Elsevier Editorial System(tm) for Gene
Manuscript Draft

Manuscript Number: GENE-D-17-00534R1

Title: Different histories of two highly variable LTR retrotransposons in sunflower species

Article Type: Research paper

Keywords: genome evolution; genome variation; Helianthus; long-terminal-repeat retrotransposon families; sunflower

Corresponding Author: Professor Lucia Natali,

Corresponding Author's Institution: University of Pisa

First Author: Flavia Mascagni

Order of Authors: Flavia Mascagni; Andrea Cavallini; Tommaso Giordani; Lucia Natali

Manuscript Region of Origin: ITALY

Abstract: In the Helianthus genus, very large intra- and interspecific variability related to two specific retrotransposons of Helianthus annuus (Helicopia and SURE) exists. When comparing these two sequences to sunflower sequence databases recently produced by our lab, the Helicopia family was shown to belong to the Maximus/SIRE lineage of the Sirevirus genus of the Copia superfamily, whereas the SURE element (whose superfamily was not even previously identified) was classified as a Gypsy element of the Ogre/Tat lineage of the Metavirus genus. Bioinformatic analysis of the two retrotransposon families revealed their genomic abundance and relative proliferation timing. The genomic abundance of these families differed significantly among 12 Helianthus species. The ratio between the abundance of long terminal repeats and their reverse transcriptases suggested that the SURE family has relatively more solo long terminal repeats than does Helicopia. Pairwise comparisons of Illumina reads encoding the reverse transcriptase domain indicated that SURE amplification may have occurred more recently than that of Helicopia. Finally, the analysis of population structure based on the SURE and Helicopia polymorphisms of 32 Helianthus species evidenced two subpopulations, which roughly corresponded to species of the Helianthus and Divaricati/Ciliares sections. However, a number of species showed an admixed structure, confirming the importance of interspecific hybridisation in the evolution of this genus. In general, these two retrotransposon families differentially contributed to interspecific variability, emphasising the need to refer to specific families when studying genome evolution.

Opposed Reviewers:

*Research Highlights

Different histories of two highly variable LTR retrotransposons in sunflower species

HIGHLIGHTS

- A pipeline was established to retrieve specific repeats from non-model species.
- Differences in retrotransposon dynamics were observed among species.
- Analyzing different families separately is necessary for understanding retrotransposon dynamics in genome evolution.

***Abbreviations list**

Abbreviations:

LTR, long terminal repeat

RE, retroelements

TE, transposable elements

RT, reverse transcriptase

Different histories of two highly variable LTR retrotransposons in sunflower species

Flavia Mascagni, Andrea Cavallini, Tommaso Giordani, Lucia Natali *

Dept. of Agriculture, Food, and Environment, University of Pisa, Via delBorghetto 80, I-56124 Pisa, Italy

Abbreviations: LTR, long terminal repeat; RE, retroelements; TE, transposable elements; RT, reverse transcriptase.

*Corresponding author
E-mail address: lucia.natali@unipi.it (L. Natali)

ABSTRACT

In the *Helianthus* genus, very large intra- and interspecific variability related to two specific retrotransposons of *Helianthus annuus* (*Helicopia* and *SURE*) exists. When comparing these two sequences to sunflower sequence databases recently produced by our lab, the *Helicopia* family was shown to belong to the *Maximus*/*SIRE* lineage of the *Sirevirus* genus of the *Copia* superfamily, whereas the *SURE* element (whose superfamily was not even previously identified) was classified as a *Gypsy* element of the *Ogre*/*Tat* lineage of the *Metavirus* genus. Bioinformatic analysis of the two retrotransposon families revealed their genomic abundance and relative proliferation timing. The genomic abundance of these families differed significantly among 12 *Helianthus* species. The ratio between the abundance of long terminal repeats and their reverse transcriptases suggested that the *SURE* family has relatively more solo long terminal repeats than does *Helicopia*. Pairwise comparisons of Illumina reads encoding the reverse transcriptase domain indicated that *SURE* amplification may have occurred more recently than that of *Helicopia*. Finally, the analysis of population structure based on the *SURE* and *Helicopia* polymorphisms of 32 *Helianthus* species evidenced two subpopulations, which roughly corresponded to species of the *Helianthus* and *Divaricati*/*Ciliares* sections. However, a number of species showed an admixed structure, confirming the importance of interspecific hybridisation in the evolution of this genus. In general, these two retrotransposon families differentially contributed to interspecific variability, emphasising the need to refer to specific families when studying genome evolution.

Keywords:

genome evolution

genome variation

Helianthus

long-terminal-repeat retrotransposon families

sunflower

1. Introduction

A large portion of plant genomes is composed of transposable elements (TEs), most of which generally belong to Class I and are called retrotransposons or retroelements (REs) because of their ‘copy and paste’ mechanism of replication, which resembles that of retroviruses (Wicker et al., 2007). The most abundant REs in plants are long terminal repeat (LTR) retrotransposons (LTR-REs); these elements are flanked by two LTRs. Between the 5'- and 3'-LTRs, there is a primer binding site and a polypurine tract that serve as the priming sites for the synthesis of minus- and plus-strand cDNAs by reverse transcriptase enzymes, respectively (Wicker et al., 2007). Autonomous REs contain one or more open reading frames (ORFs) that encode a GAG and a POL protein; the POL protein contains different domains that represent the enzymatic machinery required for retrotransposition, which includes a reverse transcriptase (RT), a protease, an RNAse, and an integrase (Boeke and Corces, 1989; Kumar and Bennetzen, 1999).

In plants, LTR-REs are subdivided into the *Copia* (*Pseudoviridae*) and *Gypsy* (*Metaviridae*) superfamilies based on the order and the sequence similarity of the enzymes within the ORFs (Wicker et al., 2007). Both superfamilies are ubiquitous throughout eukaryotes and have been present since the origin of eukaryotes (Kumar and Bennetzen, 1999). In turn, each superfamily is subdivided into three genera, *Pseudovirus*, *Hemivirus*, and *Sirevirus* for the *Copia* superfamily (Boeke et al., 2006) as well as *Metavirus*, *Errantivirus*, and *Chromovirus* for the *Gypsy* superfamily (Fauquet and Mayo, 2001). In higher plants, the LTR-RE genera consist of major evolutionary lineages (Wicker and Keller, 2007; Llorens et al., 2011). In the *Gypsy* superfamily, the *Metavirus* genus corresponds to the *Ogre*/*Tat* lineage (as described by Neumann et al., 2003), *Errantivirus* corresponds to the *Athila* lineage (described by Wright and Voytas, 2002), and *Chromovirus* to the *Chromovirus* lineage (Gorinsek et al., 2004; Llorens et al., 2011). On the other hand, the *Copia* *Pseudovirus* genus consists of many different lineages, including *AleI*/*Retrofit*/*Hopscotch*, *AleII*, *Angela*, *Bianca*, *Ivana*/*Oryco*, and *TAR*/*Tork*, as described by Wicker and Keller (2007), and the *Copia* *Sirevirus* genus consists of the *Maximus*/*SIRE* lineage (Bousios et al., 2010; Bousios and Darzentas, 2013). Within lineages, specific families of LTR-REs can be distinguished according to sequence similarity. Two LTR-REs belong to the same family if they show at least 80% sequence identity in 80% or more of their internal regions and/or their terminal repeat regions (Wicker et al., 2007).

The replicative activity of REs has produced genome diversification during species evolution, allowing insertions and recombinational losses (Kalendar et al., 2000; Neumann et al., 2006; Ammiraju et al., 2007; Hawkins et al., 2008; Morse et al., 2009). For example, unequal homologous recombination between paralogous elements on a chromosome can produce chromosomal mutations such as deletions or duplications (Ku et al., 2000).

LTR-REs are an excellent source of molecular markers in plant genomes because of their ubiquity, abundance, dispersion, and dynamism (Kalendar and Schulman, 2006). The inter-retrotransposon amplified polymorphism (IRAP; Kalendar et al., 1999) protocol can be used to analyse LTR-RE-related polymorphisms and relies on polymerase chain reaction (PCR) amplification between primers designed from one or two LTRs.

Vukich et al. (2009a) applied the IRAP protocol within the genus *Helianthus* for the first time to assess intra- and interspecific variability; these authors particularly focussed on the distinction between annual and perennial species. Two groups of LTRs, one belonging to an uncharacterised *Copia*-like RE (*Helicopia*) and the other to a putative RE of unknown nature (*SURE*), were isolated and sequenced, and primers were designed to obtain IRAP fingerprints. Jaccard's and Shannon's similarity indices (Jaccard, 1908; Shannon and Weaver, 1949) from binary matrices showed extreme variability of *Helicopia* and *SURE* elements among and within *Helianthus* species. Principal component analysis of IRAP fingerprints allowed the distinction between perennial and annual *Helianthus* species, especially for the *SURE* element.

The origin of the *Helianthus* genus was dated between 4.75 and 22.7 million years ago (MYA), and species within the genus diverged between 1.7 and 8.2 MYA (Schilling, 1997). The most recent molecular study on the evolution of the *Helianthus* genus (Timme et al., 2007) based on ribosomal external transcribed spacer sequences subdivided this genus into four sections: one consisted of the annual *H. agrestis*, the second (*Divaricati*) included perennial species and the annual *H. porteri*, the third (*Ciliares*) comprised perennial species, and the fourth (sect. *Helianthus*) contained all other annuals (including *H. annuus*). It should be noted, however, that separation between species is difficult to establish due to the recent species divergence and because many species are of hybrid origin (Rieseberg et al., 1995; Ungerer et al., 2006).

The genome of *H. annuus* was recently sequenced (Badouin et al., 2017). General surveys of LTR-REs and other repetitive DNAs in the genome of *H. annuus* had already been performed by assembling Illumina and 454 reads (Staton et al., 2012; Natali et al., 2013; Giordani et al., 2014; Mascagni et al., 2015). The resulting libraries revealed the occurrence of a number of different repeats (including LTR-RE lineages, DNA transposons, non-LTR-retrotransposons, and tandem repeats). These sequences constitute approximately 80% of the sunflower genome (i.e., all the repetitive portion of this species) (Badouin et al., 2017). The libraries are therefore representative of the repetitive DNA of this species.

The goal of this work was to establish a pipeline for characterising the specific families of repeated elements (rather than the whole RE complement as in the study by Natali et al. (2013) or LTR-RE lineages as in the study by Mascagni et al. (2015)) using high-throughput sequencing methods and applicable bioinformatic procedures, even in species whose genome has not been sequenced. Given the large variability observed in the *Helianthus* genus in polymorphism studies that focussed on *Helicopia* and *SURE* elements (Vukich et al., 2009a), we decided to analyse these two groups of LTR-REs in detail and to detect the putative evolutionary dynamics that produced the large interspecific variability related to these two retrotransposons.

2. Materials and Methods

2.1. Plant materials and DNA sequencing

The 32 species and subspecies used in these experiments are listed in Supplementary Table 1. All genotypes analysed are from United States Department of Agriculture, Agricultural Research Service, National Genetic Resources Program (ARS-GRIN). Additional data on the analysed genotypes can be found at National Germplasm Resources Laboratory homepage (http://www.ars-grin.gov/cgi-bin/npgs/acc/query.pl).

For DNA sequencing, genomic DNA was isolated from the leaflets of an individual of each of the 12 species and subspecies (Supplementary Table 2); this DNA was treated as a ‘type’ representative of the species. Of the selected species, four were annual, diploid and belonged to the section *Helianthus* (*H. annuus*, *H. argophyllus*, *H. niveus*, and *H. petiolaris*, including the two subspecies *H. petiolaris* ssp. *petiolaris* and *H. petiolaris* ssp. *fallax*), and seven were perennial and belonged to the section *Divaricati* (Timme et al., 2007). The selected *Divaricati* species included three diploid (*H. divaricatus*, *H. giganteus*, and *H. smithii*), three tetraploid (*H. hirsutus*, *H. californicus*, and *H. laevigatus*), and one hexaploid species (*H. tuberosus*). Regarding *H. annuus*, previous studies reported high variability in the repetitive component between wild and cultivated genotypes (Mascagni et al., 2015). In this study, a wild accession from Illinois was chosen to represent *H. annuus*; this particular accession exhibits average features among wild *H. annuus* genotypes (Mascagni et al., 2015).

DNA was isolated using a Nucleospin Plant Isolation kit (Macherey-Nagel) and C1 lysis buffer. This method is based on the cetyl-trimethylammonium bromide (CTAB) procedure. RNA contamination was removed by RNaseA treatment. The genomic DNA was dissolved in TE (1 mM ethylene-diamine-tetraacetic acid (EDTA), 10 mM Tris-HCl, pH 8.0) solution at 55 °C. DNA quality was assessed by visualisation after gel electrophoresis.

Paired-end libraries (insert size of 500–600 bp) were prepared from genomic DNAs using a TruSeq DNA sample kit (Illumina Inc., San Diego, CA, USA) following the standard protocol with minor modifications, after which the libraries were sequenced using an Illumina HiSeq 2000 platform. The sequence reads of two other species (*H. argophyllus* and *H. niveus*) were downloaded from the Sequence Read Archive at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/sra, accession numbers SRR2155086 and SRR2155080). All paired read sets were first checked for quality and trimmed to a length of 90 nucleotides (nt) using Trimmomatic (Bolger et al., 2014) to remove adapters and low-quality regions. To accomplish this, the following Trimmomatic parameters were used: ILLUMINACLIP:2:30:10; LEADING:15; TRAILING:15; SLIDINGWINDOW:4:15; CROP:90; and MINLEN:90. Finally, all reads containing organellar DNA sequences were removed using the software CLC-BIO Genomic Workbench 7.0.4 (CLC-BIO, Aarhus, Denmark; hereafter reported as CLC) against the chloroplast and mitochondrial sequences of *H. annuus* (NCBI reference sequences NC_007977 and KF815390, respectively).

2.2. Sequence isolation and characterisation in Helianthus *species*

The pipeline for *Helicopia* and *SURE* sequence isolation is reported in Fig. 1. In order to classify *Helicopia* and *SURE* elements, IRAP primers designed for these LTR-REs (CF, CR, U81, U82, and U89; Vukich et al., 2009a) were used to mask the available custom libraries of assembled repeated sequences of the highly inbred *H. annuus* lines HCM and HA412-HO (PI 642777) (Fig. 1) using RepeatMasker (Smit et al., 1996) under the default parameters but -div 20. The libraries used for the analysis are composed of contigs produced by assembling 454 and Illumina reads of *H. annuus*; these libraries include all repeat types reported for angiosperms and cover the whole repetitive component of the sunflower genome (Natali et al., 2013; Mascagni et al., 2015; available at the Sequence Repository website of the Department of Agriculture, Food, and Environment of the University of Pisa, http://pgagl.agr.unipi.it/sequence-repository/).

Classification of positive retrotransposon sequences was performed using BLASTX analysis against the non-redundant protein database of the NCBI at a threshold of 10^{-10}. Positive sequences were also annotated using the RepeatExplorer (Novák et al., 2010; 2013) protein domain search tool, which performed searches against the plant RepBase (Jurka et al., 2005) databases of protein domains (i.e., GAG, protease, RT, RNAseH, integrase, and chromodomain) derived from plant mobile elements, using the default parameters.

When RT domains were identified, their sequences were collected. In those cases in which the RT domain was not found, a maximum-900-nt-length sequence downstream of the forward primer was collected (Fig. 1); this sequence should include part, if not all, of the LTR.

All the collected sequences of *H. annuus* were subjected to BLASTN analysis against the genome sequence of another genotype of sunflower, XRQ (Badouin et al., 2017), in order to verify the occurrence of the sequences.

The selected *H. annuus* RT and LTR sequences were used to obtain consensus sequences of the homologous elements of the different *Helianthus* species by mapping Illumina reads of each species (Fig. 1). The resulting consensus sequences were collected and used for the analyses of *SURE* and *Helicopia* in all species.

To identify the lineage to which *SURE* and *Helicopia* belong, the translated RT domains of collected *H. annuus* sequences were aligned to RT domains of different species from the RepBase database using Clustal Omega (McWilliam et al., 2013). Afterward, phylogenetic trees were constructed using the neighbour-joining clustering method and multi-scale bootstrap resampling, which consisted of 1,000 bootstrap replications.

2.3. Analysis of retrotransposon abundance and proliferation

The abundance of the two selected REs in the genomes of *Helianthus* species was estimated by mapping the Illumina sets of reads of each species onto the consensus sequences of *SURE* and *Helicopia* of the same species and by calculating their average coverage (the sum of the bases of the aligned parts of all the reads divided by the length of the reference sequence). This parameter was chosen because it is comparable between species regardless of the length of the reference sequence and is especially useful when the total length of the related repeat is unknown.

Mapping was performed using CLC, which distributes multi-reads (i.e., reads that match multiple distinct sequences) randomly; hence, the average coverage of a single sequence is an indication of its redundancy only if multi-reads are not abundant. Mapping was performed using the following parameters: mismatch cost = 1, deletion cost = 1, insertion cost = 1, similarity fraction = 0.9, and length fraction = 0.9. Differences in abundance among species for each separate group of sequences were analysed according to Tukey’s test ($p < 0.05$).

The time course of *SURE* and *Helicopia* proliferation events in *Helianthus* species was inferred by examining the distributions of pairwise divergence values for Illumina reads aligned to the RT domains of the two RE groups, in accordance with the methods of Piegu et al. (2006) and Ammiraju et al. (2007). All Illumina 90-nt-long reads of each species were aligned to a portion (of 130 nt in length) of the respective homologous RT sequences of the same species. This reduced portion of the RT sequence was chosen in order to collect largely overlapping reads. For each species, a maximum of 100 aligned reads were collected. Afterward, pairwise divergence values between reads were determined using MEGA 7.0.18 software (Kumar et al., 2016) in accordance with the Kimura two-parameter model of sequence evolution (Kimura, 1980). Peaks of frequency distribution were interpreted as events of transposition burst. The peaks associated with lower values of divergence represented more recent proliferation events.

2.4. Analysis of population structure

IRAP bands reported by Vukich et al. (2009a) were used for analyses (Supplementary Fig. 1) and interpreted as (1) for presence or (0) for absence, assuming that each band represents a single locus (Lynch and Milligan, 1994). IRAP analysis was repeated three times, which produced three independent matrices. Non-reproducible bands were rare but were excluded from the analyses along with weak bands. Because of high IRAP variability among species and the large number of analysed genotypes, only bands that occurred in at least 20% of species were considered in some experiments.

The analysis of population structure for the detection of mixed genotypes was performed using the Bayesian method in the STRUCTURE 2.3.4 software package (Pritchard et al., 2000). The number of initial subpopulations (K) was defined from 1 to 35, and five replications were performed per run. The length of the burn-in period and the number of Markov Chain Monte Carlo replications were set to 50,000 and 100,000, respectively. The admixture model and correlated allele frequencies were chosen. The results were imported into Structure Harvester (Earl and Vonholdt, 2012) to determine the most likely number of K using the delta K (ΔK) method. In brief, Structure Harvester analyses both the logarithm of likelihood for each K (Ln P (D) = L (K)) (Rosenberg et al., 2002) and the ΔK statistic, the latter of which is based on the secondary rate of change in likelihood (ΔK= (L” (K)) / standard deviation) (Evanno et al., 2005). In this method, the probability of slope breaks at the point where the number of hypothetical K is at the maximum point of likelihood.

2.5. Data archiving

Raw reads of Illumina sequencing are accessible at NCBI SRA archive under the accession numbers SRR2919251 (*H. annuus*), SRR5713974 (*H. tuberosus*), SRR5713982 (*H. smithii*), SRR5713981 (*H. petiolaris* ssp. *fallax*), SRR5713980 (*H. petiolaris* ssp. *petiolaris*), SRR5713979 (*H. laevigatus*), SRR5713978 (*H. hirsutus*), SRR5713977 (*H. giganteus*), SRR5713976 (*H. divaricatus*), SRR5713975 (*H. californicus*), SRR2155086 (*H. argophyllus*), and SRR2155080 (*H. niveus*). All sequence collections described in this work are available at the repository sequence page of the Department of Agriculture, Food, and Environment of the University of Pisa (http://pgagl.agr.unipi.it/sequence-repository/).

3. Results

3.1. SURE *and* Helicopia *characterisation in* Helianthus *species*

The *SURE* and *Helicopia* LTR-REs display extensive variability in the *Helianthus* genus (Vukich et al., 2009a). In order to identify the superfamily, genus, and lineage to which these two elements belong and to isolate the corresponding sequences in different species of the *Helianthus* genus, a bioinformatics pipeline was established (Fig. 1). First, custom libraries of sunflower repetitive sequences (described by Natali et al. (2013) and by Mascagni et al. (2015); see Materials and Methods) were scored for the presence of primer sequences used by Vukich et al. (2009a). *SURE* primers were specific to putative LTRs isolated in accordance with the method established by Kalendar et al. (2008). *Helicopia* primers were based on an LTR sequence previously isolated by Natali et al. (2006). In most cases, all three *SURE* primers were adjacent in the same contig. In contrast, of the two *Helicopia* primers, only the CF primer was identified in most contigs, indicating that this LTR-RE family is highly variable in sequence.

The contigs containing the abovementioned primers were analysed using RepeatExplorer in order to identify DNA fragments corresponding to the RT domains of the two RE groups (Fig. 1). Eight *SURE* and five *Helicopia* RT-encoding sequences of *H. annuus* were collected. Neighbour-joining phylogenetic trees based on the multiple alignment of *Gypsy* and *Copia* RTs showed that SURE elements belong to the *Gypsy* superfamily (*Metaviridae*), *Metavirus* genus, and *Ogre*/*Tat* lineage and that *Helicopia* elements are members of the *Copia* superfamily (*Pseudoviridae*), *Sirevirus* genus, and *Maximus*/*SIRE* lineage (Fig. 2).

Whenever possible, sequences downstream of the *Helicopia* or *SURE* forward primers were also collected (Fig. 1). Based on this analysis, 23 *SURE* and 18 *Helicopia* sequences containing putative LTRs were retained. All sequences are listed in Supplementary Table 3 and were deposited in the Sequence Repository website of the Department of Agriculture, Food, and Environment of the University of Pisa. The occurrence of isolated sequences in the genome of another sunflower genotype, XRQ (Badouin et al., 2017), was verified using BLASTN analysis; all sequences were identified (Supplementary Table 4), and the probability ranged from 0 to 6.94 × e^{-137}.

The isolation of LTR and RT-encoding sequences was based on sequence similarity. According to the rules proposed by Wicker et al. (2007), two repeats belong to the same family if they share 80% (or more) sequence identity in at least 80% of their sequence. The isolated sequences fulfilled these conditions. Therefore, we attributed the isolated sequences as belonging to the *SURE* or *Helicopia* families.

To isolate corresponding LTR sequences and RT-encoding sequences from *Helianthus* wild species, Illumina reads of each species were aligned to the LTR and RT sequences of *H. annuus*, and consensus sequences for each species were built. By this method, at least five consensus sequences for each LTR and RT domain were produced for each RE and for each species (Fig. 1) and were used for subsequent analyses (Supplementary Table 5). The mean lengths of the isolated LTR and RT fragments were 127 and 492 nt for *Helicopia* and 454 and 312 nt for *SURE*, respectively.

3.2. Genomic abundance of SURE *and* Helicopia *in the* Helianthus *genus*

The relative abundance of *SURE* and *Helicopia* RT domains and LTRs was determined by mapping the Illumina reads of each species onto the isolated consensus sequences of the same species (Tenaillon et al., 2011; Natali et al., 2013; Barghini et al., 2015a; 2015b) using CLC. The CLC mapping algorithm maps multi-reads randomly among similar references, and multi-reads cannot be distinguished from exact duplicates. In these experiments, the number of multi-reads was less than 1% (data not shown). Hence, the random mapping of multi-reads did not significantly affect the abundance values of each element.

The percentages of mapping reads (corresponding to the genome proportion) of *Helicopia* LTRs ranged from 0.003% in *H. divaricatus* to 0.017% in *H. argophyllus*; those of *Helicopia* RT - encoding sequences ranged from 0.018% in *H. californicus* to 0.094% in *H. argophyllus*. The genome proportions of *SURE* LTRs ranged from 0.022% in *H. niveus* to 0.051% in *H. tuberosus*; *SURE* RT-encoding sequences ranged from 0.005% in *H. californicus* to 0.014% in *H. tuberosus*.

Fig. 3 shows the mean average coverage depth (i.e., the sum of the bases of the aligned parts of all the reads divided by the length of the reference sequence; see Materials and Methods) of each sequence type (*Helicopia* LTR and RT as well as *SURE* LTR and RT). Significant differences (according to Tukey's test) in average coverage among species regarding the *SURE* LTR and RT (see *H. niveus* vs. *H. tuberosus*) and *Helicopia* RT domain (see *H. smithii* vs. *H. tuberosus*) were recorded.

Fig. 4 shows the average coverage depth distribution of the LTRs and RT domains of *SURE* and *Helicopia* in the *Helianthus* species. Given that two LTRs occur in an RE, the average coverage of LTRs should be twice that of the corresponding coding portion (Cavallini et al., 2010). If the average coverage of LTRs is more than two fold, inter-LTR homologous recombination events may have occurred, resulting in the production of solo LTRs. Also, the occurrence of internal deletions in the retrotransposons could determine a higher number of LTRs than expected. However, inter-LTR homologous recombination is a process well known to commonly occur during genome evolution, and retrotransposon families that contain a high proportion of solo LTRs have been described in many plant species (Vicient et al., 1999).

Putative *Helicopia* LTRs were generally as abundant in the genome of each species as were *Helicopia* RT domains (Fig. 4B, C) (i.e., *Helicopia* LTRs seem under-represented in the genome of most species). It could be hypothesised that LTRs have experienced higher mutation rates than have RT-encoding domains and selected LTRs may not represent all LTRs of *Helicopia* elements.

In contrast, regarding the *Gypsy SURE* family, the average coverage median of the LTR region in each species was 2–3-fold greater than that of the RT domain (Fig. 4A, C). If LTRs accumulate more mutations than do RTs (as hypothesised for *Helicopia*), *SURE* elements might have relatively more solo LTRs than do *Helicopia* elements.

3.3. Temporal dynamics of SURE *and* Helicopia *families*

The timing of *SURE* and *Helicopia* family proliferation was inferred by analysing pairwise distances (Kimura, 1980) between paralogous RT-encoding sequences that belong to the same monophyletic groups of *Helicopia* and *SURE* elements, in accordance with the method of Piegu et al. (2006). The numbers of sequences of each species used for calculating pairwise distances are listed in Supplementary Table 6. Distances were translated into insertion dates in accordance with the methods of SanMiguel et al. (1996) and Piegu et al. (2006) but using a mutation rate of 2×10^{-8} (i.e., specific to sunflower and twice the rate calculated for synonymous substitutions in sunflower gene sequences). This mutation rate was used to keep into consideration that REs accumulate more mutations over time than do genes and to be consistent with previous analyses (Ungerer et al., 2009; Buti et al., 2011). In fact, at each insertion, the new retrotransposon copy is identical to its parental element, with the exception of mutations occurring during retrotranscription (which is error prone; Kumar and Bennetzen, 1999); additional mutations can then accumulate as time progresses.

This analysis enabled the identification of different retrotranspositional waves, mostly overlapping in terms of time between *SURE* and *Helicopia* and among species (Fig. 5). In 10 out of 12 species, *Helicopia* REs (Fig. 5A, B, C) were seemingly younger than were *SURE*s (mean insertion times of 3.31 and 4.93 MYA, respectively; Fig. 5D, E, F, G). The translation of genetic distances into insertion dates is subject to reservation; however, in our analyses, we only compared retrotransposition waves of the same RE family in different species.

Regarding *Helicopia*, proliferation seems to have begun earlier in the *Divaricati* species than in species belonging to the *Helianthus* section. All species of the *Helianthus* section showed a transpositional peak corresponding to approximately 2 MYA, and the mean insertion ages ranged from 2.2 to 3.5 MYA (Fig. 5C). The mean insertion time in *Divaricati* species was generally higher than that in species of the *Helianthus* section. *Helianthus tuberosus* showed two proliferation waves, one at approximately 4–5 MYA and the other at approximately 2 MYA (Fig. 5A), which were concurrent with the proliferation bursts observed in the other *Divaricati* species (Fig. 5B) and in the *Helianthus* section (Fig. 5C), respectively.

SURE retrotransposon dynamics were more complex, as four different proliferation profiles were observed among the analysed species (Fig. 5D, E, F, G). Retrotransposition waves largely overlapped in species of the *Helianthus* section and peaked at approximately 5 MYA, but the mean insertion times differed (ranging from 4.4–5.5 MYA) (Fig. 5G). Large differences were observed among the *Divaricati* species: *H. tuberosus*, *H. divaricatus*, and *H. smithii* showed two proliferation peaks (Fig. 5D) – the first was very ancient (approximately 10 MYA) and the second at 2–3 MYA (i.e., much more recent). *Helianthus hirsutus* and *H. californicus* showed a transpositional burst at 7–8 MYA (Fig. 5E). Only one, relatively recent, transpositional burst was observed in *H. giganteus* and *H. laevigatus* (Fig. 5F). Altogether, these results suggest an intriguing picture of species-specific increases in the abundance of these two RE families, in some cases in relatively recent times, subsequent to *Helianthus* speciation.

3.4. SURE- *and* Helicopia-*related mixed genotypes in the* Helianthus *genus*

The analysis of genome structure of the *Helianthus* genus according to the occurrence of mixed genotypes was performed using IRAPs of *SURE* and *Helicopia* elements in 32 species of the genus *Helianthus* (Vukich et al., 2009a). Using all polymorphic loci reported by Vukich et al. (2009a), subdivision in populations was not statistically supported, probably due to the very large variability.

Therefore, only bands occurring in at least 20% of species were retained and considered for this analysis.

A schematic representation of the analysed IRAP matrices is shown in Supplementary Fig. 1. This analysis included 32 sunflower species and 47 polymorphic loci, 28 of which involved *SURE*s and 19 of which were related to *Helicopia* elements. Effective analysis of the population structure and classification of species into appropriate groups were performed using the Bayesian method with STRUCTURE software (Pritchard et al., 2000). The number of initial subpopulations (K) was defined from 1 to 35, and there were five replications per run (Supplementary Fig. 1). The maximum value of ΔK was observed at K = 2, either considering all 47 loci or only 28 *SURE*-related loci. Therefore, the analysed sunflower species may consist of two subpopulations (Fig. 6). Statistical support was too low when only *Helicopia*-related loci were used.

We considered a genotype unequivocally assigned to a subpopulation when its admixture coefficient was > 80% (Qi > 0.8) for that group (Vigouroux et al., 2008; Castillo et al., 2010). Individuals with intermediate admixture coefficients (Qi < 0.8) were considered admixed. After STRUCTURE was applied, the species were classified into three groups at K = 2: (i) 12 species belonged to the *Helianthus* section; (ii) 12 species belonged to *Divaricati* and one to the *Ciliares* section; and (iii) seven admixed species comprised four that belonged to *Divaricati* and three that belonged to the *Helianthus* section (Fig. 6).

When considering only the 28 *SURE*-related loci (at K = 2), the assignment of the species to three groups was very similar, but a few differences were observed: the first group included 13 species, 12 of the *Helianthus* and one of the *Divaricati* section; the second group comprised 11 *Divaricati* and one *Ciliares* species; and the group of admixed species included four *Helianthus* and three *Divaricati* species. In particular, all *H. praecox* subspecies were admixed (Fig. 6). The observed discrepancies in genome structure when considering only the *SURE*-related loci compared with *SURE* + *Helicopia* loci might be related to different retrotransposition activity of the two RE families during species differentiation.

4. Discussion

The goal of this work was to characterise two specific families of LTR-REs of sunflower, *SURE* and *Helicopia* (Vukich et al. 2009a), and to analyse the evolutionary pathways of these families in the *Helianthus* genus. Previous studies on the repetitive component of the genome of sunflower species have focussed on global analyses of LTR-REs; studies on the behaviour of specific LTR-RE families in this genus are lacking.

SURE (*Metaviridae* (*Gypsy* superfamily), *Metavirus* genus, *Ogre*/*TAT* lineage) and *Helicopia* (*Pseudoviridae* (*Copia* superfamily), *Sirevirus* genus, *Maximus*/*SIRE* lineage) families showed no significant differences in abundance of LTRs and RT-encoding sequences within the genome of each species.

In contrast, differences between *SURE* and *Helicopia* families were found in relation to their different tendencies to be subjected to processes that imply DNA loss. Such processes may have affected the *SURE* family more than *Helicopia* (Fig. 4). In fact, the ratios between LTRs and RT-encoding sequence abundance indicate that *SURE* solo-LTRs are more abundant than are *Helicopia* ones. This might suggest that *SURE* elements are more prone to producing solo LTRs by local non-homologous recombination than are *Helicopia* elements. Processes such as DNA rearrangement and unequal homologous recombination drive DNA removal in plants by a number of mechanisms (Kalendar et al., 2000; Ma et al., 2004; Neumann et al., 2006; Ammiraju et al., 2007; Hawkins et al., 2008; Morse et al., 2009). On the other hand, it is also possible that *SURE*s had more time to accumulate solo LTRs, as *SURE*s are older than *Helicopia* ones (Fig. 5).

Differences in the proliferation time profiles between *SURE* and *Helicopia* families were also observed. *SURE* REs were on average older than were *Helicopia* in most of the analysed species (Fig. 5). It is known that proliferation bursts do not occur simultaneously for all RE families but show different timings in different RE families (Vitte and Panaud, 2003; Ammiraju et al., 2007). In sunflower, another *Copia* LTR-RE is potentially prone to a transpositional burst, as this LTR-RE is still active (i.e., it is regularly transcribed and, at low rates, is reinserted into the genome) (Vukich et al., 2009b).

Significant differences in the abundance of the two RE families were observed among species, at least for the *SURE* family, whereas the *Helicopia* family was more uniform (Fig. 3). This indicates that the equilibrium between RE amplification and loss differs among species. Such differences may have been casually produced during the evolution of *Helianthus* species. On the other hand, with the exceptions of *H. annuus* and to a minor extent *H. petiolaris* and *H. tuberosus*, all other analysed species have been reported to be distributed in relatively small areas (Heiser et al., 1969; Rogers et al., 1982). Such areas were often different among the analysed species, indicating that differences in the abundance of repetitive DNA might be correlated to the different environments in which the species live (i.e., such differences might be involved in the adaptation of the analysed genotypes to the environment).

In addition to differences in redundancy, the *SURE* family showed clear-cut differences among *Helianthus* species regarding proliferation time and proliferation profiles (Fig. 5). Differences related to the *Helicopia* family among *Helianthus* species are less defined than those related to *SURE* elements, probably because *Helicopia* proliferation is generally more recent than is *SURE* proliferation, especially in species of the *Helianthus* section (Fig. 5).

Analyses of chloroplast DNA sequences roughly dated the origin of the *Helianthus* genus to a time period between 22.7 and 4.75 MYA, and *Helianthus* species diverged between 8.2 and 1.7 MYA (Schilling, 1997). Although dating transpositional bursts is subject to reservation, *SURE* and *Helicopia* proliferation bursts seem to be concurrent with species divergence within the genus (Fig. 5).

The analysis of polymorphisms especially related to *SURE* elements showed the existence of two subpopulations in the *Helianthus* genus, roughly corresponding to the *Helianthus* section and to the *Divaricati*/*Ciliares* sections (Fig. 6). This result confirms the separation between annuals and perennials (Schilling et al., 1998; Santini et al., 2002; Natali et al., 2006) and allowed the discovery of species that have admixed structure. *Helicopia* family had a minor role compared with *SURE* in structuring the genus into two subpopulations (Fig. 6). Also this result might be explained by more recent proliferation of *Helicopia* elements compared with *SURE*s.

The presence of species with admixed genome structure is probably related to the events of interspecific hybridisation. Multiple interspecific hybridisation events have been important in the evolution of *Helianthus* species (Rieseberg et al., 1998), although dating and the extent of such events are not precisely known (Schilling, 1997). Interspecific hybridisation can involve transpositional bursts as a result of so-called genomic shock (i.e., response to the introduction of alien genetic material into a new genetic background) (McClintock, 1984). For example, the massive amplification of REs has been reported to have occurred relatively recently in *H. anomalus*, *H. deserticola*, and *H. paradoxus*, which are three species that originated by interspecific hybridisation between *H. annuus* and *H. petiolaris* (Ungerer et al., 2009). Transpositional peaks observed in the analysed *Helianthus* species for *SURE* and *Helicopia* elements may be related, at least in certain cases, to concurrent events of interspecific hybridisation.

Interestingly, *SURE* REs are more abundant in allopolyploid species (*H. tuberosus*, *H. hirsutus*, *H. laevigatus*, and *H. californicus*; Figs. 3 and 4) (i.e., in species in which an interspecific hybridisation event is shown by the presence of a multiple chromosome numbers). Genomic shock following interspecific hybridisation and subsequent chromosome doubling may have induced *SURE* proliferation. This phenomenon seems not to be true for *Helicopia*, indicating that each LTR-RE family responds specifically to genomic shock.

In general, even after analysing only two LTR-RE families, our data reveal significant differences in the evolutionary trends between these families. These differences point out the necessity, when studying retrotransposons and genome evolution, of analysing (in addition to the general characterisation of the repetitive component of the genome) genome structure separately for specific RE families. The availability of the complete genome sequence of *H. annuus* (Badouin et al., 2017) will allow the comprehensive analysis of every RE family and will establish whether the behaviour of *Helicopia* and *SURE* families in the evolution of the *Helianthus* genome and the related interspecific variability are specific to these two elements.

Funding

This research work was funded by DiSAAA-a, University of Pisa, Project "Plantomics".

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://pgagl.agr.unipi.it/sequence-repository/

References

Ammiraju, J.S., Zuccolo, A., Yu, Y., Song, X., Piegu, P., Chevalier, F., Walling, J.G., Ma, J., Talag, J., Brar, D.S., SanMiguel, P.J., Jiang, N., Jackson, S.A., Panaud, O., Wing, R.A., 2007. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J. 52, 342–351.

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.

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

Barghini, E., Natali, L., Giordani, T., Cossu, R.M., Scalabrin, S., Cattonaro, F., Šimková, H., Vrána, J., Doležel, J., Morgante, M., Cavallini, A., 2015b. LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. DNA Research 22, 91–100.

Boeke, J.D., Corces, V.G., 1989. Transcription and reverse transcription of retrotransposons. Ann. Rev. Microbiol. 43, 403–434.

Boeke, J.D., Eickbush, T.H., Sandmeyer, S.B., Voytas, D.F., 2006. Index of viruses - Pseudoviridae. ICTVdB - The Universal Virus Database, version 4, Columbia University, New York, USA.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bousios, A., Darzentas, N., Tsaftaris, A., Pearce, S.R., 2010. Highly conserved motifs in non-coding regions of *Sirevirus* retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11, 89.

Bousios, A., Darzentas, N., 2013. *Sirevirus* LTR retrotransposons: phylogenetic misconceptions in the plant world. Mobile DNA 4, 9.

Buti, M., Giordani, T., Cattonaro, F., Cossu, R.M., Pistelli, L., Vukich, M., Morgante, M., Cavallini, A., Natali, L., 2011. Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. Theor. Appl. Genet. 123, 779–791.

Castillo, A., Dorado, G., Feuillet, C., Sourdille, P., Hernandez, P., 2010. Genetic structure and ecogeographical adaptation in wild barley (*Hordeum chilense* Roemer et Schultes) as revealed by microsatellite markers. BMC Plant Biol. 10, 266.

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

Earl, D.A., Vonholdt, B.M., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet. Res. 4, 359-361.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611-2620.

Fauquet, C.M., Mayo, M.A., 2001. The 7th ICTV Report. Arch. Virol. 146, 189-194.

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

Gorinsek, B., Gubensek, F., Kordis, D., 2004. Evolutionary genomics of chromoviruses in eukaryotes. Mol. Biol. Evol. 21, 781–798.

Hawkins, J.S., Hu, G., Rapp, R.A., Grafenberg, J.L., Wendel, J.F. 2008. Phylogenetic determination of the pace of transposable element proliferation in plants: *Copia* and LINE-like elements in *Gossypium*. Genome 51, 11–18.

Heiser, C.B., Smith, D.M., Clevenger, S.B., Martin, W.C., 1969. North American sunflowers (*Helianthus*). Torrey Bot. Club Mem. 22, 1-218.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44, 223–270.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Kalendar, R., Grob, T., Regina, M., Suoniemi, A., Schulman, A.H., 1999. IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor. Appl. Genet. 98, 704–711.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA 97, 6603–6607.

Kalendar, R., Schulman, A.H., 2006. IRAP and REMAP for retrotransposon based genotyping and fingerprinting. Nature Protocols 1, 2478–2484.

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., Schulman, A.H., 2008. Cassandra retrotransposons carry independently transcribed 5S RNA. Proc. Natl. Acad. Sci. USA 105, 5833–5838.

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.

Ku, H.M., Vision, T., Liu, J.P., Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA 97, 9121–9126.

Kumar, A., Bennetzen, J.L., 1999. Plant retrotransposons. Annu. Rev. Genet. 33, 479-532.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., Latorre, A., Moya, A., 2011. The *Gypsy* database (GyDB) of mobile genetic elements: release 2.0. Nucl. Acids Res. 39, D70–D74.

Lynch, M., Milligan, B.G., 1994. Analysis of population genetic structure with RAPD markers. Mol. Ecol. 3, 91-99.

Ma, J., Devos, K.M., Bennetzen, J.L., 2004. Analyses of LTR retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14, 860–869.

Mascagni, F., Barghini, E., Giordani, T., Rieseberg, L.H., 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.

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

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., Lopez, R., 2013. Analysis tool web services from the EMBL-EBI. Nucl. Acids Res. 41, W597-W600.

Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., Kubisiak, T.L., Amerson, H.V., Carlson, J.E., Nelson, C.D., Davis, J.M., 2009. Evolution of genome size and complexity in *Pinus*. PLoS One 4, e4332.

Natali, L., Santini, S., Giordani, T., Minelli, S., Maestrini, P., Cionini, P.G., Cavallini, A., 2006. Distribution of Ty3-*gypsy*- and Ty1-*copia*-like DNA sequences in the genus *Helianthus* and other *Asteraceae*. Genome 49, 64-72.

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg L., Cavallini, A., 2013. The repetitive component of the sunflower genome as revealed by different procedures for assembling next generation sequencing reads. BMC Genomics 14, 686.

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

Neumann, P., Koblizkova, A., Navratilova, A., Macas, J., 2006. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173, 1047–1056.

Novák, P., Neumann, P., Macas, J., 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11, 378.

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.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., Panaud, O., 2006. Doubling genome size without polyploidization: dynamics of retrotransposition driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 16, 1262–1269.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rieseberg, L.H., Van Fossen, C., Desrochers, A., 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature 375, 313–316.

Rieseberg, L.H., Baird, S.J.E., Desrochers, A.M., 1998. Patterns of mating in wild sunflower hybrid zones. Evolution 52, 713-726.

Rogers, C.E., Thompson, T.E., Seiler, G.J., 1982. Sunflower species of the United States. National Sunflower Association, Bismarck, North Dakota.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. The genetic structure of human populations. Science 298, 2381-2385.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

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

Schilling, E.E., 1997. Phylogenetic analysis of *Helianthus* (*Asteraceae*) based on chloroplast DNA restriction site data. Theor. Appl. Genet. 94, 925-933.

Schilling, E.E., Linder, C.R., Noyes, R.D., Rieseberg, L.H., 1998. Phylogenetic relationships in *Helianthus* (*Asteraceae*) based on nuclear ribosomal DNA internal transcribed spacer region sequence data. Syst. Bot. 23, 177-187.

Shannon, C.E., Weaver, W., 1949. The mathematical theory of communication. University of Illinois Press, Urbana, IL.

Smit, A.F.A., Hubley, R., Green, P., 1996. RepeatMasker. Open-3.0. Available at http://www.repeatmasker.org/et al.

Staton, S.E., Bakken, B.E., Blackman, B.K., Chapman, M.A., Kane, N.C., Tang, S., Ungerer, M.C., Knapp, S.J., Rieseberg, L.H., Burke, J.M., 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72, 142–153.

Tenaillon, M.I., Hufford, M.B., Gaut, B.S., Ross-Ibarra, J., 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol. Evol. 3, 219–229.

Timme, R.E., Simpson, B.B., Randal Linder, C., 2007. High-resolution phylogeny for *Helianthus* (*Asteraceae*) using the 18S-26S ribosomal DNA external transcribed spacer. Am. J. Bot. 94, 1837–1852.

Ungerer, M.C., Strakosh, S.C., Zhen, Y., 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. 16, R872–R873.

Ungerer, M.C., Strakosh, S.C., Stimpson, K.M., 2009. Proliferation of Ty3/*Gypsy*-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC Biol. 7, 40.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., Schulman, A.H., 1999. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. The Plant Cell 11, 1769–1784.

Vigouroux, Y., Glaubitz, J.C., Matsuoka, Y., Major, M., Doebley, J., 2008. Population structure and genetic diversity of the new world maize races assessed by microsatellites. Am. J. Bot. 95, 1240–1253.

Vitte, C., Panaud, O., 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice (*Oryza sativa* L.). Mol. Biol. Evol. 20, 528-540.

Vukich, M., Schulman, A.H., Giordani, T., Natali, L., Kalendar, R., Cavallini, A., 2009a. Genetic variability in sunflower (*Helianthus annuus* L.) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. Theor. Appl. Genet. 119, 1027–1038.

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

Wicker, T., Keller, B., 2007. Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res. 17, 1072–1081.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification systemfor eukaryotic transposable elements. Nature Rev. Genet. 8, 973–982.

Wright, D.A., Voytas, D.F., 2002. *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res. 12, 122–131.

Figure Legends

Fig. 1. Pipeline for the isolation of LTRs and RT-encoding sequences of *SURE* and *Helicopia* REs of 12 species of *Helianthus*.

Fig. 2. Neighbour-joining phylogenetic tree based on multiple alignment of (A) the eight RT amino acid sequences of *SURE*- and *Gypsy*-related RT amino acid sequences of several plant species and (B) the five RT amino acid sequences of *Helicopia*- and *Copia*-related RT amino acid sequences of several plant species. Bootstrap values greater than 0.6 are shown with an asterisk.

Fig. 3. Average coverage of LTRs and RT-encoding sequences of *SURE* (left) and *Helicopia* (right) RE families of 12 species of *Helianthus* (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species). The mean ± standard error is reported for each species. Significant differences for each separate group of measurements are indicated by different letters ($p < 0.05$) according to Tukey's test.

Fig. 4. (A, B) Box and whisker plots of the average coverage of consensus LTRs and RT-encoding sequences of *Helicopia* and *SURE* REs of 12 species of *Helianthus* (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species). Boxes represent 25–75% of the values, whiskers represent the whole range of values, and lines in the box represent the median values of the distribution. (C) The ratio between the median of the average coverage of consensus LTRs and the median average coverage of the consensus RT-encoding sequences of *Helicopia* and *SURE* REs in 12 sunflower species.

Fig. 5. Timing of *Helicopia* and *SURE* retrotranspositional activity in 12 species of *Helianthus* based on the pairwise comparisons of Illumina reads that match RT-encoding sequences. Graphs A, B, and C refer to *Helicopia*; graphs D, E, F, and G, *SURE*. To facilitate comparisons, each graph combines species with similar profiles. Species of the *Divaricati* section (perennials) are shown in graphs A, B, D, E, and F; species of the *Helianthus* section (annuals), graphs C and G. The y axis reports the product of the percentage of pairwise comparisons for the average coverage of the RT sequence in each species in order to account for the extent of transpositional bursts. The mean insertion times for each species and for the analysed RE families are reported in parentheses.

Fig. 6. Proportions of the ancestry of 32 *Helianthus* species and subspecies (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species; C: section *Ciliares*, perennial species) based on K = 2 (where K is the number of initial subpopulations). The proportions were obtained using STRUCTURE software for IRAP matrices obtained from the electrophoresis of PCR products amplified using *SURE* and *Helicopia* primers (above) or *SURE* primers only (below).

Different histories of two highly variable LTR retrotransposons in sunflower species

Flavia Mascagni, Andrea Cavallini, Tommaso Giordani, Lucia Natali *

Dept. of Agriculture, Food, and Environment, University of Pisa, Via delBorghetto 80, I-56124 Pisa, Italy

Abbreviations: LTR, long terminal repeat; RE, retroelements; TE, transposable elements; RT, reverse transcriptase.

*Corresponding author
E-mail address: lucia.natali@unipi.it (L. Natali)

ABSTRACT

In the *Helianthus* genus, very large intra- and interspecific variability related to two specific retrotransposons of *Helianthus annuus* (*Helicopia* and *SURE*) exists. When comparing these two sequences to sunflower sequence databases recently produced by our lab, the *Helicopia* family was shown to belong to the *Maximus*/*SIRE* lineage of the *Sirevirus* genus of the *Copia* superfamily, whereas the *SURE* element (whose superfamily was not even previously identified) was classified as a *Gypsy* element of the *Ogre*/*Tat* lineage of the *Metavirus* genus. Bioinformatic analysis of the two retrotransposon families revealed their genomic abundance and relative proliferation timing. The genomic abundance of these families differed significantly among 12 *Helianthus* species. The ratio between the abundance of long terminal repeats and their reverse transcriptases suggested that the *SURE* family has relatively more solo long terminal repeats than does *Helicopia*. Pairwise comparisons of Illumina reads encoding the reverse transcriptase domain indicated that *SURE* amplification may have occurred more recently than that of *Helicopia*. Finally, the analysis of population structure based on the *SURE* and *Helicopia* polymorphisms of 32 *Helianthus* species evidenced two subpopulations, which roughly corresponded to species of the *Helianthus* and *Divaricati*/*Ciliares* sections. However, a number of species showed an admixed structure, confirming the importance of interspecific hybridisation in the evolution of this genus. In general, these two retrotransposon families differentially contributed to interspecific variability, emphasising the need to refer to specific families when studying genome evolution.

Keywords:
genome evolution
genome variation
Helianthus
long-terminal-repeat retrotransposon families
sunflower

1. Introduction

A large portion of plant genomes is composed of transposable elements (TEs), most of which generally belong to Class I and are called retrotransposons or retroelements (REs) because of their ‘copy and paste’ mechanism of replication, which resembles that of retroviruses (Wicker et al., 2007). The most abundant REs in plants are long terminal repeat (LTR) retrotransposons (LTR-REs); these elements are flanked by two LTRs. Between the 5'- and 3'-LTRs, there is a primer binding site and a polypurine tract that serve as the priming sites for the synthesis of minus- and plus-strand cDNAs by reverse transcriptase enzymes, respectively (Wicker et al., 2007). Autonomous REs contain one or more open reading frames (ORFs) that encode a GAG and a POL protein; the POL protein contains different domains that represent the enzymatic machinery required for retrotransposition, which includes a reverse transcriptase (RT), a protease, an RNAse, and an integrase (Boeke and Corces, 1989; Kumar and Bennetzen, 1999).

In plants, LTR-REs are subdivided into the *Copia* (*Pseudoviridae*) and *Gypsy* (*Metaviridae*) superfamilies based on the order and the sequence similarity of the enzymes within the ORFs (Wicker et al., 2007). Both superfamilies are ubiquitous throughout eukaryotes and have been present since the origin of eukaryotes (Kumar and Bennetzen, 1999). In turn, each superfamily is subdivided into three genera, *Pseudovirus*, *Hemivirus*, and *Sirevirus* for the *Copia* superfamily (Boeke et al., 2006) as well as *Metavirus*, *Errantivirus*, and *Chromovirus* for the *Gypsy* superfamily (Fauquet and Mayo, 2001). In higher plants, the LTR-RE genera consist of major evolutionary lineages (Wicker and Keller, 2007; Llorens et al., 2011). In the *Gypsy* superfamily, the *Metavirus* genus corresponds to the *Ogre*/*Tat* lineage (as described by Neumann et al., 2003), *Errantivirus* corresponds to the *Athila* lineage (described by Wright and Voytas, 2002), and *Chromovirus* to the *Chromovirus* lineage (Gorinsek et al., 2004; Llorens et al., 2011). On the other hand, the *Copia Pseudovirus* genus consists of many different lineages, including *AleI*/*Retrofit*/*Hopscotch*, *AleII*, *Angela*, *Bianca*, *Ivana*/*Oryco*, and *TAR*/*Tork*, as described by Wicker and Keller (2007), and the *Copia Sirevirus* genus consists of the *Maximus*/*SIRE* lineage (Bousios et al., 2010; Bousios and Darzentas, 2013). Within lineages, specific families of LTR-REs can be distinguished according to sequence similarity. Two LTR-REs belong to the same family if they show at least 80% sequence identity in 80% or more of their internal regions and/or their terminal repeat regions (Wicker et al., 2007).

The replicative activity of REs has produced genome diversification during species evolution, allowing insertions and recombinational losses (Kalendar et al., 2000; Neumann et al., 2006; Ammiraju et al., 2007; Hawkins et al., 2008; Morse et al., 2009). For example, unequal homologous recombination between paralogous elements on a chromosome can produce chromosomal mutations such as deletions or duplications (Ku et al., 2000).

LTR-REs are an excellent source of molecular markers in plant genomes because of their ubiquity, abundance, dispersion, and dynamism (Kalendar and Schulman, 2006). The inter-retrotransposon amplified polymorphism (IRAP; Kalendar et al., 1999) protocol can be used to analyse LTR-RE-related polymorphisms and relies on polymerase chain reaction (PCR) amplification between primers designed from one or two LTRs.

Vukich et al. (2009a) applied the IRAP protocol within the genus *Helianthus* for the first time to assess intra- and interspecific variability; these authors particularly focussed on the distinction between annual and perennial species. Two groups of LTRs, one belonging to an uncharacterised *Copia*-like RE (*Helicopia*) and the other to a putative RE of unknown nature (*SURE*), were isolated and sequenced, and primers were designed to obtain IRAP fingerprints. Jaccard’s and Shannon’s similarity indices (Jaccard, 1908; Shannon and Weaver, 1949) from binary matrices showed extreme variability of *Helicopia* and *SURE* elements among and within *Helianthus* species. Principal component analysis of IRAP fingerprints allowed the distinction between perennial and annual *Helianthus* species, especially for the *SURE* element.

The origin of the *Helianthus* genus was dated between 4.75 and 22.7 million years ago (MYA), and species within the genus diverged between 1.7 and 8.2 MYA (Schilling, 1997). The most recent molecular study on the evolution of the *Helianthus* genus (Timme et al., 2007) based on ribosomal external transcribed spacer sequences subdivided this genus into four sections: one consisted of the annual *H. agrestis*, the second (*Divaricati*) included perennial species and the annual *H. porteri*, the third (*Ciliares*) comprised perennial species, and the fourth (sect. *Helianthus*) contained all other annuals (including *H. annuus*). It should be noted, however, that separation between species is difficult to establish due to the recent species divergence and because many species are of hybrid origin (Rieseberg et al., 1995; Ungerer et al., 2006).

The genome of *H. annuus* was recently sequenced (Badouin et al., 2017). General surveys of LTR-REs and other repetitive DNAs in the genome of *H. annuus* had already been performed by assembling Illumina and 454 reads (Staton et al., 2012; Natali et al., 2013; Giordani et al., 2014; Mascagni et al., 2015). The resulting libraries revealed the occurrence of a number of different repeats (including LTR-RE lineages, DNA transposons, non-LTR-retrotransposons, and tandem repeats). These sequences constitute approximately 80% of the sunflower genome (i.e., all the repetitive portion of this species) (Badouin et al., 2017). The libraries are therefore representative of the repetitive DNA of this species.

The goal of this work was to establish a pipeline for characterising the specific families of repeated elements (rather than the whole RE complement as in the study by Natali et al. (2013) or LTR-RE lineages as in the study by Mascagni et al. (2015)) using high-throughput sequencing methods and applicable bioinformatic procedures, even in species whose genome has not been sequenced. Given the large variability observed in the *Helianthus* genus in polymorphism studies that focussed on *Helicopia* and *SURE* elements (Vukich et al., 2009a), we decided to analyse these two groups of LTR-REs in detail and to detect the putative evolutionary dynamics that produced the large interspecific variability related to these two retrotransposons.

2. Materials and Methods

2.1. Plant materials and DNA sequencing

The 32 species and subspecies used in these experiments are listed in Supplementary Table 1. All genotypes analysed are from United States Department of Agriculture, Agricultural Research Service, National Genetic Resources Program (ARS-GRIN). Additional data on the analysed genotypes can be found at National Germplasm Resources Laboratory homepage (http://www.ars-grin.gov/cgi-bin/npgs/acc/query.pl).

For DNA sequencing, genomic DNA was isolated from the leaflets of an individual of each of the 12 species and subspecies (Supplementary Table 2); this DNA was treated as a ‘type’ representative of the species. Of the selected species, four were annual, diploid and belonged to the section *Helianthus* (*H. annuus*, *H. argophyllus*, *H. niveus*, and *H. petiolaris*, including the two subspecies *H. petiolaris* ssp. *petiolaris* and *H. petiolaris* ssp. *fallax*), and seven were perennial and belonged to the section *Divaricati* (Timme et al., 2007). The selected *Divaricati* species included three diploid (*H. divaricatus*, *H. giganteus*, and *H. smithii*), three tetraploid (*H. hirsutus*, *H. californicus*, and *H. laevigatus*), and one hexaploid species (*H. tuberosus*). Regarding *H. annuus*, previous studies reported high variability in the repetitive component between wild and cultivated genotypes (Mascagni et al., 2015). In this study, a wild accession from Illinois was chosen to represent *H. annuus*; this particular accession exhibits average features among wild *H. annuus* genotypes (Mascagni et al., 2015).

DNA was isolated using a Nucleospin Plant Isolation kit (Macherey-Nagel) and C1 lysis buffer. This method is based on the cetyl-trimethylammonium bromide (CTAB) procedure. RNA contamination was removed by RNaseA treatment. The genomic DNA was dissolved in TE (1 mM ethylene-diamine-tetraacetic acid (EDTA), 10 mM Tris-HCl, pH 8.0) solution at 55 °C. DNA quality was assessed by visualisation after gel electrophoresis.

Paired-end libraries (insert size of 500–600 bp) were prepared from genomic DNAs using a TruSeq DNA sample kit (Illumina Inc., San Diego, CA, USA) following the standard protocol with minor modifications, after which the libraries were sequenced using an Illumina HiSeq 2000 platform. The sequence reads of two other species (*H. argophyllus* and *H. niveus*) were downloaded from the Sequence Read Archive at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/sra, accession numbers SRR2155086 and SRR2155080). All paired read sets were first checked for quality and trimmed to a length of 90 nucleotides (nt) using Trimmomatic (Bolger et al., 2014) to remove adapters and low-quality regions. To accomplish this, the following Trimmomatic parameters were used: ILLUMINACLIP:2:30:10; LEADING:15; TRAILING:15; SLIDINGWINDOW:4:15; CROP:90; and MINLEN:90. Finally, all reads containing organellar DNA sequences were removed using the software CLC-BIO Genomic Workbench 7.0.4 (CLC-BIO, Aarhus, Denmark; hereafter reported as CLC) against the chloroplast and mitochondrial sequences of *H. annuus* (NCBI reference sequences NC_007977 and KF815390, respectively).

2.2. Sequence isolation and characterisation in Helianthus *species*

The pipeline for *Helicopia* and *SURE* sequence isolation is reported in Fig. 1. In order to classify *Helicopia* and *SURE* elements, IRAP primers designed for these LTR-REs (CF, CR, U81, U82, and U89; Vukich et al., 2009a) were used to mask the available custom libraries of assembled repeated sequences of the highly inbred *H. annuus* lines HCM and HA412-HO (PI 642777) (Fig. 1) using RepeatMasker (Smit et al., 1996) under the default parameters but -div 20. The libraries used for the analysis are composed of contigs produced by assembling 454 and Illumina reads of *H. annuus*; these libraries include all repeat types reported for angiosperms and cover the whole repetitive component of the sunflower genome (Natali et al., 2013; Mascagni et al., 2015; available at the Sequence Repository website of the Department of Agriculture, Food, and Environment of the University of Pisa, http://pgagl.agr.unipi.it/sequence-repository/).

Classification of positive retrotransposon sequences was performed using BLASTX analysis against the non-redundant protein database of the NCBI at a threshold of 10^{-10}. Positive sequences were also annotated using the RepeatExplorer (Novák et al., 2010; 2013) protein domain search tool, which performed searches against the plant RepBase (Jurka et al., 2005) databases of protein domains (i.e., GAG, protease, RT, RNAseH, integrase, and chromodomain) derived from plant mobile elements, using the default parameters.

When RT domains were identified, their sequences were collected. In those cases in which the RT domain was not found, a maximum-900-nt-length sequence downstream of the forward primer was collected (Fig. 1); this sequence should include part, if not all, of the LTR.

All the collected sequences of *H. annuus* were subjected to BLASTN analysis against the genome sequence of another genotype of sunflower, XRQ (Badouin et al., 2017), in order to verify the occurrence of the sequences.

The selected *H. annuus* RT and LTR sequences were used to obtain consensus sequences of the homologous elements of the different *Helianthus* species by mapping Illumina reads of each species (Fig. 1). The resulting consensus sequences were collected and used for the analyses of *SURE* and *Helicopia* in all species.

To identify the lineage to which *SURE* and *Helicopia* belong, the translated RT domains of collected *H. annuus* sequences were aligned to RT domains of different species from the RepBase database using Clustal Omega (McWilliam et al., 2013). Afterward, phylogenetic trees were constructed using the neighbour-joining clustering method and multi-scale bootstrap resampling, which consisted of 1,000 bootstrap replications.

2.3. Analysis of retrotransposon abundance and proliferation

The abundance of the two selected REs in the genomes of *Helianthus* species was estimated by mapping the Illumina sets of reads of each species onto the consensus sequences of *SURE* and *Helicopia* of the same species and by calculating their average coverage (the sum of the bases of the aligned parts of all the reads divided by the length of the reference sequence). This parameter was chosen because it is comparable between species regardless of the length of the reference sequence and is especially useful when the total length of the related repeat is unknown.

Mapping was performed using CLC, which distributes multi-reads (i.e., reads that match multiple distinct sequences) randomly; hence, the average coverage of a single sequence is an indication of its redundancy only if multi-reads are not abundant. Mapping was performed using the following parameters: mismatch cost = 1, deletion cost = 1, insertion cost = 1, similarity fraction = 0.9, and length fraction = 0.9. Differences in abundance among species for each separate group of sequences were analysed according to Tukey’s test ($p < 0.05$).

The time course of *SURE* and *Helicopia* proliferation events in *Helianthus* species was inferred by examining the distributions of pairwise divergence values for Illumina reads aligned to the RT domains of the two RE groups, in accordance with the methods of Piegu et al. (2006) and Ammiraju et al. (2007). All Illumina 90-nt-long reads of each species were aligned to a portion (of 130 nt in length) of the respective homologous RT sequences of the same species. This reduced portion of the RT sequence was chosen in order to collect largely overlapping reads. For each species, a maximum of 100 aligned reads were collected. Afterward, pairwise divergence values between reads were determined using MEGA 7.0.18 software (Kumar et al., 2016) in accordance with the Kimura two-parameter model of sequence evolution (Kimura, 1980). Peaks of frequency distribution were interpreted as events of transposition burst. The peaks associated with lower values of divergence represented more recent proliferation events.

2.4. Analysis of population structure

IRAP bands reported by Vukich et al. (2009a) were used for analyses (Supplementary Fig. 1) and interpreted as (1) for presence or (0) for absence, assuming that each band represents a single locus (Lynch and Milligan, 1994). IRAP analysis was repeated three times, which produced three independent matrices. Non-reproducible bands were rare but were excluded from the analyses along with weak bands. Because of high IRAP variability among species and the large number of analysed genotypes, only bands that occurred in at least 20% of species were considered in some experiments.

The analysis of population structure for the detection of mixed genotypes was performed using the Bayesian method in the STRUCTURE 2.3.4 software package (Pritchard et al., 2000). The number of initial subpopulations (K) was defined from 1 to 35, and five replications were performed per run. The length of the burn-in period and the number of Markov Chain Monte Carlo replications were set to 50,000 and 100,000, respectively. The admixture model and correlated allele frequencies were chosen. The results were imported into Structure Harvester (Earl and Vonholdt, 2012) to determine the most likely number of K using the delta K (ΔK) method. In brief, Structure Harvester analyses both the logarithm of likelihood for each K (Ln P (D) = L (K)) (Rosenberg et al., 2002) and the ΔK statistic, the latter of which is based on the secondary rate of change in likelihood (ΔK= (L” (K)) / standard deviation) (Evanno et al., 2005). In this method, the probability of slope breaks at the point where the number of hypothetical K is at the maximum point of likelihood.

2.5. Data archiving

Raw reads of Illumina sequencing are accessible at NCBI SRA archive under the accession numbers SRR2919251 (*H. annuus*), SRR5713974 (*H. tuberosus*), SRR5713982 (*H. smithii*), SRR5713981 (*H. petiolaris* ssp. *fallax*), SRR5713980 (*H. petiolaris* ssp. *petiolaris*), SRR5713979 (*H. laevigatus*), SRR5713978 (*H. hirsutus*), SRR5713977 (*H. giganteus*), SRR5713976 (*H. divaricatus*), SRR5713975 (*H. californicus*), SRR2155086 (*H. argophyllus*), and SRR2155080 (*H. niveus*). All sequence collections described in this work are available at the repository sequence page of the Department of Agriculture, Food, and Environment of the University of Pisa (http://pgagl.agr.unipi.it/sequence-repository/).

3. Results

3.1. SURE *and* Helicopia *characterisation in* Helianthus *species*

The *SURE* and *Helicopia* LTR-REs display extensive variability in the *Helianthus* genus (Vukich et al., 2009a). In order to identify the superfamily, genus, and lineage to which these two elements belong and to isolate the corresponding sequences in different species of the *Helianthus* genus, a bioinformatics pipeline was established (Fig. 1). First, custom libraries of sunflower repetitive sequences (described by Natali et al. (2013) and by Mascagni et al. (2015); see Materials and Methods) were scored for the presence of primer sequences used by Vukich et al. (2009a). *SURE* primers were specific to putative LTRs isolated in accordance with the method established by Kalendar et al. (2008). *Helicopia* primers were based on an LTR sequence previously isolated by Natali et al. (2006). In most cases, all three *SURE* primers were adjacent in the same contig. In contrast, of the two *Helicopia* primers, only the CF primer was identified in most contigs, indicating that this LTR-RE family is highly variable in sequence.

The contigs containing the abovementioned primers were analysed using RepeatExplorer in order to identify DNA fragments corresponding to the RT domains of the two RE groups (Fig. 1). Eight *SURE* and five *Helicopia* RT-encoding sequences of *H. annuus* were collected. Neighbour-joining phylogenetic trees based on the multiple alignment of *Gypsy* and *Copia* RTs showed that SURE elements belong to the *Gypsy* superfamily (*Metaviridae*), *Metavirus* genus, and *Ogre*/*Tat* lineage and that *Helicopia* elements are members of the *Copia* superfamily (*Pseudoviridae*), *Sirevirus* genus, and *Maximus*/*SIRE* lineage (Fig. 2).

Whenever possible, sequences downstream of the *Helicopia* or *SURE* forward primers were also collected (Fig. 1). Based on this analysis, 23 *SURE* and 18 *Helicopia* sequences containing putative LTRs were retained. All sequences are listed in Supplementary Table 3 and were deposited in the Sequence Repository website of the Department of Agriculture, Food, and Environment of the University of Pisa. The occurrence of isolated sequences in the genome of another sunflower genotype, XRQ (Badouin et al., 2017), was verified using BLASTN analysis; all sequences were identified (Supplementary Table 4), and the probability ranged from 0 to 6.94 × e^{-137}.

The isolation of LTR and RT-encoding sequences was based on sequence similarity. According to the rules proposed by Wicker et al. (2007), two repeats belong to the same family if they share 80% (or more) sequence identity in at least 80% of their sequence. The isolated sequences fulfilled these conditions. Therefore, we attributed the isolated sequences as belonging to the *SURE* or *Helicopia* families.

To isolate corresponding LTR sequences and RT-encoding sequences from *Helianthus* wild species, Illumina reads of each species were aligned to the LTR and RT sequences of *H. annuus*, and consensus sequences for each species were built. By this method, at least five consensus sequences for each LTR and RT domain were produced for each RE and for each species (Fig. 1) and were used for subsequent analyses (Supplementary Table 5). The mean lengths of the isolated LTR and RT fragments were 127 and 492 nt for *Helicopia* and 454 and 312 nt for *SURE*, respectively.

3.2. Genomic abundance of SURE *and* Helicopia *in the* Helianthus *genus*

The relative abundance of *SURE* and *Helicopia* RT domains and LTRs was determined by mapping the Illumina reads of each species onto the isolated consensus sequences of the same species (Tenaillon et al., 2011; Natali et al., 2013; Barghini et al., 2015a; 2015b) using CLC. The CLC mapping algorithm maps multi-reads randomly among similar references, and multi-reads cannot be distinguished from exact duplicates. In these experiments, the number of multi-reads was less than 1% (data not shown). Hence, the random mapping of multi-reads did not significantly affect the abundance values of each element.

The percentages of mapping reads (corresponding to the genome proportion) of *Helicopia* LTRs ranged from 0.003% in *H. divaricatus* to 0.017% in *H. argophyllus*; those of *Helicopia* RT - encoding sequences ranged from 0.018% in *H. californicus* to 0.094% in *H. argophyllus*. The genome proportions of *SURE* LTRs ranged from 0.022% in *H. niveus* to 0.051% in *H. tuberosus*; *SURE* RT-encoding sequences ranged from 0.005% in *H. californicus* to 0.014% in *H. tuberosus*.

Fig. 3 shows the mean average coverage depth (i.e., the sum of the bases of the aligned parts of all the reads divided by the length of the reference sequence; see Materials and Methods) of each sequence type (*Helicopia* LTR and RT as well as *SURE* LTR and RT). Significant differences (according to Tukey's test) in average coverage among species regarding the *SURE* LTR and RT (see *H. niveus* vs. *H. tuberosus*) and *Helicopia* RT domain (see *H. smithii* vs. *H. tuberosus*) were recorded.

Fig. 4 shows the average coverage depth distribution of the LTRs and RT domains of *SURE* and *Helicopia* in the *Helianthus* species. Given that two LTRs occur in an RE, the average coverage of LTRs should be twice that of the corresponding coding portion (Cavallini et al., 2010). If the average coverage of LTRs is more than two fold, inter-LTR homologous recombination events may have occurred, resulting in the production of solo LTRs. Also, the occurrence of internal deletions in the retrotransposons could determine a higher number of LTRs than expected. However, inter-LTR homologous recombination is a process well known to commonly occur during genome evolution, and retrotransposon families that contain a high proportion of solo LTRs have been described in many plant species (Vicient et al., 1999).

Putative *Helicopia* LTRs were generally as abundant in the genome of each species as were *Helicopia* RT domains (Fig. 4B, C) (i.e., *Helicopia* LTRs seem under-represented in the genome of most species). It could be hypothesised that LTRs have experienced higher mutation rates than have RT-encoding domains and selected LTRs may not represent all LTRs of *Helicopia* elements.

In contrast, regarding the *Gypsy SURE* family, the average coverage median of the LTR region in each species was 2–3-fold greater than that of the RT domain (Fig. 4A, C). If LTRs accumulate more mutations than do RTs (as hypothesised for *Helicopia*), *SURE* elements might have relatively more solo LTRs than do *Helicopia* elements.

3.3. Temporal dynamics of SURE *and* Helicopia *families*

The timing of *SURE* and *Helicopia* family proliferation was inferred by analysing pairwise distances (Kimura, 1980) between paralogous RT-encoding sequences that belong to the same monophyletic groups of *Helicopia* and *SURE* elements, in accordance with the method of Piegu et al. (2006). The numbers of sequences of each species used for calculating pairwise distances are listed in Supplementary Table 6. Distances were translated into insertion dates in accordance with the methods of SanMiguel et al. (1996) and Piegu et al. (2006) but using a mutation rate of 2×10^{-8} (i.e., specific to sunflower and twice the rate calculated for synonymous substitutions in sunflower gene sequences). This mutation rate was used to keep into consideration that REs accumulate more mutations over time than do genes and to be consistent with previous analyses (Ungerer et al., 2009; Buti et al., 2011). In fact, at each insertion, the new retrotransposon copy is identical to its parental element, with the exception of mutations occurring during retrotranscription (which is error prone; Kumar and Bennetzen, 1999); additional mutations can then accumulate as time progresses.

This analysis enabled the identification of different retrotranspositional waves, mostly overlapping in terms of time between *SURE* and *Helicopia* and among species (Fig. 5). In 10 out of 12 species, *Helicopia* REs (Fig. 5A, B, C) were seemingly younger than were *SURE*s (mean insertion times of 3.31 and 4.93 MYA, respectively; Fig. 5D, E, F, G). The translation of genetic distances into insertion dates is subject to reservation; however, in our analyses, we only compared retrotransposition waves of the same RE family in different species.

Regarding *Helicopia*, proliferation seems to have begun earlier in the *Divaricati* species than in species belonging to the *Helianthus* section. All species of the *Helianthus* section showed a transpositional peak corresponding to approximately 2 MYA, and the mean insertion ages ranged from 2.2 to 3.5 MYA (Fig. 5C). The mean insertion time in *Divaricati* species was generally higher than that in species of the *Helianthus* section. *Helianthus tuberosus* showed two proliferation waves, one at approximately 4–5 MYA and the other at approximately 2 MYA (Fig. 5A), which were concurrent with the proliferation bursts observed in the other *Divaricati* species (Fig. 5B) and in the *Helianthus* section (Fig. 5C), respectively.

SURE retrotransposon dynamics were more complex, as four different proliferation profiles were observed among the analysed species (Fig. 5D, E, F, G). Retrotransposition waves largely overlapped in species of the *Helianthus* section and peaked at approximately 5 MYA, but the mean insertion times differed (ranging from 4.4–5.5 MYA) (Fig. 5G). Large differences were observed among the *Divaricati* species: *H. tuberosus*, *H. divaricatus*, and *H. smithii* showed two proliferation peaks (Fig. 5D) – the first was very ancient (approximately 10 MYA) and the second at 2–3 MYA (i.e., much more recent). *Helianthus hirsutus* and *H. californicus* showed a transpositional burst at 7–8 MYA (Fig. 5E). Only one, relatively recent, transpositional burst was observed in *H. giganteus* and *H. laevigatus* (Fig. 5F). Altogether, these results suggest an intriguing picture of species-specific increases in the abundance of these two RE families, in some cases in relatively recent times, subsequent to *Helianthus* speciation.

3.4. SURE- *and* Helicopia-*related mixed genotypes in the* Helianthus *genus*

The analysis of genome structure of the *Helianthus* genus according to the occurrence of mixed genotypes was performed using IRAPs of *SURE* and *Helicopia* elements in 32 species of the genus *Helianthus* (Vukich et al., 2009a). Using all polymorphic loci reported by Vukich et al. (2009a), subdivision in populations was not statistically supported, probably due to the very large variability.

Therefore, only bands occurring in at least 20% of species were retained and considered for this analysis.

A schematic representation of the analysed IRAP matrices is shown in Supplementary Fig. 1. This analysis included 32 sunflower species and 47 polymorphic loci, 28 of which involved *SURE*s and 19 of which were related to *Helicopia* elements. Effective analysis of the population structure and classification of species into appropriate groups were performed using the Bayesian method with STRUCTURE software (Pritchard et al., 2000). The number of initial subpopulations (K) was defined from 1 to 35, and there were five replications per run (Supplementary Fig. 1). The maximum value of ΔK was observed at K = 2, either considering all 47 loci or only 28 *SURE*-related loci. Therefore, the analysed sunflower species may consist of two subpopulations (Fig. 6). Statistical support was too low when only *Helicopia*-related loci were used.

We considered a genotype unequivocally assigned to a subpopulation when its admixture coefficient was > 80% (Qi > 0.8) for that group (Vigouroux et al., 2008; Castillo et al., 2010). Individuals with intermediate admixture coefficients (Qi < 0.8) were considered admixed. After STRUCTURE was applied, the species were classified into three groups at K = 2: (i) 12 species belonged to the *Helianthus* section; (ii) 12 species belonged to *Divaricati* and one to the *Ciliares* section; and (iii) seven admixed species comprised four that belonged to *Divaricati* and three that belonged to the *Helianthus* section (Fig. 6).

When considering only the 28 *SURE*-related loci (at K = 2), the assignment of the species to three groups was very similar, but a few differences were observed: the first group included 13 species, 12 of the *Helianthus* and one of the *Divaricati* section; the second group comprised 11 *Divaricati* and one *Ciliares* species; and the group of admixed species included four *Helianthus* and three *Divaricati* species. In particular, all *H. praecox* subspecies were admixed (Fig. 6). The observed discrepancies in genome structure when considering only the *SURE*-related loci compared with *SURE* + *Helicopia* loci might be related to different retrotransposition activity of the two RE families during species differentiation.

4. Discussion

The goal of this work was to characterise two specific families of LTR-REs of sunflower, *SURE* and *Helicopia* (Vukich et al. 2009a), and to analyse the evolutionary pathways of these families in the *Helianthus* genus. Previous studies on the repetitive component of the genome of sunflower species have focussed on global analyses of LTR-REs; studies on the behaviour of specific LTR-RE families in this genus are lacking.

SURE (*Metaviridae* (*Gypsy* superfamily), *Metavirus* genus, *Ogre*/*TAT* lineage) and *Helicopia* (*Pseudoviridae* (*Copia* superfamily), *Sirevirus* genus, *Maximus*/*SIRE* lineage) families showed no significant differences in abundance of LTRs and RT-encoding sequences within the genome of each species.

In contrast, differences between *SURE* and *Helicopia* families were found in relation to their different tendencies to be subjected to processes that imply DNA loss. Such processes may have affected the *SURE* family more than *Helicopia* (Fig. 4). In fact, the ratios between LTRs and RT-encoding sequence abundance indicate that *SURE* solo-LTRs are more abundant than are *Helicopia* ones. This might suggest that *SURE* elements are more prone to producing solo LTRs by local non-homologous recombination than are *Helicopia* elements. Processes such as DNA rearrangement and unequal homologous recombination drive DNA removal in plants by a number of mechanisms (Kalendar et al., 2000; Ma et al., 2004; Neumann et al., 2006; Ammiraju et al., 2007; Hawkins et al., 2008; Morse et al., 2009). On the other hand, it is also possible that *SURE*s had more time to accumulate solo LTRs, as *SURE*s are older than *Helicopia* ones (Fig. 5).

Differences in the proliferation time profiles between *SURE* and *Helicopia* families were also observed. *SURE* REs were on average older than were *Helicopia* in most of the analysed species (Fig. 5). It is known that proliferation bursts do not occur simultaneously for all RE families but show different timings in different RE families (Vitte and Panaud, 2003; Ammiraju et al., 2007). In sunflower, another *Copia* LTR-RE is potentially prone to a transpositional burst, as this LTR-RE is still active (i.e., it is regularly transcribed and, at low rates, is reinserted into the genome) (Vukich et al., 2009b).

Significant differences in the abundance of the two RE families were observed among species, at least for the *SURE* family, whereas the *Helicopia* family was more uniform (Fig. 3). This indicates that the equilibrium between RE amplification and loss differs among species. Such differences may have been casually produced during the evolution of *Helianthus* species. On the other hand, with the exceptions of *H. annuus* and to a minor extent *H. petiolaris* and *H. tuberosus*, all other analysed species have been reported to be distributed in relatively small areas (Heiser et al., 1969; Rogers et al., 1982). Such areas were often different among the analysed species, indicating that differences in the abundance of repetitive DNA might be correlated to the different environments in which the species live (i.e., such differences might be involved in the adaptation of the analysed genotypes to the environment).

In addition to differences in redundancy, the *SURE* family showed clear-cut differences among *Helianthus* species regarding proliferation time and proliferation profiles (Fig. 5). Differences related to the *Helicopia* family among *Helianthus* species are less defined than those related to *SURE* elements, probably because *Helicopia* proliferation is generally more recent than is *SURE* proliferation, especially in species of the *Helianthus* section (Fig. 5).

Analyses of chloroplast DNA sequences roughly dated the origin of the *Helianthus* genus to a time period between 22.7 and 4.75 MYA, and *Helianthus* species diverged between 8.2 and 1.7 MYA (Schilling, 1997). Although dating transpositional bursts is subject to reservation, *SURE* and *Helicopia* proliferation bursts seem to be concurrent with species divergence within the genus (Fig. 5).

The analysis of polymorphisms especially related to *SURE* elements showed the existence of two subpopulations in the *Helianthus* genus, roughly corresponding to the *Helianthus* section and to the *Divaricati*/*Ciliares* sections (Fig. 6). This result confirms the separation between annuals and perennials (Schilling et al., 1998; Santini et al., 2002; Natali et al., 2006) and allowed the discovery of species that have admixed structure. *Helicopia* family had a minor role compared with *SURE* in structuring the genus into two subpopulations (Fig. 6). Also this result might be explained by more recent proliferation of *Helicopia* elements compared with *SURE*s.

The presence of species with admixed genome structure is probably related to the events of interspecific hybridisation. Multiple interspecific hybridisation events have been important in the evolution of *Helianthus* species (Rieseberg et al., 1998), although dating and the extent of such events are not precisely known (Schilling, 1997). Interspecific hybridisation can involve transpositional bursts as a result of so-called genomic shock (i.e., response to the introduction of alien genetic material into a new genetic background) (McClintock, 1984). For example, the massive amplification of REs has been reported to have occurred relatively recently in *H. anomalus*, *H. deserticola*, and *H. paradoxus*, which are three species that originated by interspecific hybridisation between *H. annuus* and *H. petiolaris* (Ungerer et al., 2009). Transpositional peaks observed in the analysed *Helianthus* species for *SURE* and *Helicopia* elements may be related, at least in certain cases, to concurrent events of interspecific hybridisation.

Interestingly, *SURE* REs are more abundant in allopolyploid species (*H. tuberosus*, *H. hirsutus*, *H. laevigatus*, and *H. californicus*; Figs. 3 and 4) (i.e., in species in which an interspecific hybridisation event is shown by the presence of a multiple chromosome numbers). Genomic shock following interspecific hybridisation and subsequent chromosome doubling may have induced *SURE* proliferation. This phenomenon seems not to be true for *Helicopia*, indicating that each LTR-RE family responds specifically to genomic shock.

In general, even after analysing only two LTR-RE families, our data reveal significant differences in the evolutionary trends between these families. These differences point out the necessity, when studying retrotransposons and genome evolution, of analysing (in addition to the general characterisation of the repetitive component of the genome) genome structure separately for specific RE families. The availability of the complete genome sequence of *H. annuus* (Badouin et al., 2017) will allow the comprehensive analysis of every RE family and will establish whether the behaviour of *Helicopia* and *SURE* families in the evolution of the *Helianthus* genome and the related interspecific variability are specific to these two elements.

Funding

This research work was funded by DiSAAA-a, University of Pisa, Project "Plantomics".

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://pgagl.agr.unipi.it/sequence-repository/

References

Ammiraju, J.S., Zuccolo, A., Yu, Y., Song, X., Piegu, P., Chevalier, F., Walling, J.G., Ma, J., Talag, J., Brar, D.S., SanMiguel, P.J., Jiang, N., Jackson, S.A., Panaud, O., Wing, R.A., 2007. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J. 52, 342–351.

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.

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

Barghini, E., Natali, L., Giordani, T., Cossu, R.M., Scalabrin, S., Cattonaro, F., Šimková, H., Vrána, J., Doležel, J., Morgante, M., Cavallini, A., 2015b. LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. DNA Research 22, 91–100.

Boeke, J.D., Corces, V.G., 1989. Transcription and reverse transcription of retrotransposons. Ann. Rev. Microbiol. 43, 403–434.

Boeke, J.D., Eickbush, T.H., Sandmeyer, S.B., Voytas, D.F., 2006. Index of viruses - Pseudoviridae. ICTVdB - The Universal Virus Database, version 4, Columbia University, New York, USA.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bousios, A., Darzentas, N., Tsaftaris, A., Pearce, S.R., 2010. Highly conserved motifs in non-coding regions of *Sirevirus* retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11, 89.

Bousios, A., Darzentas, N., 2013. *Sirevirus* LTR retrotransposons: phylogenetic misconceptions in the plant world. Mobile DNA 4, 9.

Buti, M., Giordani, T., Cattonaro, F., Cossu, R.M., Pistelli, L., Vukich, M., Morgante, M., Cavallini, A., Natali, L., 2011. Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. Theor. Appl. Genet. 123, 779–791.

Castillo, A., Dorado, G., Feuillet, C., Sourdille, P., Hernandez, P., 2010. Genetic structure and ecogeographical adaptation in wild barley (*Hordeum chilense* Roemer et Schultes) as revealed by microsatellite markers. BMC Plant Biol. 10, 266.

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

Earl, D.A., Vonholdt, B.M., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet. Res. 4, 359-361.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611-2620.

Fauquet, C.M., Mayo, M.A., 2001. The 7th ICTV Report. Arch. Virol. 146, 189-194.

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

Gorinsek, B., Gubensek, F., Kordis, D., 2004. Evolutionary genomics of chromoviruses in eukaryotes. Mol. Biol. Evol. 21, 781–798.

Hawkins, J.S., Hu, G., Rapp, R.A., Grafenberg, J.L., Wendel, J.F. 2008. Phylogenetic determination of the pace of transposable element proliferation in plants: *Copia* and LINE-like elements in *Gossypium*. Genome 51, 11–18.

Heiser, C.B., Smith, D.M., Clevenger, S.B., Martin, W.C., 1969. North American sunflowers (*Helianthus*). Torrey Bot. Club Mem. 22, 1-218.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44, 223–270.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Kalendar, R., Grob, T., Regina, M., Suoniemi, A., Schulman, A.H., 1999. IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor. Appl. Genet. 98, 704–711.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA 97, 6603–6607.

Kalendar, R., Schulman, A.H., 2006. IRAP and REMAP for retrotransposon based genotyping and fingerprinting. Nature Protocols 1, 2478–2484.

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., Schulman, A.H., 2008. Cassandra retrotransposons carry independently transcribed 5S RNA. Proc. Natl. Acad. Sci. USA 105, 5833–5838.

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.

Ku, H.M., Vision, T., Liu, J.P., Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA 97, 9121–9126.

Kumar, A., Bennetzen, J.L., 1999. Plant retrotransposons. Annu. Rev. Genet. 33, 479-532.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., Latorre, A., Moya, A., 2011. The *Gypsy* database (GyDB) of mobile genetic elements: release 2.0. Nucl. Acids Res. 39, D70–D74.

Lynch, M., Milligan, B.G., 1994. Analysis of population genetic structure with RAPD markers. Mol. Ecol. 3, 91-99.

Ma, J., Devos, K.M., Bennetzen, J.L., 2004. Analyses of LTR retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14, 860–869.

Mascagni, F., Barghini, E., Giordani, T., Rieseberg, L.H., 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.

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

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., Lopez, R., 2013. Analysis tool web services from the EMBL-EBI. Nucl. Acids Res. 41, W597-W600.

Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., Kubisiak, T.L., Amerson, H.V., Carlson, J.E., Nelson, C.D., Davis, J.M., 2009. Evolution of genome size and complexity in *Pinus*. PLoS One 4, e4332.

Natali, L., Santini, S., Giordani, T., Minelli, S., Maestrini, P., Cionini, P.G., Cavallini, A., 2006. Distribution of Ty3-*gypsy*- and Ty1-*copia*-like DNA sequences in the genus *Helianthus* and other *Asteraceae*. Genome 49, 64-72.

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg L., Cavallini, A., 2013. The repetitive component of the sunflower genome as revealed by different procedures for assembling next generation sequencing reads. BMC Genomics 14, 686.

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

Neumann, P., Koblizkova, A., Navratilova, A., Macas, J., 2006. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173, 1047–1056.

Novák, P., Neumann, P., Macas, J., 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11, 378.

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.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., Panaud, O., 2006. Doubling genome size without polyploidization: dynamics of retrotransposition driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 16, 1262–1269.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rieseberg, L.H., Van Fossen, C., Desrochers, A., 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature 375, 313–316.

Rieseberg, L.H., Baird, S.J.E., Desrochers, A.M., 1998. Patterns of mating in wild sunflower hybrid zones. Evolution 52, 713-726.

Rogers, C.E., Thompson, T.E., Seiler, G.J., 1982. Sunflower species of the United States. National Sunflower Association, Bismarck, North Dakota.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. The genetic structure of human populations. Science 298, 2381-2385.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

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

Schilling, E.E., 1997. Phylogenetic analysis of *Helianthus* (*Asteraceae*) based on chloroplast DNA restriction site data. Theor. Appl. Genet. 94, 925-933.

Schilling, E.E., Linder, C.R., Noyes, R.D., Rieseberg, L.H., 1998. Phylogenetic relationships in *Helianthus* (*Asteraceae*) based on nuclear ribosomal DNA internal transcribed spacer region sequence data. Syst. Bot. 23, 177-187.

Shannon, C.E., Weaver, W., 1949. The mathematical theory of communication. University of Illinois Press, Urbana, IL.

Smit, A.F.A., Hubley, R., Green, P., 1996. RepeatMasker. Open-3.0. Available at http://www.repeatmasker.org/et al.

Staton, S.E., Bakken, B.E., Blackman, B.K., Chapman, M.A., Kane, N.C., Tang, S., Ungerer, M.C., Knapp, S.J., Rieseberg, L.H., Burke, J.M., 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72, 142–153.

Tenaillon, M.I., Hufford, M.B., Gaut, B.S., Ross-Ibarra, J., 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol. Evol. 3, 219–229.

Timme, R.E., Simpson, B.B., Randal Linder, C., 2007. High-resolution phylogeny for *Helianthus* (*Asteraceae*) using the 18S-26S ribosomal DNA external transcribed spacer. Am. J. Bot. 94, 1837–1852.

Ungerer, M.C., Strakosh, S.C., Zhen, Y., 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. 16, R872–R873.

Ungerer, M.C., Strakosh, S.C., Stimpson, K.M., 2009. Proliferation of Ty3/*Gypsy*-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC Biol. 7, 40.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., Schulman, A.H., 1999. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. The Plant Cell 11, 1769–1784.

Vigouroux, Y., Glaubitz, J.C., Matsuoka, Y., Major, M., Doebley, J., 2008. Population structure and genetic diversity of the new world maize races assessed by microsatellites. Am. J. Bot. 95, 1240–1253.

Vitte, C., Panaud, O., 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice (*Oryza sativa* L.). Mol. Biol. Evol. 20, 528-540.

Vukich, M., Schulman, A.H., Giordani, T., Natali, L., Kalendar, R., Cavallini, A., 2009a. Genetic variability in sunflower (*Helianthus annuus* L.) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. Theor. Appl. Genet. 119, 1027–1038.

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

Wicker, T., Keller, B., 2007. Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res. 17, 1072–1081.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification systemfor eukaryotic transposable elements. Nature Rev. Genet. 8, 973–982.

Wright, D.A., Voytas, D.F., 2002. *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res. 12, 122–131.

Figure Legends

Fig. 1. Pipeline for the isolation of LTRs and RT-encoding sequences of *SURE* and *Helicopia* REs of 12 species of *Helianthus.*

Fig. 2. Neighbour-joining phylogenetic tree based on multiple alignment of (A) the eight RT amino acid sequences of *SURE*- and *Gypsy*-related RT amino acid sequences of several plant species and (B) the five RT amino acid sequences of *Helicopia*- and *Copia*-related RT amino acid sequences of several plant species. Bootstrap values greater than 0.6 are shown with an asterisk.

Fig. 3. Average coverage of LTRs and RT-encoding sequences of *SURE* (left) and *Helicopia* (right) RE families of 12 species of *Helianthus* (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species). The mean ± standard error is reported for each species. Significant differences for each separate group of measurements are indicated by different letters ($p < 0.05$) according to Tukey's test.

Fig. 4. (A, B) Box and whisker plots of the average coverage of consensus LTRs and RT-encoding sequences of *Helicopia* and *SURE* REs of 12 species of *Helianthus* (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species). Boxes represent 25–75% of the values, whiskers represent the whole range of values, and lines in the box represent the median values of the distribution. (C) The ratio between the median of the average coverage of consensus LTRs and the median average coverage of the consensus RT-encoding sequences of *Helicopia* and *SURE* REs in 12 sunflower species.

Fig. 5. Timing of *Helicopia* and *SURE* retrotranspositional activity in 12 species of *Helianthus* based on the pairwise comparisons of Illumina reads that match RT-encoding sequences. Graphs A, B, and C refer to *Helicopia*; graphs D, E, F, and G, *SURE.* To facilitate comparisons, each graph combines species with similar profiles. Species of the *Divaricati* section (perennials) are shown in graphs A, B, D, E, and F; species of the *Helianthus* section (annuals), graphs C and G. The y axis reports the product of the percentage of pairwise comparisons for the average coverage of the RT sequence in each species in order to account for the extent of transpositional bursts. The mean insertion times for each species and for the analysed RE families are reported in parentheses.

Fig. 6. Proportions of the ancestry of 32 *Helianthus* species and subspecies (H: section *Helianthus*, annual species; D: section *Divaricati*, perennial species; C: section *Ciliares*, perennial species) based on K = 2 (where K is the number of initial subpopulations). The proportions were obtained using STRUCTURE software for IRAP matrices obtained from the electrophoresis of PCR products amplified using *SURE* and *Helicopia* primers (above) or *SURE* primers only (below).

Figure 1

Click here to download high resolution image

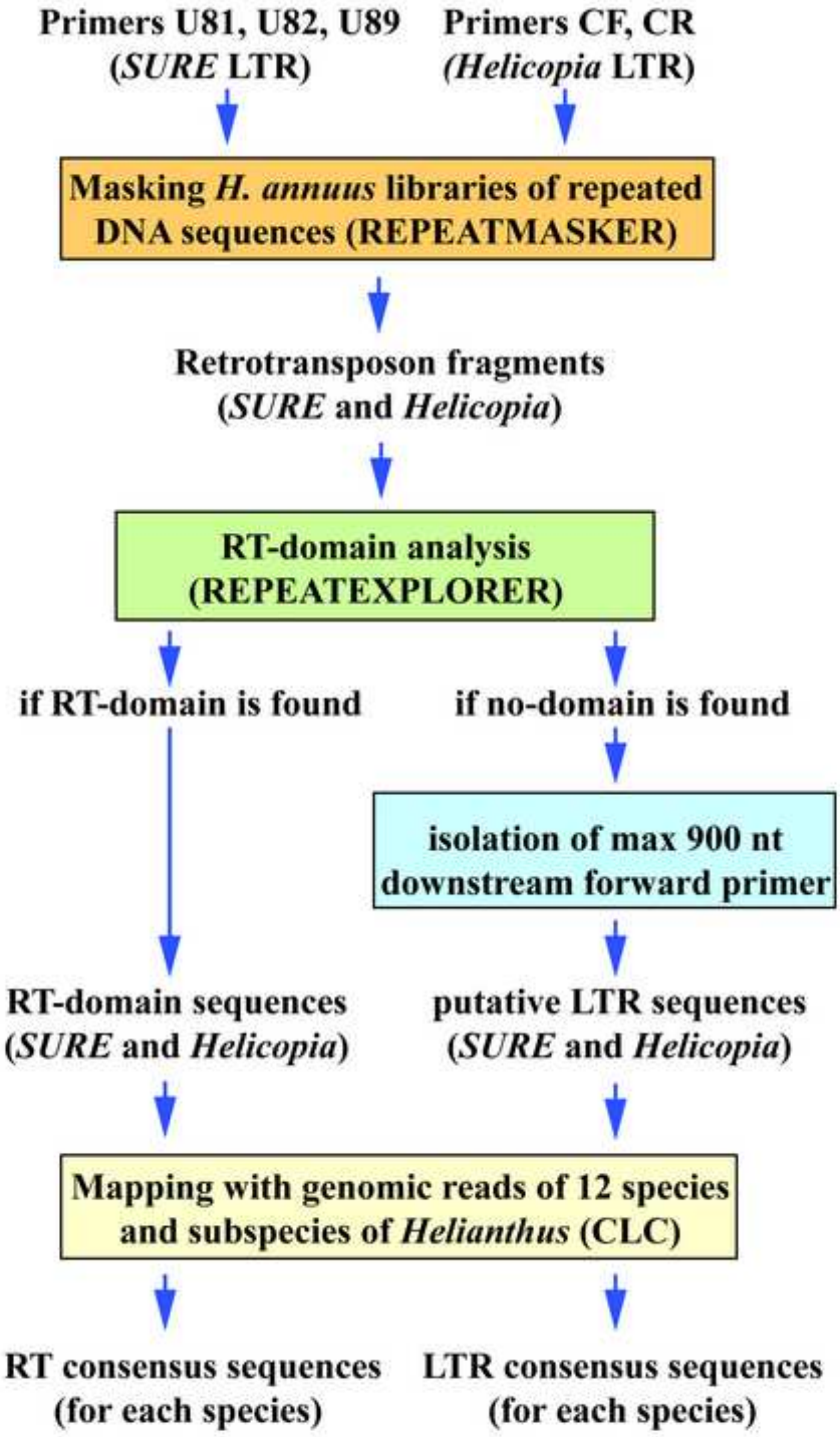

Primers U81, U82, U89 (*SURE* LTR)
Primers CF, CR (*Helicopia* LTR)
Masking *H. annuus* libraries of repeated DNA sequences (REPEATMASKER)
Retrotransposon fragments (*SURE* and *Helicopia*)
RT-domain analysis (REPEATEXPLORER)
if RT-domain is found
if no-domain is found
isolation of max 900 nt downstream forward primer
RT-domain sequences (*SURE* and *Helicopia*)
putative LTR sequences (*SURE* and *Helicopia*)
Mapping with genomic reads of 12 species and subspecies of *Helianthus* (CLC)
RT consensus sequences (for each species)
LTR consensus sequences (for each species)

Figure 2

Click here to download high resolution image

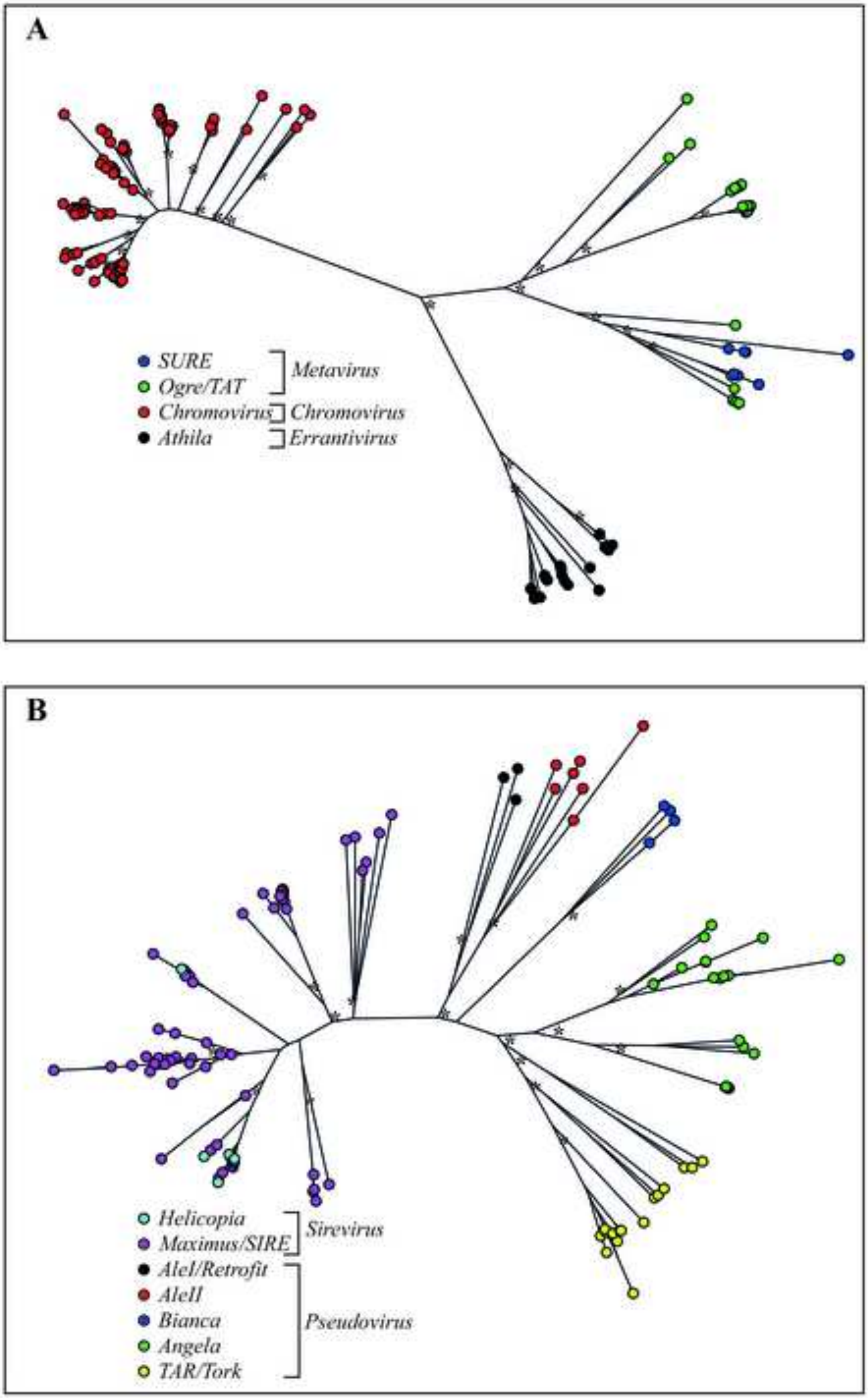


Figure 3

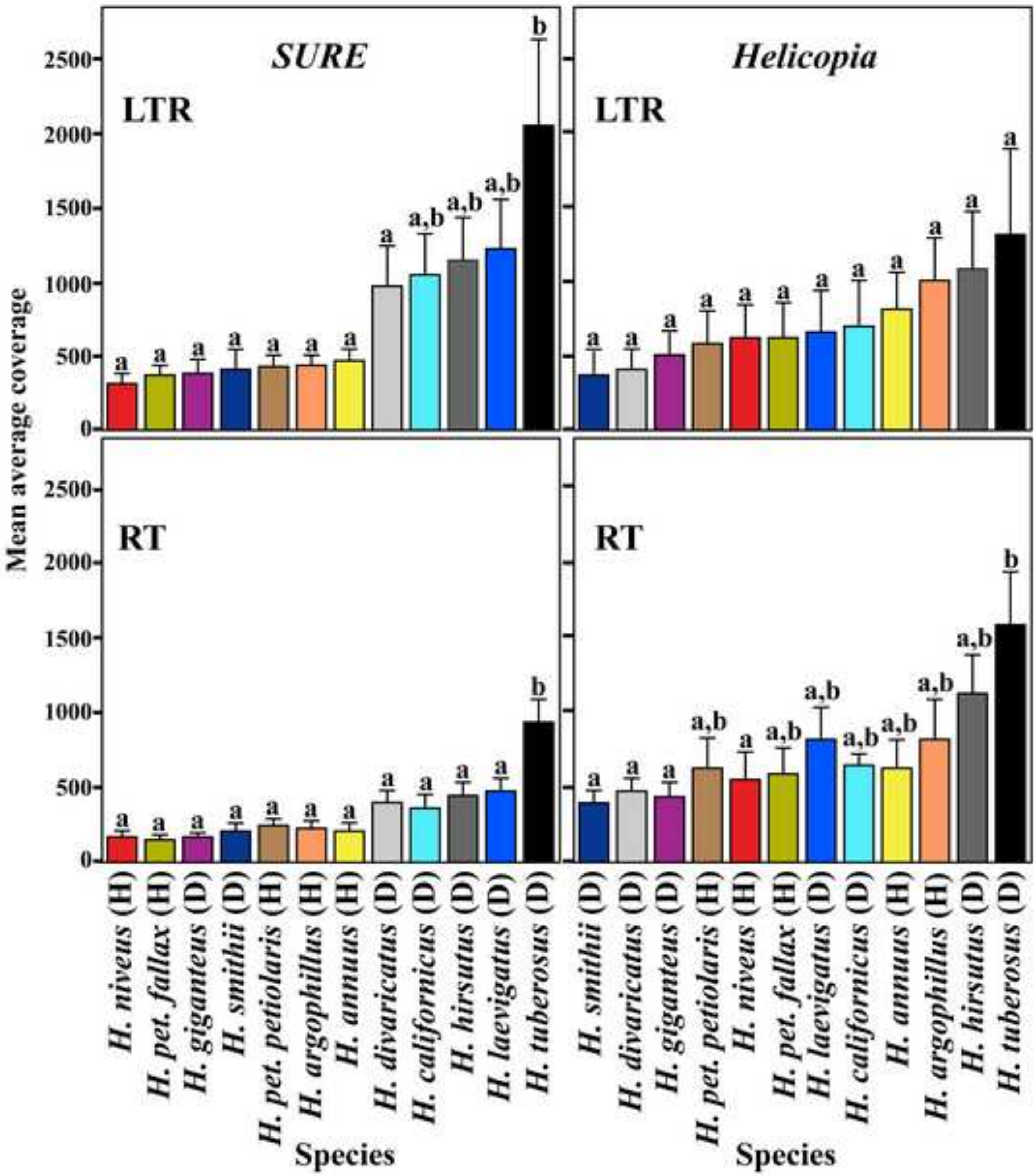

SURE
Helicopia
LTR
LTR
RT
RT
Mean average coverage
0
500
1000
1500
2000
2500
a
a
a
a
a
a
a
a
a,b
a,b
a,b
b
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
a
b
a
a
a
a,b
a
a,b
a,b
a,b
a,b
a,b
a,b
b
H. niveus (H)
H. pet. fallax (H)
H. giganteus (D)
H. smithii (D)
H. pet. petiolaris (H)
H. argophillus (H)
H. annuus (H)
H. divaricatus (D)
H. californicus (D)
H. hirsutus (D)
H. laevigatus (D)
H. tuberosus (D)
H. smithii (D)
H. divaricatus (D)
H. giganteus (D)
H. pet. petiolaris (H)
H. niveus (H)
H. pet. fallax (H)
H. laevigatus (D)
H. californicus (D)
H. annuus (H)
H. argophillus (H)
H. hirsutus (D)
H. tuberosus (D)
Species
Species

Figure 4

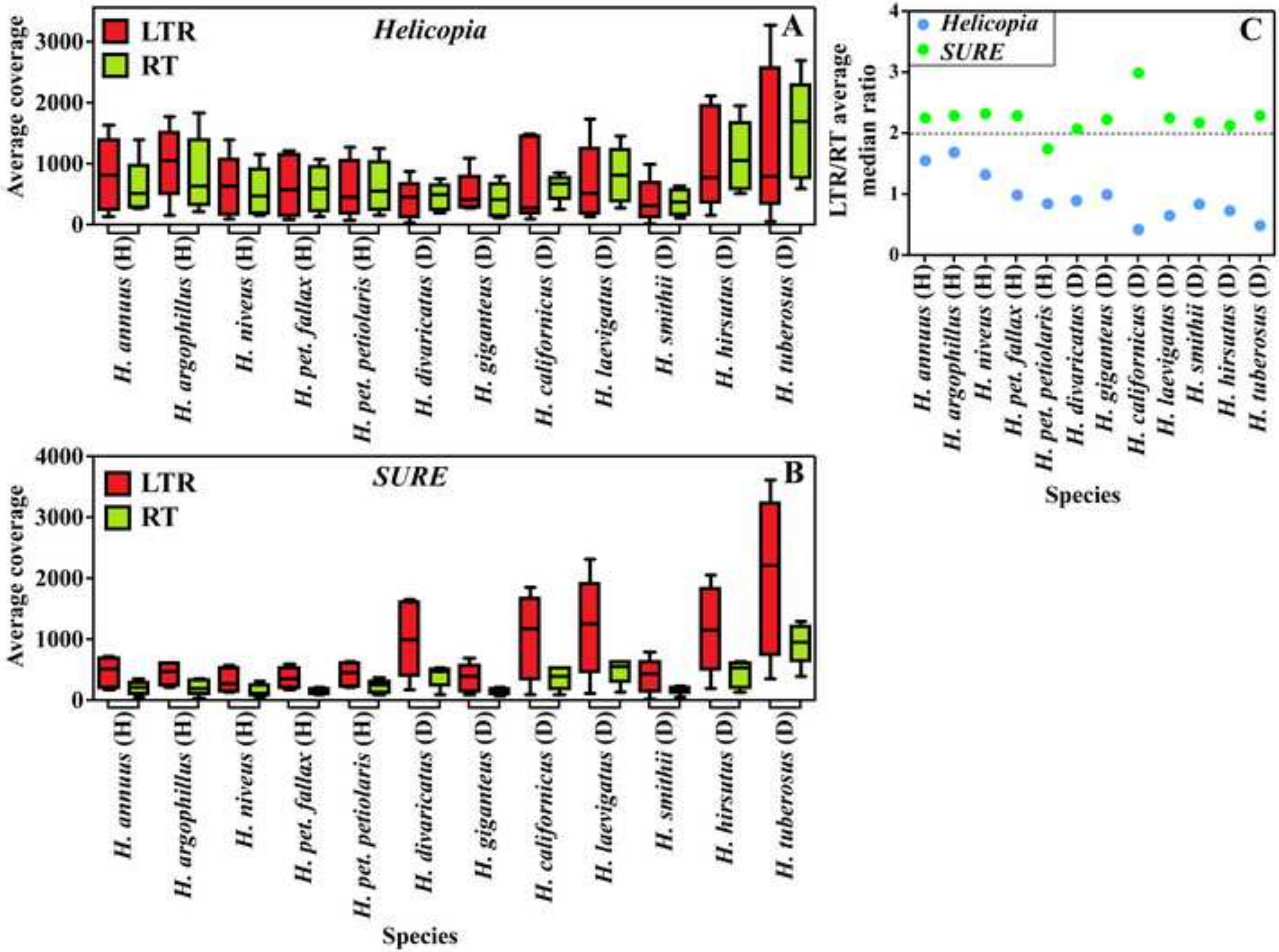

Helicopia
A
LTR
RT
Average coverage
3000
2000
1000
0
H. annuus (H)
H. argophillus (H)
H. niveus (H)
H. pet. fallax (H)
H. pet. petiolaris (H)
H. divaricatus (D)
H. giganteus (D)
H. californicus (D)
H. laevigatus (D)
H. smithii (D)
H. hirsutus (D)
H. tuberosus (D)
SURE
B
LTR
RT
Average coverage
4000
3000
2000
1000
0
H. annuus (H)
H. argophillus (H)
H. niveus (H)
H. pet. fallax (H)
H. pet. petiolaris (H)
H. divaricatus (D)
H. giganteus (D)
H. californicus (D)
H. laevigatus (D)
H. smithii (D)
H. hirsutus (D)
H. tuberosus (D)
Species
C
Helicopia
SURE
LTR/RT average median ratio
4
3
2
1
0
H. annuus (H)
H. argophillus (H)
H. niveus (H)
H. pet. fallax (H)
H. pet. petiolaris (H)
H. divaricatus (D)
H. giganteus (D)
H. californicus (D)
H. laevigatus (D)
H. smithii (D)
H. hirsutus (D)
H. tuberosus (D)
Species

Figure 5

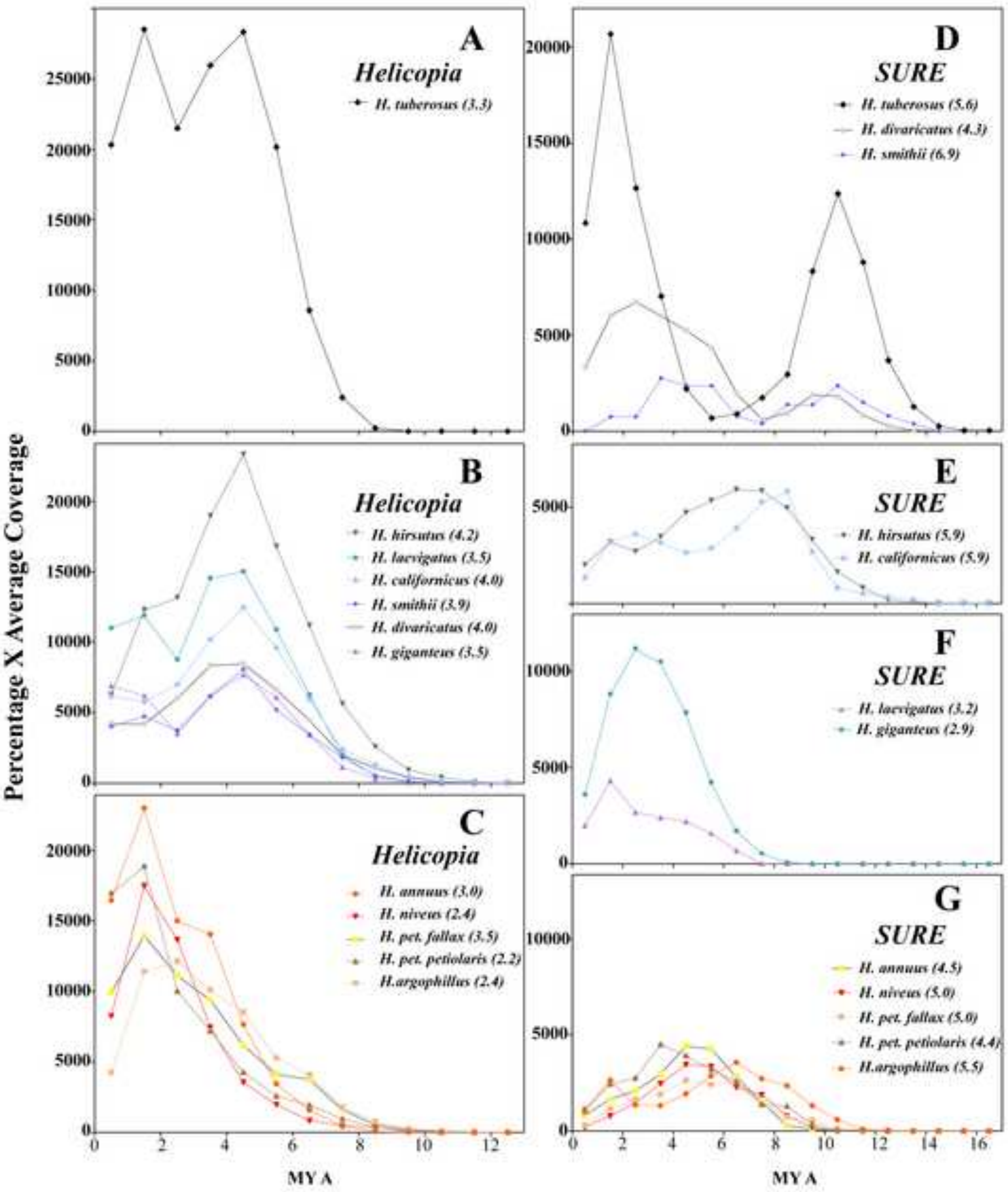
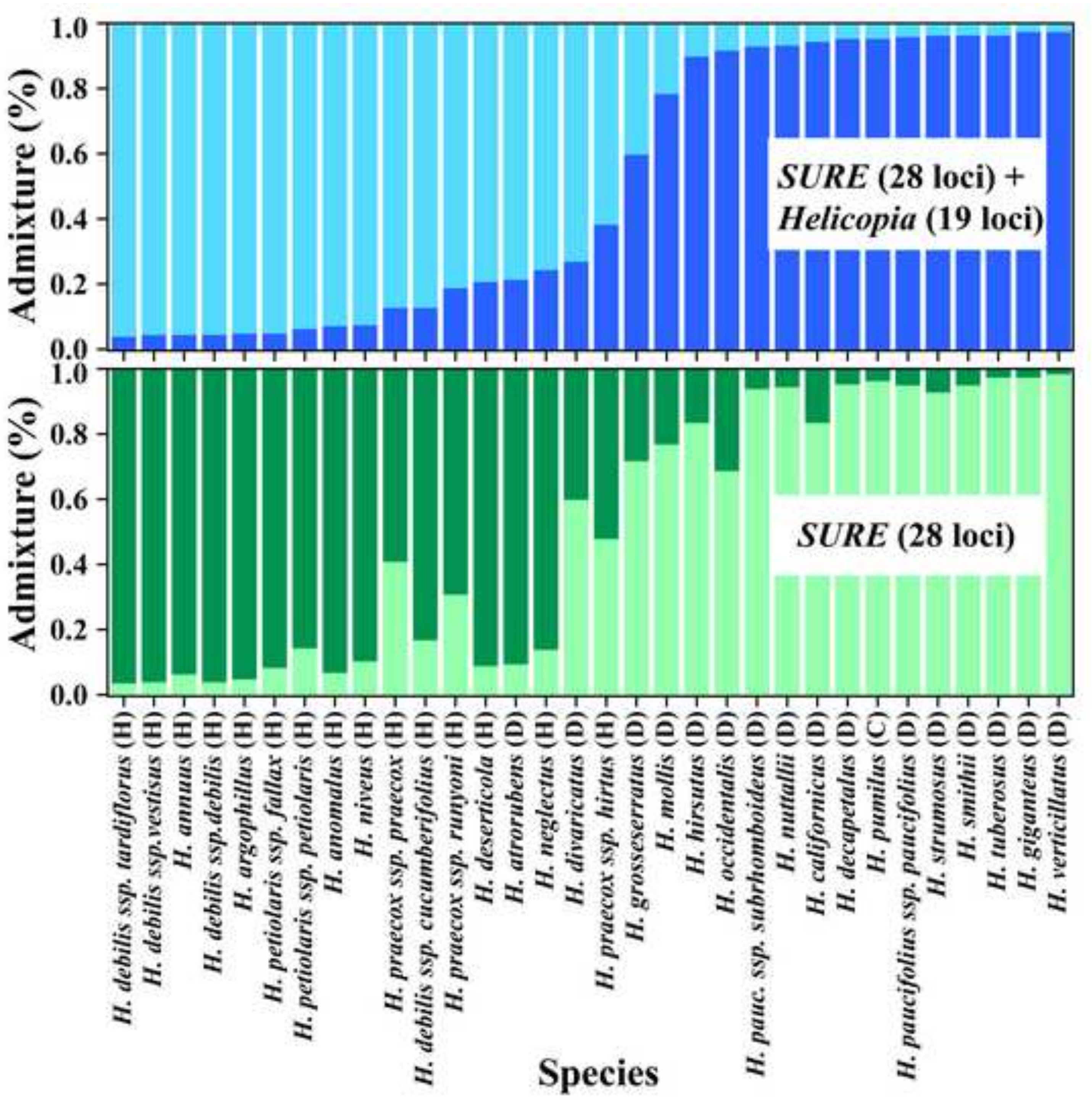

A
Helicopia
H. tuberosus (3.3)
25000
20000
15000
10000
5000
0
D
SURE
H. tuberosus (5.6)
H. divaricatus (4.3)
H. smithii (6.9)
20000
15000
10000
5000
0
B
Helicopia
H. hirsutus (4.2)
H. laevigatus (3.5)
H. californicus (4.0)
H. smithii (3.9)
H. divaricatus (4.0)
H. giganteus (3.5)
20000
15000
10000
5000
0
E
SURE
H. hirsutus (5.9)
H. californicus (5.9)
5000
F
SURE
H. laevigatus (3.2)
H. giganteus (2.9)
10000
5000
0
Percentage X Average Coverage
C
Helicopia
H. annuus (3.0)
H. niveus (2.4)
H. pet. fallax (3.5)
H. pet. petiolaris (2.2)
H.argophillus (2.4)
20000
15000
10000
5000
0
0
2
4
6
8
10
12
MY A
G
SURE
H. annuus (4.5)
H. niveus (5.0)
H. pet. fallax (5.0)
H. pet. petiolaris (4.4)
H.argophillus (5.5)
10000
5000
0
0
2
4
6
8
10
12
14
16
MY A

Figure 6

Click here to download high resolution image



Supplementary Fig. 1

Supplementary Table 1

Click here to download Supplementary material for on-line publication only: Mascagni_Suppl Table 1_R1.doc

Supplementary Table 2

Click here to download Supplementary material for on-line publication only: Mascagni_Suppl Table 2_R1.doc

Supplementary Table 3

Supplementary Table 4

Supplementary Table 5

Click here to download Supplementary material for on-line publication only: Mascagni_Suppl Table 5_R1.xlsx