DNA Research **19**, 11–21, (2012)
Advance Access Publication on 15 November 2011

doi:10.1093/dnares/dsr038

Mobilizing the Genome of Lepidoptera through Novel Sequence Gains and End Creation by Non-autonomous *Lep*1 *Helitrons*

Brad S. Coates[1,*], Richard L. Hellmich[1,2], David M. Grant[1,3], and Craig A. Abel[1,2]

USDA-ARS, Corn Insect and Crop Genetics Research Unit, 113 Genetics Laboratory, Iowa State University, Ames, IA 50011, USA[1]; Department of Entomology, Iowa State University, Ames, IA 50011, USA[2] and Department of Agronomy, Iowa State University, Ames, IA 50011, USA[3]

*To whom correspondence should be addressed. Tel. +515 294-0668. Fax. +515 294-2265.
Email: brad.coates@ars.udsa.gov

Edited by Yuji Kohara
(Received 14 June 2011; accepted 28 September 2011)

Abstract

Transposable elements (TEs) can affect the structure of genomes through their acquisition and transposition of novel DNA sequences. The 134-bp repetitive elements, *Lep*1, are conserved non-autonomous *Helitrons* in lepidopteran genomes that have characteristic 5′-CT and 3′-CTAY nucleotide termini, a 3′-terminal hairpin structure, a 5′- and 3′-subterminal inverted repeat (SIR), and integrations that occur between AT or TT nucleotides. *Lep*1 *Helitrons* have acquired and propagated sequences downstream of their 3′-CTAY termini that are 57–344-bp in length and have termini composed of a 3′-CTRR preceded by a 3′-hairpin structure and a region complementary to the 5′-SIR (3′-SIRb). Features of both the *Lep*1 *Helitron* and multiple acquired sequences indicate that secondary structures at the 3′-terminus may have a role in rolling circle replication or genome integration mechanisms, and are a prerequisite for novel end creation by *Helitron*-like TEs. The preferential integration of *Lep*1 *Helitrons* in proximity to gene-coding regions results in the creation of genetic novelty that is shown to impact gene structure and function through the introduction of novel exon sequence (exon shuffling). These findings are important in understanding the structural requirements of genomic DNA sequences that are acquired and transposed by *Helitron*-like TEs.

Key words: *Helitron*; sequence gain; genome rearrangement

1. Introduction

Helitrons are class II transposable elements (TEs) that are proposed to propagate by rolling circle replication (RCR) at the DNA level[1] and are dependent upon the function of Replicase/Helicase (RepHel) proteins for autonomous transposition.[1,2] *Helitrons* show a high degree of sequence plasticity, such that computational predications mainly rely upon the identification of conserved 5′-CT and 3′-CTRR termini,[3] a 6–20-bp stem-loop structure near the 3′-terminus,[4,5] and predicted integration between AT[1] or TT nucleotides.[3] Non-autonomous *Helitrons* are small in size due to lack of an internal protein-coding region, but often remain mobile through the retention of functional *trans*-acting RepHel proteins.[1,2] A high copy number of non-autonomous *Helitrons* within genomes likely results from evasion of host repression.[6] The observed paucities in sequence variation among *Helitrons* appears indicative of recent bursts in transposition.[7,8] Certain groups of class II DNA transposons and *Helitrons* integrate frequently in proximity to protein-coding regions[9] and can affect the structure and function of genes and gene products.[7,10] These integrations can result in the modification of transcriptional efficiency,[11] as well as introduce transcript splice variation and polyadenylation sites, changes in transcription start and stop sites, and incorporation of novel exon sequence.[12,13]

Helitrons are potent modifiers of genome structure and function due to frequent acquisition and

Published by Oxford University Press on behalf of Kazusa DNA Research Institute 2011

transposition of host genomic DNA, which oftentimes results in exon shuffling or the duplication of gene sequences.[4,14–18] The mechanism by which *Helitrons* acquire novel sequence remains largely unknown,[19] but is hypothesized to occur at the DNA level due to transposition of both intron and exon sequences. Furthermore, *Helitrons* are often chimeric constructs that have acquired DNA from multiple independent loci,[16,17] which may occur by a step-wise addition of novel 5′- and 3′-ends that are compatible with the minimal requirements for functioning during RCR.[6,10,20] These instances indicate that class II TEs and *Helitrons* participate in the rearrangement and duplication of genome regions and contribute to the evolution of novel eukaryotic genome functions.[21,22]

TE integration and excision mutations cause phenotypic variation at the individual and population scale,[23,24] may be contributing factors to speciation events[25,26] and provide the genetic novelties for local adaptation via natural selection.[27] The insect order Lepidoptera contains the second largest number of species on earth, some of which cause widespread damage to crop plants. Short repetitive sequences and TEs are important players in the generation of genetic diversity and evolution of Lepidoptera due to integrations within genes,[28,29] but the genome-wide affects of TE-derived mutations upon genetic and phenotypic variation remain relatively unknown. The silkmoth, *Bombyx mori*, is the lepidopteran model species whose ~420 Mb genome sequence assembly is composed of ~43.6% repetitive DNA,[30] with most being <500 bp.[31,32] Nearly 13% of the *B. mori* genome comprise short interspersed nuclear elements (SINEs),[33] which are class I TEs derived from tRNA, 5S rRNA or 7SL RNA-like sequences that propagate by retrotransposition.[34] In contrast, DNA-based class II TEs occupy ~3% of the *B. mori* genome (*Helitrons* 0.1%),[33] and are mostly located within introns and non-coding DNA of Lepidoptera.[35] This dearth of TE knowledge among lepidopteran species has resulted in difficulty in interpreting their role in the generation of structural and functional genome variance.

The *Lep*1 box is a conserved 99-bp repetitive DNA sequence originally described within intron and untranslated regions from eight lepidopteran species.[36] The *Lep*1 box retains homology +10 to −50 of the core repeat, which was later described as the 134-bp lepidopteran-specific common sequence 3 (LSCS3).[37] In the following, we annotate the *Lep*1 element as a *Helitron*-like TE and indicate that multiple novel DNA sequences have been acquired and propagated by the *Lep*1 *Helitron*. The conservation of predicted secondary structures between the ancestral *Helitron* and an acquired terminal region offers significant insight into the mode of *Helitron* propagation and features within the acquired sequence required for propagation. *Lep*1 *Helitrons* and acquired sequences co-localize with gene-coding regions in the *B. mori* genome, cause structural gene mutations, and are important mediators of genome-wide mutation.

2. Materials and methods

2.1. *Annotation of the Lep1 Helitron*

All annotations are made with respect to the reverse complement of the LSCS3 sequence[37] that includes the 99-bp *Lep*1 box[36] and from hereon is referred to as *Lep*1. Sequences showing homology to *Lep*1 were retrieved from the GenBank non-redundant (nr) nucleotide database via a BLASTn search that used *Lep*1 as the query (conducted 11-04-2010), with output filtered for ≥70% homology over ≥100 bp. Similarly, GenBank dbEST accessions for species of Lepidoptera were downloaded in FASTA format (10 November 2010), imported into a local databases using BioEdit,[38] queried using *Lep*1, and results filtered and aligned as described previously. All lepidopteran DNA sequence accessions that passed filter criteria were downloaded in FASTA format, imported into the MEGA 5.0 sequence alignment application,[39] and a multiple sequence alignment was made using the ClustalW algorithm (default parameters: gap opening penalty, 15; gap extension penalty, 6.66; weight matrix IUB, and transition weight, 0.5) in the MEGA 5.0 alignment module.[39]

Sequence homologies flanking the *Lep*1 sequence were identified by performing an ‘all vs. all’ search using the BLASTn algorithm using all nr and dbEST ‘hits’ to *Lep*1. The regions of intraspecific DNA sequence homology were extracted from accessions using a custom PERL script, and then used as input into the Mfold DNA secondary structure server[40] (http://mfold.rna.albany.edu/?q=mfold/DNA-Folding-Form) with the partial function and pair probabilities = 25°C.

2.2. *Estimation of Lep1 genome copy number and distribution*

Scaffolds from the *B. mori* whole genome sequence build v. 2.3 were downloaded from Kaikobase (http://sgp.dna.affrc.go.jp/pubdata/genomicsequences.html; file assembledset.txt.gz) and imported into a local database using BioEdit.[38] Build v. 2.3 was searched with *Lep*1 as the query using the BLASTn algorithm and results filtered for ‘hits’ showing ≥80% similarity over ≥50 bp. The putative Bm*Lep*1 integration positions were called the Bm*Lep*1 model v. 2.3, which was then merged with positions of the *B. mori* assembly v. 2.3 gene models (file: glean_cds_on_chr.gff at http://sgp.dna.affrc.go.jp/pubdata/genomicsequences.html). The combined features were displayed using CMap.[41] Sequence intervals for Bm*Lep*1 elements were retrieved from the assembled *B. mori* scaffold for chromosome 1 (Z chromosome) using a custom PERL script, reverse complements generated using the Sequence Manipulation Suite (SMS) at http://www.bioinformatics.org/sms2/rev_comp.html, and FASTA formatted text imported into MEGA 5.0.[39] A multiple sequence alignment was made for all *B. mori Lep*1 elements using parameters described earlier, gamma parameter estimated, and a maximum likelihood-based estimation of *Lep*1 phylogenetic relationship made using the general time reversible model of sequence evolution. Nucleotide sites were chosen using a partial deletion of missing characters (cut-off $= 0.05$), and all possible trees were interrogated using the Close-Neighbor-Interchange heuristic. Node support was acquired using 1000 bootstrap pseudoreplicates reported within a strict consensus tree.

The frequency of *Lep*1 integrations within species of Lepidoptera was investigated using full bacterial artificial chromosome (BAC) insert sequences from *Bicyclus anynana*, *B. mori*, *Heliconius melpomene*, *H. numata*, *Helicoverpa armigera*, *Papilio dardanus*, and *Spodoptera frugiperda*. These sequences were downloaded from NCBI, imported into BioEdit,[38] and a search for *Lep*1 positions performed as described previously. The mean frequency of *Lep*1 integrations were calculated manually and significance of frequency differences among species was assessed using *F*-statistics (significance threshold $\alpha = 0.05$).

2.3. *Predictions of haplotype variation caused by Lep1 elements*

Nucleotide accessions from the NCBI nr database were used to query derived protein sequences within the nr protein database using the blastx algorithm and results filtered for 'hits' showing $\geq$50% identity. Proteins derived from orthologous lepidopteran genes that were not identified within the initial *Lep*1 screen (see 2.1), but present within the blastx output were compared manually to predict instances of copy number variation within or between species. *Lep*1copy number variation at orthologous loci among species was further investigated using integration/excision variation among cadherin gene sequences from *B. mori* (gene model: BTGIBMGA013616), *H. armigera* (GenBank accession: AY714876.1), and *Ostrinia nubilalis* (DQ000165.1). Sequences were imported into a local database in BioEdit,[38] searched using with *Lep*1 as the query and alignment of exon 1, intron 1, and exon 2 using the MEGA 5.0 sequence alignment application[39] was created as described previously.

2.4. *Predictions of Lep1 modification to gene structure De novo*

Acquisition of transcribed *Lep*1 sequences was performed for *O. nubilalis*, where total RNA was isolated from whole larvae using RNAagents kit (Promega, Madison, WI) and cDNA was synthesized using the SMART cDNA Synthesis Kit (Clontech, Mountain View, CA) according to manufacturer's instructions, RACE reactions using the CDSIII primer (Clontech) with On*Lep*1-f2 (5′-TAC TRA TAT TAT AAA GCT GAA GAG TT-3′) and SMART V with On*Lep*1-r (5′-GAT AAA TGG GCT ATC TAA CAC TGA AAG-3′), and amplified according to manufacturer's instructions. PCR fragments were cloned, and sequenced using T7 and SP6 primers, and resulting sequence data were assembled into contigs using CAPs[42] as described previously. Contigs were annotated using the Blast2Go suite,[43,44] where the GenBank nr protein database was interrogated using the BLASTx algorithm.

3. Results and discussion

3.1. *Annotation of the Lep1 Helitron*

A conserved 99-bp *Lep*1 box was previously described in eight lepidopteran species[3,6] and later described as portion of a 134-bp consensus LSCS3.[3,7] Our homology-based searches of the GenBank nr/nt database resulted in the estimation of 618 regions within 210 nucleotide sequence accessions from Lepidoptera that show $\geq$70.0% similarity to *Lep*1 (32 species; mean similarity: $80.3 \pm 5.8\%$ over 120.5 ± 19.0 bp; Supplementary Table S1). Eighteen of these *Lep*1-containing accessions were annotated as microsatellite loci and 51 *Lep*1s were within introns or untranslated regions of known genes. The remaining 526 *Lep*1s were within un-annotated sequences from BAC full inserts of the lepidopteran species *B. mori*, *B. anynana*, *S. frugiperda*, *H. armigera*, *H. melpomene*, *H. erato* and *P. dardanus*. An analogous search of the GenBank dbEST database identified 443 accessions from lepidopteran species that showed $\geq$71.1% interspecific similarity with 84 of these dbEST hits being $\geq$130 bp (Supplementary Table S2). Multiple sequence alignment of GenBank nr/nt and dbEST accessions that contained full-length *Lep*1 elements resulted in a 197-bp consensus that shared *Helitron*-like 5′-CT and 3′-CTRY termini,[1] and were, respectively, designated as region H1 and region H2 of *Lep*1 (Fig. 1A). The nucleotides directly adjacent to the 5′- and 3′-ends consisted of TA or AA in 96.2% of predicted *Lep*1s and showed not discernable target site duplications which are consistent

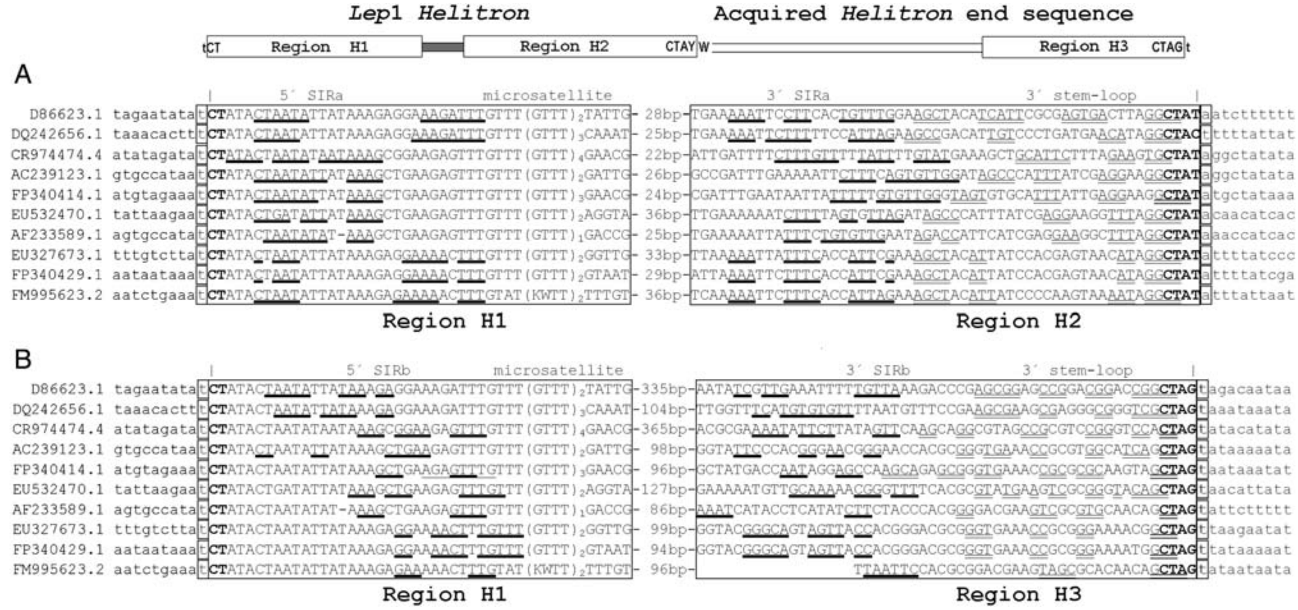


Figure 1. Alignment of *Lep*1 *Helitron* sequences from species of Lepidoptera. (A) Termini are defined by 5′-CT and 3′-CTRY motifs, and show secondary structures that are formed between a 5′-SIR that is complementary to a 3′-SIR (underline) and a 6–7-bp hairpin is present at the 3′-terminus (3′-stem-loop; double underlined). Additionally, *Lep*1s contain a hitchhiking $(GTTT)_n$ repeat microsatellite located between the SIRs. Flanking genomic sequence (small caps) that indicates integration occurs between TT or TA dinucleotides (enclosed in boxes). Corresponding secondary structures for *Bombyx mori* and *Heliconius melpomene Lep*1s are in Fig. 2. (B) Alignment of the same 5′-region of the *Lep*1 *Helitron* as in Fig. 1A, showing base secondary structure elements between an alternate 5′-SIR (5′-SIRa) and a 3′-SIR (3′-SIRa) within downstream acquired *Helitron* regions and a new 3′-stem-loop.

with previously described *Helitron* genomic integration events.[1,3,7]

Structural homology among *Helitrons* is often used for prediction and characterization, and is based upon a conserved 3′-stem-loop (hairpin) formed upstream of the 3′-CTRR terminus.[4,5,8] We predicted that a 3′-stem-loop would form at the 3′ *Lep*1 terminus in *B. mori* accession D86623.1 [AATCT ACATCATTCGCGAGTGACTTAGGCTA] (nucleotides involved in base pairing are double underlined) with a Gibbs free-energy change $(\Delta G) = -2.82\ \mathrm{kcal\ mol^{-1}}$ (Figs 1A and 2A), as well as at the 3′-terminus other lepidopteran *Lep*1s (Fig. 1A; not all data shown) including *H. melpomene* (Fig. 2B; $\Delta G = -3.67\ \mathrm{kcal\ mol^{-1}}$). In addition to 5′-CT and 3′-CTRY termini, the formation of a 3′-stem-loop (hairpin) near the 3′-terminus are hallmarks of *Helitron* TEs and suggest important roles of these structures in RCR or genome integration.[6,10,20] A second conserved stem-loop (hairpin) structure was predicted from *B. mori* and *H. melpomene Lep*1s that we designated the 5′-inverted repeat (5′-IR). This 5′-IR involves base pairing between portions of the 5′-subterminal inverted repeat (SIRa) and the microsatellite loop, and are analogous to the 5′-IR formed by the *Helitron*-like *Drosophila* interspersed nuclear element (*DINE*-1) family of TEs[7] and the lepidopteran microsatellite associated interspersed nuclear element (*MINE*-1).[35] Specifically, the *B. mori* GenBank accession D86623.1 is predicted to have a 7-bp 5′-IR formed by a portion of the 5′-SIRa sequence [ACTAATATT-7nt-GGAAAGATTTGTTT] (nucleotides involved in intramolecular base pairing are underlined; $\Delta G = -0.83\ \mathrm{kcal\ mol^{-1}}$; Fig. 2A), and the 6 bp 5′-IR in *H. melpomene Lep*1s are formed between later six nucleotides of the 5′-SIRa and the $(GTTT)_n$ microsatellite loop ($\Delta G = -0.99\ \mathrm{kcal\ mol^{-1}}$; Fig. 2B). The conservation of hairpin structures adjacent to both the 5′- and 3′-termini among 460 of 618 (74.4%) of *Lep*1s predicted among nr database accessions suggests a role in the RepHel protein recognition, nascent strand cleavage, or other portion of the RCR mechanism,[5,7,10,18,20] but further investigation is required to elucidate their role.

Structural predictions for *Lep*1 further indicated that intramolecular base pairing may occur between nucleotides immediately downstream of the 5′-CT and a region 26–37 nt upstream of the 3′-CTRY terminus and involve interaction of the SIRa. Specifically, nucleotides of the 5′-SIRa from *B. mori* accession D86623.1 [ACTAATATT-7nt-ATAAAGATTTGTTT] are predicted to pair with the 3′-SIRa [AAAAT TCTTTTCCATTAGA] located 49 bp downstream (nucleotides involved in base pairing are underlined in; $\Delta G = -5.92\ \mathrm{kcal\ mol^{-1}}$; Fig. 1A). Analogous

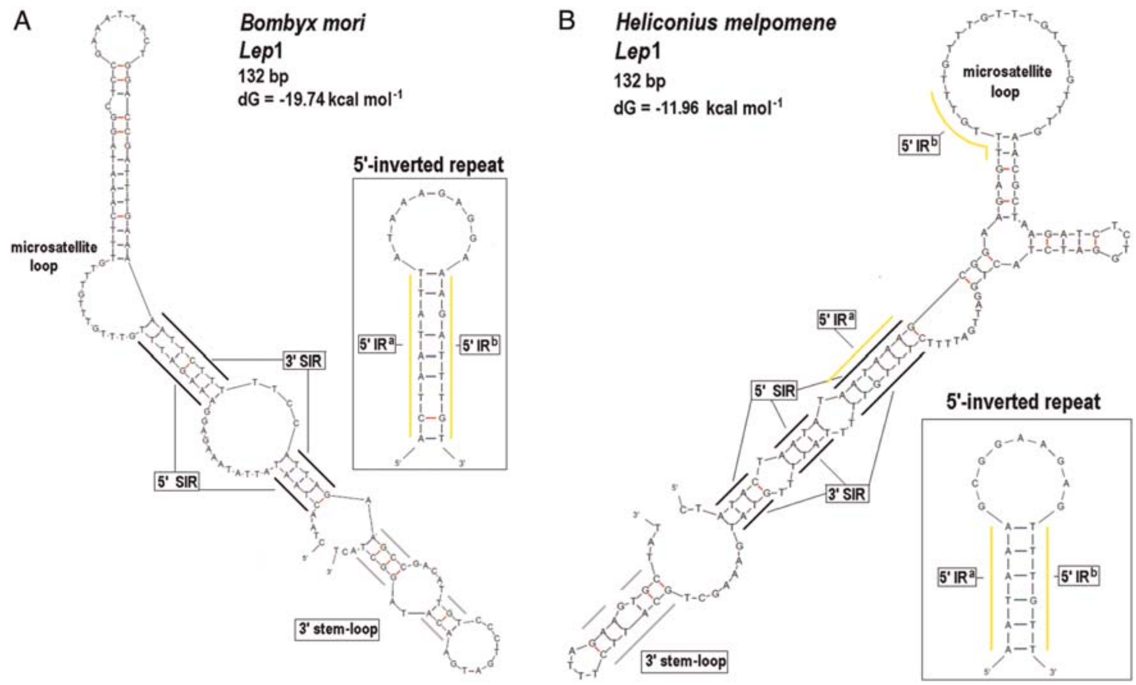


Figure 2. The secondary structure conserved among *Lep*1 elements. (A) *Bombyx mori* GenBank accession D86623.1 and (B) *Heliconius melpomene* GenBank accession CR974474.4 positions 69 560–69 971. *Helitron*-like features include 5′-CT and 3′-CTRY termini, a 3′-stem-loop is located two to three nucleotides upstream of the 3′-CTRY terminus (nucleotide with parallel grey bars), and a 5′-SIR that is complementary to a 3′-SIR (nucleotide with parallel black bars). Nucleotides involved in the formation of a 5′-IR are paralleled by yellow bars and the structure is shown within the boxed area for *H. melpomene*.

structures were predicted from *H. melpomene* (Figs 1A and 2B) as well as all other full-length *Lep*1s (Fig. 1A; all data not shown). These molecular interactions of *Lep*1 contribute to an overall secondary structure that is stable at 25°C for *B. mori* ($\Delta G = -19.74$ kcal mol^{-1}) and *H. melpomene* ($\Delta G = -11.90$ kcal mol^{-1}; Fig. 2), and is analogous to the structure formed by the *MINE*-1 *Helitron* from Lepidoptera.[35]

3.2. Annotation of acquired end sequences

Helitrons modify the structure of genomes through the acquisition and transposition of novel DNA sequence that occurs by a largely unresolved mechanism.[19,20] We described the consensus *Lep*1 element sequence from 31 accession for species of Lepidoptera that shared *Helitron*-like termini at the border of regions H1 and H2 (Fig. 1A), but also identified 62–342-bp sequence regions downstream of the 3′-CTRY terminus at the boundary of region H2 that are shared within a species. The novel shared sequences downstream of the 3′-CTRY were referred to as region H3. Specifically, the region H3 showed ≤62.3% similarity between species and ≥95.7% similarity within a species or closely related species such as between the 55- and 59-bp region H3 from *Helicoverpa* species *H. armigera* (GenBank accession: FP340429.1) and *H. zea* (EU327673.1; Fig. 1B). Among region H3 sequences from the same species or closely related species, sequence similarity terminated at a ubiquitous 3′-CTAG motif that was followed by a thymidine nucleotide (T; Figs 1B and 3), and only one variant of the H3 region was described within a species. The *Lep*1s predicted from the *B. mori* genome assembly shows two unique 87- or 335-bp sequences within region H3 that were, respectively, represented by GenBank accessions: DQ242656.1 and D86623.1 (Fig. 1B). The *B. mori Lep*1 *Helitrons* with an 87- or 335-bp region H3 were subsequently called Bm*Lep*1_87 and Bm*Lep*1_335 variants, respectively, and alignments showed no discernable homology (Fig. 3). Phylogenetic reconstruction of *B. mori Lep*1_87 and _335-bp *Helitron* variants from chromosome 1 suggested that two weakly supported clades may exist, which indicated that *Lep*1s have evolved independently within the *B. mori* genome through the acquisition of two different downstream sequences (Supplementary Fig. S2; gamma parameter rate distribution $\gamma = 3.9894$; Log likelihood $= D = -3729.34$). Independent gain of sequence mutations has previously been identified for *Helitrons* in the maize genome[6,10,20] and indicate that arthropod genomes are also modified by *Helitron* movements.

Although the sequences in region H3 share little interspecific sequence similarity, the secondary structures predicted to form appear to be identical by state

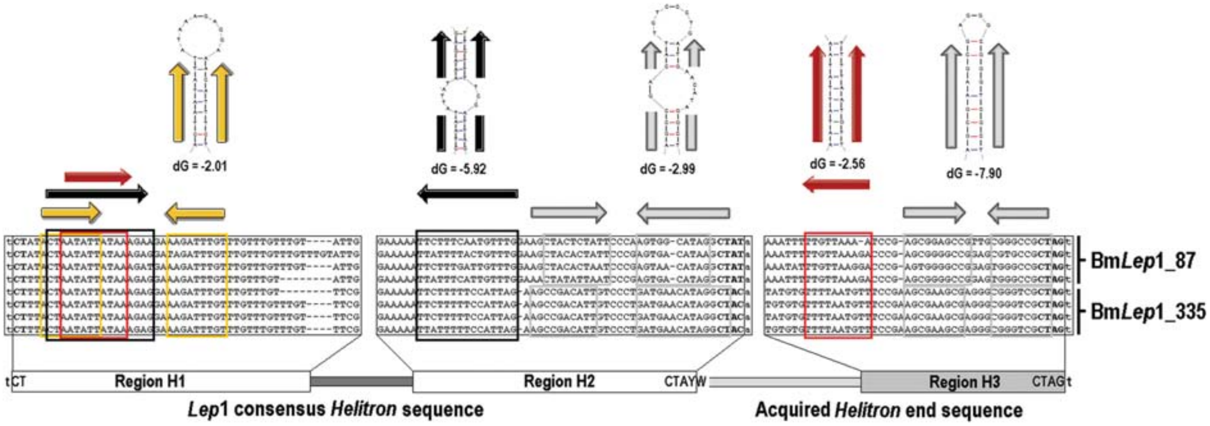


Figure 3. Alignment of nucleotides for secondary structure of *Bombyx mori Lep*1 *Helitrons*. The 5′-SIR form alternate hairpin structures with either a 5′-invered repeat (5′-IR; yellow arrows) or a 3′-SIR within the *Lep*1 consensus *Helitron* (black arrows), as well as a 3′-SIR within the acquired *Helitron* end sequence (red arrows) within both the 87- and 355-bp acquired regions. Two hairpins are formed, one at the 3′-terminus of the *Lep*1 *Helitron* and one within each of the 87- or 335-bp acquired *Helitron* end sequence (grey arrows). The 3′-termini of the ancestral *Lep*1 *Helitron* and the acquired end sequences are, respectively, composed of CTAY and CTAG motifs (yellow highlighted nucleotides).

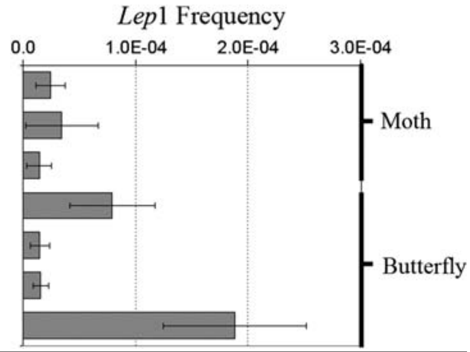
with those formed in region H2. Specifically, the 7–9-bp 3′-stem-loops (hairpins), respectively, formed in region H3 of Bm*Lep*1_87 and Bm*Lep*1_335 *Helitrons* ($\Delta G \leq -7.90$ kcal mol^{-1}) are more highly stable compared with the analogous structure in the ancestral region H2 ($\Delta G \leq -2.99$ kcal mol^{-1}; Fig. 3). This evidence may suggest that *Helitrons* are dependent upon a 3′-stem-loop directly upstream of the 3′-CTAG terminus for RCR function,[5] and that acquired sequences undergo selection for the capacity to support propagation by RCR.[20] A functional switch from use of ancestral to derived *Lep*1 *Helitron* ends may have been influenced by the comparatively higher stability we predicted for 3′-stem-loops within the *Lep*1 acquired sequence. Thereby, functional shifts could occur when equally or more efficient terminal structure are encountered by change within flanking DNA or swapped between other *Helitrons*. In contrast, reduction in 3′-stem-loop stability in region H2 also could have resulted from degradation following the relaxation of selective constraints, and would mirror the degradation of a *Helitron*-like 3′-CTAG terminus that followed accretion by maize *Helitrons*.[20]

Additionally, intramolecular base pairings we predicted between the 5′-SIRa (region H1) and 3′-SIRa (region H2; Fig. 1A) are analogously formed through interaction of nucleotides within the 5′-SIRa of the ancestral *Lep*1 region H1 and a 3′-SIRb within the derived region H3 (Fig. 1B). These interactions between the 5′-SIRb and 3′-SIRb in *B. mori Lep*1_87 and _335-bp *Helitrons* ($\Delta G \leq -2.56$ kcal mol^{-1}) is lower than between the 5′-SIRa and 3′-SIRa ($\Delta G \leq -5.92$ kcal mol^{-1}; Fig. 3), but remain consistent with the characteristic secondary structures described previously for insect *Helitrons*.[7,45] The formation of a hairpin between the 5′-SIR and 3′-SIRs within both the ancestral and acquired regions indicates that, in addition to the 3′-stem-loop structure, the base pairing between SIRs at proximal and distal ends of *Lep*1 may potentially be required from RCR. Furthermore, the requirement for a 3′-SIR within the independently acquired DNA sequences in proximity to a 3′-CTAG novel terminal motif could suggest that gain of sequence mutations by *Lep*1 may be rare. This hypothesis could be supported by our description of two *Helitron* variations in the *B. mori* genome, which contrasts with the high number of cryptic *Helitrons* from the maize genome where functional constraints appear to be relaxed.[17]

The described accretion by lepidopteran *Helitrons* involves sequences that are relatively small and added in a unidirectional fashion. DNA sequence upstream of the single 5′-CT terminal motif showed no homology among integrations in lepidopteran genomes or conserved secondary structures that would be indicative of a chimeric *Helitron*.[46] These observations are analogous to those described by Yang and Bennetzen.[8] *Lep*1 was not shown to capture entire genes, but may be due to inability to accurately describe haplotype variation from available data resources or the possible culling of the mutations from genomes due to negative affects on genome function.[47] Smaller non-autonomous *Helitrons* are better able to evade host repression[6] or show greater replication efficiency within the RCR mechanism.[48] *Lep*1s also appears to gain sequence only at the 3′ end, which contrasts with bidirectional end creation by maize *Helitrons*.[20] This directionality of *Lep*1s may result from the preferential capture that is known to occur in the same orientation as the RepHel-

Table 1. Estimated frequency of *Lep*1 *Helitrons* per mega base (Mb) in species of Lepidoptera

Species	No. BACs	Total Mb	No. *Lep*1s	*Lep*1 frequency
B. mori	50	7.5	191	$2.5 \times 10^{-5} \pm 1.3 \times 10^{-5}$
H. armigera	18	1.96	67	$3.4 \times 10^{-5} \pm 3.2 \times 10^{-5}$
S. frugiperda	12	1.47	20	$1.4 \times 10^{-5} \pm 1.1 \times 10^{-5}$
B. anynana	11	1.30	103	$7.9 \times 10^{-5} \pm 3.8 \times 10^{-5}$
H. melpomene	6	0.86	10	$1.5 \times 10^{-5} \pm 8.8 \times 10^{-6}$
H. numata	3	0.19	3	$1.6 \times 10^{-5} \pm 7.0 \times 10^{-6}$
P. dardanus	4	0.57	104	$1.8 \times 10^{-5} \pm 3.3 \times 10^{-5}$



Estimates from BAC full-insert sequences. Respective GenBank accessions and positional information is provided in Supplementary Table S1.

coding sequence of the autonomous *Helitron*, but this cannot be ascertained until further research is preformed to identify the parent TE of the non-autonomous *Lep*1 element.

3.3. *Lep1 copy number and genome distribution*

The variation in non-autonomous *Helitrons* copy number among species of Lepidoptera may result from differential effects of replicative repression via DNA methylation,[49] mutation-selection balance within the overall genome architecture,[50] or random genetic drift at the population scale. We showed that *Lep*1 *Helitron* integration densities range from 1.04×10^{-5} to 1.8×10^{-4} based upon annotation of full BAC insert sequences. These estimates indicated that *Lep*1s have an ~13-fold copy number difference across species' genomes (Table 1), and is comparable to the ~11-fold variance observed among *DINE*-1 insertions within *Drosophila* genomes.[7] The highest *Lep*1 copy number density was estimated for the butterfly *P. dardanus*, wherein *Lep*1 abundance was significantly higher than for all other species except for *B. anynana* (P-values ≥ 0.0089; Student's-t values not shown). The mean *Lep*1 integration density among butterflies ($6.9 \times 10^{-5} \pm 6.8 \times 10^{-5}$) is not significantly different from the densities estimated among moth species (mean: $2.5 \times 10^{-5} \pm 2.0 \times 10^{-5}$; F-statistic = 1.4; d.f.$_{num}$ = 1; d.f.$_{den}$ = 2; P-value = 0.3583).

The only whole genome sequence available for use to directly estimate the TE frequencies in Lepidoptera is the 432 Mb *B. mori* assembly (build v.2.3),[30] from which we estimated 5541 putative Bm*Lep*1 integrations. Each putative Bm*Lep*1 *Helitron* was assigned a unique identifier (Bm*Lep*1_000001 to Bm*Lep*1_005541), and further categorized as containing either the Bm*Lep*1_87 or Bm*Lep*1_335 variant downstream sequence described previously

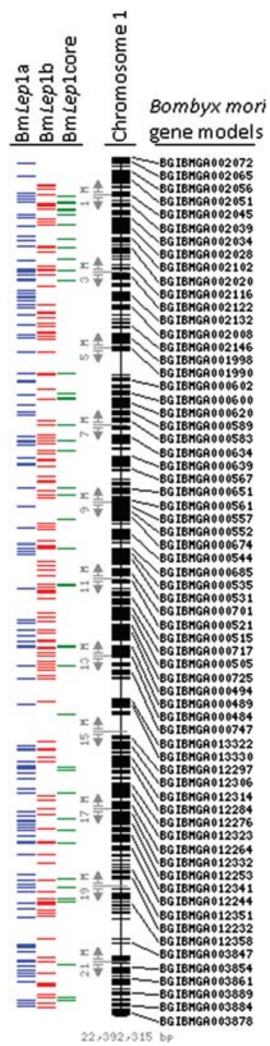


Figure 4. *Lep*1 *Helitron* integration positions on the *Bombyx mori* chromosome 1. *Lep*1 model v. 2.3 showing positions of the whole genome assembly via a CMap output. Integrations are classified into *Lep*1s that have acquired novel 87- or 335-bp end sequences (positional data for all 28 linkage groups present in Supplementary Table S3).

(Supplementary Table S3). The ancestral regions of the *Lep*1 *Helitron* (regions H1 thru H2;) within build v. 2.3 showed $\geq 94.3 \pm 1.9\%$ similarity with the *B. mori Lep*1 elements from accessions DQ242656.1 and D86623.1 (Fig. 1B). This paucity of *Lep*1 sequence evolution within the *B. mori* genome may indicate a recent burst in transposition,[7,8] or a high degree of functional conservation. The density of *B. mori Lep*1s estimated from the genome sequence (1.3×10^{-5}) is ~2-fold lower than that estimated from BAC full inserts (2.5×10^{-5}; Table 1), and highlights the error that may be associated with subsampling from BAC sequences. Mapping the positions of Bm*Lep*1 onto chromosome assemblies indicated that integrations are co-localized with protein-coding genes (Supplementary Fig. S1; chromosome 1 shown in Fig. 4), which agrees with seminal evidence of *Lep*1 being within gene intervals.[36] Furthermore, *Lep*1 proximity to *B. mori* gene-coding regions suggests that *Lep*1s may affect gene structure and function on a genome-wide scale.[9,20] The effects of *Lep*1 *Helitron* integrations upon gene structure and function are for the first time presented in Section 3.4.

3.4. *Lep*1 elements modify gene structure

The movement and propagation of TEs introduce haplotype variation,[51] and alter gene functions when integrated within coding regions of a genome.[28] TEs are present within insect ESTs[52] and mature transcripts of *D. melanogaster*,[53] and can cause alternative the modification of *cis*-regulatory function,[54] introduction of frameshift and premature stop codon mutations,[28] and be involved in exon shuffling[18] or insertion of introns.[55] Evidence that *Lep*1s co-localize with protein-coding genes in *B. mori* genome (section 3.3) suggests that integrations may affect the structure and function of lepidopteran genes, and that presence/absence mutation at orthologous loci may be a source of function genetic

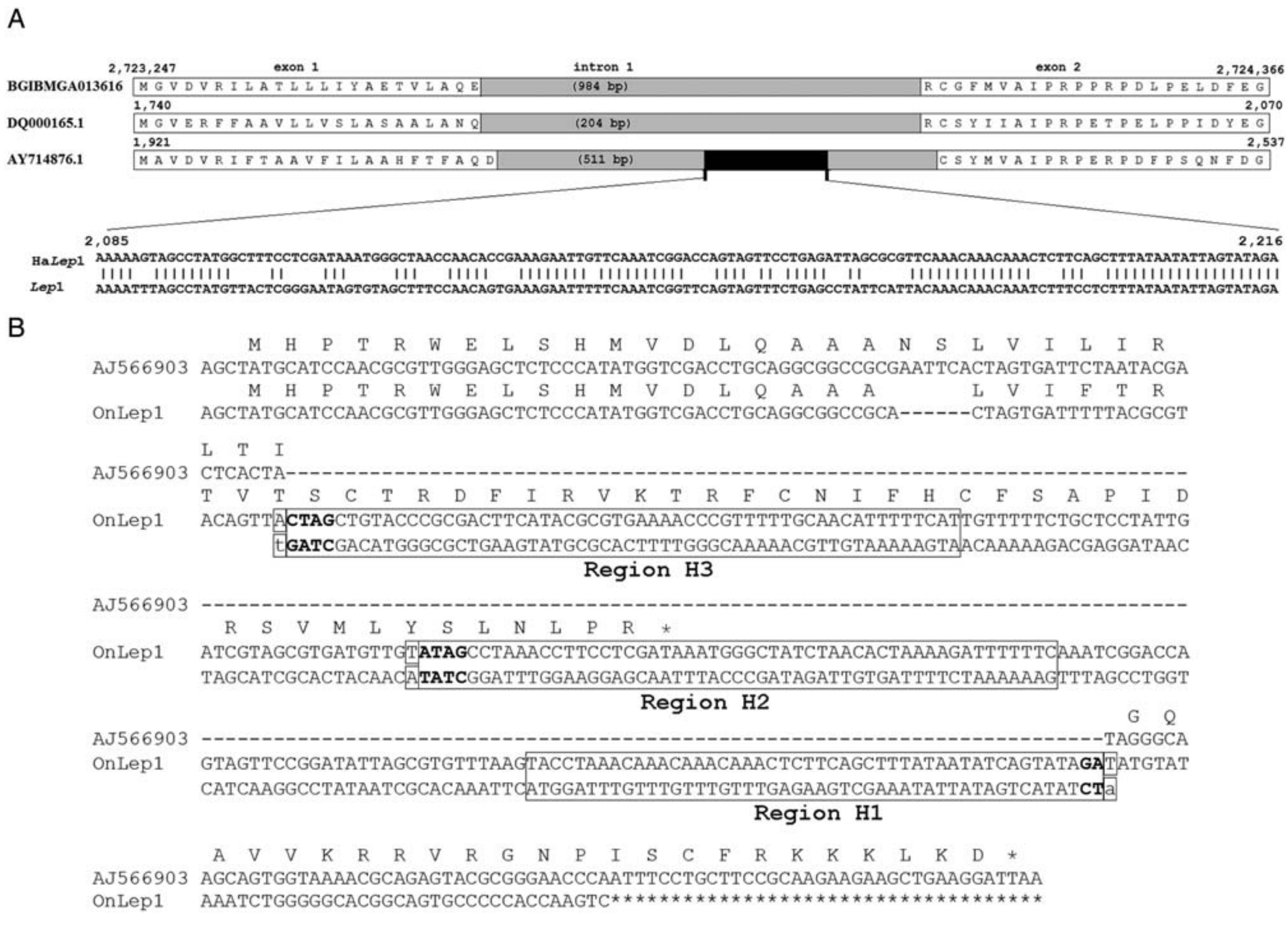


Figure 5. *Lep*1 *Helitron* copy number variation. (A) Integration variation among *Bombyx mori* BGIBMGA013616, *O. nubilalis* DQ000165.1, and *Helicoverpa armigera* AY714876.1 DNA sequence accessions for intron 1 at the cadherin locus. The annotated *H. armigera Lep*1 (Ha*Lep*1) is integrated into the minus strand (positions 2216–2085) and shows 68.7% nucleotide sequence similarity to the *Lep*1 consensus and retains 5′-CT and 3′-CTAY termini. (B) Integration of a *Lep*1 element within the *O. nubilalis* allatotropin neuropeptide precursor gene that is compared with the *Spodoptera frugiperda* homolog (AJ566903.1) (integration is underscored by <). Gaps in the alignment and missing data are represented by - and *, respectively.

variation among species. Comparison among orthologs of the cadherin gene from *B. mori*, *H. armigera*, and *O. nubilalis* genomes showed that *Helitron* copy number variation is present. Specifically, a nucleotide sequence with 68.7% similarity to the *Lep*1 consensus was predicted within the *H. armigera* cadherin intron 1, whereas orthologous introns from *B. mori* or *O. nubilalis* show no *Lep*1-like sequence (Fig. 5A). Although the effects of this integration upon gene function was not investigated, TE integrations within introns are known to affect splicing efficiencies[56] and indicated that *Lep*1s are a source of genome copy number variation between lepidopteran species.

Analogously, *Lep*1 *Helitron* integrations were described within cDNA-RACE products from 46 *O. nubilalis* clones (29.6 kb total; mean insert size: 616.1 ± 244.5 bp; GenBank accessions: JG732059–JG732089; JG744027–JG744041). Sequence from RACE products were assembled into 8 contigs and 14 singletons (3.56 ± 1.94 reads per contig), and 21 of these contigs were subsequently annotated as having a *Lep*1 integration (Supplementary Fig. S3). Functional annotation of these contigs indicated that all transcript-derived *O. nubilalis Lep1 Helitrons* were within intron or untranslated regions (data not shown), with the exception of contig04. Contig04 was predicted to show 85% amino acid similarity to the *S. frugiperda* allatotropin neuropeptide (*at2*a; GenBank accession CAD98809.1) and that a *Lep*1 *Helitron* integration had occurred within the protein-coding regions in the *O. nubilalis* ortholog (Fig. 5B). When compared to the 53 aa *S. frugiperda at2a* gene sequence, the C-terminal 37 aa of the 64 residue *O. nubilalis* ortholog was predicted to be encoded by regions H2 and H3 of an integrated *Lep*1 *Helitron* (Fig. 5B). The integration inserted a novel protein-coding sequence that contains a TAA stop codon and changed the predicted molecular weight and isoelectric point of the *O. nubilalis at*2 protein (pI ~10.87; 13.6 kDa) compared with that of *S. frugiperda* (pI = 11.4; 6.1 kDa). The affect of these changes on protein function was not investigated further, but indicated that the *Lep*1 *Helitron* can affect the structure and function of gene coding sequences in Lepidoptera.

In conclusion, a comparative genomics approach was used to identify novel sequences acquired by the highly conserved ancestral *Lep*1 *Helitron*. Although the primary sequence among gained sequences are variable, a conservation of secondary structures showed that sequence identity by state is an important factor in determining the success of acquired genomic regions for in subsequent transposition events within the genome. *Lep*1 provides insight into the structural requirements for RCR in animal *Helitrons*. Furthermore, the prevalence and preference of *Lep*1 integrations in proximity to gene-coding regions shows that this class of *Helitrons* impacts the structure and function of genomes in which they reside.

Acknowledgements: This research was a joint contribution of the United States Department of Agriculture (USDA), Agricultural Research Service and the Iowa State University outreach station, Ames, IA (Project 3543). This article reports the results of research only. Mention of a proprietary product or service does not constitute an endorsement or recommendation by USDA or Iowa State University for its use.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the USDA Current Research Information System (CRIS) project number 3625-22000-017-00D.

References

1. Kapitonov, V.V. and Jurka, J. 2001, Rolling-circle transposons in eukaryotes, *Proc. Natl. Acad. Sci. USA*, **98**, 8714–9.
2. Feschotte, C. and Mouches, C. 2000, Recent amplification of miniature inverted-repeat transposable elements in the vector mosquito *Culex pipiens*: characterization of the *Mimo* family, *Gene*, **250**, 109–16.
3. Kapitonov, V.V. and Jurka, J. 2007, *Helitrons* on a roll: eukaryotic rolling-circle transposons, *Trends Genet.*, **23**, 521–9.
4. Lai, J., Li, Y., Messing, J. and Dooner, H.K. 2005, Gene movement by *Helitron* transposons contributes to haplotype variability of maize, *Proc. Natl. Acad. Sci. USA*, **102**, 9068–73.
5. Galagan, J.E., Calvo, S.E., Cuomo, C., et al. 2005, Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzæ*, *Nature*, **438**, 1105–15.
6. Donner, H.K. and Weil, C.F. 2009, Give-and-take: interactions between DNA transposons and their host plant genomes, *Curr. Opin. Genet. Dev.*, **17**, 486–92.
7. Yang, H.P. and Barbash, D.A. 2008, Abundant and species specific *DINE*-1 transposable elements in 12 *Drosophila* genomes, *Genome Biol.*, **9**, R39.
8. Yang, L. and Bennetzen, J.L. 2009, Structure-based discovery and description of plant and animal *Helitrons*, *Proc. Natl. Acad. Sci. USA*, **106**, 12832–7.
9. Bureau, T.E. and Wessler, S.R. 1992, *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes, *Plant Cell*, **4**, 1283–94.

10. Li, Y. and Donner, H.K. 2009, Excision of Helitron transposons in maize, *Genetics*, **182**, 399–402.
11. Ohmori, Y., Abiko, M., Horibata, A. and Hirano, H.Y. 2008, A transposon, *Ping*, is integrated into intron 4 of the DROOPING LEAF gene of rice, weakly reducing its expression and causing a mild drooping leaf phenotype, *Plant Cell Physiol.*, **49**, 1176–84.
12. Varagona, M.J., Purugganan, M. and Wessler, S.R. 1992, Alternative splicing induced by insertion of retrotransposons into the maize *waxy* gene, *Plant Cell*, **4**, 811–20.
13. Benjak, A., Boue, S., Forneck, A. and Casacuberta, J.M. 2009, Recent amplification and impact of MITEs on the genome of grapevine (*Vitis vinifera* L.), *Genome Biol. Evol.*, **1**, 75–84.
14. Feschotte, C. and Wessler, S.R. 2001, Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes, *Proc. Natl. Acad. Sci. USA*, **98**, 8923–4.
15. Fu, H. and Dooner, H.K. 2002, Intraspecific violation of genetic colinearity and its implications in maize, *Proc. Natl. Acad. Sci. USA*, **99**, 9573–8.
16. Morgante, M., Brunner, S., Pea, G., Fengler, K., et al. 2005, Gene duplication and exon shuffling by *Helitron*-like transposons generate intraspecies diversity in maize, *Nat. Genet.*, **37**, 997–1002.
17. Lal, S.K. and Hannah, L.C. 2005, *Helitrons* contribute to the lack of gene colinearity observed in modern maize inbreds, *Proc. Natl. Acad. Sci. USA*, **102**, 9993–4.
18. Britten, R.J. 2004, Coding sequences of functioning human genes derived entirely from mobile element sequences, *Proc. Natl. Acad. Sci. USA*, **101**, 16825–30.
19. Bennetzen, J.L. 2005, Transposable elements, gene creation and genome rearrangement in flowering plants, *Curr. Opin. Gene Devel.*, **15**, 621–7.
20. Yang, L. and Bennetzen, J.L. 2009, Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome, *Proc. Natl. Acad. Sci. USA*, **106**, 19922–7.
21. Gupta, S., Gallavotti, A., Stryker, G.A., et al. 2005, A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes, *Plant Mol. Biol.*, **57**, 115–27.
22. Zhang, J., Yu, C., Pulletikurti, V., et al. 2009, Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize, *Genes Dev.*, **23**, 755–65.
23. McClintock, B. 1950, The origin and behavior of mutable loci in maize, *Proc. Natl. Acad. Sci. USA*, **36**, 344–55.
24. Wendel, J.F. and Wessler, S.R. 2000, Retrotransposon-mediated genome evolution on a local ecological scale, *Proc. Natl. Acad. Sci. USA*, **97**, 6250–2.
25. Binghamp, M., Kidwell, M.G. and Rubin, G.M. 1982, The molecular basis of PM hybrid dysgenesis: the role of the *P* element, a *P*-strain-specific transposon family, *Cell*, **29**, 995–1004.
26. Noor, M.A.F. and Chang, A.S. 2006, Evolutionary genetics: jumping into a new species, Curr. Biol., **16**, R890–2.
27. Gonzalez, J., Karasov, T., Messer, P.W. and Petrov, D.A. 2010, Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*, *PLoS Genet.*, **6**, e1000905.
28. Gahan, L.J., Gould, F. and Heckel, D.G. 2001, Identification of a gene associated with Bt resistance in *Heliothis virescens*, *Science*, **293**, 857–60.
29. d'Alençon, E., Sezutsu, H., Legeai, F., et al. 2010, Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements, *Proc. Natl. Acad. Sci. USA*, **107**, 7600–5.
30. International Silkworm Genome Consortium. 2008, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, *Insect Biochem. Mol. Biol.*, **38**, 1036–45.
31. Mita, K., Kasahara, M., Sasaki, S., et al. 2004, The genome sequence of silkworm, *Bombyx mori*, *DNA Res.*, **11**, 27–36.
32. Xia, Q., Zhou, Z., Lu, C., et al. 2004, A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*), *Science*, **306**, 1937–40.
33. Osanai-Futahashi, M., Suetsugu, Y., Mita, K. and Fujiwara, H. 2009, Genome-wide screening and characterization of transposable elements and their distribution in the silkworm, *Bombyx mori*, *Insect Biochem. Mol. Biol.*, **38**, 1046–57.
34. Okada, N. 1991, SINEs: short interspersed repeated elements of the eukaryotic genome, *Trends Ecol. Evol.*, **6**, 358–61.
35. Coates, B.S., Sumerford, D.V., Hellmich, R.L. and Lewis, L.C. 2010, A *Helitron*-like transposon superfamily from Lepidoptera disrupts $(GAAA)_n$ microsatellites and is responsible for flanking sequence similarity within a microsatellite family, *J. Mol. Evol.*, **70**, 278–88.
36. Yang, C., Teng, X., Zurovec, M., et al. 1998, Characterization of the *P25* silk gene and associated insertion elements in *Galleria melonella*, *Gene*, **209**, 157–65.
37. Van't Hof, A.E., Brakefield, P.M., Saccheri, I.J. and Zwaan, B.J. 2007, Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera, *Heredity*, **98**, 320–8.
38. Hall, T.A. 1999, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, *Nucl. Acids Symp. Ser.*, **41**, 95–8.
39. Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007, MEGA4: molecular evolutionary genetic analysis MEGA. software version 4.0, *Mol. Biol. Evol.*, **24**, 1596–9.
40. Zuker, M. 2003, Mfold web server for nucleic acid folding and hybridization prediction, *Nucl. Acids Res.*, **31**, 3406–15.
41. Youens-Clark, K., Faga, B., Yap, I.L., Stein, L. and Ware, D. 2009, CMap 1,01: a comparative mapping application for the Internet, *Bioinformatics*, **25**, 3042–2.
42. Huang, X. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome Res.*, **9**, 868–77.
43. Conesa, A., Götz, S., Garca-Gomez, J. M., et al. 2005, Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.

44. Götz, S., Garcia-Gomez, J.M., Terol, J., et al. 2008, High-throughput functional annotation and data mining with the Blast2GO suite, *Nucl. Acids Res.*, **36**, 3421–35.
45. Coates, B.S., Kroemer, J.A., Sumerford, D.V. and Hellmich, R.L. 2011, A novel class of miniature inverted repeat transposable elements MITEs that contain hitch-hiking GTCYn microsatellites, *Insect Mol. Biol.*, **20**, 15–27.
46. Tempel, S., Nicolas, J., Amrani, A.E. and Couee, I. 2007, Model-based identification of *Helitrons* results in a new classification of their families in *Arabidopsis thaliana*, *Gene*, **403**, 18–28.
47. Sweredowski, M., Wilson, L.D.R. and Gaut, B.S. 2008, A comparative computational analysis of nonautonomous *Helitron* elements between maize and rice, *BMC Genomics*, **9**, 467.
48. Leung, S.K. and Wong, J.T.Y. 2009, The replication of plastid microcircles involved rolling circle intermediates, *Nucl. Acids Res.*, **37**, 1991–2002.
49. Yoder, J.A., Walsh, C.P. and Bestor, T.H. 1997, Cytosine methylation and the ecology of intragenomic parasites, *Trends Genet.*, **13**, 335–40.
50. Orgel, L.E. and Crick, F.H.C. 1980, Selfish DNA-the ultimate parasite, *Nature*, **284**, 604–7.
51. Wang, W. and Kirkness, E.F. 2005, Short interspersed elements SINEs are a major source of canine genome diversity, *Genome Res.*, **15**, 1798–808.
52. Sunter, J.D., Patel, S.P., Skilton, R.A., et al. 2008, A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks from the arthropod sub-phylum Chelicerata, *Gene*, **415**, 13–22.
53. Lipatov, M., Lenkov, K., Petrov, D.A. and Bergman, C.M. 2005, Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome, *BMC Biol.*, **3**, 24.
54. van de Legemaat, L.N., Landry, J.R., Mager, D.L. and Medstrand, P. 2003, Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions, *Trends Genet.*, **19**, 530–6.
55. Giroux, M.J., Clancy, M., Baier, J., et al. 1994, *De novo* synthesis of an intron by the maize transposable element dissociation, *Proc. Natl. Acad. Sci. USA*, **91**, 12150–4.
56. Davis, M.B., Dietz, J., Standiford, D.M. and Emerson, C.P. Jr. 1988, Transposable element insertions respecify alternative exon splicing in three *Drosophila* myosin heavy chain mutants, *Genetics*, **150**, 1105–14.