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

Brad S. Coates  
*United States Department of Agriculture*, brad.coates@ars.usda.gov

Douglas V. Sumerford  
*United States Department of Agriculture*

Richard L. Hellmich  
*United States Department of Agriculture*

Leslie C. Lewis  
*United States Department of Agriculture*

Follow this and additional works at:  
http://lib.dr.iastate.edu/ent_pubs

Part of the *Agronomy and Crop Sciences Commons*, *Entomology Commons*, and the *Plant Breeding and Genetics Commons*

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/ent_pubs/124. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Entomology at Iowa State University Digital Repository. It has been accepted for inclusion in Entomology Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
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

Abstract
The European corn borer, Ostrinia nubilalis, is a serious pest of food, fiber, and biofuel crops in Europe, North America, and Asia and a model system for insect olfaction and speciation. A bacterial artificial chromosome library constructed for O. nubilalis contains 36 864 clones with an estimated average insert size of \( \geq 120 \text{ kb} \) and genome coverage of 8.8-fold. Screening OnB1 clones comprising approximately 2.76 genome equivalents determined the physical position of 24 sequence tag site markers, including markers linked to ecologically important and Bacillus thuringiensis toxin resistance traits. OnB1 bacterial artificial chromosome end sequence reads (GenBank dbGSS accessions ET217010 to ET217273) showed homology to annotated genes or expressed sequence tags and identified repetitive genome elements, O. nubilalis miniature subterminal inverted repeat transposable elements (OnMITE01 and OnMITE02), and ezi-like long interspersed nuclear elements. Mobility of OnMITE01 was demonstrated by the presence or absence in O. nubilalis of introns at two different loci. A \((GTCT)^n\) tetranucleotide repeat at the 5’ ends of OnMITE01 and OnMITE02 are evidence for transposon-mediated movement of lepidopteran microsatellite loci. The number of repetitive elements in lepidopteran genomes will affect genome assembly and marker development. Single-locus sequence tag site markers described here have downstream application for integration within linkage maps and comparative genomic studies.

Keywords
Ostrinia nubilalis, bacterial artificial chromosome library, repetitive elements

Disciplines
Agronomy and Crop Sciences | Entomology | Plant Breeding and Genetics

Comments
This article is from Genome; 52 (2009); 57-67; doi: 10.1139/G08-104

Rights
Works produced by employees of the U.S. Government as part of their official duties are not copyrighted within the U.S. The content of this document is not copyrighted.

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/ent_pubs/124
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

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

**Abstract:** The European corn borer, *Ostrinia nubilalis*, is a serious pest of food, fiber, and biofuel crops in Europe, North America, and Asia and a model system for insect olfaction and speciation. A bacterial artificial chromosome library constructed for *O. nubilalis* contains 36,864 clones with an estimated average insert size of ≥120 kb and genome coverage of 8.8-fold. Screening OnB1 clones comprising approximately 2.76 genome equivalents determined the physical position of 24 sequence tag site markers, including markers linked to ecologically important and *Bacillus thuringiensis* toxin resistance traits. OnB1 bacterial artificial chromosome end sequence reads (GenBank dbGSS accessions ET217010 to ET217273) showed homology to annotated genes or expressed sequence tags and identified repetitive genome elements, *O. nubilalis* miniaturesubterminal inverted repeat transposable elements (OnMITE01 and OnMITE02), and *ezi*-like long interspersed nuclear elements. Mobility of OnMITE01 was demonstrated by the presence or absence in *O. nubilalis* of introns at two different loci. A (GTCT)$_n$ tetranucleotide repeat at the 5’ ends of OnMITE01 and OnMITE02 are evidence for transposon-mediated movement of lepidopteran microsatellite loci. The number of repetitive elements in lepidopteran genomes will affect genome assembly and marker development. Single-locus sequence tag site markers described here have downstream application for integration within linkage maps and comparative genomic studies.

**Key words:** *Ostrinia nubilalis*, bacterial artificial chromosome library, repetitive elements.

**Re**˚sume : La pyrale du maı¨ s, *Ostrinia nubialis*, est un ravageur important des cultures pour des fins alimentaires, textiles et énergétiques en Europe, en Amérique du Nord et en Asie. De plus, il s’agit d’une espèce modèle pour l’olfaction et la spéciation chez les insectes. Une banque de chromosomes bactériens artificiels a été produite pour l’*O. nubilalis* et totalise 36,864 dont la taille moyenne des inserts est d’environ ≥120 kb pour une couverture génomique de 8,8 fois. Un criblage de clones OnB1 conférant une couverture de 2,76 équivalents génomiques a permis de déterminer la position physique de 24 marqueurs de séquence connue dont des marqueurs liés à des caractères écologiques importants ainsi qu’à la résistance à la toxine du *Bacillus thuringiensis*. Les séquences des extrémités de clones chromosomes bactériens artificiels OnB1 (acccessions GenBank dbGSS ET217010 à ET217273) présentaient de l’homologie avec des gènes annotés ou des étiquettes de séquences exprimées et ont permis d’identifier des séquences répétitives : des éléments transposables miniatures à répétitions subterminales inversées (OnMITE01 et OnMITE02) ainsi que des éléments nucléaires dispersés de grande taille) de type *ezi*. La mobilité des éléments OnMITE01 a été démontrée par leur présence ou absence au sein d’introns chez deux locus de l’*O. nubilalis*. Un motif tétranucléotidique répété (GTCT)$_n$ aux extrémités 5’ des éléments OnMITE01 et OnMITE02 fournit la preuve de la mobilité de locus microsatellites chez les insectes via le mouvement de transposons. Le nombre d’éléments répétitifs au sein des génomes des lépidoptères affectera l’assemblage des génomes et le développement de marqueurs. Les marqueurs de séquence connue à locus unique qui sont décrits dans ce travail se prêtent à des utilisations ultérieures en vue de l’intégration dans des cartes génétiques et pour des études comparatives.

**Mots-clé**s : *Ostrinia nubilalis*, banque de chromosomes bactériens artificiels, éléments répétitifs.
Introduction

Many lepidopteran insect species feed on plants that are the source of food, fiber, and biofuel products. Genome information of lepidopterans could lead to novel ways of controlling pest species worldwide (International Lepidopteran Genome Project 2001). Comparative genomics using whole genome sequences (WGSs) or linkage maps will identify conservation among lepidopteran genomes (synteny) or preservation of linear gene order (collinearity; Eckardt 2001; Sankoff and Nadeau 2003). Bombyx mori and Heliconius melpomene share similar gene content on a given chromosome (synteny), which suggested that the B. mori WGS (Mita et al. 2004; Xia et al. 2004) could be used as a reference for detecting lepidopteran genome rearrangements when shared loci are mapped (Heckel et al. 1998; Schmidt 2002; Delseny 2004; Yasukochi et al. 2006; Pringle et al. 2007). Single nucleotide polymorphism (SNP) markers developed from B. mori bacterial artificial chromosome (BAC) end sequence (BES) data were placed into 28 linkage groups and provided a connection between BAC physical positions and linkage information that will be needed for reconstruction of this lepidopteran genome (Yamamoto et al. 2006, 2008). Connections between linkage maps and physical constructs allow isolation of genome regions linked to mapped loci among lepidopterans that lack a WGS and can be accomplished by positioning single-locus markers on one or more BAC clones.

Repetitive elements often complicate the assembly of a WGS from shotgun sequence reads (Arabidopsis Genome Initiative 2000; International Human Genome Sequencing Consortium 2001). Highly repetitive regions comprise approximately 45% of the B. mori genome (Gage 1974), with a majority of units <500 bp in size (Mita et al. 2004). Proliferation of transposable elements (TEs) contributes to the increased size of eukaryotic genomes (San-Miguel and Bennetzen 1998) and affects cellular function when inserted into gene coding regions (Sommer et al. 1998; Gahan et al. 2001). TEs also have inserted within genome regions that are adjacent to lepidopteran microsatellite repeats. Subsequent oligonucleotide primers used in PCR amplification often anneal to multiple loci within a given genome and produce products comprising multiple fragments derived from more than one independently segregating locus (Van’t Hof et al. 2007). Repetitive eukaryotic genomes necessitate the use of large physical constructs such as BAC library clones assembled into a physical map to assist in WGS assembly (Yamamoto et al. 2006, 2008) because short shotgun sequence reads often do not contain a sufficient level of uniqueness to denote a single genome position. BAC libraries and ordered overlapping contigs constructed from individual clones are therefore necessary tools for lepidopteran genome research.

European corn borer, Ostrinia nubilalis, larvae feed upon cultivated crops in North America, Europe, and Asia. Larval O. nubilalis feeding damage is controlled by transgenic crops expressing Bacillus thuringiensis toxins but have developed varying levels and modes of tolerance in laboratory environments. We present a preliminary characterization of an O. nubilalis BAC library, OnB1, by positioning of sequence tag site (STS) markers and end sequencing. The STS markers that we used encompassed genes linked to Ostrinia traits such as pheromone response, diapause duration, and B. thuringiensis toxin resistance traits. Furthermore, we show that OnB1 BES data are a resource for lepidopteran comparative genomics and studies of repetitive genome elements. TEs, long interspersed nuclear elements (LINEs), and miniature subterminal inverted repeat TEs (On-MITE01 and On-MITE02) were identified from OnB1 BES reads. We documented mobility and abundance of O. nubilalis TEs and suggest their role in the propagation of tetranculeotide repeat microsatellites in lepidopteran genomes.

Materials and methods

BAC library construction

Larval O. nubilalis were obtained from the USDA-ARS, Corn Insects and Crop Genetics Research Unit, in Ames, Iowa. The colony is restocked each year using >500 females collected near Ames and consists of biotype Z-pheromone strain ecotypes. Larvae were flash frozen in liquid nitrogen and 100 individuals (approximate 1:1 ratio of males to females) were shipped on dry ice to Amplicon Express, Pullman, Washington. There, genomic DNA was prepared, partially digested with BamHI, and size selected for fragments >100 kb, which were then ligated into the BamHI site of vector pECBAC1. Escherichia coli strain XL1-Blue (Stratagene, La Jolla, California) was transformed and clones selected using the chloramphenical resistance marker. Resistant clones were picked and arrayed onto ninety-six 384-well plates (36 864 total clones) and stored in 10% glycerol at −80 °C.

BAC library pooling strategy

A pooling strategy combined DNA from BAC clones into single samples for PCR screening whereby 16 rows containing 24 clones from each of thirty 384-well plates were combined to form 480 separate row pools (RPs). The 16 RPs from each 384-well plate(s) additionally were pooled to form 30 superpools (SPs). Pooling represented approximately one third of the library, PCR-based screening of 1.38 × 10⁹ bp, or approximately 2.76-fold genome coverage. Specifically, 10 μL of glycerol stock from the 24 clones located in each row was combined to create a 240 μL RP. SPs were generated by combining 15 μL of the 16 RPs in a given 384-well plate such that each 240 μL SP corresponded to one 384-well plate. The RPs and SPs were stored in 10% glycerol at −80 °C. This pooling strategy was done for thirty 384-well plates comprising 11 520 clones or approximately 1.52 × 10⁹ bp. The RP and SP glycerol stocks were streaked onto LB agar containing 30 μg/L chloramphenical and grown overnight at 37 °C. RP and SP clones were used to inoculate 800 μL of LB broth containing 30 μg/L chloramphenical and grown overnight at 37 °C. BAC DNA was isolated by alkaline lysis with RNase treatment omitted. BAC DNA was suspended in 200 μL of 0.5× TE buffer, pH. 7.5, and stored at −20 °C prior to use.
Marker development and OnB1 PCR screening

PCR screening of SPs and RPs was used as a method to position genes of interest (STSS) on OnB1 clones and to identify homologous genome regions (since PCR markers were single locus). Single-locus STS markers were shown by pedigree analysis for *O. nubilalis* cadherin (Coates et al. 2005a), ommochrom binding proteins (obp1 and obp2) (Coates et al. 2005b), brainiac/brs5 (Coates et al. 2007), and aminopeptidase 1 (APN1) (Coates et al. 2008). Additional single focus markers were from *OnZ* (Coates and Hellmich 2003) and kettin (ket) (Dopman et al. 2005; Malausa et al. 2007). We designed a total of 19 degenerate oligonucleotide primers for PCR amplification of ribosomal loci from Lepidoptera of which 11 were successful in PCRs with *O. nubilalis* genomic DNA as template. Oligonucleotide primers were also designed from OnB1 BES data. Four markers were designed to amplify regions of OnB1 contigs OnC0003, On-0004, On-0006, and On-0007 that contained putative SNP loci (Table 1).

For each STS marker, PCR amplification of SP and appropriate RP BAC template took place in 10 μL reaction volumes containing 2.5 mmol/L MgCl₂, 50 μmol/L dNTPs, 5 ng of BAC DNA, 1.8 pmol of each primer, 2 μL of 5× thermal polymerase buffer (Promega, Madison, Wisconsin), and 0.3125U of GoTaq DNA polymerase (Promega). A BioRad Tetrade 2 thermocycler used 96 °C for 3 min and then seven initial touchdown (TD) cycles consisting of 96 °C for 20 s, 65 °C for 30 s (–2 °C per cycle for each subsequent cycle), and 72 °C for 1 min. Final PCR amplification was 35 cycles of 96 °C for 20 s, 50 °C for 30 s, and 72 °C for 1 min. PCR products for *O. nubilalis* ZOn1 and ZOnW1, cadherin, obp1, obp2, and bre5 took place according to Coates and Hellmich (2003) and Coates et al. (2005a, 2005b), respectively. Entire PCR product volumes were separated on 1.5% agarose gels and positive BAC SPs or RPs was identified by presence-absence of the gene fragment compared with positive control DNA. Once RPs were identified, the 24 clones from the respective RPs were streaked onto LB agar, cultured overnight, and DNA isolated by alkaline lysis as described previously.

### Table 1. Eight contigs constructed from 17 *Ostrinia nubilalis* BAC end sequence (BES) reads.

<table>
<thead>
<tr>
<th>Contig name</th>
<th>OnB1 BES</th>
<th>Read length (bp)</th>
<th>Contig overlap (bp)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-C0001</td>
<td>23A08-F, 23P08-F</td>
<td>365</td>
<td>217</td>
<td>None</td>
</tr>
<tr>
<td>On-C0002</td>
<td>23G02-F, 24G01-F</td>
<td>342</td>
<td>285</td>
<td>None</td>
</tr>
<tr>
<td>On-C0003</td>
<td>04B19-F, 04F23-F</td>
<td>435</td>
<td>331</td>
<td>186 A/T</td>
</tr>
<tr>
<td>On-C0004</td>
<td>04A19-F, 04P18-F</td>
<td>348</td>
<td>286</td>
<td>174 A/G</td>
</tr>
<tr>
<td>On-C0005</td>
<td>23L08-F, 23L09-F, 23E08-F</td>
<td>570</td>
<td>396</td>
<td>178 T/G</td>
</tr>
<tr>
<td>On-C0006</td>
<td>23B08-F, 23O08-F</td>
<td>400</td>
<td>319</td>
<td>None</td>
</tr>
<tr>
<td>On-C0007</td>
<td>23L07-F, 23L09-F</td>
<td>410</td>
<td>339</td>
<td>241 A/C</td>
</tr>
<tr>
<td>On-C0008</td>
<td>04D23-F, 04A23-F</td>
<td>469</td>
<td>424</td>
<td>362 A/C</td>
</tr>
</tbody>
</table>

**Note:** The read length for each clone and length of overlap of assembled contigs are given in basepair (bp). Putative single nucleotide polymorphism (SNP) loci are indicated by location in the contig followed by alternate nucleotides present. The criteria for contig assembly are ≥95% homology over ≥200 nucleotides and alignment gap penalty = 10.

BAC DNA isolation and sequencing

OnB1 clones were streaked from glycerol stocks onto LB agar containing 30 μg/L chloramphenicol and grown overnight at 37 °C. Individual clones were used to start seed cultures of 800 μL of LB broth containing 30 μg/L chloramphenicol and grown overnight at 37 °C. Larger volume 30 mL LB chloramphenicol (30 μg/L) cultures were inoculated with individual seed cultures and grown overnight at 37 °C with shaking at 200 rpm. Cells were harvested by centrifugation at 5000 rpm for 1 min. BAC DNA was purified using the BAC and Low Copy miniprep kits (GerardBIOTECH, Oxford, Ohio) and final pellets suspended in 43 μL of nuclease-free water, 5 μL of 10× buffer D, and 1 μL of XbaI and 1 μL SalI restriction enzymes (10 units/μL) (Promega). Digests were incubated overnight at 37 °C. DNA was precipitated in a 5% polyethylene glycol and 0.5 mol/L NaCl solution at 14,000 rpm for 20 min at 4 °C in a centrifuge. Pelleted DNA was washed twice with 0.5 mL of cold 70% ethanol and then dried and suspended in 12 μL of deionized water. Samples were quantified by UV absorbance at 260 nm and diluted to a final concentration of 0.1 μg/μL.

Template was denatured by incubation at 96 °C for 5 min followed by snap chilling on ice for 5 min. End sequencing of individual BACs was performed with 10 μL (1.0 μg) of XbaI- and SalI-digested BAC DNA template, 2 μL of T7-T4BAC primer (5'-ATACGACTCATATAGGGCG-3', 30 pmol/μL), and 8 μL of Dye-Terminator Cycle Sequencing mastermix (Beckman-Coulter, Fullerton, California). Primer extension reactions were carried out on an Eppendorf Mastercycler (Hamburg, Germany) with an initial denaturation at 96 °C for 2 min followed by 80 cycles of 96 °C for 30 s, 54 °C for 30 s, and 60 °C for 4 min. Primer extension products were purified by ethanol precipitation, suspended in 25 μL of deionized formamide, and separated on a CEQ 8000 genetic analysis system (Beckman-Coulter) with
method LFR-a (denature: 90 °C for 120 s, inject: 2.0 kV for 15 s, and separated: 4.2 kV for 120 min in a 50 °C capillary).

Partial DNA sequence analysis of OnB1 BACs that were identified by PCR screening was performed by the use of the separate primer extension reactions with forward and reverse primers (supplementary material S1). Individual BACs were grown in LB chloramphenical (30 μg/mL) cultures and DNA isolated as described above. DNA sequencing was performed with 10 μL (1.0 μg/mL) of XbaI- and SalI-digested BAC DNA template, 2 μL (30 pmol/μL) of the corresponding forward or reverse primer, and 8 μL of Dye-Terminator Cycle Sequence kit mastermix (Beckman-Coulter). Cycle sequencing primer extension reactions and separation on a CEQ8000 system were carried out as previously described.

DNA sequence quality trimming and annotation

Raw sequence data were analyzed and quality of sequence assessed using the PHRED quality parameter. Probability of correct base calls during sequence analysis (PHRED quality parameter \( q = -10 \times \log_{10}(p) \) where \( p \) is the estimated error probability of a base call) was calculated with the Sequence Analysis software (version 8.0) of the Beckman-Coulter CEQ8000 genetic analysis system. Sequences were trimmed when \( q < 20 \) (90% base call accuracy) and internal discrepancies were inspected and corrected visually. Vector sequence was automatically trimmed from FASTA-formatted data when exported by the CEQ8000 Sequence Analysis software.

ContigExpress software (Informax, San Francisco, California) was used to assemble contiguous and overlapping sequences among OnB1 BESs. The OnB1 BES contigs and singletons were exported in FASTA format and used as queries molecular database searches. Homology searches were performed for each OnB1 BES against GenBank nt/nt accessions using BLASTn and tBLASTx (current December 2007). Similarity of OnB1 BES reads to expressed sequence libraries marker (Glover et al. 1992). BACs carrying Z-chromosome-linked markers and construction of BAC contigs in proximity to QTL (Coates and Hellmich 2003) were identified using BLASTn searches of GenBank dbEST and dbGSS databases. Eight OnB1 contigs and 246 singletons were annotated by using BLASTn of GenBank nucleotide (nt/nt) and EST databases (dbEST). Criteria for positive matches were ≥80% homology over ≥130 nucleotides with \( E \) values ≥e-10.

Phylogenetics

GenBank accessions that showed homology to OnB1 BES reads 04K12-F and 27L07-F and pheromone binding protein (ppb) alleles were downloaded in FASTA format (Fig. 1) (supplementary material S2) and imported into the Molecular Evolutionary Genetics Analysis program (version 3.1) (Kumar et al. 2004). ClustalW alignments were performed using default parameters and corrections made manually. Subsequent maximum likelihood phylogenies were constructed using the genetic distances calculated. Bootstrap branch support for each node was generated using 1000 pseudoreplications. A linearized consensus phylogeny produced with nodes contained within ≥400 of 1000 (40%) random trees that were generated.

Results and discussion

BAC library construction

An *O. nubilalis* BAC library, OnB1, was constructed from an approximately equal proportion of Z-pheromone strain males and females that showed two mating generations per year in field conditions in Ames, Iowa (bivoltine ecotype). The haploid (1N) genome size of *O. nubilalis* is unknown, but an assumption of 500 Mb was made based on the size of the *Galleria mellonella* (Rasch 1985) and *B. mori* genomes (Gage 1974). OnB1 consists of genomic DNA inserted into the BamHI site of pEBCAC1 with an average size of 120 kb (Amplicon Express communication). A total of 36864 OnB1 clones were arrayed on ninety-six 384-well plates. Sequencing data presented later indicated that 1.85% of clones carry no genomic insert, which suggested that OnB1 contains approximately 4432 Mb, or approximately 8.8-fold, 1N genome coverage. Assuming random probability of insertion for each genomic fragment into the pEBCAC1 vector, the 95% and 99% probabilities of recovering a genomic fragment (locus) suggest a minimum of 2.95- and 1.92-fold genome coverage, respectively.

Screening STS markers

PCR screening of BAC library pools is an effective method for isolation of single loci from represented genomes (Bruno et al. 1995; Yim et al. 2007) and was used to test OnB1 representation of the *O. nubilalis* genome. PCR-based screening of 1.38 \( \times 10^9 \) bp, or approximately 2.76-fold genome coverage, isolated one or more BAC clones for 24 of 25 *O. nubilalis* STS markers (mean = 1.64 ± 0.70). Some of the genes involved in ecologically important traits were also located on OnB1 clones. The quantitative trait locus (QTL) for male pheromone response (*Resp*) is in proximity (28.1 ± 4.1 cM) to a *tpi* gene marker (Glover et al. 1990; Dopman et al. 2004), and postdipause development (*Ppd*) time of larvae is determined by male parents (Showers et al. 1972; Glover et al. 1992). BACs carrying Z-chromosome-linked *tpi*, *ket*, and *ldh* genes and the minisatellites marker *OnZ1* (Coates and Hellmich 2003) were identified (supplementary material S1). Linkage relationships between the markers *tpi*, *ket* and *ldh* and QTL *Resp* and *Ppd* have been determined previously (Dopman et al. 2005). These OnB1 clones carrying Z-chromosome regions likely will assist in the development of additional single-locus Z-chromosome-linked markers and construction of BAC contigs in proximity to QTL *Resp* and *Ppd*. A total of 29 BAC clones were identified that contained 24 of the 25 STS markers (96.3%) that were screened for by PCR. The screen used an approximate one-third coverage of the OnB1 library and was effective in recovery of one to three clones for each STS marker.

---

3 Supplementary data for this article are available on the journal Web site (http://genome.nrc.ca) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 3864. For more information on obtaining material, refer to http://cisti-icist.nrc-cncr.gc.ca/cms/unpub_e.html.
Annotation, contig assembly, and redundancy of BESs

A total of 264 GSSs were obtained from 294 randomly selected OnB1 clones using the T7 sequencing primer (90% success rate). Vector trimming identified five sequences that were vector only (1.87%). The remaining 264 sequences gave a total of 101,486 nucleotide calls for an average read length of 381.5 ± 119.9 nt with PHRED scores of >20 (Ewing et al. 1998). OnB1 sequences are accessible through the GenBank GSS database (dbGSS) (accessions ET217010 to ET217273). BLASTn search results indicated that 10 OnB1 BES reads matched lepidopteran genes or ESTs (supplementary material S2). Eight overlapping or contiguous regions (contigs) were constructed from OnB1 BES reads with two or three sequences per contig (mean = 2.13 ±
of the repetitive elements (data not shown). Approximately 45% of the \( B. \) \( mori \) genome is composed of repetitive DNA (Gage 1974) with an average size of <500 bp and present at 2.5 to 3 kb intervals (Mita et al. 2004). Yamamoto et al. (2006, 2008) used data from 115 968 BES reads to design markers that were amplified from 264 BES reads that had homology to \( O. \) \( nubilalis \) genome region encoding a transposase. The \( 04K12\) \(-\) \( F \) \((ET217030) \) sequence had homology to introns of \( pbp \) and mannose 6-phosphate isomerase (\( mpi \)), chitinase (\( chit \)), and the retrotransposon-like long interspersed nuclear element (LINE) \( ezi\) located upstream of the \( \Delta 11\alpha\) -desaturase gene.

### Table 2. Homology of \( Ostrinia \) \( nubilalis \) BAC end sequence (BES) reads to previously described \( Ostrinia \) genes in the GenBank nr/nt database.

<table>
<thead>
<tr>
<th>Queried sequence</th>
<th>Homologous sequence</th>
<th>BLASTn statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>OnB1 BES Positions</td>
<td>Gene</td>
<td>Accession</td>
</tr>
<tr>
<td>04K12-F</td>
<td>25–274</td>
<td>( pbp)–38</td>
</tr>
<tr>
<td>25–274</td>
<td>( pbp)–28</td>
<td>EF396401</td>
</tr>
<tr>
<td>25–274</td>
<td>( pbp)–6</td>
<td>EF396379</td>
</tr>
<tr>
<td>25–274</td>
<td>( pbp)–25</td>
<td>EF396398</td>
</tr>
<tr>
<td>28–286</td>
<td>( mpi)–22</td>
<td>EF396469</td>
</tr>
<tr>
<td>23L07-F</td>
<td>1–329</td>
<td>( chit)</td>
</tr>
<tr>
<td>23N03-F</td>
<td>43–264</td>
<td>( ezi)–( \Delta 11\alpha )</td>
</tr>
<tr>
<td>43–264</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113399</td>
</tr>
<tr>
<td>43–264</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113400</td>
</tr>
<tr>
<td>23M11-F</td>
<td>24–253</td>
<td>( ezi)–( \Delta 11\alpha )</td>
</tr>
<tr>
<td>24–253</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113399</td>
</tr>
<tr>
<td>24–253</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113400</td>
</tr>
<tr>
<td>24–253</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113401</td>
</tr>
<tr>
<td>24–253</td>
<td>( ezi)–( \Delta 11\alpha )</td>
<td>EF113402</td>
</tr>
</tbody>
</table>

**Note:** The criteria for match are >90% homology over >200 nucleotides with \( E \) values >1e\(^{-10}\). Homologous genes are \( Ostrinia \) pheromone-binding protein (\( pbp \)), mannose phosphate 6-isomerase (\( mpi \)), chitinase (\( chit \)), and the retrotransposon-like long interspersed nuclear element (LINE) \( ezi\) located upstream of the \( \Delta 11\alpha\) -desaturase gene.

Annotation of repetitive genome elements

BLASTn searches of the GenBank nr/nt database indicated that OnB1 sequences 04K12-F, 23L07-F, 23N03-F, and 23M11-F had significant similarity to previously described \( Ostrinia \) genes (Table 2). The 23N03-F (ET217126) and 23M11-F (ET217161) sequence reads both shared similarity to an \( Ostrinia \) genome region encoding a transposase. The 04K12-F (ET217030) sequence had homology to introns of \( pbp \) and mannose 6-phosphate isomerase (\( mpi \)) alleles, as did the 23L07-F (ET217030) to intron 2 of the \( Ostrinia \) \( furnacalis \) chitinase gene. Further investigations, presented below, indicated that the four sequences belonged to two types of repetitive genomic elements: LINEs related to the \( Ostrinia \) \( ezi\)-element (Xue et al. 2007) and mobile MITE-like sequences.

Retrotransposon-like LINEs

The OnB1 BESs 23N03-F and 23M11-F showed homology to an \( O. \) \( nubilalis \) reverse transcriptase (RT) and \( \Delta 11\alpha\) -desaturase gene fusion locus (Xue et al. 2007). The 23N03-F sequence (264 bp) has 94.6% homology to a 222 bp region of the first 21 amino acids of the RT and 157 bp of the upstream 5’-UTR. Similarly, a 230 bp region of the 539 bp 23M11-F OnB1 sequence has approximately 90% homology to the C-terminal end of the RT (amino acids 222–363 of the translated peptide AB045235). RT genes typically are encoded by retrotransposons. The \( B. \) \( mori \) genome is highly repetitive with several active and degraded retrotransposons (Mita et al. 2004; Xia et al. 2004) and might suggest that sequence reads 23N02-F and 23M11-F represent portions of a transposon-like sequence. Xue et al. (2007) indicated BLAST evidence relating the RT upstream of the \( Ostrinia \) \( \Delta 11\alpha\) -desaturase genes to LINEs from \( Caenorhabditis \) \( elegans \) and renamed the \( Ostrinia \) LINE \( ezi\)-LINEs. LINEs are autonomous non-long terminal repeat retrotransposons that typically are >5 kb and encode an RT and an endonuclease required for genome propagation by RNA intermediates (class I TEs). The copy number of the \( B. \) \( mori \) LINE element \( kaikoga \) was shown to be 200 copies based on WGS searches (Xue et al. 2007), and the \( ezi\)-\( \Delta 11\alpha \) LINE has proliferated in the \( Ostrinia \) genome based on desaturase and OnB1 BES data. Screening of 360 OnB1 clones with pri-
mers designed to PCR amplify the ezi-Δ11a elements identified 113 putative *O. nubilalis* LINE-like elements in an approximately 0.184-fold genome equivalent sampling (supplementary material S12). The library screening likely gave an underrepresentation of the ezi copy number owing to sampling error but still suggested that approximately 600 related loci may be present in the *O. nubilalis* genome.

**MITEs**

MITEs are short high copy number mobile DNA regions.
characterized by high A+T content, pronounced secondary structure, lack of an internal protein coding sequence, and the presence of terminal inverted repeats (TIRs) (Wessler et al. 1995). MITEs are found ubiquitously within genomes and have been characterized from fruit fly (Wilderm and Hollarcher 2001) and mosquito (Tu 2000) and within introns of the cytochrome p450 monooxygenase genes from *Helicoverpa zea* (Chen and Li 2007). MITEs may represent nonautonomous TEs depleted of an internal coding sequence. Cut-and-paste movements of MITEs occur by DNA intermediates (class II TEs) and is mediated by trans-acting transposases that are encoded by related autonomous TEs (Dufresne et al. 2007). Quesneville et al. (2006) identified 304 autonomous P-element-like TEs within the *Anopheles gambiae* WGS that may facilitate transposition of 2570 related P-MITEs, which indicated that MITEs have proliferated to a greater degree within genomes compared with related autonomous elements.

Two MITE-like sequences, OnMITE01 and OnMITE02, were predicted from OnB1 BES data. OnMITE01 was found within two BES reads, 04K12-F and 23D08-F, where only the 04K12-F sequence provided a full 236 bp copy. The 328 bp OnMITE02 is present in the BES 23D08-F. Characteristic of other MITEs, OnMITE01 and OnMITE02 have high A+T contents (55.5% and 63.8%, respectively) and were shown to PCR amplify from 65% and 56% of BACs, respectively. OnMITE secondary structure composed of two or three adjacent stem loop structures was indicated by Mfold output (supplementary material S1, S2, and S3) (Fig. 2) (Zuker et al. 1999). Lastly, no internal OnMITE1 or OnMITE02 protein coding sequences >50 amino acids were predicted using BLASTx or tBLASTx searches of GenBank. In contrast with most other MITEs, OnMITE01 and OnMITE02 lack characteristic TIRs (Fig. 2). Omission of TIRs from MITE-like sequences has been previously shown in the *Aedes aegypti microuli* element (Tu and Orphanidis 2001) where transposition was hypothesized to occur from subterminal inverted repeats (SIRs). SIRs may be derived from TIRs that have undergone recombination or incurred mutations or represent intermediates during genesis of novel MITE-like TEs (Tu and Orphanidis 2001). OnMITE01 and OnMITE02 appeared lack TIRs wherein OnMITE01 and OnMITE02, respectively, have the SIR sequences CT TTCC/GGAAAG and TGCCAAA/T TTTGGCA (Fig. 2).

Sequences outside the MITE-like regions of 04K12 and 20D08 showed no similarity and suggested insertion of OnMITE01 into more than one genome region. Mobility of nonautonomous MITEs in the *Ostrinia* genome was inferred by location at multiple loci and presence–absence at a single locus. Searches of the GenBank nucleotide collection (nt/nr) indicated homology between OnMITE01 and *O. nubilalis* *pbp* alleles (-6, -25, -28, and -38 from accessions EF396379, EF396398, EF396401, and EF396411, respectively; ID > 94%, E value ≤ 7 × 10^{-11}, and mpi (allele -22 from accession EF396469; ID = 97%; E value ≤ 2 × 10^{-120}).

OnMITEs insertions into *Ostrinia pbp*, mpi, and chitinase genes were within introns, whereas OnMITEs from OnB1 BES data may be located in nongenic regions. Quesneville et al. (2006) found that 14.9% of all *A. gambiae* P MITEs were within introns compared with 0.04% within exons. Prevalence of MITE insertion within introns, as compared with peptide coding regions, may have resulted from preference of trans-acting endonucleases for genome regions with high AT content (Tu 2001). Cut-and-paste movement of class II TEs results in “jumping” of respective DNA segments in or out of a locus, with the only remnant of insertion being an inverted terminal repeat created by staggered endonuclease cleavage and DNA polymerase fill-in. Target site cleave and insertion takes place at AT/TA dinucleotides for OnMITE01 and OnMITE02 (Fig. 2). The AT content of *pbp* and *mpi* introns are 68.8% and 70.3%, respectively, compared with 58.9% and 50.7% of the flanking exons. Similarly, the *O. furnacalis* chit1 intron 1 has 66.3% and flanking exons have 53.4% AT content (OnMITE sequence omitted in all comparisons). Similar to mosquito MITEs (Tu 2001), OnMITEs may insert within introns at high frequency owing to higher prevalence of AT/TA dinucleotides, but the target site specificity of the trans-acting endonuclease will be required to confirm target site preference. Approximately 83% of *P* MITEs are located outside gene coding regions in the *A. gambiae* WGS. Insertions into exon regions likely are highly selected against owing to disruption of encoded proteins, with only 0.4% found disrupting exons (Quesneville et al. 2006). Insertion into transcribed regions does take place and was indicated from *B. mori* and *Bicyclus anynana* ESTs (supplementary material S2). The presence of OnMITEs within introns of *Ostrinia* *pbp*, *mpi*, and chitinase genes may not be selectively neutral, considering that Quesneville et al. (2006) showed that mosquito MITE prevalence within introns is significantly lower than expected by random genomic distribution. OnMITE01 was present within intron 2 in 4 of 83 *O. nubilalis* *pbp* alleles (0.046%, *pbp*-6, -25, -28, and -38: Willett and Harrison 1999; Dopman et al. 2005; Malausa et al. 2007). Similarly, 1 of 31 *O. nubilalis* *mpi* alleles (*mpi*-22) had the OnMITE01 insertion (0.032%; Malausa et al. 2007). A *pbp* phylogeny constructed from intron regions flanking the insertion indicated OnMITE01 jumped into a single allele lineage (supplementary material S3). One member of the same lineage has not undergone OnMITE01 insertion (*pbp*-34, EF396407). Since flanking intron and exon regions of *pbp*-34 have not diverged from *pbp*-6, -25, -28, and -38 alleles that contained OnMITE01, recent insertion or excision of OnMITE01 from a single *pbp* allele lineage was inferred.

**Lepidopteran (GTCT)_n microsatellites hitchhiking within MITEs**

Genome regions flanking microsatellite loci within and between lepidopteran genomes often share sequence similarities and were shown to be paralogs that belonged to multilocus microsatellite families (Anderson et al. 2007; Meglécz et al. 2007; Van’t Hof et al. 2007). The origin of microsatellite flanking sequence similarity remains unknown (Meglécz et al. 2007) but has been hypothesized to result from association of minisatellites and mobile DNA elements (Zhang 2004; Van’t Hof et al. 2007). Microsatellite family members are associated with interspersed repetitive elements of *Drosophila melanogaster* and *A. gambiae* (Meglécz et al. 2007) as well as different TE families (Hoekstra et al. 1997; Fagerberg et al. 2001). Omitted from discussion to this point was...
the presence of (GTCT)$_n$ tetranucleotide microsatellite repeat arrays near the 5’ end of OnMITE01 and OnMITE02 (Fig. 2). (GTCT)$_n$ microsatellites were also found within OnMITE-like regions in several lepidopteran species including in the 5’-UTR of the M. sexta juvenile hormone binding protein precursor (AF226857) and transcribed genome regions in B. anyanana wing (DY763404 and DY64579) and B. mori embryonic EST accessions (CK564259) (supplementary material S2). Genome regions flanking lepidopteran (GTCT)$_n$ microsatellite loci from Drupadia theda, Plutella xylostella, and Utetheisa ornatrix also showed similarity to OnMITE01 (supplementary material S2). A single variant microsatellite repeat array, (GT)$_b$, was observed from a Coenonympha hero sequence in the position homologous to (GTCT)$_n$ repeats of all other MITEs.

Microsatellite loci from D. theda and P. xylostella provided the largest intraspecific microsatellite data sets (Fig. 1). We previously showed two lines of evidence that suggested that OnMITE01 and OnMITE02 are mobile; the two elements were multic locus as indicated from divergent sequences 5’ and 3’ of AT or TA target site duplications, and active transposition was inferred by the presence of absence within introns of php and mpi alleles. Similar to OnMITEs, the D. theda and P. xylostella (GTCT)$_n$ microsatellite flanking regions showed little or no homology outside the putative target site duplications (alignments not shown), which suggested that the “core” MITE-like sequences are located at different genome positions. In contrast, MITE-like “core” regions that immediately flanked D. theda and P. xylostella (GTCT)$_n$ microsatellites showed ≤0.05 nucleotide substitutions per nucleotide site within each species (Fig. 1). Active transposition of D. theda and P. xylostella (GTCT)$_n$ microsatellite loci could not be shown, since homologous sequences 5’ and 3’ of target site duplications that also lacked the putative MITE were not present in GenBank. Analogous to OnMITE01 and OnMITE02, the D. theda and P. xylostella (GTCT)$_n$ microsatellite presented in Fig. 1 may be associated with interspersed repetitive elements. Proliferation of lepidopteran MITEs (LepMITEs) that also carry microsatellite repeats within respective genomes may be responsible for the difficulties that have been incurred during single-locus marker development in several species (Zhang 2004).

Phylogenetic analysis showed that four closely related MITE-like elements are present among lepidopteran genomes (LepMITE01 to LepMITE04). Sequence comparison indicated ≤0.17 nucleotide substitutions per nucleotide site within MITE families, and LepMITE families 03 and 04 appeared to be exclusively present within P. xylostella and D. theda genomes, respectively. The low level of sequence divergence among LepMITE families suggested that they were derived from a single ancestral element. Our comparative analysis suggested that a class of lepidopteran microsatellites are contained within families of related MITE sequences and are multilocus owing to hitchhiking within these mobile units. Tetranucleotide microsatellite repeat arrays previously were encountered within a microsatellite-inducing mobile element, mini-me, from Drosophila, which was hypothesized to function in the dispersal and genesis of new microsatellite loci (Wilders and Hollocher 2001). The origin of lepidopteran microsatellites encoded by mobile MITEs is unknown, but presence in multiple species suggests that entrance may have occurred early in the lepidopteran lineage. The total number of lepidopteran MITE families that contain mobile (CTGT)$_n$ microsatellite repeats, or other microsatellite repeats, is unknown. The presence of a (GT)$_b$ repeat in the putative C. hero MITE sequence suggested that (CTGT)$_n$ microsatellites may have undergone substitution mutation such that other microsatellite repeat motifs also may be present within mobile MITEs and will provide an area of new research in lepidopteran genomics.

Conclusions

BAC end sequencing and STS marker screening were used to characterize the O. nubilalis BAC library, OnB1. Representation of OnB1 was demonstrated through identification of 20 autosomal and 4 Z-chromosome-linked genes and included genes shown to be candidate B. thurinigiensis toxin resistance genes. The OnB1 library was shown to contain clones representative of the Z chromosome involved in the determination of male Resp and larval Ppd. Therefore, if linkage mapping efforts improve QTL intervals, we expect OnB1 to serve as a genomic resource in the positional cloning of the genes underlying those QTL. Our BES data have proven to be a resource for the identification of repetitive genome elements from Lepidoptera. Two eZ-like LINEs were identified and shown to be multiplexed within OnB1. Two microsatellite-containing MITE-like sequences from OnB1 BES were shown to be mobile with the Osiris genome and had homology to other lepidopteran microsatellite sequences. In the current study, we proposed a connection between MITEs and mobile microsatellites within the genome of Lepidoptera. The presence of multilocus microsatellite families in Lepidoptera has hindered the development of single-locus markers for population and linkage analysis. Development of a repetitive element bank for Lepidoptera for use with the RepeatMasker2 algorithm (www.repeatmasker.org/webrepeatmaskerhelp.html#references) may be an effective method by which to avoid the design of PCR primers within known repetitive elements but is contingent on the future development of additional mobile element families.

Acknowledgements

This research was a joint contribution from the USDA, Agricultural Research Service, and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa (project 3543). This article reports the results of research only. Mention of a proprietary product does not constitute an endorsement or a recommendation by the USDA or Iowa State University for its use.

References


