Author/s:
Zhang, J;Zhang, F;Tay, WT;Robin, C;Shi, Y;Guan, F;Yang, Y;Wu, Y

Title:
Population genomics provides insights into lineage divergence and local adaptation within the cotton bollworm

Date:
2022-07-01

Citation:
Zhang, J., Zhang, F., Tay, W. T., Robin, C., Shi, Y., Guan, F., Yang, Y. & Wu, Y. (2022). Population genomics provides insights into lineage divergence and local adaptation within the cotton bollworm. Molecular Ecology Resources, 22 (5), pp.1875-1891. https://doi.org/10.1111/1755-0998.13581.

DR. WEE TEK TAY (Orcid ID : 0000-0002-8451-0811)
DR. YIDONG WU (Orcid ID : 0000-0003-3456-3373)

Article type : Resource Article

In preparation for: Molecular Ecology Resources

Running title: Population genomics of *Helicoverpa armigera*

Population genomics provides insights into lineage divergence and local adaptation within the cotton bollworm

Jianpeng Zhang[1], Feng Zhang[1], Wee Tek Tay[2], Charles Robin[3], Yu Shi[1], Fang Guan[1], Yihua Yang[1], Yidong Wu[1,*]

[1] College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China

[2] CSIRO Black Mountain Laboratories, Clunies Ross Street, ACT 2601, Australia

[3] School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia.

* Corresponding author. E-mail: wyd@njau.edu.cn

ABSTRACT

The cotton bollworm *Helicoverpa armigera* is a cosmopolitan pest and its diverse habitats plausibly contribute to the formation of diverse lineages. Despite the significant threat it poses to economic crops worldwide, its evolutionary history and genetic basis of local adaptation are poorly

understood. In this study, we *de novo* assembled a high-quality chromosome-level reference genome of *H. a. armigera* (contig N50 = 7.34 Mb), with 99.13% of the HaSCD2 assembly assigned into 31 chromosomes (Z-chromosome + 30 autosomes). We constructed an ultra-dense variation map across 14 cotton bollworm populations and identified a novel lineage in northwestern China. Historical inference showed that effective population size changes coincided with global temperature fluctuation. We identified nine differentiated genes in the three *H. armigera* lineages (*H. a. armigera*, *H. a. conferta*, and the new northwestern Chinese lineage), of which *per* and *clk* genes are involved in circadian rhythm. Selective sweep analyses identified a series of GO categories related to climate adaptation, feeding behavior and insecticide tolerance. Our findings reveal fundamental knowledge of the local adaptation of different cotton bollworm lineages and will guide the formulation of cotton bollworm management measures at different scales.

Keywords: cotton bollworm, population genomics, genetic structure, local adaptation

INTRODUCTION

The cotton bollworm, *Helicoverpa armigera* (Lepidoptera: Noctuidae), is a devastating pest. It can feed on a broad range of agricultural and horticultural plants, including cotton, soybean, maize and sunflower (Fitt, 1989). It is capable of long-range migration, facultative diapause and high fecundity, and these features enable the cotton bollworm to successfully colonize and flourish in various ecological niches, and therefore, it can be seen as a model species for studying local adaptation in Lepidoptera (Jones, Parry, Tay, Reynolds, & Chapman, 2019; Wu & Guo, 2005). Until its recent invasion into the Americas (Tay et al., 2013), the cotton bollworm was naturally distributed throughout Asia, Europe and Africa (known as *H. a. armigera*, *sensu* Hardwick, 1965) as well as Oceania where the subspecies, *H. a. conferta* is found. A third subspecies confined to the Canton Island (Republic of Kiribati) in the South Pacific, *H. a. commoni,* was also proposed by Hardwick (1965) based on body size and coloration differences.

Over the last decade, numerous investigations on the cotton bollworm radically reinvigorated our understanding of the population genomics of this major agricultural pest (Jones et al., 2019; Van Leeuwen, Dermauw, Mavridis, & Vontas, 2020). In particular, the previous short-read whole-genome sequencing efforts generated the draft genomes of *H. a. conferta* and its sister species *H. zea* (Pearce et al., 2017), which greatly promoted our understanding of insecticide

resistance, biotic incursion, population admixture, and insect-plant interactions (Anderson et al., 2016; Guo et al., 2020; Pearce et al., 2017; Valencia-Montoya et al., 2020). However, the inherent deficiencies, e.g., fragmentary sequences and numerous N-gap regions frustrate efforts to fully characterize genotype-phenotype relationships. Meanwhile, despite of their morphological similarity, genomic data has supported Hardwick's (1965) taxonomic classification separating *H. a. armigera* and *H. a. conferta* into two allopatric and genetically distinct subspecies (Anderson, Tay, McGaughran, Gordon, & Walsh, 2016). Thus, it is necessary to account for population structure and not rely upon genomic data sourced predominantly from a single subspecies to build a comprehensive management framework. A perfect illustration of this is the desire to understand and learn from the recent establishment of expanding *H. a. armigera* populations in the South American continent (Arnemann et al., 2019; Tay et al., 2013, 2017) and the Caribbean Islands (Jones et al., 2019). Furthermore, recent methodological and technological advances in third generation sequencing (TGS) have greatly improved the reference genomes of different agricultural pests (Fritz et al., 2018; Quan et al., 2019; Zhu et al., 2017). Compared to next generation sequencing (NGS) technology, TGS is able to generate longer reads which can span many repetitive regions and provide a more comprehensive genomic portrait (Sedlazeck, Lee, Darby, & Schatz, 2018).

Elucidating the dynamics, composition and relationships of different populations has practical benefits for pest management. In particular, unrecognized adaptive introgression may render local pest management unsuccessful. For example in 2013, it was confirmed that *H. a. armigera* had invaded Brazil (Czepak, Albernaz, Vivan, Guimarães, & Carvalhais, 2013; Tay et al., 2013). Invaded individuals were subsequently shown to hybridize with the previously isolated sister species *H. zea*, resulting in the introgression of an allele conferring resistance to fenvalerate and cypermethrin, *CYP337B3* (Anderson et al., 2018; Joußen et al., 2012; Valencia-Montoya et al., 2020; Walsh et al. 2018). Other insecticide resistance alleles and loci that counter diverse insecticides including *Bacillus thuringiensis* (Bt) toxins have been reported in different cotton bollworm populations and could also spread through populations, subspecies or species hybrids (Tabashnik & Carrière, 2017). For example, cotton bollworm in northern China has evolved high-level resistances to synthetic pyrethroid and organophosphate insecticides since the 1990s, whereas populations in northwestern China remained sensitive as recently as 2013 (Yang, Li, & Wu, 2013). A dominant point mutation conferring dominant resistance to Cry1Ac was identified in

northern China populations but it has not been detected in other regions (Jin et al., 2018). With globalization of trade and human movement, regions are less isolated and pests are more likely to spread (Tay & Gordon, 2019). Consequently, the invasion risks of noctuid pests have greatly elevated despite of strict quarantine measures worldwide.

Here, we combined single-molecule real-time (SMRT) sequencing and genetic map to assemble a chromosome-level genome of *H. a. armigera*. We re-sequenced the whole genomes of 141 cotton bollworm from China and downloaded 30 *Helicoverpa* complex data. By the population genomics approach, we explored the population structure and demographic history within the cotton bollworm. We also identified various highly divergent genomic regions that suggested climate, host plants and insecticides have contributed to shaping the cotton bollworm populations. Our research provides novel insights into adaptation by the cotton bollworm as well as valuable genomic resources for management of this major agricultural pest.

MATERIALS AND METHODS

SCD strain and sequencing for genome assembly

The SCD strain (*H. a. armigera*), used for genome assembly, was originally sampled from Africa in the 1970s and continuously reared on artificial diets in our laboratory (Nanjing, China) without exposure to any chemical pesticides or Bt toxins (Jin et al., 2018). For this study, the SCD strain was bred through three generations of successive sib-crossing to reduce its heterozygosity. Subsequently, we selected a single SCD male (ZZ) *H. a. armigera* pupa for total genomic DNA isolation using the phenol/chloroform method. The purity and concentration of the extracted DNA were measured using OneDrop™ 1000 Spectrophotometer (OneDrop Technologies). The extracted DNA was used to construct Illumina library with 350-bp insert size and Pacific Biosciences (PacBio) library with 20-kb insert size following their respective standard protocols. The short-read library was sequenced on Illumina HiSeq Xten platform using the 150-bp paired-end (PE) strategy, and seven single-molecule real-time (SMRT) long-read libraries were sequenced on PacBio Sequel platform.

Genome survey and *de novo* genome assembly

We summarized the main processes and their corresponding programs throughout genome assembly

as illustrated in Figure S1. Qualified Illumina data (Phred score > 15) were used to count *k*-mers (*k* = 17) by khist.sh built-in BBTools v38.49 (https://doi.org/jgi.doe.gov/data-and-tools/bbtools/), then genome size and heterozygosity rate were estimated using GenomeScope v1.0.0 (Vurture et al., 2017). The subreads from seven SMRT cells were merged and converted to FASTA format using SAMtools v0.1.19 (Li et al., 2009). The Canu v1.8 (Koren et al., 2017) and Falcon v1.3.0 (Chin et al., 2016) were applied to *de novo* assemble draft assemblies, respectively. The advanced options "batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50" were selected in Canu to maintain haplotype separation. These two draft assemblies had complementary contiguity, we therefore merged them using quickmerge v0.3 (Chakraborty, Baldwin-Brown, Long, & Emerson, 2016) to salvage missing and/or wrong segments. Alternative haplotypes were removed from the merged assembly using Purge Haplotigs v1.1.0 (Roach, Schmidt, & Borneman, 2018). Non-redundant sequences were polished for two rounds with short PE reads using Pilon v1.23 (Walker et al., 2014). The vector and microbe contaminations were searched using HS-BLASTN v1.0.0 (Chen, Ye, Zhang, & Xu, 2015) against the Nt and UniVec databases, with the alignment threshold set as 1e-5. Finally, we employed SeqKit v0.14.0 (Shen, Le, Li, & Hu, 2016) to remove identified contaminations (i.e., vector and microbe) from the HaSCD2 assembly.

Genetic map construction

For genetic map construction, the SCD individual was crossed with an AY individual (collected in 2011 from Anyang city, China) as illustrated in Figure S2. Genomic DNA of 100 BC1 offspring and their parents was isolated to prepare short-read libraries, and these libraries were sequenced using the Illumina HiSeq Xten platform. The short reads were mapped to the assembled HaSCD2 genome using BWA-mem v0.7.12, with '-M' parameters (Li & Durbin, 2010). SAMtools v0.1.19 was used to format intermediate files (Li et al., 2009). Genotype and variants calling were conducted by GATK v3.8 (McKenna et al., 2010), and hard filtering was executed with the following criteria: QD < 2.0; FS > 60.0; MQ < 30.0; HaplotypeScore > 13.0; MQRankSum < -12.5; ReadPosRankSum < -8.0. High-quality SNP markers were used to construct the genetic map using Lep-map3 (Rastas, 2017). Female-informative markers (i.e., heterozygous alleles in the mother and homozygous in the father) were selected to determine linkage groups (LGs) with LOD score of 15. Subsequently, male-informative markers (i.e., homozygous alleles in the mother and heterozygous in the father) were sorted and used to compute genetic distance of pairwise markers. Finally, the

contigs of HaSCD2 assembly were clustered, ordered and orientated based on the genetic map using ALLMAPS (Tang et al., 2015).

Quality assessment of the HaSCD2 assembly

We evaluated the quality of the HaSCD2 assembly in three aspects, i.e., completeness, continuity, and accuracy. First, we assessed the completeness using Benchmarking Universal Single-Copy Orthologs v3.0.2 (BUSCO) analysis against insecta_odb9 database (n=1,658) (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). Generally, the N50 and L50 values can be affected by the genome sizes and chromosome numbers. Thus, we used the chromosome-normalized N50 (CN50) and chromosome-normalized L50 (CL50) to evaluate the continuity at the level of hypothetical chromosomes (Jiao et al., 2017). Moreover, we assessed the accuracy of HaSCD2 assembly by detecting genomic variants between the reference genome and sequencing data used for *de novo* genome assembly. In theory, heterozygous variants represent the discordant alleles and homozygous variants represent the misassembled segments. Specifically, Illumina short reads were used to detect single nucleotide polymorphisms (SNPs; GQ > 30, DP > 50) and short insertions and deletions (InDels; length < 50 bp, GQ > 30, DP > 50), PacBio long reads were used to detect structural variants (SVs; length > 50 bp) using Minimap2 v2.17 (Li, 2018) and Sniffle v1.0.12a (Sedlazeck, Rescheneder, et al., 2018), loci with less than 20 supporting reads or that were flagged as 'imprecise' (i.e., ambiguous breakpoints) and/or 'unresolved' (i.e., low quality) were removed.

Repetitive elements annotation

Repetitive elements including tandem and interspersed repetitive sequences were discovered based on homology and structural characteristics. First, we *de novo* constructed a species-specific transposable element library for the HaSCD2 assembly using RepeatModeler2 v2.0.1 (Flynn et al., 2020). Known repeat families extracted from Dfam_3.2 (Hubley et al., 2016) and RepBase-20181026 databases (Bao, Kojima, & Kohany, 2015) were combined with species-specific repeat library, and RepeatMasker v4.1.0 (Tarailo-Graovac & Chen, 2009) was used to search and classify repetitive elements in the HaSCD2 assembly.

Transcriptome sequencing and analyses

Transcriptome of the SCD strain was conducted using individuals at three main developmental stages, including embryos (~2 days, 40 eggs), larvae (5 instars) and adults (2 male and 2 female

individuals). These specimens were snap-frozen in liquid nitrogen for extraction of total RNA using TRIzol (Invitrogen). The purities and concentrations of total RNA were measured using OneDrop™ 1000 Spectrophotometer (OneDrop Technologies). RNA-seq libraries were constructed using Illumina TruSeq RNA Sample Preparation Kit according to the manufacturer's protocol, and 150-bp short reads were sequenced on Illumina HiSeq 2500 platform. Low-quality sequences (Phred score < 10) and adapter sequences were trimmed in 4-bp sliding windows using Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014). The clean reads were mapped to the HaSCD2 assembly by HISAT2 v2.1.0 (Kim, Langmead, & Salzberg, 2015). StringTie v1.3.4 (Pertea et al., 2015) was used to parse sorted BAM files and to reconstruct transcripts from all transcriptome data. Redundant transcripts were removed with CD-HIT v4.8.1 (Fu, Niu, Zhu, Wu, & Li, 2012) with an identity cutoff of 0.9.

Repetitive elements mask and gene prediction

Repetitive elements identified above were masked using SeqKit v0.14.0 (Shen, Le, Li, & Hu, 2016) prior to gene annotation. We conducted *ab initio* gene prediction using BRAKER v2.1.0 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016) invoking GeneMark-ET v4.38 (Lomsadze, 2005) and Augustus v3.3 (Stanke, Steinkamp, Waack, & Morgenstern, 2004). Namely, GeneMark-ET generated initial gene structures, then AUGUSTUS integrated the predicted genes and RNA-seq information to generate the final gene predictions. The protein sequences of *Acyrthosiphon pisum* (GCF_005508785.1), *Apis mellifera* (GCF_003254395.2), *Bombyx mori* (GCF_014905235.1), *Drosophila melanogaster* (GCF_000001215.4), *Helicoverpa armigera* (GCF_002156985.1) and *Tribolium castaneum* (GCF_000002335.3) were retrieved from NCBI RefSeq database (O'Leary et al., 2016). We used MAKER v2.31.10 (Holt & Yandell, 2011) to integrate *ab initio*, transcriptome-based and homology-protein evidence. The molecular functions of annotated genes were assigned against the UniProtKB (SwissProt + TrEMBL) database using Diamond v0.9.24 (Buchfink, Xie, & Huson, 2015). Protein domains, Gene Ontology (GO) and pathways (MetaCyc and Reactome) were searched against Pfam (El-Gebali et al., 2019), PANTHER (Mi, Muruganujan, Ebert, Huang, & Thomas, 2019), Gene3D (Lewis et al., 2018), Superfamily (Wilson et al., 2009), and CDD databases (Marchler-Bauer et al., 2017) using InterProScan 5.34-73.0 (Mitchell et al., 2019). The KEGG pathway assignments were conducted using KofamKOALA (Aramaki et al., 2020) based on KOfam, a customized HMM database of KEGG Orthologs (KOs).

Collinearity

We utilized BLASTP (Camacho et al., 2009) to perform all-vs-all homologous gene searches between *H. armigera* and other lepidopteran insects (*Bombyx mori*, *S. frugiperda*, *S. exigua*, *S. litura*) with a cut-off e-value of 1e-5. The top 5 BLASTP hits for a gene were selected to detect inter-genome collinear blocks using MCScanX toolkit (Wang et al., 2012) with default settings. The links of putative collinear blocks were visualized by Circos v0.69 (Krzywinski et al., 2009).

Genomic comparisons between two subspecies

The csiro4b assembly (Pearce et al., 2017) was retrieved from NCBI with RefSeq accession GCF_002156985.1. The csiro4b assembly was assembled based on multiple sequencing libraries with different insert size and there remained 26,911 N-gaps as padding of unsolved regions, covering over 56.71 Mb genomic sequences (Figure 2c). We illustrated the process of closing N-gaps in the csiro4b assembly in Figure S3. Specifically, scaffolds in the csiro4b assembly were split into contigs at arbitrary Ns to avoid erroneous junctions, then we coalesced these contigs into pseudomolecules with padding of 100 Ns according to the HaSCD2 assembly using RaGOO v1.11 (Alonge et al., 2019). Subsequently, we located the genomic coordinates of N-gaps in the csiro4b assembly and extracted additional 500-bp downstream and upstream flanking sequences of the N-gaps. To obtain corresponding locations and sizes of these unsolved regions in the HaSCD2 assembly, we aligned PE flanking sequences to the HaSCD2 assembly using MUMmer v3.23 (Kurtz et al., 2004). N-gaps were considered as successfully closed when they simultaneously satisfied the following criteria: (1) matched bases in PE flanking sequences were ≥ 200 bp long; (2) PE flanking sequences were assigned to the same chromosome.

Field cotton bollworm collection and variant calling

During 2016-2018, we used light traps to collect 141 adult cotton bollworm from 13 locations in mainland China, consisted of Yellow River Region (YRR) of Ac (n=7), Ay (n=10), Gy (n=5), Hm (n=10), Kf (n=10), Np (n=5), Xj (n=6); Changjiang River Region (CRR) of Aq (n=8), Df (n=19), Jz (n=20); and Northwestern Region (NR) of Hmi (n=12), Kel (n=12) and Sc (n=17) (Figure 3a, Table S8). All collected samples were preserved in absolute ethanol for genomic DNA extraction and library preparation, followed by sequencing on the Illumina HiSeq Xten platform. Partial mitochondrial cytochrome oxidase subunit I (COI) sequences were retrieved from raw sequencing

reads by MitoZ (Meng, Li, Yang, & Liu, 2019) to confirm the *H. armigera* species status (Table S8). In addition, published re-sequencing resources, including 16 *H. a. conferta* from Australia, 7 *H. punctigera* from Australia and 7 *H. zea* from Brazil (n=3) and America (n=4) were also integrated for downstream analyses (Anderson et al., 2018; Pearce et al., 2017). The sequence adapters and low-quality reads (Phred score < 20) were trimmed in 4-bp sliding windows using Trimmomatic v0.36 (Bolger et al., 2014). The clean re-sequencing reads were mapped to the HaSCD2 assembly using BWA v0.7.12 with default parameters (Li & Durbin, 2010). We used SAMtools v0.1.19 (Li et al., 2009) to remove duplicated reads. GATK v4.1.4.1 (McKenna et al., 2010) was used for variant calling and for filtering out low-quality variants as described above. Finally, the high-quality SNPs were annotated using SnpEff v5.0d (Cingolani et al., 2012).

Population diversity and structure

The pairwise nucleotide diversity ($\theta\pi$), Watterson's estimator (θ_w), expected heterozygosity (H_E) and Tajima's *D*, were calculated using VCFtools v0.1.16 (Danecek et al., 2011). Loci with minor allele frequency (MAF) < 0.01 and genotyping rate < 90% were further removed for downstream population structure analyses. We employed VCF2Dis v1.42 (https://doi.org/github.com/BGI-shenzhen/VCF2Dis) to calculate pairwise distance (*p*-dist) matrices between individuals, and constructed neighbor-joining (NJ) trees using the 'neighbor' module in PHYLIP v3.695 (https://doi.org/evolution.genetics.washington.edu/phylip.html). Phylogenetic trees were visualized using FigTree (http://doi.org/tree.bio.ed.ac.uk/software/figtree/). We also performed principal component analysis (PCA) using the smartPCA module built-in EIGENSOFT v7.2.1 with no automatic outlier removal allowed (Patterson, Price, & Reich, 2006). The Tracy-Widom test was applied to demonstrate statistical significance of different eigenvectors. Finally, individual ancestries were quantified using fastStructure (Raj, Stephens, & Pritchard, 2014).

Inference of demographic history

We selected two representative individuals with high sequencing depth (> 25-fold, Table S8) to infer the historical effective population size of CRR & YRR and NR lineages using Pairwise Sequentially Markovian Coalescence (PSMC) v0.6.5 with 100 bootstrap replications (-b 100) (Li & Durbin, 2011). Other parameters were set as following: "-N25 -t15 -r5 -p 4+25*2+4+6". The

SMC++ v1.15.4 which was shown to have a better resolution in inferring demographic history of the recent past than the PSMC program (Terhorst, Kamm, & Song, 2017), was also used to infer more recent demographic history of CRR & YRR and NR lineages based on multiple individuals, and outputs from these two modelers were scaled with a generation time of 0.25 year and a mutation rate of 2.9×10^{-9} (Keightley et al., 2015).

Detection of divergent patterns and adaptive footprints

All female individuals (i.e., individuals with the heterogametic sex chromosome) were removed to avoid adverse effects on divergence computation. The male individuals with diploid Z-chromosomes have comparable sequence depth coverage between Z-chromosome and autosomes. We identified 90 male cotton bollworms for downstream analyses, including 47 individuals from CRR & YRR, 33 individuals from NR and 10 *H. a. conferta* individuals (Table S8 and Table S9). We split the HaSCD2 assembled 31 chromosomes into 10 kb sliding windows with a step size of 5 kb. Windows with more than 10 SNPs were retained for calculating inter-populations fixation index (F_{ST}) using VCFtools v0.1.16 (Danecek et al., 2011) with negative F_{ST} values treated as zero. We performed Mantel test to evaluate the relationship between genetic distance (F_{ST}) and geographic distance (km) using the 'vegan' package (https://doi.org/cran.r-project.org/web/packages/vegan/index.html). The significance of the correlation was estimated based on 9,999 permutations.

Complementary methods can be used to detect selective signatures. We calculated fixation index (F_{ST}) and nucleotide diversity ratio (θ_πA/θ_πB) for each sliding window (5-kb windows with the step size of 2 kb). The F_{ST} values were Z-transformed and the nucleotide diversity ratios were $\log_2$-transformed. Selective signatures with the highest 5% Z(F_{ST}) and the highest 5% $\log_2(\theta_\pi A/\theta_\pi B)$ were considered to be candidate selection targets of population B, whereas the lowest 5% $\log_2(\theta_\pi A/\theta_\pi B)$ were population A. Genes that overlapped more than 50% in length with candidate selection targets were defined as putative selective genes. The Gene Ontology (GO) terms of putative selective genes were retrieved from the HaSCD2 annotation, and GO enrichment analysis was carried out to identify gene sets that were significantly over-represented relatives to common genes using the R package clusterProfiler (Yu, Wang, Han, & He, 2012).

RESULTS

Chromosome-scale assembly of the cotton bollworm

Our sequencing generated 5.48 million long reads (~137-fold coverage; N50 size of 12.89 kb) using single-molecule real-time (SMRT) sequencing, and 132.95 million PE reads (~112-fold coverage) using the Illumina HiSeq Xten platform (Table S1). The genome size and heterozygosity rate of cotton bollworm were estimated to be ~352.30 Mb and ~0.84% respectively (Figure S4). The two genome assemblers, i.e., Falcon and Canu, produced two draft assemblies with total sizes of 398.01 Mb (N50 size of 4.81 Mb) and 454.70 Mb (N50 size of 4.17 Mb), respectively. After merging assemblies and discarding alternative haplotypes, we obtained a 356.67 Mb primary assembly, named as HaSCD2, that consisted of 106 contigs with a N50 length of 7.34 Mb (Table 1). Our assembly is 5.81% larger than the published *H. a. conferta* assembly (csiro4b, 337.07 Mb; Pearce et al., 2017). Importantly, our assembly is much more continuous, i.e., contig N50 of 7.34 Mb *vs*. 23.48 kb, representing an approximately 313-fold increase in average contig length (Figures 1a and 1b, Table 1).

Clean sequence data from 100 BC1 cotton bollworm offspring and their parents (~5.88-fold average coverage for offspring; ~18.83-fold average coverage for parents, Table S2) were mapped to the HaSCD2 assembly. This yielded 425,788 female-informative markers that were assigned into 31 linkage groups (LGs), and 756,597 male-informative markers that spanned a genetic distance of 1545.4 cM (Table S3). Finally, 94 of the 106 contigs (353.56 Mb) were anchored into 31 LGs, of which 351.05 Mb sequences were further oriented (Figure 1c, Table S3). Remarkably, eight chromosomes are represented by a single contig, including the Z-chromosome, the longest chromosome in HaSCD2 assembly (Table S3).

Quality assessment of HaSCD2 assembly

We confirmed that the HaSCD2 assembly was highly complete through the BUSCO assessment (97.7% completeness) (Table S4). The CN50 and CL50 values of the HaSCD2 assembly were 7.34 Mb and 1 (Table S5), indicating that nearly all chromosomes were assembled by a few contigs only. Irrespective of whether the NGS data were sourced from *H. a. armigera* or *H. a. conferta* (Anderson et al., 2018), the mapping rate and coverage breadth of the HaSCD2 assembly were significantly higher than those of the csiro4b assembly (Figures 2a and 2b), which resulted from longer length of most chromosomes (Figure 2c). The conserved synteny between cotton bollworm

and other moths confirmed the reliability of our chromosome assignment on the basis of the genetic map (Figure S5). Moreover, we only detected 129, 273 and 2,322 homozygous SNPs, InDels and SVs (covering ~0.52 Mb) and thus estimated the sequence accuracy of the HaSCD2 assembly to be ~99.85%. In particular, most heterozygous variants were restricted to hotspots in the genome, which have failed to become homozygous possibly due to closely linked lethal alleles segregating in repulsion (Figure S6). Overall these results confirm our HaSCD2 genome is more comprehensive and will be of greater utility for conducting genome-wide analyses of the cotton bollworm.

Repeat and gene annotation

In total, 96.44 Mb sequences (27.04%) were annotated as repetitive elements in the HaSCD2 assembly, of which 74.56% of repeats were classified into various repetitive families that included long interspersed nuclear elements (LINEs, 21.04%), rolling-circle transposons (RCs, 18.23%), long terminal repeats (LTRs, 14.46%), short interspersed nuclear elements (SINEs, 9.58%) and DNA transposon (8.32%) (Table S6). Compared to the csiro4b assembly (14.60%, 49.21 Mb; Pearce et al., 2017), repetitive elements annotated in HaSCD2 were increased by 95.98% in length, owing to the better performance of long reads sequencing technology in resolving repetitive sequences. In addition, the total proportion of repetitive elements in the HaSCD2 assembly was lower than the other noctuid pests with larger genome sizes, indicating that repetitive elements are an important factor affecting differences in the size of noctuid genomes (Table S7).

A total of 18,668 protein-coding genes were annotated in HaSCD2 assembly, and the mean gene length was 6,776.03 bp (Table 1). The BUSCO analysis revealed 95.0% of single-copy genes in insecta_odb9 database were identified as complete in our HaSCD2 assembly (Table S4). The gene number in the HaSCD2 assembly was slightly more than the csiro4b assembly (17,086), which suggested that the HaSCD2 offered opportunities to unmask protein-coding genes in those previously unsolved regions in the csiro4b genome (Table 1). In addition, approximately 95.37% of genes were functionally annotated, of which 9,067 and 6,689 genes were assigned with GO and KEGG accessions, respectively.

Comparisons of HaSCD2 and csiro4b assemblies

A total of 12,115 N-gaps (45.02%) were successfully localized to unique locations based on their flanking sequences matches, and corresponded to greater than 22.45 Mb in the HaSCD2 assembly

(Figure 2d). Noticeably, 44.45% of these assigned N-gaps (9.98 Mb) overlapped with various repetitive families, including long interspersed nuclear elements (LINEs, 2.07 Mb), rolling-circles (RCs, 1.65 Mb), long terminal repeat (LTRs, 1.64 Mb), short interspersed nuclear elements (SINEs, 0.89 Mb) and DNA transposons (0.84 Mb) (Figure 2e). The total proportion of repetitive sequences in N-gaps was higher than the whole genome, indicating that the performance of NGS in genome assembly was greatly hindered by repetitive sequences. On the other hand, the proportions of different repetitive families in N-gaps regions were very close to that in the whole HaSCD2 genome, revealing that N-gaps regions have no obvious bias to any repeat category (Figure 2e). Meanwhile, these assigned N-gaps overlapped with exonic (2.71 Mb), intronic (5.67 Mb) and intergenic (13.80 Mb) regions, and ~2,987 genes were improved due to at least one missing fragments in their exons (Figure S6). For example, we found a completely missing and a partially fragmentary odorant binding proteins (OBPs) in an assigned N-gap located in HaChr18 of HaSCD2 assembly (Figure 2f).

Unique lineage of cotton bollworm in northwestern China

We conducted a population genomic study of 141 individuals of cotton bollworm, which were collected from 13 locations in three cotton-producing regions of China, i.e., the Yellow River Region (YRR), the Changjiang River Region (CRR) and the Northwestern Region (NR) (Wu & Guo, 2005) (Figure 3a, Table S8). We generated a total of 746.74 Gb PE sequencing data, 88.25% of which were mapped to the HaSCD2 assembly with a mean depth of 10.21-fold (Table S8). By integrating publicly available data of 30 genomes from the *Helicoverpa* complex (see MATERIALS AND METHODS, Table S9), we identified 28,714,909 bi-allelic SNPs, including 2,891,882 non-synonymous variants (Table S10), and they covered more than 87.10% of the HaSCD2 assembly (Table S11).

After stringent quality filtering, we refined 5,227,071 high-quality SNPs to infer their evolutionary attributes. The pairwise *p*-dist matrices based on genome-wide, Z-chromosome, and non-coding SNPs, were calculated to construct neighbor-joining (NJ) trees, respectively (Figure 3b, Figure S7). All phylogenetic trees consistently showed that the cotton bollworm was clustered into three distinct groups, that is, *H. a. conferta* sampled in Australia, *H. a. armigera* sampled in CRR and YRR, and *H. armigera* sampled in NR. The result was consistent with the sampling locations of these specimens. Similarly, principal component analysis (PCA) demonstrated that the first two

eigenvectors jointly separated all individuals into three clusters, and they explained ~4.72% variation present across the 13 Chinese populations as well as a *H. a. conferta* population (Figure 3c). We further used fastStructure (Raj et al., 2014) to gain insights into the ancestry of each cotton bollworm individual (Figure 3d). When K = 2, NR lineage, rather than *H. a. conferta* subspecies, was initially separated from the other groups, which indicated their unique genetic composition and perhaps warranted them being considered their own subspecies. When K = 3, three distinct genetic clusters were in agreement with the phylogenetic tree and PCA analysis. Furthermore, asymmetric genetic admixture events seem to exist from CRR & YRR to the NR lineage (Figure 3d).

Population diversity and demographic history

As shown in Table S12, the genome-wide genetic diversity of different populations in the NR lineage was similar to that in CRR and YRR lineages. The mean Tajima's *D* values of different populations ranged from -1.22 to -0.67 (Table S12), meaning that different cotton bollworm populations have an excess of low frequency alleles. At the lineage levels, the high proportions of negative Tajima's *D* values across the whole genome (97.26% for CRR & YRR and 97.49% for NR) suggest that population expansion was the most likely cause (Biswas & Akey, 2006) (Figure 4a).

PSMC analysis revealed that CRR & YRR and NR lineages have maintained a similar and constant effective population sizes for up to *ca*. 0.2 million years ago (Mya), after which, their effective population sizes began to simultaneously increase but at different growth rates, coinciding with the most recent rise of global temperature (Figure 4b). At marine isotope stage 5 (MIS5), the effective population size of CRR & YRR exceeded that of the NR lineage. During the last glaciation (LG), the global temperature began to gradually fall and effective population sizes of the two lineages reached a plateau. Then when the temperature decreased to its lowest point the NR lineage (but not the CRR & YRR lineage) had a decrease in population size. Subsequently, the CRR & YRR and the NR lineages started to expand again as the global temperature rose. In general, historical global temperature profoundly affected the effective population sizes of the CRR & YRR and the NR lineages and may therefore likely have facilitated their divergence. In addition, the PSMC result further confirmed the assumption that the genome-wide negative Tajima's *D* values were the result of population expansions. SMC++ analysis further revealed the dramatic population expansion of the CRR & YRR lineage between ~5,000 – 8,000 years ago (Figure S8).

Genomic divergence among different lineages

The fixation indices (F_{ST}) in pairwise cotton bollworm populations ranged from 0 to 0.076 (Figure S9). Several populations from YRR showed no divergence (i.e., F_{ST} = 0) which indicated frequent gene exchange underpinned their homogenization. A Mantel test showed that the divergence levels of different populations was highly correlated with their geographical distances (Figure S10, P = 1e-04). This suggested that spatial distance has been highly influential in generating the observed genetic variation. We also compared the differentiation levels among NR, CRR & YRR, and *H. a. conferta*. We found *H. a. conferta vs.* NR exhibited the highest divergence levels (F_{ST} = 0.069), followed by *H. a. conferta vs.* CRR & YRR (F_{ST} = 0.041), while CRR & YRR *vs.* NR was lowest (F_{ST} = 0.025), which was still higher than the divergence level between the corn and rice strains of *Spodoptera frugiperda* (F_{ST} = 0.019; Gouin et al., 2017). Overall, our genomic analyses further suggested that the NR is likely to be a unique cotton bollworm lineage.

Among these three *H. armigera* lineages, the divergence levels of Z-chromosome in the male individuals were 5.38-, 2.67-, and 4.44-fold higher than autosomes in the comparisons of CRR & YRR *vs*. NR, CRR & YRR *vs*. *H. a. conferta,* and NR *vs*. *H. a. conferta*, respectively (Figure 5a). The genomic regions with the highest 1% F_{ST} value in the above comparisons, 63.55%, 34.12%, and 61.33% regions were located in the Z-chromosome. Thus, we deduced that the divergence rate of the Z-chromosome was faster than autosomes of cotton bollworm either at the level of the whole chromosomes or at outlier loci. Surprisingly, discordant trends of divergence levels were detected in Z-chromosome and autosomes among different pairwise lineages, i.e., the Z-chromosome divergence level of CRR & YRR *vs. H. a. conferta* was lower than that of CRR & YRR *vs.* NR (mean F_{ST} = 0.085 and 0.091, Wilcoxon test, W=5,687,304, P = 2.43e-05), whereas the former had obvious elevation in autosomal divergence (mean F_{ST}= 0.032 and 0.017, Wilcoxon test, W=731,287,952, P < 2.2e-16).

A total of 25 windows were identified with the highest 1% F_{ST} values among all three comparisons (Figure 5b). These windows covered 195 kb genomic regions, of which 76.92% were located on the Z-chromosome. From these highly divergent regions, nine protein-coding genes were annotated (four genes located in the Z-chromosome), including two clock genes (*per* and *clk*), a G-protein coupled receptor and Acyl-CoA desaturase (Table S13). Spatiotemporal transcriptome analysis showed that *per* and *clk* genes broadly expressed in most tissues and developmental stages (Figure

5c). Notably, the *per* and *clk* genes act in the negative feedback loop of the core circadian system and they have been shown to be associated with nocturnal flight, feeding behavior and endocrine activity (Tomioka & Matsumoto, 2015). We identified 17 and 2 non-synonymous SNPs in *per* and *clk* genes, respectively, which could contribute to the differences in facultative diapause and annual generations observed among different cotton bollworm lineages (Table S14). Expression levels of two circadian genes (*per* and *clk*) in adults were also markedly higher than in larvae and pupae (Figure 5c), which could highlight their critical roles in behavioral rhythms during adulthood, such as nocturnality and mating activity. In addition, expression levels of the two circadian genes in peripheral tissues were higher than other adult tissues, such as in antennae, abdomen and thorax (Figure 5c), emphasizing the divergence of circadian genes may be related to orientation and reproduction (Merlin, Gegear, & Reppert, 2009; Tomioka & Matsumoto, 2015).

Selection signatures on climate adaptation, feeding behavior and insecticide tolerance

Loci that feature in different cotton bollworm lineages may have roles responding to environmental factors, such as low temperature and deficient diets in NR populations, or the different pest management strategies in CRR & YRR and in *H. a. conferta* (Table S14). Using a combination of $\log_2$-transformed θ_πA/θ_πB and *Z*-transformed F_{ST} values (see MATERIALS AND METHODS), we detected 310/349 selective windows for the NR lineage when compared to CRR & YRR and *H. a. conferta*, 185/192 selective windows for CRR & YRR when compared to NR and *H. a. conferta*, and 145/269 selective windows for *H. a. conferta* when compared to NR and CRR & YRR (Figure 6a, Figure S11).

For the NR lineage, two GO categories were commonly enriched irrespective of whether *H. a. conferta* or CRR & YRR was set as the control lineage (Table S15), i.e., oxidation-reduction process (GO:0055114) and ATPase activity (GO:0016887). Another three GO categories related to energy metabolic processes were enriched in a single comparison (Table S15). These intense selection signatures were consistent with expected greater energy consumption for the NR lineage to adapt to the extreme winter, especially the process of resisting severe cold and looking for host plants in the next spring (Arrese & Soulages, 2010). A previous study of *Drosophila melanogaster* populations has shown that fluctuating temperature could induce genes involving in "ATP synthesis coupled proton transport" category (Hsu, Belmouaden, Nolte, & Schlötterer, 2020). Meanwhile, the down-regulation of genes involved in "oxidation-reduction process" and "ATP synthesis coupled

proton transport" is also considered to be the response of Antarctic midges to extreme desert habitats (Teets et al., 2012). These results suggest that diverse insects have evolved similar climate adaptation strategies.

For *H. a. conferta*, we annotated 203 and 289 genes in these selective regions when compared to NR and CRR & YRR lineages, and four GO categories were significantly overrepresented (Table S16). Two enriched GO categories, gustatory and olfactory receptors, are identified, of which gustatory receptors are the only chemosensory protein family to display radiations in lepidopteran insects with narrow host ranges (Pearce et al., 2017), which presumably indicated their important roles in insect-plant interactions (Xu, 2020). Previous research has revealed that these chemosensory genes were differentially expressed when cotton bollworm were fed on different host plants (Pearce et al., 2017). In addition, olfactory receptors are required by the insect to detect and respond to innumerable volatile odors in the field, and they are relevant to foraging, mate recognition, and oviposition site selection (Montagné, de Fouchier, Newcomb, & Jacquin-Joly, 2015). When NR was set as the control lineage, the GO term named "structural constituent of cuticle" was uniquely enriched. For lepidopteran pests, the cuticle is the first barrier to slow down the penetration of insecticide molecules within their bodies (Balabanidou, Grigoraki, & Vontas, 2018), thus natural selection on genes encoding cuticle may effectively decrease the efficiency of insecticides in the field.

We further detected a specific divergent signature between CRR & YRR and NR lineages spanning a 120 kb window on chromosome 7 (Figure 6a). In this genomic region, the Tajima's *D* values in the NR lineage were significantly higher than the CRR & YRR lineage, indicating positive selection could lead to local decrease of Tajima's *D* values (Figure 6b). The nucleotide diversity in the CRR & YRR lineage also decreased, indicating that genes in this regions experienced intense positive selection (Figure 6c). Six glutathione S-transferase (GST) genes were annotated in this region (Figure 6d, Table S17). These GSTs that belonged to the Delta clan have well-established roles in xenobiotic degradation, including insecticide and plant secondary metabolites (Cui, Rui, Yang, Wang, & Yuan, 2017; Low et al., 2010; Pearce et al., 2017; Rane et al., 2019). Our analysis of public transcriptome data (Jin et al., 2019) revealed *GSTD1h*, *GSTD1k* and *GSTD1l* were constitutively expressed in insecticide treatment and control groups. Moreover, none of these GSTs are differentially expressed in response to different insecticides (Figure 6f). In this region, we also

detected two linkage disequilibrium (LD) blocks (Figure 6e), which may suggest that at least two GST genes were targeted by the natural selection rather than a single selection with extended genetic hitchhiking. We identified 75 missense SNPs within these divergent signatures and further identified 15 loci whose alleles were specific to the CRR & YRR or the NR lineage (Table S18). Thus, these loci may provide critical cues for understanding host adaptation.

DISCUSSION

In this study, we generated a chromosome-level reference genome of *H. armigera* based on the long-read sequencing technology and an ultra-dense genetic map. This new genome offers not only significant advancements on the preceding work but also represents a near complete reference genome resource for this species. The HaSCD2 assembly had significant advances in a series of assessments, especially the CN50 value which was even superior to other recently published reference genomes of lepidopteran pests (Table S7). In addition, the number of different repetitive DNA families annotated in HaSCD2 assembly were dramatically elevated, and this provides fundamental information critical to resolve complex biological and evolutionary questions related to repetitive sequences (Klai et al., 2020; Tay, Behere, Batterham, & Heckel, 2010). In the genomic regions that failed to be assembled in the csiro4b draft genome, we found numerous genes including two OBP genes that could be biologically significant to receive odorants in the olfactory system (Figure 2f). Therefore, our assembly represents a significantly improved genome resource of this global pest, offering opportunities for more comprehensive genome-wide studies owing to its excellent features in completeness and continuity.

With applications of different molecular markers, deductions on the genetic structure of the cotton bollworm have been revised several times at different scales (Anderson et al., 2016; Behere et al., 2007; Behere, Tay, Russell, Kranthi, & Batterham, 2013; Nibouche, Buès, Toubon, & Poitout, 1998; Weeks et al., 2010). In particular, several studies based on whole-genome SNPs (Anderson et al., 2016; Anderson et al., 2018; Pearce et al., 2017) consistently revealed the subspecies boundary between *H. a. armigera* and *H. a. conferta*, and the presence of gene flow between *H. a. armigera* and *H. zea*, which emphasized the biological risks of allopatric lineages. We jointly analyzed 14 cotton bollworm populations (13 Chinese populations and 1 Australian population) to investigate genetic structure and diversity within *H. armigera* complex. We found the NR individuals collected

from northwestern China formed a unique cotton bollworm lineage, which was genetically distinct from the known *H. a. conferta* and *H. a. armigera* lineages. This conclusion may well explain why NR and CRR & YRR lineages have heterogeneous resistance levels to chemical and Bt insecticides (Yang et al., 2013). Noticeably, we found that three outliers in the PCA were sampled from Sc population that lies on the western region of Xinjiang. Previous research has demonstrated that European corn borer, *Ostrinia nubilalis*, invaded into Yili area and made secondary contact with native Asian corn borer, *O. furnacalis* (Wang et al., 2017). Thus, these three outliers may provide an important cue for the likely presence of other genetically distinct *H. armigera* lineages in the India subcontinent, Europe, and/or Africa. Variations in genomic diversity in several other insects have been attributed to factors, such as climate stress, artificial domestication and biological invasion (Ding et al., 2018; Montero-Mendieta et al., 2019; Xia et al., 2009; Wu et al., 2019). Interestingly, although suffering from severe coldness and deficient diets, NR lineage maintains a similar level of genetic diversity to the CRR & YRR lineage (Table S14). Therefore, the NR lineage exhibits remarkable genetic plasticity and capacities to adapt to intensive planting of transgenic Bt cotton and has the potential to generate new mutations conferring Bt resistances such as that observed in the northern region of China. The explicit population relationships we have identified form a basis to monitor and manage this pest complex across spatial and temporal scales.

Contemporary studies in a wide range of taxa are using genomic variation to track the biological processes of speciation and population differentiation on a landscape scale (Wolf & Ellegren, 2017). Aided by an ultra-dense variation map, we investigated the patterns of divergence in three distinct cotton bollworm lineages. We found that Z-chromosomes had higher divergence levels than autosomes in any pairwise lineage comparisons, reflecting their faster cumulative rates of differentiated loci. Similar phenomena have been reported during the speciation processes of the sibling species of birds, fish, and other moths (Presgraves, 2018; Van Belleghem et al., 2018). The Z-chromosome is less exchangeable between species, and thus more refractory to gene introgression than autosomes (Presgraves, 2018). We infer gene introgression from CRR & YRR to NR effectively reduced their interspecific divergence level in the autosomes, but the divergence level in Z-chromosome remained similar across *H. armigera* subspecies and populations. We also identified nine universally divergent genes in inter-lineage comparisons, including two circadian genes (*per* and *clk*). It is worth noting that the *per* gene was among the genes that exhibited elevated

F_{ST} in a less well-powered but independent study of Chinese and Australian *H. armigera* populations (Song et al., 2018). Previous studies developed mitochondrial COI markers or used full mitochondrial genomes to distinguish species in the *Helicoverpa* complex (Behere et al., 2008; Behere et al., 2007; Walsh et al. 2019) and track the invasion routes of *H. a. armigera* (Tay et al. 2017; Arnemann et al., 2019). However, mitochondrial markers are maternally transmitted and non-recombining compared to nuclear markers, and they have limited resolution to distinguish different lineages within *H. armigera* (Behere et al., 2007). These nine genes we listed in Table S13 have the potential to be developed as lineage-specific markers to accurately identify the different evolutionary lineages within the *H. armigera* subspecies.

The patterns of differentiation across the genomes measured by F_{ST} imply local environments could shape genomic divergence. Our analyses indicated that genes that were involved in climate adaptation, feeding behaviors, and insecticide tolerance, were also under strong positive selection. Previous research also demonstrated that different host plants could induce these GST genes to become highly expressed in detoxification and digestive tissues, such as mid-gut, Malpighian tubules and fat body (Pearce et al., 2017), which is consistent with its recent duplication and diversification evolution. The Delta clan of GSTs highlighted in this study have been inferred to be involved in the adaptation of fly species to brassica (Gloss et al., 2014). Cotton bollworm residing in the northwestern China have a unique evolutionary ancestry and local habitat (Figure 3 and Table S14), and they are still sensitive to most chemical insecticides and Bt toxins. Meanwhile, a dominant point mutation underlying resistance to Cry1Ac was only detected in northern China populations (Jin et al., 2018), and highlighted the need for concerted management practice to prevent the introgression of the dominant Cry1Ac resistance gene from the northern populations into the NR populations of *H. armigera* in China.

ACKNOWLEDGEMENTS

This study was supported by grants from the National Natural Science Foundation of China (grant no. 31930093 to YW) and the SAFEA of China (grant no. BP0719029 to YW). We thank Yanhui Lu for providing cotton bollworm samples from northwestern China, and Shuai Zhan for critical reading of the manuscript. WTT was supported by CSIRO Health & Biosecurity.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Yidong Wu, Jianpeng Zhang, Feng Zhang and Yihua Yang conceived the ideas and designed the methodology; Jianpeng Zhang, Yu Shi and Fang Guan conducted the molecular experiments; Jianpeng Zhang, Feng Zhang, Wee Tek Tay, and Charles Robin analysed the data; and Jianpeng Zhang and Yidong Wu led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

The HaSCD2 assembly and corresponding annotation file were uploaded at Figshare: https://doi.org/10.6084/m9.figshare.14913858.v1. The PacBio and Illumina sequencing reads for genome assembly and gene annotation were deposited in the NCBI database with the BioProject Number of PRJNA623133. The whole genome resequencing data of the pedigreed individuals was deposited with the BioProject Number of PRJNA730914. The raw sequencing data of field cotton bollworm was deposited with the BioProject Number of PRJNA731848, and the high-quality SNPs was at https://figshare.com/articles/dataset/population_genomics_of_cotton_bollworm/17008922.

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., … Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 224. https://doi.org/10.1186/s13059-019-1829-6

Anderson, C. J., Tay, W. T., McGaughran, A., Gordon, K. H. J, & Walsh, T. K. (2016). Population structure and gene flow in the global pest, *Helicoverpa armigera*. *Molecular Ecology*, *25*(21), 5296–5311. https://doi.org/10.1111/mec.13841

Anderson, C. J., Oakeshott, J. G., Tay, W. T., Gordon, K. H. J., Zwick, A., & Walsh, T. K. (2018). Hybridization and gene flow in the mega-pest lineage of moth, *Helicoverpa*. *Proceedings of the National Academy of Sciences*, *115*(19), 5034–5039. https://doi.org/10.1073/pnas.1718831115

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020).

KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251–2252. https://doi.org/10.1093/bioinformatics/btz859

Arnemann, J. A., Roxburgh, S., Walsh, T. K., Guedes, J., Gordon, K. H. J, Smagghe, G., & Tay, W. T. (2019). Multiple incursion pathways for *Helicoverpa armigera* in Brazil show its genetic diversity spreading in a connected world. *Scientific Reports*, *9*(1), 19380. https://doi.org/10.1038/s41598-019-55919-9

Arrese, E. L., & Soulages, J. L. (2010). Insect Fat Body: Energy, Metabolism, and Regulation. *Annual Review of Entomology*, 55(1), 207–225. https://doi.org/10.1146/annurev-ento-112408-085356

Balabanidou, V., Grigoraki, L., & Vontas, J. (2018). Insect cuticle: a critical determinant of insecticide resistance. *Current Opinion in Insect Science*, *27*, 68–74. https://doi.org/10.1016/j.cois.2018.03.001

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 4–9. https://doi.org/10.1186/s13100-015-0041-9

Behere, G. T., Tay, W. T., Russell, D. A., & Batterham, P. (2008). Molecular markers to discriminate among four pest species of *Helicoverpa* (Lepidoptera: Noctuidae). *Bulletin of Entomological Research*, *98*(6), 599–603. https://doi.org/10.1017/S0007485308005956

Behere, G. T., Tay, W. T., Russell, D. A., Heckel, D. G., Appleton, B. R., Kranthi, K. R., & 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*(1), 117. https://doi.org/10.1186/1471-2148-7-117

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*(1), e53448. https://doi.org/10.1371/journal.pone.0053448

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics*, *22*(8), 437–446. https://doi.org/10.1016/j.tig.2006.06.005

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1). https://doi.org/10.1038/nmeth.3176

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), 1–12. https://doi.org/10.1093/nar/gkw654

Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Research*, *43*(16), 7762–7768. https://doi.org/10.1093/nar/gkv784

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., … Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), 1050–1054. https://doi.org/10.1038/nmeth.4035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cui, L., Rui, C., Yang, D., Wang, Z., & Yuan, H. (2017). *De novo* transcriptome and expression profile analyses of the Asian corn borer (*Ostrinia furnacalis*) reveals relevant flubendiamide response genes. *BMC Genomics*, *18*(1), 20. https://doi.org/10.1186/s12864-016-3431-6

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*(1), 110–113. https://doi.org/10.1590/S1983-40632013000100015

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Ding, D., Liu, G., Hou, L., Gui, W., Chen, B., & Kang, L. (2018). Genetic variation in PTPN1 contributes to metabolic adaptation to high-altitude hypoxia in Tibetan migratory locusts. *Nature Communications*, *9*(1), 4991. https://doi.org/10.1038/s41467-018-07529-8

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Fitt, G. (1989). The ecology of Heliothis species in relation to agroecosystems. *Annual Review of Entomology*, *34*(1), 17–52. https://doi.org/10.1146/annurev.ento.34.1.17

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Fritz, M. L., DeYonke, A. M., Papanicolaou, A., Micinski, S., Westbrook, J., & Gould, F. (2018). Contemporary evolution of a Lepidopteran species, *Heliothis virescens*, in response to modern agricultural practices. *Molecular Ecology*, *27*(1), 167–181. https://doi.org/10.1111/mec.14430

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gloss, A. D., Vassão, D. G., Hailey, A. L., Nelson Dittrich, A. C., Schramm, K., Reichelt, M., … Whiteman, N. K. (2014). Evolution in an ancient detoxification pathway is coupled with a transition to herbivory in the drosophilidae. *Molecular Biology and Evolution*, 31(9), 2441–2456. https://doi.org/10.1093/molbev/msu201

Gouin, A., Bretaudeau, A., Nam, K., Gimenez, S., Aury, J.-M., Duvic, B., … Fournier, P. (2017). Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. *Scientific Reports*, *7*(1), 11816. https://doi.org/10.1038/s41598-017-10461-4

Guo, M., Du, L., Chen, Q., Feng, Y., Zhang, J., Zhang, X., … Liu, Y. (2021). Odorant receptors for detecting flowering plant cues are functionally conserved across moths and butterflies. *Molecular Biology and Evolution*, *38*(4), 1413–1427. https://doi.org/10.1093/molbev/msaa300

Hardwick, D. F. (1965). The corn earworm complex. *Memoirs of the Entomological Society of Canada*, *97*(S40), 5–247. https://doi.org/10.4039/entm9740fv

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. https://doi.org/10.1186/1471-2105-12-491

Hsu, S. K., Belmouaden, C., Nolte, V., & Schlötterer, C. (2020). Parallel gene expression evolution in

699 natural and laboratory evolved populations. *Molecular Ecology*, 30(4), 1–11.
700 https://doi.org/10.1111/mec.15649

701 Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., … Wheeler, T. J. (2016). The
702 Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89.
703 https://doi.org/10.1093/nar/gkv1272

704 Jiao, W.-B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., … Schneeberger, K. (2017).
705 Improving and correcting the contiguity of long-read genome assemblies of three plant species using
706 optical mapping and chromosome conformation capture data. *Genome Research*, *27*(5), 778–786.
707 https://doi.org/10.1101/gr.213652.116

708 Jin, L., Wang, J., Guan, F., Zhang, J. P., Yu, S., Liu, S. Y., … Wu, Y. D. (2018). Dominant point mutation in
709 a tetraspanin gene associated with field-evolved resistance of cotton bollworm to transgenic Bt cotton.
710 *Proceedings of the National Academy of Sciences*, *115*(46), 11760–11765.
711 https://doi.org/10.1073/pnas.1812138115

712 Jin, M., Liao, C., Chakrabarty, S., Zheng, W., Wu, K. M, & Xiao, Y. T. (2019). Transcriptional response of
713 ATP-binding cassette (ABC) transporters to insecticides in the cotton bollworm, *Helicoverpa armigera*.
714 *Pesticide Biochemistry and Physiology*, *154*, 46–59. https://doi.org/10.1016/j.pestbp.2018.12.007

715 Jones, C. M., Parry, H., Tay, W. T., Reynolds, D. R., & Chapman, J. W. (2019). Movement ecology of pest
716 *Helicoverpa*: implications for ongoing spread. *Annual Review of Entomology*, *64*(1), 1–19.
717 https://doi.org/10.1146/annurev-ento-011118-111959

718 Joußen, N., Agnolet, S., Lorenz, S., Schöne, S. E., Ellinger, R., Schneider, B., & Heckel, D. G. (2012).
719 Resistance of Australian *Helicoverpa armigera* to fenvalerate is due to the chimeric P450 enzyme
720 *CYP337B3*. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(38),
721 15206–15211. https://doi.org/10.1073/pnas.1202047109

722 Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., … Jiggins, C. D.
723 (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and*
724 *Evolution*, *32*(1), 239–243. https://doi.org/10.1093/molbev/msu302

725 Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory
726 requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

727 Klai, K., Chenais, B., Zidi, M., Djebbi, S., Caruso, A., Denis, F., … Mezghani K. M. (2020). Screening of

Helicoverpa armigera mobilome revealed transposable element insertions in insecticide resistance genes. *Insects*, *11*(12), 879. https://doi.org/10.3390/insects11120879

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., … Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2). R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., … Lees, J. (2018). Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Research*, *46*(D1), D435–D439. https://doi.org/10.1093/nar/gkx1069

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lomsadze, A. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, *33*(20), 6494–6506. https://doi.org/10.1093/nar/gki937

Low, W. Y., Feil, S. C., Ng, H. L., Gorman, M. A., Morton, C. J., Pyke, J., … Batterham, P. (2010). Recognition and Detoxification of the Insecticide DDT by *Drosophila melanogaster* Glutathione S-Transferase D1. *Journal of Molecular Biology*, *399*(3), 358–366. https://doi.org/10.1016/j.jmb.2010.04.020

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., … Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203. https://doi.org/10.1093/nar/gkw1129

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, *47*(11), e63–e63. https://doi.org/10.1093/nar/gkz173

Merlin, C., Gegear, R. J., & Reppert, S. M. (2009). Antennal circadian clocks coordinate sun compass orientation in migratory Monarch butterflies. *Science*, *325*(5948), 1700–1704. https://doi.org/10.1126/science.1176221

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419–D426. https://doi.org/10.1093/nar/gky1038

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., … Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351–D360. https://doi.org/10.1093/nar/gky1100

Montagné, N., de Fouchier, A., Newcomb, R. D., & Jacquin-Joly, E. (2015). Advances in the identification and characterization of olfactory receptors in insects. *Progress in Molecular Biology and Translational Science*, *130*, 55–80. https://doi.org/10.1016/bs.pmbts.2014.11.003

Montero-Mendieta, S., Tan, K., Christmas, M. J., Olsson, A., Vilà, C., Wallberg, A., & Webster, M. T. (2019). The genomic basis of adaptation to high-altitude habitats in the eastern honey bee (*Apis cerana*). *Molecular Ecology*, *28*(4), 746–760. https://doi.org/10.1111/mec.14986

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*(4), 438–445. https://doi.org/10.1046/j.1365-2540.1998.00273.x

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional

786 annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

787 Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12),
788 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

789 Pearce, S. L., Clarke, D. F., East, P. D., Elfekih, S., Gordon, K. H. J., Jermiin, L. S., … Wu, Y. D. (2017).
790 Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence
791 of two highly polyphagous and invasive *Helicoverpa* pest species. *BMC Biology*, *15*(1), 63.
792 https://doi.org/ 10.1186/s12915-017-0402-6

793 Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015).
794 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature*
795 *Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

796 Presgraves, D. C. (2018). Evaluating genomic signatures of "the large X-effect" during complex speciation.
797 *Molecular Ecology*, *27*(19), 3822–3830. https://doi.org/10.1111/mec.14777

798 Quan, Q., Hu, X., Pan, B., Zeng, B., Wu, N., Fang, G., … Zhan, S. (2019). Draft genome of the cotton aphid
799 *Aphis gossypii*. *Insect Biochemistry and Molecular Biology*, 105, 25–32.
800 https://doi.org/10.1016/j.ibmb.2018.12.007

801 Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population
802 structure in large SNP data sets. *Genetics*, *197*(2), 573–589.
803 https://doi.org/10.1534/genetics.114.164350

804 Rane, R. V., Ghodke, A. B., Hoffmann, A. A., Edwards, O. R., Walsh, T. K., & Oakeshott, J. G. (2019).
805 Detoxifying enzyme complements and host use phenotypes in 160 insect species. *Current Opinion in*
806 *Insect Science*, *31*, 131–138. https://doi.org/10.1016/j.cois.2018.12.008

807 Rastas, P. (2017). Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing
808 data. *Bioinformatics*, *33*(23), 3726–3732. https://doi.org/10.1093/bioinformatics/btx494

809 Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for
810 third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460.
811 https://doi.org/10.1186/s12859-018-2485-7

812 Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of
813 long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329–346.
814 https://doi.org/10.1038/s41576-018-0003-4

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, 11(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Song, S. V., Anderson, C. J., Good, R. T., Leslie, S., Wu, Y. D., Oakeshott, J. G., & Robin, C. (2018). Population differentiation between Australian and Chinese *Helicoverpa armigera* occurs in distinct blocks on the Z-chromosome. *Bulletin of Entomological Research*, *108*(6), 817–830. https://doi.org/10.1017/S0007485318000081

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*, W309–W312. https://doi.org/10.1093/nar/gkh379

Tabashnik, B. E., & Carrière, Y. (2017). Surge in insect resistance to transgenic crops and prospects for sustainability. *Nature Biotechnology*, *35*(10), 926–935. https://doi.org/10.1038/nbt.3974

Tang, H. B., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., … Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology*, *16*(1), 3. https://doi.org/10.1186/s13059-014-0573-1

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1), 1–14. https://doi.org/10.1002/0471250953.bi0410s25

Tay, W. T., Behere, G. T., Batterham, P., & Heckel, D. G. (2010). Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. *BMC Evolutionary Biology*, *10*(1), 144. https://doi.org/10.1186/1471-2148-10-144

Tay, W. T., & Gordon, K. H. J. (2019). Going global – genomic insights into insect invasions. *Current Opinion in Insect Science*, *31*, 123–130. https://doi.org/10.1016/j.cois.2018.12.002

Tay, W. T., Soria, M. F., Walsh, T., Thomazoni, D., Silvie, P., Behere, G. T., … Downes, S. (2013). A brave New World for an Old World pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. *PLoS*

844 *ONE*, *8*(11), e80134. https://doi.org/10.1371/journal.pone.0080134

845 Tay, W. T., Walsh, T. K., Downes, S., Anderson, C. J., Jermiin, L. S., Wong, T. K. F., … Gordon, K. H. J.
846 (2017). Mitochondrial DNA and trade data support multiple origins of *Helicoverpa armigera*
847 (Lepidoptera, Noctuidae) in Brazil. *Scientific Reports*, *7*(1), 45302. https://doi.org/10.1038/srep45302

848 Teets, N. M., Peyton, J. T., Colinet, H., Renault, D., Kelley, J. L., Kawarasaki, Y., … Denlinger, D. L.
849 (2012). Gene expression changes governing extreme dehydration tolerance in an Antarctic insect.
850 *Proceedings of the National Academy of Sciences*, *109*(50), 20744–20749.
851 https://doi.org/10.1073/pnas.1218661109

852 Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from
853 hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309.
854 https://doi.org/10.1038/ng.3748

855 Tomioka, K., & Matsumoto, A. (2015). Circadian molecular clockworks in non-model insects. *Current*
856 *Opinion in Insect Science*, *7*, 58–64. https://doi.org/10.1016/j.cois.2014.12.006

857 Valencia-Montoya, W. A., Elfekih, S., North, H. L., Meier, J. I., Warren, I. A., Tay, W. T., … Jiggins, C. D.
858 (2020). Adaptive introgression across semipermeable species boundaries between local *Helicoverpa*
859 *zea* and invasive *Helicoverpa armigera* moths. *Molecular Biology and Evolution*, *37*(9), 2568–2583.
860 https://doi.org/10.1093/molbev/msaa108

861 Van'T Hof, A. E., Nguyen, P., Dalíková, M., Edmonds, N., Marec, F., & Saccheri, I. J. (2013). Linkage map
862 of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): A model of industrial melanism.
863 *Heredity*, *110*(3), 283–295. https://doi.org/10.1038/hdy.2012.84

864 Van Leeuwen, T., Dermauw, W., Mavridis, K., & Vontas, J. (2020). Significance and interpretation of
865 molecular diagnostics for insecticide resistance management of agricultural pests. *Current Opinion in*
866 *Insect Science*, *39*, 69–76. https://doi.org/10.1016/j.cois.2020.03.006

867 Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C.
868 (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14),
869 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

870 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon:
871 An integrated tool for comprehensive microbial variant detection and genome assembly improvement.
872 *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112963

873 Walsh, T. K., Joussen, N., Tian, K., McGaughran, A., Anderson, C. J., Qiu, X.H., … Heckel, D. G. (2018).
874 Multiple recombination events between two cytochrome P450 loci contribute to global pyrethroid
875 resistance in *Helicoverpa armigera*. *PLoS ONE*, 13(11), e0197760.
876 https://doi.org/10.1371/journal.pone.0197760

877 Walsh, T. K., Perera, O., Anderson, C.J., Gordon, K., Czepak, C., McGaughran, A., … Tay, W. T. (2019).
878 Mitochondrial DNA genomes of five major Helicoverpa pest species from the Old and New Worlds
879 (Lepidoptera: Noctuidae). *Ecology and Evolution*, 9(5), 2933–2944. https://doi.org/10.1002/ece3.4971

880 Wang, Y., Kim, K. S., Guo, W., Li, Q., Zhang, Y., Wang, Z., & Coates, B. S. (2017). Introgression between
881 divergent corn borer species in a region of sympatry: Implications on the evolution and adaptation of
882 pest arthropods. *Molecular Ecology*, 26(24), 6892–6907. https://doi: 10.1111/mec.14387

883 Wang, Y. P., Tang, H. B., Debarry, J. D., Tan, X., Li, J., Wang, X. Y., … Paterson, A. H. (2012). MCScanX :
884 a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids*
885 *Research*, *40*(7). https://doi.org/10.1093/nar/gkr1293

886 Weeks, A. R., Endersby, N. M., Lange, C. L., Lowe, A., Zalucki, M. P., & Hoffmann, A. A. (2010). Genetic
887 variation among *Helicoverpa armigera* populations as assessed by microsatellites: A cautionary tale
888 about accurate allele scoring. *Bulletin of Entomological Research*, *100*(4), 445–450.
889 https://doi.org/10.1017/S0007485309990460

890 Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., … Gough, J. (2009). SUPERFAMILY
891 – Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids*
892 *Research*, *37*, 380–386. https://doi.org/10.1093/nar/gkn762

893 Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of
894 speciation. *Nature Reviews Genetics*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

895 Wu, K. M., & Guo, Y. Y. (2005). The evolution of cotton pest management practices in China. *Annual*
896 *Review of Entomology*, *50*(1), 31–52. https://doi.org/10.1146/annurev.ento.50.071803.130349

897 Wu, N., Zhang, S., Li, X., Cao, Y., Liu, X., Wang, Q., … Zhan, S. (2019). Fall webworm genomes yield
898 insights into rapid adaptation of invasive species. *Nature Ecology & Evolution*, *3*(1), 105–115.
899 https://doi.org/10.1038/s41559-018-0746-5

900 Xia, Q. Y., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., … Wang, J. (2009). Complete resequencing of 40
901 genomes reveals domestication events and genes in silkworm (Bombyx). *Science*, *326*(5951), 433–436.

https://doi.org/10.1126/science.1176620

Xu, W. (2020). How do moth and butterfly taste?—Molecular basis of gustatory receptors in Lepidoptera. *Insect Science*, *27*(6), 1148–1157. https://doi.org/10.1111/1744-7917.12718

Yamamoto, K., Nohata, J., Kadono-Okuda, K., Narukawa, J., Sasanuma, M., Sasanuma, S., … Mita, K. (2008). A BAC-based integrated linkage map of the silkworm Bombyx mori. *Genome Biology*, *9*(1), R21. https://doi.org/10.1186/gb-2008-9-1-r21

Yang, Y. H., Li, Y., & Wu, Y. D. (2013). Current status of insecticide resistance in *Helicoverpa armigera* after 15 years of Bt cotton planting in China. *Journal of Economic Entomology*, *106*(1), 375–381. https://doi.org/10.1603/EC12286

Yu, G. C., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhu, J., Jiang, F., Wang, X., Yang, P., Bao, Y., Zhao, W., … Cui, F. (2017). Genome sequence of the small brown planthopper, *Laodelphax striatellus*. *GigaScience*, 6(12), 1–12. https://doi: 10.1093/gigascience/gix109

Table 1. Summary statistics of HaSCD2 and csiro4b assemblies as well as their corresponding gene annotation.

Species	***H. a. armigera***	***H. a. conferta***
Genome assembly	HaSCD2	csiro4b
Assembly size (Mb)	356.67	337.07
Number of scaffolds	41	997
Number of contigs	106	24,547
Longest scaffold (Mb)	17.88	6.15
Longest contig (Mb)	17.88	0.31
N50 scaffolds length (Mb)	12.27	1.00
N50 contig length	7.34 Mb	23.48 kb
N90 scaffolds length (Mb)	7.86	0.18

N90 contig length	1.65 Mb	6.21 kb
GC (%)	36.53	36.13
Gene annotation		
Protein-coding genes	18,668	17,086
Mean protein length (aa)	475.55	442.79
Mean gene length (bp)	6,776.03	9,364.52
Exons per gene	5.97	6.05
Exon (%)	7.48	6.75
Mean exon length	239.40	219.96
Intron (%)	23.93	35.48
Mean intron length	919.69	1,368.88

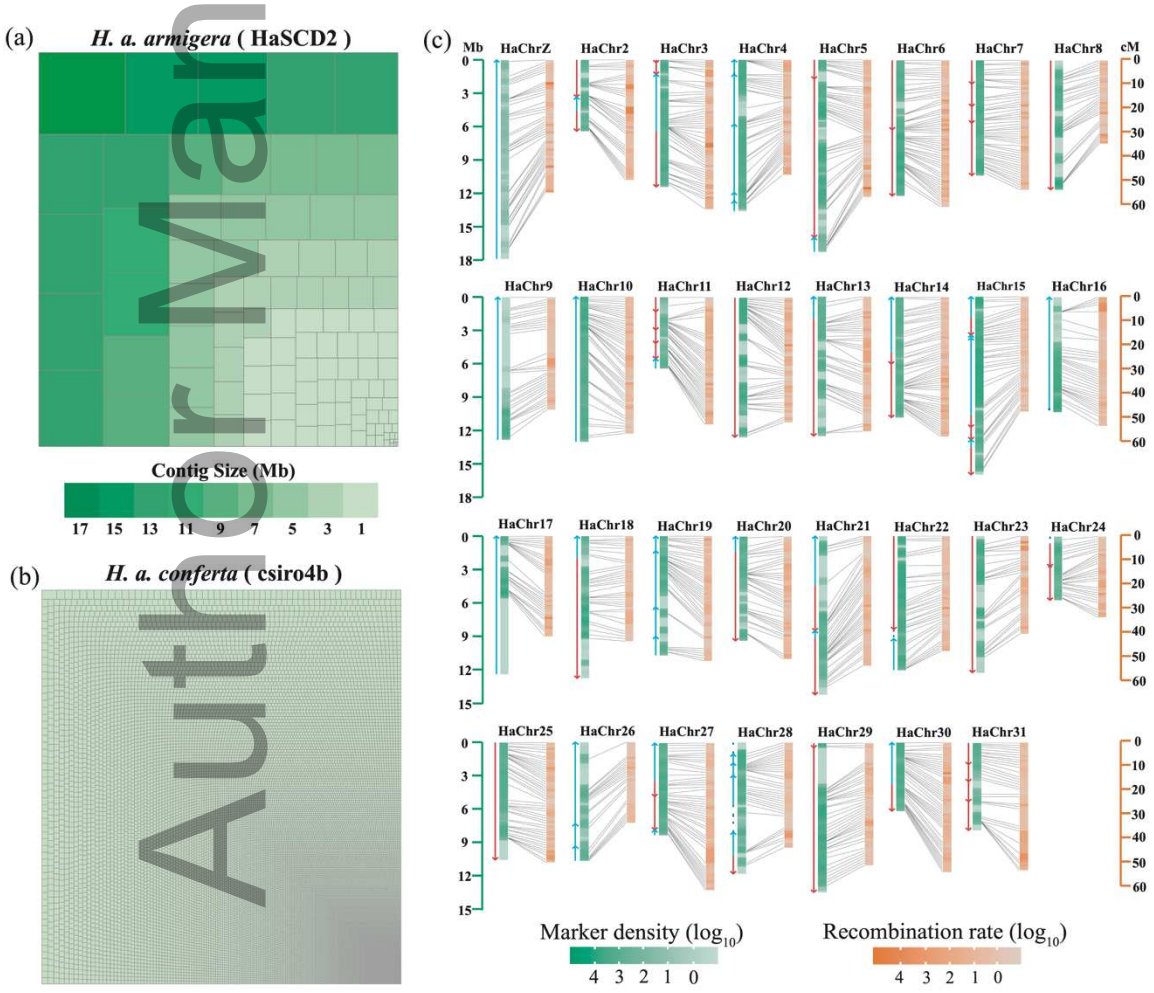

Figure 1. Genome assembly and scaffolding of the HaSCD2. (a) - (b) The differences of contigs length between HaSCD2 and csiro4b assemblies. The proportions and levels of contig lengths are represented by rectangular areas

and filled colors. (c) Assigning contigs into chromosomes with the guide of high-resolution genetic map. Markers with identical recombination distance are collapsed into bins to succinctly show their congruent relationship, and the first markers were linked with the corresponding recombination distance. Graduated colors filled in two rectangles are determined with log10-transformed marker density and recombination distances along 31 chromosomes. The 94 anchored contigs are represented by red and blue lines and their arrows indicate their orientation, and dark blue dotted lines indicate the contigs not oriented.

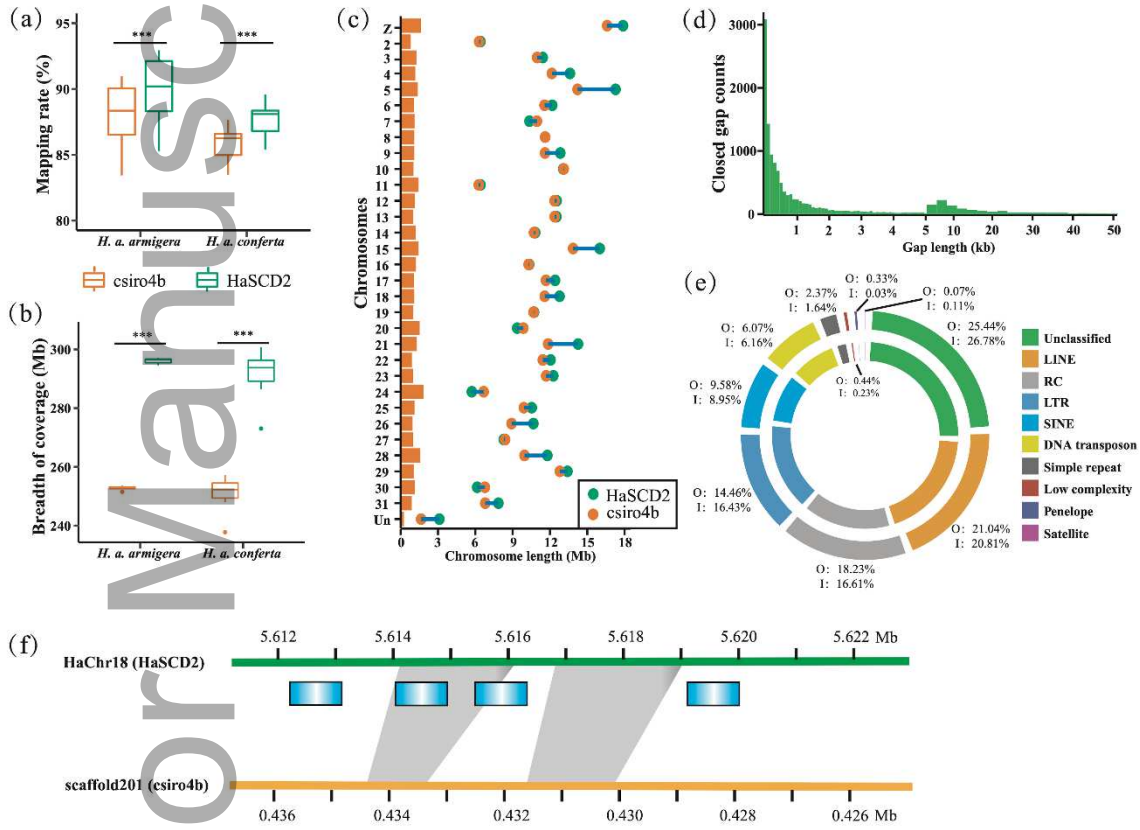

Figure 2. Genome comparisons between HaSCD2 and csiro4b assemblies. (a) - (b) Genomic mapping rate and breadth of coverage (more than 2-fold coverage) against HaSCD2 and csiro4b assemblies. (c) Comparisons of chromosomes size between HaSCD2 and csiro4b assemblies. (d) The distribution of the sizes of closed gaps in the csiro4b assembly. (e) Proportions of major repeat categories distributed in the HaSCD2 assembly. The outer circle represents the repeat in whole HaSCD2 genome and the inner circle represents the repeat in assigned N-gap regions. LINE: long interspersed nuclear elements; RC: rolling-circle transposition; LTR: long terminal repeats; SINE: short interspersed nuclear elements. (f) A case of assigned N-gap affecting the completeness of two odorant-binding proteins (OBPs). Blue rectangles represent odorant binding proteins, grey shadows represent mapping relationships between two assemblies.

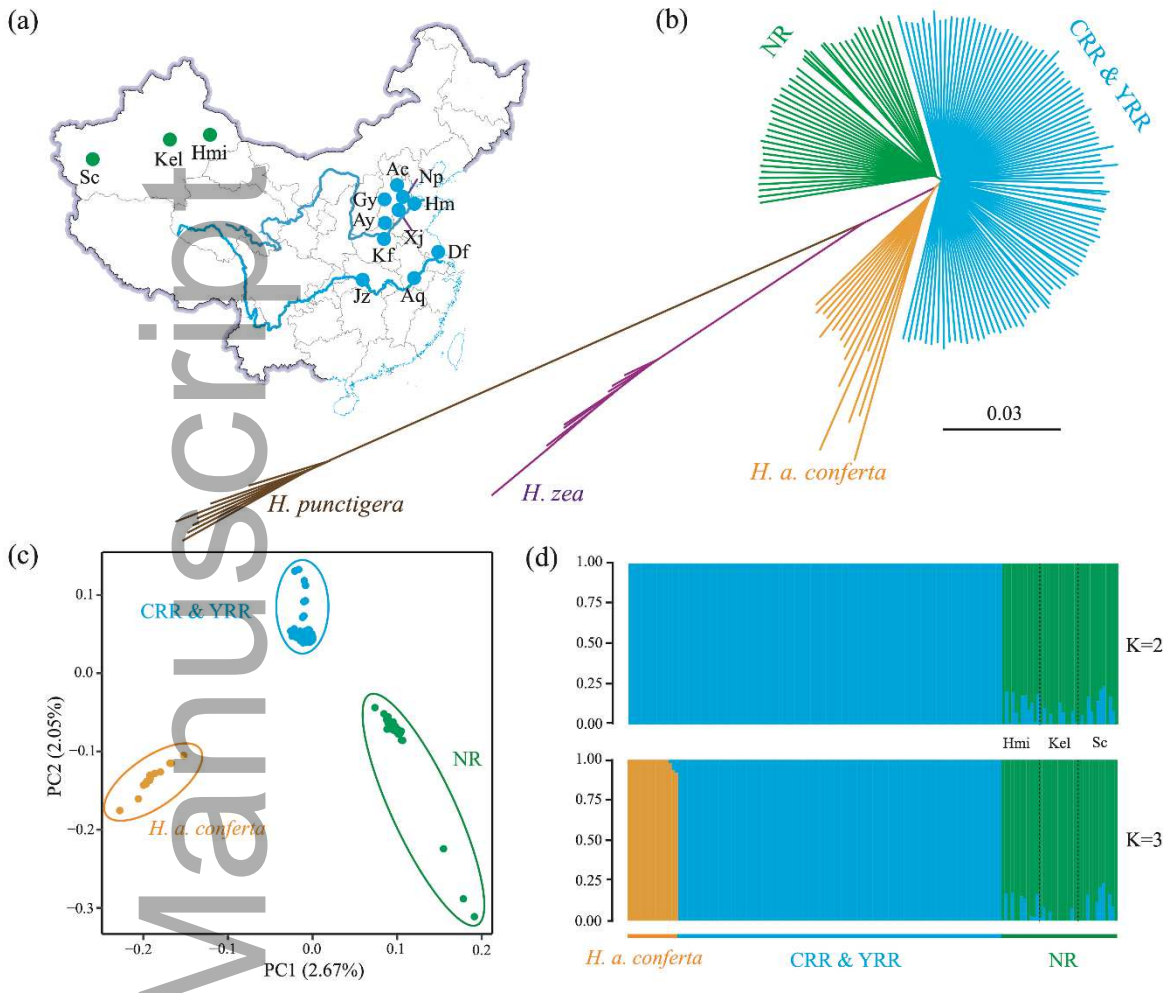

Figure 3. Geographic distribution and genetic structure of collected cotton bollworm. (a) Schematic map denoting the sampling localities in mainland China. Blue solid circles represent localities in the Yellow River Region (YRR) and the Changjiang River Region (CRR), green solid circles represent localities in the Northwestern Region (NR), two blue curves represent Yellow River (top) and Changjiang River (bottom), respectively. (b) Neighbor-joining (NJ) tree of cotton bollworm inferred from whole-genome SNPs using *H. punctigera* as outgroup. (c) Principle components analysis, PC1=2.67% and PC2=2.05% (Tracy-Widom statistics, $P <$ 4.91e-291). (d) Population admixture plot with K equals to 2 (top panel) and 3 (bottom panel).

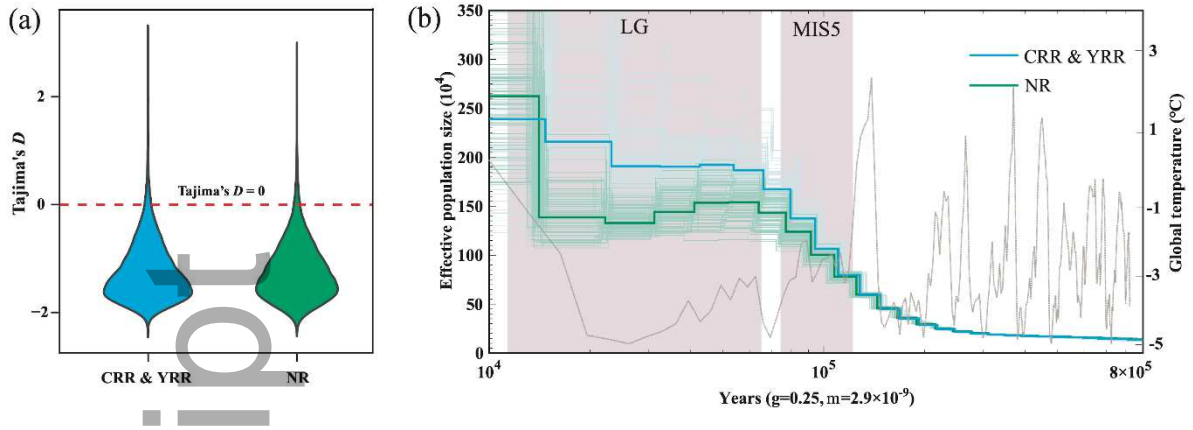

Figure 4. Inference of demographic history of cotton bollworm. (a) Violin plot shows the genome-wide distribution of Tajima's *D* values in CRR & YRR and NR lineages. The red dotted line indicates Tajima's *D* value of zero. (b) The historical effective population size of representative individuals in CRR & YRR and NR lineages. The light blue and green lines represent 100 bootstraps. The dotted grey line represents the historical global temperature. LG, last glaciation; MIS5, marine isotope stage 5.

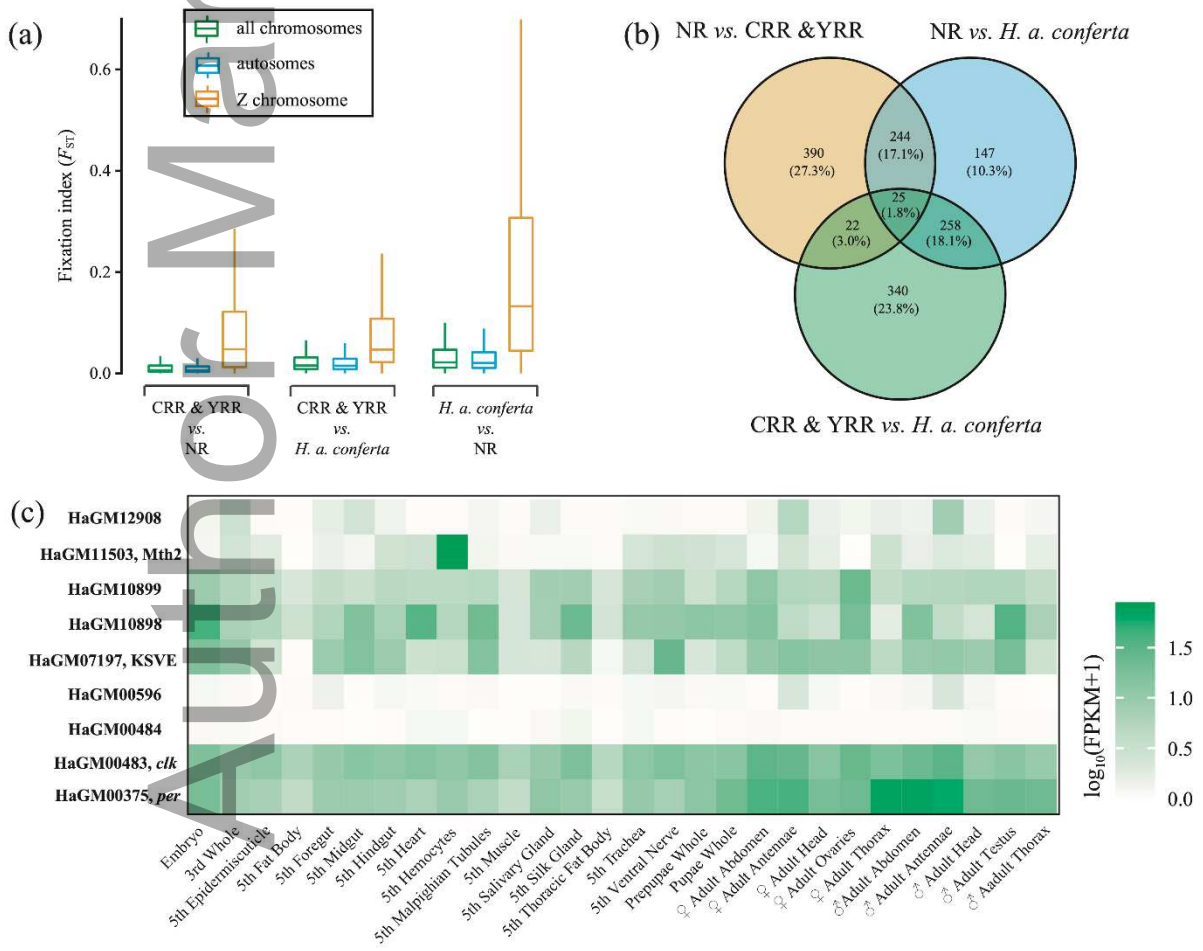

Figure 5. Divergence patterns among different cotton bollworm lineages. (a) F_{ST} values in the sex chromosome and autosomes; (b) Venn diagram shows the number of shared genes with highest 1% F_{ST} values in different

comparisons. (c) Expression levels of nine divergent genes in different tissues and developmental stages of cotton bollworm.

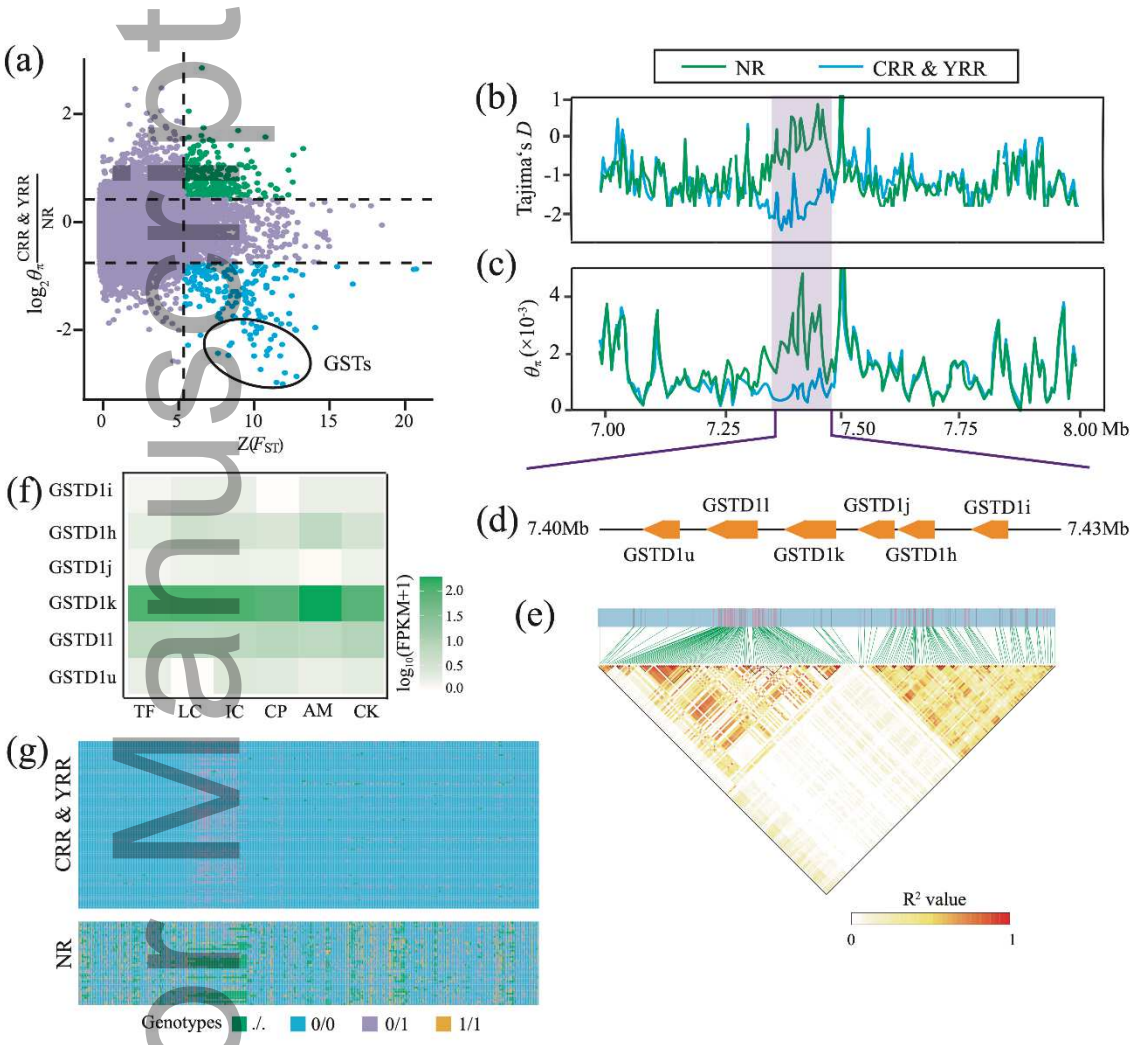

Figure 6. Selection signatures between CRR & YRR and NR lineages. (a) Distribution of $\log_2$-transformed θ_π ratio (Y axis) and *Z*-transformed F_{ST} (X axis) between CRR & YRR and NR lineages. Black lines represent their highest 1% thresholds. Green dots represent genomic regions under selection for the NR lineage and blue dots for the CRR & YRR lineage. (b) - (c) Tajima's *D* and θ_π values around selection signatures in HaChr7. Green line represents NR lineage and blue line represents CRR & YRR lineage. (d) Six glutathione S-transferase (GST) genes annotated in the selection signatures of HaChr7. (e) Patterns of LD blocks around these six GST genes in the CRR & YRR lineage. (f) Expressions of six GST genes in *H. armigera* larvae fed on artificial diet supplemented with different insecticides. AM: abamectin; CP: chlorpyrifos; IC: indoxacarb; LC: lambda-cyhalothrin; TF: tebufenozide; CK: control. (g) Genotypes of loci distributed around these six GST genes.