https://doi.org/10.56093/ijas.v90i5.104388

Indian Journal of Agricultural Sciences **90** (5): 1025–31, May 2020/Article

Genome-wide comparative and evolutionary analysis of transposable elements in eight different legume plants

PAWAN KUMAR JAYASWAL, ASHEESH SHANKER and NAGENDRA KUMAR SINGH*

ICAR-National Institute for Plant Biotechnology, Pusa, New Delhi 110 012, India

Received: 30 August 2019; Accepted: 12 September 2019

ABSTRACT

Transposable elements (TEs) are a major component of the eukaryotic genomes, which are highly dynamic in nature and significantly contribute in the expansion of genome. We have genome sequence information on several legume species but there is limited information regarding the evolutionary pattern of TEs in these. To understand the expansion of the genomes, we did comparative analysis of TEs in eight different legume species, viz. *Arachis durensis* (Adu,2.7Gb)*, Arachis ipaensis* (Aip,2.7Gb)*, Cicer arietinum* (Car,738.09 Mb)*, Cajanus cajan* (Cca,858 Mb)*, Glycine max* (Gma,1115 Mb)*, Lotus japonicas* (Lja,472Mb)*, Medicago truncatula* (Mtr,465 Mb) and *Vignaan gularis* (Van,612 Mb). Our analysis showed that, the TEs in legume genome varied between 27.86% (Lja) to 70.62% (Aip) and LTR was the most dominant category over other TEs. Two *Arachis* species from *Dalbergia* tribe differ significantly in their total TEs content (Adu: 60.23%, Aip:70.62%). Comparative analysis indicated that despite the abundance of species-specific TEs in these genome, total 2,850 copies of repeat elements were conserved among all eight selected legume species. These belonged to LTR (n=2,514), non-LTR (n=14), and DNA transposons (n= 133). Evolutionary analysis revealed that most of the conserved TEs belonging tothe same tribe were clustered together, indicating introgression of repeats via horizontal transfer process. Intra and inter tribe divergence time analysis of conserved TEs provided evidence of single and multiple duplication events in the eight legume species.

Key words: Comparative analysis, Divergence analysis, Legume species, Phylogenetic tree

Large portions of the eukaryotic genome consist of different types of repetitive elements like transposable elements (TEs) and simple repeat (mini and microsatellite) which are highly repetitive and mobile in nature and play an important role in genome evolution with cut-paste and copy-paste mechanism.TEs are highly impactful in the expansion of genome, for example, the 82 Mb genome size of the *Utricularia gibba* a carnivorous bladderwort minute plant species contains 3.15% of repeat elements (Ibarra-Laclette *et al*. 2013) while rice, maize, and jute contain 35% (Gill *et al*. 2010), 85% (Schnable *et al*. 2009), 51.9% (Sarkar *et al*. 2017) repeat sequence respectively which is broadly distributed across the chromosome of their respective species. To explore the different categories of RNA intermediate Class I and DNA intermediate Class II TEs, numbers of program have been developed which support the idea of homology as well as de-novo based repeat element estimation process which provides the valuable information about the variability of the genome structure in the long evolutionary process (Singh *et al.* 2012, Sarkar *et al*. 2017).

Present analysis is focused on the structural characterization, phylogeny, and divergence analysis of conserved TEs among the eight legume species. We implemented the various program to do the in-depth analysis of conserved repeat elements specifically in context of evolution and divergence of different families of Class I as well as Class II TEs. Overall to understand the evolutionary dynamics of conserved TEs of legume, we estimated the synonymous substitution value of the possible combination of repeat pairs which suggested about the number of the event of duplication occurred in close as well as distantly related legume species and ultimately the finding enabled us to highlight the novel aspect into repeat dynamics across the leguminosae family.

MATERIALS AND METHODS

Present study was carried out at ICAR-National Institute for Plant Biotechnology, Pusa, New Delhi during 2016-19.

Estimation of TEs in legume species and phylogenetic

*Corresponding author e-mail: nksingh4@gmail.com

tree study: To build the TE libraries of eight legume plant species, we retrieved the genome sequence data of Car (Varshney *et al.* 2013), Adu, Aip (Bertioli*et al.* 2016), *Mtr* (Young *et al.* 2011), Lja (Sato *et al.* 2008), Gma (Schmutz *et al.* 2010), Cca (Singh *et al.* 2012) and Van (Yang *et al.* 2015) from different public database. The Repeat Modeler (Benson 1999, Bao and Eddy 2002), SINE finder (SINE_Scan-v1.0, Mao and Wang 2017) and MITE-Hunter (Han and Wessler 2010) program were used for de novo repeat mining, and merged all the identified repeat families along with the Rep base library data (https://www.girinst.org/repbase/, accessed on Jan 12, 2017) to create a mega repeat database. Further, Repeat Masker program was used for the masking the repeat elements of individual species and later on all masked sequences were pulled from samtools-1.3.1 program (http://www.htslib.org/download/) and further conserved TEs were identified using the customized parameter of BLASTN (Singh *et al.* 2012).

Bayesian phylogenetic tree was developed based on conserved class I and II TEs. The conserved repeat sequences were aligned using mafft version 7 program (Katoh and Standley 2013) and the tree was developed using Mrbayes v3.2.1 (Ronquist *et al*. 2012, Jayaswal *et al*. 2019) with HKY+GAMMA substitution model and visualized in fig tree program. The divergence time was estimated between conserved TEs as:

$$T=Ks/2r$$

where T, time in million years ago; Ks, synonymous substitution; r, rate of synonymous substitution; here we considered $r=1.3\times10^{-8}$ (Ma and Bennetzen 2004). The synonymous substitution (Ks) values for each pair of repeat were computed using DnaSP program (Nei and Gojoboris 1986, Librado and Rozas 2009).

RESULTS AND DISCUSSION

Abundance and diversity of TEs in legume genome: We analyzed the abundance of TE families in eight legume species and the result (S Fig 1a) confirmed the variability in the distribution of TEs content in selected eight legume species. Results showed that both Adu (~1024.20/2700Mb) and Aip (~1337.9/2700Mb) species were from *Dalbergieae* tribe contain 59.78% and 69.02% TEs. Cca, Gma, and Van were from *Phaseoleae* tribe of legume species contain 61.31%, 48.97% and 40.01% of TEs respectively, while two species from Trifolieae tribe contain 30.85% (Mtr) as well as 45.03% (Car) of TEs. Lja species from *Loteae* tribe was having the smallest genome size in comparison of other legume species which contain only 26.42% of class I and II and unclassified TEs. The comparative bar-graph of major repeat families revealed that class I repeat elements were more abundant than class II TEs (S Fig 1b). LTR-RTs were predominant over other repeat elements and varied from 10.23% (Lja) to 58.20% (Aip) while the total masked genome sequence of non-LTR element was varied in between 1.37%(Van) – 14.12% (Mtr). The result showed CACTA, Harbinger, hAT, Helitron, MULE_MuDR and Mariner like repeat elements are most abundant in DNA TEs and masked elements varied between 2.85 (Lja) – 5.94% (Car). The analysis also showed a large portion of the total identified repeat elements were unclassified and may be species specific (Supplementary Fig.1).

Comparative abundant analysis of non-LTR and LTR TEs: In the TEs annotation process, we identified 22 different repeat families of non-LTR-RTs, viz. Ambal, Crack, CR1, CRE, Daphne, I, Jockey, Hero, L1, L2, LOA, NeSL, Nimb, Outcast, R1, R2, R4, Rex1, RTE, Tad1, Penelope and SINE which were distributed among eight legume species (S Table 1). Two species of *Arachis* were having similar types of non-LTR family distribution like Jockey, L1, L2, R1, RTE, Penelope and SINE which masked the major portion of the genome. Among all eight legume species, the repeat family of Cca was well-annotated and featured with all listed 22 non-LTR elements. Interestingly, L1 repeat element was dominant over other LINE elements across the eight legume species and proportionally Mtr (length=12,010,439 bp, 8.97% of total repeat, 2.91% of the total genome) and Lja species (length=7436648 bp, 5.97% of total repeat, 1.66% of the total genome) have the highest masked L1 repeat elements, followed by other non-LTR elements like RTE and SINE (S Table 1). SINE repeat element was abundant in all selected legume species and comparatively Gma contain the highest copy number of 6,256 (length=1,547,147 bp) followed by Mtr (n=3,798, length=601,825 bp), Adu (n=3,198, length= 2,189,063 bp), Aip (n=1,705, length=588,792 bp) and Cca (n=1300, length=146,482 bp). A very few number of SINE elements were observed in Lja (n=854, length=149,111 bp), Car (n=441, length=55,106 bp) and Van (n=190, length=19,141 bp) species. Apart from the above described repeat families some of them were species-specific and less in copy number like Ambal, Crack, Hero, Outcast, and others. Despite of difference in the genome size of the eight legume species the non-LTR repeat element showed the contrasting view in respect of copy number and size of the repeat elements. However, the analysis showed the two *Arachis* species having similar genome sizes, differ in non-LTR contain and varied in between 1.90 % (Aip) - 2.07% (Adu) of the total genome proportions of the repeat. The similar, pattern of distribution of non-LTR TEswere also reported earlier in *Solanum tuberosum* group Phureja (0.939 %) and *Solanum tuberosum* group Tuberosum (1.16 %, (Gaiero *et al*. 2019).

Like non-LTR, a total of seven LTR family, viz. BS1, BEL/Pao, Cassandra/TRIM, Caulimoviridae/Pararetrovirus, Ty1-Copia, Ty3-Gypsy, and ERV were identified among all selected species (S Table 1). Comparatively, the Ty3-Gypsy repeat element was dominant over other elements and the total proportion of masked genome coverage of Ty3-Gypsy was varied in between 3.81% (Lja) to 51.40% (Aip). Copia repeatelements of LTR was the second-largest repeat family in all legume species (S Table 1). The analysis showed LTR elements are highly species-specific and showed significant variations in copy number as well as in the masked repeat length.

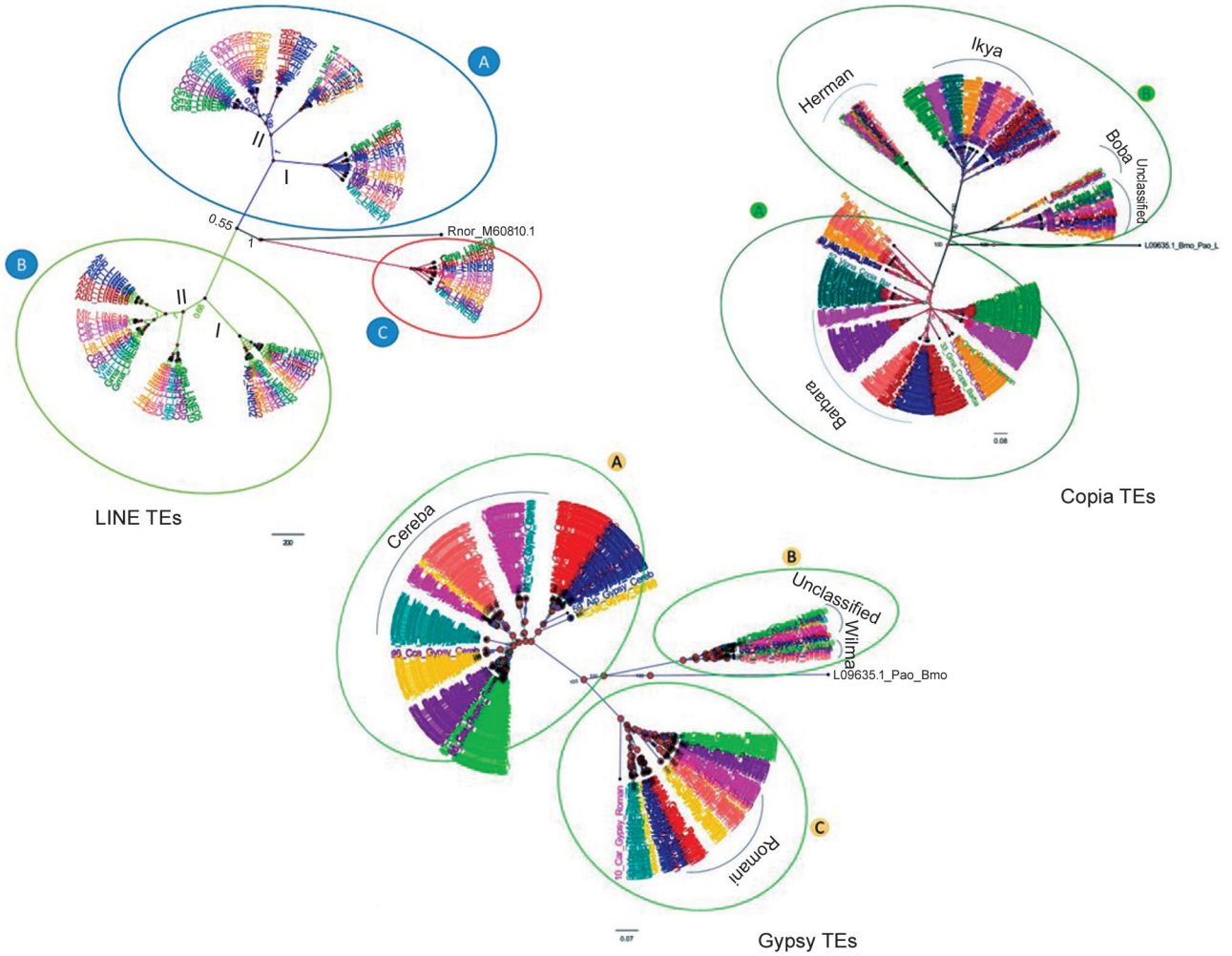

Fig 1 Phylogenetic tree of conserved Class I TEs. The consensus Bayesian phylogenetic tree showed the clustering of 112 LINEs (non-LTR), 872 Copia and 568 Gypsy (LTR-RTs) transposable elements. LINE and Gypsy elements were grouped into three major cluster while Copia was grouped into two cluster. In developed tree, lineages of each taxa color represent the individual species.

Comparative analysis of DNA TEs in genome size of legume: Total 23 different families of DNA transposons (Academ, EnSpm/CACTA, Crypton, Dada, Ginger, Harbinger, hAT, Helitron, IS3EU, ISL2EU, Kolobok, Mariner/Tc1, Merlin, Maverick, MITE, MuDR, Novosib, P, piggyBac, Sola, Transib, Zator and Zisupton) were identified and distributed across all the selected legume species. The identified DNA-TEs vary from species to species and among which Mtr containing the highest percentage of DNA TEs (6.26%) followed by Car (5.94%), Gma (5.92%), Cca (5.11%), Van (4.69%) while two species of *Arachis* contain the least percentage of DNA TEs, i.e. Aip (3.77%) and Adu (4.30%). The majority of DNA transposons of eight legume genome belong to six major super families (S Fig 1a-c). Adu and Aip species showed similar pattern of DNA TEs distribution while in other six legume species DNA TEs varied dramatically (S Table 1).Apart from the common super families some of the DNA transposable elements were species-specific for example Crypton, Dada, Ginger was present in the Car, Cca, Gma, Lja, Mtr and Van legume species only (S Table 1).

Genome sequences of all selected legume species are highly diverse in context of genome size. Our analysis showed two *Arachis* species which were having similar genome size(~2700 Mb) differ in the percentage of masked TEs by around 10% (S Fig 1) while *Lotus* and *Medicago* species comparatively smaller genome size, contain less than 32% of masked TEs. The analysis also showed a compact genome comparatively containing fewer TEs while larger genome size species proportionally masked large portion of the repeat elements (Xia *et al*. 2019).

Conserved TEs in eight legume species and Phylogenetic tree analysis: To identify the conserved TEs among eight legume species we initially mapped the individual species-specific repeat elements in reference to the Gma TEs using BLASTN program and retrieve the conserved homologous TEs. The BLAST output was parsed and filtered at 60% identity and 100 bit score value. The result showed that total 2,850 copies of TEs were conserved among all eight species among which 2,661copies of repeat elements belongs to

class I (Non-LTR-LINE:14; LTR- Copia:1,672, Gypsy:834, BS1:2, Other-LTR:6) and class II (hAT:46; Helitron:45; CACTA:14; MULE:25; other DNA TEs:3) super families while 189 repeat elements were from unclassified category. To understand the entire evolutionary process of TEs we generated a phylogenetic tree of a major family of Class I (LINE, Copia, Gypsy) and Class II (CACTA, hAT, MuDR, MULE, and Helitron) repeat elements. The detailed phylogenetic tree analysis is summarized below.

Phylogenetic tree analysis of conserved non-LTR and LTR TEs

LINE elements: LINE is widely distributed transposons in different eukaryote species. There were 14 conserved copies of LINE elements identified in eight legume species. A phylogenetic tree was developed based on 112 LINE elements (n=14×8) which was rooted with the LINE element of *Rattus norvegicus* species (accession number: M60810, length: 2018 bp). A total of 30000 trees was generated using Mrbayes program and the first 25% of the tree were discarded. The developed consensus phylogenetic tree was grouped in to three major clade A-C (Fig 1). The clade A and B contain 48 elements in each group while clade C contains 16 different LINE elements. Detail analysis of the tree showed clade A of sub-group I contain 16 andsub-group II contains 32LINE elements formed a monophyletic group. Similarly, like clade A, clade B was grouped into two subgroups I and II, which contain 32 and 16 LINE elements respectively. The result highlighted the evolution of conserved LINE elements and grouping of the lineages showed the interrelationship among the different tribe of legume TEs.

Copia elements: Among eight legume species out of 2514 conserved LTR elements, 66.50% (n=1672) were from the Copia family. A phylogenetic tree of the Copia element was developed based on conserved reverse transcriptase (RT) domain identified using HMMER program (Johnson *et al*. 2010). The analysis showed out of 1672 Copia elements, 109 copies were containing reverse transcriptase domain RVT_2, RVT_3, zf_RVT and these elements were conserved in eight legume species (n=109×8=872 Copia elements). Further, these conserved Copia elements were annotated with TREP database and out of 16 listed families four families, viz. Barbara, Boba, Hermanand Ikya were matched with the RT domain containing conserved repeat elements while some of them remained unclassified. The developed consensus Bayesian phylogenetic tree of RT Copia element was grouped into two major clades, i.e. clade A and B (Fig 1). Clade A contains 537 Barbara class of Copia repeat elements and in this clade, most of the Gma Barbara TEs (66 TEs) was clustered with Cca (50 TEs) species, similarly, 117 elements of Aipand Adu species from Dalbergieae tribe were clustered together, while some of the Barbara elements from Mtr species were grouped along with Lja, Adu and Aip species. Clade B includes a cluster of Boba, Herman,Ikya, and unclassified elements that contain 8, 31, 224, and 72 copies of repeat family respectively. All Herman and Ikya elements were grouped separately while Boba and unclassified elements were grouped together and formed a sister clade of Herman and Ikya element cluster (Fig 1).

Gypsy elements: Total 834 conserved Gypsy like LTR elements were identified in each legume species and out of which 71 elements possess RVT_1 and RVT_3 reverse transcriptase domain, which were further re-annotated with TREP database. All 71 annotated Gypsy elements were belonged to Cereba(n=49), Romani (n=18), and Wilma (n=2) category while two Gypsy elements remained unclassified. A Bayesian phylogenetic tree was developed based on 568 Gypsy elements (n=71×8) which was clustered into three major clades, called clade A, B and C. Clade A contains 392 copies of Cereba element (n=49×8), clade C contains 144 copies (n=18×8) of Romani like elements which were conserved in all the eight species, most of the lineages were clustered according to their respective tribal information like Gma-Cca and Adu-Aip. In clade B, 16 Wilma (2 from each species) and 16 unclassified Gypsy elements were clustered together and formed a monophyletic group (Fig 1). The interrelationship study of conserved elements showed the evolution of different families of Gypsy-like elements.

Phylogenetic tree analysis of conserved DNA TEs

CACTA like DNA Transposons: CACTA like DNA TEs is one of the most abundant repetitive elements distributed among the selected eight legume species. Comparative analysis showed 14 copies of CACTA like repeat elements were conserved in individual legume species (n=14×8).A Bayesian phylogenetic tree was developed (upto1053000 generations, sample freq=500, total no. of tree 2107) based on 112 CACTA DNA TEs conserved in eight legume species (S Fig 2). The consensus tree was clustered into two major clades, called clade A and B. Clade A showed the grouping of 48 elements of CACTA which was further sub-grouped into clade A-I (n=18, CACTA elements) and clade A-II (n=30, CACTA elements), similarly, clade B contains 64 different elements of CACTA which was ultimately sub-grouped into Clade B-I (23 elements) and clade B-II (41 elements). The developed tree was rooted with CACTA DNA TEs of *Daniorerio* retrieved from the repbase database. The posterior probability of the developed tree was varied between0.69 -1.0 which showed the strong clustering evidence among the CACTA elements.The cladogram showed most of the CACTA elements of Cca, Gma, and Van species from *Phaseoleae* tribe were clustered together followed by Aduand Aipspecies in clade A as well in clade B, while Lja, Mtr, and Car formed a separate clade (S Fig 3).

hAT-likeDNA Transposons: A total of 46 hAT-like DNA TEs in each legume species was conserved. For the development of Bayesian phylogenetic tree out of 46 hAT TEs 21 elements from each legume species was considered for the tree development which was having more than 500 bp nucleotide sequence in length. Total 168 hAT (n=21×8) repeat elements along with one hAT from *Daniorerio* species was considered for the phylogenetic tree development. The developed hAT tree was grouped into three major clades

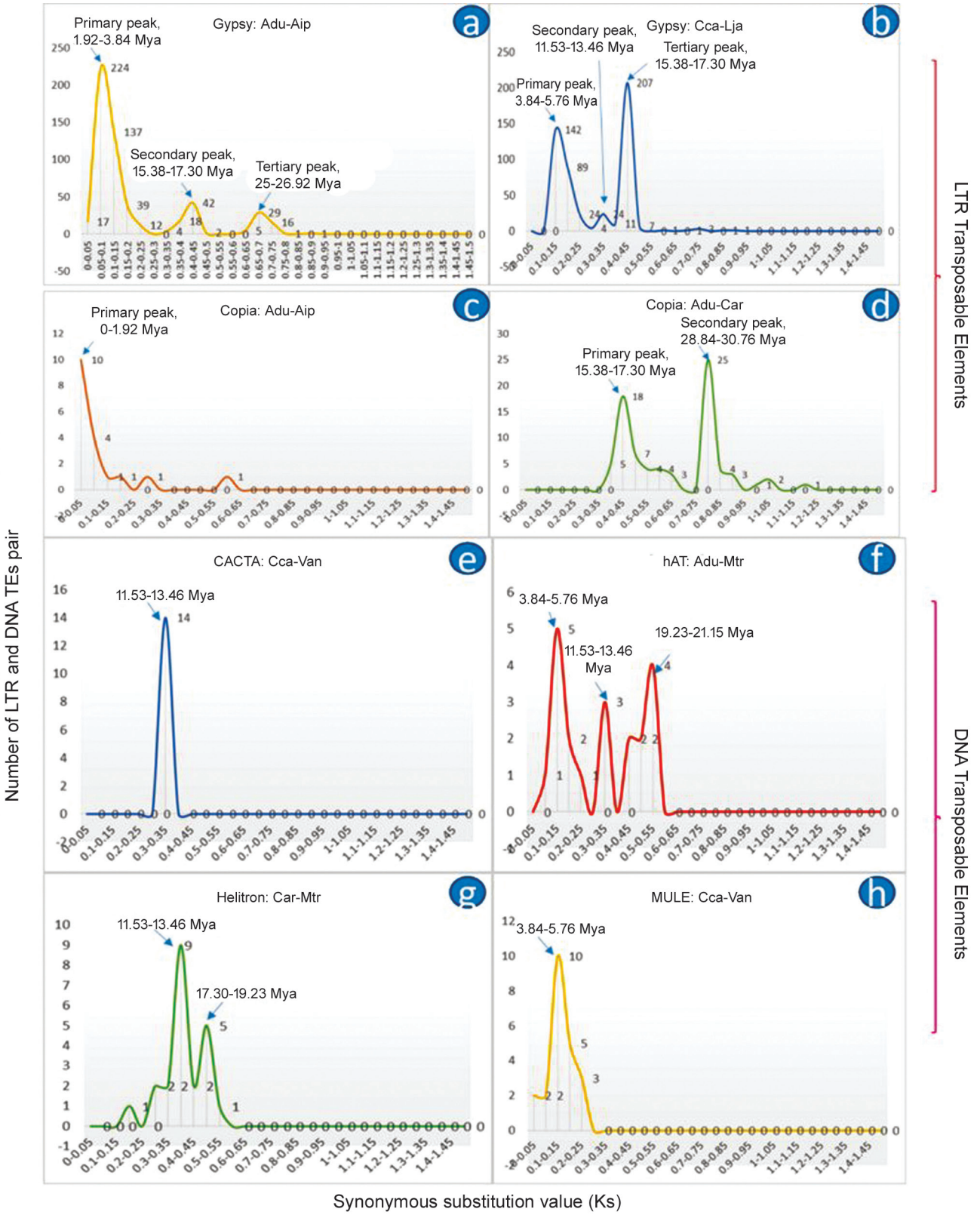

Fig 2 Graph represents the distribution plot of synonymous substitution value (Ks), and number of conserved duplicated pairs of LTR and DNA transposable elements (TEs). The Y-axis indicates the conserved duplicated TEs pair and X-axis indicates the Ks values with the intervals of 0.05. The graphs showed the primary, secondary and tertiary peak of the synonymous substitution value and their respective divergence time of Gypsy (a-b), Copia (c-d), CACTA (e), hAT(f), Helitron (g) and MULE (h) TEs

like A, B and C and each contain 56, 72 and 40 elements respectively and further, all three clades were grouped into two sub-grouped I and II (S Fig 4). The posterior probability of each node of the tree was varied between 0.50 - 1.0, which showed the convergence of the tree.

MULE like DNA Transposons: Mutator-like DNA TEs

are broadly distributed in plants, animals, and fungi species. In comparative TE analysis, we have identified 24 different MULE like DNA TEs which were conserved among all the eight legume species. A Bayesian phylogenetic tree was developed based on 192 MULE elements (n=24×8) and total of 2365 trees were generated, the consensus tree revealed the distribution and evolutionary relationship of all conserved MULE elements. The tree was grouped into two major clades A and B, which were containing 136 and 56 MULE TEs respectively. Further, clade A has grouped into four subclade I-IV, clade I contain 104 elements of MULE which was the largest group among all four sub clade while clade B grouped into I-V subgroup (S Fig 5). The developed tree revealed most of the MULE elements of the closely related species belongs to the same tribe like Gma, Cca and Van from Phaseoleae, Mtr, and Car from Trifolieae, Adu as well as Aip from Dalbergieae tribe were clustered together, which indicate the evolutionary relationship of 192 conserved MULE elements among the eight legume species.

Helitron like DNA Transposons: Total 48 copies of Helitron DNA TEs were conserved in eight legume species. Out of 48 Helitron elements, 32 copies of repeat were having more than 700 nucleotide base pair lengths which was considered for the evolutionary analysis (n=32x8). The developed Bayesian tree (no. of generation=4061500; sample freq=500; no. of tree=8124) was rooted with Helitron TEs of *Drosophila ananassae* (length=1839 bp). A consensus phylogenetic tree was grouped into the two major clades, viz. clade A (136 tips) and B (120 tips),and further clade B was sub-grouped into B1 (96 tips) as well as B2 (24 tips, S Fig 6). Each tip color of the Helitron tree represents the individual species. The posterior probability of each node was varied between 0.5 -1 which indicates the statistical support of the tree. The lineages of the tree showed the evolutionary relationship among all the conserved Helitron like DNA TEs in eight legume species.

Despite of well-known theory of dominance of species-specific TEs,we identified 2850 conserved copies of repeat elements belongs to eight legume species. The earlier analysis also showed the evidence of conserved transposable elements from non-LTR and DNA transposons elements conserved in different eukaryotic species (Gao *et al*. 2018, Karakülah and Pavlopoulou 2018). Overall, we observed all different types of repeat lineages like LTR (Gypsy, Copia), Non-LTR (LINE), DNA transposons (CACTA, hAT, helitron, MULE and others) suggesting that these conserved TEs were also present in the last common ancestor and few of them retained in all eight legume species. The phylogenetic relationships showed that lineages belong to the same tribe were grouped together and highlighted the evolution of conserved lineages in different legume species.

Inter-species divergence analysis of conserved LTR and DNA transposable element: Inter-species divergence time analysis was performed using conserved TEs in 28 pair of combination of species [n (n-1)/2, where n = no. of species]. Comparatively, Copia family was more conserved (>= 2500 bp length) in between the following combination of species: Gma, Car, Cca, Lja, Van, while Gypsy repeat elements were dominant in those combination of species in which at least one species either from the Adu or Aip (S Fig 7).For each pair of species, we developed a cumulative frequency distribution graph and analyzed the primary, secondary and tertiary peak, which reflect the duplication event occurred in TEs.

The frequency distribution graph of synonymous substitution(Ks) value showed there was three possible independent duplication events occurred in between the Adu and Aip Gypsy TEs and most of the pair of sequences fall in first peak and the corresponding synonymous substitution interval was ranged in between Ks =0.05 to 0.1. The estimated divergence time for the primary peak was1.92 to 3.84 million years ago (Mya), however, we observed the secondary as well tertiary peak for the same pair of species which was comparatively higher Ks value ranged in between 0.4 -0.45 and 0.65-0.7 respectively and both have diverged in 15.38 to 17.30 and 25 to 26.92 Mya (Fig 2a, Supplementary Fig.8). Another pair of distantly related legume species like Cca-Lja (*Phaseoleae-Lotae* tribe) which was having total 515 pair of conserved sequence (length ≥2500 bp) has formed three different peaks i.e. 0.1-0.15, 0.3 -0.35 and 0.4-0.45 and each peak contain 142, 24, and 207 pair of sequences (Fig 2b). The analysis showed the primary peak which contains 142 Gypsy elements diverged recently (3.84-5.76 Mya), secondary peak was comparatively older with divergence timeof 11.53 – 13.46 Mya while the third peak i.e. tertiary peak was oldest among all three which was diverged in between 15.38-17.30 Mya (Fig 2b).

Similarly, like Gypsy elements the distributions of Copia Ks interval plot showed the clear evidence of accumulation of conserved Copia TEs among all the 28 different possible combination of species. Some of the species pair clearly showed the primary, secondary and tertiary peaks in the frequency distribution plot, for example Aip-Cca, Cca-Lja, Car-Van, Gma-Lja (S Fig 9). Here we showed the estimated divergence time of one closely and one distantly related pair of legume species like Adu-Aip(n=38, ≥2500bp) and Adu-Car (n=154, ≥2500 bp) species. We estimated the synonymous substitution value and draw a frequency distribution graph which revealed the conserved Adu-Aip Copiasequences diverged recently 0-1.92 Mya while two distantly related legume species Adu and Car (Dalbergieae and Trifolieae tribe) formed primary (n=18, interval 0.4-0.45) as well as secondary (n=25, interval: 0.75-0.8) peak which was diverged in during 15.38-17.30 Mya and 28.84 – 30.76 Mya respectively.

Like, LTR and non-LTR TEs, we estimated the divergence time of a major group of DNA transposons. There were six major families of DNA transposons conserved among the selected legume species, however, in comparison to LTR, the frequency of conserved pair of repeat DNA transposons was low in copy number. Although, we analyzed the synonymous substitution based frequency distribution graph of CACTA, hAT, Helitron, and MULE

DNA transposons present in all 28 different pairs of species (S Fig 10-13). The cumulative frequency distribution graph of CACTA Ks value clearly showed that the species belongs to the same tribe was having only primary peak except some of the pair of legume species while inter tribe distribution plot showed the primary as well as secondary peak which clearly indicates about the occurrence of multiple event of divergence of respective DNA TEs in between the species (Fig 2e, S Fig 10), the similar trend with some exception was observed in the other DNA transposons like hAT, Helitronand MULE (S Fig 11-13). An example of distribution and divergence time analysis of all four major groups of transposons was shown in Fig 2e-h.

Transposable elements are a key component of the eukaryotic genome which play important role in the evolution of genome via a different mechanism like a rearrangement of genomic sequence, gene mutagenesis (Lisch 2013, Hirsch and Springer 2017). Genome-wide analysis of TEs in eight legume species showed a relation between the size of the genome and TEs content. In this study, we demonstrated that despite of dominance of the species-specific TEs some of the TEs were conserved across all selected eight legume species which were comparatively older and divergence events occurred in million years ago.

At the basic level of enquiry, the percent of transposable elements derived in eight legume genomes varied from 26.42-69.02%. The annotated TEs explain about the variation and accumulation of different elements in legume plants. Interestingly, high copy number of LTR and DNA transposons was conserved over non-LTR elements. Identified 2850 conserved repeat lineages revealed about the architecture, evolution and diversification of conserved repetitive elements from one to other legume species. By analyzing divergence time of homoeologous repeat families, we found that the event of divergence occurred only once in most of the legume species belonging to same tribe while inter species analysis showed the multiple event of divergence.

ACKNOWLEDGEMENTS

The financial assistance received from the Indian Council for Agricultural Research (ICAR) for Network Project on Transgenic in Crops (NPTC) and ICAR-National Professor, B P Pal Chair is gratefully acknowledged.

REFERENCES

Benson G.1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–80.

Bertioli D J. 2016. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**: 438–46.

Bao Z and Eddy S R. 2002. Automated de novo identification of repeats sequence families in sequenced genomes. *Genome Research* **12**: 1269–76.

Gaiero P. 2019. Comparative analysis of repetitive sequences among species from the potato and the tomato clades. *Annals of Botany* **123**(3): 521–32.

Gao D. 2018. Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants. *Molecular Biology and Evolution* **35**(2): 354–64.

Gill N. 2010. Dynamic oryza genomes: Repetitive DNA sequences as genome modeling agents. *Rice* **3**: 251–69.

Han Y and Wessler S R. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research* **38**(22): e199.

Hirsch C D and Springer N M. 2017. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta* **860**: 157–65.

Ibarra-Laclette. 2013. Architecture and evolution of a minute plant genome. *Nature* **498**: 94–98.

Jayaswal P K. 2019. Phylogeny of actin and tubulin gene homologs in diverse eukaryotic species. *Indian Journal of Genetics and Plant Breeding* **79**(1): 284–91.

Johnson L S. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**: 431.

Karakülah G and Pavlopoulou A. 2018. *In silico* phylogenetic analysis of hAT transposable elements in plants. *Genes* (Basel) **9**(6).

Katoh K and Standley D M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**(4): 772–80.

Librado P and Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**(11): 1451–52.

Lisch D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics* **14**(1): 49–61.

Ma J and Bennetzen J L.2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences, USA* **101**(34): 12404-10.

Mao H and Wang H. 2017. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics* **33**(5): 743–45.

Nei M and Gojobori T.1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**(5): 418–26.

Ronquist F. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**(3): 539–42.

Sato S. 2008. Genome structure of the legume, *Lotus japonicus*. *DNA Res*earch. **15**(4): 227–39.

Sarkar D. 2017. The draft genome of *Corchorus olitorius* cv. JRO-524 (Navin). *Genomics Data* **12**: 151–54.

Schmutz J. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**(7278): 178–83.

Schnable P S. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–15.

Singh NK. 2012. The first draft of the pigeonpea genome sequence. *Journal of Plant Biochemistry and Biotechnology* **21**: 98–112.

Varshney R K. 2013. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nature Biotechnology **31**(3): 240–46.

Xia E. 2019. The tea plant reference genome and improved gene annotation using long-read and paired-end sequencing data. *Scientific Data* **6**(1): 122.

Yang K. 2015. Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Sciences*, USA **112**(43): 13213–18.

Young N D. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**(7378): 520–24.