ASTER-REP, a Database of Asteraceae Sequences for Structural and Functional Studies of Transposable Elements

Authors:
Maria Ventimiglia[1], Emanuele Bosi[2], Luca Vasarelli[3], Andrea Cavallini[1], Flavia Mascagni[1*]

1) Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, 56124 Pisa (Italy)

2) Department of Earth, Environmental and Life Sciences (DISTAV), University of Genoa, Genova, Corso Europa, 26, 16132 Genova (Italy)

3) CNR, Istituto di informatica e telematica, Via Giuseppe Moruzzi, 1, 56124 Pisa (Italy)

Corresponding author:

Flavia Mascagni

Email: flavia.mascagni@unipi.it

Keywords:
Database, Transposable elements annotation, Asteraceae

Main text:

Transposable elements (TEs) are interspersed repetitive DNA sequences that can move independently within the genome through specific transposition mechanisms. In eukaryotes, TEs are divided into two classes: Class I TEs (retrotransposons, REs), which use an RNA intermediate for the transposition, and Class II TEs (DNA transposons), which move through DNA excision (Wicker et al. 2007). These classes are divided into orders and lineages, according to sequence homology and to the ability to encode their own transposition machinery or not.

Traditionally, TEs have been poorly studied also because their identification has been challenging, nevertheless these sequences have tremendous implications for the genome stability and function: first, TEs drive structural variation, as their transposition can cause deletions, inversions (*i.e.*, homologous recombination of TEs with an opposite orientation) and duplications (Cordaux and Batzer 2009), making TEs activity strongly associated with dynamics of genome reduction/expansion. Secondly, TEs may drastically affect the expression of other genes (Lisch 2013, Fambrini et al. 2018), either by disrupting the coding sequence or by altering the gene regulation when transposing into the upstream region with different effects, such as the disruption of regulatory sites, epigenetic silencing of neighboring regions, or providing new regulatory elements (Morgante et al. 2007, Slotkin and Martienssen 2007). Finally, TEs can originate novel genes through exaptation mediating neofunctionalization with a selective advantage to the host (Ventimiglia et al. 2022).

For these reasons, the detection of such elements is now considered an essential step to achieve high-quality genomes, more so for eukaryotes since TEs represent a considerable fraction of their genomes. However, the accurate identification of TEs remains a challenging and largely manual endeavor due to the diversity in structures and sequences, which greatly vary across species. There are three strategies for repetitive sequences annotation: homology-based, *ab initio*, and structural-based prediction (Bergman and Quesneville 2007, Saha et al. 2008). Of these, the homology-based is the most commonly adopted approach, although it requires libraries of well-curated, representative repeats from related organisms which might be lacking.

Here we present ASTER-REP, a comprehensive TEs database of full-length sequences belonging to species of the Asteraceae family which represents a novel resource to facilitate the homology-based annotation of TEs. We focused on the family Asteraceae for different reasons: it is an enormous clade, comprising 8% of all plant species distributed across the globe, including economically important species, such as sunflower artichoke and lettuce. On average, the genomes of Asteraceae display a high proportion of TEs (70%) representing an excellent system to study plant genome evolution (Staton and Burke 2015). However, only the sunflower, *Helianthus annuus,* has been deeply investigated focusing the TE content (Natali et al. 2013, Mascagni et al. 2015, 2020) revealing to have a large proportion of repeated sequences accounting for more of the 80% of the genome itself.

The principal aim of this database is to collect and carefully annotate full-length TEs producing a useful tool for starting studies on transposon diversity and dynamics in this important plant family. Furthermore, given its size, detailed annotation, and user-friendly implementation, ASTER-REP will be a useful platform to study transposon variability, helping to unravel the genome structure and improving transposon annotation for other plant genomes.

The ASTER-REP database (accessible at https://aster-rep.agr.unipi.it) contains original data of interspersed repeat sequences, organized with hierarchical criteria (Supplementary Figure S1). *De novo* interspersed repeat libraries were developed identifying TEs using structural-based methods on sequence assemblies of six Asteraceae species. In particular, *Helianthus annuus*, *Lactuca sativa*, and *Cynara cardunculus* var. *scolymus*, assembled at chromosome level, and *Artemisia annua*, *Carthamus tinctorius,* and *Chrysanthemum seticuspe*, assembled at contig level, were chosen (see also Supplementary Table S1).

All TEs were retrieved using structure-based software, by searching for their characteristic features, *i.e.*, high copy number, specific domains, and typical sequence motifs (see Supplementary Data). The approach used to identify and collect Asteraceae TEs has been consistently applied in all six species, harmonizing the resulting datasets. One feature that differentiates ASTER-REP from other existing TE databases is that it is composed of full-length elements.

Additional efforts were put into cross-validating all libraries belonging to different TE orders to reduce the misclassification of individual entries or the presence of any other sequence differing from the one reported in the entry (e.g., captured gene fragments or nested TEs). A total of 328,696 unique full-length TEs are presently included in the database (Supplementary Table S2).

The interface of ASTER-REP web-site allows users to search, browse, display, and retrieve sequences (Figure 1). Users can obtain the desired data with automatic search options by selecting checkboxes for species, whereas TE class, order, superfamily, and lineage are selectable from drop-down menus.

The hierarchical structure of sequence classification was exploited to allow a straightforward selection of the desired repeat pool with a single query (*i.e.*, selecting Class I, only SINE and LTR REs will be selectable). Selected entries can be then visualized and downloaded locally.

Finally, we consider the sequence similarity-based access to TE data among the most useful aspects of ASTER-REP, for which we integrated a BLAST service linked to the database sequences. Users can exploit this tool for analyzing their own sequences and/or for the study of transposons in species not belonging to the Asteraceae family. Actually, beacuse of the general conservation of sequence characteristics among plant TEs, the data collected in ASTER-REP can be used as a template to produce new and more customized datasets.

Considering its size and annotation level, ASTER-REP collection represents an important resource for repeat annotation of plant genomes and a step forward compared to unspecific sequence databases. Furthermore, the user-friendly database we implemented will facilitate the exploitation of these sequences, also providing a stable platform to be periodically updated through the integration of newly identified repeat elements from other Asteraceae genomes.

Data availability

The data underlying this article are available in ASTER-REP database at https://aster-rep.agr.unipi.it.

Funding

This work was supported by University of Pisa, Project Plantomics [grant number 569999_2017].

Disclosures

The authors declare no conflicts of interest.

References

Bergman, C.M. and Quesneville, H. (2007) Discovering and detecting transposable elements in genome sequences. Briefings in bioinformatics. 8: 382-392.

Cordaux, R. and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews genetics. 10: 691-703.

Fambrini, M., Bellanca, M., Costa Munoz, M., Usai, G., Cavallini, A. and Pugliesi, C. (2018) Ligulate inflorescence of *Helianthus* x *multiflorus*, cv. Soleil d'Or, correlates with a mis-regulation of a *CYCLOIDEA* gene characterised by insertion of a transposable element. Plant biology. 20: 956–967.

Lisch, D. (2013) How important are transposons for plant evolution? Nature Reviews Genetics. 14: 49-61.

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

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

Morgante M., De Paoli E. and Radovic S. (2007) Transposable elements and the plant pan-genomes. Current opinion in plant biology. 10: 149-155.

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

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

Saha, S., Bridges, S., Magbanua, Z.V. and Peterson, D.G. (2008) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. Tropical plant biology. 1: 85-96.

Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. Nature reviews genetics. 8: 272-285.

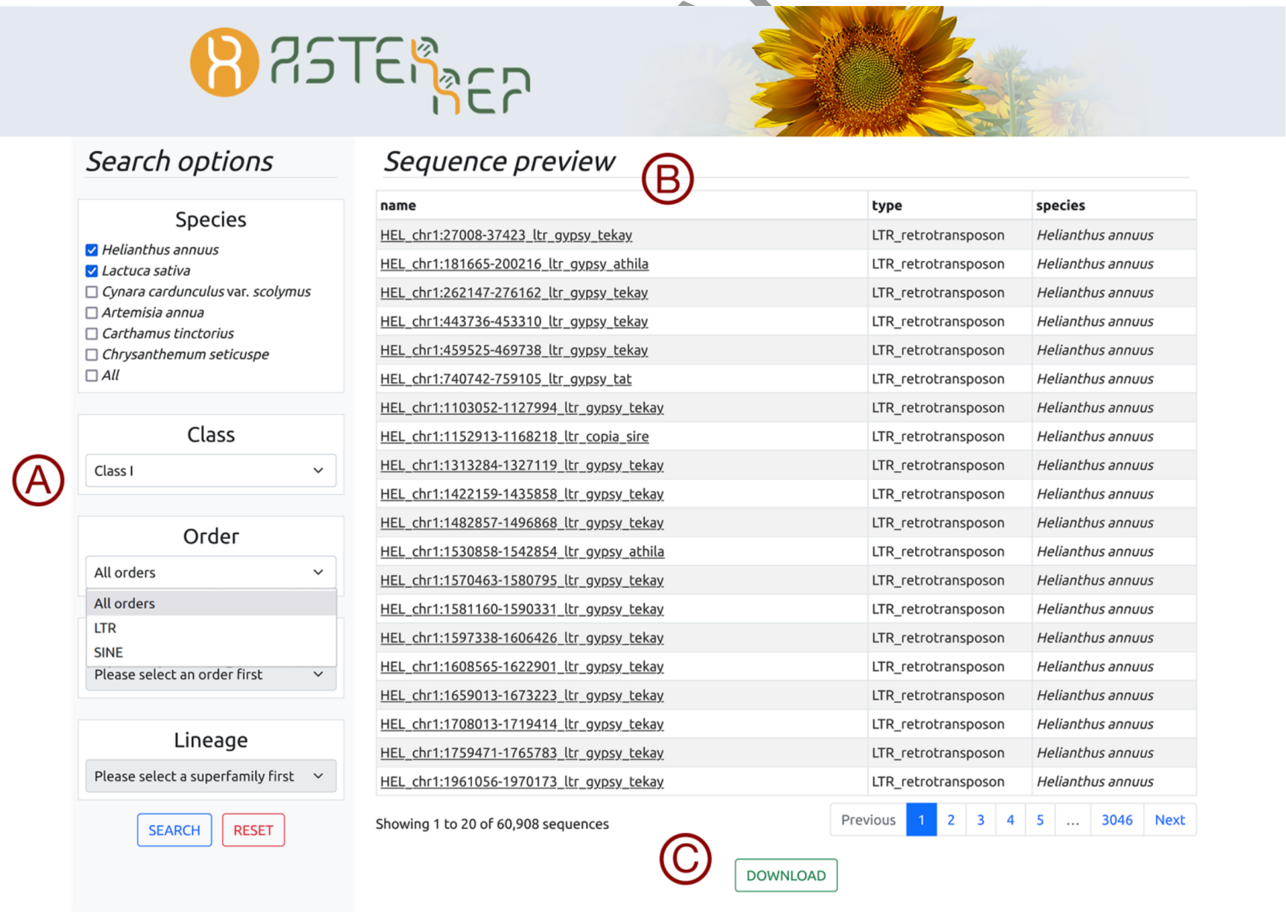
Staton, S.E. and Burke, J.M. (2015) Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. BMC genomics. 16: 1-13.

Ventimiglia, M., Marturano, G., Vangelisti, A., Usai, G., Simoni, S., Cavallini, A., Giordani, T., Natali, L., Zuccolo, A., Mascagni, F. (2022) Genome-wide identification and characterisation of exapted transposable elements in the large genome of sunflower (*Helianthus annuus* L.). The Plant Journal. doi.org/10.1111/tpj.16078

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

Legends to main figures:

Figure 1. ASTER-REP database search function. Checkboxes and drop-down menus permit the customization of search parameters (A), resulting elements are listed and visualized (B). Search results can be downloaded locally using the temporary links (C).



name	type	species
HEL_chr1:27008-37423_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:181665-200216_ltr_gypsy_athila	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:262147-276162_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:443736-453310_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:459525-469738_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:740742-759105_ltr_gypsy_tat	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1103052-1127994_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1152913-1168218_ltr_copia_sire	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1313284-1327119_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1422159-1435858_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1482857-1496868_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1530858-1542854_ltr_gypsy_athila	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1570463-1580795_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1581160-1590331_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1597338-1606426_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1608565-1622901_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1659013-1673223_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1708013-1719414_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1759471-1765783_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*
HEL_chr1:1961056-1970173_ltr_gypsy_tekay	LTR_retrotransposon	*Helianthus annuus*

Showing 1 to 20 of 60,908 sequences