The polychromatic *Helitron* landscape of the maize genome

Chunguang Du[a,1], Nadezhda Fefelova[a], Jason Caronna[a], Limei He[b], and Hugo K. Dooner[b,c,1]

[a]Department of Biology and Molecular Biology, Montclair State University, Montclair, NJ 07043; [b]Waksman Institute, Rutgers University, Piscataway, NJ 08854; and [c]Department of Plant Biology, Rutgers University, New Brunswick, NJ 08901

Contributed by Hugo K. Dooner, May 4, 2009 (sent for review April 21, 2009)

Maize *Helitron* transposons are intriguing because of their notable ability to capture gene fragments and move them around the genome. To document more extensively their variability and their contribution to the remarkable genome structure variation of present-day maize, we have analyzed their composition, copy number, timing of insertion, and chromosomal distribution. First, we searched 2.4 Gb of sequences generated by the Maize Genome Sequencing Project with our HelitronFinder program. We identified 2,791 putative nonautonomous *Helitrons* and manually curated a subset of 272. The predicted *Helitrons* measure 11.9 kb on average and carry from zero to nine gene fragments, captured from 376 different genes. Although the diversity of *Helitron* gene fragments in maize is greater than in other species, more than one-third of annotated *Helitrons* carry fragments derived from just one of two genes. Most members in these two subfamilies inserted in the genome less than one million years ago. Second, we conducted a BLASTN search of the maize sequence database with queries from two previously described agenic *Helitrons* not detected by HelitronFinder. Two large subfamilies of *Helitrons* or *Helitron*-related transposons were identified. One subfamily, termed *Cornucopious*, consists of thousands of copies of an ≈1.0-kb agenic *Helitron* that may be the most abundant transposon in maize. The second subfamily consists of >150 copies of a transposon-like sequence, termed *Heltir*, that has terminal inverted repeats resembling *Helitron* 3′ termini. Nonautonomous *Helitrons* make up at least 2% of the maize genome and most of those tested show +/− polymorphisms among modern inbred lines.

Cornucopious | fragmented genes | gene capture | *Heltir* | transposons

Helitrons are a class of eukaryotic transposable elements (TE) discovered only recently from a computational analysis of the fully sequenced genomes of *Arabidopsis*, rice, and *C. elegans* (1). These transposons vary greatly in abundance, even among closely related species. In fruit flies, for instance, they can comprise from 1 to 5% of the genome (2). Although they appear not to be present in humans (3), they can make up as much as 3% of the genome in other mammals (4). Among well studied plants, *Helitron* content has been estimated to be >2% in *Arabidopsis thaliana* (1), <0.01% in maize (5), and to range from 0.03 to 1.02% in different rice species (6). However, these values are most likely underestimates because *Helitrons* are hard to detect computationally given their lack of classical transposon structural features, such as terminal inverted repeats (TIRs) and target site duplications (TSDs).

The remarkable variation in genome structure among maize inbred lines has been attributed to recent LTR retrotransposon insertions and to the differential presence of genes or gene fragments (7, 8). Two *Helitrons*, *HelA* and *HelB*, carrying fragments from four different genes, accounted for all of the genic differences distinguishing the *bz1* locus haplotypes of inbred lines McC and B73 (9). From a comparison of allelic BAC contigs, Morgante et al. (10) calculated that ≈10,000 gene fragments were inserted differentially in the genomes of the inbreds B73 and Mo17. Eight of nine insertions analyzed appeared to be nonautonomous *Helitron*-like transposons containing gene fragments, leading the authors to conclude that *Helitrons* were responsible for most of the genic sequence diversity between those two maize inbreds. Wang and Dooner (11) compared vertically the *bz1* genomic regions of eight different inbred lines and land races that shared only 25% to 84% of their sequences and established that *Helitrons* could account for as much as 8% of the total sequence variation among haplotypes.

Most known maize *Helitrons* are relatively large (2–30 kb) nonautonomous elements containing gene fragments or pseudogenes (9, 10, 12, 13), although smaller elements (<1 kb) without gene fragments have also been identified (11). Some elements have fragments of a gene encoding a "RepHel" protein with replication initiator and DNA helicase domains (2). These elements are postulated to be derived by deletion from putative autonomous elements encoding an intact RepHel protein that presumably mobilizes *Helitrons* by rolling circle (RC) transposition (1, 10). However, evidence that this protein is involved in *Helitron* transposition is strictly circumstantial. In fact, it was recently reported that *Helitrons* can excise and leave footprints, an outcome not expected from RC transposition (14), so *Helitrons* may exhibit both copy-and-paste and cut-and-paste modes of transposition.

Whether *Helitrons* can capture an intact gene remains an open question. The presence within a *Helitron* of an almost complete maize gene has been reported for a cytidine deaminase and a P450 monooxygenase gene (15, 16), but there is no evidence that either is functional. Several *Helitrons* have been reported to produce chimeric transcripts that fuse exons from different genes (8–10, 12). Thus, *Helitrons* resemble the rice Pack-MULE TEs described by Jiang et al. (17) in producing chimeric transcripts from multiple gene fragments and their main importance may lie in their potential to create new plant genes through the rearrangement and fusion of fragments from multiple genomic loci.

Hollister and Gaut (18) investigated a class of abundant Arabidopsis *Helitron* elements named *Basho* and found that these elements had undergone rapid expansion and relatively few gene capture events. Sweredoski et al. (19) identified 23 *Helitrons* in maize and 552 *Helitrons* in rice by BLASTN searches of the nr database and concluded that there were clear differences between maize and rice with respect to gene capture. However, only a handful of maize *Helitrons* have been discovered by sequence alignments. Many more putative *Helitrons* were discovered recently by application of the HelitronFinder program to the GenBank nr database and several were confirmed empirically (20). Here, we have analyzed the *Helitron* content of the almost complete B73 inbred genome sequence generated by the Washington University in St. Louis Genome Sequencing Center. We find that the *Helitrons* of maize are more variable than those of any species analyzed to date. To gain a better understanding of their contribution to the highly polymorphic genome of present-day maize, we have analyzed

Author contributions: C.D. and H.K.D. designed research; C.D., N.F., J.C., L.H., and H.K.D. performed research; C.D., N.F., J.C., L.H., and H.K.D. analyzed data; and C.D. and H.K.D. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 19747.

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

This article contains supporting information online at www.pnas.org/cgi/content/full/0904742106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0904742106

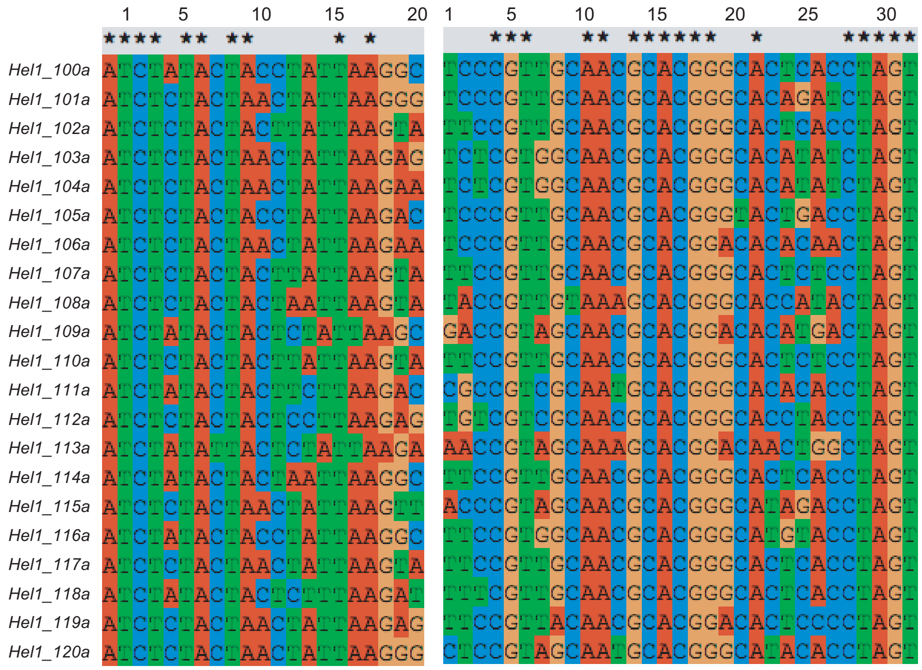
1 5 10 15 20
1 5 10 15 20 25 30
Hel1_100a ATCTATACTACCTATTAAGGC TCCCGTTGCAACGCACGGGCACTCACCTAGT
Hel1_101a ATCTCTACTAACTATTAAGGG TCCCGTTGCAACGCACGGGCACAGATCTAGT
Hel1_102a ATCTCTACTACTTATTAAGTA TTCCGTTGCAACGCACGGGCACTCACCTAGT
Hel1_103a ATCTCTACTAACTATTAAGAG TCTCGTGGCAACGCACGGGCACATATCTAGT
Hel1_104a ATCTCTACTAACTATTAAGAA TCTCGTGGCAACGCACGGGCACATATCTAGT
Hel1_105a ATCTCTACTACCTATTAAGAC TCCCGTTGCAACGCACGGGTACTGACCTAGT
Hel1_106a ATCTCTACTAACTATTAAGAA TCCCGTTGCAACGCACGGACACACAACTAGT
Hel1_107a ATCTCTACTACTTATTAAGTA TTCCGTTGCAACGCACGGGCACTCTCCTAGT
Hel1_108a ATCTCTACTACTAATTAAGTA TACCGTTGTAAAGCACGGGCACCATACTAGT
Hel1_109a ATCTATACTACTCTATTAAGC GACCGTAGCAACGCACGGACACATGACTAGT
Hel1_110a ATCTCTACTACTTATTAAGTA TTCCGTTGCAACGCACGGGCACTCTCCTAGT
Hel1_111a ATCTATACTACTTCTTAAGAC CGCCGTCGCAATGCACGGGCACACACCTAGT
Hel1_112a ATCTCTACTACTCCTTAAGAG TGTCGTCGCAACGCACGGGCACCTACCTAGT
Hel1_113a ATCTATATTACTCTATTAAGA AACCGTAGCAAAGCACGGACAACTGGCTAGT
Hel1_114a ATCTATACTACTAATTAAGGC TTCCGTTGCAACGCACGGGCACTCACCTAGT
Hel1_115a ATCTCTACTAACTATTAAGTT ACCCGTAGCAACGCACGGGCATAGACCTAGT
Hel1_116a ATCTATACTACCTATTAAGGC TTCCGTGGCAACGCACGGGCATGTACCTAGT
Hel1_117a ATCTCTACTAACTATTAAGTA TTCCGTTGCAACGCACGGGCACTCACCTAGT
Hel1_118a ATCTATACTACTCTTTAAGAT TTTCGTTGCAACGCACGGGCACTCACCTAGT
Hel1_119a ATCTCTACTAACTATTAAGAG TTCCGTTACAACGCACGGACACTCCCCTAGT
Hel1_120a ATCTCTACTAACTATTAAGGG CTCCGTAGCAATGCACGGGCATACACCTAGT

Fig. 1. Alignment of the 5′ terminal 20 bp (*Left*) and 3′ terminal 30 bp (*Right*) of the *Hel1* family. One representative of each of 21 *Hel1* subfamilies was chosen for the alignment. Asterisks above the figure mark sites with 100% consensus among all sequences.

their gene content, copy number, chromosomal distribution, and timing of insertion.

Results

Computational Identification and Analysis of Nonautonomous *Helitrons* in B73. We used HelitronFinder (20), a program designed to detect *Hel1* elements (21), to search 2.4 Gb of genome sequences from the maize inbred B73 (Release 2a.50, October, 2008). The program identifies putative *Helitrons* on the basis of conserved sequences at the *Hel1* 5′ and 3′ ends and assigns them to high, medium, or low probability categories, depending on the degree of sequence conservation. Among 16,205 BACs, we identified 2,791 putative nonautonomous *Helitrons* in the likeliest category. The maximum, minimum, median, and average sizes of the predicted *Helitrons* are 50 kb, 126 bp, 7,342 bp, and 11.9 kb, respectively. The 2,791 nonautonomous *Helitrons* make up 33.4 Mb or ≈1.39% of the sequenced B73 genome.

Alignment of the 5′ terminal 20 bp and 3′ terminal 30 bp of 21 *Helitrons* from different subfamilies (see below) revealed a high degree of conservation (Fig. 1). Analysis of the 3′ end, including the canonical palindromic sequence presumably involved in hairpin formation and transposition (1), confirmed that all of the identified *Helitrons* belong to the *Hel1* group (20). Over 70% of the identified palindromes started within 33 bp from the 3′ end.

Verification of Putative *Helitrons*. We selected nine predicted B73 *Helitrons* for verification by PCR in a panel of 13 inbreds. Presence of a PCR band of the size expected from a vacant site in at least one inbred line and its absence in B73 constitutes evidence of a +/− polymorphism in the panel (20). Primers for two predicted *Helitrons* gave nonspecific amplification and those for a third one failed to amplify fragments in any inbred, so these three were not considered further. The PCR results for the remaining six are summarized in Table 1. The polymorphic distribution of vacant and occupied sites for all six tested *Helitrons* establishes that they are not fixed residents of the maize genome and validates their prediction by HelitronFinder.

Fig. 2 shows the results of *Helitron* vacant site amplification for Silico_102: Inbred lines 4Co63, BSSS53, CML139, Mo17, McC, W22, and W23 produce a strong PCR band, showing that they lack Silico_102 at this genomic location and are, thus, polymorphic with B73. For each predicted *Helitron*, the vacant site nature of the PCR-amplified band was confirmed by sequencing. The distribution pattern for all six *Helitrons* tested differs among the 13 tested inbred lines, implying recent *Helitron* movements in maize.

An alternative approach to validate the predicted *Helitrons* is to compare duplicated regions of the maize genome. HelitronFinder predicts that there is a *Helitron* in chromosome 1 BAC AC190706. The AC190706 sequences 500 bp upstream and 400 bp downstream of the predicted *Helitron* are 100% identical with those of the homeologous region (22, 23) in chromosome 4 BAC AC212409 (Fig. S1).

Gene Fragments Captured by *Helitrons*. To identify *Helitrons* that had captured gene fragments, we performed a BLASTN search (24) against transcribed monocot sequences in the TIGR Database (25). Approximately 94% of the predicted *Helitrons* were found to carry fragments from transcribed genes. We further curated 272 of our 2,791 predicted *Helitrons*. Only 27 (10%) lacked gene fragments. Of these, the longest was 4,071 bp long, but 20 were smaller than 1 kb. Thus, most agenic *Helitrons* detected by HelitronFinder are smaller than 1 kb.

Table 1. Molecular verification of predicted *Helitrons*

Inbred line	Predicted *Helitrons*					
	Silico_101	Silico_102	Silico_103	Silico_106	Silico_107	Silico_110
4Co63	Vacant	Vacant	Occupied	Vacant	Vacant	Occupied
A188	Vacant	Occupied	Vacant	Occupied	Vacant	Occupied
BSSS53	Vacant	Vacant	Occupied	Vacant	Vacant	Occupied
B73	Occupied	Occupied	Occupied	Occupied	Occupied	Occupied
CML139	Vacant	Vacant	Occupied	Vacant	Vacant	Vacant
H99	Vacant	Occupied	Occupied	Occupied	Vacant	Occupied
I137TN	Vacant	Occupied	Occupied	Occupied	Vacant	Occupied
Ki3	Vacant	Vacant	Occupied	Vacant	Vacant	Occupied
M14	Vacant	Occupied	Occupied	Vacant	Vacant	Occupied
Mo17	Vacant	Vacant	Occupied	Occupied	Vacant	Occupied
McC	Vacant	Vacant	Occupied	Occupied	Vacant	Vacant
W22	Vacant	Vacant	Occupied	Occupied	Vacant	Vacant
W23	Occupied	Vacant	Occupied	Vacant	Vacant	Occupied

Vacant (no *Helitron*): PCR product obtained. Occupied (*Helitron* present): no PCR product. B73 is the positive control and H2O served as negative control.

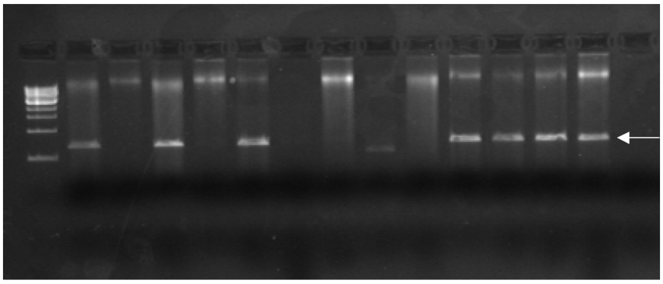

Fig. 2. PCR verification in a panel of 13 inbred lines for Silico_102. Lanes from 1 to 15 are: DNA marker, 4Co63, A188, BSSS53, B73, CML139, H99, I137TN, Ki3, M14, Mo17, McC, W22, W23, and H_2O, respectively. The arrow points to the *Helitron* vacant site band in inbreds that lack Silico_102 at this genomic location. The diffuse larger fluorescent band in some lanes corresponds to genomic DNA.

We carefully characterized a subset of 158 of the curated *Helitrons* in terms of percentage sequence identity among their 5′ and 3′ ends, copy number, size, and gene fragment content. The results obtained after removing duplicate entries from overlapping BAC sequences are summarized in Table 2. Elements sharing ≈90% end sequence identity grouped into subfamilies, which, in general, shared all or some gene fragments (Table S1). The size of the elements ranged from 0.2 kb to 39.2 kb, the number of gene fragments, from 0 to 8, and the copy number of elements within a subfamily, from 1 to 79.

Two subfamilies, *Hel1_105* and *Hel1_106*, were particularly numerous. The 79 members of the *Hel1_105* subfamily (Table S2A) contain sequences homologous to a rice gene encoding a phosphatase-like protein (Os05g0358500 NP_001055303). Most of the *Helitrons* in this subfamily carry a segment of the gene consisting of ≈180 bp from exon 2 and ≈240 bp from exon 3, but a few either lack the fragment from exon 3 or carry an additional 40-bp fragment from exon 4. The size of most *Helitrons* in this subfamily ranges from 1.5 to 2.5 kb, the sole exception being *Hel1_105Y3*, which includes an 8-kb retrotransposon. A BLASTN search of the B73 genome sequence with the 1,617-bp *Hel1_105a* element uncovered 1,505 additional members of the *Hel1_105* subfamily, making this the largest subfamily of gene-fragment-bearing *Helitrons* in B73 (Table S3A). HelitronFinder detected most of these elements, but assigned them a medium probability value because of polymorphisms in the 5′ and 3′ end sequences.

The second largest subfamily is *Hel1_106*, composed of *Helitrons* that have captured fragments of a *CesA1* cellulose synthase gene. Twenty-six members were identified among the manually curated *Helitrons* (Table S2B) and 58 additional members from a BLASTN search of the B73 sequence database using *Hel1_106a* as query (Table S3B). The average size of these elements is ≈4 kb. Most carry a 1.12-kb fragment of *CesA1* (NP_001104954.1 GI: 162460417) that extends from the middle of exon 1 to almost the 3′ end of intron 2. Some contain shorter fragments from the same *CesA* region. The much larger 39-kb *Hel1_106s* element contains, in addition, a retrotransposon insertion. In all of these cases, the introns separating adjacent exons have also been captured by *Helitrons*, allowing us to predict that the *CesA1* gene fragment was probably captured from a genomic copy in chromosome 8 (AC198758).

We manually annotated 100 additional *Helitrons* that had captured gene fragments (Table S4: *Hel1_200a* to *Hel1_311a*). We did not determine the copy number of this group because we were mainly interested in the diversity of maize genes that can contribute fragments to *Helitrons*. Some *Helitrons* appear to have two 5′ ends and only one 3′ end or vice versa (e.g., *Hel1_201a*). Although these apparently double-headed or double-tailed *Helitrons* may move as a unit, we only considered the internal *Helitron* component with single 5′ and 3′ ends in our analysis of gene content. These 100 *Helitrons* carried a variable number of gene fragments, ranging from one to nine. The size of the four *Helitrons* that captured nine gene fragments (*Hel1_303a*, *Hel1_302a*, *Hel1_299a*, and *Hel1_284a*) ranged from 13 to 34 kb. *Hel1_303a* carried fragments from nine genes, some of which are linked in the genome. The genes for flavonol synthase, TPR-like protein, and PDIL 1–5 genes are all on chromosome 9 at physical map locations of 22.55, 29.84, and 22.28 Mb, respectively. Similarly, four of the nine gene fragments in *Hel1_299a* come from genes in chromosome 3 that map at 31.10, 136.42, 31.56, and 41.70 Mb, respectively. However, each of these *Helitrons* also contained gene fragments from different chromosomes. The fragments captured by these 100 manually annotated *Helitrons* come from >200 different genes. Thus, the diversity of *Helitron*-embedded gene fragments is much greater in maize than in rice or *Arabidopsis*.

Our 272 curated *Helitrons* captured gene fragments from 376 different genes (Table S5). This indicates that maize *Helitron* sequences contain a wide variety of gene fragments. Genes encoding retrotransposon gag-pol proteins or fragments thereof are also abundant in *Helitrons*. Copies of elements with at least 70% sequence identity usually retain common gene fragments. The increase in length of related *Helitron* copies is often the result of retrotransposon insertions. Some possibly full-length genes may have been captured by *Helitrons*, such as a putative casein kinase I by *Hel1_309a*. The majority of these are putative genes and additional evidence is required to verify that they are functional.

Table 2. Characteristics of maize *Helitron* subfamilies

Helitron subfamilies	Copy number	5′ end identity, %	3′ end identity, %	Size, kb	No. of gene fragments
Hel1_100a-e	5	80–100	70–100	6.2–11.6	2–7
Hel1_101a-c	3	100	94–98	5.8–6.2	4–5
Hel1_102a	1	NA	NA	7.9	6
Hel1_103a	1	NA	NA	5.3	2
Hel1_104a	1	NA	NA	0.9	0
Hel1_105a-a4	79	61–100	78–100	1.4–10.4	1
Hel1_106a-z	26	57–100	86–100	3.8–39.2	1
Hel1_107a-b	2	100	100	8	4
Hel1_108a-b	2	100	100	3.6–9.9	3
Hel1_109a	1	NA	NA	39.4	6
Hel1_110a-c	3	100	100	8.8	3
Hel1_111a	1	NA	NA	15.1	2
Hel1_112a	1	NA	NA	6.6	1
Hel1_113a-b	2	100	100	2.1	0
Hel1_114a-b	2	85	88	8.2–11.0	2–4
Hel1_115a-b	2	100	100	2.7–23.5	1–2
Hel1_116a-b	2	100	98	13.1–13.6	7
Hel1_117a	1	NA	NA	4.5	1
Hel1_118a	1	NA	NA	16.4	5
Hel1_119a	1	NA	NA	2.1	1
Hel1_120a	1	NA	NA	7.9	5
Hel1_121a-d	4	100	100	2.6	1
Hel1_122a	1	NA	NA	10.9	7
Hel1_123a-b	2	61	90	0.5	2–3
Hel1_124a	1	NA	NA	17.6	7
Hel1_125a-h	8	95–100	82–94	0.9–7.2	0
Hel1_126a	1	NA	NA	14	8
Hel1_127a	1	NA	NA	12.3	3
Hel1_128a-b	2	95	85	0.2	0

NA, not applicable

***Helitron* Copy Numbers.** Among manually annotated *Helitrons*, smaller ones tend to have the highest number of copies (e.g., *Hel1_105*: Table 2). Outside of the two largest copy-number subfamilies, *Hel1_105* and *Hel1_106*, there are 14 subfamilies with one copy; 8 with two copies; 2 with three copies; and 1 each with four, five, and eight. Most single-copy *Helitrons* measure several kilobases and only one of them is smaller than 1 kb.

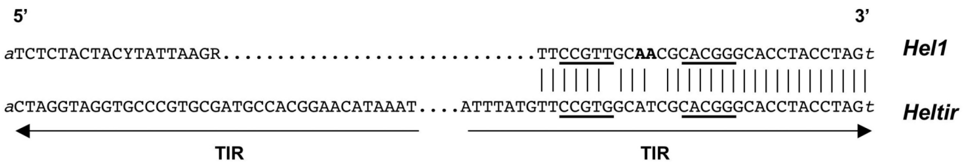
5'
3'
aTCTCTACTACYTATTAAGR...........................TTCCGTTGCAACGCACGGGCACCTACCTAGt
Hel1
aCTAGGTAGGTGCCCGTGCGATGCCACGGAACATAAAT....ATTTATGTTCCGTGGCATCGCACGGGCACCTACCTAGt
Heltir
TIR
TIR

Fig. 3. Comparison of *Hel1* and *Heltir* end sequences. The 5′ terminal 20 bp and 3′ terminal 30 bp of *Hel1* and *Heltir* elements are shown. The TIRs of the *Heltir* element are marked by an arrow and their sequence identity to the 3′ end of *Hel1* elements is indicated by vertical strokes. The palindromic sequences at the 3′ end of *Hel1* and both ends of *Heltirs* are underlined and the AA at the center of the *Hel1* palindrome is in bold.

The agenic *Helitrons Hel1-4* and *Hel1-5* discovered by Wang and Dooner (11) from a vertical comparison of *bz1* haplotypes were scored by HelitronFinder as being less probable than the 2,791 previously discussed *Helitrons* and were omitted from the above analysis. To obtain an estimate of the copy number and distribution of these *Helitrons* in B73, we conducted BLASTN searches against the current release of the maize genome sequence database (3b.50: Feb. 2009) using each of these sequences as queries. *Hel1-5* turned out to be not only the most numerous *Helitron*, but probably the most numerous transposon in maize. It is present in so many copies in the corn genome (≈3,000: Table S6) that we have renamed it *Cornucopious*. The *Cornucopious* subfamily consists of a variety of elements with different degrees of sequence identity to the original 994-bp *Hel1-5* element (11).

The 469-bp agenic *Hel1-4* element discovered in the *bz1* haplotype of A188 is not found intact in B73. *Hel1-4*-homologous sequences are present in several copies in B73, but they are split. The 5′ terminal 336 bp are present in 43 copies that cover >70% of the sequence at >90% identity. The 3′ terminal 104 bp are present in 158 copies that cover >66% of the sequence at >90% identity. Interestingly, these 158 copies are part of a larger (≈390-bp) sequence flanked by perfect 37-bp TIRs that are homologous to the 3′ end of *Hel1-4* and that we have designated *Heltir* for *Helitron*-like sequence with TIRs (Fig. 3). *Heltirs* have two *Helitron* 3′ ends, are almost invariably flanked by an A and a T (aCTAGGT..ACCTAGt), as would be expected from the canonical site of insertion of *Helitrons*, and are inserted in TA-rich regions. An example of a *Heltir* from chromosome 10 is given in Fig. S2.

Besides these two short elements, there are probably other agenic elements that were given a low probability score by HelitronFinder and could be present in multiple copies in the maize genome. The sequence variability of *Helitrons*, even at their relatively conserved termini, makes it difficult to estimate the true percentage of the maize genome that is made up by this transposon superfamily.

***Helitron* Divergence Time.** The divergence time of *Helitrons* cannot be estimated as readily as that of LTR retrotransposons because *Helitrons* lack duplicate sequences that become homogenized upon transposition. Hollister and Gaut (18) used terminal branch length (TBLs) as a proxy for estimating the time of *Helitron* insertions. We applied their method to estimate the time of insertion of elements in the two most abundant gene-capturing subfamilies, *Hel1_105* and *Hel1_106*. The majority of *Helitrons* in these two subfamilies inserted recently in the maize genome: 71% of *Hel1_105* elements and 69% of *Hel1_106* elements (Table S7) are <1 million years old and >96% of elements in both subfamilies are <5 million years old. This dating indicates that most recognizable *Helitrons* in the genome sequences of present day maize have transposed recently.

Dynamic Distribution of *Helitrons*. Most *Helitrons* appear to be proliferating actively and are able to capture genes or gene fragments from multiple locations in the genome. The BAC sequence AC183508 (Table 3) contains an interesting nested *Helitron*,

Table 3. *Helitrons* with multiple ends

Helitron	Accession no.	Start 5′_1*	Start 5′_2*	Start 5′_3*	End 3′_1*	End 3′_2*
Hel1_103a	AC190799.1	77100	76942	73315	71740	
Hel1_105 h2	AC200479.3	72973			71230	58668
Hel1_105a3	AC202878.2	99277	99572		101195	
Hel1_109a	AC149034.3	54590	51684		15101	
Hel1_201a	AC191691.3	170683	167108		160573	
Hel1_203a	AC183928.4	71437	62566		58077	
Hel1_206a	AC184164.2	130023	131805		138112	
Hel1_225a	AC185470.2	152279	161454		162631	
Hel1_231a	AC186166.3	95281	93840		92310	
Hel1_233a	AC186223.3	35296			38049	58820
Hel1_237a	AC186815.3	50586	54373		59681	
Hel1_239a	AC185211.3	90120	92586		94472	
Hel1_240a	AC186329.3	32604			42102	55571
Hel1_244a	AC177874.2	192724	183714		176844	
Hel1_245a	AC185499.2	161144			154517	154206
Hel1_248a	AC190964.2	168988			163728	160378
Hel1_257a	AC190498.1	24663	17072		15643	
Hel1_260a/261a	AC183508.3	112810	87052		78952	78616
Hel1_274a	AC187845.3	116665	117678		120258	
Hel1_276a	AC177926.2	139016			132784	120710
Hel1_285a	AC177904.2	155390			123713	105891
Hel1_289a	AC193322.3	41857			57972	58095
Hel1_292a	AC190626.3	142543	144299		149859	
Hel1_297a	AC186577.3	26879	27883		30462	
Hel1_298a	AC186577.3	39172	42750		50201	
Hel1_299a	AC186189.3	90063	86847		75250	
Hel1_301a	AC191802.3	26106			27813	68593
Hel1_306a	AC191705.3	128700	128105		127390	

*Location of the 5′ end first base or the 3′ end last base in the BAC sequence listed under the accession no. column.

Hel1_260a/*Hel1_261a* (Table S4). Although nested elements are common among LTR retrotransposons, this was the only *Helitron* nest detected in B73. However, some of the predicted *Helitrons* have multiple ends, either two or three 5′ ends or two 3′ ends (Table 3). *Hel1_103a* has three 5′ ends and 17 others have two 5′ ends, but all of them have one 3′ end. Nine *Helitrons* have two 3′ ends, but a single 5′ end. This lack of symmetry could be due to failure of the HelitronFinder program to detect a less conserved 5′ or 3′ end or, alternatively, to a real truncation of the elements. We recently found a case of two *Helitrons* adjacent to, rather than nested within, each other in the *r1* region from chromosome 10 (20) and here we have found an additional case in sequence AC187094 from chromosome 7 (Table S3A and Table S4).

To examine whether *Helitrons* were distributed randomly among the 10 maize chromosomes, we tested by χ^2 whether the distributions of the two most abundant subfamilies, *Cornucopious* and *Hel1_105*, deviated significantly from those expected based on chromosome lengths. Both of them did (Fig. S3). The calculated χ^2 values are 28.98 and 24.95, respectively, and the probability of obtaining deviations from expectation as large as or larger than those observed based on the null hypothesis is <0.01 in both cases, suggesting that the distribution of *Hel1* type *Helitrons* is nonrandom in maize. Interestingly, however, a χ^2 test of heterogeneity indicates that the two observed distributions do not differ significantly from each other ($\chi^2 = 16.86$, 9 df, $P > 0.05$)

Discussion

Nonautonomous *Helitrons* make up at least 2% of the maize genome, a much higher percentage than that reported in ref. 5. We estimate that there are ≈4,000 gene-fragment-bearing *Hel1 Helitrons* in B73 and that the average *Helitron* outside of the *Hel1_105* subfamily contains 1.75 genes. Thus, our number is in agreement with the ≈10,000 +/− gene sequence polymorphisms between B73 and Mo17 detected by Morgante et al. (10) and interpreted by them as being caused most likely by nonautonomous *Helitron*-like transposons. We are currently modifying our HelitronFinder program with the 2,791 predicted *Helitrons* as our training dataset in an attempt to make it less stringent, but still accurately predictive. A preliminary survey of the B73 sequence database with this less stringent version of HelitronFinder yields 5,000 additional predicted *Helitrons*. Furthermore, known short agenic *Helitrons*, which were not placed in the highest probability category of predicted *Helitrons* by our original HelitronFinder program, can be very numerous. A BLAST search with the agenic 0.9-kb *Hel1-5* element first identified in the *bz1* haplotype of inbred I137TN (11) revealed >3,000 copies in the B73 genome, making it the most abundant corn transposon and prompting us to rename it *Cornucopious*. Given the large number of copies of this element, we did not attempt to resolve duplicate entries in the htgs maize sequence database. Nevertheless, even assuming a 30% overlapping coverage in the sequence of the B73 genome (www.maizesequence.org/faq.html), the *Cornucopious Helitron* subfamily would still be represented by >2,000 members. It is clear from the above discussion that the percentage of *Helitrons* in the maize genome could be much higher than we report here.

The chromosomal distribution of the two largest subfamilies of *Hel1 Helitrons*, *Cornucopious* and *Hel1_105*, was found not to be proportional to chromosome length in either case. If *Helitrons* did transpose by RC replication and formed tandem arrays, as reported in *Myotis* (4), some chromosomes might accumulate a disproportionate number of a particular *Helitron* relative to other chromosomes. However, we did not observe *Helitron* concatemers and the elements in the two examples of adjacent *Helitrons* that we detected in the B73 genome differed from each other. Although the calculated χ^2 values are significant, the differences between the observed and expected distributions (Fig. S3) are not glaringly obvious and, in fact, the two observed distributions do not differ from each other. A plausible explanation of this result is that the maize genome is not yet finished and sequences from some chromosomes, such as 4 and 10, may be over-represented in the current sequence release (www.maizesequence.org/index.html).

Not all maize *Helitrons* were found to be inserted between an A and a T. We identified 18 *Helitrons* (6.6%) that were inserted into CT, GT or TT dinucleotides, instead of the characteristic AT dinucleotide. Elements inserted into a TT target site have been reported before (10). This finding has prompted us to remove the strict requirement for a specific nucleotide adjacent to the *Helitron* 5′ end from our HelitronFinder program to increase its ability to find maize *Hel1* elements.

Analysis of the gene content of *Helitrons* revealed that maize *Helitrons* have the capacity to capture a wide variety of gene fragments. These fragments showed different degrees of similarity with homologous genes found in maize or rice. The fragments captured by our set of 272 curated *Hel1 Helitrons* range in size from 28 bp to 7.6 kb and come from 376 different genes. In contrast, the gene fragments found in *Arabidopsis thaliana Basho* elements measure from 30 to 350 bp and derive from only five genes (18). Thus, maize *Helitrons* possess a greater, possibly species-specific, ability to capture gene fragments. Several of the annotated *Helitrons* may contain full-length genes. However, because many of those genes are hypothetical, putative or unknown and the corresponding full-length cDNAs have not been found, additional evidence is required to confirm that the proteins they encode are expressed and functional. Some elements captured *gag-pol* gene fragments, creating mimics of the retrotransposon insertions into *Helitrons* found by Morgante et al. (10) and by us here (e.g., *Hel1_105y3*). Other TE nests, such as *Helitrons* into LTR retrotransposons (15, 16) and DNA transposons into *Helitrons* (10) were also observed in this study.

Although *Helitron*-borne maize gene fragments were identified by their homology to rice genes, some are also homologous to previously annotated maize genes. In a few cases, the host gene contributing a specific fragment could be identified on the basis of high (>95%) sequence identity. *Helitrons* bearing fragments from multiple genes were analyzed for possible linkage of the donor genes. Some *Helitrons*, like *Hel1_303a* and *Hel1_299a*, appeared to have captured fragments from relatively closely linked donor genes, but adjacent gene fragments often came from different chromosomes. For example, *Hel1_299a* carries adjacent fragments from a chromosome 3 gene encoding a carbohydrate transporter and a chromosome 5 gene encoding an unknown protein. As noted by Morgante et al. (10), gene fragments tended to have the same transcriptional orientation. The overall ratio of 5′-3′ vs. 3′-5′ orientation of fragmented genes relative to the ends of the capturing *Helitron* was 4:1 (427:102) in 156 annotated maize elements.

The original model of gene capture by *Helitrons* (26) invoked a malfunction or deletion of the 3′ end RC terminator and the de novo formation of a terminator-like signal located downstream. Although the probability of random DNA functioning as a de novo terminator has been questioned (2), our study shows that the 3′ termini of maize *Helitrons* are more variable, their sequence identity ranging from 11% to 100%, whereas the 5′ termini are more conserved, their identity ranging from 57% to 100%. Our dataset includes several complex *Helitrons* with more than one 5′ or 3′ terminus. Their occurrence could be explained by deletion of either the 5′ or 3′ internal end from nested *Helitrons* or from closely linked *Helitrons*, as in the model of gene capture via two *Helitrons* proposed by Kapitonov and Jurka (2). Therefore, both possibilities regarding the occurrence of multiple 3′ end fragments could be valid. However, one aspect of gene capture not readily explained by either model is the strong asymmetry in the orientation of *Helitron*-captured gene fragments discussed above.

The agenic ≈0.4-kb *Hel1-4 Helitron* (11) is unusual in that related TIR-containing sequences, which we have named *Heltirs*, are found in the B73 genome in a high number of copies (≥158). Their structure and repetitive nature suggests that they are transposons.

Their TIRs are perfect, measure 37 bp, are highly similar to the conserved *Hel1-4* 3′ end, and are flanked by an A and a T, observations that suggest an unexpected relationship between *Helitrons* and TIR-flanked transposons (Fig. 3). If a TIR-binding, conventional transposase mobilizes *Heltirs* in maize, the same enzyme could be responsible for the recently reported maize *Helitron* excision footprints (14).

“Young” *Helitrons*, with an estimated time of insertion <1 million years, comprise 69% to 72% of the two largest subfamilies of gene-fragment-bearing *Helitrons*, *Hel1_105* and *Hel1_106*. “Old” *Helitrons*, with an estimated time of insertion >5 million years, comprise only 2.6 to 7.7% of the total. Our analysis reveals that an expansion of elements with a size between 1 and 2 kb occurred <1 million years ago. We found that small elements tend to have more copies, indicating that they transpose more easily. However, although a large fraction of small *Helitrons* inserted <1 mya, older members of the same subfamily are also observed, indicating that small *Helitrons* can persist in the genome for a long time. Thus, maize *Helitrons* represent a group of transposons that have actively reshaped the genome of present day maize. No statistically significant correlation between the size and age of *Helitrons* was observed. In contrast, Hollister and Gaut (18) noted that larger *Helitron* elements are less likely to persist in the *Arabidopsis* genome and suggested that this may be due to selection against the deleterious effects of inter*Helitron* ectopic recombination. Whether maize *Helitrons* can participate in such ectopic recombination events today is questionable. In one study designed to detect meiotic recombination between two pairs of *Helitrons* at allelic locations, no recombination events were detected and the elements were found to be densely methylated (27). The identification of additional new elements and the expanded curation of our dataset to include all predicted *Helitrons* will allow us to draw a clearer picture of the evolutionary dynamics of maize *Helitrons*.

With the maize genome sequencing and annotation on the way, it will be interesting to investigate further the relationship between *Helitron* age and size, the possibility of linkage among captured fragments, the distance of *Helitrons* to donor genes, and the gene density of donor chromosome segments. The sequencing of additional maize lines will reveal whether other *Helitron* transposons, different from the highly abundant *Cornucopious* of B73, have amplified to a high degree. If so, the differential escape from transposition suppression by the host genome of a specific transposon within a family would be reminiscent of the dramatic amplification of an *mPing* MITE described by Wessler and associates in four different rice strains (28).

Materials and Methods

Prediction of *Helitrons*. We identified *Helitrons* computationally with the HelitronFinder program (20). We then validated a sample of the predicted *Helitrons* by both molecular and in *silico* methods. The predicted *Helitrons* were used to retrain HelitronFinder so as to produce more generalized predictions. We annotated a portion of the predicted output *Helitrons* by blasting them against the htgs GenBank databases. Because of redundancy in the B73 GenBank entries, we individually checked the chromosomal clone location of curated *Helitrons* to eliminate potential duplicates.

Identification of Gene Fragments. We downloaded the TIGR Plant Transcript Assemblies (25) into our server. All 2,791 predicted *Helitrons* were compared by BLASTN (24) against the sequence databases of monocot ESTs, full length and partial cDNAs. We then curated a subset of 272 *Helitrons* to identify captured gene fragments. The GeneSeqer tool from MaizeGDB (http://maizegdb.org) was used to infer potential exon/intron structure in premRNA. The predicted protein sequence was compared by BLASTP against the nr/protein National Center for Biotechnology Information database. The predicted gene sequences were blasted against the *Helitron* sequences to identify the embedded gene fragments. We further conducted BLASTX searches against the nr protein sequences in the National Center for Biotechnology Information database to identify protein-coding regions with an e value cutoff of e^{-10}.

Estimation of the Time of Divergence of *Helitrons*. We adopted the terminal branch lengths method (18) to estimate the time of *Helitron* insertions, using a substitution rate of 1.3×10^{-8} per site per year for intergenic regions (29). We aligned the 3′ terminal 50 nucleotides of the *Hel1_105* and *Hel1_106 Helitrons* by ClustalX (30) and carried out a neighbor-joining phylogenetic analysis for each subfamily using the MEGA 4 (31) software package with 1,000 bootstrap replicates.

Molecular Validation of Predicted *Helitrons*. The inbreds used in the *Helitron* validation panel were: 4Co63, A188, BSSS53, B73, H99, M14, Mo17, W22, and W23, North American breeding lines; CML139, from CIMMYT; I137TN, from South Africa, and Ki3, from Thailand (32). McC carries the *Bz-McC* haplotype in a W22 background (9). Primer design, cloning, and sequencing were as described in ref. 20.

Note Added in Proof. The maize genome sequence data used in the present analysis are reported in: Schnable PS, et al. (2009) the B73 maize genome: Complexity, diversity and dynamics. *Science*, 10.1126/science.1178534.

ACKNOWLEDGMENTS. We thank Yubin Li, Qinghua Wang, and Jun Huang for valuable suggestions on the manuscript. Sequences provided by L. Yang and J. Bennetzen enabled the cross-referencing of *Helitrons* in Table S8. This work was supported by National Science Foundation Grant DBI-03-20683.

1. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* 98:8714–8719.
2. Kapitonov VV, Jurka J (2007) Helitrons on a roll: Eukaryotic rolling-circle transposons. *Trends Genet* 23:521–529.
3. Poulter RT, Goodwin TJ, Butler MI (2003) Vertebrate helentrons and other novel *Helitrons. Gene* 313:201–212.
4. Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus. Proc Natl Acad Sci USA* 104:1895–1900.
5. Messing J, et al. (2004) Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* 101:14349–14354.
6. Zuccolo A, et al. (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC Evol Biol* 7:152.
7. Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* 99:9573–9578.
8. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 17:343–360.
9. 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* 102:9068–9073.
10. Morgante M, et al. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 37:997–1002.
11. Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci USA* 103:17644–17649.
12. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a *Helitron* insertion. *Plant Cell* 15:381–391.
13. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK (2005) A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol Bio* 57:115–127.
14. Li Y, Dooner HK (2009) Excision of *Helitron* transposons in maize. *Genetics* 182:399–402.
15. Xu JH, Messing J (2006) Maize haplotype with a helitron-amplified cytidine deaminase gene copy. *BMC Genet* 7:52.
16. Jameson N, et al. (2008) *Helitron* mediated amplification of cytochrome P450 monooxygenase gene in maize. *Plant Mol Biol* 67:295–304.
17. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431:569–573.
18. Hollister JD, Gaut BS (2007) Population and evolutionary dynamics of *Helitron* transposable elements in *Arabidopsis thaliana. Mol Biol Evol* 24:2515–2524.
19. Sweredoski M, DeRose-Wilson L, Gaut BS (2008) A comparative computational analysis of nonautonomous helitron elements between maize and rice. *BMC Genomics* 9:467.
20. 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.
21. Dooner HK, Lal SK, Hannah LC (2007) Suggested guidelines for naming helitrons in maize. *Maize Genetics Coop Newslet* 81:24–25.
22. Helentjaris T, Weber D, Wright S (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics* 118:353–363.
23. Wilson WA, et al. (1999) Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* 153:453–473.
24. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res* 25:3389–3402.
25. Childs KL, et al. (2007) The TIGR Plant Transcript Assemblies database. *Nucl Acids Res* 35:D846–851.
26. Feschotte C, Wessler SR (2001) Treasures in the attic: Rolling circle transposons discovered in eukaryotic genomes. *Proc Natl Acad Sci USA* 98:8923–8924.
27. He L, Dooner HK (2009) Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for *Helitron* and retrotransposon insertions *Proc Natl Acad Sci USA* 106:8410–8416.
28. Naito K, et al. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci USA* 103:17620–17625.
29. Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* 101:12404–12410.
30. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
31. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* 9:299–306.
32. Liu K, et al. (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165:2117–2128.