

RESEARCH ARTICLE

Open Access

# Western corn rootworm (*Diabrotica virgifera virgifera*) transcriptome assembly and genomic analysis of population structure

Lex E Flagel<sup>1†</sup>, Raman Bansal<sup>2†</sup>, Randall A Kerstetter<sup>1</sup>, Mao Chen<sup>1</sup>, Matthew Carroll<sup>1</sup>, Ronald Flannagan<sup>1</sup>, Thomas Clark<sup>1</sup>, Barry S Goldman<sup>1\*</sup> and Andy P Michel<sup>2\*</sup>

## Abstract

**Background:** Western corn rootworm (WCR) is one of the most significant insect pests of maize in North America. WCR has dramatically increased its range in the last century, invading key maize production areas in the US and abroad. In addition, this species has a history of evolving traits that allow it to escape various control options. Improved genetic and genomic resources are crucial tools for understanding population history and the genetic basis of trait evolution. Here we produce and analyze a transcriptome assembly for WCR. We also perform whole genome population resequencing, and combine these resources to better understand the evolutionary history of WCR.

**Results:** The WCR transcriptome assembly presented here contains approximately 16,000 unigenes, many of which have high similarity to other insect species. Among these unigenes we found several gene families that have been implicated in insecticide resistance in other species. We also identified over 500,000 unigene based SNPs among 26 WCR populations. We used these SNPs to scan for outliers among the candidate genes, and to understand how population processes have shaped genetic variation in this species.

**Conclusions:** This study highlights the utility of transcriptomic and genomic resources as foundational tools for dealing with highly adaptive pest species. Using these tools we identified candidate gene families for insecticide resistance and reveal aspects of WCR population history in light of the species' recent range expansion.

**Keywords:** *Diabrotica virgifera virgifera*, Insect resistance, Transgenic crop, Transcriptome, Western corn rootworm

## Background

Western corn rootworm (*Diabrotica virgifera virgifera*) is a beetle (order Coleoptera) that is native to North America and the dominant maize pest in the US Corn Belt. It is estimated that each year Western Corn Rootworm (WCR) costs US farmers at least 1 billion dollars through yield losses and treatment costs [1,2]. WCR is native to North America, being first described in the western Kansas in 1867 [3]. Throughout the 20<sup>th</sup> century it spread eastward, colonizing the major maize production areas in the north central US (the Corn Belt) [3]. More recently it invaded

Europe [3,4], making it now a global pest of maize production.

The impact of WCR on maize production has been exacerbated by the fact that WCR has proved difficult to control. Over the last few decades WCR has evolved resistance to crop rotation [3,5,6], a range of chemical insecticides [7-10] and to transgenic maize expressing the Cry3Bb1 protein from *Bacillus thuringiensis* (Bt) [11]. The genetics behind some WCR resistance mechanisms have been investigated [12-14]. For example, a single nucleotide polymorphism (SNP) in the gamma-aminobutyric acid (GABA) receptor is correlated with cydodiene resistance [14]. However, for many of the most pressing resistance traits the genetic basis remains unknown.

Developing a genetic toolbox for WCR is an important step toward better understanding the genetic basis of

\* Correspondence: barry.s.goldman@monsanto.com; michel.70@osu.edu

†Equal contributors

<sup>1</sup>Monsanto Company, 700 Chesterfield Parkway W, Chesterfield, MO 63017, USA

<sup>2</sup>Department of Entomology, Ohio Agricultural Research and Development Center, 1680 Madison Ave., Wooster, OH 44691, USA

insecticide resistance in WCR. To this end, Siegfried et al. [15] produced 691 expressed sequence tags (ESTs) from WCR midgut tissues and screened them for candidate receptors for Bt. Among these ESTs, they reported one cadherin (a putative Bt receptor) and 15 cathepsin-like fragments (cysteine proteases) that are potential candidates for resistance to Bt and other toxins.

Studies in other insect species have suggested that cytochrome P450s (CYPs) are involved in resistance to several insecticides including methyl parathion, carbaryl, carbamate and organophosphates (Scharf et al. [16] and references therein; [17,18]). In WCR only 3 CYPs have been characterized, and all belong to the CYP4 clade, which is not typically associated with insecticide resistance [16]. Furthermore, a variety of other candidate genes for insecticide or Bt resistance are lacking for WCR, including esterases and other receptors [19,20]. A more complete understanding of the diversity of these important resistance associated gene families can lead to a more rapid characterization of current and future resistance mechanisms.

In addition to resistance, questions remain regarding WCR population structure and patterns of gene flow. Previous studies for North American WCR populations have used several types of molecular markers (allozymes, mtDNA sequencing, AFLPs, microsatellites, and SNPs), with all studies indicating large amounts of genetic diversity and a general lack of population structure across the Corn Belt [10,21-25]. These findings are consistent with the hypothesis that WCR has retained large amounts of genetic diversity during the species' rapid eastward range expansion. Whether or not this pattern is shared across the entire WCR genome or if certain areas contain less genetic diversity is unknown.

As the amount gene flow among populations largely dictates the spread of resistance alleles, a better understanding of the genomic variation across populations will lead to a more informed assessment of resistance monitoring and mitigation, and prolong the life of various control options. Modern DNA sequencing technologies have increased our ability to undertake large-scale sequencing projects for almost any organism, including those with complicated genomes such as WCR [26]. Therefore, our first objective for this study was to characterize the WCR transcriptome using publically available data as well as new data generated by 454 Roche pyrosequencing. We then mined this data to identify candidate genes in pathways predicted to be involved in insecticide resistance. Our second objective was to develop an expansive set of SNPs that can be used for more robust inferences of population structure, construction of linkage maps, and genomic scans of selection. To produce these SNPs we generated genomic sequences from multiple laboratory and field populations, including five populations that showed reduced response to Cry3Bb1 in lab-based single plant

bioassays. Furthermore, our ability to combine both transcriptomic and genomic data provide important information regarding genetic variation and level of selection among candidate resistance genes.

## Methods

### Sample collection, EST sequencing and assembly

For 454 EST sequencing WCR third-instar midgut tissues were isolated from 200 larvae from the Waterman population (Table 1) and stored in RNAlater (Life Technologies) prior to extraction. Total RNA extraction was performed on four sets of 50 midguts using the RNeasy mini kit (Qiagen). Approximately 500 µg of total RNA were used for two rounds of PolyA selection using the MicroPoly (A) Purist kit (Life Technologies). Following PolyA selection, cDNAs were synthesized using the SuperScript II cDNA synthesis kit (Life Technologies), using random hexamers for priming during first-strand synthesis. Finally, cDNAs were purified using a QIAquick PCR purification column (Qiagen).

Before 454-FLX sequencing, emulsion PCR reactions were conducted following the manufacturers specifications (Roche). In total, 1,271,800 454 EST reads were generated with a mean length of 297.6 bp, together totaling 378,442,184 bp of total sequence. These 454 sequences were supplemented with 17,833 Sanger sequences from NCBI (sequence IDs provided in Additional file 1) adding an additional 9,331,393 bp. The Sanger sequences derive from a variety of tissues and life cycle stages, including many previously published by Siegfried et al. [15].

Prior to assembly all 454 sequences were trimmed of their adaptors, while all Sanger reads were trimmed by searching against the NCBI UniVec database of common cloning vectors. 454 and Sanger sequences were merged and assembled with Newbler 2.3. All 454 ESTs can be found on the NCBI Sequence Read Archive (Accession SRS528965).

### Transcriptome annotation and characterization

The WCR transcriptome was annotated using Blast2GO [27]. BLASTx searches ( $e$  value  $< 10^{-3}$ ) were performed between WCR unigene sequences and the NCBI non-redundant (nr) database. Following the mapping step, gene ontology (GO) terms with  $e$  value  $< 10^{-6}$ , annotation cut-off  $> 55$ , and GO weight  $> 5$  were used for annotation. The "Aqua" classification in CateGORizer [28] was used to categorize the GO terms into different GO categories. The GO categories for WCR were compared to those from *Tribolium castaneum* (available at <http://www.b2gfar.org/showspecies?species=7070>). To find the pathways in which putative proteins of the WCR transcriptome are involved, analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) was performed using Blast2GO [29]. For comparative genomics, pairwise BLASTx searches ( $e$  value  $< 10^{-3}$ )

**Table 1 Population details for all 26 populations sampled in this study**

Population name	State	Year	Source	Diapause	Response to Cry3Bb1	$\pi$
Jones	IA	2011	Field	Yes	NA	0.0064
Plainview	MN	2011	Field	Yes	NA	0.0063
Gayville	SD	2011	Field	Yes	medium	0.0067
Humphrey	NE	2011	Field	Yes	high	0.0061
Sherman 2	CO	2011	Field	Yes	NA	0.0063
North Mankato	MN	2011	Field	Yes	medium	0.0065
Maynard	IA	2011	Field	Yes	NA	0.0064
Kit Carson 2	CO	2011	Field	Yes	high	0.0063
Onslow	IA	2011	Field	Yes	low	0.0066
Kit Carson 3	CO	2011	Field	Yes	medium	0.0065
Shell rock	IA	2011	Field	Yes	medium	0.0063
Trent	SD	2011	Field	Yes	low	0.0063
Adams	IN	2012	Field	Yes	high	0.0058
Fillmore	MN	2012	Field	Yes	NA	0.0057
Colfax	NE	2012	Field	Yes	NA	0.0059
Moody	SD	2012	Field	Yes	low	0.0057
Sherman 1	KS	2012	Field	Yes	medium	0.0058
Hansel	IA	2012	Field	Yes	NA	0.0059
Kit Carson 1	CO	2012	Field	Yes	NA	0.0059
Knox	IL	2012	Field	Yes	medium	0.0057
Hopkinton	IA	2009	Lab	No	low	0.0056
Brookings (random mated)	SD	NA	Lab	No	high	0.0055
Brookings (inbred)	SD	NA	Lab	No	high	0.005
Seneca resistant	KS	2005	Lab	No	low	0.0056
Seneca susceptible	KS	2005	Lab	No	NA	0.0056
Waterman	IL	NA	Lab	Yes	high	0.0063

For each population the state of origin, year collected, source, diapause status, response to Cry3Bb1, and intra-population genetic diversity ( $\pi$ ) are given. Diapause status indicates whether or not the population undergoes diapause. All non-diapause ("No") populations derive this phenotype from the Brookings, SD source. All field sourced populations diapause. Response to Cry3Bb1 was determined by larval recovery data in single plant bioassays under laboratory conditions. Each population was reared on Cry3Bb1 expressing (treatment) and isoline (control) maize. Populations with at least 80% relative survival on treatment when compared to control were defined as low response, while medium response populations ranged between 20% and 80%, and high response populations were less than 20% relative survival. Missing data is indicated by NA.

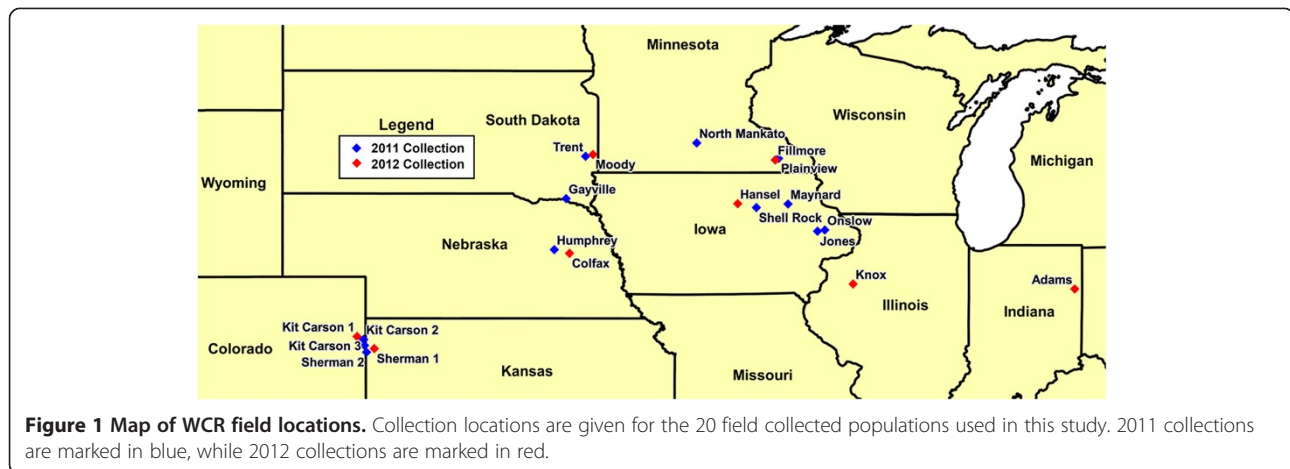
between WCR contig sequences and genomes of model insect species (*Bombyx mori*, *Drosophila melanogaster*, *Nasonia vitripennis*, *Tribolium castaneum*) were performed. Results of these BLASTx searches were also used to calculate the ortholog hit ratio ([30]; at OARDC MCIC galaxy: <http://www.oardc.ohiostate.edu/mcic/bioinformatics/bioinformatics.html>). The ortholog hit ratio was calculated by dividing the number of non-gap characters in the query (WCR unigene) by the length of the subject (model insect ortholog) [30].

#### Pooled population resequencing and SNP detection

Population genomic diversity was estimated by sampling 26 WCR populations (Figure 1; Table 1), 20 of which originated from field collections made in 2011 and 2012 (to protect the anonymity of the landowners only approximate

locations are given for field collections). Where available, we also list each population's phenotypic status for diapause and response to the Cry3Bb1 Bt toxin based on single plant bioassays under laboratory conditions [11] (M. Chen et al., in prep). The remaining 6 populations originated from both diapause and non-diapause laboratory colonies. Among these populations the Hopkinton and Seneca research colonies are resistant to Cry3Bb1 [11,31].

For each population we randomly selected 5 individuals (effectively 10 haploid chromosome sets) from mixed sex populations for Illumina sequencing (130 total individuals). All individuals within a population were combined prior to DNA extraction (5 individuals per pool). DNA was extracted using the E.Z.N.A. Insect DNA isolation kit (Omega Bio-Tek). Illumina sequencing libraries were prepared using the Nextera kit (Illumina) and sequenced on an



Illumina HiSeq 2000 genome sequencer in a  $2 \times 100$  bp paired-end configuration. After sequencing, barcode adapters were removed and all reads were quality trimmed from both ends to a minimum PHRED score of 20. All raw Illumina sequences used in this study can be found on the NCBI Sequence Read Archive (Accession SRP035322).

Paired-end Illumina genomic shotgun sequences from all populations were aligned to the combined transcriptome assembly described above using BWA (version 0.5.8; [32]), with default settings. Aligning genomic sequences to a reference transcriptome creates mismatches at exon/intron boundaries because the introns are sampled by the genomic sequences but typically spliced out of the transcripts. When BWA encounters these mismatches it issues a “soft-clip” flag [33]. To avoid SNP calling errors that might be associated with these mismatches, all reads were first filtered to remove any reads that had been soft-clipped. We also removed all reads with a “MAPQ” alignment score  $< 25$ . Finally, SNPs were extracted from these alignments using the *mpileup* program from the SAMtools package (version 0.1.18; [33]), at all sites with at least 5X coverage from nucleotides with a PHRED sequence quality  $\geq 30$ . All SNP calls can be found in Additional file 2.

To validate SNPs, we used the Cleaved Amplified Polymorphic Sequences (CAPS) technique [34]. Potential CAPS markers were identified *in silico* by screening alternate SNP haplotypes against the commercially available restriction enzyme database found in Biopython (version 1.60; [35]). We selected 27 SNPs that created a polymorphism in a recognition site for the enzyme *RsaI*. A region around each SNP was amplified and digested in up to 11 WCR individuals (average 9 per marker). Primer sequences can be found in Additional file 3. PCR reactions were performed in a 20  $\mu$ l mixture containing 1  $\mu$ l (50 ng/ $\mu$ l) DNA template, 10  $\mu$ l 2X PCR Promega master mix (with a final concentration of 0.4 mM each deoxynucleoside triphosphate, 1.5 mM MgCl<sub>2</sub> and 0.625 units of *Taq* DNA

polymerase in PCR reaction buffer pH 8.5) and 0.5 mM each primer. Amplified products for each sample were digested with 1  $\mu$ l (10 u/ $\mu$ l) *RsaI* at 37°C for 2 hrs. The resulting digest was visualized after electrophoresing in a 2% agarose gel stained with ethidium bromide.

#### Estimating population genetic parameters

We estimated nucleotide diversity within populations for each unigene using a bias-corrected estimator ( $\hat{\theta}_{\pi}^{(b)*}$ ; simply  $\pi$  throughout this manuscript) derived for pooled populations of individuals [36]. This estimator requires a minimum count ( $b$ ) for the inclusion of a particular SNP variant in the calculation. Here we used  $b = 2$  as our minimum. We also disregarded any sites with greater than  $> 40X$  coverage to minimize the inclusion of paralogous sequences or undetected copy number variants. Within each population mean  $\pi$  was approximately 0.006, however in each population we observed 16 to 30 outlier unigenes ( $\pi > 0.15$ ). These outliers were removed prior to the calculation of summary statistics, as these loci are likely impacted by alignment issues given their extreme departure from the mean. These calculations were done using a custom Python script (available on request).

We computed pairwise  $F_{ST}$  between all populations for each unigene using a bias-corrected estimator for pooled population samples as implemented in PoPoolation2 (version 1.201; [37]), again using a minimum coverage of 5 and a maximum coverage of 40 for each SNP locus. Window and step sizes were both set to  $10^9$ , forcing the program to estimate  $F_{ST}$  once for each unigene.

$F_{ST}$  is a measure of genetic differentiation relative to the net genetic diversity of the populations sampled, which makes  $F_{ST}$  sensitive to fluctuations in net genetic diversity. To complement  $F_{ST}$ , several authors have concluded that absolute metrics of genetic differentiation should also be employed [38,39]. For this study we created a simple statistic ( $\pi_{xy}$ ) to estimate absolute genetic differentiation

between populations using pooled sequence data. This statistic is an estimator of Nei's  $d_{XY}$  statistic [40], though unlike  $d_{XY}$  it uses estimated allele frequencies rather than allele counts, making it suitable for pooled sequence data. To estimate  $\pi_{xy}$  we first recorded a four element sequence profile  $b_1, b_2, b_3, b_4$ , for every nucleotide site within each pooled population, that contains the frequency of A, C, G, and T bases in the sample. Then for each unigene we calculated the average per-site expected heterozygosity among gametic unions between populations  $x$  and  $y$  (i.e.  $\pi_{xy}$ ), using the equation below.

$$\pi_{xy} = 1 - \left[ \sum_{i=1}^n \sum_{j=1}^4 x_{ib_j} y_{ib_j} \right] \cdot n^{-1}$$

Where  $n$  is the number of nucleotide sites available for a unigene,  $x_{ib_j}$  and  $y_{ib_j}$  are the frequencies of base  $b_j$  at nucleotide site  $i$  in populations  $x$  and  $y$ , respectively.  $\pi_{xy}$  requires only estimates of within population allele frequencies and not estimates of within population nucleotide diversity, unlike  $\pi$  and  $F_{ST}$ , and thus  $\pi_{xy}$  does not require bias-correction for pooled populations.

We applied the  $\pi_{xy}$  metric to calculate genetic distances between populations for principal coordinates analysis (i.e. classical multidimensional scaling). We also applied the  $\pi_{xy}$  metric to characterize the genetic divergence between combined populations with low or high responsiveness to Cry3Bb1 or diapause and non-diapause populations (Table 1). To estimate  $\pi_{xy}$  from combined populations we used the average of the nucleotide frequencies within each population to avoid biases that could arise due to uneven sequence coverage between populations. As with  $\pi$ , we required a minimum coverage of 5 and a maximum coverage of 40 within each population for a site to be considered. We also required a minimum count of 2 for the inclusion of a particular SNP variant.

For each unigene we identified  $\pi_{xy}$  and  $F_{ST}$  outliers by permuting population membership and calculating both metrics for permuted pairs. In the case of the Cry3Bb1 response phenotype there are  $\binom{11}{5} = 462$  possible permutations between high and low response populations, and all permutations were used to create the null distribution. For diapause versus non-diapause populations there are  $\binom{26}{5} = 65,780$  potential permutations, among which we randomly selected 1,000 to create the null distribution. In either case we selected the 99<sup>th</sup> percentile  $\pi_{xy}$  and  $F_{ST}$  values from among these permuted populations as our threshold for significance, after removing all unigenes with fewer than 50 available nucleotide sites. This threshold is arbitrary and lacks multiple test correction, though our goal is to produce a list of potential candidates that is inclusive, though it may include some false positives. Also, given the permutation tests outlined above, the

most confidence one can have regarding significance is 1/462 or 1/1,000, respectively.

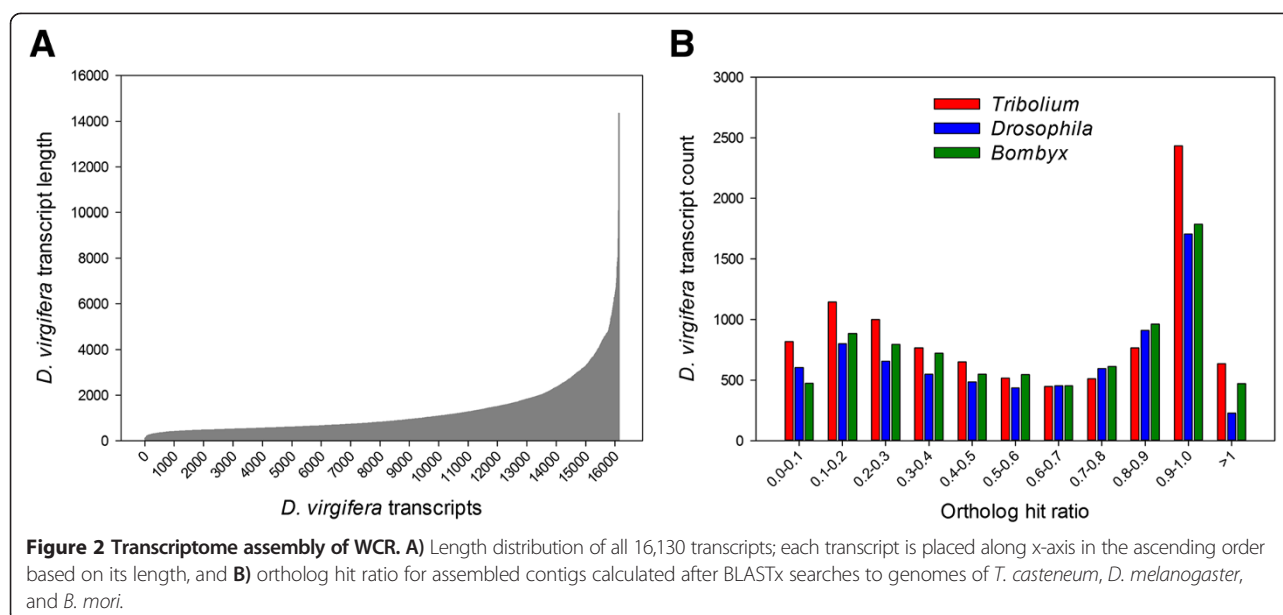
Our sampling procedure involved estimating population parameters from five individuals per population across approx. 10,000 loci. Theory suggests that multi-locus estimates of population genetic parameters can be robust even when estimated from just five individuals [41,42]. Though this theory applies specifically to direct sequencing of individuals and not pooled population sequencing, we have conducted a power analysis (Additional file 4) of our pooled sampling design, and it confirms that adequate estimates can be made from just five individuals, especially for statistics utilizing the grand mean across all loci. In light of this, our grand mean genomic estimates (i.e. mean within population  $\pi$  and between population  $F_{ST}$ ) should be quite accurate, despite the small within population sample size. Due to the small within population sample size ( $n = 5$ ), we expect our single locus estimates of divergence (i.e. the  $\pi_{xy}$  and  $F_{ST}$  outlier scans described above) to have relatively modest statistical power. These scans should be viewed as exploratory.

## Results and discussion

### De novo assembly and annotation

The *de novo* assembly of the WCR transcriptome yielded 16,130 high quality unigenes totaling 20,643,715 bp, in addition to 95,335 singleton reads. The length of unigenes varied from 99-14,348 bp with an average of 1,280 bp (Figure 2A). To determine the completeness of WCR transcriptome assembly, each unigene was compared to its putative ortholog from *T. castaneum*, *D. melanogaster*, and *B. mori*. Overall, 44-46% of the unigenes (with matches) had an ortholog hit ratio > 0.7 and 54-58% were > 0.5 (Figure 2B). On the assumption of conserved gene length between species, an ortholog hit ratio near zero indicates a poor assembly while values near one indicate a fully assembled transcriptome [30]. Following these criteria the current assembly for WCR appears to be moderately complete.

About 64% (10,359/16,130) of the WCR unigenes had one or more hits to protein sequences in the non-redundant (nr) database at GenBank while the remaining (36%) had no match (Additional file 5). The majority of top blast hits for WCR unigenes were to insects (~93%), whereas a small proportion showed top hits to non-insect arthropods (~1%), other eukaryotes (~5%), and bacteria (< 1%). As expected, more than two-thirds of the top hits for the WCR unigenes were to coleopteran sequences in the nr database such as those of *T. castaneum*, *Dendroctonus ponderosae*, *Chrysomela tremula*, *Leptinotarsa decemlineata*, *Phaedon cochleariae*, *Tenebrio molitor*, and *Batocera horsfieldi*. The WCR unigenes with no match to the nr database (unmatched) may represent novel genes with functions not yet been identified, or they could be a



product of assembly errors. Interestingly, the percentage of unmatched unigenes obtained in our study is considerably lower than those obtained for other insect transcriptomes [43-45]. An InterProScan search [46] of these unmatched WCR unigenes against the nr database revealed hits to the protein signature databases for 2,357 out of 5,762 (41%), suggesting that many have homologs in other species that were undetected using BLASTx alone. Further, 1,423 of the unmatched unigenes had an open reading frame (ORF) of at least 50% of the contig length and 302 had an ORF of at least 90% of the contig length. A pairwise comparison to the recently sequenced genome of mountain pine beetle [47] did not improve the annotation significantly as only 132 out of 5,762 unmatched WCR unigenes showed hits (data not shown).

### Comparative genomics

Using pairwise BLASTx searches to protein databases for four model insects, significant matches for a majority of WCR unigenes (10,356/16,130) were obtained. The BLASTx search to the *T. castaneum* database showed the highest number of matches for WCR unigenes ( $n = 9,693$ ), followed by searches to *N. vitripennis* ( $n = 8,981$ ), *B. mori* ( $n = 8,257$ ), and *D. melanogaster* ( $n = 7,422$ ) databases (Additional file 6). Several WCR unigenes ( $n = 6,767$ ) showed matches to all four databases. However, there were a substantial number of WCR unigenes that matched uniquely to *T. castaneum* ( $n = 754$ ), *N. vitripennis* ( $n = 288$ ) and *B. mori* ( $n = 130$ ). Analyzing the pairwise comparisons, the highest level of agreement was observed between WCR and *T. castaneum*, likely reflecting the evolutionary relatedness between these two insect species and the completeness of the *Tribolium* genome [48].

### Functional annotation

Using Blast2GO, about 46% of (7,377/16,130) WCR unigenes were annotated. The observed gene ontology (GO) terms for each domain (biological process, molecular function and cellular component) were widely distributed into different categories (Additional file 7). A comparison of percent mappings to each GO category in WCR and *T. castaneum* revealed a highly similar distribution for both insect species. The majority of contigs assigned to the 'biological process' domain were involved in cellular, regulatory, and developmental activities, while most of the contigs under the 'molecular function' domain were predicted to have catalytic, binding and transporter functions. Through the KEGG-based pathway analysis using Blast2GO, a total of 2,376 WCR unigenes were assigned to one or more of 117 pathways (Additional file 8). The majority of contig sequences were assigned to pathways for metabolism of organic compounds such as purine, pyrimidine, and glucose. The information on predicted pathways coupled with the functional annotation of the WCR transcriptome provides useful information for future genetic studies in this insect.

We further characterized potential genes that may be involved in WCR Bt resistance, specifically to Cry3Bb1 [11], focusing on detoxification and degradation (cytochrome P450s (CYPs), esterases, and cathepsins). We also identified several unigenes identified as putative receptors for Bt toxins [20,49-53] such as cadherins, ABC transporters, aminopeptidases and glycosyltransferases. (Note: glycosyltransferases are not technically themselves receptors, but are hypothesized to be responsible for glycosylation of lipids which functions as receptors [52].) Additional file 9 lists the WCR unigenes that significantly matched these

sequences in other insects, mainly from *Tribolium*. The number of putative receptors totaled 32, whereas the number of cathepsins, CYPs, and esterases were much larger ( $n = 98, 90,$  and  $70,$  respectively). These totals mark a considerable expansion beyond what was previously known for WCR candidate resistance gene targets. We further characterized two of the more important groups, the CYPs and cathepsins, because of their known relevance to resistance.

Cytochrome P450s (CYPs) are a superfamily of enzymes which are ubiquitous among organisms and play an important role in metabolizing endogenous and exogenous substances. In insects, CYP genes are classified into four clades: CYP2, CYP3, CYP4, and mitochondrial P450s [17]. Despite their importance for insecticide resistance in WCR and other insects [17,18], little information is known regarding the diversity or genetic variation of WCR-CYPs. In addition to BLASTx searches against the NCBI nr database, we identified the putative clades and families of WCR-CYPs through a BLASTp comparison to a recently published *Tribolium* CYP database [18]. The majority of WCR-CYPs had highest similarity to clade 3 ( $n = 65$ ), with a moderate number ( $n = 19$ ) showing similarity to clade 4 (Additional file 10).

The expanded 454 sequencing combined with previous resources show a considerable diversity in the number of cathepsins in WCR. We found a total of 98 unigenes that matched known cathepsins, including some with high similarity to those already described in WCR [15] (Additional file 9). A previous EST-based sequencing project [15], revealed 15 cathepsin-like fragments (a total of 171 sequences that assembled into 11 contigs and 4 singletons), corresponding to two different classes, L and B. In our transcriptome data, 18 unigenes belonged to class B and 55 matched class L; we also identified 8 unigenes in the lesser known D and F classes ( $N = 6$  and  $N = 2$ , respectively, Additional file 9). Cathepsins play a pivotal role in WCR digestion, and have been implicated in the evolution of rotation resistance. Normally, WCR feed and oviposit in corn, but rotation resistant WCR feed on soybean before oviposition. When these soybean fields are rotated to corn the following year, the rotation resistant WCR eggs hatch and larvae feed on emerging corn plants. Curzi et al. [13] found that WCR feeding on soybean leaves had higher levels of cathepsin-L like activity compared to WCR fed on corn. This expanded set of cathepsin-L like sequences and targeted resequencing of WCR populations could help enable the development of a molecular diagnostic for rotation resistant WCR.

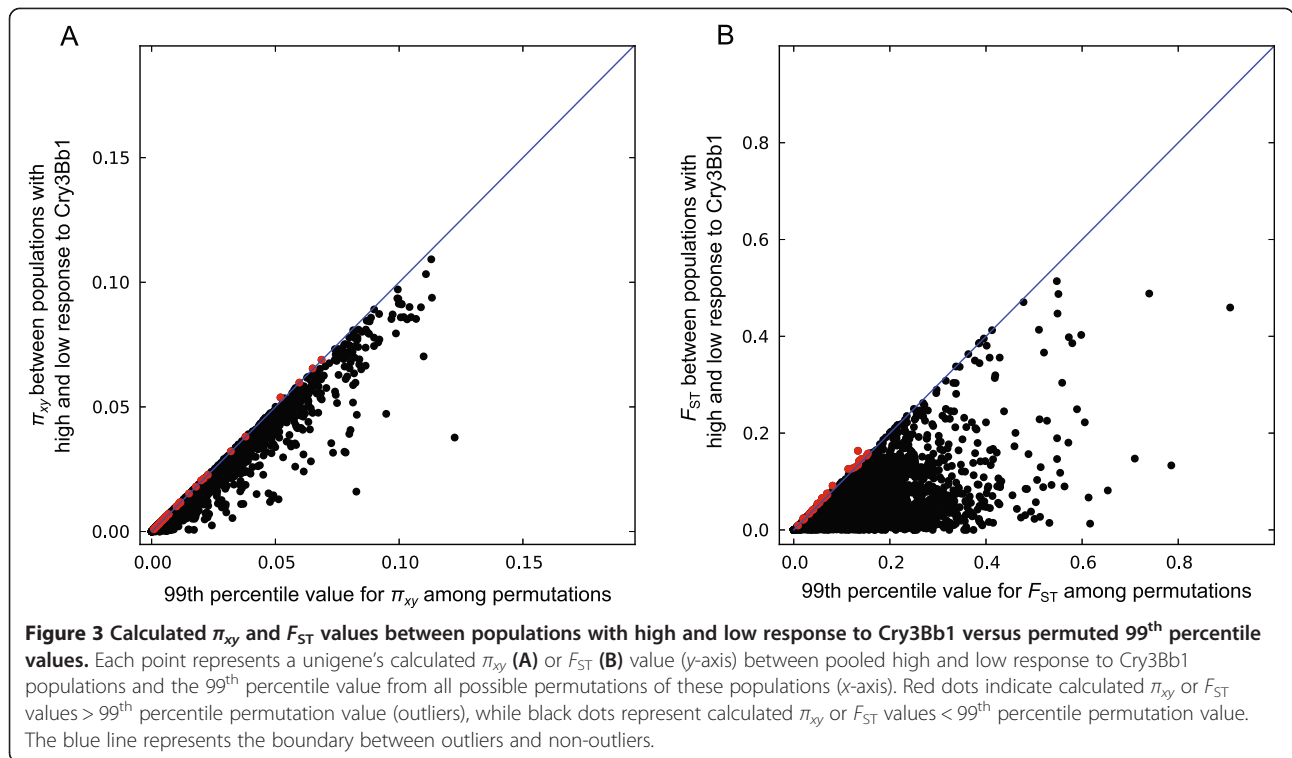
For all unigenes, including those that may be involved in WCR resistance to Cry3Bb1, we pooled SNP data from different populations with high or low response to Cry3Bb1 (Table 1) and screened for genes that were  $\pi_{xy}$  and  $F_{ST}$  outliers (see Methods). We identified 39  $\pi_{xy}$  outliers and

26  $F_{ST}$  outliers (Figure 3; Additional file 9). Only one unigene was found on both lists (unigene: DIAVI-09JAN12-CLUS09972\_1). This unigene shares sequence similarity with a gene of unknown function from *T. castaneum* (GenBank ID: XP\_976004). Among all 64  $\pi_{xy}$  and  $F_{ST}$  outlier unigenes we found two on our lists of candidates for genes potentially involved in reduced response to Cry3Bb1. These include one esterase (unigene: DIAVI-09JAN12-CLUS09401\_1) and one CYP (unigene: DIAVI-09JAN12-CLUS02697\_1; clade 3, CYP9), both of which were  $F_{ST}$  outliers. All outliers should be considered preliminary and exploratory as our tests are low-powered given the relatively small within population sample sizes, and because our strategy assumes that populations with a low response to Cry3Bb1 share resistance alleles, though there is no evidence supporting this assumption. Nonetheless, these loci may represent a promising pool of potential candidates for Cry3Bb1 resistance.

#### WCR SNP development and validation

The field collected populations (Figure 1) included a wide geography across the US Corn Belt, while the laboratory populations (Table 1) included several common research strains. Each population consisted of 5 individuals (effectively 10 haploid chromosome sets) that were pooled and sequenced to approximately 20X-30X coverage using whole genome shotgun sequencing. These sequences were aligned to the reference transcriptome for SNP discovery and estimation of population genetic parameters. In total we found 532,396 SNPs segregating among the 260 haploid chromosome sets sampled. SNP discovery saturates quickly among the WCR populations we sampled (Figure 4). After sampling 5 populations, few new SNPs are detected among the remaining populations. Those new SNPs that are found by adding more than 5 populations tend to be rare, often occurring in just one population. Thus, within our unigene collection, it is possible that our data has revealed most of the common SNPs in Corn Belt WCR.

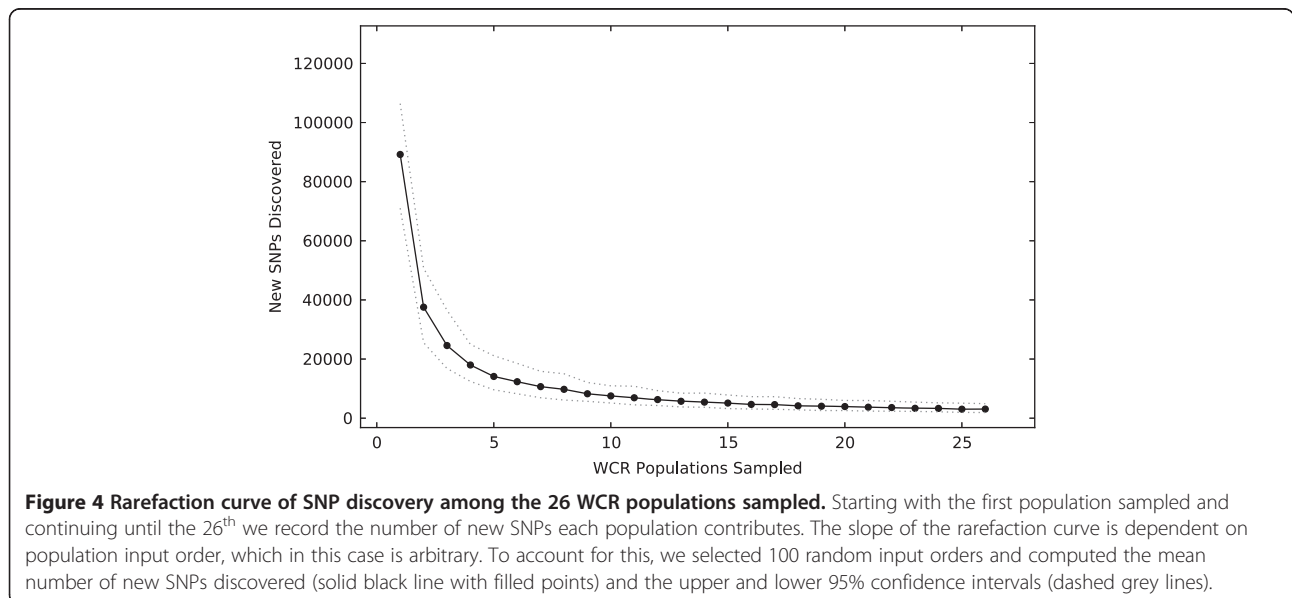
To address the validity of our SNP predictions and their utility for genotyping we identified 27 that were predicted to create a polymorphic restriction enzyme site and used this to genotype individuals from the inbred and randomly mated Brookings population using CAPs (see Methods). For 24 of the 27 SNPs the CAPs genotypes showed evidence of both allelic versions of the predicted SNP (Additional file 3). The remaining 3 SNPs were homozygous for one allele among all individuals tested, which could reflect inaccurate SNP predictions or simply the absence of one allele among the populations tested. In any case, these results demonstrate that the predicted SNPs are reliable and can be readily utilized in genotyping assays.



#### Population genomics of western corn rootworm

Using the SNPs extracted from pooled population sequencing we estimated genetic diversity within and between WCR populations. To estimate within population nucleotide diversity ( $\pi$ ) we used a bias-corrected estimator [36], suitable for pooled population sequences. For between population comparisons we created a simple statistic ( $\pi_{xy}$ ), which is an estimator of the  $d_{XY}$  statistic [40]. We focused our analyses on the 10,359 WCR

unigenes with a hit to a protein in the NCBI nr database (300,425 SNPs). For these unigenes, the mean  $\pi$  within each population can be found in Table 1. Nucleotide diversity within populations ranged from 0.005-0.007, with a mean of 0.006. As expected, the inbred Brookings colony (5 generations of full-sib mating; Chad Nielson pers. comm.) showed the lowest mean  $\pi$ , and the highest mean  $\pi$  values were found among field collected populations.



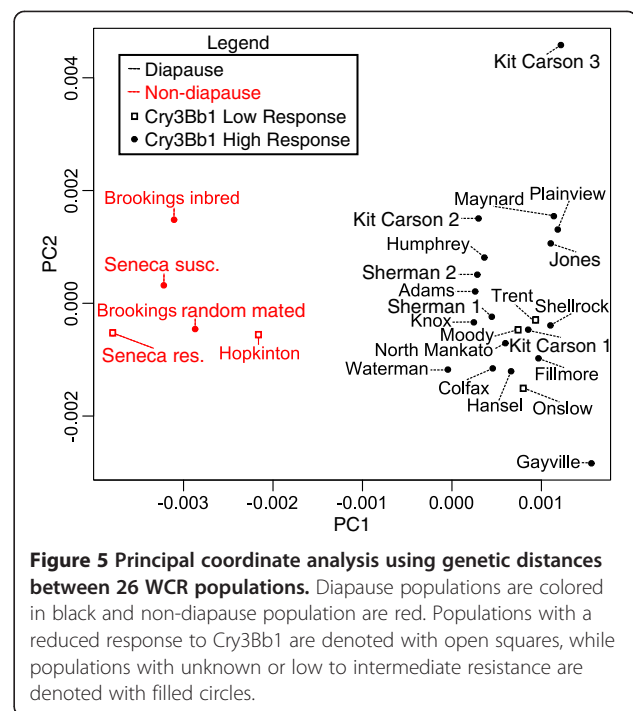


Our extensive SNP dataset reveals several interesting aspects of genetic variation in WCR. First, the inbred population had experienced 5 generations of full-sib mating, which is theoretically expected to fix approximately 41% of the sites segregating among the gametes of the founding sib-mated pair (Table 5.1 in Falconer and Mackay [54]), yet we observed only a slight reduction in  $\pi$  within the inbred Brookings population (mean  $\pi = 0.005$ ) compared to the randomly mated Brookings population (mean  $\pi = 0.0055$ ) from which it was derived. This suggests that full-sib mating has been fairly ineffective at removing variation. Second, there appears to be slightly less nucleotide diversity within the 2012 collections, when compared to the 2011 collections (t-test  $p$ -value < 0.001). This could reflect annual variation in population sizes across the Corn Belt, or perhaps some unnoticed deviation in our sampling procedures between these two years. Finally, there is little evidence for a significant reduction of genetic diversity among lab populations, suggesting that maintenance in artificial environments has not caused an erosion of genetic diversity. This finding mirrors observations by Kim et al. [55], who compared a laboratory population derived from the non-diapause Brookings population to field populations using microsatellite markers.

Among the 325 pairwise comparisons between populations,  $\pi_{xy}$  ranged between 0.0079 and 0.0097 with a mean of 0.0088. From this we observed that the levels of within population nucleotide diversity (mean  $\pi = 0.006$ ) are of a similar magnitude as between population divergence (mean  $\pi_{xy} = 0.0088$ ). This indicates that each population maintains considerable genetic variation, and that there is little evidence for substantial genetic differentiation between populations (see below).

Pairwise  $\pi_{xy}$  values were also used to generate a principal coordinates analysis among all populations. Diapause status appears to be the most important factor in population differentiation (Figure 5). The natural state for WCR is diapause, and the trait shared by all 5 non-diapause populations was inherited from a single source [56]. This non-diapause trait was developed after several generations of mass selection, and because of its complex inheritance, it has been suggested that it is likely a complex polygenic trait [56]. Consistent with this hypothesized mode of inheritance and strong artificial selection for its maintenance, we find 253 unigenes that are  $\pi_{xy}$  outliers when comparing all diapause versus all non-diapause populations (Additional file 11 and Additional file 12; see Methods). These outlier unigenes may include causal alleles for the non-diapause phenotype, but likely also include linked loci that have been differentiated between diapause and non-diapause populations by genetic hitchhiking.

To more formally address population differentiation we estimated  $F_{ST}$  pairwise between all populations for all 10,359 unigenes with a hit to the NCBI nr database.



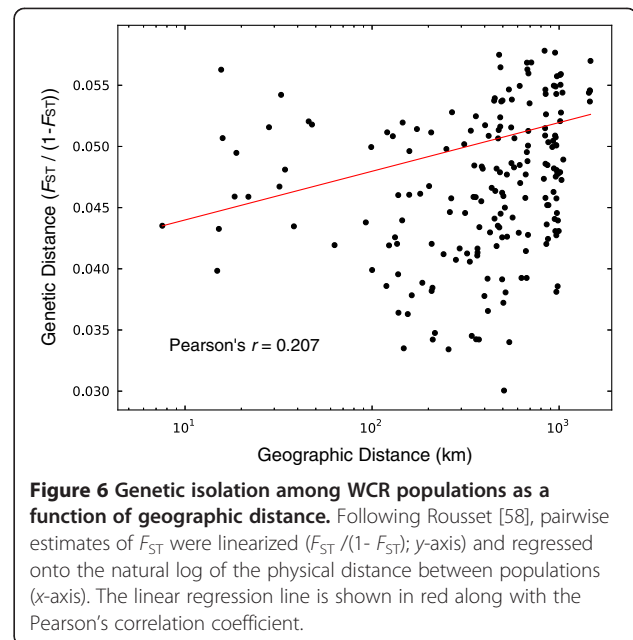
Mean  $F_{ST}$  between all populations across all unigenes was 0.052 (95<sup>th</sup> percentile bootstrap confidence intervals [0.051, 0.053]), and only 5.6% of all unigenes had  $F_{ST}$  values > 0.2 (Additional file 13). The overall pattern of low  $F_{ST}$  indicates little genetic differentiation among WCR populations, even across distances of approximately 1,500 km. These findings are consistent with earlier work in WCR, which also found low levels of genetic differentiation between field collected populations [22-25]. They are also consistent with our observation that new SNP discovery saturates fairly quickly as new Corn Belt populations are sampled (Figure 4).

Among all pairwise population contrasts, the highest  $F_{ST}$  values were found between field and lab populations, with nearly all of the most extreme values between a field population and the inbred Brookings population (data not shown). This observation is expected, as inbreeding impacts allele frequencies and will tend to increase  $F_{ST}$ . Interestingly, the smallest mean pairwise  $F_{ST}$  estimate was between the Onslow, IA and Trent, SD populations (mean  $F_{ST} = 0.029$ ; 95<sup>th</sup> percentile bootstrap confidence intervals [0.028, 0.030]). Despite being 508 km apart, these two field collected populations have both demonstrated reduced response to Cry3B1 in a lab-based single plant bioassays (Table 1). It is premature to conclude that these populations derive from a single Cry3Bb1 resistant source, though this intriguing possibility warrants further research.

In the last 75 years WCR have expanded from their native range in the US Southwest into the Corn Belt and

further eastward, recently arriving at the Eastern Seaboard [3]. Given this recent and rapid range expansion, it is not surprising that our results and previous efforts [22-25] reveal limited genetic differentiation among Corn Belt WCR populations, even across distances of approximately 1,500 km. Theoretical treatments of the genetic properties of range expansion typically find that populations near the edges of the range have reduced genetic diversity and increased  $F_{ST}$  when compared to populations near the center of the range, and that both of these features reset after colonization at a rate that is proportional to the migration rate between populations (reviewed by Excoffier et al. [57]). Among the Corn Belt WCR populations, we find high genetic variation and low  $F_{ST}$  values among the most recently colonized easterly populations sampled (Knox, IL and Adams, IN; likely colonized < 40 years ago [3]). These populations are now over 1,000 km west of the eastern edge of the species range, and our genetic observations are consistent with theoretical predictions, in that these populations are no longer in their colonizing phase and are likely gaining new genetic diversity through gene flow with more westerly populations. This pattern may also suggest that migration within the Corn Belt is fairly high; enough to spread considerable genetic diversity into populations founded in the last 40 years. Another interpretation could be that these two eastern populations were formed by a large and diverse set of founders, and that this diversity has been simply maintained since founding. Without historical samples it may be difficult to differentiate between these two hypotheses, though future studies could sample recently colonized populations in the eastern US to estimate typical founding population sizes.

We do observe a modest – albeit significant – trend of increasing genetic isolation as a function of geographic distance among field collected populations (Figure 6; Pearson's  $r = 0.207$ ; Mantel test  $p$ -value = 0.002). Thus, there may be a subtle signature of the range expansion in the form of genetic isolation by distance. Moreover, the significance of the isolation by distance is strongly influenced by the inclusion of the most easterly population (Adams, IN). If we remove this population, the correlation between genetic and physical distance falls to 0.106 and is no longer significant (Mantel test  $p$ -value = 0.101). In contrast, if we remove all 5 populations from eastern Colorado and western Kansas (the most westerly populations) the correlation between genetic and physical distance is still significant (Pearson's  $r = 0.256$ , Mantel test  $p$ -value = 0.003). The importance of this eastern population in detecting isolation by distance is consistent with incomplete homogenization following range expansion, however, because this result is driven by a single population we cannot exclude the hypothesis this population is in some way anomalous. Moreover, because recently formed Corn Belt WCR populations may not be at migration-drift



**Figure 6 Genetic isolation among WCR populations as a function of geographic distance.** Following Rousset [58], pairwise estimates of  $F_{ST}$  were linearized ( $F_{ST} / (1 - F_{ST})$ ; y-axis) and regressed onto the natural log of the physical distance between populations (x-axis). The linear regression line is shown in red along with the Pearson's correlation coefficient.

equilibrium we cannot determine if the weak pattern of genetic isolation by distance is a product of genetic drift overcoming gene flow or if it is a hold-over from population founder effects during the species eastward invasion. Based on theory and these preliminary results we predict that sampling populations further east will tend to increase the extent of detectable of isolation by distance in WCR, and may reveal important estimates of genetic diversity in recently colonized areas.

In a review article, McDonald and Linde [59] posit that knowing information about the population genetics of a pest organism can lead to a better understanding of its evolutionary potential to evade control strategies. They proposed that species with high levels of genetic diversity and high gene flow between populations have the greatest evolutionary potential to overcome control options through Darwinian evolution. In this study, and in earlier studies [18,21-25,55], we find evidence to support the hypothesis that WCR has both high levels of genetic diversity and high gene flow between populations. For example, assuming WCR has a mutation rate ( $\mu$ ) between  $10^{-8}$  or  $10^{-9}$  bp<sup>-1</sup> generation<sup>-1</sup> (a range that encompasses estimates from *D. melanogaster* and *Caenorhabditis elegans* [60,61]), the observed values of  $\pi$  imply that WCR field populations maintains approximate effective population sizes ( $N_e$ ) on the order of 0.2 to 2 million individuals (i.e.  $N_e \approx \pi / 4 \mu$ ). This level of  $N_e$  gives little indication of recurrent population bottlenecks in WCR, which gives WCR populations the ability to generate and maintain many individuals with novel mutations. This creates a large pool of standing genetic variation that may in turn increase the probability of evolving resistance to control strategies. Moreover, we show very little genetic differentiation between field

populations, and that genetic variation has spread quickly eastward following range expansion, both of which are consistent with high rates of gene flow between populations. Thus, if resistance does evolve, there are few barriers preventing it from spreading between populations. Together these two features of WCR population genetics are hypothesized to lead to a high evolutionary potential to evade controls [59], and may help to explain why WCR has a history of evolving resistance to insecticides and agricultural practices [3,10].

## Conclusions

WCR is one of the most significant agricultural pests in North America, yet at the time of publication WCR has a modest number of nucleotide sequences available in GenBank. Improved genetic and genomic resources for WCR are crucial tools in the ongoing effort to control this pest. Here we produce a large transcriptome assembly for WCR and complement these unigenes with diversity data obtained by whole genome population resequencing. Our WCR transcriptome assembly contains approximately 16,000 unigenes, including several gene families that have been implicated in insecticide resistance in other species. We identify a large pool of SNPs among 26 WCR populations, and show that these SNPs can be readily adapted for use as genotyping markers. We also use these SNPs to scan for outliers among candidate Bt resistance genes and to understand how population processes have shaped genetic variation in this species. We find two potential candidate Bt resistance genes that show strong patterns of molecular differentiation between high and low response to Cry3Bb1 populations. We also expand on past population genetic studies in WCR and show that this species has significant genetic diversity, little population structure, and a high evolutionary potential to resist control strategies. This study will provide a solid foundation for future research on the molecular genetics of WCR.

## Availability of supporting data

454 ESTs and Illumina pooled population sequences used in this study can be found on the NCBI Sequence Read Archive under the accessions SRS528965 and SRP035322, respectively.

## Additional files

**Additional file 1: GenBank IDs for public WCR EST sequences.** This spreadsheet contains the GenBank IDs for the public WCR sequences incorporated our transcriptome assembly.

**Additional file 2: Population level WCR SNP genotypes.** A comma separated values file (csv) containing the genotypic status of all 26 populations sampled at 532,396 SNP sites. Missing genotypes are denoted by a dash. The sequence of each unigenes is listed on first usage. This file has been compressed with the bzip2 compression software.

**Additional file 3: SNP genotyping validation study.** This spreadsheet contains the CAPs SNP validation information, including primer sequences, and observed restriction patterns and inferred genotypes.

**Additional file 4: Power analysis of population genetic sampling procedure.**

**Additional file 5: Blast2GO annotations.** This spreadsheet contains annotations for 10,359 WCR unigenes as identified by Blast2GO.

**Additional file 6: Comparative genomics of WCR.** A Venn diagram showing the number of WCR contigs with significant matches (unique and common) to genomes of *T. castaneum*, *D. melanogaster*, *N. vitripennis*, and *B. mori*. The significant matches ( $e$  value  $< 10^{-3}$ ) were calculated after pairwise comparisons (BLASTx) to each individual genome.

**Additional file 7: Distribution and comparison of GO categories.**

Vertical bars indicate the distribution of WCR (*Diabrotica*) and red flower beetle (*Tribolium*) GO term mappings that belong to each of the three top-level GO categories (i.e. biological process, molecular function, and cellular component).

**Additional file 8: KEGG pathway annotations.** This file contains two spreadsheets. The "Detailed" spreadsheet contains putative KEGG pathway assignments to unigenes in the WCR transcriptome, while the "Summary" spreadsheet lists a summary of the counts of all pathways identified.

**Additional file 9: Identification of putative candidate genes for Cry3Bb1 resistance.** This file contains a spreadsheet for each of several putative gene families that may be candidates for resistance to Bt toxins, including the unigenes name and descriptions, and results from  $\pi_{xy}$  and  $F_{ST}$  outlier scans between Cry3Bb1 resistant and susceptible populations. The last two spreadsheets list the results of  $\pi_{xy}$  and  $F_{ST}$  outlier scans between populations with high or low response to Cry3Bb1 for all available unigenes.

**Additional file 10: Characterization of WCR-CYP genes.** BLASTp result comparing WCR-CYP unigenes to a *Tribolium* CYP data set [18].

**Additional file 11: Identification of candidate genes for non-diapause trait.** This spreadsheet contains the results of  $\pi_{xy}$  outlier scans between diapause and non-diapause populations for all available unigenes. Along with the scan results, unigenes name and descriptions are also given.

**Additional file 12: Plot of calculated  $\pi_{xy}$  values between diapause and non-diapause populations and permuted 99<sup>th</sup> percentile values.** Each point represents a unigenes calculated  $\pi_{xy}$  value and the 99<sup>th</sup> percentile value from 1,000 permutations of the diapause and non-diapause population labels. Red dots indicate calculated  $\pi_{xy} > 99^{\text{th}}$  percentile permutation value (outliers), while black dots represent calculated  $\pi_{xy} < 99^{\text{th}}$  percentile permutation value. The blue line represents the boundary between outliers and non-outliers.

**Additional file 13: A histogram of pairwise  $F_{ST}$  between WCR populations.** This histogram represents  $F_{ST}$  estimates for all genes among all pairwise comparisons of the 26 WCR populations.

## Competing interests

LEF, RAK, M Chen, M Carroll, BSG, RF and TC are employed by Monsanto Company and some of the authors are also shareholders in Monsanto Company, though individually or collectively in no way represent any controlling interest in the affairs of Monsanto Company. Results described in the manuscript are the product of work financed by Monsanto Company, and the article processing charge will be paid by Monsanto Company. The authors have an obligation under an employment contract with Monsanto Company to assign the rights of any invention made in the course of their employment with Monsanto Company to the company. Monsanto Company has filed one or more patent applications disclosing and claiming all or some portion of the results of the work described in the manuscript. The authors are not known to have any financial competing interests in addition to those listed above.

## Authors' contributions

LEF, RB, RAK, M Chen, RF, TC, BSG, and APM conceived the study and contributed to experimental design. LEF, RB and APM analyzed the data and drafted the manuscript. M Chen and M Carroll contributed materials. All authors read and approved the final manuscript.

## Acknowledgments

Results described in the manuscript are the product of work financed by Monsanto Company. RB and APM received additional support from the Ohio Agricultural Research and Development Center, The Ohio State University. We thank Aaron Gassmann at Iowa State University, Bruce Hibbard at USDA ARS, Columbia, MO, and B. Wade French at the USDA North Central Agricultural Research Laboratory for sharing WCR colonies. We are grateful to John Cantwell, Douglas Jones, and Leanna Trail for their assistance in WCR field collections, Oliver Ilagan, Kaylee Kuban, Cara Vazquez for colony maintenance, and Nay Thane for producing the transcriptome assembly. Finally, we thank John LeDeaux and Peter Morrell for helpful comments and discussions.

Received: 15 October 2013 Accepted: 4 March 2014

Published: 14 March 2014

## References

- Chandler LD, Woodson WD, Ellsbury MM: **Corn rootworm IPM: implementation and information management with GIS.** In *Proceedings of the American Seed Trade Association 52nd Annual Corn and Soybean Research Conference: 1998; Chicago*. Edited by Wilkinson D 1998:129–143.
- Mitchell PD, Gray ME, Steffey KL: **A composed-error model for estimating pest-damage functions and the impact of the Western Corn Rootworm soybean variant in Illinois.** *Am J Agr Econ* 2004, **86**(2):332–344.
- Gray ME, Sappington TW, Miller NJ, Moeser J, Bohn MO: **Adaptation and invasiveness of Western Corn Rootworm: intensifying research on a worsening pest.** *Annu Rev Entomol* 2009, **54**(1):303–321.
- Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, Derridj S, Kim KS, Reynaud P, Furlan L, Guillemaud T: **Multiple transatlantic introductions of the Western Corn Rootworm.** *Science* 2005, **310**(5750):992.
- Krysan J, Foster D, Branson T, Ostlie K, Cranshaw W: **Two years before the hatch: rootworms adapt to crop rotation.** *Bull Entomol Soc Am* 1986, **32**:250–253.
- Sammons A, Edwards R, Bledsoe L, Boeve P, Stuart J: **Behavioral and feeding assays reveal a western corn rootworm (Coleoptera: Chrysomelidae) variant that is attracted to soybean.** *Environ Entomol* 1997, **26**:1336–1342.
- Zhu KY, Wilde GE, Higgins RA, Sloderbeck PE, Buschman LL, Shufrán RA, Whitworth RJ, Starkey SR, He F: **Evidence of evolving carbaryl resistance in western corn rootworm (Coleoptera: Chrysomelidae) in areawide-managed cornfields in north central Kansas.** *J Econ Entomol* 2001, **94**(4):929–934.
- Parimi S, Meinke LJ, French B, Chandler LD, Siegfried BD: **Stability and persistence of aldrin and methyl-parathion resistance in western corn rootworm (Coleoptera: Chrysomelidae).** *Crop Prot* 2006, **25**:269–274.
- Meinke LJ, Siegfried BD, Wright RJ, Chandler LD: **Adult susceptibility of Nebraska western corn rootworm (Coleoptera: Chrysomelidae) populations to selected insecticides.** *J Econ Entomol* 1998, **91**(3):594–600.
- Miller NJ, Guillemaud T, Giordano R, Siegfried BD, Gray ME, Meinke LJ, Sappington TW: **Genes, gene flow and adaptation of *Diabrotica virgifera virgifera*.** *Agric For Entomol* 2009, **11**(1):47–60.
- Gassmann AJ, Petzold-Maxwell JL, Keweshan RS, Dunbar MW: **Field-evolved resistance to Bt maize by Western Corn Rootworm.** *PLoS ONE* 2011, **6**(7):e22629.
- Chu C-C, Spencer JL, Curzi MJ, Zavala JA, Seufferheld MJ: **Gut bacteria facilitate adaptation to crop rotation in the western corn rootworm.** *Proc Natl Acad Sci USA* 2013, **110**:11917–11922.
- Curzi MJ, Zavala JA, Spencer JL, Seufferheld MJ: **Abnormally high digestive enzyme activity and gene expression explain the contemporary evolution of a *Diabrotica* biotype able to feed on soybeans.** *Ecol Evol* 2012, **2**:2005–2017.
- Wang H, Coates BS, Chen H, Sappington TW, Guillemaud T, Siegfried BD: **Role of a gamma-aminobutyric acid (GABA) receptor mutation in the evolution and spread of *Diabrotica virgifera virgifera* resistance to cyclodiene insecticides.** *Insect Mol Biol* 2013. Early Access; doi:10.1111/imb.12037.
- Siegfried BD, Waterfield N, French-Constant RH: **Expressed sequence tags from *Diabrotica virgifera virgifera* midgut identify a coleopteran cadherin and a diversity of cathepsins.** *Insect Mol Biol* 2005, **14**:137–143.
- Scharf ME, Parimi S, Meinke LJ, Chandler LD, Siegfried BD: **Expression and induction of three family 4 cytochrome P450 (CYP4) genes identified from insecticide-resistant and susceptible western corn rootworms, *Diabrotica virgifera virgifera*.** *Insect Mol Biol* 2001, **10**:139–146.
- Feyereisen R: **Evolution of insect P450.** *Biochem Soc Trans* 2006, **34**:1252–1255.
- Zhu F, Moural TW, Shah K, Palli SR: **Integrated analysis of cytochrome P450 gene superfamily in the red flour beetle, *Tribolium castaneum*.** *BMC Genomics* 2013, **14**:174.
- Zhu KY, Wilde GE, Sloderbeck PE, Buschman LL, Higgins RA, Whitworth RJ, Bowling RA, Starkey SR, He F: **Comparative susceptibility of western corn rootworm (Coleoptera: Chrysomelidae) adults to selected insecticides in Kansas.** *J Econ Entomol* 2005, **98**:2181–2187.
- Pigott CR, Ellar DJ: **Role of receptors in *Bacillus thuringiensis* crystal toxin activity.** *Microbiol Mol Biol Rev* 2007, **71**(2):255–281.
- Coates BS, Sumerford DV, Miller NJ, Kim KS, Sappington TW, Siegfried BD, Lewis LC: **Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis.** *J Hered* 2009, **100**(5):556–564.
- Miller NJ, Ciosi M, Sappington TW, Ratcliffe ST, Spencer JL, Guillemaud T: **Genome scan of *Diabrotica virgifera virgifera* for genetic variation associated with crop rotation tolerance.** *J Appl Entomol* 2007, **131**(6):378–385.
- Miller NJ, Kim KS, Ratcliffe ST, Estoup A, Bourguet D, Guillemaud T: **Absence of genetic divergence between Western Corn Rootworms (Coleoptera: Chrysomelidae) resistant and susceptible to control by crop rotation.** *J Econ Entomol* 2006, **99**(3):685–690.
- Szalanski A, Roehrdanz R, Taylor D, Chandler L: **Genetic variation in geographical populations of western and Mexican corn rootworm.** *Insect Mol Biol* 1999, **8**(4):519–525.
- Kim KS, Sappington TW: **Genetic structuring of Western Corn Rootworm (Coleoptera: Chrysomelidae) populations in the United States based on microsatellite loci analysis.** *Environ Entomol* 2005, **34**(2):494–503.
- Coates BS, Alves AP, Wang H, Walden KKO, French BW, Miller NJ, Abel CA, Robertson HM, Sappington TW, Siegfried BD: **Distribution of genes and repetitive elements in the *Diabrotica virgifera virgifera* genome estimated using BAC sequencing.** *J Biomed Biotechnol* 2012, **2012**:9.
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18):3674–3676.
- Hu Z, Bao J, Reecy JM: **CateGORizer: a web-based program to batch analyze gene ontology classification categories.** *Onl J Bioinform* 2008, **9**:108–112.
- Kanehisa M, Susumu G: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27–30.
- O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ: **Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*.** *BMC Genomics* 2010, **11**:310.
- Meihls LN, Higdon ML, Ellersieck MR, Tabashnik BE, Hibbard BE: **Greenhouse-selected resistance to Cry3Bb1-producing corn in three Western Corn Rootworm populations.** *PLoS ONE* 2012, **7**(12):e51055.
- Li H, Durbin R: **Fast and accurate long read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**:589–595.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPPD: **The sequence alignment/map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
- Konieczny A, Ausubel FM: **A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers.** *Plant J* 1993, **4**:403–410.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics* 2009, **25**(11):1422–1423.
- Futschik A, Schlötterer C: **The next generation of molecular markers from massively parallel sequencing of pooled DNA samples.** *Genetics* 2010, **186**(1):207–218.
- Kofler R, Pandey RV, Schlötterer C: **PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq).** *Bioinformatics* 2011, **27**:3435–3436.
- Noor MAF, Bennett SM: **Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species.** *Heredity* 2009, **103**:439–444.
- Charlesworth B: **Measures of divergence between populations and the effect of forces that reduce variability.** *Mol Biol Evol* 1998, **15**:538–543.
- Nei M, Kumar S: *Molecular Evolution and Phylogenetics*. New York: Oxford University Press; 2000.

41. Willing E, Dreyer C, van Oosterhout C: **Estimates of genetic differentiation measured by *Fst* do not necessarily require large sample sizes when using many SNP markers.** *PLoS ONE* 2012, **7**:e42649.
42. Pluzhnikov A, Donnelly P: **Optimal sequencing strategies for surveying molecular genetic diversity.** *Genetics* 1996, **144**:1247–1262.
43. Bai X, Zhang W, Orantes L, Jun T-H, Mittapalli O, Mian MAR, Michel AP: **Combining next-generation sequencing strategies for rapid molecular resource development from an invasive aphid species, *Aphis glycines*.** *PLoS ONE* 2010, **5**(6):e11370.
44. Chen Y, Cassone BJ, Bai X, Redinbaugh MG, Michel AP: **Transcriptome of the plant virus vector *Graminella nigrifrons*, and the molecular interactions of *Maize fine streak rhabdovirus* transmission.** *PLoS ONE* 2012, **7**(7):e40613.
45. Mittapalli O, Bai X, Mamidala P, Rajarapu SP, Bonello P, Herms DA: **Tissue-specific transcriptomics of the exotic invasive insect pest emerald ash borer (*Agrilus planipennis*).** *PLoS ONE* 2010, **5**(10):e13708.
46. Zdobnov EM, Apweiler R: **InterProScan - an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847–848.
47. Keeling CI, Yuen MM, Liao NY, Roderick Docking T, Chan SK, Taylor GA, Palmquist DL, Jackman SD, Nguyen A, Li M, Henderson H, Janes JK, Zhao Y, Pandoh P, Moore R, Sperling FAH, Huber DPW, Birol I, Jones SJM, Bohlmann J: **Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest.** *Genome Biol* 2013, **14**:R27.
48. Tribolium Genome Sequencing Consortium: **The genome of the model beetle and pest *Tribolium castaneum*.** *Nature* 2008, **452**(7190):949–955.
49. Contreras E, Schoppmeier M, Real MD, Rausell C: **Sodium solute symporter and cadherin proteins act as *Bacillus thuringiensis* Cry3Ba toxin functional receptors in *Tribolium castaneum*.** *J Biol Chem* 2013, **288**:18013–18021.
50. Fabrick J, Oppert C, Lorenzen MD, Morris K, Oppert B, Jurat-Fuentes JL: **A novel *Tenebrio molitor* cadherin is a functional receptor for *Bacillus thuringiensis* Cry3Aa toxin.** *J Biol Chem* 2009, **284**(27):18401–18410.
51. Gahan LJ, Pauchet Y, Vogel H, Heckel DG: **An ABC transporter mutation is correlated with insect resistance to *Bacillus thuringiensis* Cry1Ac toxin.** *PLoS Genet* 2010, **6**(12):e1001248.
52. Griffiths JS, Haslam SM, Yang T, Garczynski SF, Mulloy B, Morris H, Cremer PS, Dell A, Adang MJ, Aroian RV: **Glycolipids as receptors for *Bacillus thuringiensis* crystal toxin.** *Science* 2005, **307**(5711):922–925.
53. Atsumi S, Miyamoto K, Yamamoto K, Narukawa J, Kawai S, Sezutsu H, Kobayashi I, Uchino K, Tamura T, Mita K, Kadono-Okuda K, Wada S, Kanda K, Goldsmith MR, Noda H: **Single amino acid mutation in an ATP-binding cassette transporter gene causes resistance to Bt toxin Cry1Ab in the silkworm, *Bombyx mori*.** *Proc Natl Acad Sci* 2012, **109**(25):E1591–E1598.
54. Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*. 4th edition. Harlow, UK: Longman Group Ltd; 1996.
55. Kim KS, French BW, Sumerford DV, Sappington TW: **Genetic diversity in laboratory colonies of Western Corn Rootworm (Coleoptera: Chrysomelidae), including a nondiapause colony.** *Environ Entomol* 2007, **36**(3):637–645.
56. Branson T: **The selection of a non-diapause strain of *Diabrotica virgifera* (Coleoptera: Chrysomelidae).** *Entomol Exp Appl* 1976, **19**:148–154.
57. Excoffier L, Foll M, Petit R: **Genetic consequences of range expansion.** *Annu Rev Ecol Syst* 2009, **40**:481–501.
58. Rousset F: **Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance.** *Genetics* 1997, **145**:1219–1228.
59. McDonald BA, Linde C: **Pathogen population genetics, evolutionary potential, and durable resistance.** *Annu Rev Phytopathol* 2002, **40**:349–379.
60. Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledó JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, Baer CF: **A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes.** *Proc Natl Acad Sci* 2009, **106**:16310–16314.
61. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, Keightley PD: **Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*.** *Nature* 2007, **445**(7123):82–85.

doi:10.1186/1471-2164-15-195

**Cite this article as:** Flagel et al.: Western corn rootworm (*Diabrotica virgifera virgifera*) transcriptome assembly and genomic analysis of population structure. *BMC Genomics* 2014 **15**:195.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

