J Mol Evol (2010) 70:275–288
DOI 10.1007/s00239-010-9330-6

A *Helitron*-Like Transposon Superfamily from Lepidoptera Disrupts $(GAAA)_n$ Microsatellites and is Responsible for Flanking Sequence Similarity within a Microsatellite Family

Brad S. Coates · Douglas V. Sumerford · Richard L. Hellmich · Leslie C. Lewis

Received: 9 December 2008 / Accepted: 17 February 2010 / Published online: 9 March 2010

Abstract Transposable elements (TEs) are mobile DNA regions that alter host genome structure and gene expression. A novel 588 bp non-autonomous high copy number TE in the *Ostrinia nubilalis* genome has features in common with miniature inverted-repeat transposable elements (MITEs): high A + T content (62.3%), lack of internal protein coding sequence, and secondary structure consisting of subterminal inverted repeats (SIRs). The *O. nubilalis* TE has inserted at $(GAAA)_n$ microsatellite loci, and was named the microsatellite-associated interspersed nuclear element (*MINE*-1). Non-autonomous *MINE*-1 superfamily members also were identified downstream of $(GAAA)_n$ microsatellites within *Bombyx mori* and *Pectinophora gossypiella* genomes. Of 316 $(GAAA)_n$ microsatellites from the *B. mori* whole genome sequence, 201 (63.6%) have associated autonomous or non-autonomous *MINE*-1 elements. Autonomous *B. mori MINE*-1s a encode a helicase and endonuclease domain RepHel-like protein (BMHELp1) indicating their classification as *Helitron*-like transposons and were renamed *Helitron*1_BM. Transposition of *MINE*-1 members in Lepidoptera has resulted in the disruption of $(GAAA)_n$ microsatellite loci, has impacted the application of microsatellite-based genetic markers, and suggests genome sequence that flanks TT/AA dinucleotides may be required for target site recognition by RepHel endonuclease domains.

Keywords Microsatellite family · Transposable element · Mobile element

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9330-6) contains supplementary material, which is available to authorized users.

B. S. Coates (✉) · D. V. Sumerford · R. L. Hellmich · L. C. Lewis
USDA-ARS, Corn Insect and Crop Genetics Research Unit, 113 Genetics Lab, Iowa State University, Ames, IA 50011, USA
e-mail: brad.coates@ars.usda.gov

D. V. Sumerford · R. L. Hellmich · L. C. Lewis
Department of Entomology, Iowa State University, Ames, IA 50011, USA

Introduction

Transposable elements (TEs) are mobile genetic elements that influence chromosome structure and gene expression (Roeder et al. 1985; Witte et al. 2001; Lerat and Sémon 2007). Class II TEs are non-conservative DNA-based TEs that transpose by a "cut-and-paste" mechanism, whereby the entire insertion sequence (IS) is excised and reinserted into target genome locations (Craig 1995). Remnants of past insertion consists of target site duplications (TSDs), which are short direct repeats created via staggered cleave by transposases or integrases followed by site fill-in by DNA polymerase. Miniature inverted-repeat transposable elements (MITEs) are non-autonomous class II mobile DNA elements that are usually <500 bp, have a characteristically high A + T nucleotide content, and secondary structures that include terminal inverted repeats (TIRs; Wessler et al. 1995) or subterminal inverted repeats (SIRs; Tu 2000). MITEs lack an internal protein coding region, and mobility is mediated by *trans*-acting factors encoded by related autonomous TEs (Dufresne et al. 2006). MITEs are found ubiquitously within eukaryotic genomes, including mosquito (Tu 2000; Quesneville et al. 2006), *Drosophila* (Locke et al. 1999; Vivas et al. 1999; Miller et al. 2000; Wilder and Hollocher 2001; Yang and Barbash 2008), and Lepidoptera (Chen and Li 2007; Coates et al. 2009) and have a role in altering genome structure

including insertion at microsatellite loci (Akagi et al. 2001).

Class II TEs also include *Helitron* elements that propagate by rolling circle replication (RCR), and have structural features that include TIRs or SIRs, 3′ stem-loops, and conserved 5′ TC and 3′ CTRR termini. *Helitrons* do not create TSDs, but in contrast to other class II TEs insert between AT (Kapitonov and Jurka 2001) or TT dinucleotides (Kapitonov and Jurka 2007a). Autonomous *Helitrons* encode a RepHel protein that carries out replication initiation and helicase activities, and optional replication protein A (RPA) genes that function in ssDNA binding (Kapitonov and Jurka 2001). Similar to MITEs, non-autonomous *Helitron* elements exist and were predicted to represent all known *Helitron*-like *Drosophila* interspersed elements (*DINE*-1s) in the *D. melanogaster* genome (Kapitonov and Jurka 2007a). *Helitrons* have been described from *Aradidopsis thaiana*, *Oriza sativa*, *Caenorhabditis elegans*, and *Zea maize* (Gupta et al. 2005; Lai et al. 2005), and might be related to Geminiviruses that have integrated into plant genomes (Kapitonov and Jurka 2007b).

Microsatellite loci are composed of short nucleotide motifs repeated in tandem, and except for recent whole genome duplications, each microsatellite has arisen independently by slip strand mispairing (SSM) during DNA replication (Levinson and Gutman 1987). Microsatellites that originate at independent loci have unique flanking DNA that is used to design locus-specific genetic markers (Pemberton et al. 1995) that are desirable due to high levels of allelic variation (Tautz 1989; Weber and May 1989). Polymorphic microsatellite markers from Lepidoptera have been used for linkage mapping (Miao et al. 2005) and population genetic studies (Reddy et al. 1999; Prasad et al. 2005; Malausa et al. 2007), but problems often arise during genotype analyses. PCR coamplification of fragments primed at >1 microsatellite locus often occur due to the presence of multilocus microsatellites (microsatellite families) that share nucleotide sequence similarity at genome regions that flank the tandem repeat (Zhang 2004; Meglécz et al. 2004; Van't Hof et al. 2007). Difficulties are encountered because alleles that originate from independently segregating loci typically are scored, and render genetic markers unsuitable for population or linkage analysis (Anthony et al. 2001; Fauvelot et al. 2006; Anderson et al. 2007). Microsatellite families are known among insect species (Meglécz et al. 2007), but appear pronounced within lepidopteran species due to a yet undescribed common ancestry that is shared among genomes.

Structural effects of MITE- and *Helitron*-like elements include insertion within genes and introns (Chen and Li 2007; Yang and Barbash 2008; Kapitonov and Jurka 2007b) leading to an increase in overall gene size (Xia et al. 2004). Genome associations between TEs and microsatellite loci are also observed when integrations occur preferentially into tandem repeats (Akagi et al. 2001; Temnykh et al. 2001) or when tandem repeats hitchhike within mobile elements (Wilder and Hollocher 2001; López-Giráldez et al. 2006; Coates et al. 2009). Modification of microsatellite loci by TE insertion has been documented. The *Alu* repeats are retroelement (class I TE)-like short interspersed nucleotide elements (SINEs) that are prevalent in non-coding regions of primate genomes and are composed of a 282 bp conserved region that is similar to the 7SL rRNA gene (Jelinek et al. 1980). Subfamilies of human *Alu* elements are associated with microsatellite repeats (Zuliani and Hobbs 1990; Jurka and Pethiyagoda 1995), wherein *Alu*J subfamily members are preferentially located near $(GAAA)_n$ repeats (Yandava et al. 1997). Also, the wheat genome is composed of microsatellite loci with flanking nucleotide sequences that contain TEs (Ramsey et al. 1999), whereas rice $(TA)_n$ microsatellites are a known target for the insertion of the *Micron* family of MITEs (Akagi et al. 2001; Temnykh et al. 2001). TE associations with microsatellite loci often result in molecular markers that are non-Mendelian, show PCR co-amplification of >1 locus, or weak PCR amplification due to oligonucleotide primer competition (Economou et al. 1990; Zhang 2004; Meglécz et al. 2004; Van't Hof et al. 2007), and are problematic during the development of genetic and genomic marker loci.

In the following research we describe a superfamily composed of autonomous and non-autonomous MITE- or *Helitron*-like elements within the genomes of lepidopteran species that show structural features similar to the *DINE*-1 family of TEs from *D. melanogaster*. Due to observation of $(GAAA)_n$ repeat unit microsatellite loci flanking these TEs, the elements are referred to as the microsatellite-associate interspersed nuclear element (*MINE*-1). Autonomous *MINE*-1 members in the *Bombyx mori* genome assembly encoded helicase and endonuclease domain proteins that are similar to RepHels of autonomous *Helitrons* and may be involved in RCR. We show that the transposition of autonomous *B. mori Helitrons* (*Helitron*1_BM) and related non-autonomous *MINE*-1 elements within genomes of Lepidoptera impact chromosome structure through integration at (GAAA)n microsatellite loci, which further indicates that sequence features at the genome target site may facilitate *Helitron* integration.

Methods

Isolation and Annotation of an *Ostrinia nubilalis* Insertion Sequence

An *O. nubilalis* BAC library, OnB1, clones 04M1–04M24, 04B1–04B24, 55M1–55M24, and 50K1–50K24 were

screened for OnMITE01 presence by PCR using primers OnMITE01-F 5′-TCC YAA CTA ATA TTA TAR ATG CGA AAG-3′ and OnMITE01-R 5′-CCC GCG TGG AAT TTT GTC TG-3′ (Coates et al. 2009). PCR amplification of each BAC clone took place in 10 µl reaction volumes containing 1.5 mM $MgCl_2$, 50 µM dNTPs, 5 ng BAC DNA, 1.8 pmol of each primer, 2 µl 5× thermal polymerase buffer (Promega), and 0.3125U Go*Taq* DNA polymerase (Promega, Madison, WI, USA). A BioRad Tetrad 2 thermocycler program of 96°C for 2 min, then 32 cycles of 96°C for 20 s, 55°C for 30 s, and 72°C for 30 s (Program TD2) was used. Entire PCR product volumes were separated on 1.5% agarose gels and positive BAC identified by presence/absence of the gene fragment compared to positive control DNA (220–260 bp expected size range due to a polymorphic internal CTGT microsatellite repeat).

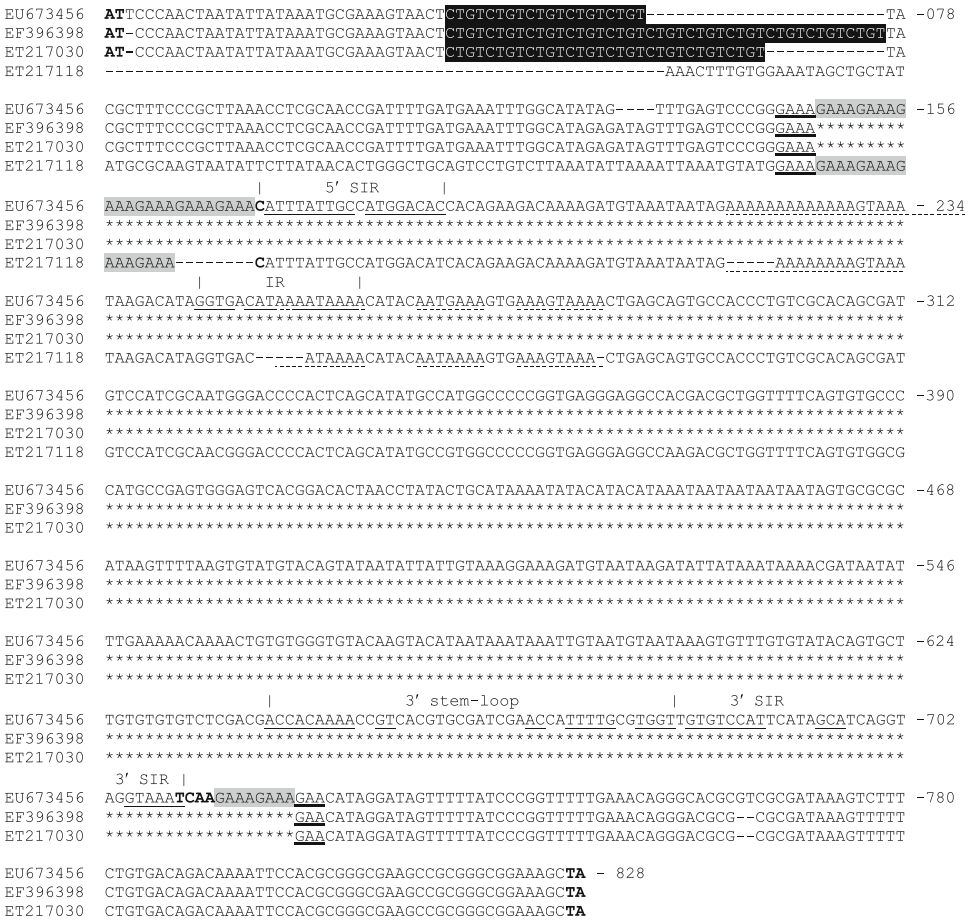
A PCR product of approximately 850 bp was observed from the OnB1 clones 04M15 compared to 198 bp that was predicted. OnMITE01-like regions were reamplified by PCR in 50 µl reactions scaled up from that described previously. Products were purified using Qiagen PCR quickspin columns (Qiagen, Valencia, CA), and ligated into the pGEM-T easy cloning vector (Promega) that was used to transform *E. coli* strain XL1 Blue (Stratagene) by electroporation. Transformants were plated on LB agar containing 20 µg/ml chloramphenical and 50 µg/ml tetracycline. Positive insert clones were incubated overnight in 0.8 ml TB containing 20 µg/ml chloramphenical, plasmid DNA purified using Zippy Plasmid Isolation Kits (Zymo Research), and inserts were sequenced using DTCS Kits (Beckman-Coulter) with T7 and SP6 primers. DTCS products were separated on a CEQ8000 DNA Sequence Analysis System (Beckman-Coulter; inject 4.2 kV for 10 s, separate 4.2 kV for 140 m). Raw sequence data were analyzed and quality of sequence assessed using the PHRED quality parameter, and was calculated in the Sequence Analysis software of the Beckman-Coulter CEQ8000 Genetic Analysis System. Sequences were trimmed when $q < 30$ (99% base call accuracy). Vector sequence was automatically trimmed from FASTA formatted data when exported by the CEQ8000 Sequence Analysis software (v. 8.0). The plasmid insert was assembled using from T7 and SP6 reads with Contig Express software (Informax, San Francisco, CA).

This sequence constructed by Contig Express was aligned with 16 OnMITE01 sequences (listed in footnote of Fig. 1), and the GAAAGAA insertion site identified compared to the sequences that lacked the transposon (from hereon called the *O. nubilalis* GAAA microsatellite-associated interspersed nuclear element; On*MINE*-1). BLASTn searches of GenBank nr, EST, and GSS databases were performed using the putative *O. nubilalis* IS (On*MINE*-1 from positions 156–721 in Fig. 1) is referred to as database search #1, and the resulting "hits" to sequence accessions with *E*-values $\geq 3 \times 10^{-50}$ and ≥80% similarity to the query sequence were downloaded in FASTA format. The On*MINE*-1 secondary structure was predicted for the contig and all downloaded sequences using MFOLD (Zuker 2003 http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi).

Frequency of TE integration into the *O. nubilalis* genome was conducted by screening of 384 OnB1 clones 72A01–72P24 by PCR, and by real-time PCR analysis. Oligonucleotide primers On*MINE*-1-F (5′-CAT TTA TTG CCA TGG ACA YCA C-3′) and On*MINE*-1-R (5′-GTG CGA CAG GGT GGC ACT-3′) were designed from alignment of On*MINE*-1 and OnB1 BAC end sequence (BESs) from GenBank GSS accessions ET217118 and ET217119 (Coates et al. 2009). The 139 bp predicted PCR fragment was PCR amplified under conditions outlined for primers OnMITE01-F and -R, except the 72°C extension time was decreased to 10 s and products were separated by 2% agarose gel electrophoresis. Additionally, two adult male and 2 adult female *O. nubilalis* were obtained from the USDA-ARS, CICGRU colony, dissected, and genomic DNA isolated individual thoracic tissue using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). Real-time PCR reactions (3 replicates per sample) included 12.5 µl iQ SYBR Green reaction mix (BioRad), 5 pmol each of primer On*MINE*-1-F and On*MINE*-1-R, and 1 µl genomic DNA diluted 1:100 in a 25 µl reaction. Reactions were cycled on an iQ thermocycler (BioRad) at 95°C for 3 m, then 40 cycles of 95°C 30 s, 60°C for 30 s, and 72°C for 20 s, and fluorescence data collected at 490 nm. All reactions were followed by melt curve analysis (55°C for 10 s, +0.5°C/cycle for 80 cycles) to ensure presence of a homogeneous population of PCR product and lack of primer dimerization. Cycle threshold (C_T) levels generated from background subtracted data using iCycler software v. 3.0.6070. On*MINE*-1 assays were repeated using 1:1,000, 1:10,000, and 1:100,000 dilutions of genomic DNA to test consistency across DNA concentrations, and resulting change in cycle threshold (ΔC_T) values plotted against Log_{10}(DNA dilution) according to (Livak and Schmittgen 2001). Mean of 3 unknown replicates was normalized to the mean C_T value of 3 replicates of the β-actin gene control amplification, and with the relative product amounts calculated using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001).

Insertion Site Specificity of an *O. nubilalis* TE

Genome sequence that flanked a subset of On*MINE*-1 integrations was obtained by inverse PCR. Six *O. nubilalis* adults (3 male and 3 female) were collected from a wild population near Ames, IA. Individual thoraces were dissected, pooled, and genomic DNA isolated using Qiagen



Note: Homologous OnMITEs not listed in Table 1 are in GenBank accessions EF396379, EF396401, EF396411, and EF396469.

Fig. 1 DNA sequence alignments of an *Ostrinia nubilalis* miniature inverted repeat transposable element (OnMITE01)-like sequence from bacterial artificial chromosome (BAC) clone OnB1-04M15 with OnMITE01 sequences in GenBank accessions EF396298 and ET217030. The AT/TA target site duplication (TSD; bold) and conserved $(CTGT)_n$ microsatellite motif of OnMITE01 are highlighted black with white text. The variant OnMITE01 sequence from OnB1-04M15 has a unique microsatellite-associated interspersed nuclear element (On*MINE*-1) inserted at OnMITE01 GAAAGAA motifs (bold underlined). The variable number of $(GAAA)_n$ microsatellite repeat units flanking the On*MINE*-1 insertion are highlighted, and subterminal inverted repeats (SIRs) of [ATTTATTGCCATGGACAC/GTGTCCATTCATA GCA(8 nt)GTAAAT] are underlined. Four imperfect direct repeats (DRs) within the central core regions are underlined with dashes, and inverted repeat (IR) complementary to the 5′ SIR, and 3′ stem-loop are indicated. "–" indicates a gap in the alignment, and "*" indicates sequence not present

DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). DNA was quantified a A^{260} on a UV spectrophotometer, and 0.5 µg digested in with *Xba*I or *Vsp*I (Promega) in a 10 µl reaction volume according to manufacturer instructions. Digested fragments were purified using Qiagen PCR Quickspin columns and eluted in 30 µl of kit elution buffer. DNA fragments were circularized by addition of 2.5 µl of 10 × T4 ligase buffer and 3 units of T4 ligase to 21.5 µl of column eluate. Oligonucleotide primers for inverse PCR On*MINE*-1_Inv1 (5′-GAC AGG GTG GCA CTG CTC-3′) and On*MINE*-1_Inv2 (5′-ACG ACG CTG GTT TTC AGT GT-3′) were designed from the full cloned contig sequence using Primer3 (Rozen and Skaletsky 2000), and synthesized at Integrated DNA technologies (Coralville, IA). Circularized-ligated genomic DNA (2 µl) was PCR amplified in 20 µl reactions that also contained 1.5 mM $MgCl_2$, 50 µM dNTPs, 2.0 pmol of each primer, 4 µl 5× thermal polymerase buffer (Promega), and 0.3125U Go*Taq* and 0.3125U *Tth* DNA polymerase (Promega). Reactions were placed on an Eppendorf Mastercycler, denatured for 2 min at 94°C, cycled 30 times at 94°C for 30 s, 60°C for 30 s, and 72°C for 1.5 min, and then incubated at 72°C for 10 min (program STS60). PCR products were visualized on 1.5% agarose, purified with the Qiagen PCR Quickspin column procedure, ligated into the pGEM-T vector, transformed into *E. coli* strain XL1-Blue, and plasmid insert DNA sequenced from T7 and SP6 primer sites as described previously. The plasmid insert was assembled using from T7 and SP6 sequencing reads with Contig Express software (Informax, San Francisco, CA) and FASTA format sequences exported.

MINE-1 Superfamily Insertion at Lepidopteran $(GAAA)_n$ Microsatellite Loci

The 5′ region of the On*MINE*-1 sequence (positions 156–216) also was used as a query against GenBank nr/nt, dbEST, and dbGSS databases, and "hits" with *E*-values $\leq 1 \times 10^{-2}$ and ≥85% sequence similarity were retained (database search #2). GenBank accessions containing >150 bp of putative *MINE*-1-like regions were imported into MEGA sequence alignment module (Tamura et al. 2007) and ClustalW alignments performed using default parameters (gap opening penalty 15, gap extension penalty 6.66, weight matrix IUB, and transition weight of 0.5). Manual adjustments were made to correct for single base shifts, and secondary structure predictions were made for regions downstream of the $(GAAA)_n$ motif using Mfold (Zuker 2003).

The *B. mori* whole genome scaffold sequence from build 2.0 was downloaded in FASTA format from Kaikobase (http://sgp.dna.affrc.go.jp/pubdata/genomicsequences.html; integretedseq.txt.gz), and was loaded into a local database using BioEdit (Hall, Ibis BioSciences, Carlsbad, CA). Our database search #3 queried the *B. mori* genome assembly with a 30 bp On*MINE*-1-like sequence identified from *B. mori* dbGSS during database search #2 (GAAAGAAA GAAAGAAATCATTTATTCGCCA; see Results and Discussion), and start position information for sequence "hits" with *E*-values >1.0 with hit length ≥20 bp were exported in tab-delimited format. The scaffolds for the *B. mori* whole genome sequence build 2.0 were also used as an input file for an imperfect tetranucleotide repeat unit microsatellite loci search using the program SciRoKo (default parameters; chunk size 50 Mb; Kofler et al. 2007), start position information for $(GAAA)_n$ microsatellites were exported in tab-delimited format, and merged with exported *MINE*-1-like sequence from database search #3. The co-occurrence of $(GAAA)_n$ and putative *MINE*-1 integrations were identified by overlap of positional overlap of predicted sequence elements. *Bombyx mori* genome sequence from +20 to −700 of $(GAAA)_n$ microsatellite sequences identified by SciRoKo were parsed from the scaffold assemblies, and ClustalW alignments and Mfold secondary structure predictions were performed as described previously.

Results and Discussion

Isolation and Annotation of an *Ostrinia nubilalis* TE

Portion of an *O. nubilalis* bacterial artificial chromosome (BAC) library (OnB1; 120 kb average insert size) was screened by PCR with oligonucleotide primers that amplified a 198 bp product specific for the mobile *O. nubilalis* MITE-like element, OnMITE01 (Coates et al. 2009). PCR products had estimated sizes ranging from 190–210 bp, and were observed from 63 of 96 (65.6%) OnB1 clones. In contrast, an OnMITE01-like element within the insert of OnB1 clone 04M15 generated a PCR product that was ~850 bp, or ~650 bp larger than expected. DNA sequence analyses indicate that the PCR product was 813 bp, and was composed of 119 bp 5′ and 106 bp 3′ ends similar to the OnMITE01 sequences EF396398 and ET217030 (Coates et al. 2009; Fig. 1). The 813 bp sequence also contains a novel 588 bp insertion sequence (IS; GenBank accession EU673456). No prior OnMITE01 or other lepidopteran MITE-like elements shows presence of an IS (Chen and Li 2007; Coates et al. 2009), but were observed within the *stowaway* family of MITEs (Petersen and Seberg 2000) and *Z. maize Helitron*-like transposons (Gupta et al. 2005). This 588 bp insertion was identified within 1 of 63 OnMITE01-like elements, which suggests the frequency of disruption is low (≤1.58%). Regardless, the presence/absence within OnMITE01-like regions indicates that the IS may be mobile over evolutionary time, or alternatively may be derived via recombination with another locus.

No other sequence within GenBank nr or EST databases met the our criteria for a positive "hit" within Database search #1 that used the full-length *O. nubilalis* IS as the query, except a BLASTn result from a dbGSS search indicates two *O. nubilalis* BAC end sequences (BES) show ≥93% sequence identity over 230 bp (*E*-value $\geq 1 \times 10^{-92}$; accessions ET217118 and ET217119; Coates et al. 2009). Compared to the putative IS, the *O. nubilalis* BES data show insertion/deletion mutations within four imperfect direct repeats (DRs) and repeat units of a $(GAAA)_4$ microsatellite (Fig. 1). The BES sequences from dbGSS do not contain flanking OnMITE01-like sequence, which suggests that the novel IS was present at >2 genome locations. Copy number of the IS was estimated from the *O. nubilalis* genome by BAC screening and quantitative real-time PCR. An approximate 0.2-fold genome equivalent from the OnB1 library (768 clones) was screened by PCR with IS-specific oligonucleotide primers, and indicates that 724 (94.3%) of BAC clones have ≥1 IS-like region. Results suggest an insertion frequency of ≥1 per 127,000 bp and a whole genome estimate of 3620 copies (=724 ÷ 0.2), but may be an underestimate due to likely presence of >1 copy per BAC. Assuming this IS has a uniform Poisson distribution among BAC inserts, a maximum likelihood estimate provided an estimated IS copy number of 4496 [$\lambda = -\log(44 \div 768) = 1.242$ IS per BAC; genome copy estimate = $((724 \div 0.2) \times 1.242)$]. Evidence for high genome prevalence was corroborated by real-time quantitative PCR data that estimates copy number at 2730 ± 82 within the *O. nubilalis* genome (data not shown). Differences in *O. nubilalis* genome copy number estimates may have resulted from variation in individual fragment intensities that were observed by gel analysis of BAC clones (not shown). This variation likely results from

template-primer mismatches, and thus affects SYBR green fluorescence detection during real-time PCR and thereby leading to an underestimate of genome copy number. Alternatively, *O. nubilalis* may show variation in IS content between haploid genomes. Since the *O. nubilalis* BAC library and real-time PCR template were derived from different sources, the discrepancy in IS copy number estimates may reflect actual differences in genome content analogous to that observed previously among *Helitron* elements within *Z. maize* inbred lines (Brunner et al. 2005).

Although the novel IS shows no homology to any TEs within GenBank, we investigated sequence and structural characteristics that may indicate the presence of a mobile genetic element. Lepidopteran genomes contain numerous small mobile elements (Mita et al. 2004), and some have inserted into introns and have led to the increased size of *B. mori* genes compared to their *Drosophila* orthologs (Xia et al. 2004). The novel 588 bp IS from *O. nubilalis* (Fig. 1) shows a high A + T content (62.3%), which is similar to *O. nubilalis* MITE-like sequences OnMITE01 (55.5%) and OnMITE02 (63.8%; Coates et al. 2009). No internal protein coding sequences >54 amino acids were predicted by BLASTx or tBLASTx searches (*E*-values $\geq$ 0.33; similarities $\leq$31%). The program Mfold (Zuker 2003) predicted a secondary structure elements formed by the IS; 5′ [ATTTATTGCCATGGACAC] and 3′ SIRs [GTGTCCATTCATAGCATCAGGTGTAAAT] (complementary bases underlined). The structure also contains a putative inverted repeat (IR) located 29 bp downstream and complementary to the 5′ SIR (Fig. 2). The change in Gibbs free energy ($\Delta G = \Delta H - T\Delta S$) associated with the secondary structure formation is predicted to be -93.08 kcal mol^{-1} (Supplementary Fig. 1). A highly negative ΔG value indicates that the stem-loop structure may spontaneously form in the given system (20°C), and is energetically favorable compared to random coiled DNA. MITEs have an estimated ΔG of -66.4 (Dufresne et al. 2006) to -87.7 kcal mol^{-1} (Bureau and Wessler 1994a, b; Casacuberta et al. 1998), and suggests that the putative IS secondary structure is as stable as prior predicted MITEs. Sequence and structural analyses indicates that we may have encountered a MITE-like mobile element, and lack of homologous sequence within GenBank suggests that it is a new class of lepidopteran TE. Presence of a $(GAAA)_n$ microsatellite that flanks representative elements in the *O. nubilalis* genome further suggests that the putative mobile element may produce GAAA TSDs or show insertion site specificity at GAAA loci.

Insertion Site Specificity of an *Ostrinia nubilalis* TE

Putative target sites for the integration of invertebrate MITEs typically are short A + T-biased genome sequences that are duplicated during class II mobile element

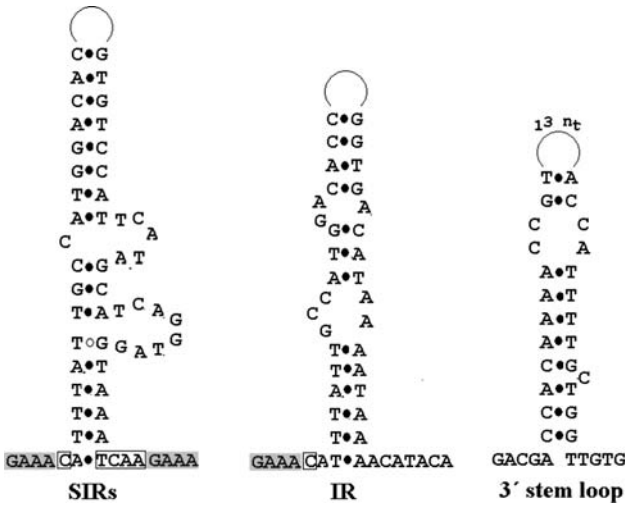


Fig. 2 Putative *Ostrinia nubilalis* microsatellite-associated interspersed nuclear element (On*MINE*-1) secondary structure elements

insertion (target site duplications; TSDs). TAYA TSDs flank the *Ades Aegypti Wukong*, *Jujin*, and *Weneng* families of elements, whereas *Nemo*1 elements appear to target and duplicate CA sequences (Tu 1997), the *Culex pipiens mimo* element has TA TSDs (Feschotte and Mouchès 2000), and similarly lepidopteran MITEs from *O. nubilalis* (Coates et al. 2009) and *Helicoverpa zea* (Chen and Li 2007) insert at TA genome positions. In contrast, *Helitron*-like TEs insert between AT or TT dinucleotides without the creation of TSDs (Kapitonov and Jurka 2001; 2007b). We show that 16 independent OnMITE01 integrations that lack the novel 588 bp IS have a conserved ancestral GAAAGAA nucleotide motif, but in contrast $(GAAA)_7$ and $(GAAA)_3$ microsatellite repeat units are flanking the 5′ and 3′ ends of the novel IS, respectively (GenBank accession EU673456; Fig. 1). Due to a microsatellite array length that is greater than that observed within the 16 ancestral OnMITE01 elements, the expansion of both 5′ and 3′ arrays likely occurred after integration or the mobile IS element may itself be composed of terminal (GAAA) repeats. This evidence suggest that the IS may be involved in the creation of GAAA microsatellites, but cannot exclude the integration within extant (GAAA)n microsatellites, so we name the IS the *O. nubilalis* microsatellite-associated interspersed nuclear element (On*MINE*-1).

Associations between MITE-like elements and microsatellites have been shown previously. Miller et al. (2000), Wilder and Hollocher (2001), and Coates et al. (2009) indicated that $(CTGT)_n$ microsatellites are mobile due to presence within MITEs, whereas the *Micron* family of MITEs within the *O. sativa* genome have preferentially inserted within $(TA)_n$ microsatellites (Akagi et al. 2001). Additional investigation of *O. sativa* $(TA)_n$ microsatellites indicated that nearly 45% were flanked by sequence that

had homology to *Micron* superfamily members (Temnykh et al. 2001). Primate *Alu* elements also are located near multiple microsatellite repeat motifs (Arcot et al. 1995), with the most ancient *AluJ* class associated with $(GAAA)_n$ repeats (Yandava et al. 1997). No prior evidence indicates that MITE or *Helitron*-like elements are preferentially integrated within microsatellite repeats of insects. To investigate a potential association between *MINE*-1 elements and $(GAAA)_n$ repeat microsatellites within the *O. nubilalis* genome we obtained DNA sequence from a plasmid library constructed from On*MINE*-1 specific inverse PCR products.

Sequence data from 47 On*MINE*-1 inverse PCR product library inserts show a mean insert size of 824 ± 72 bp (GenBank dbGSS accessions FI495597–FI495643), and were assembled into 3 contigs and 14 singletons. These 17 unique sequences contain regions homologous to the initial On*MINE*-1 element (EU673456; ≥94.3% similarity). With the exception of the 3 contigs, all On*MINE*-1 sequences show unique genome sequence flanking the putative insertion point (alignment not shown). Additionally, each On*MINE*-1 sequence contains a consensus 5′ SIR [ATTTATTGCCATGGACAC] that is invariably preceded by a cytosine and a $(GAAA)_n$ microsatellite at the 5′ terminus. Remaining portions of On*MINE*-1s contain multiple insertion/deletion mutations, which includes sequence length variation within the direct repeat (DR) region. The predicted termini of On*MINE*-1 elements are composed of 5′ C and a 3′ CTAT motifs that are similar to those conserved among *Helitron*-like TEs (5′ TC and a 3′ CTRR; Kapitonov and Jurka 2001), and suggests the elements may belong to a *Helitron*-like group of TEs. Sequence data also indicate that the number of $(GAAA)_n$ microsatellite repeat units that flank independent On*MINE*-1s integrations (loci) vary from 2–16 at the 5′ end (mean 4.28 ± 3.43) and 1–4 at the 3′ end (mean 1.47 ± 0.83). Imperfect repeat arrays also are encountered within both genomic regions that flank On*MINE*-1s. Inverse PCR product sequence data suggests that On*MINE*-1 family members show a 100% association with $(GAAA)_n$ microsatellite loci within the *O. nubilalis* genome, and that longer microsatellite arrays are biased toward presence upstream at the 5′ terminus the On*MINE*-1 elements. Evidence also suggests that On*MINE*-1s may be involved the modification of microsatellite loci when integration occurs at pre-existing $(GAAA)_n$ repeats, may carry (GAAA) repeats at 5′ and 3′ termini, or may be involved in the formation of microsatellites through a yet undescribed mechanism. Regardless, the association between On*MINE*-1 elements and $(GAAA)_n$ microsatellite loci appears to be a source of DNA sequence similarity among these genome regions in Lepidoptera. In the following sections we provide further evidence that *MINE*-1 elements are ancestral to the lepidopteran lineage and are positioned adjacent to $(GAAA)_n$ microsatellite loci.

MINE-1 Superfamily Insertion at Lepidopteran $(GAAA)_n$ Microsatellite Loci

Due to the observed co-occurrence of *MINE*-1 integrations and $(GAAA)_n$ microsatellite loci in the *O. nubilalis* genome, investigation of similar associations were conducted for all GenBank accessions for Lepidoptera and within the *B. mori* genome assembly (build 2.0; The International Silkworm Genome Consortium 2008). Our previous database search #1 that used the full-length On*MINE*-1 element sequence as a query resulted in identification of homology within two *O. nubilalis* BES accessions in GenBank (Fig. 1). It was previously shown that sequence variation among mobile elements often results in difficulties when homology-based identifications are used cross-species, whereas structure-based predications prove to be more robust in the discovery of novel TEs (Bergman and Quesneville 2007). For example, *Drosophila DINE*-1 superfamily members show a high level of intraspecific sequence variation and presence of truncated copies which cause difficulty in copy number estimations from the WGS of 12 species (Yang and Barbash 2008). Despite the high level of divergence, *DINE*-1s retain structural features that may be necessary for MITE- or *Helitron*-like mobility (Yang et al. 2006; Yang and Barbash 2008): conserved 5′ and 3′ SIRs, an inverted repeat (IR) complementary to the 5′ SIR, and a stem-loop structure in the 3′ terminal region (Locke et al. 1999; Vivas et al. 1999; Miller et al. 2000; Wilder and Hollocher 2001; Yang and Barbash 2008). Analogously, *Helitron*-like transposons have conserved 5′ TC and 3′ CTRR termini (R = A or G nucleotides; Kapitonov and Jurka 2001), but show optional requirements for 3′ terminal stem-loop structures (Galagan et al. 2005; Kapitonov and Jurka 2007b). *Helitrons* within the maize genome often show little sequence similarity outside of conserved terminal nucleotide motifs, such that their identification is often difficult and relies on comparison of haplotypes (Lai et al. 2005).

Since *Drosophila DINE*-1 elements have been predicted using genome database searches that use a short but highly conserved query sequences that correspond to the 5′ SIR (Wilder and Hollocher 2001; Yang and Barbash 2008). Since sequence downstream of *DINE*-1s are more highly polymorphic and can show little intraspecific homology that show limited use in homology-based searches (Locke et al. 1999; Vivas et al. 1999; Miller et al. 2000), we correspondingly used the $(GAAA)_n$ and 5′ SIR sequence of On*MINE*-1 (positions 156–215 in Fig. 1) to query GenBank nr, dbEST, and dbGSS in our search #2. These query results show that a sequence 86% similarity to On*MINE*-1 may be present within an intron in the *Pectinophora gossypiella* cadherin gene (GenBank Accessions in Fig. 3). This putative *P. gossypiella MINE*-1 element (Pg*MINE*-1) has a 13 bp 5′ SIR

[TTATTACTTCCAGACAC] is preceded by an imperfect $(GAAA)_2$ TTTT$(GAAA)_3$ microsatellite (Fig. 3). The 556 bp full-length Pg*MINE*-1 additionally shows a high A + T nucleotide content (65.5%), a 3′ SIR [GTGTAAGTACATAA], central core of direct repeats (DRs), and lack of an internal protein coding sequence (Figs. 3, 4). The Mfold-predicted Pg*MINE*-1 secondary structure ($\Delta G = -92.70$ kcal mol^{-1}; supplementary Fig. 2) has features analogous to those of *DINE*-1 and On*MINE*-1 family members (Fig. 4). In contrast to On*MINE*-1 and *DINE*-1 elements, the Pg*MINE*-1 may lack an IR complementary to and downstream of the 5′ SIR (Fig. 4b). The IR secondary structure may be critical for recognition by *trans*-acting factors required for transposition (Wessler et al. 1995), but lack of the IR suggests it may not be required for transposition of the Pg*MINE*-1 or that the mutation has rendered the element immobile. Additionally, the termini of Pg*MINE*-1 consisted of 5′ C and a 3′ TCAA sequences that are similar to those conserved among *Helitron*-like TEs (5′ TC and a 3′ CTRR; Kapitonov and Jurka 2001), which suggests that the Pg*MINE*-1 from accession AY707868 could be classified as a non-autonomous *Helitron*.

The database search #2 also identified 28 dbGSS accessions from *B. mori* that shares 93% similarity with the

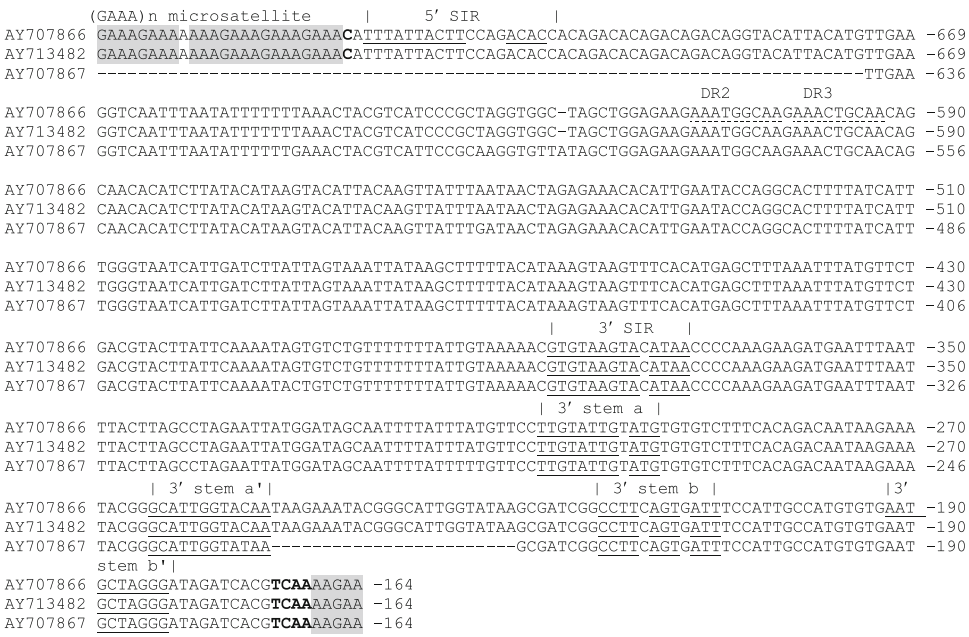


Fig. 3 A microsatellite-associated interspersed nuclear element (*MINE*-1) superfamily member integrated within an intron region of the *Pectinophora gossypiella* (Lepidoptera: Gelechiidae) cadherin gene. The Pg*MINE*-1 subterminal inverted repeats (SIRs) are underlined [TTATTACTTCCAGACAC/GTGTAAGTACATAA], imperfect direct repeats (DRs) are underlined. Positions of respective GenBank accession indicated at left of the alignment

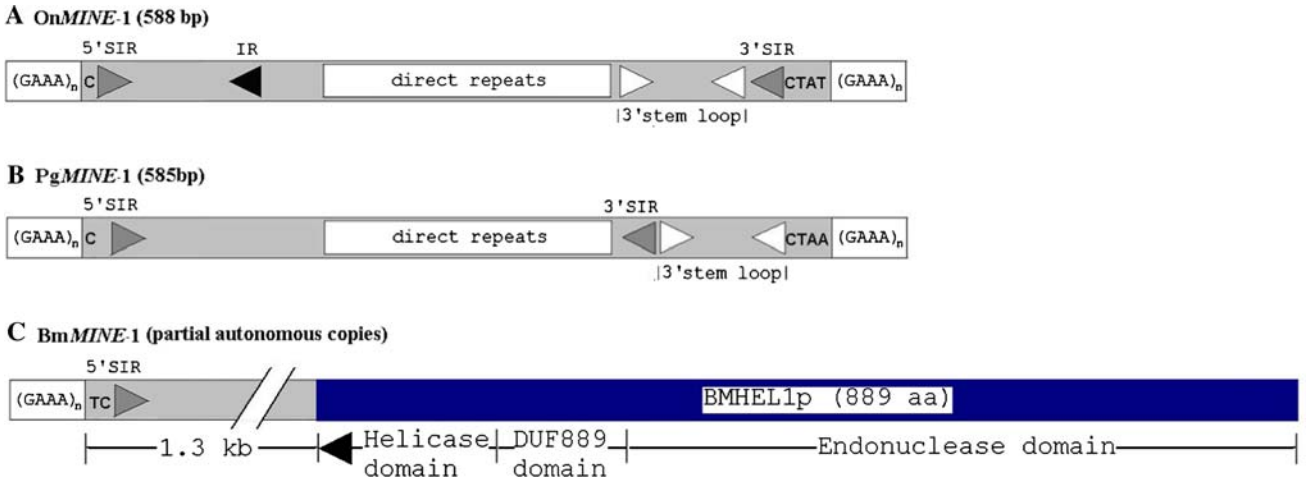


Fig. 4 Structure of microsatellite-associated interspersed nuclear element (*MINE*-1) superfamily members from lepidopteran species *Ostrinia nubilalis* (On*MINE*-1), *Pectinophora gossypiella* (On*MINE*-1), and *Bombyx mori* (Bm*MINE*-1). Feature designations are as indicated in Fig. 1, and functional domains predicted *B. mori Helitron*-like RepHel protein (BMHEL1p) encoded by the autonomous *Helitron*1_BM element are described with the text

On*MINE*-1 $(GAAA)_n$ and 5′ SIR regions, but were only 30 bp in length (*E*-values = 0.08) and consisted of a [$(GAAA)_4$T-CATTTATTCGCCA] sequence. Although this represents a motif that could be encountered by chance within the genome, similarly short regions of homology were used to define *DINE*-1-like elements in *Drosophila* (Wilder and Hollocher 2001; Yang and Barbash 2008) and the highly variable CACTA TE family from the plant species *Triticeae* (Wicker et al. 2003). The subsequent database search #3 that queried the *B. mori* genome sequence build 2 with the [$(GAAA)_4$T-CATTTATTCGCCA] sequence identified 316 unique positions that share ≥89.7% similarity, and are distributed within the genome at an average of 0.65 ± 0.24 per Mb (11.29 ± 3.93 per chromosome; Table 1; supplementary Table T1). In a parallel bioinformatic search, the program SciRoKo predicted 475 perfect or imperfect $(GAAA)_n$ repeat microsatellite loci within the *B. mori* genome (average length of 22.03 bp and average mismatch of 0.3) and are distributed once every 0.94 ± 0.28 Mb (16.96 ± 5.09 per chromosome; supplementary Table T2). The genome copy number prediction of 316 Bm*MINE*-1 integrations is lower than the 334–5,424 for *DINE*-1 elements within 12 *Drosophila* genomes (Yang and Barbash 2008), and lower than our estimate of 2730–3620 *MINE*-1 copies within the *O. nubilalis* genome. This reduced Bm*MINE*-1 content compared to the estimate from *O. nubilalis* may represent natural copy number variation between species, and be analogous to that of *DINE*-1 copy number variation in *Drosophila*. Reasons for disparity between Bm*MINE*-1 and On*MIN*E-1 copy estimates may also be due to the content within build 2. The *B. mori* WGS was sequenced by use of short shotgun sequence reads with a final 3-fold coverage depth from strain p50T (Mita et al. 2004) and 6-fold depth from strain Dazao (Xia et al. 2004), and comprises 23,156 scaffolds (Wang et al. 2005) assembled by repeat masking with the RePS assembler (Wang et al. 2002). Genome assemblies often have an under representation of heterochromatic DNA in which *Helitrons* are located, which led to underestimates of these TEs in *A. thaliana* and *C. elegans* genomes (Kapitonov and Jurka 2007b). Since repeat elements are removed from the final scaffold assemblies (replaced by Ns) a majority of the Bm*MINE*-1 elements likely were omitted, and suggests that repetitive genome elements may be underrepresented within the final genome build of *B. mori*.

Table 1 Distribution of putative *Bombyx mori MINE*-1 elements (Bm*MINE*-1) within the whole genome assembly build 2

Chr	Chr size (Mb)	*MINE*-1 Count	*MINE*-1 Count/Mb	$(GAAA)_n$ Count	$(GAAA)_n$ Count/Mb
chr01	22.40	11	0.49	19	0.85
chr02	10.47	8	0.76	10	0.96
chr03	18.21	9	0.49	16	0.88
chr04	20.92	8	0.38	17	0.81
chr05	20.92	5	0.24	10	0.48
chr06	18.95	7	0.37	9	0.47
chr07	16.28	9	0.55	17	1.04
chr08	18.90	18	0.95	18	0.95
chr09	19.01	13	0.68	21	1.10
chr10	19.76	8	0.40	11	0.56
chr11	24.13	20	0.83	33	1.37
chr12	20.51	12	0.59	18	0.88
chr13	17.99	10	0.56	20	1.11
chr14	15.63	13	0.83	18	1.15
chr15	19.51	6	0.31	13	0.67
chr16	15.24	11	0.72	20	1.31
chr17	18.38	12	0.65	16	0.87
chr18	16.69	16	0.96	22	1.32
chr19	15.22	12	0.79	17	1.12
chr20	14.64	10	0.68	10	0.68
chr21	18.46	19	1.03	17	0.92
chr22	23.37	10	0.43	11	0.47
chr23	23.13	14	0.61	20	0.86
chr24	18.49	14	0.76	25	1.35
chr25	16.68	9	0.54	18	1.08
chr26	12.28	6	0.49	16	1.30
chr27	14.47	10	0.69	17	1.18
chr28	12.35	16	1.30	16	1.30
	Total 316		475		
	Chr mean 11.29	0.65	16.96	0.97	
	SD 3.92	0.24	5.09	0.28	

Positional data for the [$(GAAA)_4$TCATTTATTCGCCA] and SciRoKo-defined $(GAAA)_n$ microsatellites indicates that *MINE*-1-like 5′ SIRs co-occur with $(GAAA)_n$ microsatellite loci in 201 of 316 instances (63.61% association; supplementary Table T3). The remaining 115 *MINE*-1-like regions of the *B. mori* genome (Bm*MINE*-1 elements) may represent TEs that are flanked by repeats too short or contained too many mismatches to be defined as microsatellites using our SciRoKo parameters. Our data from the *O. nubilalis* genome shows that 2–16 GAAA repeat units flank the 5′ SIR of On*MINE*-1 elements, which suggests that a proportion of flanking arrays would be missed by SciRoKo and that the association between *B. mori MINE*-1 (Bm*MINE*-1) elements may be underestimated. Multiple sequence alignment of 119 randomly chosen Bm*MINE*-1 elements from genome positions +20 to −750 of $(GAAA)_n$ indicate that regions ≥649 bp downstream of the microsatellite share extensive similarity (supplementary file F1). Also,113 of 119 (94.54%) Bm*MINE*-1s have a 5′ TC that is similar to *Helitron*-like TEs (Kapitonov and Jurka 2001), and show as conserved 5′ SIR [TC(A)TTTATTCRCCAR] (5′ TC in bold and optional A in parenthesis) that has a complementary 5′ IR similar to the secondary structure of On*MINE*-1 and

Drosophila DINE-1 elements. Despite the observation of short $(GAAA)_n$ microsatellite repeats at the 3′ terminus of On*MINE*-1 elements and AAGAA motif following the Pg*MINE*-1, no discernable $(GAAA)_n$ repeat is present at the 3′ terminus of the 119 aligned Bm*MINE*-1s.

Helitron-like TEs vary in size due to nested integration of retrotransposons (Du et al. 2009) or acquisition of genomic sequence by an unknown mechanism that results in obstacles in their annotation (Lal and Hannah 2005), and may be responsible for difficulties encountered in defining the 3′ end of Bm*MINE*-1 elements compared to On*MINE*-1 and Pg*MINE*-1 elements. Evidence for nested TE integration within a proportion of Bm*MINE*-1 elements was shown for an integration that starts at position 25,226 of the *B. mori* ABC transporter intron 2 (GenBank accession AB445460; Fig. 5). A BLASTn search of the GenBank nr database with positions 25,211–27,159 of the ABC transporter (query sequence containing $(GAAA)_n$ microsatellite through the beginning of exon 3) resulted in a "hit" to the *B. mori* BAC clone 048C11 (GenBank accession AP009015.1; positions 30,035–28,194). Sequence alignment indicates the two genomic regions share 99% similarity over 1.83 kb that terminates after a 3′ GTAA nucleotide motif, which suggests that Bm*MINE*-1 *Helitrons* may terminate at 3′ GTRR motifs compared to the 3′-CTRR *Helitron* motif (Kapitonov and Jurka 2001; Lai et al. 2005). Additionally, an internal ORF on the antisense strand was predicted to encode a 520 residue peptide called BMMINE-1p_1 that contains a non-LTR-like reverse transcriptase domain (RT; hit *E*-value $4e^{-36}$; cd01650; Marchler-Bauer et al. 2009) from residues 61–325 (Fig. 5; supplementary file F2). Furthermore, 28 Bm*MINE*-1 elements aligned in supplementary file F1) are predicted to encode a partial Bm*MINE*-1p_1 peptide, which suggest this coding region may be present within ~ 13.9% of the 201 integrations in the *B. mori* genome. No discernable poly(A) tail flanks the RT CDS which would suggest integration of a non-LTR element (Hutchison et al. 1989), and suggests that acquisition occurred by an yet unknown mechanism. The high prevalence of Bm*MINE*-1s predicted to encode the BMMINE-1p_1 enzyme suggests the CDS was acquired early in the *Helitron* propagation phase within the *B. mori* genome. The function of this acquired protein in the mechanism of *Helitron* mobility remains undetermined, and will be the focus of future research.

Bombyx mori Autonomous MINE-1 Helitrons

In the previous section, we predicted that Bm*MINE*-1 elements have acquired additional sequence that encodes a putatively functional enzyme but does not show characteristics of RepHel protein. Autonomous *Helitrons* are large (5,514–17,261 bp) and encode a 1000–3000 amino acid RepHel protein that contains rolling circle replication initiation (Rep) and DNA helicase (Hel) domains. RepHels are structurally diverse, yet carry out conserved endonuclease, helicase, DNA transfer, and ligation reactions during rolling circle replication (Kapitonov and Jurka 2007b). In addition to endonuclease and helicase domains, the RepHel proteins also can encode a zinc finger domain that may function during DNA binding, and animal RepHel proteins also have been shown to contain C-terminal endonuclease and cysteine proteinase domains (Poulter 2003; Zhou 2006). The functional roles of these latter two domains are yet to be elucidated, but is hypothesized that endonuclease activity may be involved in host genome integration (Kapitonov and Jurka 2001). Due to the potential omission of repetitive sequences from the *B. mori* genome assembly due to masking (Kapitonov and Jurka 2007b), *B. mori* strain Dazao contigs (GenBank accessions AADK01000001-AADK01066482) were used to replicate SciRoKo and 5′ SIR prediction as performed for the WGS. A total of 502 $(GAAA)_n$ microsatellites were predicted within strain Dazao contigs (average length of 23.35 bp and average mismatch of 0.37) distributed once every 1.28 Mb. BLASTx and tBLASTx searches with contig sequence 30 kb downstream of $(GAAA)_n$ microsatellites as queries of GenBank nr databases resulted in the identification of genome regions that may encode helicase or endonuclease domain proteins.

A total of 38 of 49 predicted ORFs encode proteins that show ≥65% identity to reverse transcriptase (RT) or endonuclease (EN) proteins (data not shown). DNA helicase-like

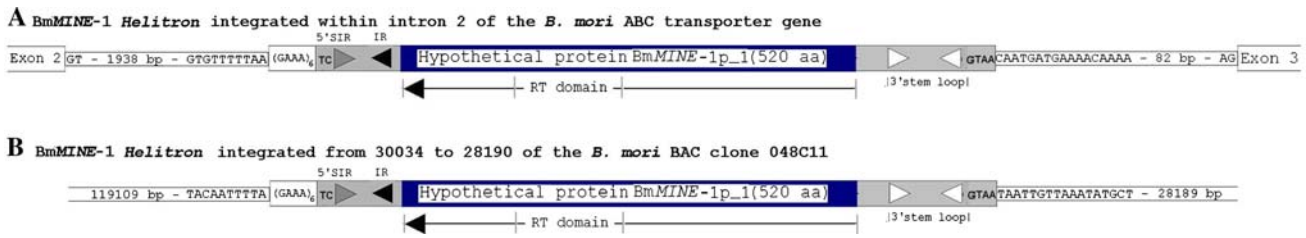


Fig. 5 Identification of *Bombyx mori* microsatellite-associated interspersed nuclear element (Bm*MINE*-1) that shows novel *Helitron*-like 5′-TC and 3′-GTAA terminal nucleotide motifs, and encode a multifunctional 520 amino acid reverse transcriptase (RT) enzyme, BMMINE-1p_1

ORFs are predicted within the *B. mori* contigs AADK 01000103 (positions 8,364–17,120) and AADK01010870 (positions 2886–5634) that are, respectively, downstream of GAA$(GAAA)_4$ACATTTATTCGCTAACGT (29.3 kb) and $(GAAA)_6$TCATTTATTCGCTAAGC motifs (1.3 kb). The 830 amino acid AADK01000103-encoded ORF shares 83% amino acid identity with an *Aedes aegypti* helicase (XP_001659679) and 73% identity with *D. melanogaster* CG4125-PA (AAF53481), and is similar to DEAD/H helicases (Bork and Koonin 1993; alignments not shown). A tBLASTx search of the *B. mori* genome build 2 indicates that the helicase within contig AADK01000103 corresponds to the predicted 1030 residue encoding gene BGIBMGA003539 (chr 5 from 14,957,972–14,971,385; 100% similarity, *E*-values $\leq 2 \times 10^{-91}$). The stop codon of BGIBMGA003539 is 29.3 kb upstream of the nearest Bm*MINE*-1 start, which is a distance greater than known *Helitron* sizes (0.5–17.7 kb; Kapitonov and Jurka 2001; Lal and Hannah 2005). Additionally, BGIBMGA003539 comparatively shows <26% similarity to RepHel domains encoded by *C. elegans* and *A. thaliana Helitrons*, which further suggests the gene is not *Helitron*-encoded.

BLAST results suggest that AADK01010870 may encode two independent proteins. First, the tBLASTx search of the GenBank nr protein database with the contig AADK01010870 sequence predicts that a *gag-pol* protein from a *B. mori* TRAS3-like non-LTR retrotransposon is encoded from positions 64–1457 (similarity ≥59%, *E*-values $\leq 6 \times 10^{-32}$; Kudo et al. 2001). Secondly, a region located on the antisense strand 1.3 kb downstream of a Bm*MINE*-1-like $(GAAA)_6$TCATTTATTCGCTAAGC motif encodes a putative helicase-like peptide that is highly similar to SF1 helicase superfamily members previously described within *Helitron*-encoded RepHel proteins from *A. thaliana* and *C. elegans* (Fig. 6). The helicase-like gene appears to lack introns, as is observed for insect *Helitrons* (Fig. 5). The predicted 889 residue-long peptide contains an ATP-dependent exonuclease V-like domain from amino acid positions 4–600 (COG0507), a DUF889-like region from position 602–748 (pfam05970), and a helicase-like domain from positions 791–853 (pfam02689; Marchler-Bauer et al. 2009; supplementary file F3). This putative RepHel protein is shorter than the typical 1000–3000 amino acid RepHel proteins due to an undefined N-terminal portion, which may be due to high levels of peptide divergence as is also observed between *A. thaliana* RepHel proteins ATHEL1p and ATHEL2p (Kapitonov and Jurka 2001) or possible truncation of the protein following integration of the TRAS3 retrotransposon we defined from positions 64–1457. Additional evidence for classification of the AADK01010870-encoded helicase as a *Helitron*-like RepHel-like protein was obtained by tBLASTx results that indicate ≥47% amino acid similarity to a *Philodina roseola Helitron* (PrHelitron; *E*-values $\leq 2 \times 10^{-89}$; GenBank accession DQ138288), and the observed level of interspecies variation similar to the 55% identity shared among the C-terminal regions of ATHEL1p and ATHEL2p (Kapitonov and Jurka 2001). This evidence supports the classification of this novel protein as a probable *Helitron*-encoded RepHel protein, which we named BMHEL1p. It is yet to be shown that this autonomous *B. mori Helitron*, *Helitron*1_BM, encoded BMHEL1p will direct transposition of Bm*MINE*-1s, but suggests that the *B. mori* genome may be continually modified by movement of these TEs and is a mode by which novel genetic variation is produced.

Conclusions

Helitron-like transposable elements from the superfamily *MINE*-1 are present as autonomous and non-autonomous forms within the genomes of Lepidoptera, and have an association with $(GAAA)_n$ microsatellites. Autonomous *Helitron*1_BM copies are comparatively rare within the *B. mori* genome, and encode a RepHel protein (BMHEL1p) that may be required for mobility for non-autonomous *MINE*-1 elements. Consistent with other *Helitrons*, genome integration of *MINE*-1 elements appears to occur between AA/TT dinucleotides but may have an additional requirement of the target site being present in or adjacent to an array of GAAA repeat units. It is known that GAA trinucleotides may form a secondary structure that interferes with efficient translation (Bidichandani et al. 1998) and that DNA secondary structure can play in role in target site recognition necessary prior to TE integration (Posey et al.

```
                |   Domain I   |    | Domain II |    | Domain III |    | Domain IV   |
Helitron1_CE  LDGPGGSGKTYLYITL-10-VACTAWTGIAANLL-37-VFIWDEASMVPKTAL-18-GKILILGGDFRQILPV-27-
Helitron1     VYGFGGTGKTFIWKTL-10-VLNVASSGIASLLL-38-LVIWDEAPMMSRFCF-20-GKILILGGDFRQILPV-27-
Helitron1_BM  LDAPGGTGKT*GVISL-11-ALALASSGIAATLL-39-LIAWDECTMAHKKSL-19-LMILL*AGDFRQTLPV-27-
XP_001602496  INGPGGSGKTFIYTTV-12-VCSMAFTGIAATLL-38-VFIWDEAPMAPRYAL-19-KIIIL*GGDFRQLLPI-28-
XP_970703     VDGPGGTGKTYLYN*V-11-YIAMAWTGMAATLL-38-LIIWDEASMAPRKAV-19-KIVVF*GGDFRQLLPV-28

              Domain IV   |   Domain VI    |   |    Domain V     |    | Domain IV   |
Helitron1_CE  LISNMR-124-GSVIILLRNLDVSAGLCNG-54-FALSINKAQGQSFSKIGL-12-YVALSRVRTKEGLIV
Helitron1     LTKNMR-149-GAPVMLLRNLDPKGGLCNG-55-FAMTINKSQGQSLEHVGL-12-YVALSRVTSKKGLKI
Helitron1_BM  LTTNMR-108-GTVIMILRNLNPG*RLCNG-53-FAMTINKSQGQSLSVCGI-12-YVACSRVGKPSALFV
XP_001602496  LSANMR-   -NCIVMLIRNISINEGLCNG-54-FAMTINKSQGQTFQRIAI-12-YVAMPRVRSWDSLKI
XP_970703     LTQNMR-   -GAIVMLLRNIDLEVGLCNG-55-FAITINKSQGQTFDKVGI-12-YVAMSR*********
```

Fig. 6 Alignment of RepHel SF1 superfamily-like helicase domains from *C. elegans Helitron*1_CE, *A. thaliana Helitron*1, and *B. mori Helitron*1_BM (BMHEL1p), and helicases from *Tribolium castaneum* (XP_970703) and *Nasonia vitripennis* (XP_001602496)

2006). This evidence suggests that $(GAAA)_n$ microsatellite loci may also form DNA secondary structures that are recognized by BMHEL1p or mediate *MINE*-1 integration. The presence of a greater number of GAAA units upstream of the 5′ SIR compared to downstream of the 3′ SIR, where protein–DNA interaction at the $(GAAA)_n$ induced DNA secondary structure may orient the BMHEL1p endonuclease domain such that cleavage and subsequent *Helitron* integration preferentially occurs downstream of the microsatellite. In contrast, microsatellite array expansion was shown to occur after On*MINE*-1 integration in *O. nubilalis*, which would suggest that secondary structures formed at the 5th end of the *Helitron* may increase the potential for slip strand mispairing to occur during DNA replication by stalling the progression of DNA polymerase I. Undoubtedly, additional investigations will be needed to elucidate the mechanism of *MINE*-1 transposition, and implications on genome structure, function, and evolution. These data provide insight into the characteristics of genome target sites that are liable for *Helitron* integration and may assist in elucidating the transposition mechanism of this class of DNA-based transposons.

Acknowledgments This research is a joint contribution from the United States Department of Agriculture Agriculture Research Service and the Iowa Agriculture and Home Economics Experiment Station, Ames (project 3543). Mention of proprietary products does not constitute an endorsement or a recommendation by USDA, or Iowa State University for its use.

References

Akagi H, Yokozeki Y, Inagaki A, Mori K, Fujimura T (2001) *Micron,* a microsatellite-targeting transposable element in the rice genome. Mol Genet Genomics 266:471–480

Anderson SJ, Gould P, Freeland JR (2007) Repetitive flanking sequences (ReFS): novel molecular markers from microsatellite families. Mol Ecol Notes 7:374–376

Anthony N, Gelembiuk G, Raterman D, Nice C, Ffrench-Constant R (2001) Brief report: isolation and characterization of microsatellite markers from the endangered Karner blue butterfly *Lycaeides melissa samuelis* (Lepidoptera). Hereditas 134:271–273

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) *Alu* repeats: a source for the genesis of primate microsatellites. Genomics 29:136–144

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8:382–392

Bidichandani SI, Ashizawa T, Patel PI (1998) The GAA triplet-repeat expansion in Friedrich ataxia interferes with transcription and may be associated with an unusual DNA structure. Am J Hum Genet 62:111–121

Bork P, Koonin EV (1993) An expanding family of helicase within the 'DEAD/H' superfamily. Nucleic Acids Res 11:751–752

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360

Bureau TE, Wessler SR (1994a) Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc Natl Acad Sci USA 91:1411–1415

Bureau TE, Wessler SR (1994b) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907–916

Casacuberta E, Casacuberta JM, Puigdomenech P, Monfort A (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. Plant J 16:79–85

Chen S, Li X (2007) Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. BMC Evol Biol 7:46

Coates BS, Sumerford DS, Hellmich RL, Lewis LC (2009) Repetitive genome elements in a European corn borer, *Ostrinia nubilalis*, bacterial artificial chromosome library was indicated by (BAC) end sequencing and development of sequence tag site (STS) markers: Implications for lepidopteran genomic research. Genome (in press)

Craig NL (1995) Unity in transposition reactions. Science 270:253–254

Du C, Fefelova N, Caronna J, He L, Dooner HK (2009) The polychromatic Helitron landscape of the maize genome. Proc Natl Acad Sci USA 106:19916–19921

Dufresne M, Hua-Van A, El Wahab HA, Ben M'Barek S, Vasnier C, Teysset L, Kema GHJ, Daboussi MJ (2006) Transposition of a fungal MITE through the action of a Tc1-like transposase. Genetics 175:441–452

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) The poly(A) tract of *Alu* repetitive elements is polymorphic in the human genome. Proc Natl Acad Sci USA 87:2951–2954

Fauvelot C, Cleary DFR, Menen SBJ (2006) Short-term impact of 1997/1998 ENSO-induced disturbance on abundance and genetic variation in a tropical butterfly. J Heredity 97:367–380

Feschotte CM, Mouchès C (2000) Evidence that a family of miniature inverted repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transpsoson. Mol Biol Evol 17:730–737

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzaglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S et al (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A.* fumigatus and A. oryzae. Nature 438:1105–1115

Gupta S, Gallavotti A, Stryker GA, Lai SK (2005) A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol 57:115–127

Hutchison C III, Hardies S, Loeb D, Shehee W, Edgell M (1989) LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, DC, pp 593–617

Jelinek WR, Toomey TP, Leinwand L, Duncan CH, Biro PA, Choudary PV, Weissman SM, Rubin CM, Houck CM, Denlinger PL, Schmid CW (1980) Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc Natl Acad Sci USA 77:1398–1402

Jurka J, Pethiyagoda C (1995) Sequences from primates: compilation and analysis. J Mol Evol 40:120–126

Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719

Kapitonov VV, Jurka J (2007a) *Helitrons* in fruit flies. Rephase reports 7:132–271

Kapitonov VV, Jurka J (2007b) *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trend Genet 23:521–529

Kofler R, Schlotterer C, Lelley T (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23:1683–1685

Kudo Y, Okazaki S, Anzai T, Fujiwara H (2001) Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. Mol Biol Evol 18:848–857

Lai J, Li Y, Messing J, Dooner H (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability in maize. Proc Natl Acad Sci USA 102:9068–9073

Lal SK, Hannah L (2005) *Helitrons* contribute to the lack of gene colinearity observed in modern maize inbreds. Proc Natl Acad Sci USA 102:9993–9994

Lerat E, Sémon M (2007) Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. Gene 396:303–311

Levinson G, Gutman GA (1987) Slipped strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Livak KL, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods 25:402–408

Locke J, Howard LT, Aippersbach N, Podemski L, Hodgetts RB (1999) The characterization of *DINE-1*, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. Chromosoma 108:356–366

López-Giráldez F, Andres O, Domingo-Roura X, Bosch M (2006) Análisis of carnivore microsatellites and their intimate association with tRNA-derived SINEs. BMC Genomics 7:269

Malausa T, Dalecky A, Ponsard S, Audiot P, Streiff R, Chaval Y, Bourguet D (2007) Genetic structure and gene flow in French populations of two *Ostrinia* taxa: host races or sibling species? Mol Ecol 16:4210–4222

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH (2009) CDD: Specific functional annotation with the conserved domain database. Nucleic Acids Res 37:D205–D210

Meglécz E, Pétenian F, Danchin E, D'Acier AC, Rasplus JY, Faure E (2004) High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. Mol Ecol 13:1693–1700

Meglécz E, Anderson SJ, Bourguet D, Butcher R, Caldas A, Cassel-Lundhagen A, d'Acier AC, Dawson DA, Faure N, Fauvelot C, Franck P, Harper G, Keyghobadi N, Kluetsch C, Muthulakshmi M, Nagaraju J, Patt A, Péténian F, Silvain JF, Wilcock HR (2007) Microsatellite flanking region similarities among different loci within insect species. Insect Mol Biol 16:175–185

Miao XX, Xub SJ, Ming JL, Li MW, Huang JH, Dai FY, Marino SW, Mills DR, Zeng P, Mita K, Jia SH, Zhang Y, Liu WB, Xiang H, Guo QH, Xu AY, Kong XY, Lin HX, Shi YZ, Lu G, Zhang X, Huang W, Yasukochi Y, Sugasaki T, Shimada T, Nagaraju J, Xiang ZH, Wang SY, Goldsmith MR, Lu C, Zhao GP, Huang YP (2005) Simple sequence repeat-based consensus linkage map of *Bombyx mori*. Proc Natl Acad Sci USA 45:16303–16308

Miller WJ, Nagel A, Bachmann J, Bachmann L (2000) Evolutionary dynamics of the *SGM* transposon family in the *Drosophila obscura* species group. Mol Biol Evol 17:1597–1609

Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimuar M, Ravikumar G, Shimomura M, Nagamura Y, Shin-I T, Abe H, Shimada T, Morishita S, Sasaki T (2004) The genome sequence of silkworm, *Bombyx mori*. DNA Res 11:27–36

Pemberton JM, Slate J, Bancroft DR, Barrett JA (1995) Nonamplifying alleles at microsatellite loci—a caution for parentage and population studies. Mol Ecol 4:249–252

Petersen G, Seberg O (2000) Phylogenetic evidence of excision of *stowaway* miniature inverted-repeat transposable elements in *Triticeae* (*Poaceae*). Mol Biol Evol 17:1589–1596

Posey JE, Pytlos MJ, Sinden RR, Roth DB (2006) Target DNA structure plays a crucial role in RAG transposition. PLoS Biol 4:E350

Poulter RT (2003) Vertebrate *Helitrons* and other novel *Helitrons*. Gene 313:201–212

Prasad MD, Muthulakshmi M, Madju M, Archak S, Mita K, Nagaraju J (2005) Survey analysis of microsatellites in the Silkworm, *Bombyx mori*: freqeuncy distribution, mutations, marker potential and their conservation in heterologous species. Genetics 169:197–214

Quesneville H, Nouaud D, Anxolabéhère D (2006) *P* elements and MITE relative in the whole genome sequence of *Anopheles gambiae*. BMC Genomics 7:214

Ramsey L, Macaulay M, Cardle L, Morgante M, degli Ivanissevich S, Maestri E, Powell W, Waugh R (1999) Intimate association of microsatellite repeats with retroelements and other dispersed repetitive elements in barley. Plant J 17:415–425

Reddy KD, Abraham EG, Nagaraju J (1999) Microsatellites in the silkworm, *Bombyx mori*: abundance, polymorphism, and strain characterization. Genome 42:1057–1065

Roeder GS, Rose AB, Pearlman RE (1985) Transposable element sequences involved in the enhancement of yeast gene expression. Proc Natl Acad Sci USA 82:5428–5432

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, NJ, pp 365–386

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetic analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

The International Silkworm Genome Consortium (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol 38:1036–1045

Tu Z (1997) Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. Proc Natl Acad Sci USA 94:7475–7480

Tu Z (2000) Molecular and evolutionary analysis of two divergent subfamilies of a novel miniature inverted repeat transposable element in the yellow fever mosquito, *Aedes aegypti*. Mol Biol Evol 17:1313–1325

Van't Hof AE, Brakefield PM, Saccheri IJ, Zwaan BJ (2007) Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera. Heredity 98:320–328

Vivas MV, Garcia-Planells J, Ruiz C, Marfany G, Paricio N, Gonzalez-Duarte R, de Frutos R (1999) *GEM*, a cluster of repetitive sequences in the *Drosophila subobscura* genome. Gene 229:47–57

Wang J, Wong GK, Ni P, Han Y, Huang X, Zhang J, Ye C, Zhang Y, Hu J, Zhang K et al (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. Genome Res 12:824–831

Wang J, Xia Q, He X, Dai M, Ruan J, Chen J, Yu G, Yuan H, Hu Y, Li R, Feng T, Ye C, Lu C, Wang J, Li S, Wong GKS, Yang H, Wang J, Xiang Z, Zhou Z, Yu J (2005) SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res 33:D399–D402

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814–821

Wicker T, Guyot R, Yahiaoui N, Keller B (2003) CACTA transposons in *Triticeae*. A diverse family of high-copy repetitive elements. Plant Phys 132:52–63

Wilder J, Hollocher H (2001) Mobile elements and the genesis of microsatellites in dipterans. Mol Biol Evol 18:384–392

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13779–13783

Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y et al (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306:1937–1940

Yandava CN, Gastier JM, Pulido JC, Brody T, Sheffield V, Murray J, Buetow K, Duyk GM (1997) Characterization of *Alu* repeats that are associated with trinucleotide and tetranucleotide repeat microsatellites. Genome Res 7:716–724

Yang HP, Barbash DA (2008) Abundant and species-specific *DINE*-1 transposable elements in 12 *Drosophila* genomes. Genome Biol 9:R39

Yang HP, Hung TL, You TL, Yang TH (2006) Genome-wide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. Genetics 173:189–196

Zhang DX (2004) Lepidopteran microsatellite DNA: redundant but promising. Trends Ecol Evol 19:507–509

Zhou Q (2006) Helitron transposons on the sex chromosome of the platyfish *Xiphphorus maculates* and their evolution in animal genomes. Zebrafish 3:39–52

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415

Zuliani G, Hobbs H (1990) A high frequency of length polymorphisms in repeated sequences adjacent to *Alu* sequences. Am J Hum Genet 46:963–969