Mobile Genetic Elements 4:5, 1–5; October 1, 2014; Published with license by Taylor & Francis Group, LLC

Mining hidden polymorphic sequence motifs from divergent plant *helitrons*

Wenwei Xiong and Chunguang Du*
Department of Biology and Molecular Biology; Montclair State University; Montclair, NJ USA

As a major driving force of genome evolution, transposons have been deviating from their original connotation as "junk" DNA ever since their important roles were revealed. The recently discovered *Helitron* transposons have been investigated in diverse eukaryotic genomes because of their remarkable gene-capture ability and other features that are crucial to our current understanding of genome dynamics. *Helitrons* are not canonical transposons in that they do not end in inverted repeats or create target site duplications, which makes them difficult to identify. Previous methods mainly rely on sequence alignment of conserved *Helitron* termini or manual curation. The abundance of *Helitrons* in genomes is still underestimated. We developed an automated and generalized tool, HelitronScanner, that identified a plethora of divergent *Helitrons* in many plant genomes. A local combinational variable approach as the key component of HelitronScanner offers a more granular representation of conserved nucleotide combinations and therefore is more sensitive in finding divergent *Helitrons*. This commentary provides an in-depth view of the local combinational variable approach and its association with *Helitron* sequence patterns. Analysis of *Helitron* terminal sequences shows that the local combinational variable approach is an efficacious representation of nucleotide patterns imperceptible at a full-sequence level.

Keywords: algorithm, bioinformatic analysis, *Helitron*, local combinational variable, sequence pattern

Abbreviations: LCV, local combinational variable; PSSM, position-specific scoring matrix

*Correspondence to: Chunguang Du; Email: duc@mail.montclair.edu

Submitted: 09/07/2014

Revised: 09/24/2014
Accepted: 09/29/2014

http://dx.doi.org/10.4161/21592543.2014.971635

Transposable elements jump around and reshape genomes through the action of transposases either encoded by themselves or other transposons from the same family. Transposons in one family share the transposase and transposition mechanism and homologous terminal/subterminal sequences. As a special kind of transposon, *Helitrons* have been widely studied in a broad range of eukaryotes because of their remarkable ability to capture genes and regulatory elements.[1-5] *Helitrons* presumably transpose by a rolling-circle mechanism because putative autonomous *Helitrons* encode proteins containing 3 conserved functional motifs that are known to be involved in bacterial and phage rolling-circle replication.[6] However, unlike other DNA transposons, *Helitrons* do not possess terminal inverted repeats or create target site duplications, which likely delayed their discovery and hindered subsequent large-scale automated annotation. Even though *Helitrons* are reported broadly in diverse genomes, the number of *Helitrons* is probably still underestimated due to their lack of canonical transposon structures.[7]

Methods to identify *Helitrons* are based on homology of a RepHel protein and terminal sequences. *Helitrons* are deemed putatively autonomous if they encode intact RepHel proteins, and non-autonomous if they lack the transposase. Autonomous ones are scarce, so automated *Helitron* identification tools mainly focus on homology of short sequences at the *Helitron* termini. Our previous HelitronFinder tool looks for *Helitron* hallmarks, including AT dinucleotide insertion site, 5′-TC, CTAG-3′, and a conserved 16- to 20-bp palindromic structure located 10–15 bp away from the 3′ termini.[8] HelSearch, another structure-based tool, first detects 3′ hairpins, retains those with multiple copies in the host genome, and manually extends toward 5′ ends to determine *Helitron* 5′ boundaries.[9] HelitronFinder

works optimally with maize but is hard to extend to other species, while HelSearch does not appear to have species limitations but requires manual inspection to identity 5′ ends. Both methods identified approximately 3 thousand *Helitrons* in maize with a 95% overlap, but neither was able to detect a highly abundant ~1-kb *Helitron* named *Cornucopious,* with thousands of copies in maize genome, that had been identified earlier from a vertical comparison of allelic haplotypes.[10,11] This failure was caused by *Cornucopious* having more divergent 3′ ends than previously known *Helitrons*. Another work combining BLAST search and hidden Markov models identified many *Helitrons* in the rice genome, but seemed not applicable to maize.[12] A model-based method searched for new *Helitron* termini by BLASTing known *Helitron* terminal consensus sequences and identified a number of *Helitrons* in *Arabidopsis thaliana*.[13] This method brought more flexibility than searching for whole homologous *Helitrons*, but was still limited to *Helitron* termini that are highly similar to known ones. There are other ad hoc methods for *Helitron* identification in diverse genomes. They rely heavily on BLAST and manual annotation. De novo transposon identification algorithms like RECON[14] and RepeatScout[15] also depend on pair-wise genome BLAST and high sequence similarity among copies of one transposon family in the host genome. Divergent

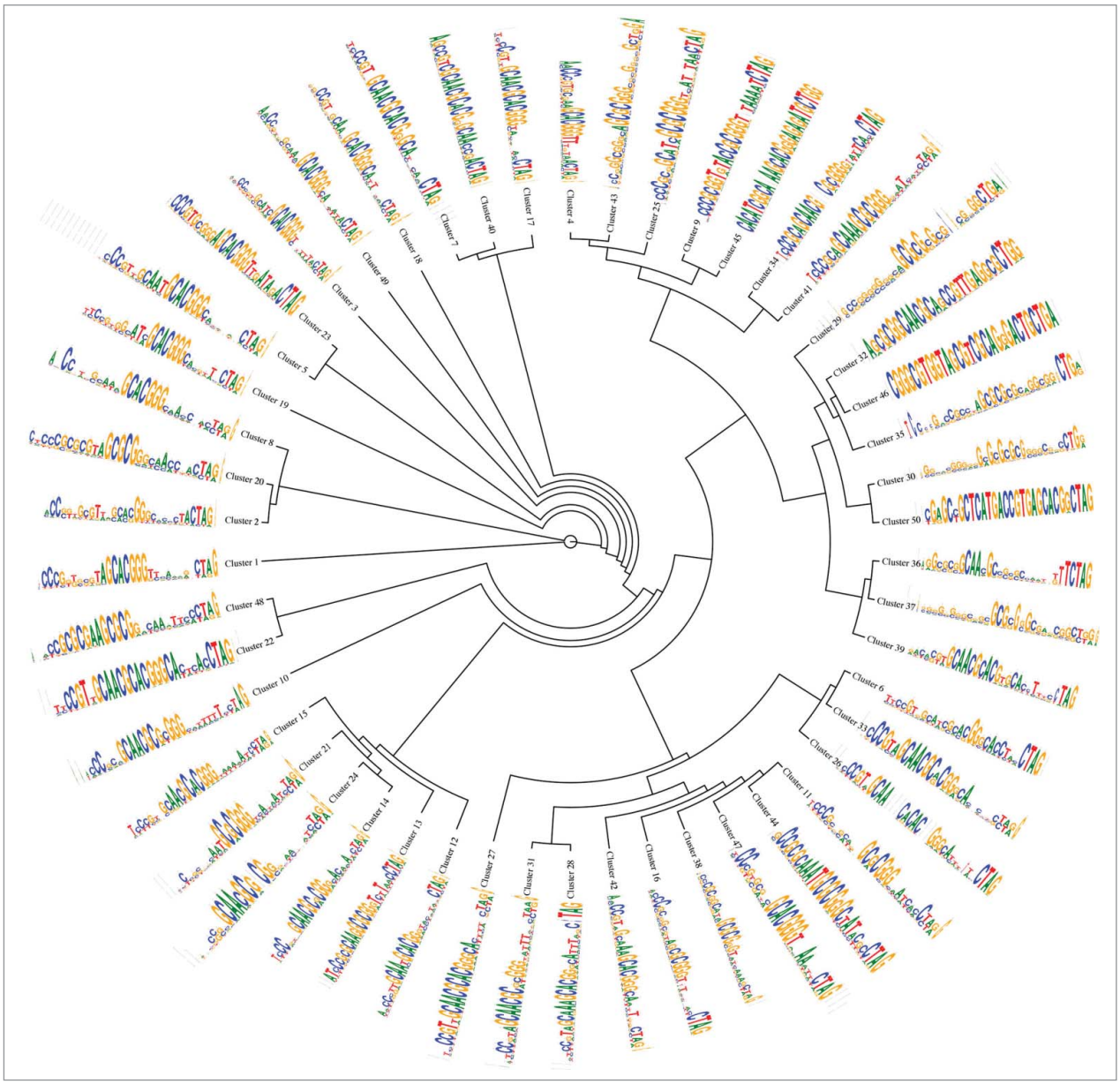

Figure 1. Divergent *Helitron* termini represented by LCVs. HelitronScanner identified 107,367 putative *Helitrons* from 39 plant genomes.[19] Their top 50 clusters of 30-bp 3′-end sequences include 39,554 *Helitrons*. Similarities of the clusters are shown by the inner dendrogram. Sequence logos of the clusters are shown in the outer ring.

Helitrons would be missed by these de novo methods because they do not align well.

BLAST has been the most valuable weapon in the arsenal of bioinformatic analysis as a result of its power in finding sequence homology at given thresholds of statistical significance.[16] However, there is no clear division in the spectrum of sequence similarity from being completely identical to not even remotely related. Divergent sequences that evolved from one ancient ancestor may appear totally unrelated in BLAST output or manual inspection, yet they behave as one functional family or bear common features when they function. In other words, although homologous sequences always lead to common functions, function resemblance does not guarantee global sequence similarity,[17] at least not in an apparent manner. The difficulties in functional bioinformatics studies are, by and large, attributed to this inconsistency. Hurdles in previous *Helitron* identification also fall into this category because of the divergent nature of *Helitrons* and the lack of common transposon features like terminal inverted repeats and target site duplication. Position-specific scoring matrix (PSSM) is a more flexible representation of sequence patterns than consensus sequences.[18] Although successfully applied in various DNA binding site prediction studies, PSSM requires a target region from a group of well-aligned sequences that are functionally related. Creating such a sequence profile for *Helitrons* would be difficult considering our current insufficient understanding of the *Helitron* transposition mechanism.

In order to automate and generalize *Helitron* identification in various species, we developed a tool, HelitronScanner,[19] using a local combination variable (LCV) approach.[20] LCVs were first extracted and refined from a training set compiled from previously published *Helitrons*. HelitronScanner searches for sequence patterns that match these LCVs. Significance of matches is measured with scores separately for the 5′ and 3′ ends of putative *Helitrons* and filtered with empirical thresholds. HelitronScanner identified a plethora of diverse *Helitrons* in many plant genomes, including those missed by previous methods, and thus should pave the way to a better understanding of the transposition mechanism of *Helitrons* and their evolutionary contribution to genome dynamics. The local combinational variable approach constitutes the key component

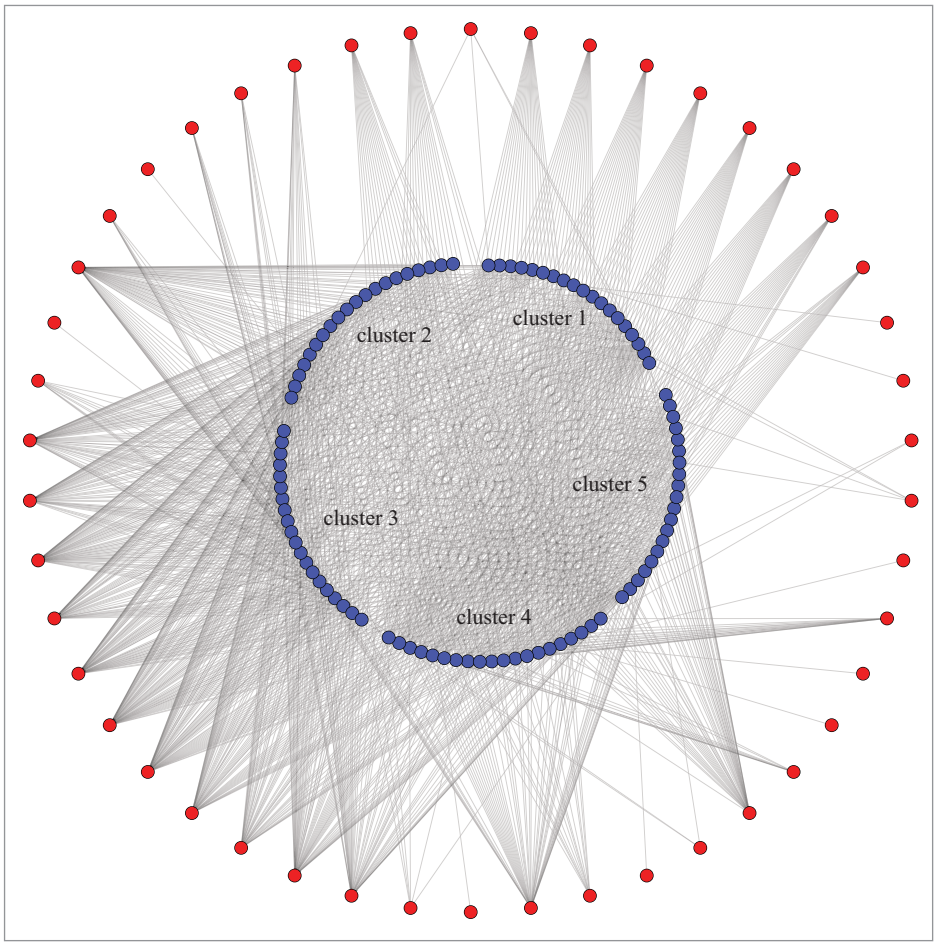

Figure 2. Connections of less frequent LCVs to *Helitrons*. *Helitrons* in the training set are clustered based on their 3′-end sequences. The top 5 clusters, each including 20 selected *Helitrons* (blue circles), are connected with 46 less frequent LCVs (red circles) they contain. The LCVs are shared by less than 30% of *Helitrons* in the training set. More frequent LCVs are not shown here to ensure better visualization.

of HelitronScanner. Compared to BLAST-derived sequence similarities, LCVs are more granular overrepresented patterns present at variable locations, not necessarily in line with the order of their original locations. How LCVs are combined in known *Helitrons* from the training set does not have to be the same as how they appear in new *Helitrons*, provided that putative *Helitrons* bear enough significance measured by the number of LCVs they contain. This relaxed constraint gives rise to the discovery of more divergent putative *Helitrons* that would otherwise be missed by BLAST or similar methods, while still demanding a certain degree of connection between known and predicted *Helitrons*. It is the LCVs that bridge functional resemblance and seemingly unrelated divergence on a whole-sequence level among *Helitrons*. Out of the 107,367 putative *Helitrons* identified by HelitronScanner from 39 plant genomes,[19] we investigated their divergence by clustering 30-bp 3′ end sequences using the cd-hit program.[21] **Figure 1** shows hierarchical relationships among the top 50 clusters, which account for 39,554 *Helitrons*, with respective sequence logos. Sequence similarity within clusters varies. For instance, *Helitron* termini are more homogeneous within cluster 32, cluster 44, cluster 46 and cluster 50 than within other clusters. Similarities among clusters are revealed by the inner dendrogram in **Figure 1**. Although the 3′-CTRR is not universal in all clusters, all clusters appear to be more conserved at the very 3′ terminus and another region a few base pairs upstream from it, which probably reflects the known 3′-end hairpin structure existing in most *Helitrons*.

In a host genome, *Helitron* copies can be almost identical or very divergent. The gradual sequence variation makes clustering *Helitrons* based on sequence similarities somewhat arbitrary in terms of chosen thresholds. Creating multiple sequence alignment profiles for each cluster of *Helitrons* is also affected by how *Helitrons* are clustered. The LCV approach does not require *Helitron* categorization before an exhaustive search for overrepresented sequence patterns in the training set. LCVs are retained during the search only if their frequency is higher than average or a preset threshold. We clustered *Helitron* 3′-terminal sequences from the training set[19] and analyzed their connections to the extracted LCVs. It is natural that most LCVs are shared within clusters. Some highly frequent LCVs are even shared in many clusters. On the other hand, *Helitrons* in one cluster may have different sets of LCVs due to sequence variation within the cluster. Generally LCVs do not coincide with *Helitron* clusters. As in **Figure 2**, we chose 20 *Helitrons* from each of the top 5 clusters (blue circles) in the training set and connected them with the LCVs (red circles) they contain. Only 46 LCVs that are shared by less than 30% of *Helitrons* in the training set were depicted here to ensure better visualization. It can be seen from that these less frequent LCVs do not exclusively reside in one cluster, which complicates a clear categorization of *Helitron* families. Given the evolutionary distance revealed by *Helitron* terminal clustering, the mostly shared LCVs among clusters may represent nucleotide patterns that are conserved throughout evolution and are likely under selection pressure.

The LCV approach does not require prior knowledge of how training sequences should be aligned or which regions are of interest, especially when

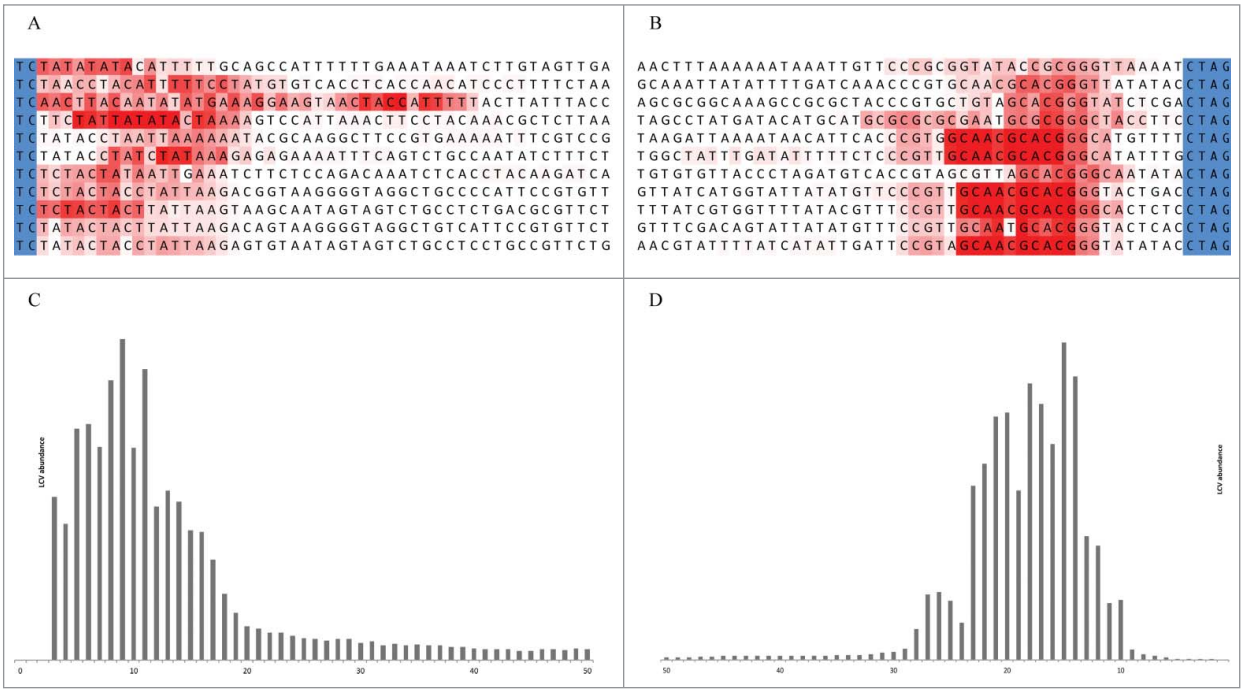

Figure 3. LCV variation and their accumulated weight in *Helitrons*. LCV distribution in *Helitron* 5′ (**A**) and 3′ (**B**) ends is depicted by nucleotides colored in red. Saturation of color is proportional to numbers of LCVs nucleotides match. The invariant 5′-TC and 3′-CTAG *Helitron* hallmarks are colored in blue. Histograms of accumulated numbers of matched LCVs in *Helitron* 5′ (**C**) and 3′ (**D**) ends show variation in conserved terminal regions.

experimental data is not available. We tried to extract LCVs without assumptions of regions of interest and found that LCVs reside only within 50-bp of both termini after testing 200-bp *Helitron* terminal sequences and 100-bp insertion sites. Most *Helitrons* share the 5′-TC and 3′-CTRR hallmarks at their termini. The LCVs are fine-grained representation of favorable combinations of nucleotides in divergent *Helitrons*. In contrast, BLAST essentially detects larger-scale sequence homology. **Figure 3** shows the distribution of 303 and 575 LCVs from *Helitron* 5′ and 3′ ends respectively over 11 representative *Helitron* terminal sequences from the training set. Nucleotides are colored in red if they match LCVs in the 50-bp region of *Helitron* 5′ (**Fig. 3A**) and 3′ (**Fig. 3B**) ends. Saturation of color of each nucleotide is proportional to the number of LCVs it matches. The blue regions at the very termini are the known 5′-TC and 3′-CTRR *Helitron* hallmarks. One LCV may reside at variable locations in different *Helitrons* and may have gaps between the conserved nucleotides. Uncolored nucleotides flanked by colored ones in **Figure 3** are gaps in LCVs. That the gapped nucleotides are less conserved and more susceptible to mutation suggests they are not as functionally crucial as the conserved nucleotides. The length of the colored region also varies with different terminal sequences. *Helitrons* with potential multiple ends are expected to have longer range of matched LCVs. Different numbers of matched LCVs in *Helitron* termini indicate location-specific weight of conserved nucleotides, as the color patterns demonstrated in **Figure 3**. Histograms of LCV abundance at *Helitron* 5′ (**Fig. 3C**) and 3′ (**Fig. 3D**) ends show distribution of overall LCV weight at each location contributed by all LCVs in all *Helitrons* from the training set. The 9th and 15th nucleotides from the 5′ and 3′ ends, respectively, appear to be overall the most conserved locations in all *Helitron* termini. LCVs at *Helitron* 3′ ends are mostly concentrated within a 30-bp range while LCVs at 5′ ends spread more broadly, which makes it harder to detect *Helitron* 5′ ends than 3′ ends in practice.

The large cache of overlooked *Helitrons* uncovered by HelitronScanner is a great resource to the research community for further study of the active roles *Helitrons* have played in genome dynamics. We are currently working on functional annotation and comparative analysis of the newly identified *Helitrons* by HelitronScanner.

Conclusion

As the key component of HelitronScanner, the LCV approach extracts granular conserved information (LCVs) at variable locations from unaligned *Helitron* sequences, and identifies *Helitrons* based on match numbers of the LCVs. HelitronScanner outperformed previous methods by utilizing LCVs collectively as definitive *Helitron* features besides known hallmarks. A large number of divergent *Helitrons* in many plant species was uncovered, which will be a great resource for research community. The results indicate that the LCV approach is more sensitive to highly divergent *Helitrons* than previous sequence alignment methods. Analysis of the overrepresented and conserved LCVs over different groups of *Helitrons* may help provide insights into the evolutionary trajectory of this unusual transposon superfamily.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Feschotte C, Pritham EJ: DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 2007; 41:331-368; PMID:18076328; http://dx.doi.org/10.1146/annurev.genet.40.110405.090448.
2. Lai J, Li Y, Messing J, Dooner HK. Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci U S A 2005; 102:9068-9073; PMID:15951422; http://dx.doi.org/10.1073/pnas.0502923102.
3. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC: The maize genome contains a Helitron insertion. Plant Cell 2003; 15:381-391; PMID:12566579; http://dx.doi.org/10.1105/tpc.008375.
4. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 2005; 37:997-1002; PMID:16056225; http://dx.doi.org/10.1038/ng1615.
5. Xu JH, Messing J: Maize haplotype with a helitron-amplified cytidine deaminase gene copy. BMC Genet 2006; 7:52; PMID:17094807; http://dx.doi.org/10.1186/1471-2156-7-52.
6. Kapitonov VV, Jurka J: Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 2001; 98:8714-8719; PMID:11447285; http://dx.doi.org/10.1073/pnas.151269298.
7. Li Y, Dooner HK: Helitron proliferation and gene-fragment capture. In Topics in Current Genetics: Plant Transposable Elements-Impact on Genome Structure and Function. Volume 24. Berlin: Springer-Verlag; 2012.
8. Du C, Caronna J, He L, Dooner HK: Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 2008; 9:51; PMID:18226261; http://dx.doi.org/10.1186/1471-2164-9-51.
9. Yang L, Bennetzen JL: Structure-based discovery and description of plant and animal Helitrons. Proc Natl Acad Sci U S A 2009; 106:12832-12837; PMID:19622734; http://dx.doi.org/10.1073/pnas.0905563106.
10. Du C, Fefelova N, Caronna J, He LM, Dooner HK: The polychromatic Helitron landscape of the maize genome. Proc Natl Acad Sci U S A 2009; 106:19916-19921; PMID:19926866; http://dx.doi.org/10.1073/pnas.0904742106.
11. Yang L, Bennetzen JL: Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A 2009; 106:19922-19927; PMID:19926865; http://dx.doi.org/10.1073/pnas.0908008106.
12. Sweredoski M, DeRose-Wilson L, Gaut BS: A comparative computational analysis of nonautonomous Helitron elements between maize and rice. BMC Genomics 2008; 9:467; PMID:18842139; http://dx.doi.org/10.1186/1471-2164-9-467.
13. Tempel S, Nicolas J, El Amrani A, Couee I: Model-based identification of Helitrons results in a new classification of their families in Arabidopsis thaliana. Gene 2007; 403:18-28; PMID:17889452; http://dx.doi.org/10.1016/j.gene.2007.06.030.
14. Bao Z, Eddy SR: Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 2002; 12:1269-1276; PMID:12176934; http://dx.doi.org/10.1101/gr.88502.
15. Price AL, Jones NC, Pevzner PA: De novo identification of repeat families in large genomes. Bioinformatics 2005; 21 Suppl 1:i351-358; PMID:15961478; http://dx.doi.org/10.1093/bioinformatics/bti1018.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990; 215:403-410; PMID:2231712; http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
17. Lee D, Redfern O, Orengo C: Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007; 8:995-1005; PMID:18037900; http://dx.doi.org/10.1038/nrm2281.
18. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 1982; 10:2997-3011; PMID:7048259; http://dx.doi.org/10.1093/nar/10.9.2997.
19. 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 U S A 2014; 111:10263-10268; PMID:24982153; http://dx.doi.org/10.1073/pnas.1410068111.
20. Xiong W, Li T, Chen K, Tang K: Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information. Nucleic Acids Res 2009; 37:5632-5640; PMID:19651875; http://dx.doi.org/10.1093/nar/gkp628.
21. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ: CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28:3150-3152; PMID:23060610; http://dx.doi.org/10.1093/bioinformatics/bts565.