A chromosome-scale genome assembly of a *Bacillus thuringiensis* Cry1Ac insecticidal protein resistant strain of *Helicoverpa zea*

Amanda R. Stahlke[1, ORCID], Jennifer Chang[2,3,4, ORCID], Luke R. Tembrock[5,6, ORCID], Sheina B. Sim[7, ORCID], Sivanandan Chudalayandi[4], Scott M. Geib[7, ORCID], Brian E. Scheffler[2], Omaththage P. Perera[8, ORCID], Todd M. Gilligan[5], Anna K. Childers[1, ORCID], Kevin J. Hackett[9], Brad S. Coates[10*, ORCID]

Affiliations:

1. USDA, Agricultural Research Service, Beltsville Agricultural Research Center, Bee Research Laboratory, 10300 Baltimore Avenue, Beltsville MD 20705, USA.
2. USDA, Agricultural Research Service, Jamie Whitten Delta States Research Center, Genomics and Bioinformatics Research Unit, 141 Experiment Station Road, Stoneville, MS 38776, USA
3. Oak Ridge Institute for Science and Education, P.O. Box 117, Oak Ridge, TN 37831, USA
4. Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, Iowa, USA, 50010
5. USDA, Animal and Plant Health Inspection Service, Plant Protection and Quarantine, Science & Technology, Identification Technology Program, 2301 Research Boulevard, Fort Collins, CO 80526, USA
6. Department of Agricultural Biology, Colorado State University, Fort Collins, CO 80523, USA
7. USDA, Agricultural Research Service, U.S. Pacific Basin Agricultural Research Center, Tropical Crop and Commodity Protection Research Unit, 64 Nowelo Street, Hilo, Hawaii 96720
8. USDA, Agricultural Research Service, Jamie Whitten Delta States Research Center, Southern Insect Management Research Unit, 141 Experiment Station Road, Stoneville, MS 38776, USA
9. USDA, Agricultural Research Service, Office of National Programs, Crop Production and Protection, 5601 Sunnyside Avenue, Beltsville, MD 20705, USA
10. USDA, Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, 819 Wallace Road, Ames, IA 50011

*Corresponding author:
Brad Coates, Ph.D.
Research Geneticist
USDA-ARS Corn Insects & Crop Genetics Research Unit
532 Science II
2310 Pammel Dr.
Ames, IA 50011
Tel: 515-294-6948
Mobile: 515-357-8496
brad.coates@usda.gov
ORCID: 0000-0001-8908-1529

Abstract:

Helicoverpa zea (Lepidoptera: Noctuidae) is an insect pest of major cultivated crops in North and South America. The species has adapted to different host plants and developed resistance to several insecticidal agents, including *Bacillus thuringiensis* (Bt) insecticidal proteins in transgenic cotton and maize. *H. zea* populations persist year-round in tropical and subtropical regions, but seasonal migrations into temperate zones increase the geographic range of associated crop damage. To better understand the genetic basis of these physiological and ecological characteristics, we generated a high-quality chromosome-level assembly for a single *H. zea* male from Bt resistant strain, HzStark_Cry1AcR. Hi-C data were used to scaffold an initial 375.2 Mb contig assembly into 30 autosomes and the Z sex chromosome (scaffold N50 = 12.8 Mb and L50 = 14). The scaffolded assembly was error-corrected with a novel pipeline, polishCLR. The mitochondrial genome was assembled through an improved pipeline and annotated. Assessment of this genome assembly indicated 98.8% of the Lepidopteran Benchmark Universal Single-Copy Ortholog set were complete (98.5% as complete single-copy). Repetitive elements comprised approximately 29.5% of the assembly with the plurality (11.2%) classified as retroelements. This chromosome-scale reference assembly for *H. zea*, ilHelZeax1.1, will facilitate future research to evaluate and enhance sustainable crop production practices.

Keywords: corn earworm, cotton bollworm, strain HzStark_Cry1AcR, agriculture

Issue Section: Genome Report

Significance: We established a chromosome-level reference assembly for *Helicoverpa zea*, an insect pest of multiple cultivated crops in the Americas. This assembly of a *Bacillus thuringiensis* insecticidal protein resistant strain, HzStark_Cry1AcR, will facilitate future research in areas such as population genomics and adaptations to agricultural control practices.

Introduction

Helicoverpa zea (Boddie), (Lepidoptera: Noctuidae) is a widespread insect in North and South America (Djaman et al. 2019; Huseth et al. 2021; **fig. 1A**). This pest causes extensive damage to plants including maize, cotton, soybean, and vegetable crops, and feeds on many weedy plants (Degrande and Omoto 2013; Cunningham and Zalucki 2014; Leite et al. 2014; Reay-Jones 2019). Relatively rapid development and short generation times (Fitt 1989), annual long-distance migrations (Reay-Jones 2019; Perera et al. 2020), highly polyphagous larvae that tolerate a range of plant secondary defensive metabolites (Niu et al. 2008; Li et al. 2020), and entry into larval diapause during adverse environmental conditions (Roach and Adkisson 1970; Reynolds et al. 2019) contribute to the severity of *H. zea* as a pest. Furthermore, resistance of *H. zea* to chemical insecticides (McCaffrey 1988; Hamadain and Chambers 2001; Jacobson et al. 2009) and *Bacillus thuringiensis* (Bt) crystalline (Cry) and vegetative insecticidal proteins (Vip) expressed in transgenic cotton (Luttrell and Jackson 2012; Rabelo et al. 2020) and maize (Reisig and Reay-Jones 2015; Dively et al. 2016; Bilbo et al. 2019; Yang et al 2019) has led to difficulties controlling crop damage. Resistance to Bt insecticidal proteins is associated with changes in midgut receptors or signal transduction pathways (Soberón et al. 2009). Evidence suggests *H. zea* Bt Cry1Ac resistance is influenced by point mutations in tetraspanin (Jin et al. 2019) or kinesin (Benowitz et al. 2021) genes, or at several loci (Taylor et al. 2021). These reductions in *H. zea* insecticide susceptibility (Yang et al. 2020; Santiago González et al. 2021) contribute to increased levels of crop damage and lower yields (Reisig and Kurtz 2018; Reay-Jones 2019), presenting a threat to global food security (Coates et al. 2015). This has been exacerbated by the introduction and spread of the sister species, *H. armigera*, into South America and the Caribbean (Czepak et al. 2013; Tay et al. 2013; Arnemann et al. 2015; Murúa et

al. 2016; Sosa-Gómez et al. 2016; Tembrock et al. 2019) where introgression of adaptive alleles has led to novel phenotypes that further complicate pest management efforts (Anderson et al. 2018; Cordeiro et al. 2020; Valencia-Montoya et al. 2020; Rios et al. 2021).

Prior *Helicoverpa* genome sequencing using short-read data produced a 341.1 Mb *H. zea* assembly, Hzea_1.0, arranged in 2,975 scaffolds, and a similar low contiguity 337.1 Mb assembly for the closely related *H. armigera*, Harm_1.0 (Pierce et al. 2017). Although these assemblies have been useful for population analyses (Valencia-Montoya et al. 2020; Taylor et al. 2021), chromosome-scale assemblies empower multiple areas of genomic research (Lewin et al. 2018; Childers et al. 2021). As such, we assembled and scaffolded a high-quality chromosome-scale genome for *H. zea*. Our approach implemented best practices for non-model genome assembly with noisy long-reads, whereby an initial assembly from a single individual produced haplotype-specific contigs, followed by purging of duplicate haplotypes prior to Hi-C scaffolding and manual curation, error correction (polishing) including a mitochondrial assembly, and filtering of possible contaminants (Rhie et al. 2021, Howe et al. 2021). This assembly is substantially more contiguous and complete compared to the prior *Helicoverpa* genome resources and serves as an exemplar for developing high quality resources to improve understanding of insecticide resistance, population dynamics, and efficacy of pest management strategies.

Results and Discussion

Genome assembly, scaffolding, annotation, and completeness

In total, 54.0 Gb of raw PacBio continuous long read (CLR) data was generated from a single *H. zea* HzStark_Cry1AcR strain male (supplementary table S1), which resulted in an estimated 148.8-fold coverage based on a 362.8±8.8 Mb flow cytometry estimated genome size (Coates et al. 2017). GenomeScope k-mer analyses of 154.6 million whole-genome shotgun Illumina reads provided an estimated genome size of 341.9 Mb, 80.2% unique (non-repetitive) content, and low heterozygosity (0.3%) (fig. 1B), the latter likely a result of using a single individual from a colony with a degree of inbreeding. Assembly of CLR reads using FALCON-Unzip produced an initial 375.2 Mb assembly consisting of 134 primary pseudo-haploid contigs and 99.4 Mb in 765 alternate contigs (supplementary table S2). After purging duplicate haplotigs, the assembly was improved to 81 primary and 801 alternate contigs (Table 1 and supplementary table S2). These 81 deduplicated primary contigs contained 4.8% fewer duplicated (D) orthologs compared to the initial FALCON-Unzip assembly when assessed with BUSCO.

Scaffolding of primary contigs using Hi-C data broke five mis-joins in three contigs resulting in 42 total scaffolds (fig. 1C) and improved overall contiguity (scaffold N50 of 12.8 Mbp in 14 scaffolds) in the final ilHelZeax1.1 assembly (fig. 1D, table 1 and supplementary table S2). Hi-C contact mapping between contigs resolved 31 scaffolds, representing the expected set of 30 autosomes and the Z chromosome based on a *H. armigera* karyotype (Sahara et al. 2013) and other noctuid assembly data (supplementary table S3). The remaining 11 scaffolds comprised 1.2 Mb and represented < 0.3% of the entire assembly. Error-correcting the scaffolds with Illumina data using the polishCLR pipeline (Stahlke et al. 2022) increased the consensus quality value (QV) from 38.9 to 42.0, and improved the contig assembly L50 and N50

to 15 and 10.9 Mb, respectively. Scaffold length distributions did not change through polishing. The final scaffolded length of 375.2 Mbp is similar to prior flow cytometry estimates (Coates et al. 2017), and larger and more contiguous than the prior 341.1 Mb Hzea_1.0 assembly (Pearce et al. 2017) or other *Helicoverpa* assemblies (supplementary table S2).

Overall, the ilHelZeax1.1 assembly exceeds the minimum reference standard of 6.C.Q40 (>1.0 Mb contig and >10.0 Mb scaffold N50) set for eukaryotic species by the Earth BioGenome Project (Lewin et al. 2018) and meets the qualifications for chromosome level (7.C.Q50; Lawniczak et al. 2022). Assessing the assembly relative to short-reads, k-mer spectra indicated a moderate degree of heterozygosity in the primary pseudo-haploid assembly, shown by a histogram peak (x = 90) corresponding to both haplotypes that is much greater compared to that for single-copy k-mers (x = 45; fig. 1B). Also, almost no redundant sequence was apparent. Corresponding BUSCOs indicated a high level of representation and minimal duplication (score of C:98.8% [S:98.5%, D:0.3%], F:0.3%, and M:0.9%; table 1), which is similar to chromosome-scale assemblies from other species in the Family Noctuidae (supplementary table S3). The NCBI Eukaryotic Genomic Annotation Pipeline predicted 16,988 genes, of which 14,922 protein coding genes gave rise to 23,683 RefSeq transcripts (table 1). Greater than 94% RefSeq models were supported by RNA-seq evidence. A total of 110.8 Mb of repeats and transposable elements were masked in the ilHelZeax1.1 assembly (supplementary table S4).

Mitochondrial genome assembly

The mitoPolishCLR pipeline (Formenti and Stahlke 2022) provided a robust assembly and annotation of the *H. zea* mitochondrial genome. The final HzStark_Cry1AcR mitochondria assembly was 15,351 bp and slightly larger than the 15,343 bp previously reported from a different *H. zea* strain (Perera et al. 2016; supplementary table S5). Size variation was accounted

for by four indels. Specifically, our assembly has a predicted 5 bp insertion at positions 11,662 to 11,666 located adjacent to *trnS2* that resulted from an AACTA duplication that accounts for the discrepancy in position of this tRNA between the two assemblies. Compared to KJ930516.1, *rpnL* in our mitochondrial genome assembly has a 10 bp insertion at positions 12,777 to 12,886 comprising five AT repeat units, and an AATATT deletion from 13,544 to 13,548. This comparison to KJ930516.1 additionally predicts that our assembly has an AT dinucleotide deletion and nine substitutions within the AT-rich control region. The GC-content in both assemblies was about 19%, and typical of insect mitochondrial genomes (Crozier and Crozer 1999). Order and orientation of the 13 protein-coding, 22 tRNA, and 2 rRNA genes in our assembly (supplementary fig. S1 and table S5) were typical of lepidopteran mitochondrial genomes, and identical to the previous *H. zea* assembly (Perera et al. 2016). Specifically, *trnM* was inverted compared to the consensus animal mitochondrial genome (Boore 1999) and a non-canonical Arginine CGA codon putatively functions in initiation of cytochrome c oxidase I (*coxI)* translation. No variation was detected within tRNA or protein coding genes compared to KJ930516.1. Absence of mitochondrial-derived k-mers in the Illumina k-mer database after two rounds of polishing indicated high accuracy and confidence in variation between our mitochondrial assembly compared and the previous accession.

Materials and Methods

Contig assemblies from continuous long read (CLR) PacBio data

Adult *H. zea* were collected near the Mississippi State University Campus in Starkville, MS, USA during 2011 and maintained in a colony at the USDA-ARS Southern Insect Management Unit (SIMRU), Stoneville, MS, USA as described previously (Gore et al. 2005). Larvae were selected on a diagnostic dose of 2.0 µg ml^{-1} purified Cry1Ac, and survivors used to

create the strain, HzStark_Cry1AcR. HzStark_Cry1AcRwas backcrossed every 5 generations to a susceptible line maintained at USDA-ARS SIMRU.

A single male pupa (homogametic with ZZ sex chromosome pair; ToLID ilHelZeax1) from HzStark_Cry1AcR was dissected laterally into eight ~20.0 µg sections. High molecular weight DNA was extracted from each section using the MagAttract HMW DNA Kit (Qiagen, Hilden, Germany) modified from manufacturer instructions to include gentle inversion to mix all components, an additional wash step, and eluting DNA from beads with three separate additions of 115.0 µl AE buffer at 37□C. Mean fragment size and quantity was estimated on a TapeStation 4200 (Agilent, Santa Clara, CA, USA; 100 bp to 48.5 kb ladder). HMW genomic DNA was sheared with a Covaris g-TUBE and a PacBio library was prepared using the SMRTbell Express Template Prep Kit 2.0 for Continuous Long Read (CLR) generation (Pacific Biosciences, Menlo Park, CA, USA). The library was size selected on a BluePippin (Sage Sciences, Beverly, MA, USA) with a 15.0 kb cutoff. The library was loaded on an 8M SMRTcell at a concentration of 35.0 pmol and run on a Sequel II (Pacific Biosciences) with Instrument Control Software Version 7.0 for a 20-hour movie time. Libraries were created from DNA of the same HzStark_Cry1AcR male using the TruSeq DNA PCR-Free Low Throughput Library Prep Kit with TruSeq DNA UD Indexes (Illumina, San Diego, CA, USA) to yield a standard paired end library with a 350±50 bp insert size. Paired end 150 bp sequence reads were generated on a NovaSeq 6000 (Illumina) at the Hudson Alpha Genome Sequencing Center (Huntsville, AL, USA). Genome size, repetitive content, and heterozygosity were estimated from 21-mer histograms of the paired-end Illumina reads generated with jellyfish (Marçais et al. 2011) and modeled in GenomeScope 2.0 (Ranallo-Benavidez et al. 2020).

For genome assembly, raw CLR subread BAM files were converted to FASTQ format using BamTools v2.5.1 (Barnett et al. 2011), then used as input for the FALCON assembler (Chin et al. 2016) using the pb-assembly conda environment v.0.0.8.1 (Pacific Biosciences; default parameters). FALCON-Unzip created primary and alternate contigs with one round of haplotype-aware polishing by Arrow (Pacific Biosciences). Duplicated sequence at the ends of primary contigs were removed with purge_dups v1.2.5 (Guan et al. 2020), with cutoffs estimated from an automatically generated histogram of k-mers. Purged primary sequences were added to the alternate haplotype contigs and purged of duplicates again.

Mitochondrial genome assembly

The mitochondrial genome was assembled and polished using a custom workflow, mitoPolishCLR (Formenti and Stahlke 2022), modified from mitoVGP (Formenti et al. 2021a) and improved for arthropods. Raw CLR reads were filtered to remove those > 1.5-times the known *H. zea* mitochondrial genome size (15,343 bp; Perera et al. 2016; GenBank accession KJ930516.1), thereby reducing low quality reads. Length filtered reads were retained if they shared homology to KJ930516.1 within results from BLASTn searches (default parameters; Camacho et al. 2009), then assembled using Canu v.2.2 (Koren et al. 2017) with parameters adjusted for organelle assemblies. Contig polishing was performed with FreeBayes, then evaluated by QV scores generated using Merqury (Rhie et al. 2020) and a 31-mer database of Illumina short reads generated by Meryl (Miller et al. 2008). After a final round of trimming the mitochondrial assembly was linearized and annotated with MitoFinder (Allio et al. 2020). Start and stop codons for protein-coding genes were manually reviewed according to guidelines for mitochondrial genomes of lepidoptera (Cameron 2014).

Scaffolding

Hi-C libaries were prepared at Dovetail Genomics (Santa Cruz, CA, USA) as prevously described (Lieberman-Aiden et al. 2009) from *H. zea* derived from the Benzon Reasearch colony (Carlisle, PA, USA) that were reared at the USDA-APHIS Forest Pest Methods Laboratory (Buzzards Bay, MA, USA). These libraries were sequenced at Dovetail on an Illumina HiSeq X (Illumina).

Hi-C data were then used for scaffolding by aligning them to contigs using bwa-mem v2.2.1 (Li 2013) with the -5SP options to allow chimeric read alignment. Matlock Hi-C processing (https://github.com/phasegenomics/matlock) and Juicebox v1.11.08 assembly tools (https://github.com/phasegenomics/juicebox_scripts) were used to convert the resulting BAM file and prepare the primary contig assembly for review. A manually curated contact-map relating the primary contigs to the aligned Hi-C reads was created in Juicebox v1.11.08. With the Hi-C contact map, we manually broke mis-joins, re-oriented inverted contigs, and joined contigs into scaffolds. The manually curated scaffold assembly was converted back to FASTA format for final polishing.

After scaffolding, the primary, alternate, and mitochondrial assemblies were merged and polished with one additional round of Arrow (Pacific Biosciences), followed by two rounds of variant identification using FreeBayes v1.0.2 (Garrison and Marth 2012). FreeBayes identified variants were filtered via Merfin (Formenti et al. 2021b) before being incorporated into a new polished consensus. The second round of Arrow-identified variants were also filtered via Merfin as the PacBio CLR and Illumina reads came from the same *H. zea* individual. Genome assembly quality scores were generated via Merqury (Rhie et al. 2020) before, between, and after each round of polishing to assess quality improvements. The polishCLR pipeline used here

(https://github.com/isugifNF/polishCLR) is a reproducible Nextflow workflow (Di Tommaso et al. 2017) inspired by the Vertebrate Genome Project assembly pipeline (Rhie et al. 2021).

Contigs derived from non-*H. zea* biological contamination (such as microbial symbionts) were screened using the BlobToolkit (Challis et al. 2020) which implements BLAST+ (Camacho et al. 2009) and Diamond (Buchfink et al. 2015), and parsed using a python script (https://github.com/sheinasim/FindContaminantsFromBlob). Final assembly completeness was assessed by BUSCO v.5.2.2 with the lepidoptera_odb10 gene set of 5,286 orthologs (Simao et al. 2015; Seppey et al. 2019; Manni et al. 2021), and composition evaluated by a 21-mer k-mer spectra using KAT v.2.4.2 (Mapleson et al. 2016).

Genome annotation

Repeats were identified using RepeatMasker v4.1.0 (Smit et al. 2005) using a combined repeat library built with *de novo* repeats using RepeatModeler v2.0.2 (Smit et al. 2008-2015) and lineage-specific Dfam v3.1 databases for Insecta and Lepidoptera (Storer et al. 2021). Scaffolded contigs from the primary assembly were submitted to the National Center for Biotechnology Information (NCBI) for automated eukaryotic genome annotation (Thibaud-Nissen et al. 2016).

Supplementary Materials

Supplementary data are available from *Genome Biology and Evolution* online.

Acknowledgements

This work was supported by the U.S. Department of Agriculture, Agricultural Research Service (USDA-ARS). The genome assembly was generated as part of the USDA-ARS Ag100Pest Initiative. The HzStark_Cry1AcR pupae for genome assembly were provided by USDA-ARS

Southern Insect Management Research Unit (SIMRU). Funding for PacBio CLR read data was provided by the USDA-ARS Corn Insects & Crop Genetics Research Unit project number 5030-22000-019-00D. This research used resources provided by the SCINet project of the USDA-ARS project number 0500-00093-001-00-D. Jennifer Chang was supported, in part, by an appointment to the Research Participation Program at USDA-ARS, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and USDA-ARS under contract number DE-AC05-06OR23100. Luke Tembrock was supported, in part, by a USDA-APHIS-PPQ cooperative agreement AP18PPQS&T00C074 to Colorado State University. Thanks members of the USDA-ARS Ag100Pest Team for sequencing and analysis support. Thanks to Sheron Simpson from the USDA-ARS Jamie Whitten Delta States Research Center, Genomics and Bioinformatics Research Unit for library preparation and sequencing of PacBio CLR libraries. Thanks to Calvin Pierce, USDA-ARS, SIMRU, Stoneville, MS for maintaining the HzStark_Cry1AcR colony, and Hannah Nadel and Lara Trozzo at the USDA Forest Pest Methods Laboratory in Buzzards Bay, MA for rearing the *H. zea* used to generate Hi-C libraries. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by USDA. USDA is an equal opportunity provider and employer.

Author Contributions

B.C., L.T., B.S., A.C., S.S., S.G., T.G., and K.H. designed and supervised the project; B.C., O.P., and L.T. prepared samples; A.S. J.C., S.C., and S.S analyzed data; B.C., A.S., and L.T.

293 interpreted results; B.C., A.S., and L.T. wrote a majority of the manuscript; All authors read and
294 approved the final manuscript.
295
296 **Data Availability**
297 Raw WGS CLR and Illumina sequence data were deposited at DDBJ/ENA/GenBank within
298 BioProject PRJNA804956, under the Sequence Read Archive (SRA) accessions SRR17965731
299 and SRR17993835, respectively. Raw Hi-C Illumina sequence data were deposited within
300 BioProject PRJNA788876, under SRA accession SRR17229576. The annotated primary
301 assembly version ilHeaZeax1.1 accession GCF_022581195.2 (BioProject PRJNA807638;
302 Annotation Release 100) and ilHeaZeax1.1 alternate haplotype assembly versions accession
303 GCA_022581175.1 (BioProject PRJNA807637) are available at NCBI. Both are under the
304 Ag100Pest umbrella project, BioProject PRJNA555319. The primary assembly and annotations
305 are also available at the i5k Workspace@NAL (Poelchau et al. 2015). The annotated
306 mitochondrial genome assembly is available in the GenBank accession OM990843.1.

Literature Cited

Allio R, Schomaker□Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020. MitoFinder: efficient automated large□scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Res.* 20(4):892-905.

Anderson CJ, Oakeshott JG, Tay WT, Gordon KH, Zwick A, Walsh TK. 2018. Hybridization and gene flow in the mega-pest lineage of moth, *Helicoverpa*. *Proc Natl Acad Sci USA.* 115(19):5034-5039.

Arnemann JA, James WJ, Walsh TK, Guedes JV, Smagghe G, Castiglioni E, Tay WT. 2016. Mitochondrial DNA COI characterization of *Helicoverpa armigera* (Lepidoptera: Noctuidae) from Paraguay and Uruguay. *Genet Mol Res.* 15(2):1-8.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools. *Bioinformatics*. 27(12):1691-1692.

Benowitz KM, Allan CW, Degain BA, Li X, Fabrick JA, Tabashnik BE., Carrière, Y. Matzkin, L.M, 2021. Novel genetic basis of resistance to Bt toxin Cry1Ac in *Helicoverpa zea*. *bioRxiv* https://doi.org/10.1101/2021.11.09.467966

Bilbo TR, Reay-Jones FP, Reisig DD, Greene JK, Turnbull MW. 2019. Development, survival, and feeding behavior of *Helicoverpa zea* (Lepidoptera: Noctuidae) relative to Bt protein concentrations in corn ear tissues. *PLoS One*. 14(8):e0221343.

Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* 27(8):1767-80.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 12(1):59-60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10(1):1-9.

Cameron S. 2014. How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. *Syst Entomol.* 39(3):400-411.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit–Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*. 10(4):1361-1374.

Childers AK, Geib SM, Sim SB, Poelchau MF, Coates BS, Simmonds TJ, Scully ED, Smith TP, Childers C, Corpuz RL, Hackett KJ, Scheffler BE. 2021. The USDA-ARS Ag100Pest Initiative: High-quality genome assemblies for agricultural pest insect research. *Insects*. 12(7):626.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 13(12):1050-1054.

341 Coates BS, Poelchau MF, Childers C, Evans JD, Handler AM, Guerrero F, Skoda SR, Hopper
342 KR, Wintermantel WM, Ling K, Hunter WB, Oppert BS, Perez De Leon AA, Hackett KJ,
343 Shoemaker DD. 2015. Arthropod genomics research in the United States Department of
344 Agriculture-Agricultural Research Service: Current impacts and future prospects. *Trends*
345 *Entomol*. 11(1):1-27.

346 Coates BS, Abel CA, Perera OP. 2017. Estimation of long terminal repeat element content in the
347 *Helicoverpa zea* genome from high-throughput sequencing of bacterial artificial chromosome
348 pools. *Genome*. 60(4):310-324.

349 Cordeiro EM, Pantoja-Gomez LM, de Paiva JB, Nascimento AR, Omoto C, Michel AP, Correa
350 AS. 2020. Hybridization and introgression between *Helicoverpa armigera* and *H. zea*: an
351 adaptational bridge. *BMC Evol Biol*. 20:1-2.

352 Crozier RH, Crozier YC. 1993. The mitochondrial genome of the honeybee *Apis mellifera*:
353 complete sequence and genome organization. *Genetics*. 133(1):97-117.

354 Cunningham JP, Zalucki MP. 2014. Understanding heliothine (Lepidoptera: Heliothinae) pests:
355 what is a host plant? *J Econ Entomol*. 107(3):881-896.

356 Czepak C, Albernaz KC, Vivan LM, Guimarães HO, Carvalhais T. 2013. First reported
357 occurrence of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in Brazil. *Pesquisa*
358 *Agropecuária Tropical*. 43:110-3.

359 Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow
360 enables reproducible computational workflows. *Nat Biotechnol*. 35(4):316-319.

361 Djaman K, Higgins C, O'Neill M, Begay S, Koudahe K, Allen S. 2019. Population dynamics of
362 six major insect pests during multiple crop growing seasons in northwestern New Mexico.
363 *Insects*. 10(11):369.

364 Degrande PE, Omoto C. 2013. Estancar prejuízos. *Cultivar Grandes Culturas Abril*. 167:32-35.

365 Dively GP, Venugopal PD, Finkenbinder C. 2016. Field-evolved resistance in corn earworm to
366 Cry proteins expressed by transgenic sweet corn. *PloS One*. 11(12):e0169115

367 Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, Brown S, Capodiferro MR,
368 Al-Ajli FO, Ambrosini R, Houde P. 2021a. Complete vertebrate mitogenomes reveal widespread
369 repeats and gene duplications. *Genome Biol*. 22(1):1-22.

370 Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, Myers EW, Jarvis ED,
371 Phillippy AM. 2021b. Merfin: improved variant filtering and polishing via k-mer validation.
372 bioRxiv doi:https://doi.org/10.1101/2021.07.16.452324

373 Formenti G, Stahlke A. 2022. Ag100Pest/mitoPolishCLR: v0.1.0-alpha-Hzea (v0.1.0-alpha).
374 Zenodo. https://doi.org/10.5281/zenodo.6235247

375 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.
376 arXiv preprint arXiv:1207.3907.

377 Gore J, Adamczyk Jr JJ, Blanco CA. 2005. Selective feeding of tobacco budworm and bollworm
378 (Lepidoptera: Noctuidae) on meridic diet with different concentrations of *Bacillus thuringiensis*
379 proteins. *J Econ Entomol*. 98(1):88-94.

380 Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing
381 haplotypic duplication in primary genome assemblies. *Bioinformatics*. 36(9):2896-2898.

382 Hamadain EI, Chambers HW. 2001. Susceptibility and mechanisms underlying the relative
383 tolerance to five organophosphorus insecticides in tobacco budworms and corn earworms.
384 *Pesticide Biochem Physiol*. 69(1):35-47.

385 Huseth AS, Koch RL, Reisig D, Davis JA, Paula-Moraes SV, Hodgson EW. 2021. Current
386 distribution and population persistence of five lepidopteran pests in US soybean. *J Integr Pest*
387 *Manag*. 12(1):11.

388 Howe K, Chow W, Collins J, Pelan S, Pointon DL, Sims Y, Torrance J, Tracey A, Wood J. 2021.
389 Significantly improving the quality of genome assemblies through curation. *GigaScience*.
390 10(1):giaa153.

391 Jacobson A, Foster R, Krupke C, Hutchison W, Pittendrigh B, Weinzierl R. 2009. Resistance to
392 pyrethroid insecticides in *Helicoverpa zea* (Lepidoptera: Noctuidae) in Indiana and Illinois. *J*
393 *Econ Entomol*. 102(6):2289-2295.

394 Jin L, Wang J, Guan F, Zhang J, Yu S, Liu S, Xue Y, Li L, Wu S, Wang X, Yang Y. 2018.
395 Dominant point mutation in a tetraspanin gene associated with field-evolved resistance of cotton
396 bollworm to transgenic Bt cotton. *Proc Natl Acad Sci USA. 115*(46):11760-11765.

397 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable
398 and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome*
399 *Res*. 27(5):722-736.

400 Lawniczak MK, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov
401 K, Blaxter ML, Bonet TM, Childers AK. 2022. Standards recommendations for the earth
402 BioGenome project. *Proc Natl Acad Sci USA.* 119(4):e2115639118.

403 Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I,
404 Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R. 2009. Comprehensive mapping of long-range
405 interactions reveals folding principles of the human genome. *Science*. 326(5950):289-293.

406 Leite NA, Alves-Pereira A, Corrêa AS, Zucchi MI, Omoto C. 2014. Demographics and genetic
407 variability of the new world bollworm (*Helicoverpa zea*) and the old world bollworm
408 (*Helicoverpa armigera*) in Brazil. *PloS One*. 9(11):e113286.

409 Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards
410 SV, Forest F, Gilbert MT, Goldstein MM. 2018. Earth BioGenome Project: Sequencing life for
411 the future of life. *Proc Natl Acad Sci USA*. 115(17):4325-4333.

412 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
413 arXiv preprint arXiv:1303.3997.

414 Li X, Zangerl AR, Schuler MA, Berenbaum MR. 2000. Cross-resistance to α-cypermethrin after
415 xanthotoxin ingestion in *Helicoverpa zea* (Lepidoptera: Noctuidae). *J Econ Entomol*. 93(1):18-
416 25.

417 Luttrell RG, Jackson RE. 2012. *Helicoverpa zea* and Bt cotton in the United States. *GM Crops*
418 *Food* 3(3):213-27.

419 Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing Genomic Data
420 Quality and Beyond. *Current Protocols*. 1(12):e323.

421 Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer
422 analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*.
423 33(4):574-576.

424 Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of
425 occurrences of k-mers. *Bioinformatics*. 27(6):764-770.

426 McCaffery AR. 1998. Resistance to insecticides in heliothine Lepidoptera: a global view. *Phil*
427 *Trans R Soc London B: Biol Sci*. *353*(1376):1735-1750.

428 Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry
429 C, Sutton A. 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*.
430 24(24):2818-2824.

431 Murúa MG, Cazado LE, Casmuz A, Herrero MI, Villagrán ME, Vera A, Sosa-Gómez DR,
432 Gastaminza G. 2016. Species from the Heliothinae complex (Lepidoptera: Noctuidae) in
433 Tucumán, Argentina, an update of geographical distribution of *Helicoverpa armigera*. *J Insect*
434 *Sci*. 16(1):61.

435 Niu G, Wen Z, Rupasinghe SG, Zeng RS, Berenbaum MR, Schuler MA. 2008. Aflatoxin B1
436 detoxification by CYP321A1 in *Helicoverpa zea*. *Arch Insect Biochem Physiol*. 69(1):32-45.

437 Pearce SL, Clarke DF, East PD, Elfekih S, Gordon KHJ, Jermiin LS, McGaughran A. 2017.
438 Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and
439 divergence of two highly polyphagous and invasive *Helicoverpa* pest species. *BMC*
440 *Biol*. 15(1):1-30.

441 Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee CY, Lin H, Lin JW, Hackett K.
442 2015. The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of
443 arthropod genomes. *Nucl Acids Res*. 43(D1):D714-719.

444 Perera OP, Walsh TK, Luttrell RG. 2016. Complete mitochondrial genome of *Helicoverpa zea*
445 (Lepidoptera: Noctuidae) and expression profiles of mitochondrial-encoded genes in early and
446 late embryos. *J Insect Sci*. 16(1):40.

447 Perera OP, Fescemyer HW, Fleischer SJ, Abel CA. 2020. Temporal variation in genetic
448 composition of migratory *Helicoverpa zea* in peripheral populations. *Insects*. 11(8):463.

449 Rabelo MM, Paula-Moraes SV, Pereira EJ, Siegfried BD. 2020. Demographic performance of
450 *Helicoverpa zea* populations on dual and triple-gene Bt cotton. *Toxins*. 12(9):551.

451 Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for
452 reference-free profiling of polyploid genomes. *Nat Comm*. 11(1):1-0.

453 Reay-Jones FP. 2019. Pest status and management of corn earworm (Lepidoptera: Noctuidae) in
454 field corn in the United States. *J Integr Pest Manag*. 10(1):19.

455 Reisig DD, Reay-Jones FP. 2015. Inhibition of *Helicoverpa zea* (Lepidoptera: Noctuidae) growth
456 by transgenic corn expressing Bt toxins and development of resistance to Cry1Ab. *Environ*
457 *Entomol*. 44(4):1275-1285.

458 Reisig DD, Kurtz R. 2018. Bt resistance implications for *Helicoverpa zea* (Lepidoptera:
459 Noctuidae) insecticide resistance management in the United States. *Environ Entomol*.
460 47(6):1357-1364.

461 Reynolds JA, Nachman RJ, Denlinger DL. 2019. Distinct microRNA and mRNA responses
462 elicited by ecdysone, diapause hormone and a diapause hormone analog at diapause termination
463 in pupae of the corn earworm, *Helicoverpa zea*. *General Compar Endocrinol*. 278:68-78.

464 Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality,
465 completeness, and phasing assessment for genome assemblies. *Genome Biol*. 21(1):1-27.

466 Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
467 Fungtammasan A, Kim J, Lee C. 2021. Towards complete and error-free genome assemblies of
468 all vertebrate species. *Nature*. 592(7856):737-746.

469 Rios DA, Specht A, Roque□Specht VF, Sosa□Gómez DR, Fochezato J, Malaquias JV,
470 Gonçalves GL, Moreira GR. 2021. *Helicoverpa armigera* and *Helicoverpa zea* hybridization:
471 constraints, heterosis, and implications for pest management. *Pest Manag Sci*.
472 https://doi.org/10.1002/ps.6705
473
474 Roach SH, Adkisson PL. 1970. Role of photoperiod and temperature in the induction of pupal
475 diapause in the bollworm, *Heliothis zea*. *J Insect Physiol*. 16(8):1591-1597.

476 Santiago González JC, Kerns DL, Head GP, Yang F. Status of Cry1Ac and Cry2Ab2 resistance
477 in field populations of *Helicoverpa zea* in Texas, USA. *Insect Sci*. https://doi.org/10.1111/1744-
478 7917.12947

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness In: Kollmar M, editor. *Gene prediction*. Humana, New York, NY. p. 227-245.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31(19):3210-3212.

Smit AFA, Hubley R, Green P. 2013-2015. *RepeatMasker Open-4.0*. http://www.repeatmasker.org

Smit AF, Hubley R, Green P. 2008–2015. RepeatModeler Open-1.0. http://www.repeatmasker.org

Soberón M, Gill SS, Bravo A. 2009. Signaling versus punching hole, how do *Bacillus thuringiensis* toxins kill insect midgut cells? *Cell Mol Life Sci*. 66(8):1337–1349.

Sosa-Gómez DR, Specht A, Paula-Moraes SV, Lopes-Lima A, Yano SA, Micheli A, Morais EG, Gallo P, Pereira PR, Salvadori JR, Botton M. 2016. Timeline and geographical distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera, Noctuidae: Heliothinae) in Brazil. *Revista Brasileira de Entomologia*. 60:101-104.

Stahlke AR, Chang J, Chadlayandi S, Rosen BD, Childers AK, Severin A. 2022. polishCLR: a Nextflow workflow for polishing PacBio CLR genome assemblies. bioRxiv. https://doi.org/10.1101/2022.02.10.480011

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*. 12(1):1-4.

Tay WT, Soria MF, Walsh T, Thomazoni D, Silvie P, Behere GT, Anderson C, Downes S. 2013. A brave new world for an old world pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. *Plos One*. 8(11):e80134.

Taylor KL, Hamby KA, DeYonke AM, Gould F, Fritz ML. 2021. Genome evolution in an agricultural pest following adoption of transgenic crops. *Proc Natl Acad Sci USA*. 118(52):e2020853118.

Tembrock LR, Timm AE, Zink FA, Gilligan TM. Phylogeography of the recent expansion of *Helicoverpa armigera* (Lepidoptera: Noctuidae) in South America and the Caribbean Basin. *Ann Entomol Soc Am*. 112(4):388-401.

Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts PA, Murphy TD, Pruitt KD, Souvorov A. 2016. P8008 the NCBI eukaryotic genome annotation pipeline. *J Animal Sci*. 94(suppl_4):184.

512 Valencia-Montoya WA, Elfekih S, North HL, Meier JI, Warren IA, Tay WT, Gordon KH,
513 Specht A, Paula-Moraes SV, Rane R, Walsh TK. 2020. Adaptive introgression across
514 semipermeable species boundaries between local *Helicoverpa zea* and invasive *Helicoverpa*
515 *armigera* moths. *Mol Biol Evol*. 37(9):2568-2583.

516 Xiao H, Ye X, Xu H, Mei Y, Yang Y, Chen X, Yang Y, Liu T, Yu Y, Yang W, Lu Z. 2020.0The
517 genetic adaptations of fall armyworm *Spodoptera frugiperda* facilitated its rapid global dispersal
518 and invasion. *Mol Ecol Res*. 20(4):1050-68.

519 Yang F, González JC, Williams J, Cook DC, Gilreath RT, Kerns DL. 2019. Occurrence and ear
520 damage of *Helicoverpa zea* on transgenic *Bacillus thuringiensis* maize in the field in Texas, US
521 and its susceptibility to Vip3A protein. *Toxins*. 11(2):102.

522 Zhang F, Zhang J, Yang Y, Wu Y. 2020. A chromosome-level genome assembly for the beet
523 armyworm (*Spodoptera exigua*) using PacBio and Hi-C sequencing. bioRxiv
524 https://doi.org/10.1101/2019.12.26.889121

Table 1

Helicoverpa zea genome assembly and annotation metrics

Metric	Value
Alternate contig assembly	
Total number	801
Assembly size (bp)	133,873,018
L50/N50 (bp)	90/464,725
L90/N90 (bp)	417/49,392
Largest contig (bp)	1,812,199
Primary assembly (ilHelZeax1.1)[1]	
Scaffolds	
Total number	42
Size (bp)	375,165,593
L50/N50 (bp)	14/12,841,304
L90/N90 (bp)	27/9,278,413
Largest (bp)	18,805,280
Contigs	
Total number	65
Size (bp)	375,163,385
L50/N50 (bp)	15/10,871,041
L90/N90 (bp)	34/3,915,976
Largest (bp)	15,512,169
BUSCOs	
Complete (C)	5227 (98.8%)
Complete single copy (S)	5209 (98.5%)
Complete duplicated (D)	18 (0.3%)
Fragmented (F)	17 (0.3%)
Missing (M)	42 (0.9%)
Gene structural annotations	
Genes (total)	16.988
Protein coding	14.933
Non-coding	1,812
Pseudogenes	242
RefSeq mRNAs	23,683
RefSeq proteins	23,696
Masked repeats bp (%)[2]	110,825,260 (29.5%)
Repeat content (%)[3]	23.6%

1. *Helicoverpa zea* WGS Project: GenBank accession: JAKYJW000000000
2. See RepeatMasker results in supplementary table S3

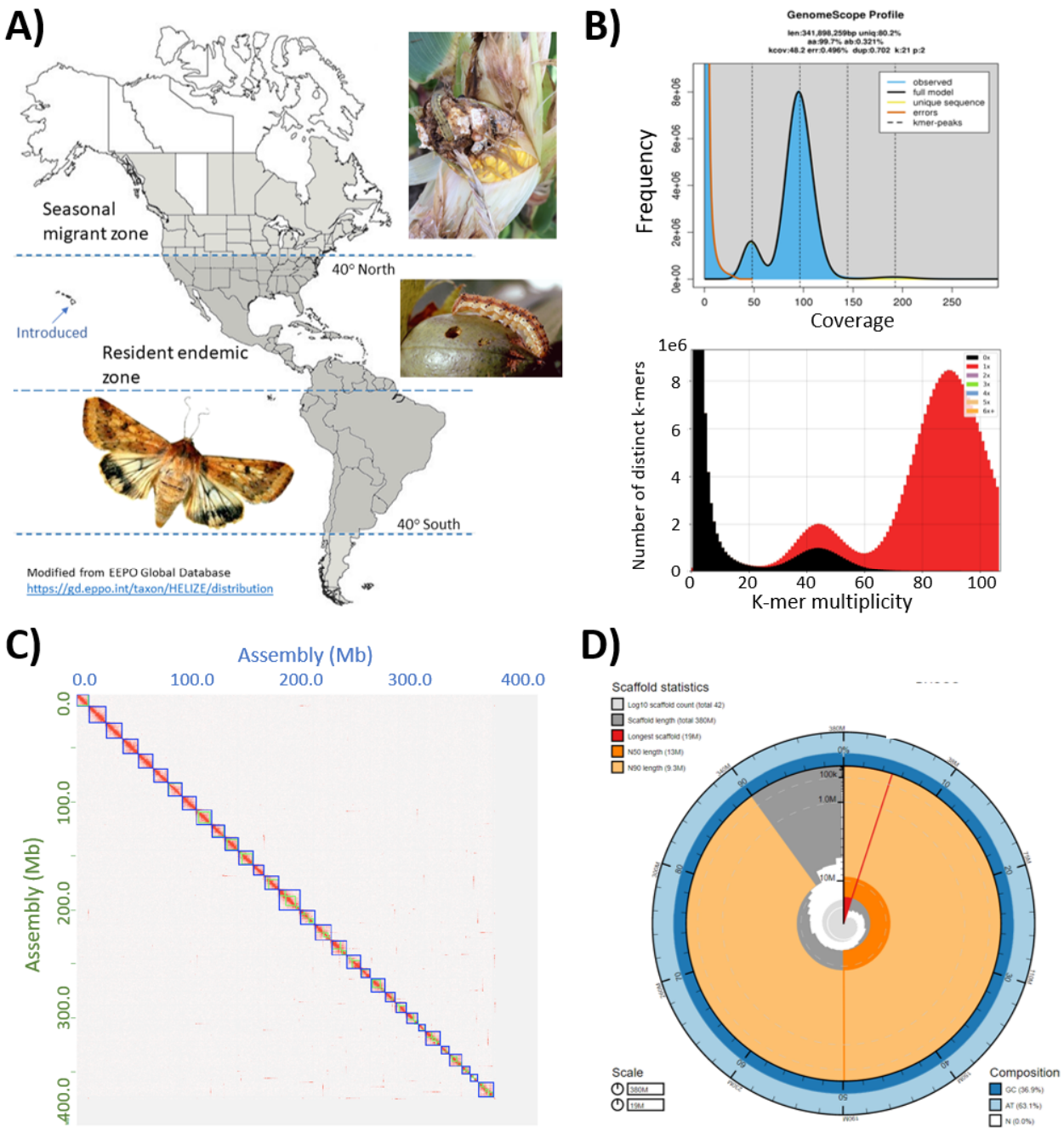

Fig. 1. Features of *Helicoverpa zea* biology and genome assembly. *(A)* Approximate geographic distribution and examples of crop damage, *(B)* K-mer analyses of short reads for estimated genome size, heterozygosity, and duplication level. *(C)* Hi-C contact map adjoining and ordering 31 pseudochromosomes. *(D)* Snailplot indicating metrics of 31 pseudochromomes within the entire assembly of 42 scaffolds.