HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes

Wenwei Xiong[a], Limei He[b], Jinsheng Lai[c], Hugo K. Dooner[b,d,1], and Chunguang Du[a,1]

[a]Department of Biology and Molecular Biology, Montclair State University, Montclair, NJ 07043; [b]Waksman Institute, Rutgers, the State University of New Jersey, Piscataway, NJ 08854; [c]National Maize Improvement Center, China Agricultural University, Beijing 100083, China; and [d]Department of Plant Biology, Rutgers, the State University of New Jersey, New Brunswick, NJ 08801

Contributed by Hugo K. Dooner, June 6, 2014 (sent for review January 28, 2014)

Transposons make up the bulk of eukaryotic genomes, but are difficult to annotate because they evolve rapidly. Most of the unannotated portion of sequenced genomes is probably made up of various divergent transposons that have yet to be categorized. *Helitrons* are unusual rolling circle eukaryotic transposons that often capture gene sequences, making them of considerable evolutionary importance. Unlike other DNA transposons, *Helitrons* do not end in inverted repeats or create target site duplications, so they are particularly challenging to identify. Here we present HelitronScanner, a two-layered local combinational variable (LCV) tool for generalized *Helitron* identification that represents a major improvement over previous identification programs based on DNA sequence or structure. HelitronScanner identified 64,654 *Helitrons* from a wide range of plant genomes in a highly automated way. We tested HelitronScanner's predictive ability in maize, a species with highly heterogeneous *Helitron* elements. LCV scores for the 5′ and 3′ termini of the predicted *Helitrons* provide a primary confidence level and element copy number provides a secondary one. Newly identified *Helitrons* were validated by PCR assays or by in silico comparative analysis of insertion site polymorphism among multiple accessions. Many new *Helitrons* were identified in model species, such as maize, rice, and *Arabidopsis*, and in a variety of organisms where *Helitrons* had not been reported previously to our knowledge, leading to a major upward reassessment of their abundance in plant genomes. HelitronScanner promises to be a valuable tool in future comparative and evolutionary studies of this major transposon superfamily.

transposition | algorithm | computational tool | bioinformatic analysis

Although transposable elements constitute the bulk of most sequenced eukaryotic genomes, their annotation has been hindered by their rapid evolutionary divergence. It is conceivable that a large fraction of the unannotated genome of most eukaryotes is made up of as yet unrecognized transposons. To date, elements have been assigned to a superfamily largely on the basis of terminal sequence homology to other elements that still encode vestiges of that superfamily's transposase (1). *Helitrons* are particularly challenging to identify because, unlike other DNA transposons, they do not end in inverted repeats or create target site duplications. These novel eukaryotic transposons were discovered only recently from a comparative bioinformatic analysis of several plant and animal genomes (2). *Helitrons* have attracted widespread attention because their remarkable ability to capture gene sequences, and intergenic regions containing potential regulatory elements, makes them of considerable potential evolutionary importance (3–10). Among carefully studied genomes, *Helitron* content has been estimated to be approximately 2% in *Arabidopsis* and maize (2, 11, 12) and 4.23% in silkworm (9). However, these values are most likely underestimates because *Helitrons* are hard to detect computationally given their lack of classical transposon structural features. As has been suggested (13), the number of reported *Helitrons* probably constitutes just the tip of the iceberg.

The first *Helitron* computational searching tool, HelitronFinder, was developed by us for the purposes of analyzing the *Helitron* content of maize (14). This tool was based on conserved sequences at the termini of most *Helitrons* (5′-TC and CTAG-3′) and a conserved 16- to 20-bp palindromic structure located 10–15 bp upstream of the 3′ terminus. Using HelitronFinder, we identified almost 3,000 new *Helitrons* in the B73 maize genome (12). HelSearch (15), another computational tool, is very similar to HelitronFinder in terms of identifying the 3′ end of *Helitrons*. Both programs look for the hairpin structure and the CTRR 3′ terminus. The difference is that users of HelSearch have to manually search for the 5′ end of *Helitrons*, whereas HelitronFinder can identify the 5′ end automatically. When the two groups compared the predicted *Helitrons* in maize by using HelSearch and HelitronFinder, more than 95% of the *Helitron* candidates identified by both programs were identical (11, 12). However, the *Cornucopious* element, which consists of thousands of copies of an ∼1.0-kb *Helitron* that may be the most abundant transposon in maize, was overlooked by both HelSearch and HelitronFinder because of a more divergent 3′ end. Another early computational work used a combination of BLAST search and hidden Markov models to identify many new *Helitrons* in the rice genome, but very few in maize (16). A method based on separate exhaustive searches for *Helitron* 5′ and 3′ end consensus sequences identified a number of new *Helitrons* in *Arabidopsis thaliana* (17). Because there is no comprehensive list of all *Helitron* termini identified to date, this method also has limitations in finding *Helitrons* with more diverse termini.

A more efficient and general way to identify *Helitrons* from plant and animal genomes is needed. The key to achieve this objective is to find sequence patterns applicable to most known *Helitrons* and extensible to unknown ones. However, as a special type of transposon, *Helitrons* do not have an identifiable

Significance

Helitrons are unusual rolling-circle eukaryotic transposons with a remarkable ability to capture gene sequences, which makes them of considerable evolutionary importance. Because *Helitrons* lack typical transposon features, they are challenging to identify and are estimated to comprise at most 2% of sequenced genomes. Here, we describe HelitronScanner, a generalized tool for their identification based on a motif-extracting algorithm proposed initially in a study of natural languages. HelitronScanner overcomes the divergence of *Helitron* termini among species by using conserved nucleotides at potentially variable locations. Many new *Helitrons* were identified in all organisms examined, resulting in a major reassessment of their abundance in eukaryotic genomes. In maize, they make up >6% of the genome and are the most abundant DNA transposons identified.

Author contributions: W.X. and C.D. designed research; W.X., L.H., and C.D. performed research; J.L. contributed new reagents/analytic tools; W.X., L.H., H.K.D., and C.D. analyzed data; and W.X., H.K.D., and C.D. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: dooner@waksman.rutgers.edu or duc@mail.montclair.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1410068111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1410068111

deterministic functional structure and their conserved termini are so diverse among different families that BLAST-based methods will miss distantly related *Helitron* families. Thus, an effective motif discovery algorithm that can extract representative patterns from diverse clusters of *Helitrons* without prior knowledge is crucial to the success of *Helitron* identification. Many motif discovery algorithms have been developed by extracting representative patterns from various kinds of datasets. An unsupervised motif extraction algorithm (18) that can distill hierarchically structured patterns from corpuses of strings recursively without prior knowledge was proposed initially in a study of natural languages and was later applied to biological problems (19). It is superior to other grammar induction methods that need prior knowledge to carry out their inferences. A method for discovering conserved sequence motifs from families of aligned protein sequences named EMOTIF (20) generates a set of motifs with a wide range of specificities and sensitivities and can generate motifs that describe possible subfamilies of a protein superfamily. Another iterative statistical approach aims to develop a tool to determine potential phosphorylation sites in proteins of interest by relying on the intrinsic alignment of phospho-residues and the extraction of motifs through iterative comparison with a dynamic statistical background (21).

Here, we develop HelitronScanner, a generalized computational tool for identifying *Helitrons* from plant genomes. HelitronScanner identifies divergent *Helitrons* that were missed by both HelSearch and HelitronFinder (11, 12) and detects new *Helitrons* that would be missed by the model-based method (17). HelitronScanner relies on a local combinational variable (LCV) algorithm that has been used to extract patterns from sequences of diverse protein families varying in length and function (22). The application of these LCV patterns to all genomes available in Phytozome (23) led to the discovery of many new *Helitrons,* resulting in a major reassessment of the fraction of plant genomes that is comprised of *Helitrons*. HelitronScanner may help to unravel the transposition mechanism of *Helitrons* by providing the research community with a powerful tool for their identification.

Results

Helitron Terminal Features Represented by LCVs. Based on an improved LCV algorithm (22), HelitronScanner aims to extract more definitive *Helitron* features than the few previously identified: the TC dinucleotide at the 5′ end, the hairpin structure and CTRR (R = A or G) sequence at the 3′ end, and the A and T residues flanking the 5′ and 3′ ends, respectively. More than 5,000 double-ended *Helitron* sequences in maize, *Arabidopsis*, rice, sorghum, *Caenorhabditis elegans*, and *Medicago truncatula* genomes were taken to create two sets of 100-bp slices from both *Helitron* ends, which were then clustered separately by using cd-hit (24) at 90% similarity to remove redundant sequences and, thus, avoid LCV bias. A training set was created from 2,846 seed sequences of 5′-end clusters and 2,048 of 3′-end clusters to extract LCVs (*Methods*), and 1,613 double-ended *Helitrons* were assembled from these paired seed sequences to evaluate LCVs (see *Confidence Level by LCV Scores* below). LCVs from *Helitron* termini convey local conserved information in the sense that no global multiple sequence alignment is required that, in turn, overcomes the drawback of overall *Helitron* variation and gives rise to a more generalized set of features for *Helitron* identification in a broader array of organisms. LCVs combined together as identification features reveal more terminal conservation than previous methods that focus on overall sequence alignment.

Sequence logos of 303 LCVs from the *Helitron* 5′-terminal 50 bp (Fig. 1*A*) and 575 LCVs from the 3′-terminal 50 bp (Fig. 1*B*) display patterns in sequence conservation, compared with sequence logos from the 5,676 raw sequences of the same regions (Fig. 1 *C* and *D*). *Helitron* 5′ ends appear to be AT-rich, whereas 3′ ends contain a higher content of Cs and Gs between 11 and 30 bp from the endpoint. It is interesting that the dinucleotide AA at the 21 and 20 positions in Fig. 1*B* coincides with the dinucleotide AA often found in the middle of the CG base-paired hairpin loops (14) and that the hairpin structure at the 3′ end is reflected in LCVs without any prior knowledge input. Because of significant sequence diversity among *Helitrons*, there is little conserved sequence information other than 5′-TC and 3′-CTAG (Fig. 1 *C* and *D*).

The LCV algorithm extracts overrepresented patterns based on an iteration of exhaustive enumeration of oligonucleotides. The patterns may have nonconserved locations and do not have to be vertically aligned to a fixed location, which brings great flexibility for motif finding at the cost of speed. After trials of LCV extraction, we determined from the location of generated LCVs that the *Helitron's* conserved regions are 50 bp at 5′ and 3′ ends, so LCVs were extracted from them. The 10 most frequent LCVs from 5,676 *Helitrons* are shown in Table 1. LCV notation follows the syntax of regular expression in computer science, i.e., a dot denotes any nucleotide, brackets denote alternative nucleotides, and numbers within braces denote occurrences. For instance, *AA.G.ACG.{9}CT[AG]{2}* represents a motif comprised of 2 As, a nucleotide of any kind (i.e., N), 1 G, 1 N, ACG, 9 Ns, and the CTRR (R for A or G) ending. An LCV with nonconserved regions of fixed length, represented by a single number within braces, provides more discriminative information than an LCV with nonconserved regions of variable length. Among the 5,676 *Helitrons* in the training set, the frequencies of 3′ LCVs are higher than those of 5′ LCVs, indicating that *Helitron* 3′ ends are more conserved than 5′ ends, in agreement with our previous study (14).

Confidence Level by LCV Scores. The sum of the LCV scores of the predicted *Helitron's* 5′ and 3′ ends is taken as the primary criterion of the prediction's confidence level. Higher scores indicate more specificity of the combination of LCV features and, thus, provide a higher confidence level. However, too stringent a threshold may cause more diverse *Helitrons* to be missed. After analyzing the distribution of LCV scores among the 1,613 *Helitrons* in the training set, a threshold of 5 for the LCV score at each end was chosen initially as a compromise between sensitivity and specificity (*SI Appendix*, Fig. S1). Statistically, 95% of the *Helitrons* in the training set had scores ≥10.

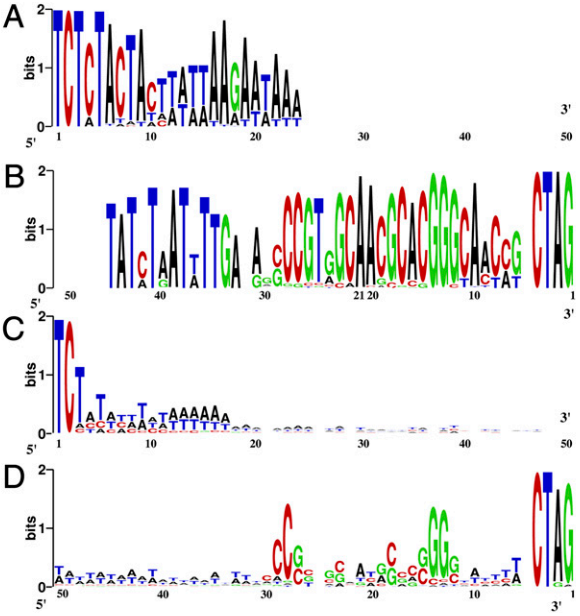

Fig. 1. *Helitron* terminal conservation shown in sequence logos. Sequence logos of: 303 LCVs generated from 50 bp at *Helitron* 5′ termini (*A*); 575 LCVs generated from 50 bp at *Helitron* 3′ termini (*B*); raw sequences of *Helitron* 5′ termini (*C*); raw sequences of *Helitron* 3′ termini (*D*). LCV location variation (*A* and *B*) has been normalized according to their frequency.

Table 1. Top 10 LCVs on 5′ and 3′ ends from 5676 *Helitrons* in the training set

LCVs from *Helitron* 5′ ends	Occurrences	LCVs from *Helitron* 3′ ends	Occurrences
TCTCTACTA	2,336	AA.G.ACG.{9}CT[AG]{2}	2,646
TCT.TACTA.T	1,993	CGT.GCAA.{15}CT[AG]{2}	2,537
TCT.TACTAC	2,039	G.AA.GC.CG.{9}CT[AG]{2}	2,563
TC.{2}TACTACT	1,843	CAA.GC.CG.{9}CT[AG]{2}	2,554
TCT.TAC.ACT	1,823	GC.A.GC.CG.{9}CT[AG]{2}	2,588
TCT.TA.TACT	1,767	AA.GCACG.{9}CT[AG]{2}	2,549
TCT.TACT.CT	1,734	G.AA.G.ACG.{9}CT[AG]{2}	2,566
TCT.T.CTACT	1,735	GC.A.G.ACG.{9}CT[AG]{2}	2,577
TC.CTACTA.T	1,553	GCAA.GC.C.{10}CT[AG]{2}	2,491
TC.{9}TATTAAG	1,556	C.A.GCACG.{9}CT[AG]{2}	2,551

***Helitron* Identification in Angiosperms.** Using an LCV score threshold of 5 for each end, we ran HelitronScanner against a wide range of plant genome sequences from Phytozome version 9.0 (23) and identified 107,367 *Helitrons*. The LCV scores assigned by HelitronScanner to the identified *Helitrons* are an indicator of prediction confidence. *SI Appendix*, Fig. S2 shows the variation of predicted *Helitron* number from all plant genomes under different LCV score thresholds. The number of *Helitrons* decreases dramatically with more stringent LCV score thresholds. There are 107,367 *Helitrons* using an LCV score ≥10 as the cutoff criterion, 33,530 (or 31.2% of original) at ≥20, 12,439 (11.6%) at ≥30, 7,812 (7.3%) at ≥40, and 5,164 (4.8%) at ≥50. Higher LCV scores provide a higher prediction confidence, whereas more stringent thresholds lead to the loss of a fraction of true *Helitrons* (see *In Silico Verification of Helitrons* below).

SI Appendix, Table S1 shows the number of *Helitrons* in each organism, their genome abundance, and size distribution. Because HelitronScanner identifies *Helitron* 5′ and 3′ termini separately and the 3′ termini are more conserved, the *Helitron* number is mainly determined by the 3′ termini, whereas the 5′ termini are more diverse and abundant. For each species, we then estimated a false positive rate (FPR) at the chosen LCV threshold by running HelitronScanner on a randomized version of its genome (*Methods*). The FPR varied from species to species, and several of those lacking close relatives in the training set often had an estimated FPR >50%. A more stringent 3′-end threshold score of 10 was chosen for those species to balance detection power against FPR (*SI Appendix*, Table S2). Plant species, such as algae and bryophytes, still giving an estimated FPR >45% under the new threshold, were dropped from the list. Table 2 gives the number of *Helitrons* predicted by HelitronScanner and their percentage in angiosperm genomes, after correcting for estimated false positives.

A thorough search for *Helitrons* in the maize genome by HelitronFinder (12) and HelSearch (11) identified 2,791 and 1,930 *Helitrons*, respectively, which comprised approximately 2% of the genome. In contrast, HelitronScanner identified 31,233 *Helitrons* or 6.6% of the maize genome, including 92% of those identified by HelitronFinder. To verify new *Helitrons* uniquely identified by HelitronScanner, we carried out tests of insertion site polymorphisms by PCR assays and in silico comparisons. The results indicate that HelitronScanner is efficacious in identifying authentic *Helitrons* (see details in *Helitron verification*). HelSearch (11) identified 281 complete *Helitrons* in *A. thaliana*, 230 in *Medicago*, 651 in rice, and 608 in sorghum, compared with HelitronScanner, which identified 609, 1,142, 4,634, and 2,082 *Helitrons* in the respective genomes. Three rice accessions (*japonica*, *indica*, and *glaberrima*) were used to test *Helitron* insertion polymorphisms: 248 of the *Helitrons* predicted by HelitronScanner were verified, in contrast to 179 of those predicted by HelSearch.

HelitronScanner also identified *Helitrons* in many plant species where *Helitrons* had not been previously reported (organisms marked with an asterisk in Table 2), including a wide range of monocots and eudicots. The *Helitron* abundance revealed by our

Table 2. *Helitron* identification in plant genomes

Organism	Genome scanned, MB	Helitrons	Genome percentage, %
*Aquilegia coerulea**	302	112	0.1
Arabidopsis lyrata	207	2,262	3.6
Arabidopsis thaliana	120	609	1.6
Brachypodium distachyon	272	2,558	4.1
Brassica rapa	284	3,314	4.3
*Capsella rubella**	135	1,317	3.3
*Carica papaya**	343	394	0.5
*Citrus clementina**	301	124	0.2
*Citrus sinensis**	319	345	0.5
*Cucumis sativus**	203	14	0.0
*Eucalyptus grandis**	691	136	0.1
*Fragaria vesca**	207	225	0.5
*Manihot esculenta**	533	548	0.5
Medicago truncatula	419	1,142	1.2
Mimulus guttatus	322	2,208	2.9
Oryza sativa ssp. japonica	374	4,634	4.0
*Panicum virgatum**	1,358	6,340	0.5
*Phaseolus vulgaris**	521	1,238	1.2
Populus trichocarpa	434	216	0.2
*Ricinus communis**	351	199	0.2
*Setaria italica**	406	977	0.9
*Solanum lycopersicum**	782	344	0.2
*Solanum tuberosum**	706	1,619	0.9
Sorghum bicolor	739	2,082	1.0
Thellungiella halophila	243	78	0.1
*Theombroma cacao**	346	386	0.5
Zea mays	2,066	31,233	6.6

*Plants with no previously reported *Helitrons*.

study is consistent with previous suggestions that the percentage of *Helitrons* in genomes was most likely underestimated (12). We found no sign of correlation between *Helitron* abundance and genome sizes.

Helitron Verification by PCR Assays in Multiple Maize Lines. Maize has the most variable genome structure yet described (25, 26) and one of the highest contents of transposons (85%) among fully sequenced genomes, components of the genome that have shaped its architecture over time (27). Different maize inbreds are highly polymorphic in their transposon content and distribution and provide valuable germplasm for the validation of computationally predicted *Helitrons* (14). Often, a *Helitron* is present in some lines while absent in others. Here, we also adopted the plus/minus polymorphism criterion to validate the authenticity of predicted *Helitrons*.

We randomly picked 15 high-score (LCV > 20) and 4 medium-score (LCV = 11–20) *Helitrons* predicted exclusively by HelitronScanner (i.e., not by HelitronFinder or HelSearch) that were flanked by single-copy regions in the B73 reference genome (28). PCR primers for flanking sequences were designed and plus/minus variation was tested in different inbred lines (*SI Appendix*, Table S3 from hel_pcr_001 to hel_pcr_019). Of the 19 *Helitrons*, 13 exhibited vacant sites (i.e., only flanking sequences amplified) in maize lines other than B73, thus verifying their authenticity (Column "Validated"; TRUE entries in *SI Appendix*, Table S3). A PCR band image of five validated *Helitrons* is also shown in *SI Appendix*, Fig. S3. *Helitrons* not validated here by the PCR assay are not necessarily false positives because they may be absent in inbred lines not included in our small panel.

A supporting criterion of a *Helitron's* authenticity is element copy number. As seen in *SI Appendix*, Table S3, either a high LCV score or a high copy number (>4 copies in the host genome) is a strong predictor of *Helitron* authenticity.

In Silico Verification of _Helitrons_. Applying the same concept as in the PCR verification, we also compared *Helitron* insertion sites and their flanking sequences in silico among multiple sequenced accessions where true *Helitrons* might exhibit plus/minus polymorphism. By BLASTing *Helitron* flanking sequence joints against other sequenced accessions, we verified *Helitrons* in silico based on the presence of vacant sites. *SI Appendix*, Fig. S4 shows examples of *Helitron* polymorphisms detected in maize, rice, and *Arabidopsis*. As can be seen, the flanking regions are highly conserved, so vacant sites lacking *Helitron* insertions are easily identified. We tested maize *Helitrons* predicted in the fully assembled genome of the inbred B73 (28) against contigs of the Mo17 inbred (http://bo.csam.montclair.edu/du/software/helitronscanner). *SI Appendix*, Fig. S4*A* shows that a 1,572-bp *Helitron* identified on chromosome 1 of B73 was absent in Mo17. In rice, the fully assembled genomes of the two subspecies *japonica* (29) and *indica* (30) of *Oryza sativa* and contig data of *Oryza glaberrima* from Arizona Genomics Institute were used to test *Helitron* polymorphism. *SI Appendix*, Fig. S4*B* shows that a 2,009-bp *Helitron* identified on chromosome 1 of *japonica* was absent in *glaberrima*. In *Arabidopsis*, the ecotypes Columbia, C24, and Bur-0 from the *A. thaliana* 1001 Genomes Project were investigated. As seen in *SI Appendix*, Fig. S4*C*, a 547-bp *Helitron* was present in Col and C24, but not in Bur-0. The identified *Helitron* and its flanking region are highly conserved in Col and C24. Compared with other programs, HelitronScanner identified many inactive *Helitrons* that are too divergent to have been detected previously, which, in turn, caused a decrease in the *Helitron* verification rate. For example, HelSearch detected 651 full-length *Helitrons* and at least 6,947 elements with conserved 3′ ends in the rice subspecies *japonica*. The detection power of HelitronFinder, however, is highly confined to the maize genome.

To validate *Helitrons* efficiently, artificial sequences of 50 bp were made up by joining 25-bp flanking sequences on both sides of the *Helitron* and BLASTed against genome sequences for evidence of plus or minus polymorphisms. The mega-BLAST task of nucleotide blastn was chosen to search for highly similar sequences, and a minimum 80% coverage of the query sequence was required to support the presence of the joint, i.e., vacant, flanking sequences. Within one organism, the presence of vacant sequences in accessions other than the one with the predicted *Helitron* validated the *Helitron*'s authenticity. We compared numbers of validated *Helitrons* identified by HelitronScanner with those identified by HelitronFinder or HelSearch, two widely used computer programs, in maize, rice, and *Arabidopsis* (Table 3). A large number of maize *Helitrons* were missed by the previous methods, showing the efficacy of HelitronScanner in identifying *Helitrons* in the highly polymorphic maize genome. In rice, HelitronScanner also predicted 69 more validated *Helitrons* than HelSearch. However, in *Arabidopsis*, HelitronScanner identified three fewer validated *Helitrons* than HelSearch simply because these *Helitrons* had 3′ LCV scores below 5, the HelitronScanner threshold chosen to avoid a high false positive rate.

We analyzed the LCV scores of 1,616 *Helitrons* validated in maize, rice and *Arabidopsis* to assess retrospectively our selection of a threshold (*SI Appendix*, Table S4). Because *Helitrons* scoring lower than 10 were not included, based on the distribution of scores in the training set and PCR assays, the minimum score is 10. Of the validated *Helitrons,* only 15 (0.9%) have an LCV score of 10, 152 (9.4%) have a score of 11–20, and the vast majority (89.7%) have scores >20 (Fig. 2). The maximum, average, and median LCV scores are 75, 48, and 56, respectively. In particular, 813 (50.3%) of the validated *Helitrons* have scores higher than 50. The score distribution indicates that validated *Helitrons* mainly have high scores and that low-scoring *Helitrons* are rare. Copy number is a complementary indicator when *Helitron* scores are low, because most *Helitrons* have at least two copies. Of the validated *Helitrons,* only 30 are singleton and their scores are all >10. Therefore, a pragmatic guideline would be to accept *Helitrons* with LCV scores >20 and multiple copies, reject singleton *Helitrons* with LCV scores ≤10, and analyze intermediate cases further.

Evolutionary Distance Revealed by LCVs. LCVs are overrepresented patterns attributed to *Helitrons*. We analyzed the more conserved 3′ termini by showing how LCVs are shared among species in terms of evolutionary distance (Fig. 3). We looked for the presence of the extracted 575 LCVs from the 3′ ends in the 1,616 in silico-validated *Helitrons* (1,352 from maize, 248 from rice, and 16 from *Arabidopsis*). A 1,616 × 575 matrix was generated based on the matching condition "1 for true and 0 for false" of every *Helitron* against every LCV, which was then decomposed by principal component analysis. Each *Helitron* was projected onto the top-two principal components (PC) 1 and PC2, which accounted for 15.03% and 3.07% of total sample variance, respectively (Fig. 3). Although *Helitrons* in the monocots (maize and rice) overlap, those in the eudicot *Arabidopsis* are distinctive. This interrelationship reveals that LCVs cannot only serve collectively as the deterministic feature of *Helitrons*, but can also convey evolutionary relationships.

More interestingly, the pattern GC.CG.{9}CTRR is the most shared sequence feature among all studied species. In particular, 959 (70.9%) of 1,352 validated (i.e., polymorphic and, therefore, recently transposed) maize *Helitrons* and 156 (62.9%) of 248

Table 3. _Helitrons_ validated by insertion polymorphism from HelitronScanner and other sources

	Helitrons			HelitronFinder
Species	from	Validated in	Scanner	HelSearch
Maize	B73	Mo17	1,352	328*
Rice	Japonica	Indica, glaberrima	248	179†
Arabidopsis	Col-0	C24, Bur-0	16	19†

*HelitronFinder (11).

†HelSearch (12).

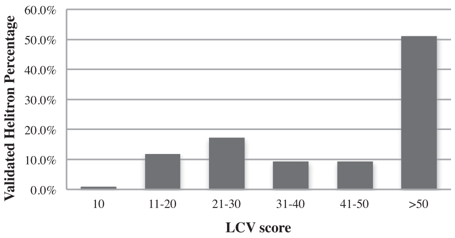

Fig. 2. Score distribution of 1,616 validated *Helitrons*.

validated rice *Helitrons* contain this pattern, which may reflect a crucial characteristic for *Helitron* transposition.

Discussion

HelitronScanner identifies *Helitrons* in an automated way, uncovering many new *Helitrons* in sequenced organisms. For example, maize *Helitrons* constitute 6.6% of the genome rather than the previously estimated 2% (11, 12). The generality of HelitronScanner is supported by the identification of new *Helitrons* in many plants where they had not been previously reported (Table 2). *Helitron* abundance varied greatly among sequenced genomes. The 3′ ends of identified *Helitrons* are more conserved; consequently, *Helitrons* tend to have unique 3′ ends but multiple 5′ ends, as has been found in prior studies.

HelitronScanner outperforms previous methods in identifying not only a larger number, but also new types, of *Helitrons*. In our previous study, the most abundant *Helitron* family, *Cornucopious*, was estimated to have >2,000 copies in the maize genome (12). These elements were not identified by either HelitronFinder or HelSearch, but by a manual BLAST search with the agenic 0.9-kb *Hel1-5* element first identified in the *bz1* haplotype of inbred I137TN (26). Introduction of the more flexible LCVs allowed HelitronScanner to detect 2,058 copies of *Cornucopious* in maize. This outcome strongly supports the generality of HelitronScanner because none of the *Cornucopious Helitrons* were included in the training dataset. *Helitrons* identified exclusively by HelitronScanner were missed by HelitronFinder because their 3′-ends diverged from the pattern CG.{3,5}A{1,2}.{3,5}[CG]G.{9}CTRR required by HelitronFinder. The discordance included mismatches of the leading CG dinucleotide or absence of A's in the middle of the hairpin loop. Only the CG or GG dinucleotide 9 bp away from the CTRR appears to be more conserved, but this loose constraint would match too many sequences in the genome for practical uses. In contrast, the combinatorial power of LCVs provides a good balance between sensitivity and specificity for *Helitron* identification and gives HelitronScanner the flexibility to incorporate potentially better features once additional *Helitrons* are discovered and validated.

The LCV algorithm is effective in drawing representative patterns from *Helitrons*, considering that hairpin structures and other conserved features identified in our earlier work are detected de novo: more than 90% of Helitrons detected by HelitronScanner had hairpins. However, the absence of assumptions in HelitronScanner provides for a more thorough search of the genome with small intrinsic patterns (i.e., LCVs) that allow gaps and tolerate variability, thereby collectively defining *Helitrons* in a more flexible way. The LCVs drawn from known *Helitrons* are overrepresented sequence patterns that contain conserved nucleotides critical for *Helitron* amplification during diversification. HelitronScanner works in a finer-grained level of *Helitron* similarities, compared with methods based on overall terminal consensus (14, 15) or a model-based method that essentially exhausts all possible combinations of known *Helitron* termini (17). In other words, our LCV approach uncovers far more combinations of conserved patterns because each *Helitron* end has hundreds of significant LCVs and, therefore, identifies more divergent *Helitrons*.

To estimate a false positive rate of medium-to-low-score predictions in maize, we first tested copy numbers of 100 randomly chosen maize *Helitrons* that were exclusively identified by HelitronScanner, had scores ranging from 10 to 20, and were distinct from each other at less than 90% similarity. Only 17 or 17% of them were singletons in the maize genome. This test provides an approximate FPR of medium-scoring *Helitrons*, even though bona fide single-copy *Helitrons* do exist. We then ran HelitronScanner against randomized genomes (*Methods*) that share nucleotide composition and overall size with the maize B73 genome (*SI Appendix*, Fig. S5). The FPR decreases from 28% down to 8% as the threshold increases from 5 to 10. We choose a threshold of 5 for each Helitron end so as to generate a broad range of divergent *Helitrons* at a cost of a slightly higher false positive rate. Although our previous HelitronFinder has a low error rate of 0.13%, it only detected 3,405 Helitrons in the maize genome, compared with 31,233 Helitrons by HelitronScanner, after removing the estimated 27.5% false ones. For species more distant than those in the training set, we increased the threshold to 10 for the 3′ end and kept a threshold of 5 for the 5′ end so as to achieve lower false positive rates while not rejecting too many true *Helitrons*.

Because *Helitrons* are highly divergent, different sets of LCVs occur preferentially at particular *Helitron* locations in different species. In the maize genome, the top LCV pattern GACCG.GAGC.{4}CTRR is 4 bp away from the 3′-end CTRR, whereas another pattern GAGC.G.TC.{12,16}CTRR is farther away and its location is more uncertain. In rice, another important monocot organism, the top LCV patterns are GCACGGGC.{7,8}CTRR and CGT.GCAA.{14,17}CTRR. The eudicots *Arabidopsis lyrata* and *A. thaliana* share the top LCV pattern TA.C.CGGGT.{6,7}CTRR and most other top LCVs, most likely because of their close phylogenetic relationship. Most interestingly, the pattern GC.CG.{9}CTRR is shared by the majority of *Helitrons* from all species studied. No longer universal pattern than that emerged, again conforming with the great variability of *Helitrons*. The same pattern was also observed in most *Helitrons* validated by in silico comparisons of plus/minus polymorphism, suggesting that it may be crucial to the mechanism of *Helitron* transposition.

Methods

A Two-Layered Workflow of HelitronScanner. HelitronScanner consists of a two-layered workflow using LCVs generated from *Helitron* termini collectively as *Helitron* definitive features (Fig. 4). The iterative process of generating LCVs as sequence patterns from known *Helitrons* is referred to as the first layer. The second layer predicts putative *Helitrons* with locations and scores from input DNA sequences in Fasta format. The key component of HelitronScanner involves matching matrices created for both *Helitron* termini by matching LCVs (L items) to DNA sequences (Q items, each for a putative *Helitron*). Each element $M_{i,j}$ of the Q-by-L matching matrix M is either 1 or 0 depending on

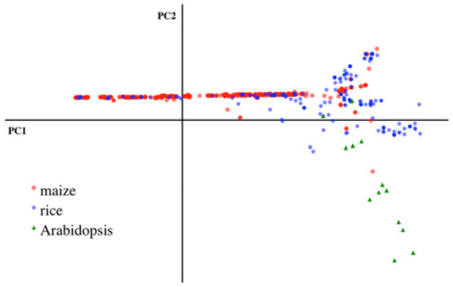

Fig. 3. Clustering of validated *Helitrons* by LCV principal components. Five hundred seventy-five LCVs from the 3′ ends of *Helitrons* in the training set were matched against 1,616 in silico-validated *Helitrons*, including 1,352 maize *Helitrons* (red circles) 248 rice *Helitrons* (blue circles), and 16 *Helitrons* from *A. thaliana* (green triangles). All *Helitrons* were projected to the top two significant principal components, PC1 and PC2, which account for 15.03% and 3.07% of total sample variance, respectively.

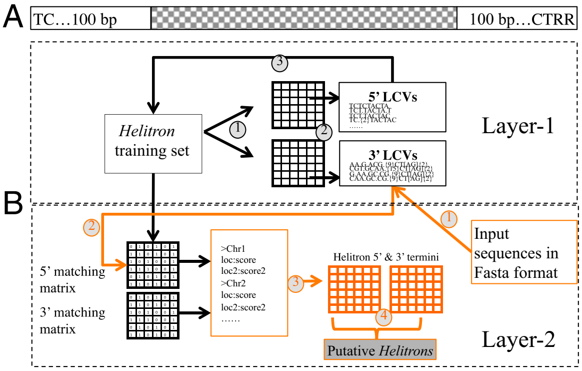

Fig. 4. Workflow of the two-layered HelitronScanner tool. (*A*) *Helitron* structural regions. LCVs were extracted from both *Helitron* terminal 100-bp regions. (*B*) Workflow of HelitronScanner. In layer-1, there are three steps (numbers in black circles) for extracting definitive features from known *Helitrons* in the training set. In step 1, 100-bp *Helitron* sequence slices from the 5′ and 3′ ends are clustered to remove redundancy and, thus, to avoid LCV bias. In step 2, two sets of LCVs are generated separately in an iteration of different thresholds. In step 3, two matching matrices (one for each *Helitron* end) are created by applying these LCVs to known *Helitrons,* representing the distribution of conserved patterns in the training set (see *SI Appendix*, Fig. S1 for LCV distribution in the training set). Layer-2 predicts putative *Helitrons* in four steps (numbers in orange circles). In step 1, the input sequences in Fasta format are scanned by using the two sets of LCVs generated in layer-1. Two matching matrices are created in step 2, similar to step 3 in layer-1 for known *Helitrons.* Scores for both ends and their sum are calculated in step 3, along with matched locations in the input sequences. In step 4, after pairing two ends within a length range, putative *Helitrons* are drawn from the input sequences, with scores representing prediction confidence.

whether the *i*th LCV matches the jth sequence. Score S_j for the *j*th sequence is the number of matches or 1s in the *j*th row of the matching matrix. Each putative *Helitron* has two terminal scores from the 5′ and 3′ end matching matrices and a total score reflecting the prediction confidence.

Extracting LCVs from the Training Set. The LCV algorithm used here was first used in a study of DNA-binding helix-turn-helix motifs (22) and applied to other protein structure studies (31, 32). Here, we optimized the original LCV algorithm so that it is suitable to extract sequence motifs from DNA sequences.

To get overrepresented sequence patterns for *Helitron* identification, we created a training set consisting of two groups of 100-bp slices from the 5′ and 3′ ends of 5,676 published *Helitrons* (Fig. 4*A*). Then, we clustered these sequences by the cd-hit program (24) at 90% similarity and removed redundant ones to avoid LCV bias. Two sets of LCVs were extracted iteratively until at least 95% of sequences in the training set could be covered. See *SI Appendix, Methods* for details.

Identifying New *Helitrons*. HelitronScanner searches for matches of extracted LCVs in input DNA sequences and identifies *Helitrons* by pairing the 5′ and 3′ ends that meet the adjustable thresholds of LCV scores (see *SI Appendix, Methods*).

***Helitron* Copy Numbers.** *Helitron* copy number is defined as the number of hits with at least 90% sequence similarity obtained from BLASTing the 3′ terminal 50 bp of a given *Helitron* against its host genome sequence.

Estimation of False Positive *Helitrons*. For each species, randomized genomes were created by shuffling nucleotides within 1-Mb size sliding windows. *Helitrons* predicted on these randomized genomes are regarded as false positives and are excluded proportionally in Table 2 and *SI Appendix*, Table S2.

Data Access. The HelitronScanner tool with user manual and LCVs extracted from the training set are freely available at https://sourceforge.net/p/helitronscanner and http://bo.csam.montclair.edu/du/software/helitronscanner.

ACKNOWLEDGMENTS. We thank Yubin Li and Qinghua Wang for valuable suggestions on the manuscript. This work was supported by National Science Foundation (NSF) Grant 0923950 (to H.K.D. and C.D.) and NSF Grant 0959461 (to C.D.).

1. Wicker T, et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8(12):973–982.
2. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* 98(15):8714–8719.
3. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a helitron insertion. *Plant Cell* 15(2):381–391.
4. Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci USA* 102(25):9068–9073.
5. Xu JH, Messing J (2006) Maize haplotype with a helitron-amplified cytidine deaminase gene copy. *BMC Genet* 7:52.
6. Morgante M, et al. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 37(9):997–1002.
7. Hollister JD, Gaut BS (2007) Population and evolutionary dynamics of *Helitron* transposable elements in Arabidopsis thaliana. *Mol Biol Evol* 24(11):2515–2524.
8. Novick PA, Smith JD, Floumanhaft M, Ray DA, Boissinot S (2011) The evolution and diversity of DNA transposons in the genome of the Lizard Anolis carolinensis. *Genome Biol Evol* 3:1–14.
9. Han MJ, et al. (2013) Identification and evolution of the silkworm helitrons and their contribution to transcripts. *DNA Res* 20(5):471–484.
10. Brunner S, Pea G, Rafalski A (2005) Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J* 43(6):799–810.
11. Yang L, Bennetzen JL (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc Natl Acad Sci USA* 106(47):19922–19927.
12. Du C, Fefelova N, Caronna J, He L, Dooner HK (2009) The polychromatic Helitron landscape of the maize genome. *Proc Natl Acad Sci USA* 106(47):19916–19921.
13. Li Y, Dooner HK (2012) *Helitron Proliferation and Gene-Fragment Capture. Topics in Current Genetics: Plant Transposable Elements-Impact on Genome Structure and Function* (Springer, Berlin), Vol 24.
14. Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics* 9:51.
15. Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci USA* 106(31):12832–12837.
16. Sweredoski M, DeRose-Wilson L, Gaut BS (2008) A comparative computational analysis of nonautonomous helitron elements between maize and rice. *BMC Genomics* 9:467.
17. Tempel S, Nicolas J, El Amrani A, Couée I (2007) Model-based identification of Helitrons results in a new classification of their families in Arabidopsis thaliana. *Gene* 403(1-2): 18–28.
18. Solan Z, Horn D, Ruppin E, Edelman S (2005) Unsupervised learning of natural languages. *Proc Natl Acad Sci USA* 102(33):11629–11634.
19. Kunik V, et al. (2007) Functional representation of enzymes by specific peptides. *PLOS Comput Biol* 3(8):e167.
20. Nevill-Manning CG, Wu TD, Brutlag DL (1998) Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci USA* 95(11):5865–5871.
21. Schwartz D, Gygi SP (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* 23(11): 1391–1398.
22. Xiong W, Li T, Chen K, Tang K (2009) Local combinational variables: An approach used in DNA-binding helix-turn-helix motif prediction with sequence information. *Nucleic Acids Res* 37(17):5632–5640.
23. Goodstein DM, et al. (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* 40(Database issue, D1):D1178–D1186.
24. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.
25. Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* 99(14):9573–9578.
26. Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc Natl Acad Sci USA* 103(47): 17644–17649.
27. Feschotte C, Pritham EJ (2009) A cornucopia of Helitrons shapes the maize genome. *Proc Natl Acad Sci USA* 106(47):19747–19748.
28. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.
29. Goff SA, et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296(5565):92–100.
30. Yu J, et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296(5565):79–92.
31. Sun JM, Li TH, Cong PS, Tang SN, Xiong WW (2012) Retrieving backbone string neighbors provides insights into structural modeling of membrane proteins. *Mol Cell Proteomics* 11(7):016808.
32. Tang S, et al. (2013) PlantLoc: An accurate web server for predicting plant protein subcellular localization by substantiality motif. *Nucleic Acids Res* 41(Web Server issue):W441–W447.