Bulletin of Entomological Research (2018) **108**, 817–830
© Cambridge University Press 2018

doi:10.1017/S0007485318000081

Population differentiation between Australian and Chinese *Helicoverpa armigera* occurs in distinct blocks on the Z-chromosome

S.V. Song[1,2], C. Anderson[3,4], R.T. Good[1,2], S. Leslie[1,5,6], Y. Wu[7], J.G. Oakeshott[4] and C. Robin[1,2]*

[1]School of Biosciences, University of Melbourne, Victoria, Australia: [2]Bio21 Institute, Parkville, Victoria, Australia: [3]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK: [4]Land and Water Flagship, Commonwealth Scientific and Industrial Research Organisation, Australian Capital Territory, Australia: [5]School of Mathematics and Statistics, University of Melbourne, Victoria, Australia: [6]Centre for Systems Genomics, University of Melbourne, Victoria, Australia: [7]College of Plant Protection, Nanjing Agricultural University, Nanjing, China

Abstract

Over the last 40 years, many types of population genetic markers have been used to assess the population structure of the pest moth species *Helicoverpa armigera*. While this species is highly vagile, there is evidence of inter-continental population structure. Here, we examine Z-chromosome molecular markers within and between Chinese and Australian populations. Using 1352 polymorphic sites from 40 Z-linked loci, we compared two Chinese populations of moths separated by 700 km and found virtually no population structure ($n = 41$ and $n = 54$, with <1% of variation discriminating between populations). The levels of nucleotide diversity within these populations were consistent with previous estimates from introns in Z-linked genes of Australian samples ($\pi = 0.028$ vs. 0.03). Furthermore, all loci surveyed in these Chinese populations showed a skew toward rare variants, with ten loci having a significant Tajima's *D* statistic, suggesting that this species could have undergone a population expansion. Eight of the 40 loci had been examined in a previous study of Australian moths, of which six revealed very little inter-continental population structure. However, the two markers associated with the *Cyp303a1* locus that has previously been proposed to be a target of a selective sweep, exhibited allele structuring between countries. Using a separate dataset of 19 Australian and four Chinese moths, we scanned the molecular variation distributed across the entire Z-chromosome and found distinct blocks of differentiation that include the region containing *Cyp303a1*. We recommend some of these loci join those associated with insecticide resistance to form a set of genes best suited to analyzing population structure in this global pest.

Keywords: CYP303A1, *Helicoverpa armigera conferta*

(Accepted 2 January 2018; First published online 5 February 2018)

Introduction

Helicoverpa armigera is one of several polyphagous moths classified within the so-called 'mega pest' lineage of heliothine

*Author for correspondence:
Tel: +61 3 8344 2349
E-mail: crobin@unimelb.edu.au

moths which also includes *Chloridea virescens*, *Helicoverpa zea* and *Helicoverpa punctigera*. It causes billions of dollars of crop damage annually. The genus *Helicoverpa* most probably evolved within Australasia, as the most basal lineage within it (*H. punctigera*) is only found in Australia, and the closest sister genus, *Australothis* is also found there (Matthews, 1999; Cho *et al*., 2008). Five (*H. armigera*, *H. assulta*, *H. hardwicki*, *H. prepodes*, and *H. punctigera*) of the approximately 20 *Helicoverpa* species described are found in Australia (Gordon *et al*., 2010; Mitchell & Gopurenko, 2016). It is less clear where the species *H. armigera* arose. A typical way of inferring origins, the center of nucleotide diversity, does not provide a clear answer in the literature published thus far (Nibouche *et al*., 1998; Zhou *et al*., 2000; Behere *et al*., 2007). The age of the *H. armigera* diaspora is also somewhat clouded, although mounting evidence suggests that it arose after the divergence from the New World sibling species, *H. zea*, which has been estimated as 1.5–2 mya (Mallet *et al*., 1993; Behere *et al*., 2007; Pearce *et al*., 2017).

Until recently, the distribution of *H. armigera* was limited to Europe, Asia, Africa, and Australia. However, the species was detected in South America around 2012 and is currently spreading through the Americas, causing significant damage (Czepak *et al*., 2013; Tay *et al*., 2013; Murúa *et al*., 2014; Arnemann *et al*., 2016; Sosa-Gómez *et al*., 2016). Many countries currently control this pest with strategies that combine crops genetically modified to produce *Bacillus thuringiensis* insecticidal proteins with practices such as the planting of 'refuges' aimed to slow the spread of recessive resistance alleles (Tabashnik *et al*., 2004). The recent range expansion of *H. armigera* into areas likely to have heterogeneous insecticide exposures and control strategies may increase the likelihood of insecticide resistance alleles proliferating. Resistant alleles may then flow back to populations in the ancestral range. Thus, tracking gene flow of insecticide resistance alleles is important to control this damaging species throughout its range (Daly, 1993; Fitt, 1994).

Population genomics offers a new way to identify insecticide resistance alleles because they may be associated with selective sweeps. Under the scenario where a resistance allele reaches high frequency (or fixation) in a population, the polymorphisms in neighboring gene regions will exhibit atypical patterns such as reduced within-species variation, elevated levels of differentiation among populations, extended haplotypes, and a skewed frequency spectrum such that a greater fraction of the variants are rare (Nielsen, 2005). Studies in *Drosophila melanogaster* underscore the potential of this approach, with genome-wide scans for positive selection successfully recovering strong signals at known resistance loci (Garud *et al*., 2015; Battlay *et al*., 2016) despite the fact that *D. melanogaster* is not itself a direct target of insecticides. Thus, these findings hold much promise for organisms such as *H. armigera*, which would be expected to face much stronger selective pressures as targeted pest species.

As high-throughput sequencing technologies continue to be refined, increasingly sophisticated approaches are available to reduce the complexities in genomic data yet capture the inherent architectures that are unique to each species. These reduced-representation sequencing technologies include RAD-Seq (Baird *et al*., 2008; Rašić *et al*., 2014) and genotyping-by-sequencing (GBS; Elshire *et al*., 2011). Recently, Anderson *et al*. (2016) investigated worldwide population genomic variation in *H. armigera* with two different ways of sampling alleles from portions of the genome. In one, they examined 21,043 SNPs located across the genome among 216 individuals using a GBS approach. In the other, they examined the variation among 50 individuals in specific fully sequenced regions aligning to 2.3 Mb of 20 BAC clones. They found clear evidence for inter-continental population structure in their full mitochondrial sequence dataset and in their GBS and BAC-aligned sequence datasets. This prompted them to resurrect an idea that *H. armigera* from Australia had its own subspecies, named *H. armigera conferta*, that differs from *H. armigera armigera* in Africa, Europe, and Asia (Common, 1953; Matthews, 1999). Several groups have used mtDNA to examine population structure, but these earlier datasets show a shallow star phylogeny and little structure (Behere *et al*., 2007; Tay *et al*., 2013; Leite *et al*., 2014; Mastrangelo *et al*., 2014). Microsatellite studies have also been used to examine this issue but interpretations are confounded by non-Mendelian patterns of the markers attributable to a high frequency of null alleles and to associations with transposable elements (Zhang, 2004; Tay *et al*., 2010). Behere *et al.* (2013) primarily used autosomal exon-primed intron-crossing (EPIC) markers to address population structure issues between crops in India, and they too showed little compelling evidence for population structure.

Our previous study of the Z-chromosome of Australian *H. armigera* indicated that the amount of variation observed within populations was very high relative to other organisms (nucleotide diversity, $\pi \sim 0.02$) and that recombination between sites must also be high as linkage disequilibrium (LD) decayed rapidly, with $E(r^2)$ approaching 0.2 within 200 bps (Song *et al*., 2015). Given that these metrics will differ due to the varying degrees of influence that neutral and non-neutral processes have on a finite population, we wanted to explore the robustness of the estimated parameters and examine if these observations could be extended to non-Australian populations. *H. armigera* follows a ZZ/ZW sex determination system with females being the heterogametic sex. As in Song *et al.* (2015), a crucial aspect of the design we follow here is the focus on Z-linked loci because sequencing females provides empirical observations of haplotypes and overcomes the limitations of imputation solutions to the gametic-phase problem (in an individual heterozygous at two loci, the gametic phase is either AB/ab or Ab/aB and which one it is affects LD calculations), which are particularly inaccurate when LD decays rapidly (Slatkin, 2008).

Here, we use a Z-linked EPIC dataset and a Z-chromosome-wide scan to identify candidate loci that are informative for inter-continental population structure and to assess the *conferta* sub-species hypothesis. The first dataset employs targeted resequencing to examine nucleotide diversity and LD at 40 EPIC markers in female samples from two collection sites in China. The use of 454 pyrosequencing of PCR amplicons provides us with relatively long reads (~600 bp), and by analyzing only female moths which contain only a single Z-chromosome, we cleanly sample a single allele per individual. The 40 markers include eight loci characterized in Australian samples from our previous study (Song *et al*., 2015), which allows us to explore the question of differentiation between inter-continental populations of *H. armigera*. The second dataset uses previously reported whole-genome sequencing from Anderson *et al.* (2016) to scan the Z-chromosome for regions that show high levels of differentiation between Australian and Chinese individuals. This dataset provides us with an independent sample of moths from the two countries to investigate how patterns of differentiation are distributed across the chromosome while further exploring the question of inter-continental population structure.

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Materials and methods

DNA samples

Our analysis of polymorphic sites in the 40 Z-linked EPIC markers in the Chinese samples was based on 41 female adults from Nanpi (Hebei Province, Yellow River Valley) (38°2′N, 116°42′E) and 54 female adults from Yancheng (Jiangsu Province, Changjiang River Valley) (33°20′N, 120°9′E) collected in 2011. DNA was isolated by column purification. The provenance of the Australian samples has been reported in Song *et al.* (2015). Briefly, they consist of eggs collected from MacIntyre Valley (Queensland) which were reared to adults in the laboratory, after which only female samples were used.

Samples used in the Z-chromosome-wide analysis comprised 19 Australian and four Chinese individuals used in Anderson *et al.* (2016). In this case, *H. armigera* were collected as adults using emergence traps placed among cotton plants in New South Wales, Australia, and as caterpillars from cotton in Shandong, China. DNA was isolated by column purification.

Library preparation

For the EPIC markers, Z-linked loci were identified in the silkworm *Bombyx mori* using the published GLEAN cDNA dataset for this species and coding sequences were extracted based on the Z-chromosome (chromosome 1) scaffold numbers (nscaf1690, nscaf2210, nscaf2734, nscaf3040, and nscaf3068 from http://silkworm.genomics.org.cn/). The *H. armigera* orthologues of these *B. mori* loci were identified though BLAST searches against a repository containing contigs from an *H. armigera* reference strain (Pearce *et al.*, 2017). Ultimately an independent assembly of the *Helicoverpa* genome confirmed that these 40 loci were indeed on the Z-chromosome of *H. armigera*. Intron–exon boundaries on the *H. armigera* contig sequences were identified with EXONERATE under the *cdna2genome* model using the silkworm cDNA as input queries. Loci were selected to include different regions of the Z-chromosome and to incorporate clusters of loci, where the distance separating two loci was <50 kb (fig. 1) to allow for the possibility of detecting long-range LD.

The 454 Universal Tailed Amplicon (Roche) sequencing design was used where sequences from each individual were given unique barcodes. Briefly, the strategy utilizes two successive rounds of PCR to produce amplicons appropriate for the sequencing platform. The first round of PCR employs fusion primers consisting of the universal tail sequence and a gene-specific primer. The second round of PCR is carried out with fusion primers comprising the 454 sequencing primers, a custom 5-bp MID (multiplex identifier) sequence, and a sequence which targets the tail sequences introduced in the first round. Primer sequences are provided in Supplementary table S1 and loci have been named according to the *B. mori* gene they are based on; those with the suffix BGIBMGA refer to *B. mori* orthologs using the names assigned by the silkworm sequencing consortium (http://silkworm.genomics.org.cn).

First-round amplification with locus-specific primers was carried out in a reaction volume of 20 µl. Cycling conditions were typically 35 cycles of 94°C for 20 s, 58°C for 20 s, 68°C for 1 min. To estimate the size and intensity of bands, 3 µl of each product was visualized on agarose gels with 5 µl of DNA markers. Each amplicon was assigned a score of intensity ranging from 2 (strongest) to 10 (weakest) prior to pooling by sample barcodes (MIDs) for library preparation and sequencing. To compensate for different yields (to avoid oversampling of particular amplicons during sequencing), pooling was carried out in the following manner: 2 µl of an amplicon was included if it had a score of 2, and up to 10 µl was included for weaker products.

Bead purification was carried out with MagNA buffer (Rohland & Reich, 2012) and 1 µl of each pool was used as a template for second-round amplification in individual reaction volumes of 50 µl. Second-round amplification was carried out in duplicate with ten cycles of 94°C for 30 s, 68°C for 1.5 min to incorporate the MID barcodes and adapter sequences. The barcoded pools were combined into a single library then purified and concentrated by bead purification, gel excision, and column purification prior to quantification. Sequencing was performed by the Ramaciotti Centre for Genomics (University of New South Wales, Australia) using Roche 454 GS FLX Titanium chemistry for amplicon sequencing (XLR70).

For the Z-chromosome-wide dataset, library preparation has been described in Anderson *et al.* (2016). Briefly, 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, Australia).

Sequence diversity and LD

For the EPIC dataset, sequence reads were first sorted by MIDs (individual barcodes) and assembled *de novo* in GENEIOUS R7 (Biomatters, Auckland, New Zealand) using a threshold of 10% for maximum mismatches per read. Contigs from all individuals were then mapped to the reference *H. armigera* amplicon sequences using BLASTN. Mapped contigs comprised 116,069 reads, representing 83% of the total number of reads. Multiple sequence alignments for each locus were performed using SEAVIEW 4.0 (Gouy *et al.*, 2010) and CLUSTALX (Larkin *et al.*, 2007). Sequences were assessed at two levels in the process of error correction: at the individual (MID) level and at the population (aligned to reference) level. Where low coverage at the individual level resulted in base ambiguities, this was resolved by comparing to other sequences in the population, i.e. if all other individuals were monomorphic at that site, the ambiguity was manually edited to match the population. Base ambiguities observed at segregating sites were edited to match the majority (>50%) of individuals in the population. This approach was adopted to counteract the possible inflation of rare SNPs in the population which could result in deviations from the expected site frequency spectrum under neutrality. The filtered dataset contained a total of 40 amplicons distributed across 29 *H. armigera* contigs that map to five *B. mori* Z-linked scaffolds (fig. 1) with an average of 58 individuals per locus.

Analyses of polymorphism and LD were carried out using DNASP 5.10.01 (Rozas *et al.*, 2003) with alignment files indicated as haploid Z-chromosome. Nucleotide diversity (π) and Tajima's *D* were estimated using a total of 2351 sites. A second estimate of π and θ was also made using only the first 100 bases from the 5′ ends of the sequence alignments (with indels excluded) to explore the possibility that diversity could be inflated by sequencing errors toward the 3′ ends of longer reads (Gilles *et al.*, 2011). LD was estimated as the square of the correlation coefficient, r^2, using only

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

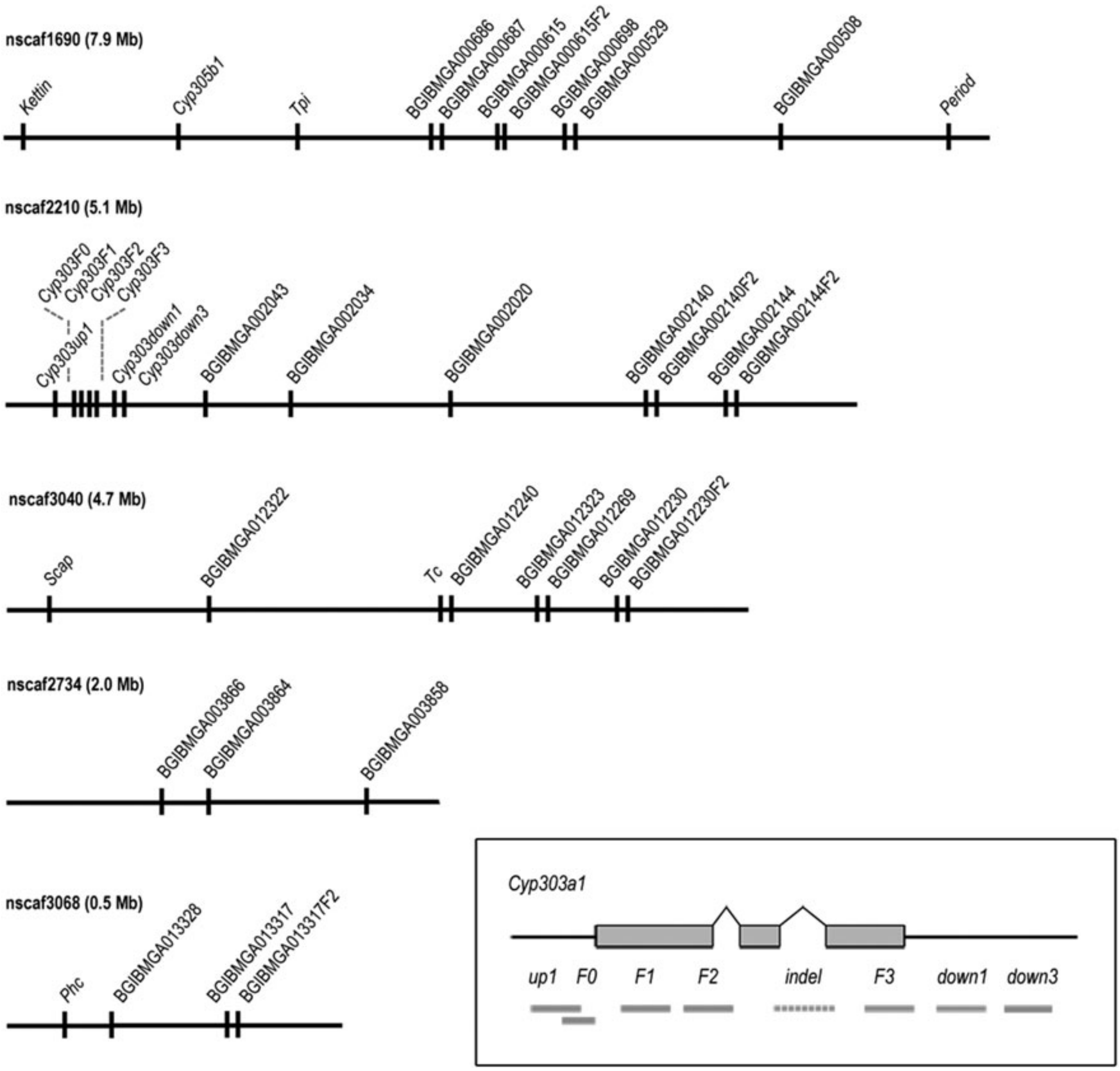

Fig. 1. Map of EPIC amplicons used in this study. Positions of contigs are based on the five Z-chromosome scaffolds of *Bombyx mori*. Numbers in brackets after the scaffold names refer to the lengths of the scaffolds in megabases. Amplicons are named according to the gene they are based on; those with the suffix BGIBMGA refer to *B. mori* orthologs using the names assigned by the silkworm sequencing consortium (http://silkworm.genomics.org.cn). Inset: Amplicon design for the *Cyp303a1* locus showing the location of a 200bp insertion/deletion (indel) polymorphism.

parsimony-informative sites, with the decay of LD over physical distance modeled on the expectations of Hill & Weir (1988) and implemented using the non-linear least-squares (*nls*) function in R (R Core Team, 2014).

The data analyzed in the frequency spectrum tests above included eight haplotypes (spanning four loci) that were highly diverged from other alleles at those loci. Notably, three of the haplotypes came from a single individual collected from Yancheng, MID-95. At two loci where we had sequences from *H. assulta*, the divergent haplotype of MID-95 resembled the *H. assulta* sequences. We thus excluded MID-95 from all analyses on the likely basis that it was an *assulta* specimen.

For the Z-chromosome-wide dataset, data handling has been described in Anderson *et al.* (2016). Briefly, raw sequence reads were aligned to the Z-chromosome of the *H. armigera* genome using BBMAP v.33.43 (http://sourceforge.net/projects/bbmap/). Quality trimming of reads was carried out when at least two consecutive bases fell below Q10, and only uniquely aligning reads were included in the analysis. UnifiedGenotyper in GATK v. 3.3-0 (McKenna *et al.*, 2010) was used to estimate genotypes, implementing a heterozygosity value of 0.01.

Analyses of population structure

Analysis of molecular variance, F_{ST}*, and STRUCTURE analyses*

Analysis of molecular variance (AMOVA) and F_{ST} calculations were performed on the EPIC dataset using functions implemented in the R packages *adegenet* v2.0.0, *hierfstat* v0.04-22, and *Poppr* v2.2.0 (Goudet, 2005; Jombart & Ahmed, 2011; Kamvar *et al.*, 2014). We used STRUCTURE version 2.3.4 (Pritchard *et al.*, 2000) to assess the degree of population stratification between Nanpi and Yancheng, and between

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

MacIntyre Valley (Australia) and China. Two approaches were initially considered in the definition of a locus: haplotypes, for which the dataset is well suited due to the phased (hemizygous) nature of the sequences; and sites, which increase the number of loci available for analysis. A haplotype is made up of many polymorphic sites within the same amplicon. Using this 'haplotype approach', our dataset thus contains only 40 loci for analysis. The 'site approach' treats each polymorphic site as a locus, and there were 2351 such sites in our dataset. Singleton sites (where only a single instance of a variant allele was observed) were then excluded as they are uninformative with respect to distinguishing between populations, reducing the number of loci to 1352 in the final analysis. In cases where sequences were not available for all individuals, a value of −9 was assigned to denote missing data.

The haplotype approach avoids any possible issues with non-independence between sites due to LD, but has the disadvantage of collapsing multiple polymorphic sites into a single haplotype, thus reducing the number of loci for analysis and consequently the power to assign individuals to a specific cluster. A second limitation was the high levels of diversity in all populations, leading to an increase in the uncertainty surrounding the clustering of singleton and low-frequency haplotypes. We attempted to reduce the number of singleton haplotypes by grouping together individuals that differed from a 'core haplotype' at polymorphic sites not represented elsewhere (singleton sites) but even so, overall haplotype diversity remained high, and this motivated the site approach. While a STRUCTURE analysis would not typically consider sites within the same amplicon to be independent markers, the rapid decay of LD in *H. armigera* suggested that a valid analysis could be performed without violating the assumptions of independence between loci (Falush *et al.*, 2003). To assess the robustness of the site approach, a series of 'thinned' datasets were generated by randomly selecting 50% of the 1352 sites available, and subjected to the same STRUCTURE analyses as the full dataset (described below). Ten such datasets were generated, all of which showed a similar outcome to the full dataset, indicating that linked SNPs are not over-represented despite the short distances separating each locus (site). Only the results of the 'site' analysis are presented here as they proved to be more informative than the haplotype analysis.

All STRUCTURE analyses were carried out under the model incorporating admixture and independent allele frequencies between populations, without using prior population information. A series of analyses was run using values of K (the number of genetically defined populations, which may be unknown) from 1 to 5 using the same parameters. Ten replicates were run for each K value with 10,000 iterations for the burn-in period followed by an additional 10,000 iterations after the burn-in. In cases where a choice lay between $K = 2$ and $K > 2$ to explain most of the structure in the data, the Evanno method (Evanno *et al.*, 2005) was used to formally evaluate the most likely K value via STRUCTURE HARVESTER (Earl & vonHoldt, 2011) by choosing the value of K that corresponds to the largest value of ΔK. For the Nanpi–Yancheng comparison where the choice lay between $K = 1$ and $K = 2$, no formal evaluation was applied as it is not possible to obtain a value of ΔK between $K = 0$ and $K = 1$; the most likely value of K was inferred from the graphical results.

The output of a STRUCTURE analysis is typically presented in the form of a bar chart illustrating the number of distinct populations as well as assignments of individuals to populations (such as those depicted in fig. 2). However, these visualizations represent the output of only a single run. We performed an analysis to explore the correlation between a single STRUCTURE run vs. the results of 100 runs (Supplementary fig. S1). The results show that there is a moderate to high level of reproducibility in the assignment of individuals to a population across multiple runs. We therefore conclude that in most cases, the graphical output of a single, randomly selected run is a reasonable representation of the results of multiple runs.

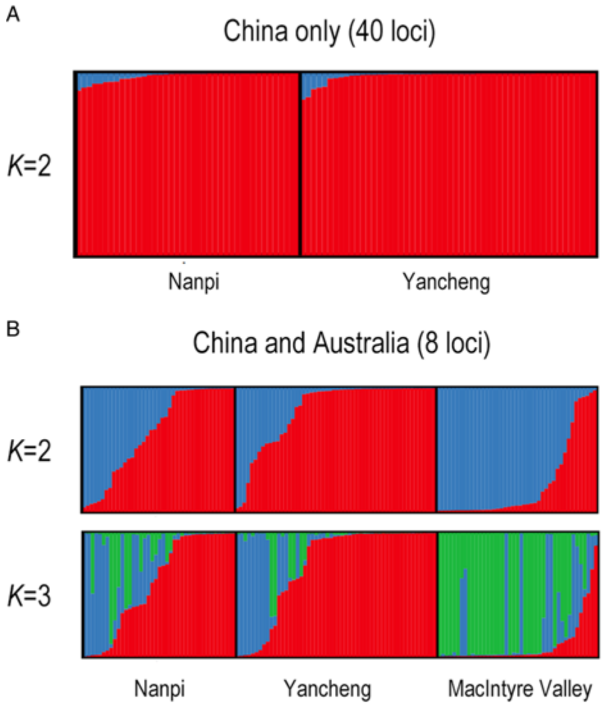

Fig. 2. (A) STRUCTURE plots using 40 loci for the Nanpi and Yancheng populations, with $K = 2$. Each bar (column) represents an individual. Colors and bar heights represent the inferred ancestries of an individual. (B) STRUCTURE plots using eight loci for the Nanpi, Yancheng, and MacIntyre Valley populations. Each bar (column) represents an individual. Colors and bar heights represent the inferred ancestries of an individual, so admixed individuals that have a heritage derived from mixed sources are represented by columns with more than one color. Only $K = 2$ and $K = 3$ are shown as they represent the values most likely to explain the major structure in the dataset. In the $K = 2$ visualization, MacIntyre Valley has more 'blue' individuals while China has more 'red'. In the $K = 3$ visualization, MacIntyre Valley has more 'green' individuals compared with the primarily red and blue landscape of the Chinese individuals, suggesting that some alleles observed in Australia are rare (or not found) in China. Similarly, the higher incidence of 'red-only' individuals in China suggests that some alleles that are common here are rare in Australia.

Z-chromosome-wide sliding window analysis of weighted F_{ST}

Imputation of missing bases in the Z-chromosome-wide dataset was performed using default parameters in *Beagle* (Browning & Browning, 2007). LD-based pruning was conducted using *Plink* v.1.07 (Purcell *et al.*, 2007) with the command '--indep 50 5 2'. Non-negative weighted F_{ST} and Tajima's D were calculated across sliding windows using *vcftools* v.0.1.14

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

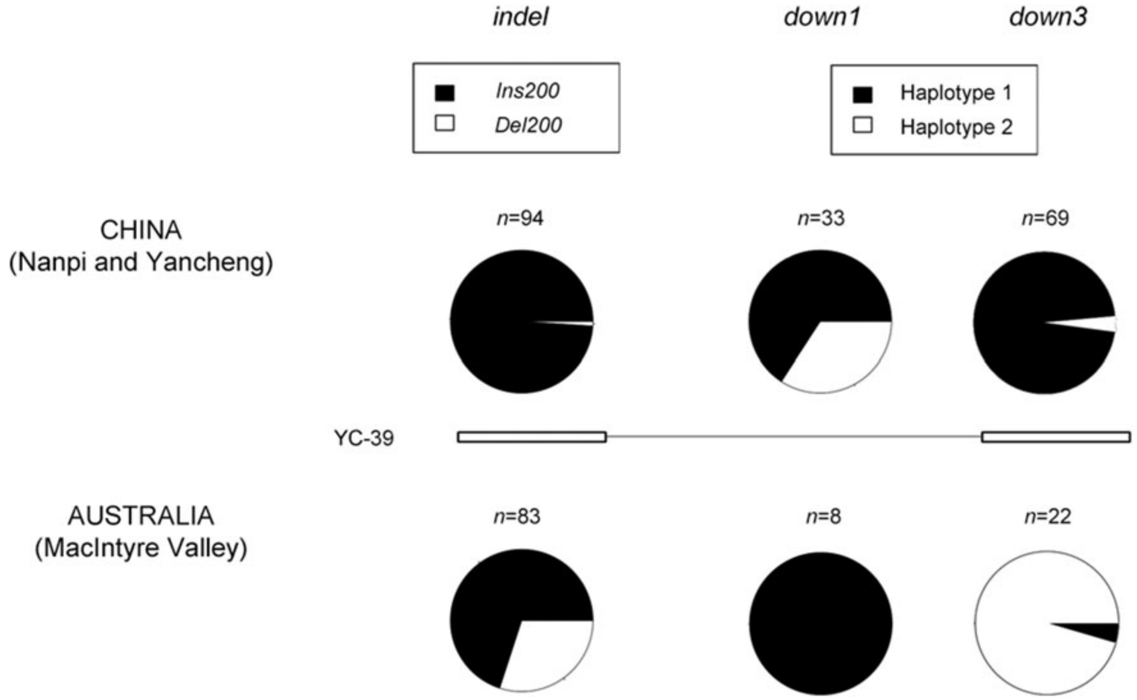

Fig. 3. Major haplotypes observed in the sequenced regions of the *Cyp303a1* locus in Australia. At the *Cyp303indel* locus, the *Del200* haplotype occurs at approximately 30% frequency in Australia but only appears in one individual, YC-39 from China. Downstream 1 kb of the indel, two major haplotypes are observed in China. At the *down3* locus, a different haplotype dominates in the two countries, and individual YC-39 carries a haplotype that is commonly found in Australia. The haplotype at the *down3* locus is primarily defined by two sites, from a total of 14 polymorphic sites at this locus (Supplementary fig. S4).

(Danecek *et al.*, 2011) where each window contained 500 SNPs. Sliding window analyses were plotted using *R* and *ggplot2* (Wickham, 2009). A series of 1000 permutations was carried out on the empirical dataset by randomly partitioning the samples into two groups, each containing 19 and four individuals. The same weighted F_{ST} analysis was then carried out on these permuted datasets whereby the 95th percentile of the F_{ST} values obtained in each window and plotted. The aim of this exercise was to assess the extent to which the empirical observations deviated from the expected distribution of F_{ST} values, given the small sample sizes and the disparity between the number of individuals in each group. Outlier loci were defined as those with an F_{ST} value above the 95th percentile of all F_{ST} values in the empirical dataset.

Results

Sequence and haplotype diversity in China

Measures of nucleotide diversity within the Nanpi population and the Yancheng population were similar. Values of π ranged from 0.003 to 0.130 nucleotide differences per site (Supplementary table S2), averaging 0.028 across 40 loci, while haplotype diversity ranged from 0.5 to 1 (within amplicon lengths of 120–660 bp), similar to that observed in Australian populations (Song *et al.*, 2015). LD decayed in a manner similar to that of the Australian populations, with the average r^2 (the square of the correlation coefficient between pairs of polymorphic sites) falling below 0.2 for sites separated by 200 bp or more (Supplementary fig. S2).

There was little population structure between Nanpi and Yancheng. F_{ST}, a traditional measure of population structure, was very low (0.03) and an AMOVA indicated that <1% of the variation in the samples discriminated between the two Chinese populations, despite them being 700 km apart. Likewise, analyses using the STRUCTURE program provided little support for population differentiation (fig. 2A, Supplementary table S3). We performed two separate analyses on the combined Nanpi and Yancheng datasets; the 'haplotype' and the 'site' analysis. In the former, haplotypes were determined for each of the 40 loci; many haplotypes were found and all were at low frequencies (many haplotypes occur only once in a population). This approach therefore reduces the power to assign individual moths to a particular genetically defined cluster. By contrast, the 'site' analysis treats sites as independent (i.e. assumes no LD). As each site is limited to a maximum of four states (we have not considered gaps) with most sites being bi-allelic, the estimated allele frequencies resulting from this treatment are substantially higher, increasing the ability of the STRUCTURE algorithm to strongly assign an individual to one cluster or another. Additionally, the use of polymorphic sites rather than haplotypes increases the number of loci available for analysis by 30-fold (40 loci vs. 1352 sites). The STRUCTURE plot shows that Nanpi and Yancheng are both dominated by a single 'genetically defined population' (red in fig. 2A). These results indicate that the geographical separation of the two moth collections is not manifest in the genetic data surveyed here.

Given the lack of evidence for population differentiation, the following analyses in this section were performed by treating Nanpi and Yancheng as a single population. Initially, to

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Table 1. Nucleotide diversity and Tajima's *D* for all sites, and for the first 100 bases.

Locus	π	π_{100}	θ	θ_{100}	Tajima's *D*
BGIBMGA000508 (22)	0.009	0.021	0.018	0.033	−1.69
BGIBMGA000529 (56)	0.082	0.114	0.114	0.144	−0.98
BGIBMGA000615 (68)	0.026	0.016	0.072	0.044	−1.92*
BGIBMGA000615F2 (47)	0.035	0.021	0.069	0.036	−1.49
BGIBMGA000686 (61)	0.030	0.061	0.069	0.113	−1.68
BGIBMGA000687 (66)	0.026	0.023	0.040	0.036	−1.13
BGIBMGA000698 (61)	0.029	0.021	0.079	0.051	−1.70
BGIBMGA002020 (19)	0.037	0.027	0.043	0.034	−0.54
BGIBMGA002034 (54)	0.044	0.031	0.090	0.092	−1.61
BGIBMGA002043 (46)	0.019	0.012	0.029	0.009	−1.22
BGIBMGA002140 (50)	0.032	0.005	0.063	0.018	−1.63
BGIBMGA002140F2 (90)	0.029	0.031	0.050	0.041	−1.37
BGIBMGA002144 (28)	0.048	0.045	0.084	0.081	−1.65
BGIBMGA002144F2 (57)	0.032	0.020	0.066	0.048	−1.48
BGIBMGA003858 (10)	0.009	0.010	0.014	0.011	−1.65
BGIBMGA003864 (52)	0.046	0.040	0.062	0.042	−0.74
BGIBMGA003866 (93)	0.067	0.053	0.092	0.076	−0.91
BGIBMGA012230 (19)	0.004	0.002	0.009	0.006	−2.15**
BGIBMGA012230F2 (83)	0.005	0.002	0.015	0.012	−2.00*
BGIBMGA012240 (88)	0.027	0.007	0.043	0.014	−1.01
BGIBMGA012269 (13)	0.032	0.045	0.048	0.089	−1.00
BGIBMGA012322 (83)	0.027	0.035	0.062	0.058	−1.42
BGIBMGA012323 (31)	0.125	0.095	0.202	0.145	−0.43
BGIBMGA013317 (24)	0.036	0.002	0.056	0.005	−1.24
BGIBMGA013317F2 (84)	0.006	0.006	0.015	0.008	−1.77
BGIBMGA013328 (68)	0.006	0.001	0.011	0.002	−1.86*
Cyp303down1 (33)	0.021	0.032	0.029	0.041	−0.98
Cyp303down3 (69)	0.009	0.004	0.026	0.012	−2.04*
Cyp303F0 (75)	0.013	0.015	0.046	0.043	−2.21*
Cyp303F1 (88)	0.004	0.007	0.015	0.014	−2.30**
Cyp303F2 (83)	0.005	0.004	0.026	0.024	−2.63***
Cyp303F3 (25)	0.016	0.012	0.018	0.021	−0.49
Cyp303up1 (78)	0.023	0.029	0.061	0.057	−1.98*
Cyp305b1 (36)	0.020	0.011	0.027	0.012	−0.94
Kettin (93)	0.007	0.012	0.024	0.049	−2.15*
Period (87)	0.015	0.012	0.032	0.025	−1.34
Phc (59)	0.039	0.018	0.055	0.037	−0.72
Scap (79)	0.034	0.026	0.062	0.085	−1.46
Tc (89)	0.008	0.009	0.014	0.012	−1.37
Tpi (48)	0.048	0.025	0.061	0.036	−0.33
Average	0.028	0.024	0.050	0.043	

Figures in brackets after the locus name represent the total number of sequences surveyed.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

address the possibility that diversity could be inflated due to sequencing errors, π and θ (measures of genetic diversity estimated from the average number of pairwise differences and the number of segregating sites, respectively) were estimated using the first 100 bases (hereafter referred to as π_{100} and θ_{100}) from the 5′ end (excluding indels) where sequence quality is expected to be superior (Table 1). These values were compared to values obtained from the full-length datasets. While there was considerable variance between the two estimates at individual loci, the values averaged across all loci were similar, suggesting increased polymorphism in longer sequences due to sequencing error does not inflate estimates substantially.

As the differences between θ and θ_{100} (and π and π_{100}) were small, Tajima's *D* was estimated using all sites. A negative value was observed at every locus, indicating an excess of rare variants in the combined population. At ten of the 40 loci, these values were significantly different from neutral expectations. Five of the ten are located around the *Cyp303a1* locus, which we have deliberately chosen to sample at a higher density here compared with other loci, as this is a region previously reported to harbor signatures of a selective sweep. Of the remaining five loci with a significantly negative Tajima's *D*, two (BGIBMGA012230 and BGIBMGA012230F2) are located in the same protein-coding gene (BGIBMGA012230) separated by a distance of 4 kb. This gene is predicted to code for a subunit of the CCR4-Not protein complex, a global regulator of gene expression (Collart & Panasenko, 2012). The third locus is in the *Kettin* gene, which codes for a highly conserved protein involved in insect flight muscle development (Lakey *et al.*, 1993). The fourth, BGIBMGA000615 shares some sequence similarity with CG32030 in *D. melanogaster* (http://flybase.org/) which contains a formin domain. Members of the formin family of proteins have been characterized as playing a role in cytokinesis and cytoskeletal control (Wallar & Alberts, 2003). The fifth, BGIBMGA013328 is a TUDOR-SN protein containing staphylococcal nuclease-like (SN) and Tudor domains. The silkworm TUDOR-SN is thought to be involved in the formation of stress granules (RNA-protein

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

complexes that form when translation initiation is impaired during a stress response) and interacts with the components of the RNAi pathway (Zhu *et al.*, 2012, 2013).

Analysis of intercontinental population structure

To investigate the extent of population structure between *H. armigera* from different continents, eight loci that had been Sanger-sequenced in Australian populations previously (Song *et al.*, 2015) were included among the markers sequenced in the Nanpi and Yancheng populations: *Cyp303down1*, *Cyp303down3*, *Cyp305b1*, *Period*, *Phc*, *SCAP*, *Tc*, and *Tpi*. An analysis was performed to assess two competing hypotheses, namely, whether the samples from Nanpi, Yancheng, and MacIntyre Valley clustered into two or three groups. Replicate runs of the STRUCTURE analysis favored models involving two genetically defined groups, $K = 2$ (fig. 2B, Supplementary table S4). This result is consistent with the lack of differentiation between the two Chinese populations as seen in the data above. The differences between the Australian and Chinese samples are subtle with the two genetically defined groups being reciprocally more abundant in the two countries (fig. 2B) and admixed individuals being present in all three populations. If the number of genetic groups is increased to three, then one class (green) is more abundant in Australia. Omitting the *Cyp303a1*-associated loci produced a similar pattern. These patterns should not be overinterpreted however as the lack of distinct clustering could also arise due to a lack of power from an insufficient number of markers and/or missing data (Pritchard *et al.*, 2000).

F_{ST} between the Australian and Chinese samples was generally low (0.09), and an AMOVA indicated that 88% of the variation in the samples could be explained by the variation within samples from the same country. The *Cyp303a1*-associated loci exhibited the highest F_{ST} values, with 0.24 and 0.69 at *Cyp303down1* and *Cyp303down3*, respectively. F_{ST} at the remaining six loci did not exceed 0.2.

The Cyp303a1 *locus*

At seven of the eight loci, diversity in the MacIntyre Valley population did not markedly differ from the Chinese populations, although haplotype diversity appeared slightly elevated in the Australian samples (Table 2). Notably, *Cyp303down1* was an exception (the Australian population had about tenfold lower nucleotide diversity than the Chinese populations), consistent with other lines of evidence (extended LD, reduced nucleotide diversity, and a skewed frequency spectrum) that support the occurrence of a sweep at or around this locus in Australian populations (Song *et al.*, 2015). We explored the patterns around the *Cyp303a1* locus in the Chinese populations, in particular the frequency of the *Del200* haplotype (the swept allele in Australia which is characterized by an intronic 200 bp deletion) and the molecular signatures downstream of the indel polymorphism. Analysis of the amplicon lengths revealed that only 1/94 individuals carried the deletion, and sequencing confirmed the presence of the *Del200* haplotype in this individual, YC-39 (fig. 3). Downstream 1 kb of the indel, the most abundant haplotype was one that was shared between Australia and China but approximately 27% of the Chinese individuals had a second haplotype that was not observed in MacIntyre Valley (although the small sample size does not preclude the possibility that this second haplotype is present at a low-to-intermediate frequency in Australia). This locus, *Cyp303down1* also exhibited a pattern of extended LD ($r^2 > 0.2$ beyond 400 bps) that differed from that of other loci where r^2 typically declines to below 0.2 within 200 bps (Supplementary fig. S3).

A similar pattern of differentiation was observed between the Australian ($n = 22$) and Chinese ($n = 69$) populations in the region 3 kb downstream of the *Cyp303* coding sequence (fig. 4). The major haplotype in Australia (haplotype 2) was seen in two individuals from China, one of which was YC-39. As YC-39 was also the sole carrier of the *Del200* allele among the 94 Chinese samples in this dataset, we infer that YC-39 carries a haplotype that originated from Australia, and conclude that this locus is highly informative for determining whether an *H. armigera* individual originated from Australia or China. The *Cyp303down3* haplotype is characterized by 14 polymorphic sites, three of which show a distinct difference in allele frequencies between Australia and China (Supplementary fig. S4).

Are there other Z-loci that show strong differentiation between Australian and Chinese populations?

The strong discordance between the inter-continental population structure observed at *Cyp303a1* and the other loci surveyed motivated us to extend the study to more loci. We wished to know whether the *Cyp303a1* region was unique – perhaps because of its selective sweep, and whether there were other loci that showed strong inter-continental structure. We therefore turned our attention to the whole-genome sequence dataset generated by Anderson *et al.* (2016), and examined the Z-chromosome sequences from 19 Australian and four Chinese samples in that dataset. In particular, we measured the weighted F_{ST} (fig. 4) across the *H. armigera* Z using sliding windows of 500 SNPs (see Materials and methods). The plot has two striking features relative to such analyses from other species (e.g. Reinhardt *et al.*, 2014): firstly, the F_{ST} values are distributed over a surprisingly large range (as high as $F_{\mathrm{ST}} = 0.6$), and secondly, the extremely high F_{ST} values are clustered along the chromosome so that distinct peaks are observed in the plot. This completely independent analysis is concordant with the EPIC dataset to the extent that the *Cyp303a1* locus is highly differentiated between Australia and China and only one of the six loci (the *period* gene) that showed minimal population differentiation in the EPIC dataset, shows any Australia–China differentiation in the re-sequencing dataset.

As the whole Z-chromosome dataset had an unbalanced sampling design with four moths sampled from China and 19 from Australia, we used a permutation approach to evaluate the extent to which the empirical dataset deviated from the distribution of F_{ST} values that could be expected given the small sample sizes. Four chromosomes were sampled at random from the combined set of 23 and F_{ST} was calculated. This was repeated 1000 times to give 1000 datasets. Regions in the empirical dataset (represented by black dots in fig. 3) that fall outside of a distribution generated by 1000 permutations (gray dots) cannot be attributed to population sampling biases. It is noteworthy that the 95th percentile of the empirical dataset is substantially higher than that of the permuted datasets (represented by black and gray dashed lines in the figure, respectively) supporting the contention of population differentiation between Australia and China.

There are 302 windows that cross a threshold defined by the top 95% most differentiated windows. These loci are therefore of interest as they provide useful markers of population differentiation. The 302 sliding windows can be grouped

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Table 2. Haplotype diversity and nucleotide diversity for Nanpi, Yancheng, and MacIntyre Valley at eight loci.

Locus	Population	Sample size (*n*)	Haplotype diversity	Nucleotide diversity (π)
Cyp303down1	Nanpi	15	0.96	0.025
	Yanch.	18	0.93	0.015
	MV	8	0.64	0.002
Cyp303down3	Nanpi	26	0.71	0.012
	Yanch.	43	0.65	0.008
	MV	22	0.65	0.005
Cyp305b1	Nanpi	17	0.95	0.020
	Yanch.	19	0.94	0.021
	MV	21	0.97	0.024
Period	Nanpi	37	0.76	0.021
	Yanch.	50	0.84	0.020
	MV	17	0.97	0.020
Phc	Nanpi	25	0.92	0.042
	Yanch.	34	0.89	0.036
	MV	16	0.95	0.037
Scap	Nanpi	34	0.95	0.036
	Yanch.	45	0.90	0.032
	MV	11	0.95	0.045
Tc	Nanpi	38	0.81	0.009
	Yanch.	51	0.68	0.007
	MV	19	0.99	0.014
Tpi	Nanpi	17	0.95	0.052
	Yanch.	31	0.95	0.044
	MV	19	0.96	0.059

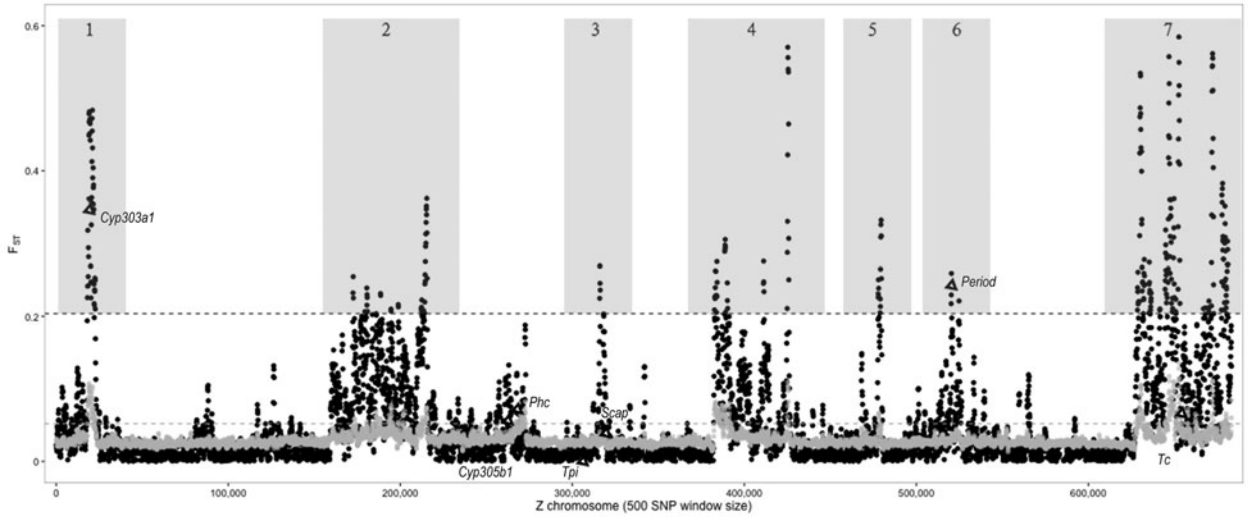

Fig. 4. Sliding window analysis of weighted F_{ST} across the *Helicoverpa armigera* Z-chromosome for Australian ($n = 19$) and Chinese individuals ($n = 4$). Gray dots represent the 95th percentile of F_{ST} values in each window for 1000 random permutations of the dataset while black dots represent the values from the empirical dataset. The gray dashed line indicates the 95th percentile of the F_{ST} values for the permuted datasets while the black dashed line indicates the 95th percentile of the F_{ST} values for the empirical dataset. Seven broad visually defined regions showing high population structure are numbered and indicated by gray shading. The loci that were studied in the initial dataset are shown and their values in this dataset are shown as triangles.

into 32 contiguous regions (ranging in size from 9 to 268 kb) and further simplified into seven broad arbitrarily-defined regions based on visual inspection. We used the recently available *H. armigera* genome sequence (Pearce *et al.*, 2017) to identify the loci in these differentiated regions – these genes are listed in Table 3 along with credible homology-based functional annotations we can assign to them. Notably, the list includes members of the *ABC* transporter gene family which have been implicated in insecticide resistance (Srinivas *et al.*, 2004; Buss & Callaghan, 2008). The list also includes *period* (used as an EPIC marker in this study) and *CCR4-Not* which shows a significantly negative Tajima's *D* value in the Chinese populations.

Discussion

Our analysis of the EPIC data found that Australian and Chinese populations of *H. armigera* harbored similar levels of

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Table 3. Functional annotations for F_{ST} outlier loci identified from the sliding window analysis across the *Helicoverpa armigera* Z-chromosome. HAOG numbers refer to the *H. armigera* official gene set annotated in Pearce *et al.* (2017). Region numbers correspond to those shown in fig. 3.

HAOGS reference	Annotation	Region
HaOG205589	BMORI:facilitated trehalose transporter Tret1-like	1
HaOG203061	BMORI:macrophage migration inhibitory factor-like	1
HaOG203059	BMORI:mitochondrial thiamine pyrophosphate carrier-like	1
HaOG216810	BMORI:patj homolog	1
HaOG213506	BMORI:putative zinc finger protein 724-like	1
HaOG203065	BMORI:RNA pseudouridylate synthase domain-containing protein 1-like	1
HaOG214178	BMORI:T-box protein H15-like	1
HaOG203062	BMORI:ubiquitin-conjugating enzyme E2 G2-like	1
HaOG203060	CELEG:P91119 Probable $3',5'$-cyclic phosphodiesterase pde-5 BMORI:probable $3',5'$-cyclic phosphodiesterase pde-5-like	1
HaOG200007	CYP303A1_Ha	1
HaOG203063	DMELA:A1Z6X0 CG12164	1
HaOG203066	DMELA:Q8SYS6 RE37593p BMORI:$\alpha-1$,3-mannosyl-glycoprotein 4-β-*N*-acetylglucosaminyltransferase B-like	1
HaOG213125	DMELA:Q7Z020 Transient receptor potential cation channel subfamily A member 1 BMORI:transient receptor potential cation channel subfamily A member 1-like	2
HaOG213127	DMELA:Q9VSQ2 CG13310 BMORI:uncharacterized protein LOC101744434	2
HaOG211956	BMORI:tRNA-specific adenosine deaminase 1-like isoform X2	2
HaOG213135	DMELA:A8JUT1 Furrowed, isoform B BMORI:sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1-like	2
HaOG214578	BMORI:rho GTPase-activating protein 8-like	2
HaOG206575	BMORI:titin2	2
HaOG214577	DMELA:Q7YZH1 PHD finger protein rhinoceros BMORI:LOW QUALITY PROTEIN: PHD finger protein rhinoceros-like	2
HaOG214580	DMELA:Q9VS49 CG8600, isoform A	2
HaOG214579	DMELA:Q9W0Q2 Peptidyl-prolyl cis-trans isomerase BMORI:peptidylprolyl isomerase	2
HaOG204759	DMELA:A1ZAJ7 CG15615	3
HaOG216091	BMORI:chromatin-remodeling complex ATPase chain Iswi-like	4
HaOG209512	BMORI:filaggrin-like	4
HaOG216089	BMORI:sodium-dependent noradrenaline transporter	4
HaOG200348	HaABCB1 ALT:HaABC-B-01-1-F	4
HaOG200349	HaABCB2 ALT:HaABC-B-01-2-F	4
HaOG200350	HaABCB3 ALT:HaABC-B-01-3-F	4
HaOG216075	DMELA:A8JMD5 CG34356 BMORI:SCY1-like protein 2-like	4
HaOG216074	BMORI:EF-hand calcium-binding domain-containing protein 1-like	4
HaOG216055	DMELA:P45447 Ecdysone-induced protein 78C BMORI:ecdysone-inducible protein E75-like	4
HaOG209024	BMORI:niemann-Pick C1-like protein 1-like	4
HaOG208377	BMORI:succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial-like	4
HaOG208378	BMORI:succinate dehydrogenase cytochrome b560 subunit, mitochondrial-like	4
HaOG207326	BMORI:insulin-like growth factor 2 mRNA-binding protein 1-like isoform X1	5
HaOG215518	BMORI:pogo transposable element with ZNF domain-like	5
HaOG207329	BMORI:exportin-1-like	5
HaOG212018	BMORI:LIM domain-containing protein jub-like	5
HaOG214354	BMORI:methylcrotonoyl-CoA carboxylase subunit α, mitochondrial-like	5
HaOG207328	BMORI:rho guanine nucleotide exchange factor 11-like	5
HaOG207327	DMELA:E1JJM0 FI20063p1	5
HaOG215851	DMELA:P07663 Period circadian protein BMORI:period	6
HaOG203153	DMELA:O76933 Pentaxin-like protein BMORI:uncharacterized protein LOC101736996 isoform X1	7
HaOG203152	DMELA:Q7KW14 Coiled-coil domain-containing protein CG32809 BMORI:coiled-coil domain-containing protein AGAP005037-like	7
HaOG209864	DMELA:Q8MQJ5 CG31122, isoform B	7
HaOG203160	BMORI:protein bric-a-brac 1-like, partial	7
HaOG215757	BMORI:leucine-rich repeat serine/threonine-protein kinase 1-like	7
HaOG203170	BMORI:protein Daple-like	7
HaOG204924	BMORI:protein ELYS-like	7
HaOG213826	BMORI:threonine dehydratase, mitochondrial-like	7
HaOG203168	DMELA:E1JI71 CG32352, isoform E BMORI:biorientation of chromosomes in cell division protein 1-like 1-like	7
HaOG203164	DMELA:O76867 EG:100G10.7 protein BMORI:paraplegin-like	7
HaOG213827	DMELA:Q7KST5 CG8129, isoform A	7
HaOG203169	DMELA:Q9VRT9 CG13293 BMORI:uncharacterized protein LOC101746024	7
HaOG203162	DMELA:Q9W3N7 CG18624, isoform A	7
HaOG203161	HSAPI:A5YKK6 CCR4-NOT transcription complex subunit 1 BMORI:CCR4-NOT transcription complex subunit 1-like	7

Table 3. (*Cont.*)

HAOGS reference	Annotation	Region
HaOG203166	HSAPI:Q13946 High affinity cAMP-specific 3′,5′-cyclic phosphodiesterase 7A BMORI:uncharacterized protein LOC101739482	7
HaOG205332	BMORI:Fanconi anemia group M protein-like	7
HaOG205326	BMORI:carbonic anhydrase 1-like	7
HaOG205325	BMORI:H/ACA ribonucleoprotein complex non-core subunit NAF1-like	7
HaOG205324	DMELA:A8JUV8 Terribly reduced optic lobes, isoform G BMORI:basement membrane-specific heparan sulfate proteoglycan core protein-like	7
HaOG205330	DMELA:Q9W4Y2 PDF receptor BMORI:neuropeptide receptor B2	7
HaOG207317	DMELA:P47825 Transcription initiation factor TFIID subunit 4 BMORI:LOW QUALITY PROTEIN: transcription initiation factor TFIID subunit 4-like	7
HaOG207316	DMELA:Q8IQN2 CG5284, isoform B BMORI:H(+)/Cl(−) exchange transporter 3-like	7
HaOG207314	BMORI:LOW QUALITY PROTEIN: adenosine deaminase CECR1-like	7
HaOG207312	BMORI:neuropeptide receptor A18	7
HaOG207310	BMORI:olfactory receptor 60	7
HaOG207311	BMORI:trypsin-1-like	7
HaOG207313	DMELA:Q2HPH2 Peptide receptor GPCR BMORI:neuropeptide receptor A18	7
HaOG207315	DMELA:Q9VT28 Protein furry BMORI:LOW QUALITY PROTEIN: microtubule-associated serine/threonine-protein kinase 2-like	7
HaOG207296	BMORI:β carbonic anhydrase 1-like	7
HaOG207303	BMORI:G-protein-signaling modulator 2-like	7
HaOG207295	BMORI:mucin-17-like	7
HaOG207299	BMORI:ociad protein isoform 1	7
HaOG207293	BMORI:tetratricopeptide repeat protein 26-like isoform X1	7
HaOG207301	DMELA:O61734 Protein cycle BMORI:Cycle like factor b	7
HaOG207304	DMELA:Q9NFR7 Rapsynoid BMORI:G-protein-signaling modulator 2-like	7
HaOG207300	DMELA:Q9W1X9 OCIA domain-containing protein 1 BMORI:ociad protein isoform 2	7

nucleotide diversity. Furthermore, the average pairwise nucleotide divergence between the populations in the EPIC dataset is the same as that calculated in the independent dataset of Anderson *et al.* (2016) where $\pi = 0.028$. Haplotype diversity generally appears to be slightly elevated in the Australian samples (excluding *Cyp303a1* from the comparison), although there are instances where nucleotide diversity of the Nanpi population exceeds that of MacIntyre Valley. The elevated Australian diversity motivated us to examine all the datasets in the literature to consider whether nucleotide diversity could provide a clue as to where the ancestral *H. armigera* populations arose. Revisiting the mtDNA data in Behere *et al.* (2007) reveals that a Ugandan population of *H. armigera* has higher nucleotide diversity than the Australian population, although haplotype diversity of Australia still exceeds that of Uganda. Anderson *et al.* (2016) found more mitochondrial diversity in Australia but the larger number of Australian samples distorts the interpretation slightly. More data would be required to determine whether there is a geographic region that harbors more diversity than all others.

Another result from our EPIC analyses is the consistent signal of a negative Tajima's *D*, indicating an excess of rare variants in the Chinese populations. This result contrasts with that of Anderson *et al.* (2016) who report that Tajima's *D* was positive in eight of the nine populations they examined, the exception being the Australian population. Even when singleton sites are removed from our analysis, Tajima's *D* remains negative for 29 of the 40 loci, suggesting that our data are robust and sequencing errors do not overly inflate the overall negative pattern. We have two hypotheses to explain the discordance between the datasets. Firstly, all our loci are on the Z-chromosome, whereas the data from Anderson *et al.* (2016) come from across the genome (which consists of 31 chromosomes of roughly similar size). Sex chromosomes are subjected to more efficient selection than autosomes because recessive mutations are exposed in the heterogametic sex (Charlesworth *et al.*, 1987). Furthermore, female heterogamety means that Z-linked loci are subjected to higher mutation rates than autosomal loci because the Z-chromosome spends two-thirds of its time in males where spermatogenesis is more mutagenic than oogenesis (Vicoso & Charlesworth, 2006; Sackton *et al.*, 2014). Consequently, the combination of increased efficacy of selection and the increase in mutation rate could produce a relatively greater excess of rare variants on the Z-chromosome relative to the autosomes.

The second explanation is that the Tajima's *D* analysis of Anderson *et al.* (2016) is affected by the small sample size of populations (most are between three and five samples per population). Since Tajima's *D* test is an allele frequency spectrum test, power is much lower with small sample sizes. If we exclude these populations on the basis of small sample size, we are left with the one sample where Anderson *et al.* (2016) have a moderate depth for their Tajima's *D* test, the Australian population ($n = 17$) which has a negative Tajima's *D* value. This set of individuals is in fact a subset of the 19 examined here, but is genome-wide rather than limited to the Z-chromosome, as our analysis is. Thus, we interpret the negative Tajima's *D* as an evidence of a population expansion in the evolutionary history of *H. armigera.* It should be noted that such an expansion would be a much older event and not the recent incursion into the Americas (where insufficient time has passed to leave behind such molecular signatures). It is not clear if this expansion relates to the spread of agriculture or to an even older event such as the availability of new niches during favorable climatic conditions in the Pleistocene (the estimated period of divergence with *H. zea*). More data and

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

sophisticated population models would be required to date the expansion.

Our data also speak to the question of whether *H. armigera* from Australia belong to a separate subspecies (Common, 1953). Most of the eight loci where we have EPIC data from China and Australia show no population differentiation between the two countries; the exceptions are the *Cyp303a1*-associated markers. Our previous study (Song *et al.*, 2015) which characterized the *Cyp303a1 Ins200* and *Del200* haplotypes revealed extremely diverged alleles – so diverged that they may be considered 'comet alleles' (Staubach *et al.*, 2012) which may have arisen from introgression with some other unknown species. This suggests two contrasting models to explain the results of the genome-wide STRUCTURE analysis reported by Anderson *et al.* (2016). In one model, the signal for population differentiation comes from a smallish number of highly diverged loci (like *Cyp303a1*) that are distributed patchily across Australian genomes. These may derive from an introgression event in a way analogous to the 'Neanderthal' introgression footprints in non-African modern humans (Green *et al.*, 2010). In the other model, the Australian *H. armigera* differ by small amounts diffused throughout the genome, reflecting variants that accumulated in Australian moths when gene flow to other parts of the world was less than it appears to be now; in other words, a more conventional isolation-by-distance model. The gene flow that established such a difference would have to be less than what is observed currently based on reports such as the repeated incursions of *H. armigera* into South America in recent years.

The F_{ST} scan across the Z-chromosome offers some support for the 'Neanderthal' model. There appear to be distinct clusters of elevated F_{ST} arising from a substantially lower baseline. Under an isolation-by-distance model, we may have seen more intermediate values of F_{ST}. These inter-population divergence patterns will be influenced by many factors including recombination rate heterogeneity and the timing and abundance of local selective sweeps. Future studies should test other global populations for Z-linked population structure and should contrast these within-species data to between-species divergence with closely related species.

At a pragmatic level, we have identified multiple loci including *Cyp303a1* as useful markers to identify gene flow from Australia. We believe that the Chinese YC-39 individual, which shows a distinctly Australian haplotype at regions close to *Cyp303a1*, tells us that inter-continental gene flow can occur. Such gene flow would contribute to a general lack of population structure in worldwide populations of *H. armigera*, and the most informative loci to track moth movements will be those that have recently increased in frequency due to strong positive selection locally. Most likely among those will be insecticide resistance genes, which currently include cadherin, ABC transporters, *Cyp337b3*, and *kdr* genes (Martinez-Torres *et al.*, 1997; Head *et al.*, 1998; Gahan *et al.*, 2001, 2010; Joußen *et al.*, 2012). In fact, *Cyp303a1*, which Song *et al.* (2015) found to show the hallmarks of a selective sweep, and which belongs to the cytochrome P450 multigene family that includes classic detoxifying enzymes, may actually be a resistance allele itself. We note that Daly & Fisk (1998) report a Z-linked endosulfan resistance in Australian populations, so *Cyp303a1* may be a candidate gene worth testing for this trait. Similarly, the *ABCB* subfamily of *ABC* transporters is of particular interest as they encompass a class of proteins known as P-glycoproteins which have been shown to be involved in transporting xenobiotics out of the cell (reviewed in Buss & Callaghan, 2008).

Conclusions

Our study has led us to the following four conclusions. Firstly, specific Z-linked loci (such as *Cyp303a1*) provide useful markers for inter-continental population structure in *H. armigera*. Secondly, an Australian haplotype at the *Cyp303a1* gene is found in China and the length of this haplotype indicates that there has been recent gene flow from Australia to China. Thirdly, our deep sampling of Chinese allelic diversity reveals that low frequency, diverse haplotypes exist at multiple Z-loci within Chinese populations, and even when highly diverged haplotypes are excluded from the analyses, there is a skew toward rare variants in the dataset generally that we interpret as an evidence of population expansion. Fourthly, remarkable localized clusters of high F_{ST} values occur across the *H. armigera* Z-chromosome, perhaps supporting the proposition that the *H. armigera conferta* subspecies may originate from an introgression event from an unknown species.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S0007485318000081

References

Anderson, C.J., Tay, W.T., McGaughran, A., Gordon, K. & Walsh, T.K. (2016) Population structure and gene flow in the global pest, *Helicoverpa armigera*. *Molecular Ecology* **25**(21), 5296–5311.

Arnemann, J.A., James, W.J., Walsh, T.K., Guedes, J.V.C., Smagghe, G., Castiglioni, E. & Tay, W.T. (2016) Mitochondrial DNA *COI* characterization of *Helicoverpa armigera* (Lepidoptera: Noctuidae) from Paraguay and Uruguay. *Genetics and Molecular Research* **15**, 15028292.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376, doi: 10.1371/journal.pone.0003376.

Battlay, P., Schmidt, J.M., Fournier-Level, A. & Robin, C. (2016) Genomic and transcriptomic associations identify a new insecticide resistance phenotype for the selective sweep at the *Cyp6g1* locus of *Drosophila melanogaster*. *G3 (Bethesda)* **6**, 2573–2581.

Behere, G., Tay, W., Russell, D., Heckel, D., Appleton, B., Kranthi, K. & Batterham, P. (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, G.T., Tay, W.T., Russell, D.A., Kranthi, K.R. & 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. ISSN 1932-6203. doi: 10.1371/journal.pone.0053448.

Browning, S. R. & Browning, B. L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084–1097.

Buss, D. S. & Callaghan, A. (2008) Interaction of pesticides with p-glycoprotein and other ABC proteins: a survey of the possible importance to insecticide, herbicide and fungicide resistance. *Pesticide Biochemistry and Physiology* **90**, 141–153.

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Charlesworth, B., Coyne, J.A. & Barton, N.H. (1987) The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist* **130**, 113–146.

Cho, S., Mitchell, A., Mitter, C., Regier, J., Matthews, M. & Robertson, R. (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.

Collart, M.A. & Panasenko, O.O. (2012) The CCR4-Not complex. *Gene* **492**, 42–53.

Common, I. (1953) The Australian species of *Heliothis* (Lepidoptera: Noctuidae) and their pest status. *Australian Journal of Zoology* **1**, 319–344.

Czepak, C., Albernaz, K.C., Vivan, L.M., Guimarães, H.O. & Carvalhais, T. (2013) First reported occurrence of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in Brazil. *Pesquisa Agropecuária Tropical* **43**, 110–113.

Daly, J.C. (1993) Ecology and genetics of insecticide resistance in *Helicoverpa armigera*: interactions between selection and gene flow. *Genetica* **90**, 217–226.

Daly, J.C. & Fisk, J.H. (1998) Sex-linked inheritance of endosulphan resistance in *Helicoverpa armigera*. *Heredity* **81**, 55–62.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. & McVean, G. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

Earl, D.A. & vonHoldt, B.M. (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359–361.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. & Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379. doi: 10.1371/journal.pone.0019379.

Evanno, G., Regnaut, S. & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Falush, D., Wirth, T., Linz, B., Pritchard, J.K., Stephens, M., Kidd, M., Blaser, M.J., Graham, D.Y., Vacher, S., Perez-Perez, G.I., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M. & Suerbaum, S. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585.

Fitt, G.P. (1994) Cotton pest management: Part 3. An Australian perspective. *Annual Review of Entomology* **39**, 543–562.

Gahan, L.J., Gould, F. & Heckel, D.G. (2001) Identification of a gene associated with *Bt* resistance in *Heliothis virescens*. *Science* **293**, 857–860.

Gahan, L.J., Pauchet, Y., Vogel, H. & Heckel, D.G. (2010) An ABC transporter mutation is correlated with insect resistance to *Bacillus thuringiensis Cry*1*ac* toxin. *PLoS Genetics* **6**, e1001248. doi: 10.1371/ journal.pgen.1001248.

Garud, N.R., Messer, P.W., Buzbas, E.O. & Petrov, D.A. (2015) Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics* **11**, e1005004. doi: 10.1371/journal.pgen. 1005004.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T. & Martin, J.F. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**, 245.

Gordon, K.H.J., Tay, W.T., Collinge, D., Williams, A. & Batterham, P. (2010) Genetics and molecular biology of the major crop pest genus *Helicoverpa*. pp. 219–238 *in* Goldsmith, M.R. & Marec, F. (*Eds*) *Molecular Biology and Genetics of the Lepidoptera*. CRC Press, Boca Raton, FL, USA ISBN 978-1-4200-6020-1.

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Resources* **5**, 184–186.

Gouy, M., Guindon, S. & Gascuel, O. (2010) Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**, 221–224.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Pääbo, S. (2010) A draft sequence of the Neandertal genome. *Science* **328**, 710–722.

Head, D.J., McCaffery, A.R. & Callaghan, A. (1998) Novel mutations in the *para*-homologous sodium channel gene associated with phenotypic expression of nerve insensitivity resistance to pyrethroids in Heliothine lepidoptera. *Insect Molecular Biology* **7**, 191–196.

Hill, W.G. & Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**, 54–78.

Jombart, T. & Ahmed, I. (2011) *Adegenet* 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071.

Joußen, N., Agnolet, S., Lorenz, S., Schöne, S.E., Ellinger, R., Schneider, B. & Heckel, D.G. (2012) Resistance of Australian *Helicoverpa armigera* to fenvalerate is due to the chimeric P450 enzyme CYP337B3. *Proceedings of the National Academy of Sciences* **109**, 15206–15211.

Kamvar, Z.N., Tabima, J.F. & Grünwald, N.J. (2014) *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281. ISSN 2167-8359. doi: 10.7717/peerj.281.

Lakey, A., Labeit, S., Gautel, M., Ferguson, C., Barlow, D.P., Leonard, K. & Bullard, B. (1993) Kettin, a large modular protein in the Z-disc of insect muscles. *The EMBO Journal* **12**, 2863–2871.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948.

Leite, N.A., Alves-Pereira, A., Corrêa, A.S., Zucchi, M.I. & Omoto, C. (2014) Demographics and genetic variability of the New World bollworm (*Helicoverpa zea*) and the Old World bollworm (*Helicoverpa armigera*) in Brazil. *PLoS ONE* **9**, e113286. ISSN 1932-6203. doi: 10.1371/journal.pone. 0113286.

Mallet, J., Korman, A., Heckel, D.G. & King, P. (1993) Biochemical genetics of *Heliothis* and *Helicoverpa* (Lepidoptera: Noctuidae) and evidence for a founder event in *Helicoverpa zea*. *Annals of the Entomological Society of America* **86**, 189–197.

Martinez-Torres, D., Devonshire, A.L. & Williamson, M.S. (1997) Molecular studies of knockdown resistance to pyrethroids: cloning of domain II sodium channel gene sequences from insects. *Pesticide Science* **51**, 265–270.

Mastrangelo, T., Paulo, D.F., Bergamo, L.W., Morais, E.G.F., Silva, M., Bezerra-Silva, G. & Azeredo-Espin, A.M.L. (2014)

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press

Detection and genetic diversity of a heliothine invader (Lepidoptera: Noctuidae) from north and northeast of Brazil. *Journal of Economic Entomology* **107**, 970–980.

Matthews, M. (1999) Heliothine moths of Australia: a guide to pest bollworms and related noctuid groups. *Monographs on Australian Lepidoptera, Vol. 7*. Melbourne, CSIRO Publishing. ISBN 0-643-06305-6.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303.

Mitchell, A. & Gopurenko, D. (2016) DNA barcoding the Heliothinae (Lepidoptera: Noctuidae) of Australia and utility of DNA barcodes for pest identification in *Helicoverpa* and relatives. *PLoS ONE* **11**, e0160895. ISSN 1932-6203. doi: 10.1371/journal.pone.0160895.

Murúa, M.G., Scalora, F.S., Navarro, F.R., Cazado, L.E., Casmuz, A., Villagrán, M.E., Lobos, E. & Gastaminza, G. (2014) First record of *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Argentina. *Florida Entomologist* **97**, 854–856.

Nibouche, S., Bues, 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.

Nielsen, R. (2005) Molecular signatures of natural selection. *Annual Review Of Genetics* **39**, 197–218.

Pearce, S.L., Clarke, D.F., East, P.D., Elfekih, S., Gordon, K.H.J., Jermiin, L.S., McGaughran, A., Oakeshott, J.G., Papanikolaou, A., Perera, O.P., Rane, R.V., Richards, S., Tay, W.T., Walsh, T.K., Anderson, A., Anderson, C.J., Asgari, S., Board, P.G., Bretschneider, A., Campbell, P.M., Chertemps, T., Christeller, J.T., Coppin, C.W., Downes, S.J., Duan, G., Farnsworth, C.A., Good, R.T., Han, L.B., Han, Y. C., Hatje, K., Horne, I., Huang, Y.P., Hughes, D.S.T., Jacquin-Joly, E., James, W., Jhangiani, S., Kollmar, M., Kuwar, S.S., Li, S., Liu, N.Y., Maibeche, M.T., Miller, J.R., Montagne, N., Perry, T., Qu, J., Song, S.V., Sutton, G.G., Vogel, H., Walenz, B.P., Xu, W., Zhang, H.J., Zou, Z., Batterham, P., Edwards, O.R., Feyereisen, R., Gibbs, R.A., Heckel, D.G., McGrath, A., Robin, C., Scherer, S.E., Worley, K.C. and Wu, Y.D. (2017) Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive *Helicoverpa* pest species. *BMC Biology* **15**, 63.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M. J. & Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575.

Rašić, G., Filipović, I., Weeks, A. R. & Hoffmann, A. A. (2014) Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics* **15**, 275.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing.

Reinhardt, J.A., Kolaczkowski, B., Jones, C.D., Begun, D.J. & Kern, A.D. (2014) Parallel geographic variation in *Drosophila melanogaster*. *Genetics* **197**, 361–373.

Rohland, N. & Reich, D. (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* **22**, 939–946.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Sackton, T.B., Corbett-Detig, R.B., Nagaraju, J., Vaishna, L., Arunkumar, K.P. & Hartl, D.L. (2014) Positive selection drives faster-Z evolution in silkmoths. *Evolution* **68**, 2331–2342.

Slatkin, M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477–485.

Song, S.V., Downes, S., Parker, T., Oakeshott, J.G. & Robin, C. (2015) High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep. *Heredity* (*Edinb*) **115**, 460–470.

Sosa-Gómez, D.R., Specht, A., Paula-Moraes, S.V., Lopes-Lima, A., Yano, S.A.C., Micheli, A., Morais, E.G.F., Gallo, P., Pereira, P.R. V.S., Sal-vadori, J.R., Botton, M., Zenker, M.M. & Azevedo-Filho, W.S. (2016) Timeline and geographical distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera, Noctuidae: Heliothinae) in Brazil. *Revista Brasileira de Entomologia* **60**, 101–104.

Srinivas, R., Udikeri, S. S., Jayalakshmi, S. K. & Sreeramulu, K. (2004) Identification of factors responsible for insecticide resistance in *Helicoverpa armigera*. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **137**, 261–269.

Staubach, F., Lorenc, A., Messer, P.W., Tang, K., Petrov, D.A. & Tautz, D. (2012) Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics* **8**, e1002891.

Tabashnik, B.E., Gould, F. & Carrière, Y. (2004) Delaying evolution of insect resistance to transgenic crops by decreasing dominance and heritability. *Journal of Evolutionary Biology* **17**, 904–912.

Tay, W., Behere, G., Batterham, P. & Heckel, D. (2010) Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. *BMC Evolutionary Biology* **10**, 144.

Tay, W.T., Soria, M.F., Walsh, T., Thomazoni, D., Silvie, P., Behere, G.T., Anderson, C. & Downes, S. (2013) A brave new world for an Old World pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. *PLoS ONE* **8**, e80134. doi: 10.1371/journal.pone.0080134.

Vicoso, B. & Charlesworth, B. (2006) Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics* **7**, 645–653.

Wallar, B.J. & Alberts, A.S. (2003) The formins: active scaffolds that remodel the cytoskeleton. *Trends in Cell Biology* **13**, 435–446.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, vol. **1**. New York, Springer, p. 3.

Zhang, D.X. (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends in Ecology & Evolution* **19**, 507–509.

Zhou, X., Faktor, O., Applebaum, S.W. & Coll, M. (2000) Population structure of the pestiferous moth *Helicoverpa armigera* in the Eastern Mediterranean using RAPD analysis. *Heredity* **85**, 251–256.

Zhu, L., Tatsuke, T., Li, Z., Mon, H., Xu, J., Lee, J.M. & Kusakabe, T. (2012) Molecular cloning of BmTUDOR-SN and analysis of its role in the RNAi pathway in the silkworm, *Bombyx mori* (Lepidoptera: Bombycidae). *Applied Entomology and Zoology* **47**, 207–215.

Zhu, L., Tatsuke, T., Mon, H., Li, Z., Xu, J., Lee, J.M. & Kusakabe, T. (2013) Characterization of Tudor-SN-containing granules in the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology* **43**, 664–674.

https://doi.org/10.1017/S0007485318000081 Published online by Cambridge University Press