Genetica

A computational genome-wide analysis of long terminal repeats retrotransposon expression in sunflower roots (Helianthus annuus L.)
--Manuscript Draft--

Manuscript Number:	GENE-D-19-00051R1
Full Title:	A computational genome-wide analysis of long terminal repeats retrotransposon expression in sunflower roots (Helianthus annuus L.)
Article Type:	Original Article
Keywords:	LTR-retrotransposons; sunflower; root; retrotransposon expression; Illumina cDNA libraries; LTR-retrotransposon lineages
Corresponding Author:	Andrea Cavallini Universita degli Studi di Pisa Pisa, I am not in the U.S. or Canada ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universita degli Studi di Pisa
Corresponding Author's Secondary Institution:	
First Author:	Flavia Mascagni
First Author Secondary Information:	
Order of Authors:	Flavia Mascagni
	Alberto Vangelisti
	Gabriele Usai
	Tommaso Giordani
	Andrea Cavallini
	Lucia Natali
Order of Authors Secondary Information:	

Abstract:	Long terminal repeats (LTR) retrotransposons have a major role in determining genome size, structure and function, thanks to their ability to transpose. We performed a meta-analysis of LTR-retrotransposon expression in roots of sunflower plantlets treated with different plant hormones, chemicals and NaCl. By using Illumina cDNA libraries, available from public repositories, we measured the number of reads matching the retrotranscriptase domains isolated from a whole genome library of retrotransposons. LTR-retrotransposons resulted in general barely expressed, except for 4 elements, all belonging to the AleII lineage, which showed high transcription levels in roots of both control and treated plants. The expression of retrotransposons in treated plants was slightly higher than in the control. Transcribed elements belonged to specific chromosomal loci and were not abundant in the genome. A few elements resulted differentially expressed depending on the treatment. Results suggest that, although most retrotransposons are not expressed, the transcription of such elements is related to their abundance, to their position in the chromosome and to their lineage.

Suggested Reviewers:	Andrea Zuccolo Scuola Superiore Sant'Anna a.zuccolo@santannapisa.it Dr. Zuccolo has a large experience in plant structural genomics and retrotransposons
	Matteo Buti Universita degli Studi di Firenze

	mbuti78@gmail.com He has large experience in plant retrotransposons
	Concepcion Muñoz Díez Universidad Nacional de Cordoba g62mudim@uco.es She has large experience in plant structural genomics
Response to Reviewers:	We wish to thank the Editor and the Reviewers for the useful suggestions to improve the manuscript. In the revised manuscript, changes are evidenced in red. In the paragraphs that follow, we describe our responses to reviewers' comments. Comments of reviewers are copied from the e-mail sent us by the editor. Reviewer #1: In this manuscript, Mascagni, Vangelisti and colleagues report LTR retrotransposon expression in sunflower roots. The overall idea of this manuscript deserves merit, but the manuscript still need some reviews. Authors should clarify why they did not use directly the genome annotation for RNA-seq mapping. RESPONSE: Our library of sunflower retrotransposons is regularly updated and validated and was used also in recent articles of ours (Mascagni et al 2015; 2017; Vangelisti et al. 2019). The library was built with a de novo identification approach for repeats, and sequence reliability was also validated by specific Sanger sequencing. On the other hand, a de novo assembled repeat library (as our library) is free by common mis-assembly issues which often affect whole genome reference sequences as, for example, repeat collapsing due to the incorrect gauging of the number of repeat copies in a genome (so providing a reference with less repeat copies than expected) and/or collapsing different haplotypes in genome assemblies (which can result in sequence errors due to differences between homologous chromosomes). For these reasons we think that using the genome reference sequence would not significantly improve a study of this kind, and we would prefer to maintain the use of our LTR-RE library. We have now validated our library also comparing it to the genome reference assembly (see also response #4 to Reviewer 2) and reported this point in the Materials and Methods and in the Results sections. I would also recommend that authors include the information if the selected nuclear genes are used for normalization in expression studies. RESPONSE: Nuclear genes were used for normalization. This information was included in the Materials and Methods. Authors should also include some comparative analysis with similar reports in sunflower (i.e. https://www.ncbi.nlm.nih.gov/pubmed/29304730). After these modifications, the manuscript can be re-analyzed. RESPONSE: Change made Reviewer #2: The article by conducted the analysis of expression of retrotransposons in roots of sunflower under different treatments. I feel the topic is interesting, but I have some major concerns, especially related to the methodologies used for analysis. In my opinion, it can be accepted for publication for Genetica after major revision. 1. The public Illumina datasets used in the current study were sequenced using paired-end library. Therefore, the authors utilizing RPKM or reads per million as expression quantification is inaccurate. Please reanalyze the data and use FPKM instead. RESPONSE: We agree with reviewer 2, calculating FPKM is more accurate in order to analyse paired-end dataset of reads. Actually, as stated in CLC Genomics Workbench manual, paired end data for RNA-seq analysis are by default scheme counted as FPKM instead of RPKM. Hence, using CLC by default, we already considered FPKM values for all the libraries. We apologize for the mistake, the manuscript has been changed accordingly and all "RPKM" have been replaced with "FPKM". Concerning the number of mapped reads per million, we modified tables and figures using FPKM. 2. What is the genetic variation of the retro-transcriptase domains for each lineage? RESPONSE: In Helianthus, retro-transcriptase domains show large sequence conservation. We reported these data as supplementary information to Mascagni et al. (2017, BMC Genomics 18:634 - Additional file S2). It was observed a clear-cut separation of RT-encoding sequences according to their lineage, even among RT sequences from different species of Helianthus.

How did the authors deal with reads that were mapped to multiple locations? They were randomly assigned or ignored for down-stream analysis? If randomly assigned, it may homogenize and reduce the differences

RESPONSE: This information was included in the text. CLC-BIO assigns randomly non-uniquely matching reads (i.e., reads that align at more than one position with an equally good score) to different sequences.In our RNA-seq experiments, the number of non-specific reads was quite low (for example, in control leaves was 12.59 ± 0.63 and similar percentages were observed for all treatments). Such multireads were included in the analysis because this strategy correctly estimates the expression of paralogous RT sequences (see Mortazavi et al. 2008). We agree that differences may be reduced, but, consequently such difference, when found, become more reliable.

3. The default parameter of RepeatExplorer is 90% minimum similarity and 55% minimum identity. The authors lowered the criterion to 60% and 40% for domain identification, will this increase false positive detection?

RESPONSE: Such relaxed parameters, compared to default ones (i.e., 90% minimum similarity and 55% minimum identity) were used because Repbase database contains only a few sunflower LTR-REs and using such parameters allowed us to isolate a larger number of RT sequences. However, we validated all isolated RT sequences by BLAST analysis against the Nucleotide Sequence Database at NCBI and maintained only validated sequences in the dataset. This point is now reported in the Materials and Methods section.

4. Also, a whole genome set of assembled sequences from 2013 was used to perform domain search. Will it be better to use the current available sunflower genome to infer RE domain as the genome was used to localize the expressed REs.

RESPONSE: Our assembly was regularly updated and validated in the last years (verifying that isolated sequences actually encoded RTs and the other LTR-RE proteins) and used also in recent articles of ours (Mascagni et al 2015; 2017; Vangelisti et al. 2019). We verified the occurrence of isolated RT sequences in the 2017 genome sequence and found that 2,118 over 2,134 sequences occurred in the genome assembly, with a 96.3% average sequence identity. We also validated the other domains (encoding integrase and RNAseH) and obtained similar results. Hence we preferred maintaining our library (see also the first response to Reviewer#1). We reported the new validation data in the Results section.

Some minor points:
Page 3, paragraph 2, line 3. What does "In this case" refer to? Did you mean "In the latter case"?
RESPONSE: Change made

Page 5, paragraph 2, lines 8-12, please label SRR # for each library. For example, SRR4996797 for gibberellic acid (GA3, 10 μM). In this way, it is easy to download the sequencing data if someone wants to repeat the work.
RESPONSE: Change made

Add the accession # for the chloroplast and mitochondrial sequences used to filter out organelle DNAs.
RESPONSE: Change made.

Figures 2-4, use FPKM as expression unit.
RESPONSE: Change made

A computational genome-wide analysis of long terminal repeats retrotransposon expression in sunflower roots (*Helianthus annuus* L.)

Flavia Mascagni . **Alberto Vangelisti** . **Gabriele Usai** . **Tommaso Giordani** . **Andrea Cavallini** . **Lucia Natali**

_____

Flavia Mascagni e Alberto Vangelisti have contributed equally to this work.

_____

Andrea Cavallini   andrea.cavallini@unipi.it   * Corresponding author

Lucia Natali   lucia.natali@unipi.it   * Corresponding author

Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, I-56124 Pisa, Italy

Abstract

Long terminal repeats (LTR) retrotransposons have a major role in determining genome size, structure and function, thanks to their ability to transpose. We performed a meta-analysis of LTR-retrotransposon expression in roots of sunflower plantlets treated with different plant hormones, chemicals and NaCl. By using Illumina cDNA libraries, available from public repositories, we measured the number of reads matching the retrotranscriptase domains isolated from a whole genome library of retrotransposons. LTR-retrotransposons resulted in general barely expressed, except for 4 elements, all belonging to the *AleII* lineage, which showed high transcription levels in roots of both control and treated plants. The expression of retrotransposons in treated plants was slightly higher than in the control. Transcribed elements belonged to specific chromosomal loci and were not abundant in the genome. A few elements resulted differentially expressed depending on the treatment. Results suggest that, although most retrotransposons are not expressed, the transcription of such elements is related to their abundance, to their position in the chromosome and to their lineage.

Introduction

Transposons are mobile DNA sequences, widespread throughout eukaryotic species, which can change their position in the genome through transposition, a process operated by enzymes encoded by the transposon itself. Among transposons, retrotransposons (REs) change their location through a replicative mechanism that involves the transcription of an RNA intermediate followed by retro-transcription and insertion in the genome (Wicker et al. 2007). This mechanism has allowed REs to become the largest portion of genomes in most eukaryotic species (SanMiguel et al. 1998; Vicient et al. 1999).

In plants, the most abundant REs are characterized by two long terminal repeats (LTRs) at their ends. Long terminal repeats contain promoter elements, polyadenylation signals, and enhancers, which regulate RE transcription (Bennetzen 2000). The portion of RE internal to the two LTRs contains two protein-encoding domains, the Gag and the Pol. Gag proteins are similar to virus-like particles, Pol contains enzyme domains necessary for transposition, as the reverse

transcriptase (RT), which produces a double-stranded DNA, the RNAseH, the protease, and the integrase (Bennetzen 2000). A primer binding site and a poly-purine tract are other structural features of LTR-REs involved in the transposition process (Bennetzen 2000).

Transposition of LTR-REs (retrotransposition) starts with the transcription of the element. LTR-RE RNAs can be polyadenilated (when destined to be translate to produce RE enzymes) or not (when subjected to be reverse-transcribed) (Chang et al. 2013; Meignin et al. 2003). In the latter case, the transcripts are retro-transcribed by the RT and double-stranded DNAs are produced, after the RNA templates are degraded by the RNAseH; then, the integrase inserts the double-stranded DNAs into the host genome. LTR-REs transposition is limited by the host genome due to its potential mutagenic action. The mechanisms to control transposon activity include chromatin silencing by methylation of histones and cytosine residues (Dieguez et al. 1998) and RNA silencing, which produces both RNA degradation and chromatin specific methylation (Slotkin and Martienssen 2007; Lisch 2013; Ito 2013).

Retrotransposition determines large variations in the genome size and structure of plants, even at intraspecific level (Springer et al. 2009; Vitte et al. 2014). Besides structural changes in the genome, retrotransposition can often cause changes in gene expression regulation, with consequent phenotypic changes (Slotkin and Martienssen 2007; Butelli et al. 2012; Falchi et al. 2013; Lisch 2013).

Plant LTR-REs are mostly subdivided into two superfamilies, *Gypsy* and *Copia* (Wicker et al. 2007), according to the order of encoding domains within the Pol and on sequence similarity. Superfamilies have been further subdivided into lineages, based on the structure of the element, on DNA sequence similarity, and on the occurrence of specific sequence motifs. Such lineages can be recognised in many plant species.

Sequence similarity of LTR-REs among species is often minimal and limited to the coding regions (Wicker et al. 2007). In Angiosperms, many *Gypsy* and *Copia* lineages have been identified (Wicker and Keller 2007; Llorens et al. 2011; Barghini et al. 2015; Usai et al. 2017; Buti et al. 2018; Neumann et al. 2019). The main *Gypsy* lineages are *Chromovirus*, a lineage of REs carrying a chromodomain at the 5' end of the coding portion, which is especially abundant in centromeres (Gorinsek et al. 2004; Llorens et al. 2011); *Athila*, reported also in Gymnosperms (Neumann et al. 2019); and *Ogre,* represented by large elements with an open reading frame located upstream of the gag gene (Neumann et al. 2003). *Chromovirus* lineage, on its turn, is often subdivided into four sublineages, *Galadriel*, *Tekay*, *CRM*, and *Reina* (Weber et al. 2013). Also *Copia* LTR-REs can belong

to many different lineages, the most diffused being *Ale*, on its turn often distinguished into *AleI/Retrofit/Hopscotch* and *AleII*, *Ivana*, *Angela*, *Bianca*, *TAR* and *Tork* (often considered as an unique lineage), and *SIRE* (Wicker and Keller 2007; Neumann et al. 2019).

During the last years, LTR-REs of sunflower (*Helianthus annuus*), a species belonging to the largest Angiosperm family (Asteraceae) have been the subject of many studies. General surveys of LTR-REs and other repetitive DNAs in the genome of *H. annuus* have been performed using Illumina and 454 sequencing techniques and assembling the relative sequence reads (Staton et al. 2012; Natali et al. 2013; Giordani et al. 2014). These studies revealed the occurrence of many different transposable sequences, which amount approximately at 80% of the sunflower genome, recently fully sequenced (Badouin et al. 2017).

The importance of LTR-REs in the evolution of *Helianthus* genus was shown by comparing the abundance of different lineages in *Helianthus* species. Retrotransposition on the one hand and recombinational loss on the other have produced changes in the genome size among species and even within *H. annuus* (Mascagni et al. 2017a; 2017b). The extent of interspecific and intraspecific LTR-RE abundance was related to the different accumulation of specific superfamilies and lineages: for example *Gypsy* REs are much more abundant than *Copia* REs (Mascagni et al. 2015; 2018). At lineage level, *Chromoviruses* are by far the most represented *Gypsy* REs in the sunflower genome; among *Copia* lineages the most abundant belong to *SIRE* lineages, while other lineages are rare (Mascagni et al. 2015).

In this study we measured the expression (i.e., the first phase of retrotransposition) of the LTR-REs of sunflower in relation to their lineage in order to evaluate if different lineages are differently transcribed. With this aim, we performed a meta-analysis using a comprehensive library of RT encoding sequences previously identified in sunflower LTR-REs (Vangelisti et al. 2019). The expression of these RT sequences was measured by mapping with a number of Illumina cDNA read libraries, available in a public repository, produced by Badouin et al. (2017) from RNAs isolated by roots of plantlets of sunflower subjected to different treatments (hormones, chemicals, salt) to mimic abiotic and biotic stresses.

Materials and methods

Sequence collection

Reverse transcriptase, integrase, and RNAseH encoding sequences from LTR-retrotransposons of *H. annuus* were identified in a sequence set, representing a whole-genome set of assembled sequences (Natali et al. 2013) and collected. Briefly, genomic Illumina DNA reads from sunflower inbred line HA412-HO were obtained and assembled following several procedures (Natali et al. 2013). The occurrence of the domain sequences in the available genome sequence of *H. annuus* (Badouin et al. 2017) was verified by BLAST analysis.

The collection of assembled sequences were then submitted to RepeatExplorer (Novak et al. 2013) protein domain search tool. Repbase database enclosed in the RepeatExplorer tool was exploited to perform domain searches. The following parameters were used: 60% minimum similarity, 40% minimum identity; proportion of the hit length from the length of the database sequence = 0.3; maximum allowed frameshifts = 3. Such relaxed parameters, compared to default ones (i.e., 90% minimum similarity and 55% minimum identity) were used because Repbase database contains only a few sunflower LTR-REs and using such parameters allowed us to isolate a larger number of RT sequences. However, we validated all isolated RT sequences by BLAST analysis against the nr Sequence Database at NCBI and against an in-house library of LTR-REs (Mascagni et al. 2015).

In addition three genes of *H. annuus* were downloaded from NCBI repository (https://www.ncbi.nlm.nih.gov/); these three sequences encode a Z-carotene desaturase (Z-Des, FR671183.1, Giordani et al. 2011), a Lipid Transfer Protein (LTP, FR671365.1, Giordani et al. 2011) and an Actin (HanXRQChr11g0323331, Badouin et al. 2017). The expression of RT domain sequences and of the three selected genes was analysed in *H.annuus* plantlets root after different treatments. Overall, 16 Illumina cDNA libraries from roots of HanXRQ line, publicly available (Badouin et al. 2017), were downloaded from SRA repository (https://www.ncbi.nlm.nih.gov/sra/, accession nr. SRP092742). Ten libraries were obtained from roots of plantlets treated for 14 days with auxin (IAA, 0.1 µM, SRA code SRR4996845), ethylene (ETH, 0.25 µM, SRR4996811), gibberellic acid (GA3, 10 µM, SRR4996797), salicylic acid (SA, 0.05 µM, SRR4996847), kinetin (KIN, 0.5 µM, SRR4996844), abscisic acid (ABA, 10 µM, SRR4996816), strigolactones (STRI, 1 µM, SRR4996826), brassinosteroid (BRA, 1 µM, SRR4996803), polyethylene glycol (PEG, 100 g/L, SRR4996815) and NaCl (100 mM, SRR4996819). Conversely, the 6 remaining libraries were obtained from untreated sunflower roots (ID codes SRR4996805, SRR4996818, SRR4996820, SRR4996828, SRR4996846, SRR4996852) and used as control (Badouin et al. 2017).

Mapping procedure and statistical analysis

Quality check of cDNA reads was performed by FastQC (v. 0.11.3), and overall quality was improved by Trimmomatic (Bolger et al. 2014) removing Illumina adapters and trimming the sequences (HEADCROP:10; CROP:56; MINLEN:50) (Table 1).

Additional ribosomal filter for all libraries was performed by mapping against *H. annuus* rDNA sequences downloaded from NCBI repository. Default parameters were used for ribosomal filtering except for length fraction = 0.5 and similarity fraction = 0.8.

Trimmed cDNA reads were mapped onto the RT libraries and the three chosen gene sequences using the CLC Genomics Workbench (v. 9.5.3; CLC-BIO, Aarhus, Denmark) using stringent parameters (length fraction = 0.9 and similarity fraction = 0.9) but retaining mismatch penalties = 1 and gap penalties = 1. The software assigns randomly non-uniquely matching reads (i.e., reads that align at more than one position with an equally good score) to different sequences. Raw counts per sequence after alignment were normalized as fragments per kilobase per million reads mapped (FPKM, Trapnell et al. 2010), including the selected genes in the analysis, in order to establish gene and RT expression level. A single transcript was considered expressed if showed at least 1 mapped fragment per million in at least one library. Fold change was reported as positive when values where higher in treated than in control plants, and as negative for the opposite case, leading to "+" and "-" annotation for expressed elements. Pairwise analysis for differentially expressed genes between control and treatments was performed using Baggerley's statistical test, based on count proportion and working on binomial distribution (Baggerley et al. 2003). A sequence was considered differentially expressed when absolute log fold change > 1 and False Discovery Rate (FDR) corrected *p*-Value < 0.05.

In order to assess genomic abundance of REs, DNA Illumina reads of *Helianthus annuus* inbred line HanXRQ were downloaded from SRA public repository (SRR5004633). Paired end reads were trimmed removing adapters and improving overall quality with following parameters: SLIDINGWINDOW:4:20, HEADCROP:15, MINLEN:86.

Additional filter was made for organelle DNAs, aligning reads on chloroplast and mitochondrial sequences of *H. annuus* (NCBI accession numbers NC_007977 and KF815390, respectively). Unmapped reads were retained.

Finally, trimmed reads of 86 bp were mapped onto reference retrotransposon domains library using CLC with same parameters described above.

Localization of expressed REs along the sunflower genome

Each of the 17 linkage groups (LGs) of the currently available sunflower genome sequence (HanXRQ inbred line, Badouin et al. 2017), were subdivided into 3-Mbp-long genome regions. Then, the DNA reads mapped onto the RE domain library (see above) were collected using the CLC Genomics Workbench (v. 9.5.3). Such reads were used for masking the 3-Mbp-long fragments of the HanXRQ genome using RepeatMasker (http://www.repeatmasker.org) under default parameters, in order to perform genome localization of RE sequences. A putative sunflower centromeric sequence, HAG002P01 (Cavallini et al. 2010) was also used for masking the fragmented genome, separately, using the same procedure. The number of masked bases was then counted for each of the 3 Mbp fragment using another in-house perl script.

In other analyses, the cDNA reads (from roots of control or treated plantlets) matching the retrotransposon domains library (see above) were collected and mapped onto the 3-Mbp-long genome regions using the CLC Genomics Workbench (v. 9.5.3) in order to localize the expressed REs in the genome.

Results

Overall expression of LTR-REs in sunflower roots

The expression of LTR-REs in the roots of sunflower was measured by mapping Illumina cDNA reads onto a collection of sunflower RT encoding sequences, identified in a whole genome set of assembled sequences (Natali et al. 2013; Vangelisti et al. 2019) and validated through BLASTX analysis against the NCBI nr sequence database and an in-house sunflower LTR-RE library (Mascagni et al. 2015). After validation, the collection of RT domains was composed of 2,134 sequences. We performed a BLASTN search of these sequences against the available sunflower genome assembly (Badouin et al. 2017) and found that 2,118 over 2,134 sequences occurred in the genome assembly, with a 96.3% average sequence identity. We also validated and analysed the expression of other domains (encoding integrase and RNAseH) and obtained similar results (data not reported).

Retrotranscriptase-encoding sequences were subdivided as belonging to 7 lineages of *Copia* and 3 lineages of *Gypsy* elements. Illumina cDNA libraries were obtained by Badouin et al. (2017) from roots of control plantlets (6 libraries) and from roots of plantlets treated with abscisic acid, ethylene, brassinosteroids, gibberellic acid, indole-acetic acid, kinetin, NaCl, polyethylene glycol, salicylic acid and strigolactones, respectively (10 libraries).

The expression level of each RT was evaluated by measuring the FPKM and compared with those of three single copy gene sequences, encoding two proteins involved in lipid metabolism (a LTP and a Z-Des) and a housekeeping protein (an actin). CLC-BIO assigns randomly nonspecifically matched reads, i.e., reads aligning at more than one position with an equally good score. However, in our experiments, the number of non-specific reads was quite low. For example, the average percentage of non-specific matches in control leaves was 12.59 ± 0.63. Similar percentages were observed for all treatments (data not shown). Such multireads were included in the analysis because this strategy correctly estimates the expression of paralogous RT sequences (see Mortazavi et al. 2008).

Table 1 reports the percentage of cDNA reads matching 1,133 RTs of the *Copia* superfamily and 1,001 RTs of the *Gypsy* superfamily, and the percentages of reads mapping onto the three selected genes. Overall LTR-RE expression was very low compared to the three single copy gene sequences. *Copia* REs were much more expressed than *Gypsy* ones and increased their overall expression in stressed plants.

To estimate the expression level of each RT sequence we arbitrarily established FPKM threshold: we considered as untranscribed or barely expressed those sequences mapped by less than 1 fragment per million, which corresponded, in our experiments, to FPKM = 1,500. FPKM values higher than 1,500 were subdivided into 3 classes: from 1,500 to 15,000, RT sequences were considered as lowly expressed; from 15,000 to 40,000, as expressed; and > 40,000 (i.e. higher than that of the most expressed reference gene, encoding the actin) as highly expressed.

Considering single RT sequences, the vast majority of RTs of both *Copia* and *Gypsy* superfamilies resulted barely expressed or untranscribed (Table 2). In the roots of control plants, only 28 RTs (20 *Copia* and 8 *Gypsy*) showed an FPKM > 1,500 and only 3 RTs (of the *Copia* superfamily) showed FPKM > 40,000. In the roots of treated plants (considering all treatments), the expression levels of single elements were similar to those of the roots of control plants (Table 2). The average FPKMs of the actin sequence were 34,124 and 23,165, those of Z-des sequence

3,983 and 5,505, and those of LTP sequence 552 and 4,140, in the controls and in the treated plants, respectively.

The localization of expressed RTs in the 17 chromosomes of *Helianthus annuus* was determined by mapping the LTR-RE-matching Illumina cDNA reads on the available genome sequence of *H. annuus* HanXRQ line (Badouin et al. 2017) (Figure 1). In Figure 1 are also reported the chromosomal localization of LTR-REs and of centromeres, determined by masking the *H. annuus* linkage groups with the library of retroelements of sunflower from which RT sequences were isolated (Mascagni et al. 2015) and a putative centromeric tandem repeat (Cavallini et al. 2010). It can be observed that, in roots of both control and treated plants, cDNA mapping identified specific chromosome regions, often localized at the extremities of chromosomes and identical in control and treated plants. Such regions never corresponded to centromeres and were generally regions with the lowest abundance of LTR-REs.

The expression of LTR-RE RTs was analysed also at lineage level. Figure 2 shows the number of elements subdivided among untranscribed (FPKM < 1,500), lowly expressed (1,500 < FPKM < 15,000), expressed (15,000 < FPKM < 40,000) and highly expressed (FPKM > 40,000) in roots of control and treated plants. In either control and treated plants, the vast majority of LTR-REs were untranscribed, a low expression level was measured for a few elements of 6 lineages (*AleII*, *Bianca*, *Ivana*, *SIRE*, *Chromovirus* and *Ogre*) in control plants and 7 lineages (the same as control plants plus *TAR/Tork*) in roots of treated plants. A higher expression level was observed only for 4 LTR-REs of the lineage *AleII* of the *Copia* superfamily.

Figure 3 reports the relationship between abundance in the genome and expression of a RT encoding sequence, subdivided per lineage, in roots of control plants. It can be observed that the seven most expressed elements, belonging to *AleII* (5 sequences) and *Bianca* (1 sequence) are barely abundant in the genome.

Analysis of expressed LTR-REs

Assuming at least one mapped fragment per million reads in at least one library (of roots of control or treated plants) as the threshold to define a RT-encoding sequence as expressed, only 49 RT-sequences were identified as expressed. All 49 sequences occurred in the present genome assembly of sunflower (Badouin et al. 2017), with a 96.6% average sequence identity. Their expression levels in roots of control and differently treated plants are reported in Fig. 4,

subdivided into five classes. The expressed RT sequences belong to 8 lineages, 6 of the *Copia* superfamily (*AleII*, 14 sequences; *SIRE*, 10; *AleI/Retrofit*, 6; *Ivana/Oryco*, 4; *TAR/Tork*, 3; *Bianca*, 1) and 2 of the *Gypsy* superfamily (*Chromovirus* and *Ogre/TAT*, each with 5 sequences). Four RT sequences (all belonging to the *AleII* lineage) are highly expressed in control roots as well as in all treatments (Figure 4).

The FPKM values of RT sequences in each culture treatment was compared with those of control roots and the occurrence of differential expression was established when FPKM fold change was > 2 with FDR < 0.05. The occurrence of differential expression for each RT sequence and each treatment is indicated by the symbol + in case of overexpression and - in case of repression in Figure 4. In some cases RT sequences were overexpressed, in other they were repressed. Concerning the different treatments, indole acetic acid induce a general repression of RT sequences. On the other hand, salicylic acid induces the over-expression of many REs. The effect of the other treatments is different depending on the LTR-RE. Gibberellin, poly-ethylene glycol and NaCl induced differential expression only for 3, 5 and 6 RT sequences, respectively.

Discussion

The occurrence of RE transcripts in different tissues and organs of plants have been reported in many species, both constitutively or following exposition to abiotic and biotic stresses (Grandbastien 2015). Certain LTR-REs are expressed without apparent inducing conditions, i.e. constitutively, for example in *Citrus sinensis*, in sunflower, in rice and in the poplar (Rico-Cabanas and Martínez-Izquierdo 2007; Vukich et al. 2009; Gao et al. 2015; Giordani et al. 2016). On the contrary, other elements are constitutively transcribed only at very low level, for example in grasses (Vicient et al. 2001; Ishiguro et al. 2014). In any case, the expression of a LTR-RE does not imply that such element will complete retrotransposition by inserting in a new genome site. A complete retrotranspositional process has been reported in a few cases: for example, the expression of *Tnt*1 and *Tto*1 in *Nicotiana* and *Tos*17 in rice were described in tissue cultures of those species, followed by their subsequent insertion in the genome (Grandbastien 1998). Complete retrotransposition was also observed for one *Copia* element of sunflower, in normally cultivated plants (Vukich et al. 2009).

In this study, we reported an analysis of the expression of RT-encoding domains. We are conscious that the expression of a protein-encoding domain is only an indication that the whole

LTR-RE to which that domain belongs is actually expressed. On the other hand, analysis of complete elements (which are generally transcribed at low rates) would have needed cDNA sequence coverages largely higher than those available and currently used in RNA-seq studies.

The results reported in this study suggest that LTR-REs are not generally expressed in sunflower roots. As a matter of fact, only three elements were more expressed than an actin-encoding gene in control roots and only two in roots of treated plants. In some cases, LTR-REs have been shown to be activated at transcriptional and possibly transpositional level by different treatments mimicking abiotic and biotic stresses (Kashkush et al. 2003; Liu et al. 2004; He et al. 2012; Ito et al. 2013; Voronova et al. 2014). For example, transcription of LTR-REs of both *Copia* and *Gypsy* superfamilies were induced by gibberellin and salicylic acid in cotton and pine (Hawkins et al. 2008; Fan et al. 2014); ABA, cytokinins and auxins induced strawberry FaRE1 (He et al. 2010); tissue culture, jasmonate and fungal elicitors activated *Tto1* in tobacco (Takeda et al. 1999; Liu et al. 2004).

In our study, the expression level was basically the same in roots of plantlets treated with different hormones, with salts, or with chemicals as in the controls. In these treatments, the sum of the Illumina reads mapping onto the RT library increased but remained substantially very low. Moreover, it is worth noting that the same few RT sequences which were expressed in treated plantlets were expressed also in controls.

In general, it can be deduced that LTR-REs are substantially inactive in the roots of sunflower, although a few elements showed detectable expression rates. These elements, which are expressed in roots of both treated and control plantlets, are not silenced and may have a mutagenic potential, if their transcription were followed by retrotranscription and insertion in the genome.

Genome localization of expressed RT domains suggest that the expression of an element is related also to its chromosomal localization. In fact, we observed specific chromosome regions with peaks of RT mapping cDNA reads. The abundance of LTR-REs in these regions was minor than in the other and, probably for this reason, these regions are less subject to repression of LTR-RE transcription. For example, these regions are relatively distant from putative centromeres, which are especially abundant in LTR-REs. Chromosomal regions related to LTR-RE expression are preferentially located at chromosome ends, that in sunflower and other species are preferentially occupied by *Copia* LTR-REs (Santini et al. 2002; Cavallini et al. 2010).

Our results showed that the expression of *Copia* REs is by far larger than that of *Gypsy* elements. Moreover, in sunflower, all elements expressed at high level belonged to the *Copia* superfamily. Many of the LTR-REs expressed in other species are actually of the *Copia* superfamily (Ma et al. 2008). In sunflower, *Copia* elements are much less abundant than *Gypsy* and this might explain such difference in expression. In fact, LTR-RE abundance and transcription are apparently inversely correlated. Abundant LTR-REs were only slightly transcribed or completely untranscribed, whereas rare REs were actively expressed. It is commonly accepted that the more abundant is an element the more easily it is recognized and subjected to RNA silencing (Meyers et al. 2001; Yamazaki et al. 2001; Lisch 2013). Even in the *Helianthus* genus, Qiu and Ungerer (2018) found similar results. In three wild species, *H. agrestis*, *H. carnosus* and *H. porteri*, they found that *Gypsy* elements are much more abundant than *Copia* ones, however the most expressed LTR-REs belonged to the *Copia* superfamily and especially to barely represented families. The lack of correlation between LTR-RE abundance and transcription rate also indicates that the presence of RT domain sequences in the Illumina libraries was not due to DNA contamination of RNA samples.

Interestingly, all RT domains expressed at relatively high level belonged to *AleII* lineage, indicating that, besides chromosomal localization and genome abundance, also the "genotype" of the LTR-RE might play a role in its activation. LTR-RE lineage-depending expression was also reported in cotton (Hawkins et al. 2006). In the case of tobacco, both *Tnt1* and *Tto1* (which are induced by tissue culture) belong to the *TAR/Tork* lineage (Neumann et al. 2019). It is presumable that different lineages are specifically activated in different species.

Although in roots of plantlets subjected to different treatments, expressed LTR-REs are generally the same as in control plantlets; for many of these LTR-REs the different treatments induced over-expression or repression, probably because of the occurrence, in the LTRs, of *cis*-regulatory motifs recognized in specific stresses (as observed in the LTR of the HaCRE1 element of sunflower (Buti et al. 2009)). For example, in our analysis, IAA induced a general repression of LTR-REs. This hormone is important in the control of root growth and differentiation (Aloni et al. 2006). Reduced expression of LTR-REs during IAA treatment might be related to the repression of these elements in order to making sure that their activation does not interfere with a regular root development.

In conclusion, our study shows that LTR-REs are sporadically transcribed in sunflower roots, even in plants subjected to treatments mimicking abiotic and biotic stresses. However, a few elements, all belonging to the *AleII* lineage of the *Copia* superfamily, are expressed at high levels,

indicating that the RE lineage affect LTR-RE expression and that the process of retrotransposition might naturally occur in roots of sunflower. These few elements will be the candidates for further studies in order to ascertain the occurrence of new insertions of LTR-REs in the genome.

Compliance with ethical standards

Conflict of interest   All authors declare that they do not have conflict of interest.

Ethical approval   This article does not contain any study with human subjects or animals performed by any of the authors.

References

Aloni R, Aloni E, Langhans M, Ullrich CI (2006) Role of cytokinin and auxin in shaping root architecture: regulating vascular differentiation, lateral root initiation, root apical dominance and root gravitropism. Ann Bot 97:883–893

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. (2017) The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546:148–152

Baggerley K, Deng L, Morris J, Aldaz C (2003) Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics 19:1477-1483

Barghini E, Mascagni F, Natali L, Giordani T, Cavallini A (2015) Analysis of the repetitive component and retrotransposon population in the genome of a marine angiosperm, *Posidonia oceanica* (L.) Delile. Mar Genomics 24:397–404

Bennetzen JL (2000) Transposable elements contributions to plant gene and genome evolution. Plant Mol Biol 42:251–269

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

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P (2012) Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24:1242-1255

Buti M, Giordani T, Vukich M, Gentzbittel L, Pistelli L, Cattonaro F, et al. (2009) HACRE1, a recently inserted *Copia*-like retrotransposon of sunflower (*Helianthus annuus* L.). Genome 52:904-911.

Buti M, Moretto M, Barghini E, Mascagni F, Natali L, Brilli M, et al. (2018) The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). GigaScience 7:1-14

Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V, et al. (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. Theor Appl Genet 120:491–508

Chang W, Jääskeläinen M, Li S, Schulman AH (2013) BARE retrotransposons are translated and replicated via distinct RNA pools. PLoS ONE 8:e72270

Dieguez MJ, Vaucheret H, Paszkowski J, Mittelsten Scheid O (1998) Cytosine methylation at CG and CNG sites is not a prerequisite for the initiation of transcriptional gene silencing in plants, but it is required for its maintenance. Mol Gen Genet 259:207–215

Falchi R, Vendramin E, Zanon L, Scalabrin S, Cipriani G, Verde I, et al. (2013) Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. Plant J 76:175-187

Fan FH, Cui BW, Zhang T, Ding GJ, Wen XP (2014) LTR-retrotransposon activation, IRAP marker development and its potential in genetic diversity assessment of masson pine (*Pinus massoniana*). Tree Genet Genomes 10:213–222

Gao X, Zhou J, Li J, Zou X, Zhao J, Li Q, et al. (2015) Efficient generation of marker-free transgenic rice plants using an improved transposon-mediated transgene reintegration strategy. Plant Physiol 167:11-24

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

Giordani T, Cossu RM, Mascagni F, Marroni F, Morgante M, Cavallini A, Natali L (2016) Genome-wide analysis of LTR-retrotransposons expression in leaves of *Populus x canadensis* water-deprived plants. Tree Genet Genomes 12:75

Giordani T, Buti M, Natali L, Pugliesi C, Cattonaro F, Morgante M, Cavallini A (2011) An analysis of sequence variability in eight genes putatively involved in drought response in sunflower (*Helianthus annuus* L.). Theor Appl Genet 122:1039–1049

Gorinšek B, Gubenšek F, Kordiš D (2004) Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21:781-798

Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3:181–189

Grandbastien MA (2015) LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim Biophys Acta 1849:403–416

Hawkins JS, Kim HR, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252-1261

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF (2008) Phylogenetic determination of the pace of transposable element proliferation in plants: copia and LINE-like elements in *Gossypium*. Genome 51:11–18

He P, Ma Y, Zhao G, Dai H, Li H, Chang L, Zhang ZH (2010) FaRE1: a transcriptionally active *Ty1-copia* retrotransposon in strawberry. J Plant Res 123:707–714

He P, Ma Y, Dai HY, Li LG, Liu YX, Li H, Zhao GL, Zhang ZH (2012) Characterization of the hormone and stress-induced expression of FaRE1 retrotransposon promoter in strawberry. J Plant Biol 55:1–7

Ishiguro S, Ogasawara K, Fujino K, Sato Y, Kishima Y (2014) Low temperature-responsive changes in the anther transcriptome's repeat sequences are indicative of stress sensitivity and pollen sterility in rice strains. Plant Physiol 164:671–682

Ito H (2013) Small RNAs and regulation of transposons in plants. Genes Genet Syst 88:3–7

Ito H, Yoshida T, Tsukahara S, Kawabe A (2013) Evolution of the ONSEN retrotransposon family activated upon heat stress in Brassicaceae. Gene 518:256–261

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102–106

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

Liu Z, Han PF, Tan M, Shan XH, Dong YZ, Wang XZ, Fedak G, Hao S, Liu B (2004) Activation of a rice endogenous retrotransposon Tos17 in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. Theor Appl Genet 109:200–209

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. (2011) The *Gypsy* Database (GyDB) of mobile genetic elements: release 2.0. Nucl Acid Res 39:70-74

Ma Y, Sun HY, Zhao GL, Dai HY, Gao XY, Li H, Zhang ZH (2008) Isolation and characterization of genomic retrotransposon sequences from octoploid strawberry (*Fragaria × ananassa* Duch.). Plant Cell Rep 27:499–507

Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, 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, Cavallini A, Giordani T, Natali L (2017a) Different histories of two highly variable LTR retrotransposons in sunflower species. Gene 634:5-14

Mascagni F, Giordani T, Ceccarelli M, Cavallini A, Natali L (2017b) Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). BMC Genomics 18:634

Mascagni F, Vangelisti A, Giordani T, Cavallini A, Natali L (2018) Specific LTR-retrotransposons show copy number variations between wild and cultivated sunflowers. Genes 9:433

Meignin C, Bailly JL, Arnaud F, Dastugue B, Vaury C (2003) The 5' untranslated region and gag product of Idefix, a long terminal repeat retrotransposon from *Drosophila melanogaster*, act together to initiate a switch between translated and untranslated states of the genomic mRNA. Mol Cell Biol 23:8246–8254

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Meth 5:621-628

Natali L, Cossu RM, Barghini E, Giordani T, Buti M, Mascagni F, et al. (2013) The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. BMC Genomics 14:686

Neumann P, Pozarkova D, Macas J (2003) Highly abundant pea LTR retrotransposon *Ogre* is constitutively transcribed and partially spliced. Plant Mol Biol 53:399–410

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

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792-793

Qiu F, Ungerer MC (2018) Genomic abundance and transcriptional activity of diverse *gypsy* and *copia* long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biology 18:6

Rico-Cabanas L, Martínez-Izquierdo JA (2007) CIRE1, a novel transcriptionally active *Ty1-copia* retrotransposon from *Citrus sinensis*. Mol Genet Genom 277:365

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Santini S, Cavallini A, Natali L, Minelli S, Maggini F, Cionini PG (2002) *Ty1/copia*-and *Ty3/gypsy*-like DNA sequences in *Helianthus* species. Chromosoma 111:192–200

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Rev Genet 8:272

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734

Staton SE, Bakken BH, Blackman BK, Chapman MA, Kane NC, Tang S, et al. (2012) The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J 72:142–153

Takeda S, Sugimoto K, Otsuki H, Hirochika H (1999) A 13-bp cis-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. Plant J 18:383–393

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol 28:511-515

Usai G, Mascagni F, Natali L, Giordani T, Cavallini A (2017) Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. Tree Genet Genomes 13:96

Vangelisti A, Mascagni F, Giordani T, Sbrana C, Turrini A, Cavallini A, Giovannetti M, Natali L (2019) Arbuscular mycorrhizal fungi induce the expression of specific retrotransposons in roots of sunflower (*Helianthus annuus* L.). PLoS ONE 14: e0212371

Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125:1283–1292

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769–1784

Vitte C, Fustier MA, Alix K, Tenaillon MI (2014) The bright side of transposons in crop evolution. Briefings Funct Genom 13:276-295

Voronova A, Belevich V, Rungis D, Jansons A (2014) Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (*Pinus sylvestris* L.) genome. Tree Genet Genomes 10:937–951

Vukich M, Giordani T, Natali L, Cavallini A (2009) *Copia* and *Gypsy* retrotransposons activity in sunflower (*Helianthus annuus* L.). BMC Plant Biol 9:150

Weber B, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T (2013) Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mob DNA 4:8

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. (2007) A unified classification system for eukaryotic transposable elements. Nature Rev Genet 8:973-982

Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, Yamamoto S, et al. (2001) The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. Mol Genet Genomics 265:336–344

Table 1 Percentage of cDNA reads matching 1,133 RTs of the *Copia* superfamily and 1,001 RTs of the *Gypsy* superfamily, and three selected genes encoding an actin, a Z-carotene-desaturase (Z-des) and a lipid-transfer-protein (LTP)

| cDNA libraries | Total number of reads | % of mapped reads | | | | | |
		RTs (overall)	*Copia* RTs	*Gypsy* RTs	Actin	Z-des	LTP
Controls	218,878,770	0.022	0.019	0.003	0.006	0.0003	0.00002
Indol-acetic acid	63,006,388	0.009	0.008	0.001	0.002	0.0001	6.35×10^{-06}
Kinetin	54,445,226	0.034	0.031	0.003	0.005	0.0004	1.47×10^{-05}
Gibberellin	52,020,586	0.040	0.036	0.004	0.005	0.0004	0
Ethylene	41,012,542	0.027	0.024	0.003	0.004	0.0003	4.39×10^{-05}
Salt	21,813,710	0.027	0.023	0.004	0.005	0.0003	6.88×10^{-05}
Polyethylene glycol	37,849,316	0.023	0.020	0.003	0.005	0.0003	1.06×10^{-05}
Abscisic acid	43,105,042	0.038	0.036	0.002	0.004	0.0009	0.001926
Salicylic acid	45,066,892	0.027	0.024	0.003	0.004	0.0003	6.66×10^{-06}
Strigolactones	49,097,546	0.033	0.031	0.003	0.004	0.0004	8.15×10^{-06}
Brassinosteroids	54,559,166	0.027	0.024	0.003	0.005	0.0005	5.50×10^{-06}
All treatments	461,976,414	0.028	0.025	0.003	0.004	0.0004	0.000192

Table 2 Mean FPKM of RT-sequences by using 6 and 10 cDNA Illumina libraries from roots of control and treated plants, respectively

| Mean FPKM | Control plants | | Treated plants | |
	Number of *Copia* RTs	Number of *Gypsy* RTs	Number of *Copia* RTs	Number of *Gypsy* RTs
> 40,000	3	0	2	0
15,000 - 40,000	1	0	2	0
1,500 - 15,000	16	8	17	7
< 1,500	1,113	993	1,112	994

LEGENDS FOR FIGURES

Fig. 1 Distribution of Illumina cDNA reads from roots of control (in green) and treated plants (in red) after mapping the 17 chromosomes of the sunflower genome. The distribution of LTR-REs on the chromosomes (in blue) was obtained by masking the genome of sunflower with the library of LTR-REs used for RT-encoding domains isolation (Natali et al. 2013). The 17 chromosomes of sunflower were also masked with a putative centromeric sequence (in black) and red arrows indicate the most probable centromere position in each chromosome, corresponding to the peaks of highest frequency of the putative centromeric sequence. The space of each chromosome is proportional to its length in nucleotides

Fig. 2 Number of RT-encoding domains belonging to different lineages, subdivided into untranscribed or barely expressed (FPKM < 1,500, in blue), lowly expressed (1,500 < FPKM < 15,000, in light blue), expressed (15,000 < FPKM < 40,000, in pink), and highly expressed (FPKM > 40,000, in red), in roots of control (above) and treated plants (below)

Fig. 3 Relationship between average coverage of a RT domain in the sunflower genome and FPKM in roots of control plants

Fig. 4 Expression of reverse transcriptase-encoding sequences in roots after treatment of *Helianthus annuus* plants with auxin (IAA), ethylene (ETH), gibberellic acid (GA3), salicylic acid (SA), kinetin (KIN), abscisic acid (ABA), strigolactones (STRI), brassinosteroids (BRA), polyethylene-glycol (PEG) and NaCl, and of control plants (C) (Badouin et al. 2017). The corresponding libraries were collected from the public database (see the materials and methods section). All the sequences reported were considered as expressed because mapped by than one read per million in at least one cDNA library. ID codes refer to the set of sunflower RT-encoding sequences (Natali et al. 2013). The occurrence of differential expression between treatments and control is indicated by the symbol "+" in case of significant over-expression and "-" in case of significant under-expression (fold change > 2 or < -2, respectively, with FDR-corrected $p < 0.05$)

Figure 2

Figure 3