Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 604076, 9 pages
doi:10.1155/2012/604076

Research Article

Distribution of Genes and Repetitive Elements in the *Diabrotica virgifera virgifera* Genome Estimated Using BAC Sequencing

Brad S. Coates,[1,2] Analiza P. Alves,[3] Haichuan Wang,[3] Kimberly K. O. Walden,[4] B. Wade French,[5] Nicholas J. Miller,[3] Craig A. Abel,[1] Hugh M. Robertson,[4] Thomas W. Sappington,[1,2] and Blair D. Siegfried[3]

[1] *Corn Insects and Crop Genetics Research Unit, ARS, USDA, Ames, IA 50011, USA*
[2] *Department of Entomology, Iowa State University, Ames, IA 50011, USA*
[3] *Department of Entomology, University of Nebraska, Lincoln, NE 68583, USA*
[4] *University of Illinois, Champaign-Urbana, IL 61801, USA*
[5] *North Central Agricultural Research Laboratory, Brookings, ARS, USDA, SD 57006, USA*

Correspondence should be addressed to Brad S. Coates, brad.coates@ars.usda.gov

Received 29 February 2012; Accepted 16 May 2012

Academic Editor: Yong Lim

Feeding damage caused by the western corn rootworm, *Diabrotica virgifera virgifera*, is destructive to corn plants in North America and Europe where control remains challenging due to evolution of resistance to chemical and transgenic toxins. A BAC library, DvvBAC1, containing 109,486 clones with 104 ± 34.5 kb inserts was created, which has an ~4.56X genome coverage based upon a 2.58 Gb (2.80 pg) flow cytometry-estimated haploid genome size. Paired end sequencing of 1037 BAC inserts produced 1.17 Mb of data (~0.05% genome coverage) and indicated ~9.4 and 16.0% of reads encode, respectively, endogenous genes and transposable elements (TEs). Sequencing genes within BAC full inserts demonstrated that TE densities are high within intergenic and intron regions and contribute to the increased gene size. Comparison of homologous genome regions cloned within different BAC clones indicated that TE movement may cause haplotype variation within the inbred strain. The data presented here indicate that the *D. virgifera virgifera* genome is large in size and contains a high proportion of repetitive sequence. These BAC sequencing methods that are applicable for characterization of genomes prior to sequencing may likely be valuable resources for genome annotation as well as scaffolding.

1. Introduction

Bacterial artificial chromosome (BAC) libraries are composed of physical constructs that contain large genomic DNA inserts and provide a tool for the molecular genetic research of organisms of interest. For instance, anonymous genetic markers linked to genes that control insecticide resistance traits have been identified on BAC clones, and, following subsequent sequencing of cloned inserts, allowed the characterization of gene(s) that influence the expression of these traits [1]. Furthermore, sequence data from BAC inserts provide a means to evaluate genome structure, including the estimation of repetitive element densities [2] and the relative gene content of a species [3]. BAC clones are also useful for the construction of physical maps that represent contiguous sequence from an entire genome or genomic regions, and these assemblies have proven useful for determination of minimum tiling paths prior to BAC-by-BAC sequencing of large or highly repetitive genomes [4]. Scaffolding takes advantage of paired BAC end sequence (BES) data which provide direct physical linkages between sequence tags [5] and may assist in the scaffolding of contigs assembled from mate paired reads from next generation sequencing technologies.

The western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae), is a beetle native to North

America, which is adapted to feeding on a limited number of grasses including corn [6]. The ancestral geographic range of *D. virgifera virgifera* extended from present day Mexico into the Southwest United States and Great Plains, but an eastward range expansion began in the 1940s that coincided with the widespread cultivation of continuously planted corn in the central United States [7, 8]. *D. virgifera virgifera* was accidentally introduced into Central Europe in the early 1990s [9], and subsequent transatlantic and intra-European introductions have contributed to its contemporary geographic range in Europe [10, 11]. *D. virgifera virgifera* has one generation per year. Individuals overwinter as diapausing eggs which hatch in the spring, and subterranean larvae feed on corn roots [12]. Root damage caused by *D. virgifera virgifera* reduces the plant's ability to absorb soil nutrients and compromises structural stability [13]. Upon pupation and emergence from the soil, adult corn rootworm beetles can persist in fields for up to 4 weeks, can reduce seed pollination rates through feeding damage to corn silks (stamen) and can vector maize chlorotic mottle virus [14] and stalk rot fungus [15].

Larval feeding damage can be suppressed by systemic seed treatments, soil-applied insecticides, or transgenic corn hybrids that express *Bacillus-thuringiensis*-(Bt) derived insecticidal proteins. However, resistance to both chemical and Bt toxins are documented [16–19]. *D. virgifera virgifera* populations have also been managed by an alternating corn-soybean crop rotation (grass-legume rotation) [20], which negates the need for insecticide applications. This control strategy is based on a strong female preference to oviposit in soil at the base of corn plants, and the cospecialization of larvae for feeding on grass roots. However, female *D. virgifera virgifera* phenotypes have evolved that are no longer specifically attracted to cornfields but will lay eggs near other plants [21–23]. In the subsequent crop production year, progeny of these variant *D. virgifera virgifera* females will emerge in and damage first-year corn crops. This adaptive loss of adult fidelity in oviposition behavior has defeated the use of corn-soybean rotation as an effective control practice in many corn growing regions of the United States [22].

The propensity for corn rootworm to adapt to control measures has raised concern among producers, scientists, and regulatory agencies, and the need to investigate the underlying genetic mechanisms for adaptation is critical to developing sustainable pest management approaches [7, 24, 25]. In anticipation of a recently initiated whole genome sequencing (WGS) effort for *D. virgifera virgifera* that aims to build a foundation for future genetic and genomics research [25], we have determined the haploid genome size and have estimated gene and repetitive fraction densities from BAC sequencing data. These data and resources will facilitate annotation and contig scaffolding efforts of the upcoming WGS project.

2. Materials and Methods

2.1. Genome Size Estimation. Three males from the inbred nondiapausing *D. virgifera virgifera* colony at USDA-ARS, North Central Agricultural Research Laboratory in Brookings, SD [26] were starved for 24 hr, and homogenized with a razor blade in 0.5 ml chopping buffer (15 mM HEPES, 1 mM EDTA, 80 mM KCl, 20 mM NaCl, 300 mM sucrose, 0.2% Triton X-100, 0.5 mM spermine tetrahydrochloride, 0.25 mM PVP). Homogenate was filtered through 20 um nylon mesh, centrifugated at 100 ×g for 5 minutes, and nuclei suspended in 0.5 ml propidium iodide (PI) staining buffer (10 mM $MgSO_2$, 50 mM KCl, 5 mM HEPES, 0.1% DL-dithiothreitol, 2.5% Triton X-100, 100 ug/mL propidium iodide). Nuclei from the *Zea mays* inbred line B73 (genome size 2.5 Gb) and *Glycine max* line Williams 83 (1.115 Gb) were similarly prepared. Propidium iodide stained nuclei were analyzed on a BD Biosciences (San Jose, CA, USA) FACSCanto flow cytometer equipped with a 488 nm laser and 610/620 emission filter. Estimates for standards (B73 and Williams 83) and *D. virgifera virgifera* were performed in triplicate. The estimated *D. virgifera virgifera* genome size was calculated from PI signals [27] and converted to pg estimates [28].

2.2. BAC Library Construction, End Sequencing, and Annotation. Genomic DNA was extracted from ~100 individuals of the *D. virgifera virgifera* nondiapausing strain, pooled, and fractionated by partial digestion with *Hind*III, fragments between ~100 and 150 kb were excised, and these inserts were ligated into the pCC1 BAC vector (Epicentre, Madison, Wl, USA). Constructs were used to transform the *Escherichia coli* strain DH10B T1 by electroporation. Transformants were plated on LB agar (12.5 μg mL^{-1} chloramphenicol, 80 μg mL^{-1} X-Gal, and 100 0.5 mM IPTG) and a total of 110,592 BAC clones were arrayed on 288 individual 384-well plates to comprise the DvvBAC1 library. The mean insert size within DvvBAC1 was estimated by contour-clamped homogeneous electric field (CHEF) electrophoresis of *Not*I digested BAC DNA from 96 clones on a 0.9% agarose in 0.5X TAE buffer gel ramp run with a pulse time 5–15 s at 5 V/cm for 24 hrs and 4°C. Insert size estimates were made by comparison to the MidRange II PFG Marker (New England Biolabs, Ipswich, MA, USA), and the fold genome coverage of DvvBAC1 was estimated according to Clark and Carbon [29]. BAC DNA from 1152 DvvBAC1 clones (plates 217, 218, and 227) were purified and sequenced and annotated as described by Coates et al. [2], and sequence data deposited into the GenBank genome survey sequence (GSS) database (accession numbers JM104642–JM106797).

2.3. BAC Screening and Full Insert Sequencing. DNA from DvvBAC1 clones were pooled into matrix, row, and column pools according to Yim et al. [30] and used in PCR reactions as described by Coates et al. [31]. DNA from DvvBAC1 clones was purified using the Large Construct Purification Kit (Qiagen, Valencia, CA, USA) according to the manufacturers instructions, and DNA preparations run on 0.8% agarose gel electrophoresis. BAC DNA was used to create individual mid-tagged libraries (RL1 to RL10) and each was sequenced on Roche GS-FLX at the William H. Keck Center for Comparative and Functional Genomics at the University of Illinois.

Cross-match (http://www.phrap.org/), Roche-provided sff tools (http://454.com/products/analysis-software/index.asp) and custom Java scripts were used to identify trim sequencing adaptors within sff file data and remove sequences of <50 nucleotides or with homopolymer stretches ≥60% of the raw read length. Processed sequence data was assembled into contigs using the Roche GS De Novo Assembler v 1.1.03 using default parameters (seed step: 12, seed length: 16, min overlap length: 40, min overlap identity: 90%, alignment identity score: 2, and alignment difference score: 23).

Cloning vector sequence was identified using VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) and masked using Maskseq [32]. Contigs assembled from contaminating *Escherichia coli* DNA were identified by querying against the K-12 reference genome (GenBank accession NC_000913) using the blastn algorithm, and contigs that produced *E*-values $\leq 10^{-15}$ were removed manually. The remaining filtered BAC contigs were annotated using the MAKER 2 genome annotation pipeline [33] using coding sequence evidence from 17,778 *D. virgifera virgifera* ESTs (GenBank dbEST accessions EW761110.1–EW777358.1 [34] and CN497248.1–CN498776.1 [35]), protein homology by blastx searches of the UniProt/Swiss-Prot databases, and *T. castaneum* gene models using the AUGUSTUS web server [36]. Prior to any annotation, RepeatMasker and RepeatRunner were used to identify retroelement-like regions within the BAC full inserts by running against predefined RepBase and RepeatRunner te_proteins provided by MAKER 2 [33]. MAKER 2 output was imported into the Apollo Genome Annotation and Curation Tool [37], where additional annotations were performed via blastx searches of the NCBI nr protein database (*E*-values $\leq 10^{-15}$).

Contigs from clones 142B02 and 156M20 represent partially overlapping homologous sequence and were assembled into a single reference using CAP3 [38] (default parameters). Processed 454 read data from libraries RL003 and RL007 were mapped to this assembled reference using the program LASTZ [39] (default parameters). LASTZ output was made in Sequence Alignment/Map (SAM) format which was used to create an indexing sorted alignment file (.BAI file) with the command line index tool from SAMtools [40]. Mapped read data was visualized using BAMview in the Artemis Genome Viewer [41].

2.4. Comparative Genomics and Annotation of Repetitive Elements. Genomic and EST sequences were separately aligned for cadherin orthologs from *D. virgifera virgifera* (GenBank accessions; mRNA EF531715.1 with DNA EF541349.1) and *T. castaneum* (gene model XM_966295.2 with DNA scaffold NC_007417 positions 19,140,127 to 19,1330,052) using the program Splign with the discontinuous megablast option [42]. Splign tab-delimited output was used to estimate mean intron and exon size. Additionally, a *de novo* prediction of *D. virgifera virgifera* repetitive sequence was made by assembling BAC end sequence (BES) data using CAP3 [38] (default parameters), and subsequently used to query the *D. virgifera virgifera* cadherin genomic DNA sequence (EF541349.1) for putative repetitive sequence using the blastn algorithm. Blastn output was filtered for *E*-values $\geq 10^{-40}$. *De novo* prediction of *D. virgifera virgifera* repetitive sequences were also made within our assembled BAC full inserts (GenBank accessions JQ581035–JQ581043) by querying accession EF541349.1 using identical parameters. BAC insert regions with similarity to the EF541349.1 sequence were excised from BES contigs and BAC full insert sequences (using a custom PERL script), mapped to EF541349.1 using LASTZ [39] (default parameters) and output handled as described previously.

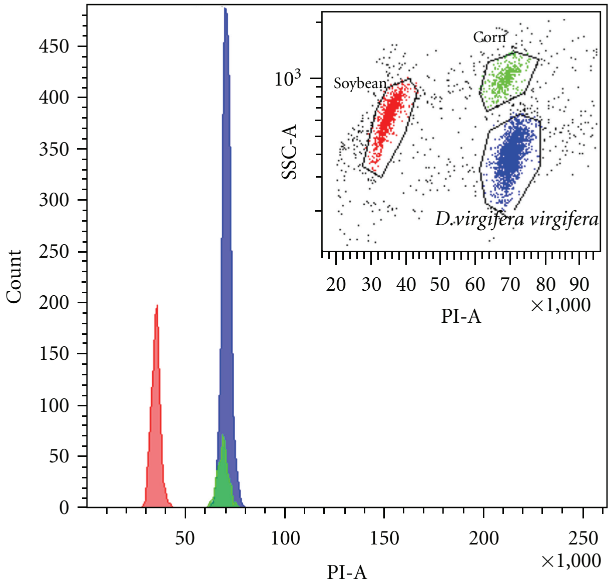
A computational prediction of short repetitive DNA elements known as miniature inverted repeat transposable elements (MITEs) was made for *D. virgifera virgifera* BES contigs and singletons as well as GenBank accession EF541349.1 and *T. castaneum* scaffold NC_007417 positions 19,140,127 to 19,1330,052 using the MITE Uncovering SysTem (MUST; http://csbl1.bmb.uga.edu/ffzhou/MUST/) [43] (default parameters except max DR length = 4 and Min MITE length = 150). The secondary structures of putative MITEs were confirmed by using the Mfold DNA Server (http://mfold.rna.albany.edu/?q=mfold/DNA-Folding-Form) [44] with conditions 25°C and 1.0 mM Mg^{2+}.

3. Results

3.1. Genome Size Estimation. The *D. virgifera virgifera* haploid genome size was estimated at $71,144 \pm 537$ fluorescent units from propidium iodide (PI) stained nuclei, which compared to $69,319 \pm 491$ and $35,631 \pm 687$ units for the internal standards of known genome size, *Zea mays* (2.50 Gbp) and *Glycine max* (1.115 Gbp), respectively. Populations of nuclei from *Z. mays* and *D. virgifera virgifera* produced overlapping PI signals on a flow cytometer, but the size scatter component (SSC-A) indicative of nucleosome densities was used to separate the signals of independent PI readings (Figure 1). Subsequent calculations of PI to genome size ratios indicate an estimated *D. virgifera virgifera* haploid genome size of ∼2.58 Gb or 2.80 pg.

3.2. BAC Library Construction, End Sequencing, and Annotation. Blue-white screening indicated the ligation efficiency with the pCC1 vector was ∼99.25% and arraying of clones onto 384-well plates with ∼99.75% of the wells being successfully filled (Amplicon Express, personal comm.). From these data, ∼109,486 genomic clones were estimated within the 288 × 384 well plates of DvvBAC1. Insert DNA was isolated, digested with *Not*I, and separated by CHEF electrophoresis from 93 of 96 DvvBAC1 clones (96.9%), which indicated a mean pCC1 insert size of 104.4 ± 34.5 kb (Figure 2; not all data shown). From these data, we estimate that $11,496 \pm 3,758$ Mbp are within DvvBAC1, translating to ∼4.56 ± 1.49-fold genome coverage (1.49- and 0.97-fold genome coverage at 95 and 99% probability thresholds, resp.).

Paired end sequencing of 1152 DvvBAC1 clones generated 2304 raw reads, of which 2156 produced high quality sequence data (PHRED scores ≥20; NCBI dbGSS; accessions JM104642–JM106797). Paired BAC end sequence (BES) data was obtained from 1037 of the 1152 clones (90.0%). Filtering for reads >100 bp resulted in 1999 sequences averaging 579.0 ± 141.1 bp (1.17 Mb total; ~0.05% of the 2.58 Gb *D. virgifera virgifera* genome). Functional annotations were obtained for 599 of 1999 filtered BES reads (30.0%) using blastx results, of which 167 sequences received 620 gene ontology (GO) annotations (3.75 ± 1.83 GO annotations per annotated sequence; see Table S1 in Supplementary Material available online at doi:10.1155/2012/604076 which provides a list of putative genes, biochemical functions, and pathway assignments). At level 2, the distribution of GO terms among biological process (P), cellular component (C), and molecular function (F), respectively, showed cellular process, cellular component, and binding activity as most prominent (Figure S1). A total of 447 unique InterPro annotations were made (Table S1; 12 most frequent are listed in Table 1). Predicted functional gene annotations within catalytic activities at GO level F were corroborated by 154 reverse transcriptase (IPR015706 and IPR000477), 12 endonuclease/exonuclease (IPR005135), 17 ribonuclease (IPR012337 and IPR002156), and 5 integrase (IPR017853) annotations in the InterProScan output. An analogous blastn search indicated that 210 sequences (14.3%) showed ≥68.0% similarity to the complete *D. virgifera virgifera* cadherin gene (GenBank Accession EF531715.1; *E*-values $\leq 1.31 \times 10^{-11}$), and 23 (1.6%) showed ≥69.0% similarity to the *D. virgifera virgifera*, D. barberi, and D. virgifera zeae microsatellite sequences (*E*-values $\leq 1.14 \times 10^{-11}$; Figure S2). In addition, a total of 45 annotations of DvvBAC1 BES reads (3.1% of total) indicate an origin from the proteobacterial endosymbiont, *Wolbachia* (Figure S2; Table S1).



Population	Number of events	Percentage of parent	PI-A mean	Genome (Gb)
🟥 Soybean	978	13.8	35,631	1.115
🟩 Corn	397	5.6	69,319	2.5
🟦 *D.virgifera virgifera*	2,453	34.6	71,144	2.58

Figure 1: Flow cytometry estimate of the *D. virgifera virgifera* genome size compared to internal standards of *Zea mays* (inbred line B73; 2.500 Gb) and *Glycine max* (isoline Williams 83; 1.115 Gb).

3.3. BAC Screening and Full Insert Sequencing. Screening of DvvBAC1 identified clones containing sequence from eight EST markers (5.29 ± 2.98 hits; range of 1 to 9 hits per marker; data not shown). Eight of the 9 BAC inserts (88.9%) were successfully sequenced on the Roche GS-FLX. After raw data filtering, a total of 240,586 reads were assembled into 39 contigs that contained 642.0 kb of sequence (16.5 ± 18.9 kb per contig; Table 2). The annotation of BAC inserts using MAKER 2 predicted 37 putative genes and 48 retrotransposon-like protein coding intervals with 3 and 31 of these sequences supported by EST evidence, respectively.

Contigs from clones 142B02 and 156M20 represent homologous genomic regions within different clones and provide a measure of haplotype variation within the library. Sequences from these two clones shared 11 endogenous and 5 retroelement-like protein coding sequences, which represent homologous genome intervals from unique BAC inserts. Six contigs totaling 31.9 kb were aligned (Figure 3), and haplotype variation between inserts was shown via 3 SNPs within the 22.5 kb of CAP3 aligned sequence (SNP frequency $\sim 1.3 \times 10^{-4}$), protein coding sequences were 100% conserved, and no indels were present. Compared to the consensus, 2564 and 5467 bp regions were not represented within clones 142B02 and 156M20, respectively, and was verified by mapping reads to the CAP3 scaffolds (Figure 3). *Hind*III restriction site mapping showed that cut sites used in cloning may not have been the cause of sequence disparity. Additionally, the entire pCC1 cloning vector sequence was sequenced and masked from both clone 142B02 and 156M20 assemblies, indicating that insert boundaries did not give rise to the two gaps. Retroelement-like sequences were annotated within the two haplotype sequence gaps. These results also suggest that structural variation based on the integration/excision or random deletion of repetitive DNA elements may exist among *D. virgifera virgifera* haplotypes.

3.4. Comparative Genomics and Annotation of Repetitive Elements. Comparison of the cadherin gene intron and exon structure from the 94.6 kb *D. virgifera virgifera* and 7.1 kb *T. castaneum* orthologs indicated that the ~13.3-fold increase in the former is accounted for by intron sequence. Specifically, the *T. castaneum* cadherin has a mean intron size of 0.085 ± 0.189 kb, compared to 2.9 ± 1.5 kb in the *D. virgifera virgifera* cadherin, whereas respective total exon sizes of 4.9 kb (mean 180 ± 72.8) and 5.4 kb (mean 173 ± 71.2) were similar between species (Tables S2 and S3). The *de novo* prediction of repetitive elements by alignment of *D. virgifera virgifera* BES data and BAC full insert sequences resulted in 226 contigs and 1089 singletons (mean length of 761.0 ± 236.3 bp; maximum 2002 bp, mean depth = 3.3 ± 4.3 reads). Mapping *de novo* repetitive genome regions (150 from BES

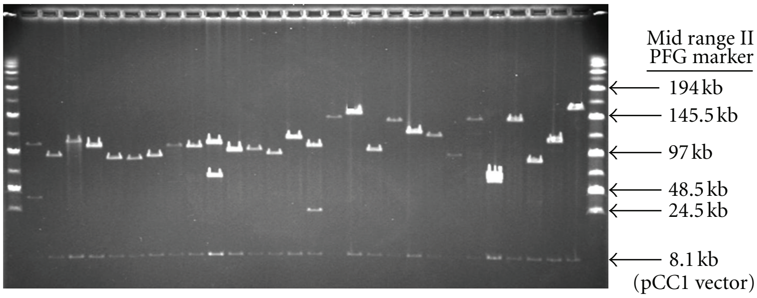


Figure 2: Estimated BAC genomic insert sizes using contour-clamped homogeneous electric field (CHEF) electrophoresis. DNA preparations were digested with *Not*I prior to separation on a 0.9% agarose gel in 0.5X TAE buffer for 24 h at 4°C.

Table 1: The number of InterPro accessions obtained during annotation of *D. v. virgifera* BAC end sequences.

InterPro entry	Number	InterPro functional description(s)
IPR015706	87	RNA-directed DNA polymerase (reverse transcriptase)
IPR000477	67	Reverse transcriptase
IPR009072	18	Histone-fold
IPR012337	17	Ribonuclease H-like
IPR007125	16	Histone core
IPR000558	14	Histone H2B
IPR005135	12	Endonuclease/exonuclease/phosphatase
IPR011991	8	Winged helix-turn-helix transcription repressor
IPR001878	8	Zinc finger, CCHC-type
IPR005819	5	Histone H5
IPR001584	5	Integrase, catalytic core
IPR002156	5	Ribonuclease H domain
IPR005818	5	Histone H1/H5

data (mean 287.4 ± 131.9 bp) and 146 from BAC full inserts (348.2 ± 201.0 bp)) to the *D. virgifera virgifera* cadherin gene sequence resulted in alignments mostly within introns, where the greatest read depth of 37 and 42 were in introns 2 and 12, respectively (Figure 4). Annotation of *de novo* assembled repetitive sequences indicated that 36 (15.9%) encoded reverse transcriptase, gag-pol, or other retrovirus-associated proteins. Histone-like proteins were encoded by contig 87 (histone H1), contig 110 and 149 (histone H2a), contigs 11 and 173 (histone H2b), and contig 214 (histone H3; remaining data not shown).

Predictions of transposable elements by MUST indicated 88 putative MITE-like sequences with direct repeats (DRs) of 2 nucleotides were located within the *D. virgifera virgifera* cadherin gene (Table S4), where 22, 18, and 11 of the DRs involved AT/TA, AA, and TT dinucleotides. Putative MITEs that occupy a total of 2.4 kb (mean = 278.9 ± 124.5 bp) are composed of 65.7 ± 0.1% A or T nucleotides and have predicted terminal inverted repeat (TIR) lengths of 11.9 ± 5.5 bp. Positions of MITE-like inserts were predicted to be within intron regions (Figure 4). Comparatively, the *T. castaneum* cadherin gene contained 12 putative MITE-like elements that were all predicted within intron regions (Table S5).

4. Discussion

4.1. Genome Size Estimation. The haploid *D. virgifera virgifera* genome size of 2.58 Gb (2.80 pg) is one of the largest estimated among beetle species ([45]; mean 0.891 ± 0.795 pg), which range from ∼0.15 for *Oryzaephilus surinamensis* (Coleoptera: Silvanidae) [46] to 3.40 Gb for *Chrysolina carnifex* (Coleoptera: Chysromelidae) [45]. Genome size heterogeneity among beetle species does not appear to be correlated with organism "complexity" (*C*-value paradox) [47], specialization [48], or increased gene content [49]. The relation between repetitive DNA content and genome size in Coleoptera is only available for the model species *Tribolium castaneum* (Coleoptera: Tenebrionidae), where the ∼0.200 Gb genome has an estimated 5110 repetitive elements [30] which comprise ∼13 of the 0.160 Gb assembled sequence [49]. In contrast, our data suggest that the proportion of the *D. virgifera virgifera* genome consisting of repetitive DNA is much higher.

4.2. BAC Library Construction, End Sequencing, and Annotation. BAC libraries are genomic tools that are useful for the isolation of genes linked to a trait [50] as well as the generation of end sequences that provide estimates of genome structure and TE densities [2, 51, 52]. Despite their utility in genomics research, only one coleopteran BAC library has previously been reported, for *T. castaneum* [53]. The prediction of gene-coding regions from BAC end sequence from nonmodel species rely on functional annotation by homology-based identification with related genes in model organisms. This can result in vague or inaccurate gene definitions for nonmodel species [54], such as our *D. virgifera virgifera* dbGSS dataset. Despite the relative dearth in gene discovery by *D. virgifera virgifera* BES, 179 novel protein coding regions were annotated which will provide a resource for annotation of future WGS efforts. Studies with similarly low genome sequence coverage have been useful for initial descriptions of functional and repetitive elements [55].

Table 2: Summary of contigs per BAC that were assembled from Roche-454 sequencing data.

Mid-tag library	BAC clone	Marker	Raw data (reads/kb)	Assembled data (reads/kb)	GenBank accession	Contig size (kb)
RL001	40F02	1304	23,300/10,298	21,525/97.8	JQ581035	19.6 ± 22.2
RL002	89B10	1224	8,701/7,408	7,277/118.7	JQ581036	29.7 ± 20.6
RL003	142B02	1203	29,444/12,894	19,447/30.5	JQ581037	4.3 ± 3.6
RL005	191G22	1125	64,410/28,909	8,348/104.0	JQ581038	17.3 ± 18.8
RL006	163F14	1304	9,495/3,523	22,530/101.7	JQ581039	25.4 ± 7.2
RL007	156M20	1203	16,427/7,179	9,435/24.6	JQ581040	4.1 ± 2.8
RL008	222P02	1345	43,702/20,431	0/0.0	FAILED	NA
RL009	213A05	1411	25,196/11,345	18,868/74.5	JQ581041	74.5 ± 0.0
RL010	188M01	1300	19,912/9,238	18,478/90.6	JQ581042	15.1 ± 9.7
		Total	240,587/111,225	125,908 (15,738 ± 6286)/642.3 (80.3 ± 34.9)		13.6 ± 20.1

NA: not applicable due to DNA sequencing failure.

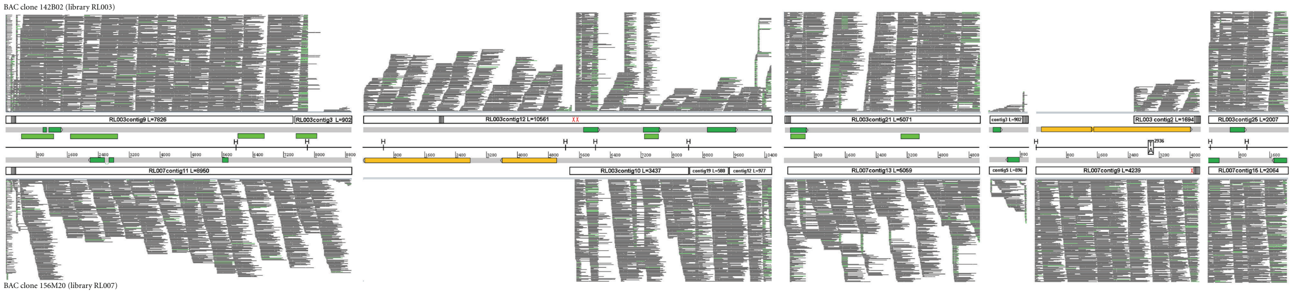


Figure 3: Comparison of haplotypes between assembled full BAC insert sequences of clones 142B02 and 156 M20. Homologous regions are aligned and representative read depths are indicated above for 142B02 and below for 156M20. Annotated genes (dark green), expressed sequence tag (EST; light green), and repetitive element sequences (orange) are indicated. Microsatellite repeat motifs are shown as (||||).

Proteins encoded by DNA-based TEs and retrotransposons totaled ~16.0% of BES reads and outnumbered endogenous genes by ~1.8-fold. Extrapolation suggests that retroelement-like TE genes might occupy ~0.41 Gb of the 2.58 Gb genome. Compared to *T. castaneum* which has ~3.7% of the genome assembly occupied by LTR- and non-LTR-retrotransposons [30], the *D. virgifera virgifera* genome may have an ~4.3-fold higher retroelement content. Our investigations also indicate that small nonautonomous miniature inverted repeat transposable elements (MITEs) are present within the *D. virgifera virgifera* genome.

4.3. BAC Screening and Full Insert Sequencing. The Roche-454 GS-FLX provides a robust method for rapid sequence generation, from which single end read data were sufficient to assemble 8 of the 9 BAC plasmids we sequenced. Assembly of *D. virgifera virgifera* BAC inserts into an average of ~5 contigs per clone and encompassing 80.3 kb of total sequence was greater than that obtained following assembly of BACs from barley [23]. Annotations indicated that the number of TE-derived genes in assembled contigs were 1.3-fold higher than endogenous protein coding genes. This result differs from our estimate from BES data but may be influenced by sample number or by the effect of large TE-derived gene sizes on the probability of sampling from BES data. Regardless, full BAC insert sequences indicate that the *D. virgifera virgifera* genome is comprised of a high proportion of TE-derived sequence but also suggests that DNA-based and retroelement-like TEs are localized within intergenic space. This preliminary genome sequencing evidence suggests that genic regions of the *D. virgifera virgifera* genome can be assembled from short single-end NGS read data, but the use of longer read lengths and paired-end or mate-pair NGS strategies may result in increased contig size and/or scaffolding by the spanning of repetitive elements.

Comparison of the homologous regions within contigs from clones 142B02 and 156M20 provided a direct measure of haplotype variation within DvvBAC and also within the *D. virgifera virgifera* nondiapause strain. SNP variation between haplotypes was low, which may be the result of a genetic bottleneck and subsequent inbreeding within the colony. These results are consistent with a microsatellite marker-based estimate of 15–39% allele diversity reduction in the nondiapause colony compared to wild populations [56]. Comparison of *D. virgifera virgifera* haplotypes suggested that local genome variation based upon insertion/deletion of large DNA regions may occur. Evidence suggests that these variations are not likely due to differences in read depth or effects of cloning due to variation in *Hind*III restriction sites. Interestingly, retroelement-like sequences were annotated within regions of haplotype variation and may indicate that microsynteny is altered through TE integration. Analogous

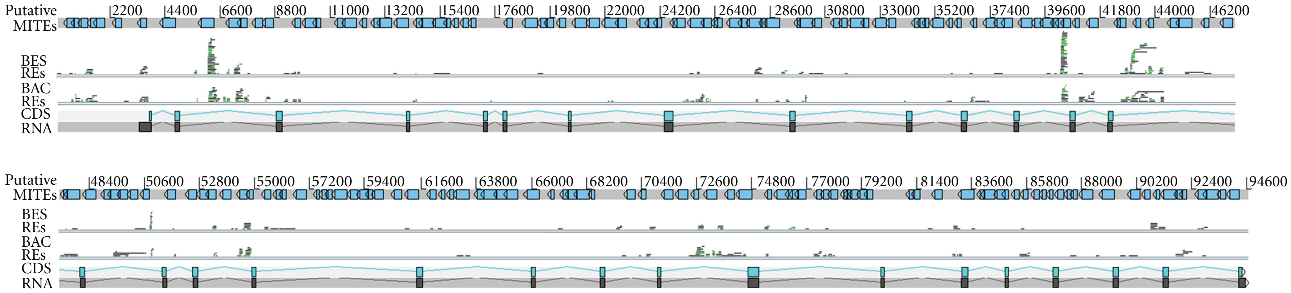


Figure 4: Identification of putative miniature inverted repeat transposable elements (MITEs) (blue rectangles indicating direction), and *de novo* mapped repetitive elements identified from BAC end sequences (BES REs) and BAC full insert sequences (BAC REs), within the gene protein coding sequence (CDS) and transcript sequence (RNA).

haplotype variation was caused by movement of *Helitron*-like TEs in maize and SINEs in canine genomes. Our results similarly suggest that retroelement movement may be a source of haplotype variation in the *D. virgifera virgifera* genome but will require further investigation to realize the extent to which movements affect genome structure and function.

4.4. Comparative Genomics and Annotation of Repetitive Elements. Compared to *T. castaneum*, the orthologs of intronless histone encoding genes show no size increase within the *D. virgifera virgifera* genome, although intron-containing genes tend to show a dramatic increased size in *D. virgifera virgifera*. For example, the 94.6 kb *D. virgifera virgifera* cadherin gene is ∼13.3-fold larger than the *T. castaneum* ortholog despite the coding regions being approximately the same length. Mapping of BES reads and computational prediction of MITE-like elements within the *D. virgifera virgifera* cadherin gene indicated that TEs and other repetitive elements have inserted within intron regions and are the cause of the comparative increase in gene size. TE integrations within introns are known to affect splicing efficiencies [57], but this remains to be investigated in *D. virgifera virgifera*. As stated previously, the insertion of large retroelements within gene coding regions was not predicted. The insertion of a repetitive DNA in the *D. virgifera virgifera* cadherin 5′-UTR suggests that the movement of TEs within the genome could alter gene expression and regulation. TE integrations are also known to cause chromosomal changes that alter gene expression [58]. The accumulation of these changes across the genome can lead to differential selection among local environments [59] or even contribute to the evolution of new species [60]. Knowledge of TE composition within a genome is a fundamental step in the study of relationships between structure and function that may form a basis for future comparative studies. We defined 296 small repetitive DNA elements and 48 large retroelement-like coding sequences within the *D. virgifera virgifera* genome. Although these elements were defined from only 1.15 Mb of genomic sequence, these predictions represent an initial resource for understanding the proliferation and phenotypic effects of repetitive DNA. The DvvBAC1 library has proven useful for the description of gene and repetitive element densities in the *D. virgifera virgifera* genome and will be a tool for the investigation of the genetic basis of problematic insecticide resistance and behavioral traits expressed by this crop pest species.

Acknowledgments

This research was a joint contribution from the USDA Agricultural Research Service (CRIS Project 3625-22000-017-00D), Iowa Agriculture and Home Economics Experiment Station, Iowa State University, Ames, IA (Project 3543), the University of Nebraska, and the University of Illinois at Champaign-Urbana. This paper reports the results of research only. Mention of a proprietary product does not constitute an endorsement or a recommendation by the parties herein for their use.

References

[1] B. Grisart, W. Coppieters, F. Farnir et al., “Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition,” *Genome Research*, vol. 12, no. 2, pp. 222–231, 2002.

[2] B. S. Coates, D. V. Sumerford, R. L. Hellmich, and L. C. Lewis, “Repetitive genome elements in a European corn borer, *Ostrinia nubilalis*, bacterial artificial chromosome library were indicated by bacterial artificial chromosome end sequencing and development of sequence tag site markers: implications for lepidopteran genomic research,” *Genome*, vol. 52, no. 1, pp. 57–67, 2009.

[3] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein, “454 sequencing put to the test using the complex genome of barley,” *BMC Genomics*, vol. 7, article 275, 2006.

[4] K. Osoegawa, A. G. Mammoser, C. Wu et al., “A bacterial artificial chromosome library for sequencing the complete human genome,” *Genome Research*, vol. 11, no. 3, pp. 483–496, 2001.

[5] J. M. Kelley, C. E. Field, M. B. Craven et al., “High throughput direct end sequencing of BAC clones,” *Nucleic Acids Research*, vol. 27, no. 6, pp. 1539–1546, 1999.

[6] J. L. Krysan and T. F. Branson, “Biology, ecology and distribution of *Diabrotica*,” in *Proceedings of the International Maize Virus Disease Colloquium and Workshop*, O. H. Wooster, D. T. Gordon, J. K. Knoke, L. R. Nault, and R. M. Ritter, Eds., pp. 144–150, August 1982.

[7] M. E. Gray, T. W. Sappington, N. J. Miller, J. Moeser, and M. O. Bohn, "Adaptation and invasiveness of western corn rootworm: intensifying research on a worsening pest," *Annual Review of Entomology*, vol. 54, pp. 303–321, 2009.

[8] L. J. Meinke, T. W. Sappington, D. W. Onstad et al., "Western corn rootworm (*Diabrotica virgifera virgifera* LeConte) population dynamics," *Agricultural and Forest Entomology*, vol. 11, no. 1, pp. 29–46, 2009.

[9] F. Baca, "New member of the harmful entomofauna of Yugoslavia *Diabrotica virgifera virgifera* LeConte (Coleoptera: Chrysomelidae)," *Zaštita Bilja*, vol. 45, no. 2, pp. 125–131, 1994.

[10] N. Miller, A. Estoup, S. Toepfer et al., "Multiple transatlantic introductions of the western corn rootworm," *Science*, vol. 310, no. 5750, p. 992, 2005.

[11] M. Ciosi, N. J. Miller, K. S. Kim, R. Giordano, A. Estoup, and T. Guillemaud, "Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various reductions of genetic diversity," *Molecular Ecology*, vol. 17, no. 16, pp. 3614–3627, 2008.

[12] H. C. Chiang, "Bionomics of the northern and western corn rootworms," *Annual Review of Entomology*, vol. 18, pp. 47–72, 1973.

[13] A. L. Kahler, A. E. Olness, O. R. Sutter, C. D. Dybing, and O. J. Devine, "Root damage by western corn rootworm and nutrient content in maize," *Agronomy Journal*, vol. 77, pp. 769–774, 1985.

[14] S. G. Jensen, "Laboratory transmission of maize chlorotic mottle virus by three species of corn rootworms," *Plant Disease*, vol. 69, no. 10, pp. 864–868, 1985.

[15] R. L. Gilbertson, W. M. Brown, E. G. Ruppel, and J. L. Capinera, "Association of corn stalk rot *Fusarium* spp. and western com rootworm beetles in Colorado," *Phytopathology*, vol. 76, no. 12, pp. 1309–1314, 1986.

[16] S. A. Lefko, T. M. Nowatzki, S. D. Thompson et al., "Characterizing laboratory colonies of western corn rootworm (Coleoptera: Chrysomelidae) selected for survival on maize containing event DAS-59122-7," *Journal of Applied Entomology*, vol. 132, no. 3, pp. 189–204, 2008.

[17] L. N. Meihls, M. L. Higdon, B. D. Siegfried et al., "Increased survival of western corn rootworm on transgenic corn within three generations of on-plant greenhouse selection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 49, pp. 19177–19182, 2008.

[18] K. J. Oswald, B. W. French, C. Nielson, and M. Bagley, "Selection for Cry3Bb1 resistance in a genetically diverse population of nondiapausing western corn rootworm (Coleoptera: Chrysomelidae)," *Journal of Economic Entomology*, vol. 104, no. 3, pp. 1038–1044, 2011.

[19] A. J. Gassmann, J. L. Petzold-Maxwell, R. S. Keweshan, and M. W. Dunbar, "Field-evolved resistance to Bt maize by Western corn rootworm," *PLoS ONE*, vol. 6, no. 7, Article ID e22629, 2011.

[20] J. L. Krysan, D. E. Foster, T. F. Branson, K. R. Ostlie, and W. S. Cranshaw, "Two years before the hatch: rootworms adapt to crop rotation," *Bulletin of the Entomological Society of America*, vol. 32, no. 4, pp. 250–253, 1986.

[21] J. T. Shaw, J. H. Paullus, and W. H. Luckmann, "Corn rootworm oviposition in soybeans," *Journal of Economic Entomology*, vol. 71, no. 2, pp. 189–191, 1978.

[22] A. E. Sammons, C. R. Edwards, L. W. Bledsoe, P. J. Boeve, and J. J. Stuart, "Behavioral and feeding assays reveal a western corn rootworm (Coleoptera: Chrysomelidae) variant that is attracted to soybean," *Environmental Entomology*, vol. 26, no. 6, pp. 1336–1342, 1997.

[23] M. E. O'Neal, C. D. DiFonzo, and D. A. Landis, "Western corn rootworm (Coleoptera: Chrysomelidae) feeding on corn and soybean leaves affected by corn phenology," *Environmental Entomology*, vol. 31, no. 2, pp. 285–292, 2002.

[24] T. W. Sappington, B. D. Siegfried, and T. Guillemaud, "Coordinated *Diabrotica* genetics research: accelerating progress on an urgent insect pest problem," *American Entomologist*, vol. 52, no. 2, pp. 90–97, 2006.

[25] N. J. Miller, S. Richards, and T. W. Sappington, "The prospects for sequencing the western corn rootworm genome," *Journal of Applied Entomology*, vol. 134, no. 5, pp. 420–428, 2010.

[26] T. F. Branson, "The selection of a non-diapause strain of *Diabrotica virgifera* (Coleoptera: Chrysomelidae)," *Entomologia Experimentalis et Applicata*, vol. 19, no. 2, pp. 148–154, 1976.

[27] J. Doležel and J. Bartoš, "Plant DNA flow cytometry and estimation of nuclear genome size," *Annals of Botany*, vol. 95, no. 1, pp. 99–110, 2005.

[28] J. Doležel, J. Bartoš, H. Voglmayr, J. Greilhuber, and R. A. Thomas, "Nuclear DNA content and genome size of trout and human," *Cytometry Part A*, vol. 51, no. 2, pp. 127–129, 2003.

[29] L. Clarke and J. Carbon, "A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome," *Cell*, vol. 9, no. 1, pp. 91–99, 1976.

[30] Y. S. Yim, P. Moak, H. Sanchez-Villeda et al., "A BAC pooling strategy combined with PCR-based screenings in a large, highly repetitive genome enables integration of the maize genetic and physical maps," *BMC Genomics*, vol. 8, article 47, 2007.

[31] B. S. Coates, D. V. Sumerford, N. J. Miller et al., "Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis," *Journal of Heredity*, vol. 100, no. 5, pp. 556–564, 2009.

[32] P. Rice, L. Longden, and A. Bleasby, "EMBOSS: the European molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[33] B. L. Cantarel, I. Korf, S. M. C. Robb et al., "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes," *Genome Research*, vol. 18, no. 1, pp. 188–196, 2008.

[34] L. M. Knolhoff, K. K. O. Walden, S. T. Ratcliffe, D. W. Onstad, and H. M. Robertson, "Microarray analysis yields candidate markers for rotation resistance in the western corn rootworm beetle, *Diabrotica virgifera virgifera*," *Evolutionary Applications*, vol. 3, no. 1, pp. 17–27, 2010.

[35] B. D. Siegfried, N. Waterfield, and R. H. Ffrench-Constant, "Expressed sequence tags from *Diabrotica virgifera virgifera* midgut identify a coleopteran cadherin and a diversity of cathepsins," *Insect Molecular Biology*, vol. 14, no. 2, pp. 137–143, 2005.

[36] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern, "AUGUSTUS: a web server for gene finding in eukaryotes," *Nucleic Acids Research*, vol. 32, pp. W309–W312, 2004.

[37] S. E. Lewis, S. M. J. Searle, N. Harris et al., "Apollo: a sequence annotation editor," *Genome Biology*, vol. 3, no. 12, Article ID R0082, 2002.

[38] X. Huang and A. Madan, "CAP3: a DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[39] R. S. Harris, *Improved pairwise alignment of genomic DNA [Ph.D. thesis]*, The Pennsylvania State University, 2007.

[40] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[41] K. Rutherford, J. Parkhill, J. Crook et al., "Artemis: sequence visualization and annotation," *Bioinformatics*, vol. 16, no. 10, pp. 944–945, 2000.

[42] Y. Kapustin, A. Souvorov, T. Tatusova, and D. Lipman, "Splign: algorithms for computing spliced alignments with identification of paralogs," *Biology Direct*, vol. 3, article 20, 2008.

[43] Y. Chen, F. Zhou, G. Li, and Y. Xu, "A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in *Geobacter uraniireducens* Rf4," *Genetics*, vol. 179, no. 4, pp. 2291–2297, 2008.

[44] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.

[45] E. Petitpierre, C. Segarra, and C. Juan, "Genome size and chromosomal evolution in leaf beetles (Coleoptera, Chrysomelidae)," *Hereditas*, vol. 119, no. 1, pp. 1–6, 1993.

[46] K. Sharaf, L. Horová, T. Pavlíček, E. Nevo, and P. Bureš, "Genome size and base composition in *Oryzaephilus surinamensis* (Coleoptera: Sylvanidae) and differences between native (feral) and silo pest populations in Israel," *Journal of Stored Products Research*, vol. 46, no. 1, pp. 34–37, 2010.

[47] C. A. Thomas Jr., "The genetic organization of chromosomes," *Annual Review of Genetics*, vol. 5, pp. 237–256, 1971.

[48] R. Hinegardner, "Evolution of genome size," in *Molecular Evolution*, F. Ayala, Ed., pp. 179–199, Sinauer, Sunderland, Mass, USA, 1976.

[49] Tribolium Genome Sequencing Consortium, "The genome of the model beetle and pest *Tribolium castaneum*," *Nature*, vol. 452, no. 6782, pp. 949–955, 2008.

[50] S. Wang, M. D. Lorenzen, R. W. Beeman, and S. J. Brown, "Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome," *Genome Biology*, vol. 9, no. 3, article R61, 2008.

[51] L. Mao, T. C. Wood, Y. Yu et al., "Rice Transposable elements: a survey of 73,000 sequence-tagged-connectors," *Genome Research*, vol. 10, no. 7, pp. 982–990, 2000.

[52] S. R. Cornman, M. C. Schatz, S. J. Johnston et al., "Genomic survey of the ectoparasitic mite Varroa destructor, a major pest of the honey bee Apis mellifera," *BMC Genomics*, vol. 11, no. 1, article 602, 2010.

[53] M. D. Lorenzen, Z. Doyungan, J. Savard et al., "Genetic linkage maps of the red flour beetle, *Tribolium castaneum*, based on bacterial artificial chromosomes and expressed sequence tags," *Genetics*, vol. 170, no. 2, pp. 741–747, 2005.

[54] Y. Pauchet, P. Wilkinson, H. Vogel et al., "Pyrosequencing the *Manduca sexta* larval midgut transcriptome: messages for digestion, detoxification and defence," *Insect Molecular Biology*, vol. 19, no. 1, pp. 61–75, 2010.

[55] D. A. Rasmussen and M. A. F. Noor, "What can you do with 0.1× genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae)," *BMC Genomics*, vol. 10, article 382, 2009.

[56] K. S. Kim, B. W. French, D. V. Sumerford, and T. W. Sappington, "Genetic diversity in laboratory colonies of western corn rootworm (Coleoptera: Chrysomelidae), including a nondiapause colony," *Environmental Entomology*, vol. 36, no. 3, pp. 637–645, 2007.

[57] M. B. Davis, J. Dietz, D. M. Standiford, and C. P. Emerson, "Transposable element insertions respecify alternative exon splicing in three *Drosophila myosin* heavy chain mutants," *Genetics*, vol. 150, no. 3, pp. 1105–1114, 1998.

[58] M. G. Kidwell, "Transposable elements and the evolution of genome size in eukaryotes," *Genetica*, vol. 115, no. 1, pp. 49–63, 2002.

[59] J. González, T. L. Karasov, P. W. Messer, and D. A. Petrov, "Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*," *PLoS Genetics*, vol. 6, no. 4, 2010.

[60] M. A. F. Noor and A. S. Chang, "Evolutionary Genetics: jumping into a New Species," *Current Biology*, vol. 16, no. 20, pp. R890–R892, 2006.