RESEARCH ARTICLE

CicerSpTEdb: A web-based database for high-resolution genome-wide identification of transposable elements in *Cicer* species

Morad M. Mokhtar[1]*, Alsamman M. Alsamman[1], Haytham M. Abd-Elhalim[2], Achraf El Allali[1]*

1 African Genome Center, Mohammed VI Polytechnic University, Ben Guerir, Morocco, **2** Agricultural Genetic Engineering Research Institute, Agricultural Research Center, Giza, Egypt

* Achraf.ELALLALI@um6p.ma (AEA); morad.mokhtar@um6p.ma (MMM)

Abstract

Recently, *Cicer* species have experienced increased research interest due to their economic importance, especially in genetics, genomics, and crop improvement. The *Cicer arietinum*, *Cicer reticulatum*, and *Cicer echinospermum* genomes have been sequenced and provide valuable resources for trait improvement. Since the publication of the chickpea draft genome, progress has been made in genome assembly, functional annotation, and identification of polymorphic markers. However, work is still needed to identify transposable elements (TEs) and make them available for researchers. In this paper, we present CicerSpTEdb, a comprehensive TE database for *Cicer* species that aims to improve our understanding of the organization and structural variations of the chickpea genome. Using structure and homology-based methods, 3942 *C. echinospermum*, 3579 *C. reticulatum*, and 2240 *C. arietinum* TEs were identified. Comparisons between *Cicer* species indicate that *C. echinospermum* has the highest number of LTR-RT and hAT TEs. *C. reticulatum* has more Mutator, PIF Harbinger, Tc1 Mariner, and CACTA TEs, while *C. arietinum* has the highest number of Helitron. CicerSpTEdb enables users to search and visualize TEs by location and download their results. The database will provide a powerful resource that can assist in developing TE target markers for molecular breeding and answer related biological questions.

Database URL: http://cicersptedb.easyomics.org/index.php

Citation: Mokhtar MM, Alsamman AM, Abd-Elhalim HM, El Allali A (2021) CicerSpTEdb: A web-based database for high-resolution genome-wide identification of transposable elements in *Cicer* species. PLoS ONE 16(11): e0259540. https://doi.org/10.1371/journal.pone.0259540

Editor: Khalil Kashkush, Ben-Gurion University, ISRAEL

Received: July 23, 2021

Accepted: October 20, 2021

Published: November 11, 2021

Data Availability Statement: All relevant data are within the manuscript, Supporting Information files, and CicerSpTEdb is freely available at http://cicersptedb.easyomics.org/.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Transposable elements (TEs) are mobile DNA sequences that can move and integrate themselves in another location throughout the genome [1]. Based on the transposition systems, TEs were classified into two classes [2]. Class I is known as retrotransposons, and Class II is known as DNA transposons. Retrotransposons utilize a copy and paste system, while DNA transposons use the cut and paste systems to transpose along the genome [2]. Retrotransposons are divided into two sub-classes, the long terminal repeat-retrotransposons (LTR-RT) and the non-LTR retrotransposons [2]. More evidence documented that TEs contribute to the reshaping of plant genomes and play important roles in regulating, altering, and creating new genes, as well as its essential role in the evolutionary dynamics of host genomes [3–5]. Many reports discussed in detail the impacts of TEs in both the genome and the epigenome [6, 7], the creation of pseudo-gene [8], the alteration [8], and transcriptional silencing [9, 10]. Moreover, TEs affect the development of both vertebrates [11] and plants [12]. In rice, maize, wheat, and barley, there is a correlation between the insertion of TEs near genes and the increased mutation rates in regulatory regions and coding sequences [13]. TEs represent a large percentage of plant genomes, such as rice 40% [14], maize, and wheat 85% [15, 16]. In plants, researchers have found evidence that TEs affect agronomic traits for maize [17], grape [18], foxtail millet [19], blood oranges [20], apples [21, 22] and others. Since TEs play an important role in genome variations, their genetic variation could be considered advantageous for crop breeding [23–26].

The *Cicer* genus contains 45 species with nine annual and 36 perennial species. Only chickpea (*Cicer arietinum* L.) is cultivated in 49 countries on a large scale. Currently, only two annual *Cicer* species (*Cicer reticulatum* and *Cicer echinospermum*) are in the primary and secondary gene pools and crossable to chickpea [27]. Chickpea is one of the most important Fabaceae crops. It has special significance to food security in developing countries due to its potential nutritional and health benefits [28]. According to 2019 FAO statistics [29], about 13.7 million hectares were cultivated with chickpea in more than 47 countries, yielding about 14.2 million tons. As a member of the Fabaceae family, chickpeas can restore soil fertility by fixing atmospheric nitrogen [30]. Because of its value for the economy and human nutrition, chickpea-related research has also increased interest, especially in crop improvement, genetic, genomics, and basic biological studies [31].

A significant achievement in *Cicer* species genomics was attained as a result of publishing the genomic sequences of *Cicer arietinum* [30–32], *Cicer reticulatum* [33], and *Cicer echinospermum* [34]. Chickpea genomic studies aim to improve our knowledge of genome organization, structural variations, genome evolution, and the basic biology of legume crops. Advances in bioinformatics and sequencing technologies have led to the fast creation of large-scale sequencing and genotyping data sets for chickpea [35–37]. The integrated study of massive phenotypic and genomics data opens the door for discovering new genes, functional elements, and biological processes correlated with several economic traits [38].

The fast-growing of chickpea omics data led to the establishment of many functional genomics databases, including the microsatellites markers database "CicArMiSatDB" [39], the SNP and InDels database "CicArVarDB" [40], and the transcriptome database "CTDB" [41]. In recent years, functional genomic elements such as miRNAs [42], transcription factors [38], long non-coding RNAs [43], and transposable elements [44] were discovered for several plant species. For chickpea, miRNAs were identified in 2014 [45], transcription factor in 2016 [46], and long non-coding RNAs in 2017 [47].

At present, several multiple-species TE databases are available and are exemplified by Repbase [48], GyDB [49], PlantsDB [50], and RepetDB [51]. Species-specific TE databases are also available such as RetrOryza [52], BmTEdb [53], BrassicaTED [54], MnTEdb [55], FmTEMDb [56], PlanTE-MIR DB [57], SPTEdb [58], and ConTEdb [59]. Despite advancements in functional genome annotation of chickpea, no database for chickpea TEs has been established. Chickpea TEs need to be clearly identified in detail and made available to researchers. Studying these valuable genomic elements should accelerate the improvement of this important crop and become a new area of research in chickpea.

Genome-wide identification of TEs in the *Cicer* species and the establishment of comprehensive TE databases are key resources for the accurate characterization of genes and other genomic elements. Here, we used the Extensive *de-novo* TE Annotator (EDTA) pipeline [60] as both structure and homology-based methods to identify, classify and annotate TEs in *Cicer* species. All identified TEs were deposited for browsing and visualization in the developed *Cicer* species Transposable Elements database (CicerSpTEdb). CicerSpTEdb will represent an open resource that will allow researchers to improve our knowledge of the origin, organization, structural variations, and evolution of the *Cicer* species, including the chickpea genome. The CicerSpTEdb database will also provide an essential resource to other related legume crops. In addition, we hope that CicerSpTEdb aid plant breeders in developing TE target markers for molecular breeding and help the research community in general answer related biological questions.

Materials and methods

Genomic data

We retrieved the chickpea (*C. arietinum*) reference genome sequence of Kabuli type cultivar CDC-Frontier (ASM33114v1) from the NCBI FTP server [61] in both fasta and gff formats. Due to the unavailability of annotations, only fasta files were downloaded for *Cicer reticulatum* (GCA_002896235.1) and Cicer *echinospermum* (GCA_002896215.2) from NCBI FTP server [61].

Identification of TEs

We conducted the intact transposon identification and characterization for the 530, 657, and 715 Mbps representing the *Cicer arietinum*, *Cicer echinospermum*, *and Cicer reticulatum* reference genome, respectively, using EDTA pipeline [60]. EDTA pipeline combines tools for the structure, homology-based, and *de novo* identification methods. The EDTA pipeline combines LTRharvest [62], LTR_FINDER [63], LTR_retriever [64], Generic Repeat Finder [65], TIR-Learner [66], HelitronScanner [67], and RepeatModeler [68]. The parameters of each tool are described in S1 File.

Estimation of LTR-RT insertion time

ClustalW [69] was used to alignment the 5′and 3′ LTRs of each intact LTR-RTs to estimate the insertion time of LTR-RTs. The nucleotide substitutions/divergence among LTRs (K) were computed by applying the Kimura-2-parameter model [70] using the KaKs_Calculator program [71]. Using an evolutionary rate (r) of 1.5×10^{-8} substitutions per synonymous site per year [72–74] and the formula $T = k/2r$ the insertion time was estimated [70].

Identification of TEs positioned inside or nearby genes

Perl scripts were used to differentiate the predicted TEs according to the localization in the genome sequence. The goal was to identify TEs that are positioned within or nearby genes according to the genome annotation. For nearby genes, 10 kbp upstream genes were used to detect TEs located inside this region. We performed gene ontology on all genes that house TEs using UniProtKB [75].

Protein-protein interaction analysis

The amino acid sequences of genes that contain or are close to TEs were used for protein-protein interaction analysis using the STRING database [76]. Cytoscape 3.8.2 software and the STRING-App were used for PPI networks analysis and visualization [77, 78].

Database construction

JBrowse [79] was embedded in our developed database to map and visualize identified TEs across the reference genome. The CicerSpTEdb database was designed as an interactive web application using CSS, Perl, MySQL, PHP, HTML, and JavaScript. (Fig 1) illustrates the framework used to identify TEs in *Cicer* species and develop the proposed CicerSpTEdb.

Results and discussion

TEs identification and annotation have been formed for numerous plant genomes through extensive efforts and manual identification, e.g., Arabidopsis [80], rice [14], and maize [81]. Despite the increase in the number of sequenced plant genomes, manual identification remains labor-intensive, and automated TE annotation is needed [60]. Intact TEs are the complete structures of TEs that can transpose throughout the genome [82]. Most sequenced plant genomes have had annotated TEs. However, it is important to predict which of these TEs are still viable for mobility. Thus, in the present investigation, we focused only on the analysis of intact TEs. Using the EDTA pipeline [60], several approaches were applied to identify intact TEs in the *C. arietinum*, *C. reticulatum*, and *C. echinospermum* genomes. EDTA consist of LTRharvest [62] and LTR_FINDER [63] for LTR identification. False discoveries are filtered by LTR_retriever [64]. In addition, TIR-Learner [66] is used for TIR candidates identification, and HelitronScanner [67] is used to recognize Helitron candidates.

TEs identification in *Cicer* species

As a result, a total of 794 intact LTR-RTs were identified in *C. arietinum*, including 521 Copia, 80 Gypsy, and 193 unknown LTRs. For DNA TEs, we identified 775 Mutator followed by 245 CACTA, 215 hAT, 156 helitrons, 28 Tc1 Mariner, and 27 PIF Harbinger (Table 1). S1 Table includes details of the identified TEs, *including the* chromosome/scaffold id, TE start and end position in the genome, TE corresponding superfamily, and TE length.

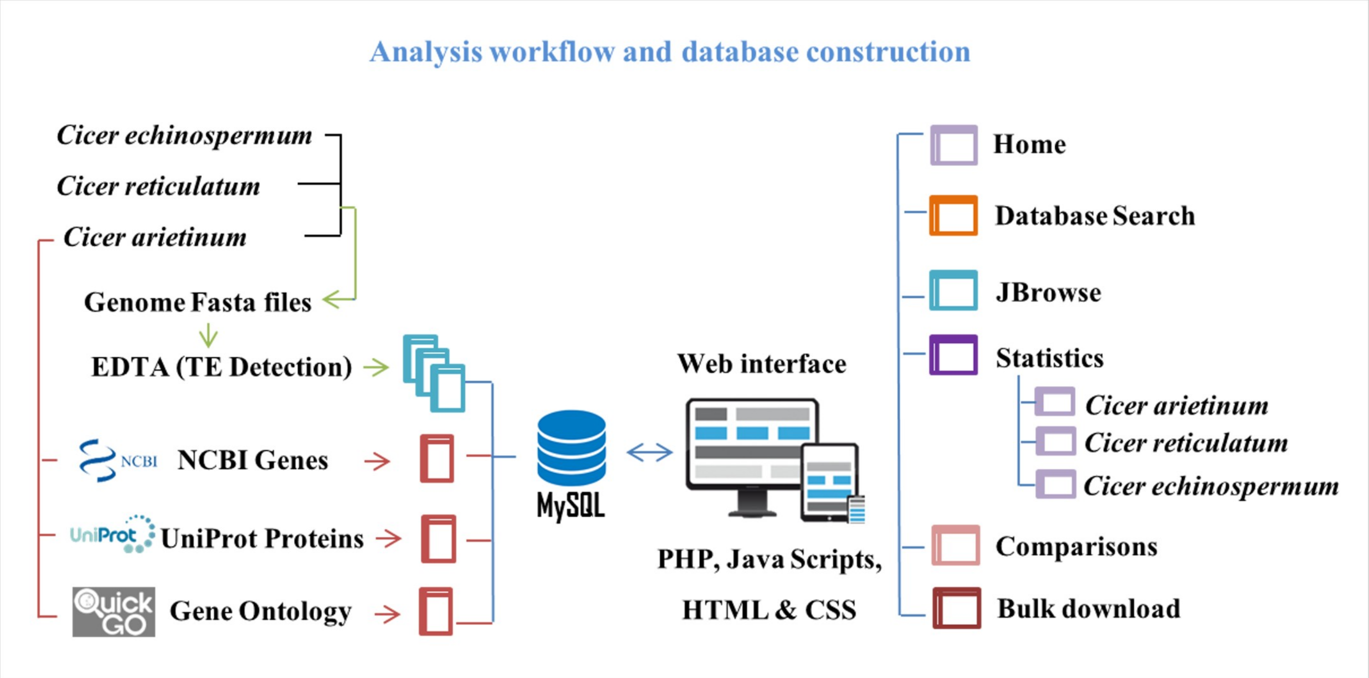


Fig 1. The overall framework used for identifying and characterizing TEs in *Cicer* species and the steps involved in creating CicerSpTEdb.

https://doi.org/10.1371/journal.pone.0259540.g001

Table 1. Summary of the intact TEs identified in *Cicer* species.

TE superfamily	*Cicer arietinum*			*Cicer reticulatum*			*Cicer echinospermum*		
	No. elements	**Total length (bp)**	**Percentage of the genome (%)**	**No. elements**	**Total length (bp)**	**Percentage of the genome (%)**	**No. elements**	**Total length (bp)**	**Percentage of the genome (%)**
Copia	521	2692203	0.507	1097	7294492	1.02	1404	10055403	1.53
Gypsy	80	466654	0.088	267	1983455	0.277	423	3515110	0.535
Unknown LTR	193	578105	0.109	407	1514753	0.212	445	1766078	0.269
CACTA	245	594115	0.112	329	822381	0.115	291	757423	0.115
hAT	215	211394	0.04	211	231880	0.032	263	274877	0.042
Helitron	156	1498699	0.282	151	1142623	0.16	152	1438982	0.219
Mutator	775	1001091	0.189	1025	1397305	0.195	892	864817	0.132
PIF Harbinger	27	21975	0.004	56	100906	0.014	40	52275	0.008
Tc1 Mariner	28	40588	0.008	36	56626	0.008	32	43869	0.007
Total	2240	7104824	1.339	3579	14544421	2.033	3942	18768834	2.857

https://doi.org/10.1371/journal.pone.0259540.t001

Interestingly, Varsheny et al. [30] reported that approximately 49.41% of the *C. arietinum* genome is composed of TEs and unclassified repeats, including 617,505 repeat retrotransposons and 197,959 DNA transposons. However, our investigation produced lower numbers and found that the intact TEs represent only 2240 elements, approximately 1.3% of the whole genome. This difference could be because many of them will not be intact TEs and may be nested elements or fragmented.

(Fig 2) shows the distribution and histogram of TEs across eight *C. arietinum* chromosomes. As shown, the distribution of TEs superfamily were 217, 179, 187, 200, 230, 240, 200, 76 elements for chromosomes 1, 2, 3, 4, 5, 6, 7, and 8, respectively. Chromosome 6 (CA6) had the highest presence of TEs (240 elements), including 57 Copia, 55 Mutator, 29 unknown, 27 CACTA, 27 hAT, 21 helitrons, 12 Gypsy, 9 PIF, and 3 Tc1_mariner.

The analysis of *C. reticulatum* TEs revealed that 14.54 Mb (approximately 2% of the genome) were intact TEs. The highest copy number of LTR-RT was 1097 Copia, followed by 193 unknown LTRs, and 80 Gypsy. In addition, 1808 *C. reticulatum* DNA transposons were detected and consist of 1025 Mutator, 329 CACTA, 211 hAT, 151 helitrons, 56 PIF Harbinger, and 36 Tc1 Mariner (Table 1 and S2 Table). For *C. echinospermum*, a total of 3942 intact TEs were detected and covered 18.76 Mb (approximately 2.8% of the genome). Out of these, 2272 LTR-RTs included 1404 Copia, 423 Gypsy, and 445 unknown LTRs. *C. reticulatum* has more Mutator, PIF Harbinger, Tc1 Mariner, and CACTA TE types (Table 1 and S3 Table). The present investigation revealed that *C. reticulatum*, *C. echinospermum*, and *C. arietinum* have a higher copy number of Copia superfamily than Gypsy, which is consistent with previous reports of flax [73], grape [83], cocoa [84], and cucumber [85].

To our knowledge, there are no TE reports for both *C. echinospermum and C. reticulatum* to allow the discussion of our new findings. In addition, the draft genomes of *C. reticulatum*, *C. echinospermum*, and *C. arietinum* were partially sequenced. The sizes of their available sequences are 530, 657, and 715 Mb, respectively. This variation in size could be correlated with the variation of the identified copy number of TEs in these genomes.

Estimation of LTR-RT insertion time

It is deemed that the 5′ and 3' LTRs are the same at transposition time for each LTR-RT. Consequently, based on nucleotide substitutions/divergence among LTR-RT, the 5' and 3' LTRs accumulated through ages were applied to estimate the insertion time [86–88]. In the present

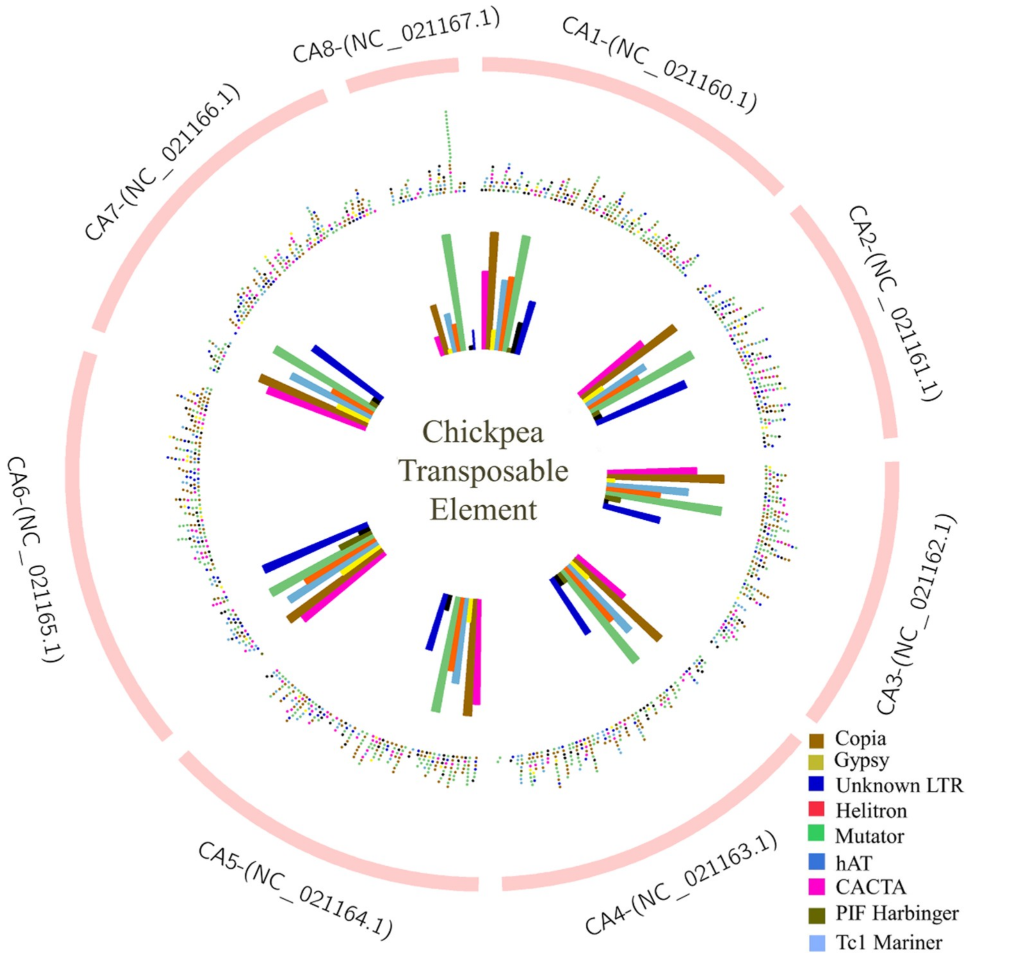


Fig 2. The distribution of intact TEs across eight chickpea chromosomes. The outermost circle in pink-colored represents eight *C. arietinum* chromosomes (CA1 to CA8). The middle circle illustrates the distribution of TE types in different colors, and the innermost circle shows the histogram of TE types in each chromosome.

https://doi.org/10.1371/journal.pone.0259540.g002

investigation, the 5′ and 3' LTR nucleotide substitutions were used to estimate the identified intact LTR-RT insertion time across *Cicer* species.

For *C. arietinum*, the minimum and maximum assumed age after discarding outliers using boxplot analysis ranged from 0 to 4.4 million years (MY) with an average of 0.94 MY. The unknown elements were older than Copia and Gypsy (4.3, 4.2, and 2.9 MY, respectively). The average insertion ages of the unknown, Copia and Gypsy elements are 1.3, 1.3, and 0.85 MY, respectively (Fig 3). Interestingly, about 23.3% of Copia, 21.7% of Gypsy, and 23.4% of unknown elements have estimated ages of < 0 MY, and they may still be active elements. While the proportions of insertion times that are more than 1.2 MY were 56.5%, 53.1%, and 40% of Gypsy, unknown, and Copia, respectively (Fig 3).

For *C. reticulatum*, the estimated age ranged from 0 to 100 MY with an average of 33.7 MY. The unknown elements are younger than Copia and Gypsy elements. The Copia, Gypsy, and unknown elements' average insertion times were 33.4, 40, and 28.7 MY, respectively. Overall, about 10.5% of Copia, 8.9% of Gypsy, and 7.5% of unknown elements were estimated to have an age of < 0 MY. While the proportions of insertion times that are more than 1.2 MY were 86%, 85.6%, and 81.9% of Gypsy, Copia, and unknown elements, respectively (Fig 3).

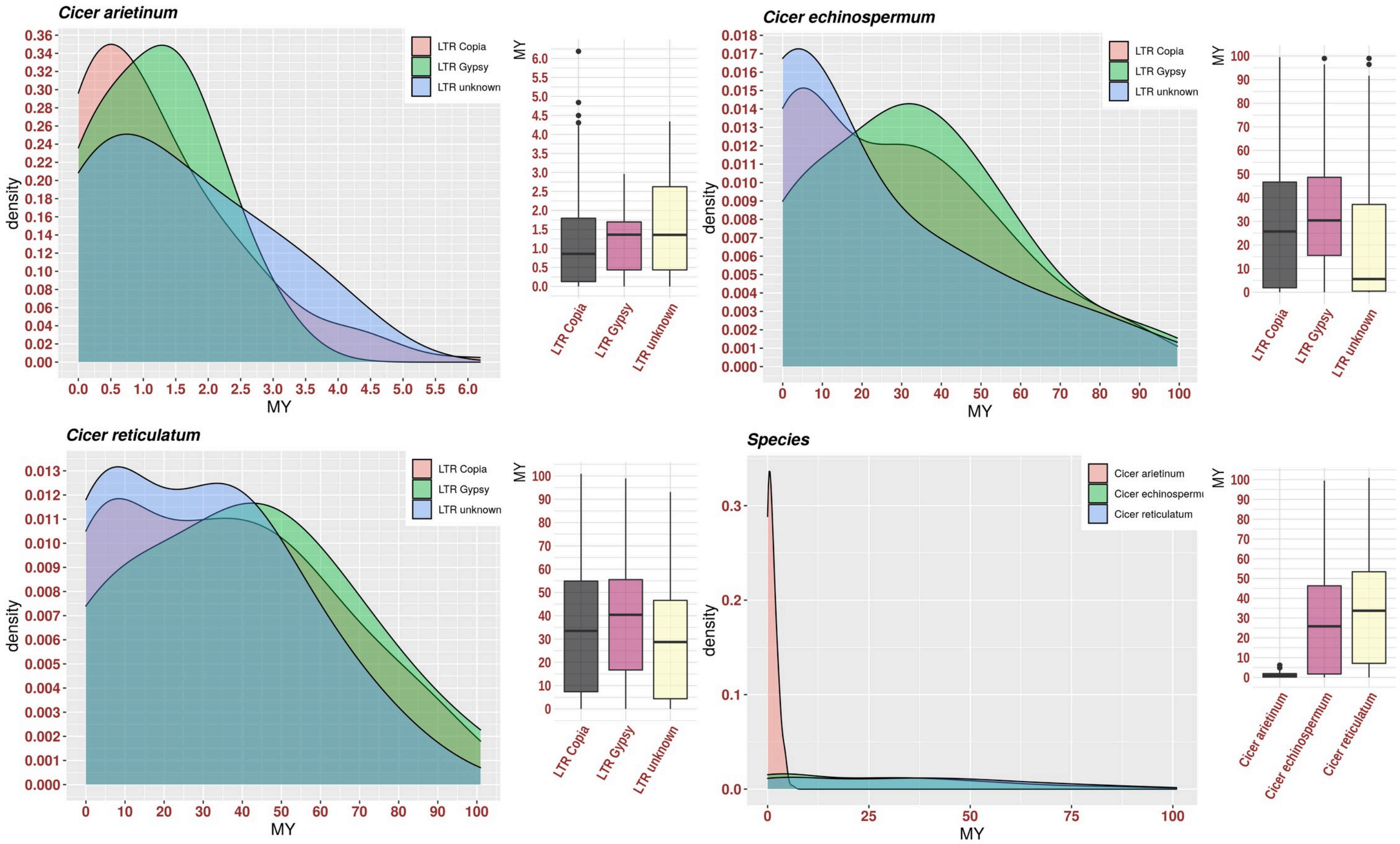


Fig 3. Estimation of *Cicer* species LTR-RT insertion age in millions of years (My).

https://doi.org/10.1371/journal.pone.0259540.g003

For *C. echinospermum*, the estimated age ranged from 0 to 99.5 MY with an average of 25.8 MY. The unknown elements are younger than Copia and Gypsy elements. The average insertion time of the Copia, Gypsy, and unknown elements were 30.3, 5.5, and 25.7 MY, respectively. Overall, about 6.6% of Gypsy, 12.5% of Copia, and 21.7% of unknown elements had an estimated age of $<$ 0 MY. While the proportions of insertion times that are more than 1.2 MY were 87.5%, 78.7%, and 62.6% of Gypsy, Copia, and unknown elements, respectively (Fig 3).

The chromosomes number of the *C. arietinum*, *C. echinospermum*, and *C. reticulatum* were the same 2n = 16. Therefore *C. reticulatum* was in the primary gene pools and recognized as the wild ancestor of the *C. arietinum*. In addition, genetic studies revealed that the *C. echinospermum* was closely and in secondary gene pools of *C. arietinum* [27, 89]. The estimation of *Cicer* species LTR-RT insertion time revealed that the wild species *C. echinospermum* and *C. reticulatum* were older than the cultivated species *C. arietinum* (Fig 3). Based on estimated LTR-RT age, *C. arietinum* may be derived/split from their wild progenitor *C. reticulatum* ~ 4.4–6 MY.

TEs length distribution

The lengths of *C. arietinum* intact TEs ranged from 80 bp to 19.7 kb for both DNA and LTR transposons. The average sizes of various superfamilies were Gypsy 5.8 kb, Copia 5.1 kb, unknown LTR 2.9 kb, Helitron 9.6 kb, CACTA 2.4 kb, Tc1 Mariner 1.4 kb, Mutator 1.2 kb, hAT 0.9 kb, and PIF Harbinger 0.8 kb (Fig 4). For *C. echinospermum* TEs, lengths ranged from

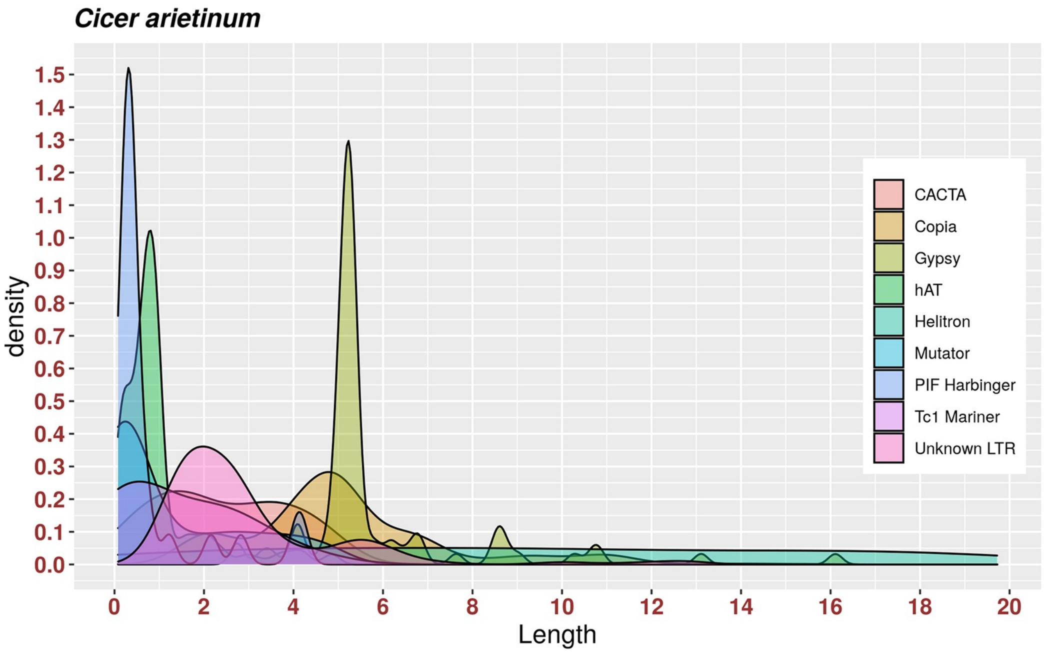


Fig 4. The distribution of *C. arietinum* TEs superfamilies according to their length.

https://doi.org/10.1371/journal.pone.0259540.g004

80 bp to 19.6 kb for DNA and LTR transposons. The average sizes of various superfamilies were Gypsy 8.3 kb, Copia 7.1 kb, unknown LTR 3.9 kb, Helitron 9.4 kb, CACTA 2.6 kb, Tc1 Mariner 1.3 kb, PIF Harbinger 1.3 kb, hAT 1 kb, and Mutator 0.9 kb (S1 Fig). For *C. reticulatum*, TE lengths ranged from 80 bp to 21.7 kb for both DNA and LTR transposons. The average sizes of various superfamilies were Gypsy 7.4 kb, Copia 6.6 kb, unknown LTR 3.7 kb, Helitron 7.5 kb, CACTA 2.4 kb, PIF Harbinger 1.8 kb, Tc1 Mariner 1.5 kb, Mutator 1.3 kb, and hAT 1 kb (S2 Fig).

Identification of TEs positioned inside or nearby genes

The transposition of TEs across the genome may affect both nearby genes' expression and genes unlinked to the insertion. TEs can affect genes through the movement, duplication, and recombination processes, creating new genes or altering the gene structure [86]. Furthermore, they may alter the expression of nearby genes by inserting themself within *cis*-regulatory elements or by presenting a new *cis*-regulatory element that may act as gene enhancers or repressors [90]. Due to the unavailability of annotation for *C. reticulatum* and *C. echinospermum*, only TEs identified in *C. arietinum* were subject to further analysis to determine TEs that are inside or nearby genes.

Overall, 1162 *C. arietinum* intact TEs (about 51.8%) were positioned inside (TE-gene chimeras) or nearby genes. Only 20 TEs were found within pseudo-genes (Table 2). From these elements, 426 (approximately 36.6%) were LTR-RT, and 736 (approximately 63.3%) were DNA transposons. For LTR-RT, the Copia superfamily was overrepresented, followed by unknown elements and gypsy superfamilies with 250, 140, and 36 elements, respectively. However, DNA transposons included 326 Mutator, 173 hAT, 150 CACTA, 38 Helitron, 28 Tc1 Mariner, and 21 PIF Harbinger elements.

More evidence documented that TEs construct the chimeric genes (TE-gene chimeras) in plants [8, 91]. Previous eukaryotic reports revealed that one thousand human proteins contain

Table 2. Summary of the *C. arietinum* TEs that are positioned inside or nearby genes.

TEs superfamily	Number of elements				
	Inside genes (TE-gene chimeras)	Inside pseudo-genes	Near genes (0-2kb)	Near genes (2-10kb)	Other genomic regions
Copia	100	5	29	116	271
Gypsy	27	1	1	7	44
Unknown LTR	91	2	3	44	53
CACTA	54	5	14	77	95
hAT	70	3	28	72	42
Helitron	9		10	19	118
Mutator	92	4	53	177	449
PIF Harbinger	9		1	11	6
Tc1 Mariner	19		1	8	
Total	471	20	140	531	1078

https://doi.org/10.1371/journal.pone.0259540.t002

TEs [92, 93], and few expressed genes house TEs in Drosophila [94]. In addition, approximately 1.2% of Arabidopsis proteins were constructed from TE-gene chimeras [95]. Consistent with our results, previous studies reported that Class I TEs favor transposition inside gene-poor heterochromatic regions [96]. In comparison, euchromatin regions have more Class II TEs that prefer to transposition inside or nearby genes [97–99]. The finding that TE elements inside and nearby genes in *C. arietinum* are overrepresented by Copia than Gypsy is consistent with previous studies in maize [100], Arabidopsis [95], and sugarcane [101].

Regarding class II TEs, Mutator was overrepresented, followed by hATs, while Helitrons were underrepresented. Our results agree with Lockton et al. [95] and Leonardo et al. [73], who found that hATs were overrepresented in Arabidopsis, while Helitrons were underrepresented in flax. Interestingly, the Mutator superfamily was overrepresented inside and closely to genes in *C. arietinum*. More evidence documented that the association between Mutator elements and genes supports TE-mediated gene transposition in rice [91] and Arabidopsis [102]. Finally, the distance between identified TEs near genes and genes ranged from 3 to 9.9 kb (S4 Table). From these TEs, 140 elements were located within 2kb near genes, among this 53 Mutator, 29 Copia, 28 hAT, 14 CACTA, ten Helitron, three Unknown_LTR, 1 Gypsy, 1 PIF_-Harbinger, and 1 Tc1_Mariner.

Functional classification by gene ontology analyses

To determine whether the genes housing TEs in their sequence were disrupted or still have functions, UniProtKB [75] was used to map and classify 441 TE-gene chimeras according to their function (GO terms). Only 366 genes were successfully mapped to 482 UniProtKB IDs and assigned to GO terms. These GO terms include 315 genes assigned to 486 molecular functions GO terms (S3 Fig), 213 genes assigned to 393 biological processes GO terms (Fig 5), and 215 genes assigned to 325 cellular components GO terms (S4 Fig). The 393 GO terms assigned to biological processes were distributed among 164 cellular process, 134 metabolic process, 41 biological regulation, 29 response to the stimulus, 19 localization, two reproductive processes, two developmental processes, one flower development, and one response to another organism (Fig 5). Molecular function analysis showed the overrepresented TE-gene chimeras were catalytic activity, binding, and transporter activity. In contrast, the underrepresented TE-gene chimeras were DNA-binding transcription factor activity and Structural constituent of ribosome (S3 Fig). Based on these results, we can infer that a high percentage of TE-gene chimeras are still functional in various biological processes in *C. arietinum*. However, to determine their level of activity, further experimental validation still needs to be performed.

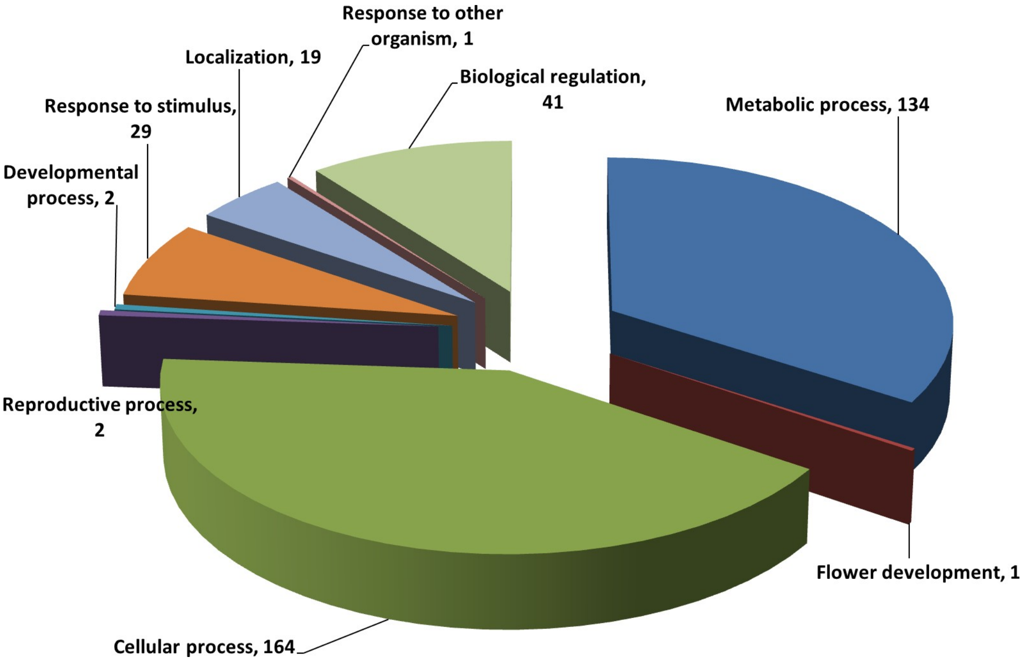


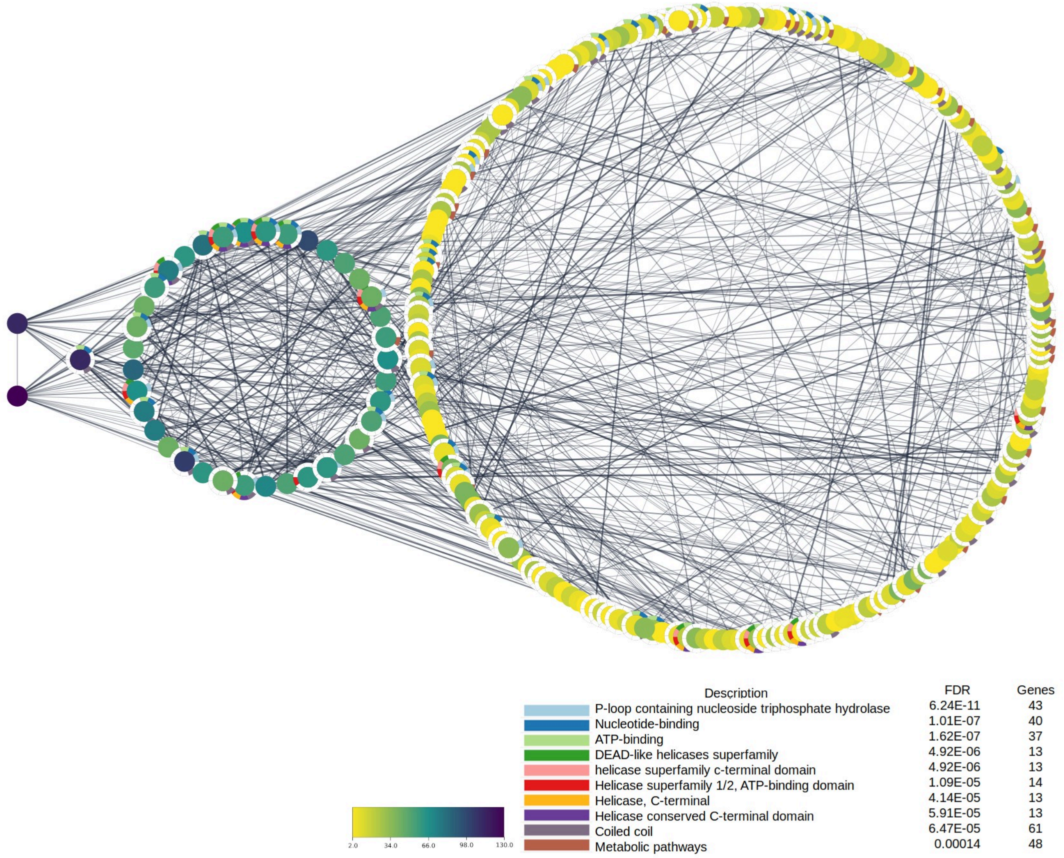
Fig 5. Gene ontology of 213 *C. arietinum* TE-gene chimeras assigned to 393 biological processes GO terms.

https://doi.org/10.1371/journal.pone.0259540.g005

Protein-protein interaction analysis

Protein-protein interaction (PPI) analysis is an instrumental analysis tool. It can show how a group of genes interact in the cellular system and their activity level, thus showing their biological importance. Furthermore, it adds more information about the type of connection these proteins have and the biological pathways that they control. We examined the protein interaction activity of chickpea genes that contain or are close to TEs. The STRING database retrieved the interaction information of 619 proteins, from which high interactive proteins could be identified [76]. PPI analysis was carried out for genes that contain, or are close to, TEs, as well as for all genes collectively (Fig 6, S5 and S6 Figs).

The most interactive gene was the DNA-directed RNA polymerase II subunit (RPB1), which contains nearby TEs (Fig 6). The RPB1 gene is an essential component of the RNA polymerase transcription machinery, catalyzing the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates [103]. Several research articles have discussed the relationship between RPB1 and TEs [104]. The carboxy-terminal domain of eukaryotic RPB1 has a heptad-repeat structure that is intrinsically disordered. These repeats regulate the length of the RPB1 C-terminal domain, which in turn controls transcription activation by influencing transcription cycle coordination. Such a relationship could impact the regulator's system of essential cellular functions [105]. Phosphoglycerate kinase one gene (PGK1) also revealed many nearby TEs compared to other chickpea genes (S5 Fig). The activity of TEs influences polyploidization modifications in plant genomes, affecting the copy number and the content of genes. Due to its single-copy status per diploid chromosome in several plant species, the PGK1 gene has been widely used to reveal the evolutionary history of complex genomes [106, 107]. The String database offers the ability to analyze PPI networks depending on their biological pathways. It has revealed that the most gene-enriched pathways are those linked to Neclotide binding and metabolic pathways, and mostly, these genes are linked by lab experiments, published articles (text mining), or their genome physical distance (neighborhood) (S6 Fig).



	Description	FDR	Genes
🟦	P-loop containing nucleoside triphosphate hydrolase	6.24E-11	43
🟦	Nucleotide-binding	1.01E-07	40
🟩	ATP-binding	1.62E-07	37
🟩	DEAD-like helicases superfamily	4.92E-06	13
🟥	helicase superfamily c-terminal domain	4.92E-06	13
🟥	Helicase superfamily 1/2, ATP-binding domain	1.09E-05	14
🟧	Helicase, C-terminal	4.14E-05	13
🟪	Helicase conserved C-terminal domain	5.91E-05	13
🟪	Coiled coil	6.47E-05	61
🟫	Metabolic pathways	0.00014	48

Fig 6. Protein-protein interaction for TE-gene chimeras.

https://doi.org/10.1371/journal.pone.0259540.g006

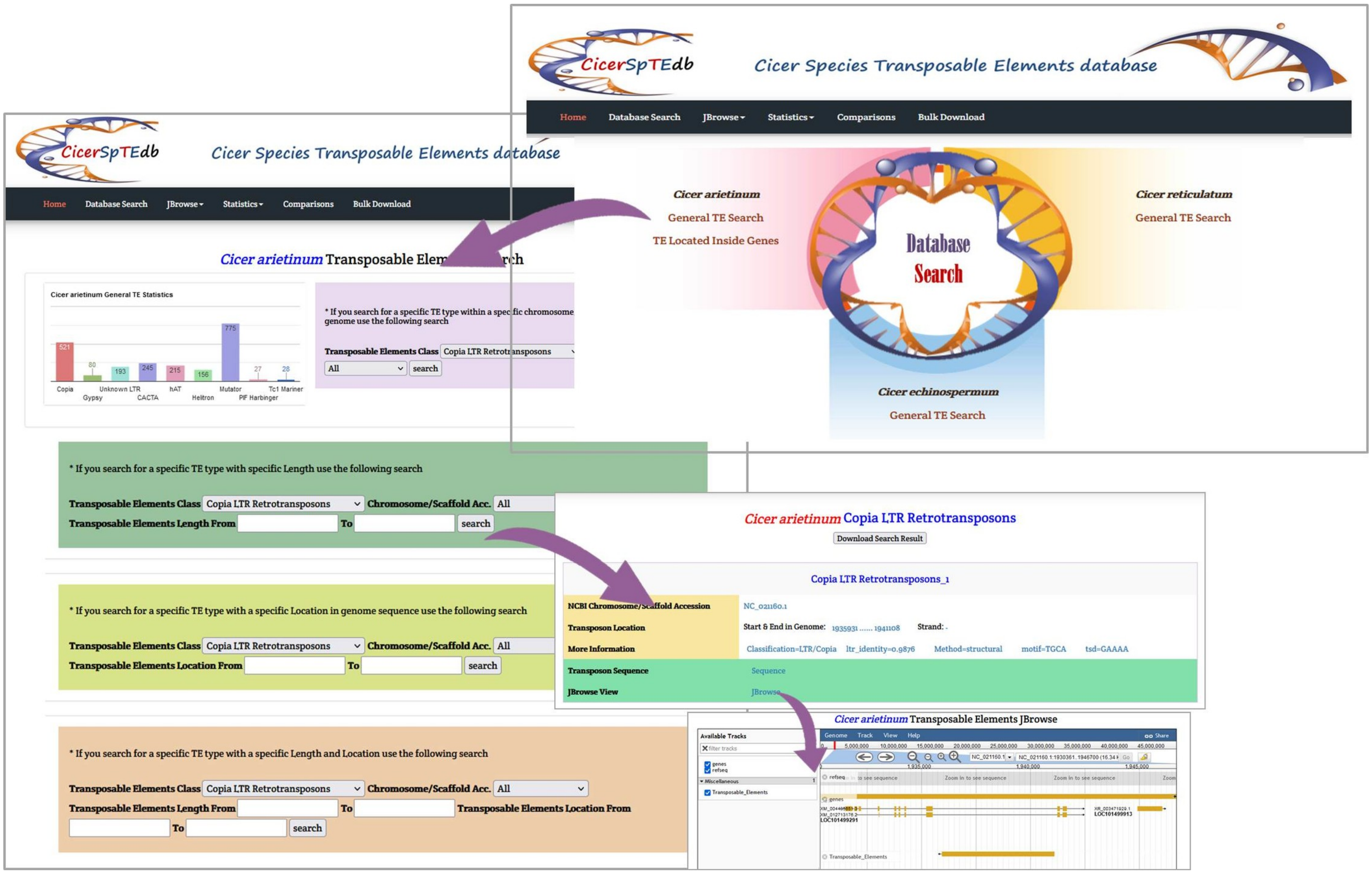
Out of these results, we can point out that TEs affect distinctive genes with high interplay activity and consequently impact a widespread biological process in the chickpea genome.

CicerSpTEdb web interface

The *Cicer* Transposable Elements database (CicerSpTEdb) is accessible through a user-friendly portal (http://cicersptedb.easyomics.org/index.php). The website allows users to explore and understand the Cicer transposable elements. The database offers comprehensive details of TEs and their features in the genome, especially for chickpea. The CicerSpTEdb interface allows users to search, browse, compare, and download TEs interactively. From the homepage, users can capture the essential information about CicerSpTEdb and access relative external databases and software. The navigation bar allows access to six sections for browsing and retrieving data, including Home, Database Search, JBrowse, Statistics, Comparisons, and Bulk Download.

The Database Search page

From anywhere on CicerSpTEdb's interface, users can access the Database Search page through the top bar that links to the main search page. The latter provides links to access two separate *Cicer arietinum* pages. The first page allows a general search of TEs, while the second option provides detailed information on TEs located within genes. In addition, links to access



* If you search for a specific TE type within a specific chromosome genome use the following search

Transposable Elements Class: Copia LTR Retrotransposons; All

* If you search for a specific TE type with specific Length use the following search

Transposable Elements Class: Copia LTR Retrotransposons; Chromosome/Scaffold Acc.: All
Transposable Elements Length From To

* If you search for a specific TE type with a specific Location in genome sequence use the following search

Transposable Elements Class: Copia LTR Retrotransposons; Chromosome/Scaffold Acc.: All
Transposable Elements Location From To

* If you search for a specific TE type with a specific Length and Location use the following search

Transposable Elements Class: Copia LTR Retrotransposons; Chromosome/Scaffold Acc.: All
Transposable Elements Length From To Transposable Elements Location From
To

Cicer arietinum Copia LTR Retrotransposons

Copia LTR Retrotransposons_1

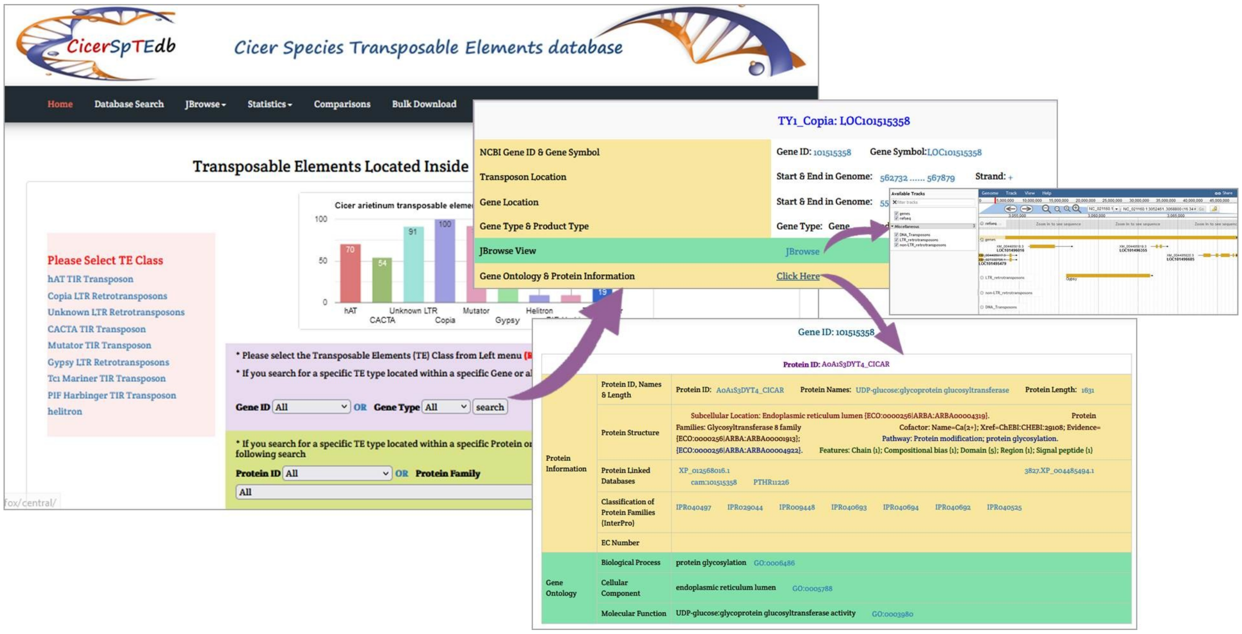
NCBI Chromosome/Scaffold Accession	NC_021160.1
Transposon Location	Start & End in Genome: 1935931 ...... 1941108 Strand: -
More Information	Classification=LTR/Copia ltr_identity=0.9876 Method=structural motif=TGCA tsd=GAAAA
Transposon Sequence	Sequence
JBrowse View	JBrowse

Fig 7. CicerSpTEdb's TE general search page.

https://doi.org/10.1371/journal.pone.0259540.g007

Cicer reticulatum and *Cicer echinospermum* TE general search pages are also available. The top section of the TE general search page allows users to see a statistical chart of all identified TEs by type. The main section is divided into four sub-sections. It allows users to 1) search by TE type within a specific chromosome/scaffold or within the whole-genome, 2) search by TE type in a specific chromosome/scaffold or whole-genome with specific TE length, 3) search by TE type with a specific location inside the genome, 4) search by TE type with a specific length and location in the genome. The search results appear on a new page and include NCBI chromosome/scaffold accession, transposon start, end, length, and corresponding strand in the genome, TE structure details, download TE sequence, and a JBrowse link. The results can be exported by clicking the download button (Fig 7).

On the page dedicated to TEs located within genes, the left section is used to select TE types, and the top part of the page allows users to see a statistical chart of all identified TEs located inside genes. The main section is divided into four sub-sections that allow searching by different keys. Users can search using the gene ID and gene type (gene or pseudo-gene) or using either the protein ID, protein family, or enzyme EC number. The results are displayed on a separate page and include gene details (gene ID, symbol, location, type, and product), JBrowse link, and link to gene ontology and protein information. The link will redirect the user to a new page that contains all accessible protein information such as protein ID, names,



Transposable Elements Located Inside

Please Select TE Class
hAT TIR Transposon
Copia LTR Retrotransposons
Unknown LTR Retrotransposons
CACTA TIR Transposon
Mutator TIR Transposon
Gypsy LTR Retrotransposons
Tc1 Mariner TIR Transposon
PIF Harbinger TIR Transposon
helitron

- Please select the Transposable Elements (TE) Class from Left menu (R
- If you search for a specific TE type located within a specific Gene or al

Gene ID All OR Gene Type All search

- If you search for a specific TE type located within a specific Protein or following search

Protein ID All OR Protein Family
All

TY1_Copia: LOC101515358

NCBI Gene ID & Gene Symbol	Gene ID: 101515358 Gene Symbol: LOC101515358
Transposon Location	Start & End in Genome: 562732 ...... 567879 Strand: +
Gene Location	Start & End in Genome: 55
Gene Type & Product Type	Gene Type: Gene
JBrowse View	JBrowse
Gene Ontology & Protein Information	Click Here

Gene ID: 101515358

Protein ID: A0A1S3DYT4_CICAR

Protein Information	Protein ID, Names & Length	Protein ID: A0A1S3DYT4_CICAR Protein Names: UDP-glucose:glycoprotein glucosyltransferase Protein Length: 1631
	Protein Structure	Subcellular Location: Endoplasmic reticulum lumen {ECO:0000256|ARBA:ARBA00004319}. Protein Families: Glycosyltransferase 8 family Cofactor: Name=Ca(2+); Xref=ChEBI:CHEBI:29108; Evidence={ECO:0000256|ARBA:ARBA00001913}; Pathway: Protein modification; protein glycosylation. {ECO:0000256|ARBA:ARBA00004922}. Features: Chain (1); Compositional bias (1); Domain (5); Region (1); Signal peptide (1)
	Protein Linked Databases	XP_012568016.1 3827.XP_004485494.1 cam:101515358 PTHR11226
	Classification of Protein Families (InterPro)	IPR040497 IPR029044 IPR009448 IPR040693 IPR040694 IPR040692 IPR040525
	EC Number	
Gene Ontology	Biological Process	protein glycosylation GO:0006486
	Cellular Component	endoplasmic reticulum lumen GO:0005788
	Molecular Function	UDP-glucose:glycoprotein glucosyltransferase activity GO:0003980

Fig 8. CicerSpTEdb's search page for TEs located within genes.

https://doi.org/10.1371/journal.pone.0259540.g008

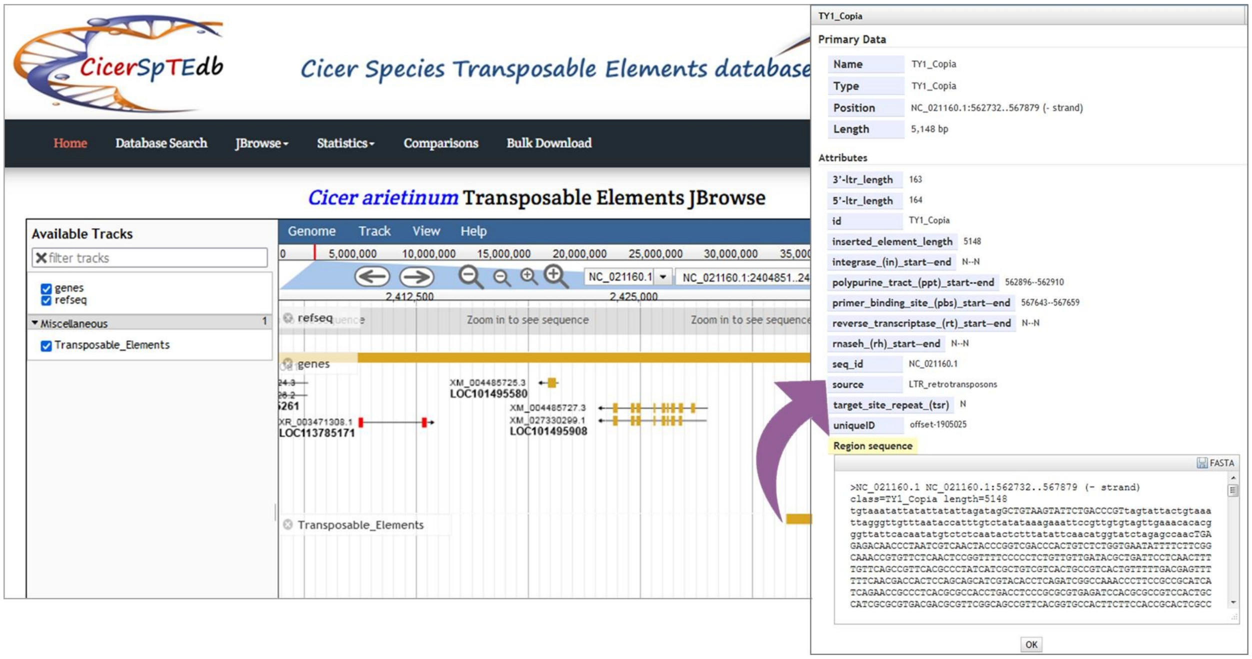
length, structure, family, EC number, externally linked databases, and gene ontology. All external databases are cross-linked, and the results can be exported by clicking the download button (Fig 8).

The JBrowse page

JBrowse is a powerfully interactive genome visualization tool established to illustrate the coordinates of TEs in the genome. By clicking the JBrowse from the top bar of any page, a visualization window will display all chickpea coordinates, including reference sequence, genes, and identified transposons. Users can retrieve any TE's data (name, position in the genome, length, described information, and sequence) by clicking on it in the JBrowse graphic interface. In addition, the JBrowse page offers an important function that allows users to browse all genes around TEs and the genes that TEs are positioned inside. The latter visualization function is an easy way to build a clear idea of each TE and understand the interaction between TEs and the surrounding genes (Fig 9).

Other pages

The Statistics page was created to provide researchers with a visualization of several statistics computed from the data in CicerSpTEdb. Users can access the Statistics page through a drop-down menu that links to three pages, one for each studied genome (S7 Fig). The comparisons page was created to provide researchers with a visual comparison of identified TEs between *C. arietinum*, *C. reticulatum*, and *C. echinospermum*. Users can access the comparison page through the top bar from any page (S8 Fig). The Bulk Download page was created to allow researchers to download all stored data in CicerSpTEdb. The Bulk Download page allows users to select the species, transposons, and data type (fasta or gff3 files) from organism name, TE type, and data type drop-down menus (S9 Fig).



Name	TY1_Copia
Type	TY1_Copia
Position	NC_021160.1:562732..567879 (- strand)
Length	5,148 bp

3'-ltr_length	163
5'-ltr_length	164
id	TY1_Copia
inserted_element_length	5148
integrase_(in)_start--end	N--N
polypurine_tract_(ppt)_start--end	562896--562910
primer_binding_site_(pbs)_start--end	567643--567659
reverse_transcriptase_(rt)_start--end	N--N
rnaseh_(rh)_start--end	N--N
seq_id	NC_021160.1
source	LTR_retrotransposons
target_site_repeat_(tsr)	N
uniqueID	offset-1905025

```
>NC_021160.1 NC_021160.1:562732..567879 (- strand)
class=TY1_Copia length=5148
tgtaaatattatattatattagatagGCTGTAAGTATTCTGACCCGTtagtattactgtaaa
ttagggttgtttaataccatttgtctatataaagaaattccgttgtgtagttgaaacacacg
ggttattcacaatatgtctctcaatactctttatattcaacatggtatctagagccaacTGA
GAGACAACCCTAATCGTCAACTACCCGGTCGACCCACTGTCTCTGGTGAATATTTTCTTCGG
CAAACCGTGTTCTCAACTCCGGTTTTCCCCCTCTGTTGTTGATACGCTGATTCCTCAACTTT
TGTTCAGCCGTTCACGCCCTATCATCGCTGTCGTCACTGCCGTCACTGTTTTTGACGAGTTT
TTTCAACGACCACTCCAGCAGCATCGTACACCTCAGATCGGCCAAACCCTTCCGCCGCATCA
TCAGAACCGCCCTCACGCGCCACCTGACCTCCCGCGCGTGAGATCCACGCGCCGTCCACTGC
CATCGCGCGTGACGACGCGTTCGGCAGCCGTTCACGGTGCCACTTCTTCCACCGCACTCGCC
```

Fig 9. CicerSpTEdb's JBrowse page.

https://doi.org/10.1371/journal.pone.0259540.g009

Conclusion

CicerSpTEdb is the first comprehensive database designated to *Cicer* species transposable elements. This database contains 9761 TEs that combines DNA transposon and LTR retrotransposons. Moreover, the proposed database is available through an easy-to-use interface to allow researchers to search, browse, and download the identified TEs in *C*. *echinospermum*, *C*. *reticulatum*, and *C*. *arietinum*. We propose to continuously update the database and improve its applications to achieve its goals. We expect CicerSpTEdb to provide a valuable resource that can be used to improve our knowledge of the origin, organization, structural variations, and evolution of the *Cicer* species genomes and other related legume crops. CicerSpTEdb should help researchers develop TEs target markers for molecular breeding and to answer any related biological questions.

Supporting information

S1 Fig. The distribution of *C*. *echinospermum* TEs superfamilies according to their length.
(TIF)

S2 Fig. The distribution of *C*. *reticulatum* TEs superfamilies according to their length.
(TIF)

S3 Fig. Gene ontology of 315 *C*. *arietinum* TE-gene chimeras assigned to 486 molecular functions GO terms.
(TIF)

S4 Fig. Gene ontology of 215 *C*. *arietinum* TE-gene chimeras assigned to 325 cellular components GO terms.
(TIF)

S5 Fig. Protein-protein interaction for genes nearby TEs.
(TIF)

S6 Fig. Protein-protein interaction for both TE-gene chimeras and genes nearby TEs.
(TIF)

S7 Fig. CicerSpTEdb's statistics page.
(TIF)

S8 Fig. CicerSpTEdb's comparisons page.
(TIF)

S9 Fig. CicerSpTEdb's bulk download page.
(TIF)

S1 Table. The details of the identified TEs in *C. arietinum*, including the chromosome/scaffold id, TE start and end position in the genome, TE corresponding superfamily, and TE length.
(XLSX)

S2 Table. The details of the identified TEs in *C. reticulatum*, including the chromosome/scaffold id, TE start and end position in the genome, TE corresponding superfamily, and TE length.
(XLSX)

S3 Table. The details of the identified TEs in *C. echinospermum*, including the chromosome/scaffold id, TE start and end position in the genome, TE corresponding superfamily, and TE length.
(XLSX)

S4 Table. The details of the identified TEs that located near genes in *C. arietinum*.
(XLSX)

S1 File. The programs and their parameters used for identification of TEs in *Cicer* species.
(DOCX)

Acknowledgments

The authors acknowledge the African Supercomputing Center at Mohamed VI Polytechnic University for the supercomputing resources (https://ascc.um6p.ma/) made available for conducting the research reported in this paper.

Availability

CicerSpTEdb is freely available at http://cicersptedb.easyomics.org/index.php

Author Contributions

Conceptualization: Morad M. Mokhtar, Haytham M. Abd-Elhalim.

Data curation: Morad M. Mokhtar, Haytham M. Abd-Elhalim.

Formal analysis: Morad M. Mokhtar.

Funding acquisition: Achraf El Allali.

Methodology: Alsamman M. Alsamman, Haytham M. Abd-Elhalim.

Resources: Achraf El Allali.

Software: Alsamman M. Alsamman.

Supervision: Achraf El Allali.

Validation: Achraf El Allali.

Visualization: Morad M. Mokhtar, Alsamman M. Alsamman.

Writing – original draft: Morad M. Mokhtar.

Writing – review & editing: Achraf El Allali.

References

1. Finnegan DJ. Transposable elements in eukaryotes. Int Rev Cytol. Elsevier; 1985; 93: 281–326. https://doi.org/10.1016/s0074-7696(08)61376-5 PMID: 2989205
2. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. Nature Publishing Group; 2007; 8: 973–982. https://doi.org/10.1038/nrg2165 PMID: 17984973
3. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol. Springer; 2000; 42: 251–269. https://doi.org/10.1023/A:1006344508454 PMID: 10688140
4. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. Nature Publishing Group; 2008; 9: 397–405. https://doi.org/10.1038/nrg2337 PMID: 18368054
5. Bucher E, Reinders J, Mirouze M. Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr Opin Plant Biol. Elsevier; 2012; 15: 503–510. https://doi.org/10.1016/j.pbi.2012.08.006 PMID: 22940592
6. Sahebi M, Hanafi MM, van Wijnen AJ, Rice D, Rafii MY, Azizi P, et al. Contribution of transposable elements in the plant's genome. Gene. Elsevier; 2018; 665: 155–166. https://doi.org/10.1016/j.gene.2018.04.050 PMID: 29684486
7. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. Annual Reviews; 2014; 65: 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811 PMID: 24579996
8. Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, et al. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell. American Society of Plant Biologists; 2006; 18: 1791–1802. https://doi.org/10.1105/tpc.106.041905 PMID: 16829590
9. Lisch D. Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol. Annual Reviews; 2009; 60: 43–66. https://doi.org/10.1146/annurev.arplant.59.032607.092744 PMID: 19007329
10. Lin L, Tang H, Compton RO, Lemke C, Rainville LK, Wang X, et al. Comparative analysis of Gossypium and Vitis genomes indicates genome duplication specific to the Gossypium lineage. Genomics. Elsevier; 2011; 97: 313–320. https://doi.org/10.1016/j.ygeno.2011.02.007 PMID: 21352905
11. Yandım C, Karakülah G. Expression dynamics of repetitive DNA in early human embryonic development. BMC Genomics. 2019; 20: 439. https://doi.org/10.1186/s12864-019-5803-1 PMID: 31151386
12. Sun F, Guo W, Du J, Ni Z, Sun Q, Yao Y. Widespread, abundant, and diverse TE-associated siRNAs in developing wheat grain. Gene. Elsevier; 2013; 522: 1–7. https://doi.org/10.1016/j.gene.2013.03.101 PMID: 23562726
13. Wicker T, Yu Y, Haberer G, Mayer KFX, Marri PR, Rounsley S, et al. DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nat Commun. 2016; 7: 12790. https://doi.org/10.1038/ncomms12790 PMID: 27599761
14. Project IRGS. The map-based sequence of the rice genome. Nature. 2005; 436: 793–800. https://doi.org/10.1038/nature03895 PMID: 16100779
15. Meyers BC, Tingey S V, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. Cold Spring Harbor Lab; 2001; 11: 1660–1676. https://doi.org/10.1101/gr.188201 PMID: 11591643
16. Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, et al. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics. Oxford University Press; 2008; 180: 1071–1086. https://doi.org/10.1534/genetics.108.092304 PMID: 18780739

17. Doebley J, Stec A, Gustus C. teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics. Genetics Soc America; 1995; 141: 333–346. Available: https://www.genetics.org/content/141/1/333.short https://doi.org/10.1093/genetics/141.1.333 PMID: 8536981

18. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. Science (80-). Citeseer; 2004; 304: 982. https://doi.org/10.1126/science.1095011 PMID: 15143274

19. Kawase M, Fukunaga K, Kato K. Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions. Mol Genet Genomics. Springer; 2005; 274: 131–140. https://doi.org/10.1007/s00438-005-0013-8 PMID: 16133169

20. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. Am Soc Plant Biol; 2012; 24: 1242–1255. https://doi.org/10.1105/tpc.111.095232 PMID: 22427337

21. Yao J-L, Dong Y-H, Morris BAM. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. Proc Natl Acad Sci. National Acad Sciences; 2001; 98: 1306–1311. https://doi.org/10.1073/pnas.031502498 PMID: 11158635

22. Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun. Nature Publishing Group; 2019; 10: 1–13. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

23. Paszkowski J. Controlled activation of retrotransposition for plant breeding. Curr Opin Biotechnol. Elsevier; 2015; 32: 200–206. https://doi.org/10.1016/j.copbio.2015.01.003 PMID: 25615932

24. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to global change: a transposable element—epigenetics perspective. Trends Ecol Evol. Elsevier; 2016; 31: 514–526. https://doi.org/10.1016/j.tree.2016.03.013 PMID: 27080578

25. Thieme M, Lanciano S, Balzergue S, Daccord N, Mirouze M, Bucher E. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biol. 2017; 18: 134. https://doi.org/10.1186/s13059-017-1265-4 PMID: 28687080

26. Thieme M, Bucher E. Chapter Six—Transposable Elements as Tool for Crop Improvement. In: Mirouze M, Bucher E, Gallusci P, editors. Plant Epigenetics Coming of Age for Breeding Applications. Academic Press; 2018. pp. 165–202. https://doi.org/10.1016/bs.abr.2018.09.001

27. Toker C, Berger J, Eker T, Sari D, Sari H, Gokturk RS, et al. Cicer turcicum: A New Cicer Species and Its Potential to Improve Chickpea. Front Plant Sci. Frontiers; 2021; 12: 587. https://doi.org/10.3389/fpls.2021.662891 PMID: 33936152

28. Jukanti AK, Gaur PM, Gowda CLL, Chibbar RN. Nutritional quality and health benefits of chickpea (Cicer arietinum L.): a review. Br J Nutr. Cambridge University Press; 2012; 108: S11–S26. https://doi.org/10.1017/S0007114512000797 PMID: 22916806

29. FAOSTAT [Internet]. [cited 10 Jan 2021]. Available: http://www.fao.org/faostat/en/#data/QC

30. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe a G, et al. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol. 2013; 31: 240–246. https://doi.org/10.1038/nbt.2491 PMID: 23354103

31. Parween S, Nawaz K, Roy R, Pole AK, Venkata Suresh B, Misra G, et al. An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). Sci Rep. 2015; 5: 12806. https://doi.org/10.1038/srep12806 PMID: 26259924

32. Jain M, Misra G, Patel RK, Priya P, Jhanwar S, Khan AW, et al. A draft genome sequence of the pulse crop chickpea (C icer arietinum L.). Plant J. Wiley Online Library; 2013; 74: 715–729. https://doi.org/10.1111/tpj.12173 PMID: 23489434

33. Gupta S, Nawaz K, Parween S, Roy R, Sahu K, Kumar Pole A, et al. Draft genome sequence of Cicer reticulatum L., the wild progenitor of chickpea provides a resource for agronomic trait improvement. Dna Res. Oxford University Press; 2017; 24: 1–10. https://doi.org/10.1093/dnares/dsw042 PMID: 27567261

34. NCBI. No Title [Internet]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCA_002896215.2

35. Varshney RK, Thudi M, Roorkiwal M, He W, Upadhyaya HD, Yang W, et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nat Genet. Nature Publishing Group; 2019; 51: 857–864. https://doi.org/10.1038/s41588-019-0401-3 PMID: 31036963

36. Gaur R, Verma S, Pradhan S, Ambreen H, Bhatia S. A high-density SNP-based linkage map using genotyping-by-sequencing and its utilization for improved genome assembly of chickpea (Cicer arietinum L.). Funct Integr Genomics. Springer; 2020; 20: 763–773. https://doi.org/10.1007/s10142-020-00751-y PMID: 32856221

37. Varshney RK, Thudi M, Muehlbauer FJ. The chickpea genome: An introduction. The Chickpea Genome. Springer; 2017. pp. 1–4. https://doi.org/10.1007/978-3-319-66117-9_1

38. Lehti-Shiu MD, Panchy N, Wang P, Uygun S, Shiu S-H. Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. Biochim Biophys Acta (BBA)-Gene Regul Mech. Elsevier; 2017; 1860: 3–20. https://doi.org/10.1016/j.bbagrm.2016.08.005

39. Doddamani D, Katta MA, Khan AW, Agarwal G, Shah TM, Varshney RK. CicArMiSatDB: the chickpea microsatellite database. BMC Bioinformatics. 2014; 15: 212. https://doi.org/10.1186/1471-2105-15-212 PMID: 24952649

40. Doddamani D, Khan AW, Katta MAVSK, Agarwal G, Thudi M, Ruperao P, et al. CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea. Database. 2015; 2015. https://doi.org/10.1093/database/bav078 PMID: 26289427

41. Verma M, Kumar V, Patel RK, Garg R, Jain M. CTDB: An Integrated Chickpea Transcriptome Database for Functional and Applied Genomics. PLoS One. 2015; 10: e0136880. https://doi.org/10.1371/journal.pone.0136880 PMID: 26322998

42. Li S, Liu L, Zhuang X, Yu Y, Liu X, Cui X, et al. MicroRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in Arabidopsis. Cell. Elsevier; 2013; 153: 562–574. https://doi.org/10.1016/j.cell.2013.04.005 PMID: 23622241

43. Vieira LM, Grativol C, Thiebaut F, Carvalho TG, Hardoim PR, Hemerly A, et al. PlantRNA_Sniffer: a SVM-based workflow to predict Long Intergenic non-coding RNAs in plants. Non-coding RNA. Multidisciplinary Digital Publishing Institute; 2017; 3: 11. https://doi.org/10.3390/ncrna3010011 PMID: 29657283

44. Satheesh V, Fan W, Chu J, Cho J. Recent advancement of NGS technologies to detect active transposable elements in plants. Genes Genomics. Springer; 2021; 1–6. https://doi.org/10.1007/s13258-020-01012-9 PMID: 33111208

45. Jain M, Chevala VVSN, Garg R. Genome-wide discovery and differential regulation of conserved and novel microRNAs in chickpea via deep sequencing. J Exp Bot. Oxford University Press UK; 2014; 65: 5945–5958. https://doi.org/10.1093/jxb/eru333 PMID: 25151616

46. Gayali S, Acharya S, Lande NV, Pandey A, Chakraborty S, Chakraborty N. CicerTransDB 1.0: a resource for expression and functional study of chickpea transcription factors. BMC Plant Biol. 2016; 16: 169. https://doi.org/10.1186/s12870-016-0860-y PMID: 27472917

47. Singh U, Khemka N, Rajkumar MS, Garg R, Jain M. PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. Nucleic Acids Res. 2017; 45: e183–e183. https://doi.org/10.1093/nar/gkx866 PMID: 29036354

48. Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. Karger Publishers; 2005; 110: 462–467. https://doi.org/10.1159/000084979 PMID: 16093699

49. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. Oxford University Press; 2010; 39: D70–D74. https://doi.org/10.1093/nar/gkq1061 PMID: 21036865

50. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, et al. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. Oxford University Press; 2012; 41: D1144–D1151. https://doi.org/10.1093/nar/gks1153 PMID: 23203886

51. Amselem J, Cornut G, Choisne N, Alaux M, Alfama-Depauw F, Jamilloux V, et al. RepetDB: a unified resource for transposable element references. Mob DNA. Springer; 2019; 10: 1–8. https://doi.org/10.1186/s13100-018-0144-1 PMID: 30622655

52. Chaparro C, Guyot R, Zuccolo A, Piégu B, Panaud O. RetrOryza: a database of the rice LTR-retrotransposons. Nucleic Acids Res. Oxford University Press; 2007; 35: D66–D70. https://doi.org/10.1093/nar/gkl780 PMID: 17071960

53. Xu H-E, Zhang H-H, Xia T, Han M-J, Shen Y-H, Zhang Z. BmTEdb: a collective database of transposable elements in the silkworm genome. Database. Oxford Academic; 2013; 2013. https://doi.org/10.1093/database/bat055 PMID: 23886610

54. Murukarthick J, Sampath P, Lee SC, Choi B-S, Senthil N, Liu S, et al. BrassicaTED-a public database for utilization of miniature transposable elements in Brassica species. BMC Res Notes. BioMed Central; 2014; 7: 1–11. https://doi.org/10.1186/1756-0500-7-1 PMID: 24382056

55. Ma B, Li T, Xiang Z, He N. MnTEdb, a collective resource for mulberry transposable elements. Database. Oxford Academic; 2015; 2015. https://doi.org/10.1093/database/bav004 PMID: 25725060

56. Yadav CB, Bonthala VS, Muthamilarasan M, Pandey G, Khan Y, Prasad M. Genome-wide development of transposable elements-based markers in foxtail millet and construction of an integrated database. DNA Res. 2014; 22: 79–90. https://doi.org/10.1093/dnares/dsu039 PMID: 25428892

57. Lorenzetti APR, De Antonio GYA, Paschoal AR, Domingues DS. PlanTE-MIR DB: a database for transposable element-related microRNAs in plant genomes. Funct Integr Genomics. Springer; 2016; 16: 235–242. https://doi.org/10.1007/s10142-016-0480-5 PMID: 26887375

58. Yi F, Jia Z, Xiao Y, Ma W, Wang J. SPTEdb: a database for transposable elements in salicaceous plants. Database. Oxford Academic; 2018; 2018. https://doi.org/10.1093/database/bay024 PMID: 29688371

59. Yi F, Ling J, Xiao Y, Zhang H, Ouyang F, Wang J. ConTEdb: a comprehensive database of transposable elements in conifers. Database. Oxford Academic; 2018; 2018. https://doi.org/10.1093/database/bay131 PMID: 30576494

60. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. BioMed Central; 2019; 20: 1–18. https://doi.org/10.1186/s13059-018-1612-0 PMID: 30606230

61. NCBI FTP server [Internet]. [cited 6 May 2020]. Available: ftp://ftp.ncbi.nlm.nih.gov/

62. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. Springer; 2008; 9: 1–14. https://doi.org/10.1186/1471-2105-9-1 PMID: 18173834

63. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007; 35: W265–W268. https://doi.org/10.1093/nar/gkm286 PMID: 17485477

64. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. American Society of Plant Biologists; 2018; 176: 1410–1422. https://doi.org/10.1104/pp.17.01310 PMID: 29233850

65. Shi J, Liang C. Generic Repeat Finder: a high-sensitivity tool for genome-wide de novo repeat detection. Plant Physiol. American Society of Plant Biologists; 2019; 180: 1803–1815. https://doi.org/10.1104/pp.19.00386 PMID: 31152127

66. Su W, Gu X, Peterson T. TIR-learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. Mol Plant. Elsevier; 2019; 12: 447–460. https://doi.org/10.1016/j.molp.2019.02.008 PMID: 30802553

67. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci. National Acad Sciences; 2014; 111: 10263–10268. https://doi.org/10.1073/pnas.1410068111 PMID: 24982153

68. Smit AFA, Hubley R, Green P. RepeatModeler Open-1.0 (2008–2015). In: Seattle, USA: Institute for Systems Biology. [Internet]. 2015 p. 2018. Available: http://www.repeatmasker.org/RepeatModeler/

69. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. Oxford university press; 1994; 22: 4673–4680. https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417

70. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. Springer; 1980; 16: 111–120. Available: https://link.springer.com/content/pdf/10.1007/BF01731581.pdf PMID: 7463489

71. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. Elsevier; 2006; 4: 259–263. https://doi.org/10.1016/S1672-0229(07)60007-2 PMID: 17531802

72. Koch MA, Haubold B, Mitchell-Olds T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol. Oxford University Press; 2000; 17: 1483–1498. https://doi.org/10.1093/oxfordjournals.molbev.a026248 PMID: 11018155

73. González LG, Deyholos MK. Identification, characterization and distribution of transposable elements in the flax (Linum usitatissimum L.) genome. BMC Genomics. 2012; 13: 644. https://doi.org/10.1186/1471-2164-13-644 PMID: 23171245

74. Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, de Mattos Fontes MR, et al. Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol. BioMed Central; 2015; 15: 1–16. https://doi.org/10.1186/s12870-014-0410-4 PMID: 25592487

75. UniProtKB [Internet]. Available: https://www.uniprot.org/uploadlists/

76. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2016; 45: D362–D368. https://doi.org/10.1093/nar/gkw937 PMID: 27924014

77. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. American Chemical Society; 2019; 18: 623–632. https://doi.org/10.1021/acs.jproteome.8b00702 PMID: 30450911

78. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003; 13: 2498–2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

79. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016; 17: 66. https://doi.org/10.1186/s13059-016-0924-1 PMID: 27072794

80. Kaul Samir and Koo Hean L and Jenkins Jennifer and Rizzo Michael and Rooney Timothy and Tallon Luke J and Feldblyum, et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408: 796–815. https://doi.org/10.1038/35048692 PMID: 11130711

81. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science (80-). 2009; 326: 1112–1115. https://doi.org/10.1126/science.1178534 PMID: 19965430

82. Su W, Ou S, Hufford MB, Peterson T. A Tutorial of EDTA: Extensive De Novo TE Annotator. Plant Transposable Elem. Springer; 2021; 55–67. https://doi.org/10.1007/978-1-0716-1134-0_4 PMID: 33900591

83. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007; 449: 463–467. https://doi.org/10.1038/nature06148 PMID: 17721507

84. Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, et al. The genome of Theobroma cacao. Nat Genet. 2011; 43: 101–108. https://doi.org/10.1038/ng.736 PMID: 21186351

85. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, et al. The genome of the cucumber, Cucumis sativus L. Nat Genet. 2009; 41: 1275–1281. https://doi.org/10.1038/ng.475 PMID: 19881527

86. Zhao D, Ferguson AA, Jiang N. What makes up plant genomes: The vanishing line between transposable elements and genes. Biochim Biophys Acta—Gene Regul Mech. 2016; 1859: 366–380. https://doi.org/10.1016/j.bbagrm.2015.12.005 PMID: 26709091

87. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998; 20: 43–45. https://doi.org/10.1038/1695 PMID: 9731528

88. Da Costa ZP, Cauz-Santos LA, Ragagnin GT, Van Sluys M-A, Dornelas MC, Berges H, et al. Transposable element discovery and characterization of LTR-retrotransposon evolutionary lineages in the tropical fruit species Passiflora edulis. Mol Biol Rep. 2019; 46: 6117–6133. https://doi.org/10.1007/s11033-019-05047-4 PMID: 31549373

89. Ladizinsky G, Adler A. The origin of chickpea Cicer arietinum L. Euphytica. 1976; 25: 211–217. https://doi.org/10.1007/BF00041547

90. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011; 472: 115–119. https://doi.org/10.1038/nature09861 PMID: 21399627

91. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004; 431: 569–573. https://doi.org/10.1038/nature02953 PMID: 15457261

92. Nekrutenko A, Li W-H. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 2001; 17: 619–621. https://doi.org/10.1016/s0168-9525(01)02445-3 PMID: 11672845

93. Britten R. Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci. National Academy of Sciences; 2006; 103: 1798–1803. https://doi.org/10.1073/pnas.0510007103 PMID: 16443682

94. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. BMC Biol. 2005; 3: 24. https://doi.org/10.1186/1741-7007-3-24 PMID: 16283942

95. Lockton S, Gaut BS. The Contribution of Transposable Elements to Expressed Coding Sequence in Arabidopsis thaliana. J Mol Evol. 2009; 68: 80–89. https://doi.org/10.1007/s00239-008-9190-5 PMID: 19125217

96. Pereira V. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thalianagenome. Genome Biol. 2004; 5: R79. https://doi.org/10.1186/gb-2004-5-10-r79 PMID: 15461797

97. Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics. 1995; 140: 315–324. https://doi.org/10.1093/genetics/140.1.315 PMID: 7635296

98. SanMiguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. Nested Retrotransposons in the Intergenic Regions of the Maize Genome. Science (80-). 1996; 274: 765–768. https://doi.org/10.1126/science.274.5288.765 PMID: 8864112

99. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009; 19: 1419–1428. https://doi.org/10.1101/gr.091678.109 PMID: 19478138

100. Bennetzen JL. The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol. 1996; 4: 347–353. https://doi.org/10.1016/0966-842x(96)10042-1 PMID: 8885169

101. Rossi M, Araujo PG, Van Sluys M-A. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). Genet Mol Biol. SciELO Brasil; 2001; 24: 147–154. https://doi.org/10.1590/S1415-47572001000100020

102. Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, et al. Transposon-Mediated Expansion and Diversification of a Family of ULP-like Genes. Mol Biol Evol. 2006; 23: 1254–1268. https://doi.org/10.1093/molbev/msk015 PMID: 16581939

103. Doyle O, Corden JL, Murphy C, Gall JG. The distribution of RNA polymerase II largest subunit (RPB1) in the Xenopus germinal vesicle. J Struct Biol. 2002; 140: 154–166. https://doi.org/10.1016/s1047-8477(02)00547-6 PMID: 12490164

104. Kunze R, Saedler H, Lönnig W-E. Plant Transposable Elements. In: Callow JA, editor. Classic Papers. Academic Press; 1997. pp. 331–470. https://doi.org/10.1016/S0065-2296(08)60284-0

105. Sawicka A, Villamil G, Lidschreiber M, Darzacq X, Dugast-Darzacq C, Schwalb B, et al. Transcription activation depends on the length of the RNA polymerase II C-terminal domain. EMBO J. 2021; 40: e107015. https://doi.org/10.15252/embj.2020107015 PMID: 33555055

106. Peng Y, Zhou P, Zhao J, Li J, Lai S, Tinker NA, et al. Phylogenetic relationships in the genus Avena based on the nuclear Pgk1 gene. PLoS One. Public Library of Science; 2018; 13: 1–18. https://doi.org/10.1371/journal.pone.0200047 PMID: 30408035

107. Tang C, Qi J, Chen N, Sha L-N, Wang Y, Zeng J, et al. Genome origin and phylogenetic relationships of Elymus villosus (Triticeae: Poaceae) based on single-copy nuclear Acc1, Pgk1, DMC1 and chloroplast trnL-F sequences. Biochem Syst Ecol. 2017; 70: 168–176. https://doi.org/10.1016/j.bse.2016.11.011