIR elements included in the TE library used for ETE identification in *H. annuus.*

Transposable elements	No. of elements	Genomic abundance (% of

				mapped reads)
Order	Superfamily	Lineage		
LTR-RE			**6,163**	**31.03**
	Copia		**1,745**	**8.28**
		Ale	443	0.89
		Alesia	7	0.002
		Angela	266	1.04
		Bianca	65	0.21
		Ikeros	462	1.35
		Ivana	197	0.34
		SIRE	179	3.98
		TAR	50	0.20
		Tork	76	0.27
	Gypsy		**4,418**	**22.75**
		Chromovirus|CRM	88	0.15
		Chromovirus| Reina	190	0.25
		Chromovirus|Tekay	407	9.26
		non-Chromovirus|OTA|Athila	603	2.40
		non-Chromovirus|OTA|Tat	3,130	10.69
TIR			**7,472**	**15.10**
	CACTA		**924**	**1.90**
	hAT		**2,197**	**2.00**
	Mutator		**4,076**	**10.43**
	PIF-Harbinger		**205**	**0.52**
	Tc1-Mariner		**70**	**0.25**

Table 2 Result of the validation process for putative ETEs in *H. annuus* genome. Evaluated genes were subdivided according to the TE superfamily from which they putatively originated.

	Copia	***Gypsy***	**LTR-RE_NC**	***CACTA***	***hAT***	***Mutator***	***PIF-Harbinger***	***Tc1-Mariner***	**TIR_NC**	**TOTAL**
Nr. of evaluated genes	8,357	4,867	4,883	96	482	108	616	39	568	20,016
Nr. of excluded genes	6,460	3,827	4,488	71	384	89	585	28	554	16,486
Nr. of candi-date ETEs	1,897	1,040	395	25	98	19	31	11	14	3,530

(of which having max score)	(505)	(302)	(91)	(5)	(35)	(6)	(4)	(3)	(3)	(954)

Table 3 Most represented PFAM codes in differentially expressed *H. annuus* ETEs. The number of ETEs carrying each PFAM domain is reported.

PFAM accession code	**PFAM corresponding name**	**Number of ETEs showing PFAM domain**
PF01535	PPR repeat	87
PF13041	PPR repeat family	81
PF00069	Protein kinase domain	51
PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	28
PF00249	Myb-like DNA-binding domain	27
PF12854	PPR repeat (additional variant)	22
PF13855	Leucine rich repeat	22
PF00646	F-box domain	21
PF07714	Protein tyrosine and serine/threonine kinase	21
PF13812	Pentatricopeptide repeat domain	21

Figures & Figure legends

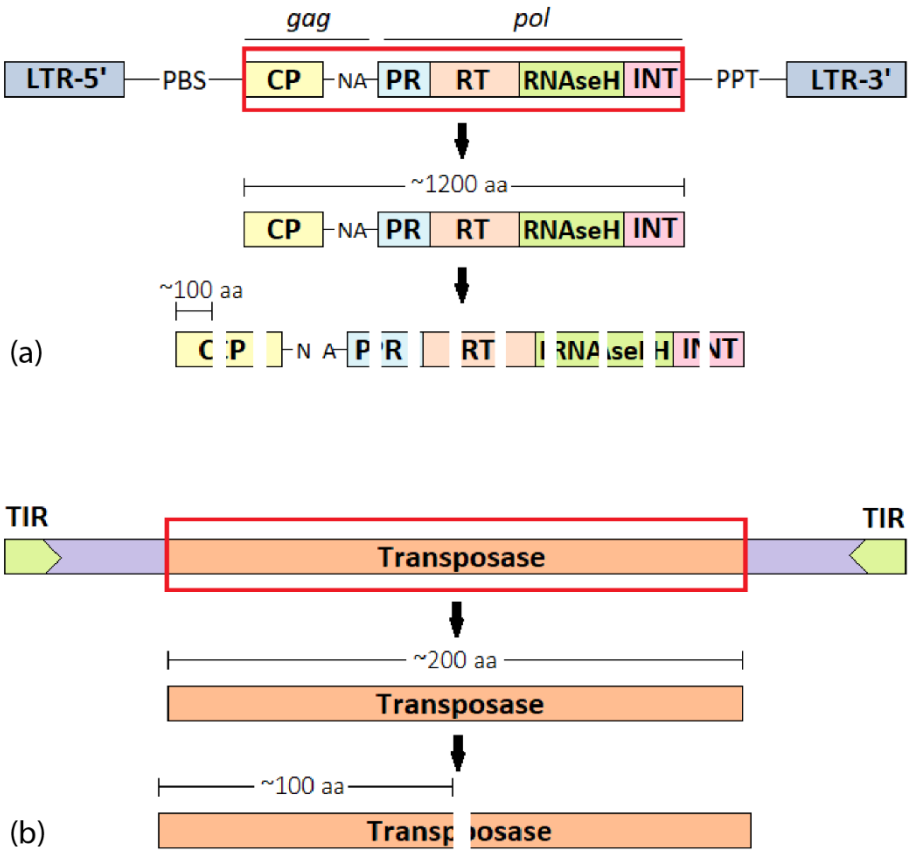

Figure 1 Representation of the strategy adopted to split GAG-POL (a) and TPase (b) amino acid sequences into fragments of 100 amino acids each.

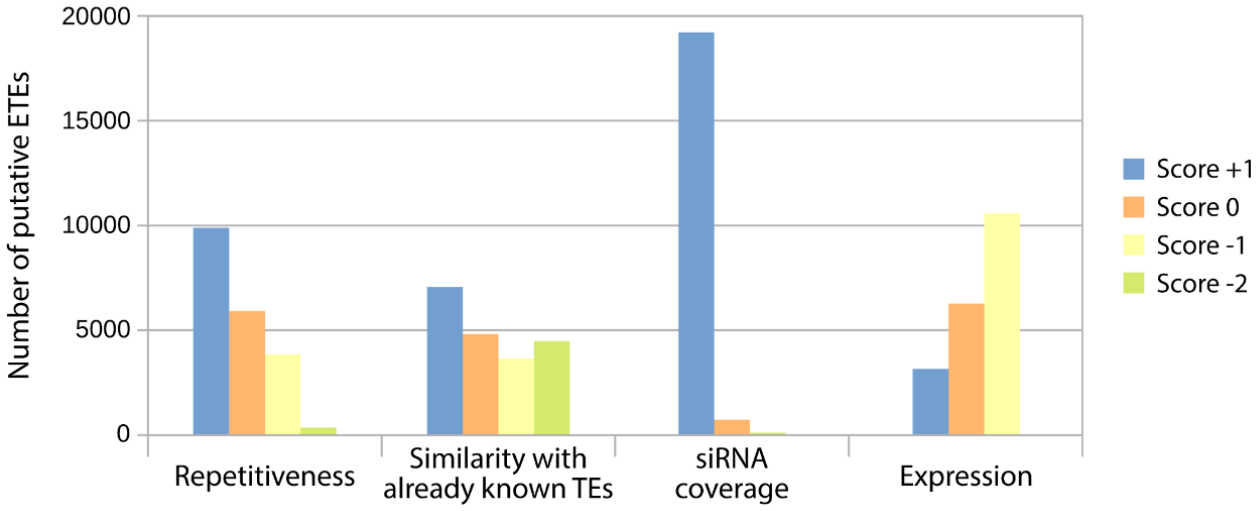

Figure 2 Scores assigned to putative ETEs for each term of validation in *H. annuus* genome.

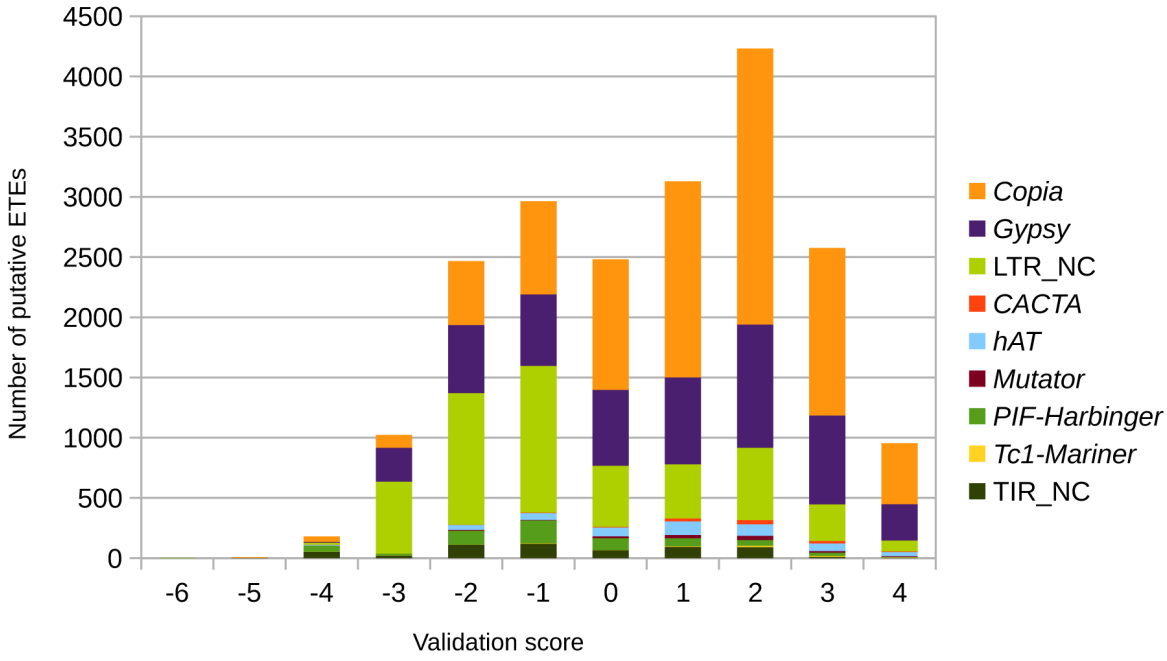

Figure 3 Distribution of the validation scores of putative ETEs in *H. annuus* genome; colours shown in the legend refer to the superfamilies that likely generated the putative ETEs.

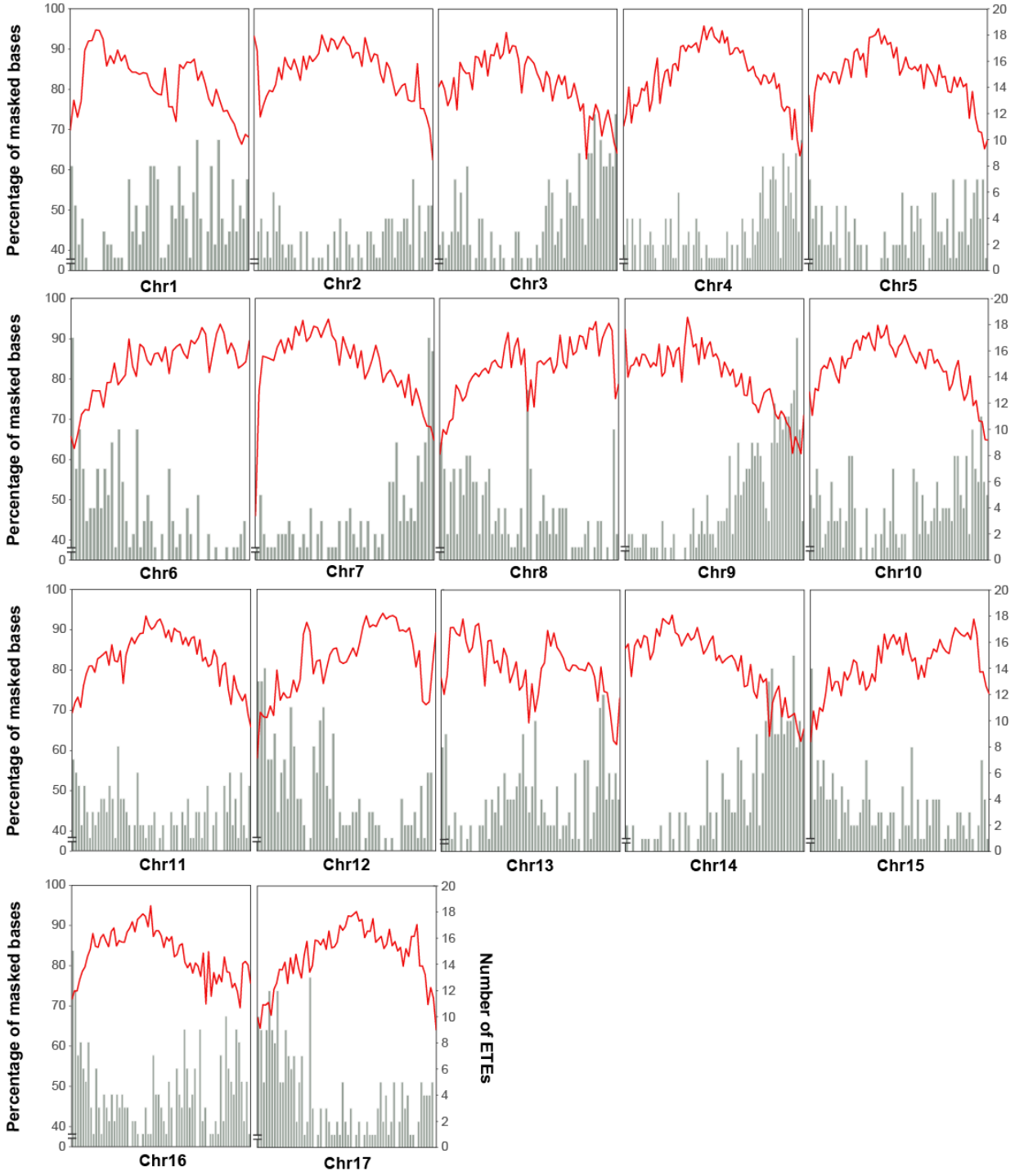

Figure 4 Genomic localisation of ETEs and TEs in *H. annuus* genome. For each chromosome, the histograms report the ETE counts in 3 Mbp intervals, with the red line reporting the percentage of TE masked bases.

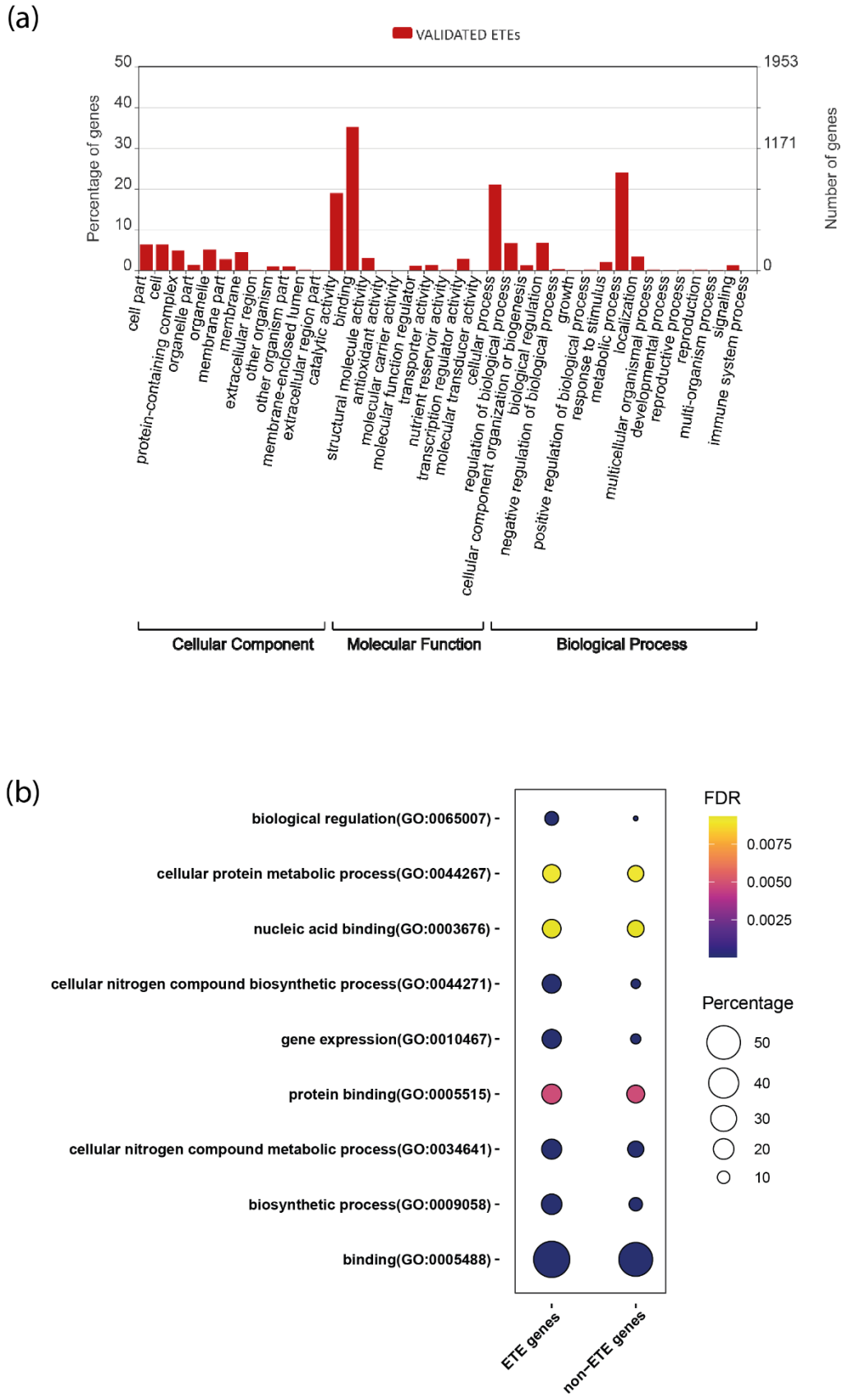

Figure 5 GO term distribution for validated ETEs in *H. annuus* genome (a). Enrichment analysis between non-ETE genes and validated ETEs in sunflower. On the x-axis GO terms with

respective id between brackets are shown. Colours scale is based on adjusted PValue (FDR), whereas balloon size reflects percentage distribution of gene per each GO class. Only significant GOs overrepresented in ETEs ($P < 0.05$) are reported (b).

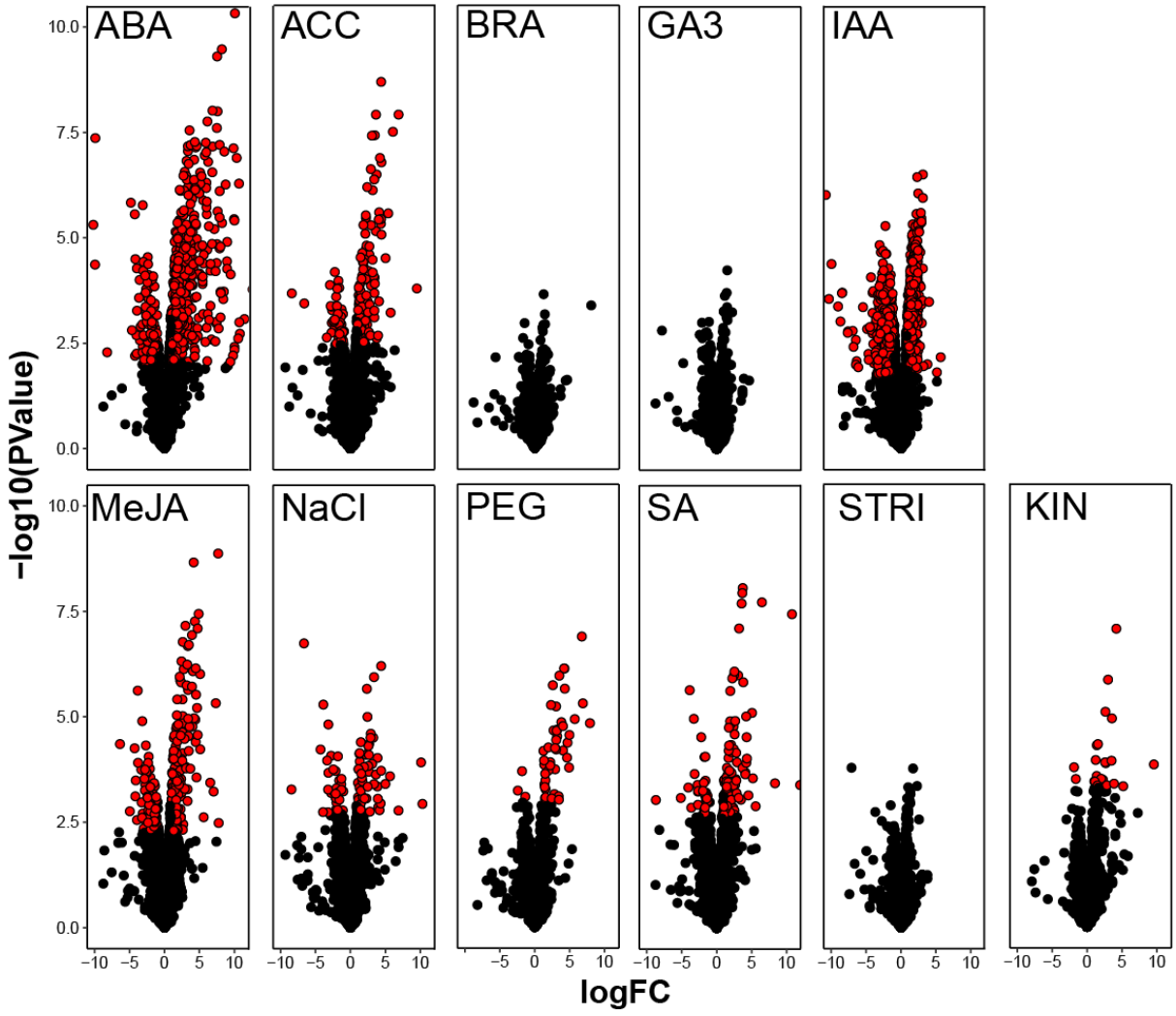

Figure 6 Volcano plots showing ETEs differentially expressed in roots of *H. annuus* plantlets treated with different substances (ABA: abscisic acid; ACC: ethylene; BRA: brassinosteroids; GA3: gibberellic acid; IAA: auxin; MeJA: methyl jasmonate; NaCl: sodium chloride; PEG: polyethylene glycol; SA: salicylic acid; STRI: strigolactones; KIN: kinetin).

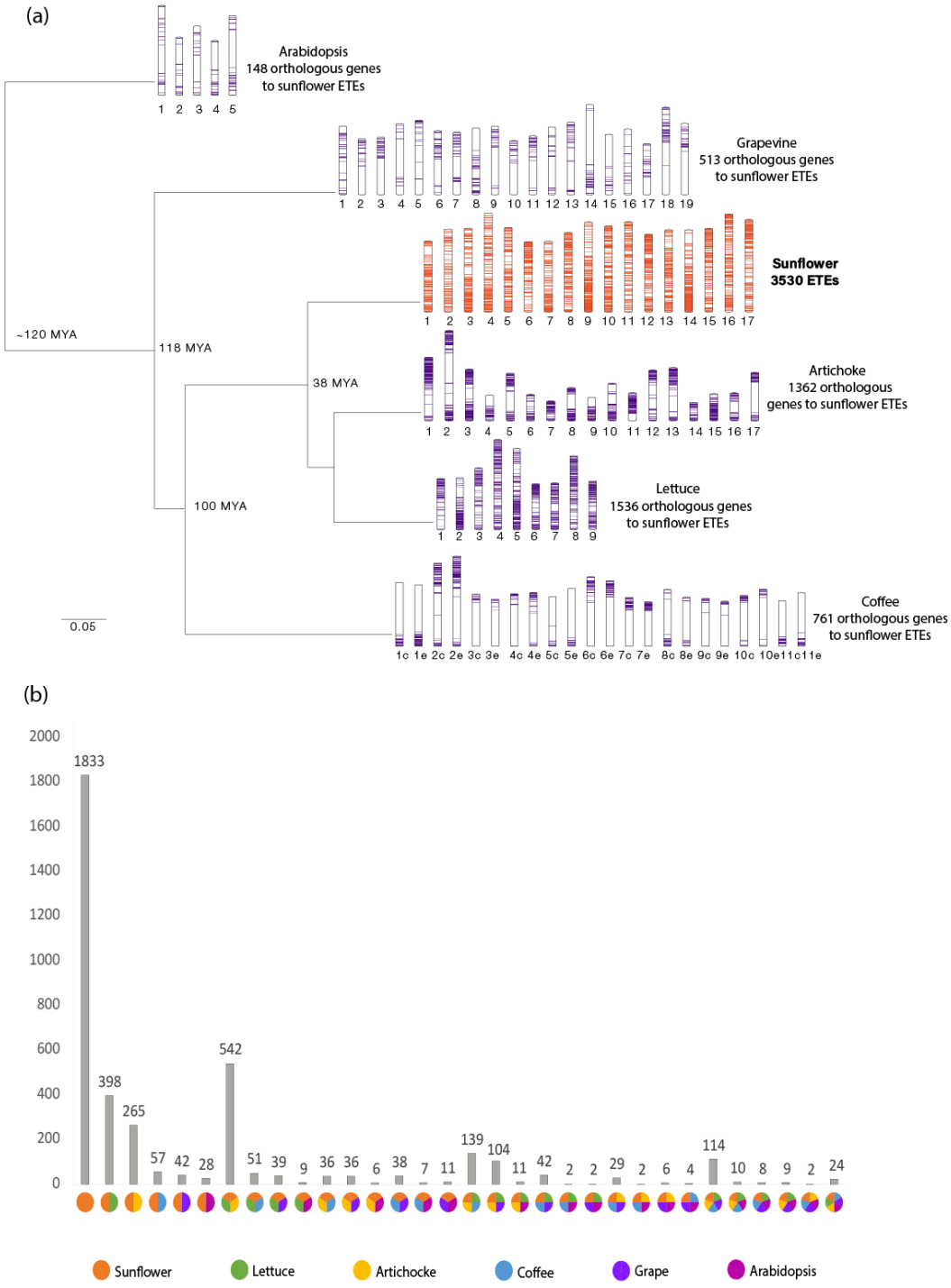

Figure 7 Evolutionary conservation of ETE orthologs. The phylogenetic tree describes the relationships among sunflower (*H. annuus*), artichoke (*Cynara cardunculus* var. *scolymus*), lettuce (*Lactuca sativa*), coffee (*Coffea arabica*), grapevine (*Vitis vinifera*), and *Arabidopsis*

thaliana. In each leaf, the ideograms report the location of ETE orthologs in each chromosome (a). The histogram reports the number of ETE orthologs that are shared in each combination of species (b).