Proliferation and copy number variation of BEL-like long terminal repeat retrotransposons within the Diabrotica virgifera virgifera genome

Brad S. Coates  
*Iowa State University*, brad.coates@ars.usda.gov

Lisa M. Fraser  
*Iowa State University*, lmfraser@iastate.edu

B. Wade French  
*United States Department of Agriculture*

Thomas W. Sappington  
*Iowa State University*, tsapping@iastate.edu

Follow this and additional works at: [http://lib.dr.iastate.edu/ent_pubs](http://lib.dr.iastate.edu/ent_pubs)

Part of the Animal Sciences Commons, Entomology Commons, Genetics Commons, and the Genomics Commons

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/ent_pubs/204](http://lib.dr.iastate.edu/ent_pubs/204). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Proliferation and copy number variation of BEL-like long terminal repeat retrotransposons within the Diabrotica virgifera virgifera genome

Brad S. Coates a,b,*, Lisa M. Fraser c,1, B. Wade French d,2, Thomas W. Sappington a,b,3

a United States Department of Agriculture, Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Iowa State University, Ames, IA 50011, USA
b Iowa State University, Department of Entomology, Ames, IA 50011, USA
c Iowa State University, Biomedical Sciences Department, Ames, IA 50011, USA
d United States Department of Agriculture, Agricultural Research Service, North Central Agricultural Research Laboratory, Brookings, SD 57006, USA

A R T I C L E   I N F O

Article history:
Accepted 26 September 2013
Available online 26 October 2013

Keywords:
Genome variation
Integration/excision mutation
Corn rootworm

A B S T R A C T

The proliferation of retrotransposons within a genome can contribute to increased size and affect the function of eukaryotic genes. BEL/Pao-like long-terminal repeat (LTR) retrotransposons were annotated from the highly adaptable insect species Diabrotica virgifera virgifera, the Western corn rootworm, using survey sequences from bacterial artificial chromosome (BAC) inserts and contigs derived from a low coverage next-generation genome sequence assembly. Eleven unique D. v. virgifera BEL elements were identified that contained full-length gag–pol coding sequences, whereas 88 different partial coding regions were characterized from partially assembled elements. Estimated genome copy number for full and partial BEL-like elements ranged from ~8 to 1582 among individual contigs using a normalized depth of coverage (DOC) among Illumina HiSeq reads (total genome copy number ~8821). BEL element copy number was correlated among different D. v. virgifera populations ($R^2 = 0.9846$), but individual element numbers varied $\leq 1.68$-fold and the total number varied by ~527 copies. These data indicate that BEL element proliferation likely contributed to a large genome size, and suggest that differences in copy number are a source of genetic variability among D. v. virgifera.

1. Introduction

Transposable elements (TEs) are mobile “selfish” DNAs that have propagated within and can comprise a significant proportion of eukaryotic genomes (SanMiguel et al., 1996). TEs are categorized as RNA-based class I (retrotransposons) or DNA based class II elements, which are mobilized by reverse transcription of an mRNA-intermediate or by conservative DNA “cut-and-paste” mechanisms, respectively (Kidwell and Lisch, 2001). Based on gene content, mode of integration and phylogenetic relationship among encoded proteins, retrotransposons are further classified into long terminal repeat (LTRs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (together also known as non-LTR), direct inverted repeat (DIR)-like, and Penelope-like elements (Wicker et al., 2007). LTR retroelements retain retroviral-like genes and a conserved mechanism for genome insertion (Boeke and Stoye, 1997). They also contain internal RNA Pol II promoters and termination signals that are necessary for the transcription of an mRNA intermediate containing gag and pol genes. The gag gene encodes a structural virus-like particle (VLP), and the resultant VLP proteins oligomerize to encapsulate the retroelement mRNAs. In contrast to retroviruses, encapsulated retrotransposon mRNAs cannot cross nuclear or cell membranes due to loss of the envelope (env) gene. The pol gene encodes a multifunctional enzyme with reverse transcriptase (RT), ribonuclease H (RH), protease (PR) and integrase (IN) activities. Based upon the order of genes (or functional domains) LTR retrotransposons are categorized as Ty1/copia elements with the order IN-RT, and Ty3/gypsy or BEL/Pao elements with the RT and IN domains inverted.

LTR retrotransposon integration is directed by the IN domain, and non-random distribution of integrations near gene-rich regions in
eukaryotic genomes suggests target-site specificity (Kim et al., 1998). For example, yeast Ty1 and Ty3 retroelements are preferentially located upstream of RNA Pol III promoters, where IN may interact with the Pol III transcriptional machinery or proteins associated with open chromatin (Bachman et al., 2005; Mou et al., 2006; Yieh et al., 2000, 2002). Movement of TEs within a genome can change chromosome structure (Eichler and Sankoff, 2003; Frischotte and Pritham, 2007; Kidwell and Lisch, 2001) and modulate the expression of nearby protein coding genes (Feschotte, 2008). Actively transcribed retrotransposons remain mobile within a genome, and are potent modulators of gene function by producing genomic variation at points of integration. Specifically, the integration of transposable elements (TEs) can cause mutations that give rise to phenotypic variation among individuals within populations (McClintock, 1950; Wendel and Wessler, 2000), and may generate genetic novelties that contribute to local adaptations (Gonzalez et al., 2010). TEs can also cause insertional knockout of a gene following integration within an open reading frame (Gahan et al., 2001), or decreased splicing efficiency and aberrant polyadenylation when integrated in an intron (Beeman et al., 1996; Kaer et al., 2011). Additionally, integrations within transcripational regulatory elements (promoters and enhancers) can affect the expression of nearby genes (Chung et al., 2007). Understanding TE structure, evolution and interactions within eukaryotic genomes is therefore important in uncovering roles in genome evolution and species adaptation. This is exemplified in the TE-induced mutations that are linked to insecticide resistance traits in arthropods (Aminetzach et al., 2005; Chung et al., 2007; Fabrick et al., 2011; Gahan et al., 2001), and supports the premise that TEs have genome-wide effects on genetic diversity and subsequent phenotypic variation (Kaminker et al., 2002).

The Western corn rootworm, Diabrotica virgifera virgifera (Coleoptera: Chrysomelidae), is a highly destructive insect pest of cultivated corn in most of the United States and Central Europe (Gray et al., 2009). Subterranean larvae damage roots (Chiang, 1973), which decreases the plant’s structural stability and capacity to absorb water and soil nutrients (Kahler et al., 1985). Adult beetles feed on corn silks and can reduce corn pollination rates as well as transmit plant pathogens (Gilbertson et al., 1986; Jensen, 1985). Protection of corn crops from D. v. virgifera is difficult in part due to recurring selection of populations for survival when exposed to large-scale pest management practices that rely on different insecticide modes of action, including organochlorine, organophosphate and carbamate chemistries (Gray et al., 2009; Meinke et al., 1998, 2009; Metcalf, 1986; Parimi et al., 2006). Recently, resistance to commercial corn hybrids that express a Bacillus thuringiensis (Bt) toxin has been documented in some D. v. virgifera populations (Gassmann et al., 2011). The genetic basis of many resistance traits in D. v. virgifera remains undetermined, but resistance to the organochlorine insecticide, aldrin, has been linked to a point mutation within the gamma-aminobutyric acid (GABA) receptor (Wang et al., 2013). Organophosphate resistance traits were partially linked to increased carboxylesterase activities in adult D. v. virgifera (Miota et al., 1998; Zhou et al., 2003), but the mechanism(s) underlying this adaptation remains unknown. The capacity of D. v. virgifera larvae and adults to rapidly adapt to insecticides highlights the need to investigate underlying genetic and genomic mechanisms (Gray et al., 2009; Sappington et al., 2006), with the anticipated benefits of developing sustainable management practices.

The D. v. virgifera genome size has been estimated as 2.58Gb (Coates et al., 2012). Sequence analysis indicated that ~16% of BAC ends consisted of retroelement protein coding sequence (RT, PR, RH and IN protein domains) and that retroelements comprise a large portion of assembled BAC insert sequence (Coates et al., 2012). Mobility of TEs in the D. v. virgifera genome was inferred from haplotype variation among BAC ends from full DvvBAC1 inserts were downloaded from the National Center for Biotechnology Information: JQ581035 to JMS81042 (Coates et al., 2012), KC248067 to KC248069 (Wang et al., 2013), and KC962414 to KC962431 (Coates et al., unpublished). BAC inserts were sequenced on a Roche 454 sequencer, and read data were assembled into contigs using the Newbler assembler (Roche Applied Sciences, Penzberg, Germany). The entire BAC insert could not be reconstructed into a single contig for all clones, so some HTGS entries are working drafts of unordered fragments (phase 1 submissions; see Coates et al., 2012).

Contigs of genomic DNA were also constructed by assembling short shotgun reads from a low-coverage sequencing of the D. v. virgifera genome. Specifically, 50 adult non-diapausing (ND) strain D v. virgifera beetles (Branson, 1976) were obtained from the United States Department of Agriculture — Agricultural Research Service, North Central Agricultural Research Laboratory (USDA-ARS, NCARL), Brookings, SD, and 80 wild adults were collected on 25 July 2012 from a cornfield at the Eastern Soil and Water Research Farm, Brookings, SD (BRSD). Beetles from both populations were maintained separately. Live beetles were pooled by population, starved for 2 days, then ground to a powder in liquid nitrogen. DNA was extracted from ~5mg of material using QiaGen G20 gravity flow columns according to manufacturer instructions (Qiagen, Valencia, CA). DNA quality and quantity were determined at the Iowa State University DNA Sequence and Synthesis Facility (ISU-DSSF) on an Agilent 2100 BioAnalyzer. Randomly sheared ~500 bp fragments were generated from ~2.0 μg of DNA and used to create L006 (ND strain) and L007 (BRSD) indexed sequencing libraries (ISU-DSSF) on an Agilent 2100 BioAnalyzer. Randomly sheared ~500 bp fragments were generated from ~2.0 μg of DNA and used to create L006 (ND strain) and L007 (BRSD) indexed sequencing libraries using the Illumina paired-end sample prep kit according to manufacturer instructions (Illumina, San Diego, CA). Single end 100-cycle reads were used to generate DNA sequence data from approximately equal proportions of each library in a single Illumina HiSeq 2000 flow cell at the ISU-DSSF. FASTQ formatted output was trimmed of read data with PHRED quality scores <20 using the script, TrimmingReads.pl (NGSToolKit, Patel and Jain, 2012). A total of 30 million reads (~3.0 Gbp or ~1.17× genome coverage) from each of the trimmed FASTQ sequence files were loaded in the Velvet assembler (Zerbino and Birney, 2008) and assembled using the de Bruijn graph method with a hash size (k-mer) of 21, no coverage cutoff, and a minimum contig length (−min_contig_lgth) of 1000.

Contigs from HTGS accessions and the de novo genome assembly were used to create a local BLAST database (dbDvvBACs). Contigs of HTGS accessions and the de novo genome assembly were used to create a local BLAST database (dbDvvBACs).
Contigs with significant “hits” were retrieved from the Velvet output file using the PERL script “get_selected_sequences.pl” (downloaded at http://alrlab.research.pdx.edu/aquifaeces/scripts/get_selected_sequences.pl). Longest open reading frames were predicted by ORF Finder (http://www.bioinformatics.org/sms2/orf_find.html) using the parameters of an ATG start codon, minimum length ≥ 300 codons in standard genetic code and in all six possible reading frames. Each putative BEL/Pao ORF as well as sequence encompassing up to 1000 bp upstream of the start codon and 1000 bp downstream of the stop codons were used as input for the program LTR Finder (Zhao and Wang 2007; default parameters). In instances where LTR Finder predicted no LTR retrotransposons, LTR regions were identified by alignment of DNA sequences located immediately upstream and downstream of start and stop codons using ClustalW. HTGS accessions with putative D. v. virgifera LTR retrotransposons were loaded into the Artemis genome viewer (Rutherford et al., 2000), and ORFs, direct flanking repeats, and target site duplications were annotated manually. Functional protein domains within putative D. v. virgifera BEL/Pao gag-pol proteins were identified by query of the conserved domain database (CDD; Marchler-Bauer et al., 2012). ORFs were stored in the file DvBelORFs.fasta (Supplementary File S1).  

2.2. Phylogenetic reconstruction

Reverse transcriptase (RT) domains were manually trimmed from amino acid sequences of previously described BEL/Pao LTR retrotransposons (see accessions in the previous section; representing Pao, Saci-1, BEL, Tas and Sazu subfamilies). Additionally, the RT domains were trimmed from the derived protein coding sequence of 37 putative BEL/Pao-like elements predicted within assembled D. v. virgifera contigs. Sequences were loaded into the MEGA 5.0 alignment utility (Tamura et al., 2011), and a multiple sequence alignment was constructed using ClustalW default parameters and the Gonnet protein weight matrix. The gamma shape and best model parameters were estimated by maximum likelihood (ML) using MEGA 5.0. Phylogenetic trees were constructed from exported .meg files using the ML method using complete deletion of gapped data under the general reverse transcriptase model (rtREV; Dimmic et al., 2002) + G (gamma distributed rate variation). Bootstrap support at each node was based on 1000 pseudoreplicates of the aligned data (Felsenstein, 1985). The RT domain from the D. melanogaster gypsy element (P10401.1) served as the outgroup.

2.3. Estimation of BEL copy number and expression

BEL-like retrotransposon copy number within the D. v. virgifera genome was estimated by mapping whole genome shotgun sequence read data from the ND laboratory strain and BRSD field population to 99 contigs in the DvBelORFs.fasta file (Supplementary File S1). The cDNA sequence from the D. v. virgifera cadherin (cad; GenBank accession EF531715.1) and forager 1 (for1; DQ91742.1) genes were used as internal references for normalizing mapped Illumina read data derived from the ND strain and BRSD field population libraries, and provided single-copy controls to normalize the estimates of BEL retrotransposon copy number. Specifically, 50 million Illumina reads from ND strain and BRSD FASTQ files (previously trimmed of segments with PHRED quality scores <20) were aligned to the 99 D. v. virgifera contigs encoding BEL/Pao-like ORFs as well as to the cad and for1 cDNA sequences using the short read aligner Bowtie (Langmead et al., 2009; mismatches allowed (−v) 3; output in SAM format (−S)). The copy numbers of LTR retrotransposons were estimated from a ratio of the average coverage of mapped Illumina reads across each contig to that of the single-copy genes cad and for1 (i.e., mean coverage of contig containing a BEL-like element/mean coverage of single-copy gene). Copy number was also estimated as reads per kilobase per million mapped (RPKM) according to Tenailleon et al. (2011), and RPKM values for BEL elements were normalized using the ratio of RPKM to single-copy genes as described above.

Variation in copy number of each BEL element between ND strain and BRSD populations was estimated by plotting the estimated ratio-normalized copy number for each of the 99 BEL-containing contigs. This was performed separately for the mean read depths and RPKM estimates for each library. Estimates of BEL copy number normalized by the cad gene were used to calculate a ratio of BEL copy number within each population based on mean read depth. A ratio was similarly calculated using RPKM values. The ratio of copy number between populations was regressed on the log2 scaled difference between the ND strain vs. BRSD population.

Expressed sequence tag (EST) data for D. v. virgifera head and midgut tissues were downloaded from GenBank dbEST accessions EW761110. 1–EW777358.1 (Knolhoff et al., 2010), and CN497248.1–CN498776.1 (Siegfried et al., 2005), and used to construct a blast database (dbDvEST). dbDvEST was queried with translated protein sequence from D. v. virgifera genomic contigs encoding BEL/Pao element gag-pol functional domains using the tblastn algorithm, and results were filtered for E-values ≥ 10−20 and percent identities ≥ 40. ESTs showing >98% similarity to a characterized BEL element from the non-diapausing strain were considered to be derived from the same mobile and multicopy locus.

3. Results and discussion

3.1. Identification and annotation of LTR retrotransposons

TE integrations have been implicated in the alteration of gene structure and function that has led to the evolution of insecticide resistance traits. Initial annotation of TE structure and abundance are fundamental for future studies of genome evolution and investigations of TE roles in shaping genetic variability in a species. To address this knowledge gap in the highly adaptable corn pest D. v. virgifera, we characterized the diversity and prevalence of BEL-like LTR retrotransposons within its genome. BEL/Pao, Ty1/copia and Ty4/gypsy groups are the main types of LTR-retrotransposon among animal species (Xiong and Eckbush, 1990), where BEL/Pao retroelements represent the most recently described clade identified through phylogenetic analysis of gag-pol coding regions (Frame et al., 2001). BEL/Pao element diversity remains relatively unknown, and their description herein provides insight into BEL/Pao structure, functional constraint, and diversification in eukaryotic genomes.

BEL/Pao elements were successfully identified and annotated from an assembly of the D. v. virgifera genome from low coverage read data and BAC full insert sequences. De novo assembly of an ~1.17× coverage depth of the D. v. virgifera genome from the ND strain (NCBI BioProject PRJNA205954; BioSample SAMN02183419: ND_pool2012; NCBI short read archive (SRA) run SRR921383) and a Brookings, SD field population (NCBI BioProject PRJNA206802: BioSample SAMN02190085; SRA run SRR921430) resulted in the assembly of 11.7 million reads into 4173 contigs (314,773 bp total contig length; range 1000 to 73,314 bp; N50 2646 bp). Due to the use of low genome coverage and unpaired read data during assembly, scaffolding or assembling larger portions of the D. v. virgifera genome were not attempted. Assembly parameters and number of reads affect contig number and length. Although not investigated in this study, the relaxation of assembly parameters probably increased contig length and decreased contig number, but also increased the probability of assembling non-homologous reads. Query of the 4173 D. v. virgifera genomic contigs with previously described BEL/Pao gag-pol protein sequences resulted in “hits” to 99 contigs with amino acid similarities ≥ 37.6% (contig size 1482 to 9296 bp; ORF size 434 to 1866 amino acids; Supplementary File S1). Similarly, queries of previously assembled BAC inserts (Coates et al., 2012; Wang et al., 2013) identified regions of clones 213A05 and 278L20 with ≥ 38% identity to the T. castaneum
BEL12_AG TcBEL LTR transposon (GenBank accession XP_001809963.1; remaining results not shown).

A total of 11 full-length and 88 partial BEL/Pao coding sequences were identified from D. v. virgifera genomic contigs and BAC inserts. Neither the LTR Finder utility nor alignment of flanking sequences with ClustalW predicted direct repeats upstream or downstream of the ORF for most elements, and was therefore difficult in assembling of non-unique sequence from the directly repeated LTR sequences. In instances where LTRs assembled, unique block deletions and single nucleotide polymorphisms were present, suggesting an increased age since integration. Full-length ORFs encoded putative gag-pol-like peptides with functional domains in the order GAG-PR-RT-RH-IN (Supplementary File S1), which is consistent with other known BEL/Pao and Ty3/gypsy elements (Frame et al., 2001; Xiong and Eickbush, 1990), while differing from the Ty1/copia order of GAG-PR-IN-RT-RH. An LTR promoter/primer binding site (PBS) predicted to bind tRNAYtctTTT was identified upstream of the GAG domain. Additionally, a 15 bp polypurine tract (PPT), which functions during reverse transcription of gag-pol mRNAs, was annotated (Fig. 1). All putative translated D. v. virgifera gag-domains contained a characteristic BEL/Pao triple Cys–His box motif (CX2CX3–4HXCX5C, CX2CX3–4HX2–5C, and CX2–4CX3HX4H) (Fig. 2A), and not the single zinc fingers found in Ty3/gypsy and Ty1/copia gag-domains (Bowen and McDonald, 1999; Copeland et al., 2005; Frame et al., 2001). BEL/Pao protease (PR) domains are characterized by unique ALLDXG and LIGXD His box motifs (Abe et al., 2001), and D. v. virgifera elements varied at the first position of the ALLDXG motif (Fig. 2B). BEL/Pao RT domains are presumably derived from Ty3/gypsy elements (Frame et al., 2001) and share a unique TVDN active site instead of F/YXDD found among other LTR-retrotransposons (Felder et al., 1994; Frame et al., 2001). The D. v. virgifera BEL/Pao element RT active site, YWVD, is similar to that of the BEL-like Ninja element from D. simulans (Ogura et al., 1996) and BmPao from B. mori (Xiong et al., 1993) (Fig. 3). Seven conserved amino acid blocks characteristic of the BEL/Pao element RT domain (Xiong et al., 1993) were identified in the putative D. v. virgifera BEL/Pao elements as well (Fig. 3). Finally, IN domains of D. v. virgifera elements contained a BEL/Pao-like Cys–His box and a conserved catalytic triad (DX65DX49E) (Fig. 4; Copeland et al., 2005). The spacing between catalytic residues within D. v. virgifera IN domains was DX(60)DX40 to 53E among all elements, except for those from contig0034 and contig2516 that had a DX40DX...4E and DX93DX...40E spacing, respectively (Fig. 4).

3.2. Phylogenetic reconstruction

Phylogenetic reconstruction of the BEL/Pao lineage was conducted using RT domains from 37 unique D. v. virgifera elements, as well as representatives from five BEL/Pao subfamilies (Pao, Sinbad/Saci-1, BEL, Tas, and Suzu) previously identified by phylogenetic analysis of RT domains by Copeland et al. (2005). A 157 amino acid long consensus alignment was generated (Supplementary File S2), from which a gamma parameter for among site rate variation was estimated by maximum likelihood as 1.2832 (likelihood = −13,138.012). The resulting ML tree placed constituent members for the RT domain alignment into the major subfamilies of the BEL/Pao LTR retrotransposons; Pao, BEL, Saci-1, Tas, and Suzu (Fig. 5). Tree topology was similar to that previously described among the major subfamilies defined by Copeland et al. (2005). Thirty six of the 37 unique D. v. virgifera RT domains were grouped with previously described BEL subfamily members Max, Moose, Roo, and T. castaneum BEL12-AG (TcBEL) with 72% of bootstrap iterations supporting this clade. Subgroups within the D. v. virgifera BEL-lineage lacked strong bootstrap support, perhaps reflecting low sequence homology and thus rapid sequence divergence among RT domains.

3.3. Estimation of BEL copy number and expression

3.3.1. Genome copy number variation between populations

Ty3/gypsy elements are typically more prevalent in eukaryotic genomes than BEL/Pao-like and Ty1/copia elements with intermediate and low copy numbers, respectively (Felder et al., 1994; Xiong et al., 1993). Estimating TE numbers from species that lack whole genome sequence assemblies presents unique challenges. NGS data has been used to determine unique copy number variants (CNVs) in genomes by mapping reads to a reference genome, but this is an evolving field where best practices or pipelines are not yet established (Teo et al., 2012). Depth of coverage (DOC) estimates have been applied to predict CNVs by assuming that sequence data is generated uniformly across a genome (following a Poisson distribution) and depth of aligned reads is proportional to genome copy numbers (Xie and Tammi, 2009). Our DOC analyses of the 99 contigs with BEL coding regions resulted in mean read depths among individual BEL elements ranging from 12.3 to 2024.5 (normalized 9.6 to 1581.6) and 11.0 to 1842.5 (normalized 8.3 to 1395.8) from ND strain and BRSD libraries, respectively (Supplementary Table S3). GC nucleotide content can lead to under- or overrepresentation of sequence reads acquired by NGS technologies, and the need for GC-corrections has been proposed (Abzyov et al., 2011). However, the DOC (mean read depth) for the single copy D. v. virgifera cod gene (35.68% GC content) of 1.28 and 1.32 respectively for ND strain and BRSD libraries was not significantly different than the 1.18 and 1.41 mean read depths for the single copy D. v. virgifera for1 gene (41.79% GC), respectively estimated from ND strain and BRSD libraries. Because the DOC observed between single copy genes was similar, we reasoned
that GC correction was not required. Subsequent normalization of data based on DOG for the single copy cod gene suggested that individual contigs containing BEL elements have copy numbers from ~10 to 1582. Based on the difference in total genome copy number between the ND strain \( (n=82) \) and the reference \( (n=8249) \), we estimated a total CNV between populations of ~527.

Fold difference in DOG (mean read depth) estimated between individual BEL-like contigs ranged from 1.278 to ~1.688 between ND and the reference. The estimated relative BEL element copy numbers for variant BEL elements (represented as individual contigs) have fold change. The estimated relative BEL element copy numbers for variant BEL elements (represented as individual contigs) have fold change. The estimated relative BEL element copy numbers for variant BEL elements (represented as individual contigs) have fold change. The estimated relative BEL element copy numbers for variant BEL elements (represented as individual contigs) have fold change. The estimated relative BEL element copy numbers for variant BEL elements (represented as individual contigs) have fold change.

Fig. 2. Multiple alignment of amino acid sequences of \( D. v. \) virgifera BEL-like retrotransposons. A) Conserved triple Cys–His box motifs (\( CX_2CX_3CX_4HX_3 \)) and \( CX_2CX_3HX_4 \) of zinc fingers within the gag gene-encode nucleocapsid protein are underlined, and B) ALLDXG and LIGXD motifs within the protease (PR) domain are underlined and catalytic residues are marked with a (+). Residues showing 100% conservation are shaded in black, and chemically similar amino acids in a tl e a s th a l fo f

member copy numbers in \( Z. \) mays and \( Z. \) luxurians genomes are also highly correlated (Tenaillon et al., 2011), whereas Hawkins et al. (2006) described differential amplification of \( copia \)-like elements among species of \( Gossypium \). Correlation between copy numbers of BEL-like contigs in \( D. v. \) virgifera does not conflict with the high total CNV between populations described above, rather it suggests that copy numbers for variant BEL elements (represented as individual contigs) have not diverged significantly over evolutionary time. Transposition of the BEL-like \( \) Rho element of \( D. \) melanogaster causes intraspecies variation in copy number (Papacét al., 2007), and associated genome-wide effects of such mutations were negatively correlated with overall fitness within replicated \( D. \) melanogaster lines (Pasyukova et al., 2004). Our data suggest that BEL element CNV is likely present. Identifying the
Fig. 3. Alignment of reverse transcriptase (RT) domains from BEL-like LTR retrotransposon gag-pol proteins. Identical residues among previously described Pao and BEL elements (see Fig. 2 for accession numbers) and derived D. v. virgifera peptides are shaded in black, and chemically similar amino acids in at least half of the sequences are shaded in gray. The RT catalytic domain are marked (+), and seven conserved RT blocks (Xiong et al., 1993) are underlined.

Fig. 4. Alignment of the integrase (IN) domain from BEL-like LTR retrotransposon gag-pol proteins. Identical residues among previously described Pao and BEL elements (see Fig. 2 for accession numbers) and derived D. v. virgifera peptides are shaded in black, and chemically similar amino acids in at least half of the sequences are shaded in gray. Residues of a conserved Cys–His box motif of a zinc-finger are indicated with (*), and those within the catalytic domain are marked (+).
3.3.2. Transcription of BEL elements

Expressed sequence tags represent gene transcripts in cDNA libraries constructed from a specific tissue, developmental stage, phenotype, or experimental treatment group (Adams et al., 1991). Published *D. v. virgifera* EST datasets from adult head (Knolhoff et al., 2010) and larval midgut tissues (Siegfried et al., 2005) were examined for evidence of BEL element transcription. Results revealed that 18 transcripts expressed in the head (EW765187, EW765400, EW766938, EW767051, EW767234, EW767949, EW768879, EW768975, EW769268, EW770478, EW770495, EW770524, EW770537, EW771419, EW771663, EW771738, EW777129) had ≥40% identity (E-values ≤ 1 × 10^{-36}) to *D. v. virgifera* of BEL LTR retrotransposon proteins. The dbEST accessions EW767051.0, EW770495.1 and EW771419.1 showed ≥98.7% nucleic acid similarity to the BEL element assembled in contig0042, suggesting that this BEL element is actively transcribed in the genome. None of the other transcripts showed high similarity to BEL-like retrotransposons assembled from ND or BRSD libraries. Since the DOC (estimated copy number) of BEL-like elements represented by contig0042 was high, this suggests that transcription may be very low from any given element, and that overall these transcripts are rare within EST datasets. This further suggests that genomic silencing may not be completely effective in suppressing the proliferation of TE integrations in the genome (Lisch and Bennetzen, 2011; Slotkin and Martienssen, 2007). Alternatively, TEs may have been domesticated after integration into genes by subsequent gain of function within gene regulatory networks (Feschotte, 2008; Volff, 2006). The role that transcribed LTR elements play in *D. v. virgifera* genome function remains unknown, although co-option into various gene pathways might explain the observed tolerance of TE proliferation in this species.

In contrast to our ability to assign the origin of 3 transcripts to a particular BEL element, the remaining 15 putative BEL retrotransposon-derived ESTs showed ≤77.8 similarities to BEL elements identified in our low coverage de novo assembly. This might result from sequence differences in BEL elements in the populations from Illinois used to construct the head EST library (Knolhoff et al., 2010). A more likely explanation may reside in the incomplete assembly of BEL elements, as highlighted by the 88 partial elements annotated in the current study. The low-coverage genome assembly may be insufficient for the assembly of low copy number BEL elements. Regardless, search results from EST data resources suggest that LTR retrotransposons are transcribed and may still be actively mobile in the *D. v. virgifera* genome.

3.4. Conclusions

BEL-like LTR retrotransposons have proliferated to a high copy number within the *D. v. virgifera* genome, and their presence in EST sequence data suggests that several elements are actively transcribed and may remain mobile. Variation in BEL copy number was detected between two populations, which further suggests that these retroelements are capable of modifying genome structure and may have consequences for gene function. The role of host suppression in TE proliferation, as well as
as the potential function of TEs in the creation of genetic novelty and the
subsequent evolution of D. v. virgifera phenotypes, will be the subjects of
future research efforts.

Supplementary data to this article can be found online at http://dx.

Conflict of interest

The authors have no conflicts of interest to declare.

Acknowledgments

This research was a joint contribution from the USDA, Agricultural
Research Service (CRIS Project 3625-22000-017-00D), and the Iowa
Agriculture and Home Economics Experiment Station, Iowa State
University, Ames, IA (Project 3543). The authors would like to thank
USDA-ARS Technicians, Robert Gunnarson and Chad Nielson, for
performing associated laboratory protocols, and collecting and rearing
of beetles. This article reports the results of research only. Mention of
a proprietary product does not constitute an endorsement or a recom-
mandation by USDA for their use.

References

Abe, H., et al., 2001. Novel Pao-like retrotransposons (Kamikaze and Yamato) from the
silkworm species Bombyx mori and B. mandarina: common structural features of

Abyzov, A., et al., 2011. CNVnator: an approach to discover, genotype, and characterize
typical and atypical CNVs from family and population genome sequencing. Genome
Res. 21, 974–984.

Adams, M.D., et al., 1991. Complementary DNA sequencing: expressed sequence tags and
human genome project. Science 252, 1651–1656.


Bachman, N., Gelbart, M.E., Tsukiyama, T., Boeke, J.D., 2005. TFIIB subunit Bdp1p is
required for periodic integration of the Ty1 retrotransposon and targeting of Iw2p
to S. cerevisiae DNA. Genes Dev. 19, 955–964.

Beeman, R.W., et al., 1996. Woot, an active gypsy-class retrotransposon in the flour beetle,
Tribolium castaneum, is associated with a recent mutation. Genetics 143, 417–426.


Xie, C., Tamm, M.T., 2009. CNV-Seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 10, 80.


