In-Silico Identification of LTR type Retrotransposons and Their Transcriptional Activities in Solanum Tuberosum

Chandra Bhan Yadav, Himansu Narayan Singh

Abstract - Eukaryotes genomes contains large amount of mobile genetic elements. More than two million EST (expressed sequence tags) sequences have been sequenced from potato crop plant and this amount of ESTs allowed us to analyze the transcriptional activity of the potato transposable elements. We predicted the full length LTR from potato genomic database using LTR finder software. Maximum number of full length Gypsy type LTRs were present on chromosome 03 (197) and Copia type retrotransposons on chromosome number 01 (172). We have also investigated the transcriptional activities of LTR type retrotransposons in different potato organs based on the systematic search of more than two million expressed sequence tags. At least 0.86% potato ESTs show sequence similarity with LTR type retrotransposons. According to these data, the patterns of expression of each LTRs (Gypsy & Copia) is variable among various tissue specific EST libraries. In general, transcriptional activity of the Gypsy-like retrotransposons is higher compared to Copia type. Transcriptional activity of several transposable elements is especially high in Flower, Callus and root tissues. The use of powerful high-throughput sequencing technologies allowed us to elucidate the transcriptional activation in various cells of potato. In this study, we observed that Gypsy and Copia like retrotransposons have a considerable transcriptional activity in some tissues which indicate that the transposition is more frequent in various tissues specific EST libraries.

Index terms: Retrotransposons, LTR, Solanum tuberosum, Gypsy, Copia

I. INTRODUCTION

Besides genes, there are other substructures present within the genome which could play important role in biological regulations called repeat. These sequences are resided within the genome and these genomic patterns are the repeat element and latter on named as transposable elements (TEs). Barbara McClintock first identified the existence of repetitive elements in the maize genome which is known as "jumping genes". It has been identified in many eukaryotes and almost all plant species. TEs are small segments of the chromosome that can 'jump' from one location to another location called TE transposition. On the basis of their mode of replication and transposition the TEs are separated into two categories, class I TEs, or retrotransposons produce RNA intermediates that, by the fraction action of reverse transcriptase (RT), are copied in to DNA and then inserted into new locations within the genome, while other TEs class II TEs, or DNA transposons move directly by a "cut and paste" mechanism [1].

Manuscript received on March, 2013.

Dr. Chandra Bhan Yadav (Ph.D.) Division of Plant Pathology, Indian Agricultural Research Institute, New Delhi-110012, India.

Mr. Himansu Narayan Singh Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, India.

Class I elements are also called retrotransposons, or retroelements, and comprise two main types: (1) LTR retrotransposons, flanked by long terminal repeats (LTRs), and (2) non-LTR elements [such as Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs)]. Retrotransposons are the most abundant mobile elements in plant genomes [2], as the replicative mode of retroelement transposition enables the LTR retrotransposon to accrue in high copy number. Indeed, in some grasses, LTR retrotransposons represent up to 90% of the genome [2 & 3]. They constitute more than 50% of the maize genome [4 & 5], 14% of the Arabidopsis genome (The Arabidopsis Genome Initiative 2000) and up to 90% of the wheat genome [6].

TEs make up a significant proportion of the genomes of higher plants and their activation can have a range of effects, including genome evolution structural and functional alterations in gene expression, gene deletion and insertion [7]. Therefore, inactivation of TEs can be crucial for the survival of the host organisms. Because TEs encode enzymes required for their own maintenance and jumping, suppressing these gene products at a transcriptional or a post-transcriptional level would be the most effective ways to inactivate TEs. TEs are enriched in the centromeric region, which is highly methylated and packed into heterochromatin in the genome sequence of Arabidopsis (The Arabidopsis Genome Initiative 2000). Maize genome has also been shown that the most of the TEs are restricted to the methylated heterochromatin region [5 & 8]. Carlos M Vicient (2010) stated that TEs and other repetitive sequences are the primary targets of DNA methylation which play an important role in structural and functional alteration in plants [9].

Publically available EST and genomic database are useful tools to identify the TE/transposable elements in plants and their validation for an association with specific trait because it transposes themselves into new location of the genome through mRNA. The use of powerful high-throughput sequencing methodologies allowed us to elucidate the extent and character of repetitive element transcription in plant cells. In plants, retrotransposons appear to be the most abundant and the most transcriptionally active (Arabidopsis Gnome Initiative 2000; The Rice Chromosome 10 Sequencing Consortium 2003). RNA transcript from Ty 1-copia-type and Ty3-gypsy-type retro elements have a poly tail A and, therefore, have the potential to be included in the mRNA populations. A search for transposable elements in the ESTs databases has been attempted by Macas J. et al [10] in *Pisum sativum* showed that at least some elements are transcribed, most probably due to their association with genic regions.

Echenique et al [11] were also use of ESTs database to estimate the presence of TEs in the RNA population such as retrotransposons (Ty 1-copia-type and Ty3-gypsy-type retro elements) within Triticeae EST databases provides an indirect estimation of the patterns of transcriptional activity of these repetitive elements and is important to improve the annotation of genomic sequences used to search these EST databases. Carlos M Vicient (2010) analyze the transcriptional activity of the maize transposable elements based on EST databases and found 1.5% maize ESTs show sequence similarity with transposable elements.

Retrotransposons have the potential to alter the genomic landscape and change gene expression when they amplify or integrate into new sites in the host genome. Instead of negative roles of genome regulation, retrotransposons are also play a important role in the evolution of genes and genomes at many aspects by translocations, gene and segmental duplications leading to gene family expansions that may further undergo selection and diversification [7]. Translocation of the retrotransposons can also cause alteration in gene expression patterns since many of them contain elements for transcriptional regulation. Another important factor epigenetic factor, by which activity of transposable elements (TEs) can influence the regulation of genes; though, this regulation is confined to the genes, promoters, and enhancers that neighbor the TE. As we know TEs are major producers of small RNAs that act to maintain the TE in an epigenetically silenced state. In plants, and perhaps in animals, heterochromatin modifications are targeted by the activity of small RNAs. Two major types of small silencing RNAs have been described in plants: MicroRNAs (miRNAs), and short interfering RNAs (siRNAs). miRNAs are small non-coding RNA molecules which regulates the expression of genes at the post-transcriptional level [12]. Plant miRNA biogenesis is the complex process in which it produces a long primary transcript (pri-miRNA) that then undergoes two cleavage events, the first giving a precursor (pre-miRNA) that folds into a hairpin structure, the second extracting the mature 19–24 nt miRNA from the stem of the hairpin reviewed by Wei et al. (2009) [13]. These mature miRNA were then incorporated into the RNA-induced silencing complex (RISC), where it directs transcriptional repression of cognate mRNA targets [14]. Endogenous siRNAs have also been extensively characterized in *Arabidopsis thaliana*, where they are processed by Dicer proteins from long, and perfectly double stranded RNA (dsRNA) precursors. The endogenous dsRNA precursors are most often produced by RNA-dependent RNA polymerases (RDRs). The majority of expressed small RNAs in *A. thaliana* depend on the activity of two RDR proteins, implying that siRNA production from RDR-dependent dsRNA precursors is rampant in plants. This retrotransposons mediated epigenetic mechanism has been recognized to be ubiquitous in plants and important in an increasing variety of biological processes, including development, their own biogenesis, and biotic and abiotic stress responses [15]. TE/Retrotransposons also play important role in miRNA mediated silencing mechanism in the regulation of male and female gametophyte development [9].

Potato (*Solanum tuberosum*) is the most important food crop in all over world, the average yield of potatoes around the world is far below its physiological potential of 120 tons/ha [16]. The potato genome and accompanying gene regulatory elements are powerful resources for understanding this complex system and advancing in mechanism of molecular regulation efforts in this crop. Increasing amounts of genomic sequence as well transcriptomic data is available for potato crop species to identify the different classes of retrotransposons. In view of this, an extensive study on computationally identification full length retrotransposons and their transcriptional activation among various tissues specific libraries have been carried out in potato. In doing so we generate a catalogue of LTR type retrotransposons which are transcriptinally active in leaf tissue in response to various biotic and abiotic stress.

II. MATERIALS AND METHODS

1. Data resources

Potato genomic sequence build ITAG2.40 were downloaded from Sol Genomics Network project (http://solanaceae.plantbiology.msu.edu/data/PGSC_DM_v3_2.1.11_pseudomolecule .zip). 235,967 EST sequences are available in NCBI database, were used in this study to identify the transtriptionally active retrotransposons and these sequences were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=Solanum_tuberosum#).

2. Prediction of full length LTR (Long Terminal Repeats) retrotransposons

An efficient algorithm for de novo identification of full length LTR retrotransposons was developed by Zhao and Hao [17]. They implemented this algorithm in LTR-Finder software which includes a robust parameter set encompassing both structural constraints and quality controls. The structural feature of the full-length LTR element was also confirmed by the LTR-FINDER program (http://tlife.fudan.edu.cn/ltr_finder/). LTR-FINDER uses a suffixarray- based algorithm to construct all exact matching pairs, which are extended to long highly similar pairs [17]. Smith–Waterman algorithm has been used to demarcate the LTR boundaries. We choose ‘‘Use ps_scan to predict IN (core), IN(c-term) and RH,’’ and selected the database of ‘‘*Arabidopsis thaliana* (2004)’’ to predict PBS. We also choose to check the putative LTRs for the conserved TG-CA termini and choose ‘‘at least two of TSR, PBS and PPT’’ to ensure the specificity (Fig. 1). The following parameters were used: LTR sequence length from 100 to 2,000 bp and maximum distance between LTRs of 10,000 bp. The sequences between two putative LTRs were subsequently analysed by BLASTX and BLASTN searches (E value threshold, 10−5) against public non-redundant databases at GenBank and against REPBASE [18]. All LTR sequences were further verified with RepeatMasker (developed by A.F.A. Smit, R. Hubley, and P. Green; http://www.repeatmasker.org/) by masking the repetitive region against Repdatabase [19].

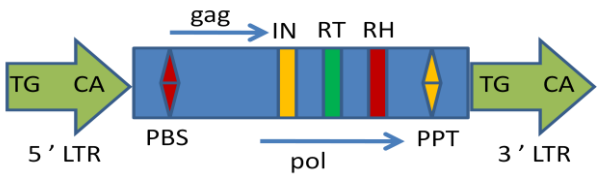

Fig. 1 The general structure of a full-length LTR retrotransposons. Based on this features, LTR prediction software has been able to detect the specific regions in the query sequence such as Long terminal repeats (LTR), PBS (primer binding sites) integrase (IN), reverse transcriptase (RT), and RNAase-H, PPT (polypurine tracts) [17].

3. Comparison between putative full length LTRs derived from genomic sequences and transcriptionally active LTR derived from ESTs

To compare the full length putative LTR with EST database, we have performed BLAST analysis locally by using the LTR sequences of each putative full-length LTR as queries against tissue specific EST database. The occurrence of sequences with at least 80% similarity with 80% coverage to putative LTRs in EST databases of *Solanum tuberosum* was searched by BLASTN algorithm (E value threshold, 10^{-5}).

III. RESULTS AND DISCUSSION

In this study, we identified the full length LTR type retrotransposons and their distribution among potato chromosome. We have also investigated the transcriptional activities of retrotransposons in various potato organs.

1. Identification and classification of full length retrotransposons

Intact full length LTR retrotransposons have been identified from the entire *Solanum tuberosum* genome consist of 12 chromosome ranges from 43 MB to 81 MB (Table 1). The *Arabidopsis* LTR database was used for LTR prediction with LTR Finder program. The full length LTRs was defined as one that contains two LTRs (5' & 3' LTR) and PBS & PPT which were of 20 bp and 15 bp in length respectively (Fig. 1). This software analysed the LTRs and internal domains, which could be divided into 10 domains: (1) TG in 5' end of 5' LTR, (2) CA in 3'end of 5' LTR, (3) TG in 5' end of 3' LTR, (4) CA in 3' end of 3' LTR, (5) TSR, (6) PBS, (7) PPT, (8) RT, (9) IN, (10) RH. A detail of alignment of a predicted LTR is given in Fig. 2.

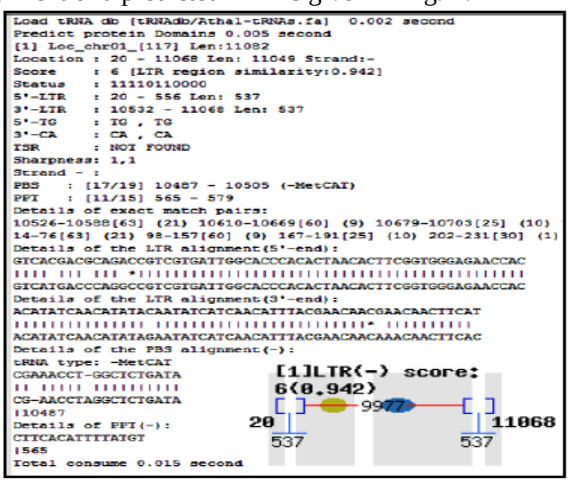

Fig. 2 Prediction of full length LTR retrotransposons (11.08 Kb) from genomic sequence (Chromosome 1) of potato using LTR finder software. The size of 5' LTR and 3' is 537 bp in length.

The promoter binding site (PBS) is a region complementary to tRNA 3' terminal sequences used during reverse transcription at a later stage of the retrotransposons life cycle and the 3' end of the retro element contains a purine rich sequence called poly-purine tract (PPT). On the whole, 3695 intact full length LTRs were predicted with LTRstruct/par and further validated with LTR finder software.

These full length LTRs were further analysed for coding region. The sequence between two LTRs are coding region ie. gag region which is a gene that code capsid like protein; pol region is a gene coding for protease, integrase and reverse transcriptase enzymes; the env region contains the gene coding for envelop protein. Some randomly selected LTRs were also confirmed with BLASTX analysis against nonredandant database of NCBI using BLAST2GO software. For this, the sequences from coding region of retrotransposons were extracted using perl script. In our analysis, only LTR type retrotransposons (Copia & Gypsi) were considered for further classification. The size of full length LTR was ranges from 1.3 Kb to 15.6 Kb with a mean length of 8970.42Kb. Out of 3695, 2036 (55%) retrotransposons were confirmed as full length intact LTR type retrotransposons. Full length Retrotransposons includes 5' LTR and 3' LTR flanking the internal region of the LTR. Both the LTRs started with TG and ended with CA. Moreover, putative LTR with two or one of the above described some typical LTR features such as PPT, PBS and TSD were identified. 3'LTR and 5'LTR sequences of all LTRs were extracted using perl script. The recorded putative LTR had a mean length of 643.72 bp, but large length variability was observed which ranges from 0.01 Kb to 2.7 Kb. The full-length LTRs were further validated for the coding region whether retro encoding enzymes are present. For this, internal region of the LTRs were compared with the repeat database by blast analysis (e value threshold, 10). LTRs were further classified as belonging to Gypsy and Copia type superfamilies according to blast analysis of their internal region (i.e., between 5' LTR & 3' LTR) in comparisons with *Arabidopsis* repeat databases.

Maximum number of intact LTRs were present at chromosome number 1 where as minimum was on chromosome 12. Table 1 showed the number of full-length Copia type & Gypsy type identified in the potato genome and their distribution among potato chromosomes.

Table: 1 Number of full-length LTR-retrotransposons on potato chromosome

Potato Chrom.	Types of retrotros		Total	Chrom. length (bp)	% of LTR
	Copia	Gypsy			
Chr 01	172	168	340	81,482,218	20
Chr 02	75	115	190	47,066,339	11
Chr 03	51	197	248	47,880,312	15
Chr 04	56	137	193	64,339,883	11
Chr 05	48	146	194	47,045,015	12

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Chr 06	30	55	85	54,975,350	5
Chr 07	17	16	33	53,427,889	2
Chr 08	22	93	115	43,648,432	7
Chr 09	29	64	93	53,640,104	6
Chr 10	19	61	80	52,313,507	5
Chr 11	32	57	89	42,253,929	5
Chr 12	17	7	24	59,100,875	1

Other transposable elements such as non LTR (LINE & SINE) and DNA transposons were excluded after blast analysis with repeat database and shortlisted manually in excel sheet. In this we found that full-length LTR elements (454 LTRs), in some cases, showed the presence of coding sequences with similarity to non-LTR retrotransposons (35 LINE & 22 SINE elements), to DNA transposons (26 elements), or to helitrons (16 elements). These elements possibly originated by insertion of such sequences in previously existing LTR sequence.

2. *Representation of retrotransposons into potato EST databases and their transcriptional activity*

The transcriptional activity of intact retrotransposons has been estimated using the number of ESTs in a large potato transcript database. 235,967 EST sequences derived from various tissues (~89 librries) are deposited in the NCBI EST database (st-dbEST). These potato ESTs were categorised according to library names (Auxillary buds, Callus, Flower, Leaf, Mixed tissue, Root, Sprout, Stolon and Tuber cells). Such a large amount of sequences provides an opportunity to identify the transcriptionally active retrotransposons in this species. We used the full length retrotransposons as query against the potato EST database using BLASTN algorgthm. 0.86% of the total potato ESTs (2036) out of 235,967) showed significant sequence homology with retrotrasnposons. Considering the different retrotransposons classes separately, the average number of Retros-ESTs obtained for Gypsi-like LTR retrotransposons is near about two folds higher than Copia-like LTR retrotransposons. Relatively high numbers of transcriptionally active retrotransposons have been observed in leaf specific tissues (Fig. 3).

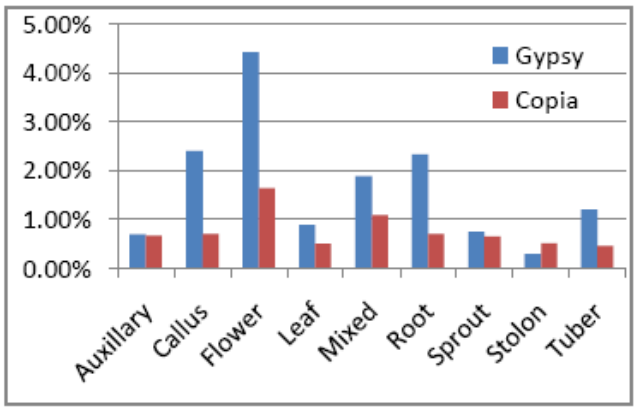

Fig. 3 Copia type & Gypsi type LTRs which are transcript -ionally active in various tissues.

There are several reports are available regarding the transposition and transcriptional activities of retrotransposons using EST database. The retrotransposons are transposes via mRNA and reside to the new location of the genome which alters the structural and functional activity of the genome [9]. The transposition frequency in the host genome is regulated by the expression of retrotransposons in plants and in other eukaryotic organisms. The evolution of control mechanisms for the transcription and transposition of retrotransposons in the host genome may be crucial to minimize their possible deleterious effects on the host [7]. The transposition activity of retrotransposons is known to vary in maize and barley described by SanMiguel et al. [5] & Gribbon et al. [20].

3. *Presence of ESTs similar to LTR and their corresponding full length LTR in leaf specific libraries*

Representation of transcriptionally active retrotransposons in leaf tissue (obtained from healthy, biotic and abiotic stressed leaf). This is interesting because this organ can potentially play impotant role in photosynthesis, allowing a possible event for carbohydrate fixation and general metabolism. However, "Leaf" group is composed by sequences derived from more than 20 cDNA libraries obtained at various conditions such as healthy leaf, leaf treated with Abiotic and Biotic factors. So, we decided to examine in more detail the origin of the TE-ESTs and their co-relation with full length LTRs (Fig. 4).

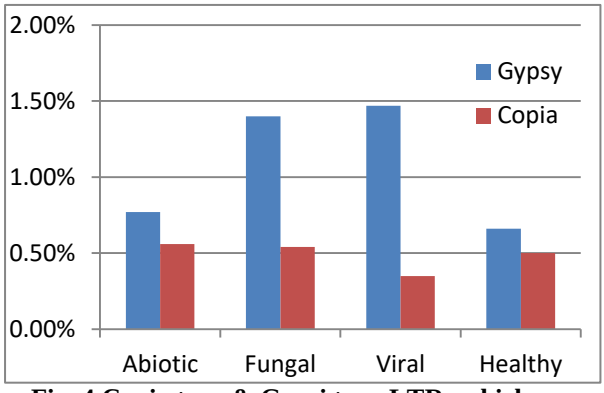

Fig. 4 Copia type & Gypsi type LTRs which are trancriptionally active in leaf specific tissues (Abiotic stressed, Fungus infected Virus infected & Healthy leaf).

We observed the distributions of the TE-ESTs which were originated in the leaf library derived from biotic and abiotic stress treated leaf (Fig. 4). Transcriptionally active retrotransposons are higher in the biotic and abiotic stress treated leaf category as compared to healthy leaf. Gypsy–like retrotransposons were maximum in the leaf specific EST library.

Transposition of retrotransoposons (transcriptionally active retrotransoposons) in the genome may leads to accumulation of mutations and become transpositionally inactive. However, even partial or rearranged TE copies may retain their capacity to initiate transcription. Cells have active mechanisms to protect their genome integrity against TE activity including transcriptional silencing [21] and short-interfering RNAs (siRNAs) [22]. Under certain circumstances some TEs can escape this cell control and transcribe and, sometimes, transpose [23]. For example, different TE families are transcribed in response to biotic or abiotic stresses or in cell culture [24-26]. In addition to these "stress response" transcription, increasing data demonstrate that some TEs may have at least low transcriptional activities under normal circumstances in plant life.

Expression and transposition of the retrotransposons are regulated by interacting with miRNA mediated silencing or by epigenetic methylation mechanism as a study on maize showed that expression and transposition of the Suppressor-mutator (Spm) transposon of maize is controlled by interacting epigenetic and autoregulatory mechanisms. miRNA derived from plant retro element and their regulatory role were reviewed by Jones-Rhoades et al [12].

IV. CONCLUSION

Genomic and EST Databases has become an essential tool for research in the vegetable crop plants. A well characterized EST database of potato would provide a framework for other annotations and biological features that are derived from the genes. High-throughput sequencing, bioinformatic, and functional genomic methods, genome sequences can be annotated with gene models depicting the exon, intron, and untranslated region structure of genes, functional descriptions of genes, numerous alignment results, promoter annotations, protein interactions, expression data allelic variation, genetic maps and a new adventives area of the study transposable elements.

REFERENCE

1. Roderick et al., "Retrotransposon Sequence Variation in Four Asexual Plant Species". J. Mol. Evol. 62:375–387, 2006
2. Feschotte et al., "Plant transposable elements: Where genetics meets genomics". Nat. Rev. Genet. 3:329–341, 2002.
3. Bennetzen J. E. and Kellogg A. "Do plants have a one-way ticket to genomic obesity?" Plant Cell 9:1509-1514, 1997.
4. Meyers et al., "Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome". Genome Res. 11:1660–1676, 2001.
5. SanMiguel et al., "The paleontology of intergene retrotransposons of maize". Nat. Genet. 20:43–45, 1998.
6. Flavell et al., "Retrotransposonbased insertion polymorphisms (RBIP) for high throughput marker analysis". Plant J. 16:643–650, 1998.
7. Kumar A., and Bennetzen J.L. "Plant retrotransposons". Annu. Rev. Genet. 33:479–532, 1999.
8. Rabinowicz et al., "Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome". Nat. Genet. 23:305-308, 1999.
9. Vianey et al., "Control of female gamete formation by a small RNA pathway in *Arabidopsis"*. Nature 464:628–632, 2010.
10. Macas et al., **"**Zaba: a novel miniature transposable element present in genomes of legume plants**"**. Mol. Genet. Genomics 269:624-31, 2003.
11. Echenique et al., "Frequencies of Ty1-copia and Ty3- gypsy retroelements within the Triticeae EST databases". Theor. Appl. Genet. 104:840–844, 2002.
12. Jones-Rhoades et al., "MicroRNAS and their regulatory roles in plants". Annu Rev Plant Biol 57:19–53, 2006.
13. Wei et al., "Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust". Genome Biol 10: R6, 2009.
14. Bartel et al. "MicroRNAs: Genomics, biogenesis, mechanism, and function". Cell 116:281–297, 2004.
15. Khraiwesh et al., "Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants". Biochim Biophys Acta. 1819(2):137-48, 2011.
16. Massa1 et al., "The Transcriptome of the Reference Potato Genome Solanum tuberosum Group Phureja Clone DM1-3 516R44". PLoS One 6:10 | e26801, 2011.
17. Zhao and Wang **"**LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**"**. Nucleic Acids Res. 35:W265-W268**,** 2007.
18. Jurka et al., "Repbase Update, a database of eukaryotic repetitive elements". Cytogenet. Genome Res. 110(1-4): 462-467, 2005.
19. Zhou and Ying Xu "RepPop: a database for repetitive elements in *Populus trichocarpa"* BMC Genomics 1471-2164-10-14, 2009.
20. Gribbon et al., "Phylogeny and transpositional activity of Ty1- copia group retrotransposons in cereal genomes". Mol. Gen. Genet. 261:883–891, 1999.
21. Tanurdzic et al., "Epigenomic consequences of immortalized plant cell suspension culture". PLoS Biol. 6:2880-2895, 2009.
22. Kasschau et al., "Genome-wide profiling and analysis of Arabidopsis siRNAs". PLoS Biol. 5:e57, 2007.
23. Picault et al., "Identification of an active LTR retrotransposon in rice". Plant J. 58:754-765, 2009.
24. Pouteau et al., "Specific expression of the tobacco Tnt1 retrotransposon in protoplasts". EMBO J. 10:1911-1918, 1991.
25. Hirochika "Activation of tobacco retrotransposons during tissue culture". EMBO J. 12:2521-2528, 1993.
26. Mhiri et al., "The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress". Plant Mol. Biol. 33:257-266, 1997.

AUTHORS PROFILE

Dr. Chandra Bhan Yadav received his B.Sc. degree from Awadh University, Faizabad (2000). He obtained his M.Sc. (Botany) from Purvanchal University, Jaunpur (2002). He has completed his M.Phil. (Botany) degree from Department of Botany, University of Delhi (2005). He was awarded Ph. D. degree in Botany from University of Delhi in 2012. Currently he is working as Postdoctoral Fellow at Division of Plant Pathology, Indian Agricultural Research Institute, New Delhi, India. He has published four papers in International journals (Molecular Breeding, African Journal of Biotechnology, Biologia Plantarun and *In-vitro* Cellular and Developmental Biology-Plant). He has also presented more than six posters as well as oral presentations in National and International conferences. His area of interest is large scale genome analysis and their use for dissecting the complex trait during plant developments.

Mr. Himanshu N Singh, received his M.Tech. degree in Bioinformatics stream from Jamia Hamdard, New Delhi in 2011. Currently he is working as Research Assistant at Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, India. His area of interest is large scale genome analysis using computational approach.