8-2012

High-Throughput Transcriptome Sequencing for SNP and Gene Discovery in a Moth

Nicholas J. Miller
University of Nebraska-Lincoln, nmiller11@iit.edu

Jing Sun
Iowa State University

Thomas W. Sappington
USDA-Agricultural Research Service, tom.sappington@ars.usda.gov

MOLECULAR ECOLOGY AND EVOLUTION

High-Throughput Transcriptome Sequencing for SNP and Gene Discovery in a Moth

NICHOLAS J. MILLER,[1] JING SUN,[2] AND THOMAS W. SAPPINGTON[3]

Environ. Entomol. 41(4): 997–1007 (2012); DOI: http://dx.doi.org/10.1603/EN11216

ABSTRACT The western bean cutworm, *Striacosta albicosta* (Smith) (Leptidoptera: Noctuidae) is a pest of corn (*Zea mays* L.) and dry beans that underwent a dramatic range expansion in North America during the 1st decade of the 21st century. Research into the population genetics of this species has been hindered by a lack of genetic markers. The transcriptome of adult male *S. albicosta* was partially sequenced using Illumina sequencing-by-synthesis. Assembly of the sequence reads yielded 16,847 transcript sequences, of which 6,631 could be assigned a putative function. A search for single nucleotide polymorphisms (SNPs) identified 2,487 candidate SNPs distributed among 1,265 transcripts. A panel of 108 candidate SNPs was selected for empirical testing, of which 68 proved to be assayable polymorphisms that are suitable for population studies. This work provides significant genetic resources for studying *S. albicosta* and demonstrates the power of applying of second-generation sequencing to previously understudied species.

KEY WORDS *Striacosta albicosta*, western bean cutworm, single nucleotide polymorphism, second-generation sequencing, transcriptome

Genetic markers, especially DNA markers, have become indispensable tools for population and ecological genetics. Populations of a huge variety of insects have been studied using DNA markers, including many species that are detrimental to human well-being because they are pests of agriculture or urban environments or vectors of disease (Behura 2006). Periodically, the significance of an arthropod pest increases dramatically and unexpectedly. Often, this situation arises when a species is introduced into a new location and becomes invasive (e.g., Miller et al. 2005, Lombaert et al. 2010, Suhr et al. 2010) or suddenly expands its existing range into new territory or ecological niches (Samarasekera et al. 2012). These events can provoke a need for rapid research on the species in question, including a demand for markers to use in genetic studies.

A recent example of an insect that has suddenly become the subject of intensified interest is the western bean cutworm, *Striacosta albicosta* (Smith). This noctuid has long been recognized as an occasional pest of corn (*Zea mays* L.) and dry beans, largely confined to the Great Plains (Hoerner 1948; Hagen 1962, 1963). At the turn of the 21st century, large, economically damaging populations of *S. albicosta* were observed for the first time in the United States east of the Missouri River (O'Rourke and Hutchinson 2000). This event heralded the start of an explosive eastward range expansion. By 2009, *S. albicosta* was to be found as far east as Pennsylvania (Tooker and Fleischer 2010).

The spectacular range expansion of *S. albicosta* has, unsurprisingly, prompted a number of hypotheses regarding its cause. Dorhout and Rice (2010) have suggested that the adoption of Cry1Ab-expressing transgenic corn has opened up an ecological niche because *S. albicosta* is much less sensitive to the toxin than is *Helicoverpa zea* (Boddie). Both species are ear feeders, but the more agressive *H. zea* larva typically will kill any other lepidopteran larva on the same plant (Dorhout and Rice 2010). Another suggestion is that the range expansion is associated with a recent escape from pathogens that otherwise kept populations low and relatively sedentary (Dorhout 2007). An alternative hypothesis is that the species recently managed to cross a physical barrier associated with the Missouri River (Miller et al. 2009a).

No definitive evidence has been found in support of any of these hypotheses. However, a study of mitochondrial DNA variation found no evidence of reduced genetic diversity in newly-established eastern populations. Reduced diversity would be expected for a bottleneck associated with a small pioneer population being transported by chance beyond a stable physical barrier (Miller et al. 2009a). The mitochondrial DNA study therefore favored hypotheses that invoke a loss of constraint to mass eastward movement of *S. albicosta.* Although this study provided some

This article reports the results of research only. Mention of a proprietary product does not constitute an endorsement or a recommendation by the USDA for its use.

[1] Corresponding author: Nicholas Miller, Department of Entomology, Entomology Hall, University of Nebraska-Lincoln, Lincoln, NE 68583-0816. (e-mail: nmiller4@unl.edu).

[2] Department of Entomology, Iowa State University, Ames, IA.

[3] USDA–ARS Corn Insects and Crop Genetics Research Unit, Genetics Laboratory, Iowa State University, Ames, IA.

0046-225X/12/0997–1007$04.00/0

clues to the cause of the range expansion by *S. albicosta*, the power of mitochondrial DNA sequencing to detect subtle changes in genetic diversity is limited because the mitochondrial genome is a single genetic locus. Mitochondrial DNA sequencing was selected as a genetic marker system because, at the time, no nucleic acid sequence data were available for *S. albicosta.* Under such circumstances mitochondrial DNA is appealing because it can be examined using universal polymerase chain reaction (PCR) primers that will amplify regions of the mitochondrial genome in most arthropods (Simon et al. 1994).

The obligatory selection of a less-informative marker system for our earlier study highlights a fundamental dilemma when investigating the genetics of previously understudied species. The preferred genetic markers for population genetic studies are usually multiple independent single locus codominant markers such as microsatellites and single nucleotide polymorphisms (SNPs) (Miller et al. 2009b). These markers generally require a significant amount of time and resources to be expended to first identify and then develop assays for polymorphic loci. This problem is particularly acute when studying Lepidotera, for which the development of reliable microsatellite markers is often extremely challenging. Lepidopteran genomes contain mobile elements that generate duplicate microsatellite loci, including associated flanking sequences throughout the genome (Meglécz et al. 2007, Coates et al. 2009). In contrast, genetic markers that are initially easier and cheaper to develop have less power to detect demographic changes and population structure. The power of mitochondrial DNA is limited because all sites on the mitochondrial genome are linked and therefore not independent. Anonymous "fingerprinting" techniques like amplified fragment length polymorphism and random amplification of polymorphic DNA suffer because they generate markers with dominant inheritance and the (sometimes unrealistic) assumption usually must be made that individuals within predefined groups are mating at random (Bonin et al. 2007).

Recent advances in nucleic acid sequencing, commonly referred to as next- or second-generation sequencing, have the potential to ease the dilemma of genetic marker selection by greatly accelerating the discovery of many codominant genetic markers. Here, we describe the application of second-generation sequencing to the transcriptome of western bean cutworm as a rapid and relatively inexpensive way to discover SNP markers and gene sequences from a previously under-studied insect.

Methods

Insect Sampling. Adult male *S. albicosta* were collected from four locations in Yuma and Phillips counties, northeast Colorado on the night of 21st-22nd July 2008. Four wire-mesh 75-cm-diameter Hartstack cone traps (Hartstack et al. 1979, Reardon et al. 2006), baited with *S. albicosta* sex pheromone (Trécé, Adair, OK) were placed at each location and left overnight. Captured moths were anesthetized with carbon dioxide and transferred to round plastic containers (80 mm in diameter by 80 mm deep), the lids of which had windows covered in wire mesh to allow ventilation. The containers were placed in insulated coolers with ice packs to minimize moth mortality. The insects were transported to the USDA-ARS CICGRU laboratory in Ames, IA. Moths that survived transport were placed in 1.5-ml microcentrifuge tubes, flash-frozen in liquid nitrogen, and stored at −80°C until required for RNA extraction.

Preparation and Sequencing of Normalized cDNA. Sixteen adult male *S. albicosta* were ground together under liquid nitrogen using a pestle and mortar. The pulverized tissue was mixed with 16-ml TRIzol reagent (Invitrogen, Carlsbad, CA) and transferred to a 50-ml disposable centrifuge tube. The suspension was centrifuged at 4,000 × g, 4°C for 15 min to pellet exoskeleton particles and other insoluble debris. Total RNA was extracted from the supernatant by following the manufacturer-supplied protocol for TRIzol and was resuspended in 50-μl nuclease-free water. The concentration and purity of the RNA solution was estimated using a Nanodrop UV-spectrophotometer (Thermo Scientific, Wilmington, DE).

Messenger RNA was purified from 3.3 mg total RNA by using an Ambion Poly(A)Purist kit (Invitrogen), by following the manufacturer's protocol, and dissolved in 30-μl nuclease-free water. The concentration of the mRNA solution was estimated using a Nanodrop UV-spectrophotometer. Double-stranded cDNA was synthesized from 1-μg mRNA by using a MINT kit (Evrogen, Moscow, Russia) and then normalized using a TRIMMER kit (Evrogen) by following the manufacturer's protocols. The normalized cDNA sample was sent to the Iowa State University DNA Facility for sequencing using a Genome Analyzer II (Illumina, San Diego, CA). The ISU DNA Facility processed the cDNA sample by following Illumina's published protocols for genomic DNA and sequenced the sample in two lanes of a 75-base single read run. Sequence read data were deposited with the NCBI Sequence Read Archive (Accession SRA010931).

Sequence Assembly. Before assembly, low-quality reads were removed using a custom Python (van Rossum and de Boer 1991) script that made use of the Biopython libraries (Cock et al. 2009) to filter the FASTQ-formatted read data. The script removed bases from both ends of the reads until the ends were composed of five consecutive bases with quality scores of 10 or more. After removing low-quality ends, the script removed reads that were <35 nucleotides long or contained any windows of five nucleotides where the mean quality score was below 15. The Cross_match (Gordon et al. 1998) program was used to identify regions of the quality-filtered reads that matched the sequences of the adapters and primers from the MINT and TRIMMER kits. Matching regions were then clipped from the reads using a custom Python script and 8,752,259 clipped reads of 35 nucleotides or more were retained for assembly.

Reads were assembled using a combination of SSAKE (Warren et al. 2007) and CAP3 (Huang and Madan 1999). Both programs were run using their default parameters. The combination of the two assembly programs was necessary to circumvent the memory limitations imposed on SSAKE by the 32-bit Perl interpreter that was used at the time. Assembly was conducted in two rounds. In the first round, the reads were split into batches of $\leq 10^6$ and each batch was assembled with SSAKE. Two output files were obtained for each batch: the consensus sequences of assembled contigs and the sequences of singlet reads that were not included in a contig. The contig consensus sequences from each batch were then combined into a single input file and further assembled using CAP3. In the second round, the CAP3 output was used to provide "seed" sequences to SSAKE. The seeds were extended using the singlet sequences left over from the first round of SSAKE assembly. Finally, the contig consensus sequences from the second round of SSAKE assembly were used as input for a second round of CAP3 assembly. Assembled sequences shorter than 100 bases were discarded.

The fate of each read during the assembly process was tracked using a combination of the output from SSAKE, a custom SQLite3 (available from www.sqlite.org) database and several custom python scripts. This procedure resulted in a list of reads that had been incorporated into each contig. The assembly was further refined by aligning reads to the consensus sequence of their contig using Consed version 19 (Gordon et al. 1998), as described in the program's manual.

Annotation. Gene Ontology (GO) terms were associated with assembled *S. albicosta* transcript sequences by using Blast2GO (Conesa et al. 2005). A BLASTX (Altschul et al. 1997) search of the NCBI nonredundant protein database was performed for each *S. albicosta* sequence with a minimum E-value of 10^{-3} and a minimum HSP length of 25. Blast2GO then was used to infer GO terms for the *S. albicosta* sequences based on the annotations of the corresponding BLASTX hits. The initial GO annotations were improved by running the optional "ANNEX" analysis provided by Blast2GO (Gotz et al. 2008). To obtain an overview of the distribution of high-level GO terms, the GOA GO Slim was applied using the functionality provided by Blast2GO.

The BLASTX hits furnished by the Blast2GO analysis were also used to identify the likely organismal source (i.e., *S. albicosta* versus associated microorganisms) of the sequences. For each transcript sequence for which one or more BLASTX hits were obtained, the accession number of the single highest-scoring hit was exported from Blast2GO. A custom Python/Biopython script was used to query the NCBI databases and retrieve the taxonomy id associated with each accession number. The script then queried the NCBI taxonomy database to recover, where applicable, the superkingdom, kingdom, phylum, class, order, family, genus and scientific name for each taxonomy id.

Verification of Selected Assembled Sequences. A set of ten assembled sequences were selected to evaluate the accuracy of the assembly. The sequences were chosen from sequences that were annotated as being of insect origin and ≥500 bases long. Primer three (Rozen and Skaletsky 2000) was used to design PCR primers to amplify a region between 500 and 1,000 basepairs in length from each of the selected sequences. The target amplicons were amplified by PCR and purified by one of two methods. Amplicons derived from sequences EZ579877, EZ579885, EZ579916, EZ579933, and EZ579952 were amplified in reactions of 20-μl volume containing five ng normalized cDNA (see above) as a template, forward and reverse primers at 0.2 μM, dNTPs at 0.2 mM each, 1.5 mM $MgCl_2$, 0.5 U GoTaq Flexi DNA polymerase (Promega, Madison, WI) in one X GoTaq Flexi Buffer. Temperature cycling conditions for PCR were: initial denaturation at 95°C for 2 min followed by 30 cycles of 95°C for 30 s, 56°C for 45 s, 72°C for 1 min, and a final extension step of 72°C for 5 min. PCR products were purified using an IBI Gel/PCR DNA Fragments extraction kit (IBI, Peosta, IA). Amplicons derived from sequences EZ579865, EZ579872, EZ579873, EZ579936, and EZ579964 were amplified in 20-μl reactions of one X Qiagen Multiplex PCR Master Mix (Qiagen, Valencia, CA) containing 0.2 μM each primer and five ng normalized cDNA. Temperature cycling conditions were 95°C for 15 min followed by 35 cycles of 94°C for 30 s, 56°C for 90 s, and 72°C for 90 s, with a final extension step of 72°C for 10 min. PCR products were purified by pooling three reactions per amplicon, which were added to 440 μl 10 mM Tris-HCl (pH 8.0). The diluted PCR products were then reconcentrated by centrifugation through a Microcon Ultracell YM-100 spin column (Millipore, Billerica, MA) at 500 X g for 24 min followed by addition of a further 500 μl 10 mM Tris and an additional centrifugation of 500 X g for 24 min.

Purified PCR products were ligated into pGEM-T Easy plasmid vector (Promega) according to the supplier's instructions. The plasmid/PCR ligation was used to transform *Escherichia coli* XL1-Blue cells by electroporation. Plasmid DNA was isolated from four or five isolated *E. coli* colonies for each of the 10 PCR products. Plasmids were purified using Zyppy Plasmid Miniprep Kits (Zymo Research, Orange, CA) and used as template for Sanger sequencing in both directions using standard T7 and SP6 sequencing primers with a GenomeLab DTCS Quick Start Kit (Beckman Coulter, Fullerton, CA). Sequencing reaction products were analyzed by capillary electrophoresis by using a CEQ 8000 instrument (Beckman Coulter). Sequencing reads were assembled to reconstruct the sequences of the 10 amplicons by using the programs Phred, Phrap and Consed (Gordon et al. 1998).

SNP Detection and Testing. Searching for putative SNPs was restricted to sequences that met two conditions. Only sequences that were annotated as being of insect origin were used, to avoid SNPs in sequences originating from moth-associated micro-organisms. Of these, only sequences ≥500 nucleotides in length

were searched for SNPs, to maximize the chances that a SNP assay could be developed. This subset of sequences was used as a reference to which the original sequencing reads were mapped using Maq (Li et al. 2008). The list of putative SNPs provided by Maq was filtered to remove likely false positives and problematic SNPs using the following criteria for inclusion: consensus quality at the SNP position >50, average number of hits for reads at the SNP position ≤1, minimum consensus quality in the three positions flanking the SNP >20 and no other credible SNPs within 50 nucleotides.

Before testing putative SNPs experimentally, an additional filter was applied to minimize the incidence of false-positive candidates because of paralogous sequences. The sequences bearing putative SNPs were used as query sequences in a BLASTX search of the GLEAN-predicted *Bombyx mori* L. peptide set. Sequences that gave a hit to multiple *B. mori* peptides were excluded from further testing because they were likely to be members of conserved gene families.

The SNP-detection analysis was performed by mapping reads against the consensus sequences produced by SSAKE and cap3, before the refinement of the assembly. Because only trivial differences were observed between consensus sequences before and after read mapping, the positions of the already-detected SNPs were converted to their positions on the refined contig sequences, rather than repeating the entire SNP detection analysis unnecessarily. Conversion of SNP locations was done by aligning the initial and refined consensus sequence for each contig using "needle," part of the EMBOSS (Rice et al. 2000) suite. The aligned sequences were parsed using a custom Python script to convert SNP locations on the initial contig to the corresponding position on the refined contig.

Empirical testing of putative SNPs was done using the Sequenom MassARRAY system operated by the Genomic Technologies Facility, Iowa State University. The Sequenom Assay Design 3.1 software was used to design multiplex MassARRAY assays from a list of 273 candidate SNPs. Three multiplex assays of 36 candidate SNPs each (Table 1) were selected for testing. Testing was performed on a panel of 96 adult male moths collected from two sites in Yuma County, CO. DNA was extracted from individual moths by using a Cetyl trimethylammonium bromide-based method (Marçon et al. 1999) and resuspended in 100-μl TE buffer. The concentration of each DNA sample was estimated using a Nanodrop UV-spectrophotometer and diluted to 10 ng/μl with water, whereupon it was transferred to the Genomic Technologies facility for MassARRAY analysis.

The Genepop computer program (Rousset 2008) was used to compute allele frequencies and to perform exact tests of genotypic differentiation between the two sample locations, of deviations from Hardy–Weinberg genotypic proportions and of deviations from linkage equilibrium between pairs of loci. The procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) was used to control the false discovery rate for tests at multiple loci, using a q-value of 0.05. The unbiased neutral distribution of SNPs, conditional on observing both alleles in a sample of 96 diploid individuals was simulated using CoaSim (Mailund et al. 2005). Ten thousand SNP loci were simulated with minor allele frequencies between 1/192 and 0.5.

Results

Assembly. Two lanes of sequencing on an Illumina Genome Analyzer II yielded 13,231,040 reads of 75 nucleotides each. Filtering out low-quality reads and adaptor and primer sequences left 8,752,259 reads between 35 and 75 nucleotides in length (mean 61.7). An initial assembly of the filtered reads using SSAKE and cap3 yielded 16,850 contiguous sequences ≥100 nucleotides in length. Refining the assembly by mapping reads back to their corresponding consensus sequences revealed three contigs for which the constituent reads could not be mapped satisfactorily to the consensus sequence, leaving 16,847 acceptable contigs in the final assembly. Read mapping also resulted in the truncation of the consensus sequences of 2,840 contigs (16.8%) by a few (mean 9.18) bases because reads could not be mapped to one end of the consensus sequence. Most contigs were relatively short (Fig. 1), the median length was 173 nucleotides. However, the total number of contigs was large enough that 2,527 contigs ≥500 nucleotides long were obtained. The consensus sequences of the contigs in the final assembly were deposited with the NCBI Transcriptome Shotgun Assembly database under accession numbers EZ579864–EZ596710.

Annotation. In total, 6,631 sequences (39.4%) had one or more significant BLASTX hits to the NCBI nonredundant protein database. Sequences that produced a BLASTX hit were significantly longer than those that did not (Wilcoxon-rank-sum $W = 52813264$, $P < 2.2 \times 10^{-16}$). Blast2GO was able to assign one or more GO terms to 4,139 sequences (24.6%). A wide range of GO terms were represented, with 59 of the 66 terms in the GOA GO-Slim (89%) associated with one or more sequences. The GO terms associated with each sequence are given in Supplemental File S1 and gene descriptions assigned by blast2go are given in Supplemental File S2.

The majority of best BLASTX hits were to sequences from insects. In total, 5,306 sequences (80%) had their best hit to a database sequence of arthopod origin, of which 5,248 were insect sequences. The next most well represented phylum was Microsporidia, which contributed 732 hits (11%) followed by Chordata with 259 hits (3.9%). The remaining database hits were to sequences from a wide variety of sources including bacteria, archea, viruses, fungi, plants, and animals. The distribution of phyla from which the best BLASTX hits came is illustrated in Fig. 2 and full details are provided in Supplemental File S3.

Assembly Validation. PCR primers were designed to amplify portions of ten of the assembled contigs

Table 1. Details of western bean cutworm SNP loci tested using the Sequenom MassARRAY system

Accession	Position	Segregating bases	Sequenom multiplex assay	Validation result	Minor allele frequency
EZ582923	335	C/T	1	Monomorphic	0
EZ586136	144	A/G	1	Pass	0.355
EZ583491	218	G/A	1	Bad assay	
EZ589028	350	T/A	1	Pass	0.408
EZ588114	602	G/A	1	Pass	0.245
EZ581071	228	A/G	1	Bad assay	
EZ582479	675	A/T	1	Pass	0.043
EZ589446	597	G/A	1	Pass	0.253
EZ584512	406	C/T	1	Bad assay	
EZ587139	605	T/C	1	Monomorphic	0
EZ582979	616	T/C	1	Bad assay	
EZ582821	733	C/T	1	Pass	0.237
EZ580337	463	G/T	1	Bad assay	
EZ596238	375	T/A	1	non Hardy–Weinberg	0.416
EZ583763	530	A/G	1	Pass	0.484
EZ596199	448	G/A	1	Pass	0.444
EZ580320	466	A/G	1	Monomorphic	0
EZ584270	288	A/G	1	Pass	0.005
EZ583570	179	G/A	1	Pass	0.172
EZ587712	712	G/A	1	Bad assay	
EZ589824	160	G/A	1	Pass	0.464
EZ583827	590	T/C	1	Monomorphic	0
EZ585585	430	A/G	1	Pass	0.231
EZ593529	339	C/T	1	Bad assay	
EZ591511	683	T/C	1	Pass	0.016
EZ586502	749	C/G	1	Pass	0.237
EZ583798	515	A/G	1	Pass	0.151
EZ595986	460	A/G	1	Pass	0.367
EZ596544	739	G/A	1	Pass	0.057
EZ581779	601	G/A	1	Bad assay	
EZ582480	468	T/C	1	Pass	0.489
EZ580138	467	A/T	1	Pass	0.032
EZ583057	565	T/C	1	Pass	0.090
EZ584803	331	G/A	1	Pass	0.194
EZ583859	772	A/G	1	Pass	0.071
EZ582464	392	G/C	1	Bad assay	
EZ584879	133	T/C	2	Bad assay	
EZ582039	493	G/A	2	Pass	0.112
EZ596004	342	G/A	2	Pass	0.021
EZ595136	259	G/A	2	Pass	0.134
EZ580375	363	T/C	2	Monomorphic	0
EZ580294	608	C/T	2	Pass	0.156
EZ585939	420	C/T	2	Pass	0.005
EZ583794	565	A/G	2	Pass	0.138
EZ581148	649	T/A	2	Pass	0.101
EZ580153	447	G/A	2	Pass	0.085
EZ579885	397	C/T	2	Pass	0.231
EZ580163	844	G/A	2	Pass	0.037
EZ582889	638	T/C	2	Pass	0.016
EZ580742	389	C/T	2	Monomorphic	0
EZ580556	307	C/T	2	Pass	0.377
EZ592313	561	G/A	2	Pass	0.034
EZ581729	559	G/A	2	Bad assay	
EZ583314	611	C/T	2	non Hardy–Weinberg	0.091
EZ587221	510	C/T	2	Pass	0.269
EZ581948	380	G/A	2	Monomorphic	0
EZ589139	555	G/A	2	Pass	0.29
EZ581913	513	C/T	2	Pass	0.051
EZ580650	330	T/C	2	Pass	0.044
EZ580647	772	T/C	2	Pass	0.102
EZ583455	211	C/T	2	Pass	0.468
EZ582582	260	A/G	2	non Hardy–Weinberg	0.3
EZ587844	657	G/A	2	Bad assay	
EZ581706	275	C/T	2	Pass	0.29
EZ584485	367	T/C	2	Pass	0.256
EZ581845	493	A/G	2	Pass	0.258
EZ582963	506	A/G	2	Bad assay	
EZ585447	567	T/C	2	Bad assay	
EZ584256	639	G/T	2	Bad assay	
EZ581223	474	C/A	2	non Hardy–Weinberg	0.117
EZ580019	94	G/T	2	non Hardy–Weinberg	0.212

Continued on following page

Table 1. Continued

Accession	Position	Segregating bases	Sequenom multiplex assay	Validation result	Minor allele frequency
EZ581778	950	A/G	2	non Hardy–Weinberg	0.378
EZ581379	408	G/T	3	Pass	0.345
EZ593488	389	C/T	3	Pass	0.279
EZ582115	290	G/A	3	Bad assay	
EZ582061	760	G/A	3	Bad assay	
EZ582842	240	C/A	3	Pass	0.145
EZ590074	167	C/T	3	Pass	0.238
EZ583961	480	T/A	3	Pass	0.471
EZ582238	549	A/T	3	Pass	0.369
EZ582388	421	C/T	3	non Hardy–Weinberg	0.134
EZ582943	644	G/A	3	Monomorphic	0
EZ580614	208	G/A	3	Pass	0.4
EZ581065	507	T/C	3	Pass	0.035
EZ582648	284	G/A	3	Pass	0.353
EZ583901	268	G/A	3	Pass	0.174
EZ585108	326	G/A	3	Pass	0.141
EZ579981	623	A/G	3	Bad assay	
EZ580545	753	T/C	3	Pass	0.018
EZ591759	296	G/T	3	Pass	0.011
EZ582897	398	A/T	3	Pass	0.23
EZ583780	514	A/G	3	non Hardy–Weinberg	0.089
EZ580823	892	G/A	3	Pass	0.483
EZ582766	499	C/T	3	non Hardy–Weinberg	0.123
EZ585760	929	A/G	3	Bad assay	
EZ581163	451	G/T	3	Bad assay	
EZ585488	329	C/T	3	Pass	0.241
EZ583018	758	T/C	3	Pass	0.464
EZ581542	327	T/C	3	Pass	0.377
EZ582121	610	G/A	3	Pass	0.098
EZ581402	428	C/T	3	Bad assay	
EZ581488	871	T/C	3	Pass	0.067
EZ582275	285	C/T	3	Bad assay	
EZ582641	540	C/T	3	Bad assay	
EZ583984	224	T/C	3	Pass	0.053
EZ582430	505	T/A	3	Pass	0.142
EZ588606	738	T/A	3	Pass	0.088
EZ584861	198	G/C	3	Pass	0.037

(Table 2). All 10 primer sets produced an amplicon that agarose gel electrophoresis indicated was of the size predicted from the contig consensus sequences. Cloning and sequencing of the amplicons revealed that nine of the amplicons were of precisely the length predicted from their corresponding consensus sequence. The amplicon produced by primers designed from contig EZ579873 differed from the expected length by a single nucleotide. The pair of primers designed to amplify contig EZ579865 also produced a second, shorter amplicon. The reduced length of this amplicon was because of an internal deletion (with respect to the consensus sequence) of 203 nucleotides (Fig. 3). Sequence identity between the amplicons

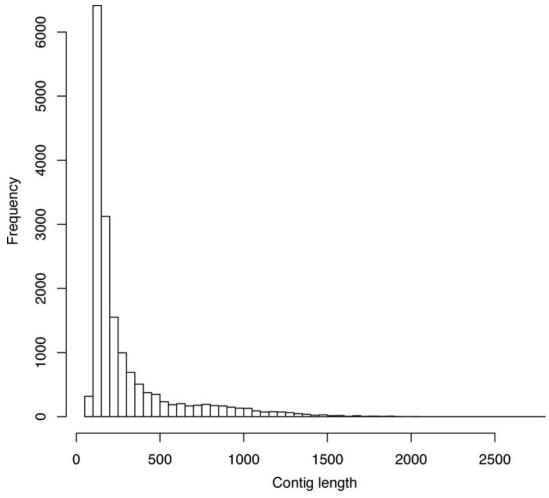

Fig. 1. Length distribution of assembled western bean cutworm transcript contigs.

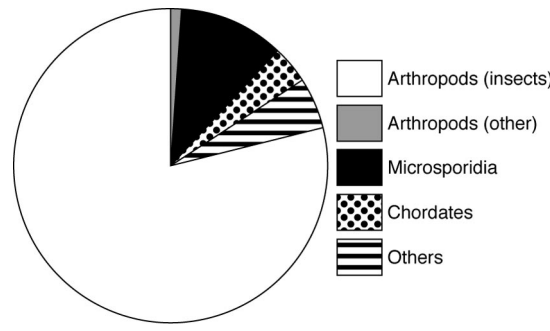

Fig. 2. Taxonomic distribution of best BLASTX hits for assembled transcript contigs.

Table 2. Comparison of ten PCR amplicon sequences to the contig consensus sequences from which their PCR primers were designed

Accession	Expected amplicon length	Observed amplicon length	% identity to contig
EZ579865	539	539 (336)	99.4
EZ579872	631	631	99.5
EZ579873	742	741	99.2
EZ579877	701	701	99.9
EZ579885	862	862	99.8
EZ579916	712	712	99.0
EZ579933	772	772	97.9
EZ579936	781	781	98.2
EZ579952	457	457	98.5
EZ579964	900	900	99.6

and their respective contigs was high, ranging from 97.9 to 99.9%.

SNP Detection and Testing. An initial, unfiltered search for putative SNPs in 1,780 contigs >500 bp, using Maq identified 55,222 candidates, distributed over 1,776 contigs. Applying an initial quality filter reduced the number of credible candidate SNPs to 2,487, distributed over 1,265 contigs. BLASTX searches against the predicted silkworm peptides set showed that 273 contigs did not hit to more than one silkworm peptide and therefore were unlikely to be members of conserved gene families.

The most centrally positioned candidate SNPs in 108 of the 273 contigs were selected for empirical validation and assayed in three multiplex reactions (36 SNPs each) by using the Sequenom MassArray platform. Details of each candidate SNP and the outcome of empirical testing are given in Table 1. Candidate SNPs were rejected on the grounds of a bad assay either if no or few individuals in the test panel generated data or if a majority of individuals had genotype calls that the Sequenom analysis software designated "Aggressive" or "Low Probability" or that required user intervention and could not be resolved. Eighty-five of 108 candidate SNPs gave reliable assay data.

Eight candidate SNPs were found to be monomorphic in the test panel of 96 individuals (Table 1). After controlling the False Discovery Rate (FDR) to $q = 0.05$, no locus showed significant genotypic differentiation between the two sites from which the test individuals were collected. This result allowed all 96 individuals to be pooled into a single sample for the purpose of testing for Hardy-Weinberg genotypic proportions. Controlling the FDR with $q = 0.05$, nine SNPs were identified with genotype frequencies that differed significantly from Hardy–Weinberg expectations (Table 1). The distribution of minor allele frequencies (i.e., the frequency of the less common allele) among loci that were both polymorphic and conformed to Hardy–Weinberg expectations is illustrated in Fig. 4. The distribution of minor allele frequencies deviated slightly from a uniform distribution between 0 and 0.5, with a slight excess of alleles at low frequencies. This represents an excess of SNPs with high-frequency minor alleles compared with the unbiased neutral allele frequency spectrum (Fig. 5). After controlling the FDR with $q = 0.05$, two pairs of SNP loci exhibited significant deviations from linkage equilibrium. These were contig EZ589028, position 350 with contig EZ588114, position 602 and contig EZ589028, position 350 with contig EZ589139, position 555.

Discussion

The length distribution of contigs in our assembly was skewed heavily toward short sequences. Thus, most of the contigs represent fragments of expressed sequences rather than full-length transcripts. This length distribution is in keeping with some other as-

```
EZ579865       TAGACCGCTTTTCACGAATGCAAGTGAAACAATTTTTTTCTTTCTCTTCATTTCCATCCGTAGTCCACGCCAAACACCTATTAAGAGCCTCTTCCCATCG
Short Amplicon ------------------------------------------------------------GTAGTCCACGCCAAACACCTATTAAGAGCCTCTTCCCATCG
Long Amplicon  ------------------------------------------------------------GTAGTCCACGCCAAACACCTATTAAGAGCCTCTTCCCATCG

EZ579865       GACACGTCTGATATCACGTTCTTCGGAAGTCATTGCTGGCGAAAAGGTATCTGCTGGCGGGCTTGGTGTAGACGTGGGCCACACTCGCAGTGCCCGGCCC
Short Amplicon GACACGTCTGATATCACGTTCTTCGGAAGTCATTGCTGGCGACAAAGTGTCTGCTGGCGGGCTTGGTGTAGACGTGGGCCACACTCGCAGTGCCCGGCCC
Long Amplicon  GACACGTCTGATATCACGTTCTTCGGAAGTCATTGCTGGCGAAGAAGTATCTGCTGGCGGGCTTGGTGTAGACGTGGGCCACACTCGCAGTGCCCGGCCC

EZ579865       GCTACTATAGCAGCTCCCAGCGCTGTACTTTCCATCATTAGAGGTCGAATCACAGGGATTCCCAGAATGTCTGCTTGCATTTGCATCAAAAGAGTGTTTT
Short Amplicon GCTACTATAGCAGCTCCCAGCGCTGTACTTTCCATCATTAGAGGTCGAAT----------------------------------------------------
Long Amplicon  GCTACTATAGCAGCTCCCAGCGCTGTACTTTCCATCATTAGAGGTCGAATCACAGGGATTCCCAGAATGTCTGCTTGCATTTGCATCAAAAGAGTGTTTT

EZ579865       GTGCCATACCACCGTCGGCAAGTAGTCGCCCCACTGGCGCACAGTCCTGTGCCATCGCTGTCGCCACACTACGCGTTTGATGACACACCGACTCCAAAGC
Short Amplicon ----------------------------------------------------------------------------------------------------
Long Amplicon  GTGCCATACCACCGTCGGCAAGTAGTCGCCCCACTGGCGCACAGTCCTGTGCCATCGCTGTCGCCACACTACGCGTTTGATGACACACCGACTCCAAAGC

EZ579865       TGCCCTGACTATGTGTTCACGACGTGTCTTGCCACTAATTCCACACATTATACCTCTAGCGTCTTTCCTCCAATAAGGAGAGTATAAACCGTTAAATGCA
Short Amplicon ------------------------------------------------------CTCTAGCGTCTTTCCTCCAATAAGGAGAGTATAAACCGTTAAATGCA
Long Amplicon  TGCCCTGACTATGTGTTCACGACGTGTCTTGCCACTAATTCCACACATTATACCTCTAGCGTCTTTCCTCCAATAAGGAGAGTATAAACCGTTAAATGCA

EZ579865       GGTACGAAACAAATGCTTCCTTCTTCATCATCAGCAACCTCTGCTCCCATCGTCTCAGATTCTTGTATATCAGTGATCAGTTCCAAGCCGTCTCGTAGCC
Short Amplicon GGTACGAAACAAATGCTTCCTTCTTCATCATCAGCAACCTCTGCTGCCATCGTCTCAGATTCTTGTATATCAGTGATCAGTTCCAAGCCGTCTCGTAG--
Long Amplicon  GGTACGAAACAAATGCTTCCTTCTTCATCATCAGCAACCTCTGCTGCCATCGTCTCAGATTCTTGTATATCAGTGATCAGTTCCAAGCCGTCTCGTAG--

EZ579865       ACTGTAATGTGTCACCT
Short Amplicon -----------------
Long Amplicon  -----------------
```

Fig. 3. Alignment of the sequences of two PCR amplicons to the sequence of contig EZ 579865, from which PCR primers were designed. Discrepant nucleotides are shaded.

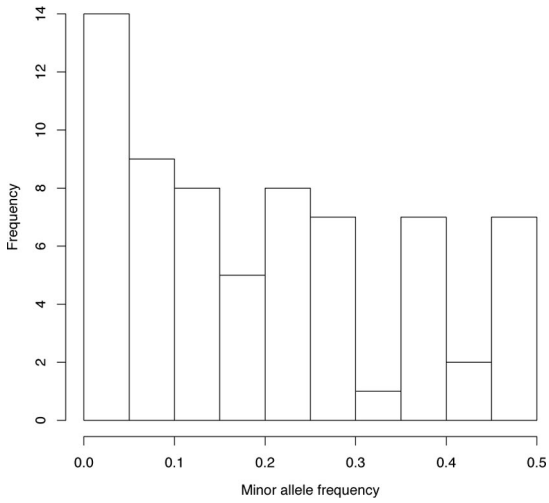

Fig. 4. Distribution of minor allele frequencies at 68 SNP loci.

semblies of insect transcriptomes based on short-read Illumina sequencing. The median contig length of our *S.albicosta* assembly (173 bp) fell between those of recent assemblies of the transcriptomes of *Bemisia tabaci* (Gennadius) (357 bp) (Wang et al. 2010) and *Acyrthosiphon pisum* (Harris) (92 bp) (G. R. Burke and N. A. Moran, unpublished NCBI submission, project accession PRJNA52531). Because the primary purpose of the assembly reported here was SNP discovery, the fact that most contigs were short and therefore likely to be partial transcript sequences was not a major concern. Nevertheless, the ability to assemble full-length transcript sequences from cost-effective short sequence reads is desirable. A significant issue that confronted us when assembling our sequences was that all of the available assembly programs designed for short read data were developed with a focus on assembling genome sequences. Genomes and transcriptomes differ in important ways, most notably in their expected read coverage. Thus, transcriptomes may violate some of the assumptions built into algorithms for assembling genomes from short reads (e.g., Chaisson and Pevzner 2008, Zerbino and Birney 2008). Fortunately, a number of tools are being developed to extend assembly software designed for genomes to account for the peculiarities of transcriptomes (Robertson et al. 2010, Grabherr et al. 2011). Recently, Crawford et al. (2010), demonstrated that the contig length of a short-read based transcriptome assembly of *Anopheles funestus* Giles could be significantly improved by taking advantage of modest amounts of preexisting traditional ESTs, in combination with peptide sequences derived from the *Anopheles gambiae* Giles genome. Taking a different approach, Chen et al. (2010), were able to assemble a transcriptome of *Locusta migratoria* (L.) with reasonably long contigs by carrying out paired end sequencing at much greater depth (roughly 21-fold) than we used for *S. albicosta.* Evidently, the state of the art in short-read based transcriptome sequencing is advancing at a remarkable rate and researchers sequencing new transcriptomes in the future may anticipate considerably more complete assemblies than we achieved for *S. albicosta.*

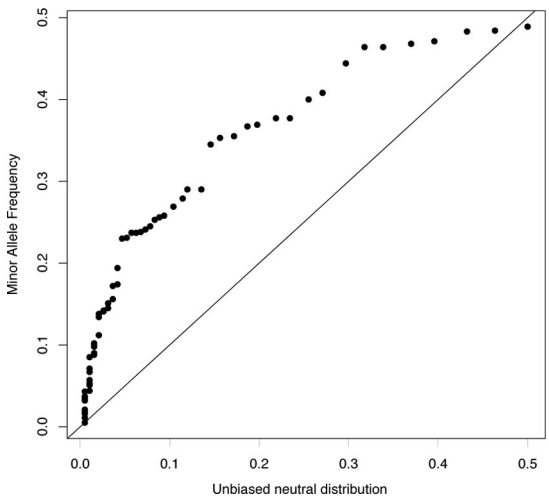

Fig. 5. Quantile-quantile plot of minor allele frequencies at 68 SNP loci versus the unbiased neutral minor allele frequency spectrum, based on 10,000 coalescent simulations.

A striking feature of our annotated sequences was that >10% appeared to be of microsporidian, rather than insect origin. *Striacosta albicosta* is commonly infected by a microsporidium in the genus *Nosema* (Su 1976), so the discovery of microsporidian transcripts is not a great surprise. Nevertheless, this result highlights the fact that an effort to sequence the transcriptome of a multicellar organism is often an exercise in sequencing the meta-transcriptome of the target organism plus its associated microorganisms. It is not uncommon to identify appreciable numbers of nonarthropod sequences when sequencing the transcriptomes of whole insects (Zhang et al. 2010, Xue et al. 2010, Karatolos et al. 2011, Bai et al. 2011), although the source of these sequences (pathogens, symbionts or contaminants) is not always clear. When the motivation for transcriptome sequencing is to develop genetic markers, it is especially important to perform a taxonomic analysis of the resulting sequences. Such an analysis is the only practical way to be confident that any candidate makers do indeed originate from the genome of the organism of interest.

Unsurprisingly, given that we chose to sequence the transcriptome of complete adult insects, our sequences were annotated with a wide variety of putative functions, processes and cellular locations. Potentially, our choice of whole insects risked our missing transcripts that are tissue-specific or expressed at low levels because they could be diluted by housekeeping transcripts that are expressed in most cells. Normalizing the cDNA before sequencing helped to mitigate this risk. Although many of the transcripts were annotated with housekeeping functions others appeared to be tissue-specific (Supplementary Files S1 and S2), including transcripts with complete or

nearly complete open reading frames (ORFs). For example, genes involved in sensory perception included four complete odorant-binding protein ORFs (EZ581712, EZ583454, EZ584056, and EZ585325) and two nearly complete opsin ORFs (EZ583488 and EZ585684). Immunity related genes included a complete defensin ORF (EZ596498) and nearly complete hemolin (EZ580554) and petidoglycan recognition protein ORFs (EZ580007). Many more short contigs also were annotated with putative specialist roles. These short contigs represent fragments of genes but the full length transcripts could, in principle, be easily obtained using standard methods such as RACE (Rapid Amplification of cDNA Ends).

The main objective of the work reported in this paper was the identification of SNPs that could be used for genetic studies of *S. albicosta.* Hundreds of candidate SNPs were identified Nearly two-thirds of the candidate SNPs that were tested proved to be genuine polymorphisms with Hardy-Weinberg genotypic proportions that should be suitable for population genetics research on the species. One SNP marker, on contig EZ589028 at position 350 exhibited significant linkage disequilibrium with two other loci and should therefore be avoided for use in population genetic analyses for which independence between loci is assumed.

The allele frequency spectrum of the SNP markers reported here showed a bias toward loci with high-frequency minor alleles (i.e., high heterozygosity), compared with the unbiased neutral distribution. This ascertainment bias is common in SNPs and stems from the fact that SNPs are identified from a finite (often small) panel of individuals, which biases against the detection of rare alleles (Morin et al. 2004). The issue of ascertainment bias might be considered as a reason to favor microsatellites over SNPs when codominant markers are required. However, it should be noted that it is common practice to test new microsatellite markers on a panel of individuals and only use those loci deemed polymorphic in large scale population studies (e.g., Miller et al. 2000, Coates et al. 2005, Kim and Sappington 2005, Torres-Leguizamon et al. 2009). This practice will cause microsatellite data to be subject to ascertainment biases in a similar manner to SNP data.

The work presented here represents a relatively inexpensive route to obtaining genetic markers. We estimate that the total costs of consumables and services to collect insects, prepare and sequence normalized cDNA and test the chosen candidate SNPs were under US$10,000. Equally, the amount of hands-on labor required to identify and test SNPs was modest, amounting to a few person-weeks to collect insects, prepare the normalized cDNA for sequencing and prepare DNA for SNP validation of the Sequenom system. The use of core facilities for high throughput sequencing and genotyping significantly reduced the amount of labor on our part. By far the most time consuming elements of the work reported here were the assembly, annotation and associated bioinformatic analyses and the verification of the assembly by traditional Sanger sequencing of selected genes. The time required for the former of these activities is being mitigated by the steady development of faster, more user-friendly transcriptome assembly software, whereas the need for validation will likely decline as knowledge accumulates on strategies to produce reliable assemblies. Even accounting for the time required for assembly and validation, the approach to SNP discovery reported here is highly efficient compared with our previous experiences developing microsatellite markers and SNPs using Sanger sequencing.

In the last few years, there has been a rapid adoption of second generation sequencing as a tool for genetic marker discovery. A number of publications have demonstrated that sequencing cDNA is an effective strategy for discovering microsatellites (e.g., Mikheyev et al. 2010, Bai et al. 2010, Angeloni et al. 2011) and SNPs (e.g., Novaes et al. 2008, Bai et al. 2010, Crawford et al. 2010, Blanca et al. 2011, Angeloni et al. 2011). Thus far, transcriptome sequencing for marker discovery has mostly relied on 454 pyrosequencing, which has the advantage of comparatively long reads that facilitate contig assembly. Nevertheless, as we have shown here, the shorter-read, but much cheaper Illumina sequencing-by-synthesis technology also can provide transcript contigs that are adequate for both gene and molecular marker discovery.

Acknowledgments

We thank Bryony Bonning and Sijun Liu for their help and advice on preparing normalized cDNA. We also thank Steven Canon and Nathan Weeks for their advice on sequence assembly and SNP detection. Two anonymous reviewers provided valuable comments on an earlier version of the manuscript.

References Cited

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Angeloni, F., C.A.M. Wagemaker, M.S.M. Jetten, H.J.M. Op den Camp, E. M. Janssen-Megens, K.-J Francoijs, H. G. Stunnenberg, and N. J. Ouborg. 2011. De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. Mol. Ecol. Resour. 11: 662–674.

Bai, X., W. Zhang, L. Orantes, T.-H. Jun, O. Mittapalli, M.A.R. Mian, and A. P. Michel. 2010. Combining next-generation sequencing strategies for rapid molecular resource development from an invasive aphid species, *Aphis glycines.* PLoS ONE 5: e11370.

Bai, X., P. Mamidala, S. P. Rajarapu, S. C. Jones, and O. Mittapalli. 2011. Transcriptomics of the bed bug (*Cimex lectularius*). PLoS ONE 6: e16336.

Behura, S. K. 2006. Molecular marker systems in insects: current trends and future avenues. Mol. Ecol. 15: 3087–3113.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to mul-

tiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57: 289–300.

Blanca, J., J. Cañizares, C. Roig, P. Ziarsolo, F. Nuez, and B. Picó. 2011. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12: 104.

Bonin, A., D. Ehrich, and S. Manel. 2007. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol. Ecol. 16: 3737–3758.

Chaisson, M. J., and P. A. Pevzner. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18: 324–330.

Chen, S., P. Yang, F. Jiang, Y. Wei, Z. Ma, and L. Kang. 2010. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. PLoS ONE 5: e15633.

Coates, B. S., R. L. Hellmich, and L. C. Lewis. 2005. Polymorphic CA/GT and GA/CT microsatellite loci for *Ostrinia nubilalis* (Lepidoptera: Crambidae). Mol. Ecol. Notes 5: 10–12.

Coates, B., D. Sumerford, R. Hellmich, and L. Lewis. 2009. 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 52: 57–67.

Cock, P.J.A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Crawford, J. E., W. M. Guelbeogo, A. Sanou, A. Traoré, K. D. Vernick, N. Sagnon, and B. P. Lazzaro. 2010. De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-Seq technology. PLoS ONE 5: e14202.

Dorhout, D. L. 2007. Ecological and behavioral studies of the western bean cutworm (Lepidoptera: Noctuidae) in corn. M.S. thesis, Iowa State University, Ames, IA.

Dorhout, D. L., and M. E. Rice. 2010. Intraguild competition and enhanced survival of western bean cutworm (Lepidoptera: Noctuidae) on transgenic Cry1Ab (MON810) *Bacillus thuringiensis* corn. J. Econ. Entomol. 103: 54–62.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8: 195–202.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo, and A. Conesa. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420–3435 doi: http://dx.doi.org/10.1093/nar/gkn176.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Hagen, A. F. 1962. The biology and control of the western bean cutworm in dent corn in Nebraska. J. Econ. Entomol. 55: 628–631.

Hagen, A. F. 1963. Evaluation of populations and control of the western bean cutworm in field beans in Nebraska. J. Econ. Entomol. 56: 222–224.

Hartstack, A. W., J. A. Witz, and D. R. Buck. 1979. Moth traps for the tobacco budworm. J. Econ. Entomol. 72: 519–522.

Hoerner, J. L. 1948. The cutworm *Loxagrotis albicosta* on beans. J. Econ. Entomol. 41: 631–635.

Huang, X., and A. Madan. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9: 868–877.

Karatolos, N., Y. Pauchet, P. Wilkinson, R. Chauhan, I. Denholm, K. Gorman, D. R. Nelson, C. Bass, R. H. ffrench-Constant, and M. S. Williamson. 2011. Pyrosequencing the transcriptome of the greenhouse whitefly, *Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. BMC Genomics 12: 56.

Kim, K. S., and T. W. Sappington. 2005. Polymorphic microsatellite loci from the western corn rootworm (Insecta: Coleoptera: Chrysomelidae) and cross-amplification with other *Diabrotica* spp. Mol. Ecol. Notes 5: 115–117.

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–1858.

Lombaert, E., T. Guillemaud, J.-M. Cornuet, T. Malausa, B. Facon, and A. Estoup. 2010. Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. PLoS ONE 5: e9743.

Mailund, T., M. Schierup, C. Pedersen, P. Mechlenborg, J. Madsen, and L. Schauser. 2005. CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics 6: 252.

Marçon, P.C.R.G., D. B. Taylor, C. E. Mason, R. L. Hellmich, and B. D. Siegfried. 1999. Genetic similarity among pheromone and voltinism races of *Ostrinia nubilalis* (Hübner) (Lepidoptera: Crambidae). Insect Mol. Biol. 8: 213–221.

Meglécz, E., S. J. Anderson, D. Bourguet, R. Butcher, A. Caldas, A. Cassel-Lundhagen, A. C. d' Acier, D. A. Dawson, N. Faure, C. Fauvelot, et al. 2007. Microsatellite flanking region similarities among different loci within insect species. Insect Mol. Biol. 16: 175–185.

Mikheyev, A. S., T. Vo, B. Wee, M. C. Singer, and C. Parmesan. 2010. Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: performance and a comparison with AFLP-derived distances. PLoS ONE 5: e11212.

Miller, N. J., A. J. Birley, and G. M. Tatchell. 2000. Polymorphic microsatellite loci from the lettuce root aphid, *Pemphigus bursarius.* Mol. Ecol. 9: 1951–1952.

Miller, N., A. Estoup, S. Toepfer, D. Bourguet, L. Lapchin, S. Derridj, K. S. Kim, P. Reynaud, L. Furlan, and T. Guillemaud. 2005. Multiple transatlantic introductions of the western corn rootworm. Science 310: 992.

Miller, N. J., D. L. Dorhout, M. E. Rice, and T. W. Sappington. 2009a. Mitochondrial DNA variation and range expansion in western bean cutworm (Lepidoptera: Noctuidae): no evidence for a recent population bottleneck. Environ. Entomol. 38: 274–280.

Miller, N. J., T. Guillemaud, R. Giordano, B. D. Siegfried, M. E. Gray, L. J. Meinke, and T. W. Sappington. 2009b. Genes, gene flow and adaptation of *Diabrotica virgifera virgifera.* Agric For. Entomol. 11:47–60.

Morin, P. A., G. Luikart, R. K. Wayne, Fred W. Allendorf, Charles F. Aquadro, Tomas Axelsson, Mark Beaumont, Karen Chambers, Gregor Durstewitz, Thomas Mitchell-Olds, et al. 2004. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19: 208–216.

Novaes, E., D. Drost, W. Farmerie, G. Pappas, D. Grattapaglia, R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis,* an uncharacterized genome. BMC Genomics 9: 312.

O'Rourke, P. K., and W. D. Hutchinson. 2000. First report of the western bean cutworm, *Richia albicosta* (Smith) (Lepidoptera: Noctuidae) in Minnesota corn. J. Agric. Urban Entomol. 17: 213–217.

Reardon, B. J., D. V. Sumerford, and T. W. Sappington. 2006. Impact of trap design, windbreaks, and weather on captures of European corn borer (Lepidoptera: Crambidae) in pheromone-baited traps. J. Econ. Entomol. 99: 2002–2009.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16: 276–277.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, et al. 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7: 909–912.

Rousset, F. 2008. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol. Ecol. Resour. 8: 103–106.

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

Samarasekera, N. G., N. V. Bartell, B. S. Lindgren, J.E.K. Cooke, C. S. Davis, P.M.A. James, D. W. Coltman, K. E. Mock, and B. W. Murray. 2012. Spatial genetic structure of the mountain pine beetle (*Dendroctonus ponderosae*) outbreak in western Canada: historical patterns and contemporary dispersal. Mol. Ecol. 21: 2931–2948.

Simon, C., F. Frati, A. Beckenbach, B. Crespi, H. Liu, and P. Flook. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. 87: 651–701.

Su, P. P. 1976. Life cycle of *Nosema loxagrotidis* sp. N. (Microsporida: Nosematidae) and its development in *Loxagrotis albicosta* (Smith) (Lepidoptera: Noctuidae). M.S. thesis, University of Nebraska, Lincoln.

Suhr, E. L., D. J. O'Dowd, S. W. McKechnie, and D. A. Mackay. 2010. Genetic structure, behaviour and invasion history of the Argentine ant supercolony in Australia. Evol. Appl. 4: 471–484.

Tooker, J. F., and S. J. Fleischer. 2010. First report of western bean cutworm (*Striacosta albicosta*) in Pennsylvania. Crop management online: doi:10.1094/CM-2010-0616-01-RS.

Torres-Leguizamon, M., M. Solignac, D. Vautrin, C. Capdevielle-Dulac, S. Dupas, and J.-F. Silvain. 2009. Isolation and characterization of polymorphic microsatellites in the potato tuber moth *Tecia solanivora* (Povolny, 1973) (Lepidoptera: Gelechiidae). Mol. Ecol. Resour. 3: 102–104.

van Rossum, G., and J. de Boer. 1991. Interactively testing remote servers using the Python programming language. CWI Quarterly 4: 283.

Wang, X.-W., J.-B. Luan, J.-M. Li, Y.-Y. Bao, C.-X. Zhang, and S.-S. Liu. 2010. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11: 400–400.

Warren, R. L., G. G. Sutton, S.J.M. Jones, and R. A. Holt. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500–501.

Xue, J., Y.-Y. Bao, B.-ling Li, Y.-B. Cheng, Z.-Y. Peng, H. Liu, H.-J. Xu, Z.-R. Zhu, Y.-G. Lou, J.-A. Cheng, et al. 2010. Transcriptome analysis of the brown planthopper *Nilaparvata lugens.* PLoS ONE 5: e14233.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829.

Zhang, F., H. Guo, H. Zheng, T. Zhou, Y. Zhou, S. Wang, R. Fang, W. Qian, and X. Chen. 2010. Massively parallel pyrosequencing-based transcriptome analyses of small brown planthopper (*Laodelphax striatellus*), a vector insect transmitting rice stripe virus (RSV). BMC Genomics 11: 303–303.

Received 27 August 2011; accepted 24 May 2012.