MOLECULAR ECOLOGY

Population structure and gene flow in the global pest, Helicoverpa armigera

Journal:	*Molecular Ecology*
Manuscript ID	MEC-16-0531
Manuscript Type:	Original Article
Date Submitted by the Author:	05-May-2016
Complete List of Authors:	Anderson, Craig; University of Stirling, BES; CSIRO Tay, Wee Tek; CSIRO Mcgaughran, Angela; CSIRO; University of Melbourne, School of BioSciences Gordon, Karl; CSIRO Walsh, Tom; CSIRO
Keywords:	Hybridization, Insects, Invasive Species, Population Genetics - Empirical, Ecological Genetics

Population structure and gene flow in the global pest, *Helicoverpa armigera*

Anderson, CJ. [1,2]*, Tay, WT.[2], McGaughran, A.[2,3], Gordon, K.[2], Walsh,TK.[2].

1. Biological and Environmental Sciences, University of Stirling, Stirling, FK9 4LA, UK.

2. CSIRO, Black Mountain Laboratories, Acton, ACT, 2601, Australia.

3. University of Melbourne, School of BioSciences, Melbourne, VIC, 3010, Australia.

Keywords: population genomics, gene flow, pest, moth, GBS

*Corresponding author. Email: Craig.Anderson@stir.ac.uk

Running title: Population structure in the pest, *H. armigera*

Abstract

Helicoverpa armigera is a major agricultural pest that presents a wide distribution across much of the Old World. This species is hypothesised to have spread to the New World 1.5 million years ago, founding a population that is at present a distinct species called *Helicoverpa zea*. In 2013, *H. armigera* was found to have re-entered South America via Brazil and subsequently spread throughout the continent. The source of the recent incursion is unknown and population structure in *H. armigera* is poorly resolved, but a basic understanding would highlight potential biosecurity failures and determine the recent evolutionary history of region specific lineages. Here, we integrate several end points derived from high-throughput sequencing to assess gene flow in *H. armigera* and *H. zea* from populations across six continents. We first assemble mitochondrial genomes to demonstrate the phylogenetic relationship of *H. armigera* with other Heliothine species, as well as the lack of distinction between populations. We subsequently use *de novo* genotyping by sequencing and whole genome sequences, aligned to bacterial artificial chromosomes, to assess levels of admixture. Primarily, we find that European individuals are most similar to Brazilian *H. armigera* and also identify a potential hybrid between *H. armigera* and *H. zea*. We also demonstrate the occurrence of an *H. armigera* subspecies that is generally endemic to Australia. While structure among the bulk of populations remains unresolved, we present distinctions that are pertinent to future investigations as well as to the biosecurity threat posed by *H. armigera*.

Introduction

Identifying population structure and patterns of gene flow in any species is often the basis for understanding the complexities of current and historical relationships (Martin *et al.* 2013; Nadachowska-Brzyska *et al.* 2013; Sankararaman *et al.* 2014). This information can have a number of important implications for conservation, management of invasive species and predicting the spread of novel phenotypes within a species distribution (Kirk *et al.* 2013; Prado-Martinez *et al.* 2013;

Malinsky *et al.* 2015). For example, one phenotype that can spread through populations is resistance to pesticides and understanding the population genetic factors underpinning this process can help to predict the spread and minimise damage imposed by pest species presenting this unwanted phenotype (Jin *et al.* 2015).

Helicoverpa armigera is one of the most significant pests of agriculture, with a wide range of suitable hosts and climatic conditions, rapid rates of reproduction, and a capacity for long distance dispersal across it's largely Old World distribution (Fitt 1989; McCaffery 1998; Feng *et al.* 2005). Recent modelling work has identified that the potential value of crops exposed to *H. armigera* totals approximately US $78 billion p.a. (Kriticos *et al.* 2015). Though *H. armigera* has typically been confined to the Old World (Europe, Africa, Asia and Australasia), it was identified in Brazil in 2013, before subsequently being identified as far north as Puerto Rico (2014) and Florida (2015) (Tay *et al.* 2013; Czepak *et al.* 2013; Hayden & Brambila 2015; Kriticos *et al.* 2015). The New World is typically the range of a closely related species, *Helicoverpa zea*, which is hypothesised to be the product of an ancient incursion by *H. armigera* approximately 1.5-2 million years ago (Behere *et al.* 2007). The two species are capable of hybridising in the laboratory (Hardwick 1965), however natural occurrences have yet to be recorded (Laster & Sheng 1995; Laster & Hardee 1995).

Previous work aiming to identify genetic variation with insight into population structure has lacked resolution in *H. armigera*. Mitochondrial markers are capable of determining variation at the species level and have been used, sometimes in tandem with genomic markers, in an attempt to distinguish populations. Taxonomic characterisation of two *H. armigera* strains has highlighted a potential distinction between Australian populations (*H. armigera conferta*) and the "Rest of the world" (*H. armigera armigera*) (Matthews 1999). However, much research (discussed in Behere *et al.* (2013)) has shown limited evidence of consistent structure on local or global scale.

Genetic variation derived from recent selection events can be useful for inferring population structure over regional scales and in turn, effectively allow for the inference of associated phenotypes. Where pesticides have been widely implemented to control *H. armigera*, resistance has been selected for and spread rapidly through populations in response to a broad range of treatments (Gunning *et al.* 2005; Yang *et al.* 2013; Tay *et al.* 2015). Of the many cases of resistance, that among populations of *H. armigera* to the pyrethroid, fenvalorate, has been particularly well studied. Introduced in the late 1970s, resistance to fenvalorate became established in Australia within six years and is now extremely common around the world (McCaffery 1998). The mechanism was identified as a chimeric P450, *CYP337B3* (Joußen *et al.* 2012), and subsequent analysis has identified a number of different haplotypes associated with geographic localities that imply independent evolutionary events and should be useful for inferring population structure (Walsh *et al.* submitted; Joußen *et al.* 2012; Rasool *et al.* 2014).

Recently developed molecular methods have been used to increase ability to identify population structure by massively increasing the number of markers available for analysis (Andrews *et al.* 2016). In particular, genotyping by sequencing (GBS) methods such as restriction associated DNA sequencing (RADseq), offer a powerful means for identifying genetic variation associated with specific phenotypes (Davey *et al.* 2011). These data, especially when mapped to a reference genome, have proven to be capable of identifying candidate genotype-phenotype associations and are underpinned by a series of increasingly sophisticated analytical tools (Patterson *et al.* 2006; Falush *et al.* 2007; Allendorf *et al.* 2010; Catchen *et al.* 2013b; Sousa & Hey 2013; Veeramah & Hammer 2014).

Given the technological developments within the field of molecular ecology, and in light of recent developments concerning the spread of *H. armigera*, we have used contemporary sequencing methods in an attempt to resolve *H. armigera* population structure and gene flow across the globe. We use high-throughput sequencing data for *H. armigera* and several other Heliothine species to confirm the phylogenetic relationship between species and populations through whole mitochondrial genome sequencing. We subsequently use *de novo* GBS to determine population structure and gene flow at a continental scale in *H. armigera* and *H. zea*. Finally, we use publically available bacterial artificial chromosome (BAC) sequences to determine possible signals of gene flow, population structure and other evolutionary processes that can be inferred from across 20 BACs (approximately 2.3 Mb of the genome), including the region where the locus involved in fenvalorate resistance, *CYP337B3*, is found. Overall, this work offers insights into the potential sources of recent *H. armigera* incursions into Brazil, as well as describing global population structure in *H. armigera* and evidence of hybridisation among *H. armigera* and *H. zea*.

Methods

Sample collection and DNA extraction

Heliothine moths, including *H. armigera*, were collected between 2007 and 2014 from 16 different countries around the world across various climatic zones and altitudes (Tables S1 and S2), many of which are described in Behere *et al.* (2007); and Tay *et al.* (2013). Samples were collected as larvae from wild and crop host plants, as adult moths via light/pheromone traps, or as larvae after bioassay, and preserved in ethanol (>95%) or RNAlater, or stored at -20°C prior to DNA extraction. DNA was extracted from samples using DNeasy blood and tissue kits (Qiagen), before being quantified with a Qubit 2.0.

Species Identification

The species status of several preserved specimens was confirmed by mitochondrial gene (COI and Cytb) sequencing, either from previous work (Walsh *et al.* submitted ; Behere *et al.* 2007; Tay *et al.* 2013,) or, where new samples were available, by amplifying and sequencing the same regions. PCR amplification followed the protocols of Behere et al. (2007) and Tay et al. (2013), using the primers

Harm-COI-F02/R02 and Harm-Cytb-F02/R02. PCR products were sequenced at Macrogen (Seoul, Korea) and the Biological Resources Facility (Australian National University, Canberra, Australia). Assembly of DNA trace sequences was performed using CLC Genomics Workbench v. 8.0 (www.clcbio.com).

Genotyping by Sequencing

GBS library preparation and sequencing was outsourced to Cornell University. Information regarding the samples used and sequencing output is recorded in the supplementary material (Table S1). Briefly, 50 ng of gDNA was digested using PstI, before being sequenced using an Illumina Hiseq. A negative control was included with each plate. Raw data were assessed for quality and processed using Stacks v. 1.30 (Catchen *et al.* 2013b). Briefly, process_radtags was used to demultiplex samples, trim to 90 bp and assess the quality of reads before being forwarded to denovo_map, which was run using default settings. The Populations module was then run, limiting the output to loci existing in at least 5% of the population with at least 5x coverage. The Populations module was used to output SNP data in Plink and Structure formats, the latter of which was limited to handling a single SNP per locus, chosen at random to account for linkage. Population level statistics were calculated using the Populations module and included pairwise F_{ST}, which was summarised using PCA conducted in Minitab v. 1.7 (www.minitab.com). Minitab was also used to calculate Pearson's correlation coefficient between principle components and missing data. For these analyses, two discrete samples of *H. zea* were defined based on their sample collection date being either putatively before or after the invasion of *H. armigera* (denoted as "Brazil zea" and "Brazil zea 2", respectively).

Whole Genome Sequencing

Nextera libraries were produced following the manufacturer's instructions and sequence was generated as 100 bp PE reads (Illumina HiSeq 2000, Biological Resources Facility, Australian National

University, Canberra, Australia, as well as at Beijing Genomics Institute, Hong Kong). Sample and sequencing data are included in the supplementary material.

Mitochondrial genome assembly and analysis

Raw sequence reads obtained from whole genome sequencing were aligned to the *H. armigera* mitochondrial genome using BBMap v. 33.43 (http://sourceforge.net/projects/bbmap/), permitting a minimum identity of 0.6 and allowing for a minimum quality threshold equivalent to Q10 over two consecutive bases before reads were trimmed. Reads were assembled using mira v. 4 (Chevreux *et al.* 2004) before mitobim v. 1.7 (Hahn *et al.* 2013) was used to iteratively map and assemble whole mitochondrial sequences. Heterozygous bases were removed, sequences were aligned using MAFFT v. 7.017 (Katoh 2002) and sequences were trimmed using the Gblocks v. 0.91b online server (http://molevol.cmima.csic.es/castresana/Gblocks_server.html) (Talavera & Castresana 2007). Statistical selection of nucleotide substitution models were estimated using jModelTest and the SYM model was implemented. A phylogenetic tree was then estimated using MrBayes v. 3.2.2 (Ronquist & Huelsenbeck 2003) via Geneious v. 8.1.7 (www.geneious.com), using a GTR substitution model as identified in jModelTest v. 2.1.7 (Posada 2008). Run parameters were: a chain length of 1,100,000 and a burn-in length of 100,000 over 4 heated-chains (chain temp 0.2). Pairwise nucleotide distance between species was measured using default parameters in MEGA v. 6.0 (Tamura *et al.* 2013). A haplotype network was calculated and generated using Popart v. 1.7 (Forster & Ro 1994; Leigh & Bryant 2015). Mitochondrial variants were then used to test for genetic divergence through calculation of pairwise differences for Φst (a measure related to F_{ST} used for haplotype data) using Arlequin v. 3.5.2.2 (Excoffier *et al.* 1992), with significance assessed using 20,000 permutations. Geographic structure among populations was also tested in Arlequin, using analysis of molecular variance (AMOVA).

Alignment and processing of whole genome sequencing data

Raw reads were aligned to BAC sequences, originally derived from *H. armigera* and available on NCBI (accessions in supplementary document), using BBMap. Reads were trimmed when quality in at least 2 bases fell below Q10. Only uniquely aligning reads were included in the analysis, to prevent spuriously inferring evolutionary processes occurring independently on each BAC. Outputted BAM files were sorted before duplicate reads were removed and files were annotated with read groups using Picard v. 1.138 (http://picard.sourceforge.net). BAC reference sequences were indexed using Samtools v. 1.1.0 (Li *et al.* 2009). UnifiedGenotyper in GATK v. 3.3-0 (McKenna *et al.* 2010) was used to estimate genotypes across all individuals simultaneously, implementing a heterozygosity value of 0.01. Variant call format files containing SNP calls were reformatted into Plink format using VCFtools v. 0.1.12b (Danecek *et al.* 2011). When linkage disequilibrium (LD)-based pruning was necessary, Plink v. 1.07 (Purcell *et al.* 2007) was used to filter one of a pair of SNPs using a pairwise LD threshold (r^2=0.5) within windows of 50 SNPs, moving forwards 5 SNPs per iteration.

Structure Analysis

The software, Structure v. 2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2007), was used on both GBS and BAC data to implement a model-based clustering method for inferring population structure. For all analyses, an initial run of 1,000 burn-in was followed by 1,000 repetitions of data collection, with K = 1 to estimate the allele frequency distribution (lambda), where K is the assumed number of populations. For the GBS data, 2,671 SNPs derived from all *H. armigera* and *H. zea* populations were tested across runs (20,000 burn-in, 20,000 data collection) implementing values of K from 1-5, with 8 replicates of each K-value, before Structure Harvester v. 0.6.94 (Earl & vonHoldt 2012) was used to identify the most appropriate value of K via implementation of the Evanno method (Evanno *et al.* 2005). The optimal K for each analysis was implemented in a final run (100,000 burn-in, 100,000 data collection) and results were plotted using Distruct2.pl (https://githum.com/crytpic0/distruct2). Separate analyses focusing on all *H. armigera* populations (6,713 loci), and then on *H. armigera* without individuals from Australia (6,868 loci) were also run, testing K in the ranges of 1-5 and 1-7, respectively. Structure analysis of BAC-aligned whole genome sequences was performed in a similar manner, but focused only on all populations of *H. armigera* to limit computational requirements, testing values of K between 1 and 7.

Principle component analysis and the D statistic

GBS and BAC-aligned data in Plink format was converted to eigenstrat format using the convertf module from EIGENSOFT v. 6.0.1 (Patterson *et al.* 2006). SNPs generated from GBS data were randomly attributed to one of 64 pseudo-chromosomes to obtain standard errors and calculate Z-scores. PCA was performed using the smartPCA module, implementing the LSQ option, with no automatic outlier removal allowed. A Tracy-Widom distribution was used to infer statistical significance for principle components, with a threshold of $1x10^{-12}$. Additionally, the “missing data” option was implemented to assess patterns of population structure where genotypes are present/ absent in the GBS data.

The *D* statistic is a measure of admixture between populations that is robust to biases associated with SNP ascertainment and demographic history, as well as the outlier incorporated (Patterson *et al.* 2012). The *D* statistic can be calculated using AdmixTools v. 3.0 (Patterson *et al.* 2012), where the estimation makes use of tree-like histories (explained in detail in Durand et al. 2011). In this instance, the program incorporates a structure of (Out group, *x*; *y*, *H. zea* (USA)), where *x* and *y* are combinations of *H. armigera* populations as defined by country of origin. Under the assumptions of the model, there is no assumed gene flow between the out group and *H. zea*, but there is potential gene flow allowable between either *x* and *y* or *y* and *H. zea*, which results in positive or negative *D*, respectively, with *D* = 0 indicating a lack of gene flow. Calculation of *D* uses biallelic SNPs and is accompanied by a Z-score that is considered to be significant when greater than three times its standard error. Individuals of *Helicoverpa punctigera* were included in the GBS sequencing and serve as an out group for this data, though *Helicoverpa assulta* was used as the out group for the resequencing data. The more recent divergence of *H. assulta* will result in capability for fulfilling the tree-like model across an increased number of SNPs. This test for admixture has been demonstrated to remain robust despite the use of increasingly distant out groups (Patterson *et al.* 2012).

Testing for selection across the BAC containing CYP337B3

Samples were screened for the presence of *CYP337B3* using the primers described in Joußen *et al*. (2012) and Walsh *et al*. (submitted). Heterozygote/homozygote status was determined through relevant band detection on 1.5-2% agarose gels containing 1% (w/v) of GelRed (Biotium) and visualised under UV light. Sanger sequences were generated from these short fragments for a subset of samples following the PCR amplification protocol of Joußen *et al*. (2012) and Walsh *et al*. (submitted).

VCFtools was used to calculate π and Tajima's D in *CYP337B3*-positive individuals, as defined via Sanger sequencing and PCA cluster membership, in sliding windows of 2,500 bp that progressed by 1,250 bp across biallelic sites. Results were plotted in R v. 3.1.2 (R Core Team 2014) using ggplot2 v. 1.0.1 (Wickham 2009), while gene annotations were derived via tblastx (Altschul *et al.* 1990) using default settings, and visualised with CLC Genomics workbench v. 8.0.

Results

Mitochondrial phylogeny

After aligning and removing sites with missing data, 12,248 bp of the 15,539 bp full length reference genome was used to infer phylogenetic relationships (Fig. 1). Phylogenetic analysis of the data demonstrates that we are clearly capable of determining species-level membership of various individuals included in this analysis, and faithfully follows the tree determined by Cho et al. (2008). Pairwise distances further demonstrate the relative difference between each of the species (Table

S3), with both *H. armigera* from Australia and the remaining global sample registering scores of 0.03 and 0.05 against *H. zea* and *H. puntigera*, respectively.

Inference of population structure and gene flow using mitochondrial data

Calculation of the whole mitochondrial genome haplotype network (Fig. 2) shows some clustering of Australian individuals in the left side of the network, however, the remaining populations are poorly resolved. For example, Australian individuals also appear closer to Asian and African samples throughout the network (Fig. 2). Inferences of gene flow between populations (i.e. Φst; Table S4) broadly support the haplotype network. Principle components 1 and 2 account for a total of 65.59% of the variation in the data and demonstrate differentiation between *H. armigera* from Australia and New Zealand from other populations (Fig. 3). In support of this, AMOVA implicated strong, significant genetic structure, with 74.15% ($P < 0.00001$) of the variance apportioned among populations and 25.85% apportioned within (Fig. S3).

Inference of population structure using de novo GBS

Principle coordinate analysis

Genotyping by sequencing (GBS) data from populations of *H. armigera*, *H. zea* and *H. punctigera* were used to determine if genetic variation from across the nuclear genome could provide insight into population structure and gene flow. Using data from 14,548 loci, 21,043 SNPs were used to reveal improved resolution in population structure and gene flow relative to mitochondrial data (Fig. 4). The first two eigenvalues are significant as inferred by the Tracy-Widom test ($P \leq 3.1 \times 10^{-23}$), with the greatest variation (6.61%) demonstrating a distinction between *H. armigera* and *H. zea*, while the second PC (1.89%), defines two discrete groups of *H. armigera*. This likely reflects *H. armigera* subspecies, *H. armigera armigera* (“Rest of World”) and *H. armigera conferta* (Australia), as described by Matthews (1999). *H. armigera armigera* populations cannot be resolved into populations using this data. Of note in Figure 4 is the presence of an individual collected in China clustering among Australian *H. armigera*, as well as an individual identified by mitochondrial markers as *H. zea* from Brazil that falls between the main cluster *of H. zea* and the two *H. armigera* clusters, potentially representing a hybrid between the two species. When presence/absence of markers was analysed using PCA, the first 6 PCs were found to be significant under the Tracy-Widom statistic ($P \leq 9.57x10^{-21}$, Fig. S1). Both PC1 and PC2 were negatively correlated with the amount of missing data in each sample, with Pearson's correlation coefficient as -0.56 and -0.82, respectively ($P < 0.0001$). Interrogation of subsequent PCs is suggestive of potential differentiation between populations, though there remains a degree of overlap between *H. armigera armigera* individuals.

Structure

Using SNPs from the GBS analysis, the number of genetic clusters inferred by eigenstrat support the Structure results (Fig. 5). Using the method established by Evanno et al. (2005), we determined that K=2 best fit the data and clearly defined the split between *H. armigera* and *H. zea*. Expected heterozygosity is a useful measure representing genetic diversity between individuals in the same cluster, and in this instance was greatest in the cluster relating most to *H. armigera* (red, 0.043) over that defining *H. zea* (blue, 0.020). The individual belonging to "Brazil zea 2" that was hypothesised to be a hybrid between *H. armigera* and *H. zea* (see above) was found to have the highest membership to the red cluster of any *H. zea* (0.367). K=2 was again selected following the removal of *H. zea*, but in this instance, identified variation between *H. armigera armigera* (i.e., "Rest of world") and *H. armigera conferta* (i.e., Australian *H. armigera*). Membership of Australian samples indicates a relatively large degree of gene flow (i.e., where red and blue colours are present in the same bar in the figure) between the two sub-species, though expected heterozygosity was greater for the cluster that best defined *H. armigera armigera* (red, 0.073) over that predominantly associated with *H. armigera conferta* (blue, 0.057). The Chinese individual found clustering with Australian *H. armigera* in Figure 4 has the second highest membership towards the red (*H. armigera armigera*) cluster than any other sample in this analysis. Subsequently, K=4 provided the largest delta K in an analysis of populations with high membership to *H. armigera armigera*, though no cluster clearly supports geographic partitioning. Expected heterozygosity is highest for cluster 4 (red) *H. armigera armigera* (0.085), which represents genotypes most frequently seen among populations from China, India and Uganda, at 91.2%, 83.6% and 75.6%, respectively. Brazilian individuals are most commonly found in cluster 3 (blue, 52.5%), but only 22% to cluster 4 (red). Only Brazil and Uganda are frequently found within cluster 2 (green).

Population genetic statistics

F_{ST} was calculated as corrected AMOVA F_{ST} in Stacks and is a pairwise measure across variable SNPs that does not consider fixed sites (Fig. S2, Table S5). PCA demonstrates that, across PC1 (87.2%), populations are clearly differentiated at a species level, with *H. armigera* from Brazil tending away from other populations of *H. armigera* and towards *H. zea*; this could reflect hybridisation or similar populations of origin for sampled individuals (Fig. S2). Of the other *H. armigera* populations, the Australian population has the lowest observable variation in allele frequency to Brazilian *H. armigera* as measured by F_{ST}. This is also emulated on PC2 (6.7%), though the reasons for the distribution across PC2 are not clear. *H. zea* sampled from before and after the incursion of *H. armigera* do not appear to differ extensively (Fig. S2).

Table 1 represents a summary of population genetic statistics derived from variant calls for samples analysed using GBS markers. Nucleotide diversity (π, equivalent to expected heterozygosity) is approximately the same in all *H. armigera* (including Brazil). In *H. zea*, π is lower compared to *H. armigera* but is similar for both Brazilian populations and the population from the USA. Observed heterozygosity is highest in Australia, and lower in Brazilian individuals, which present similar values to all populations of *H. zea*. A positive inbreeding coefficient (F_{IS}) identifies an excess in homozygosity that is indicative of genetic isolation of subpopulations by means of non-random mating or cryptic population structure and recent hybridization (Catchen *et al.* 2013a). In this instance, F_{IS} is low in *H. zea*, but higher in *H. armigera*, with the highest F_{IS} observable among Brazilian *H. armigera*. The most recent sampling of *H. zea* from Brazil has higher F_{IS}, comparable with levels seen in *H. armigera*, which might simply reflect the local variation. The number of private alleles, representing a basic measure of genetic distinctiveness in a population, is much higher in Brazilian *H. armigera*; a more geographically diverse global sample set might reduce this number, or it may be inflated due to the higher number of individuals sampled relative to other populations (Kalinowski 2004).

Measurement of gene flow using the D statistic

21,043 SNPs were incorporated into calculation of levels of gene flow between populations via the *D* statistic (Table 2). Overall, gene flow between Chinese and other populations of *H. armigera* is most commonly significant under a range of tree-like scenarios. For example, *D* is most negative (-0.259) under the model of (Out group, *x; y*, *H. zea* (USA)), when x is Australia and y is China, representing gene flow between Australia and China. *D* is also negative between Chinese and Brazilian populations of *H. armigera* (-0.270), though Ugandan and Indian populations appear to maintain similar levels of gene flow (*D*= -0.259 and -0.251, respectively) and appear to share similar levels of gene flow with Chinese populations themselves *(D*= -0.262 and -0.237, respectively). When considering gene flow between populations of *H. zea* and Brazilian *H. armigera*, there are no significant levels detected and both collections of Brazilian *H. zea* appear to be similar. *D* is only significantly positive between Brazilian *H. zea* and *H. zea* from USA, and is comparable in both instances tested. Furthermore, *D* is far higher in this instance than in any comparisons between *H. armigera* populations, which is likely reflective of the population founder event that this species is hypothesised to have undergone (Behere *et al.* 2007).

Inference of population structure using whole genome sequencing data aligned to BACs

Principle coordinate analysis

To observe the effects of increased resolution provided from alignment over BACs that accounted for a total of 2.3 Mb of the *H. armigera* genome, we conducted whole genome sequencing across a number of individuals representing several geographic locations (Table S2). Initially, we aligned reads to all BACs deposited on NCBI, before then looking at population structure derived from a BAC containing a chimeric P450 gene that is considered to be under selection (Joußen *et al.* 2012). Initial insights into population specific genetic variation in all BACs were provided by PCA (Fig. 6), though only PC1 was considered significant ($P=6.83x10^{-48}$) and primarily reflects the distinction between *H. armigera* and *H. zea*. A single outlier from China and two from India were manually removed from this analysis after falling far from all other samples, though having been previously identified as *H. armigera* using a mitochondrial marker. Several individuals from the geographically farthest relative sampling points (Brazil, Senegal and Europe) are placed farthest from a clearly discernible Australian cohort. When considering each of the BACs individually (supplementary document), this distribution is maintained across several BACs, including BACs 4, 8 and 18, which most clearly define this relationship. When SNPs from these three BACs are pruned to account for LD across the BACs, both PC1 and PC2 are to be considered significant, with PC2 accounting for variation across *H. armigera* populations ($P \leq 7.28x10^{-14}$).

Structure

Using the method established by Evanno et al. (2005), we determined that K=2 best fit data for analysis of population structure and defined the split between *H. armigera armigera* and *H. armigera conferta* (Fig. 7). Expected heterozygosity is approximately similar for both cluster 1 (red) and cluster 2 (blue) at 0.1167 and 0.1286, respectively. Australian individuals are most clearly identified in cluster 1, accounting for 84.2% of this population. While the majority of individuals from New Zealand and China are placed in geographically proximal clusters, some individuals are also found to have degrees of membership to cluster 1. K=4 also provided a notably high delta K, where certain clusters are associated with specific geographic regions. Expected heterozygosity is highest for cluster 3 (blue, 0.1392), which represents genotypes most frequently seen among *H. armigera*, but least frequently in Chinese, Ugandan and Indian populations (25.9%, 7.1% and 1.9% respectively). Cluster 1 has the next highest expected heterozygosity (yellow, 0.1358), and is most frequently found among 84.2% of Indian *H. armigera* (86%). Australia and New Zealand (52.4% and 30%, respectively) observe high degrees of membership to cluster 4 (red, expected heterozygosity is 0.0945), while European individuals have the greatest membership to cluster 2 (green, 76.5%). Brazilian individuals have the highest membership to cluster 3 (52.2%), but only 21.7% to cluster 1 and 26.1% to cluster 2 (Fig. 7).

Population genetic statistics

Values for nucleotide diversity and Tajima's D were calculated across BACs for all species sequenced (Table 3, Table S6). Nucleotide diversity, when considered in tandem with Tajima's D can be prescriptive of specific evolutionary scenarios, such as purifying selection (low heterozygosity, negative value of D) or a bottleneck (low heterozygosity and positive value of D), and gauges the frequency of variants under the neutral model of evolution (Kimura 1968). The highest value for nucleotide diversity across all BACs belongs to Australian *H. armigera* (0.018; in agreement with GBS data), and is closely followed by the remainder of the *H. armigera* populations, the lowest of which are of European origin (0.012). Generally, nucleotide diversity is higher among *H. armigera* populations than in other species, though the highest seen in other species belongs jointly to *H. zea* from Brazil, *H. assulta*, and *H. punctigera*, at 0.01 each, while the lowest is observed in *Heliothis virescens* (Table S6). In respect of this, Tajima's D is generally positive for all populations of *H. armigera*, except for Australian individuals. The only other instance whereby Tajima's D is negative is apparent in *H. punctigera*, which is endemic to Australia. The highest Tajima's D values belong to European *H. armigera* and *H. assulta*, at 0.32 and 0.47, respectively.

Measurement of gene flow using the D statistic

The highest level of *D* between *H. armigera* populations signifies gene flow between Australian and New Zealand populations (*D*= -0.204) (Table 4). The next highest is between Australia and China (*D*= -0.182). Brazilian populations present the lowest levels of gene flow (*D*= -0.138), and are followed by individuals representing *H. armigera* from Senegal (*D*= -0.149). When focusing on gene flow into or from Brazilian samples, *D* is highest for European samples (*D*= -0.180), with Senegalese *H. armigera* found to have the next greatest level of gene flow (*D*= -0.164), whereas populations from Australia and New Zealand have the lowest estimated levels (*D*= -0.138 and -0.143, respectively). Gene flow into *H. zea* from the USA is strongest from Brazilian *H. zea* (*D*= 0.595) and there is no evidence demonstrating greater levels of gene flow between *H. armigera* from Brazil into Brazilian *H. zea*, over *H. zea* from the USA.

Selection across the CYP337B3 BAC

Individuals genotyped for the haplotypes of the *CYP337B3* chimeric gene located on BAC 33J17 (JQ995292.1) formed clusters associated with the discrete selection events postulated to have occurred in the development of resistance to fenvalerate (Fig. 8). For example, clusters reflecting African, Asian and Australian origins for each of the haplotypes generally support geographical origins of samples, with the first 2 principle components considered significant via Tracy-Widom statistic ($P \leq 5.66 \times 10^{-21}$). This approach groups Brazilian samples with the Asian haplotype, while there is evidence that certain samples that are distinguishable from the most populous clusters are in fact heterozygotes, either in that they are heterozygous for the presence of the chimeric gene or are heterozygous for specific CYP337B3 haplotypes. This is supported by genotypes derived from Sanger sequencing, which were able to bin samples under haplotypes described in Walsh et al. (Submitted) (Table S7).

With regard to population genetic statistics across the BAC derived from clusters of individuals identified in eigenstrat that corroborate with Sanger sequencing, we were able to compare the evolutionary processes affecting genes associated with resistance to the pesticide fenvalerate (Fig. 9). On average, African samples had the most genetic diversity (0.026), in comparison to Asian (0.022) and Australian individuals (0.018). Extremely low nucleotide diversity was found proximal to the B3 gene (yellow bars in the figure legend), signifying selection at this site in all three populations. Tajima's D is variable across the BAC, with an average of 0.92 for African individuals, 0.41 for Asian and -0.46 for Australian populations. The lowest figure for Tajima's D was observed among Australian individuals (-2.17, 81.25 kb), which is the approximate region in which the B3 exons lie. Though this isn't apparent unless the size of the sliding windows is reduced to 1,250 bp (Fig. S4), Tajima's D is highest (2.86) at the same location in African samples, as well as towards the end of the BAC (Tajima's D= 2.04) and is likely the result of multiple alleles for this gene occurring in Africa. Subsequently, the next highest Tajima's D value occurs at 88.75 kb in Australian samples, only a short distance from B3. High Tajima's D in the region following from the B3 exons is likely the result of relatively high levels of nucleotide diversity seen in all populations, though only Asian and African populations see high nucleotide diversity preceding them.

Discussion

Source of the Brazilian incursion

Previous attempts have demonstrated that resolving population structure and gaining insight into gene flow between populations of *H. armigera* is difficult, even at a continental scale. Here, we have demonstrated that high-throughput sequencing methods are capable of resolving genetic variation associated with specific geographic localities. Specifically, we show that European *H. armigera* are most similar to Brazilian samples and we are unable to see a pattern in individuals collected in Africa and Asia. The first molecular identification of *H. armigera* in Brazil suggested that, even from a limited collection, samples were likely derived from a number of maternal lineages that are prevalent throughout the Old World (Tay *et al.* 2013). Subsequent work identified the distribution of variation at the B3 allele, which contributes towards fenvalerate resistance (Walsh *et al.* submitted), represented regional dominance of specific variants that were likely independently generated throughout Australasia, Africa and Asia. Of these, alleles found most frequently in Asia were seen in both African and Australian populations, but dominated those documented in Brazil, which also bore a single African allele.

We find that our data broadly agree with previous findings; variation across the BAC containing the *CYP337B3* gene demonstrates that all *H. armigera* sampled in Brazil bear the variant considered to have originated in Asian populations, with other individuals broadly sorted into the most prevalent geographic variants observed by Walsh et al. (submitted). While populations remain poorly resolved when considering variants mapped across other BACs, sophisticated analytical techniques have demonstrated that the greatest levels of gene flow occur between European and Brazilian *H. armigera*. Individuals sampled from Europe were also found to possess *CYP337B3* alleles most commonly found in Asia and Africa, likely the result of high degrees of admixture from these regions. This is spatially intuitive given that one would expect African and Asian populations to mix in this region and that the west coast of Africa, is the closest continental landmass to Brazil. The latter point is validated through the levels of gene flow inferred between Brazilian and Senegalese *H. armigera* (outputting the second highest value of *D* in these analyses). Showing that the incursion is likely sourced from a diverse population, has important implications in that the invasive population will possess greater genetic variation that will increase adaptive potential (Lavergne & Molofsky 2007). This is supported with levels of nucleotide diversity in the Brazilian individuals that are similar to other populations in both high-throughput data sets. We also see the highest levels of inbreeding coefficient, which, in tandem with high nucleotide diversity is synonymous with an incursion by a diverse population (Blackburn *et al.* 2015).

Population structure in H. armigera

Within our analyses, we are clearly able to distinguish two discrete populations of *H. armigera*. This reflects a distinction between *H. armigera armigera* and *H. armigera conferta* as has been previously identified by taxonomists (Hardwick 1965; Matthews 1999). *H. armigera conferta* has been considered to be restricted to Australasia, with individuals being distinguishable from the rest of the global *H. armigera* population in our analyses. Australian populations share the most genetic variation with individuals from New Zealand, and to a lesser extent, China. Notably, a single Chinese individual shares a large degree of genetic variation with Australian *H. armigera*. The distinction between subspecies is not as extreme in individuals from New Zealand and implies that gene flow from China to New Zealand is stronger than that into Australia. This may suggest differences in issues effecting biosecurity or perhaps that Australian populations are less susceptible to invasion. Tajima's *D* across BACs belonging to *H. armigera* from New Zealand is at a similar level to other well-established *H. armigera* populations, but is only negative in Australian individuals and *H. punctigera* and might highlight Australia-specific evolutionary processes that could be imposed by region-specific climactic events or agricultural practices. Analogous results are seen in an attempt to identify gene flow between populations of *H. armigera*, made by Song et al. (2015) who analysed 9 sex chromosome-linked EPIC markers, which use primers in adjacent exons to span intronic sequences. These authors found little population structure among geographic regions similar to those assessed in our study but at one locus in particular, they identified a distinction between Australian *H. armigera* and those from Africa and Asia and suggested that there is likely a significant degree of gene flow between Australian and Chinese populations. The fact that we see a Chinese individual bearing a greater resemblance to Australian *H. armigera* suggests that the distinction between *H.*

armigera armigera and *H. armigera conferta* may act as a potential model for hybridisation between Heliothine species in the New World. Primarily, the comparison offers an example of what can be expected when two genetically distinct populations from differing climates, subject to alternative agricultural practices and pest management, come into contact. Admixture between subspecies in the Asia-Pacific region could provide insight into the spread of resistance genes, the biosecurity measures that are able to restrict movement, and what evolutionary patterns may be expected in the New World. At the very least, through our data, we recognise gene flow between these regions and acknowledge that biosecurity would gain insight as to movements across this region with continued genotyping efforts.

While we are unable to resolve meaningful population structure between China, India and Uganda across GBS or whole-genome sequences, they remain relatively distinguishable from Brazilian populations, with particular insight demonstrable in the large number of private alleles and in patterns of missing genotypes observed in the GBS data. Missing data in GBS may be caused by an interruption of the recognition site of restriction enzymes, biases introduced during library preparation, sequencing biases, or inadequate coverage. While we feel it would be inappropriate to comment upon population structure using this missing data at this time, there may be signals of variation associated with specific geographic regions when the whole genome is analysed.

Interactions between populations of *H. armigera* over large spatial ranges as we've demonstrated are unsurprising. Many noctuid moths are, in fact, facultative migrants that are capable of using wind flow to take flight in response to environmental conditions (Bowden & Johnson 1976; Nibouche *et al.* 1998; Jones *et al.* 2015). Recent experiments using *H. armigera* have recorded that individuals are capable of flying between 20 and 40 km in a single night using tethered laboratory simulations (Jones *et al.* 2015). Within the mitochondrial genome tree presented here, we can see that *H. gelotopoeon*, found only in South America, shares a common ancestor with *H. punctigera*, whose distribution is restricted exclusively to Australia (Fitt 1989), thus insinuating the spread of noctuides across the Pacific Ocean, previously. Though a number of insect species are capable of similar feats of extensive migratory flight, many are unable to maintain population density in South America. This includes the Painted lady, *Vanessa cardui,* and the locust, *Locusta migratoria* (Rosenberg & Burt; Stefanescu *et al.* 2013), while a single incursion of the desert locust, *Schistocerca gregaria*, is likely the source of several contemporary species now found in the New World (Lovejoy *et al.* 2006). In an example most pertinent to the spread of *H. armigera* into the New World, monarch butterflies (*Danaus plexippus*), have dispersed across the Atlantic and Pacific oceans, as demonstrated by Zhan *et al*. (2014). These authors used whole genome sequencing of individuals from a number of populations along the species distribution to elaborate upon demographic history and were able to highlight candidate genes associated with migration and colour morphology. While the intercontinental spread of migratory species has been documented in several instances, it remains difficult to distinguish the natural spread from an anthropogenic cause, though geographic distances will play a great role in the interpretation. Further work using archived *H. armigera*, caught early on in the incursion, might therefore be useful for identifying not only the source of the incursion. This would perhaps define the basis for migratory performance while simultaneously defining the role of humans in the spread of H. armigera into the New World.

Hybridisation between H. armigera and H. zea

The single, clearest distinction that we're able to make throughout these analyses is that between *H. armigera* and *H. zea,* though this may become more difficult if *H. armigera* is able to successfully spread and hybridise throughout the New World. Even the clearly observable distinction made via mitochondrial markers may, in time, represent a series of haplogroups or be lost entirely, as is observable in modern humans, who show genomic evidence for admixture with Neanderthals though no mitochondrial haplogroups exist (Ghirotto *et al.* 2011; Sankararaman *et al.* 2012). It is possible that *H. zea* has facilitated the arrival of *H. armigera* into the New World, as we have presented putative evidence of a naturally occurring hybrid within our GBS data, which is characterised as having mitochondrial DNA originating from *H. zea*, but shares a considerable proportion of genomic DNA between genotypes clustering with both *H. armigera* and *H. zea*. Similar patterns were used recently to distinguish naturally occurring hybrids between golden jackals (*Canis aureus*) and domestic dogs (*Canis familiaris*) (Galov *et al.* 2015). The use of laboratory crosses to confirm patterns of heredity in the Galov *et al.* (2015) study greatly supported the author's inferences and as such, understanding the capability of *H. armigera* and *H. zea* to hybridise across multiple generations cannot be understated and should serve as a principle goal for future research.

Sequencing the genomes of these two species will provide not only an insight into the relative ancestry of individuals in the future, but will allow for interpretation of what makes *H. armigera* such a successful pest. Genomic analyses will also identify additional regions of the genome that are likely to be under selection among emerging populations of *H. armigera* and this will have implications as to the magnitude of the species as a pest in the New World. For example, using a range of methods that make use of high-throughput sequencing, it is possible to estimate the extent of linkage disequilibrium across the genome and estimate the likelihood of introgression with species-specific haplotypes as in works analysing humans and Neanderthals (Sankararaman *et al.* 2012, 2014). Therefore, the *H. armigera/H. zea* model has the potential to act, not only as an exemplary evolutionary model for incursive and introgressive processes, but also as an important indicator of susceptibility in global biosecurity. Indeed, recent work in monarch butterflies made use of whole-genome sequencing combined with phenotypes to provide insights into migratory behaviour and morphology (Zhan *et al.* 2014). A similar approach would improve the power of such analyses in pest species, although gaining phenotypic data relevant to resistance across a global sampling effort remains a logistical problem. Through the use of whole-genome sequencing, and based upon a wealth of previous work into the bases of resistance, we've been able to show that tight regions around the *CYP337B3* gene are under selection in all of the haplogroups assessed. This likely reflects that the populations are under selection and represents the value of monitoring populations so as to infer evolutionary processes. For example, if the population has recently undergone a bottleneck, then the impact of measures to control for pesticide resistance might have longer lasting effects if implemented in this population rather than a population undergoing purifying selection. Genes associated with resistance to pesticides play a definitive role in monitoring gene flow, are the most relevant for managing agricultural practices, and will play a key role in observing interactions between *H. armigera* and *H. zea* in the New World.

Conclusion

For the majority of *H. armigera* populations assessed, population structure remains unclear following interrogation with a number of analyses based upon high-throughput sequencing. Though we were unable to resolve populations with more than 12 kb of the mitochondrial genome, we are able to make a number of distinctions that are clearly important to the future of research in this area. Primarily, we are able to suggest that Brazilian *H. armigera* likely originated from Europe or West Africa, based upon whole genome sequences aligned to BACs. Using the same data, we were able to distinguish a subspecies of *H. armigera* generally endemic to Australasia and distinct from other populations. This inference was discernible in *de novo* GBS data, and was also apparent when a gene associated with resistance to fenvalorate was examined. Further analyses demonstrated that this gene is likely under selection and supports the perspective that agriculturally relevant genes can be used to not only monitor the spread of resistance but also to differentiate populations. Importantly, we highlight a potential example of natural hybridisation between *H. armigera* and *H. zea*, which sets a strong precedent for future research in establishing the capability of these species to hybridise. The end points presented here supply a series of provocative insights into the recent evolutionary history of a destructive pest species, but superior insight will only become apparent as genomic resources for these species become available.

Acknowledgments

We thank Andreas Zwick for providing the *Helicoverpa hardwickii* for these analyses. This work was funded under the CSIRO OCE postdoctoral scheme.

Author Contributions

CA and TKW wrote the manuscript. WTT and TKW organised sample collection. CA, WTT and TKW prepared DNA, made libraries and analysed the data. All authors provided intellectual input and contributed to organising the work. All authors edited the manuscript and endorse its submission.

References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews. Genetics*, **11**, 697–709.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–10.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

Behere GT, Tay WT, Russell D a *et al.* (2007) Mitochondrial DNA analysis of field populations of Helicoverpa armigera (Lepidoptera: Noctuidae) and of its relationship to H. zea. *BMC evolutionary biology*, **7**, 117.

Behere GT, Tay WT, Russell DA, Kranthi KR, Batterham P (2013) Population genetic structure of the cotton bollworm Helicoverpa armigera (Hübner) (Lepidoptera: Noctuidae) in India as inferred from EPIC-PCR DNA markers. *PloS one*, **8**, e53448.

Blackburn TM, Lockwood JL, Cassey P (2015) The influence of numbers on invasion success. *Molecular ecology*, **24**, 1942–53.

Bowden J, Johnson CG (1976) Migrating and other terrestrial insects at sea. In: *Marine Insects* (ed Cheng L), pp. 97–118. North-Holland Publishing Company, Oxford.

Catchen J, Bassham S, Wilson T *et al.* (2013a) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular ecology*, **22**, 2864–83.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013b) Stacks: an analysis tool set for population genomics. *Molecular ecology*, **22**, 3124–40.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome research*, **14**, 1147–59.

Cho S, Mitchel A, Mitter C *et al.* (2008) Molecular phylogenetics of heliothine moths (Lepidoptera: Noctuidae: Heliothinae), with comments on the evolution of host range and pest status. *Systematic Entomology*, **33**, 581–594.

Czepak C, Albernaz KC, Vivan LM, Guimarães HO, Carvalhais T (2013) First reported occurrence of Helicoverpa armigera (Hubner) (Lepidoptera: Noctuidae) in Brazil. *Pesquisa Agropecuaria Tropical*, **43**, 110–113.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, **27**, 2156–8.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, **12**, 499–510.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, **28**, 2239–2252.

Earl D a., vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, **14**, 2611–20.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes*, **7**, 574–578.

Feng HQ, Wu KM, Ni YX, Cheng DF, Guo YY (2005) High-altitude windborne transport of Helicoverpa armigera (Lepidoptera : Noctuidae) in mid-summer in northern China. *Journal of Insect Behavior*, **18**, 335–349.

Fitt GP (1989) The Ecology of Heliothis Species in Relation to Agroecosystems. *Annual Reveviews Entomology*, **34**, 17–52.

Forster P, Ro A (1994) Median-Joining Networks for Inferring Intraspecific Phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Galov A, Fabbri E, Caniglia R *et al.* (2015) First evidence of hybridization between golden jackal (Canis aureus) and domestic dog (Canis familiaris) as revealed by genetic markers. *Royal Society Open Science*, **2**, 150450.

Ghirotto S, Tassi F, Benazzo A, Barbujani G (2011) No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans. *American journal of physical anthropology*, **146**, 242–52.

Gunning R V, Dang HT, Kemp FC, Nicholson IC, Moores GD (2005) New resistance mechanism in Helicoverpa armigera threatens transgenic crops expressing Bacillus thuringiensis Cry1Ac toxin. *Applied and environmental microbiology*, **71**, 2558–63.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. *Nucleic acids research*, **41**, e129.

Hardwick DF (1965) The Corn Earworm Complex. *Memoirs of the Entomological Society of Canada*, **97**, 1–247.

Hayden J, Brambila J (2015) *Florida Department of Agriculture and Consumer Services Division of Plant Industry*.

Jin L, Zhang H, Lu Y *et al.* (2015) Large-scale test of the natural refuge strategy for delaying insect resistance to transgenic Bt crops. *Nature biotechnology*, **33**, 169–74.

Jones CM, Papanicolaou A, Mironidis GK *et al.* (2015) Genomewide transcriptional signatures of migratory flight activity in a globally invasive insect pest. *Molecular ecology*, **24**, 4901–11.

Joußen N, Agnolet S, Lorenz S *et al.* (2012) Resistance of Australian Helicoverpa armigera to fenvalerate is due to the chimeric P450 enzyme CYP337B3. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 15206–11.

Kalinowski ST (2004) Counting Alleles with Rarefaction: Private Alleles and Hierarchical Sampling Designs. *Conservation Genetics*, **5**, 539–543.

Katoh K (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kimura M (1968) Evolutionary Rate at the Molecular Level. *Nature*, **217**, 624–626.

Kirk H, Dorn S, Mazzi D (2013) Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evolutionary Applications*, **6**, 842–856.

Kriticos DJ, Ota N, Hutchison WD *et al.* (2015) The potential distribution of invading Helicoverpa armigera in North America: is it just a matter of time? *PloS one*, **10**, e0119618.

Laster ML, Hardee DD (1995) Intermating Compatibility Between North American Helicoverpa zea and Heliothis armigera (Lepidoptera: Noctuidae) from Russia. *Journal of Economic Entomology*, **88**, 77–80.

Laster M, Sheng C (1995) Search for Hybrid Sterility for Helicoverpa-Zea in Crosses between the North-American Heliothis-Zea and Helicoverpa-Armigera (Lepidoptera, Noctuidae) from China. *Journal of Economic Entomology*, **88**, 1288–1291.

Lavergne S, Molofsky J (2007) Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 3883–8.

Leigh JW, Bryant D (2015) popart : full-feature software for haplotype network construction (S Nakagawa, Ed,). *Methods in Ecology and Evolution*, **6**, 1110–1116.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–9.

Lovejoy NR, Mullen SP, Sword GA, Chapman RF, Harrison RG (2006) Ancient trans-Atlantic flight explains locust biogeography: molecular phylogenetics of Schistocerca. *Proceedings. Biological sciences / The Royal Society*, **273**, 767–74.

Malinsky M, Challis RJ, Tyers AM *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome research*, **23**, 1817–28.

Matthews M (1999) *Heliothine moths of Australia. A guide to pest bollworms and related noctuid groups.* CSIRO Publishing, Collingwood, Australia.

McCaffery AR (1998) Resistance to insecticides in Heliothine Lepidoptera: a global view. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **353**, 1735–1750.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–303.

Nadachowska-Brzyska K, Burri R, Olason PI *et al.* (2013) Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS genetics*, **9**, e1003942.

Nibouche S, Buès R, Toubon J-F, Poitout S (1998) Allozyme polymorphism in the cotton bollworm Helicoverpa armigera (Lepidoptera: Noctuidae): comparison of African and European populations. *Heredity*, **80**, 438–445.

Patterson N, Moorjani P, Luo Y *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–93.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS genetics*, **2**, e190.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular biology and evolution*, **25**, 1253–6.

Prado-Martinez J, Sudmant PH, Kidd JM *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–5.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**, 559–75.

R Core Team (2014) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasool A, Joußen N, Lorenz S *et al.* (2014) An independent occurrence of the chimeric P450 enzyme CYP337B3 of Helicoverpa armigera confers cypermethrin resistance in Pakistan. *Insect biochemistry and molecular biology*, **53**, 54–65.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rosenberg J, Burt PJA Windborne displacements of Desert Locusts from Africa to the Caribbean and South America. *Aerobiologia*, **15**, 167–175.

Sankararaman S, Mallick S, Dannemann M *et al.* (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**, 354–7.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. *PLoS genetics*, **8**, e1002947.

Song S V, Downes S, Parker T, Oakeshott JG, Robin C (2015) High nucleotide diversity and limited linkage disequilibrium in Helicoverpa armigera facilitates the detection of a selective sweep. *Heredity*, **115**, 460–70.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature reviews. Genetics*, **14**, 404–14.

Stefanescu C, Páramo F, Åkesson S *et al.* (2013) Multi-generational long-distance migration of insects: studying the painted lady butterfly in the Western Palaearctic. *Ecography*, **36**, 474–486.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, **56**, 564–77.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution*, **30**, 2725–9.

Tay WT, Mahon RJ, Heckel DG *et al.* (2015) Insect Resistance to Bacillus thuringiensis Toxin Cry2Ab Is Conferred by Mutations in an ABC Transporter Subfamily A Protein. *PLoS genetics*, **11**, e1005534.

Tay WT, Soria MF, Walsh T *et al.* (2013) A brave new world for an old world pest: Helicoverpa armigera (Lepidoptera: Noctuidae) in Brazil. *PLoS ONE*, **8**, e80134.

Veeramah KR, Hammer MF (2014) The impact of whole-genome sequencing on the reconstruction of human population history. *Nature reviews. Genetics*, **15**, 149–62.

Walsh T, Joußen N, Tian K *et al.* Multiple recombination events between two cytochrome P450 loci contribute to global pyrethroid resistance in Helicoverpa armigera. *Submitted*.

Wickham H (2009) *ggplot2*. Springer New York, New York, NY.

Yang Y, Li Y, Wu Y (2013) Current Status of Insecticide Resistance in <I>Helicoverpa armigera</I> After 15 Years of Bt Cotton Planting in China. *Journal of Economic Entomology*, **106**, 375–381.

Zhan S, Zhang W, Niitepõld K *et al.* (2014) The genetics of monarch butterfly migration and warning colouration. *Nature*, **514**, 317–21.

Data Accessability

Alignment of mtDNA genomes: Dryad entry XXX

VCF of B3 BAC alignments: Dryad entry XXX

Plink format data for the GBS data: Dryad entry XXX

Tables

Table 1. Summary genetic statistics at variant positions in populations of *H. armigera* and *H. zea* assessed using GBS, including nucleotide diversity (π) and Wright's inbreeding coefficient (*Fis*).

Population	Private Alleles	Observed Heterozygosity	π	*Fis*
Australia	339	0.027	0.029	0.009
China	243	0.025	0.028	0.011
India	253	0.024	0.029	0.019
Uganda	318	0.024	0.028	0.020
Brazil	767	0.019	0.027	0.068
USA zea	88	0.015	0.015	0.002
Brazil zea	120	0.017	0.016	-0.001
Brazil zea 2	122	0.013	0.016	0.012

Table 2. Evidence for gene flow between populations of heliothine moth determined from GBS data. Significant calculations of *D* are in bold.

D (*H. puntigera*, x ; y, *H. zea* (USA))			
Population x	Population y	*D*	Z-score
Australia	China	**-0.259**	**-3.88**
Australia	Brazil	-0.198	-2.75
Australia	Uganda	-0.198	-2.78
Australia	India	-0.196	-2.90
Australia	Brazil zea	0.023	0.48
Australia	Brazil zea 2	0.078	2.22
China	Brazil	**-0.270**	**-3.74**
Uganda	Brazil	**-0.259**	**-3.95**
India	Brazil	**-0.251**	**-3.44**
Australia	Brazil	-0.198	-2.75
Brazil zea 2	Brazil	**0.846**	**42.34**
Brazil zea	Brazil	**0.852**	**41.14**
China	India	**-0.262**	**-3.88**
China	Uganda	**-0.237**	**-3.31**
India	Uganda	-0.206	-2.82
Brazil	Brazil zea	0.004	0.08

Table 3. Summary genetic statistics, including nucleotide diversity (π), for populations of *H. armigera* assessed using whole genome sequencing data aligned to all BACs.

Population	Tajima's *D*	π	Number of Samples
Australia	-0.370	0.018	17
Brazil	0.020	0.016	5
China	0.140	0.014	4
Europe	0.320	0.012	4
India	0.030	0.017	5
Madagascar	0.230	0.014	3
New Zealand	0.280	0.016	3
Senegal	0.190	0.016	3
Uganda	0.170	0.015	4

Table 4. Evidence for gene flow between populations of Heliothine moth determined from whole-genome sequences aligned to BACs. Significant calculations of *D* are in bold.

D (*H. Assulta*, x ; y, *H. zea* (USA))			
Population x	Population y	*D*	Z-score
New Zealand	Australia	**-0.204**	**-9.45**
China	Australia	**-0.182**	**-9.32**
Uganda	Australia	**-0.172**	**-12.81**
Europe	Australia	**-0.171**	**-11.13**
Madagascar	Australia	**-0.158**	**-12.04**
India	Australia	**-0.156**	**-6.98**
Senegal	Australia	**-0.149**	**-10.63**
Brazil	Australia	**-0.138**	**-9.67**
Brazil zea	Australia	**0.595**	**28.81**
Brazil	Europe	**-0.180**	**-10.37**
Brazil	Senegal	**-0.164**	**-11.86**
Brazil	China	**-0.163**	**-11.09**
Brazil	Uganda	**-0.160**	**-9.55**
Brazil	India	**-0.155**	**-9.49**
Brazil	Madagascar	**-0.149**	**-10.33**
Brazil	New Zealand	**-0.143**	**-9.27**
Brazil	Australia	**-0.138**	**-9.67**
Brazil	Brazil zea	0.007	0.64

Figures

Fig. 1. Bayesian phylogenetic tree derived from 12,248 bp of the mitochondrial genome, showing the relationship of *H. armigera* with other heliothine species. Bootstrap values are shown above nodes and subtrees beyond species distinction have been contracted for clarity. *Spodoptera frugiperda* is the out group.

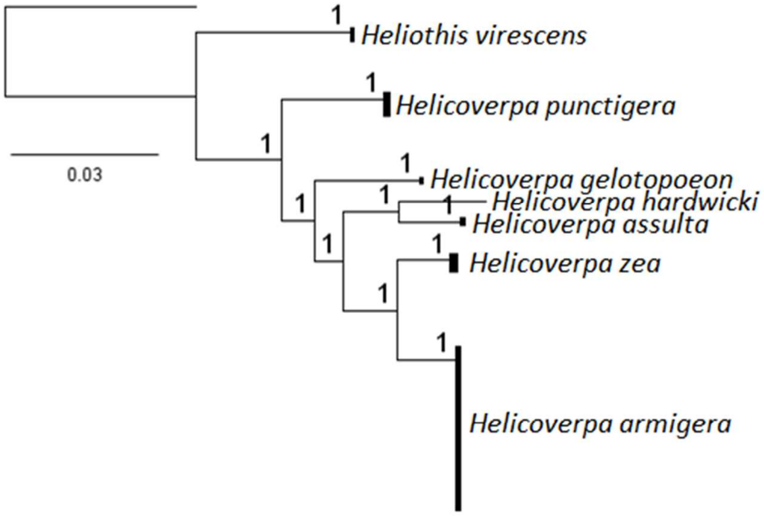


Fig. 2. Mitochondial haplotype network of *H. armigera* from populations around the world. Hatch marks symbolise missing haplotypes and small, open circles represent hypothetical intermediates.

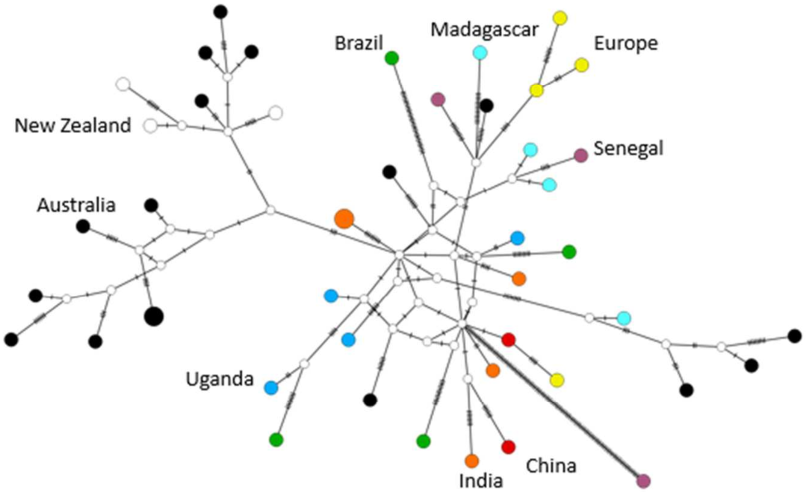


Fig. 3. Principle component anlysis of Φst derived from mitochondrial variation in *H. armigera* populations. The amount of variance explained by each component is noted on their respective axes.

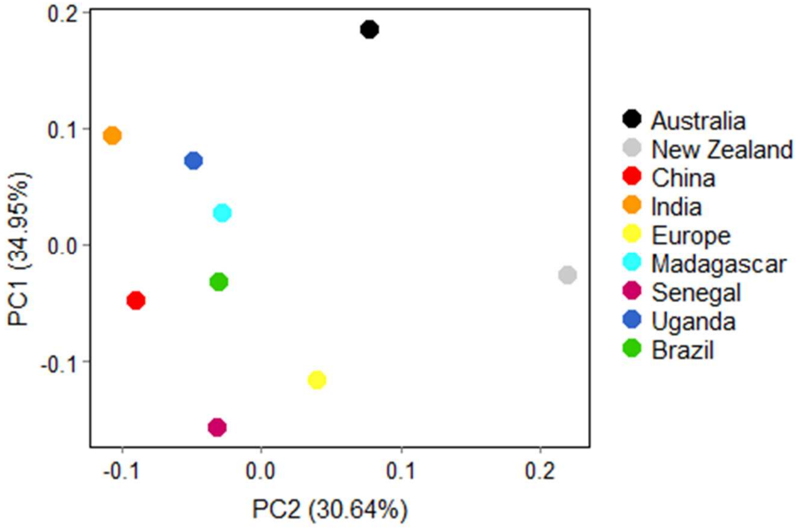


Fig. 4. Principle component analysis of GBS data for populations of *H. armigera* (*n*=217) and *H. zea* (*n*=62). The amount of variance explained by each component is noted on their respective axes.

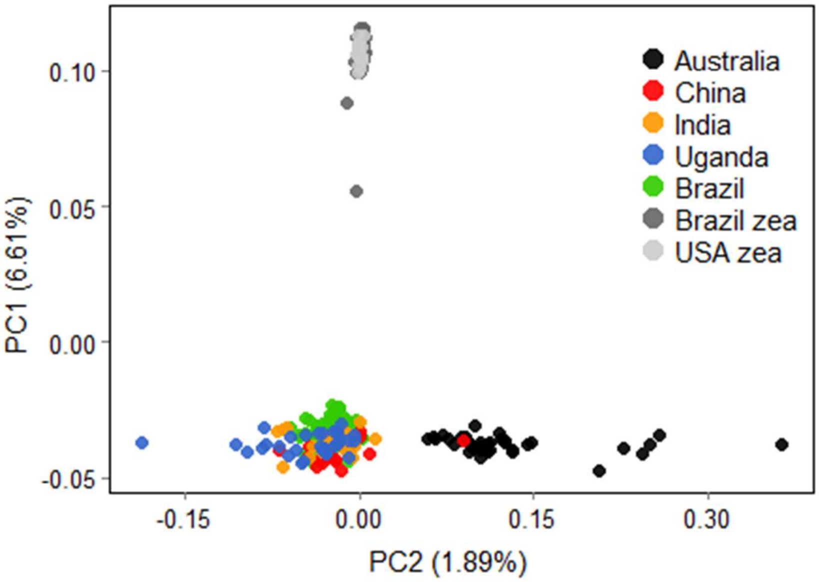


Fig. 5. Structure results from GBS data for *H. armigera* (*n*=217) and *H. zea* (*n*=62), highlighting the distinctions between *H. armigera and H. zea* (top), *H. armigera armigera* and *H. armigera conferta* (middle) and populations of *H. armigera armigera* (bottom). The grey colour reflects samples not used in the analysis and values of K found best to fit the data are next to their respective analyses.

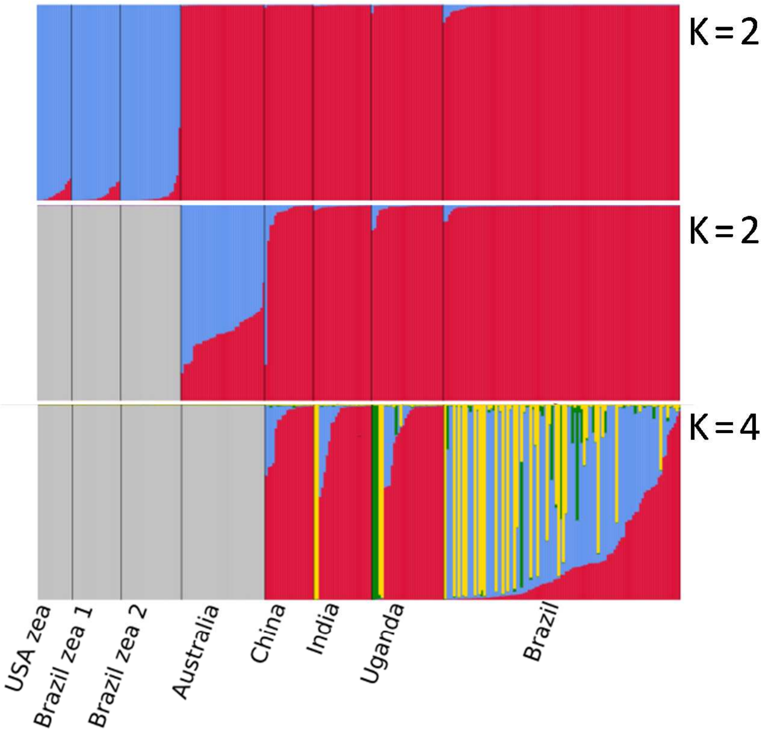


Fig. 6. Principle component analysis of *H. armigera* (*n*=50) and *H. zea* (*n*=8) variants derived from whole-genome sequences aligned to BACs. The amount of variance explained by each component is noted on their respective axes.

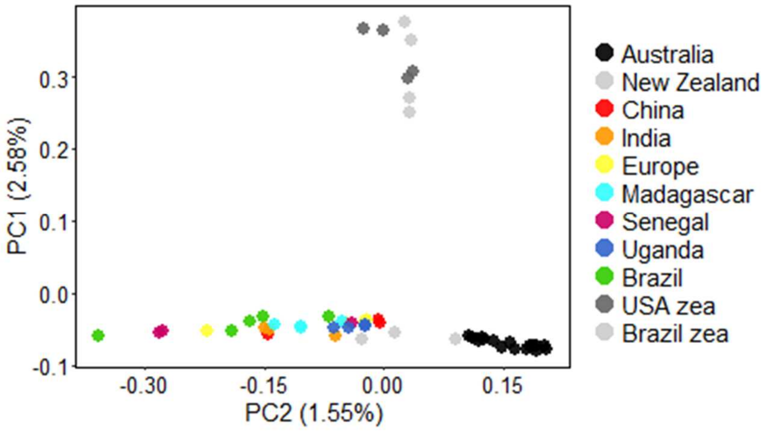


Fig. 7. Structure results derived from whole-genome sequences aligned to BACs, highlighting the distinctions between *H. armigera armigera* and *H. armigera conferta* (top) and populations of *H. armigera armigera* (bottom). Values of K found best to fit the data are next to their respective analyses.

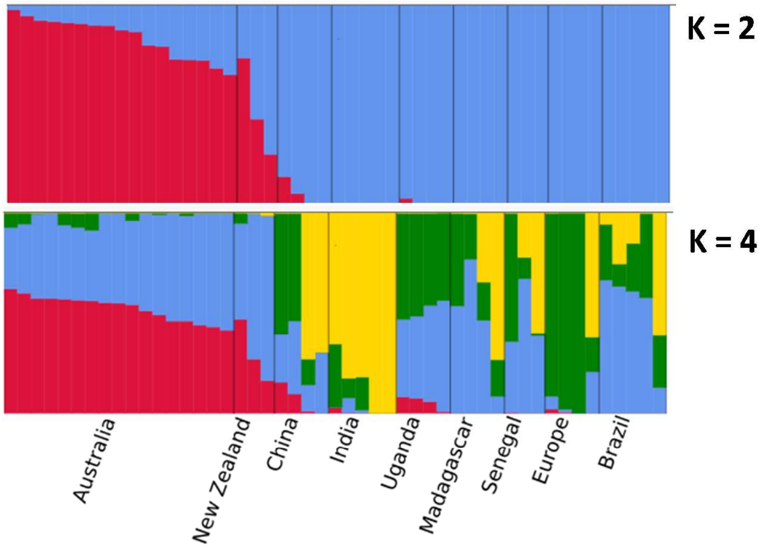
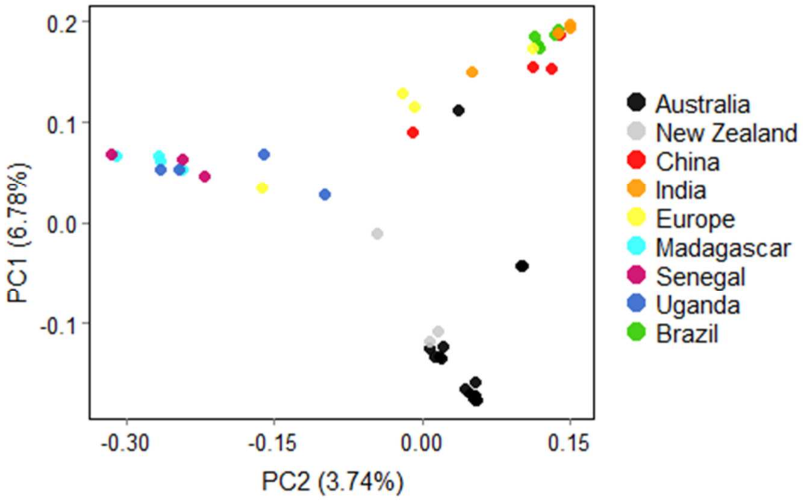
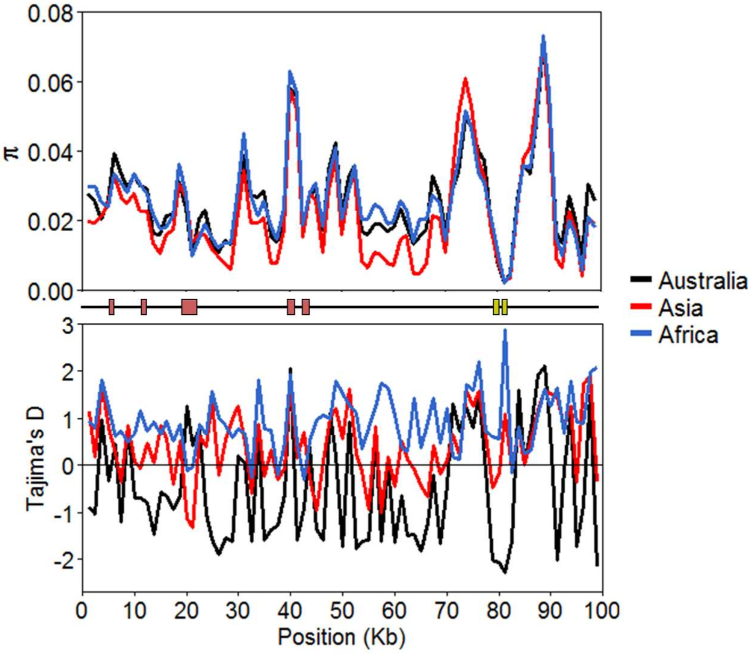


Fig. 8. Principle component analysis of *H. armigera* sequencing data aligned to the 33J17 BAC (JQ995292.1) containing the CYP337B3 gene implicated in pyrethroid resistance. The amount of variance explained by each component is noted on their respective axes.



814 Fig. 9. Nucleotide diversity (π, top) and Tajima's D (bottom) calculated across sliding windows of the 33J17 BAC (JQ995292.1) for individuals found to be homozygous for haplotypes of the chimeric P450 gene associated with fenvalerate resistance. The location of gene bodies are indicated between the plots, with those in red identified as potential reverse transcriptases and those in yellow as exons of *CYP337B3v1*.



819