Genetic Population Analyses of Invasive Agricultural Arthropod Pests *Drosophila suzukii* and *Tuta absoluta*

By

KYLE M. LEWALD

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

INTEGRATIVE GENETICS AND GENOMICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Joanna Chiu, Chair

Graham Coop

Jeffrey Ross-Ibarra

Committee in Charge

2023

Table of Contents

Acknowledgements

Thank you, Dr. Joanna Chiu, for mentoring me and giving me the opportunity to learn on a project that was outside the lab's usual scope of research. I appreciate your commitment to always meet with me whenever I needed advice on research direction. You have always been supportive of my career goals and constantly sought out opportunities to help me meet those goals. Thank you as well to my committee members Dr. Jeffrey Ross-Ibarra and Dr. Graham Coop for providing expertise and suggestions throughout my thesis work. Your advice greatly strengthened my understanding of population genetics.

I would like to thank all the friends I've made in my IGG cohort as well as all my Chiu lab members for being such a great support group. Whether it was talking about our research, career plans, or just enjoying each other's company, you all made graduate school so much easier to navigate and enjoy.

Thank you to my family members, in particular my parents, for supporting me throughout my academic journey and always being there when I needed. The values you instilled in me in my childhood have been driving forces in my life.

Finally, thank you to Michelle, my girlfriend, who has been at my side since high school. You have always supported me in graduate school and have been willing to listen to me talking about the latest scientific discovery even when you probably were not that interested. I could not have made it this far without you.

Abstract

Tuta absoluta and *Drosophila suzukii* are two agricultural pest insects that have both rapidly spread worldwide from their original ecosystems. Both cause serious economic losses, with the South American native *T. absoluta* targeting tomato agriculture and the Asian native *D. suzukii* targeting soft berry fruits. To inform current management strategies and prevent further introduction of these species, it is necessary to gain insights into current population structure and migration history of these species. Additionally, as *T. absoluta* has yet to be detected in North America, effective molecular diagnostics are needed to improve quarantine and monitoring efforts.

In chapter one, we sequenced whole genomes of hundreds of *D. suzukii* samples collected worldwide. We identified two major population clusters in the United States and found that West coast *D. suzukii* populations originated from a combination of migration events from Hawaii and Asia. We saw no strong loss in genetic diversity in invasive populations relative to Asian populations, suggesting ongoing migration is alleviating any bottleneck effects. In chapter two, we sequenced dozens of whole genomes of *T. absoluta* across Latin America and Spain and found three populations in Latin America. Using population simulation approaches, we found that these populations diverged prior to human agriculture of tomatoes. Additionally, we detected signals of selective sweeps near genes relevant to insecticide resistance and metabolism. In chapter three, we developed two molecular diagnostics to distinguish *T. absoluta* from two morphologically similar species already present in the United States. The probe-based quantitative PCR diagnostic can differentiate between *T. absoluta*, *Phthorimaea operculella*, and *Keiferia lycopersicella* using a thermocycler, while the RPA-Cas12a diagnostic can identity presence of *T. absoluta* in an isothermal reaction with minimal lab equipment requirements. We expect that these analyses and genomic resources will be of use to the agricultural research community in developing new strategies for controlling *T. absoluta* and *D. suzukii*.

Introduction

Globalization of human trade has led to great economic advances but has also resulted in serious environmental and economic threats in the form of the transport of organisms across large geographic distances. When a transported species ends up flourishing in its new home to the detriment of the pre-existing ecosystem or to our economic interests, we term this species "invasive". In the world of entomology, invasion biology is often associated with many serious problems, including vectoring animal and plant diseases, agricultural crop losses, and competition with native species. Focusing on plant pests, the most significant factors in the spread of invasive pests are the international trade networks of live plants, forestry products, and seeds (Chapman *et al.* 2017). Thus, in the modern era of global shipping and trade, the cost of preventing and controlling invasive pests has only grown higher, with the cost of management and damage growing threefold every decade since 1970 (Diagne *et al.* 2021).

An important aspect of invasion biology studies is the understanding of population history and structure. Knowledge of where an invasive species originated can allow government agencies to set up quarantine and shipping policies to prevent further introduction of the species, and prevent repeated re-introductions that increase transfer of genetic diversity to the invaded region, boosting the invasive species' chance of success (Estoup and Guillemaud 2010). Additionally, understanding whether distinct populations exist in an invaded area, or whether migrations are occurring between populations, can allow management efforts to customize tailored approaches to controlling each population (Fitzpatrick *et al.* 2012). Historically, molecular markers such as microsatellites, amplified fragment length polymorphisms, or mitochondrial sequences were used to differentiate populations and reconstruct population or species history (Bashasab *et al.* 2006; Darling and Blum 2007). However, cost reduction of whole-genome sequencing technologies have enabled relatively cheap access to hundreds of thousands

of single nucleotide polymorphism (SNP) markers. Non-coding SNPs distributed across the genome are ideal for accurately representing population history, as they are assumed to be unaffected by selective pressures (Trask *et al.* 2011). As long as individuals sequenced are sampled randomly across the regions of interest, using a large number of genomic SNPs from a relatively small number of individuals can be sufficient for assessing population relationships with one another (Willing *et al.* 2012). Thus, the large increase in number of markers that can be interrogated is driving the ability to recognize increasingly complex invasion patterns (Garnas *et al.* 2016).

In my first chapter, I perform a population analysis of *Drosophila suzukii*, an invasive vinegar fly originally from Asia (Peng 1937). *D. suzukii* was detected in Spain and California in 2008 (Hauser *et al.* 2009; Calabria *et al.* 2012), and within a few years rapidly spread worldwide. This pest is a particular problem to soft-flesh berries such as blueberries and raspberries, due to the serrated ovipositor of the female that allows it to lay eggs under the skin of fresh fruit (Walton *et al.* 2016). I used whole genome sequencing of hundreds of individual *Drosophila suzukii* flies to identify population structure and found evidence of two major populations in the United States, composed of an Eastern and Western group, but no evidence of North to South differentiation. We followed up this discovery with a migration analysis and found evidence that the Western U.S. populations originated from an Asian migration event with subsequent migrations from Hawaii. We additionally found evidence of back-migration from the Western U.S. to Asia, indicating that *Drosophila suzukii* is being exchanged in both directions, likely by agricultural shipping. Significantly, invasive populations in the U.S. or Europe showed no loss in genetic diversity relative to Asian ancestors, suggesting migrating populations are not undergoing bottlenecks or that migrations are ongoing, continually supplying genetic diversity to invaded regions (Lewald *et al.* 2021).

In my second chapter, I perform a similar analysis of population structure on *Tuta absoluta*, a moth responsible for massive tomato crop losses worldwide. *T. absoluta* was an agricultural pest in South

America throughout the late 20th century before it was detected in Spain in 2006 (EPPO 2008) before subsequently spreading across Europe, Africa, and Asia . For this analysis I focused on understanding the population structure of *T. absoluta* in Latin America. I first assembled and annotated a new genome using long-read technology to improve gene annotation and variant calling. Using whole genome sequencing of individuals collected across Latin America, I found evidence for three major clusters, with the Chilean populations identified the source of the initial European invasion in 2006. Selection statistics and diversity levels show several genomic regions have experienced recent selective sweeps. Using the new genome annotations, I found these regions contain genes important to insecticide resistance, immunity, and metabolism.

In addition to understanding population-level dynamics of invasive species, genomic data can also be used to develop molecular diagnostics, either for the purpose of identifying the species or identifying the geographic source of an individual (Venette and Hutchison 2021). Having access to highly reliable detection methods of an invasive species is key to any eradication events, both in detecting its presence, but also in detecting its absence so that prevention efforts can be efficiently used elsewhere (Tobin *et al.* 2014). While identification by morphology is typically used for insects, this strategy can become challenging if the insect is extremely small or is morphologically similar to other species already present in the surveilled area. Molecular diagnostics have the advantage of not requiring expert knowledge of the species' traits, and can often be faster, cheaper, and more accurate (Stouthamer *et al.* 1999; Gariepy *et al.* 2007).

In my third chapter, I used available genomic data to develop a molecular diagnostic to identify *T. absoluta.* Unlike *D. suzukii*, *T. absoluta* has not yet been detected in the United States. Monitoring farms and greenhouses for the appearance of *T. absoluta* is a critical step in preventing the establishment of this pest, but identification is hampered by morphologically identical species already in the U.S. that inhabit the same ecological niche. Thus, I develop two types of molecular diagnostics that enable

detection of *T. absoluta* from DNA samples. The first relies on quantitative probe-based PCR, while the second diagnostic uses a newly developed CRISPR-Cas12a system, which enables rapid detection with simpler equipment than quantitative real-time PCR.

In summary, my thesis provides genomic resources and a knowledge of migration history and patterns for these two agricultural pests with significant economic implications. This information, combined with our new molecular diagnostics, should enable agricultural agencies to make better decisions for shipping and quarantine policy, and allow for rapid effective responses to future invasion events.

References

Bashasab, F., Vijaykumar, K. Kambalpally, B. Patil, and M. Kuruvinashetti, 2006 DNA-based marker systems and their utility in entomology. Entomol. Fenn. 17: 21–33.

Calabria, G., J. Máca, G. Bächli, L. Serra, and M. Pascual, 2012 First records of the potential pest species *Drosophila suzukii* (Diptera: Drosophilidae) in Europe: First record of *Drosophila suzukii* in Europe. J. Appl. Entomol. 136: 139–147.

Chapman, D., B. V. Purse, H. E. Roy, and J. M. Bullock, 2017 Global trade networks determine the distribution of invasive non-native species. Glob. Ecol. Biogeogr. 26: 907–917.

Darling, J. A., and M. J. Blum, 2007 DNA-based methods for monitoring invasive species: a review and prospectus. Biol. Invasions 9: 751–765.

Diagne, C., B. Leroy, A.-C. Vaissière, R. E. Gozlan, D. Roiz *et al.*, 2021 High and rising economic costs of biological invasions worldwide. Nature 592: 571–576.

EPPO, 2008 First record of *Tuta absoluta* in Spain. EPPO Report. Serv.

Estoup, A., and T. Guillemaud, 2010 Reconstructing routes of invasion using genetic data: why, how and so what? Mol. Ecol. 19: 4113–4130.

Fitzpatrick, B. M., J. A. Fordyce, M. L. Niemiller, and R. G. Reynolds, 2012 What can DNA tell us about biological invasions? Biol. Invasions 14: 245–253.

Gariepy, T. D., U. Kuhlmann, C. Gillott, and M. Erlandson, 2007 Parasitoids, predators and PCR: the use of diagnostic molecular markers in biological control of Arthropods. J. Appl. Entomol. 131: 225–240.

Garnas, J. R., M.-A. Auger-Rozenberg, A. Roques, C. Bertelsmeier, M. J. Wingfield *et al.*, 2016 Complex patterns of global spread in invasive insects: eco-evolutionary and management consequences. Biol. Invasions 18: 935–952.

Hauser, M., S. Gaimari, and M. Damus, 2009 *Drosophila suzukii* new to North America. Fly Times 12–15.

Lewald, K. M., A. Abrieux, D. A. Wilson, Y. Lee, W. R. Conner *et al.*, 2021 Population genomics of *Drosophila suzukii* reveal longitudinal population structure and signals of migrations in and out of the continental United States. G3 jkab343.

Peng, F. T., 1937 On some species of *Drosophila* from China. Annot. Zool. Jpn. 16: 20–27.

Stouthamer, R., J. Hu, F. J. P. M. van Kan, G. R. Platner, and J. D. Pinto, 1999 The utility of internally transcribed spacer 2 DNA sequences of the nuclear ribosomal gene for distinguishing sibling species of Trichogramma. BioControl 43: 421–440.

Tobin, P. C., J. M. Kean, D. M. Suckling, D. G. McCullough, D. A. Herms *et al.*, 2014 Determinants of successful arthropod eradication programs. Biol. Invasions 16: 401–414.

Trask, J. A. S., R. S. Malhi, S. Kanthaswamy, J. Johnson, W. T. Garnica *et al.*, 2011 The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*). Primates 52: 129–138.

Venette, R. C., and W. D. Hutchison, 2021 Invasive Insect Species: Global Challenges, Strategies & Opportunities. Front. Insect Sci. 1: 650520.

Walton, V. M., H. J. Burrack, D. T. Dalton, R. Isaacs, N. Wiman *et al.*, 2016 Past, present and future of *Drosophila suzukii:* distribution, impact and management in United States berry fruits. Acta Hortic. 87–94.

Willing, E.-M., C. Dreyer, and C. van Oosterhout, 2012 Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. PLOS ONE 7: e42649.

Chapter 1: Population genomics of *Drosophila suzukii* reveal longitudinal population structure and signals of migrations in and out of the continental United States

Kyle M. Lewald[1], Antoine Abrieux[1], Derek A. Wilson[1], Yoosook Lee[2], William R. Conner[1], Felipe Andreazza[3], Elizabeth H. Beers[4], Hannah J. Burrack[5], Kent M. Daane[6], Lauren Diepenbrock[7], Francis A. Drummond[8], Philip D. Fanning[8], Michael T. Gaffney[9], Stephen P. Hesler[10], Claudio Ioriatti[12], Rufus Isaacs[13], Brian A. Little[14], Gregory M. Loeb[10], Betsey Miller[15], Dori E. Nava[3], Dalila Rendon[15], Ashfaq A. Sial[14], Cherre S. Bezerra da Silva[15], Dara G. Stockton[10, 11], Steven Van Timmeren[13], Anna Wallingford[10, 16], Vaughn M. Walton[15], Xingeng Wang[17], Bo Zhao[5], Frank G. Zalom[1], Joanna C. Chiu[1*]

[1] Department of Entomology and Nematology, College of Agricultural and Environmental Sciences, University of California, Davis, CA, USA

[2] Florida Medical Entomology Laboratory, University of Florida Institute of Food and Agricultural Sciences, Vero Beach, FL, USA

[3] Laboratory of Entomology, Embrapa Clima Temperado, BR 392 Km 78, Caixa Postal 403, Pelotas, RS 96010-971, Brazil

[4] Tree Fruit Research and Extension Center, Washington State University, Wenatchee, WA, USA

[5] Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA

[6] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA

[7] UF IFAS Citrus Research and Education Center, University of Florida, Lake Alfred, FL, USA

[8] School of Biology and Ecology, University of Maine, Orono, ME, USA

[9] Horticultural Development Department, Teagasc, Ashtown, Dublin 15, Ireland

[10] Department of Entomology, Cornell AgriTech, Cornell University, Geneva, NY, USA

[11] Present affiliation: USDA-ARS, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center, Hilo, HI; USA

[12] Technology Transfer Centre, Fondazione Edmund Mach, Via E. Mach, 1, 38010 San Michele all'Adige (TN), Italy

[13] Department of Entomology, Michigan State University, East Lansing, MI, USA

[14] Department of Entomology, University of Georgia, Athens, GA, USA

[15] Department of Horticulture, Oregon State University, Corvallis, OR, USA

[16] Present affiliation: Department of Agriculture, Nutrition & Food Systems, University of New Hampshire, Durham, NH

[17] USDA Agricultural Research Service, Beneficial Insects Introduction Research Unit, Newark, DE, USA

*Corresponding author: Joanna C. Chiu, email: jcchiu@ucdavis.edu

Contributions and Acknowledgements

K.M.L., A.A., Y.L. and J.C.C. conceived the study. A.A. and D.A.W. performed nucleic acid extraction, library preparation and quality control. K.M.L. and W.R.C. performed bioinformatic and population genomics analysis. K.M.L. and J.C.C. wrote the manuscript with the input from all authors. All other authors performed field collections, provided samples for WGS, and edited the manuscript.

We thank Dr. Graham Coop and Dr. Jeffrey Ross-Ibarra for advice on population genomics analysis and Ernest Lee for advice on bioinformatics analysis. This project was supported by USDA SCRI 2015-51181-

24252 and USDA SCRI 2020-67013-30976 to JCC and co-authors; USDA APHIS Cooperative Agreement 17-8130-0194-CA to HJB. KML was supported by UC Davis Summer GSR Award.

Abstract

Drosophila suzukii, or spotted-wing drosophila, is now an established pest in many parts of the world, causing significant damage to numerous fruit crop industries. Native to East Asia, *D. suzukii* infestations started in the United States (U.S.) a decade ago, occupying a wide range of climates. To better understand invasion ecology of this pest, knowledge of past migration events, population structure, and genetic diversity is needed. In this study, we sequenced whole genomes of 237 individual flies collected across the continental U.S., as well as several sites in Europe, Brazil, and Asia, to identify and analyze hundreds of thousands of genetic markers. We observed strong population structure between Western and Eastern U.S. populations, but no evidence of any population structure between different latitudes within the continental U.S., suggesting there is no broad-scale adaptations occurring in response to differences in winter climates. We detect admixture from Hawaii to the Western U.S. and from the Eastern U.S. to Europe, in agreement with previously identified introduction routes inferred from microsatellite analysis (Fraimout *et al.* 2017). We also detect potential signals of admixture from the Western U.S. back to Asia, which could have important implications for shipping and quarantine policies for exported agriculture. We anticipate this large genomic dataset will spur future research into the genomic adaptations underlying *D. suzukii* pest activity and development of novel control methods for this agricultural pest.

Key Words: *Drosophila suzukii*, spotted-wing drosophila, invasion genomics, population structure, genetic diversity

Introduction

Over the past decade, *Drosophila suzukii* (Matsumura), also known as the spotted-wing drosophila or the Asian vinegar fly, has become an incredibly invasive pest species and a threat to soft-skinned fruit agricultural production worldwide (dos Santos *et al.* 2017). Unlike the large majority of Drosophilidae (Diptera), which preferentially breed in decaying plant material, female *D. suzukii* possess a serrated ovipositor, enabling them to lay eggs in fresh ripening soft-skinned fruits (Walsh *et al.* 2011; Walton *et al.* 2016). First described in Japan as an agricultural pest of cherries, *D. suzukii* was primarily distributed across East Asia until researchers found wild specimens in Hawaii in 1980 (Peng 1937; Kanzawa 1939; Kaneshiro 1983). In 2008, *D. suzukii* was detected in California, and by 2009 was widespread across the Western U.S. coast (Hauser *et al.* 2009; Bolda *et al.* 2010). In the Eastern U.S., *D. suzukii* first appeared in Florida in 2009 (Steck *et al.* 2009), before again rapidly spreading across the entire east coast within a few years. Meanwhile in Europe, *D. suzukii* was first detected in Spain and Italy in 2008 and rapidly spread across Europe, appearing in France, Switzerland, Austria, Germany, and Belgium by 2012. Subsequently, *D. suzukii* arrived in South America when it was detected in Brazil in 2013 (Deprá *et al.* 2014), Argentina in 2014 (Cichon *et al.* 2015), and Chile in 2015 (Medina-Muñoz *et al.* 2015). Its rapid spread across continents suggests that human transportation is likely a major factor, as eggs laid in fresh fruit are difficult to detect before shipment. Once established in a new continent, *D. suzukii* rapidly disperse to neighboring regions, aided by its ability to adapt to a wide range of climates through phenotypic plasticity (Shearer *et al.* 2016). In the Western U.S. coastal states alone, estimated economic losses were as high as 511 million dollars per year, assuming a 20% average yield loss (Bolda *et al.* 2010). Thus, there is much interest in understanding the patterns of migration and origin of these invasive populations, as these data can be used to inform shipping and quarantine policies and to identify routes of entry.

Previous research on the population genomics of *D. suzukii* was performed using a relatively small number of molecular markers. Adrion et al. (2014) used six X-linked gene fragments from flies collected across the world, and detected signals of differentiation between European, Asian, and U.S. populations. However, they found no evidence of differentiation within the 12 U.S. populations sampled, possibly due to the limited power provided from a small number of markers. A follow-up study using 25 microsatellite loci of samples collected between 2013-2015 greatly improved estimations of migration patterns worldwide; the authors found evidence for multiple invasion events from Asia into Europe and the U.S. as well as an East-West differentiation in the 7 populations sampled in the continental U.S. (Fraimout *et al.* 2017). However, using microsatellites alone may miss more subtle signals of population structure compared to genome-wide datasets, as increasing the number of independent loci genotyped increases accuracy of population parameter estimates, even when the number of biological samples is low (Trask *et al.* 2011; Willing *et al.* 2012; Rašić *et al.* 2014). With the advent of affordable whole-genome sequencing (WGS), it has become feasible to sequence hundreds of individuals to study population genomics, enabling improved inference of population structure using hundreds of thousands to millions of single nucleotide polymorphism (SNP) markers (Soria-Carrasco *et al.* 2014; Wu *et al.* 2019; Lee *et al.* 2019). A study of *D. suzukii* in Hawaii used double digest restriction-site associated DNA sequencing to identify several thousand SNPs and observed population structure between islands (Koch *et al.* 2020). However, a comprehensive survey of *D. suzukii* in the continental U.S. using a large number of SNPs enabled by WGS has not been conducted.

In this study, we leverage the power of WGS to individually sequence hundreds of *D. suzukii* samples to determine whether U.S. populations are stratified along a north-south cline corresponding to varying winter climates, as well as to detect whether migration is freely occuring between the Eastern and Western U.S. In addition, we include several populations from Asia, Europe, and Brazil to determine frequency and source of international migrations and compare genetic diversity between invasive and

native populations. We expect these analyses and the large sequencing dataset will be of value in developing policies and furthering research into mitigating the harmful effects of *D. suzukii* worldwide.

Methods

Sample collection and genomic DNA extraction

We received either flash-frozen or ethanol-preserved samples of *D. suzukii* for genomic analysis. Japanese samples were obtained from the Ehime Japanese Stock Center (strain #E-15003; MTY3; originally collected in Ehime, Japan). Hawaiian samples were taken from a small lab population maintained in vials that was established in 2009 from wild-caught samples in Oahu, Hawaii. All other samples were field-collected. Ethanol-preserved samples were re-hydrated in 100uL water prior to DNA extraction. Flies were individually disrupted using a 3mm diameter steel bead in a TissueLyser (Qiagen, Germantown, MD) for 30 seconds at 30Hz in 100uL of 2mg/mL Proteinase K in PK buffer (MagMAX™, Thermofisher Scientific, Pleasanton, CA) before being spun down in a centrifuge for 1 minute at 10,000rpm and incubated for 2 hours at 56°C. 100uL of MagMAX DNA lysis buffer was added to each sample, followed by a 10 minute incubation, before proceeding to DNA purification using a BioSprint DNA Blood Kit on a BioSprint 96 Workstation (Qiagen), using protocol “BS96 DNA Tissue” as per manufacturer’s instructions.

Library preparation and sequencing

Illumina WGS libraries were prepared with either the Kappa HyperPlus Kit (Roche, South San Francisco, CA) (lanes 2-4) or Qiaseq FX DNA Library Kit (Qiagen) (lanes 5-8) using 50 ng of input DNA following the manufacturer’s instructions with few exceptions. With the Kappa HyperPlus Kit, DNA was fragmented at 30°C for 20 minutes and incubated with adapters for 1 hour. A 0.6X and 0.7X size selection with AmPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN) was added after 5 cycles of PCR amplification with an Eppendorf Master Cycler Pro (ThermoFisher Scientific). With the Qiagen FX kit, DNA was fragmented at 30°C for 15 minutes and amplified with 7 cycles of PCR. In both cases, DNA library

concentration and fragment size were quantified on a Qubit 2.0 fluorometer with the Qubit dsDNA HS assay kit (ThermoFisher Scientific) and a Bioanalyzer High-Sensitivity DNA chip (Agilent, Santa Clara, CA). Paired-end 150 base-pair sequencing was performed by Novogene, Inc. (Sacramento, CA) on the Illumina HiSeq 4000 platform.

Genome alignment

Raw Illumina reads were inspected for quality using fastqc version 0.11.5 (Babraham Institute, Cambridge, UK), and trimmed for low quality bases and adapter sequences using Trimmomatic version 0.35 (Bolger *et al.* 2014), using the following parameters: Leading qscore threshold = 10, trail score threshold = 10, minimum read length = 36, and illuminaclip=2:30:10. Reads were then aligned to the *D. suzukii* reference genome obtained from Dr. Benjamin Prud'homme (now available at GenBank accession GCA_013340165.1) (Paris *et al.* 2020) using bwa-mem version 0.7.9a (Li 2013), sorted by samtools-sort version 1.3.1 (Wellcome Trust Sanger Institute, London, UK), de-duplicated with picardtools-MarkDuplicates version 2.7.1 (Broad Institute, Cambridge, MA), and indexed with samtools-index version 1.3.1. Samtools-stats was used to obtain summary statistics of BAM files. Based on consistently low mapping rates for all samples (<70%), the Dandong, China population and one of two Watsonville, California, U.S. collections were excluded from analysis.

COX2 sequence analysis to confirm species identification

The *D. suzukii* mitogenome sequence and ten *D. pulchrella* COX2 sequences were downloaded from NCBI. The *Drosophila subpulchrella* mitogenome was identified by running BLAST with the *D. suzukii* mitogenome against the *D. subpulchrella* genome assembly (GCA_014743375.2), and annotated using MITOS2 (Bernt *et al.* 2013). COX2 sequences from all our *D. suzukii* samples were identified by aligning raw reads to the *D. suzukii* mitogenome (GenBank accession KU588141.1), filtering out any read pairs where one of the reads was unmapped (samtools view –f 2 –F 4). Variants were called with Freebayes version 1.1.0 in haploid mode (Garrison and Marth 2012), and fasta sequences were extracted with

bcftools-consensus version 1.10.2 (Wellcome Trust Sanger Institute). Publicly available COX2 sequences of *D. pulchrella, D. suzukii, D. biarmipes, D. lutescens, D. mimetica,* and *D. melanogaster* were downloaded from GenBank.

All COX2 sequences were aligned with the ClustalOmega web portal (Madeira *et al.* 2019) resulting in a 720 base pair alignment. Forty-seven haplotypes were identified using DNASP version 6.12.03 (Rozas *et al.* 2017). MEGA version 10.1.8 was used to identify the best nucleotide substitution model based on the Bayesian Inference Criterion score (Kumar *et al.* 2018). While the best scoring model was the Tamura 3-parameter model (T92) + invariant sites (+I) + gamma distributed rates (+G), we decided to use the second best scoring model T92+G, as combining +I and +G may be problematic due to correlated parameters (Jia *et al.* 2014). Using MEGA, initial trees for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura 3 parameter model and then selecting the topology with superior log likelihood value. Bootstrap percentages were generated from 500 replicate runs. Five categories were used in the discrete Gamma distribution to model evolutionary rate differences among sites (+*G*, parameter = 0.1066).

Principal component and admixture proportions analysis

Genotype likelihoods (GLs) were estimated from aligned reads with ANGSD version 0.935 (Korneliussen *et al.* 2014) using the samtools model with the following parameters: minimum mapping quality = 20, minimum base quality = 20, uniquely mapping reads only, excessive mismatch adjustment coefficient = 50. Using a p-value cutoff of 1E-6 and a minimum minor allele frequency = 0.05, ANGSD identified 4,955,596 SNPs. Linkage disequilibrium was estimated from the GLs and SNP list using ngsLD version 1.1.1 (Fox *et al.* 2019) to a maximum SNP pairwise distance of 100kb, and a 1% subsample was used to estimate and plot LD decay using the accompanying Rscript “fit_LDdecay.R” with the following parameters: bin size = 100bp, bootstraps = 100, fit level = 10, and a recombination rate = 2.3cM, based

on the average rate in *D. melanogaster* (Fiston-Lavier *et al.* 2010; Comeron *et al.* 2012). Based on these plots, GLs were pruned to one SNP per 1kb or 5kb, leaving 209,243 and 49,127 SNPs, respectively.

Principal component analysis was performed using PCAngsd version 1.0 (Meisner and Albrechtsen 2018). PCAngsd also reports the optimal number of clusters that describes the population structure (k), based on a minimum average partial (MAP) test. For each region, PCangsd was run 5 times with a random starting seed value and the soft upper search bound of alpha = 500; the run with the highest log likelihood was kept. Admixture proportions were then estimated using NGSadmix version 32 (Skotte *et al.* 2013). For each sample set, the run with the highest log likelihood from 5 independent runs was kept. Analysis of each region used k based on the number of optimal PCs + 1 reported from PCangsd. Analysis of all samples combined as well as 4 subsampled datasets were performed with k ranging from 3 to 10. As there was no notable difference in results using the 1kb or 5kb pruned dataset, only the 1kb pruned results are reported here.

Treemix admixture graphs and F3/F4 statistics

Based on admixture proportion estimates and PCA, samples were grouped into the following clusters: Eastern U.S., Western U.S., Hawaii, Brazil, Ireland, Italy, South Korea, and Japan. One Eastern U.S. population from Alma Research Farm, Georgia (AR) was excluded as it unexpectedly clustered with the Western U.S. populations. To root the population trees, two sister species of *D. suzukii* were downloaded and aligned to the *D. suzukii* genome; *D. biarmipes* (SRA accession SRX097584) and *D. subpulchrella* (SRA accession SRX8970519). GLs and SNPs were called with ANGSD as described above and pruned by only keeping SNPs found originally in the 1kb pruned dataset used for PCA and admixture analysis. As X-linked and autosomal SNPs may have different phylogenetic signals, X-linked SNPs were excluded. As Treemix requires genotypes to be called, PCAngsd was used to call genotypes from the GLs with a 95% posterior probability cutoff using estimated inbreeding coefficients as a prior. When looking at the distribution of fraction of missing genotypes per site, we observed a peak at 10%, and decided to

exclude any sites with greater than 20% missing data across samples, consistent with cutoffs in other studies (Brandenburg *et al.* 2017; Zecca *et al.* 2019). We also excluded sites for which data are completely missing within any one cluster as required by Treemix, leaving 29,145 SNPs for analysis.

Treemix version 1.13 (Pickrell and Pritchard 2012) was used to generate population admixture graphs with inferred migrations. Between 0 to 10 migrations were tested, each with 100 bootstraps calculated using a resampling block size of 500 SNPs, with global tree rearrangements and standard error estimation of migration weights enabled. The bootstrap run with maximum likelihood for each migration tested was used for plotting. To estimate support for migration edges, Treemix was also used to calculate F3 and F4 statistics using a resampling block size of 500 SNPs to estimate standard error and Z-scores. The F3 statistic tests if population A's allele frequencies are a result of mixture of allele frequencies from populations B and C. A significantly negative value of F3(A;B,C) supports admixture of B or C into A. The F4 statistic measures correlations in allele frequencies between populations A and B versus populations C and D. F4(A,B;C,D) is expected to be zero under no admixture. Assuming the tree ((A,B),(C,D)) exists, a significantly positive value suggests gene flow between A and C or B and D, while a significantly negative value suggests gene flow between B and C or A and D. By setting one of these populations to be an outgroup where no admixture is expected, it is possible to infer which population pair experienced admixture. We used a Z score cutoff of 2 or -2 to determine if a value was significantly positive or negative.

Estimation of F_{ST}, nucleotide diversity, and Tajima's D

Site allele frequencies were estimated for each cluster from the largest 20 contigs of the genome assembly with ANGSD version 0.933 using the samtools GL model, filtering for a minimum mapping quality = 20 and minimum base quality = 20. The subprogram realSFS from ANGSD was used to estimate a two-dimensional folded SFS for all pairs of clusters and to calculate the global weighted Fst for each cluster pair. RealSFS was also used to estimate one-dimensional folded SFS for each cluster in order to

calculate nucleotide diversity and Tajima's D per site. The command ANGSD/thetaStat do_stat was used to bin these summary statistics in 20kb windows across the genome.

Data Availability

Raw Illumina reads have been deposited to the NCBI Sequence Read Archive (SRA) and can be found under BioProject accession number PRJNA705744. Scripts used to run all analyses can be found at www.github.com/ClockLabX/Dsuz-popgen.

Results

Population structure exists between continents as well as within the U.S. and Europe

To determine if population structure exists in *D. suzukii* living in recently invaded locations, we sequenced wild caught individual *D. suzukii* flies collected from the continental U.S., Brazil, Ireland, Italy, South Korea, and China, as well as a laboratory strain from Hawaii and Japan (Figure 1). After aligning sequences to the reference genome, we found that average read coverage was low for some individuals and populations, with mean coverage per cluster ranging from 5-11X. As low coverage can cause biases in genotype calling, we used methods that implemented genotype likelihoods wherever possible.

We first used PCA and admixture proportion estimates to search for signs of population structure. When examining our Asian samples, we were surprised to discover that all the Namwon, South Korea samples as well as one Sancheong, South Korea sample clustered tightly with the Kunming, China population, rather than with the rest of the Sancheong samples (Figure S2C, Figure S3D). As several sister species to *D. suzukii* with similar morphological appearances occupy the same geographic ranges (Takamori *et al.* 2006), we performed a phylogenetic analysis using the mitochondrial COX2 gene sequence to evaluate species identity (Figure S4). Based on phylogenetic inference, we determined that the Namwon, South Korea samples; Kunming, China samples; and one Sancheong, South Korea sample may actually be *D. pulchrella*. For this reason, these samples were excluded from further analyses.

As sampling was heavily concentrated in the U.S., we first conducted PCA and admixture proportion estimation on each broad geographic region separately before analyzing all populations together (Figure S2, Figure S3). Among the Eastern U.S. samples, PCA did not separate samples by state or latitude, and no distinct populations emerged in admixture plotting at multiple clustering values (k). Among the Western U.S. samples, both the first principal component and varying values of k for admixture proportions separates Hawaii from the other sample sites; however, higher values of k and principal components do not further partition the remaining Western U.S. samples. Thus, it appears there is likely no strong population structure in a north to south cline in the U.S. Using a similar approach, we see that in the European samples, collections from Ireland and Italy partition as separate clusters in the first PC and when k=2 in admixture plotting. We also observe that samples from Asia partition into Japan and South Korea, which is unsurprising as the Japanese samples originate from a lab population.

We then used PCA to analyze all samples together to examine how differentiated invasive populations were from each other and from the ancestral Asian samples (Figure 2A). As subtler signals can be obscured by unequal population sampling (McVean 2009; Toyama *et al.* 2020), we also analyzed a reduced dataset by subsampling 5 individuals from each region (Figure 2B-C). When using all samples, the first principal component separates Eastern and Western U.S. populations, with Asian and European samples in-between. Samples from Pelotas, Rio Grande do Sul, Brazil, appear more related to Eastern U.S. samples, although one individual clusters more with the Western U.S. flies. We also noticed that all samples collected from the Alma Research Farm (AR), Georgia clustered with the Western rather than Eastern U.S. samples, despite two other Georgia sites nearby that followed the expected pattern. The second principal component then separates the European samples. When the data is subsampled to 5 individuals per cluster, the first and second components strongly separate Hawaii and Japan, respectively (Figure 2B); this signal was likely obscured by the large number of U.S. samples when all

samples are analyzed together but is expected as these two populations were lab strains and have likely experienced significant genetic drift relative to wild relatives.

The observations made from PCA are largely recapitulated when using subsampled data to estimate admixture at varying levels of k (Figure 2C). At k=3, we observe Japanese and Hawaiian samples form their own clusters, while all the wild collections form a third cluster. As k is increased up to 7, we see the appearance of Europe, Brazil, Eastern U.S., and South Korea samples as their own clusters, before samples from Europe are split into Ireland and Italy at k=8. We notice increased variability in cluster assignment in the U.S. populations, particularly when subsampling, which likely reflects the large sample size and high within-population diversity (Figure S5). However, analysis using all individuals still clearly support Eastern and Western U.S. samples as distinct genetic populations (Figure S6). In addition, we also see that the AR Georgia population again cluster with the Western U.S. As we were unsure if this could be the result of a very recent migration or mislabeled samples, we decided to exclude this population from further analyses.

To further quantify the amount of differentiation present between regions, we estimated Fst values between regions using the 20 largest contigs, spanning all 4 chromosomes and covering 54% (145Mb) of the reference genome (Figure 3A). Three general levels of differentiation were apparent based on this analysis. As expected, Fst between Hawaii or Japan to any wild population was high (>0.30). Irish and Italian populations had intermediate levels of differentiation with the other wild populations and with each other (0.15-0.30), while Fst values between Brazil, South Korea, and both U.S. clusters were lower (0.05-0.10). These groupings broadly match those observed from PCA (Figure 2A-B).

Repeat migrations to and from the U.S. have occurred over large geographic distances

While PCA and admixture proportion estimates were able to identify population clusters, they are unable to provide more detailed depictions of population history or migration events. To estimate the

population history of these invasive populations, we used Treemix to generate a population admixture graph with inferred migration events based on co-variance of allele frequencies between clusters, testing models allowing between 0 to 10 migrations (m) (Figure S7). We found that the model using six migrations captured the most variance of the data (99.6%) (Figure 4). Residuals of the model at m=6 are within +/- 5 standard errors between populations, suggesting the model fits the data well, despite the variance of Hawaii with itself appearing less well modeled (Figure S8). The strongest signal of admixture was found in the Western U.S., with an estimated Hawaiian admixture proportion of 41.0% (SE = 6.9%, $p < 0.05$), and was also observed in most models (m = 4-8,10). To formally test for admixture, we used the F3 admixture statistic in the form F3(Western U.S.; Hawaii, popX) where popX represents any third population, and found significantly negative values for all populations (Z score < -2), strongly supporting admixture of Hawaii into the Western U.S. We also used the F4 statistic, using the form F4(A, B; C, outgroup) such that a negative value supports "B" and "C" admixture, while a positive value supports "A" and "C" admixture, assuming no migration occurred between the outgroup and either A or B. Using either *D. biarmipes* or *D. subpulchrella* as the outgroup, the tests F4(Western U.S., Brazil; Hawaii, outgroup) and F4(Western U.S., Eastern U.S.; Hawaii, outgroup) were significantly positive (Z score > 2), again supporting this admixture. Thus, the Western U.S. population sampled is composed of nearly equal ancestry from a Hawaiian ancestor and the common ancestor of the U.S./Brazil populations. As Treemix assigns the edge with smaller weight to be the "migrant" edge by default, it may be unidentifiable whether the Hawaiian ancestor or the U.S./Brazil common ancestor should be called the migration source.

We also observed two countries with U.S. admixture in the m=6 model. Ireland had an Eastern U.S. admixture of 25.3% (SE = 2.7%, $p < 0.05$), although at varying values of "m" the source of this admixture fluctuates between the Eastern U.S., Brazil, or the Eastern U.S./Brazil ancestor. However, in all cases the admixture strength and significance remain consistent. While no F3 statistic support was found, the F4

statistics (Western U.S., Brazil; Ireland, outgroup) and (Western U.S., Eastern U.S.; Ireland, outgroup) were significantly negative, supporting Ireland's Eastern U.S./Brazilian and European ancestry. As the U.S./Brazilian admixture weight is much less than the European admixture weight, this was likely due to a migration event from the Americas into Irish populations.

The other out-of-U.S. admixture event, from the Western U.S. to South Korea (admixture 23.1%, SE = 3.6%, $p < 0.05$), was seen when m=6, 8, and 10. F3 statistics (South Korea; Western U.S., Italy/Ireland/Japan) all have significantly negative values, and the F4 statistics (Western U.S., Eastern U.S.; South Korea, outgroup) and (Western U.S., Brazil; South Korea, outgroup) are significantly positive, supporting a Western U.S./South Korea admixture. However, using 9 migration edges Treemix reported the reverse direction; as F3 and F4 statistics cannot easily infer directionality, more heavily sampling of the Asian populations or alternate methods may be needed to determine whether flow is occuring in both directions.

Invasive populations have experienced little loss in genetic diversity

To determine if invasive populations have experienced loss in genetic diversity, we used the software ANGSD to estimate average pairwise nucleotide diversity in 20 kb increments across the 20 largest contigs of genome for each population. Invasive populations can sometimes exhibit reduced levels of diversity early on in their history due to a founder effect (Nei *et al.* 1975), while ancestral populations tend to have the greatest amount of diversity as they have had many generations to accumulate mutations. A Welch one-way test ($F=1590.9$, $p < 0.05$) found a significant difference in mean pairwise nucleotide diversity between clusters. We then used pairwise Games-Howell tests and found each cluster to be significantly different ($p < 0.05$), except for the Eastern U.S., Brazil, and Italy when compared to each other. As Asia is the ancestral home of *D. suzukii*, it is no surprise that South Korean wild populations exhibit the highest diversity levels (Figure 3B). Similarly, the lab populations from Japan and Hawaii have half as much pairwise diversity as the wild South Korean population, consistent with a

small lab population size. The invasive populations display an intermediate level between these extremes.

To assess whether invasive populations may have experienced a bottleneck or population shrinkage, we also estimated Tajima's D in the same genomic intervals. Extremely positive values (>2) can indicate a loss of rare alleles, which can occur during a population shrinkage, while extremely negative values (<-2) can indicate a recent bottleneck followed by rapid expansion (Tajima 1989). A Welch one way test again indicated significant differences in mean Tajima's D between clusters ($F=45598$, p-value < 0.05), and pairwise Games-Howell tests found all clusters to be statistically different ($p < 0.05$) except for Western U.S. against Brazilian flies. Strains from Hawaii and Japan both had high genome-wide levels of Tajima's D, indicative of a loss of rare alleles that can occur during a population shrinkage (Figure 3C). The remaining populations had neutral values of D, except for Ireland's relatively high value. Based on this, we conclude there are no strong signals for a recent bottleneck, although the high genome-wide D value for Ireland suggests a recent population shrinkage. As our Irish samples were collected in 2016 only one year after its discovery in Ireland, we could be observing the founder's effect in action (Gaffney 2017).

Discussion

Based on population allele frequencies, we have shown that *D. suzukii* exhibit population structure based on region and invasion history. In the New World populations, we find that Eastern and Western U.S. samples appear to be distinct populations. While this could be the result of continuous population variation from East through Central to the West coast, it is more likely the case that the two populations experience little gene flow due to strong geographic barriers such as the Sierra Nevada or Rocky Mountain ranges, and the fact that key target fruit crops such as cherries, raspberries, blueberries, and strawberries are primarily grown in states that we sampled ("Noncitrus Fruits and Nuts 2019 Summary" 2020). Any genetic exchange between these regions would likely be the result of human activity, such as could be the case with samples collected from Alma Research Farm, Bacon County, Georgia (AR)

clustering with the Western U.S. populations. As other nearby collections (AL, BD) failed to share this signal, the Alma research population could represent a recent and isolated migration event. Otherwise, we see little evidence of migration events or admixture between the Eastern and Western U.S., which is somewhat surprising as the country's supply of fresh blueberries, cherries and caneberries are concentrated in a few states (Pacific Northwest, Michigan, Maine) and shipped across the country ("Noncitrus Fruits and Nuts 2019 Summary" 2020). However, recent changes to cultural management such as more frequent harvesting and post-harvest chilling may be responsible for disrupting the *D. suzukii* lifecycle and limiting cross-country transport (Schöneberg *et al.* 2021).

While we were able to detect population structure between eastern and western locations in the U.S., we were surprised to discover a lack of structure on a finer scale, either based on latitude or simple geographic distance, given the large number of loci analyzed. In a similar study using 3,484 SNPs in 246 Hawaiian *D. suzukii* samples, researchers were able to identify three distinct populations roughly seperated by islands (Koch *et al.* 2020). The fact that *D. suzukii* has been present in Hawaii since 1980, in addition to the isolation by island, are likely the strongest factors in providing enough genetic drift to create such differentiation. As the continental U.S. *D. suzukii* have only been present since 2008, it may be too early for finer structure to have developed. Alternatively, continual dispersion and transportation of *D. suzukii* around the U.S. may be hindering the development of more local structure.

Several studies have reported a low probability of *D. suzukii* surviving when exposed to freezing temperatures, based on cold survival assays of wild-caught specimens (Dalton *et al.* 2011; Stephens *et al.* 2015), suggesting that flies collected in cold-winter regions such as Washington, Michigan, Maine, and New York could be annual migrants to the area from nearby warmer locations. The lack of north-south population structure supports the hypothesis that flies are regularly re-migrating into colder climates after the harsh winters have ended. Alternatively, flies could be tolerating winters by surviving inside human structures (Stockton *et al.* 2019), or by having evolved resistance to freezing temperatures

(Stockton *et al.* 2020). Studies using *D. suzukii* collected from different locations have reported different levels of rapid cold-hardening response, suggesting there could be regional selection present (Jakobs *et al.* 2015; Everman *et al.* 2018; Stockton *et al.* 2020). If populations in northern regions undergo strong seasonal fluctuations in allele frequencies, such as has been demonstrated in wild *D. melanogaster* populations collected in Pennsylvania (Bergland *et al.* 2014), by only sampling sites in the summer we may be missing signals of population differentiation between the north and south. Likely, some combination of these factors is responsible for the success of *D. suzukii* in these regions, and further studies will be needed to identify the causes. North-south clines in specific traits such as diapause and circadian rhythms have been previously identified in drosophilids and could be at play here as well (Schmidt *et al.* 2005; Tyukmaeva *et al.* 2011). Further analyses using methods such as those recently used to detect SNPs correlated with invasive success (Olazcuaga *et al.* 2020) could be applied to this dataset to find signals of selection.

Fst values between populations from the U.S., Brazil, and South Korea were low and agree with previously published Fst estimates based on Pool-Seq data; Olazcuaga *et al.* 2020 observed that Fst between U.S., European, Asian, and Brazilian populations varied between 8.86% to 9.02%. However, we were surprised to see that our Italian and Irish samples had much higher values of Fst compared to the other populations, and even to each other. This discrepancy could be due to the small sample sizes we had from Europe; in this scenario, pooling larger number of samples can improve power to estimate Fst, and we instead rely on comparing the relative Fst values between populations for our analysis. High Fst values between our Japanese and Hawaiian populations were expected, however, as these have likely experienced strong drift during their time in captivity.

In general, we find that our treemix and migration results largely coincide with the proposed invasion pathway inferred from microsatellites (Fraimout *et al.* 2017), as well as a recent pre-print that re-analyzed invasion pathways with pooled sequencing data (Gautier *et al.* 2021). We see that European

and U.S./Brazilian populations form two distinct clades, emphasizing these regions were invaded by two independent migrations from Asia. Hawaii is the first population to diverge in the Americas, followed by the Western U.S., then the Eastern U.S. and Brazil. Additionally, in the Western U.S., we detected a strong signal of admixture from Hawaii, which could be due to multiple or ongoing migration events. We also detected signals of admixture from the Eastern U.S./Brazil to Ireland, which matches the predicted initial invasion pathway and suggests multiple migration events. Unique to our analysis, we recover support for admixture of Western U.S. samples in Asia, suggesting that migrations could be ongoing in both directions. Invasive species transport is strongly associated with international trade of live plants and plant products (Chapman *et al.* 2017), and indeed agricultural export data supports the possibility of this migration as Japan receives almost 15% of all U.S. blueberry exports, and Oregon recently became the first state to begin shipping blueberries to South Korea in 2012 (Evans and Ballen 2014). It should be noted that while Treemix infers direction of migration, the model can occasionally infer the incorrect direction, particularly when populations are closely related without an available outgroup (Pickrell and Pritchard 2012). More sampling of Asian populations are likely needed to confirm the direction of this admixture.

In conjunction with evidence of this widespread ongoing migration, we observed nucleotide diversity levels of all invasive populations (excluding lab populations) to be only moderately below that of the wild South Korean population, a trend also observed in Fraimout et al. (2017). Typically, recent invasion events are characterized by reduced diversity relative to the ancestral populations due to founder or bottleneck effects (Dlugosch and Parker 2008). However, successive invasion events can provide relief from any initial bottlenecks by providing increased genetic diversity. This has been observed to occur in multiple animal studies (Johnson and Starks 2004; Kolbe *et al.* 2004), and could lead to increased ability to adapt and evolve to new climates. Correspondingly, in our analysis we did not find populations with broadly low values of Tajima's D, suggesting little bottleneck effect. As measures to reduce impacts of

invasive species are often hindered by repeated migrations (Garnas *et al.* 2016), it will be important to enforce that fruits being exported and imported internationally are free of live *D. suzukii* as required by the U.S. Department of Agriculture, even though this species is already internationally established.

We anticipate that the genomic data provided here will prove useful in many fields of biology beyond the scope of this study. Knowledge of genetic variation and alternate alleles present within a species can be informative for the design of probes and micro RNAs (miRNAs), such as for the purpose of creating gene drives to control invasive species. Gene drive mechanisms to eliminate *D. suzukii* have been experimentally tested on multiple lines to ensure the miRNAs are broadly effective (Buchman *et al.* 2018), but a large dataset of wild population sequencing allow researchers to more confidently select target sites that are non-variable and thus susceptible to Cas9 targeting (Schmidt *et al.* 2020). Drury *et al*., (2017) demonstrated that minor natural polymorphisms in target sites reduce gene drive effectiveness in flour beetles, and tools have been developed to help researchers design gRNAs accounting for population variation (Chen *et al.* 2020). Similarly, with the recent development of a CRISPR-Cas9 editing and RNAi knockdown protocols for *D. suzukii* (Murphy *et al.* 2016; Li and Scott 2016; Taning *et al.* 2016; Ahmed *et al.* 2020), prior knowledge of allelic variation will allow researchers to design targeting oligonucleotides more precisely to avoid loci with variability. Most recently, our dataset has been used to study sensory receptor evolution in *D. suzukii*, giving insights into its evolution toward becoming a major agricultural pest (Durkin *et al.* 2021). Other future uses of this trove of genomic data could involve insecticide resistance studies or the development of diagnostic assays for rapid detection in the field.

Figures

Figure 1: Sampling sites of *D. suzukii* populations. Sampling sites from the U.S., Europe, Brazil, and Asia. Labels indicate population code; colors and symbols depict population clusters as determined using PCangsd and NGSadmix. Note site AR has been labeled as "West US" based on clustering results. Between 5-10 flies per site were collected for WGS.

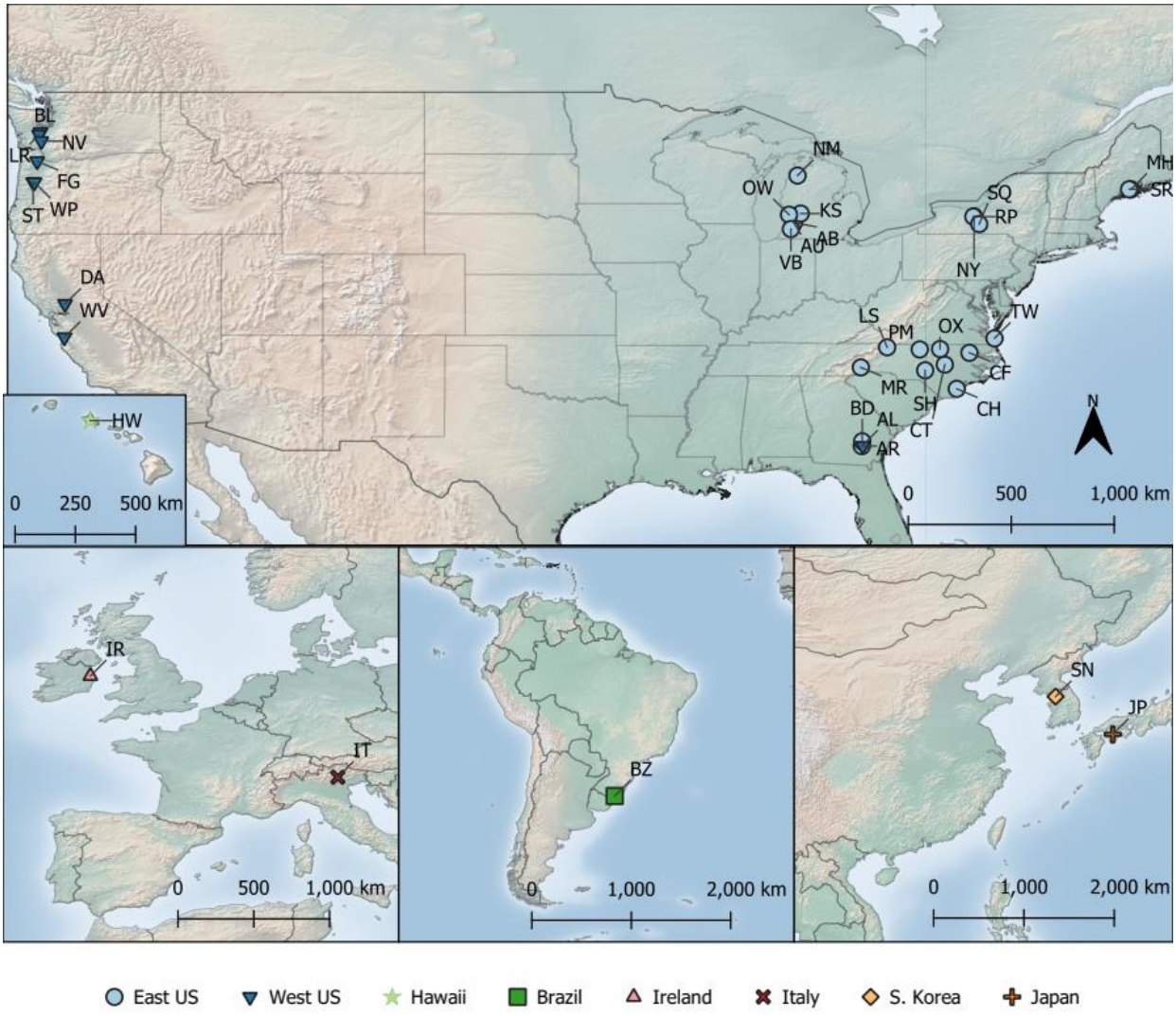


Figure 2: Population structure of *D. suzukii* populations. **(A)** First 2 principal components plotted of all samples based on 154,271 SNPs. Note several "Eastern US" samples representing Alma Research Farm, Georgia, as well as one Brazilian sample, clustering with the Western US. Percent variation of the data captured by each component indicated in axis labels. **(B)** First 2 principal components of sub-sampled dataset, using 5 individuals per cluster. **(C)** Posterior probability of cluster identity using NGSadmix calculated from 152,876 SNPs, using between 3 to 8 clusters. Samples labeled by name and cluster.

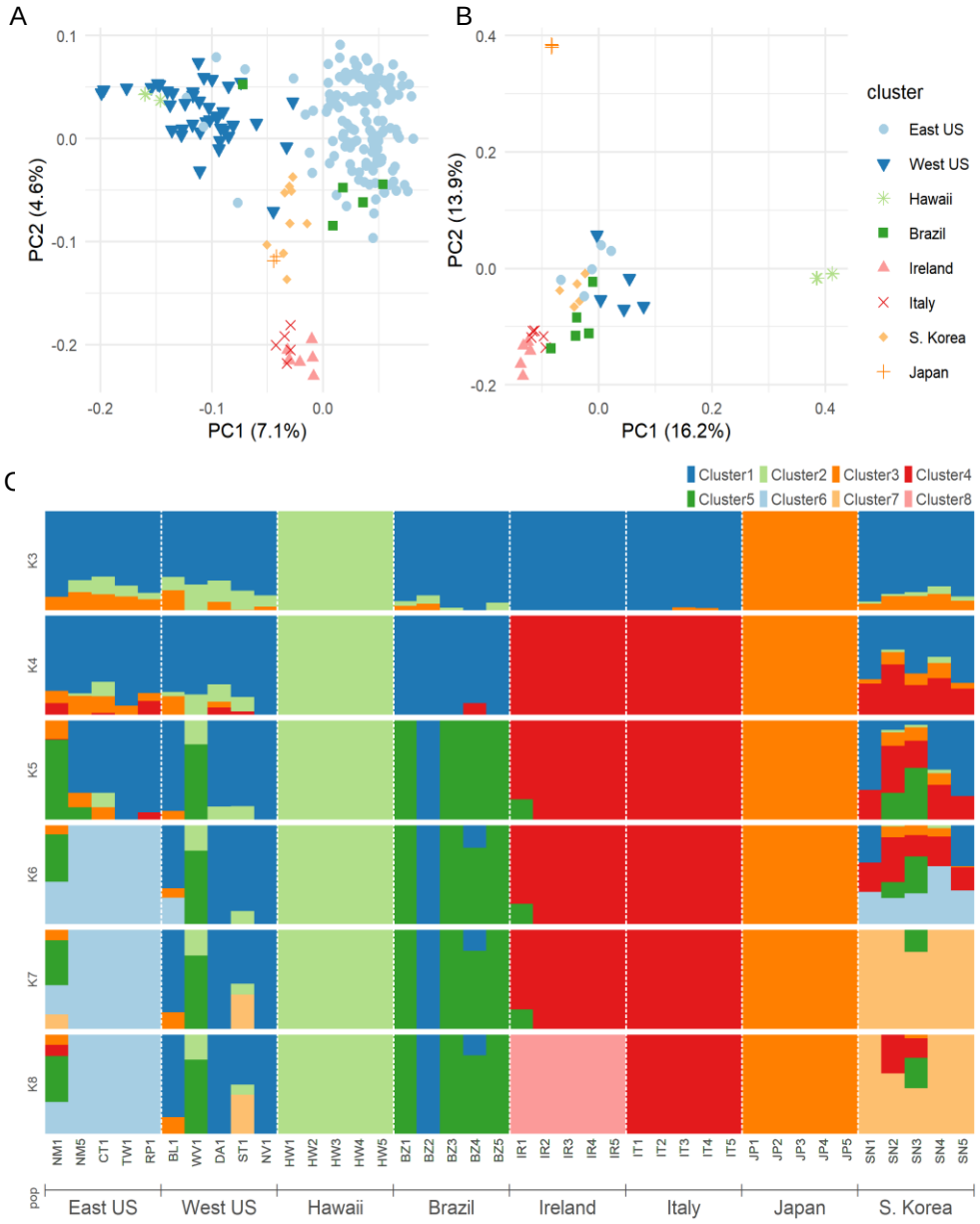


Figure 3: Population summary statistics of sampled *D. suzukii* populations. **(A)** Pairwise weighted Fst calculated from the largest 20 contigs of the reference genome between populations. **(B)** Pairwise nucleotide diversity distribution and **(C)** Tajima's D distribution, calculated in 20 kb intervals across the largest 20 contigs of the reference genome (7237-7238 blocks). Boxplots depict median, 1st and 3rd quartiles. Average values labeled along the x-axis.

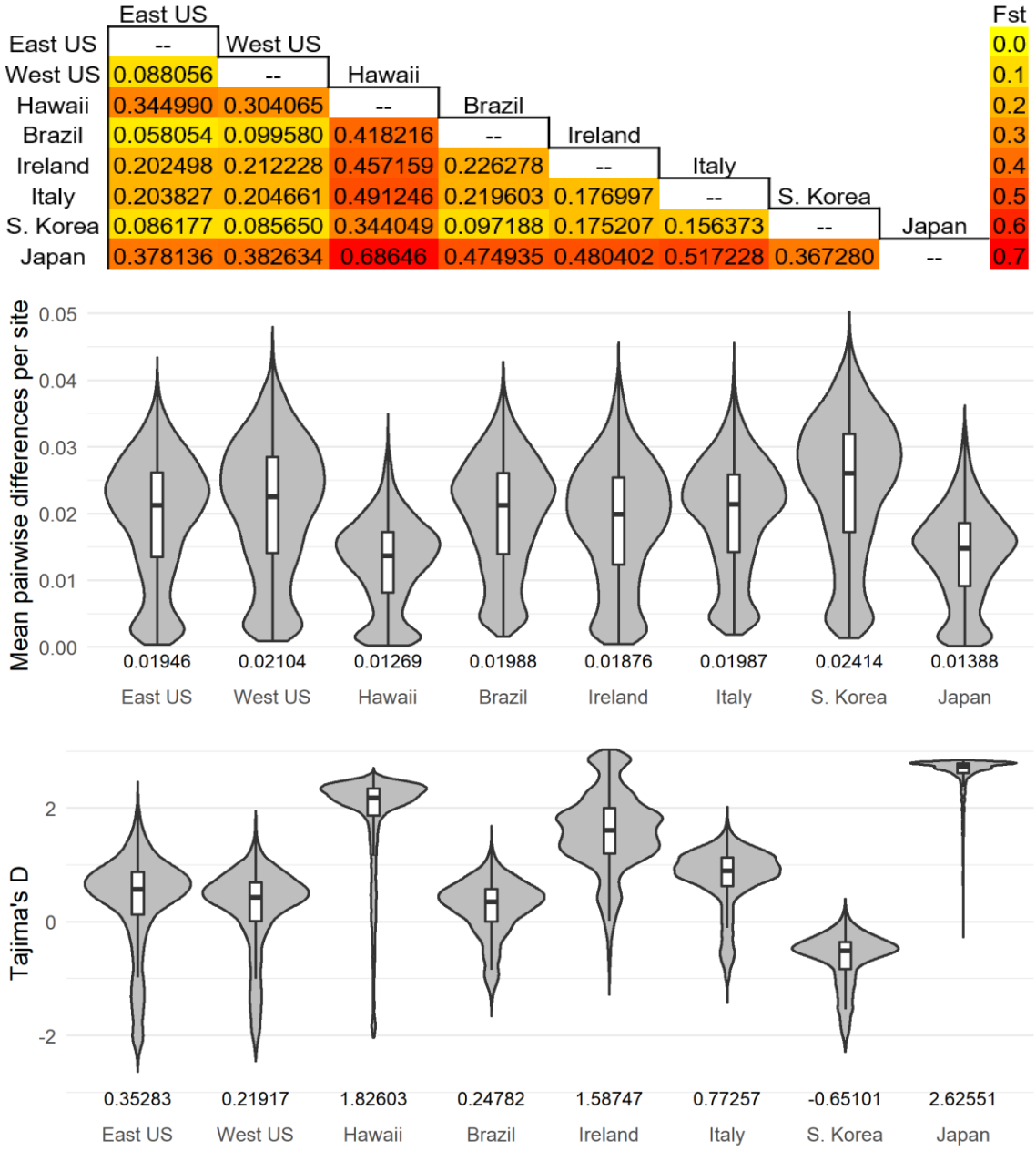
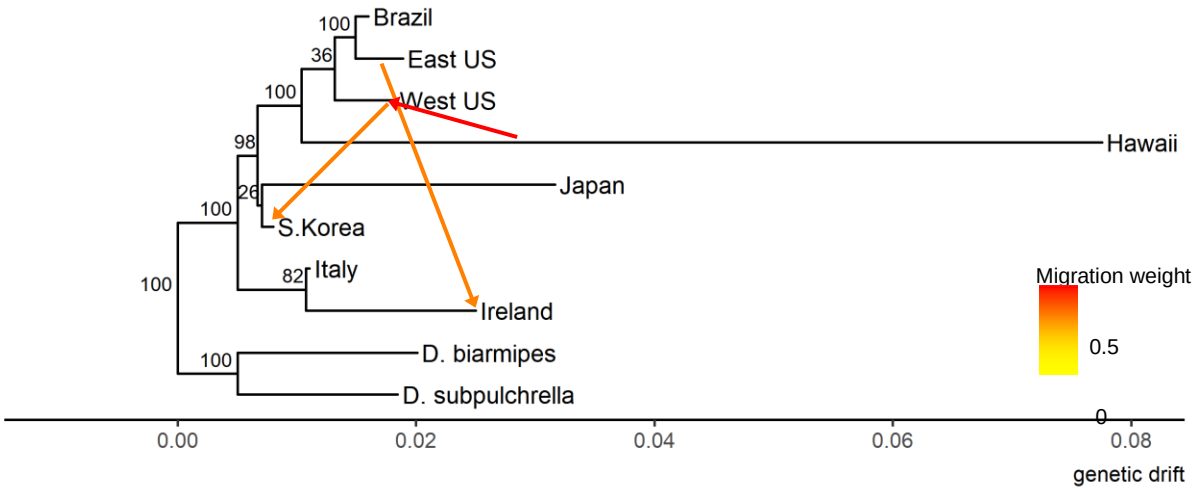


Figure 4: Maximum likelihood admixture graph. Graph is based on allele frequencies and allowed up to 5 migrations. The 3 strongest migrations are shown, colored by admixture proportion; Hawaii to Western U.S. (0.410, SE=0.069, p=1.10E-9), Eastern U.S. to Ireland (0.253, SE=0.027, p=0.0), and Western U.S. to S. Korea (0.231, SE=0.036, p=5.4E=11). Nodes labeled with jackknife bootstrap confidence percentages obtained from 100 replicates.



References

Ahmed, H. M. M., F. Heese, and E. A. Wimmer, 2020 Improvement on the genetic engineering of an invasive agricultural pest insect, the cherry vinegar fly, *Drosophila suzukii*. BMC Genet. 21: 139.

Bergland, A. O., E. L. Behrman, K. R. O'Brien, P. S. Schmidt, and D. A. Petrov, 2014 Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila* (D. Bolnick, Ed.). PLOS Genet. 10: e1004775.

Bernt, M., A. Donath, F. Jühling, F. Externbrink, C. Florentz *et al.*, 2013 MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69: 313–319.

Bolda, M. P., R. E. Goodhue, and F. G. Zalom, 2010 Spotted Wing Drosophila: potential economic impact of a newly established pest. Agric. Resour. Econ. Update Univ. Calif. Giannini Found. 13: 5–8.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. 30: 2114–2120.

Brandenburg, J.-T., T. Mary-Huard, G. Rigaill, S. J. Hearne, H. Corti *et al.*, 2017 Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts (N. M. Springer, Ed.). PLOS Genet. 13: e1006666.

Buchman, A., J. M. Marshall, D. Ostrovski, T. Yang, and O. S. Akbari, 2018 Synthetically engineered *Medea* gene drive system in the worldwide crop pest *Drosophila suzukii*. Proc. Natl. Acad. Sci. 115: 4725–4730.

Chapman, D., B. V. Purse, H. E. Roy, and J. M. Bullock, 2017 Global trade networks determine the distribution of invasive non-native species. Glob. Ecol. Biogeogr. 26: 907–917.

Chen, C.-L., J. Rodiger, V. Chung, R. Viswanatha, S. E. Mohr *et al.*, 2020 SNP-CRISPR: A web tool for SNP-specific genome editing. G3 Genes Genomes Genet. 10: 489–494.

Cichon, L., D. Garrido, and J. Lago, 2015 Primera deteccion de *Drosophila suzukii* (Matsumura, 1939) (Diptera: Drosophilidae) en frambuesas del Valle de Rio Negro, Argentina. Libro Resumenes IX Congr. Argent. Entomol. Posadas Misiones 270.

Comeron, J. M., R. Ratnappan, and S. Bailin, 2012 The Many Landscapes of Recombination in *Drosophila melanogaster*. PLOS Genet. 8: e1002905.

Dalton, D. T., V. M. Walton, P. W. Shearer, D. B. Walsh, J. Caprile *et al.*, 2011 Laboratory survival of *Drosophila suzukii* under simulated winter conditions of the Pacific Northwest and seasonal field trapping in five primary regions of small and stone fruit production in the United States. Pest Manag. Sci. 67: 1368–1374.

Deprá, M., J. L. Poppe, H. J. Schmitz, D. C. De Toni, and V. L. S. Valente, 2014 The first records of the invasive pest *Drosophila suzukii* in the South American continent. J. Pest Sci. 87: 379–383.

Dlugosch, K. M., and I. M. Parker, 2008 Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol. Ecol. 17: 431–449.

Drury, D. W., A. L. Dapper, D. J. Siniard, G. E. Zentner, and M. J. Wade, 2017 CRISPR/Cas9 gene drives in genetically variable and nonrandomly mating wild populations. Sci. Adv. 3: e1601910.

Durkin, S. M., M. Chakraborty, A. Abrieux, K. M. Lewald, A. Gadau *et al.*, 2021 Behavioral and genomic sensory adaptations underlying the pest activity of *Drosophila suzukii*. Mol. Biol. Evol. msab048.

Evans, E. A., and F. H. Ballen, 2014 An Overview of US Blueberry Production, Trade, and Consumption, with Special Reference to Florida: University of Flordia Institute of Food and Agricultural Sciences FE952, 8 p.

Everman, E. R., P. J. Freda, M. Brown, A. J. Schieferecke, G. J. Ragland *et al.*, 2018 Ovary development and cold tolerance of the invasive pest *Drosophila suzukii* (Matsumura) in the central plains of Kansas, United States. Environ. Entomol. 47: 1013–1023.

Fiston-Lavier, A.-S., N. D. Singh, M. Lipatov, and D. A. Petrov, 2010 *Drosophila melanogaster* recombination rate calculator. Gene 463: 18–20.

Fox, E. A., A. E. Wright, M. Fumagalli, and F. G. Vieira, 2019 ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinforma. Oxf. Engl. 35: 3855–3856.

Fraimout, A., V. Debat, S. Fellous, R. A. Hufbauer, J. Foucaud *et al.*, 2017 Deciphering the routes of invasion of *Drosophila suzukii* by means of ABC random forest. Mol. Biol. Evol. 34: 980–996.

Gaffney, M., 2017 Spotted Wing Drosophila in Ireland: An increasing threat to the Irish soft fruit sector: Teagasc.

Garnas, J. R., M.-A. Auger-Rozenberg, A. Roques, C. Bertelsmeier, M. J. Wingfield *et al.*, 2016 Complex patterns of global spread in invasive insects: eco-evolutionary and management consequences. Biol. Invasions 18: 935–952.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio.

Gautier, M., R. Vitalis, L. Flori, and A. Estoup, 2021 f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat:, 2021.05.28.445945 p.

Hauser, M., S. Gaimari, and M. Damus, 2009 *Drosophila suzukii* new to North America. Fly Times 12–15.

Jakobs, R., T. D. Gariepy, and B. J. Sinclair, 2015 Adult plasticity of cold tolerance in a continental-temperate population of *Drosophila suzukii*. J. Insect Physiol. 79: 1–9.

Jia, F., N. Lo, and S. Y. W. Ho, 2014 The Impact of Modelling Rate Heterogeneity among Sites on Phylogenetic Estimates of Intraspecific Evolutionary Rates and Timescales. PLOS ONE 9: e95722.

Johnson, R. N., and P. T. Starks, 2004 A surprising level of genetic diversity in an invasive wasp: *Polistes dominulus* in the Northeastern United States. Ann. Entomol. Soc. Am. 97: 732–737.

Kaneshiro, K. Y., 1983 Minutes, notes, and exhibitions: *Drosophila (Sophophora) suzukii* (Matsumura). 24: 179.

Kanzawa, T., 1939 Studies on *Drosophila suzukii* Mats. 49.

Koch, J. B., J. R. Dupuis, M.-K. Jardeleza, N. Ouedraogo, S. M. Geib *et al.*, 2020 Population genomic and phenotype diversity of invasive *Drosophila suzukii* in Hawai'i. Biol. Invasions.

Kolbe, J. J., R. E. Glor, L. Rodríguez Schettino, A. C. Lara, A. Larson *et al.*, 2004 Genetic variation increases during biological invasion by a Cuban lizard. Nature 431: 177–181.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15: 356.

Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura, 2018 MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35: 1547–1549.

Lee, Y., H. Schmidt, T. C. Collier, W. R. Conner, M. J. Hanemaaijer *et al.*, 2019 Genome-wide divergence among invasive populations of *Aedes aegypti* in California. BMC Genomics 20: 204.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Li, F., and M. J. Scott, 2016 CRISPR/Cas9-mediated mutagenesis of the white and Sex lethal loci in the invasive pest, *Drosophila suzukii*. Biochem. Biophys. Res. Commun. 469: 911–916.

Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur *et al.*, 2019 The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47: W636–W641.

McVean, G., 2009 A Genealogical Interpretation of Principal Components Analysis (M. Przeworski, Ed.). PLoS Genet. 5: e1000686.

Medina-Muñoz, M. C., X. Lucero, C. Severino, N. Cabrera, D. Olmedo *et al.*, 2015 *Drosophila suzukii* arrived in Chile. Drosoph. Inf. Serv. 98: 75.

Meisner, J., and A. Albrechtsen, 2018 Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210: 719–731.

Murphy, K. A., C. A. Tabuloc, K. R. Cervantes, and J. C. Chiu, 2016 Ingestion of genetically modified yeast symbiont reduces fitness of an insect pest via RNA interference. Sci. Rep. 6: 1–13.

Nei, M., T. Maruyama, and R. Chakraborty, 1975 The bottleneck effect and genetic variability on populations. Evolution 29: 1–10.

Noncitrus Fruits and Nuts 2019 Summary, 2020 USDA Natl. Agric. Stat. Serv. 100.

Olazcuaga, L., A. Loiseau, H. Parrinello, M. Paris, A. Fraimout *et al.*, 2020 A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure (N. Singh, Ed.). Mol. Biol. Evol. 37: 2369–2385.

Paris, M., R. Boyer, R. Jaenichen, J. Wolf, M. Karageorgi *et al.*, 2020 Near-chromosome level genome assembly of the fruit pest *Drosophila suzukii* using long-read sequencing. Sci. Rep. 10: 11227.

Peng, F. T., 1937 On some species of *Drosophila* from China. Annot. Zool. Jpn. 16: 20–27.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genet. 8: e1002967.

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

Rozas, J., A. Ferrer-Mata, J. C. Sánchez-DelBarrio, S. Guirao-Rico, P. Librado *et al.*, 2017 DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol. Biol. Evol. 34: 3299–3302.

dos Santos, L. A., M. F. Mendes, A. P. Krüger, M. L. Blauth, M. S. Gottschalk *et al.*, 2017 Global potential distribution of *Drosophila suzukii* (Diptera, Drosophilidae) (C. Wicker-Thomas, Ed.). PLOS ONE 12: e0174318.

Schmidt, H., T. C. Collier, M. J. Hanemaaijer, P. D. Houston, Y. Lee *et al.*, 2020 Abundance of conserved CRISPR-Cas9 target sites within the highly polymorphic genomes of Anopheles and Aedes mosquitoes. Nat. Commun. 11: 1425.

Schmidt, P. S., L. Matzkin, M. Ippolito, and W. F. Eanes, 2005 Geographic variation in diapause incidence, life-history traits, and climatic adaptation in *Drosophila melanogaster*. Evolution 59: 1721–1732.

Schöneberg, T., M. T. Lewis, H. J. Burrack, M. Grieshop, R. Isaacs *et al.*, 2021 Cultural Control of Drosophila suzukii in Small Fruit—Current and Pending Tactics in the U.S. Insects 12: 172.

Shearer, P. W., J. D. West, V. M. Walton, P. H. Brown, N. Svetec *et al.*, 2016 Seasonal cues induce phenotypic plasticity of *Drosophila suzukii* to enhance winter survival. BMC Ecol. 16: 11.

Skotte, L., T. S. Korneliussen, and A. Albrechtsen, 2013 Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics 195: 693–702.

Soria-Carrasco, V., Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman *et al.*, 2014 Stick insect genomes reveal natural selection's role in parallel speciation. Science 344: 738–742.

Steck, G. J., W. Dixon, and D. Dean, 2009 Spotted Wing Drosophila, *Drosophila suzukii* (Matsurmura) (Diptera: Drosophilidae), a fruit pest new to North America: Florida Dept of Agriculture and Consumer Services, 3 p.

Stephens, A. R., M. K. Asplen, W. D. Hutchison, and R. C. Venette, 2015 Cold hardiness of winter-acclimated *Drosophila suzukii* (Diptera: Drosophilidae) adults. Environ. Entomol. 44: 1619–1626.

Stockton, D. G., A. K. Wallingford, G. Brind'amore, L. Diepenbrock, H. Burrack *et al.*, 2020 Seasonal polyphenism of spotted-wing drosophila is affected by variation in local abiotic conditions within its invaded range, likely influencing survival and regional population dynamics. Ecol. Evol. 10: 7669–7685.

Stockton, D., A. Wallingford, D. Rendon, P. Fanning, C. K. Green *et al.*, 2019 Interactions between biotic and abiotic factors affect survival in overwintering *Drosophila suzukii* (Diptera: Drosophilidae). Environ. Entomol. 48: 454–464.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Takamori, H., H. Watabe, Y. Fuyama, Y. Zhang, and T. Aotsuka, 2006 *Drosophila subpulchrella*, a new species of the *Drosophila suzukii* species subgroup from Japan and China (Diptera: Drosophilidae). Entomol. Sci. 9: 121–128.

Taning, C. N. T., O. Christiaens, N. Berkvens, H. Casteels, M. Maes *et al.*, 2016 Oral RNAi to control *Drosophila suzukii*: laboratory testing against larval and adult stages. J. Pest Sci. 89: 803–814.

Toyama, K. S., P.-A. Crochet, and R. Leblois, 2020 Sampling schemes and drift can bias admixture proportions inferred by structure. Mol. Ecol. Resour. 20: 1769–1785.

Trask, J. A. S., R. S. Malhi, S. Kanthaswamy, J. Johnson, W. T. Garnica *et al.*, 2011 The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*). Primates 52: 129–138.

Tyukmaeva, V. I., T. S. Salminen, M. Kankare, K. E. Knott, and A. Hoikkala, 2011 Adaptation to a seasonally varying environment: a strong latitudinal cline in reproductive diapause combined with high gene flow in *Drosophila montana*. Ecol. Evol. 1: 160–168.

Walsh, D. B., M. P. Bolda, R. E. Goodhue, A. J. Dreves, J. Lee *et al.*, 2011 *Drosophila suzukii* (Diptera: Drosophilidae): Invasive pest of ripening soft fruit expanding its geographic range and damage potential. J. Integr. Pest Manag. 2: G1–G7.

Walton, V. M., H. J. Burrack, D. T. Dalton, R. Isaacs, N. Wiman *et al.*, 2016 Past, present and future of *Drosophila suzukii:* distribution, impact and management in United States berry fruits. Acta Hortic. 87–94.

Willing, E.-M., C. Dreyer, and C. van Oosterhout, 2012 Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. PLOS ONE 7: e42649.

Wu, N., S. Zhang, X. Li, Y. Cao, X. Liu *et al.*, 2019 Fall webworm genomes yield insights into rapid adaptation of invasive species. Nat. Ecol. Evol. 3: 105–115.

Zecca, G., M. Labra, and F. Grassi, 2019 Untangling the Evolution of American Wild Grapes: Admixed Species and How to Find Them. Front. Plant Sci. 10: 1814.

Supplemental Figures

Figure S1: Linkage disequilibrium decay in longest contig for all samples. Linkage disequilibrium measured by r^2 measured pairwise for each SNP within 100kb of each other (0-30kb shown in plot). 95% confidence shaded intervals displayed based on 100 bootstrap replicates. A 1% subsample of the r^2 values was used to fit the model of decay.

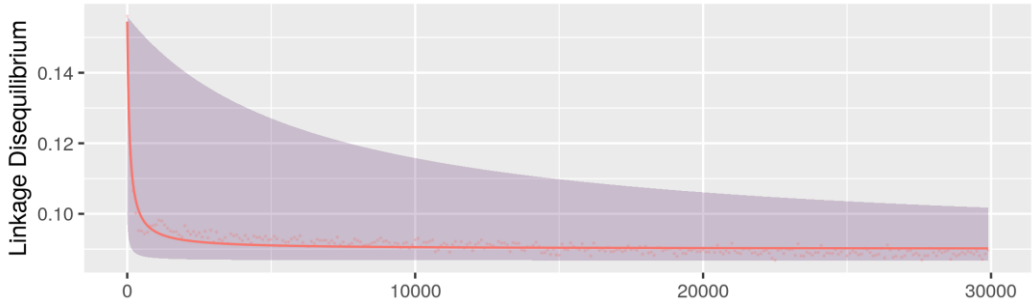


Figure S2: Admixture proportions estimated for each region individually. Samples are labeled by location code, followed by U.S. state abbreviation or country. Brazil not plotted as only one location was sampled. **(A)** Eastern U.S. samples, 182,786 sites used. **(B)** Western U.S. samples (including Hawaii), 136,929 sites used. **(C)** Asian samples, 97,134 sites used. **(D)** European samples, 77,645 sites used.

Figure S3: PCA calculated for each region individually. Percent variance captured by each principal component indicated in axis labels. **(A)** Eastern U.S. samples, 183,243 sites used. **(B)** Western U.S. samples, (including Hawaii), 139,075 sites used. **(C)** European samples, 90,624 sites used. **(D)** Asian samples, 103,312 sites used.

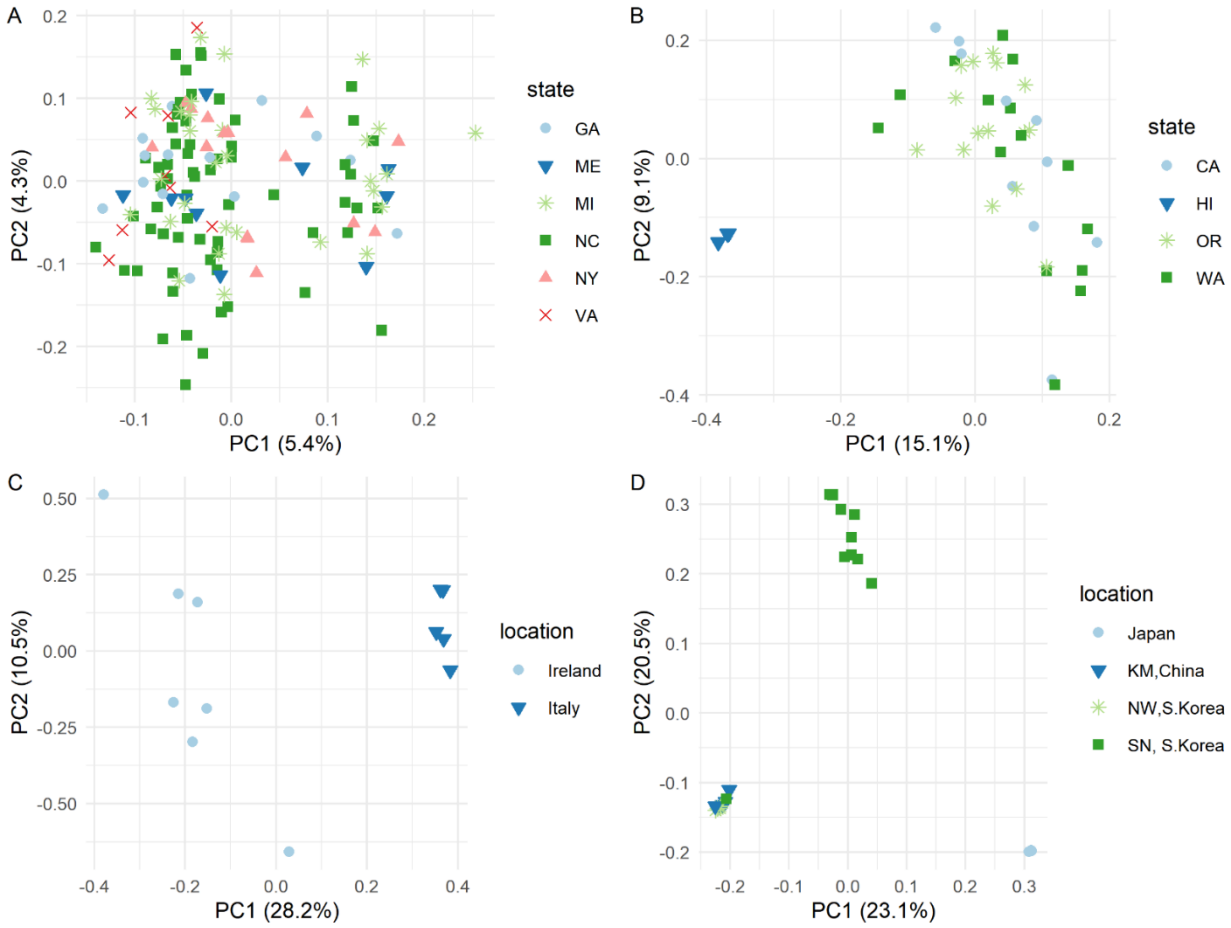


Figure S4: Phylogenetic tree using COX2 gene haplotypes. Maximum likelihood tree of COX2 rooted on *D. melanogaster* (Dmel01) using the Tamura 3-parameter model + G, with bootstrap fractions greater than 0.5 from 500 replicate runs displayed next to branch points. Branch lengths measure number of substitutions per site. 70 variable sites were analyzed from a total of 720 positions in the alignment. The following abbreviations were used for species name. Dsuz = *D. suzukii*; Dpul = *D. pulchrella*; Dsub = *D. subpulchrella*; Dbia = *D. biarmipes*; Dlut = *D. lutescens*; Dmim = *D. mimetica*; Dmel = *D. melanogaster*. Note that haplotypes Dsuz01 (Genbank HQ631606.1), Dsuz03 (pop NW), Dsuz04 (pop KM), and Dsuz05(pop NW) are likely not actually *D. suzukii*.

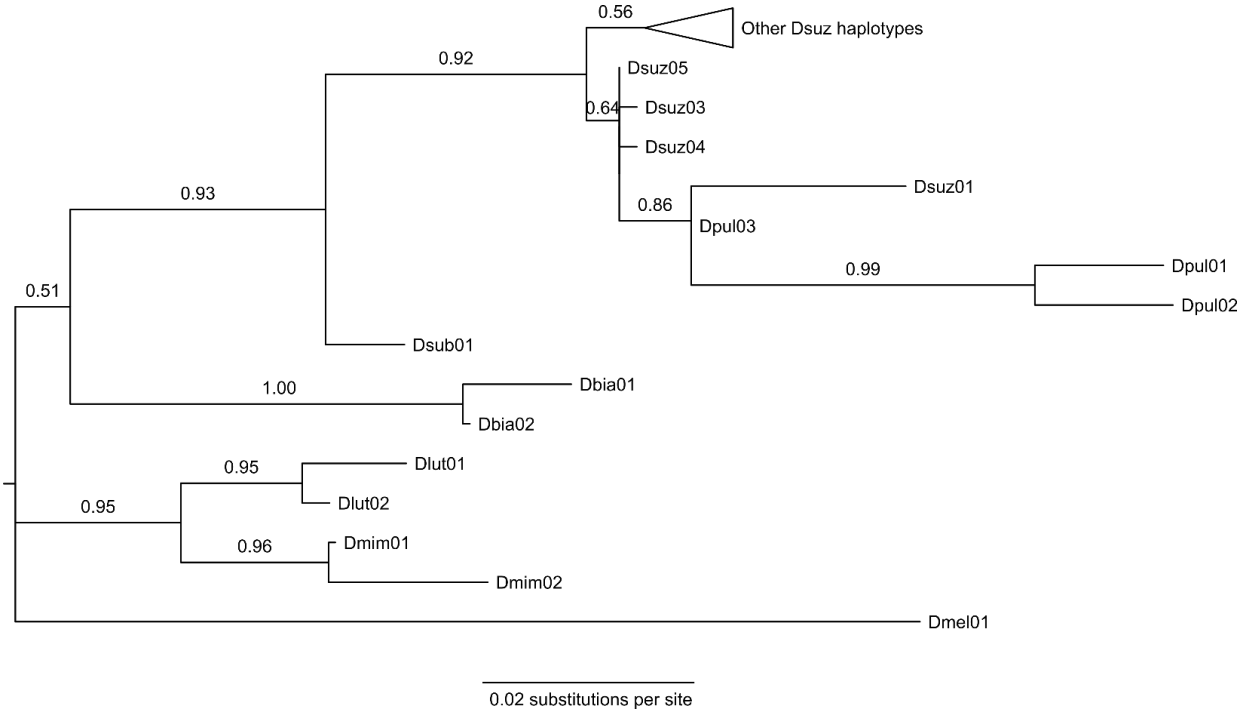


Figure S5: Admixture proportions estimated from subsampled clusters. **(A-C)** 5 random individuals were sampled from each population cluster for each analysis. Samples labeled by name, followed by population cluster. For the fourth subsample, see Figure 2C. 144,661-152,424 sites used.

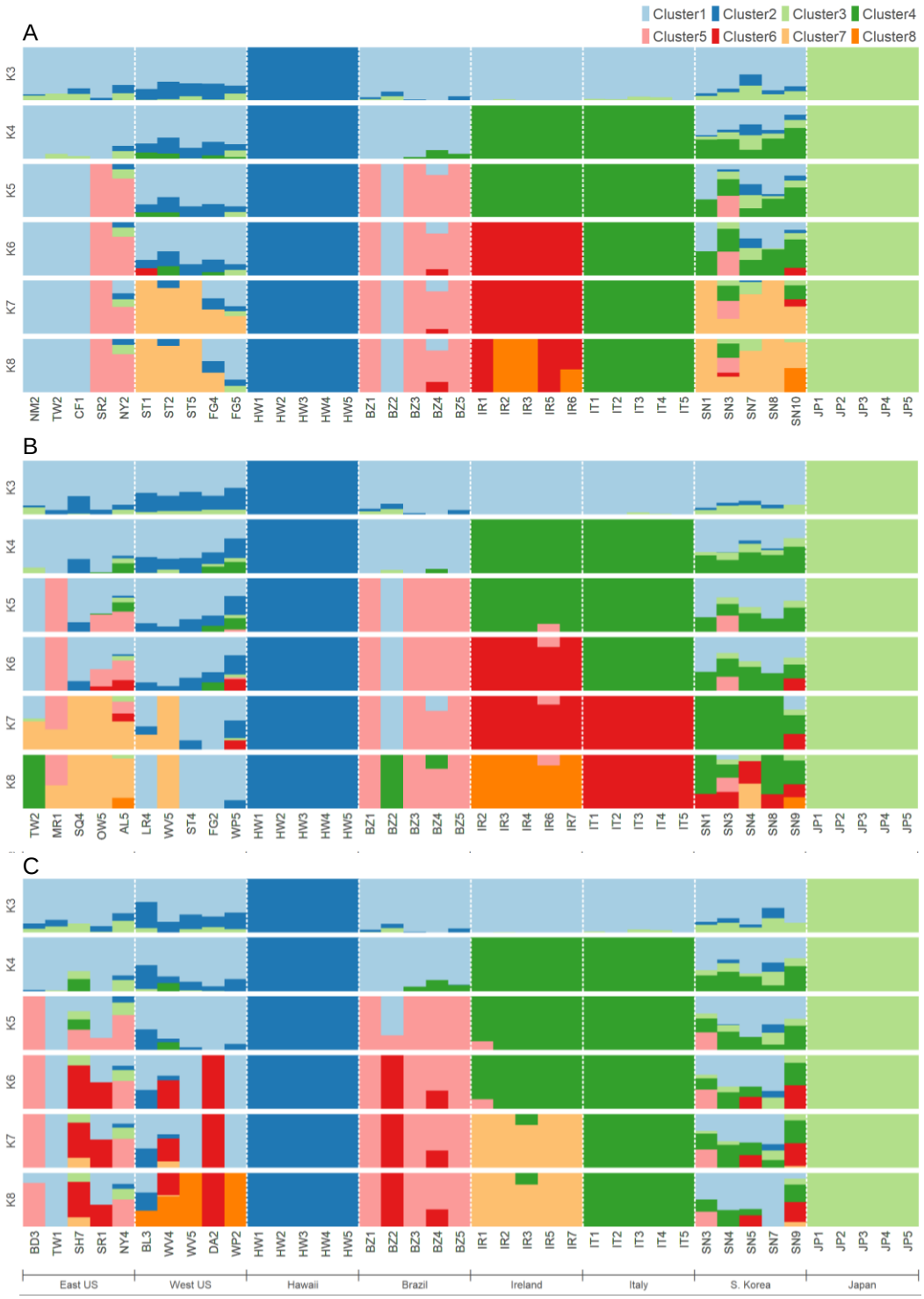

A
Cluster1 Cluster2 Cluster3 Cluster4
Cluster5 Cluster6 Cluster7 Cluster8
K3
K4
K5
K6
K7
K8
NM2 TW2 CF1 SR2 NY2 ST1 ST2 ST5 FG4 FG5 HW1 HW2 HW3 HW4 HW5 BZ1 BZ2 BZ3 BZ4 BZ5 IR1 IR2 IR3 IR5 IR6 IT1 IT2 IT3 IT4 IT5 SN1 SN3 SN7 SN8 SN10 JP1 JP2 JP3 JP4 JP5
B
TW2 MR1 SQ4 OW5 AL5 LR4 WV5 ST4 FG2 WP5 HW1 HW2 HW3 HW4 HW5 BZ1 BZ2 BZ3 BZ4 BZ5 IR2 IR3 IR4 IR6 IR7 IT1 IT2 IT3 IT4 IT5 SN1 SN3 SN4 SN8 SN9 JP1 JP2 JP3 JP4 JP5
C
BD3 TW1 SH7 SR1 NY4 BL3 WV4 WV5 DA2 WP2 HW1 HW2 HW3 HW4 HW5 BZ1 BZ2 BZ3 BZ4 BZ5 IR1 IR2 IR3 IR5 IR7 IT1 IT2 IT3 IT4 IT5 SN3 SN4 SN5 SN7 SN9 JP1 JP2 JP3 JP4 JP5
East US
West US
Hawaii
Brazil
Ireland
Italy
S. Korea
Japan

Figure S6:Admixture proportions estimated from all samples combined. Up to 10 clusters (k) were used.

Samples are labeled by state if applicable, followed by population cluster. 206,093 sites used.

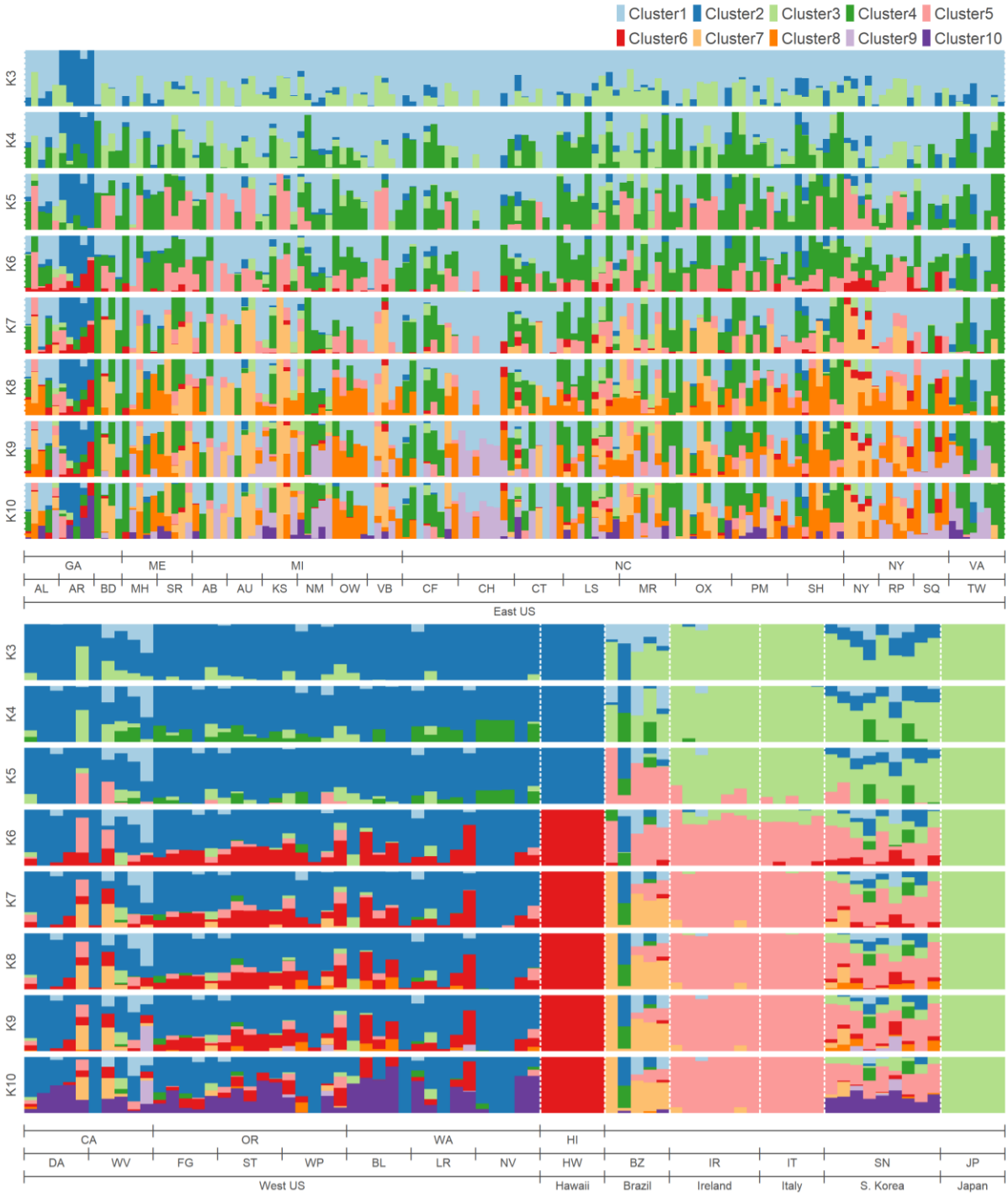


Figure S7: Trees inferred by treemix with 0 to 10 migrations. Migration edges have not been plotted for readability. Bootstrap replicate values label each branch from 100 bootstrapped runs. X-axis measures genetic drift. Fraction variance of the data captured by the model is indicated in bottom right of each plot. Labels “D.sub” and “D.bia” stand for D.subpulchrella and D. biarmipes, respectively.

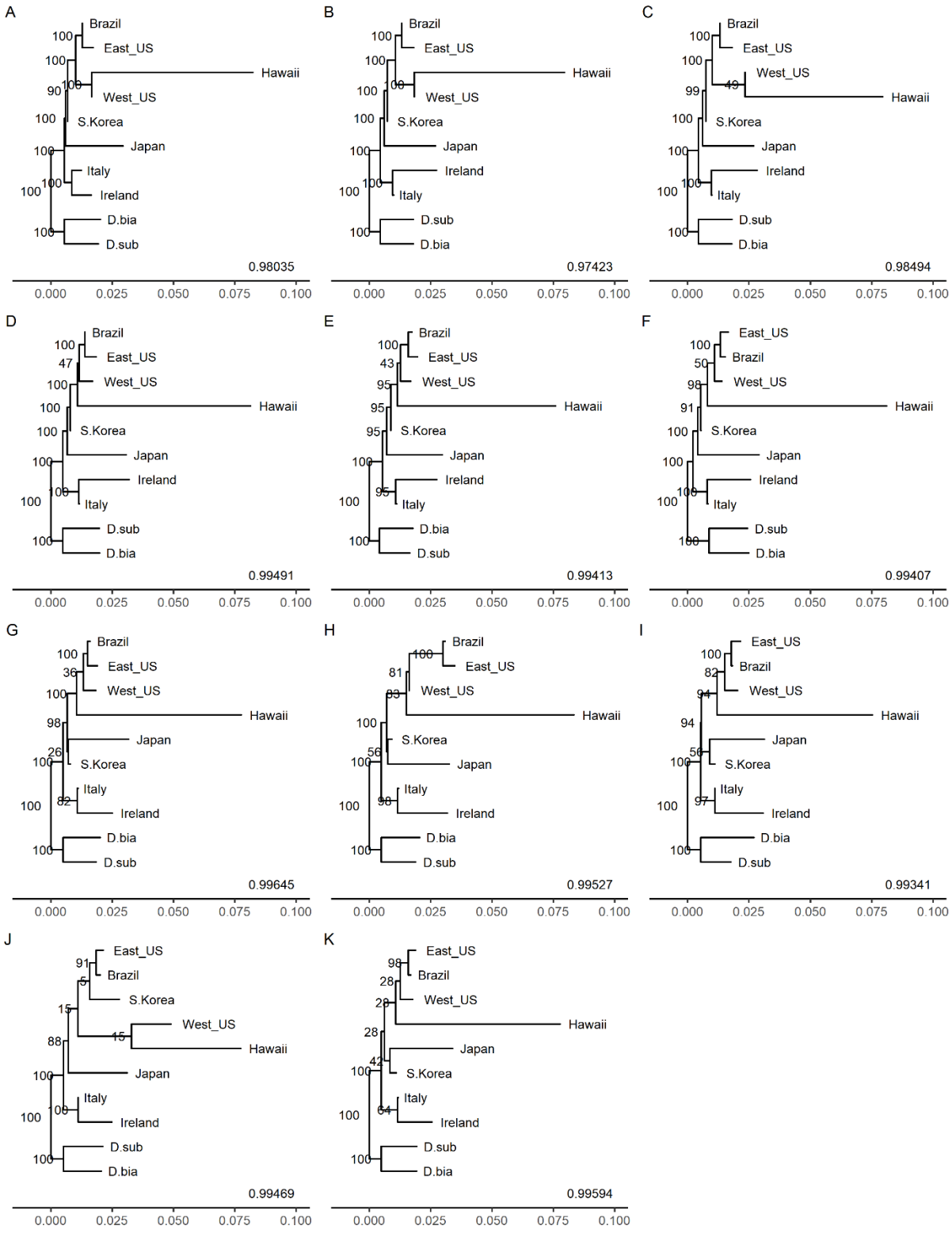
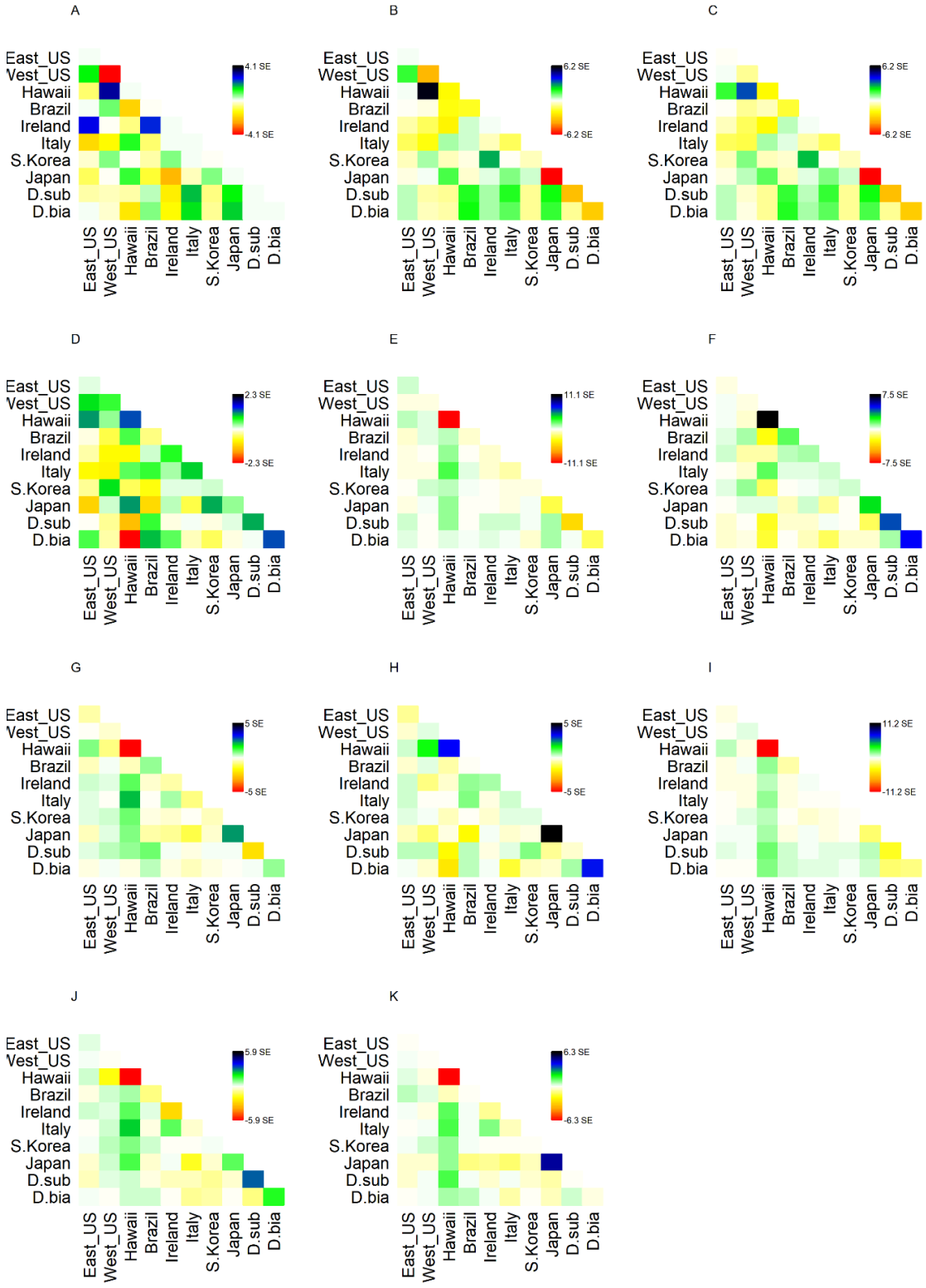

A
Brazil
East_US
Hawaii
West_US
S.Korea
Japan
Italy
Ireland
D.bia
D.sub
0.98035
B
Brazil
East_US
Hawaii
West_US
S.Korea
Japan
Ireland
Italy
D.sub
D.bia
0.97423
C
Brazil
East_US
West_US
Hawaii
S.Korea
Japan
Ireland
Italy
D.sub
D.bia
0.98494
D
Brazil
East_US
West_US
Hawaii
S.Korea
Japan
Ireland
Italy
D.sub
D.bia
0.99491
E
Brazil
East_US
West_US
Hawaii
S.Korea
Japan
Ireland
Italy
D.bia
D.sub
0.99413
F
East_US
Brazil
West_US
Hawaii
S.Korea
Japan
Ireland
Italy
D.sub
D.bia
0.99407
G
Brazil
East_US
West_US
Hawaii
Japan
S.Korea
Italy
Ireland
D.bia
D.sub
0.99645
H
Brazil
East_US
West_US
Hawaii
S.Korea
Japan
Italy
Ireland
D.bia
D.sub
0.99527
I
East_US
Brazil
West_US
Hawaii
Japan
S.Korea
Italy
Ireland
D.bia
D.sub
0.99341
J
East_US
Brazil
S.Korea
West_US
Hawaii
Japan
Italy
Ireland
D.sub
D.bia
0.99469
K
East_US
Brazil
West_US
Hawaii
Japan
S.Korea
Italy
Ireland
D.sub
D.bia
0.99594
0.000
0.025
0.050
0.075
0.100

Figure S8: Plots of standard error of residuals between populations based on models generated by treemix from 0 to 10 migration edges. High residuals indicate the model underestimates the data's co-variance, which could be a sign more migrations are needed. Low residuals may indicate the two populations are too close in the graph due to unmodeled migration elsewhere.


A
East_US
West_US
Hawaii
Brazil
Ireland
Italy
S.Korea
Japan
D.sub
D.bia
4.1 SE
-4.1 SE
B
6.2 SE
-6.2 SE
C
6.2 SE
-6.2 SE
D
2.3 SE
-2.3 SE
E
11.1 SE
-11.1 SE
F
7.5 SE
-7.5 SE
G
5 SE
-5 SE
H
5 SE
-5 SE
I
11.2 SE
-11.2 SE
J
5.9 SE
-5.9 SE
K
6.3 SE
-6.3 SE

Chapter 2: *Tuta absoluta* genome assembly and population analysis of Latin America

Kyle M. Lewald[1], Christine A. Tabuloc[1], Kristine E. Godfrey[2], Judit Arnó[3], Clérison R. Perini[4], Jerson C. Guedes[4], Joanna C. Chiu[1]

[1]Department of Entomology and Nematology, University of California, Davis.

[2]Contained Research Facility, University of California, Davis.

[3]IRTA, Cabrils, Spain.

[4]Department of Phytosanitary Defense, Federal University of Santa Maria, Brazil.

Contributions and Acknowledgements

K.M.L. and J.C.C. conceived the study. K.M.L. and C.A.T. performed nucleic acid extraction and library preparation for population samples. K.E.G., J.A., C.R.P., and J.C.G. provided field and colony collections. K.M.L. performed genome assembly and population analyses. K.M.L. and J.C.C. wrote the manuscript with input from all authors.

Thank you to Robert Munch (QB3, UC Berkeley) and John Alterio (PacBio, Menlo Park) for their advice and work on DNA extraction, library preparation, and long-read genome sequencing for the genome assembly. Thank you to Dr. Graham Coop and Dr. Jeffrey Ross-Ibarra for providing advice on population analyses.

Abstract

Tuta absoluta is one of the largest threats to tomato agriculture worldwide. Native to South America, it has rapidly spread throughout Europe, Africa, and Asia over the past two decades. To understand how *Tuta absoluta* has been so successful and to improve containment strategies, high quality genomic resources and an understanding of population history is critical. Here, we describe a highly contiguous annotated genome assembly, as well as a genome-wide population analysis of samples collected across Latin America. The new genome assembly has an L50 of 17 with only 132 contigs. Based on hundreds of thousands of SNPs, we detect three major population clusters in Latin America with some evidence of admixture along the Andes Mountain range. Based on coalescent simulations, we find these clusters diverged from each other tens of thousands of generations ago prior to domestication of tomatoes. We further identify several genomic loci under selection related to insecticide resistance, immunity, and metabolism. This data will further future research toward genetic control strategies and inform future containment policies.

Introduction

Tuta absoluta (also known as *Phthorimaea absoluta* (Chang and Metz 2021)) is a worldwide economic pest of tomatoes and other solanaceous crops. A member of the gelechiid family, this moth lays eggs on the above-ground portion of the plant, where the hatched larvae will spend their lives creating "mines" throughout the plant tissue before pupating and emerging as adults (Godfrey *et al.* 2018). At a reproduction rate of up to 10 generations per year, untreated infestations will eventually result in complete death of the plant, leading to up to 100% agricultural loss. Although a large effort has been made to develop and implement integrated pest management (IPM) programs across different world regions (Desneux *et al.* 2022), typical treatments have included heavy use of a variety of insecticides (Siqueira *et al.* 2000), leading to the rapid appearance of insecticide resistance. As tomatoes represent a massive economic industry, with an estimated 252 million metric tons of tomatoes harvested in 2020 (FAOSTAT 2020), there is a serious need to understand the invasive biology of this insect and to develop tools for detection and prevention.

Tuta absoluta was originally detected in Peru in 1917 (Meyrick 1917) but was not recorded as an agricultural pest until the 1960s and 70s when it was discovered in tomato fields in Chile, Argentina, and Venezuela; by the 1990s it was widespread across South America. In 2006, *Tuta absoluta* appeared in Spain (EPPO 2008); since then it has rapidly colonized Europe, Asia, and Africa. It is generally believed that the Peruvian highlands is the ancestral home of *T. absoluta,* and that the rapid colonization to the rest of Latin America was due to the introduction of *Tuta absoluta* by human transport of contaminated fruit, although few studies have confirmed this (Desneux *et al.* 2011). Previous research using mitochondrial and microsatellite DNA markers found some evidence of population structure, as well as evidence that the European invasion originated from a single population in central Chile (Cifuentes *et al.* 2011; Guillemaud *et al.* 2015). However, determination of higher-resolution population structure, migration events, divergence times, and population size can benefit from using a larger number of

markers, such as what is produced from genome-wide sequencing studies (Trask *et al.* 2011; Willing *et al.* 2012; Rašić *et al.* 2014; Koch *et al.* 2020). Additionally, few genetic studies have been conducted to understand how *Tuta absoluta* has performed so successfully as an agricultural pest beyond targeted examinations of known insecticide resistance alleles. One reason for this has been the lack of a highly contiguous genome with annotated genes. A short-read based assembly has been previously published for the purpose of developing molecular diagnostics (Tabuloc *et al.* 2019); however, it is highly fragmented and duplicated.

In this study, we addressed these issues by using long-read sequencing technology to produce a highly contiguous genome assembly for *Tuta absoluta*. We then use short-read technology to sequence genomes of individuals collected across Latin America, as well as a Spanish population, to identify single nucleotide polymorphisms (SNPs) in an unbiased manner. We use these SNPs to detect population structure and estimate population history parameters to understand how and when *Tuta absoluta* spread across Latin America. Finally, we use genome scanning statistics to identify genes putatively under selection that may explain *Tuta absoluta's* success as an agricultural pest. We expect the genome assembly and population data will be an asset toward developing new strategies to manage this pest.

Methods

High Molecular Weight DNA extraction

For genome assembly, a single *Tuta absoluta* larva was collected from a colony originally sourced from the Institute of Agrifood Research and Technology (IRTA), Cabrils, Spain and held in the Contained Research Facility in UC Davis and frozen on dry ice. The larva was pulverized in liquid nitrogen with a pestle in a 2 mL microcentrifuge tube using 740 mL of lysis buffer (80 mM EDTA pH 8, 324 mM NaCl, 0.68% SDS, 8 mM Tris-HCl pH 8, 80ug/mL RNase A (NEB, Ipswich, MA), 135 ug/mL Prot K(NEB)). After a 37℃ overnight incubation step, 240 µL of 5M NaCl was added and gently mixed in by rocking before centrifuging at 10,000 RCF, 4℃, for 15 minutes. Supernatant was transferred using a wide-bore pipette

to a 2 mL DNA low-bind tube (Eppendorf, Enfield CT), precipitated with 1 mL of 100% ethanol, and centrifuged at 10,000 RCF, 4℃, for 5 minutes. The DNA pellet was washed with 500 µL of ice-cold 70% ethanol twice before air-drying for 5 minutes. Dry pellet was resuspended in DEPC-treated water and allowed to dissolve for 1 hour at room temperature before being stored at 4℃ for no more than 2 weeks. Absorbance ratios were measured with a Nanodrop Lite (ThermoFisher Scientific, Waltham, MA), DNA concentration was measured with a Qubit 4 Fluorometer using a dsDNA High-Sensitivity Assay (ThermoFisher Scientific), and DNA fragment size was measured with a Tapestation genomic DNA ScreenTape (Agilent, Santa Clara, CA). Approximately 700 ng of DNA was sent to QB3-Berkeley for library preparation and PacBio HiFi sequencing with 1 SMRTcell.

Genome Assembly and Assessment

Raw subreads were collapsed into Circular Consensus Sequence (CCS) reads using ccs version 6.0.0 (PacBio, Menlo Park, CA). K-mer histograms were made with jellyfish version 2.2.6 (Marçais and Kingsford 2011) using 31-mers, then visualized with GenomeScope version 2.0 (Ranallo-Benavidez *et al.* 2020). GC content vs k-mer frequency was calculated from the jellyfish histograms using kat version 2.4.2 (Mapleson *et al.* 2016) and visualized with R. CCS reads were initially assembled using hifiasm version 0.14 or HiCanu version 2.11 (Nurk *et al.* 2020; Cheng *et al.* 2021) with default parameters. HiCanu assembly was separated into a primary and alternate haplotig set using purge-dups version 1.2.5 4 (Guan). The primary hifiasm assembly was purged with purge_dups; alternate haplotigs from this purge were purged with the alternate hifiasm assembly, and re-purged with purge_dups to discard repeats, high-coverage, or nested haplotigs. Merqury version 1.1 (Rhie *et al.* 2020) was used to assess genome assembly quality and completeness between the two assemblers and between pre- and post-purging. K-mer size of 20 was used for building the Meryl database from the raw CCS reads. Copy number k-mer plots were generated using Merqury's provided R scripts. BUSCO version 5.1.2 (Manni *et al.* 2021) was also used in genome mode to assess genome ortholog completeness using the

Lepidoptera OrthoDB-10 database (Kriventseva *et al.* 2019). To detect contigs that were contaminant DNA and not of *Tuta absoluta* origin, blastn version 2.12.0 (Camacho *et al.* 2009) was used with the "nt" database (downloaded August 3rd, 2021) under the following parameters: word size=20, max target sequences = 10. Taxonomic information was downloaded for each subject match from the NCBI Taxonomy database using the "rentrez" package in R. To corroborate blast results, Phyloligo version 1.0 (Mallet *et al.* 2017) was used to generate a Euclidian distance matrix between contigs based on k-mer distribution with k-mer length 4. The R packages "ape" version 5.6-1 (Paradis and Schliep 2019) and "ggtree" version 2.2.4 (Yu *et al.* 2017) were used to generate and visualize the contig tree using the BIONJ algorithm.

Repeat masking

The decontaminated hifiasm primary genome assembly was supplied as a database to RepeatModeler version 2.0.2a (Flynn *et al.* 2020) to produce a custom repeat library, with the long terminal repeat module enabled. RepeatMasker version 4.1.2 (Smit *et al.* 2021) was used to produce GFF annotation files of repeat coordinates. The Dfam transposable element database provided with RepeatMasker was merged with the RepBase RepeatMasker Edition database version 20181026 (Bao *et al.* 2015) to mask the genome once, and the custom RepeatModeler library was used to mask the genome separately. The resulting GFF files were merged and sorted with bedtools version 2.30 sort (Quinlan and Hall 2010), then used to soft-mask the assembly with bedtools maskfasta.

Gene model annotation

Six RNAseq datasets produced by Camargo *et al.* 2015 covering all life stages of *Tuta absoluta* (egg, four larval instars, adult) were downloaded from Bioproject PRJNA291932 for gene model annotation. Reads were checked for quality with FastQC version 0.11.9 ("FastQC" 2019), trimmed with Trimmomatic version 0.39 (Bolger *et al.* 2014), and aligned to our primary assembly using STAR version 2.7.9a (Dobin *et al.* 2013) with default parameters. In addition, protein databases were downloaded from Lepbase

(Challi *et al.* 2016) and OrthoDB-arthropoda (Kriventseva *et al.* 2019). The soft-masked genome was annotated twice with BRAKER2 version 2.1.5 (Brůna *et al.* 2021), once with the aligned RNA data, and again with the merged protein data. The two resulting GTF gene model files, as well as the GFF gene model hints files, were supplied to TSEBRA version 1.0.2 (Gabriel *et al.* 2021) to be merged into a single GTF output. Gffread version 0.12.6 (Pertea and Pertea 2020) was used to remove mRNAs with missing start or stop codons, in-frame stop codons, or that were redundant.

Functional gene annotation

Entap version 0.10.7 (Hart *et al.* 2020) was used to annotate gene models with names and predicted functions. Entap was configured to perform frame selection and filtering. The Lepbase, Refseq-Invertebrate, UniprotKB/Swiss-Prot, and UniprotKB/TrEMBL (O'Leary *et al.* 2016; Challi *et al.* 2016; The UniProt Consortium 2021) protein databases were used for gene identity search and the EggNOG database (Huerta-Cepas *et al.* 2019) for gene ontology, protein domain, and pathway annotation. As EnTAP was designed for transcriptome annotation, gene model coordinates were not referenced to the genome assembly. To correct this, the output GFF gene model annotation was converted to a GTF using gffreads, then converted to an alignment-GFF format using the script "gtf_to_alignment_gff3.pl" from Transdecoder (Brian and Papanicolaou), and finally mapped back to the genome assembly coordinates using the Transdecoder script "cdna_alignment_orf_to_genome_orf.pl".

Population sample DNA extraction and alignment

We used the same DNA extracted from *Tuta absoluta* collected from South America, Costa Rica, and Spain by Tabuloc *et al.* 2019. DNA libraries were made using the KAPA Hyperplus Kit (Roche, Basel, Switzerland). 150 basepair paired-end sequencing was performed by Novogene on the Illumina HiSeq 4000. Raw reads were trimmed of adapter sequences using scythe version 0.991 (Buffalo 2014) and were quality-filtered using sickle version 1.33 (Joshi and Fass 2011) using default settings. FastQC was used to inspect read quality before and after filtering. Reads were mapped to the genome assembly

using bwa mem version 0.7.17 (Li 2013), and duplicates were marked with samtools markdup version 1.14 (Danecek *et al.* 2021).

Population structure analysis

Angsd version 0.935 (Korneliussen *et al.* 2014) was used to estimate genotype likelihoods from the 78 contigs longer than 100kb, using a single nucleotide polymorphism (SNP) filter threshold of $p < 10^{-6}$, a minimum minor allele frequency of 0.05, and a minimum map and base quality of Q=20. SNPs were then pruned to every 500 base pairs. PCA and admixture analysis were performed using PCAngsd version 1.0 (Meisner and Albrechtsen 2018) and NGSadmix version 32 (Skotte *et al.* 2013) as described in Lewald *et al.* 2021. PCAngsd was also used to output inbreeding coefficients for each sample.

Treemix

We called genotypes from the genotype likelihoods calculated for population structure analysis using PCAngsd, with a 95% confidence threshold and with inbreeding values estimated from PCAngsd as priors. We removed loci that were missing data in more than 20% of samples, or loci with missing data in all individuals within a single sampling location. A custom R script was used to convert PCAngsd-format genotypes into a Treemix-formatted allele counts table. Treemix version 1.13 (Pickrell and Pritchard 2012) was run with 100 bootstraps, a window block size of 500 SNPs, 0 to 5 migration edges, and rooted on the CR (Costa Rica) population. The “global rearrangements” and “standard error calculation” options were also enabled. Treemix’s “threepop” subprogram was used to calculate F3 statistics between populations, using a 500 SNP window block size for standard error estimation.

Population summary statistics and Population Branch Statistic

Summary statistics were calculated with Angsd on the largest 78 contigs. The site allele frequencies (SAF) were calculated for each region (North, Andes, Central, and Spain) with the -doMaf 1 option and using individual inbreeding coefficients as priors, and no minor allele frequency or SNP filtering was used. To estimate nucleotide diversity and Tajima’s D, the global folded 1-dimensional site frequency

spectrum (1D-SFS) was calculated using Angsd realSFS for each population using the SAF in 100Mb pieces of the genome with a maximum of 400 iterations in the EM cycle. The 1D-SFS was summed across the genome for each population and realSFS saf2theta was used to estimate thetas per site. Angsd thetaStat do_stat was then used to calculate theta and Tajima's D in 20kb windows, with a step of either 20kb or 5kb.

To estimate Fst between regions, realSFS was used to calculate the global folded 2-dimensional site frequency spectrum (2D-SFS) between each pair of regions in 100Mb pieces of the genome with 400 EM cycle iterations. 2D-SFS was summed across the genome, and realSFS Fst index was then used to estimate per-site Fst, and realSFS fst stats was used to estimate the global Fst values between regions. To estimate the Population Branch Statistic (PBS), the Andes, Central, and North regions were supplied at once to realSFS Fst index and realSFS fst stats2, to produce PBS values for each region in sliding 5kb windows with a 500bp step across the genome. Windows with less than 4kb of sequence data were excluded from the analysis. To compare allele frequencies of key SNPs between populations, we repeated site allele frequency estimation, but forced the reference allele to be the "major" allele.

Population Modeling

The 2D-SFS was estimated between North, Andes, and Central regions using the same procedure as for summary statistics but excluded all genic regions (gene model boundaries plus an additional 1kb on flanking sides). Additionally, the Villa Alegre, Chile (VA) population samples were excluded from the analysis. Output 2D-SFS was converted to the Fastsimcoal version 2708 (Excoffier *et al.* 2013) format using a custom R script.

Fastsimcoal was used to estimate model parameters from the 2D-SFS data under several possible models. In the simplest model, an ancestral population is allowed to split two times, each with its own population size (5 total parameters). The exponential growth model adds a growth event to each population with unique exponential growth rates (9 total parameters). Finally, the resized population

model replaces the exponential growth with an instantaneous change in population size (11 total parameters). For each model, parameter estimation was run 100 independent times, using 1 million simulations and 100 EM loops per run. SFS categories with less than 10 counts were excluded.

To compare models to each other and the data, 1Mb of DNA was simulated 100 times under each model using Fastsimcoal with a mutation rate of $2.9x10^{-9}$ basepairs/generation (Keightley *et al.* 2015) and a recombination rate of $2.97x10^{-8}$ cM/Mb (Yamamoto *et al.* 2008), based on estimates from *Heliconius melpomene* and *Bombyx mori*, respectively. The resulting genotypes were used to estimate the r^2 - measure of linkage disequilibrium (LD) between SNPs less than 10kb away using ngsLD version 1.1.1 (Fox *et al.* 2019). To estimate LD from the sequencing data, the same minor allele frequency files used for Fastsimcoal parameter estimation were used to call genotype likelihoods in each region on the largest 29 contigs. NgsLD was run on a 10% subset of these genotype likelihoods to a max distance of 10kb. Estimates of LD decay rates, maximum, and minimum LD were calculated from a 1% subset of r^2 values from simulations and a 10% subset of r^2 values from data using the provided fit_LDdecay.R script with the following parameters: fit_bin_size = 100, recombination rate = 2.97, fit_boot = 100, fit_level = 10.

Results

New Pacbio *Tuta absoluta* genome assembly improves gene annotation and contiguity

Before performing any population analyses, we decided to produce a high-quality reference genome based on long-read technology. New protocols for PacBio HiFi sequencing allow for low DNA input, which was critical in our case as Lepidopterans are notoriously heterozygous and using DNA pooled from many individuals would make genome assembly challenging. We sequenced a single moth originating from a laboratory colony at the Institute of Agrifood Research and Technology (IRTA), Cabrils, Spain, and obtained 16.2Gbp of sequence after collapsing circular consensus reads. Based on k-mer analysis with GenomeScope, the genome is 2.9% heterozygous and has 38.8% repeat content. GC vs k-mer plots

show that there is likely no mass contamination from other species or microbes (Figure S9). The k-mer based haploid length estimate is 524 Mbp, which is close to the 564 Mbp estimate based on flow cytometry (Paladino *et al.* 2016).

To assemble reads, we used the HiFi assembler hifiasm and compared quality and contiguity metrics using Merqury. The accurate long reads allow for the ability to separately assemble maternal and paternal haplotypes at heterozygous regions. While hifiasm attempts to separate assembled contigs into primary and alternate haplotypes, we found that the primary assembly still had high haplotype retention based on its 990 Mbp length and the large peak of raw read k-mers that appear twice in the assembly (Figure S2A). We decided to further remove haplotigs using purge_dups, which shrunk the primary assembly size to 650.6 Mbp and eliminated the 2x raw read k-mer peak (Figure S2B). Additionally, BUSCO analysis using the OrthoDB Lepidoptera gene set found the percent of complete, duplicated BUSCOs dropped from 48.5% in the unpurged assembly to 6.2% in the purged assembly (Figure S10C). This means that improperly retained alternate haplotypes have been removed from the primary assembly. When we examine the raw read k-mer multiplicity in the primary assembly, we see a peak of k-mers that map only to the primary or alternate assembly at k=13, which corresponds to the heterozygous portions of the genome (Figure 5B). We also see a peak at k=26 in k-mers that are shared between the primary and alternate assemblies, which matches the expectation of a diploid genome with double the read coverage in any homozygous regions of the genome. The frequencies of k-mers missing from either assembly are low, and represent k-mers from sequencing read errors.

To identify which contigs came from contaminant DNA, we used the BLAST nt database, as well as a k-mer distance tree (Figure 5A) and found multiple contaminants. We identified a *Wolbachia* genome contig, several tomato contigs, and many microsporidian contigs, primarily from the *Nosema* genus (a common insect fungal parasite), all of which were expected. We also identified multiple contigs that matched to both *Nosema* and Lepidopteran queries in BLAST. Based on the distance tree's clustering of

these contigs with other *Nosema*-only contigs, as well as the GC content (Figure S11A), we decided to exclude these from the assembly. We also noticed four contigs that matched only to Lepidopteran contigs but clustered with Nosema sequences in k-mer content. One of these, ptg000311, matched to Lepidopteran mitochondrial sequences and likely represents the *T. absoluta* mitogenome. The remaining three matched to the same *Papilo xuthus* genome assembly (PRJNA291600) and is likely the result of inaccurate annotation in the BLAST database, as no decontamination steps were taken during its assembly (Nishikawa *et al.* 2015). In addition, these contigs' GC and repeat content profiles were distinct from all other Lepidopteran contigs (Figure S11), so we excluded them from the assembly. Finally, one contig matched to a mouse-eared bat (*Myosis)* mitochondrial genome; this was possibly contamination from the sequencing facility and was excluded as well.

After removing these contigs, our primary assembly contained 132 contigs all longer than 10 kb with a final length of 635.9 Mbp. 70% of the genome was captured in the 30 longest contigs (L70); as *Tuta absoluta* has 29 chromosomes, this suggests our assembly is approaching chromosome-level contigs (Figure 5C). This represents a significant improvement from the previously published *Tuta absoluta* genome which consists of 81,653 contigs and a length of 906 Mbp (Tabuloc *et al.* 2019).

To generate gene models and functionally annotate genes, we used RepeatModeler and RepeatMasker to soft-mask the genome for repeats using both known Lepidopteran repeat sequences, as well as *de novo* sequences identified from our assembly. We followed with BRAKER2 to identify gene models, and functionally annotated models with EnTAP, based on multiple protein databases (RefSeq Invertebrate, UniProt SwissProt and TrEMBL, Lepbase, and EggNOG) and published *Tuta absoluta* RNAseq datasets. Of the 19,570 gene models identified by BRAKER2, 17,183 were identified as complete by gffread and EnTAP, with 14,019 transcript models matching to a gene name or functional annotation.

Three distinct *Tuta absoluta* populations exist in Latin America

We analyzed whole-genome sequencing data from individuals previously collected from field and greenhouse sites across South America and Costa Rica, as well as a lab colony from Spain (Tabuloc *et al.* 2019) (Figure 6A). Mapping rates to the new genome assembly ranged between 70%-90%, although sequencing depth per individual was low (between 1X to 17X) (Figure S12). One population from Argentina (MP, Mar del Plata, Buenos Aires State) had extremely low mapping rates and read depth, so we excluded it from further analysis. Wherever possible, we used methods based on genotype likelihoods, rather than genotype calls, to account for uncertainty that results from the low read depth.

To investigate population structure in our samples, we used Principal Component Analysis (PCA) and admixture estimation based on allele frequencies from over 900,000 SNPs. The first two PCs captured 18.7% and 16.3% of the total variance in the data, with the remaining PCs each capturing less than 5% of the total data variance (Figure 6C). Samples primarily cluster together based on collection site but also formed three distinct regional groups (Figure 6B). Samples from Chile, Peru, and Ecuador form an "Andes" cluster west of the Andes Mountains; samples from Brazil, Uruguay, Paraguay, and Argentina form a "Central" cluster, east of the Andes Mountains; and samples from Columbia and Costa Rica form a "North" cluster. Spanish samples grouped tightly with the Andes cluster, particularly the VA (Villa Alegre, Chile) site.

When three clusters were allowed in admixture estimation, samples group into the same three clusters as in PCA, while at four clusters, the Spanish samples become their own group, with VA samples sharing a large proportion of admixture. Compared to other Andes populations, the VA samples are more differentiated from Central and North sites as well, with little signal of admixture at all levels of k tested. The other Andes populations (AR, LC, LJ, and RI) all had low admixture proportions from Central at k=2, although at k=3 we see that all RI (Riobamba, Ecuador) samples exhibited admixture from the North populations. This suggests that the non-VA Andes populations are more closely related to Central

populations than VA, and that VA could represent an admixture between the population that gave rise to the Spanish lineage and the other Andes populations. Additionally, we see that RI represents an intermediate population between the Andes and North, which makes sense given its geographic location between the two clusters.

To further quantify population structure between these clusters, we calculated nucleotide diversity, Tajima's D, and Fst using genotype likelihoods (Figure S13). For all clusters, nucleotide diversity was approximately 2%, which is fairly high compared to most Lepidopterans (Mackintosh *et al.* 2019). If we look at the weighted Fst, we see differentiation between clusters is high, particularly between North and all other clusters. The combination of high diversity levels and high Fst could mean these regions diverged from each other a long time ago, prior to the detection of *Tuta absoluta* by growers across Latin America in the 1960s to 1980s. If divergence had occurred recently, we might expect reduced diversity levels in invasive populations relative to the ancestral population.

Treemix confirms clustering and detects migration events to Ecuador and Chile

To detect potential migration events between populations, we used Treemix to build a maximum likelihood (ML) tree based on allele frequencies, as well as predict migration edges and calculate F3 statistics. As Treemix was designed to take allele count data per population, we called genotypes using PCAngsd using a 95% accuracy cutoff and counted alleles within each sampling location. After filtering out loci with missing data, 47,535 SNPs were available for use. In general, the tree topography aligns with results from PCA and admixture analyses. We see sampled sites cluster into the same three clusters, North, Andes, and Central, with the Spanish samples sister to the VA (Chile) site (Figure 7). In agreement with Fst estimates, North populations have experienced more genetic drift from the Andes and Central populations, compared to the Andes and Central populations with each other.

Interestingly, the RI (Ecuador) site does not form a clade with other Central populations but descends from the common ancestor of the Central/Andes group. Based on admixture analysis that showed low

levels of admixture in RI from the North, the position of RI in the tree could be further evidence that Ecuador represents an intermediate mixing zone between populations north and south of it. To investigate further, we re-ran Treemix allowing between one to five migration events (m=0 to 5) and calculated the F3 statistic between all combinations of three populations to see if admixture was supported. At m=2,4 and 5, Treemix reported a strong migration (between 11%-25%) from the Spain or Spain/VA branch to RI, while at m=3 and 5, Treemix reported a weak migration event (2%-7%) from the North to RI. F3 statistics F_3(RI; CH,SP) and F_3(RI; CR,SP) were significantly negative (Table 2), indicating that a simple bifurcating tree does not explain RI's relationship with CH, CR, and SP. While Treemix infers a migration from the Spanish branch to the RI branch, it is important to remember that this migration is inferred to have occurred somewhere along the branch between the current day Spanish population and the most recent common ancestor of Spain and VA (Chile). This migration could have occurred early in the branch, when the population was still in Chile, or late in the branch, when the population moved to Spain. Based on fresh tomato trade between the two countries, in 2006 Chile shipped over 29,000 kg of fresh tomatoes to Spain while importing none back (*Spain Vegetables: tomatoes, fresh or chilled imports from Chile in 2006* 2006). This makes admixture from Spain back to RI unlikely and suggests that RI contains admixture from North and Chilean populations.

In addition to admixture in RI, Treemix and F3 statistics also detected admixture in AR (Chile) from the Spanish population. At m=1,2,3, and 4, Treemix detected a migration edge from the Spanish branch to AR with migration weight varying between 9% to 17%. F3 statistics of AR and Spain with any population from North or Central resulted in a significantly negative value, providing strong evidence of a migration event from a Spanish ancestor like the signal detected with RI. This suggests that the admixture signal we see in AR is from the same Chilean population that gave rise to the Spanish invasion and RI admixture.

Divergence of *Tuta absoluta* predates modern tomato agriculture

Based on the high levels of nucleotide diversity and Fst between the Andes, Central, and North clusters, we hypothesized that the three regions may have diverged many generations ago, before the appearance and detection of Tuta absoluta in agricultural crops throughout South America in the mid-20th century. This would suggest a model in which Tuta absoluta may have adapted from local, wild host plants to nearby tomato fields independently, rather than a single population that became adapted to tomatoes and was spread through human activity. To investigate this, we calculated the folded two-dimension site frequency spectrum (2D-SFS) between populations and estimated parameter values under various population models using maximum likelihood coalescent methods (Figure S14, Table S1). We excluded the VA samples from the Andes cluster to avoid potential modeling issues due to VA appearing to originate from a distinct ancestor than other Andes populations. The simplest model allows for two population splits with constant population sizes, while the exponential growth model adds an exponential growth rate to each population. As exponential growth may not be appropriate if divergence times are long, we also tested a model with a simple resizing event for each population at some point in time. We used a post-hoc comparison of simulated linkage disequilibrium decay rates between models to test model fit. We found that while all three models simulated decay rates within the 95% confidence interval of the Andes population data, none simulated decay rates that overlapped with Central and North decay rate estimates, although the resizing population model was closest. (Figure S15). The lack of fit suggests there are additional complex historical events not well captured in these models. Under the resizing population model, divergence of the North occurred 252,383 generations ago (95% CI: 243,535-326,583), followed by a Central population divergence 187,034 generations ago (95% CI: 181,668-235,424). Reports of *Tuta absoluta* generation times can be as high as 6 to 12 or more generations per year ("*Tuta absoluta*" 2005; Mansour *et al.* 2018; de Campos *et al.* 2021), dating these divergence events to tens of thousands of years ago. This suggests that *Tuta*

absoluta was already present across Latin America prior to the 1960s, and as tomato agriculture surged, adapted locally to the new host plant.

Population Branch Statistic Screening identifies several peaks under selection

The Population Branch Statistic (PBS) is an Fst-based statistic that uses Fst data between 3 populations to calculate the population-specific allele frequency changes. Regions of the genome with abnormally high PBS may be under strong selective forces, causing the loci allele frequencies to change faster than expected by drift. We calculated PBS across the genome for all three populations and found several peaks in contigs 2, 9, 15, and 22 that were exceptionally high and broad, particularly in the North cluster (Figure 9A). The peak in contig 9 contained the gene *paralytic* (*para*, *T. absoluta* gene g15590), a neuronal sodium channel protein that is the active target of pyrethroid insecticides (Dong *et al.* 2014). While PBS peaks in the North population between 13.1 and 13.2Mb on contig009, we note that the allelic diversity was low in the Andes and Central clusters relative to the North (Figure 9B). We calculated allele frequencies of known resistance-inducing mutations in each cluster (Dong *et al.* 2014), and found one mutation, an alanine to leucine substitution at position 1014, was fixed in the Central and Andes, while at 41% frequency in the North (Figure 9C). In addition, we found low to intermediate frequencies of other resistance alleles, including M918T, T929I, V1016G, L925M, and I254T.

A similar selective sweep signal was also seen in the PBS hotspot on contig 2 (8.54Mb-8.58Mb), with high PBS and diversity levels in the North, and a large region of low allelic diversity in the Andes and Central (Figure S16A). Several genes were captured in this interval, including *NADH:Ubiquinone oxidoreductase subunit A8 (ndufa8)*; *mucin-5AC-like*, a gene putatively related to human *mucin-5AC*, and two hemomucin genes involved in hemocyte adhesion and innate immunity in insects. The PBS hotspot on contig 15 contained *cryptochrome-2 (cry2)*, encoding a key component of the circadian clock (Figure S16B). Finally, the large hotspot on contig 22 contained 69 genes, including multiple copies of *juvenile hormone binding protein*, ribosomal proteins, gustatory response genes, rhodopsins, and

acetylcholinesterase (*ache*), a gene implicated in organophosphate insecticide resistance (Figure S16C). Interestingly, based on alignment data it appears the Central and Andes populations may have two copies of *ache*, while North populations only have one. We looked for known mutations conferring organophosphate resistance and found moderate frequencies in all three clusters as well (Table S2).

Discussion

Using whole genome sequencing data, we found that *Tuta absoluta* samples collected from 11 locations in Latin America clustered into three basic regions, comprised of a North, Andes, and Central group. In addition, we see that Spanish populations likely originated from a Central Chilean source based on their low level of Fst with the Andes and location on the ML tree. Previous analyses with mitochondrial sequences were unable to differentiate populations (Cifuentes *et al.* 2011); however, analyses using microsatellite data was able to identify these same three clusters and suggest a Central Chilean source for the European migration as well (Guillemaud *et al.* 2015). In agreement with this conclusion, looking at fresh tomato export data we see Chile is a worldwide exporter, shipping 12 tons of tomatoes an average distance of 11,700 km in 2018, while most other countries in South America tend to export within the continent (*Worldwide Production of Tomatoes* 2018).

The Andes Mountains represent an obvious geographic barrier that would separate the Central population from the North and Andes populations. Population structure between Andes and North populations may be due to factors related to the changing latitude, including temperature and daylength. While PCA groups our Ecuador samples (RI) with Andes populations, admixture analysis and Treemix both provided some evidence that Ecuador may represent an admixture zone between the two regions. The weighted Fst between the North and Andes is also lower than between North and Central, suggesting that North and Andes are indeed more closely related. This general Fst pattern was also observed based on microsatellite analyses (Guillemaud *et al.* 2015). Sequencing of more samples from

Peru and Ecuador might be needed to further elucidate the extent of an admixture zone between these clusters.

While *Tuta absoluta* was first discovered in Peru in 1917, its native range is not well established. One hypothesis is that *Tuta absoluta* migrated out of the Andes region and across South America through the 1960s-80s because of human transport by agricultural shipping. This aligns with the surge in domestic tomato agriculture in South America at the same time (Minami 1980). However, based on the similar nucleotide diversity levels between clusters, as well as high levels of Fst, we hypothesized it might be more likely that this migration across South America may have happened prior to tomato commercialization, with populations of *Tuta absoluta* later adapting to the appearance of commercial tomato agriculture. Based on our simple 3 population model, it appears the ancestral population diverged twice tens of thousands of years ago. Relative to the ancestral population size, the combined effective population size is roughly three times larger, although this is heavily weighted toward a very large Andes population, relative to the North and Central regions. The fact that the estimated Andes population size is nearly 10 times larger than that of Central or North populations, ,as well as its slightly higher level of genetic diversity, could suggest that the Andes cluster represents the ancestral population range. Given that the wild ancestors of tomatoes and potatoes are also native to the Andes region (Spooner *et al.* 2005; Peralta and Spooner 2006), the Andes region would be an ideal place to search for native parasitoids for biocontrol. To date, biocontrol methods in South America have relied on non-natives or generalist parasites, with only a handful of native specialists identified in the literature and none that are commercially available (Salas Gervassio *et al.* 2019; Desneux *et al.* 2022). Thus, knowledge of *Tuta absoluta*'s native range may help focus efforts to identify more natural parasitoids.

Using the PBS, we found multiple genomic windows under apparent selective forces. Not surprisingly, one of the highest PBS windows contained the *para* gene, which encodes a sodium ion channel that is targeted by pyrethroids. The extremely low allele diversity in the Andes and Central populations relative

to the North suggest a hard selective sweep occurred here. Heavy pyrethroid use in Brazil led to the appearance of resistant strains starting in the 1990s (Siqueira *et al.* 2000). We found that Central and Andes populations were completely fixed for the L1014F mutation, one of the most common causes of knock-down-resistance (kdr) to pyrethroids (Dong *et al.* 2014), while the North had an intermediate frequency. A study looking at Brazilian populations found a similar pattern of fixed L1014 (Silva *et al.* 2015), while another study looking at multiple populations in South America also found the same pattern of L1014F fixation in Central and Andes populations but not in the North (Haddi *et al.* 2012). While both studies also found M918T and T929I at elevated frequencies in all populations, we additionally detected the resistance allele V1016G in the North. We also found L925M in Central and I254T in Andes. While these have not been characterized as resistance alleles, mutations at these same positions have been shown to confer resistance in *Drosophila melanogaster* (I254N) and *Bemisia tabaci* (L925I) (Pittendrigh *et al.* 1997; Morin *et al.* 2002). These new appearances indicate selection is ongoing in all three clusters.

Other regions under selection were less obvious. We found one region in contig 2 containing *Ndufa8* and several hemomucin/mucin genes with a similar low genetic diversity in the Central and Andes populations and high PBS in North, indicating a hard selective sweep. *Ndufa8* produces a nuclear-encoded subunit of the NADH dehydrogenase complex I, part of the electron transport chain in the mitochondria used to generate ATP. Mutations here are known to cause mitochondrial complex I deficiency in humans (Yatsuka *et al.* 2020), although a few studies have found evidence of positive selection occurring in other species, potentially related to metabolism (Kozell *et al.* 2020; Lee *et al.* 2020). The hemomucin genes are a component of the insect immune system that are involved in endocytosis (Schmidt *et al.* 2010). Selection here could be in response to the increased use of parasitoids and predators such as *Trichogramma evanescens and Nesidiocoris tenuis* as an biological control alternative to insecticides (Han *et al.* 2019).

The elevated PBS region in contig 22 was relatively large at approximately 1Mb in size, containing over 60 genes. Interestingly, we noticed two copies of *ache* contained within this window, which codes for *acetylcholinesterase*, a gene which encodes a protein which degrades the neurotransmitter acetylcholine (Mutero *et al.* 1994). As this enzyme is the main target of organophosphates and carbamate insecticides, alleles in *ache* have been documented to confer resistance. Using the amino acid numbering scheme based on *T. californica* (Massoulié *et al.* 1992), the resistance allele A201S has previously been reported to be present in European populations (Haddi *et al.* 2017), and we found this allele to be at moderate to high frequency in all three regions. We also found moderate frequencies of the mutation F290V and F290N. The F290V resistance allele has been documented in mosquitoes and moths (Cassanelli *et al.* 2006; Alout *et al.* 2009; Carvalho *et al.* 2013), while other mutations such as F290Y have been documented in *Drosophila* and *M. domestica* (Mutero *et al.* 1994; WALSH *et al.* 2001). Interestingly, based on mapping read depth the North population only contained a single copy of *ache*. As duplication of *ache* has been implicated in improved organophosphate resistance (Alout *et al.* 2009; Sonoda *et al.* 2014), this large structural duplication may be the reason for elevated PBS levels across such a large interval. Follow-up work with long-read methods or higher sequencing coverage will be needed to confirm the presence of structural duplication at this locus.

We expect that addition of a new contiguous genome assembly with annotations will be of benefit to the *Tuta absoluta* and Lepidopteran research community. Previous studies have worked to develop potential RNA interference (RNAi) strategies to use as an alternative to traditional pesticides (Camargo *et al.* 2015). Work is also being conducted to develop Cas9 gene-editing techniques for *Tuta absoluta* to facilitate future genetics studies (Ji *et al.* 2022). These developments in combination with an accurate assembly and gene annotations will allow for accelerated research towards understanding *Tuta absoluta* biology and methods to contain its economic impacts and spread.

Figures

Figure 5: Genome assembly assessment. (A) K-mer based distance tree of contigs in the primary genome assembly before de-contamination, labeled by NCBI BLAST matches in the “nt” database. Contig ptg000311l matched to insect mitochondrial sequences. Contigs ptg000311l, pg000280l, and ptg000281l were contigs annotated as “Lepidoptera” by BLAST but clustered with microsporidian sequences in the tree. (B) K-mer multiplicity plot of input CCS reads against the assembly. “Read only” indicates k-mers that only appear the raw reads; “primary only” “alt only” indicate read k-mers that appear in only one of the 2 haplotig sets; “shared” indicates k-mers that appear in both haplotig sets. (C) Cumulative length of contigs in primary assembly, ordered from longest to shortest. L50 and L70 indicate smallest number of contigs that contain 50% or 70% of the assembly length.

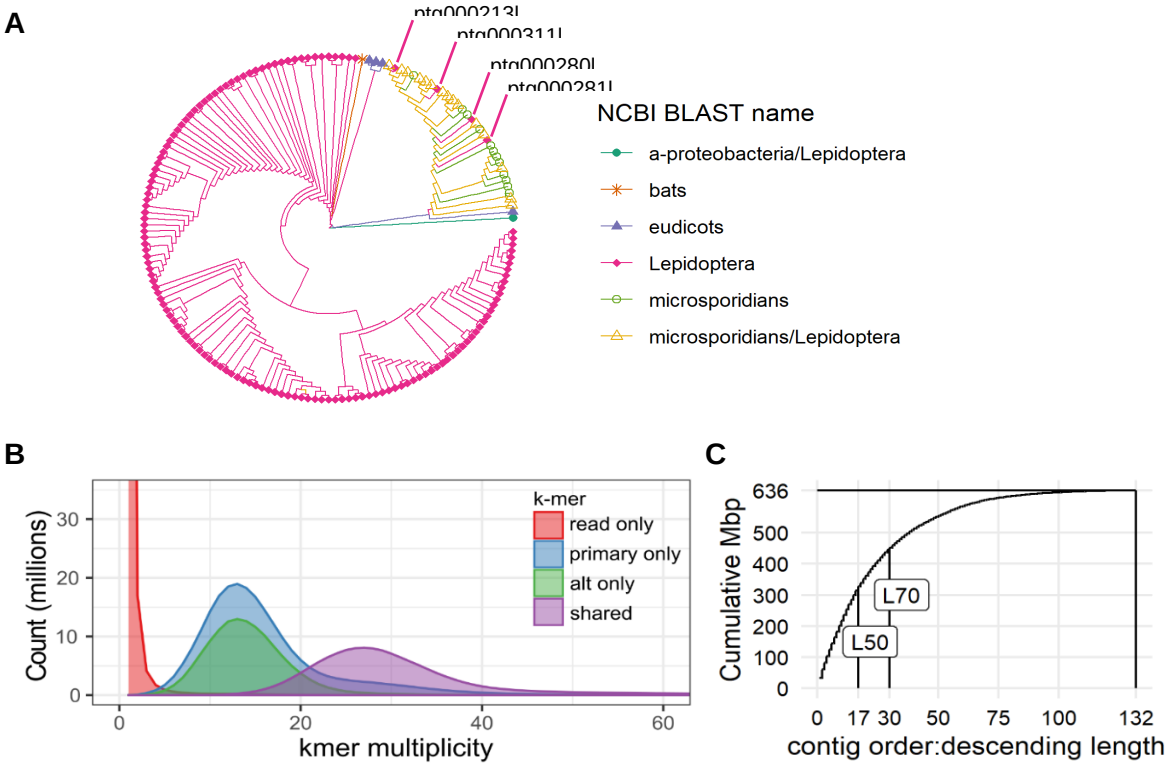


Figure 6: Sampling sites and population structure of *Tuta absoluta* individuals. (A) Map of sampling locations. Between 4 to 8 individuals were sequenced per location. Legend indicates grouping as determined from PCA and admixture analyses. (B) PCA plot of the first two principal components (PCs), based on genotype likelihoods from 933,060 sites. (C) Percent variance captured by PCs in descending order. (D) Admixture analysis for 2 to 4 clusters. Colored bars indicate the posterior probability of an individual belonging to a given cluster. Location codes: AR=Arica, Chile; CH=Chia, Columbia; CN=Campo Nove, Paraguay; CR=Costa Rica; LC=La Curva, Peru; LJ=La Joia, Peru; OV=Ouro Verde, Brazil; RI = Riobamba, Ecuador; RO=Rocha, Uruguay; SE = Santiago del Estero, Argentina; SP = Barcelona, Spain; VA = Villa Alegre, Chile.

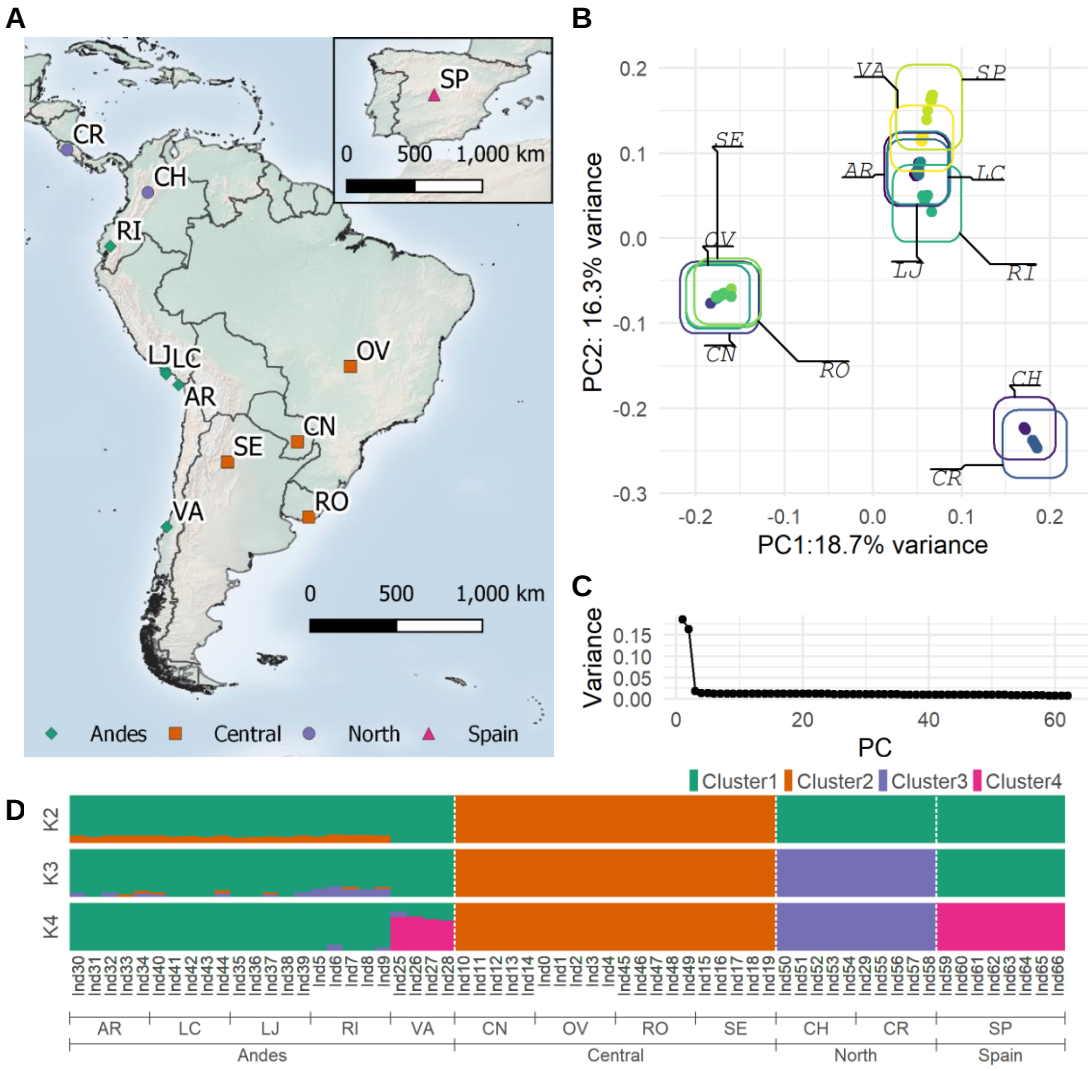


Figure 7: Maximum likelihood tree from Treemix with no migration edges called rooted on the Costa Rica (CR) samples, based on 47,535 SNPs. Confidence values were based on 100 jackknife bootstraps with 500 SNP bins. X-axis represents genetic drift distance between populations.

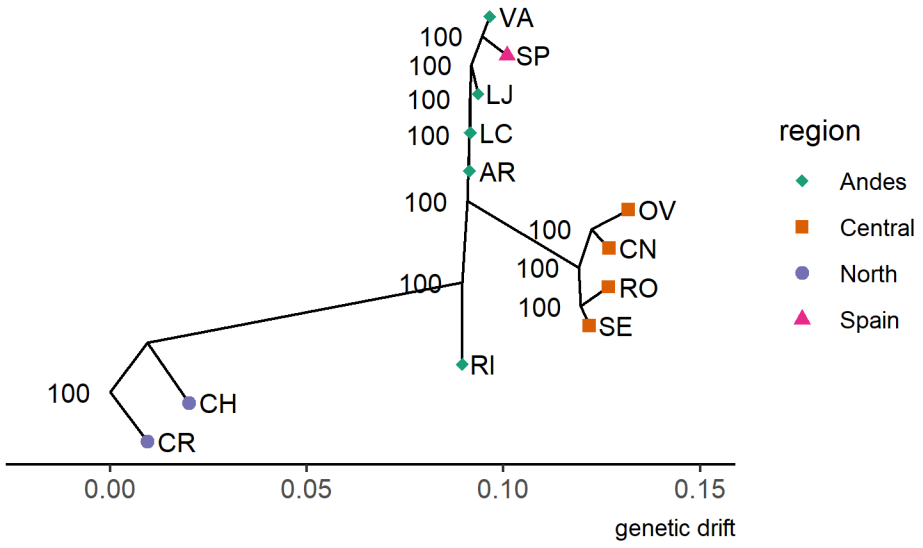


Figure 8: Coalescent simulations used to estimate model parameters. (A) Model for a three-population split with population resizing events for each population. (B) Maximum likelihood estimates of each parameter in the model indicated by a red dot. Violin plot depicts distribution of parameter values from 100 parametric bootstraps, with upper and lower boundary lines indicating the 95% interval. All point estimates were within the 95% bootstrap intervals except N.ancestral, N.final.Andes, N.final.Central, and N.final.North.

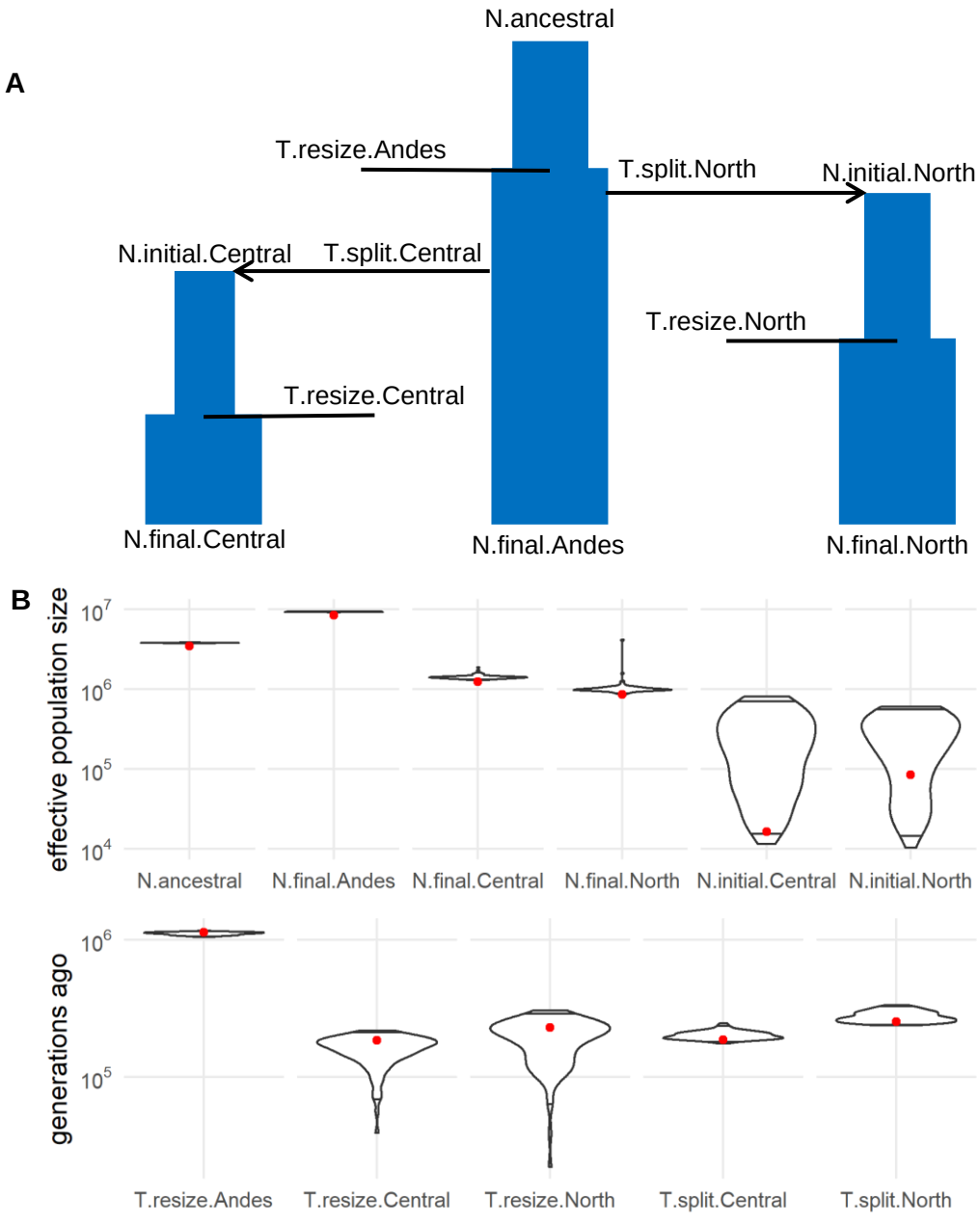
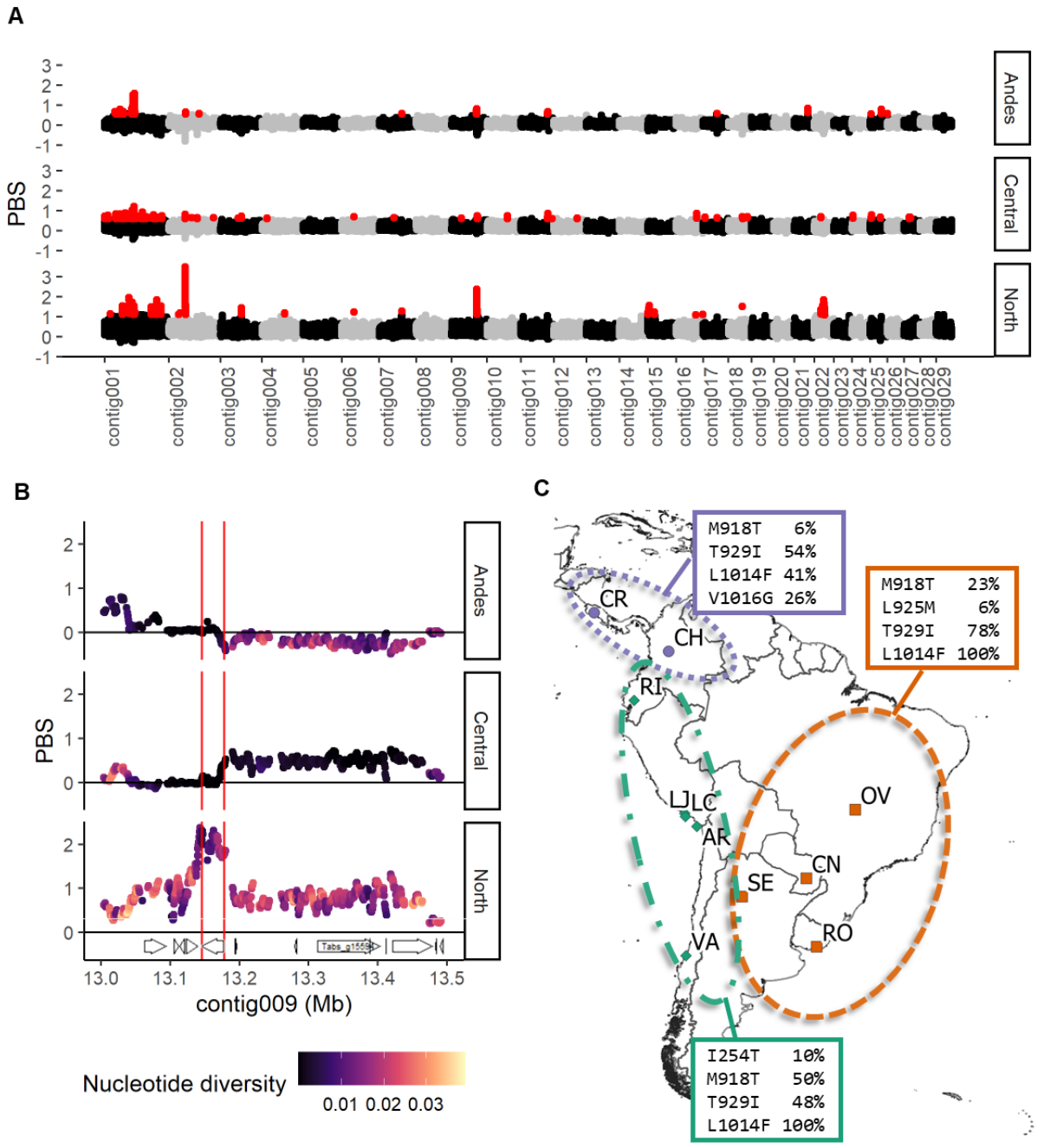


Figure 9: Selection signals in *Tuta absoluta*. (A) PBS values in each region calculated across the largest 29 contigs in 5kb intervals with 500bp steps. Red points indicate the highest 0.1% PBS values. (B) Plot of PBS and genetic diversity (pi) at contig009. Red vertical lines border the PARA gene model. (C) Map of sampling locations with allele frequencies of known amino acid substitutions that confer pyrethroid resistance within each cluster. Note the L1014F allele, which was found in all Central and Andes samples, while at 41% in North samples.


A
PBS
Andes
Central
North
contig001
contig002
contig003
contig004
contig005
contig006
contig007
contig008
contig009
contig010
contig011
contig012
contig013
contig014
contig015
contig016
contig017
contig018
contig019
contig020
contig021
contig022
contig023
contig024
contig025
contig026
contig027
contig028
contig029
B
PBS
contig009 (Mb)
13.0
13.1
13.2
13.3
13.4
13.5
Nucleotide diversity
0.01 0.02 0.03
C
M918T 6%
T929I 54%
L1014F 41%
V1016G 26%
M918T 23%
L925M 6%
T929I 78%
L1014F 100%
I254T 10%
M918T 50%
T929I 48%
L1014F 100%
CR
CH
RI
LJLC
AR
SE
VA
OV
CN
RO

Tables

Table 1: Migration events with predicted weights from Treemix between populations, under different models allowing between 1 to 5 migration events (m). * indicates the migration significantly improved model fit ($p < 0.05$) based on the Wald statistic using jackknife estimates. Inferred migrations originate somewhere between the "start-node" and the start-node's most previous branchpoint, and terminate somewhere between the end-node and the end-node's most previous branchpoint. As an example, a migration starting from "SP" occurs somewhere on the branch between "SP" and the most recent common ancestor of "SP" and "VA".

m	Weight (%)	start-node	end-node
1	8.8*	SP	AR
2	11.2*	SP	AR
2	15.1*	VA,SP	RI
3	17.7*	SP	AR
3	17.3*	SP	LC
3	2.0*	CR,CH	RI
4	10.6*	SP	AR
4	25.4*	VA,SP	RI
4	2.2*	OV	RI
4	13.6	SP	LJ
5	21.3*	SP	LC,AR
5	11.0*	SP	RI
5	24.2*	SP	LJ
5	6.7*	CR	RI
5	2.9*	LC,AR	OV

Table 2: Significant F3 statistics. Population comparisons for which F3 statistics were significant (Z score <-3) are reported.

Pop(A;B,C)	**F3**	**SE**	**Z score**
AR;CH,SP	-0.00165	0.000171	-9.63768
AR;CR,SP	-0.00178	0.000171	-10.4413
AR;CN,SP	-0.00078	0.00015	-5.15567
AR;OV,SP	-0.00082	0.000145	-5.63104
AR;RO,SP	-0.00076	0.000151	-5.05319
AR;SE,SP	-0.00077	0.00014	-5.47246
LC;CH,SP	-0.00146	0.000193	-7.56011
LC;CR,SP	-0.00154	0.000211	-7.31075
LC;OV,SP	-0.00054	0.00018	-3.00356
RI;CH,SP	-0.00285	0.000948	-3.00128
RI;CR,SP	-0.00296	0.000965	-3.06668

References

Alout, H., P. Labbé, A. Berthomieu, N. Pasteur, and M. Weill, 2009 Multiple duplications of the rare ace-1 mutation F290V in Culex pipiens natural populations. Insect Biochemistry and Molecular Biology 39: 884–891.

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6: 11.

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting FST: The impact of rare variants. Genome Res. 23: 1514–1521.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Brian, H., and A. Papanicolaou TransDecoder.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics 3:.

Buffalo, V., 2014 Scythe: A 3'-end adapter contaminant trimmer.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Camargo, R. de A., R. H. Herai, L. N. Santos, F. M. M. Bento, J. E. Lima *et al.*, 2015 De novo transcriptome assembly and analysis to identify potential gene targets for RNAi-mediated control of the tomato leafminer (Tuta absoluta). BMC Genomics 16: 635.

de Campos, M. R., P. Béarez, E. Amiens-Desneux, L. Ponti, A. P. Gutierrez *et al.*, 2021 Thermal biology of Tuta absoluta: demographic parameters and facultative diapause. J Pest Sci 94: 829–842.

Carvalho, R. A., C. Omoto, L. M. Field, M. S. Williamson, and C. Bass, 2013 Investigating the Molecular Mechanisms of Organophosphate and Pyrethroid Resistance in the Fall Armyworm Spodoptera frugiperda. PLOS ONE 8: e62268.

Cassanelli, S., M. Reyes, M. Rault, G. Carlo Manicardi, and B. Sauphanor, 2006 Acetylcholinesterase mutation in an insecticide-resistant population of the codling moth Cydia pomonella (L.). Insect Biochemistry and Molecular Biology 36: 642–653.

Challi, R. J., S. Kumar, K. K. Dasmahapatra, C. D. Jiggins, and M. Blaxter, 2016 Lepbase: the Lepidopteran genome database. 056994.

Chang, P. E. C., and M. A. Metz, 2021 Classification of Tuta absoluta (Meyrick, 1917) (Lepidoptera: Gelechiidae: Gelechiinae: Gnorimoschemini) Based on Cladistic Analysis of Morphology. went 123: 41–54.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li, 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18: 170–175.

Cifuentes, D., R. Chynoweth, and P. Bielza, 2011 Genetic study of Mediterranean and South American populations of tomato leafminer Tuta absoluta (Povolny, 1994) (Lepidoptera: Gelechiidae) using ribosomal and mitochondrial markers. Pest Management Science 67: 1155–1162.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan *et al.*, 2021 Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Desneux, N., P. Han, R. Mansour, J. Arnó, T. Brévault *et al.*, 2022 Integrated pest management of Tuta absoluta: practical implementations across different world regions. J Pest Sci 95: 17–39.

Desneux, N., M. G. Luna, T. Guillemaud, and A. Urbaneja, 2011 The invasive South American tomato pinworm, Tuta absoluta, continues to spread in Afro-Eurasia and beyond: The new threat to tomato world production. Journal of Pest Science 84: 403–408.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dong, K., Y. Du, F. Rinkevich, Y. Nomura, P. Xu *et al.*, 2014 Molecular biology of insect sodium channels and pyrethroid resistance. Insect Biochemistry and Molecular Biology 50: 1–17.

EPPO, 2008 First record of *Tuta absoluta* in Spain. EPPO Reporting Service 1(001):

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLOS Genetics 9: e1003905.

FastQC, 2019.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark *et al.*, 2020 RepeatModeler2 for automated genomic discovery of transposable element families. PNAS 117: 9451–9457.

Fox, E. A., A. E. Wright, M. Fumagalli, and F. G. Vieira, 2019 ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics 35: 3855–3856.

Gabriel, L., K. J. Hoff, T. Brůna, M. Borodovsky, and M. Stanke, 2021 TSEBRA: transcript selector for BRAKER. BMC Bioinformatics 22: 566.

Godfrey, K., F. Zalom, and J. Chiu, 2018 Tuta Absoluta, The South American Tomato Leafminer: University of California, Agriculture and Natural Resources ANR Publication 8589.

Guan, D. F. purge_dups.

Guillemaud, T., A. Blin, I. Le Goff, N. Desneux, M. Reyes *et al.*, 2015 The tomato borer, Tuta absoluta, invading the Mediterranean Basin, originates from a single introduction from Central Chile. Sci Rep 5: 8371.

Haddi, K., M. Berger, P. Bielza, D. Cifuentes, L. M. Field *et al.*, 2012 Identification of mutations associated with pyrethroid resistance in the voltage-gated sodium channel of the tomato leaf miner (Tuta absoluta). Insect Biochemistry and Molecular Biology 42: 506–513.

Haddi, K., M. Berger, P. Bielza, C. Rapisarda, M. S. Williamson *et al.*, 2017 Mutation in the *ace-1* gene of the tomato leaf miner (*Tuta absoluta*) associated with organophosphates resistance. J. Appl. Entomol. 141: 612–619.

Han, P., Y. Bayram, L. Shaltiel-Harpaz, F. Sohrabi, A. Saji *et al.*, 2019 Tuta absoluta continues to disperse in Asia: damage, ongoing management and future challenges. J Pest Sci 92: 1317–1327.

Hart, A. J., S. Ginzburg, M. (Sam) Xu, C. R. Fisher, N. Rahmatpour *et al.*, 2020 EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Molecular Ecology Resources 20: 591–604.

Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund *et al.*, 2019 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47: D309–D314.

Ji, S.-X., S.-Y. Bi, X.-D. Wang, Q. Wu, Y.-H. Tang *et al.*, 2022 First Report on CRISPR/Cas9-Based Genome Editing in the Destructive Invasive Pest Tuta Absoluta (Meyrick) (Lepidoptera: Gelechiidae). Front Genet 13: 865622.

Joshi, N., and J. Fass, 2011 Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Keightley, P. D., A. Pinharanda, R. W. Ness, F. Simpson, K. K. Dasmahapatra *et al.*, 2015 Estimation of the Spontaneous Mutation Rate in Heliconius melpomene. Mol Biol Evol 32: 239–243.

Koch, J. B., J. R. Dupuis, M.-K. Jardeleza, N. Ouedraogo, S. M. Geib *et al.*, 2020 Population genomic and phenotype diversity of invasive Drosophila suzukii in Hawai‘i. Biol Invasions 22: 1753–1770.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15: 356.

Kozell, L. B., D. Lockwood, P. Darakjian, S. Edmunds, K. Shepherdson *et al.*, 2020 RNA-Seq Analysis of Genetic and Transcriptome Network Effects of Dual-Trait Selection for Ethanol Preference and

Withdrawal Using SOT and NOT Genetic Models. Alcoholism: Clinical and Experimental Research 44: 820–830.

Kriventseva, E. V., D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias *et al.*, 2019 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 47: D807–D811.

Lee, H., J. Kim, J. A. Weber, O. Chung, Y. S. Cho *et al.*, 2020 Whole Genome Analysis of the Red-Crowned Crane Provides Insight into Avian Longevity. Molecules and Cells 43: 86–95.

Lewald, K. M., A. Abrieux, D. A. Wilson, Y. Lee, W. R. Conner *et al.*, 2021 Population genomics of *Drosophila suzukii* reveal longitudinal population structure and signals of migrations in and out of the continental United States. G3 Genes|Genomes|Genetics jkab343.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio].

Mackintosh, A., D. R. Laetsch, A. Hayward, B. Charlesworth, M. Waterfall *et al.*, 2019 The determinants of genetic diversity in butterflies. Nat Commun 10: 3466.

Mallet, L., T. Bitard-Feildel, F. Cerutti, and H. Chiapello, 2017 PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies (J. Hancock, Ed.). Bioinformatics 33: 3283–3285.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, 2021 BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes (J. Kelley, Ed.). Molecular Biology and Evolution 38: 4647–4654.

Mansour, R., T. Brévault, A. Chailleux, A. Cherif, K. Grissa-Lebdi *et al.*, 2018 Occurrence, biology, natural enemies and management of Tuta absoluta in Africa. entomologia 38: 83–112.

Mapleson, D., G. Garcia Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, 2016 KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics btw663.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

Massoulié, J., J. L. Sussman, B. P. Doctor, H. Soreq, B. Velan *et al.*, 1992 Recommendations for Nomenclature in Cholinesterases, pp. 285–288 in *Multidisciplinary Approaches to Cholinesterase Functions*, edited by A. Shafferman and B. Velan. Springer US, Boston, MA.

Meisner, J., and A. Albrechtsen, 2018 Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210: 719–731.

Meyrick, E., 1917 I. Descriptions of South American Micro-Lepidoptera. Transactions of the Royal Entomological Society of London 65: 1–52.

Minami, K., 1980 THE HISTORY OF TOMATO PRODUCTION FOR INDUSTRY IN SOUTH AMERICA. Acta Hortic. 87–92.

Morin, S., M. S. Williamson, S. J. Goodson, J. K. Brown, B. E. Tabashnik *et al.*, 2002 Mutations in the Bemisia tabaci para sodium channel gene associated with resistance to a pyrethroid plus organophosphate mixture. Insect Biochemistry and Molecular Biology 32: 1781–1791.

Mutero, A., M. Pralavorio, J. M. Bride, and D. Fournier, 1994 Resistance-associated point mutations in insecticide-insensitive acetylcholinesterase. Proc Natl Acad Sci U S A 91: 5922–5926.

Nishikawa, H., T. Iijima, R. Kajitani, J. Yamaguchi, T. Ando *et al.*, 2015 A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. Nat Genet 47: 405–409.

Nurk, S., B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon *et al.*, 2020 HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30: 1291–1305.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad *et al.*, 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733–D745.

Paladino, L. Z. C., M. E. Ferrari, J. P. Lauría, C. L. Cagnotti, J. Šíchová *et al.*, 2016 The Effect of X-Rays on Cytological Traits of Tuta absoluta (Lepidoptera: Gelechiidae). flen 99: 43–53.

Paradis, E., and K. Schliep, 2019 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526–528.

Peralta, I. E., and D. M. Spooner, 2006 History, Origin, and Early Cultivation of Tomato (Solanaceae), pp. 1–24 in *Genetic Improvement of Solanaceous Crops*, Science Publishers, Enfield, NH.

Pertea, G., and M. Pertea, 2020 GFF Utilities: GffRead and GffCompare.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genetics 8: e1002967.

Pittendrigh, B., R. Reenan, R. H. ffrench-Constant, and B. Ganetzky, 1997 Point mutations in the Drosophila sodium channel gene para associated with resistance to DDT and pyrethroid insecticides. Mol Gen Genet 256: 602–610.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz, 2020 GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11: 1432.

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

Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy, 2020 Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21: 245.

Salas Gervassio, N. G., D. Aquino, C. Vallina, A. Biondi, and M. G. Luna, 2019 A re-examination of Tuta absoluta parasitoids in South America for optimized biological control. J Pest Sci 92: 1343–1357.

Schmidt, O., K. Söderhäll, U. Theopold, and I. Faye, 2010 Role of Adhesion in Arthropod Immune Recognition. Annual Review of Entomology 55: 485–504.

Silva, W. M., M. Berger, C. Bass, V. Q. Balbino, M. H. P. Amaral *et al.*, 2015 Status of pyrethroid resistance and mechanisms in Brazilian populations of Tuta absoluta. Pesticide Biochemistry and Physiology 122: 8–14.

Siqueira, H. Á. A., R. N. C. Guedes, and M. C. Picanço, 2000 Insecticide resistance in populations of Tuta absoluta (Lepidoptera: Gelechiidae). Agricultural and Forest Entomology 2: 147–153.

Skotte, L., T. S. Korneliussen, and A. Albrechtsen, 2013 Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics 195: 693–702.

Smit, A. F. A., R. Hubley, and P. Green, 2021 RepeatMasker.

Sonoda, S., X. Shi, D. Song, P. Liang, X. Gao *et al.*, 2014 Duplication of acetylcholinesterase gene in diamondback moth strains with different sensitivities to acephate. Insect Biochemistry and Molecular Biology 48: 83–90.

Spain Vegetables: tomatoes, fresh or chilled imports from Chile in 2006, 2006 Worldbank World Integrated Trade Solution.

Spooner, D. M., K. McLean, G. Ramsay, R. Waugh, and G. J. Bryan, 2005 A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc Natl Acad Sci U S A 102: 14694–14699.

Tabuloc, C. A., K. M. Lewald, W. R. Conner, Y. Lee, E. K. Lee *et al.*, 2019 Sequencing of *Tuta absoluta* genome to develop SNP genotyping assays for species identification. J Pest Sci.

The UniProt Consortium, 2021 UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49: D480–D489.

Trask, J. A. S., R. S. Malhi, S. Kanthaswamy, J. Johnson, W. T. Garnica *et al.*, 2011 The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*). Primates 52: 129–138.

Tuta absoluta, 2005 EPPO Bulletin 35: 434–435.

WALSH, S. B., T. A. DOLDEN, G. D. MOORES, M. KRISTENSEN, T. LEWIS *et al.*, 2001 Identification and characterization of mutations in housefly (Musca domestica) acetylcholinesterase involved in insecticide resistance. Biochemical Journal 359: 175–181.

Willing, E.-M., C. Dreyer, and C. van Oosterhout, 2012 Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. PLOS ONE 7: e42649.

Worldwide Production of Tomatoes, 2018 Food and Agriculture Organization of the United Nations.

Yamamoto, K., J. Nohata, K. Kadono-Okuda, J. Narukawa, M. Sasanuma *et al.*, 2008 A BAC-based integrated linkage map of the silkworm Bombyx mori. Genome Biol 9: R21.

Yatsuka, Y., Y. Kishita, L. E. Formosa, M. Shimura, F. Nozaki *et al.*, 2020 A homozygous variant in NDUFA8 is associated with developmental delay, microcephaly, and epilepsy due to mitochondrial complex I deficiency. Clin Genet 98: 155–165.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam, 2017 ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8: 28–36.

Supplemental Figures

Figure S9: (A) GenomeScope profile of Pacbio CCS reads. (B) GC percent vs k-mer frequency plot of CCS reads, excluding k-mers with frequency less than or equal to 5 (to ignore unique k-mers due to sequencing errors).

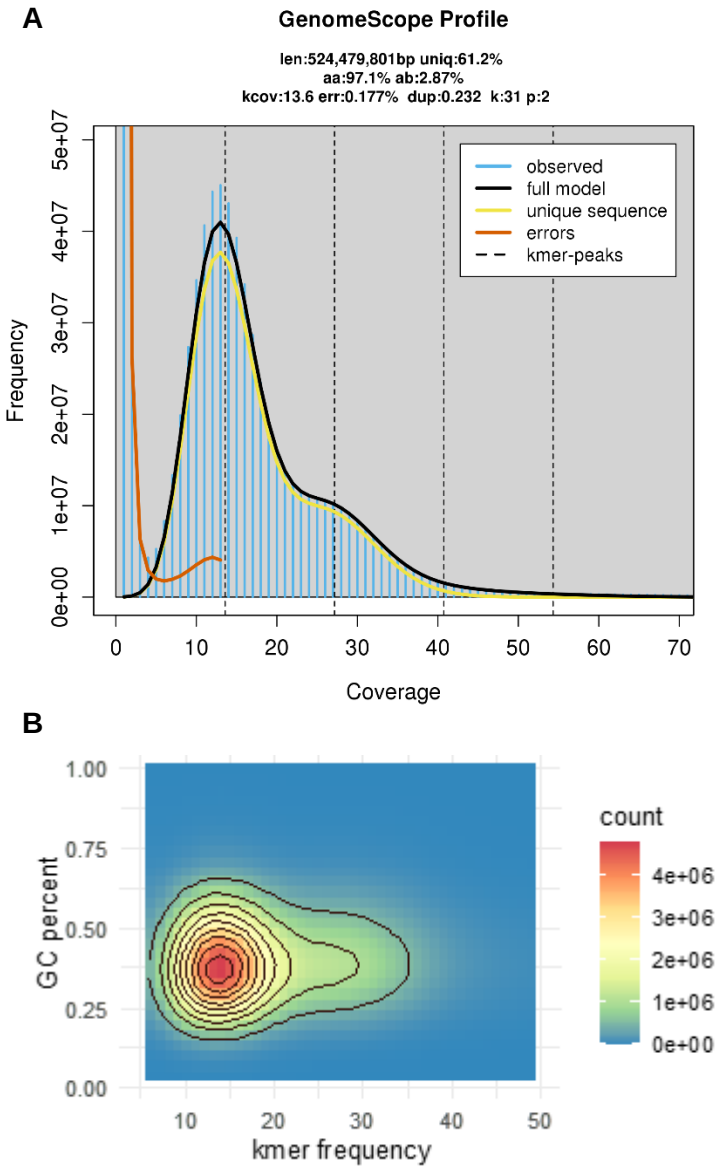
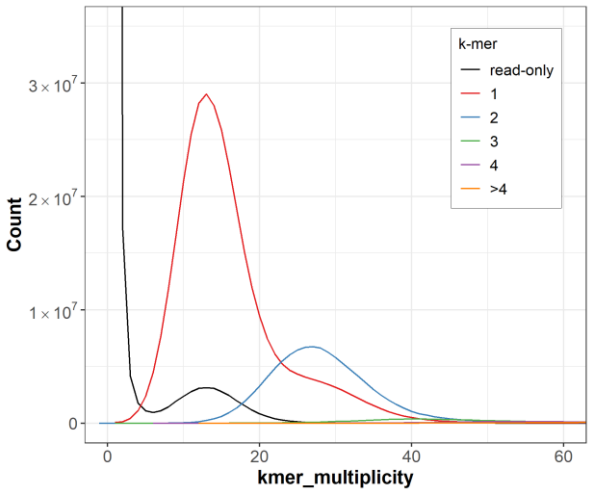
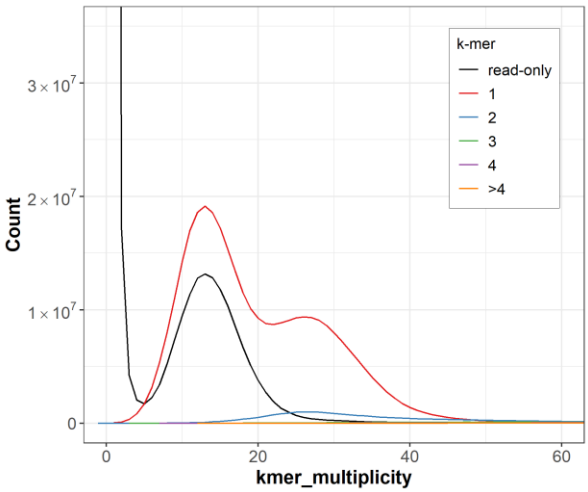


Figure S10: K-mer multiplicity plots by Merqury of the hifiasm primary assembly (A) before and (B) after purging retained haplotigs. (C) Table of Lepidopteran BUSCO scores of the purged and unpurged hifiasm primary assembly, compared to the previously published Tuta absoluta assembly.

A



B



C

BUSCO	Hifiasm-purged	Hifiasm-unpurged	Old assembly
Complete-single copy	4867 (92.1%)	2642 (50.0%)	3660 (69.2%)
Complete-duplicated	327 (6.2%)	2563 (48.5%)	1252 (23.7%)
Fragmented	27 (0.5%)	25 (0.5%)	208 (3.2%)
Missing	65 (1.2%)	56 (1.0%)	166 (3.2%)

Figure S11: Decontamination results of the primary assembly. (A) Blobplot of primary contigs showing best BLAST matches vs GC % and mean read depth. The three “Lepidopteran” contigs at the 30% GC position were contigs ptg000311l, pg000280l, and ptg000281l, which appear to be microsporidian contamination that has been mislabeled as Lepidopteran. B) Repeat content of each contig in the primary assembly. Note contigs ptg000213, ptg000280, and ptg000281 have abnormally low repeat content.

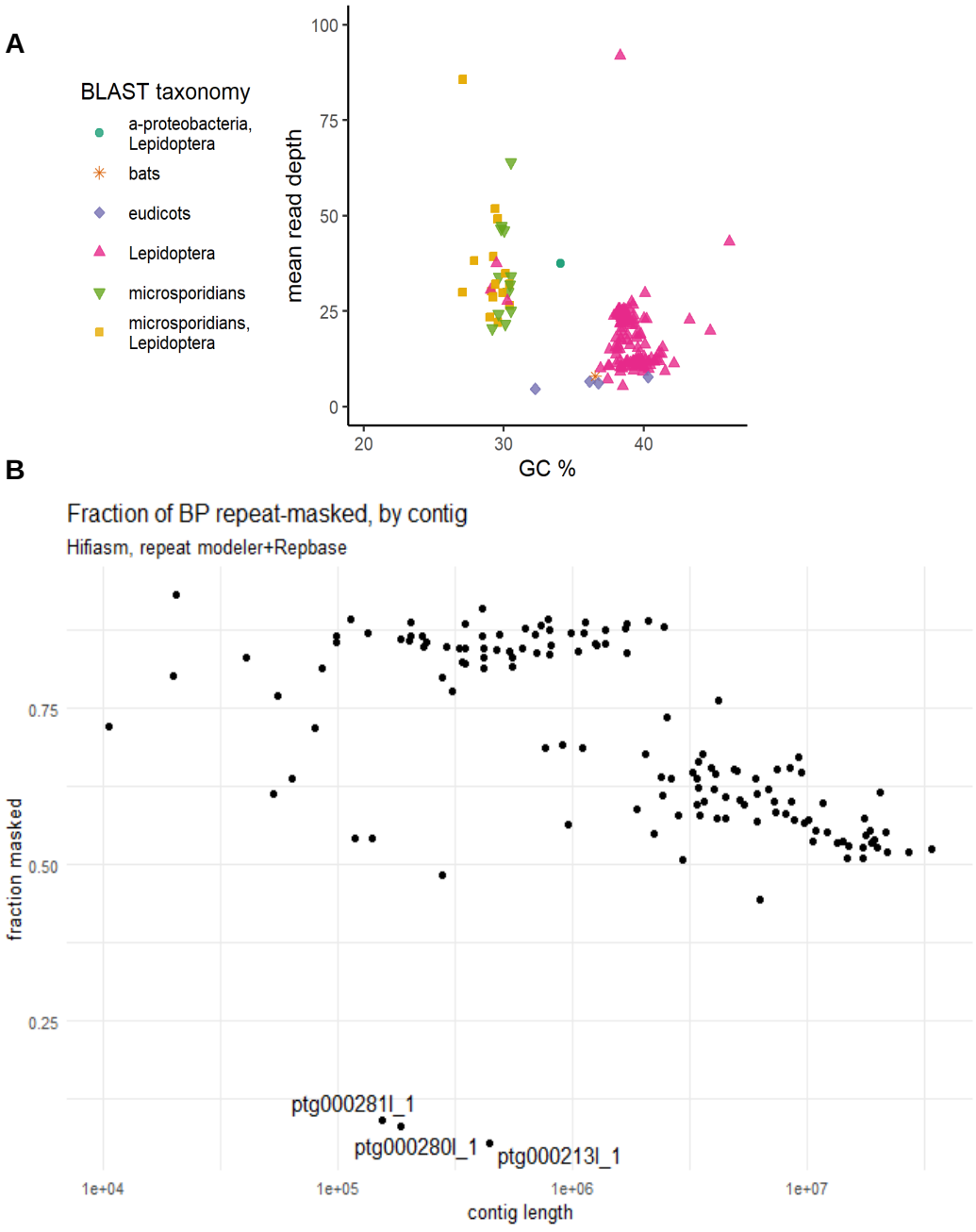


Figure S12: (A) Mapping rates and (B) median read coverage for Illumina reads from population sampling. Median read coverage was calculated at GC=39%, as this is the average GC content of the genome.

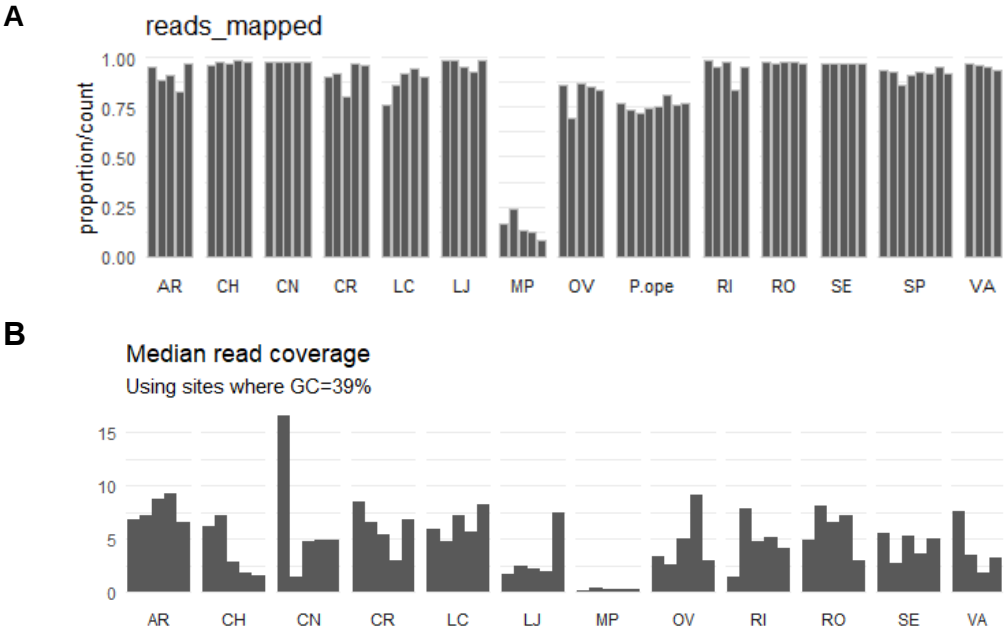


Figure S13: Summary statistics for the 4 key populations. (A) Pairwise nucleotide diversity and (B) Tajima's D for each cluster, averaged across the genome in 20kb windows. N represents the number of windows included. Pairwise t-tests with a Holm's correction method were used to compare means. **** indicates p-value < 0.0001. (C) Unweighted Fst and (D) weighted Fst calculated by Angsd. Weighted Fst is typically considered more accurate as it is less biased when using many rare, population-specific SNPSs (as is the case when genotyping by whole-genome sequencing) (Bhatia et al. 2013).

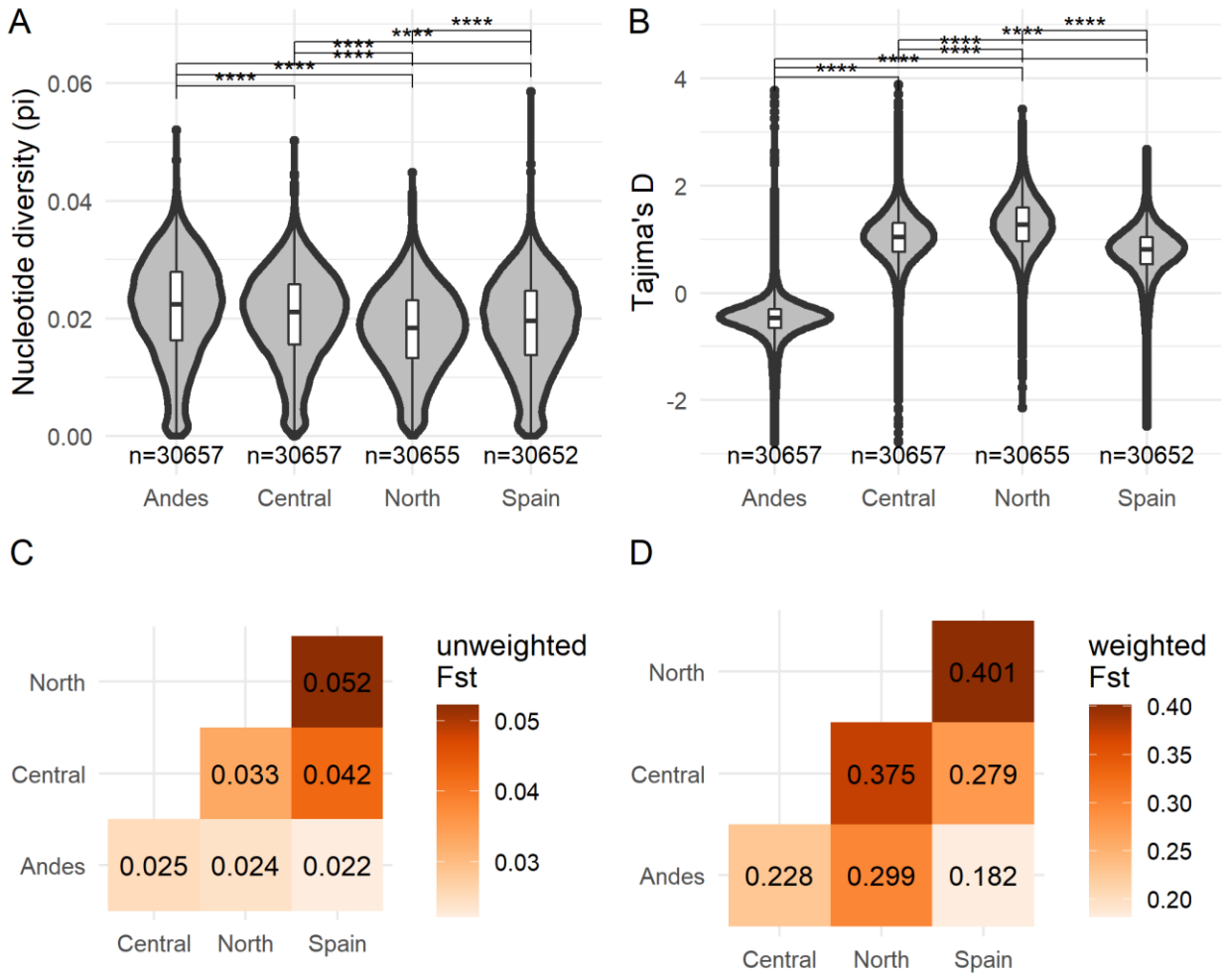


Figure S14: Three population models used to estimate parameter values. (A) M1: Model with two population splits and constant population size. (B) M2: Model with exponential growth. (C) M3: Model with population resizing events instead of exponential growth.

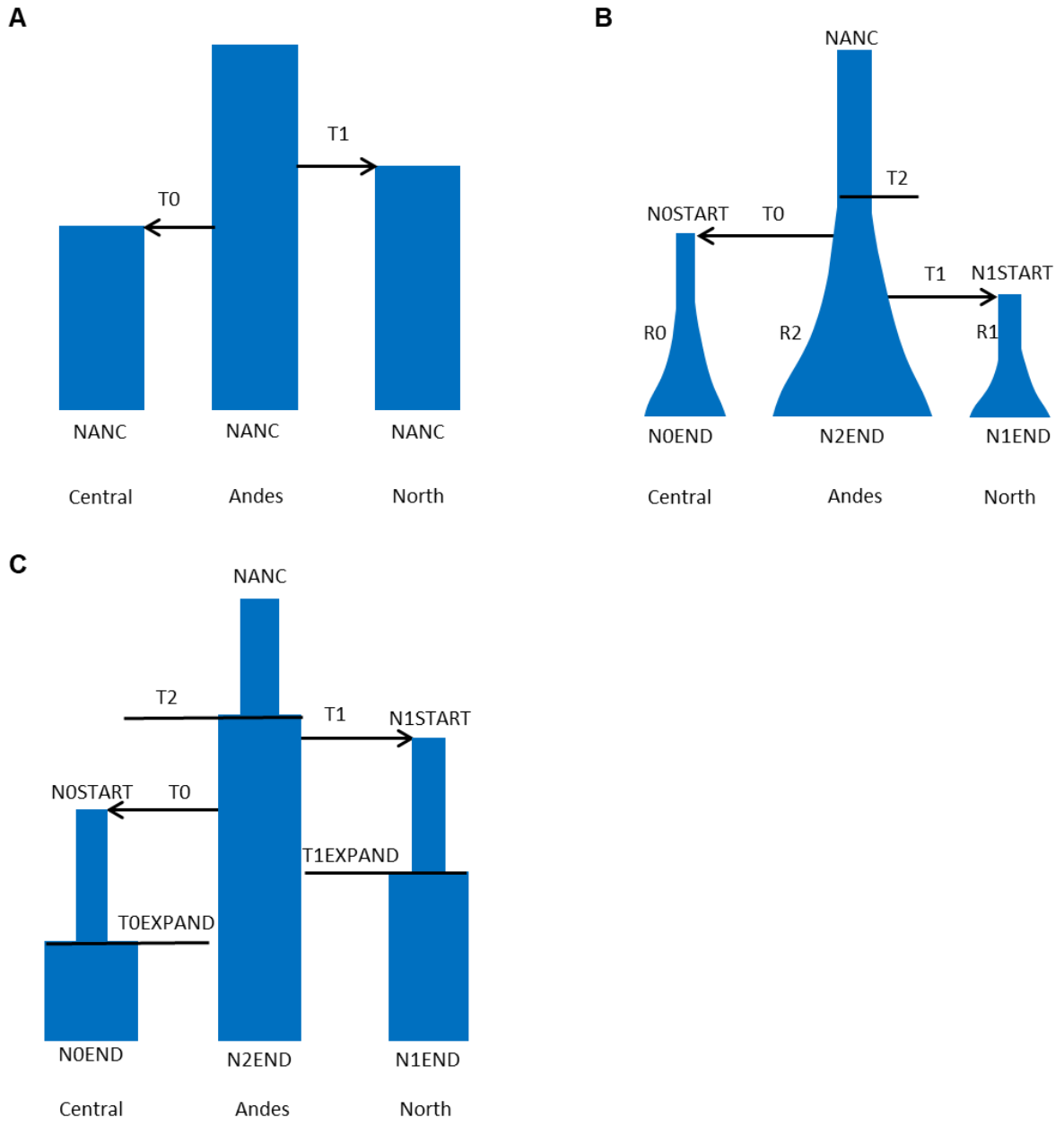


Figure S15: Linkage disequilibrium decay rates, minimums, and maximums over a 10kb interval calculated from genotype likelihoods or from 100 independent simulations under three different population history models. Values estimated from the data are shown by the red points with 95% confidence intervals. Pairwise t-tests with a Holm's correction method were used to compare means. **** indicates p-value < 0.0001.

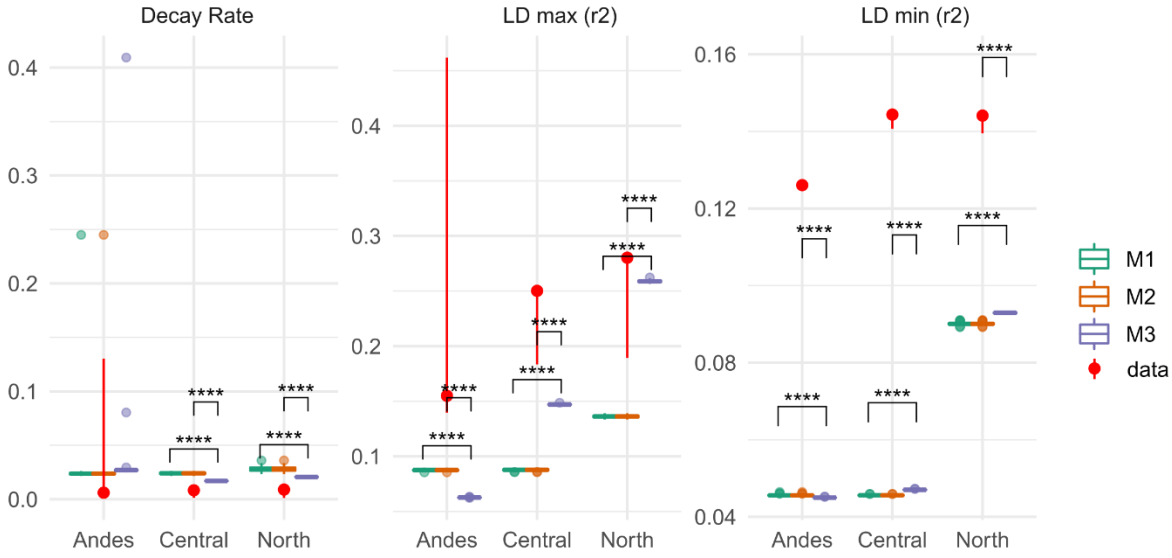
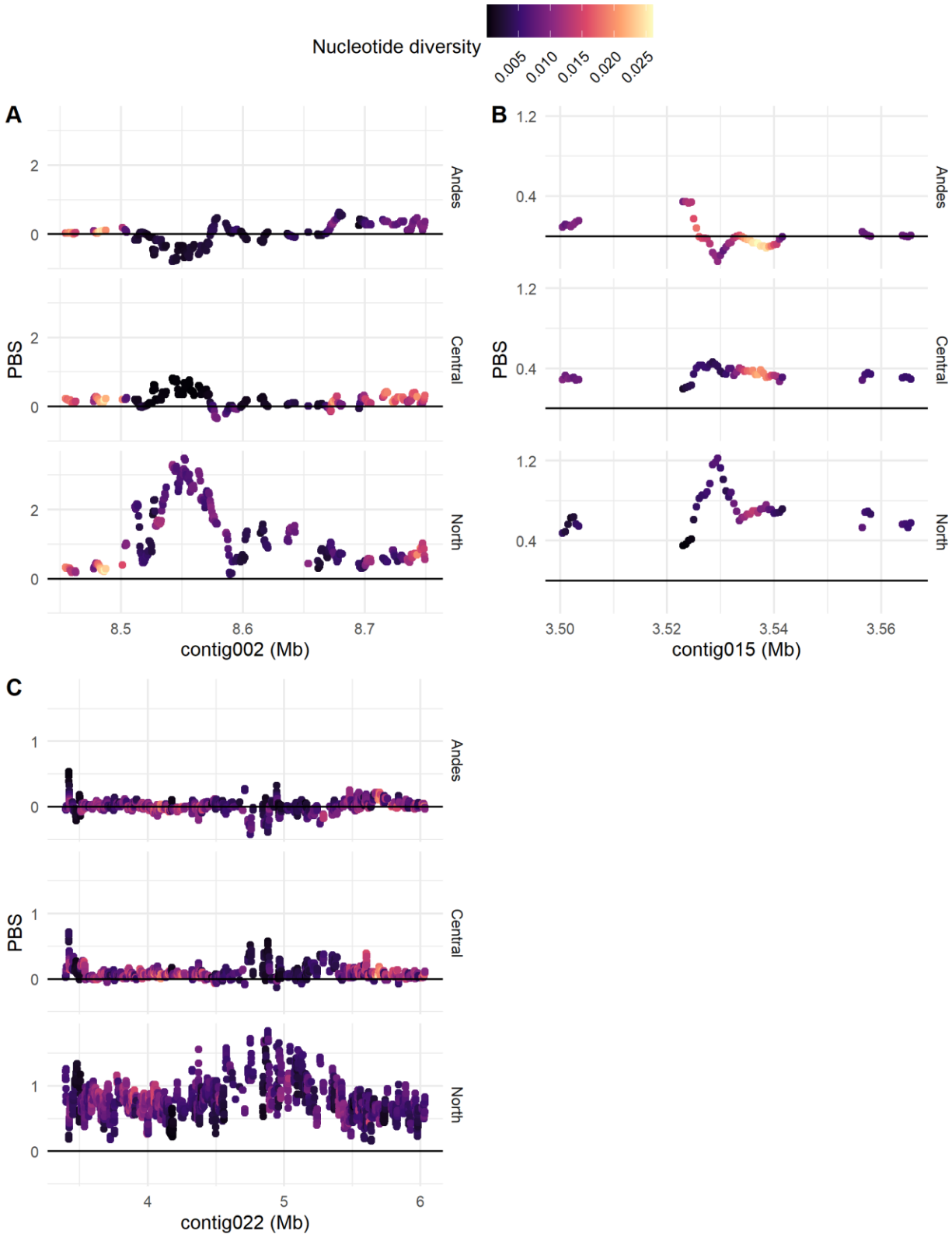


Figure S16: PBS and nucleotide diversity at three other PBS hotspots, with values averaged over 5kb intervals in 500bp steps.


Nucleotide diversity
0.005
0.010
0.015
0.020
0.025
A
B
C
PBS
Andes
Central
North
contig002 (Mb)
contig015 (Mb)
contig022 (Mb)

Supplemental Tables

Table S1: Maximum likelihood parameter estimates under the three population history models tested.

See Figure S14 for model details.

parameter	unit	M1	M2	M3
NANC	effective pop size	3324287	3943689	3433648
N0START	effective pop size	-	187146	16394
N1START	effective pop size	-	121328	84783
N0END	effective pop size	-	2282860	1237043
N1END	effective pop size	-	2659059	857884
N2END	effective pop size	-	9698709	8346208
T0	generations ago	523870	159990	187034
T1	generations ago	1028654	239766	252383
T2	generations ago	-	1082411	1134632
T0EXPAND	generations ago	-	-	184870
T1EXPAND	generations ago	-	-	230046
R0	exponential growth rate	-	-1.56E-05	-
R1	exponential growth rate	-	-1.29E-05	-
R2	exponential growth rate	-	-8.31E-07	-

Table S2: Allele frequency of mutations in either copy of ACHE1. Both A201S and F290V have been reported to confer organophosphate resistance

	g8300			g9292		
mutation	North	Central	Andes	North	Central	Andes
A201S	53%	0%	7%	-	51%	71%
F290V	38%	0%	7%	-	44%	20%
F290N	0%	27%	5%	-	0%	0%

Chapter 3: Probe-based quantitative PCR and RPA-Cas12a molecular diagnostic to detect the tomato pest *Tuta absoluta*

Kyle M. Lewald[1], Wenqi Song[1], Daniel Eweis-LaBolle[1], Cindy Truong[1], Kristine E. Godfrey[2], Joanna C. Chiu[1*]

[1]Department of Entomology and Nematology, College of Agricultural and Environmental Sciences, University of California Davis, One Shields Ave, Davis, CA 95616, USA.

[2]Contained Research Facility, University of California, Davis, 555 Hopkins Rd, Davis, CA 95616, USA

***Corresponding author**: Joanna C. Chiu, Email: jcchiu@ucdavis.edu.

Contributions and Acknowledgements

KML: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing - original draft; Writing - review & editing. KEG: Conceptualization; Resources; Writing – review & editing. JCC: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing - review & editing. WS: Investigation; Methodology; Writing - original draft; Writing - review & editing. DEL: Investigation; Methodology; Writing – original draft; Writing – review & editing. CT: Investigation; Methodology; Writing - review & editing.

This work is funded by NIFA 2020-67013-30976 to JCC. We thank Dr. Evan Braswell at USDA APHIS for providing eDNA samples. Schematic diagrams for methods were created with BioRender.com (license issued to UC Davis Chiu lab).

Abstract

The tomato pest *Tuta absoluta* Meyrick is highly invasive but has not yet invaded North America. However, several morphologically similar species are already present, making detection of *T. absoluta* presence and invasion challenging. We designed a quantitative PCR molecular diagnostic to differentiate *T. absoluta, Phthorimaea operculella* (Zeller)*, or Keiferia lycopersicella* (Walsingham) (Lepidoptera: Gelechiidae) DNA. Additionally, we developed an RPA-Cas12a molecular diagnostic that allows for the isothermal detection of *T. absoluta* DNA, eliminating the need for a thermocycler. These results can be visualized simply using a UV light source and cell phone camera. We expect these diagnostics to improve quarantine and prevention measures against this serious agricultural threat.

Introduction

The tomato plant (*Solanum lycopersicum* L.) represents a massive economic industry worldwide, with an estimated 252 million metric tons of tomatoes harvested in 2020 (FAOSTAT 2020). This production is concentrated in a few major producing countries, with 70% of the world's production currently accounted for by China, the European Union, India, the USA, and Turkey (Costa and Heuvelink 2007). However, this industry is quickly becoming threatened by the invasive gelechiid moth *Tuta absoluta* Meyrick, commonly known as the tomato leafminer. *T. absoluta* (also known as *Phthorimaea absoluta* (Chang and Metz 2021)) is an agricultural pest of the nightshade family, including peppers, eggplants, and potatoes, but primarily represents a serious threat to tomatoes. Left untreated, the larvae will consume leaf tissue, bore into flower buds, or burrow into fruit with infestations causing crop losses as high as 80 to 100% (Desneux *et al.* 2010; Biondi *et al.* 2018).

While *T. absoluta* was first identified in Peru in 1917 (Meyrick 1917), it was not considered an agricultural pest until the 1960s when *T. absoluta* was detected in farms in Argentina, causing significant crop loss (Bahamondes and Mallea 1969). From there it rapidly spread throughout South America, causing severe agricultural losses everywhere it was found. In 2006, *T. absoluta* was identified in a greenhouse in Spain, marking the first intercontinental migration of the pest, likely due to anthropological transportation (EPPO 2008). From Spain, it rapidly spread throughout Europe, Sub-Saharan Africa, the Middle East, and most recently into Asia (Bloem and Spaltenstein 2011; Sridhar *et al.* 2014; Hossain *et al.* 2016; Zhang *et al.* 2020). Without rapid identification and quarantine strategies, *T. absoluta* will likely continue to spread to North America.

Preventing initial establishment of an invasive species in a risk ecosystem can be an effective strategy but requires vigilant monitoring of imported products as well as an ability to rapidly detect the presence of the target (Tobin *et al.* 2014). In the United States, policies such as federal orders issued by the Animal and Plant Health Inspection Service have been used to prevent accidental introduction of *T.*

absoluta by regulating import of tomato fruit or propagation material from infested countries (Osama El-Lissy 2019). Correctly detecting and identifying the insect, however, remains challenging due to the presence of other gelechiid moths such as *P. operculella* (Zeller) and *Keiferia lycopersicella* (Walsingham) in the United States, which look nearly identical to *T. absoluta* and are also commonly found on tomato and potato crops (Gilboa and Podoler 1995). Distinguishing these species morphologically requires dissection of adult male genitalia, which requires entomological expertise and increases waiting times for identification. Molecular diagnostics, on the other hand, would allow for accurate and rapid identification without entomological expertise. To date, several molecular diagnostics have been developed, all using Polymerase Chain Reaction (PCR) to amplify target DNA presence before detecting signal either by fluorescence, gel electrophoresis, or mass spectrometry (Sint *et al.* 2016; Tabuloc *et al.* 2019; Zink *et al.* 2020). A potential limitation of these PCR-based diagnostics is the requirement for a thermocycler, as well as specialized equipment to visualize results. For field detection, a system that requires minimal specialized equipment would be ideal.

The CRISPR-Cas (Clustered Regularly-Interspaced Short Palindromic Repeats and CRISPR-ASociated proteins) system has recently become an attractive option for molecular diagnostics (Chen *et al.* 2018; Aman *et al.* 2020). CRISPR-Cas originated as a bacterial immunity system, in which the Cas protein complexes with a CRISPR RNA (crRNA) that is complementary to a target DNA sequence (Zetsche *et al.* 2015). When the Cas-crRNA complex binds to its target it cleaves it, causing a double-strand break in the DNA. While Cas9 is the most commonly used Cas protein for its genomic editing capabilities, the Cas12a enzyme was discovered to exhibit indiscriminate single-stranded DNA nuclease (ssDNAse) activity upon binding to its target DNA (Chen *et al.* 2018). If a single-strand oligonucleotide probe modified with a fluorophore and quencher is present when Cas12a-crRNA binds its DNA target, Cas12a will proceed to also cleave the single-strand probe, separating the fluorophore from the quencher and allowing a fluorescent signal to be measured. As ssDNAse activity is proportional to the number of target DNA

molecules and can occur isothermally between 15℃ to 50℃, using an isothermal DNA amplification method such as Recombinase Polymerase Amplification (RPA) to amplify target DNA allows for sensitive detection with Cas12a at constant temperature, eliminating the need for an expensive thermocycler. Additionally, detection of probe cleavage is flexible as several methods have been published on alternative visualization methods, including using flow strip assays, gold-conjugated DNA probes, and naked-eye fluorescence detection (Li *et al.* 2019; Wang *et al.* 2020; Yuan *et al.* 2020).

In this study, we leveraged prior sequencing data (Tabuloc *et al.* 2019) to develop two molecular diagnostic assays. We created a multiplexed probe-based quantitative PCR (qPCR) assay able to identify whether a sample contains *T. absoluta, K. lycopersicella,* or *P. operculella* DNA based on single-nucleotide polymorphisms (SNPs) between species. In addition, we created RPA-Cas12a assays to detect the presence of *T. absoluta* DNA, using either a fluorescent reader or a simple UV illuminator paired with a cell phone camera. This will allow for more rapid field detection of *T. absoluta*.

Methods

Sample collection

T. absoluta, *K. lycopersicella*, and *P. operculella* DNA samples were obtained from Tabuloc et al. 2019. Briefly, DNA was extracted from laboratory-reared *T. absoluta* adults and larvae collected in 95% ethanol. DNA was extracted from *K. lycopersicella* adults and larvae that were collected on dry ice and stored in -80°C from colonies maintained at the UC Davis Contained Research Facility that were originally from a Florida colony. DNA was extracted from *P. operculella* adults and larvae collected on dry ice and stored in -80°C from colonies maintained at the UC Davis Contained Research Facility that were originally collected from Kern County, California. Environmental DNA (eDNA) samples containing either local Florida bycatch spiked with *T. absoluta* DNA, as well as mock community environmental DNA (eDNA) samples were provided as a gift from Craig Bateman (Florida Museum of Natural History). DNA extractions from individual insects were performed as described in Tabuloc et al. 2019.

Quantitative PCR assay design

We assessed 21 single nucleotide polymorphism (SNP) markers differentiating *T. absoluta*, *K. lycopersicella* and *P. operculella* identified in Tabuloc et al. 2019 for qPCR genotyping potential using RealTimeDesign software (LGC Biosearch Technologies). Our analysis resulted in seven potential qPCR assays, each containing a set of universal amplifying primers and two distinguishing fluorescent oligonucleotide probes. We used a serial dilution qPCR (from 1 to 1/525 dilution) on all seven sets of amplifying primers on all three species to assess primer efficiency. Reactions were performed with 300nM of forward and reverse primers with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, Hercules, CA) on a CFX96 real time PCR machine (Bio-Rad). We checked for off-target products when multiplexing pairs of assay primers using melt curve analysis as well as gel electrophoresis. From the seven loci tested, we selected two to develop assays and ordered single-strand BHQplus probes (LGC Biosearch Technologies, Middlesex, UK) unique to each of the four alleles (two alleles per locus). Each probe was tagged with a distinct 5' fluorophore and 3’ quencher. Assay 1 distinguishes *T. absoluta* from *K. lycopersicella* and *P. operculella*, while assay 2 distinguishes *P. operculella* from *K. lycopersicella* and *T. absoluta*.

Quantitative PCR assay

All subsequent multiplexed qPCR genotyping assays were conducted on the CFX96 in technical triplicate using qPCRBIO Probe Mix No-ROX (PCR Biosystems; Wayne, PA), 200nM of each probe, 400nM forward and reverse primers and 1 uL of extracted DNA in a final volume of 20uL in the following PCR conditions; 2 minutes at 95°C, followed by 40 cycles of 95°C for 5 seconds and 63°C for 30 seconds. Relative fluorescent units (RFU) for FAM, CAL Gold 540, CAL Red 610, and Quasar 670 channels were recorded after each cycle for all samples using the CFX Maestro software (Bio-Rad), and end-point RFU was used to make SNP calls. DNA concentrations for the purpose of dilution testing were quantified with a Nanodrop Lite (Thermo Fisher Scientific, Pleasanton, CA, USA).

Lb-Cas12a protein purification

The plasmid pMBP-LbCas12a was a gift from Jennifer Doudna (Addgene plasmid # 113431 ; http://n2t.net/addgene:113431 ; RRID:Addgene_113431) (Chen *et al.* 2018). We chemically transformed pMBP-LbCas12a into BL21 *E. coli* cells and cultured them in Lysogeny Broth with 125 µg/mL ampicillin in a 30°C 225 RPM shaker to an OD600 of 0.7 before adding 0.2mM IPTG and shaking at 16°C overnight to induce protein expression. Resulting bacterial culture was pelleted for 30 minutes at 2600 rcf in a 4°C centrifuge, resuspended in lysis buffer (50mM Tris-HCl ph 7.5, 500mM NaCl, 1mM DTT, 0.5mM PMSF, 5% glycerol), lysed with a Sonic Dismembrator sonicator (Thermo Fisher), and passed through a 0.45µm filter. We applied the histidine-tagged Cas12a to a nickel column on an NGC chromatography system (Bio-Rad; Hercules, CA), washed with a wash buffer (lysis buffer + 20mM imidazole) and eluted with elution buffer (lysis buffer + 250mM imidazole). We inspected fractions on an SDS-PAGE 12% gel, then pooled and incubated fractions overnight at 4°C with TEV protease (1mg TEV to 10mg Cas12a) to cleave off the histidine and maltose binding protein tag. Cas12a was buffer-exchanged back into wash buffer using a desalting column on the NGC and re-eluted through a nickel column to remove impurities non-specifically binding to the column. Resulting fractions were inspected by SDS-PAGE, pooled, quantified with a Nanodrop Lite (ThermoFisher Scientific; Waltham, MA), flash frozen in liquid nitrogen, and stored at -80°C. Cas12a enzyme was tested for activity using a crispr RNA directed against M13 bacteriophage DNA as described in Chen et al. 2018.

RPA-Cas12a assay design

To identify loci distinguishing *T. absoluta* from *K. lycopersicella* and *P. operculella* for a Cas12a-based diagnostic, we used a custom python script to search a genome alignment file between all three species (obtained from Tabuloc et al. 2019) for loci with a "TTTV" PAM site in the *T. absoluta* sequence and at least 2 mismatches in the first three bases following the PAM site between *T. absoluta* and the other two species (Script S1). We used BLAST (Camacho *et al.* 2009) with the "nt" database to exclude any sequences that might be contaminants, such as Wolbachia DNA. We selected 10 loci to design crispr

RNAs (crRNAs) as well as flanking sets of forward and reverse amplification primers designed to produce an approximately 200bp product. We used the TwistAmp Basic Liquid kit (TwistDx, Maidenhead, UK) to perform isothermal amplification, following manufacturer's recommendations but in 25 µL (not 50uL) reactions. We used a final concentration of 1.8mM dNTPs and 1.8mM of the forward and reverse primers, respectively. Reactions were incubated at 37°C for 20 minutes before being quenched with a DNA loading dye containing EDTA to quench the reaction prior to gel electrophoresis. To select the best amplification primers for each locus, we tested one forward primer with three possible reverse primers and checked the amplification products on an agarose gel. We then used the reverse primer from the best reaction to test against three possible forward primers, and again used an agarose gel to determine the optimal forward and reverse primer.

RPA-Cas12a assay

To perform the RPA-Cas12a assay, we performed RPA as described above using 1 µL of DNA per sample. Importantly, we add all reagents into the bottom of the PCR strip tubes except the magnesium acetate, which is instead added to the strip tube caps. Once all samples are ready, we close the caps and gently mix to incorporate the magnesium acetate to ensure amplification initiates simultaneously for all samples. At all steps, we used sterilized filter pipette tips to avoid contamination from DNA or RNA nucleases.

To perform the Cas12a detection, 2 µL of RPA product was added to a final volume of 20 µL containing 20 mM Tris-HCl (pH 7.5), 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 1 mM DTT, 300 nM Cas12a enzyme, 300 nM of crRNA 200 nM of oligonucleotide reporter. As the exact sequence of the oligonucleotide reporter is not critical since Cas12a's ssDNAse activity is non-specific, we used the Cal Red 610 probe previously designed for our qPCR assay. We incubated the reactions at 37°C using a Bio-Rad CFX96 Real Time PCR machine for 90 minutes and recorded fluorescence every minute. Final endpoint RFU was used to

identify presence of *T. absoluta*. DNA concentrations for the purpose of dilution testing were quantified with a Qubit fluorometer (Thermo Fisher Scientific).

For visual detection, samples were placed in PCR strip tubes after the 90-minute incubation and illuminated in a gel imaging box using 302nm UV light. Photos were taken through the viewing window using a Google Pixel 5A phone using default camera settings.

Results

Probe-based qPCR assay accurately differentiates *T. absoluta, K. lycopersicella* and *P. operculella*

To develop an assay that can identify whether a sample contains DNA from *T. absoluta*, *K. lycopersicella*, or *P. operculella*, we screened through previously identified SNPs (Tabuloc *et al.* 2019) that distinguish the three species based on a whole genome alignment. We selected two loci and designed primers to amplify each locus from all three species, as well as dual-labeled fluorophore-quencher oligonucleotide probes specific to each allele at each locus (Figure 10A, Table S3). Assay 1 targets a SNP in the coding region of *eif-4a*, a canonical translation initiation factor conserved across eukaryotes. A high FAM signal indicates the *T. absoluta* variant, while a high CAL Gold 540 signal indicates the *K. lycopersicella* or *P. operculella* variant. Assay 2 targets a SNP in the coding region of *toll-6*, a gene involved in neuron development during embryogenesis. A high Quasar 670 signal indicates the *P. operculella* variant, while a high CAL Red 610 signal indicates the *K. lycopersicella* or *T. absoluta* variant. Based on the presence or absence of each fluorophore's signal at the end of the qPCR amplification, it is possible to deduce the species identity (Figure 10A). Both assays can be run multiplexed in a single reaction and takes approximately three hours from DNA extraction to results, allowing for rapid diagnostics.

We tested both assays using DNA extracted from individuals obtained from lab-maintained colonies. While the target qPCR loci were checked for conservation in eight individuals of each species in Tabuloc et al. 2019, we additionally tested *T. absoluta* samples obtained from greenhouses and fields in Costa

Rica, Brazil, Columbia, Ecuador, Paraguay, Argentina, Chile, and Peru, to ensure no genomic variations existed in field populations. In assay 1, we see strong FAM signal and low Cal Gold 540 signal in all *T. absoluta* samples, and the reverse in all other samples (Figure 10B). Similarly, in assay 2 we see strong Quasar 670 signal and low Cal Red 610 signal in *P. operculella* samples, and the reverse in all other samples except one *P. operculella* sample that was misidentified as *K. lycopersicella* in the assay (Figure 10C). We noticed some samples with lower signal overall in both channels of assay 2; these samples all came from the same PCR plates, and likely represent an experimental batch artefact. By plotting the Cal Gold 540 signal from assay 1 against the Cal Red 610 signal, it is possible to rapidly discriminate all three species in one graph. To ease categorization, allelic discrimination software such as that included in CFX Maestro (Bio-Rad) can be used to classify samples automatically.

Quantitative PCR assays are highly sensitive at low DNA concentrations

While the previous tests were conducted using DNA extracted from individual moths, it would be advantageous for a diagnostic assay that could detect *T. absoluta* presence in pooled collections where target DNA concentrations may be extremely low. To determine the lower bounds of starting genomic DNA needed to positively identify *T. absoluta*, we tested our qPCR assays on a serial dilution of genomic DNA extracted from a single sample of *T. absoluta, K. lycopersicella*, and *P. operculella*, and plotted endpoint RFU for each fluorophore (Figure 11). For all fluorescent probes as little as 0.02 ng of DNA was needed to identify the target DNA.

Often, samples collected from the field are not preserved well, and may be degraded and morphologically indistinguishable from other samples, such as with a liquid-based traps. To determine whether our qPCR assays would be suitable for detecting *T. absoluta* in this environmental DNA (eDNA) collection, we tested the assays on several mock environmental samples meant to mimic collections from bulk field collections. We tested both pooling DNA from multiple species into one sample, as well as spiking DNA extracted from trap bycatch with *T. absoluta* DNA (Figure 12). As a negative control we

used a mixture of DNA of non-*T. absoluta* DNA, including *P. operculella* and *K. lycopersicella.* In assay 1, all samples with *T. absoluta* DNA had higher FAM signal than the negative control, as expected (Figure 12A). The high Cal Gold 540 signal in the negative control indicates that the *K. lycopersicella* and *P. operculella* DNA was detected. In assay 2, all samples containing *T. absoluta* DNA except one of the bycatch samples showed the expected high Cal Red 610 levels (Figure 12B). The *T. absoluta*-negative sample showed elevated signal in both Cal Red 610 and Quasar 670 resulting from the presence of *K. lycopersicella* and *P. operculella* DNA, as expected. Based on this, our assay is generally sensitive to *T. absoluta* DNA in the presence of contaminating samples, but the low concentrations of target DNA that may be present in liquid trap collections may make detection challenging.

RPA-Cas12a assay detects *Tuta absoluta* without need for a thermocycler or specialized detection equipment

While the qPCR assays are sensitive and accurate, they require a real-time PCR machine, which may be cost-prohibitive. To avoid requiring specialized equipment, we used RPA to amplify target DNA and Cas12a's single-strand DNase activity to detect the target sequence. As both RPA and Cas12a can operate at a constant temperature, this removes the need for a thermocycler (Figure 13A). As the Cas12a complex can bind to targets with a few mismatches (Chen *et al.* 2018), we tested a dozen loci with at least two mismatches immediately adjacent to the PAM site that would differentiate *T. absoluta* from *K. lycopersicella* and *P. operculella.* After testing primers for RPA, we selected two nuclear markers and one mitochondrial marker. Nuclear marker 1 targets the exon of an uncharacterized gene that is predicted to be localized to the Golgi while nuclear marker 2 targets the intron of *bcr*, a predicted GTPase-activating protein with unknown function. The mitochondrial marker was designed from Cytochrome Oxidase I (CO1), a common gene used for species identification. We performed Cas12a-based detection for each locus using a crRNA designed to only recognize the *T. absoluta* sequence colony samples originally from Spain and found that while nuclear marker 1 and COI marker properly distinguished all *T. absoluta* samples, nuclear marker 2 failed to identify 7 of the 28 Spanish *T. absoluta*

colony samples tested (Figure 13B-C). Based on gel electrophoresis and Sanger sequencing, 5 samples failed to amplify while the other two contained a 9 base pair insertion at the crRNA binding site (data not shown). We tested nuclear marker 1 and the COI marker with the same Latin American *T. absoluta* samples tested in the qPCR assays and found a positive result for all samples. Importantly, we recorded no false positives across all *P. operculella* and *K. lycopersicella* samples tested.

To assess the sensitivity of the RPA-Cas12a assays for nuclear marker 1 and the COI marker, we tested a dilution series of *T. absoluta* DNA from 1 to $1x10^{-5}$ ng input, comparing it to a 1 ng input of *K. lycopersicella* or *P. operculella* DNA. In both markers there was a significant difference in signal between *K. lycopersicella* and the 1 and 0.1 ng of *T. absoluta* DNA, but not at lower concentrations or with *P. operculella* (Figure 14A-B). While these RFU measurements were quantified using a real-time PCR machine, we wanted to see at what concentrations detection could be achieved by eye. We placed the same samples analyzed for the nuclear marker 1 dilution series under a gel imager box and using a cell phone camera were able to positively identify the 1 ng and 0.1 ng *T. absoluta* samples, with no signal from other dilutions or negative controls (Figure 14C).

Discussion

When considering which method is best suited for *T. absoluta* identification, a key consideration is sensitivity. Based on our analyses, we find that our qPCR diagnostics can reliably identify species down to 0.02 ng of input DNA, while our RPA-Cas12a assay was capable down to 0.1 ng input DNA. Assuming a haploid genome size of 564Mb from flow cytometry estimates (Gandhi Gracy *et al.* 2019), this represents approximately 35 copies of target DNA for qPCR detection and approximately 170 copies of target DNA for RPA-Cas12a detection. Both should be sufficient for scenarios where DNA is extracted from a single insect or small pool of insects. However, in situations where target DNA concentrations are extremely low, such as the case would be with eDNA collections, the more sensitive qPCR method is preferred.

Between the two working markers for the RPA-Cas12a study, both the COI and nuclear marker 1 were equally sensitive. By eye, the COI marker appeared to detect *T. absoluta* signal at lower concentrations; however, the increased variability at lower concentrations made the results significantly indistinguishable from non-target samples. This noise may appear because RPA-Cas12a assays are more sensitive to technical experimental variation, as both RPA and Cas12a reactions can occur at room temperature. Slight changes in incubation time can cause fluctuations in RFU values between replicates, highlighting the importance of having positive and negative controls. Additionally, as crRNA is sensitive to ribonuclease degradation, greater care is needed to ensure its degradation, such as using filter pipette tips, practicing proper lab hygiene, and using nuclease-free reagents.

A few methods have been published using multiplexed PCR, qPCR, and digital droplet PCR to detect *T. absoluta* (Sint *et al.* 2016; Zink *et al.* 2020). We find our qPCR assay is equally sensitive as the previous qPCR method, with the added advantage of being able to identity an additional two moth species which are morphologically nearly identical to *T. absoluta*, occupy similar host crops, and are already established in the United States. The digital droplet PCR remains the most sensitive assay so far but requires specialized equipment that may not be widely adopted currently. An alternative method using multiplexed PCR combined with mass spectrometry has also been previously developed, allowing the simultaneous interrogation of 21 SNPs in a single reaction to differentiate *T. absoluta*, *P. operculella*, and *K. lycopersicella* (Tabuloc *et al.* 2019). This assay has the advantage of being robust to possible population variation at these loci but again requires specialized equipment and training to operate.

One of the main advantages of our RPA-Cas12a assay is its reduced need for specialized equipment. Beyond freezers to store reagents, the main equipment needed is a temperature-controlled heat block and a fluorescent imaging device equipped with a light or laser of appropriate wavelength. This could include something as complex as a qPCR machine or plate reader, but we found that using a simple UV illuminator in combination with a smartphone camera was equally effective for detecting a positive

signal in the RPA-Cas12a assay as measuring fluorescence with a qPCR machine. Compared to the qPCR diagnostic, the RPA-Cas12a diagnostic does require additional user time, as amplification and detection occur in two distinct steps, although a one-pot reaction variant has been described (Wang *et al.* 2020). While we used an oligonucleotide with a Cal Red 610 fluorophore paired with the appropriate quencher as the substrate for Cas12a-mediated ssDNAse activity, the fluorophore/quencher molecules can easily be swapped for other fluorophores that absorb and emit at different wavelengths, making this protocol adaptable to the equipment available to the user. It is possible in the future to replace the fluorometric detection method with a colorimetric detection by using biotin and FAM-labeled oligonucleotides applied to a lateral flow assay. This flexibility has been implemented in other detection assays and lets Cas12a-based detection systems stand out from qPCR methods, despite the slight reduction in sensitivity (Soh *et al.* 2022).

While *T. absoluta* has not yet been detected in North America, demographic modeling suggests low to moderate levels of invasion are possible in Mexico, the California Central Valley, and the southeastern USA (Ponti *et al.* 2021). Rapid reliable detection of an invasive pest is one of the key components of a successful eradication program, meaning it is crucial that molecular diagnostics are available (Tobin *et al.* 2014). Even if eradication is not possible, molecular diagnostics reduce the need for expert entomologists to hand identify specimens, allowing detection to be done faster, cheaper, and more accurately (Stouthamer *et al.* 1999; Gariepy *et al.* 2007). This is especially relevant for *T. absoluta* testing as tomato crops in the United States already contain the presence of *P. operculella* and *K. lycopersicella*, which are both nearly identical morphologically. Once detected, agencies can decide between strategies including eradication, quarantine, or continued monitoring (Venette *et al.* 2021). We expect that the molecular diagnostics presented here will add to the toolkit available to institutions to rapidly monitor for the appearance of *T. absoluta*.

Figures

Figure 10: A) Flowchart of the qPCR diagnostic assay, with a table detailing the species identification based on signals from both assays. B-D) Endpoint relative fluorescence units from two multiplexed assays. All samples run in technical duplicates. Assay 1 (B) uses a probe tagged with FAM to detect the *T. absoluta* SNP and a probe tagged with Cal Gold 540 to detect the *K. lycopersicella/P. operculella* SNP. Assay 2 (C) uses the Quasar 670-tagged probe to detect *P. operculella* and Cal Red 610-tagged probe to detect *T. absoluta/K. lycopersicella*. By plotting the Cal Gold 540 and Cal Red 610 channels on 1 plot (D), all three species can be differentiated quickly.

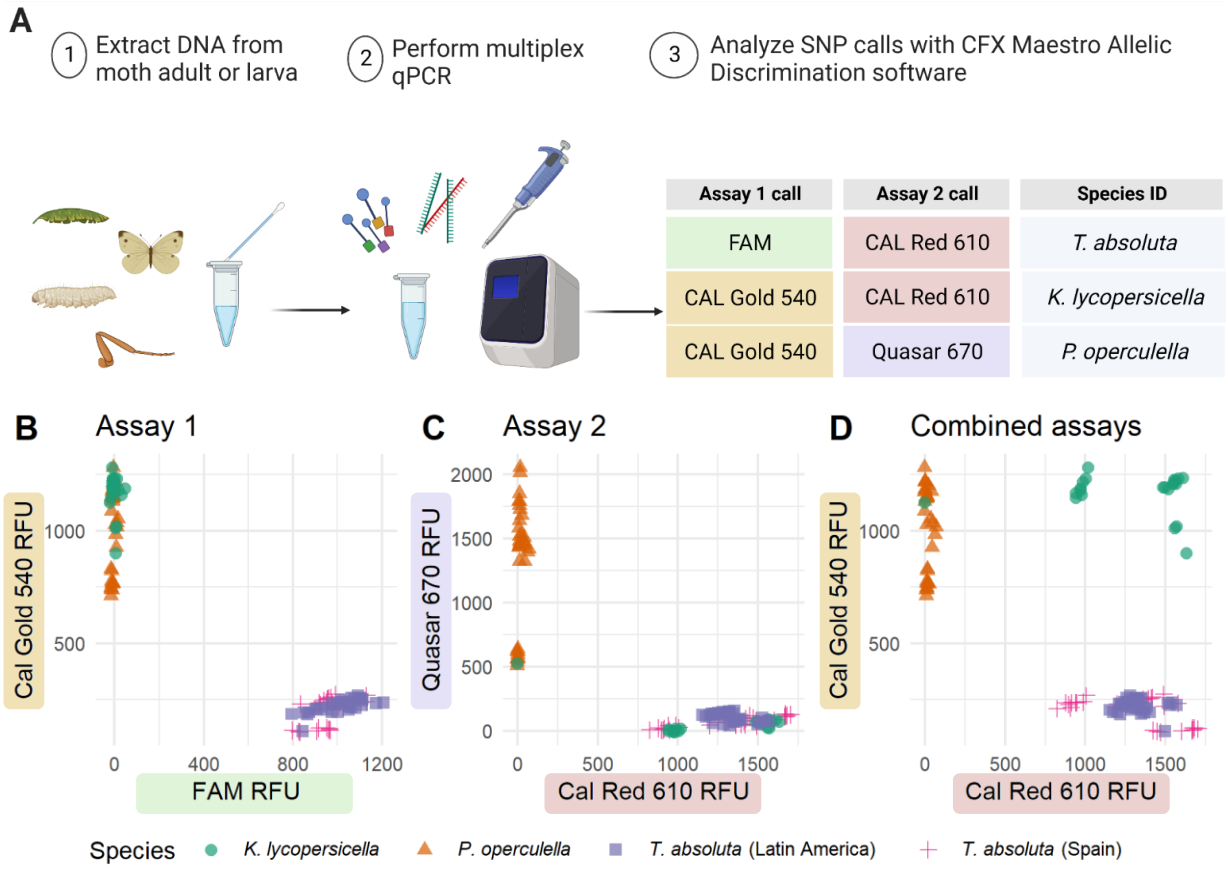


Figure 11: Dilution series of qPCR assays. A-D) qPCR end RFU signal for each probe against species at various dilutions from 20 ng to $2x10^{-3}$ ng total DNA. All samples were run in technical triplicate. Error bars represent 1 SE. Black line indicates RFU in the no-template control (NTC).

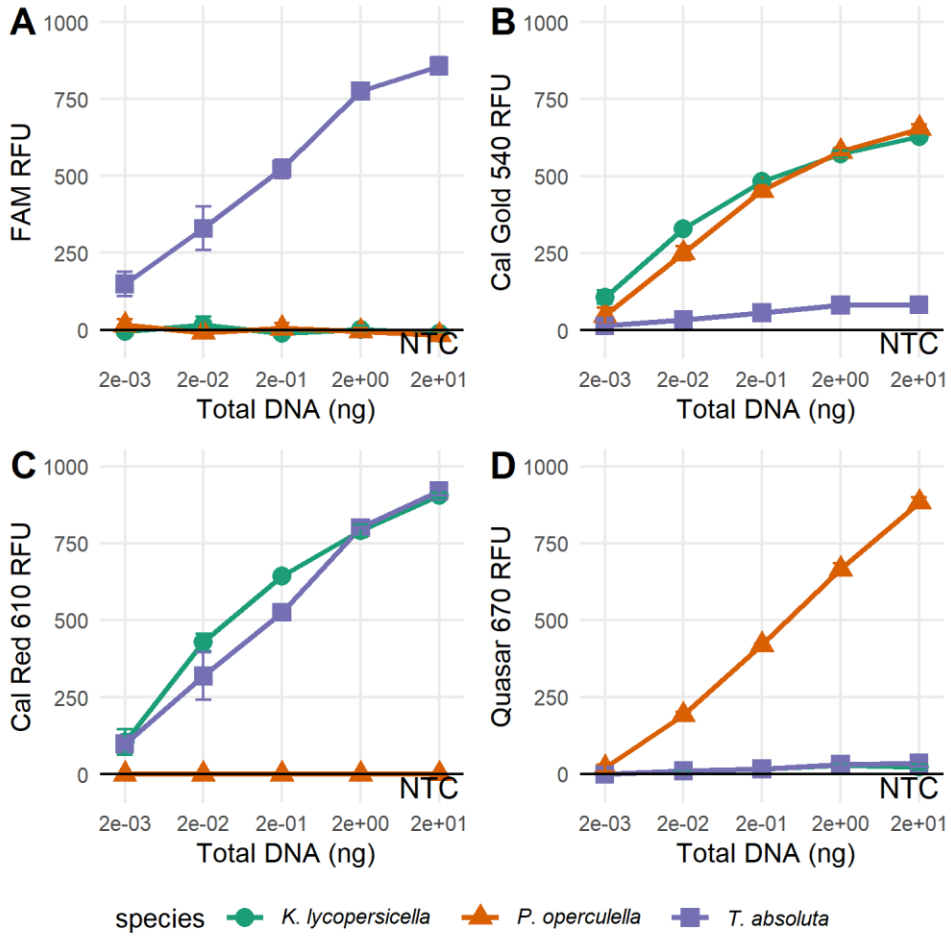


Figure 12: Species identification of mock environmental DNA samples using the qPCR assay 1 (A) and assay 2 (B). Positive control was 1 ng of *T. absoluta* DNA. Negative control was a mixture of DNA from *P. operculella*, and *K. lycopersicella*. The mock community sample consisted of DNA extracted from a single adult of: *T. absoluta*, *Plodia interpunctella*, *Anastrepha ludens, Diaphorina citri,* and *Tribolium confusum*. The bycatch samples were DNA extracted from local insect bycatch in Florida, with an artificially added *T. absoluta* adult. Numbers labeling bycatch datapoints correspond to matched samples between assay 1 and assay 2.

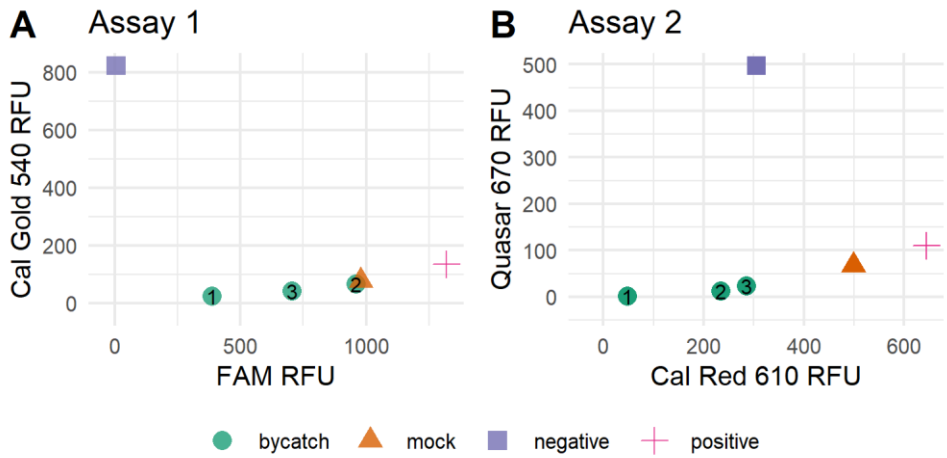


Figure 13: (A) Flowchart of the RPA-Cas12a diagnostic assay. (B) RPA-Cas12a assay results of two nuclear loci (nuclear marker 1 and nuclear marker 2) and one mitochondrial locus (COI marker). Nuclear marker 2 was not tested on Latin American samples as it was not found to reliably identify *T. absoluta* initially with Spanish colony samples. RFU values of each sample were normalized across plates by subtracting the no-template control (NTC) RFU per plate. Mean RFU values were compared using pairwise 2-tailed student t-tests with the holm's correction. **** indicates corrected p-value < $1x10^{-4}$.

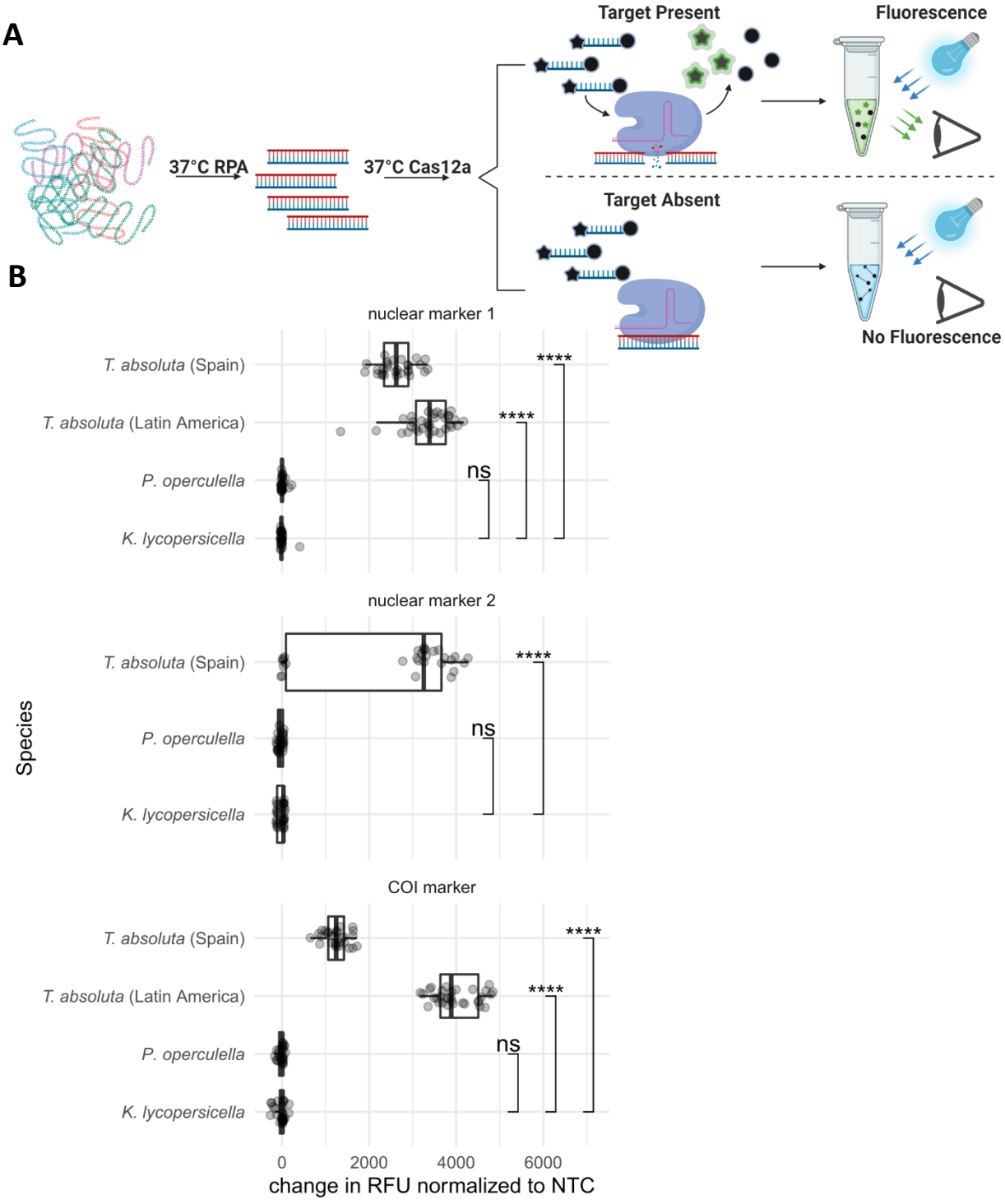
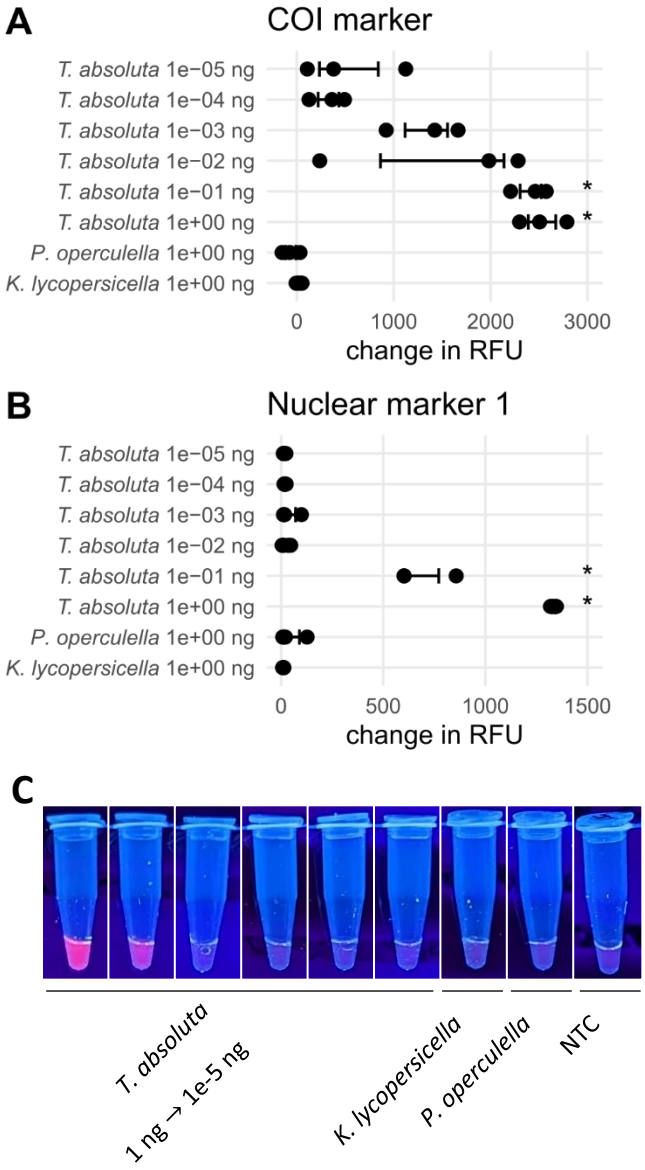

A
37°C RPA
37°C Cas12a
Target Present
Fluorescence
Target Absent
No Fluorescence
B
nuclear marker 1
nuclear marker 2
COI marker
T. absoluta (Spain)
T. absoluta (Latin America)
P. operculella
K. lycopersicella
Species
****
ns
0
2000
4000
6000
change in RFU normalized to NTC

Figure 14: Dilution series of RPA-Cas12a assay for (A) mitochondrial COI assay and the (B) nuclear marker 1 assay using 1 ng to $1x10^{-5}$ ng of *T. absoluta* DNA. 1 ng of *K. lycopersicella* and *P. operculella* DNA was used as a negative control. All RFU values are normalized by subtracting the RFU from a no-template control (NTC). RPA was performed in technical triplicate per sample. Error bars represent 1 SE. * indicates significant difference ($p<0.05$) from *K. lycopersicella* sample. C) Cell phone images of the same samples plotted in panel (B) in a UV illuminator using a 302nm wavelength. A representative for each triplicate was used.



References

Aman, R., A. Mahas, and M. Mahfouz, 2020 Nucleic Acid Detection Using CRISPR/Cas Biosensing Technologies. ACS Synth. Biol. 9: 1226–1233.

Bahamondes, L. A., and A. R. Mallea, 1969 Biologia en Mendoza de *Scrobipalpula absoluta* (Meyrick) Povolny (Lepidoptera:Gelechiidae), especie nueva para la Republica Argentina. Rev Fac Cs Agr Mendoza 15: 96–104.

Biondi, A., R. N. C. Guedes, F.-H. Wan, and N. Desneux, 2018 Ecology, Worldwide Spread, and Management of the Invasive South American Tomato Pinworm, Tuta absoluta: Past, Present, and Future. Annu. Rev. Entomol. 63: 239–258.

Bloem, S., and E. Spaltenstein, 2011 New Pest Response Guidelines: Tomato Leafminer (*Tuta absoluta*): Emergency and Domestic Programs, 176 p.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Chang, P. E. C., and M. A. Metz, 2021 Classification of Tuta absoluta (Meyrick, 1917) (Lepidoptera: Gelechiidae: Gelechiinae: Gnorimoschemini) Based on Cladistic Analysis of Morphology. Proc. Entomol. Soc. Wash. 123: 41–54.

Chen, J. S., E. Ma, L. B. Harrington, M. Da Costa, X. Tian *et al.*, 2018 CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science 360: 436–439.

Costa, M., and E. Heuvelink, 2007 Today's worldwide tomato production. Int. Suppliers Guide.

Desneux, N., E. Wajnberg, K. a. G. Wyckhuys, G. Burgio, S. Arpaia *et al.*, 2010 Biological invasion of European tomato crops by Tuta absoluta: ecology, geographic expansion and prospects for biological control. J. Pest Sci. 83: 197–215.

EPPO, 2008 First record of *Tuta absoluta* in Spain. EPPO Report. Serv.

FAOSTAT, 2020 Worldwide Production of Tomatoes: Food and Agriculture Organization of the United Nations.

Gandhi Gracy, R., B. R. Basavaarya, B. Kariyanna, C. G. Arunkumara, S. K. Jalali *et al.*, 2019 The relationship between genome size, morphological parameters and diet breadth in insect species. Biocatal. Agric. Biotechnol. 20: 101188.

Gariepy, T. D., U. Kuhlmann, C. Gillott, and M. Erlandson, 2007 Parasitoids, predators and PCR: the use of diagnostic molecular markers in biological control of Arthropods. J. Appl. Entomol. 131: 225–240.

Gilboa, S., and H. Podoler, 1995 Presence-Absence Sequential Sampling for Potato Tuber worm (Lepidoptera: Gelechiidae) on Processing Tomatoes: Selection of Sample Sites According to Predictable Seasonal Trends. J. Econ. Entomol. 88: 1332–1336.

Hossain, M. S., M. Y. Mian, and R. Muniappan, 2016 First Record of *Tuta absoluta* (Lepidoptera: Gelechiidae) from Bangladesh. J. Agric. Urban Entomol. 32: 101–105.

Li, Y., H. Mansour, T. Wang, S. Poojari, and F. Li, 2019 Naked-Eye Detection of Grapevine Red-Blotch Viral Infection Using a Plasmonic CRISPR Cas12a Assay. Anal. Chem. 91: 11510–11513.

Meyrick, E., 1917 I. Descriptions of South American Micro-Lepidoptera. Trans. R. Entomol. Soc. Lond. 65: 1–52.

Osama El-Lissy, 2019 Federal Order: Tomato Leaf Miner: Animal and Plant Health Inspection Service DA-2019-18, 6 p.

Ponti, L., A. P. Gutierrez, M. R. de Campos, N. Desneux, A. Biondi *et al.*, 2021 Biological invasion risk assessment of Tuta absoluta: mechanistic versus correlative methods. Biol. Invasions 23: 3809–3829.

Sint, D., M. Sporleder, C. Wallinger, O. Zegarra, J. Oehm *et al.*, 2016 A two-dimensional pooling approach towards efficient detection of parasitoid and pathogen DNA at low infestation rates. Methods Ecol. Evol. 7: 1548–1557.

Soh, J. H., E. Balleza, M. N. A. Rahim, H.-M. Chan, S. M. Ali *et al.*, 2022 CRISPR-based systems for sensitive and rapid on-site COVID-19 diagnostics. Trends Biotechnol. 40: 1346–1360.

Sridhar, V., A. Chakravarthy, R. Asokan, L. Vinesh, K. Rebijith *et al.*, 2014 New record of the invasive South American tomato leaf miner, Tuta absoluta (Meyrick) (Lepidoptera: Gelechiidae) in India. Pest Manag. Hortic. Ecosyst. 20: 148–154.

Stouthamer, R., J. Hu, F. J. P. M. van Kan, G. R. Platner, and J. D. Pinto, 1999 The utility of internally transcribed spacer 2 DNA sequences of the nuclear ribosomal gene for distinguishing sibling species of Trichogramma. BioControl 43: 421–440.

Tabuloc, C. A., K. M. Lewald, W. R. Conner, Y. Lee, E. K. Lee *et al.*, 2019 Sequencing of Tuta absoluta genome to develop SNP genotyping assays for species identification. J. Pest Sci. 92: 1397–1407.

Tobin, P. C., J. M. Kean, D. M. Suckling, D. G. McCullough, D. A. Herms *et al.*, 2014 Determinants of successful arthropod eradication programs. Biol. Invasions 16: 401–414.

Venette, R. C., D. R. Gordon, J. Juzwik, F. H. Koch, A. M. Liebhold *et al.*, 2021 Early Intervention Strategies for Invasive Species Management: Connections Between Risk Assessment, Prevention Efforts, Eradication, and Other Rapid Responses, pp. 111–131 in *Invasive Species in Forests and Rangelands of the United States*, edited by T. M. Poland, T. Patel-Weynand, D. M. Finch, C. F. Miniat, D. C. Hayes, et al. Springer International Publishing.

Wang, Y., Y. Ke, W. Liu, Y. Sun, and X. Ding, 2020 A One-Pot Toolbox Based on Cas12a/crRNA Enables Rapid Foodborne Pathogen Detection at Attomolar Level. ACS Sens. 5: 1427–1435.

Wang, X., P. Ji, H. Fan, L. Dang, W. Wan *et al.*, 2020 CRISPR/Cas12a technology combined with immunochromatographic strips for portable detection of African swine fever virus. Commun. Biol. 3: 62.

Yuan, C., T. Tian, J. Sun, M. Hu, X. Wang *et al.*, 2020 Universal and Naked-Eye Gene Detection Platform Based on the Clustered Regularly Interspaced Short Palindromic Repeats/Cas12a/13a System. Anal. Chem. 92: 4029–4037.

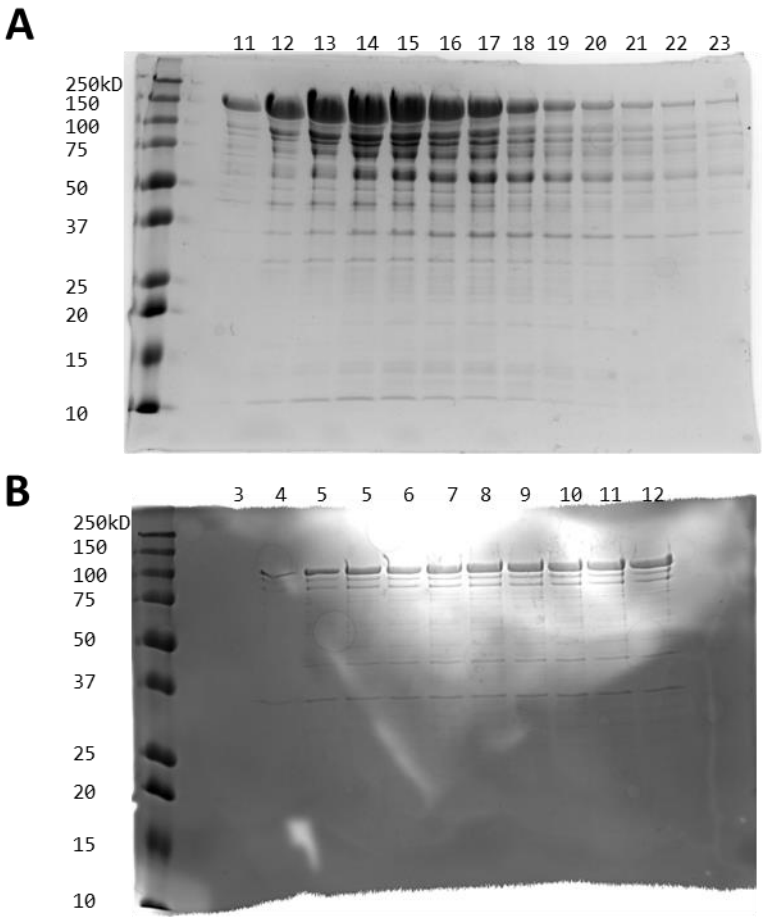
Zetsche, B., J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova *et al.*, 2015 Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163: 759–771.

Zhang, G., D. Ma, Y. Wang, Y. Gao, W. Liu *et al.*, 2020 First report of the South American tomato leafminer, Tuta absoluta (Meyrick), in China. J. Integr. Agric. 19: 1912–1917.

Zink, F. A., L. R. Tembrock, A. E. Timm, and T. M. Gilligan, 2020 A Real-Time PCR Assay for Rapid Identification of Tuta absoluta (Lepidoptera: Gelechiidae) (C. Tamborindeguy, Ed.). J. Econ. Entomol. 113: 1479–1485.

Supplemental Figures

Figure S17: Cas12a enzyme purification. A) Coomassie-stained gel of eluted fractions #11-#23 collected by liquid chromatography prior to TEV cleave of MBP and histone tag. Expected size of Cas12a+MBP tag is 189kD. B) Coomassie-stained gel of eluted fractions #3-12 after TEV cleavage of MBP and histone tag. Expected size of Cas12a is 147kD.



Supplemental Tables

Table S3: Oligonucleotides and Probes used for qPCR and RPA-Cas12a assays. For crRNAs, the conserved hairpin loop sequence is indicated in bold, and the recognition sequence indicated in plain face. For qPCR probes, the differentiating SNP is indicated in bold and flanked with brackets.

Assay	**Description**	**Sequence (5'->3')**
qPCR Assay 1	forward primer	TGTGGCCAACCTCATCTAAGG
qPCR Assay 1	reverse primer	CGGTACTGGAGTCGTCTCCTAA
qPCR Assay 1	*T. absoluta* probe	FAM-ACGCTTTCCCG**[A]**TTTATA-BHQ1Plus
qPCR Assay 1	*K. lycopersicella/P. operculella* probe	CALGold540-CGCTTTCCCG**[G]**TTTATA-BHQ1Plus
qPCR Assay 2	forward primer	GTTTCTTTCACCCCACACTATCAC
qPCR Assay 2	reverse primer	GCCATAAATATCGATCCGGACCTA
qPCR Assay 2	*P. operculella* probe	Quasar670-CACGTTTTCAG**[T]**AACACTT-BHQ2Plus
qPCR Assay 2	*K. lycopersicella/T. absoluta* probe	CALRed610-ACGTTTTCAG**[C]**AACACTTT-BHQ2Plus
RPA/Cas12a Nuclear Marker 1	forward RPA primer	GCTTTAACTGTTTAATCTGTTCATCCATTG
RPA/Cas12a Nuclear Marker 1	reverse RPA primer	CTGAAAATGTTGATGATGCCTTAAAGAACG
RPA/Cas12a Nuclear Marker 1	crRNA probe	**UAAUUUCUACUAAGUGUAGAU**UAUAGUUCACAGAGUCUUGA
RPA/Cas12a Nuclear Marker 2	forward RPA primer	TGTGCCAAATTGTTTGTCCCTGAACAAACATA
RPA/Cas12a Nuclear Marker 2	reverse RPA primer	GATGACCTTCAGGCTTTTAATACACTAGATTG
RPA/Cas12a Nuclear Marker 2	crRNA probe	**UAAUUUCUACUAAGUGUAGAU**UACUAUUUCUAGUAUUGAAA
RPA/Cas12a Mitochondrial Marker	forward RPA primer	CATTTAGCTGGTATTTCATCGATTTTAGGAGCTAT
RPA/Cas12a Mitochondrial Marker	reverse RPA primer	CCTGGTAAAATTAAAATATAAACTTCAGGATGTCC
RPA/Cas12a Mitochondrial Marker	crRNA probe	**UAAUUUCUACUAAGUGUAGAU**CUCCUUCUUUUAUCAUUGCC

Supplemental Scripts

Script S1: Python script used to extract potential crRNA targets from a mauve alignment file between *T. absoluta, K. lycopersicella,* and *P. operculella.*

```
from Bio import AlignIO
import sys
import re

##parameters
#distance to right and left to capture from probe binding region
binsize = 100

#create a new file to write output to.
sys.stdout = open("PAM_search_100bp_flanking.txt", "w")

#  open alignment file
xmfa = AlignIO.parse("Klyc-Pope-Tuta.xmfa", "mauve")
#  loop to identify potential crRNA binding regions (TTTN, with 20bp following).
loopindex = 0

for alignment in xmfa:  # goes through each alignment in xmfa file
    loopindex = loopindex + 1
    if len(alignment) == 3:  # makes sure all 3 species are aligned
        tutaseq = str(alignment[2].seq)
        for PAM in re.finditer("TTT[ACG][ACGT-]{20}", tutaseq):
            # goes through each possible PAM binding site in the sequence
            # grab klyc and pope seq at same positions as in tuta
            probestart = PAM.start()+4
            probeend = PAM.end()
            klycseq = str(alignment[0].seq)
            popeseq = str(alignment[1].seq)
            # count mismatches in the first 3 bases of the probe binding site
            klycmismatch = sum(c1 != c2 for c1, c2 in
zip(tutaseq[probestart:probestart+3], klycseq[probestart:probestart+3]))
            popemismatch = sum(c1 != c2 for c1, c2 in
zip(tutaseq[probestart:probestart+3], popeseq[probestart:probestart+3]))
            #count total mismatches in sequences compared to Tuta
            TotKlycMismatch = sum(c1 != c2 for c1, c2 in zip(tutaseq, klycseq))
            TotPopeMismatch = sum(c1 != c2 for c1, c2 in zip(tutaseq, popeseq))

            #only look at regions with more than 1 mismatches to Tuta.
            if klycmismatch > 1 and popemismatch > 1:
                # grab region before and after binding site. Skip if not at least
"binsize" distance flanking probesite.                regionstart = probestart -
binsize
                regionend = probeend + binsize
                if regionstart < 0 or regionend > len(tutaseq):
                    continue
                # skip if there is 10+ N's in a row
                if "NNNNNNNNNN" in tutaseq[regionstart:regionend] or "NNNNNNNNNN" in
popeseq[regionstart:regionend] or "NNNNNNNNNN" in klycseq[regionstart:regionend]:
                    continue
                # skip if there are 10+ blank spots in a row anywhere
                if "----------" in tutaseq[regionstart:regionend] or "----------" in
popeseq[regionstart:regionend] or "----------" in klycseq[regionstart:regionend]:
                    continue
                #skip if there are ANY Ns in corresponding Klyc and Pope probe sequence
```

```
                if "N" in popeseq[probestart:probeend] or "N" in
klycseq[probestart:probeend]:
                    continue
                #count total mismatches in binsize region flanking probe, compared to
Tuta
                TotKlycMismatch = sum(c1 != c2 for c1, c2 in
zip(tutaseq[regionstart:regionend], klycseq[regionstart:regionend]))
                TotPopeMismatch = sum(c1 != c2 for c1, c2 in
zip(tutaseq[regionstart:regionend], popeseq[regionstart:regionend]))

                # print output file
                print("> ", alignment[0].id.replace("/", ":"), "+", alignment[0].name,
",probe.region:", probestart + 1, "-", probeend + 1, ",(",TotKlycMismatch,")", sep="")
                print(klycseq[regionstart:probestart], "*",
klycseq[probestart:regionend], sep="")
                print("> ", alignment[1].id.replace("/", ":"), "+", alignment[1].name,
",probe.region:", probestart + 1, "-", probeend + 1, ",(",TotPopeMismatch,")", sep="")
                print(popeseq[regionstart:probestart], "*",
popeseq[probestart:regionend], sep="")
                print("> ", alignment[2].id.replace("/", ":"), "+", alignment[2].name,
",probe.region:", probestart + 1, "-", probeend + 1, ",(0)",sep="")
                print(tutaseq[regionstart:probestart], "*",
tutaseq[probestart:regionend], sep="")
                print("=alignment ", loopindex, sep="")
sys.stdout.close()
```