

Development and application of genomic approaches for resolving complex traits among insect pests

A dissertation by
Christopher Michael Ward

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics
School of Biological Sciences
The University of Adelaide
May 2021



THE UNIVERSITY
of ADELAIDE

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship, the Commonwealth Hill Postgraduate Scholarship, the Grains Research Development Corporation Postgraduate Research Scholarship, and from the School of Biological Sciences, University of Adelaide.

Christopher Ward

Acknowledgements:

I dedicate this thesis and all work over the course of my PhD to my mother, Ann Ward, who passed away during the final stages of this work. You imbued one of the greatest gifts a mother could – a love of the natural world.

To Simon Baxter, my primary supervisor. I would like to thank you for the chance and initially having me on as an undergraduate researcher and for supporting my growth as a researcher and encouraging me to branch out into different research fields. For the many hours spent providing critical feedback on work and research ideas I am forever indebted.

My co-supervisors, Iain Searle and Jimmy Breen, thank you for providing diverse and expert opinions on my projects and providing me with the means to take on alternate projects.

David Heckel, Heiko Vogel and the MPI for chemical ecology entomology department for making me feel welcomed in Jena for my research stay and providing helpful comments about experiments.

Many of the aspects of the *P. xylostella* genome assembly and its adaptation to *Pisum sativum* would not have been possible without funding (stipend, travel and operating) from the Grains Research Development Corporation. I would also like to thank Hugh MacLachlan (Commonwealth Hill Foundation) for generously providing stipend funds (Commonwealth Hill Scholarship) to carry out my PhD.

Thanks also to all other researchers who have helped me carry out or supported my research in some form, in no particular order: Amanda Choo, Kym Perry, Biko Kahare, Thu Nguyen, Stephen Pederson, Hien To, Kevin Powis, Greg Baker, Peter Crisp, and Alastair Ludington.

Finally, I would like to thank my partner Wen Yi (Jaysie) Chan for always being supportive throughout the course of my PhD and for listening to me talk about insects for the millionth time.

Table of Contents

Summary	5
Chapter 1: Introduction	7
Chapter 2: ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files	32
Chapter 3: Genomic evolutionary analysis in R with gear	36
Chapter 4: Assessing genomic admixture between cryptic <i>Plutella</i> moth species following secondary contact	49
Chapter 5: Estimation of molecular dates separating four <i>Plutella</i> species	64
Chapter 6: A haploid diamondback moth (<i>Plutella xylostella</i> L.) genome assembly resolves 31 chromosomes and identifies a diamide resistance mutation	72
Chapter 7: Adaptation of a major insect pest species to a new host plant is underpinned by a complex genetic mechanism and dynamic transcriptional response	88
Chapter 8: White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter	130
Chapter 9: Discussion and closing words	146
Appendix A: Contributing author publications	156
Appendix B: Supplementary information for Chapter 6	195
Appendix C: Supplementary information for Chapter 7	211

Summary:

Adaptation to sudden and dramatic environmental change can occur if the genetic variation within a population has the capacity to respond to a selective pressure. One salient example is the insect-plant arms race, whereby plants counter herbivory through evolving chemical deterrents and insects respond by developing strategies to disarm them. Brassicaceae plants produce toxic glucosinolate metabolites to deter insect predation, yet the diamondback moth, *Plutella xylostella* L., coevolved a counteradaptation to circumvent this defense and became a brassica specialist and world-wide pest. In Kenya (c. 1999), a *P. xylostella* population underwent a surprising host plant range expansion and was found infesting sugar-snap pea crops (*Pisum sativum*; Fabaceae), raising concerns this surprising adaptive phenotype could undergo selection elsewhere. This thesis primarily focuses on the evolution and diversification of the *Plutella* genus and the genetic basis of a host plant range expansion in *P. xylostella*.

Advances in genome sequencing technologies are improving opportunities to identify the genetic basis of adaptive traits, but also produce an increased volume of data leading to difficulties during data handling and processing. Here I developed two bioinformatic R package, *ngsReports* (**Chapter 2**) and *geaR* (**Chapter 3**) to overcome analytical bottlenecks encountered during this research, which facilitate quality assessment and analysis of high throughput genome sequence datasets. *ngsReports* interprets high-throughput Next Generation Sequencing (NGS) metrics, largely generated from Illumina platforms, enabling aggregation and visualization of quality control logs to rapidly identify sub-standard sequence data. *geaR* is a computationally inexpensive approach to perform established population genetic analyses using the Genomic Data Structure (GDS) format. Both packages were required for quality control and analysis of insect genomic datasets (**Chapters 4-8**).

Plutella xylostella recently evolved resistance to diamide insecticides in Australia, although its cryptic ally, *P. australiana*, is highly susceptible and causes very little pest pressure on agriculture. Hybridization occurs between these species in laboratory crosses, raising concern that insecticide resistance alleles could be transferred interspecifically. To test this hypothesis, I examine whole genomes of *P. xylostella* and *P. australiana* populations collected across two consecutive years and developed sensitive methods for identifying gene flow between species (**Chapter 4**). Subsequent crosses between diamide resistant *P. xylostella* and susceptible *P. australiana* generated hybrid pupa which were used to produce a haploid chromosome-level (n=31) genome assembly of *P. xylostella* through trio binning. This process identified a point

mutation within the insecticide's target, the Ryanodine Receptor, causing field evolved diamide resistance in Australia for the first time (**Chapter 6**).

The *P. xylostella* genome also established a resource for investigating the genetic basis of larval host plant range expansion (**Chapter 7**). Sequencing genetic crosses between Kenyan pea-adapted and wild-type *P. xylostella* strains revealed host adaptation is polygenic and polymorphic within the pea-strain despite ~17 years of laboratory captivity and sustained selection. Differential expression of larval midgut transcriptomes revealed an array of detoxification associated genes that responded to host plant diet. Similar approaches using head capsule tissue identified differential expression of olfaction and gustation pathway genes, and indicated decreased expression of gustatory and olfactory related genes were associated with pea adaptation.

Computational approaches used to assess geneflow between *Plutella* species were re-applied to identify the genetic basis of the *white pupa* phenotype in *Bactrocera dorsalis* (oriental fruit fly) (**Chapter 8**). The *white pupa* genetic sexing marker has been used for decades to separate and discard females in rearing factories that supply males for pest control, yet the genetic basis remained unknown. The *B. dorsalis* white pupa phenotype was introgressed into the genetic background of *B. tryoni* (Queensland Fruit Fly) and genome-wide analysis with *geaR* identified the causal locus. After scanning for mutations across the *white pupa* locus, a single frame shift mutation was identified in a *Major Facilitatory Superfamily* gene. This provided a strong candidate for the *white pupa* gene and subsequent CRISPR/Cas9 mediated knock out in *B. tryoni* recapitulated the phenotype, confirming its role in puparium pigmentation.

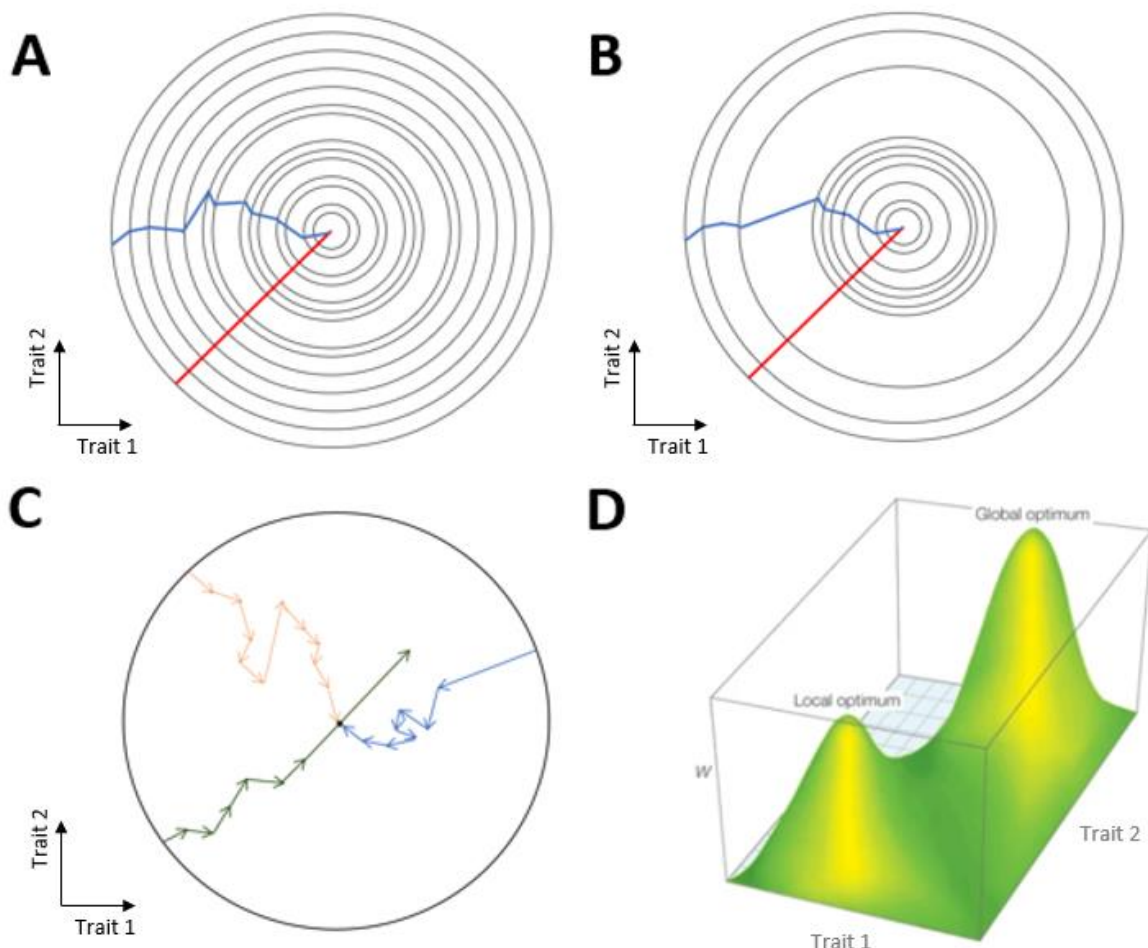
This thesis elucidates the genetic basis of agriculturally important phenotypes in major crop pest species and provides knowledge to advance integrated pest management strategies and crop protection. Bioinformatic packages developed provide a useful tool for researchers to carry out quality control of NGS datasets and evolutionary analysis of variant data when sample sizes are large. Furthermore, whole genome sequencing, RNA sequencing and a chromosome-level assembly of *P. xylostella* provide a powerful resource for future investigation of adaptive phenotypes in this agricultural pest.

1 Chapter I: Introduction

2 1.1 Genetic basis of adaptation

3 In its simplest form, evolution proceeds as natural selection and genetic drift alter allele
4 frequencies within a population culminating in the adaptation to an environment (Gilman and
5 Kozak 2015, Nonaka et al. 2015). In *On the Origin of Species* (1859), Darwin proposed that
6 adaptation to an environment occurred through a successive sequence of small phenotypic
7 changes which accumulate over evolutionary time. He hypothesised that “she [evolution] can
8 never take a leap, but must progress by the shortest and slowest steps” paving the way for a
9 gradualist view of adaptation (Suzuki 2017). Gradualism, or micromutationalism, proposes a
10 pre-genetic view of adaptation whereby a phenotype comes to fit an environment through a
11 series of infinitesimally small steps (Maynard Smith 1982). At the start of the neo-Darwinian
12 synthesis, as the individual gene was being championed as the unit of selection, R.A. Fisher
13 proposed his infinitesimal model (Fisher 1918) which argues that adaptive phenotypes result
14 from an infinitely large number of small effect mutations (**Figure 1A**) and therefore attempting
15 to directly quantify the effect of a single mutation would be folly. This was then superseded by
16 Fisher (1930) with the geometric model. Fisher proposed that adaptive change results in
17 differently sized effects on phenotype, although large effect mutations are far more infrequent
18 than their counterparts and are often deleterious. His model was elegant in its simplicity with
19 only three parameters necessary: phenotypic complexity of the organism, phenotypic fitness
20 to environment and effect size of the mutation. At its core, the geometric model argues that
21 adaptation effect size and phenotypic complexity are negatively correlated. In other words,
22 adaptive mutations in phenotypically complex organisms have small phenotypic effect size
23 and are more favourable than large mutations. Though as Kimura (1983) pointed out half a
24 century later, the probability of a mutation being favourable and its likelihood of substitution
25 are vastly different. Kimura proposed larger effect mutations are more likely to be adaptive
26 than first suggested by Fisher due to their higher probability of fixation. This was further
27 supported when modern experiments began investigating adaptation via artificial and natural

28 selection experiments, with a growing number studies finding large effect mutations underlying
29 adaptive change (Shrimpton and Robertson 1988, Paterson et al. 1991, Feyereisen 1995,
30 McKenzie and Batterham 1998). Though, this may be due to selection bias, where organisms
31 with large effect mutations are analysed simply because their phenotypes are pronounced and
32 quantifiable. To reconcile these observations, Orr (1998) reframed Fishers geometric model
33 using an adaptive walk approach. He emphasized that to avoid being deleterious, large effect
34 mutations are more likely to occur near the start of the walk, before sufficient adaptation to the
35 environment has occurred (**Figure 1C**). This demonstrated not only that large effect mutations
36 play a central role in adaptation, but that the distribution of fixed mutations during adaptation
37 is roughly exponential when migration is minimal (Orr 1998).



38
39 **Figure 1:** Genetic theories of adaptation. Graphs in A, B, and C represent simplistic adaptive walks for
40 a simple two-trait species. The species begins the adaptive walk at the perimeter of the outermost circle

41 and the optimal phenotype is found at the centre. A) Paths of adaptation under the infinitesimal model,
42 whereby a large number of small mutations are necessary to reach the optimal phenotype (blue lines),
43 making rapid adaptation to environment (red line). B) Paths of adaptation under Fisher's Geometric
44 Model. In contrast to the infinitesimal model (A), large effect mutations occur early in the adaptive walk,
45 providing rapid adaptive leaps towards the optimal phenotype. C) Two adaptive walks (blue and green)
46 under Fisher's Geometric Model with the extensions outlined in Orr (1998), where each arrow
47 represents a mutation of size equal to the length of the arrow. Larger steps at the start of an adaptive
48 walk (blue arrows) are far more effective and less likely to be deleterious than later during the walk. For
49 example, green arrows show a mutation exceeding the optimal and orange arrows show that large
50 mutations can negatively effect fitness through 'directional' steps away from the optimum that would not
51 have occurred if the effect size was small. D) A fitness landscape depicting all possible adaptations of
52 a two trait species (x and z axis) against fitness (w, y axis). This has been adapted from Orr (2005).

53

54 Theoretical models of phenotypic effect size also provide insights into the number of genes
55 responsible for adaptive change. In accordance with its predictions on mutation effect size,
56 Fisher's Geometric Model (FGM) predicts an exponential distribution of gene effect sizes
57 (**Figure 1B and C**). Over half a century later, Lande (1983) sought to understand the
58 processes underlying gene effect size distribution, proposing that effect size is proportional to
59 selection strength and persistence. Yet as Lande himself pointed out, his model assumes that
60 weak selection is the rule in nature, which may not be as true as once thought (Endler 1986,
61 Kingsolver et al. 2001). The FGM and Lande's model are by no means mutually exclusive,
62 with both predicting an abundance of small effect genes underlying adaptation. Therefore,
63 Lande's model can be seen as a pre-Orr (1998) extension of Fisherian theory to explain how
64 populations conform to the FGM. Experimental evidence supporting gene number in adaption
65 is mixed and provides no clear insights into effect size frequency (Manolio et al. 2009,
66 Pritchard and Di Rienzo 2010, Pritchard et al. 2010, Boyle et al. 2017, Marouli et al. 2017,
67 Csilléry et al. 2018) with the interesting discovery that both small effect and large effect
68 mutations can cause the same phenotype (Crow 1957, Paterson et al. 1988, Zan et al. 2017).
69 Among insects, a growing body of research is continuing to uncover both mono- and polygenic

70 mechanisms underlying adaptation. For example, insecticide resistance is generally conferred
71 by a single mutation (Baxter et al. 2010, Gahan et al. 2010, Jouraku et al. 2020, Zuo et al.
72 2020), whereas the genetic mechanism underlying host plant preference is usually controlled
73 by multiple genes (Sheck and Gould 1996, Henniges-Janssen et al. 2011).

74

75 **1.2 Fitness landscapes and genetic variation**

76

77 Adaptive walks move through the ‘hills’ and ‘valleys’ of a fitness (or adaptive) landscape to
78 select adaptations that best suit an environment. In their simplest form, adaptive landscapes
79 describe the relationship between mutational change at a single-locus and evolutionary fitness
80 (**Figure 1D**). On a theoretical level, natural selection ‘walks’ through the landscape of all
81 possible mutations to find the optimal fit state for that locus. Though in practice, genes are
82 pleiotropic in nature with optimal mutations for one gene potentially causing deleterious effects
83 to another. Kaufman’s NK model (Kauffman and Weinberger 1989) investigated the effect of
84 pleiotropy by incorporating multiple (N) genotypes in epistasis with K genes, either in the same
85 genetic neighbourhood or distributed across the genome. The fitness landscape of the NK
86 model is rugged, with multiple different combinations of loci producing many local maxima.
87 Consequently, alleles under the same environmental stressors do not always have the same
88 fitness in different genetic backgrounds. These local maxima have lower fitness than the
89 optimal combination of alleles, yet are ‘fit-enough’ to survive in the environment leading a
90 population to easily become trapped in a globally sub-optimal position in the landscape.

91

92 When faced with novel stressors, adaptive genotypes will be positively selected for, and are
93 likely to increase their frequency over generational time. However, there has been a long-
94 standing debate on whether the majority of mutations are new (*de novo* mutation) or already
95 present in the population gene pool (standing variation) (Hermisson and Pennings 2005, Peter
96 et al. 2012). In the past, many models such as Fisher’s Geometric Model and its extension by
97 Kimura (1983) and Orr (1998), assumed that *de novo* mutation was the major fuel of

98 adaptation. Yet, other theoretical and empirical work showed that standing variation
99 contributed to many observed adaptations (Orr and Betancourt 2001, Barrett and Schluter
100 2008, Sheng et al. 2015). Under the same selective pressures, the probability of fixation for
101 beneficial *de novo* mutations is far lower, especially if the mutation is recessive (James 1965,
102 Lande 1983, Orr and Betancourt 2001). This is due to the negative pressures recessive
103 beneficial alleles face when going through a selective sweep: Haldane's sieve (Turner 1981,
104 Charlesworth 1992) – the bias against recessive mutations and genetic drift (Kimura 1968,
105 1983) – the removal of low frequency variants through random survival. Adaptation from
106 standing variation can overcome both challenges by having multiple copies present in the
107 population (Orr and Betancourt 2001) that may have benefited from 'pre-testing' under a
108 similar stressor (Barrett and Schluter 2008). Interestingly, theoretical work supports selection
109 from standing variation as more prevalent, unless the neutral mutation rate is low (Hermisson
110 and Pennings 2005). Standing variation as the major fuel of selection in natural populations is
111 also supported by an increasing body of literature (Jones et al. 2012, Reid et al. 2016, Lai et
112 al. 2019), though there are still many examples of selection of adaptive genotypes from *de*
113 *novo* mutation (van't Hof et al. 2011, Acuna-Hidalgo et al. 2016, Hawkins et al. 2018).

114

115 **1.3 Strategies for identifying adaptive traits**

116 Classical genetic crosses between two pure breeding isolates fixed for alternate phenotypes,
117 and their subsequent backcrossing or sibling mating, was instrumental in understanding
118 heritability by Mendel (Wynn 2007). Not only are crosses useful for deciphering whether a trait
119 is monogenic or polygenic in diploid species, but specific regions of the genome encoding a
120 trait of interest can be identified when aided by crossing over during meiosis between non-
121 sister chromosomes to generate recombinants (Mott 2001). Next Generation Sequencing
122 (NGS) has provided new opportunities for identifying the genes driving adaptive phenotypes
123 through the ability to sequence whole genomes of populations or pedigrees and quickly

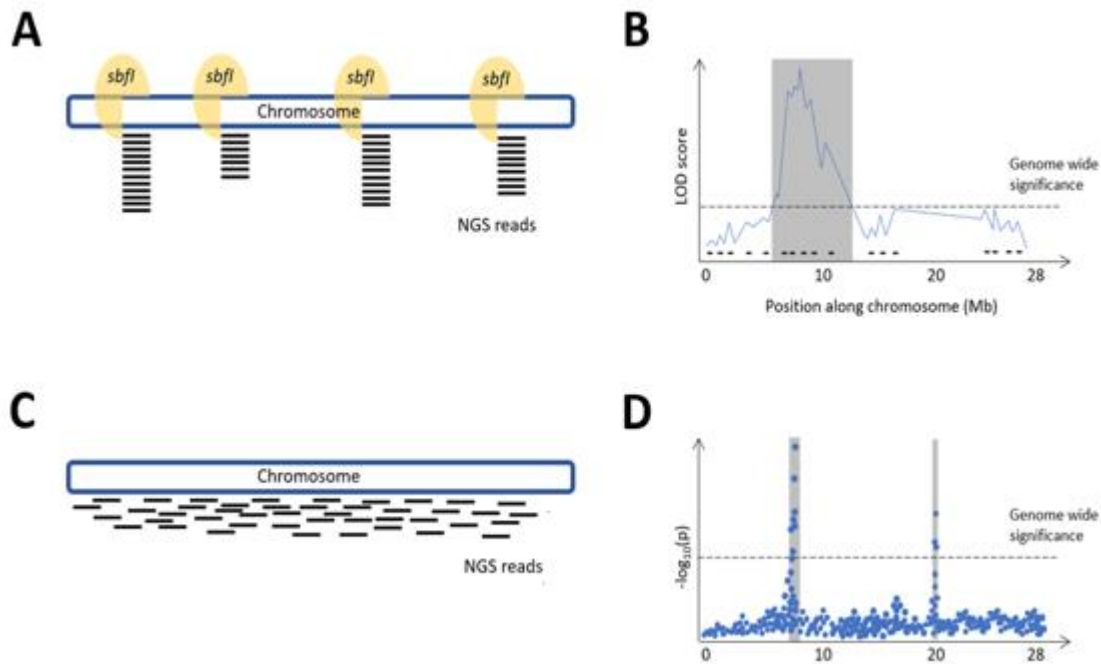
124 develop thousands of segregating markers across a genome (Solberg Woods 2014, Jamann
125 et al. 2015, Gu et al. 2018).

126

127 Early quantitative trait locus (QTL) studies, such as those in *Drosophila melanogaster*
128 (Zimmerman et al. 2000, Robin et al. 2002), focused on identifying the genetic basis of
129 monogenic Mendelian traits due to their ease of identification. Yet, next generation sequencing
130 empowered QTL and genome wide association studies (GWAS) to identify regions of the
131 genome associated with complex phenotypes. This gave the first genetic insights into highly
132 complex traits in human disease (Manolio et al. 2009, Stefansson et al. 2009) and
133 domesticated animal breeding (Cadieu et al. 2009, Goddard and Hayes 2009) with the
134 interesting discovery that significantly associated loci make up only a small fraction of the
135 predicted heritability for many traits (Manolio et al. 2009, Eichler et al. 2010). Referred to as
136 'missing heritability,' many small effect loci lacked genome wide significance causing them to
137 evade detection. Research was further confounded when, in contrast to mendelian traits,
138 complex traits were driven by non-coding variation that is likely to effect gene regulation, rather
139 than protein function directly (Ward and Kellis 2012).

140

141 Multiple sequencing methods can be used to identify QTL across the genome using both DNA
142 and RNA sequencing. Reduced representation genome sequencing methods, such as
143 Restriction-site Associated DNA sequencing (RAD-seq) (Baird et al. 2008), are powerful and
144 cost-effective strategies to carry out genetic analysis on populations or pedigrees (Davey and
145 Blaxter 2010, Peterson et al. 2012). Nuclear genomes from individuals are digested using a
146 restriction enzyme (e.g. *Sbf1*) and ligated to barcoded adapters to facilitate sequencing and
147 enable analysis of small genomic regions at each cut site (**Figure 1A**). Ultimately, only a small
148 fraction of the genome is sequenced (Baird et al. 2008), which enables RAD-seq to generate
149 genetic markers distributed across the genome for use in genotype/phenotype association
150 analysis among pedigrees and population genetic analysis.



151

152 **Figure 2:** Common strategies used to identify trait loci with next generation sequencing (NGS)
 153 technologies. **A)** Restriction-site Associated DNA sequencing (RAD-seq) involves targeted sequencing
 154 of short DNA regions flanking restriction enzyme cut sites (e.g. *SbfI*). Marker density is dependent on
 155 the frequency of cut sites across the genome. Common genomic regions are sequenced among closely
 156 related individuals or pedigrees, which represents a small proportion of the genome and decreases the
 157 cost of sequencing large numbers of individuals. **B)** RAD-seq markers generated from a pedigree with
 158 segregating phenotype can be used for genome-wide QTL mapping studies. The blue line represents
 159 LOD scores calculated for RAD-seq marker distributed across the chromosome (black dashes). The
 160 genetic region with LOD scores higher than the genome wide significance threshold (usually calculated
 161 by carrying out a permutation test with un-ordered phenotype measurements) are highlighted in grey.
 162 The resulting QTL is large as all markers significantly linked to the causal mutation will have similar
 163 allele frequencies. **C)** Schematic representation of whole genome shotgun sequence reads for a
 164 chromosomal region of a single individual. Chromosomes are randomly sheared and sequenced to
 165 provide relatively even coverage of all non-repetitive genomic regions. **D)** A genome wide association
 166 study (GWAS) involves sequencing a population of individuals and calculating the probability of
 167 genotypes associated with specific phenotypes.

168

169 RAD-seq markers are generally used to calculate logarithm of the odds (LOD) scores across
 170 the genome to identify regions genetically linked to a specific phenotype (**Figure 2B**). This has
 171 been utilized to identify both discrete and complex traits such as spinosad resistance in
 172 *Plutella xylostella* (Baxter et al. 2011), pyrethroid resistance in *Cymex lectularius* (Fountain et

173 al. 2016), assortative mating traits in *Heliconius* butterflies (Merrill et al. 2019), agronomic traits
174 in *Setaria italica* (Wang et al. 2017), and growth traits in Asian seabass (Wang et al. 2015).
175 However, as only limited genomic regions are sequenced, discovery of QTLs requires
176 sufficiently high linkage disequilibrium with RAD-seq molecular markers and a robust
177 reference genome to identify candidate genes surrounding associated markers. QTL resulting
178 from RAD-seq are generally large, as all markers significantly linked to the causal mutation
179 will have similar allele frequencies (**Figure 2B**).

180

181 Genome wide association studies (GWAS) are beneficial for associating sites linked with or
182 causing phenotypes among organisms with near-complete genome assemblies, and those
183 where genetic crosses are not feasible (eg. large generation time). While RAD-seq relies on
184 identifying molecular markers in linkage disequilibrium with a measurable phenotype, GWAS
185 can investigate each variant in the genome for an association. By sequencing the whole
186 genomes of many individuals (**Figure 2C**), single variants associated with the phenotype can
187 be statistically identified by fitting a generalized (Chu et al. 2020) or mixed (Zhang et al. 2010)
188 linear model. The probability of association is then calculated for each variant enabling direct
189 association with a given phenotype decreasing the range of QTL significantly (**Figure 2D**).
190 However, GWAS is not feasible for many species requiring genotypes from thousands of
191 individuals to generate clear statistical significance for complex traits (Hong and Park 2012),
192 making it cost-prohibitive for most organisms without reference panels available, such as UK
193 BioBank (Sudlow et al. 2015) and the *Drosophila* Genetic Reference Panel (Mackay et al.
194 2012). Although more advanced methods now exist to identify complex QTL (Wang et al. 2016,
195 Zhang et al. 2020), linear mixed models have been used to disentangle the genetic basis of
196 plant height in barley (Alqudah et al. 2016), insecticide resistance (Green et al. 2019) and
197 starvation resistance (Huang et al. 2014) in *D. melanogaster* along with type 2 diabetes risk
198 (Sladek et al. 2007) and age-related macular degeneration (Klein et al. 2005) in humans.

199

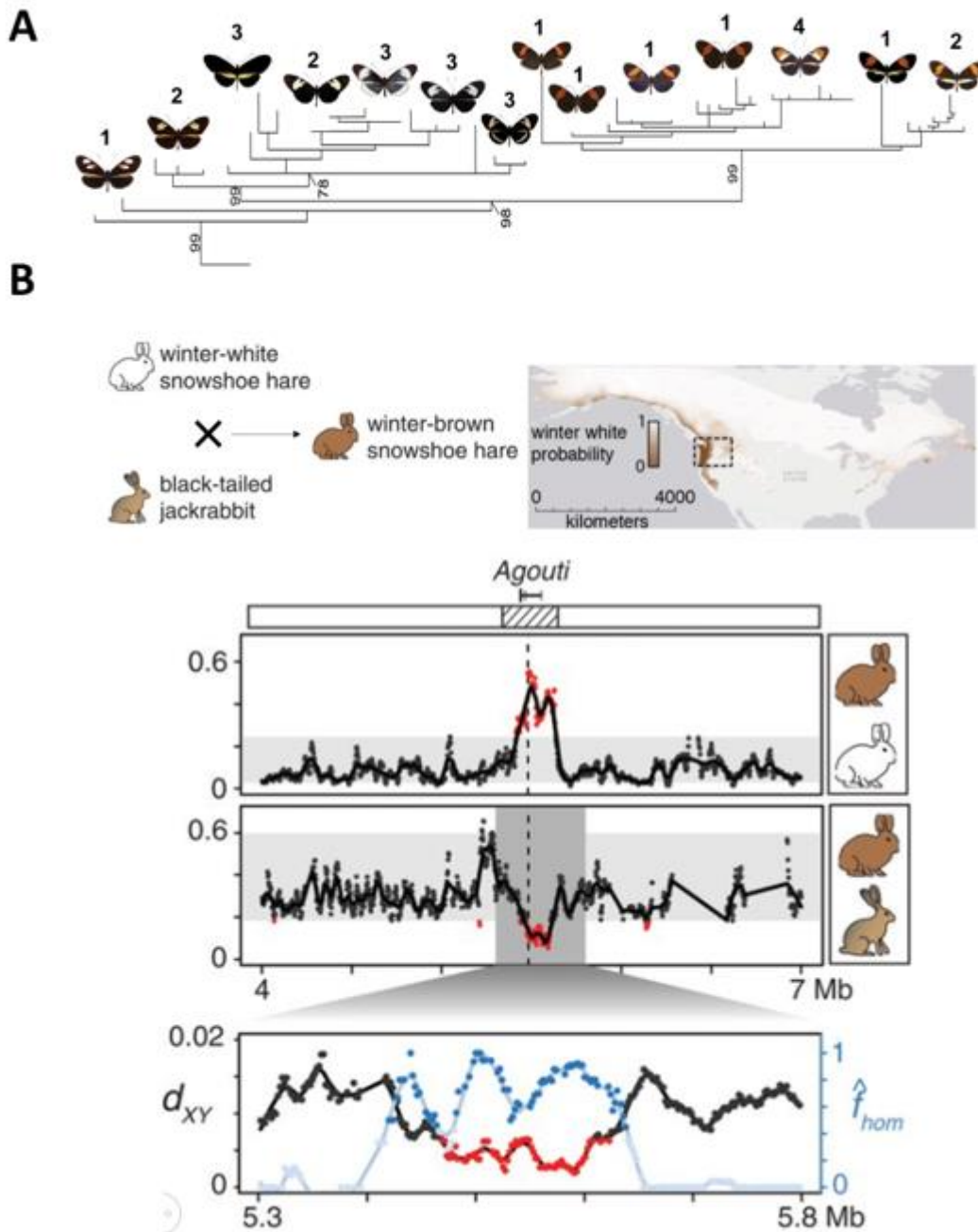
200 Phenotypic variation is not limited to changes in protein coding genes. Many extreme changes
201 in phenotype have been accounted for by changes in gene regulatory networks and gene
202 expression (Rose et al. 2016, Schweizer et al. 2016, Guo et al. 2017). For example, differential
203 expression of olfaction and gustation related genes have been shown in insect host plant races
204 reared on their non-native host plant (Kopp et al. 2008, Shiao et al. 2015, Eyres et al. 2016).
205 Furthermore, feeding assays in both hemi (Eyres et al. 2016, Xiao et al. 2021) and
206 holometabolous (Shiao et al. 2015, Orsucci et al. 2018) insects revealed differential
207 expression of chemosensory genes, suggesting they have a pivotal role in host plant
208 preference. More recent studies have sought to couple differential gene expression and
209 traditional QTL or GWAS analysis to identify trait loci for complex gene regulation phenotypes
210 referred to as expression QTLs (eQTLs) (Fagny et al. 2017, Mizuno and Okada 2019, Liu et
211 al. 2020).

212

213 **1.4 Hybridization and transference of beneficial alleles between species**

214 Adaptive mutations do not necessarily originate within the species in which they are identified.
215 Hybridization and flow of genetic variation (introgression) from a donor species into a recipient
216 can transfer adaptive advantage. The transfer of advantageous pre-adapted alleles from one
217 taxon into another removes the need of novel variation to arise in a population through *de*
218 *novo* mutation. Hybridisation with gene flow has been found to be at the root of many major
219 adaptations in extant taxa from rodenticide resistance between mice species (Song et al.
220 2011) to aposematic wing patterns in *Heliconius* butterflies (**Figure 3A**) (Mavárez et al. 2006,
221 Pardo-Diaz et al. 2012, Wallbank et al. 2016). Jones et al. (2018) examined two morphs of
222 snowshoe hare, one with a white overwintering coat for camouflage in the snow, and the other
223 brown (**Figure 3B**). Using whole genome sequence data, they showed introgression of a
224 melanism pathway gene *Agouti* from black-tailed jackrabbits swept to fixation in regions where
225 snow was scarce during winter. This conferred adaptive advantage in warmer climates where
226 the white winter coat may impose a fitness cost of reduced camouflage. Examples can also

227 be found throughout diverse organisms driving local adaptation to climate change in European
 228 white oak (Leroy et al. 2020), response to high altitude in humans (Huerta-Sánchez et al.
 229 2014), and insecticide resistance in mosquito vectors (Norris et al. 2015) emphasizing
 230 introgression's pivotal role in adaptive change.
 231



232

233 **Figure 3:** Interspecies hybridisation facilitates transfer of adaptive alleles from one species to another.
234 **A)** Phylogenetic reconstruction of sequence datasets from multiple *Heliconius* butterfly species
235 revealed the expected species topology was not observed at the “HmB453k” locus. Species clustered
236 according to phenotype, suggesting adaptive introgression of wing patterning loci was responsible for
237 their similarity and not convergent evolution. Species are denoted with numbers above images. 1) *H.*
238 *melpomene*, 2) *H. timereta*, 3) *H. cydno* and 4) *H. heurippa*. Butterfly images are placed at approximate
239 locations for their corresponding branch. Figure modified from Pardo-Diaz et al. (2012). **B)** Snowshoe
240 hares experience a winter molt, replacing brown coats with white. Populations living in regions with
241 limited snow cover have experienced adaptive introgression of a winter brown coat phenotype from the
242 black-tailed jackrabbit. Overwintering colour is geographically structured with the proportion of winter-
243 brown morphs greatly increasing within coastal populations where the temperature is warmer. Absolute
244 genetic distance (d_{XY}) and the introgression estimator f_{hom} were used to show that white and brown
245 morphs of the snowshoe hare are highly divergent at the agouti locus, whereas brown morphs were
246 highly similar to the black-tailed jackrabbit. Figure adapted from Jones et al. (2018).

247
248 Hybridisation following secondary contact between native and invasive species provides an
249 opportunity for pre-adapted alleles to be directly transferred from a donor to recipient species.
250 In some cases, adaptive introgression has facilitated the exchange of insecticide resistance
251 alleles between pyrethroid resistant *Helicoverpa armigera* moths and local *H. zea* populations
252 in Brazil (Valencia-Montoya et al. 2020). Strong pressure from pyrethroid use in Australian
253 agriculture has selected for resistance phenotypes in the invasive species *Plutella xylostella*
254 (Shelton et al. 1993, Zalucki et al. 2012, Qin et al. 2018, Xia et al. 2018), yet a recently
255 identified cryptic sister species, *P. australiana*, is highly susceptible (Perry et al. 2018).
256 Hybridisation between the two species can occur under laboratory conditions (Perry et al.
257 2018), raising concerns for the exchange of beneficial alleles in the field. Both *P. xylostella*
258 (Endersby et al. 2006, Perry et al. 2020) and *P. australiana* (Perry et al. 2018) lack population
259 structure across Australia, highlighting intraspecific extensive geneflow and the potential for
260 beneficial mutations to rapidly spread across the continent.

261
262 Introgression has been identified between many species (Mallet 2005, Mariac et al. 2006,
263 Pardo-Diaz et al. 2012, Jones et al. 2018). Genome wide summary statistics such as Nei’s d_{XY}
264 (**Figure 3B**) and F_{ST} , which characterize genetic differentiation between two populations,

265 along with Nei's π (genetic diversity within a population) have been useful in identifying high
266 levels of introgression (Neafsey et al. 2010, Nadachowska-Brzyska et al. 2013, Smith and
267 Kronforst 2013). Yet, identification of low levels of introgression remains challenging. Genomic
268 regions introgressed following interspecific hybridization between two sympatric species
269 show lower genetic distance than allopatric populations of the recipient species. However,
270 these reveal little about the rate of introgression between taxa – termed the effective migration
271 rate (Martin and Jiggins 2017). The effective migration rate is difficult to estimate using
272 traditional population metrics due to population dynamics such as incomplete lineage sorting
273 and interspecific differences in effective population size (Martin et al. 2015, Peter 2016, Martin
274 and Jiggins 2017). In recent years, formalized f -statistics have been developed to tackle this
275 challenge (Patterson et al. 2012, Martin et al. 2015), for example f_d , which is a strong estimator
276 for the effective migration rate (Martin et al. 2015). F -statistics utilize the frequency of derived
277 and ancestral alleles at biallelic sites across the genome to construct 'drift trees' which identify
278 genomic regions with incongruent evolutionary history, where an ingroup species shares more
279 genetic drift with a sympatric outgroup than members of the same species sampled from an
280 allopatric population (Patterson et al. 2012, Peter 2016). Due to the changes in local
281 evolutionary history caused by admixture, phylogenetic reconstruction can also be a powerful
282 tool to investigate local introgression. Tools such as TWISST (Martin and Van Belleghem
283 2017) and Saguaro (Zamani et al. 2013) use local topology across the genome to identify
284 regions with different topologies compared to the consensus species tree. Both topology and
285 drift-based estimates of introgression have been utilized to dissect highly complex
286 introgression patterns in *Heliconius* butterflies (Edelman et al. 2019), chickens (Lawal et al.
287 2020) and *Arabidopsis lyrta* (Marburger et al. 2019).

288

289 This research integrates aspects of classical genetics, introgression, quantitative trait
290 mapping, genomics, bioinformatics, and software development to investigate adaptive or
291 selectable phenotypes among insect pests which are of economic importance to agricultural
292 industries. Due to its importance for understanding insecticide resistance evolution, complex

293 migration patterns and host plant olfaction, I utilized *Plutella xylostella* as a model species for
294 most of the aims in this dissertation. Greatly increasing the robustness of genomic resources
295 for *P. xylostella* along with our understanding of agriculturally important trait loci in insect pest
296 species and their evolution.

297

298

299 **Thesis outline**

300 **Aim 1: Develop computational packages to interpret high-throughput sequence**
301 **libraries metrics and perform memory efficient evolutionary analysis among large**
302 **genomic datasets.**

303

304 The R programming language has become a popular platform for biologists to analyse and
305 interpret large datasets and to produce quality graphical outputs. **Chapters 2 and 3** present
306 R packages designed for quality control assessment of large Illumina sequencing datasets
307 and for performing evolutionary genomic analyses. For more than a decade, individual Illumina
308 *fastq* sequencing runs have been routinely analysed by sequencing centres and
309 computational biologists using the fastQC
310 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) application, which generates
311 graphical displays of quality control logs. It has provided a standard benchmarking strategy
312 that visualising the quality of one file per display. In **Chapter 2**, I describe the R package
313 *ngsReports* that enables simultaneous visualization of tens to hundreds of log files for quality
314 control and systematic identification of bias among Illumina sequence data (Ward et al. 2019).
315 This tool was applied to all Illumina sequence datasets described within this thesis.

316

317 **Chapter 3** describes the **Genomic and Evolutionary Analysis R** package, *geaR*, which utilizes
318 genome wide genotype data to carry out evolutionary analysis using a modular object-oriented
319 design in the Genomic Data Structure (GDS) format (Ward et al. 2020). Tests of genetic
320 diversity (d_{XY} , π , F_{ST}) and admixture (f_4 , f_d) can be carried out on partitioned genomes using

321 a single function, regardless of sample ploidy or number of observed alleles. *geaR* also
322 provides tree building functions and the ability to output DNA sequence using genomic
323 coordinates. Although other population genetics packages exist in the R programming
324 language, many lack the scalability necessary for whole genome datasets. Furthermore, direct
325 integration of GRange objects allows the user to limitlessly customize genomic windows or
326 features included in the analysis. These functions were used for data analysis in **Chapters 4**
327 **- 8.**

328

329

330 **Aim 2: Investigate the extent of hybridization, admixture and introgression among i)**
331 **sympatric pest species and ii) interspecific crosses.**

332

333 Hybridization and genomic admixture between closely related species can transfer pre-
334 adapted traits between lineages and lead to adaptive introgression. In **Chapter 4**, I use *geaR*
335 to estimate the divergence between Australian and Hawaiian *Plutella xylostella* populations
336 and its cryptic ally, *P. australiana*, then assess evidence of gene flow between sympatric
337 Australian collections. This involved constructing a novel test for phylogenetic incongruence
338 that was able to detect gene flow, even at low frequencies in simulated data (Ward and Baxter
339 2018).

340

341 Mass releases of sterilized male insects, in the frame of sterile insect technique programs,
342 have helped suppress insect pest populations since the 1950s. Genetic sexing strains of the
343 horticultural pest *Bactrocera dorsalis* have been bred to express dimorphic phenotypes, where
344 female pupae are white and male pupae are brown. Colour variation enables selective removal
345 of female pupae prior to sterilization and release, yet until now, the gene responsible for white
346 pupae remained unknown. In **Chapter 8** I utilize an interspecific hybrid line of *B. tryoni* carrying
347 the *B. dorsalis* white pupa locus to identify regions of the genome associated with the wp
348 phenotype. Whole genome scans of f_d , d_{XY} and tree topology reveal a single region of the

349 genome associated with *wp*. Coding sequence within the *wp* locus were then extracted to
350 identify a transmembrane transporter gene for functional analysis (Ward et al. 2021).

351

352 **Aim 3: Improve evolutionary and genomic resources for the major Brassica pest**
353 ***Plutella xylostella* L.**

354

355 *Plutella xylostella* is considered the only major agricultural pest of the 26 Member *Plutella*
356 genus. In **Chapter 5**, I construct and compare mitochondrial genome assemblies four *Plutella*
357 species and one Yponomeutid outgroup species, *Acrolepiopsis assectella*. Molecular dating
358 of mitochondrial genomes date the *Plutella* crown node to 7.05-4.10 MYA and estimate when
359 *P. xylostella* diverged from its most recent common ancestor.

360

361 The first draft genomes of *P. xylostella* were sequenced using short read Illumina technology
362 and released in 2013, however, complex repeats, structural variation and technical limitations
363 led to fragmented genome assemblies. In order to perform comprehensive genomic analysis,
364 a chromosome level reference genome was required. **Chapter 6** describes single paired
365 mating between a *P. xylostella* male and *P. australiana* female which generated a single
366 female hybrid pupa that was then sequenced to ~260X coverage using Pacific Biosciences
367 SMRT cells. Short read Illumina data (~30X) was used to separate the paternal *P. xylostella*
368 from the maternal *P. australiana* haplotype allowing for assembly of a haploid *P. xylostella*
369 genome. After scaffolding the assembly using Hi-C linked reads, I resolved a robust genome
370 of ~328 Mb distributed across 31 chromosome length scaffolds. Gene prediction and
371 annotation using a publicly available RNAseq datasets estimated this genome contained
372 19,003 protein coding genes and the data enabled identification of a mutation causing
373 resistance to diamide insecticides.

374

375

376 **Aim 4. Analysis of a complex host plant range expansion in *P. xylostella***

377

378 In 1999 Kenyan populations of *P. xylostella* (DBM-P) underwent a surprising host plant range
379 expansion to include *Pisum sativum* (sugar snap pea). The resources generated throughout
380 this thesis were primarily developed to identify the genetic basis of this phenotype. In **Chapter**
381 **7**, I carry out crosses between the DBM-P strain and a wild type laboratory reference then
382 perform QTL mapping with reduced representation genome sequencing data and
383 transcriptome profiling of larval tissues to understand adaptation to *P. sativum*.

384

385 I then carried out no-choice feeding assays using the same two DBM strains on either a
386 *Brassica napus* or *P. sativum* host. After dissecting either head or midgut tissue from individual
387 larvae, each tissue type was pooled separately and sequenced. Transcriptome profiling of
388 head tissue revealed clear differential expression in gustatory receptors and odorant binding
389 proteins along with enrichment of genes ontologies synaptic signalling and response to stimuli.
390 Midgut tissue showed differential expression of a host of genes involved in secondary
391 metabolism and genes involved in a response to oxidative stress.

392

393 Finally, I incorporate both analyses to compare genomic positions of differentially expressed
394 genes and QTL. I then discuss this in the context of previously published work and theoretical
395 models of adaptation.

396

397 **Appendix A:**

398 Published manuscripts I contributed to during my PhD, but did not fall under a major aim,
399 are included in Appendix A. Manuscripts included in Appendix A are:

- 400 1) Choo A, Nguyen TN, **Ward CM**, Chen IY, Sved J, Shearman D, Gilchrist AS, Crisp P,
401 Baxter SW. Identification of Y-chromosome scaffolds of the Queensland fruit fly reveals
402 a duplicated *gyf* gene paralogue common to many *Bactrocera* pest species. *Insect*
403 *molecular biology*. 2019 Dec;28(6):873-86.
404
- 405 2) You M, Ke F, You S, Wu Z, Liu Q, He W, Baxter SW, Yuchi Z, Vasseur L, Gurr GM,
406 **Ward CM**, Cerda H, Yang G, Peng L, Jin Y, Xie M, Cai L, Douglas CJ, Isman MB,
407 Goettel MS, Song Q, Fan Q, Wang-Pruski G, Lees DC, Yue Z, Bai J, Liu T, Lin L,
408 Zheng Y, Zeng Z, Lin S, Wang Y, Zhao Q, Xia X, Chen W, Chen L, Zou M, Liao J, Gao

409 Q, Fang X, Yin Y, Yang H, Wang J, Han L, Lin Y, Lu Y, Zhuang M. Variation among
410 532 genomes unveils the origin and evolutionary history of a global insect herbivore.
411 *Nature communications*. 2020 May 8;11(1):1-8.
412

413 3) Nguyen TN, Mendez V, **Ward CM**, Crisp P, Papanicolaou A, Choo A, Taylor PW,
414 Baxter SW. Disruption of duplicated yellow genes in *Bactrocera tryoni* modifies
415 pigmentation colouration and impacts behaviour. *Journal of Pest Science*. 2020 Nov
416 24:1-6.
417
418

419 **Appendix B:**

420 Supplementary figures and tables for Chapter 6.

421

422 **Appendix C:**

423 Supplementary tables for Chapter 7.

424

425 **References:**

- 426 Acuna-Hidalgo, R., J. A. Veltman, and A. Hoischen. 2016. New insights into the
427 generation and role of de novo mutations in health and disease. *Genome*
428 *Biology* **17**:241.
- 429 Alqudah, A. M., R. Koppolu, G. M. Wolde, A. Graner, and T. Schnurbusch. 2016.
430 The genetic architecture of barley plant stature. *Frontiers in Genetics* **7**:117.
- 431 Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U.
432 Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and
433 genetic mapping using sequenced RAD markers. *PLOS ONE* **3**:e3376.
- 434 Barrett, R. D., and D. Schluter. 2008. Adaptation from standing genetic variation.
435 *Trends in ecology & evolution* **23**:38-44.
- 436 Baxter, S. W., M. Chen, A. Dawson, J.-Z. Zhao, H. Vogel, A. M. Shelton, D. G.
437 Heckel, and C. D. Jiggins. 2010. Mis-Spliced Transcripts of Nicotinic
438 Acetylcholine Receptor $\alpha 6$ Are Associated with Field Evolved Spinosad
439 Resistance in *Plutella xylostella* (L.). *PLOS Genetics* **6**:e1000802.
- 440 Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, C. D.
441 Jiggins, and M. L. Blaxter. 2011. Linkage Mapping and Comparative
442 Genomics Using Next-Generation RAD Sequencing of a Non-Model
443 Organism. *PLOS ONE* **6**:e19315.
- 444 Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An expanded view of complex traits:
445 from polygenic to omnigenic. *Cell* **169**:1177-1186.
- 446 Cadieu, E., M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M.
447 VonHoldt, A. Rhue, A. Boyko, A. Byers, A. Wong, D. S. Mosher, A. G.
448 Elkahoun, T. C. Spady, C. André, K. G. Lark, M. Cargill, C. D. Bustamante, R.
449 K. Wayne, and E. A. Ostrander. 2009. Coat Variation in the Domestic Dog Is
450 Governed by Variants in Three Genes. *Science* **326**:150.
- 451 Charlesworth, B. 1992. Evolutionary rates in partially self-fertilizing species. *The*
452 *American Naturalist* **140**:126-148.

- 453 Chu, B. B., K. L. Keys, C. A. German, H. Zhou, J. J. Zhou, E. M. Sobel, J. S.
454 Sinsheimer, and K. Lange. 2020. Iterative hard thresholding in genome-wide
455 association studies: Generalized linear models, prior weights, and double
456 sparsity. *GigaScience* **9**.
- 457 Crow, J. F. 1957. Genetics of Insect Resistance to Chemicals. *Annual Review of*
458 *Entomology* **2**:227-246.
- 459 Csilléry, K., A. Rodríguez-Verdugo, C. Rellstab, and F. Guillaume. 2018. Detecting
460 the genomic signal of polygenic adaptation and the role of epistasis in
461 evolution. *Molecular Ecology* **27**:606-612.
- 462 Darwin, C. 1859. On the origin of species by means of natural selection, or, the
463 preservation of favoured races in the struggle for life / by Charles Darwin.
464 M.A., Fellow of the Royal, Geological, Linnæan, etc., societies ; author of
465 'Journal of researches during H.M.S. Beagle's voyage round the world". John
466 Murray, London.
- 467 Davey, J. W., and M. L. Blaxter. 2010. RADSeq: next-generation population
468 genetics. *Briefings in Functional Genomics* **9**:416-423.
- 469 Edelman, N. B., P. B. Frandsen, M. Miyagi, B. Clavijo, J. Davey, R. B. Dikow, G.
470 García-Accinelli, S. M. Van Belleghem, N. Patterson, and D. E. Neafsey.
471 2019. Genomic architecture and introgression shape a butterfly radiation.
472 *Science* **366**:594-599.
- 473 Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H.
474 Nadeau. 2010. Missing heritability and strategies for finding the underlying
475 causes of complex disease. *Nature Reviews Genetics* **11**:446-450.
- 476 Endersby, N. M., S. W. McKechnie, P. M. Ridland, and A. R. Weeks. 2006.
477 Microsatellites reveal a lack of structure in Australian populations of the
478 diamondback moth, *Plutella xylostella* (L.). *Molecular Ecology* **15**:107-118.
- 479 Endler, J. A. 1986. Natural selection in the wild. Princeton University Press.
- 480 Fagny, M., J. N. Paulson, M. L. Kuijjer, A. R. Sonawane, C.-Y. Chen, C. M. Lopes-
481 Ramos, K. Glass, J. Quackenbush, and J. Platig. 2017. Exploring regulation in
482 tissues with eQTL networks. *Proceedings of the National Academy of*
483 *Sciences* **114**:E7841.
- 484 Feyereisen, R. 1995. Molecular biology of insecticide resistance. *Toxicology Letters*
485 **82-83**:83-90.
- 486 Fisher, R. A. 1918. XV.—The Correlation between Relatives on the Supposition of
487 Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*
488 **52**:399-433.
- 489 Fisher, R. A. 1930. The genetical theory of natural selection. Oxford University
490 Press, Oxford.
- 491 Fountain, T., M. Ravinet, R. Naylor, K. Reinhardt, and R. K. Butlin. 2016. A Linkage
492 Map and QTL Analysis for Pyrethroid Resistance in the Bed Bug *Cimex*
493 *lectularius*. *G3: Genes|Genomes|Genetics* **6**:4059-4066.
- 494 Gahan, L. J., Y. Pauchet, H. Vogel, and D. G. Heckel. 2010. An ABC Transporter
495 Mutation Is Correlated with Insect Resistance to *Bacillus thuringiensis* Cry1Ac
496 Toxin. *PLOS Genetics* **6**:e1001248.
- 497 Gilman, R. T., and G. M. Kozak. 2015. Learning to speciate: The biased learning of
498 mate preferences promotes adaptive radiation. *Evolution; International*
499 *Journal of Organic Evolution* **69**:3004-3012.
- 500 Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in
501 domestic animals and their use in breeding programmes. *Nature Reviews*
502 *Genetics* **10**:381-391.

503 Green, L., P. Battlay, A. Fournier-Level, R. T. Good, and C. Robin. 2019. Cis- and
504 trans-acting variants contribute to survivorship in a naïve *Drosophila*
505 melanogaster population exposed to ryanoid insecticides. *Proceedings of the*
506 *National Academy of Sciences* **116**:10424.

507 Gu, X. H., D. L. Jiang, Y. Huang, B. J. Li, C. H. Chen, H. R. Lin, and J. H. Xia. 2018.
508 Identifying a Major QTL Associated with Salinity Tolerance in Nile Tilapia
509 Using QTL-Seq. *Marine Biotechnology*.

510 Guo, Y., S. Fudali, J. Gimeno, P. DiGennaro, S. Chang, V. M. Williamson, D. M.
511 Bird, and D. M. Nielsen. 2017. Networks Underpinning Symbiosis Revealed
512 Through Cross-Species eQTL Mapping. *Genetics* **206**:2175-2184.

513 Hawkins, N. J., C. Bass, A. Dixon, and P. Neve. 2018. The evolutionary origins of
514 pesticide resistance. *Biological reviews of the Cambridge Philosophical*
515 *Society* **94**:135-155.

516 Henniges-Janssen, K., A. Reineke, D. G. Heckel, and A. T. Groot. 2011. Complex
517 inheritance of larval adaptation in *Plutella xylostella* to a novel host plant.
518 *Heredity* **107**:421-432.

519 Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: molecular population
520 genetics of adaptation from standing genetic variation. *Genetics* **169**:2335-
521 2352.

522 Hong, E. P., and J. W. Park. 2012. Sample size and statistical power calculation in
523 genetic association studies. *Genomics & informatics* **10**:117-122.

524 Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, A. M. Tarone, L. Turlapati,
525 T. Zichner, D. Zhu, and R. F. Lyman. 2014. Natural variation in genome
526 architecture among 205 *Drosophila melanogaster* Genetic Reference Panel
527 lines. *Genome research* **24**:1193-1208.

528 Huerta-Sánchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang,
529 X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, Huasang, J. Luosang, Z. X. P.
530 Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J.
531 Wang, J. Wang, and R. Nielsen. 2014. Altitude adaptation in Tibetans caused
532 by introgression of Denisovan-like DNA. *Nature* **512**:194-197.

533 Jamann, T. M., P. J. Balint-Kurti, and J. B. Holland. 2015. QTL Mapping Using High-
534 Throughput Sequencing. Pages 257-285 in J. M. Alonso and A. N.
535 Stepanova, editors. *Plant Functional Genomics: Methods and Protocols*.
536 Springer New York, New York, NY.

537 James, J. 1965. Simultaneous selection for dominant and recessive mutants.
538 *Heredity* **20**:142-144.

539 Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R.
540 Swofford, M. Pirun, M. C. Zody, and S. White. 2012. The genomic basis of
541 adaptive evolution in threespine sticklebacks. *Nature* **484**:55-61.

542 Jones, M. R., L. S. Mills, P. C. Alves, C. M. Callahan, J. M. Alves, D. J. R. Lafferty, F.
543 M. Jiggins, J. D. Jensen, J. Melo-Ferreira, and J. M. Good. 2018. Adaptive
544 introgression underlies polymorphic seasonal camouflage in snowshoe hares.
545 *Science* **360**:1355.

546 Jouraku, A., S. Kuwazaki, K. Miyamoto, M. Uchiyama, T. Kurokawa, E. Mori, M. X.
547 Mori, Y. Mori, and S. Sonoda. 2020. Ryanodine receptor mutations (G4946E
548 and I4790K) differentially responsible for diamide insecticide resistance in
549 diamondback moth, *Plutella xylostella* L. *Insect Biochemistry and Molecular*
550 *Biology* **118**:103308.

551 Kauffman, S. A., and E. D. Weinberger. 1989. The NK model of rugged fitness
552 landscapes and its application to maturation of the immune response. *Journal*
553 *of Theoretical Biology* **141**:211-245.

554 Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**:624-626.

555 Kimura, M. 1983. *The neutral theory and molecular evolution*. Cambridge University
556 Press, Cambridge.

557 Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. Hill,
558 A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection
559 in natural populations. *The American Naturalist* **157**:245-261.

560 Klein, R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K.
561 Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L.
562 Ferris, J. Ott, C. Barnstable, and J. Hoh. 2005. Complement Factor H
563 Polymorphism in Age-Related Macular Degeneration. *Science* **308**:385.

564 Lai, Y.-T., C. K. L. Yeung, K. E. Omland, E.-L. Pang, Y. Hao, B.-Y. Liao, H.-F. Cao,
565 B.-W. Zhang, C.-F. Yeh, C.-M. Hung, H.-Y. Hung, M.-Y. Yang, W. Liang, Y.-C.
566 Hsu, C.-T. Yao, L. Dong, K. Lin, and S.-H. Li. 2019. Standing genetic variation
567 as the predominant source for adaptation of a songbird. *Proceedings of the*
568 *National Academy of Sciences* **116**:2152.

569 Lande, R. 1983. The response to selection on major and minor mutations affecting a
570 metrical trait. *Heredity* **50**:47-65.

571 Lawal, R. A., S. H. Martin, K. Vanmechelen, A. Vereijken, P. Silva, R. M. Al-Atiyat, R.
572 S. Aljumaah, J. M. Mwacharo, D.-D. Wu, Y.-P. Zhang, P. M. Hocking, J.
573 Smith, D. Wragg, and O. Hanotte. 2020. The wild species genome ancestry of
574 domestic chickens. *BMC Biology* **18**:13.

575 Leroy, T., J.-M. Louvet, C. Lalanne, G. Le Provost, K. Labadie, J.-M. Aury, S.
576 Delzon, C. Plomion, and A. Kremer. 2020. Adaptive introgression as a driver
577 of local adaptation to climate in European white oaks. *New Phytologist*
578 **226**:1171-1182.

579 Liu, Y., X. Liu, Z. Zheng, T. Ma, Y. Liu, H. Long, H. Cheng, M. Fang, J. Gong, X. Li,
580 S. Zhao, and X. Xu. 2020. Genome-wide analysis of expression QTL (eQTL)
581 and allele-specific expression (ASE) in pig muscle identifies candidate genes
582 for meat quality traits. *Genetics Selection Evolution* **52**:59.

583 Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S.
584 Casillas, Y. Han, M. M. Magwire, and J. M. Cridland. 2012. The *Drosophila*
585 *melanogaster* genetic reference panel. *Nature* **482**:173-178.

586 Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in ecology &*
587 *evolution* **20**:229-237.

588 Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter,
589 M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti. 2009. Finding
590 the missing heritability of complex diseases. *Nature* **461**:747-753.

591 Marburger, S., P. Monnahan, P. J. Seear, S. H. Martin, J. Koch, P. Paajanen, M.
592 Bohutínská, J. D. Higgins, R. Schmickl, and L. Yant. 2019. Interspecific
593 introgression mediates adaptation to whole genome duplication. *Nature*
594 *Communications* **10**:5218.

595 Mariac, C., T. Robert, C. Allinne, M.-S. Remigereau, A. Luxereau, M. Tidjani, O.
596 Seyni, G. Bezançon, J.-L. Pham, and A. Sarr. 2006. Genetic diversity and
597 gene flow among pearl millet crop/weed complex: a case study. *Theoretical*
598 *and Applied Genetics* **113**:1003-1014.

599 Marouli, E., M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood, T. R. Kjaer, R. S.
600 Fine, Y. Lu, C. Schurmann, and H. M. Highland. 2017. Rare and low-
601 frequency coding variants alter human adult height. *Nature* **542**:186-190.

602 Martin, S. H., J. W. Davey, and C. D. Jiggins. 2015. Evaluating the Use of ABBA–
603 BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*
604 **32**:244-257.

605 Martin, S. H., and C. D. Jiggins. 2017. Interpreting the genomic landscape of
606 introgression. *Current Opinion in Genetics & Development* **47**:69-74.

607 Martin, S. H., and S. M. Van Belleghem. 2017. Exploring evolutionary relationships
608 across the genome using topology weighting. *Genetics* **206**:429-438.

609 Mavárez, J., C. A. Salazar, E. Bermingham, C. Salcedo, C. D. Jiggins, and M.
610 Linares. 2006. Speciation by hybridization in *Heliconius* butterflies. *Nature*
611 **441**:868-871.

612 Maynard Smith, J. 1982. *Evolution now: a century after Darwin*. W. H. Freeman and
613 Co., San Francisco.

614 McKenzie, J., and P. Batterham. 1998. Predicting insecticide resistance:
615 mutagenesis, selection and response. *Philosophical Transactions of the Royal*
616 *Society of London. Series B: Biological Sciences* **353**:1729-1734.

617 Merrill, R. M., P. Rastas, S. H. Martin, M. C. Melo, S. Barker, J. Davey, W. O.
618 McMillan, and C. D. Jiggins. 2019. Genetic dissection of assortative mating
619 behavior. *PLOS Biology* **17**:e2005902.

620 Mizuno, A., and Y. Okada. 2019. Biological characterization of expression
621 quantitative trait loci (eQTLs) showing tissue-specific opposite directional
622 effects. *European Journal of Human Genetics* **27**:1745-1756.

623 Mott, R. 2001. *Quantitative Trait Loci (QTL) Mapping Methods*. eLS. John Wiley &
624 Sons, Ltd.

625 Nadachowska-Brzyska, K., R. Burri, P. I. Olason, T. Kawakami, L. Smeds, and H.
626 Ellegren. 2013. Demographic divergence history of pied flycatcher and
627 collared flycatcher inferred from whole-genome re-sequencing data. *PLoS*
628 *Genet* **9**:e1003942.

629 Neafsey, D. E., B. M. Barker, T. J. Sharpton, J. E. Stajich, D. J. Park, E. Whiston, C.-
630 Y. Hung, C. McMahan, J. White, and S. Sykes. 2010. Population genomic
631 sequencing of *Coccidioides* fungi reveals recent hybridization and transposon
632 control. *Genome research* **20**:938-946.

633 Nonaka, E., R. Svanbäck, X. Thibert-Plante, G. Englund, and Å. Brännström. 2015.
634 Mechanisms by Which Phenotypic Plasticity Affects Adaptive Divergence and
635 Ecological Speciation. *The American Naturalist* **186**:E126-E143.

636 Norris, L. C., B. J. Main, Y. Lee, T. C. Collier, A. Fofana, A. J. Cornel, and G. C.
637 Lanzaro. 2015. Adaptive introgression in an African malaria mosquito
638 coincident with the increased usage of insecticide-treated bed nets.
639 *Proceedings of the National Academy of Sciences* **112**:815.

640 Orr, H. A. 1998. The population genetics of adaptation: The distribution of factors
641 fixed during adaptive evolution. *Evolution* **52**:935-949.

642 Orr, H. A. 2005. The genetic theory of adaptation: a brief history. *Nature Reviews*
643 *Genetics* **6**:119-127.

644 Orr, H. A., and A. J. Betancourt. 2001. Haldane's sieve and adaptation from the
645 standing genetic variation. *Genetics* **157**:875-884.

646 Pardo-Diaz, C., C. Salazar, S. W. Baxter, C. Merot, W. Figueiredo-Ready, M. Joron,
647 W. O. McMillan, and C. D. Jiggins. 2012. Adaptive Introgression across
648 Species Boundaries in *Heliconius* Butterflies. *PLOS Genetics* **8**.

649 Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln,
650 E. S. Lander, and S. D. Tanksley. 1991. Mendelian factors underlying
651 quantitative traits in tomato: comparison across species, generations, and
652 environments. *Genetics* **127**:181-197.

653 Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln, and S. D.
654 Tanksley. 1988. Resolution of quantitative traits into Mendelian factors by
655 using a complete linkage map of restriction fragment length polymorphisms.
656 *Nature* **335**:721-726.

657 Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck,
658 T. Webster, and D. Reich. 2012. Ancient admixture in human history.
659 *Genetics* **192**:1065-1093.

660 Perry, K. D., G. J. Baker, K. J. Powis, J. K. Kent, C. M. Ward, and S. W. Baxter.
661 2018. Cryptic *Plutella* species show deep divergence despite the capacity to
662 hybridize. *BMC Evolutionary Biology* **18**:77.

663 Perry, K. D., M. A. Keller, and S. W. Baxter. 2020. Genome-wide analysis of
664 diamondback moth, *Plutella xylostella* L., from Brassica crops and wild host
665 plants reveals no genetic structure in Australia. *Scientific Reports* **10**:12047.

666 Peter, B. M. 2016. Admixture, population structure, and F-statistics. *Genetics*
667 **202**:1485-1501.

668 Peter, B. M., E. Huerta-Sanchez, and R. Nielsen. 2012. Distinguishing between
669 Selective Sweeps from Standing Variation and from a De Novo Mutation.
670 *PLOS Genetics* **8**:e1003011.

671 Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012.
672 Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery
673 and Genotyping in Model and Non-Model Species. *PLOS ONE* **7**:e37135.

674 Pritchard, J. K., and A. Di Rienzo. 2010. Adaptation—not by sweeps alone. *Nature*
675 *Reviews Genetics* **11**:665-667.

676 Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The genetics of human adaptation:
677 hard sweeps, soft sweeps, and polygenic adaptation. *Current biology*
678 **20**:R208-R215.

679 Qin, C., C. H. Wang, Y. Y. Wang, S. Q. Sun, H. H. Wang, and C. B. Xue. 2018.
680 Resistance to Diamide Insecticides in *Plutella xylostella* (Lepidoptera:
681 Plutellidae): Comparison Between Lab-Selected Strains and Field-Collected
682 Populations. *J Econ Entomol* **111**:853-859.

683 Reid, N. M., D. A. Proestou, B. W. Clark, W. C. Warren, J. K. Colbourne, J. R. Shaw,
684 S. I. Karchner, M. E. Hahn, D. Nacci, and M. F. Oleksiak. 2016. The genomic
685 landscape of rapid repeated evolutionary adaptation to toxic pollution in wild
686 fish. *Science* **354**:1305-1308.

687 Robin, C., R. F. Lyman, A. D. Long, C. H. Langley, and T. F. Mackay. 2002. hairy: a
688 quantitative trait locus for *Drosophila* sensory bristle number. *Genetics*
689 **162**:155-164.

690 Rose, A. M., A. Z. Shah, G. Venturini, A. Krishna, A. Chakravarti, C. Rivolta, and S.
691 S. Bhattacharya. 2016. Transcriptional regulation of PRPF31 gene expression
692 by MSR1 repeat elements causes incomplete penetrance in retinitis
693 pigmentosa. *Scientific Reports* **6**:19450.

694 Schweizer, N., T. Viereckel, C. J. A. Smith-Anttila, K. Nordenankar, E. Arvidsson, S.
695 Mahmoudi, A. Zampera, H. Wärner Jonsson, J. Bergquist, D. Lévesque, Å.
696 Konradsson-Geuken, M. Andersson, S. Dumas, and Å. Wallén-Mackenzie.
697 2016. Reduced Vglut2/Slc17a6 Gene Expression Levels throughout the
698 Mouse Subthalamic Nucleus Cause Cell Loss and Structural Disorganization

699 Followed by Increased Motor Activity and Decreased Sugar Consumption.
700 eNeuro **3**:ENEURO.0264-0216.2016.

701 Sheck, A., and F. Gould. 1996. The genetic basis of differences in growth and
702 behavior of specialist and generalist herbivore species: selection on hybrids of
703 *Heliothis virescens* and *Heliothis subflexa* (Lepidoptera). *Evolution* **50**:831-
704 841.

705 Shelton, A. M., J. A. Wyman, N. L. Cushing, K. Apfelbeck, T. J. Dennehy, S. E. R.
706 Mahr, and S. D. Eigenbrode. 1993. Insecticide Resistance of Diamondback
707 Moth (Lepidoptera: Plutellidae) in North America. *Journal of economic*
708 *entomology* **86**:11-19.

709 Sheng, Z., M. E. Pettersson, C. F. Honaker, P. B. Siegel, and Ö. Carlborg. 2015.
710 Standing genetic variation as a major contributor to adaptation in the Virginia
711 chicken lines selection experiment. *Genome Biology* **16**:219.

712 Shrimpton, A., and A. Robertson. 1988. The isolation of polygenic factors controlling
713 bristle score in *Drosophila melanogaster*. II. Distribution of third chromosome
714 bristle effects within chromosome sections. *Genetics* **118**:445-459.

715 Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent,
716 A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A.
717 Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D.
718 Meyre, C. Polychronakos, and P. Froguel. 2007. A genome-wide association
719 study identifies novel risk loci for type 2 diabetes. *Nature* **445**:881-885.

720 Smith, J., and M. R. Kronforst. 2013. Do *Heliconius* butterfly species exchange
721 mimicry alleles? *Biology Letters* **9**:20130503.

722 Solberg Woods, L. C. 2014. QTL mapping in outbred populations: successes and
723 challenges. *Physiological Genomics* **46**:81-90.

724 Song, Y., S. Endepols, N. Klemann, D. Richter, F. R. Matuschka, C. H. Shih, M. W.
725 Nachman, and M. H. Kohn. 2011. Adaptive Introgression of Anticoagulant
726 Rodent Poison Resistance by Hybridization between Old World Mice. *Curr.*
727 *Biol.* **21**:1296-1301.

728 Stefansson, H., R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D.
729 Rujescu, T. Werge, O. P. H. Pietiläinen, O. Mors, P. B. Mortensen, E.
730 Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason,
731 T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Børglum, A. Hartmann,
732 A. Fink-Jensen, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, Y.
733 Böttcher, J. Olesen, R. Breuer, H.-J. Möller, I. Giegling, H. B. Rasmussen, S.
734 Timm, M. Mattheisen, I. Bitter, J. M. Réthelyi, B. B. Magnusdottir, T.
735 Sigmundsson, P. Olason, G. Masson, J. R. Gulcher, M. Haraldsson, R.
736 Fossdal, T. E. Thorgeirsson, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, B.
737 Franke, E. Strengman, L. A. Kiemeny, R. S. Kahn, D. H. Linszen, J. van Os,
738 D. Wiersma, R. Bruggeman, W. Cahn, L. de Haan, L. Krabbendam, I. Myin-
739 Germeys, I. Melle, S. Djurovic, L. Abramova, V. Kaleda, J. Sanjuan, R. de
740 Frutos, E. Bramon, E. Vassos, G. Fraser, U. Ettinger, M. Picchioni, N. Walker,
741 T. Touloupoulou, A. C. Need, D. Ge, J. Lim Yoon, K. V. Shianna, N. B. Freimer,
742 R. M. Cantor, R. Murray, A. Kong, V. Golimbet, A. Carracedo, C. Arango, J.
743 Costas, E. G. Jönsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nöthen, M.
744 Rietschel, P. M. Matthews, P. Muglia, L. Peltonen, D. St Clair, D. B. Goldstein,
745 K. Stefansson, D. A. Collier, R. †Genetic, and P. Outcome in. 2009. Common
746 variants conferring risk of schizophrenia. *Nature* **460**:744-747.

747 Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P.
748 Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman,

749 A. Young, T. Sprosen, T. Peakman, and R. Collins. 2015. UK Biobank: An
750 Open Access Resource for Identifying the Causes of a Wide Range of
751 Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**:e1001779.
752 Suzuki, T. K. 2017. On the Origin of Complex Adaptive Traits: Progress Since the
753 Darwin Versus Mivart Debate. *J Exp Zool B Mol Dev Evol* **328**:304-320.
754 Turner, J. R. 1981. Adaptation and evolution in *Heliconius*: a defense of
755 NeoDarwinism. *Annual Review of Ecology and Systematics* **12**:99-121.
756 Valencia-Montoya, W. A., S. Elfekih, H. L. North, J. I. Meier, I. A. Warren, W. T. Tay,
757 K. H. J. Gordon, A. Specht, S. V. Paula-Moraes, R. Rane, T. K. Walsh, and C.
758 D. Jiggins. 2020. Adaptive Introgression across Semipermeable Species
759 Boundaries between Local *Helicoverpa zea* and Invasive *Helicoverpa*
760 *armigera* Moths. *Molecular Biology and Evolution* **37**:2568-2583.
761 van't Hof, A. E., N. Edmonds, M. Dalíková, F. Marec, and I. J. Saccheri. 2011.
762 Industrial Melanism in British Peppered Moths Has a Singular and Recent
763 Mutational Origin. *Science* **332**:958-960.
764 Wallbank, R. W. R., S. W. Baxter, C. Pardo-Diaz, J. J. Hanly, S. H. Martin, J. Mallet,
765 K. K. Dasmahapatra, C. Salazar, M. Joron, N. Nadeau, W. O. McMillan, and
766 C. D. Jiggins. 2016. Evolutionary Novelty in a Butterfly Wing Pattern through
767 Enhancer Shuffling. *PLOS Biology* **14**:e1002353.
768 Wang, J., Z. Wang, X. Du, H. Yang, F. Han, Y. Han, F. Yuan, L. Zhang, S. Peng, and
769 E. Guo. 2017. A high-density genetic map and QTL analysis of agronomic
770 traits in foxtail millet [*Setaria italica* (L.) P. Beauv.] using RAD-seq. *PLOS*
771 *ONE* **12**:e0179717.
772 Wang, L., Z. Y. Wan, B. Bai, S. Q. Huang, E. Chua, M. Lee, H. Y. Pang, Y. F. Wen,
773 P. Liu, F. Liu, F. Sun, G. Lin, B. Q. Ye, and G. H. Yue. 2015. Construction of a
774 high-density linkage map and fine mapping of QTL for growth in Asian
775 seabass. *Scientific Reports* **5**:16358.
776 Wang, S.-B., J.-Y. Feng, W.-L. Ren, B. Huang, L. Zhou, Y.-J. Wen, J. Zhang, J. M.
777 Dunwell, S. Xu, and Y.-M. Zhang. 2016. Improving power and accuracy of
778 genome-wide association studies via a multi-locus mixed linear model
779 methodology. *Scientific Reports* **6**:19444.
780 Ward, C. M., R. A. Aumann, M. A. Whitehead, K. Nikolouli, G. Leveque, G. Gouvi, E.
781 Fung, S. J. Reiling, H. Djambazian, M. A. Hughes, S. Whiteford, C. Caceres-
782 Barrios, T. N. M. Nguyen, A. Choo, P. Crisp, S. B. Sim, S. M. Geib, F. Marec,
783 I. Häcker, J. Ragoussis, A. C. Darby, K. Bourtzis, S. W. Baxter, and M. F.
784 Schetelig. 2021. White pupae phenotype of tephritids is caused by parallel
785 mutations of a MFS transporter. *Nature Communications* **12**:491.
786 Ward, C. M., and S. W. Baxter. 2018. Assessing Genomic Admixture between
787 Cryptic *Plutella* Moth Species following Secondary Contact. *Genome Biology*
788 *and Evolution* **10**:2973-2985.
789 Ward, C. M., A. J. Ludington, J. Breen, and S. W. Baxter. 2020. Genomic
790 evolutionary analysis in R with *geaR*. *bioRxiv*:2020.2008.2006.240754.
791 Ward, C. M., T.-H. To, and S. M. Pederson. 2019. *ngsReports*: a Bioconductor
792 package for managing FastQC reports and other NGS related log files.
793 *Bioinformatics* **36**:2587-2588.
794 Ward, L. D., and M. Kellis. 2012. Interpreting noncoding genetic variation in complex
795 traits and human disease. *Nature Biotechnology* **30**:1095-1106.
796 Wynn, J. 2007. Alone in the Garden: How Gregor Mendel's Inattention to Audience
797 May Have Affected the Reception of His Theory of Inheritance in
798 "Experiments in Plant Hybridization". *Written Communication* **24**:3-27.

799 Xia, X., B. Sun, G. M. Gurr, L. Vasseur, M. Xue, and M. You. 2018. Gut Microbiota
800 Mediate Insecticide Resistance in the Diamondback Moth, *Plutella xylostella*
801 (L.). *Frontiers in Microbiology* **9**.

802 Zalucki, M. P., A. Shabbir, R. Silva, D. Adamson, L. Shu-Sheng, and M. J. Furlong.
803 2012. Estimating the economic cost of one of the world's major insect pests,
804 *Plutella xylostella* (Lepidoptera: Plutellidae): just how long is a piece of string?
805 *J Econ Entomol* **105**:1115-1129.

806 Zamani, N., P. Russell, H. Lantz, M. P. Hoepfner, J. R. S. Meadows, N. Vijay, E.
807 Mauceli, F. di Palma, K. Lindblad-Toh, P. Jern, and M. G. Grabherr. 2013.
808 Unsupervised genome-wide recognition of local relationship patterns. *BMC*
809 *Genomics* **14**:347.

810 Zan, Y., Z. Sheng, M. Lillie, L. Rönnegård, C. F. Honaker, P. B. Siegel, and Ö.
811 Carlborg. 2017. Artificial Selection Response due to Polygenic Adaptation
812 from a Multilocus, Multiallelic Genetic Architecture. *Molecular Biology and*
813 *Evolution* **34**:2678-2689.

814 Zhang, Y.-W., C. Lwaka Tamba, Y.-J. Wen, P. Li, W.-L. Ren, Y.-L. Ni, J. Gao, and
815 Y.-M. Zhang. 2020. mrMLM v4.0: An R Platform for Multi-locus Genome-wide
816 Association Studies. *Genomics, Proteomics & Bioinformatics*.

817 Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J.
818 Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler. 2010. Mixed
819 linear model approach adapted for genome-wide association studies. *Nature*
820 *Genetics* **42**:355-360.

821 Zimmerman, E., A. Palsson, and G. Gibson. 2000. Quantitative trait loci affecting
822 components of wing shape in *Drosophila melanogaster*. *Genetics* **155**:671-
823 683.

824 Zuo, Y.-Y., H.-H. Ma, W.-J. Lu, X.-L. Wang, S.-W. Wu, R. Nauen, Y.-D. Wu, and Y.-
825 H. Yang. 2020. Identification of the ryanodine receptor mutation I4743M and
826 its contribution to diamide insecticide resistance in *Spodoptera exigua*
827 (Lepidoptera: Noctuidae). *Insect Science* **27**:791-800.
828

Chapter 2

ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files.

Ward, C. M., To, T. H., & Pederson, S. M. (2020). ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. **Bioinformatics**, 36(8), 2587-2588.

Statement of Authorship

Title of Paper	ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	<i>Bioinformatics</i> , Volume 36, Issue 8, 15 April 2020

Principal Author

Name of Principal Author (Candidate)	Christopher Ward		
Contribution to the Paper	Contributed to package design and code. Wrote the first version of the manuscript and reviewed subsequent versions.		
Overall percentage (%)	40		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	07/04/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Thu-Hien To		
Contribution to the Paper	Contributed code. 10%		
Signature		Date	03 April 2021

Name of Co-Author	Stephen Pederson		
Contribution to the Paper	Conceived package. Contributed to package design and code. Reviewed manuscript. 50%		
Signature		Date	7/04/21

Please cut and paste additional co-author panels here as required.

Sequence analysis

ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files

Christopher M. Ward¹, Thu-Hien To² and Stephen M. Pederson ^{2,*}

¹Department of Molecular and Biomedical Science and ²Bioinformatics Hub, School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 7, 2019; revised on October 17, 2019; editorial decision on December 11, 2019; accepted on December 12, 2019

Abstract

Motivation: High throughput next generation sequencing (NGS) has become exceedingly cheap, facilitating studies to be undertaken containing large sample numbers. Quality control (QC) is an essential stage during analytic pipelines and the outputs of popular bioinformatics tools such as FastQC and Picard can provide information on individual samples. Although these tools provide considerable power when carrying out QC, large sample numbers can make inspection of all samples and identification of systemic bias a challenge.

Results: We present ngsReports, an R package designed for the management and visualization of NGS reports from within an R environment. The available methods allow direct import into R of FastQC reports along with outputs from other tools. Visualization can be carried out across many samples using default, highly customizable plots with options to perform hierarchical clustering to quickly identify outlier libraries. Moreover, these can be displayed in an interactive shiny app or HTML report for ease of analysis.

Availability and implementation: The ngsReports package is available on Bioconductor and the GUI shiny app is available at <https://github.com/UofABioinformaticsHub/shinyNgsreports>.

Contact: stephen.pederson@adelaide.edu.au

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The next generation sequencing (NGS) boom has provided researchers unparalleled resources to answer fundamental questions in environmental, agricultural and biomedical research. As the cost of sequencing has decreased steadily, there has been a dramatic increase in data being generated, leading to many challenges for data handling.

Quality control (QC) of raw and processed data is arguably the most important stage in bioinformatic analysis and pipeline optimization. One of the most heavily used tools to aid in the identification of systematic and user-induced bias for NGS data is FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), which provides a series of detailed statistics and plots for an individual fastq or bam file. In addition, many aligners, such as Bowtie2 (Langmead and Salzberg, 2012), HISAT2 (Kim *et al.*, 2015) and STAR (Dobin *et al.*, 2013), along with other tools such as cutadapt (Martin, 2011) and AdapterRemoval (Schubert *et al.*, 2016) form integral parts of processing pipelines and provide useful summaries to stdout/stderr which can be saved to disk. However, collating and interpreting such output across multiple files can be tedious and prone to human error, as most tools output these logs on a per-sample basis. The decrease in cost and increase in popularity of

NGS-based experimental designs, further contributes to the growing need for software to collate, manipulate and visualize QC logs and reports across experiments with large sample numbers.

Few software tools are currently available to tackle this problem, with the notable exception of MultiQC (Ewels *et al.*, 2016). Although highly useful at investigating multiple samples, custom summaries and reports are often required for researchers working within an R environment. The R package fastqcr (<https://github.com/kassambara/fastqcr>) provides some of this functionality; however, import is limited to FastQC logs and visualization is difficult for many samples at the same time. Therefore, we present ngsReports, an R package to transparently parse FastQC output and common NGS log files into the R environment. We seek to provide a platform to easily and accurately compare outputs at various processing stages interactively, and allow simple, customizable report generation during standard data processing pipelines.

2 Materials and methods

2.1 Data analysis

To aid in access, manipulation and storage of data, we created a novel S4 class for single FastQC reports (FastqcData), while



Fig. 1. Screenshot from the additional shiny app, which provides interactive and customizable plots for each FastQC module. FastQC reports can be input directly into the shiny app, with the option to output an HTML report. Here, we show a clustered heatmap with dendrogram displaying the difference between the observed GC content and the estimated theoretical GC content for the *Homo sapiens* genome. The sidebar of the plot shows the PASS/WARN/FAIL status attributed to the ‘GC content’ module by FastQC. The line plot for a single sample can be obtained by clicking a single sample in the shiny app, to provide more detailed information about any samples of concern

multiple reports are collated into list-like extensions (FastqcDataList). S4 objects can be passed directly into plotting functions for simple generation of static and interactive figures, with the latter being enabled under the plotly framework (<https://plotly-r.com>). Raw data for each FastQC module is easily accessible by passing a FastqcDataList into the function `getModule()` and can then be freely manipulated using R packages such as those in the tidyverse (Wickham et al., 2019).

We also provide a simple wrapper function, `importNgsLogs()`, to automatically parse and import log files of popular bioinformatics tools such as: cutadapt (Martin, 2011), Trimmomatic (Bolger et al., 2014), Bowtie (Langmead et al., 2009) & Bowtie 2 (Langmead and Salzberg, 2012), Picard: Mark Duplicates (<https://broadinstitute.github.io/picard/>), STAR (Dobin et al., 2013), samtools flagstat (Li et al., 2009) and BUSCO (Waterhouse et al., 2018).

2.2 Visualization

Highly customizable default plotting functions for each module of FastQC are currently supported. Plots for each module can be generated for single or multiple files in either line (Fig. 1 and Supplementary Fig. S1) or heatmap format (Fig. 1 and Supplementary Fig. S2). Heatmaps provide an intuitive method to identify differential patterns between libraries which may be obscured when viewing single files or line plots alone. In order to aid in identification of clusters, hierarchical clustering can be carried out allowing for fast identification of outlier libraries and batch effects (Fig. 1). When the number of samples is too large to produce an informative heatmap, principle component analysis can be used for individual FastQC modules (Supplementary Fig. S3). Furthermore, libraries can be grouped using hierarchical clustering from the factoMineR package (Lê et al., 2008) or by passing user-specified groups such as libraries sequenced on the same lane.

Imported log files from commonly used alignment tools can be visualized as bar plots (Supplementary Fig. S4) using `plotAlignmentSummary()`. Likewise assembly statistics generated using quast (Gurevich et al., 2013) can also be assessed as a parallel coordinate plot using `plotAssemblyStats()` to quickly compare multiple assemblies (Supplementary Fig. S5). A list of supported tools for each function can be found in the packages github README (<https://github.com/UofABioinformaticsHub/ngsReports>).

By default, plots are rendered using ggplot2 (Wickham, 2016) and remain compatible with the addition of themes, annotations and geoms using standard ggplot2 nomenclature. All plots can additionally be produced as interactive html plots with the `usePlotly` argument providing hover text and zoom functionality.

2.3 Shiny app and HTML report

Modules in FastQC reports can be aggregated into a single HTML report using the function `writeHtmlReport()` and the default template (Supplementary File S1), or a user-specified RMarkdown file. In addition to the HTML report, we also provide a GUI written using RShiny (<https://github.com/UofABioinformaticsHub/shinyNgsreports>). This provides an interactive user interface to view and customize plots for each module within the FastQC report (Fig. 1). Furthermore, the shiny app displays arguments for clustering, zooming, data type and organism selection and on-click inspection of individual line plots for libraries (Fig. 1).

3 Conclusion

Costs involved in library preparation and sequencing for NGS applications are decreasing, allowing for much larger studies to be undertaken. The functional programming capabilities of the R programming language provide an ideal environment for data manipulation and analysis. The methods provided in ngsReports constitute a powerful tool for generic and bespoke aggregation, analysis and visualization of NGS QC and log data.

Acknowledgements

The authors would like to thank James Breen, Alastair Ludington and Pei Qin Ng for early usage and testing, Dan Kortschak for early discussions on package design and the manuscript reviewers for providing valuable feedback on both the package and manuscript.

Funding

C.M.W. is supported by The Commonwealth Hill Trust and The Grains Research Development Corporation (Grant 9175870).

Conflict of Interest: none declared.

References

- Bolger,A.M. et al. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.
- Dobin,A. et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.
- Ewels,P. et al. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32, 3047–3048.
- Gurevich,A. et al. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29, 1072–1075.
- Kim,D. et al. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12, 357–360.
- Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9, 357–359.
- Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10, R25.
- Lê,S. et al. (2008) FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.*, 25, 1–18.
- Li,H. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
- Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17, 10.
- Schubert,M. et al. (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes*, 9, 88.
- Waterhouse,R.M. et al. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, 35, 543–548.
- Wickham,H. et al. (2019) Welcome to the Tidyverse. *J. Open Source Softw.*, 4, 1686.
- Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Chapter 3

Genomic evolutionary analysis in R with gear.

Ward, C. M., Ludington, A. J., Breen, J., & Baxter, S. W. (2020). Genomic evolutionary analysis in R with gear. **Unpublished.**

Statement of Authorship

Title of Paper	Genomic evolutionary analysis in R with gear		
Publication Status	<input type="checkbox"/> Published	<input type="checkbox"/> Accepted for Publication	
	<input type="checkbox"/> Submitted for Publication	<input checked="" type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style	
Publication Details	Currently publicly available on bioRxiv Genomic evolutionary analysis in R with gear Christopher M. Ward, Alastair J. Ludington, James Breen, Simon W. Baxter bioRxiv 2020.08.06.240754; doi: https://doi.org/10.1101/2020.08.06.240754		

Principal Author

Name of Principal Author (Candidate)	Christopher M Ward		
Contribution to the Paper	Conceived the research, wrote the package, proposed improvements and reviewed code. Wrote the first version of the manuscript and edited subsequent versions.		
Overall percentage (%)	80		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/5/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alastair Ludington		
Contribution to the Paper	10% Wrote the package, proposed improvements and reviewed code. Edited draft versions of the manuscript.		
Signature		Date	21/04/21

Name of Co-Author	James Breen		
Contribution to the Paper	5% Proposed improvements and reviewed code (with C.W, A.L, and S.B). Edited draft versions of the manuscript.		
Signature		Date	21/4/2021

Name of Co-Author	Simon Baxter
Contribution to the Paper	5% Conceived the research, proposed improvements. Edited draft versions of the manuscript.
Signature	
	Date 20/5/2021

1 **Genomic evolutionary analysis in R with gear.**

2 Christopher M. Ward^{1*}, Alastair J. Ludington¹, James Breen^{2,3,4}, Simon W. Baxter⁵

3 ¹ School of Biological Sciences, University of Adelaide, Australia

4 ² Bioinformatics Platform, South Australian Health & Medical Research Institute (SAHMRI), Australia

5 ³ Robinson Research Institute, University of Adelaide, Australia

6 ⁴ Faculty of Health & Medical Sciences, University of Adelaide, Australia

7 ⁵ Bio21 Institute, School of BioSciences, University of Melbourne, Australia

8 * Corresponding author: christopher.ward@adelaide.edu.au

9

10 **Abstract:**

11 The analysis and interpretation of datasets generated through sequencing large numbers of
12 individual genomes is becoming commonplace in population and evolutionary genetic
13 studies. Here we introduce gearR, a modular R package for evolutionary analysis of genome-
14 wide genotype data. The package leverages the Genomic Data Structure (GDS) format,
15 which enables memory and time efficient querying of genotype datasets compared to
16 standard VCF genotype files. gearR utilizes GRange object classes to partition an analysis
17 based on features from GFF annotation files, select codons based on position or
18 degeneracy, and construct both positional and coordinate genomic windows. Tests of
19 genetic diversity (eg. d_{XY} , π , F_{ST}) and admixture (f_4 , \hat{f}_d) along with tree building and
20 sequence output, can be carried out on partitions using a single function regardless of
21 sample ploidy or number of observed alleles. The package and associated documentation
22 are available on GitHub at <https://github.com/CMWbio/gearR>.

23 Keywords: Evolution, Population Genomics, R Package, Admixture

24 **Introduction:**

25 Improvements in genome sequencing technologies has led to increased production of data
26 at lower relative cost per base (Schwarze et al. 2020). Genome-wide sequencing datasets
27 with hundreds of samples can be produced for population genomic analysis, allowing
28 researchers to investigate population and evolutionary history at an unprecedented scale.
29 However, due to file size and data complexity downstream problems during data storage and
30 analysis can arise. The most common format for handling genome-wide SNP data is the
31 Variant Call Format (VCF), which has historically had a large memory overhead when being
32 read into an R environment. To resolve this, the Genomic Data Structure (GDS) format has
33 allowed all genotype and metadata to be compressed into a queryable, on-disk file that
34 substantially reduce memory requirements and decrease analysis time (Zheng et al. 2017).
35 The GDS format provides an efficient format for filtering SNP data in order to perform
36 Principal Component Analysis, estimate genetic relatedness and tests for genetic
37 association (Zheng et al. 2012).

38 GDS files use GRange objects from the GenomicRanges package (Lawrence et al. 2013) to
39 define loci to query from file and import into R. In their most basic form, GRange class
40 objects define genomic loci based on reference position. Although widely used throughout
41 Bioconductor, GRange objects, to our knowledge, have not been utilized in the same
42 manner to define loci for evolutionary analyses.

43 Few R packages attempt to carry out genome-wide investigation of genotype data. Most
44 packages focus on the analysis of single or multi-locus data, with the notable exception of
45 PopGenome (Pfeifer et al. 2014). However, one limitation of PopGenome is customizability
46 of how the target genome is partitioned, and which sites are selected for analysis. Most
47 tools, including PopGenome, allow datasets to be partitioned into sliding or tiled windows
48 based on reference or SNP position. PopGenome also provides methods to split data into
49 GFF attributes, however selection of bespoke partitions not possible. This makes calculating
50 population metrics on specific codon positions (eg. four-fold or zero-fold degenerate sites) or
51 analysing many non-contiguous loci difficult and time consuming.

52 To overcome these issues, here we present the R package gear, which leverages the GDS
53 format, to efficiently construct GRanges containing genome-wide or local loci of interest and
54 to carry out common tasks for evolutionary analysis on genome-wide genotype data using a
55 single function. Furthermore, we provide methods to partition the genome based on
56 annotation, codon position or degeneracy through utilizing data in GFF files, or using
57 reference genome coordinates or genotype position.

58 **Features:**

59 ***Input data***

60 Genotype input files are required to be in GDS format, enabling high compressibility
61 compared to gzipped VCFs (>5X smaller on disk), efficient querying and the capability to
62 work on large datasets with a reduced memory footprint (Zheng et al. 2017). Conversion of
63 sample genotypes in the VCF format to a Genomic Data Structure (GDS) format can be
64 performed using the SeqArray package (Zheng et al. 2017) before analysis with gear.
65 Genotypes called at any level of ploidy can be utilized in gear, which includes whole
66 genome sequence data generated from pools of two or more individuals.

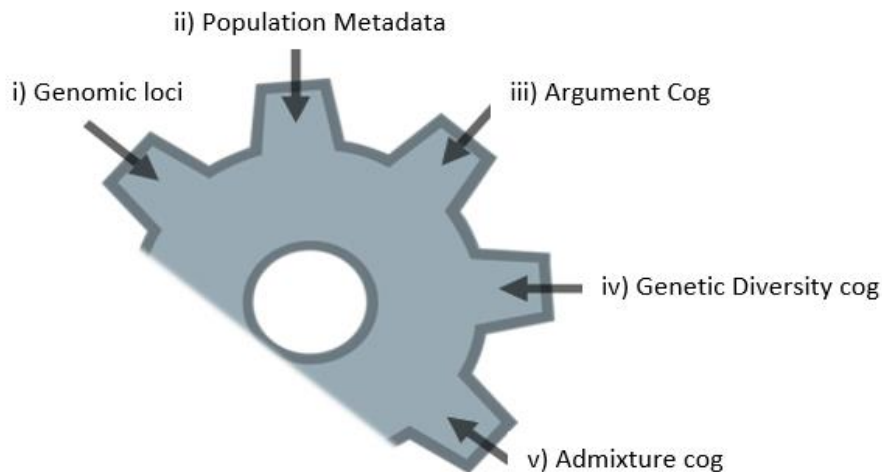
67 ***Partitioning the genome using GRanges***

68 The gear package utilizes GRange objects to define partitions for the analysis, for example,
69 segmenting a genome into 10-kb windows. This allows users to define their own GRanges
70 for the analysis or build them with provided functions. Currently, users are able to generate
71 both coordinate (based on reference coordinate) and positional (based on genotype number)
72 windows using *makeWindows()* or *makeSnpWindows()* functions. Sequence features, such
73 as protein coding regions, can be extracted from a GFF with *getFeatures()*.

74 Many evolutionary analyses seek to calculate population metrics over different codon
75 positions. To make this as simple as possible, gear provides methods to index a reference

76 genome according to codon position with *buildCodonDB()*, which can either be stored in
77 memory as a GRangesList object or an SQLite database (DB) on disk to limit static memory
78 usage. Users may then filter codons based on degeneracy (0-fold or 2-fold) and position
79 using the function *filterCodonDB()*. A codon DB can also be passed to the function
80 *validate4fold()* to select 4-fold degenerate sites across the genome that are empirically
81 supported in the GDS file. This is done by querying the GDS to i) remove codons with
82 missing data, ii) select 4-fold degenerate codons, iii) remove all those where codon positions
83 one or two have variation and iv) select third positions.

84 GRange objects generated using *gear* can then be combined using *mergeLoci()* to further
85 customize partitions. For example, genome-wide tiled windows can be combined with four-
86 fold degenerate sites to output either genomic windows that contain only 4-fold sites or all
87 sites excluding 4-fold degenerate sites.



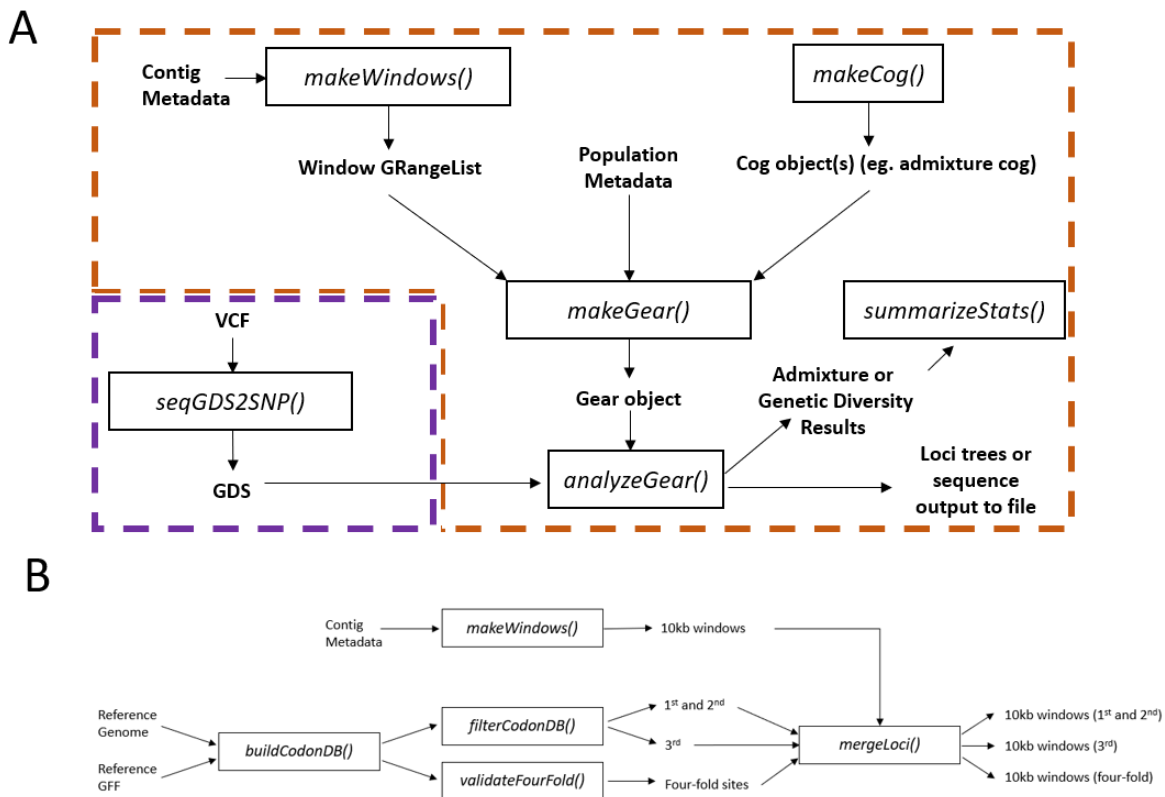
88 Figure 1: Structure of the of the gear S4 object: i) Genomic loci (GRanges) to carry out
89 analyses across, ii) Population metadata encoding sample names to the population/species
90 they belong to, iii) A cog containing general arguments for all analysis, iv) a cog specifying
91 that the Genetic Diversity module should be carried out and v) a cog specifying that the
92 Admixture cog should be carried out on the dataset.

93

94 **Setting up an analysis: cogs and gears**

95 *gear* operates through two S4 classes, the 'cog' and 'gear' (Figure 1). Cogs, built using
96 *makeCog()*, specify multiple analyses to carry out (see Table 1) setting parameters specific
97 to each analysis. A single gear object can then be constructed, using *makeGear()*, which
98 contains all of the specified cogs for analysis, along with the genomic loci and population

99 metadata (Figure 2A). The *analyzeGear()* function then performs all analyses on the same
 100 set of genomic loci and samples, greatly reducing run time compared to sequential
 101 execution.



102

103 Figure 2: A) Basic analysis workflow in *gearR* to carry out analysis on windowed genomic
 104 loci. Functions specific to *gearR* are within the orange boundary and external functions in
 105 purple. After converting the VCF to GDS format using *SeqArray*, contig metadata (contig
 106 length) is used to construct windows across the genome. A dataframe containing population
 107 metadata defining population to sample grouping is then constructed. This is used, along
 108 with windows for the analysis and cogs, to construct the *gear* class object. The analysis is
 109 then carried out on the *gear* object and outputs depend on which cogs were specified. B)
 110 Workflow used to generate partition schemes for examples in Figure 3. First 10kb windows
 111 were generated from contig metadata. This was followed by generation of a codon database
 112 that indexes codon position in a reference genome. The codon database was then passed to
 113 the function *filterCodonDB* to output separate loci-sets for four codon partition schemes:
 114 1st+2nd; 3rd; 0-fold and 2-fold. *validateFourFold()* was also used to select 4-fold degenerate
 115 sites that are supported by genotypes in the GDS file. Each of these codon loci-sets can
 116 then be passed to the function *mergeLoci()*, along with the 10kb windows, to combine loci
 117 into 10kb windows that contain only the selected codon types.

118 **Analysis types**

119 Four different cogs can be generated to carry out an analysis: i) genetic diversity, ii)
 120 admixture, iii) outputLoci and iv) outputTrees. Genetic diversity allows the calculation of a
 121 range of population metrics (Table 1), most of which rely on genetic distance which is
 122 calculated based on the hamming distance between haplotypes at all sites within the locus.
 123 The admixture cog utilizes outgroup polarized allele frequency at all biallelic sites within the
 124 locus to calculate f_4 (Patterson et al. 2012) and \hat{f}_d (Martin et al. 2015) statistics. The
 125 package also enables users to output data in fasta format for each individual (or sample
 126 pool) using outputLoci or as distance trees using outputTrees. Haplotypes are used in
 127 diversity calculations and are output to file according to the phase within the supplied GDS
 128 file, not calculated by gearR.

129 Outputs of both genetic diversity and admixture cogs can be summarized using
 130 `summarizeStats()` which calculates a mean and median values for each statistic across all
 131 loci using a block jack-knife approach.

132 Table 1: Analysis types and functionality available to apply at each locus.

Cog type	Functionality
Genetic diversity	Nucleotide diversity (π), genetic distance (d_{XY}), maximum distance (d_{max}), minimum distance (d_{min}), ancestral distance (d_a), Υ_{ST} , relative node distance (RND), minimum relative node distance (RND_{min}), and G_{min}
Admixture	f_4 and \hat{f}_d
Output loci	<i>fasta</i> format output to file
Output trees	<i>newick</i> format distance trees and metadata output to file

133

134 **Parallelization**

135 All functions allow operations to be run in parallel by leveraging methods in the `furrr`
 136 (<https://github.com/DavisVaughan/furrr>) and parallel R packages.

137 **Handling of missing data:**

138 Missing data in windows is handled in two ways: i) sites can be filtered based on the
 139 percentage of samples with a missing genotype (./.) and ii) windows can be filtered based on

140 the percentage of missing sites out of the total window size. Finally, missing data within the
141 distance matrix are handled as non-informative, whereby sites are removed in a pairwise
142 manner from the distance calculation.

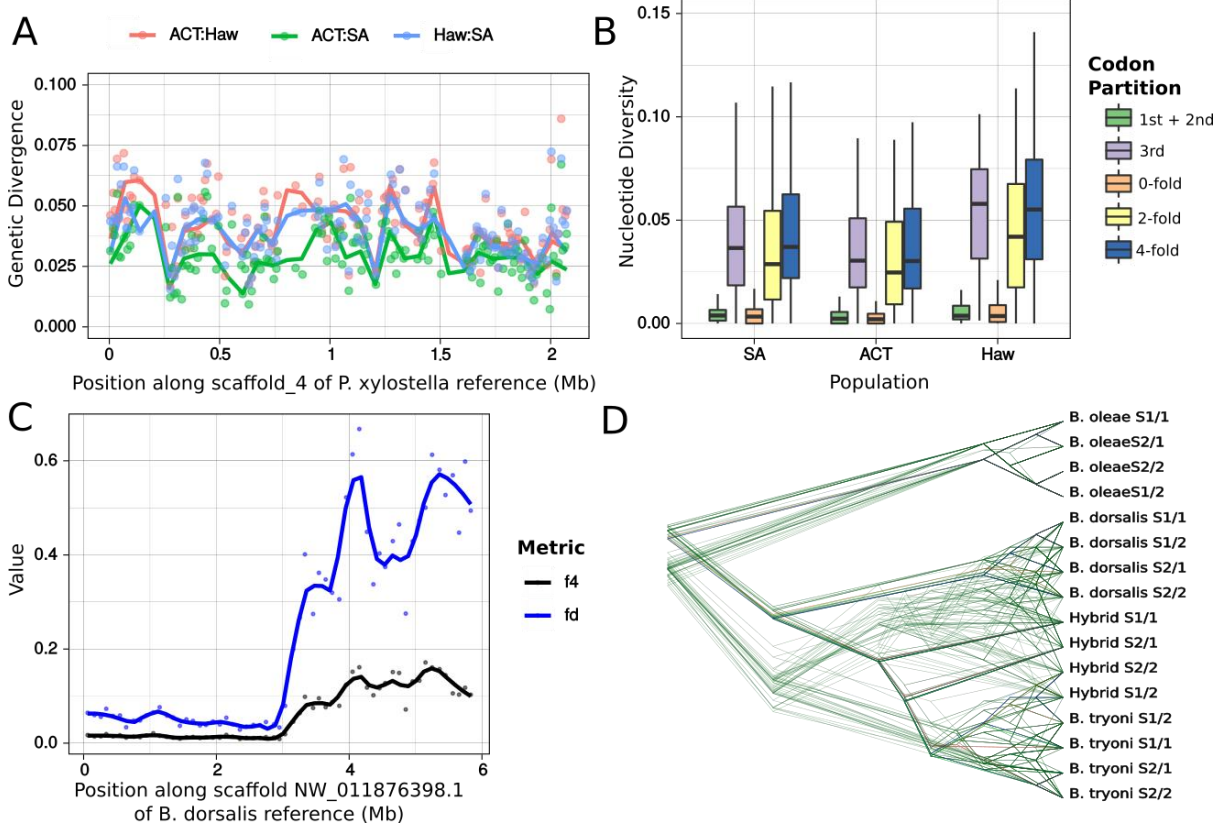
143

144

145 **Carrying out an analysis using gear:**

146 gear has successfully been used to calculate genome wide diversity metrics between
147 populations containing 532 moth genomes (You et al. 2020) and to identify introgressed
148 regions between two *Bactrocera* fly species (Ward et al. 2020). Below we outline two
149 example analyses using gear. Code for each of examples and other common workflows can
150 be found on the wiki (<https://github.com/CMWbio/gear/wiki>).

151 In our first example we use a subset of the data from Ward and Baxter (2018) containing
152 three populations of diamondback moth collected from Australian Capital Territory, Australia;
153 South Australia, Australia and Hawaii, USA. Second. Following the general workflow shown
154 in Figure 2A, we converted the called genotypes to the GDS format using SeqArray,
155 constructed partitions, built cogs, combined those cogs into a gear and then carried out the
156 analysis. We constructed our partitions for the analysis by generating a GRange object
157 containing only scaffold_4. The analysis will use this GRange object to construct six different
158 partition schemes based on 10kb tiled windows (workflow shown in Figure 2B): i) all sites, ii)
159 only 1st+2nd codon positions, iii) windows only 3rd codon positions, iv) only 0-fold degenerate
160 sites, v) only 2-fold degenerate sites and vi) only 4-fold degenerate sites. Partition i) was
161 then used to calculate pairwise genetic distance (d_{XY}) between each population across the
162 scaffold (Figure 3A) and partitions ii-vi) were used to calculate within population nucleotide
163 diversity across the whole genome (Figure 3B).



165

166 Figure 3: Example workflows to carry out with gearR: Panels A and B use *P. xylostella* data
 167 from Ward and Baxter (2018), panels C and D use *Bactrocera* from Ward *et al* (2020). A)
 168 Absolute genetic distance (d_{XY}) was calculated between pairwise comparisons of three
 169 populations for 10kb tiled windows across scaffold_4 of the diamondback moth reference
 170 genome. C) The five loci-sets constructed using the workflow in Figure 2B were used to
 171 calculate nucleotide diversity (π) at 1st+2nd, 3rd, 0-fold, 2-fold and 4-fold codon sites across
 172 scaffold_4 of the diamondback moth reference genome C) Admixture metrics f_4 and \hat{f}_d
 173 calculated on 100kb windows across scaffold NW_011876398.1 of the *B. dorsalis* reference
 174 genome. D) Distance trees for each 100kb window across NW_011876398.1 output using
 175 the *outputTrees* cog showing a mixture of discordant and concordant topologies. Plots A), B)
 176 and C) were generated using ggplot2 (Wickham 2009) and D) using densitree (Bouckaert
 177 2010).

178 For a second example we will identify one of the introgressed regions from Ward *et al.*
 179 (2020). This will use data from a single scaffold (NW_011876398.1) of the *B. dorsalis*
 180 reference genome (GCF_000789215.1) for two samples of *B. tryoni*, *B. dorsalis*, *B. oleae*
 181 and a *B. dorsalis/B. tryoni* hybrid line. Using the same methodology as the first example, we

182 constructed a 100kb tiled window partition scheme. However, for this analysis we used the
183 admixture cog to calculate f_4 and \hat{f}_d admixture metrics showing clear evidence for
184 introgression at the 3' end of the scaffold (Figure 3C). We also used the outputTrees cog to
185 output distance trees for each of these windows to illustrate both the congruent and
186 incongruent topologies resulting from partial admixture on NW_011876398.1 (Figure 3D).

187

188 **Conclusion**

189 Genome-wide datasets with many individuals are becoming the norm in population genetic
190 studies, increasing the need for tools to efficiently carry out analyses on genotype data. The
191 functional programming capabilities of the R programming language provide an intuitive
192 environment for users to carry out calculation and visualization of population and
193 evolutionary genomics metrics. The methods provided in gearR allow users easily and
194 effectively partition the genome for generic and bespoke analysis of genome-wide genotype
195 data regardless of sample ploidy and number of observed alleles.

196

197 **Acknowledgements**

198 We would like to thank Simon Martin for allowing us to use his python scripts as a reference
199 for the first draft of this package. CMW is funded by the Commonwealth Hill Trust and The
200 Grains Research Development Corporation (Grant 9175870).

201 **Data Availability**

202 All data is available in the referenced publications.

203

204 **References**

- 205 Bouckaert, R. R. 2010. DensiTree: making sense of sets of phylogenetic trees.
206 *Bioinformatics* **26**:1372-1373.
- 207 Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan,
208 and V. J. Carey. 2013. Software for computing and annotating genomic ranges.
209 *PLoS computational biology* **9**:e1003118-e1003118.
- 210 Martin, S. H., J. W. Davey, and C. D. Jiggins. 2015. Evaluating the use of ABBA-BABA
211 statistics to locate introgressed loci. *Molecular biology and evolution* **32**:244-257.

212 Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T.
213 Webster, and D. Reich. 2012. Ancient Admixture in Human History. *Genetics*
214 **192**:1065.

215 Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher. 2014. PopGenome: an
216 efficient Swiss army knife for population genomic analyses in R. *Molecular biology*
217 *and evolution* **31**:1929-1936.

218 Schwarze, K., J. Buchanan, J. M. Fermont, H. Dreau, M. W. Tilley, J. M. Taylor, P. Antoniou,
219 S. J. L. Knight, C. Camps, M. M. Pentony, E. M. Kvikstad, S. Harris, N. Popitsch, A.
220 T. Pagnamenta, A. Schuh, J. C. Taylor, and S. Wordsworth. 2020. The complete
221 costs of genome sequencing: a microcosting study in cancer and rare diseases from
222 a single center in the United Kingdom. *Genetics in Medicine* **22**:85-94.

223 Ward, C., R. Aumann, M. Whitehead, K. Nikolouli, G. Leveque, G. Gouvi, E. Fung, S.
224 Reiling, H. Djambazian, M. Hughes, S. Whiteford, C. Caceres-Barrios, T. Nguyen, A.
225 Choo, P. Crisp, S. Sim, S. Geib, F. Marec, I. Häcker, J. Ragoussis, A. Darby, K.
226 Bourtzis, S. Baxter, and M. Schetelig. 2020. White pupae genes in the Tephritids
227 *Ceratitis capitata*, *Bactrocera dorsalis* and *Zeugodacus cucurbitae*: a story of parallel
228 mutations. *BioRxiv*:2020.2005.2008.076158.

229 Ward, C. M., and S. W. Baxter. 2018. Assessing Genomic Admixture between Cryptic
230 *Plutella* Moth Species following Secondary Contact. *Genome biology and evolution*
231 **10**:2973-2985.

232 Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing
233 Company, Incorporated.

234 You, M., F. Ke, S. You, Z. Wu, Q. Liu, W. He, S. W. Baxter, Z. Yuchi, L. Vasseur, G. M. Gurr,
235 C. M. Ward, H. Cerda, G. Yang, L. Peng, Y. Jin, M. Xie, L. Cai, C. J. Douglas, M. B.
236 Isman, M. S. Goettel, Q. Song, Q. Fan, G. Wang-Pruski, D. C. Lees, Z. Yue, J. Bai,
237 T. Liu, L. Lin, Y. Zheng, Z. Zeng, S. Lin, Y. Wang, Q. Zhao, X. Xia, W. Chen, L.
238 Chen, M. Zou, J. Liao, Q. Gao, X. Fang, Y. Yin, H. Yang, J. Wang, L. Han, Y. Lin, Y.
239 Lu, and M. Zhuang. 2020. Variation among 532 genomes unveils the origin and
240 evolutionary history of a global insect herbivore. *Nature Communications* **11**:2321.

241 Zheng, X., S. M. Gogarten, M. Lawrence, A. Stilp, M. P. Conomos, B. S. Weir, C. Laurie, and
242 D. Levine. 2017. SeqArray—a storage-efficient high-performance data format for
243 WGS variant calls. *Bioinformatics* **33**:2251-2257.

244 Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. A high-
245 performance computing toolset for relatedness and principal component analysis of
246 SNP data. *Bioinformatics* **28**:3326-3328.

247

Chapter 4

Assessing genomic admixture between cryptic *Plutella* moth species following secondary contact.

Ward, C. M., & Baxter, S. W. (2018). Assessing genomic admixture between cryptic *Plutella* moth species following secondary contact. **Genome Biology and Evolution**, 10(11), 2973-2985.

Statement of Authorship

Title of Paper	Assessing Genomic Admixture between Cryptic <i>Plutella</i> Moth Species following Secondary Contact
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Christopher M Ward, Simon W Baxter, Assessing Genomic Admixture between Cryptic <i>Plutella</i> Moth Species following Secondary Contact, <i>Genome Biology and Evolution</i> , Volume 10, Issue 11, November 2018, Pages 2973–2985, https://doi.org/10.1093/gbe/evy224

Principal Author

Name of Principal Author (Candidate)	Christopher Ward
Contribution to the Paper	Conceived research, carried out quantitative analysis and interpreted results. Wrote the first version of the manuscript and edited subsequent versions.
Overall percentage (%)	80
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 18/5/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Simon Baxter
Contribution to the Paper	20% Conceived research, carried out quantitative analysis and interpreted results. Wrote the first version of the manuscript and edited subsequent versions.
Signature	Date 20/5/2021

Assessing Genomic Admixture between Cryptic *Plutella* Moth Species following Secondary Contact

Christopher M. Ward and Simon W. Baxter*

Department of Molecular and Biomedical Science, School of Biological Sciences, University of Adelaide, Australia

*Corresponding author: E-mail: simon.baxter@adelaide.edu.au.

Accepted: October 12, 2018

Data deposition: This project has been deposited at GenBank under the accessions SRR6023624, SRR6505218–SRR6505233, SRR6505268–SRR6505279, and MG787473.1. Nucleotide alignments will be provided upon request to the corresponding author.

Abstract

Cryptic species are genetically distinct taxa without obvious variation in morphology and are occasionally discovered using molecular or sequence data sets of populations previously thought to be a single species. The world-wide Brassica pest, *Plutella xylostella* (diamondback moth), has been a problematic insect in Australia since 1882, yet a morphologically cryptic species with apparent endemism (*P. australiana*) was only recognized in 2013. *Plutella xylostella* and *P. australiana* are able to hybridize under laboratory conditions, and it was unknown whether introgression of adaptive traits could occur in the field to improve fitness and potentially increase pressure on agriculture. Phylogenetic reconstruction of 29 nuclear genomes confirmed *P. xylostella* and *P. australiana* are divergent, and molecular dating with 13 mitochondrial genes estimated a common *Plutella* ancestor 1.96 ± 0.175 Ma. Sympatric Australian populations and allopatric Hawaiian *P. xylostella* populations were used to test whether neutral or adaptive introgression had occurred between the two Australian species. We used three approaches to test for genomic admixture in empirical and simulated data sets including 1) the f_3 statistic at the level of the population, 2) pairwise comparisons of Nei's absolute genetic divergence (d_{XY}) between populations, and 3) changes in phylogenetic branch lengths between individuals across 50-kb genomic windows. These complementary approaches all supported reproductive isolation of the *Plutella* species in Australia, despite their ability to hybridize. Finally, we highlight the most divergent genomic regions between the two cryptic *Plutella* species and find they contain genes involved with processes including digestion, detoxification, and DNA binding.

Key words: introgression, hybridization, admixture, cryptic species, *Plutella xylostella*, *Plutella australiana*.

Introduction

Cryptic species lack conspicuous variation in visible traits, yet can show high levels of ecological, behavioral, and genetic divergence, particularly when they arise in allopatry (Stuart et al. 2006; Bickford et al. 2007; Pfenninger and Schwenk 2007). Morphological resemblance of two or more distinct species can occur when environmental pressures maintain phenotypes or cause convergence, and through introgression of traits by interspecies hybridization (Bickford et al. 2007). Consequently, cryptic species are often overlooked, leading to both underestimates of species richness and overestimates of their geographic range (Stuart et al. 2006; Vodă et al. 2015).

Reproductive barriers can maintain boundaries between sympatric congeneric animal species (cryptic or noncryptic) using a range of isolating mechanisms such as olfaction, pheromone cues, and mating calls (Jones and Hamilton 1998;

Andersson et al. 2007), host plant preference or mating timing (Hänniger et al. 2017), and endosymbiont infection (Shoemaker et al. 1999; Bordenstein et al. 2001). Although these factors can impose reproductive isolation barriers and restrict hybridization, assortative mating does not always occur (Mallet et al. 2007). Interspecific hybridization of two species within the same genera has been found to occur at similar rates across the animal kingdom, after taxonomic groups are adjusted for species richness (Schwenk et al. 2008). While hybridization between related species has been well documented, the process of distinguishing between adaptive introgression and regions of historic population structure has been challenging (Martin et al. 2015).

Closely related allopatric or sympatric species without gene flow should exhibit genetic divergence across the genome, whereas species with gene flow should show lower levels of

divergence across broad regions relative to the frequency of interbreeding and how recently it occurred. Detecting hybridization is possible through the use of informal statistical tests on genetic variation, including principle component analysis (Patterson et al. 2006) and Bayesian STRUCTURE model analysis (Pritchard et al. 2000). While these tests can provide results indicative of admixture, they cannot distinguish between introgression, interlineage sorting, or homoplastic genetic drift. Patterson et al. (2012) formalized statistical approaches to estimate admixture based on allele frequencies across multiple populations, namely the f_3 and f_4 statistics (D -statistic), which assess the likelihood of hybridization. The f_4 statistic has identified introgression between sympatric *Heliconius* butterfly species (Martin et al. 2013; Zhang et al. 2016) and hominids (Patterson et al. 2012), as allele frequencies across these genomes did not always agree with the expected species tree, or neutral drift.

Hybridization and introgression of genetic variation from a donor species into a recipient can have adaptive advantages. The transfer of advantageous preadapted alleles from one species into another removes the reliance of new traits arising through mutation in the recipient. Examples include the transfer of rodenticide resistance between mice (Song et al. 2011), coat color alleles among jackrabbits and hares (Jones et al. 2018), aposematic wing patterns in *Heliconius* butterflies (Mavárez et al. 2006; Pardo-Díaz et al. 2012; Wallbank et al. 2016) and insecticide resistance genes in *Anopheles* mosquitoes (Lee et al. 2013; Norris et al. 2015).

The diamondback moth, *Plutella xylostella* (L.) (Lepidoptera: Plutellidae), is the most destructive pest of Brassicaceous agricultural crops, including broccoli, cabbage, and canola (Furlong et al. 2013). They are able to cause *en masse* defoliation, malformed, and improper plant growth (Zalucki et al. 2012), and often develop resistance to insecticides making pest control an ongoing challenge. *Plutella xylostella* were first documented in Australia in the 1880s (Tyron 1889), yet an endemic and phenotypically cryptic species, *P. australiana* (Landry and Hebert), was only recently identified through high divergence of mitochondrial COI barcode sequences (8.6%) and morphologically distinct genitalia (Landry and Hebert 2013). The discovery was surprising, as *P. australiana* was not detected in previous molecular studies of *P. xylostella* yet, is dispersed across eastern Australia (Endersby et al. 2006; Delgado and Cook 2009).

Insecticide susceptibility appears to limit *P. australiana*'s pest potential among cultivated brassica crops, however, introgression of insecticide resistance loci from *P. xylostella* could have serious consequences for agriculture. *Plutella xylostella* and *P. australiana* can hybridize in experimental laboratory crosses, despite their contrasting infection rates of endosymbiotic *Wolbachia* (Ward and Baxter 2017; Perry et al. 2018), which are known to cause reproductive incompatibility in some cases (Sasaki and Ishikawa 2000; Duplouy et al. 2013). *Wolbachia* infection is fixed among *P. australiana* yet extremely low in

Australian *P. xylostella* (1.5%). Although the strength of reproductive barriers in the field is unknown, limited numbers of SNP markers widely dispersed across the nuclear genome previously identified genetic structure between sympatric populations of *P. xylostella* and *P. australiana* (Perry et al. 2018). Due to *P. australiana*'s apparent endemism and the relatively recent invasion of *P. xylostella* into Australia, we assessed the capacity for sympatric Australian *Plutella* species to exchange beneficial traits through disassortative mating and introgression in the field through analyzing whole genomes.

Materials and Methods

Specimen Collection and Genome Sequencing

Plutella xylostella and *P. australiana* were collected from canola (*Brassica napus*) fields using light traps at Cook, Australian Capital Territory (ACT), (−35.262, 149.058) in October 2014 and from direct larval sampling at Ginninderra Farm, ACT, (−35.187, 149.053) in December 2015. Larvae from Calca, South Australia, (SA) (−33.049, 134.373) and Bairds Bay, (SA) (−33.023, 134.279) were collected in June 2014 from mixed stands of sand rocket (*Diploaxis tenuifolia*) and wall rocket (*D. muralis*). Larval collections were reared through to pupation then frozen, to eliminate samples infected with parasitoids. A single *P. australiana* moth was also collected from Richmond, New South Wales (−33.597, 15.740) using a light trap. Large populations of *P. xylostella* larvae were also collected from *Brassica* vegetable farms on three Hawaiian Islands in August 2013, including Kunia, Oahu (21.465, −158.064), Kula, Maui (20.791, −156.337) and Waimea on Hawaii Island (20.028, −155.636), and reared for one generation. Genomic DNA purification was performed using phenol extractions, treated with RNaseA, precipitated with ethanol, and resuspended in TE buffer (10 mM Tris, 0.1 mM EDTA). Species identification was performed using a PCR-RFLP diagnostic assay of the mitochondrial COI gene (Perry et al. 2018). Genome sequencing was performed using the Illumina HiSeq2500 or NextSeq platforms at the Australian Genome Research Facility and the Australian Cancer Research Facility.

Processing Genome Sequence Data

Summary statistics of Illumina sequence reads were generated with FastQC (Andrews 2010) and visualized using the R package ngsReports (Ward et al. 2018). Trimmomatic v 0.32 (Bolger et al. 2014) was used with the parameters (TRAILING: 15 SLIDINGWINDOW: 4: 15) to trim adapter, quality filter, and retain paired reads. The *P. xylostella* reference genome (You et al. 2013) was downloaded from NCBI (GCA_000330985.1). Stampy v1.0.21 (Lunter and Goodson 2011) was used to align the paired reads to the reference with the parameters (−gatkcgarrworkaround, −substitutionrate = 0.01) which produced Sequence Alignment/Map (SAM) files

that were converted to binary format (BAM) and indexed then sorted using SAMtools v1.2 (Li et al. 2009). PCR and optical duplicates were removed using Picard Tools v1.61 (<http://broadinstitute.github.io/picard/>). BAM summary statistics including average read depth per site called, coverage of the genome, percent missing data, total number of reads and read quality were generated using SAMtools v1.2.

Genotype Variant Calling

Variant calling was performed using the Genome Analysis ToolKit (GATK) v3.3 (DePristo et al. 2011). GATK: HaplotypeCaller was used to generate gVCF records, containing variant and invariant sites across the genome, on a per sample basis. The HaplotypeCaller parameter heterozygosity (likelihood of a site being nonreference) for each species was estimated by SAMtools v1.2, indicating *P. xylostella* from Hawaii was most similar to the reference genome (heterozygosity: *P. australiana* = 0.0497; *P. xylostella* Australia = 0.0348; *P. xylostella* Hawaii = 0.0272). Individual gVCF records were combined using GATK: Genotype GVCF and filtered using BCFtools (Li et al. 2009) to a minimum individual depth greater than five reads per base with no greater than 40% of samples missing genotypes at any one site.

Nei's mean intrapopulation nucleotide diversity, π , (Nei and Li 1979) was calculated using egglib (De Mita and Siol 2012). The mean and standard error in π and jackknifing was performed using the R package bootstrap (Canty and Ripley 2017). Pairwise F_{ST} and Tajima's D was calculated across 50-kb windows using VCFtools (Danecek et al. 2011) and minimum distances between populations (km's) determined with <http://www.movable-type.co.uk/scripts/latlong.html>, last accessed October 25, 2018.

Phylogenetic Reconstruction of *Plutella* Mitochondrial and Nuclear Genomes

All quality filtered variant and invariant sites called against the mitochondrial reference genome (GenBank KM023645) were extracted using BCFtools and converted to a FASTA alignment using the R programming language. Maximum likelihood phylogenetic inference using the nuclear genome consensus, heterozygous sites were replaced with IUPAC ambiguity codes, alignment was performed with exaML (Kozlov et al. 2015) with GTR+GAMMA bootstrap resampling ($n = 100$; GTR+GAMMA) was then carried out using RAxML v8.2.4 to provide node confidence. The phylogeny was then rooted using the midpoint method in FigTree (v1.4.3, <http://tree.bio.ed.ac.uk/software/figtree>).

De Novo Assembly of Mitochondrial Genomes and *Plutella* Split Time Estimates

De novo assembly of *Plutella* mitochondrial genomes was performed using NOVOPlasty v2.6.3 (Dierckxsens et al.

2017). A sequence read that mapped to the *P. xylostella* mitochondrial COI gene was used as the seed to initiate assembly. Genomes circularized by NOVOPlasty were then annotated through homology to the *P. xylostella* mitochondrial reference gene annotation (GenBank KM023645) with Geneious v10.0.6. Potential misassemblies were investigated by mapping individual raw reads to the appropriate de novo assembly on a per sample basis using BWA-MEM (Li 2013). Mapped reads were then used as fragments in Pilon (Walker et al. 2014) to correct the assembly. The sample with the greatest total length (15,962 bp), *Paus ACT14.1*, was used to produce a reference for the mitochondrial genome of *Plutella australiana* (Genbank accession MG787473.1).

The mitochondrial split time between *P. xylostella* and *P. australiana* was estimated using 13 mitochondrial protein coding genes extracted from 20 *Plutella* samples with circularized genomes plus *Prays oleae* (accession no. NC_025948.1) and *Leucoptera malifoliella* (accession no. JN790955.1). Nucleotide alignments were made for each gene using MAFFT (Katoh et al. 2002), substitution models were determined using JModelTest2 (Darriba et al. 2012) and alignments were then imported into BEAUTi (Drummond et al. 2012). We set the clock model to strict with 0.0177 substitutions Myr^{-1} according to Papadopoulou et al. (2010). Substitution models were unlinked to allow each sequence to coalesce independently with the Yule speciation model. MCMC sampling was carried out over 1000000 trees sampling every 1000 using BEAST2 v 2.4.7 (Bouckaert et al. 2014). Sampled trees from the chain were checked using Tracer v 1.6 (Rambaut et al. 2018) to determine burn in. Densitree was then used to superimpose MCMC trees to determine the internal node height ranges.

Data Simulation

Coalescent local trees with a total chromosomal length of 25 Mb were simulated for 24 individuals, including eight samples from an outgroup (O) and two ingroups (I_1 and I_2) using the Markovian Coalescent Simulator, MaCS (Chen et al. 2009). A coalescent model for the most recent common ancestor of I_1 and I_2 was set to $0.4 \times 4N$ generations ago and the root to $1.5 \times 4N$ generations ago, providing the topology $((I_1, I_2), O)$. Simulated divergence was determined using mean d_{XY} values from *Plutella* samples (see fig. 4). Two approaches were used to simulate introgression events from I_2 to O or from O to I_2 . First, Introgression was simulated as a single *en masse* admixture event at $0.01 \times 4N$ generations ago with admixture frequencies (f) of $f = 0, 0.05, 0.1, 0.2, \text{ and } 0.3$. Second, introgression was simulated over five distinct breakdowns in assortative mating (0.01, 0.008, 0.006, 0.004, and $0.002 \times 4N$ generations ago) and $f = 0, 0.05, 0.1, 0.2, \text{ and } 0.3$. Each simulation was carried out with a constant population recombination rate ($4Nr$) of 0.001. Sequences were generated from the coalescent trees using SeqGen (Rambaut and Grassly 1997) with the Hasegawa–Kishino–Yano substitution

model (Hasegawa et al. 1985) and a branch scaling factor of 0.01.

Admixture

The F3-Statistic

A formal test for admixture was calculated using the three population test, the f-statistic (f3) (Reich et al. 2009; Patterson et al. 2012). Three possible combinations of tip structures were assessed with the ingroups and outgroup namely f3(I_1 , I_2 ; O), f3(I_1 , O; I_2), f3(I_2 , O; I_1). Cases without introgression are expected to return positive f3 values while negative values indicate introgression has occurred from a donor to a recipient population, forming an intermediate ancestor of both source populations. Block jack-knife F3 estimation was carried out using PopStats (<https://github.com/pontussk/popstats>).

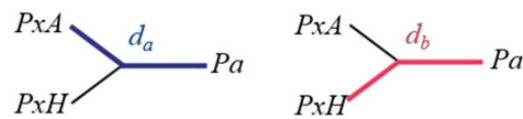
Absolute Divergence (d_{XY})

Nei's absolute divergence, d_{XY} , was used to calculate the mean number of nucleotide differences between two populations across nonoverlapping 50-kb windows with *egglib_sliding_windows.py* (<https://github.com/johnomics>). Comparisons of d_{XY} were made first with simulated data sets, using the five admixture frequencies ($f=0$, 0.05, 0.1, 0.2, and 0.3) between I_2 and O, O and I_2 and I_1 and I_2 . The d_{XY} values were summarized by transforming them into density plots to visualize the distribution and frequency across the simulated genome. This provided expected d_{XY} patterns under a range of admixture frequencies. Average d_{XY} was then calculated between 1) *P. australiana* (O) and Australian *P. xylostella* (I_2) individuals, 2) *P. australiana* (O) and Hawaiian *P. xylostella* (I_1) and 3) Hawaiian *P. xylostella* (I_1) and Australian *P. xylostella* (I_2). Histograms were plotted after setting the maximum value to 1 using the *geom_density* function in ggplot2 (Wickham 2009).

Tree-Tip Distance Proportions

Maximum likelihood phylogenetic reconstruction was performed with RAxML (Stamatakis 2014) using nonoverlapping 50-kb genomic windows generated with the python script *genoToSeq.py* (<https://github.com/simonhmartin>). Phylogenies of empirical data each used four individuals, including one *P. xylostella* and one *P. australiana* individual from sympatric Australia populations (SA14, ACT14, or ACT15), and two *P. xylostella* individuals from Hawaii (HO13.1 and HH13.2). Each tree was then converted to a distance matrix using APE (Paradis et al. 2004) and pairwise distances between tips were determined in the R programming language using the equation,

$$\text{equation 1 } \frac{d_a}{d_a + d_b} = \text{Proportion}_{ab}$$



where d_a is the branch distance between the tree-tip of an Australian *P. xylostella* individual (*PxA*) and a *P. australiana* individual (*Pa*) and d_b is the averaged branch distance between two Hawaiian *P. xylostella* (*Pxyl* HO13.1 and *Pxyl* HH13.2) and a *P. australiana* individual. Two individuals from Hawaii were used to reduce bias from this ingroup source. The Proportion_{ab} values for each 50-kb window were expected to be ~ 0.5 if the phylogeny was concordant with the species tree. Values much >0.5 indicate genomic windows more similar between *P. australiana* and Hawaiian *P. xylostella*, while values much <0.5 indicate genomic windows more similar between *P. australiana* and Australian *P. xylostella* and are candidate admixed regions. Only genomic windows with $>20\%$ of sites genotyped were analyzed.

For comparison, simulated data from 24 individuals and five admixture frequencies ($f=0$, 0.05, 0.1, 0.2, and 0.3) described earlier was divided into 50-kb windows ($n=500$). Each 50-kb window was then subdivided into 64 separate alignments containing four simulated samples; the same two I_1 individuals in each case, (reflecting the use of the same two *P. xylostella* samples from Hawaii in the empirical data) and nonredundant pairs of I_2 and O individuals. Four tip unrooted maximum likelihood phylogenies were produced for each alignment using RAxML, then Proportion_{ab} calculated and plotted using a bin width of 0.05.

Analysis of Discordant Tree-Tip Distances

After plotting tree-tip distance proportions, the tails of each distribution was investigated for symmetry by counting the number of 50-kb windows above or below each mean at three thresholds (mean ± 0.05 , 0.10, and 0.15). Windows below the mean (mean $- 0.15$) were further investigated by calculating d_{XY} across 10-kb windows, sliding by 2 kb. These genomic regions indicate greater similarity between *P. australiana* and Australian *P. xylostella* than the average and d_{XY} plots were visually inspected for signs of introgression.

Identification of Divergent Genomic Windows between *P. australiana* and *P. xylostella*

Both F_{ST} and d_{XY} were calculated across aligned 50-kb genomic windows between all *P. xylostella* samples (from Australia plus Hawaii) and *P. australiana*. Annotated protein coding genes were extracted from the most divergent 1% of 50-kb windows for each statistic and BLAST against the DBM gene list available from DBM-DB (Tang et al. 2014). To identify their

Table 1Summary of Sequence Coverage, Percentage of Sites Genotyped, and Mean Nucleotide Diversity of *Plutella* Populations (170 Mb)

Population	Year Collected	Species	Number of Samples	Average Coverage Per Site	Sites Genotyped (%)	Mean Nucleotide Diversity	(±SD)
Australian Capital Territory (ACT)	2014	<i>P. xylostella</i>	2	17	91.5	0.0151	(0.0042)
		<i>P. australiana</i>	4	12.5	71.8	0.0170	(0.0052)
Australian Capital Territory (ACT)	2015	<i>P. xylostella</i>	2	18	92.2	0.0150	(0.0045)
		<i>P. australiana</i>	4	13.5	72.3	0.0168	(0.0048)
South Australia (SA)	2014	<i>P. xylostella</i>	4	23.5	91.8	0.0157	(0.0040)
		<i>P. australiana</i>	4	17.5	65.9	0.0174	(0.0051)
Hawaii	2013	<i>P. xylostella</i>	8	13.125	91.7	0.0200	(0.0044)

molecular function, InterPro and UniProt annotations were obtained for each BLAST hit.

Results

Alignment of *Plutella* Species to the Reference Genome

The genomes of 29 *Plutella* samples were sequenced using short read Illumina platforms, including eight *P. xylostella* from Hawaii, eight *P. xylostella* from Australia, and 13 *P. australiana*. Samples from Australia were classified into three populations based on collection location and year for analysis (ACT2014, ACT2015, SA2014). A single *P. australiana* individual from Richmond, NSW, was also sequenced (supplementary table S1, Supplementary Material online). Resequenced genomes were mapped to the ~393 Mb *P. xylostella* reference genome (You et al. 2013), but just 170 Mb of non-N bases were retained after stringent quality filtering. Sequence coverage across the 170 Mb alignment ranged from 9- to 25-fold per individual and ~70% of these sites were genotyped in *P. australiana* samples compared with ~92% for Australian and Hawaiian populations of *P. xylostella* (table 1).

The highest levels of nucleotide diversity were observed within Hawaiian *P. xylostella* samples (table 1). However, endemic *P. australiana* populations showed higher levels of nucleotide diversity than Australian *P. xylostella*, which may have undergone a population genetic bottleneck when colonization occurred. Mutation-drift equilibrium of these populations was determined using Tajima's D (D_T). *Plutella xylostella* collected from Australia were under equilibrium (D_T 95% CI = -0.6046375 to +0.9435148) whereas those collected from Hawaii showed largely negative values (D_T 95% CI = -1.88 to -0.039) which may be the result of a recent population size expansion or higher than expected abundance of rare alleles. The frequency of rare alleles in *P. australiana* was also common, although the D_T 95% confidence interval overlapped with zero (D_T 95% CI = -1.18 to +0.42).

Pairwise comparisons between populations and species were then used to assess genetic structure with F_{ST} . The three Australian *P. xylostella* populations showed no genetic structure between geographic location (SA vs. ACT) or year (2014

vs. 2015) (combined average of $F_{ST} = 0.003 \pm 0.003$), as has been previously reported with microsatellite data (Endersby et al. 2006). However, much higher levels of differentiation were observed when compared with Hawaiian *P. xylostella*, supporting the expectation of genetic isolation (average of $F_{ST} = 0.108 \pm 0.01$). The average pairwise F_{ST} values were slightly lower between *P. australiana* and Hawaiian *P. xylostella* ($F_{ST} = 0.501 \pm 0.002$) than *P. australiana* and Australian *P. xylostella* ($F_{ST} = 0.532 \pm 0.013$) (table 2).

Phylogenetic Inference of *Plutella* Species

A maximum likelihood phylogeny using ~170 Mb of the nuclear genome showed two clear *Plutella* species groups with deep divergence between species. *Plutella xylostella* from Hawaii and Australia formed reciprocally monophyletic sister clades with 100% bootstrap support while *P. australiana* genomes formed a single clade, although generally had lower levels of internal branch support (fig. 1). Branch distances were shorter between the internal nodes of *P. australiana* and Hawaiian *P. xylostella* than Australian *P. xylostella*, suggesting the two *P. xylostella* clades have diverged substantially since their most recent common ancestor.

Plutella australiana Mitochondrial Genome and Dating

We carried out de novo assembly and annotation of the *P. australiana* mitochondrial genome which has a total length of 15,962 bp (GenBank accession MG787473.1) compared with 16,014 bp of *P. xylostella* (Dai, Zhu, Qian, et al. 2016). Using sequence homology to the *P. xylostella* mitochondrial genome we annotated two rRNAs, 13 protein coding mitochondrial genes and 22 t-RNA, which showed a conserved gene order for Lepidopteran mitochondrial genomes (Dai, Zhu, Zhao, et al. 2016). The nucleotide sequence of 13 protein coding mitochondrial genes from 22 *Plutella* samples were then used to estimate the mitochondrial split time between *P. xylostella* and *P. australiana* at 1.96 Ma (95% confidence interval ± 0.175 Myr, fig. 2 and supplementary table S2, Supplementary Material online). *Prays oleae* and *Leucoptera malifoliella* were used as the outgroups. The topology of the 13 mitochondrial genes used to date the split (supplementary

Table 2Matrix of the Minimum Distance between Collection Sites (km's, Above Diagonal) and Pairwise F_{ST} Values of Each *Plutella* Population (below diagonal)

Species and Population	<i>Pxyl</i> .ACT.2014	<i>Pxyl</i> .ACT.2015	<i>Pxyl</i> .SA.2014	<i>Pxyl</i> .Hawaii.2013	<i>Paus</i> .ACT.2014	<i>Paus</i> .ACT.2015	<i>Paus</i> .SA.2014
<i>Pxyl</i> .ACT.2014		8.3	1372	8470	0	8.3	1372
<i>Pxyl</i> .ACT.2015	0.000		1371	8466	8.3	0	1371
<i>Pxyl</i> .SA.2014	0.007	0.006		9480	1372	1371	0
<i>Pxyl</i> .Hawaii.2013	0.102	0.103	0.119		8470	8466	9480
<i>Paus</i> .ACT.2014	0.521	0.520	0.548	0.499		8.3	0
<i>Paus</i> .ACT.2015	0.523	0.523	0.549	0.500	0.001		8.3
<i>Paus</i> .SA.2014	0.525	0.525	0.551	0.503	0.003	0.008	

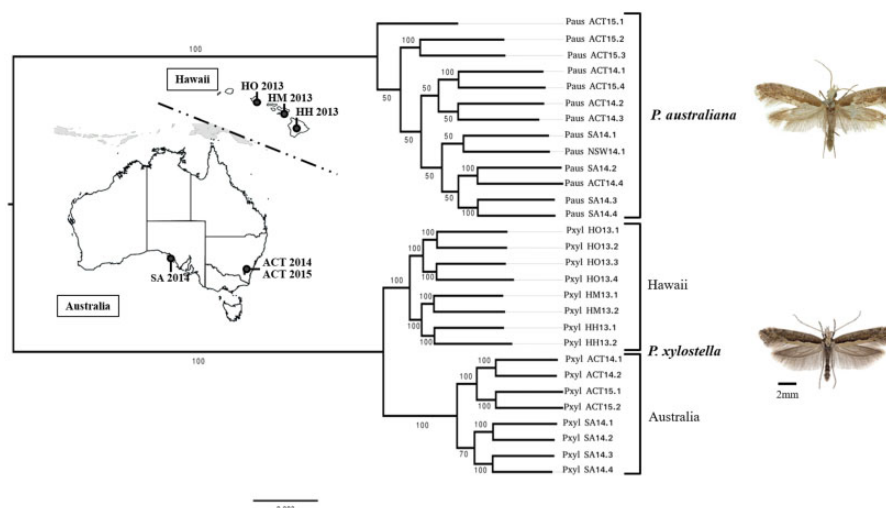


FIG. 1.—Maximum likelihood phylogeny of *Plutella xylostella* and *P. australiana* generated using a 170-Mb concatenated alignment of the nuclear genome. Bootstrap support ($n = 100$) is shown at each node. The inner maps show population locations and year collected for samples from Australia (SA, South Australia; ACT, Australian Capital Territory) and Hawaii, USA (HO, Hawaii Oahu; HM, Hawaii Maui; HH, Hawaii, Hawaii Island). See table 2 for distances between collection locations. Insect photographs were provided by Paul Hebert (*P. australiana*) and Jean-François Landry (*P. xylostella*).

table S3, Supplementary Material online) also supported two clear *Plutella* species groups with an average of 4.95% divergence. *Plutella xylostella* from Hawaii showed higher mitochondrial diversity than samples from Australia. Reduced mitochondrial diversity may have been caused by a founder effect when Australia was colonized (Perry et al. 2018).

Assessing Admixture between Australian *Plutella* Species The F_3 -Statistic

A formal test for genomic admixture was calculated using the three-population f_3 -statistic (f_3), first with simulated data sets to assess the level of sensitivity we could reasonably achieve, and second with empirical data. Simulated introgression frequencies of $f = 0.0, 0.05, 0.1, 0.2,$ and 0.3 were applied from a donor to a recipient. Introgression from ingroup 2 (I_2) into the outgroup (O) increased similarity between these groups, yet also reduced genetic differences between the outgroup and ingroup 1 (I_1). Despite the outgroup becoming more similar to I_2 , O still contained a large proportion of divergent loci which tends to confound the f_3 statistic making negativity

difficult to achieve, even with high levels of introgression (Peter 2016). Consequently, this approach failed to indicate shared ancestry through a negative f_3 -statistic (fig. 3A). Next, introgression from O into I_2 was simulated, to assess sensitivity of introgression from *P. australiana* into Australian *P. xylostella*. Negative values were detected for mixing frequencies of $\geq 20\%$ ($f = 0.2$), indicating high rates of recent hybridization are required to detect introgression using the f_3 -statistic (fig. 3B). Interestingly, spreading the total proportion of introgression to five equidistant time-points along the branch did not increase the detectability of admixture (supplementary fig. S1, Supplementary Material online). This suggests the f_3 is more dependent on the admixture frequency than the divergence between discordant and concordant regions.

Applying the f_3 -statistic to empirical data failed to identify negativity in any tip order between Australian *P. xylostella* and *P. australiana* (fig. 3C). Results for the f_3 -statistic were lower when assessing introgression between Hawaiian *P. xylostella* and *P. australiana* than from between the two sympatric Australian species, consistent with the nuclear phylogeny showing Hawaiian samples are more similar to *P. australiana*.

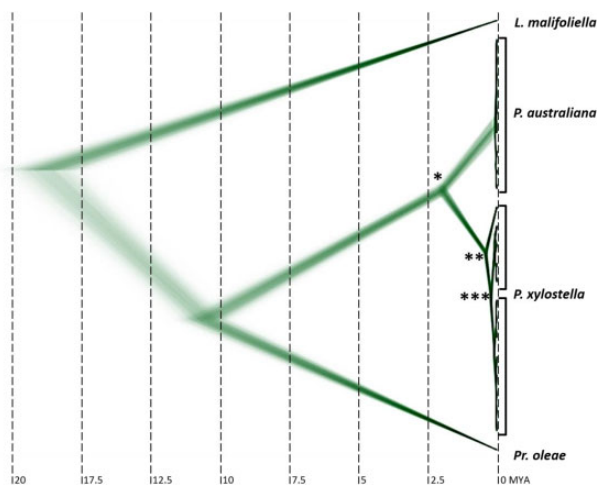


FIG. 2.—Superimposed MCMC trees of 13 protein coding mitochondrial genes used to estimate the split time of *Plutella xylostella* ($n = 13$) and *P. australiana* ($n = 9$) at 1.96 ± 0.175 Ma (*). The internal node of the *P. xylostella* clade was estimated at 0.37 ± 0.057 Ma (**). The split of Hawaiian and Australian *P. xylostella* haplotypes was estimated at 0.078 ± 0.024 Ma (***). *Prays oleae* (accession no. NC_025948.1) and *Leicoptera malifoliella* (accession no. JN790955.1) were used as outgroups.

A lower $f_3(P_{xyl}$ Hawaii, *Pau.*; *Pxyl* Australia) value was estimated for SA 2014 than ACT 2014 and ACT 2015, which may be due to differences in nucleotide diversity between the Australian populations (table 1), as f_3 is decreased proportional to the frequency of minor alleles in the target population. As f_3 did not detect recent admixture events with two closely related ingroup tips, further tests were used to investigate introgression using smaller genomic windows.

Absolute Divergence (d_{XY})

Nei's measure of absolute divergence (d_{XY}) (Nei 1987) was used to compare genetic similarity between populations using 50-kb genomic windows for both simulated and empirical data sets. In all cases, population wide comparisons of d_{XY} were performed between; 1) two ingroup populations (I_1 and I_2), 2) ingroup 1 and the outgroup (I_1 and O), and 3) ingroup 2 and the outgroup (I_2 and O). Comparisons returning values approaching zero indicate high levels of similarity and a recent allelic split time. Low d_{XY} values are expected between ingroup samples (I_1 and I_2), or in cases where introgression may be occurring between an ingroup and outgroup.

Absolute divergence in simulated populations was calculated for each 50-kb window ($n = 500$), again for admixture occurring at $f = 0.0, 0.05, 0.1, 0.2,$ and 0.3 . The distribution of d_{XY} values obtained for each comparison were plotted as histograms normalized for density by rescaling such that the maxima of the distribution is 1 (fig. 4 and supplementary figs. S2 and S3, Supplementary Material online). Introgression either from I_2 into O or from O into I_2 produced a decrease in

absolute divergence across the genome, providing a benchmark for comparisons with empirical data. Admixture in the direction O to I_2 provided a much clearer genome wide signal than the reverse direction, (I_2 to O) indicating it would be easier to detect introgression from *P. australiana* into Australian *P. xylostella* than the reverse.

Based on the whole genome phylogeny (fig. 1), we expected mean d_{XY} between *P. australiana* and Hawaiian *P. xylostella* to be slightly lower than Australian *P. xylostella*. The d_{XY} distribution of empirical 50-kb windowed data provided no support of widespread introgression (fig. 4C), as values comparing the *P. australiana* outgroup with either *P. xylostella* from Hawaii (I_1) or Australia (I_2) did not deviate from their expected values (table 3). This suggests concordance with the whole genome tree topology.

Phylogenetic Tree-Tip Distance Proportions

The f_3 -statistic and d_{XY} were both used to test for introgression within populations. Next, we used the tree-tip distance proportion to assess whether evidence for introgression could be detected between individual sample pairs. Simulated genomes with introgression from I_2 into O, or O into I_2 at the rates $f = 0.0, 0.05, 0.1, 0.2,$ and 0.3 were divided into 50-kb sequence alignments, as described earlier. Further subdivision was then performed so each 50-kb window contained just four sequences; two ingroup 1, one ingroup 2, and one outgroup sequence. This was repeated 64 times for each 50-kb window, then maximum likelihood phylogenetic reconstruction performed for each alignment. Based on the whole-genome topology, we expected the outgroup to be a similar distance from both ingroup 1 and ingroup 2, unless introgression had occurred and shortened the distance between samples.

A proportion of the branch distance between I_2 and O (d_a) and I_1 and O (d_b) was then calculated for each phylogeny using equation 1, normalizing values within the range 0–1. Tree-tip distance proportions are presented as histograms to graph the distribution (x axis), and normalized density (y axis). Although introgression from I_2 into O (fig. 5A and supplementary fig. S4, Supplementary Material online) and from O into I_2 (fig. 5B) were both detected using distance proportions, patterns did vary based on the direction of admixture. Clearer signals of admixture were evident in the direction O to I_2 (supplementary table S4, Supplementary Material online), as this made ingroup 2 less similar to ingroup 1. Given I_1 and I_2 recently split, admixture from I_2 into O is also expected to make the outgroup more similar to I_1 , decreasing detectability. Simulating introgression over five equally spaced events was effective at detecting admixture using tree-tip distance proportions (supplementary fig. S5, Supplementary Material online).

This method was then applied to empirical data to identify genomic windows that were discordant with the species tree.

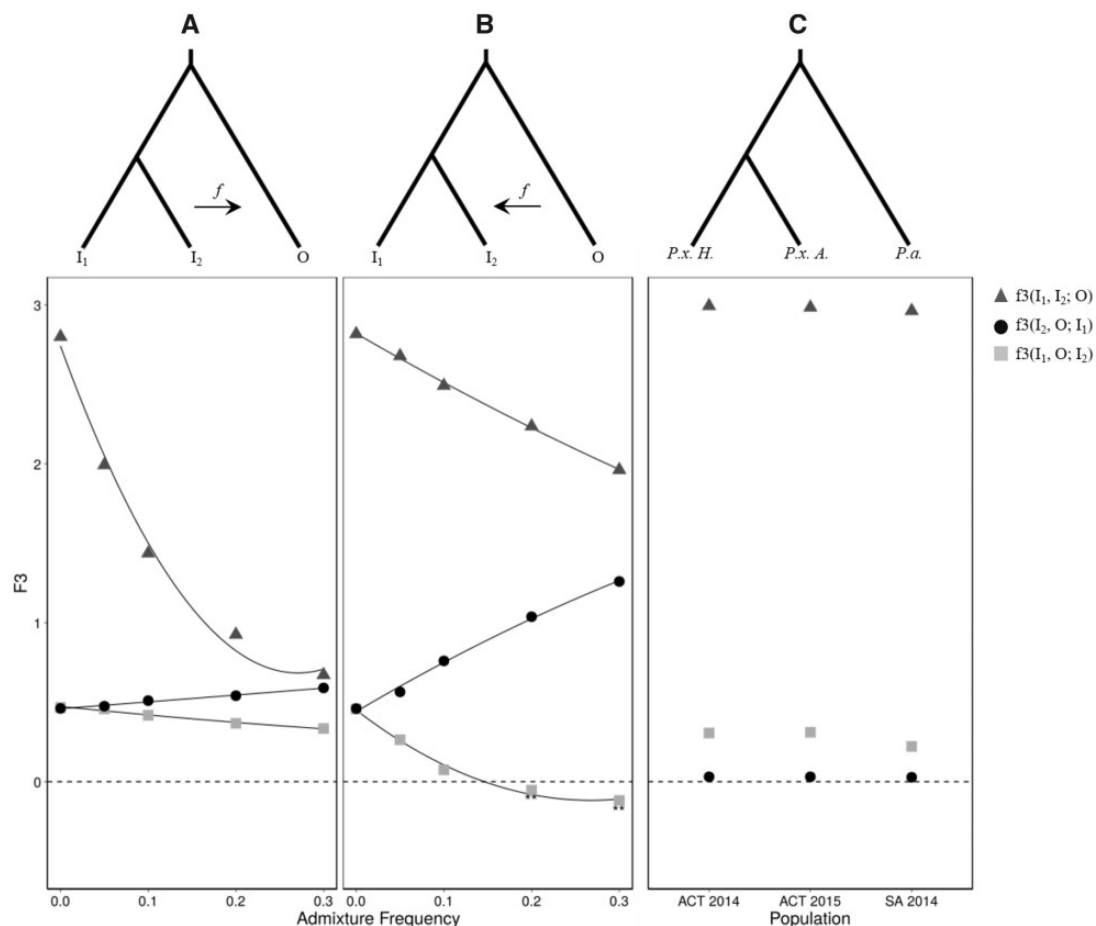


FIG. 3.—The three population f_3 -statistic (f_3). (A) Admixture from ingroup 2 (I_2) to the outgroup (O) was simulated as a single event with frequencies of $f=0, 0.05, 0.1, 0.2$, and 0.3 . Evidence for hybridization and admixture could not be clearly detected in this direction, as shown by the gray boxes for $f_3(I_1, O; I_2)$, which did not reach negative values. For comparison, the f_3 -statistic for $(I_2, O; I_1)$ and $(I_1, I_2; O)$ were plotted with circles and triangles, respectively. (B) Simulated admixture from O into I_2 did produce a significant f_3 statistic at a mixing frequency >0.2 , as indicated by the values <0 $f_3(I_1, O; I_2)$. (C) The f_3 -statistic was then applied to empirical data, testing for admixture in three possible scenarios between *Plutella xylostella* from Hawaii (*P.x. H.*), *P. xylostella* from Australia (*P.x. A.*, I_2) and *P. australiana* (*P.a.*, O). This suggests that, as no f_3 values were <0 , if admixture was occurring it could not be detected using this method.

Similar to the simulated data sets, genomic alignments were divided into nonoverlapping 50-kb contiguous windows, then further subdivided into alignments of one *P. xylostella* and one *P. australiana* individual from sympatric Australian populations, plus two consistant *P. xylostella* from Hawaii. This produced 32 different sample combinations for each 50-kb window, including eight combinations from ACT 2014, eight from ACT 2015 and 16 from SA 2014. An average of 7276 (± 293) maximum likelihood phylogenies were then produced for each of the 32 sample combinations to identify potential admixture that was not fixed in the population. A near-symmetric and unimodal distribution of tree-tip distance proportions was observed in all cases, with the ranges of each density curve showing a large degree of overlap (fig. 5C and supplementary table S5, Supplementary Material online). Mean tree-tip distance proportions for each comparison

were consistent within and between populations ACT 2014 (0.5088–0.5104), ACT 2015 (0.5102–0.5117), and SA 2014 (0.5086–0.5099). All proportion means were >0.5 showing the Hawaiian *P. xylostella* had on average shorter branch lengths to *P. australiana*, consistent with the nuclear genome phylogeny (fig. 1). Under a widespread admixture hypothesis windows with distance proportions below the mean (*P. australiana* closer to *P. xylostella* from Australia) should be much more frequent than above. The number of windows above and below three distances from the mean (0.05, 0.1, 0.15) was similar (supplementary table S6, Supplementary Material online) suggesting no clear evidence to support widespread admixture within *P. xylostella* and *P. australiana* individuals from three sympatric populations.

Despite lack of support for widespread hybridization and genome-wide introgression, we further investigated the tails

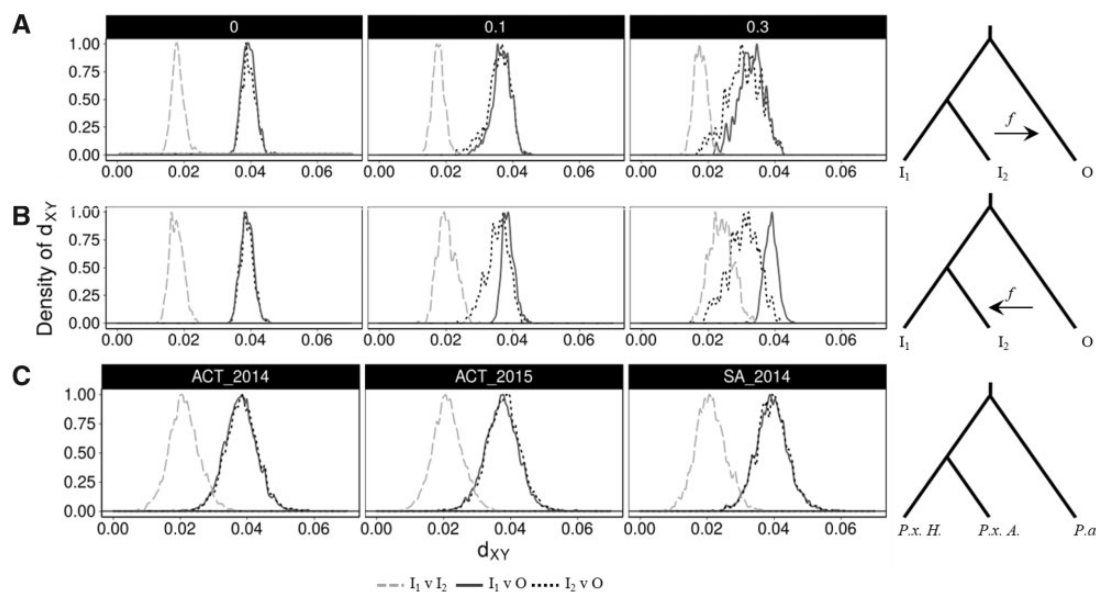


Fig. 4.—Absolute genetic divergence (d_{XY}) between populations. Each plot is a histogram summarizing pairwise comparisons of 50-kb windows across the genome, rescaled such that the maxima is 1. Simulated data uses mixing frequencies of $f=0.0, 0.1,$ and 0.3 (see [supplementary figs. S2 and S3, Supplementary Material](#) online, for additional admixture frequencies) (A) Simulated d_{XY} comparisons assessing of admixture from I_2 into O or (B) O into I_2 . In the absence of hybridization ($f=0$) the simulated ingroups show the lowest levels of divergence (dashed line), while the distance between each ingroup and the outgroup are relatively similar (dotted and solid lines). Increasing levels of admixture ($f=0.1, 0.3$) alters histogram shape as I_2 and O d_{XY} values become smaller. (C) d_{XY} summaries between *Plutella australiana* (*P.a.* in the phylogeny schematic) and *P. xylostella* from Australia (*P.x. A.*) or Hawaii (*P.x. H.*) do not deviate. This indicates d_{XY} was not able to detect hybridization at the population level.

Table 3

Confidence Intervals (95%) for d_{XY} Comparisons of Populations

	<i>Pxyl. Hawaii vs. Pxyl. Australia</i>	<i>Pxyl. Hawaii vs. Paus. Australia</i>	<i>Pxyl. Australia vs. Paus. Australia</i>
ACT 2014	0.02101–0.02122	0.03788–0.03812	0.03828–0.03852
ACT 2015	0.02102–0.02123	0.03721–0.03744	0.03771–0.03795
SA 2014	0.02087–0.02111	0.03898–0.03926	0.03926–0.03955

of the tree-tip distance proportions at a distance 0.15 below the mean for each Australian population. These scaffolds ($n=21$) have the shortest branch lengths between *P. australiana* and sympatric Australian *P. xylostella*, relative to the branch length proportions between *P. australiana* and Hawaiian *P. xylostella* ([supplementary table S7, Supplementary Material](#) online). Sliding window d_{XY} was performed on each of these scaffolds with 10-kb windows (sliding by 2 kb), revealing just one region on scaffold KB207303.1 where Australian *P. xylostella* and *P. australiana* are more similar than between Hawaiian and Australian *P. xylostella*. Historical admixture between Australian *Plutella* species is one possible explanation for this result, although the region does not contain any protein coding genes ([supplementary fig. S6A, Supplementary Material](#) online). A region on scaffold, KB207380.1, was identified in the tree-tip distribution tail in 12/32 comparisons however d_{XY} indicated admixture across this region was unlikely ([supplementary fig. S6B, Supplementary Material](#) online).

Genomic Regions with High Interspecies Divergence

The two *Plutella* species investigated in this study have been shown to have contrasting biologies and pest potential (Perry et al. 2018), and although they can hybridize in laboratory crosses, we found no evidence for widespread admixture among wild samples. This prompted us to ask which 50-kb genomic regions are most divergent between these species, and what kinds of genes do they encode? First, absolute divergence (d_{XY}) between all *P. xylostella* and all *P. australiana* individuals was used to identify the top 1% most divergent genomic windows ([supplementary table S8, Supplementary Material](#) online). These included fifty-one 50-kb windows dispersed across 41 unique scaffolds and showed 33–61% greater absolute divergence than the genome-wide average ($d_{XY}=0.0369$). Second, the top 1% of genomic windows showing highest divergence in nucleotide diversity (F_{ST}) were also identified, showing values 70–110% higher than the mean ($F_{ST}=0.356$). These two estimates of divergence only detected one 50-kb region common to both d_{XY} and

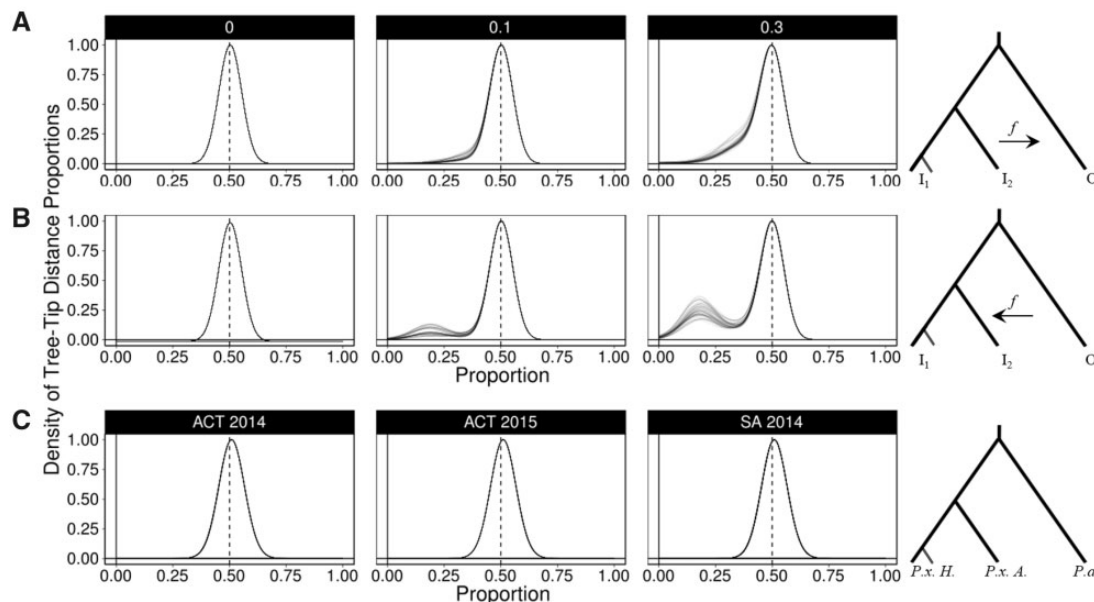


FIG. 5.—Histogram summaries of tree-tip distance proportions, depicting the phylogenetic distance between ingroup and outgroup sequences. A ratio of 0.5 indicates the outgroup sample has the same branch distance to both ingroup samples. Ratios close to zero indicate very short branch lengths between ingroup two and the outgroup, and are candidate regions for introgression. (A) Simulated introgression from ingroup two to the outgroup at mixing frequencies of $f=0, 0.1$ and 0.3 . (B) Simulated introgression from the outgroup into ingroup two at mixing frequencies of $f=0, 0.1$ and 0.3 . (C) Empirical data for the three sympatric Australian *Plutella* populations (ACT 2014, ACT 2015, SA 2014). Each panel summarizes 7276 (± 293) branch distance ratio calculations. The branch leading to ingroup 1 was standardized using the same two Hawaiian *P. xylostella* individuals for each of these comparisons.

F_{ST} (KB207411.1; 400,001...450,000 bp). Most windows with the highest d_{XY} had relatively low F_{ST} values, suggesting the two species share similar levels of polymorphism across these regions. Nonredundant protein coding genes ($n=176$) within these divergent genomic windows contained genes required for feeding including digestion (eg. chymotrypsin, trypsin, aminopeptidase-N), detoxification (eg. cytochrome P450s, carboxylesterases) and also gene regulation (zinc finger proteins) (supplementary table S9, Supplementary Material online). However, Tajima's D showed most of these windows were within the genome wide 95% confidence intervals, indicating these regions are not likely to be under directional or balancing selection (supplementary figs. S7 and S8, Supplementary Material online).

Discussion

The discovery of cryptic species can often be inadvertent and arise from sequencing mitochondrial or nuclear amplicons (Stuart et al. 2006; Landry and Hebert 2013), as well as whole genomes (Janzen et al. 2017). The fortuitous identification of *Plutella australiana* was unexpected and raised initial concern over its pest status and whether specific management practices were required. *Plutella australiana* populations collected from across southern Australia (Perry et al. 2018) and Australian *P. xylostella* (Endersby et al. 2006) lack genetic

structure, showing these species are highly mobile. Adaptive introgression of advantageous traits from one of these species into the other could potentially spread across the Australian continent. Despite high levels of movement, we sampled from sites where *Plutella* populations co-occur to attempt to detect either historical admixture or very recent hybridization.

The physical genome size of *P. xylostella* is estimated at 339 Mb (Baxter 2011) while the reference genome assembly is 393 Mb (You et al. 2013) and includes sequencing gaps totaling ~ 50 Mb. After aligning all resequenced genomes to the *P. xylostella* reference, only 170 Mb was retained in this analysis, which is likely to be caused in part by the sequence gaps and also high levels of genetic diversity (You et al. 2013). Mapping *P. australiana* sequence reads to the *P. xylostella* genome is affected by mapping bias, as the most divergent loci will not map to this reference. This causes all branches to be shortened toward the reference, underestimating the divergence between *P. australiana* and *P. xylostella* in the whole genome phylogeny. However, the introduction of this bias is unavoidable as the only reference genome within the superfamily Yponomeutoidea is currently *Plutella xylostella*.

Plutella australiana were more similar to *P. xylostella* samples from Hawaii than Australia, based on shorter phylogenetic branch lengths for nuclear genomes and subsequent tree-tip distance proportions, lower F_{ST} values and lower f_3 -statistics. A better understanding of migration or transport routes enabling *P. xylostella* to colonize the world would

help explain why this is the case. Several studies have found Australian *P. xylostella* mtDNA genomes have very low levels of diversity (Saw et al. 2006; Juric et al. 2017; Perry et al. 2018), which is indicative of a population bottleneck (and other factors), while we found Hawaiian mtDNA genomes to be quite diverse. This suggests the Hawaiian Islands may have been colonized by a larger founding population, or multiple, independent invasions while Australia may have simply been colonized by a derived population of *P. xylostella* (Juric et al. 2017).

Mitochondrial diversity between the two *Plutella* species was originally found to be ~8.2%, based on sequencing COI amplicons (Landry and Hebert 2013), although the level of diversity across all thirteen protein coding genes is less (4.95%). This level of diversity was not sufficient to result in complete reproductive isolation between the two sister species when reared in the laboratory (Perry et al. 2018). Using the 13 mitochondrial genes, we estimated the split time of *P. xylostella* and *P. australiana* to be ~1.96 Myr. To date, *P. australiana* has only been detected in Australia, yet this relatively recent split questions whether *P. australiana* did evolve within Australia. This would require a considerable migratory event some 2 Mya from the ancestral *Plutella* source population to Australia, and no further migration. Future molecular screening of *P. xylostella* may identify cryptic *P. australiana* in other countries.

Phylogenies of genes or genomic windows can deviate from an expected consensus topology or species tree and can be used to identify genomic regions that may be of biological interest. For example, genomic regions subject to incomplete lineage sorting (Scally et al. 2012), horizontal gene transfer (Moran and Jarvik 2010) and adaptive introgression (Wallbank et al. 2016) all produce discordant phylogenies. Despite simulated data detecting minor levels of introgression using phylogenetic tree-tip distances across the genome, we found few discordant distances between individual *Plutella* samples across 50-kb genomic windows. The methods used here were not sufficient to reject small regions of decreased d_{XY} between Australian *Plutella*, which may be signals of past admixture. Future work into the evolutionary history of *Plutella* moths and sequencing outgroup genomes of *Plutella* species, will enable further analysis of these regions using the D statistic and f_d (Martin et al. 2015).

The most divergent genomic windows between the two *Plutella* species identified using d_{XY} or F_{ST} showed little evidence for current selection and may potentially contain genes that underwent selection after speciation. These genes may reflect different abilities to evade host plant defenses or host plant preference, as many are involved with digestion and detoxification. Using absolute genetic divergence (d_{XY}) to identify the most divergent genomic windows between *P. australiana* and *P. xylostella* may also be highlighting loci that are highly polymorphic or rapidly evolving. Further understanding of *Plutella* biology including mating timing,

evolutionary history, host plant preference and behaviour may provide further insight into these divergent loci.

Plutella australiana and *P. xylostella* are likely to have been in secondary contact in Australia for over 125 years (>1000 generations). Despite this, we found no support for widespread admixture, and although we cannot predict the amount of time these species have spent in geographic isolation, strong reproductive barriers are apparent in the field. Furthermore, *P. xylostella* and *P. australiana* will be a useful system to investigate the genetic basis of biological differences between cryptic species from an agricultural perspective.

Supplementary Material

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

Acknowledgments

This work was funded by the Australian Research Council (grant numbers DP120100047, FT140101303) and Grains Research and Development Corporation (UOA1711-004RSX). C.M.W. is supported by The Commonwealth Hill Trust and Grains Research and Development Corporation. Supercomputing resources were provided by the Phoenix HPC service at the University of Adelaide. We thank Kym Perry (University of Adelaide, Australia), Kevin Powis (South Australian Research and Development Institute, Australia), and Ron Mau (College of Tropical Agriculture and Human Resources, University of Hawaii) for sample collection along with Paul Hebert (Centre for Biodiversity Genomics) and Jean-François Landry (Canadian National Collection of Insects, Arachnids, and Nematodes) for the images of *P. australiana* and *P. xylostella*, respectively. We also thank two anonymous reviewers for their comments on a previous version of this manuscript.

Literature Cited

- Andersson J, Borg-Karlson A-K, Vongvanich N, Wiklund C. 2007. Male sex pheromone release and female mate choice in a butterfly. *J Exp Biol*. 210(Pt 6):964–970.
- Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [last accessed 25th October 2018].
- Baxter SW. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One* 6. doi: 10.1371/journal.pone.0019315.
- Bickford D, et al. 2007. Cryptic species as a window on diversity and conservation. *Trends Ecol Evol*. 22(3):148–155.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
- Bordenstein SR, O'Hara FP, Werren JH. 2001. Wolbachia-induced incompatibility precedes other hybrid incompatibilities in *Nasonia*. *Nature* 409(6821):707–710.

- Bouckaert R, et al. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 10. doi: 10.1371/journal.pcbi.1003537.
- Canty A, Ripley B. 2017. boot: bootstrap R (S-Plus) functions.: R package version 1.3-20.
- Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res.* 19(1):136–142.
- Dai LS, Zhu BJ, Qian C, et al. 2016. The complete mitochondrial genome of the diamondback moth, *Plutella xylostella* (Lepidoptera: plutellidae). *Mitochondrial DNA A DNA Mapp Seq Anal.* 27(2):1512–1513.
- Dai LS, Zhu B-J, Zhao Y, Zhang C-F, Liu C-L. 2016. Comparative mitochondrial genome analysis of *Eligma narcissus* and other Lepidopteran insects reveals conserved mitochondrial genome organization and phylogenetic relationships. *Sci Rep.* 6:26387.
- Danecek P, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.
- Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 9(8):772.
- De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* 13:27.
- Delgado AM, Cook JM. 2009. Effects of a sex-ratio distorting endosymbiont on mtDNA variation in a global insect pest. *BMC Evol Biol.* 9:49.
- DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.
- Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45(4):e18.
- Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29(8):1969–1973.
- Duploup A, et al. 2013. Draft genome sequence of the male-killing *Wolbachia* strain wBol1 reveals recent horizontal gene transfers from diverse sources. *BMC Genomics* 14(1):20.
- Endersby NM, McKechnie SW, Ridland PM, Weeks AR. 2006. Microsatellites reveal a lack of structure in Australian populations of the diamondback moth, *Plutella xylostella*(L.). *Mol Ecol.* 15(1):107–118.
- Furlong MJ, Wright DJ, Dosdall LM. 2013. Diamondback moth ecology and management: problems, progress, and prospects. *Annu Rev Entomol.* 58:517–541.
- Hänniger S, et al. 2017. Genetic basis of allochronic differentiation in the fall armyworm. *BMC Evol Biol.* 17(1):68.
- Hasegawa M, Kishino H, Yano T-a. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22(2):160–174.
- Janzen DH, et al. 2017. Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proc Natl Acad Sci U S A.* 114(31):8313–8318.
- Jones MR, et al. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science* 360(6395):1355–1358.
- Jones T, Hamilton J. 1998. A role for pheromones in mate choice in a lekking sandfly. *Anim Behav.* 56(4):891–898.
- Juric I, Salzburger W, Balmer O. 2017. Spread and global population structure of the diamondback moth *Plutella xylostella* (Lepidoptera: plutellidae) and its larval parasitoids *Diadegma semiclausum* and *Diadegma fenestrale* (Hymenoptera: ichneumonidae) based on mtDNA. *Bull Entomol Res.* 107(02):155–164.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14):3059–3066.
- Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31(15):2577–2579.
- Landry J-F, Hebert P. 2013. *Plutella australiana* (Lepidoptera, Plutellidae), an overlooked diamondback moth revealed by DNA barcodes. *ZooKeys* 327:43–63.
- Lee Y, et al. 2013. Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci U S A.* 110(49):19854–19859.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In: arXiv: [q-bio.GN]. eprint arXiv:1303.3997v2 [q-bio.GN] [last accessed 25, October 2018].
- Li H, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
- Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21(6):936–939.
- Mallet J, Beltrán M, Neukirchen W, Linares M. 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evol Biol.* 7:28.
- Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol Biol Evol.* 32(1):244–257.
- Martin SH, et al. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23(11):1817–1828.
- Mavárez J, et al. 2006. Speciation by hybridization in *Heliconius* butterflies. *Nature* 441(7095):868–871.
- Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* 328(5978):624–627.
- Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.
- Nei and Li 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS.* 76(10):5269–5273.
- Norris LC, et al. 2015. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proc Natl Acad Sci U S A.* 112(3):815–820.
- Papadopoulou A, Anastasiou I, Vogler AP. 2010. Revisiting the insect mitochondrial molecular clock: the Mid-Aegean trench calibration. *Mol Biol Evol.* 27(7):1659–1672.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.
- Pardo-Diaz C, et al. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet.* 8. doi: 10.1371/journal.pgen.1002752
- Patterson N, et al. 2012. Ancient admixture in human history. *Genetics* 192(3):1065–1093.
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2(12):e190.
- Perry KD, et al. 2018. Cryptic *Plutella* species show deep divergence despite the capacity to hybridize. *BMC Evol Biol.* 18(1):77.
- Peter BM. 2016. Admixture, population structure and *F*-statistics. *Genetics* 202(4):1485–1501.
- Pfenniger M, Schwenk K. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evol Biol.* 7. doi: 10.1186/1471-2148-7-121
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.
- Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13(3):235–238.
- Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology.* syy032 doi:10.1093/sysbio/syy032.
- Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* 461(7263):489–494.
- Sasaki T, Ishikawa H. 2000. Transinfection of *Wolbachia* in the mediterranean flour moth, *Ephesia kuehniella*, by embryonic microinjection. *Heredity* 85(2):130–135.

- Saw J, Endersby N, McKechnie S. 2006. Low mtDNA diversity among widespread Australian diamondback moth *Plutella xylostella* (L.) suggests isolation and a founder effect. *Insect Sci.* 13(5):365–373.
- Sally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175.
- Schwenk K, Brede N, Streit B. 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philos Trans R Soc B* 363(1505):2805–2811.
- Shoemaker DD, Katju V, Jaenike J. 1999. Wolbachia and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria*. *Evolution* 53(4):1157–1164.
- Song Y, et al. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr Biol.* 21(15):1296–1301.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
- Stuart BL, Inger RF, Voris HK. 2006. High level of cryptic species diversity revealed by sympatric lineages of Southeast Asian forest frogs. *Biol Lett.* 2(3):470–474.
- Tang W, et al. 2014. DBM-DB: the diamondback moth genome database. *Database* 2014(0):bat087.
- Tyron H. 1889. In: Report on Insect and Fungus Pest No.1. Brisbane (Australia): Department of Agriculture, Queensland, James C. Beal. Government Printer.
- Vodă R, Dapporto L, Dincă V, Vila R. 2015. Cryptic matters: overlooked species generate most butterfly beta-diversity. *Ecography* 38(4):405–409.
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.
- Wallbank RWR, et al. 2016. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.* 14(1):e1002353.
- Ward CM, Baxter SW. 2017. Draft genome assembly of a Wolbachia endosymbiont of *Plutella australiana*. *Genome Announc.* 5. doi: 10.1128/genomeA.01134-17
- Ward CM, To H, Pederson SM. 2018. ngsReports: an R package for managing FastQC reports and other NGS related log files. *bioRxiv.* doi: 10.1101/313148
- Wickham H. 2009. ggplot2: elegant graphics for data analysis. New York (NY): Springer. p. 1–212.
- You M, et al. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat Genet.* 45(2):220–225.
- Zalucki MP, et al. 2012. Estimating the economic cost of one of the world's major insect pests, *Plutella xylostella* (Lepidoptera: plutellidae): just how long is a piece of string? *J Econ Entomol.* 105(4):1115–1129.
- Zhang W, Dasmahapatra KK, Mallet J, Moreira GRP, Kronforst MR. 2016. Genome-wide introgression among distantly related *Heliconius* butterfly species. *Genome Biol.* 17(25): doi: 10.1186/s13059-016-0889-0.

Associate editor: Nancy Moran

Chapter 5

Estimation of molecular dates separating four *Plutella* species.

Ward, C. M., Landry, J.-F. & Baxter, S. W. Estimation of molecular dates separating four *Plutella* species. **Submitted to the Proceedings of the "8th International Conference on Management of the Diamondback Moth and Other Crucifer Insect Pests."**

Statement of Authorship

Title of Paper	Estimation of molecular dates separating four <i>Plutella</i> species
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Conference paper submitted to Eighth International Conference on Management of the Diamondback Moth and Other Crucifer Insect Pests

Principal Author

Name of Principal Author (Candidate)	Christopher Ward			
Contribution to the Paper	Conceived research (with S.B), carried out quantitative analysis (with S.B) and interpreted results (with S.B). Wrote the first version of the manuscript and edited subsequent versions (with S.B).			
Overall percentage (%)	70			
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 60%;"></td> <td style="width: 20%;">Date</td> <td style="width: 20%;">18/5/2021</td> </tr> </table>		Date	18/5/2021
	Date	18/5/2021		

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Jean-Francois Landry			
Contribution to the Paper	5% Provided samples for sequencing. Revised the manuscript. Contribution signed for by Simon Baxter.			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 60%;">N/A</td> <td style="width: 20%;">Date</td> <td style="width: 20%;"></td> </tr> </table>	N/A	Date	
N/A	Date			

Name of Co-Author	Simon Baxter			
Contribution to the Paper	25% Conceived research (with S.B), carried out quantitative analysis (with S.B) and interpreted results (with S.B). Wrote the first version of the manuscript and edited subsequent versions (with S.B) / /			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 60%;"></td> <td style="width: 20%;">Date</td> <td style="width: 20%;">20/5/2021</td> </tr> </table>		Date	20/5/2021
	Date	20/5/2021		

Estimation of molecular dates separating four *Plutella* species

Ward, C.M.

SCHOOL OF BIOLOGICAL SCIENCES, UNIVERSITY OF ADELAIDE, ADELAIDE AUSTRALIA
christopher.ward@adelaide.edu.au

Landry J.-F.

CANADIAN NATIONAL COLLECTION OF INSECTS, ARACHNIDS, AND NEMATODES, AGRICULTURE AND AGRI-FOOD CANADA, OTTAWA RESEARCH AND DEVELOPMENT CENTRE, OTTAWA, CANADA.
jean-francois.landry@canada.ca

Baxter, S.W.

SCHOOL OF BIOSCIENCES, UNIVERSITY OF MELBOURNE, MELBOURNE, AUSTRALIA
simon.baxter@unimelb.edu.au

ABSTRACT

Plutella xylostella L., the diamondback moth, has successfully colonized agricultural regions across the world, and is now recognized as the foremost pest of cruciferous crops. In contrast, other members of the *Plutella* genus have limited observed pest potential and dispersal. Here we sequence and assemble genomes of three neglected species, *P. armoraciae* and *P. porrectella* and the leek moth, *Acrolepiopsis assectella*, then in conjunction with existing data from *P. xylostella* and *P. australiana*, apply a molecular clock to estimate their divergence. All *Plutella* and *Acrolepiopsis* mitochondrial genomes show similar sizes (~15-16 kb) and a conserved gene order, that includes two rRNA, 13 protein coding and 22 tRNA genes. *Plutella armoraciae* and *P. porrectella* show an average 91.95 and 91.08% homology to *P. xylostella* across mitochondrial protein coding genes. We then construct a chronogram using complete COX1 and 16S sequences, dating the crown node of *Plutella* to 5.46 (95% HPD 7.05 – 4.10) MYA, and find the topology is concordant with a species tree constructed from 104 nuclear single copy orthologs. This work provides complete mitochondrial genomes and nuclear genome resources for future comparison within the genus *Plutella* to understand why *P. xylostella* has become such widespread and serious pest, relative to its sister taxa.

Keywords

Plutella, mitochondrial genome, molecular clock, genome

INTRODUCTION

The genus *Plutella* are microlepidopterans within the superfamily Yponomeutoidea (Sohn et al. 2013) containing a total of 26 recognized species worldwide (Baraniak 2007; Sølvi et al. 2018). *Plutella* moths are specialist herbivores of crucifers, which include many plants of economic importance such as canola and cabbage (Bonnemaison 1965; Smith and Sears 1984; Zalucki et al. 2012). Mitochondrial gene trees of cytochrome oxidase I (COX1) have previously provided species topologies of 11 *Plutella* species (Landry and Hebert 2013; Sølvi et al. 2018). However, extensive molecular research has only been carried out on a single species of economic importance, *P. xylostella* (Talekar and Shelton 1993; Furlong et al. 2013).

Plutella xylostella, diamondback moth, is an invasive agricultural pest species throughout the world with an annual estimated cost of \$4-5 billion (Zalucki et al. 2012; Furlong et al. 2013). Recognized as the major insect pest of cruciferous crops, *P. xylostella* routinely develops resistance to insecticides (Zalucki et al. 2012; Furlong et al. 2013), and is highly dispersive (Chapman et al. 2002; Wei et al. 2013a; Fu et al. 2014; Perry et al. 2018). Lack of population structure at the continent level occurs in Australia (Endersby et al. 2006; Perry et al. 2018) and China (Wei et al. 2013a; Fu et al. 2014), enabling resistance alleles to spread between populations. Generational time is highly dependent on temperature, with generations per year ranging from greater than 10 in tropical climates to 2-4 in regions with harsh winters (Bonnemaison 1965; Harcourt 2012). Due to its economic importance, extensive research has been carried out on the life history traits (Bigger and Fox 1997; Philips et al. 2014; Garrad et al. 2015), ecology (Talekar and Shelton 1993; Furlong et al. 2013; Philips et al. 2014) and genetics (Heckel et al. 2007; You et al. 2013; Ward and Baxter 2018) of *P. xylostella*. In contrast, its neglected allies generally show low pest potential and consequently aspects of their biology and evolutionary history are underrepresented in the scientific literature.

Recent research has focused on a cryptic ally of *P. xylostella* (Landry and Hebert 2013), *P. australiana*, revealing key biological and genetic differences (Perry et al. 2018; Ward and Baxter 2018). *Plutella australiana* lacks population structure across its native range of Australia, has low pest potential and shows low tolerance to many insecticides (Perry et al. 2018). *Plutella xylostella* and *P. australiana* do have the ability to hybridize in laboratory crosses (Perry et al. 2018) raising concerns that insecticide resistance alleles may be exchanged. However, genome-wide scans have failed to identify widespread gene flow, supporting reproductive isolation in the field (Ward and Baxter 2018).

(Pryszcz and Gabaldón 2016). Insectav9 Benchmarking Universal Single Copy Orthologs were then annotated on the assembled genomes and the *P. xylostella* reference genome DBM FJ v1.1 (You et al. 2013) using BUSCOv3 (Waterhouse et al. 2017).

Nuclear divergence and phylogenetic inference

Single copy orthologs with complete open reading frames were aligned with Geneious v11.0 (Kearse et al. 2012) using amino-acid sequence to inform the alignment. Maximum likelihood gene trees were constructed for each of the single copy orthologs using RAxML (Stamatakis 2014) with a GTR

substitution model and gamma rate heterogeneity. The species tree was then inferred from gene trees using ASTRAL2 (Zhang et al. 2018).

Genetic distance between *P. xylostella* and the four other species assembled was calculated for each of the single copy ortholog alignments using Geneious v11.0 (Kearse et al. 2012). Codons within the alignment were partitioned based on their position: 1st plus 2nd codon positions, 3rd position of all codons and 3rd position of codons that show four-fold degeneracy.

2016a) mitochondrial genomes with Geneious v11.0 using the MAFFT algorithm. BEAST2 v2.5.1 (Bouckaert et al. 2018) was used to estimate the mitochondrial divergence date. Substitution models for each gene were estimated using the bModelTest package (Bouckaert and Drummond 2017). Unlinked lognormal relaxed clocks were applied to both COXI and 16S with priors set to 0.0177 (± 0.0019) and 0.0064 (± 0.0009) substitutions Myr⁻¹ according to Papadopoulou et al. (2010). Two independent MCMC chains of 2.5×10^8 were carried out, sampling every 10,000. Tracer v1.6 (Rambaut et al. 2018) as then used to visually inspect log files for model mixture and to determine burn-in (16%).

RESULTS AND DISCUSSION

Mitochondrial genome description

The complete circular mitochondrial genomes of *P. armoraciae*, *P. porrectella* and *A. assectella* were 15,569, 16,196 and 15,369 bp long, respectively (Figure 1), which is similar to the length of *P. xylostella* (16,014 bp) (Dai et al. 2016a) and *P. australiana* (15,962 bp) (Ward and Baxter 2018). The genomes of each of the species contained two rRNA, 13 protein coding and 22 tRNA genes in the conserved gene order for lepidopterans (Sun et al. 2017) (Figure 1).

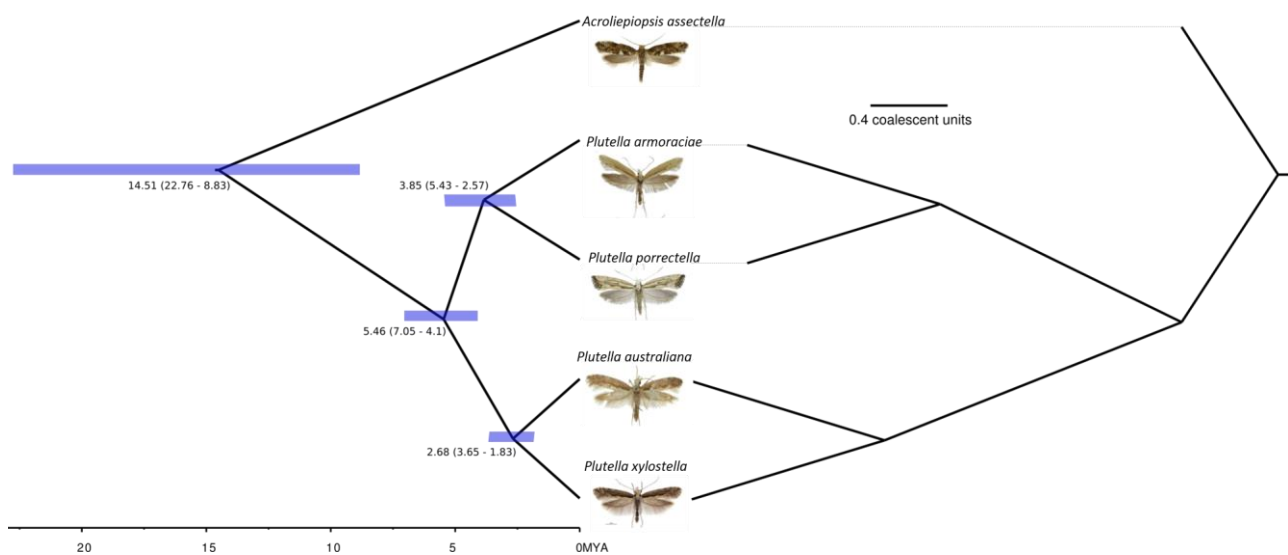


Figure 2: Mitochondrial (left) chronogram and nuclear species tree (right) for four *Plutella* species and the outgroup *Acrolepiopsis assectella*. LEFT: COX1 and 16S were used to estimate the mitochondrial divergence date of four *Plutella* species. Divergence is shown in millions of years and the 95% highest posterior density (HPD) intervals are indicated with brackets. RIGHT: 104 nuclear single copy orthologs were used to construct a maximum likelihood phylogeny using Astral III. All nodes had complete (1.0) posterior probability support.

Estimation of species divergence dates

Cytochrome oxidase I and 16S rRNA genes were extracted from each of the assembled mitochondrial genomes and aligned to their corresponding genes in *P. australiana* (MG787473) (Ward and Baxter 2018) and *P. xylostella* (KM023645) (Dai et al.

Each of the genomes contained an A-T rich D-loop of variable length. D-loop size correlated with total mitochondrial genome size, the longest being *P. porrectella* (1,179 bp). *Plutella armoraciae* and *A. assectella* had similar D-loop sizes of 579 and 490 bp. The A-T skew of the D-loop was similar

between all species with an average of 93.3% A-T (± 0.0086), consistent with reports from other lepidopterans (Dai et al. 2016b; Meng et al. 2016; Sun et al. 2017).

Protein coding genes made up 71.9% (11,198 bp), 69.1% (11,198 bp) and 72.6% (11,159 bp) of the total genome size in *P. armoraciae*, *P. porrectella* and *A. assectella*, respectively. Twelve mitochondrial protein coding genes contained a conventional start codon (ATN), whereas COX1 is initiated with CGA. Genetic distance between the mitochondrial genome of *P. xylostella* and genomes the other species analysed here was 9.43% (*P. porrectella*) and 8.06% (*P. armoraciae*). As expected, the outgroup *A. assectella*, showed the highest overall genetic distance to *P. xylostella* (16.8%).

Mitochondrial divergence of *Plutella*

Genetic distances between DNA sequences of individual mitochondrial protein coding gene were compared. The genetic distance between *P. xylostella* and *P. armoraciae* ranged from 6.19-11.9%, with an average of 8.05%. The genetic distance between *P. xylostella* and *P. porrectella* was 5.93-17.26% and had an average distance of 8.92% (Table 1). In both sequenced *Plutella* species, ATP8 showed the highest genetic distance to *P. xylostella*. The average genetic distance to *P. xylostella* among protein coding gene regions was lower than the complete mitochondrial genome, reported above (Table 1).

A relaxed molecular clock was applied to cytochrome oxidase I (0.0177 ± 0.0019 MYA) and 16S (0.0064 ± 0.0009 MYA) nucleotide alignments in order to date the mitochondrial divergence time. The topology agreed with mitochondrial gene trees produced by Landry and Hebert (2013) and Sølvi et al. (2018).

Plutella xylostella and *P. australiana* are were shown to be in reciprocal monophyly with *P. armoraciae* and *P. porrectella*. An estimated crown node age of 5.46 (95% HPD 7.05 – 4.10) MYA was generated using COX1 and 16S clock rates (Figure 2). Median divergence times of 2.68 (3.65 – 1.83) and 3.85 (5.43 – 2.57) MYA (95% HPD) were estimated for the *P. australiana/P. xylostella* and *P. armoraciae/P. porrectella* splits (Figure 2). The *P. australiana/P. xylostella* divergence date overlaps with the previously reported 1.96 (± 0.175) MYA (Ward and Baxter 2018). The discrepancy in divergence dates is likely due to the application of a relaxed log normal clock informed by both the 16S and COX1 substitution rates in this study.

Nuclear divergence of *Plutella*

Mitochondrial and nuclear genomes commonly show discordant topologies and divergence times

(Shaw 2002, Fisher-Reid and Wiens 2011, Zheng et al. 2011, Wallis et al. 2017). Therefore, we carried out nuclear genome assembly on the data and constructed a Maximum-Likelihood phylogeny from single copy orthologs.

Table 1: Genetic distance (%) and alignment length of the DNA sequences for the 13 mitochondrial protein coding genes to *P. xylostella*.

Gene name	Alignment Length	Genetic distance to <i>P. xylostella</i>		
		<i>Plutella armoraciae</i>	<i>Plutella porrectella</i>	<i>Acrolepiopsis assectella</i>
ATP6	678	6.64	7.96	15.63
ATP8	168	11.9	17.26	16.07
COX1	1531	7.7	9.86	12.34
COX2	679	6.63	7.51	10.75
COX3	789	8.88	9.38	14.83
CYTB	1152	9.03	7.64	14.94
ND1	942	8.07	8.07	13.06
ND2	1019	8.06	7.28	19.92
ND3	354	6.21	5.93	14.97
ND4	1339	7.02	7.99	14.94
ND4L	291	6.19	7.56	13.4
ND5	1724	7.83	7.19	15.16
ND6	540	10.49	12.36	23.52
Average	862	8.05	8.92	15.49

Nuclear genome assemblies of *P. armoraciae*, *P. armoraciae* and *A. assectella* were all highly fragmented with N_{50} values, (the shortest contig length needed to cover 50% of the genome) ranging from 4,017-6,157 bp. Of 1,658 complete benchmarking single copy orthologs (BUSCO) genes present in the insecta orthoDB v9 dataset, 931 were recovered in *P. armoraciae*, 1,163 in *P. porrectella*, 1,225 in *A. assectella*, and 1,300 in *P. australiana*. Only 383 BUSCO genes were both single copy and present across all five genome assemblies. These were then filtered such that all contained complete open reading frames, leaving 104 BUSCOs spanning 76.5 kb for analysis. Gene trees were then constructed using each of the 104 single copy orthologs and used to reconstruct a species tree. The species tree showed an identical topology to the mitochondrial chronogram (Figure 2). Genetic distance across codon partitions were then calculated between each of the species and *P. xylostella* (Table 2). This revealed *P. xylostella* nuclear genes share greater homology with *P. porrectella* than *P. armoraciae* across all codon partitions (Table 2), in contrast to the mitochondrial genome (Table 1). Genetic distance was greatest in the third position of four-fold degenerate codons (Table 2) as these are essentially neutral (Obbard et al. 2012).

Table 2: Genetic distance (%) of each codon partition ($1^{st}+2^{nd}$, 3rd) to *P. xylostella* for nuclear single copy ortholog nucleotide alignments. Genetic distance at the

3rd position of four-fold degenerate codons are shown separately.

Species	Genetic distance to <i>Plutella xylostella</i>		
	1 st +2 nd	3 rd	3 rd 4-fold
<i>P. australiana</i>	1.22	9.47	14.42
<i>P. armoraciae</i>	3.98	21.45	31.39
<i>P. porroctella</i>	3.97	21.15	30.86
<i>A. assectella</i>	13.3	47.24	61.9

CONCLUSION

Complete circularized mitochondrial genomes of two *Plutella* species and an outgroup Yponomeutoidean were successfully assembled and annotated. All genomes were similar in size to *P. xylostella* and contained two rRNA, 13 protein coding genes and 22 tRNA conserved throughout Insecta. The topology of the *Plutella* species sequenced were congruent between the mitochondrial and nuclear genomes. Furthermore, the topology was consistent with published literature. We also place the crown node age of *Plutella* at 5.46 (95% HPD 7.05 – 4.10) MYA.

Acknowledgements

We thank Paul Abram (Agriculture & Agri-Food Canada, Agassiz Research and Development Centre, British Columbia) and Peter Mason (Agriculture & Agri-Food Canada, Ottawa Research and Development Centre) for founding the *P. armoraciae* culture used, as well as providing unpublished results on the life history of *P. armoraciae*. CMW is funded by The Commonwealth Hill Trust and Grains Research Development Council.

References

Baraniak, E. 2007. Taxonomic revision of the genus *Plutella* Schrank, 1802 (Lepidoptera: Plutellidae) from the Palearctic Region with notes on its phylogeny. *Pismo Entomologiczne*.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, C. D. Jiggins, and M. L. Blaxter. 2011. Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. *PLOS ONE* 6:e19315.

Bernt, M., A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritsch, A. Pütz, M. Middendorf, and P. F. Stadler. 2013. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* 69:313-319.

Bigger, D. S., and L. R. Fox. 1997. High-density populations of diamondback moth have broader host-plant diets. *Oecologia* 112:179-186.

Bonnemaison, L. 1965. Insect Pests of Crucifers and Their Control. *Annual Review of Entomology* 10:233-256.

Bouckaert, R., T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. D. Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. Ogilvie, L. d. Plessis, A. Poppinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. 2018. BEAST 2.5: An Advanced

Software Platform for Bayesian Evolutionary Analysis. *BioRxiv*:474296.

Bouckaert, R. R., and A. J. Drummond. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC evolutionary biology* 17:42.

Chapman, J. W., D. R. Reynolds, A. D. Smith, J. R. Riley, D. E. Pedgley, and I. P. Woivod. 2002. High-altitude migration of the diamondback moth *Plutella xylostella* to the U.K.: a study using radar, aerial netting, and ground trapping. *Ecological Entomology* 27:641-650.

Dai, L.-S., B.-J. Zhu, C. Qian, C.-F. Zhang, J. Li, L. Wang, G.-Q. Wei, and C.-L. Liu. 2016a. The complete mitochondrial genome of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *Mitochondrial DNA Part A* 27:1512-1513.

Dai, L.-S., B.-J. Zhu, Y. Zhao, C.-F. Zhang, and C.-L. Liu. 2016b. Comparative Mitochondrial Genome Analysis of *Eligma narcissus* and other Lepidopteran Insects Reveals Conserved Mitochondrial Genome Organization and Phylogenetic Relationships. *Scientific Reports* 6:26387.

Dierckxsens, N., P. Mardulyn, and G. Smits. 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45:e18-e18.

Endersby, N. M., S. W. McKechnie, P. M. Ridland, and A. R. Weeks. 2006. Microsatellites reveal a lack of structure in Australian populations of the diamondback moth, *Plutella xylostella* (L.). *Molecular Ecology* 15:107-118.

Fisher-Reid, M. C., and J. J. Wiens. 2011. What are the consequences of combining nuclear and mitochondrial data for phylogenetic analysis? Lessons from *Plethodon* salamanders and 13 other vertebrate clades. *BMC evolutionary biology* 11:300-300.

Fu, X., Z. Xing, Z. Liu, A. Ali, and K. Wu. 2014. Migration of diamondback moth, *Plutella xylostella*, across the Bohai Sea in northern China. *Crop Protection* 64:143-149.

Furlong, M. J., D. J. Wright, and L. M. Dossall. 2013. Diamondback Moth Ecology and Management: Problems, Progress, and Prospects. *Annual Review of Entomology* 58:517-541.

Garrad, R., D. T. Booth, and M. J. Furlong. 2015. The effect of rearing temperature on development, body size, energetics and fecundity of the diamondback moth. *Bulletin of Entomological Research* 106:175-181.

Harcourt, D. G. 2012. Biology of the Diamondback Moth, *Plutella maculipennis* (Curt.) (Lepidoptera: Plutellidae), in Eastern Ontario. II. Life-History, Behaviour, and Host Relationships. *The Canadian Entomologist* 89:554-564.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)* 28:1647-1649.

Landry, J.-F., and P. D. N. Hebert. 2013. *Plutella australiana* (Lepidoptera, Plutellidae), an overlooked diamondback moth revealed by DNA barcodes. *ZooKeys* 327:43-63.

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

Meng, Z., C. Lei, X. Chen, and S. Jiang. 2016. Complete mitochondrial genome sequence of *Heliconius melpomene rosina* (Insecta: Lepidoptera: Nymphalidae). *Mitochondrial DNA Part A* 27:3911-3912.

Obbard, D. J., J. Maclennan, K.-W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins. 2012. Estimating Divergence Dates and Substitution Rates in the *Drosophila* Phylogeny. *Molecular Biology and Evolution* 29:3459-3473.

Papadopoulou, A., I. Anastasiou, and A. P. Vogler. 2010. Revisiting the Insect Mitochondrial Molecular Clock: The Mid-Aegean Trench Calibration. *Molecular biology and evolution* 27:1659-1672.

Perry, K. D., G. J. Baker, K. J. Powis, J. K. Kent, C. M. Ward, and S. W. Baxter. 2018. Cryptic *Plutella* species show deep divergence despite the capacity to hybridize. *BMC Evolutionary Biology* 18:77.

- Philips, C. R., Z. Fu, T. P. Kuhar, A. M. Shelton, and R. J. Cordero. 2014. Natural History, Ecology, and Management of Diamondback Moth (Lepidoptera: Plutellidae), With Emphasis on the United States. *Journal of Integrated Pest Management* 5:D1-D11.
- Pryszcz, L. P., and T. Gabaldón. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research* 44:e113-e113.
- Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* 67:901-904.
- Shaw, K. L. 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: What mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences* 99:16122.
- Smith, D. B., and M. K. Sears. 1984. Life history of *Plutella porrectella*, a relative of the Diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *The Canadian Entomologist* 116:913-917.
- Sohn, J.-C., J. C. Regier, C. Mitter, D. Davis, J.-F. Landry, A. Zwick, and M. P. Cummings. 2013. A Molecular Phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and Its Implications for Classification, Biogeography and the Evolution of Host Plant Use. *PLOS ONE* 8:e55066.
- Søli, G., L. Aarvik, and T. Magnussen. 2018. *Plutella polaris* Zeller, 1880 (Lepidoptera, Plutellidae) rediscovered at Svalbard, Norway, with comments on its taxonomic position. *Nota Lepidopterologica* 41:129-137.
- Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30:1312-1313.
- Sun, Y., C. Chen, J. Gao, M. N. Abbas, S. Kausar, C. Qian, L. Wang, G. Wei, B.-J. Zhu, and C.-L. Liu. 2017. Comparative mitochondrial genome analysis of *Daphnis nerii* and other lepidopteran insects reveals conserved mitochondrial genome organization and phylogenetic relationships. *PLOS ONE* 12:e0178773.
- Talekar, N. S., and A. M. Shelton. 1993. Biology, Ecology, and Management of the Diamondback Moth. *Annual Review of Entomology* 38:275-301.
- Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* 9:e112963.
- Wallis, G. P., S. R. Cameron-Christie, H. L. Kennedy, G. Palmer, T. R. Sanders, and D. J. Winter. 2017. Interspecific hybridization causes long-term phylogenetic discordance between nuclear and mitochondrial genomes in freshwater fishes. *Molecular Ecology* 26:3116-3127.
- Ward, C. M., and S. W. Baxter. 2018. Assessing Genomic Admixture between Cryptic *Plutella* Moth Species following Secondary Contact. *Genome Biology and Evolution* 10:2973-2985.
- Ward, C. M., T.-H. To, and S. M. Pederson. 2020. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. *Bioinformatics* 36:2587-2588.
- Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* 35:543-548.
- Wei, S.-J., B.-C. Shi, Y.-J. Gong, G.-H. Jin, X.-X. Chen, and X.-F. Meng. 2013a. Genetic Structure and Demographic History Reveal Migration of the Diamondback Moth *Plutella xylostella* (Lepidoptera: Plutellidae) from the Southern to Northern Regions of China. *PLOS ONE* 8:e59654.
- Wei, S.-J., B.-C. Shi, Y.-J. Gong, Q. Li, and X.-X. Chen. 2013b. Characterization of the mitochondrial genome of the diamondback moth *Plutella xylostella* (Lepidoptera: Plutellidae) and phylogenetic analysis of advanced moths and butterflies. *DNA and cell biology* 32:173-187.
- You, M., Z. Yue, W. He, X. Yang, G. Yang, M. Xie, D. Zhan, S. W. Baxter, L. Vasseur, G. M. Gurr, C. J. Douglas, J. Bai, P. Wang, K. Cui, S. Huang, X. Li, Q. Zhou, Z. Wu, Q. Chen, C. Liu, B. Wang, X. Li, X. Xu, C. Lu, M. Hu, J. W. Davey, S. M. Smith, M. Chen, X. Xia, W. Tang, F. Ke, D. Zheng, Y. Hu, F. Song, Y. You, X. Ma, L. Peng, Y. Zheng, Y. Liang, Y. Chen, L. Yu, Y. Zhang, Y. Liu, G. Li, L. Fang, J. Li, X. Zhou, Y. Luo, C. Gou, J. Wang, J. Wang, H. Yang, and J. Wang. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics* 45:220-225.
- Zalucki, M. P., A. Shabbir, R. Silva, D. Adamson, L. Shu-Sheng, and M. J. Furlong. 2012. Estimating the Economic Cost of One of the World's Major Insect Pests, *Plutella xylostella* (Lepidoptera: Plutellidae): Just How Long Is a Piece of String? *Journal of Economic Entomology* 105:1115-1129.
- Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics* 19:153.
- Zheng, Y., R. Peng, M. Kuro-o, and X. Zeng. 2011. Exploring Patterns and Extent of Bias in Estimating Divergence Time from Mitochondrial DNA Sequence Data in a Particular Lineage: A Case Study of Salamanders (Order Caudata). *Molecular biology and evolution* 28:2521-2535.
- Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29:2669-2677.

Chapter 6

A haploid diamondback moth (*Plutella xylostella* L.) genome assembly resolves 31 chromosomes and identifies a diamide resistance mutation.

Ward, C. M., Perry, K.D., Baker, K., Powis, K., Heckel D.G. & Baxter, S. W. A haploid diamondback moth (*Plutella xylostella* L.) genome assembly resolves 31 chromosomes and identifies a diamide resistance mutation. **Currently under review at Insect Biochemistry and Molecular Biology.**

Statement of Authorship

Title of Paper	A haploid diamondback moth (<i>Plutella xylostella</i> L.) genome assembly resolves 31 chromosomes and identifies an incompletely recessive diamide resistance mutation		
Publication Status	<input type="checkbox"/> Published	<input type="checkbox"/> Accepted for Publication	<input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
	<input checked="" type="checkbox"/> Submitted for Publication		
Publication Details	Under review at Insect Biochemistry and Molecular Biology		

Principal Author

Name of Principal Author (Candidate)	Christopher Ward		
Contribution to the Paper	Conceived research, carried out quantitative analysis and interpreted results. Provided samples for sequencing. Wrote the first version of the manuscript and edited subsequent versions.		
Overall percentage (%)	60		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/5/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Kym D Perry		
Contribution to the Paper	10% Conceived research, carried out quantitative analysis and interpreted results. Revised the manuscript.		
Signature		Date	9/4/2021

Name of Co-Author	Greg Baker		
Contribution to the Paper	5% Conceived research. Provided samples for sequencing. Revised the manuscript.		
Signature		Date	7/4/2021

Name of Co-Author	Kevin Powis		
-------------------	-------------	--	--

Contribution to the Paper	5% Conceived research. Carried out quantitative analysis and interpreted results. Reviewed the manuscript.
Signature	Date 6/4/21

Name of Co-Author	David Heckel
Contribution to the Paper	5% Conceived research. Revised the manuscript.
Signature	Date 7/4/2021

Name of Co-Author	Simon Baxter
Contribution to the Paper	15% Conceived research, carried out quantitative analysis and interpreted results. Provided samples for sequencing. Wrote the first version of the manuscript and edited subsequent versions.
Signature	Date 20/5/2021



A haploid diamondback moth (*Plutella xylostella* L.) genome assembly resolves 31 chromosomes and identifies a diamide resistance mutation

C.M. Ward^{a,1}, K.D. Perry^b, G. Baker^b, K. Powis^b, D.G. Heckel^c, S.W. Baxter^{d,*}

^a School of Biological Sciences, University of Adelaide, 5005, Australia

^b South Australian Research and Development Institute, Urrbrae, 5064, Australia

^c Department of Entomology, Max Planck Institute for Chemical Ecology, 07745, Jena, Germany

^d Bio21 Institute, School of BioSciences, University of Melbourne, 3052, Australia

ARTICLE INFO

Keywords:

Insecticide
Trio binning
Chlorantraniliprole
Flubendiamide
Cyclaniliprole
BUSCO
Annotation
Lepidoptera
Hi-C
Comparative genomics

ABSTRACT

The diamondback moth, *Plutella xylostella* (L.), is a highly mobile brassica crop pest with worldwide distribution and can rapidly evolve resistance to insecticides, including group 28 diamides. Reference genomes assembled using Illumina sequencing technology have provided valuable resources to advance our knowledge regarding the biology, origin and movement of diamondback moth, and more recently with its sister species, *Plutella australiana*. Here we apply a trio binning approach to sequence and annotate a chromosome level reference genome of *P. xylostella* using PacBio Sequel and Dovetail Hi-C sequencing technology and identify a point mutation that causes resistance to commercial diamides. A *P. xylostella* population collected from brassica crops in the Lockyer Valley, Australia (LV-R), was reselected for chlorantraniliprole resistance then a single male was crossed to a *P. australiana* female and a hybrid pupa sequenced. A chromosome level 328 Mb *P. xylostella* genome was assembled with 98.1% assigned to 30 autosomes and the Z chromosome. The genome was highly complete with 98.4% of BUSCO Insecta genes identified and RNAseq informed protein prediction annotated 19,002 coding genes. The LV-R strain survived recommended field application doses of chlorantraniliprole, flubendiamide and cyclaniliprole. Some hybrids also survived these doses, indicating significant departure from recessivity, which has not been previously documented for diamides. Diamide chemicals modulate insect Ryanodine Receptors (RyR), disrupting calcium homeostasis, and we identified an amino acid substitution (I4790K) recently reported to cause diamide resistance in a strain from Japan. This chromosome level assembly provides a new resource for insect comparative genomics and highlights the emergence of diamide resistance in Australia. Resistance management plans need to account for the fact that resistance is not completely recessive.

1. Introduction

The extensive level of genome size variation reported among insects (Gregory and Johnston, 2008; He et al., 2016; Westerman et al., 1987) has been attributed to duplication events (Li et al., 2018) and the proliferation of transposable or repetitive elements (Dufresne and Jeffery, 2011; Petersen et al., 2019). These features have traditionally constrained the ability to sequence and annotate high quality reference genomes, particularly large genomes (Li et al., 2019). Practical aspects including the ability to collect, culture and inbreed insect strains to remove heterozygosity (Zhan et al., 2011) have also caused problems with genome assembly, and low DNA yields from small insect species

often require pooling multiple specimens prior to sequencing which can increase diversity and heterozygosity (Richards and Murali, 2015). Solutions are emerging for these complex technical and biological obstacles, including advances in quality of long read technology that enable reads to span repetitive elements (Chaisson et al., 2015; Mahmoud et al., 2019), construction of sequencing libraries from low concentrations of DNA (Choo et al., 2020; Kingan et al., 2019), computational advances in assembly software (Kolmogorov et al., 2019; Nurk et al., 2020) and reduced dependence on obtaining DNA from inbred diploids. The requirement of generating inbred insect strains for reference genomes has recently been challenged by a trio binning strategy, which involves crossing two genetically diverse parents to produce hybrid offspring

* Corresponding author.

E-mail address: simon.baxter@unimelb.edu.au (S.W. Baxter).

¹ Present address: The Australian Wine Research Institute, Glen Osmond, South Australia, 5046, Australia.

with two distinct haploid genomes. Long sequence reads are generated for the hybrid (Koren et al., 2018), and parents are sequenced with short read Illumina technology to enable partitioning of the progeny reads into two haploid bins for independent assembly. This approach has been highly successful in cattle (Low et al., 2020) and insects (Yen et al., 2020).

Here we focus on the most destructive insect pest of brassica crops, the diamondback moth, *Plutella xylostella* (L.) which costs billions of dollars annually in terms of control and lost agricultural yield (Zalucki et al., 2012). *Plutella xylostella* has often evolved field resistance to synthetic and biological insecticides through mutations in insecticidal receptors, including diamides (Guo et al., 2014; Jouraku et al., 2020), pyrethroids (Schuler et al. 1998) and Bt toxins (Tabashnik et al. 1990), making the pest difficult to control. South America has recently been proposed as the species origin (You et al., 2020), from which it has invasively colonized all continents except Antarctica (Juric et al., 2017; Zalucki et al., 2012) through migration and human assisted dispersal. Molecular analysis of mitochondrial (Juric et al., 2017) and nuclear (You et al., 2020) markers have revealed very little or no population genetic structure at the continental level (Endersby et al., 2006; Perry et al., 2020; You et al., 2020), highlighting the ability for extensive regional movement.

Two parallel projects assembled *Plutella xylostella* reference genomes in 2013 (Jouraku et al., 2013; You et al., 2013) and produced useful databases for analysis (Tang et al., 2014). Jouraku et al. (2013) used the Roche 454 system to sequence DNA isolated from fourth instar larvae of a Bt-toxin susceptible strain, PXS, and generated an assembly with 88, 530 contigs, an N50 of 2273 bp and total size of ~186 Mb which was considerably smaller than the 338.7–339.4 Mb (± 1.1) genome estimate (Baxter et al., 2011). You et al. (2013) used a hybrid assembly approach combining both short read Illumina data and fosmid clones to generate a 394 Mb assembly with 1819 scaffolds and a larger N50 of 737 kb. Of these scaffolds, 171 (~111.9 Mb) were assigned to a chromosome using linkage mapping (Baxter et al., 2011), which represented around one third of the total genome. Polymorphic variation and repetitive sequences confounded the assembly process, despite efforts to sequence a single inbred individual (You et al., 2013). The discrepancy between observed (394 Mb) and expected (~339 Mb) genome size may have been caused by assembly gaps within scaffolds and the retention of divergent haplotype sequences. A third 385.6 Mb *P. xylostella* genome referenced “pacbioV1”, has been made available by the University of Liverpool through the database Lepbase (http://ensembl.lepbase.org/Plutella_xylostella_pacbioV1/Info/Index), and contains 3307 contigs with an N50 of 447 kb, and although details about this reference have not been published, it has been a useful resource for several studies (Harvey-Samuel et al., 2020; Jouraku et al., 2020).

Collectively these assemblies have facilitated numerous genome wide studies that have, for example, characterized gene families (Xia et al., 2015; You et al., 2015; Yu et al., 2015), helped understand the evolution of this pest (You et al., 2020) and enabled identification of insecticide resistance mutations (Liu et al., 2020), including diamide chemicals. Diamide insecticides are classed as Group 28 ryanodine receptor (RyR) modulators (<https://irac-online.org/modes-of-action/>) and have become increasingly important for managing lepidopteran pests, including *P. xylostella*. The ryanodine receptor (RyR) is a tetrameric calcium channel located in the endoplasmic and sarco-reticulum of neuromuscular tissues and enables the release of calcium, which is required for muscle contraction (Cordova et al., 2006). Insects contain a single RyR gene and the protein has six helical transmembrane domains at the C-terminal that form the calcium ion-conducting pore (Douris et al., 2017). Diamide insecticides interact with and activate RyR's, causing feeding cessation, muscle contraction, paralysis and death (Tohnishi et al., 2005). Ligand binding studies indicate a common or closely coupled binding sites exist for anthranilic (e.g. chlorantraniliprole) and phthalic acid (flubendiamide) diamides in Lepidoptera (Isaacs et al., 2012; Qi and Casida, 2013).

Field resistance to diamide insecticides has been reported in at least six Lepidoptera (Richardson et al., 2020), including *Plutella xylostella* (Cho et al., 2018; Guo et al., 2014; Troczka et al., 2012, 2017). Several point mutations that cause amino acid substitutions within the RyR C-terminal are known to result in diamide resistance in *P. xylostella*, including G4946E, I4790M and I4790K (Guo et al., 2014; Jouraku et al., 2020; Troczka et al., 2012). Based on the rabbit RyR1 structure (Yan et al., 2015), the G4946E mutation (helix S4) is in close proximity to the I4790M mutation (helix S2). Radioligand binding studies (Steinbach et al., 2015), calcium release assays with Sf9 cells (Troczka et al., 2015) and the development of *Drosophila* models expressing these mutations (Douris et al., 2017) support the link between the G4946E substitution and reduced binding of chlorantraniliprole and flubendiamide. Bioassays have shown strains fixed for G4946E are resistant to chlorantraniliprole and flubendiamide, with LC₅₀ values ranging from 23 mg/L to >1000 mg/L (Guo et al., 2014; Steinbach et al., 2015; Troczka et al., 2012, 2015), and cyantraniliprole has reduced efficacy (Jouraku et al., 2020). The I4790K mutation confers high-level resistance to a broader range of anthranilic (chlorantraniliprole and cyantraniliprole) and phthalic (flubendiamide) diamides (Jouraku et al., 2020), and I4790M provides moderate resistance to flubendiamide (Wang et al., 2020).

High-level resistance of *P. xylostella* to diamides was detected in Australia for the first time in 2018, following control failure with chlorantraniliprole in brassica vegetable crops in the Lockyer Valley, Queensland. In contrast to the widespread occurrence of insecticide resistance among Australian populations of *P. xylostella*, the Australian sister species, *Plutella australiana*, is highly susceptible to insecticides including Group 3A (alpha-cypermethrin), 6 (emamectin benzoate) and 28 (chlorantraniliprole) (Perry et al., 2018). *Plutella australiana* is found across Australia but was only identified in 2013 through mitochondrial amplicon sequencing, as it is morphologically indistinguishable from *P. xylostella* (Landry and Hebert, 2013). Genomic analysis suggests they do not hybridize in the wild, yet can be crossed in the laboratory using no-choice single pair mating experiments when supplied with canola seedlings as a stimulus. They diverged around 1.96 (± 0.175) MYA (Ward and Baxter, 2018) and can be distinguished at the molecular level as their mitochondrial and nuclear genomes show approximately 4% and 5% sequence divergence, respectively (Perry et al., 2018).

Here we apply a trio binning strategy to sequence and assemble a chromosome level haploid genome of *P. xylostella*. A single hybrid pupa generated from a cross between a diamide resistant *P. xylostella* male and a *P. australiana* female was sequenced using PacBio long-read technology and scaffolded using Hi-C. The genome contains 30 autosomes and the Z chromosome, and enabled identification of a RyR mutation associated with resistance to diamides chlorantraniliprole, cyantraniliprole and flubendiamide.

2. Materials and methods

2.1. Insect strains

An insecticide susceptible *P. xylostella* reference strain, Waite Susceptible (WS), has been reared on cabbage without insecticide exposure for 29 years (~335 generations). Field collections of *P. xylostella* were obtained from Tent Hill and Mt. Sylvia in the vegetable growing Lockyer Valley (LV) region of Queensland, Australia, during 2018 and combined to form a single laboratory population. The LV population was selected for three generations on cabbage plants sprayed directly with 0.1% v/v Coragen® insecticide, then 1% v/v Coragen® for 12 generations. We refer to this chlorantraniliprole resistant population as LV-R (Lockyer Valley, Resistant). *Plutella australiana* were collected using light traps at Angle Vale and Urrbrae, South Australia, in 2015 (Perry et al., 2018) and maintained on *Brassica napus*. *Plutella australiana* cultures were reared in laboratory cages at 26 \pm 2.0 °C and *P. xylostella* cultures at 22 \pm 2.0 °C, with a 14:10 (L:D) hour photoperiod. Both species were reared at the South Australian Research and Development Institute in separate

buildings to reduce the risk of culture cross-contamination.

2.2. Insect crosses and bioassays

Ten single pair crosses between *Plutella xylostella* LV-R males and *Plutella australiana* females were established in 100 mL plastic cups containing a canola (var. Stingray) seedling. After ~14 days, neonates were observed in one cage causing feeding damage to the canola seedling. Larvae were transferred to larger containers with fresh canola leaves and reared to pupation.

Reciprocal crosses were performed between the *P. xylostella* WS and LV-R strains ($\text{♀} \times \text{♂}$ $n = 100$, $\text{♂} \times \text{♀}$ $n = 100$) in 45 cm³ cages which contained young cabbage plants. Dose response bioassays were performed using 3rd instar larvae of the two parental strains and F₁ progeny from their two reciprocal crosses. Cabbage leaf discs were embedded in a 90 mm × 14 mm plastic Petri dishes (Techno Plas) containing 25 mL of 1% agar and ten 3rd instar larvae were placed onto each leaf. Applications of chlorantraniliprole (Coragen®, 200 g/L chlorantraniliprole, DuPont™), flubendiamide (Belt®, 480 g/L, Bayer™) or cyclaniliprole (NUL-3445, 50 g/L, Nufarm™), were applied using a Potter Spray Tower (Burkard Manufacturing Co. Ltd.) at an application rate of 3.499 ± 0.165 mg cm⁻². Four replicate plates were tested at eight concentrations of each insecticide, plus controls without insecticide. Two parameter log-logistic regression models were fitted using the R package drc v3.0-1 (Ritz et al., 2016) and plotted with ggplot2 (Wickham, 2016). Backcross progeny, generated from crossing F₁ heterozygous males with homozygous LV-R females, were assayed on cabbage leaf discs sprayed with Coragen® (14 mg/L) or control plates without insecticide.

2.3. Genomic DNA isolation and sequencing

Genomic DNA was isolated from a *P. xylostella* male and *P. australiana* female using the DNeasy Blood and Tissue Kit (Qiagen), after they had produced neonate hybrid progeny. High molecular weight DNA was isolated from a single female hybrid pupa two days after pupation by homogenization with a micropestle in a 1.7 mL microfuge tube containing DNA isolation buffer (0.1N NaCl, 50 mM Tris pH 8.0, 1 mM DTT, 10 mM EDTA, 0.2% SDS). Homogenate was incubated with 20 µL Proteinase K (20 mg/mL) and 2 µL RNase A (10 mg/mL) then DNA isolated using phenol:chloroform:isoamyl alcohol (25:24:1, Sigma-Aldrich) and ethanol precipitation. A low DNA input Sequel library was produced without size selection and whole genome sequencing was performed on a single hybrid pupa using six PacBio SMRT cells (*P. xylostella* reads SAMN17576273), and the two parents (SAMN17576274, SAMN17576275) using Illumina short read sequencing (Australian Genome Research Facility, AGRF). Hi-C libraries were prepared using the Dovetail™ Hi-C Kit using approximately 30 *P. xylostella* LV-R male pupae and sequenced on an Illumina NovoSeq (AGRF, SAMN17576276).

DNA was extracted from single individuals of *P. armoraciae* (SAMN17577696) and *P. porrectella* (SAMN17577695) using the Qiagen DNeasy kit according to manufacturer's protocol and sequenced using an Illumina NextSeq High Output kit (2 × 150bp) to an estimated 20–30X coverage, based on the assumption of a similar genome size to *P. xylostella*. Quality control of all raw data was carried out using FastQC v0.11.8 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and ngsReports v1.8 (Ward et al., 2020).

2.4. Genome assembly and scaffolding using Hi-C

PacBio circular consensus sequencing (CCS) reads were separated into *P. australiana* and *P. xylostella* haplotype genomes using parental Illumina reads with Canu v1.5 (Koren et al., 2017). The *P. xylostella* sequence data was then passed to Flye v2.8.3 (Kolmogorov et al., 2019) as corrected PacBio reads and assembled. Draft contigs were polished with pilon v1.23 (Walker et al., 2014) using *P. xylostella* paternal short read data mapped to the contigs with BWA MEM v0.7.17 (Li, 2013).

BUSCO v3 (Simão et al., 2015) was run using the Insecta and Endopterygota gene sets on the draft contigs.

Hi-C short reads were trimmed and properly paired using trimmomatic (Bolger et al., 2014), mapped to the polished contigs using the Juicer pipeline v1.6 (Durand et al., 2016) and *P. xylostella* genome scaffolded with the 3D-DNA pipeline v180419 (Dudchenko et al., 2017). Juice Box Tools identified misjoins belonging to other chromosome blocks and these were manually excised and retained as unplaced contigs. Contigs within each chromosome block were reverse-complemented where necessary to match Hi-C orientation across each chromosome. Quality control of the complete assembly and Hi-C interactions was further carried out by remapping the paired Hi-C reads to the scaffolded reference using BWA MEM v0.7.17 (Li, 2013) with the arguments -L0 in single end mode. BAMs were then passed to Hi-C explorer v3.6 (Ramírez et al., 2018) to generate quality reports and to visualize contact matrices. Gap closing was then carried out using Illumina polished PacBio reads with TGS-GapCloser v1.0.3 (Xu et al., 2020). We refer to this genome assembly as PxLV.1 (*Plutella xylostella* Lockyer Valley, version 1) and the nucleotide sequence is available through NCBI (JAHIBW000000000, BioProject PRJNA704777).

Co-linearity between PxLV.1 and an existing *P. xylostella* linkage map generated using Restriction site Associated DNA (RAD) Illumina sequencing (Baxter et al., 2011) was assessed using BLASTn (Altschul et al., 1997). BLAST hits were determined to be unambiguously assigned if there was only one hit with greater than 80% identity and query coverage. BLAST markers with identity and query coverage below 80% or those with multiple high-quality hits were marked as low confidence and not used to determine chromosome orientation or colinearity. Symap v5.0.6 (Soderlund et al., 2011) was used to determine synteny and identify chromosome fusions between *Bombyx mori* (Kawamoto et al., 2019), *Cydia pomonella* (Cpom.V2, accession GCA_003425675.2) and *P. xylostella* PxLV.1. Chromosome numbers were assigned based a published linkage map (Baxter et al., 2011).

2.5. Genome annotation

Genome wide repeat content was masked using Repeat Masker v4.1.1 (Cho et al., 2018; Kim et al., 2019; Smit, 2010) with a *de novo* library modeled using Repeat Modeler v2.0.1 (Smit and Hubley, 2008) and the Repbase library (20181026) (Bao et al., 2015). Simple repeats were also modeled and masked using TanTan v26 (Frith, 2011). The masked genome was then input into the Funannotate pipeline (<http://doi.org/10.5281/zenodo.2604804>, version 1.5.3) for gene prediction and annotation. Training was performed using publicly available *P. xylostella* RNAseq data (SRR179062, SRR179508, SRR179509, SRR179510, SRR179511) which was trimmed using Trimmomatic v0.39 (Bolger et al., 2014), mapped to the genome using Hisat2 v2.2.1 (Kim et al., 2019) and genome guided transcriptome assembly carried out using Trinity v2.11.0 (Haas et al., 2013). Transcripts were then mapped to the genome using BLAT v36 (Kent, 2002) and input into PASA v2.4.1 (Haas et al., 2003) to inform initial gene models.

Drosophila melanogaster proteins were obtained from Flybase (Dmel Release 6.18) and *Bombyx mori* proteins from SilkBase-Nov.2016 (Kawamoto et al., 2019) then aligned to the genome using DIAMOND v2.0.7 (Buchfink et al., 2015) and exonerate v2.4.0 (Slater and Birney, 2005) to enable gene prediction. Protein alignments were then compared to PASA gene predictions to select the best supported models. These models were then used as training hints for Augustus v3.2.2 (Stanke et al., 2006), GlimmerHMM v3.0.4 (Majoros et al., 2004), GeneMark-ES v4.62 (Lukashin and Borodovsky, 1998), and SNAP v2013_11_29 (Korf, 2004). High quality Augustus predictions (HiQ) were then used to perform CodingQuarry v2.0 (Testa et al., 2015) prediction. Gene models from each program were passed to Evidence Modeler v1.1.1 (Haas et al., 2008) with weighting 6 (PASA), 2 (Augustus HiQ), 1 (Augustus), 1 (GlimmerHMM), 1 (GeneMark-ES), 1 (SNAP) and 1 (CodingQuarry) to construct consensus models. Consensus models

were filtered to remove repetitive element gene models (eg. LINES) and models less than 100b in length. tRNAs were predicted using tRNA-scan-SE v2.0.7 (Lowe and Eddy, 1997).

PASA was used to add UTRs and establish exon/intron boundaries using assembled transcriptome data. RNAseq was quasi-mapped to PxlV.1 using Salmon v1.4.0 (Patro et al., 2017) to identify isoforms and to further refine intron-exon boundaries. Gene models were translated into protein coding sequence and pfam domains were identified with HMMer v3.1b2 (Finn et al., 2011) and BUSCO groups with BUSCO v3. InterProScan5 5.48–83 (Jones et al., 2014) was used to identify IPR domains, PANTHER groups, and GO annotations. EggNogMapper v2.0.1 (Huerta-Cepas et al., 2019) was used to assign COGs to genes and Swiss/UniProt database gene symbols.

2.6. Determining completeness and accuracy of the genome annotation

In order to determine the completeness and accuracy of the annotation three analyses were carried out. First, protein coding gene (PCG) datasets from *B. mori* (version SILKBASE Nov.2016) (Kawamoto et al., 2019), *C. pomonella* (version GCA_003425675.2) (Wan et al., 2019) and *P. xylostella* (version DBM_FJ_V1.1) (You et al., 2013) were compared with PxlV.1 annotated PCGs to determine completeness of the annotation using the BUSCO v3 Insecta and Endopterygota odbv9 (Zdobnov et al., 2017) gene sets. Second, open reading frames identified among the Trinity transcriptome assembly were translated and BLAST against PxlV.1 annotated PCGs with an 80% identity cut off. Query coverage was then used to determine the percentage of the query present in the annotated protein and plotted. Third, complete single copy BUSCOs shared between the *B. mori* and PxlV.1 annotations plus all 1:1 orthologs identified using reciprocal best BLAST hits were compared to investigate protein length and homology.

2.7. Sanger sequencing of PCR amplicons

DNA was isolated from moths, pupae or larvae using the DNeasy Blood and Tissue isolation kit (Qiagen). The cytochrome oxidase I gene (COI) was PCR amplified using MyTaq (Bioline) with primers PxC0IF (5'-TCAACAAATCATAAAGATATTGG- 3') and PxC0IR (5'-TAAACTT-CAGGGTGACCAAAAAATCA- 3') according to Perry et al. (2018). Primers for the Ryanodine Receptor were designed using Primer3 Plus (UNTERGASSER et al., 2012), RyR_ex113_F1 (5'-GTGAAGAA-GACGAGGACCCG- 3') and RyR_ex114_R1 (5'-ATGACGAGCTTGCC-CAGTG- 3') and amplified using 2 µL of MyTaq 5x buffer, 0.2 µL per primer (10 mM stock), ~10 ng DNA and 0.1 µL of MyTaq polymerase (Bioline). Samples were amplified at 95 °C for 2 min, then 35 cycles at 95 °C for 10 s, 60 °C for 15 s, 72 °C for 30 s and a final extension at 72 °C for 5 min. PCR amplicons were Sanger sequenced (AGRF).

2.8. Population and evolutionary genetic analysis

Genomes for *Plutella australiana*, *P. porrectella*, *P. armoraciae* and 116 *P. xylostella* individuals from You et al. (2020) (Table S1) were mapped to the *P. xylostella* mitochondrial genome (KM023645) using BWA-MEM v0.7.17 (Li, 2013). Genotypes were called using BCFtools v1.11 (Li, 2011) with ploidy set to $n = 1$ and filtered for a minimum depth of 10. Filtered genotypes were converted to the Genomic Data Structure (GDS) format using SeqArray v1.3 (Zheng et al., 2017) and imported into gear v0.1 (<https://github.com/CMWbio/gear>). RNAseq data (DRR191278) from the KA17 strain (Jouraku et al., 2020) was assembled using Trinity v2.11.0 (Haas et al., 2013) to obtain mitochondrial protein coding genes (12 were present in the transcriptome assembly ND1, ND2, ND3, ND4, ND4L, ND5, ND6, CYTB, COX1, COX2, COX3, ATP6) and RyR coding sequence. Alignments of mitochondrial and RyR sequences were performed using Geneious v10.3 and a phylogeny was constructed with RAxML-NG v1.0.1 based on concatenated mitochondrial genes using a GTR + G model. The “-all” argument was specified to generate bootstrap

replicates for the topology.

3. Results

3.1. Chromosome level assembly of a single *P. xylostella* haplotype

Interspecific crosses between a female *P. australiana* (ZW sex chromosomes) and male *P. xylostella* (ZZ) generated F₁ hybrids and a single female pupa was sequenced on six Pacific Biosciences SMRT cells, producing 90.3 Gbp of data with a median read length of 11 kb. The data was separated into either *P. xylostella* or *P. australiana* origin with k-mer based trio binning (Koren et al., 2018) using Illumina sequence data generated from each parent (~65X, Fig. 1). *Plutella xylostella* and *P. australiana* diverged ~2 million years ago and their nuclear genomes show around 5% divergence (Ward and Baxter, 2018) enabling almost all PacBio reads of *Plutella* origin (99.952%) to be unambiguously assigned to a parental haplotype. The remaining ~0.05% of unassigned sequence reads ($n = 1157$) had a short 1320 bp median read length and were discarded. Reads assigned to *P. xylostella* ($n = 1,163,016$; 9.12 Gb) were used to generate a haploid assembly containing 443 contigs, an N₅₀ of 2.03 Mb, longest contig of 9.35 Mb and total genome length of 328 Mb. Although assembly size is approximately 10 Mb less than the expected size based on flow cytometry (~338 Mb), it contained 98.4% Insecta and 96.3% Endopterygota complete BUSCO gene sets, suggesting the coding proportion of the genome is highly complete.

Hi-C linking reads scaffolded 98.1% of the assembled genome size into 30 autosomes and a single Z chromosome (Fig. S1). The remaining unassigned 352 sequences, referred to as ‘debris’ scaffolds (Durand et al., 2016), ranged from 595 bp to 570 kbp (median = 12446 bp). The N₅₀ for the chromosome level genome (including ‘debris’) was 11.06 Mb and the 16.12 Mb Z chromosome was the longest scaffold. Chromosome level scaffolds were mostly comprised of a few (4–7) mega-base length contigs, for example ~94% of the Z chromosome’s (Chr1) total length was contained within just 4 contigs (Fig. S1, Fig. S2). Manual correction was used to remove misassembled regions of the genome that showed high levels of contact with a different chromosome. After manual

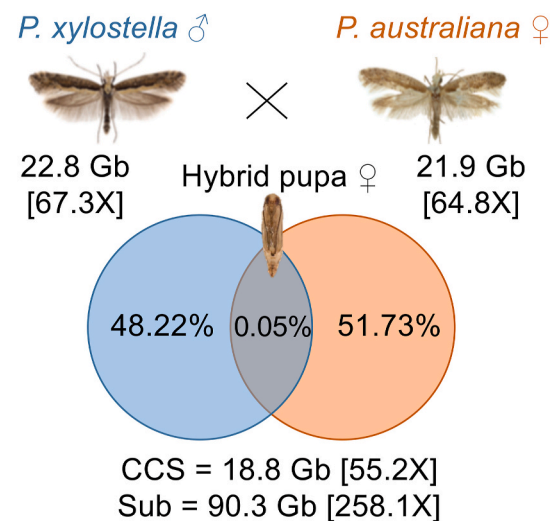


Fig. 1. A *P. xylostella* male was crossed to a *P. australiana* female and a single F₁ female hybrid pupae sequenced using a trio binning strategy. Genomic DNA from both parents was sequenced using Illumina short reads (150 bp, paired end) to a depth of more than 60-fold based on a *P. xylostella* genome size of 339 Mb (Baxter et al., 2011). The parental genomes were used to separate PacBio Sequel reads into two haplotypes, except 0.048% that could not be distinguished between species and were excluded from the assembly. The *P. xylostella* haplotype contained 1,163,016 reads and the *P. australiana* haplotype contained 1,247,498 reads. CCS, circular consensus sequencing. Sub, full-pass subreads.

curation, the Hi-C contact matrix showed no incongruent regions between chromosome assignment and contact peaks with 100 kb resolution (Fig. 2A). Structural variation within chromosomes is common in *P. xylostella* strains (You et al., 2013), therefore small regions of incongruence within contigs were not modified as this may represent polymorphic variation within the LV-R strain. Cumulative summation of PxLV.1, *Bombyx mori* (Kawamoto et al., 2019) and *Cydia pomonella* (Wan et al., 2019) scaffold length shows that they all share similar slope profiles (Fig. 2B). Despite variation in overall genome size, this indicates that a small number of scaffolds accounts for most of the genome in all three of these assemblies.

Next, the PxLV.1 genome was compared with an existing *P. xylostella* linkage map containing 2878 sequence markers 46 bp in length that were generated from unrelated strains (Baxter et al., 2011). Almost all markers had at least one BLAST hit in the genome ($n = 2476$), although many were low confidence matches or had multiple high quality BLAST hits ($n = 1440$). A total of 1036 markers were unambiguously assigned to a single locus (Figs. 2C), 98.9% for which were on the expected chromosome and 95.6% in co-linear orders. This linkage map was then used to orient and name each of the 31 chromosomes. Synteny analysis

with *B. mori* ($n = 28$) confirmed a homologous relationship between their respective chromosomes (Fig. 2D). The *C. pomonella* genome was used along with PxLV.1 to confirm previously reported fusions in *B. mori* chromosomes 11, 23 and 24 (Fig. 3A). Chromosome fusions observed in *C. pomonella* chromosomes 1, 2 and 3 (Wan et al., 2019) are not present in *P. xylostella* or *B. mori*, supporting their occurrence after the most recent common ancestor of Tortricidae and Obectomera lepidopterans (Fig. 3B).

BUSCO analysis of complete single copy genes in the scaffolded genome was consistent with the contig assembly (96.3% Endopterygota and 98.4% Insecta), and similar to values obtained from the 453 Mb *Bombyx mori* genome (96.5% Endopterygota, 98.4% Insecta) and the ~630 Mb *C. pomonella* genome (92.3% Endopterygota, 95.8% Insecta). This statistic is an improvement on the DBM_FJ_V1.1 genome which had 86.2% Endopterygota and 89.0% Insecta complete BUSCOs, with high levels of duplication (Fig. S3).

3.2. PxLV.1 genome annotation

Gene models predicted for the *P. xylostella* PxLV.1 reference genome

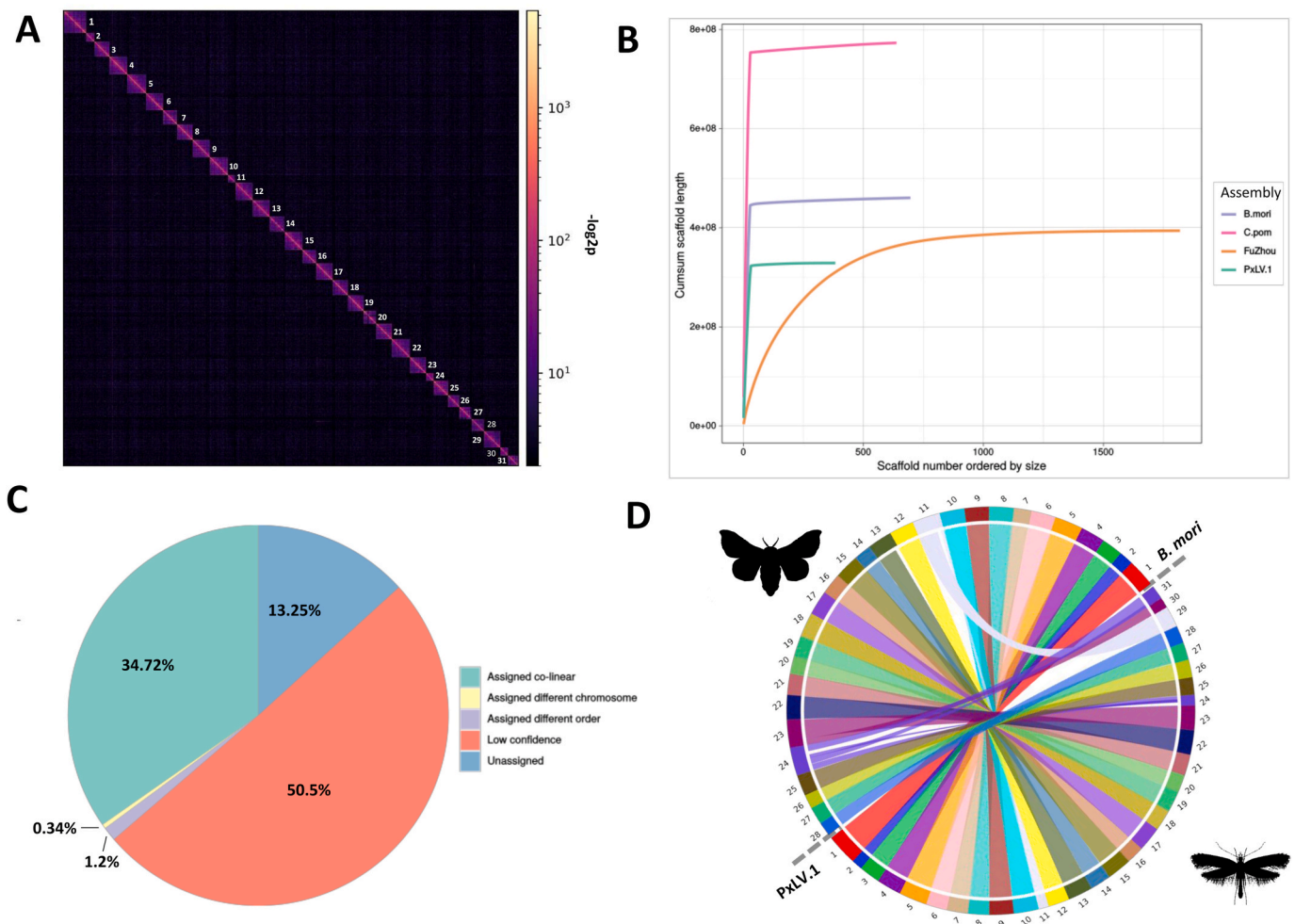


Fig. 2. Trio binning assembly of a single *P. xylostella* haplotype. A) Hi-C contact matrix of PxLV.1 showing $-\log_{2p}$ (natural logarithm of the given value plus one) counts plotted in 100 kb bins across the genome and delineation of 31 chromosomes (Chromosome 1 is the sex chromosome). B) Cumulative sum of scaffold lengths for *B. mori*, *C. pomonella*, *P. xylostella* PxLV.1 and *P. xylostella* DBM_FJ_V1.1 genomes. The PxLV.1 genome shows similar initial slopes to the chromosome level assemblies of *C. pomonella* and *B. mori* and fewer scaffolds than the DBM_FJ_V1.1 genome. C) Linkage map markers from Baxter et al. (2011) were BLAST against the PxLV.1 reference genome. *Unassigned* markers ($n = 378$) did not have a BLAST hit with at least 80% query coverage. *Low confidence* markers ($n = 1440$) had multiple hits each with >80% identity and >80% query coverage, or single hits with <80% identity. *Assigned markers* ($n = 991$) had unambiguous assignment to a single locus with >80% identity, >80% sequence coverage and were co-linear with the genome. *Assigned different order* ($n = 34$) were on the expected chromosome but not co-linear. *Assigned different chromosome* markers ($n = 11$) were not assigned to the expected chromosome. D) High levels of chromosomal synteny were observed between *Bombyx mori* ($n = 28$) and the PxLV.1 genome ($n = 31$), except for three *B. mori* chromosomes (11, 22, and 24) known to have undergone fusions.

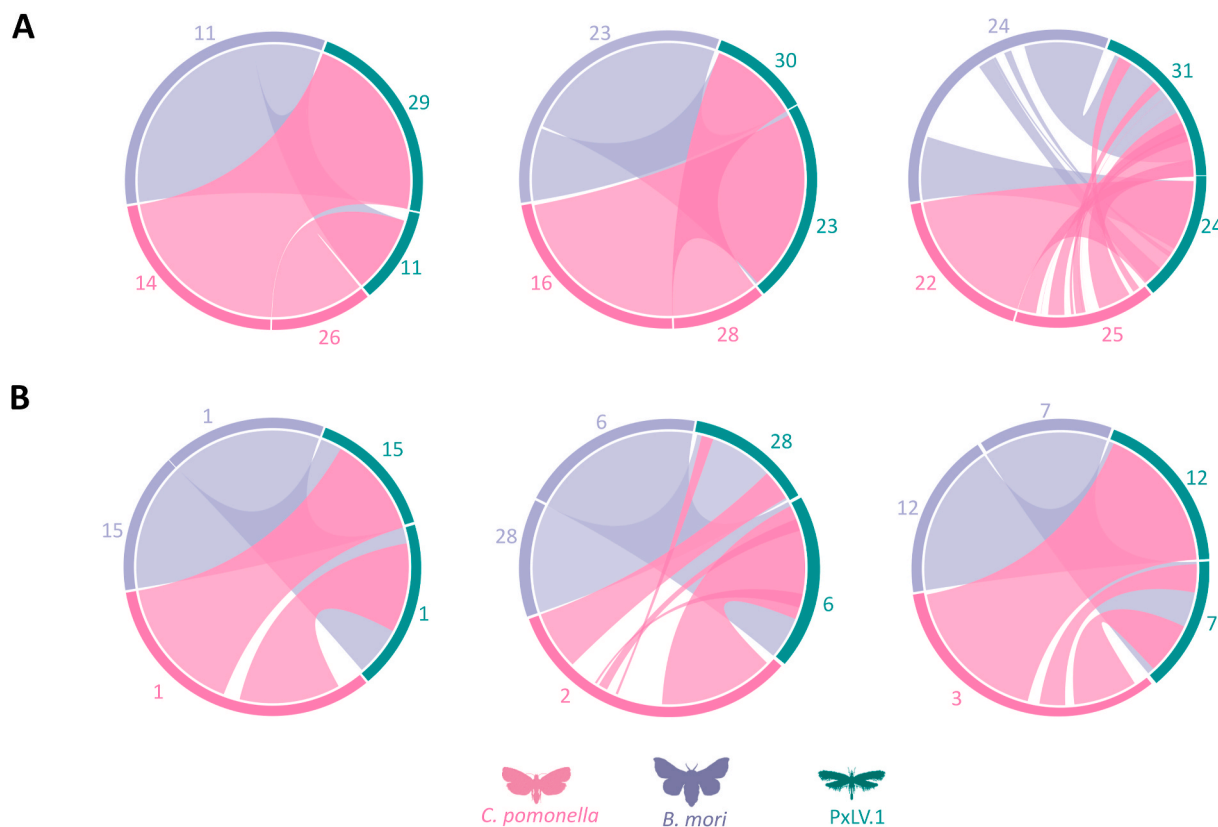


Fig. 3. Synteny plots comparing chromosomal fusions among *Bombyx mori* and *Cydia pomonella* with *P. xylostella*. A) Chromosome fusions present in *B. mori* (purple) chromosome 11, 23 and 24 are not present in either *C. pomonella* (pink) or *P. xylostella* PxLV.1 (green). B) Chromosome fusions present in *C. pomonella* chromosomes 1, 2 and 3 correspond to *P. xylostella* PxLV.1 and *B. mori* chromosomes 1|15, 6|28 and 7|12. *Plutella xylostella* has maintained the ancestral state of 31 chromosomes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

were developed using RNAseq transcriptome datasets (egg, larvae, pupae and moth) obtained from the NCBI Sequence Read Archive. RNAseq informed and *ab initio* based predictions with the funannotate pipeline annotated 19,002 protein coding genes and 4373 tRNAs. InterProScan, phmmer, EggNogg and Uni/SwissProt databases predicted functional annotations for 15,277 of the protein coding genes, which is similar to other lepidopterans (Kawamoto et al., 2019; Lu et al., 2019; Wan et al., 2019). Gene models predicted on the 31 chromosomes had an average length of 9210 bp (Table S2) and accounted for just over half of chromosome derived length (165.1 Mb, 51.1%). Interestingly, average exon (310 bp) and intron (1, 617 bp) length (Table S2) were greater than estimates in *B. mori* (exon = 290 bp, intron = 1, 330 bp) despite *B. mori*'s larger total genome size. Genes contained an average of six exons which is almost identical to the 6.1 exons per genes in the *B. mori* genome. BUSCO analysis of the annotation showed 94.7% of Insecta and 93.3% of Endopterygota genes were complete and single copy in which were similar to completeness estimates of the *B. mori* annotation (Fig. 4A, Fig. S4).

The quality and accuracy of gene annotations were then assessed against a separate *de novo* transcriptome assembly of all *P. xylostella* RNAseq datasets used for annotation. Translations of *de novo* transcript reading frames were compared against PxLV.1 predicted proteins using BLASTp to determine the query coverage of transcripts in the annotation, with an 80% identity cut off. Complete gene predictions are expected to have high query coverage as the whole transcript is contained within a single prediction, whereas fragmented genes show low query coverage as they only contain part of the transcript. The majority of transcripts showed >95% query coverage within the annotation, indicating most gene models contained full transcripts (Fig. 4B). Annotated *P. xylostella* complete single copy BUSCO genes (Endopterygota) were then compared to *B. mori* orthologues to broadly assess similarity in

length of these conserved proteins by dividing the length of PxLV.1 gene by the length of their *B. mori* ortholog. Of the 2442 genes in the Endopterygota dataset, 1436 were single copy in both the *B. mori* and PxLV.1 genome, 80.68% of which differed in length by less than 10% (Fig. 4C). Next, single copy orthologs shared between *P. xylostella* and *B. mori* were identified using the reciprocal best BLAST hit method, with sequence identity greater than 80% (n = 8561). Log2 transformation of the ratio revealed 63.17% of all orthologs differed in length by less than 10% (Fig. 4C).

3.3. The PxLV.1 genome contains RyR substitution I4790K

The automated gene prediction for the *P. xylostella* Ryanodine Receptor was annotated as four neighboring genes. Manual correction, guided by a full length mRNA sequence from *P. xylostella* reference strain "Roth" (GenBank accession JN801028), generated a 145,152 bp gene model on chromosome 31 that included 5-prime and 3-prime untranslated regions, 121 exons and 15,492 bp of coding sequence that was predicted to produce a 5164 amino acid protein (GenBank accession KAG7295227.1). Comparing PxLV.1 and Roth RyR sequences identified 10/5164 amino acid substitutions (Table S3), including one isoleucine to lysine (I4790K) substitution known to cause diamide resistance among a *Plutella* strain KA17 that was recently isolated in Japan (Jouraku et al., 2020). No other mutations known to cause diamide resistance were identified (Richardson et al., 2020).

Diamide resistance causing I4790K mutations in Australian populations may have arisen independently through *de novo* mutation or be derived from a common ancestor with East Asian I4790K alleles. These two alternate hypotheses are interesting in their own right, with the latter suggesting a recent migration of DBM may have occurred from East Asia. To investigate this, phylogenetic approaches were used to

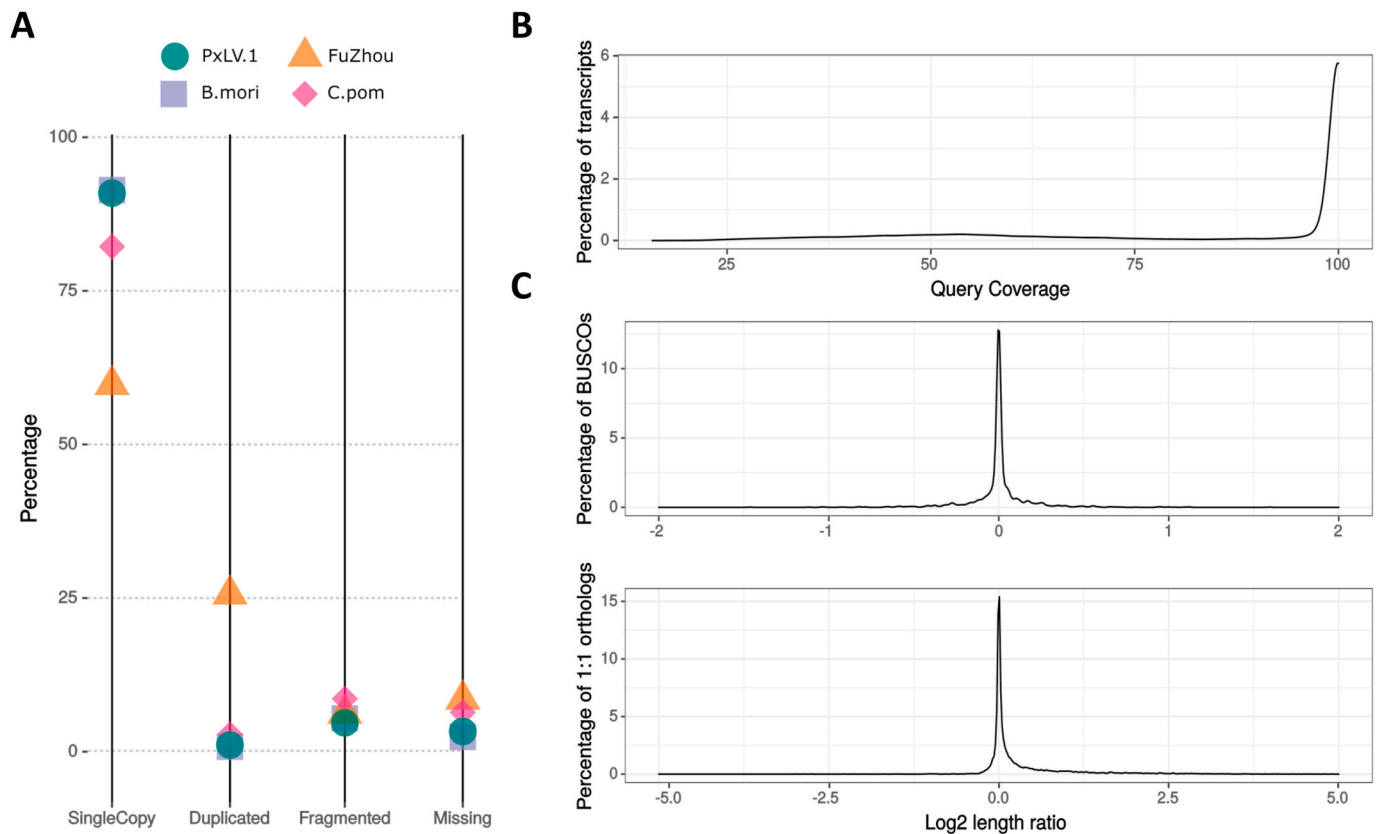


Fig. 4. RNAseq informed prediction of protein coding genes in the PxLV.1 genome. A) Parallel coordinate plot of single copy, duplicated, fragmented and missing BUSCO (Endopterygota) genes for *B. mori*, *C. pomonella* and *P. xylostella* strains LV-R (PxLV.1) and Fuzhou (DBM_FJ_V1.1). B). Query coverage (with an 80% identity cutoff) of translated proteins from a *de novo* *P. xylostella* RNAseq assembly were used to as a measure of gene completeness. High query coverage for BLASTp hits indicated that the complete open reading frame of a PxLV.1 annotated protein is contained within assembled transcripts. C) Log2 ratios comparing the length of protein orthologs between *B. mori* (SILKBASE-Nov.2016) and annotated *P. xylostella* protein predictions from the PxLV.1 genome. Upper. Comparison of 1436 complete and single copy BUSCO genes (Endopterygota) genes. Lower. Orthologs identified using reciprocal best BLAST hits ($n = 8561$). Values of zero indicate *B. mori* and *P. xylostella* orthologs were the same length. Values greater than (or less than) zero indicate *P. xylostella* orthologs were longer (or shorter) than *B. mori*.

compare RyR sequences from the Australian LV-R strain and Japanese KA17 strain to determine similarity between these populations. Assembling RNAseq data from KA17 produced a RyR contig that spanned exons 111 to 116 (803 bp) and contained the I4790K codon. The PxLV.1 genome and KA17 strain shared the same residue substitution in codon 4790 of exon 113, ATA to AAA, however there were 28 synonymous SNPs between the two strains among these six exonic regions (Fig. S5A). Phylogenetic reconstruction using this 803 bp RyR sequence from PxLV.1, KA17, the *P. xylostella* strain Roth (JX467684.1, JN801028.1) and two accessions of unknown origin (NM_001309073.1, JF927788.1) were then performed to assess whether the I4790K variant may have arisen from the same ancestral mutation (Fig. S5B). The partial gene tree indicates PxLV.1 and KA17 are sister taxa and suggests these alleles share a more recent common ancestor than the other sequences analyzed.

Genomes from 116 *P. xylostella* individuals collected from six continents by You et al. (2020) (Table S1), plus three *Plutella* outgroup species, were aligned to the PxLV.1 reference and genotyped for RyR codon 4790. As sequence coverage was very low for most individuals, only one allele with the best supported genotype based on called sequence depth was recorded for each individual. Of the 116 *P. xylostella* samples, six had insufficient read depth across RyR codon 4790 to generate confident genotype calls. The wild type susceptible Isoleucine codon (ATA) was present in *P. xylostella* ($n = 100$) across all continents whereas a synonymous ATT isoleucine codon was present in eight *P. xylostella* individuals from North America, South America, Europe, Asia and Africa. A single individual from Russia (EU) contained a

non-synonymous substitution (Fig. S6, I4790M), which has also been shown to confer moderate resistance to group 28 insecticides (Wang et al., 2020). Finally, an individual from Indonesia (AS) carried an I4790T substitution (Table S1) which has not been previously reported in *P. xylostella* but is present in at least two *Tuta absoluta* strains collected in Brazil (equivalent amino acid position 4746) (Roditakis et al., 2017). Genotypes for the outgroup species showed that *Plutella australiana* had the Isoleucine ATA codon, while *P. armoraciae* and *P. porrectella* contained a synonymous ATC codon. Based on our conservative genotyping strategy, we failed to identify any individual *Plutella* moths from this broad sampling dataset that carried the I4790K mutation. To assess the frequency of the I4790K mutation across Australia, 47 *P. xylostella* individuals from six distinct populations collected in 2014 (Perry et al., 2018) were PCR amplified and sequenced. All individuals were homozygous for the wild type isoleucine (ATA) at codon position 4790 (Table S4).

More than 60 mitochondrial cytochrome oxidase I (COI) haplotypes have been reported for *P. xylostella* (Juric et al., 2017) and Australian populations predominantly carry one of two common variants, PxCOI01 with a frequency of ≈ 0.75 and PxCOI04 a frequency of ≈ 0.22 (Perry et al., 2018). Four randomly selected LV-R individuals were sequenced and both PxCOI01 ($n = 2$) and PxCOI04 ($n = 2$) haplotypes were identified. To further investigate the relatedness of LV-R and KA17 strains, phylogenetic reconstruction using DNA sequence from mitochondrial protein coding genes was carried out on a concatenated partitioned alignment from 113/116 *P. xylostella* individuals from You et al. (2020) (three had insufficient sequence coverage), the LV-R paternal genome,

three outgroup *Plutella* species and KA17 (only 12 of the 13 protein coding genes were recovered). The mitochondrial phylogeny showed clear separation of clades by geography with high bootstrap support and placed South American samples in monophyly at the base of the *P. xylostella* clade. The KA17 sample was monophyletic with haplotypes present in Africa, Europe and Asia (AF, EU, AS) and LV-R was nested within the Oceanic clade with other Australian collections (Fig. 5, Fig. S7).

3.4. The I4790K mutation causes incompletely recessive resistance to diamides

Genetic crosses have associated RyR amino acid substitution I4790M with resistance to flubendiamide (Wang et al., 2020) and I4790K with resistance to cyantraniliprole (Jouraku et al., 2020). To test for genetic linkage between the 4790 K mutation and chlorantraniliprole resistance, single pair crosses between an LV-R male (4790 K) and WS female (I4790) were performed and F₁ males heterozygous for I4790K were backcrossed with individual LV-R females. Backcross progeny from two families were reared on either cabbage leaf discs embedded in agar (63/64 larvae survived) or cabbage discs treated with 14 mg/L of chlorantraniliprole (80/216 survived). The insecticide concentration applied was approximately >130-times higher than the LC₅₀ previously established for the WS strain and ~7-times higher than the LC₉₉ dose (Perry et al., 2018). Genotyping assays were performed on 22 control and 62 bioassay survivors and a significant association between the 4790 K genotype and survivorship was observed (X² = 21.733, p-value < 3e-06). Out of 62 backcross progeny treated with chlorantraniliprole, 57 were homozygous for the lysine substitution (K/K) and only five were heterozygous (I/K), representing a strong yet imperfect association between genotype and insecticide resistance phenotype (Table S5).

To assess whether the resistance phenotype was subject to a level of

incomplete dominance, dose-response bioassays were performed with chlorantraniliprole, flubendiamide and cyantraniliprole using the resistant LV-R strain, susceptible WS strain and F₁ progeny produced from reciprocal crosses between the two parental strains (LV-Rf x WSm and WSf x LV-Rm). The LC₅₀ levels observed in LV-R exceeded the recommended field application rates of all three insecticides assessed, predicting that control failure would likely occur using any of these chemicals against a *P. xylostella* population with these genotypes (Table 1, Fig. 6). The degree of dominance was calculated according to Stone (1964) using LC₅₀ data, producing incompletely recessive values between -0.638 and -0.456 (Table 1) (values of -1 are recessive and +1 are completely dominant). However, we observed that a small proportion of F₁ heterozygotes were able to survive field application rates of each chemical in these laboratory bioassays. The LC₉₉ values for F₁ heterozygotes presented in Fig. 6 are all higher than recommended application rates. The dose mortality curves for chlorantraniliprole predicted the concentration applied to backcross progeny (14 mg/L) was only sufficient to kill only 75.06% (66.81–83.30) of heterozygotes, and this is the likely reason genetic crosses failed to produce a more significant association between the resistance phenotype and RyR genotype (Table S6).

4. Discussion

4.1. A haploid assembly of the *P. xylostella* genome

Structural variation and nucleotide diversity between haplotypes are problematic for genome assembly processes (Patel et al., 2018), as they have the potential to assemble alleles of the same gene into paralogous gene models (Roach et al., 2018) and result in assemblies with inflated genome sizes. Many of the founding insect genome projects sequenced inbred (Mita et al., 2004) or isogenic (Adams et al., 2000) reference strains with reduced polymorphic variation relative to wild populations

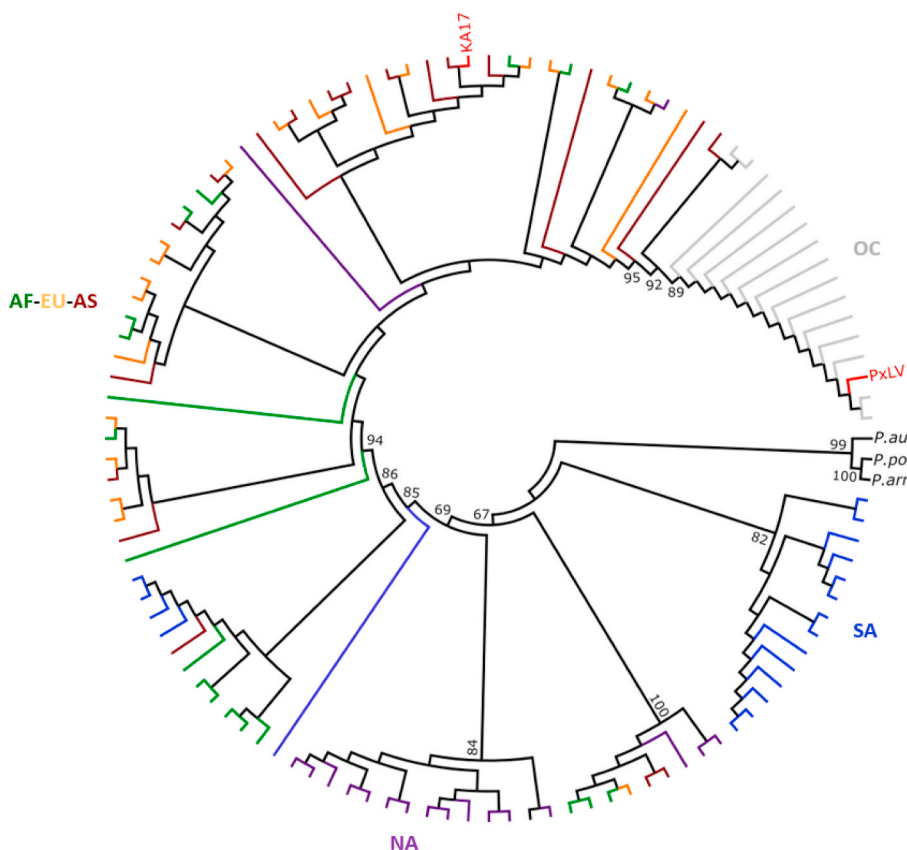


Fig. 5. The mitochondrial genomes of diamide resistant strains LV-R and KA17 are not closely related. A bootstrapped maximum likelihood phylogeny was constructed from a concatenated alignment of 12 mitochondrial protein coding genes (ND1, ND2, ND3, ND4, ND4L, ND5, ND6, CYTB, COX1, COX2, COX3, ATP6) for 113 *P. xylostella* samples (You et al., 2020), strains LV-R and KA17, and three outgroup *Plutella* species *P. australiana* (*P. aus*), *P. porrectella* (*P. por*) and *P. armoricaria* (*P. arm*). Diamide resistant strains with the I4790K mutation KA17 and LV-R are highlighted in red and show strong bootstrap support separating the two samples. Bootstrap support (n = 10,000) for informative internodes are indicated. *Plutella xylostella* samples are colored based on population of collection; South America (SA, blue), North America (NA, purple), Asia (AS, green), Africa (AF, orange), Oceania (OC, grey), Europe (EU, maroon) and Europe (EU, orange). Phylogenies with scaled branch lengths are reported in Fig. S7. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1

Bioassays of three diamide insecticides for *P. xylostella* strains Waite Susceptible (WS), Lockyer Valley (LV-R) and F₁ progeny of reciprocal crosses between WS and LV-R.

Chemical	^a Strain	^b Number	Slope ± SE	^c LC ₅₀	95% CL	RR	^d D
Chlorantranilprole (Coragen)	WS	360	-1.534 ± 0.163	0.105	(0.084–0.131)	1	
	LV-R	361	-0.903 ± 0.169	36044	(17436–74510)	342750	
	LV-Rf x WSm	360	-0.643 ± 0.069	3.374	(2.037–5.588)	32.081	-0.4555446
	WSf x LV-Rm	361	-0.727 ± 0.08	2.319	(1.464–3.674)	22.051	-0.5143794
Flubendiamide (Belt)	WS	362	-1.483 ± 0.154	0.331	(0.264–0.414)	1	
	LV-R	361	-0.497 ± 0.155	1226728	(60874–24720757)	3710761	
	LV-Rf x WSm	360	-0.736 ± 0.081	5.742	(3.632–9.079)	17.371	-0.6226975
	WSf x LV-Rm	363	-0.628 ± 0.068	5.105	(3.106–8.39)	15.443	-0.6382457
Cyclanilprole (NUL-3445)	WS	361	-1.488 ± 0.155	0.034	(0.028–0.043)	1	
	LV-R	360	-1.285 ± 0.134	640	(501–819)	18559	
	LV-Rf x WSm	362	-0.879 ± 0.102	0.274	(0.186–0.405)	7.944	-0.5760009
	WSf x LV-Rm	360	-0.87 ± 0.102	0.202	(0.137–0.297)	5.846	-0.6379439

^a “f”, female. “m”, male.

^b Number of larvae tested in bioassays.

^c Chemical concentration predicted to kill 50% of individuals.

^d Level of dominance in heterozygotes, ranging between -1 (completely recessive) and +1 (completely dominant). $D = (2X_2 - X_1 - X_3)/(X_1 - X_3)$, where X₁, X₂ and X₃ are logarithms of LC₅₀ values for strains LV-R, F₁ hybrids and WS.

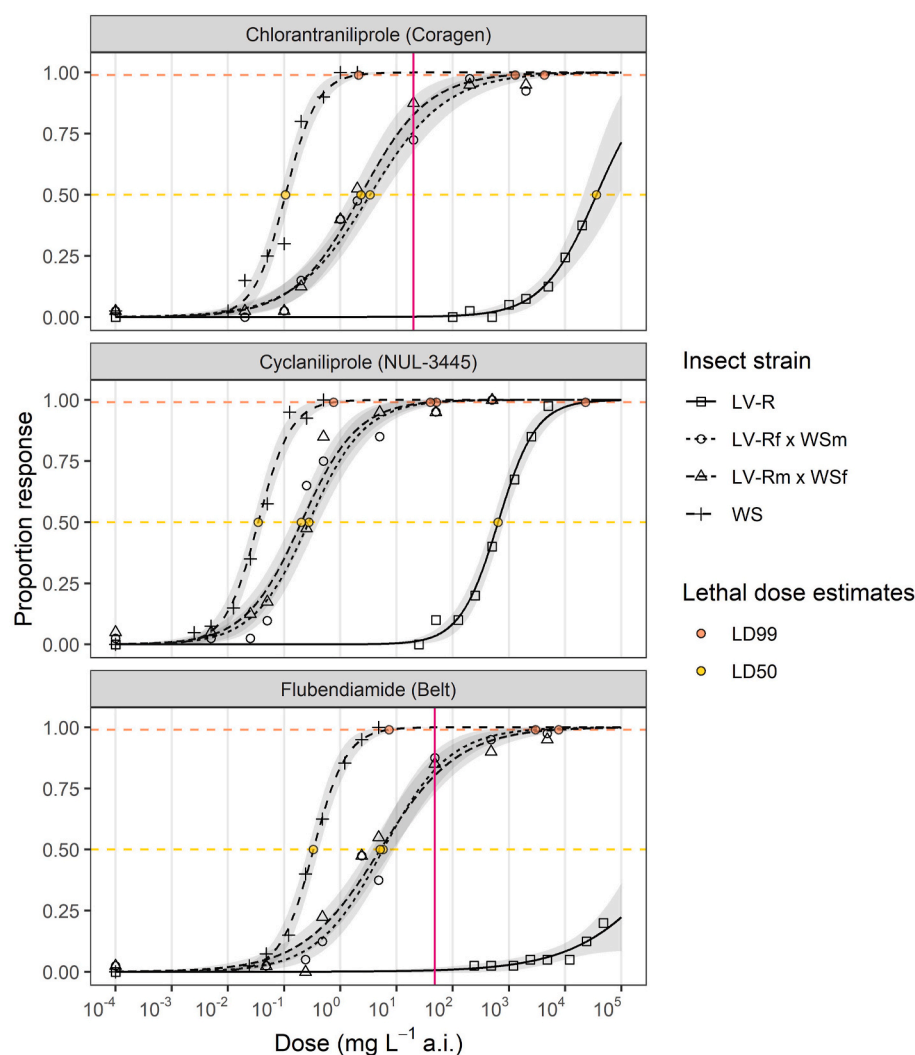


Fig. 6. Dose response bioassays curves comparing the laboratory susceptible reference strain (WS), the chlorantranilprole selected strain (LV-R) and two reciprocal F₁ crosses (LV-Rf x WSm and WSf x LV-Rm). The dashed horizontal yellow lines intersect with 50% mortality responses (LC₅₀ values) and dashed orange lines intersect with 99% mortality response (LC₉₉) values. The recommended field application rate for chlorantranilprole (Coragen®) is 20 g of active ingredient (a.i.) per hectare (20 mg/L) and flubendiamide (Belt®) is 48 g a.i. per hectare (48 mg/L), and are represented with vertical pink bars. Upper and lower confidence limits are shaded grey. Cyclanilprole (NUL-3445) is not currently registered for control of *P. xylostella* and therefore field application rates are not indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

in attempts to improve the quality of assemblies and length of contigs. While inbreeding insects such as *Drosophila melanogaster* has been successful at reducing variation and producing isogenic lines that can be propagated (Lack et al., 2016), accumulation of deleterious alleles

during the inbreeding process have often prevented establishing stable strains of other insects for sequencing and analysis. Several *P. xylostella* laboratory reference strains have been successfully propagated as large colonies for decades (Perry et al., 2018; Troczka et al., 2015; Zhao et al.,

2000), yet attempts to inbreed the Fuzhou-S strain to reduce heterozygosity presented a difficult problem when generating material for one of the first the *P. xylostella* reference genomes, as it could not be cultured indefinitely (You et al., 2013). High levels of polymorphic variation (Perry et al., 2018; Song et al., 2015; You et al., 2013, 2020; Zucchi et al., 2019) occur in *P. xylostella* genomes and heterozygosity in Australian field populations has been estimated around 0.009–0.010 using reduced representation sequencing (Perry et al., 2018) and 0.0348 with whole genome shotgun sequencing (Ward and Baxter, 2018). Insects with large populations and polymorphic genomes may maintain deleterious mutations through genetic drift, and the consequence of inbreeding could therefore impose high fitness costs.

Sequencing reference genomes using trio binning approaches relies on crossing two outbred parental strains, or two distinct species, to produce a hybrid with high genetic diversity from which true haploid genomes can be assembled (Koren et al., 2018; Yen et al., 2020). In this study, a trio binning strategy assembled a robust and highly contiguous contig assembly of a single *P. xylostella* haplotype that was scaffolded into 31 unambiguous chromosomes using Hi-C linked reads. The PxLV.1 reference genome represents a true haploid, chromosome level assembly of a basal Lepidopteran with the ancestral karyotype ($n = 31$). Synteny between *Bombyx mori*, *Cydia pomonella* and PxLV.1 chromosomes was high, as previously reported for Lepidoptera (Ahola et al., 2014; d'Alençon et al., 2010).

Chromosomes within the PxLV.1 assembly were able to identify breakpoints for previously reported chromosome fusions in *B. mori* (Baxter et al., 2011) and *C. pomonella* (Wan et al., 2019). Previous genome projects for *P. xylostella* identified putative expansions among some gene families (You et al., 2013) and facilitated characterization of genes with relevance to pest biology (You et al., 2015; Yu et al., 2015). The *P. xylostella* PxLV.1 reference genome provides a resource for future investigation of gene families, and as it was assembled from a single haplotype, gene content of this reference provides a useful benchmark to compare additional assembled genomes from diploid populations or strains in the future.

The assembled PxLV.1 reference was ~10 Mb less than the expected *P. xylostella* genome size estimated with flow cytometry performed on nuclei stained with propidium iodide (Baxter et al., 2011), which may have been caused by genome size variation between different insect strains, limitations of the PacBio library construction process and/or the inability to assign all sequence reads to a single parental species. The flow cytometry was performed using strain Geneva 88, a laboratory reference collected from Geneva New York in 1988 then maintained on artificial diet, and variation in transposable elements and repeat content may result in some variability of genome size relative to the Australian LV-R population (Biémont, 2008). The library construction process for trio binning with PacBio long read sequencing required high molecular weight DNA from an individual hybrid pupae, yet due to the small size of *Plutella*, only a limited concentration of ~1500 ng could be obtained. Despite successful library construction, size selection was not performed prior to sequencing due to concerns that the library yield recovered would be insufficient for six SMRT cells. Consequently, the median PacBio read length was relatively short and this may have restricted the capacity to separate, resolve and integrate long repetitive elements and telomeres, reducing the assembled genome size. A small proportion (0.048%) of reads could not be differentiated between *P. xylostella* and *P. australiana* and were excluded from the PxLV.1 assembly, which is also likely to account for a minor proportion of the ~10 Mb size difference. Despite this, Benchmarking Universal Single Copy Ortholog (BUSCO) content of the genome showed high levels of completeness relative to chromosomal assemblies of other lepidopterans. The genome annotation showed similar quality to the model Lepidopteran *B. mori*, although manual curation of annotations may be necessary for complete characterization of specific gene families.

4.2. Point mutations in RyR codon 4790

Ligand binding studies indicate common or closely coupled binding sites for anthranilic and phthalic diamides (Tohnishi et al., 2005). Two mutations at the RyR amino acid position 4790 have previously been reported in *P. xylostella*, including substitution of the conserved isoleucine to methionine (I4790M) reported from China (Guo et al., 2014) or isoleucine to lysine (I4790K) in a strain from Japan (Jouraku et al., 2020). Wang et al. (2020) used CRISPR/Cas9 homology directed repair to create a methionine codon in *P. xylostella* and this mutation (I4790M) causes increased levels of resistance to flubendiamide, yet only minor increases in tolerance to chlorantraniliprole and cyantraniliprole. The wild type *Drosophila melanogaster* RyR encodes a methionine at this codon, and engineering an isoleucine substitution confirmed an increase in susceptibility to flubendiamide of ~15-fold (Douris et al., 2017). Substituting the hydrophobic isoleucine for a positively charged and structurally unfavorable lysine at amino acid position 4790 disrupts diamide interaction and causes high levels of resistance in *P. xylostella* (Jouraku et al., 2020).

European or African populations have previously been proposed as sources of *P. xylostella* (Juric et al., 2017; Kfir, 1998; Talekar and Shelton, 1993), until You et al. (2020) indicated a South American origin for the species based on data generated from 532 whole genomes. *Plutella xylostella* only dispersed into Oceanic regions (including Australia) in the 1800's (You et al., 2020) and Australian populations contain just two predominant mitochondrial haplogroups (Juric et al., 2017; Perry et al., 2018) that are also present throughout Oceania, and at low frequencies in Asia (Juric et al., 2017). Phylogenetic reconstruction of mitochondrial genes extracted from whole genomes sequenced by You et al. (2020) revealed the LV-R mitochondrial haplotype is monophyletic with other derived Oceanic lineages and distinct from the KA17 haplotype. High bootstrap support was found for a basal South American clade and a mixed Afro-Eurasian clade, to which KA17 belongs. Low mitochondrial diversity present in Australian diamondback moth populations suggest migration events from are rare (Juric et al., 2017; Perry et al., 2018). However, insect migrants carrying insecticide resistance alleles at nuclear loci have the potential to interbreed and inherit local mitochondrial genomes, making them indistinguishable from native populations when compared using molecular assays.

The RyR coding sequence of LV-R and KA17 strains differed at multiple synonymous sites surrounding codon 4790, yet they represented most closely related alleles from limited publicly available sequence data. The LV-R and KA17 alleles may therefore be derived from a single ancestral I4790K mutation, present in the standing genetic variation before commercialization of diamides. Sequencing the genome of KA17 will enable comparative analysis to be performed with the PxLV.1 genome and provide opportunities for addressing the mutational origin using a molecular clock.

4.3. Degree of dominance of I4790K

Clear associations between genotype and diamide resistance phenotypes were previously established for *Plutella* strains homozygous for 4790 M (Wang et al., 2020) or 4790 K (Jouraku et al., 2020) mutations using larval bioassays of backcross progeny. Progeny homozygous for 4790 M survived on artificial diet treated with 0.15 $\mu\text{g}/\text{cm}^2$ flubendiamide (~6-fold higher than the susceptible LC_{50}) and no heterozygous genotypes were recovered (Wang et al., 2020). Jouraku et al. (2020) linked 4790 K with resistance using F_2 crosses selected with the field application rate of cyantraniliprole (51.5 mg/L). Associations with chlorantraniliprole resistance and the 4790 K mutation in this study were strong, yet imperfect when backcross progeny were selected with 14 mg/L (>130-fold higher than WS susceptible LC_{50}). Randomized or preferential combinations of the ryanodine tetramer among I4790K heterozygotes may influence diamide tolerance, as six basic combinations are possible from proteins produced by two different alleles

(excluding isoforms of the same allele). Each RyR channel could contain between zero and four proteins with the diamide resistant lysine substitution in a heterozygous individual. Variation in expression of each allele (Jouraku et al., 2020) could bias the ratio of the tetrameric assemblies further.

The degree of dominance (D) of a phenotype is often calculated using the LC₅₀ values of the wild type, homozygous mutant and heterozygote. For example, Steinbach et al. (2015) determined that D for RyR G4946E heterozygotes (Sudlon strain) was almost completely recessive, with values of -0.85 and -0.78 for flubendiamide and chlorantraniliprole, respectively (-1.0 is completely recessive). Values of D in CRISPR/Cas9 knock-in mutants I4790M were similar, although the level of resistance was low (Wang et al., 2020). Values of D obtained in this study are incompletely recessive, and ranged between -0.7 and -0.5 , depending on insecticide, when calculated with LC₅₀. However, low levels of larval survivorship were still achieved among heterozygotes at high insecticidal concentrations. Visualization of complete dose-response bioassays for heterozygotes (e.g. Fig. 6) may provide valuable insights into survivorship among other strains, beyond what can be achieved using calculations to estimate dominance.

These findings have important implications for *P. xylostella* management. Resistance ratios of homozygous resistant insects at the LC₅₀ of the three tested diamide chemistries ranged from 18,559 to 3,710,761 mg/L, indicating control failure will occur on Brassica farms using these chemistries. Due to the incomplete recessive nature of inheritance of the I4790K mutation, the two heterozygote strains showed approximately 2–3 orders of magnitude reduced susceptibility to the three diamide chemistries at the LC₉₉ level. Furthermore, the registered field application rates for the chlorantraniliprole and flubendiamide products are below the LC₉₉ estimates generated under ideal laboratory conditions. Therefore, a loss of efficacy against heterozygote insects is expected to occur with these chemistries used under field conditions. Because even a single resistance allele copy confers a significant selective advantage, frequencies of these alleles are likely to increase in local areas under continued selection pressure with diamide insecticides. To design suitable resistance management strategies, more understanding is required of frequencies in field populations, geographic spread and any fitness costs associated with the I4790K mutation.

4.4. Diamide resistance among Australian field populations

Resistance to insecticides is widespread throughout Australian *P. xylostella* populations, with high levels of resistance reported to older Group 1 and Group 3 chemistries (Baker and Kovaliski, 1999; Endersby et al., 2003). Field resistance is also present to Group 5 (Furlong et al., 2008), Group 6 (Rahman et al., 2010), and Group 22 (Eziah et al., 2008; Furlong et al., 2008) chemistries, along with evidence of Group 6/Group 28 cross resistance (Baker, 2013) and moderate Group 28 resistance (Perry et al., 2018). Although diamide insecticides have been widely used to control diamondback moth (Baker, 2013), the high levels of diamide resistance observed in the LV-R strain are yet to be described in field collections from other regions of Australia.

Genetic analysis of *P. xylostella* populations from across Australia show a lack of population structure using both genome sequencing (Perry et al., 2018, 2020) and microsatellite (Endersby et al., 2006) approaches. Using a subsample of the populations collected in 2014 by Perry et al. (2020), we were unable to detect resistant I4790K alleles from around Australia. Despite this, general usage of diamide insecticides could allow the RyR 4790 K allele to spread, creating an opportunity to use the 4790 K allele as a marker to investigate whether patterns of migration and movement of *P. xylostella* occur out of the Lockyer Valley, across the Australian continent. Therefore, repeating field genotyping will determine if allele frequencies of 4790 K have increased since 2014 and help assess whether additional populations carry this allele.

Recommended field concentrations of insecticide sprayed onto crops

generally exceed LC₉₉ values identified from laboratory bioassays. However, the field application rates for flubendiamide, chlorantraniliprole and cyclaniliprole were considerably less than the LC₉₉ for I4790K heterozygous individuals. It is important to note that bioassays conducted in the laboratory are generally under idealized conditions, and repeating these experiments using enclosed field cages warrants consideration. Due to the high level of resistance observed in 4790 K individuals, it may become increasingly important to monitor survival using the field rates in bioassays, rather than LC₉₉ or discriminating dose values to construct a more accurate model of field survivability. Molecular based assays designed to detect resistant alleles in the field will provide an important monitoring system to track the frequency of diamide resistance in Australia.

Acknowledgements

We thank Grant Cutler and Geoff Cornwell for providing diamondback moth field populations, collected in November 2018.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ibmb.2021.103622>.

Funding

This research was supported by the Australian Research Council (Grant FT140101303), the Grains Research and Development Corporation (Grant 9175870) and by the Max-Planck-Gesellschaft.

Author contributions

Wrote manuscript CMW, SWB.
Revised manuscript KDP, GB, KP, DGH.
Performed crosses CMW, SWB.
Performed bioassays KP, GB, KDP.
Computational analysis CMW.
Established and maintained insect colonies KP, GB, KDP.
Molecular analysis SWB.
Obtained funding SWB, CMW, GB, DGH.

References

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.-H.C., Blazee, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Miklos, A.B., Abril, J.F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., Pablos, B.D., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D.C., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.-Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R.-F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A.,

- Myers, E.W., Rubin, G.M., Venter, J.C., 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.
- Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Commun.* 5, 1–9.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Baker, G., 2013. Mechanism and Management of Insecticide Resistance in Australian Diamondback Moth, Chemicals & Pesticides. South Australian Research and Development Institute, Horticulture Australia Ltd, Sydney, Australia.
- Baker, G.J., Kovaliski, J., 1999. Detection of insecticide resistance in *Plutella xylostella* (L.) (Lepidoptera: plutellidae) populations in South Australian crucifer crops. *Aust. J. Entomol.* 38, 132–134.
- Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 11.
- Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D., Blaxter, M.L., 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One* 6, e19315.
- Biémont, C., 2008. Within-species variation in genome size. *Heredity* 101, 297–298.
- Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.
- Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611.
- Cho, S.-R., Kyung, Y., Shin, S., Kang, W., Jung, D.H., Lee, S., Park, G.-H., Kim, S., Cho, S. W., Kim, H., Koo, H., Kim, G., 2018. Susceptibility of field populations of *Plutella xylostella* and *Spodoptera exigua* to four diamide insecticides. *Kor. J. Appl. Entomol.* 57, 43–50.
- Choo, L.Q., Bal, T.M., Choquet, M., Smolina, I., Ramos-Silva, P., Marlétaz, F., Kopp, M., Hoarau, G., Peijnenburg, K.T., 2020. Novel genomic resources for shelled pteropods: a draft genome and target capture probes for *Limacina bulimoides*, tested for cross-species relevance. *BMC Genom.* 21, 1–14.
- Cordova, D., Benner, E.A., Sacher, M.D., Rauh, J.J., Sopa, J.S., Lahm, G.P., Selby, T.P., Stevenson, T.M., Flexner, L., Gutteridge, S., Rhoades, D.F., Wu, L., Smith, R.M., Tao, Y., 2006. Anthranilic diamides: a new class of insecticides with a novel mode of action, ryanodine receptor activation. *Pestic. Biochem. Physiol.* 84, 196–214.
- d'Alençon, E., Sezutsu, H., Legeai, F., Permal, E., Bernard-Samain, S., Gimenez, S., Gagneur, C., Cousserans, F., Shimomura, M., Brun-Barale, A., Flutre, T., Couloux, A., East, P., Gordon, K., Mita, K., Quesneville, H., Fournier, P., Feyereisen, R., 2010. Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc. Natl. Acad. Sci. Unit. States Am.* 107, 7680.
- Douris, V., Papapostolou, K.M., Ilias, A., Roditakis, E., Kounadi, S., Riga, M., Nauen, R., Vontas, J., 2017. Investigation of the contribution of Ryr target-site mutations in diamide resistance by CRISPR/Cas9 genome modification in *Drosophila*. *Insect Biochem. Mol. Biol.* 87, 127–135.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95.
- Dufresne, J., Jeffery, N., 2011. A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Res.* 19, 925–938.
- Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., Aiden, E. L., 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3, 95–98.
- Endersby, N.M., McKechnie, S.W., Ridland, P.M., Weeks, A.R., 2006. Microsatellites reveal a lack of structure in Australian populations of the diamondback moth, *Plutella xylostella* (L.). *Mol. Ecol.* 15, 107–118.
- Endersby, N.M., Ridland, P.M., Zhang, J., 2003. Reduced susceptibility to permethrin in diamondback moth populations from vegetable and non-vegetable hosts in southern Australia. *Urania* 19, 191–196.
- Eziah, V.Y., Rose, H.A., Clift, A.D., Mansfield, S., 2008. Susceptibility of four field populations of the diamondback moth *Plutella xylostella* L. (Lepidoptera: yponomeutidae) to six insecticides in the Sydney region, New South Wales, Australia. *Aust. J. Entomol.* 47, 355–360.
- Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37.
- Frith, M.C., 2011. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* 39, e23–e23.
- Furlong, M.J., Spafford, H., Ridland, P.M., Endersby, N.M., Edwards, O.R., Baker, G.J., Keller, M.A., Paull, C.A., 2008. Ecology of diamondback moth in Australian canola: landscape perspectives and the implications for management. *Aust. J. Exp. Agric.* 48, 1494–1505.
- Gregory, T.R., Johnston, J., 2008. Genome size diversity in the family Drosophilidae. *Heredity* 101, 228–238.
- Guo, L., Liang, P., Zhou, X., Gao, X., 2014. Novel mutations and mutation combinations of ryanodine receptor in a chlorantraniliprole resistant population of *Plutella xylostella* (L.). *Sci. Rep.* 4, 6924.
- Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr., R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R., 2008. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7.
- Harvey-Samuel, T., Norman, V.C., Carter, R., Lovett, E., Alphey, L., 2020. Identification and characterization of a Masculinizer homologue in the diamondback moth, *Plutella xylostella*. *Insect Mol. Biol.* 29, 231–240.
- He, K., Lin, K., Wang, G., Li, F., 2016. Genome sizes of nine insect species determined by flow cytometry and k-mer analysis. *Front. Physiol.* 7, 569.
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattai, T., Jensen, L.J., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314.
- Isaacs, A.K., Qi, S., Sarpong, R., Casida, J.E., 2012. Insect ryanodine receptor: distinct but coupled insecticide binding sites for [N-C(3)H(3)]chlorantraniliprole, flubendiamide, and [(3)H]ryanodine. *Chem. Res. Toxicol.* 25, 1571–1573.
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.
- Jouraku, A., Kuwazaki, S., Miyamoto, K., Uchiyama, M., Kurokawa, T., Mori, E., Mori, M. X., Mori, Y., Sonoda, S., 2020. Ryanodine receptor mutations (G4946E and I4790K) differentially responsible for diamide insecticide resistance in diamondback moth. *Plutella xylostella* L. *Insect biochemistry and molecular biology* 118, 103308.
- Jouraku, A., Yamamoto, K., Kuwazaki, S., Urio, M., Suetetsugu, Y., Narukawa, J., Miyamoto, K., Kurita, K., Kanamori, H., Katayose, Y., Matsumoto, T., Noda, H., 2013. KONAGAbase: a genomic and transcriptomic database for the diamondback moth, *Plutella xylostella*. *BMC Genom.* 14, 464.
- Juric, I., Salzburger, W., Balmer, O., 2017. Spread and global population structure of the diamondback moth *Plutella xylostella* (Lepidoptera: plutellidae) and its larval parasitoids *Diadegma semiclausum* and *Diadegma fenestrata* (Hymenoptera: ichneumonidae) based on mtDNA. *Bull. Entomol. Res.* 107, 155–164.
- Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., Fujiyama, A., Kiuchi, T., Yamamoto, K., Shimada, T., 2019. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 107, 53–62.
- Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Kfir, R., 1998. Origin of the diamondback moth (Lepidoptera: plutellidae). *Ann. Entomol. Soc. Am.* 91, 164–167.
- Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915.
- Kingan, S.B., Heaton, H., Cudini, J., Lambert, C.C., Baybayan, P., Galvin, B.D., Durbin, R., Koriach, J., Lawniczak, M.K., 2019. A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes* 10, 62.
- Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546.
- Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P., Phillippy, A.M., 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36, 1174–1182.
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736.
- Korf, I., 2004. Gene finding in novel genomes. *BMC Bioinform.* 5, 59.
- Lack, J.B., Lange, J.D., Tang, A.D., Corbett-Detig, R.B., Pool, J.E., 2016. A thousand fly genomes: an expanded *Drosophila* genome nexus. *Mol. Biol. Evol.* 33, 3308–3313.
- Landry, J.-F., Hebert, P.D., 2013. *Plutella australiana* (Lepidoptera, Plutellidae), an overlooked diamondback moth revealed by DNA barcodes. *ZooKeys* 43–63.
- Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., Walters, J.R., 2019. Insect genomes: progress and challenges. *Insect Mol. Biol.* 28, 739–758.
- Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993.
- Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM arXiv preprint arXiv:1303.3997.
- Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J., Barker, M.S., 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci. Unit. States Am.* 115, 4713–4718.
- Liu, Z., Fu, S., Ma, X., Baxter, S.W., Vasseur, L., Xiong, L., Huang, Y., Yang, G., You, S., You, M., 2020. Resistance to *Bacillus thuringiensis* Cry1Ac toxin requires mutations in two *Plutella xylostella* ATP-binding cassette transporter paralogs. *PLoS Pathog.* 16, e1008697.
- Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D.M., Rosen, B.D., Kronenberg, Z.N., Kingan, S.B., Tseng, E., 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.* 11, 1–14.
- Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964.
- Lu, S., Yang, J., Dai, X., Xie, F., He, J., Dong, Z., Mao, J., Liu, G., Chang, Z., Zhao, R., Wan, W., Zhang, R., Li, Y., Wang, W., Li, X., 2019. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* 8.
- Lukashin, A.V., Borodovsky, M., 1998. GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res.* 26, 1107–1115.

- Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., Sedlazeck, F.J., 2019. Structural variant calling: the long and the short of it. *Genome Biol.* 20, 246.
- Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879.
- Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., Kadono-Okuda, K., Yamamoto, K., Ajimura, M., Ravikumar, G., Shimomura, M., Nagamura, Y., Shin, I.T., Abe, H., Shimada, T., Morishita, S., Sasaki, T., 2004. The genome sequence of silkworm, *Bombyx mori*. *DNA Res.* 11, 27–35.
- Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Lomsdson, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., Koren, S., 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305.
- Patel, S., Lu, Z., Jin, X., Swaminathan, P., Zeng, E., Fennell, A.Y., 2018. Comparison of three assembly strategies for a heterozygous seedless grapevine genome assembly. *BMC Genom.* 19, 57.
- Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419.
- Perry, K.D., Baker, G.J., Powis, K.J., Kent, J.K., Ward, C.M., Baxter, S.W., 2018. Cryptic *Plutella* species show deep divergence despite the capacity to hybridize. *BMC Evol. Biol.* 18, 77.
- Perry, K.D., Keller, M.A., Baxter, S.W., 2020. Genome-wide analysis of diamondback moth, *Plutella xylostella* L., from Brassica crops and wild host plants reveals no genetic structure in Australia. *Sci. Rep.* 10, 12047.
- Petersen, M., Armişen, D., Gibbs, R.A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., Misof, B., 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol. Biol.* 19, 11.
- Qi, S., Casida, J.E., 2013. Species differences in chlorantraniliprole and flubendiamide insecticide binding sites in the ryanodine receptor. *Pestic. Biochem. Physiol.* 107, 321–326.
- Rahman, M.M., Baker, G., Powis, K.J., Roush, R.T., Schmidt, O., 2010. Induction and transmission of tolerance to the synthetic pesticide emamectin benzoate in field and laboratory populations of diamondback moth. *J. Econ. Entomol.* 103, 1347–1354.
- Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., Manke, T., 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* 9, 1–15.
- Richards, S., Murali, S.C., 2015. Best practices in insect genome sequencing: what works and what doesn't. *Current Opinion in Insect Science* 7, 1–7.
- Richardson, E.B., Troczka, B.J., Gutbrod, O., Davies, T.G.E., Nauen, R., 2020. Diamide resistance: 10 years of lessons from lepidopteran pests. *J. Pest. Sci.* 93, 911–928.
- Ritz, C., Baty, F., Streibig, J.C., Gerhard, D., 2016. Dose-response analysis using R. *PLoS One* 10, e0146021.
- Roach, M.J., Schmidt, S.A., Borneman, A.R., 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* 19, 460.
- Roditakis, E., Steinbach, D., Moritz, G., Vasakis, E., Stavrakaki, M., Ilias, A., García-Vidal, L., Martínez-Aguirre, M.d.R., Bielza, P., Morou, E., Silva, J.E., Silva, W.M., Siqueira, H.A.A., Iqbal, S., Troczka, B.J., Williamson, M.S., Bass, C., Tzarakarakou, A., Vontas, J., Nauen, R., 2017. Ryanodine receptor point mutations confer diamide insecticide resistance in tomato leafminer, *Tuta absoluta* (Lepidoptera: gelechiidae). *Insect Biochem. Mol. Biol.* 80, 11–20.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212.
- Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* 6, 31.
- Smit, A.F., 2010. Repeat-masker open-3.0.
- Smit, A.F., Hubley, R., 2008. RepeatModeler open-1.0. Available from.
- Soderlund, C., Bomhoff, M., Nelson, W.M., 2011. SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 39, e68–e68.
- Song, S.V., Downes, S., Parker, T., Okeshott, J.G., Robin, C., 2015. High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep. *Heredity* 115, 460–470.
- Stanke, M., Keller, O., Gunduz, L., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439.
- Steinbach, D., Gutbrod, O., Lümmen, P., Matthiesen, S., Schorn, C., Nauen, R., 2015. Geographic spread, genetics and functional characteristics of ryanodine receptor based target-site resistance to diamide insecticides in diamondback moth, *Plutella xylostella*. *Insect Biochem. Mol. Biol.* 63, 14–22.
- Talekar, N.S., Shelton, A.M., 1993. Biology, ecology, and management of the diamondback moth. *Annu. Rev. Entomol.* 38, 275–301.
- Tang, W., Yu, L., He, W., Yang, G., Ke, F., Baxter, S.W., You, S., Douglas, C.J., You, M., 2014. DBM-DB: the diamondback moth genome database. *Database* 2014, bat087. <https://doi.org/10.1093/database/bat087>.
- Testa, A.C., Hane, J.K., Ellwood, S.R., Oliver, R.P., 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genom.* 16, 170.
- Tohnishi, M., Nakao, H., Furuya, T., Seo, A., Kodama, H., Tsubata, K., Fujioka, S., Kodama, H., Hirooka, T., Nishimatsu, T., 2005. Flubendiamide, a novel insecticide highly active against lepidopteran insect pests. *J. Pestic. Sci.* 30, 354–360.
- Troczka, B., Zimmer, C.T., Elias, J., Schorn, C., Bass, C., Davies, T.G., Field, L.M., Williamson, M.S., Slater, R., Nauen, R., 2012. Resistance to diamide insecticides in diamondback moth, *Plutella xylostella* (Lepidoptera: plutellidae) is associated with a mutation in the membrane-spanning domain of the ryanodine receptor. *Insect Biochem. Mol. Biol.* 42, 873–880.
- Troczka, B.J., Williams, A.J., Williamson, M.S., Field, L.M., Lümmen, P., Davies, T.G.E., 2015. Stable expression and functional characterisation of the diamondback moth ryanodine receptor G4946E variant conferring resistance to diamide insecticides. *Sci. Rep.* 5, 14680.
- Troczka, B.J., Williamson, M.S., Field, L.M., Davies, T.G.E., 2017. Rapid selection for resistance to diamide insecticides in *Plutella xylostella* via specific amino acid polymorphisms in the ryanodine receptor. *Neurotoxicology* 60, 224–233.
- Walker, B.J., Abee, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963.
- Wan, F., Yin, C., Tang, R., Chen, M., Wu, Q., Huang, C., Qian, W., Rota-Stabelli, O., Yang, N., Wang, S., Wang, G., Zhang, G., Guo, J., Gu, L., Chen, L., Xing, L., Xi, Y., Liu, F., Lin, K., Guo, M., Liu, W., He, K., Tian, R., Jacquin-Joly, E., Franck, P., Siegwart, M., Ometto, L., Anfora, G., Blaxter, M., Meslin, C., Nguyen, P., Daliková, M., Marec, F., Olivares, J., Maugin, S., Shen, J., Liu, J., Guo, J., Luo, J., Liu, B., Fan, W., Feng, L., Zhao, X., Peng, X., Wang, K., Liu, L., Zhan, H., Liu, W., Shi, G., Jiang, C., Jin, J., Xian, X., Lu, S., Ye, M., Li, M., Yang, M., Xiong, R., Walters, J.R., Li, F., 2019. A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance. *Nat. Commun.* 10, 4237.
- Wang, X., Cao, X., Jiang, D., Yang, Y., Wu, Y., 2020. CRISPR/Cas9 mediated ryanodine receptor 14790M knockin confers unequal resistance to diamides in *Plutella xylostella*. *Insect Biochem. Mol. Biol.* 125, 103453.
- Ward, C.M., Baxter, S.W., 2018. Assessing genomic admixture between cryptic *Plutella* morphotypes following secondary contact. *Genome Biology and Evolution* 10, 2973–2985.
- Ward, C.M., To, T.-H., Pederson, S.M., 2020. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. *Bioinformatics* 36, 2587–2588.
- Westernman, M., Barton, N.H., Hewitt, G.M., 1987. Differences in DNA content between two chromosomal races of the grasshopper *Podisma pedestris*. *Heredity* 58, 221–228.
- Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
- Xia, X., Yu, L., Xue, M., Yu, X., Vasseur, L., Gurr, G.M., Baxter, S.W., Lin, H., Lin, J., You, M., 2015. Genome-wide characterization and expression profiling of immune genes in the diamondback moth, *Plutella xylostella* (L.). *Sci. Rep.* 5, 9877.
- Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., Zhang, Y., 2020. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* 9.
- Yan, Z., Bai, X., Yan, C., Wu, J., Li, Z., Xie, T., Peng, W., Yin, C., Li, X., Scheres, S.H.W., Shi, Y., Yan, N., 2015. Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. *Nature* 517, 50–55.
- Yen, E.C., McCarthy, S.A., Galarza, J.A., Generalovic, T.N., Pelan, S., Nguyen, P., Meier, J.I., Warren, I.A., Mappes, J., Durbin, R., Jiggins, C.D., 2020. A haplotypresolved, de novo genome assembly for the wood tiger moth (*Arctia plantaginis*) through trio binning. *GigaScience* 9 (8), gia0088.
- You, M., Ke, F., You, S., Wu, Z., Liu, Q., He, W., Baxter, S.W., Yuchi, Z., Vasseur, L., Gurr, G.M., 2020. Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nat. Commun.* 11, 1–8.
- You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., Zhan, D., Baxter, S.W., Vasseur, L., Gurr, G.M., Douglas, C.J., Bai, J., Wang, P., Cui, K., Huang, S., Li, X., Zhou, Q., Wu, Z., Chen, Q., Liu, C., Wang, B., Li, X., Xu, X., Lu, C., Hu, M., Davey, J.W., Smith, S.M., Chen, M., Xia, X., Tang, W., Ke, F., Zheng, D., Hu, Y., Song, F., You, Y., Ma, X., Peng, L., Zheng, Y., Liang, Y., Chen, Y., Yu, L., Zhang, Y., Liu, Y., Li, G., Fang, L., Li, J., Zhou, X., Luo, Y., Gou, C., Wang, J., Wang, J., Yang, H., Wang, J., 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* 45, 220–225.
- You, Y., Xie, M., Ren, N., Cheng, X., Li, J., Ma, X., Zou, M., Vasseur, L., Gurr, G.M., You, M., 2015. Characterization and expression profiling of glutathione S-transferases in the diamondback moth, *Plutella xylostella* (L.). *BMC Genom.* 16, 152.
- Yu, L., Tang, W., He, W., Ma, X., Vasseur, L., Baxter, S.W., Yang, G., Huang, S., Song, F., You, M., 2015. Characterization and expression of the cytochrome P450 gene family in diamondback moth, *Plutella xylostella* (L.). *Sci. Rep.* 5, 8952.
- Zalucki, M.P., Shabbir, A., Silva, R., Adamson, D., Shu-Sheng, L., Furlong, M.J., 2012. Estimating the economic cost of one of the world's major insect pests, *Plutella xylostella* (Lepidoptera: plutellidae): just how long is a piece of string? *J. Econ. Entomol.* 105, 1115–1129.
- Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simão, F.A., Ioannidis, P., Seppey, M., Loetscher, A., Kriventseva, E.V., 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45, D744–D749.
- Zhan, S., Merlin, C., Boore, J.L., Reppert, S.M., 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell* 147, 1171–1185.
- Zhao, J.-Z., Collins, H.L., Tang, J.D., Cao, J., Earle, E.D., Roush, R.T., Herrero, S., Escriche, B., Ferré, J., Shelton, A.M., 2000. Development and characterization of diamondback moth resistance to transgenic broccoli expressing high levels of Cry1C. *Appl. Environ. Microbiol.* 66, 3784–3789.
- Zheng, X., Gogarten, S.M., Lawrence, M., Stulp, A., Conomos, M.P., Weir, B.S., Laurie, C., Levine, D., 2017. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* 33, 2251–2257.
- Zucchi, M.I., Cordeiro, E.M.G., Allen, C., Novello, M., Viana, J.P.G., Brown, P.J., Manjunatha, S., Omoto, C., Pinheiro, J.B., Clough, S.J., 2019. Patterns of genome-wide variation, population differentiation and SNP discovery of the red banded stink bug (*Piezodorus guildinii*). *Sci. Rep.* 9, 14480.

Chapter 7

Adaptation of a major insect pest species to a new host plant is underpinned by a complex genetic mechanism and dynamic transcriptional response.

Ward, C. M., Breen, J., Heckel D.G. & Baxter, S.W (2020). Adaptation of a major insect pest species to a new host plant is underpinned by a complex genetic mechanism and dynamic transcriptional response. **Unpublished.**

Statement of Authorship

Title of Paper	Adaptation of a major insect pest species to a new host plant is underpinned by a complex genetic mechanism and dynamic transcriptional response
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	N/A

Principal Author

Name of Principal Author (Candidate)	Christopher Ward			
Contribution to the Paper	Conceived research, carried out quantitative analysis and interpreted results. Wrote the first version of the manuscript and edited subsequent versions.			
Overall percentage (%)	85			
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 60%;"></td> <td style="width: 10%;">Date</td> <td style="width: 30%;">18/5/2021</td> </tr> </table>		Date	18/5/2021
	Date	18/5/2021		

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Jimmy Breen			
Contribution to the Paper	5% Conceived research. Interpreted results.			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 60%;"></td> <td style="width: 10%;">Date</td> <td style="width: 30%;">21/4/2021</td> </tr> </table>		Date	21/4/2021
	Date	21/4/2021		

Name of Co-Author	David G. Heckel		
Contribution to the Paper	5% Conceived research. Interpreted results. Reviewed the manuscript.		
Signature		Date	10/5/2021

Name of Co-Author	Simon W. Baxter		
Contribution to the Paper	5% Conceived research. Interpreted results. Reviewed the manuscript.		
Signature		Date	20/5/2021

1 **Adaptation of a major insect pest species to a new host plant is**
2 **underpinned by a complex genetic mechanism and dynamic**
3 **transcriptional response**

4 Christopher M. Ward¹, Jimmy Breen², David G. Heckel³, Simon W. Baxter⁴

5 ¹ School of Biological Sciences, University of Adelaide, Australia

6 ² South Australian Genomics Center, Australia

7 ³ Department of Entomology, Max Planck Institute for Chemical Ecology, Germany

8 ⁴ School of BioSciences, University of Melbourne, Australia

9

10 **Running title:** Genomic and transcriptomic insights into host plant preference

11

12 **Key Words:** *Plutella*, Diamondback moth, Host plant preference, Adaptation, RNAseq,
13 Glutathione-S-Transferases, UDP-glucurotransferases

14

15

16 **Abstract:**

17 Host plant range expansion of specialist insect herbivores represents a key event in adaptive
18 speciation and pest management. In 1999, Kenyan populations of the *Brassicaceae*
19 specialist insect pest diamondback moth were found to be infesting sugar-snap pea crops
20 (*Pisum sativum*), presenting a troubling problem for agriculture. Here we investigated the
21 genetic and transcriptomic mechanisms underlying this surprising adaptation. Through
22 utilizing reduced representation genome sequencing of both back and intercross pedigrees
23 we constructed a robust linkage map across the diamondback moth's 31 chromosomes. We
24 then investigated allele frequency across the linkage map to identify five putative QTLs that
25 segregate with survival on *P. sativum*. However, QTLs were not fixed between crosses
26 suggesting the mechanism is highly complex. Transcriptomic response was then
27 investigated in a tissue specific manner for larval midgut and head capsules revealing a
28 complex metabolic response to oxidative stress when diamondback moth fed on the new
29 host plant. In head tissue, we identified host plant specific expression patterns in odorant
30 binding proteins and enrichment of response to chemical stimuli GO terms between *Pisum*
31 feeding and *Brassica* feeding diamondback moth strains. Collectively, these complementary
32 genomic approaches increase our understanding of how insects select their hostplant and
33 the complexity of adaptations required to increase host plant range.

34 **Introduction:**

35 Insect herbivores and their host plants engage in an on-going evolutionary competition for
36 dominance. Host plants develop complex chemical and physical defences to deter or
37 prevent insect herbivory, while insect larvae must counter these defence strategies (Ehrlich
38 and Raven 1964). Plants have evolved a diverse range of insecticidal metabolites that
39 disrupt vital processes within the insect (Kortbeek et al. 2019), including synaptic-signalling
40 (Yang and Guthrie 1969), digestion (Houseman et al. 1992), and protein function (Wink and
41 Schimmer 2018). The importance of this 'co-evolutionary' arms race was outlined by Ehrlich
42 and Raven (1964) through assessment of lepidopteran host plant usage, noting that the
43 diversity of secondary metabolites present across plants is likely to be the main driver for
44 insect biodiversity (Mello and Silva-Filho 2002, Becerra 2007, 2015). Since then, a large
45 body of literature investigating genetic mechanisms underlying host plant secondary
46 metabolite production and their detoxification by insect herbivores has developed (Ratzka et
47 al. 2002, Wittstock et al. 2004, Heidel-Fischer and Vogel 2015, Schweizer et al. 2017). Yet,
48 little is understood about how insect herbivores adapt to new host plants.

49 Holometabolous insects undergo metamorphosis and as a by-product face a unique set of
50 challenges when adapting to a new host plant (Anholt 2020). Adult females must first be

51 willing to oviposit, then neonate larvae must initiate feeding and overcome host chemical or
52 physical defence mechanisms to ultimately complete development. Over evolutionary time,
53 change in host plant can result in either host plant range expansion or host plant switches,
54 with the latter leading to inability to utilize the original host. Despite the difficulty of the
55 adaptation, host plant preference is not static with many insects undergoing host plant range
56 expansions and switches (Pashley and Martin 1987, Komazaki 1990, Emelianov et al. 1995,
57 Henniges-Janssen et al. 2011, Atijegbe et al. 2020). Yet, little is known about the genetic
58 mechanisms underlying these adaptations in insects, though genetic research has
59 implicated olfactory (sense of smell) and gustatory (sense of taste) genes in host plant
60 acceptance (Haile and Hofsvang 2002) and rejection (Yang et al. 2020c).

61

62 Exposure to novel chemical plant defence compounds can cause insect herbivores to
63 undergo allelochemical and oxidative stress (Weinhold et al. 1990), resulting plasticity of
64 detoxification enzymes (Müller et al. 2017, Schweizer et al. 2017, Orsucci et al. 2018a,
65 Breeschoten et al. 2019). Detoxification can be categorised as phase I metabolic reactions,
66 consisting of reduction or oxidation and hydrolysis reactions, or phase II metabolism which
67 affects the solubility of secondary metabolites via conjugation of hydrophilic endogenous
68 compounds (Heidel-Fischer and Vogel 2015, Kant et al. 2015). Adaptations in phase I or
69 phase II metabolism have been associated with range expansion of insect host plant usage
70 (Müller et al. 2017, Orsucci et al. 2018a). For example, expansion of Cytochrome P450 gene
71 families has been suggested to underly multiple host plant range expansions in insects
72 (Calla et al. 2017). Phase II conjugation of glucuronate and glutathione via UDP-
73 glucurotransferases and glutathione-S-transferases have also been implicated in the
74 response to allelochemical and oxidative stresses imposed by novel secondary metabolites
75 (Matzkin 2008, Hu et al. 2019, Huang et al. 2019, Cui et al. 2020).

76 *Plutella xylostella* (L.), diamondback moth, is the major pest of Brassicales crop plants that
77 has rapidly evolved resistance to all commonly used insecticide chemistries. Thought to
78 have originated in South America (You et al. 2020), diamondback moth spread through
79 human transport of Brassicales crops and now infests every continent in the world bar
80 Antarctica (Furlong et al. 2013). Diamondback moth's success as a pest is driven by its large
81 population size and short generational time. This allows for an average of five to seven
82 generations per year in temperate climates (Philips et al. 2014), yet in warmer equatorial
83 areas with continuous cropping cycles, as many as 15 generations have been observed
84 (Capinera 2001). The severity of diamondback moth infestations in *Brassica* crops costs
85 billions of dollars in crop yield reduction and control measures annually (Zalucki et al. 2012).

86 Although anecdotal evidence existed (Gupta and Thorsteinson 1960), diamondback moth
87 infesting non-Brassicales plants was not reported until 2002. Löhr and Gathu (2002)
88 described *P. xylostella* populations in Kenya that underwent a host plant range expansion to
89 infest *Pisum sativum* crops while maintaining the ability to survive on native Brassica hosts.
90 No-choice feeding assays of *P. xylostella* collected from neighbouring Brassica plants had
91 low survivorship on species of the family Fabaceae (Pulses), but the diamondback moth
92 *Pisum* strain (DBM-P) was fit and fecund with strikingly high survival rates. Selection
93 experiments on field populations of *Brassica* feeding DBM revealed that survivorship on *P.*
94 *sativum* could be increased after only six generations (Löhr and Gathu 2002).

95 DBM-P larvae preferentially feed on the *P. sativum* host plants, while retaining acceptance of
96 the original (Henniges-Janssen et al. 2014), yet oviposition experiments revealed DBM-P
97 females lay few eggs on *P. sativum* leaves and preference for Brassicales remains.
98 Henniges-Janssen et al. (2010) concluded that DBM-P is in the early stages of a host plant
99 range expansion, providing a unique opportunity to investigate genetic factors underlying
100 larval host plant acceptance. After crossing the DBM-P strain into a wild-type strain (Waite),
101 cross progeny descended from a female DBM-P individual were able to survive at a higher
102 frequency than those with a male DBM-P parent suggesting the phenotype may be a
103 maternally derived (Henniges-Janssen et al. 2011). Rearing DBM-P individuals for two
104 further generations on a *Brassica* removed the observed maternal effect. Screening
105 backcross progeny using molecular markers led Henniges-Janssen et al. (2011) to identify
106 five out of 31 chromosomes that statistically associated with the ability for progeny to survive
107 on *P. sativum*. However, unrelated pedigrees contained different combinations of
108 chromosomes, ranging from two to five, suggesting the genetic basis of host plant range
109 expansions was not fixed in this population (Henniges-Janssen et al. 2011).

110 Here we reassess the genetic basis underlying this complex host plant range expansion
111 using next generation sequencing methods. We utilize a reduced representation sequencing
112 approach to identify regions of the genome that segregate with the adaptation to *Pisum*
113 phenotype in both backcross and intercross pedigrees. Next, we wanted to investigate
114 expressional change in detoxification and chemosensation genes in response to host plant.
115 Therefore, we carried out tissue specific RNAseq on larval midguts and head capsules to
116 determine the transcriptomic effect of *Pisum* feeding. Taken together, these two approaches
117 provide complementary insights into host plant detoxification and selection in a major
118 *Brassica* crop pest.

119

120

121

122 **Methods:**

123 *Insect strains:*

124 The diamondback moth reference strain Waite, DBM-W, was originally collected from the
125 Waite campus, University of Adelaide, Australia, reared on *Brassica* species for more than
126 200 generations then transferred to the Max Planck Institute for Chemical Ecology, Germany
127 and reared on *B. napus* for 132 generations. The diamondback moth *Pisum* (DMB-P)
128 feeding strain, DBM-P was founded from a collection of the original Kenyan strain described
129 by Löhr and Gathu (2002) and rearing at the Max Plank Institute for Chemical Ecology,
130 Germany on *Pisum sativum* for more than 120 generations. Both strains were kept under the
131 same conditions in a temperature controlled room set to 21°C with a 16:8 light:dark
132 photoperiod. Detailed rearing methods have been described by Henniges-Janssen et al.
133 (2011).

134 *Insect crosses and tissue collection:*

135 The DBM-P strain was reared on *Brassica napus* for two generations before experimental
136 crosses to minimise potential maternal effects that increase survival of F₁ hybrids derived
137 from female DBM-P individuals (Henniges-Janssen et al. 2011).

138 Single pair reciprocal crosses between virgin DBM-P and DBM-W individuals were
139 performed in cylindrical containers approx. 11 cm in diameter and sealed with perforated
140 lids. *Brassica napus* leaves harvested from ~ 12-week-old plants were placed over the
141 containers to stimulate copulation and oviposition. The F₁ progeny were reared to pupation
142 then placed in 1.7 mL tubes with ventilated lids until eclosion.

143 Male informative backcrosses were performed between individual F₁ males and individual
144 DBM-P females.

145 Inter-crosses were carried out between reciprocal sibling pairs. Eggs collected from each
146 intercross were split into two groups for no-choice feeding assays using either *P. sativum* or
147 *B. napus*. Survivorship to adult was recorded to determine heritability (ie. dominance).
148 Individuals reared on *B. napus* were then crossed and their progeny (F₂) also split into two
149 groups and reared on *P. sativum* or *B. napus*.

150 All crosses were carried out 21°C with a 16:8 light:dark photoperiod. The number of embryos
151 that hatched, pupated and survival to adulthood were recorded for each cross at each
152 generation.

153 DNA extraction from cross progeny was performed using the Zymo 96 well DNA extraction
154 plate (company) and DNA from F₀ and F₁ cross parents extracted using the DNeasy Blood
155 and Tissue Kit (Qiagen). DNA quality and yield were determined using the Qubit 2.0
156 fluorimeter (Invitrogen).

157

158 *Genotyping and linkage map construction:*

159 Reduced representation short read libraries were prepared by Diversity Arrays Technology
160 Pty Ltd (<http://www.diversityarrays.com/>) and Illumina sequence data assessed for quality
161 using fastqc (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and ngsReports
162 (Ward et al. 2020). Quality trimming was carried out using trimmomatic (Bolger et al. 2014)
163 under default settings and reads mapped to the PxLV.1 genome (Chapter 6) using BWA
164 mem (David et al. 2013) and sorted using SAMtools (Li et al. 2009). Variants were called
165 using BCFtools (Li 2011). A filtered variant set was then generated by removing sites with
166 GQ<20, DP<5, and 80% of sites genotyped.

167 The filtered variant set was passed to LepMap3 (Rastas 2017) to construct linkage maps for
168 each cross type. Initial linkage maps used the informativeMask=13 argument to build the
169 map from both sites that are informative in only male and intercross informative markers. A
170 LOD limit of 10 was used along with a distortion LOD value of 1. After the initial map was
171 built, male only markers that were not in the linkage map (singles) were added using
172 JoinSingles. SeparateChromosomes2 was then run for each of the linkage groups using the
173 informativeMask=1 argument to output only male informative sites to the final map. The ends
174 of the map were then manually investigated for large cM anomalies according to discussion
175 by the author on the LepMap3 manual page ([https://sourceforge.net/p/lep-
176 map3/discussion/](https://sourceforge.net/p/lep-map3/discussion/)). These anomalies were then removed from the ends of linkage maps
177 where possible.

178 *Identification of candidate DBM-P loci:*

179 Genotypes from the linkage map were converted to rQTL format using the R programming
180 language and passed to rQTL (Broman et al. 2003) to calculate LOD scores for each
181 marker. A binary score (1 = reared on *P. sativum*, 0 = reared on *B. napus*) was used to test
182 for genotype/phenotype associations with the multiple QTL model function (mqmscan).

183 DBM-P derived allele frequencies were calculated for each marker in the F₁×F₁ and
184 backcross maps by polarizing the allele frequency using the DBM-P F₀ genotype. DBM-P
185 derived allele frequency was plotted for each chromosome using ggplot2.

186

187

188 *Transcriptome sequencing*

189 No-choice feeding assays of i) DBM-W on *B. napus* and ii) DBM-P on either *P. sativum* or *B.*
190 *napus* were carried out by transferring eggs to leaf cuttings in ventilated plastic containers at
191 21°C with a 16:8 light:dark photoperiod. Additional leaves were added as required until
192 development to the 4th larval instar. Three subsamples of approx. 50 individuals of each
193 condition (e.g. DBM-P on *P. sativum*) were removed from the same container and all
194 replicates were sourced from the same container.

195 Before tissue dissections all individuals were first placed at 4°C for 5 mins.

196 Head capsule tissue was dissected under a light microscope by placing a single larva onto a
197 petri-dish with a small amount of RNAlater solution (Sigma-Aldrich) and removing tissue
198 using a clean scalpel blade placed at the base of the head capsule. The scalpel blade and
199 petri-dish were cleaned and then treated with RNaseZap (Sigma) between individuals to
200 remove potential contamination from other tissues and RNAases introduced during cleaning.
201 After each head capsule was dissected, they were immediately added to a 1.7 mL
202 Eppendorf tube containing 1 mL of RNAlater solution up to a total of 20 head capsules per
203 tube.

204 Midgut tissues were obtained by creating incisions to remove the head and anal proleg/plate.
205 Separation of midguts from the larval body was then carried out using forceps. Individual
206 midguts were transferred to a 1.7mL Eppendorf tube containing 1 mL of RNAlater solution,
207 with 5 midguts per tube.

208 Total RNA was extracted using the Analytic Jena RNA 2-column kit (Analytic Jena) and
209 Illumina Poly-A selected libraries for three biological replicates of each tissue pool were sent
210 for Next Generation Sequencing at the Max Planck-Genome-centre, Cologne, Germany.

211 *Differential gene expression analysis:*

212 STAR (Dobin et al. 2013) was used to map trimmed reads to the PxLV.1 reference genome
213 using the v1.0 annotation (Chapter 6). A counts matrix was then constructed for each gene
214 using SubRead featureCounts (Liao et al. 2013). edgeR (Robinson et al. 2010) was used to
215 carry out differential expression analysis with counts per million (CPM) correction and
216 Fishers exact significance. Gene Ontology (GO) terms were retrieved from the PxLV.1
217 annotation (Chapter 6) and used to carry out enrichment analysis on molecular function GO

218 terms was carried out using the topGO R package (Alexa and Rahnenfuhrer 2010) using a
219 classic Fisher model.

220 Members of gene families involved with detoxification or metabolism and chemosensation
221 were compiled to determine host plant dependent differential expression between the DBM-
222 P and DBM-W strains. Candidate detoxification and chemosensory genes were identified by
223 searching gene functional annotation for GO terms along with PFAM and IPR domains
224 involved in either chemosensation or detoxification. A full list of GO terms, PFAM domains
225 and IPR domains along with their respective designation (eg gustatory receptors) can be
226 found in Supplementary Table 1 and 2.

227

228 **Results:**

229 ***Baseline survivorship of DBM strains reared on P. sativum or B. napus***

230 Virgin diamondback moths from the DBM-*Pisum* strain (P) or the DBM-Waite strain (W) were
231 paired in the following combinations: P x P (n=5), P♂ x W♀ (n=5), W♂ x P♀ (n=6), and W x
232 W (n=5). Eggs from these crosses were then evenly divided and reared on either *P. sativum*
233 or *B. napus* (Figure 1A). Progeny from the four cross categories showed no significant
234 difference in survivorship to adulthood when reared on *B. napus*, suggesting similar fitness
235 between the DBM-P and DBM-W (Figure 1B). When reared on *P. sativum* leaves, high
236 mortality was observed in all cross types except for DBM-P reciprocal crosses (Figure 1B).
237 Progeny from the WxW cross showed low levels of survivorship (0.083±0.058 survival,
238 Figure 1A) which was inconsistent with previous reports using this strain (Henniges-Janssen
239 et al. 2011), suggesting some phenotypic variation may be present in the DBM-W strain.

240 F₁ hybrids of the reciprocal cross between DBM-P and DBM-W showed significantly
241 (Students t-test p<<0.01) decreased survivorship when reared on *P. sativum* relative to *B.*
242 *napus*, confirming adaptation to *P. sativum* is recessive (Figure 1B). Greater survivorship
243 than reciprocal WxW crosses when reared on *P. sativum* was also observed (Figure 1B)
244 indicating some loci may have dominant effects, however this result was not significant.

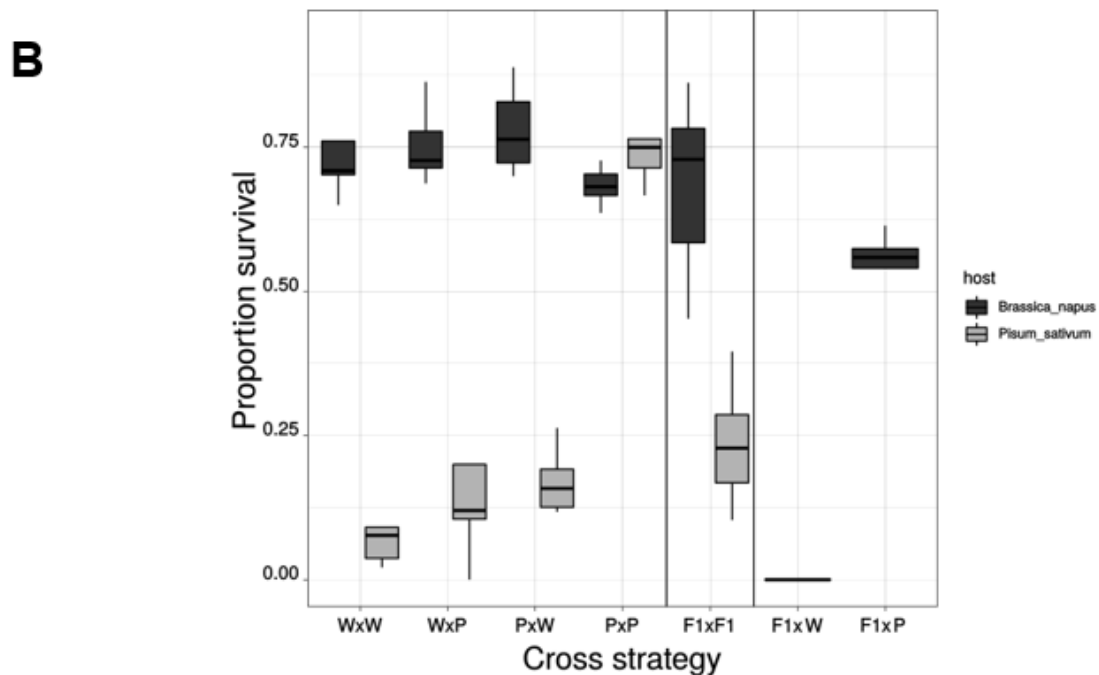
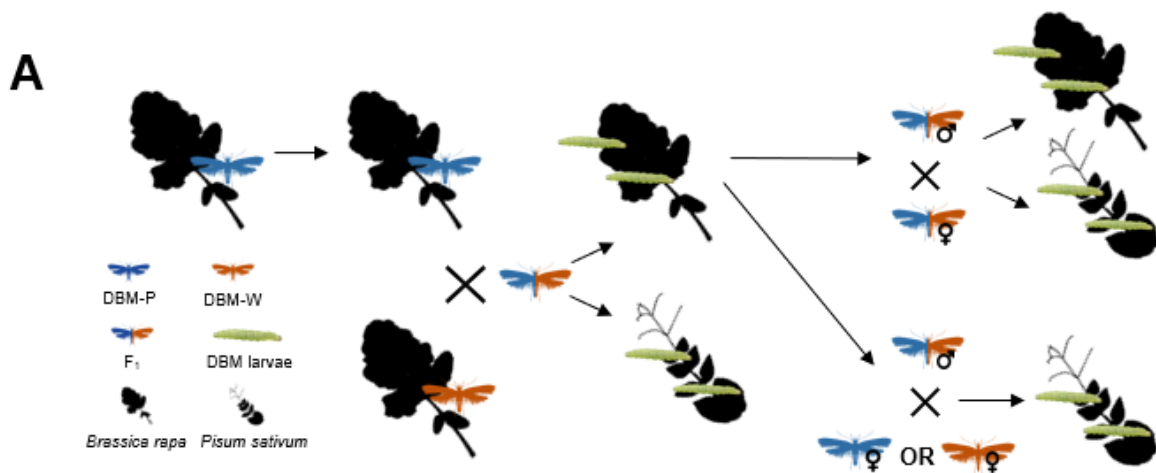
245 ***Inter- and back-crosses confirm oligogenic inheritance of the phenotype***

246 Next, single paired sibling crosses were performed using the progeny of F₀ reciprocal
247 crosses (Figure 1A). F₂ eggs were split between *P. sativum* and *B. napus* hosts (Figure 1A)
248 and feeding assays showed the F₂ individuals had low survivorship to adulthood on pea,
249 (0.31±0.079 survival, Figure 1B), compared with *B. napus* (0.84±0.12 survival, Figure 1B).
250 Interestingly this is greater than the survivorship reported by Henniges-Janssen et al. (2011)

251 (0.10-0.11) suggesting that other mutations may have arisen in the DBM-P strain since the
 252 original study (~100 generations).

253 F_1 progeny of the reciprocal crosses ($P \text{♂} \times W \text{♀}$ and $W \text{♂} \times P \text{♀}$) were further utilized in
 254 backcrosses to generate male informative lines (Figure 1A). Interestingly, a mean of
 255 0.58 ± 0.07 survivorship was observed in the progeny from backcrosses into the DBM-P
 256 strain, whereas no survivors were recorded in Waite strain backcrosses (Figure 1B). This in
 257 contrast to the survivorship observed in the $W \times W$ reciprocal crosses suggesting phenotypic
 258 plasticity may play a role in the DBM-W strains survival on *P. sativum*.

259



260

261 **Figure 1:** Survivorship of diamondback moth *Pisum* strain (DBM-P) and Waite strain (DBM-W) larvae
262 when reared on *Brassica napus* or *Pisum sativum*. A) Crossing schematic for F₁X_F₁ and backcrosses
263 carried out to generate pedigrees for survivorship and linkage analysis. F₁ progeny were then reared
264 on *B. napus* or *P. sativum*. Progeny reared on *B. napus* were used to set up F₁X_F₁ and back crosses.
265 B) Survivorship from egg to adulthood (only adult survivorship shown) for cross strategies outlined in
266 (A). F₀ reciprocal crosses between DBM-P (P) and DBM-W (W) individuals along with the F₁X_F₁ and
267 back crosses.

268

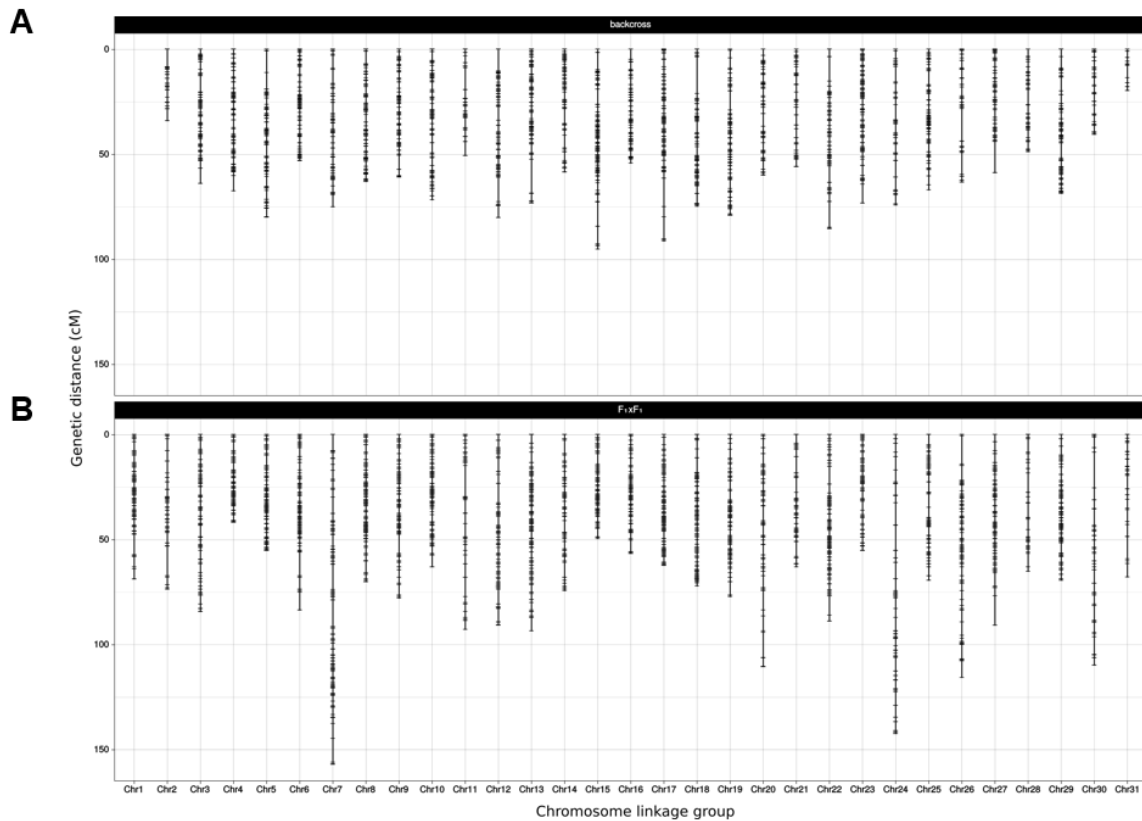
269 **Construction of two high-density linkage maps resolves 31 robust linkage groups**

270 As Henniges-Janssen et al. (2011) identified multiple allele combinations associated with
271 survival *P. sativum* in the DBM-P, we used pedigrees derived from the same F₀
272 (grandparental) cross. In brief, reduced representation genome sequencing was carried out
273 for: 1) two intercross pedigrees from a single P♂ x W♀ grandparental cross, 2) two
274 intercross pedigrees from a single W♂ x P♀ grandparental cross, and 3) three male
275 informative backcross pedigrees from a single P♂ x W♀ grandparental cross.

276 A total of 325 individuals from the seven pedigrees were sequenced (including parents and
277 grandparents), generating >23,000 markers distributed across the genome at an average of
278 12-fold depth per site against the PxLV.1 genome. Markers were used to construct two
279 linkage maps, one for the four F₁X_F₁ families and one for the three backcross families. Both
280 linkage maps were resolved into linkage groups that correspond to the 30 autosomes and Z
281 chromosome in the diamondback moth genome (Figure 2). Although the Z chromosome was
282 unresolved in the backcross map due to marker selection during linkage construction (Figure
283 2A), linkage groups in the backcross map averaged 64.8 (sd = 16.3) cM with a total length of
284 1943.9 cM. Furthermore, backcross linkage maps showed high levels of linearity between
285 genetic (cM) and physical (Mbp) against the PxLV.1 reference genome (Figure 3).

286 In contrast to the backcross map, the F₁X_F₁ linkage map was generated with molecular
287 marks of both maternal and paternal origin allowing resolution of the Z chromosome (Figure
288 2B). Linkage groups averaged 80.5 (sd =25.6) cM with a total length of 2495.9 cM. Yet, as
289 lepidopteran females do not undergo crossing over, multiple arbitrarily placed markers were
290 observed at the ends of each chromosome. Trimming non-co-linear markers (female
291 informative intercross markers) from the start and end of F₁X_F₁ linkage groups, greatly
292 increased R² values across trimmed chromosomes (Figure 4). Although differences were
293 observed between the F₁X_F₁ and backcross maps (Figure 2, Figure 3, Figure 4), linkage
294 groups from both maps were largely co-linear with PxLV.1 chromosomes.

295



297

298 **Figure 2:** Pedigree linkage maps generated with DArTseq reduced-representation genotyping. A) A
299 linkage map constructed using three related backcross families. A DBM-W male was crossed with a
300 DBM-P female, then individual F₁ brothers backcrossed to a DBM-P female. The backcross progeny
301 were reared on *P. sativum* plants, survivors genotyped, then SNP markers were mapped to PxLV.1
302 reference genome for chromosomal assignment. A consensus linkage map of all three crosses was
303 produced with a total centimorgan map length of xyz. B) Four single pair crosses were produced
304 between DBM-W and DBM-P strains, then individual F₁ males and F₁ females were paired. F₂
305 progeny were reared on both *B. napus* and *P. sativum* host plants and survivors retained. A
306 consensus linkage map was constructed using one F₂ cross from each of the four grandparental
307 lines.

308

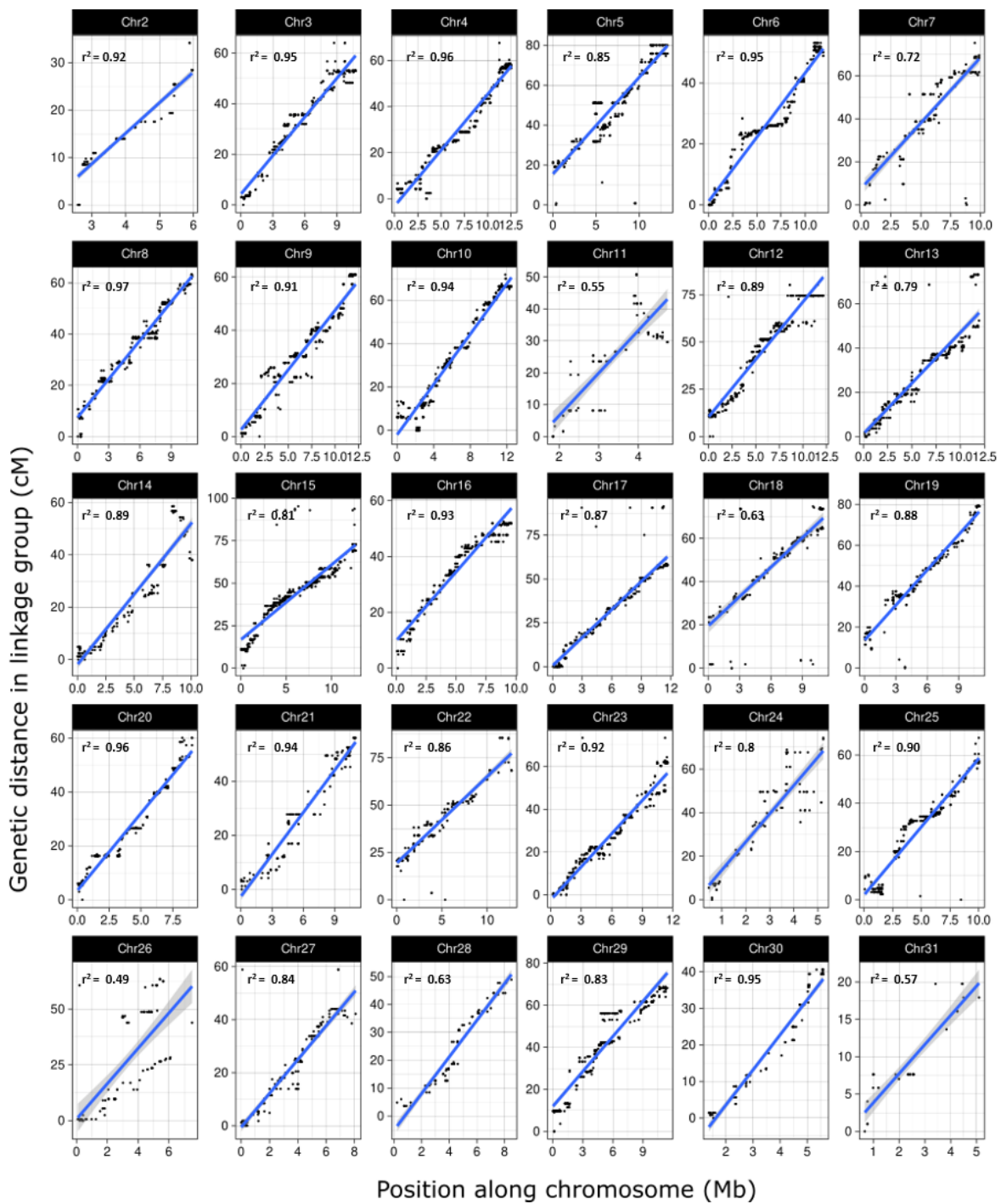
309

310

311

312

313

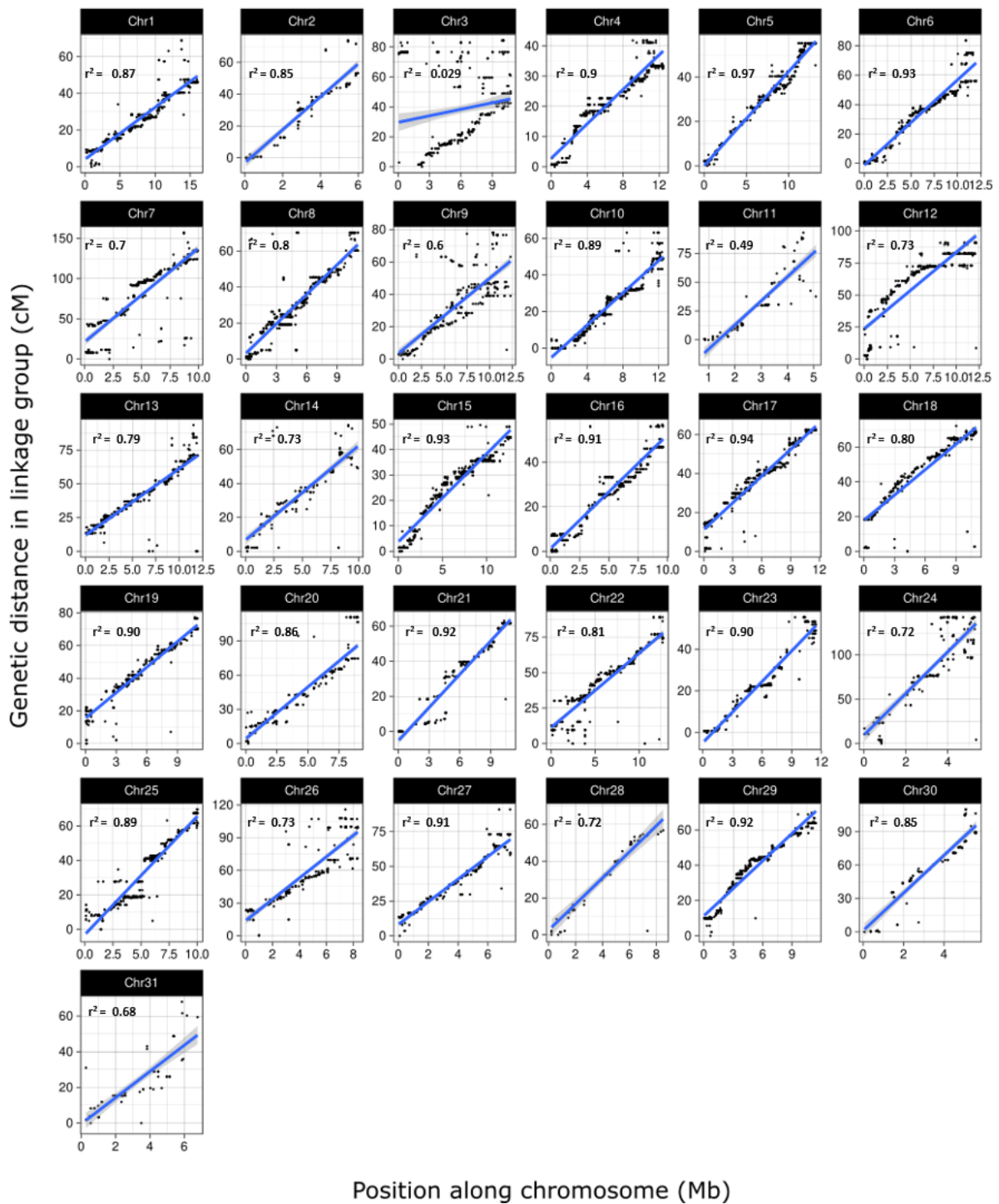


315

316 **Figure 3:** Collinearity plots comparing physical (x-axis) and recombinational (y-axis) marker positions
 317 of linkage maps produced from backcross progeny for 30 autosomes. Blue lines show a linear
 318 regression with r^2 values indicated in the top left corner. Average r^2 values was 0.84 with most having
 319 r^2 values greater than 0.7 suggesting clear co-linearity with PxLV.1 chromosomes.

320

321



322

323 **Figure 4:** Collinearity plots comparing physical (x-axis) and recombinational (y-axis) marker positions
 324 of linkage maps produced from F1x F1 progeny for 31 autosomes. Blue lines show a linear regression
 325 with r^2 values indicated in the top left corner. Average r^2 values was 0.77 with most having r^2 values
 326 greater than 0.7 suggesting clear co-linearity with PxLV.1 chromosomes..

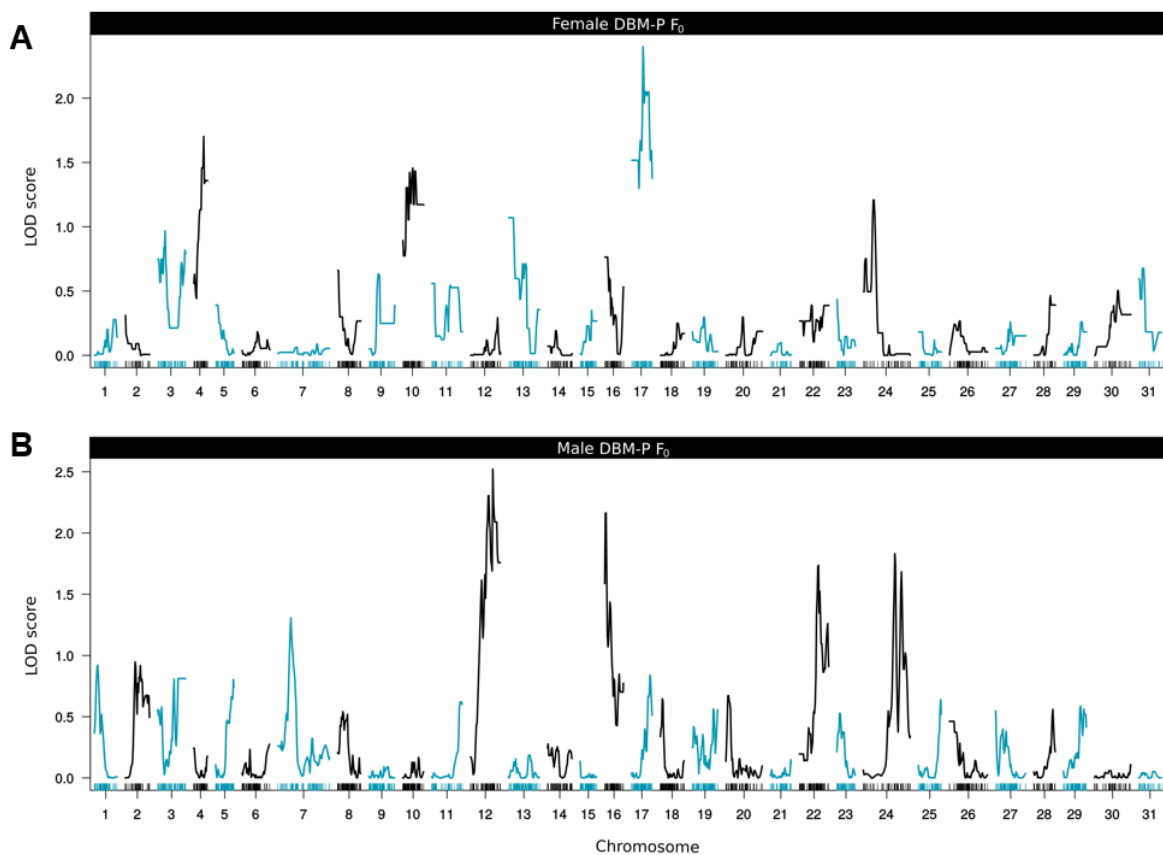
327

328

329 **Genetic linkage and overrepresentation of DBM-P derived alleles in $F_1 \times F_1$ crosses**

330 Genotype data was extracted from the four-family consensus $F_1 \times F_1$ linkage map to test for
331 genetic association with the *Pisum* survival phenotype. To increase power of the test, $F_1 \times F_1$
332 crosses derived from the same grandparental F_0 pedigree were analyzed together resulting
333 in two genome wide scans: 1) ♀ DBM-P F_0 (Figure 5A) and 2) ♂ DBM-P F_0 (Figure 5B). Little
334 overlap in QTL was observed between the two pedigree types with weak genetic linkage
335 (LOD ≥ 1.5) across chromosomes 4, 10, and 17 in crosses with a ♀ DBM-P F_0 (Figure 5A)
336 and chromosomes 12, 16, 22 and 24 in crosses with a ♂ DBM-P F_0 (Figure 5B).

337



338

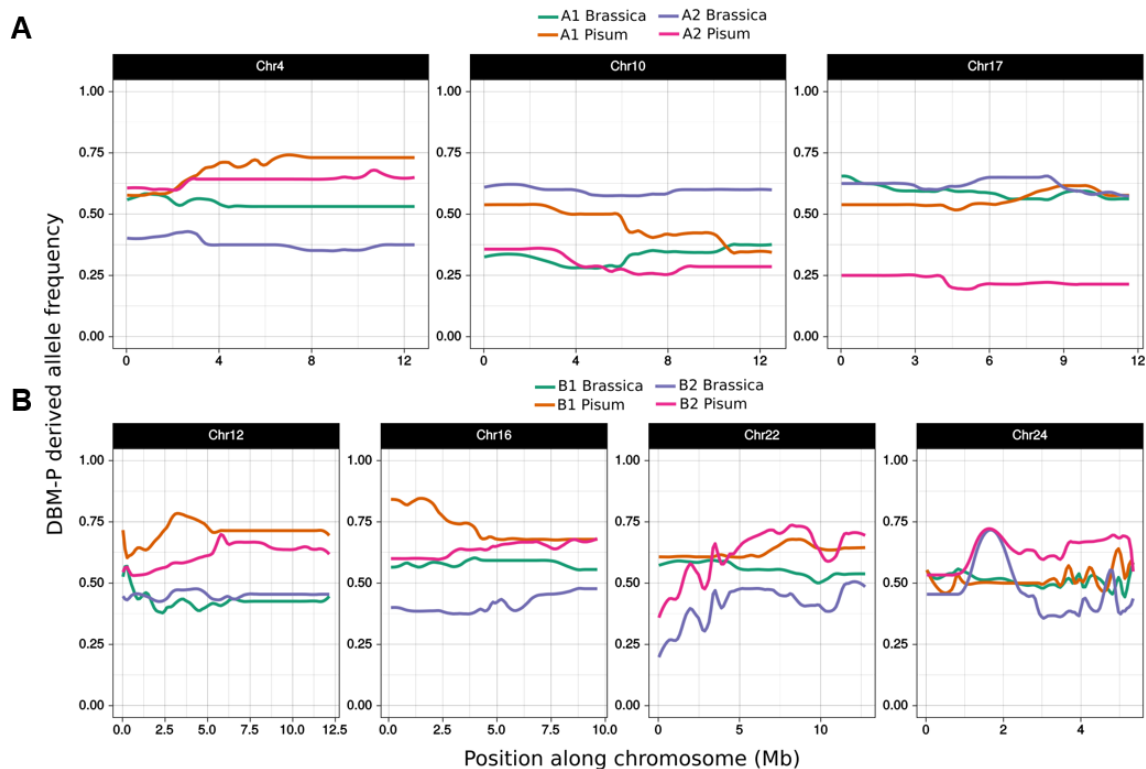
339 **Figure 5:** Genome wide QTL analysis for *P. xylostella* survival on pea, generated from $F_1 \times F_1$ reduced
340 representation genome sequencing markers. Two crosses in each direction of the F_0 reciprocal cross
341 were pooled for the analysis to increase sample number. PxLV.1 chromosomes alternate blue (odd
342 numbers) and black (even numbers). A) LOD scores calculated across the genome using the rQTL
343 additive multiple QTL model for $F_1 \times F_1$ cross progeny with the same female DBM-P individual in the F_0
344 (grandparental) generation of both crosses. B) LOD scores calculated across the genome using the
345 rQTL additive multiple QTL model for $F_1 \times F_1$ cross progeny with the same male DBM-P individual in
346 the F_0 (grandparental) generation of both crosses.

347

348 Allele frequencies among F₁XF₁ crosses were proportionately expected to be DBM-P $f \sim 0.5$
 349 and DBM-W $f \sim 0.5$ at neutral loci. Abrupt or irregular departures from these expected
 350 frequencies may occur in regions of the genome under strong selection from *P. sativum*
 351 plants. Allele frequency was then plotted across chromosomes genetically linked to the
 352 *Pisum* feeding phenotype. Peaks identified in ♀ DBM-P F₀ pedigrees (Figure 5A) showed
 353 clear differences in DBM-P derived allele frequency between individuals reared on *B. napus*
 354 or *P. sativum* (Figure 6). The QTL on chromosome 4 was well supported in each of the two
 355 crosses in the ♀ DBM-P F₀ pedigree (Figure 6A). However, LOD peaks on chromosome 10
 356 and 17 appear to be due to a reduction of DBM-P derived alleles among *P. sativum*
 357 survivors (Figure 6A).

358 In the pedigree with a ♂ DBM-P F₀, both crosses showed high levels of DBM-P derived allele
 359 frequency on chromosome 12 and 22 along with support for loci on chromosome 16 and 24
 360 in only one cross (Figure 6B). Interestingly, DBM-P derived alleles in *P. sativum* reared
 361 individuals were not fixed at any loci, suggesting a high proportion of alleles from the DBM-
 362 W strain are able to survive. Furthermore, QTL identified on chromosome 4, 22 and 24
 363 appeared to also show decreases in DBM-P allele frequency when reared on *B. napus*
 364 which may be indicative of a trade-off for survivorship on *P. sativum* at these loci (Figure 6).

365



366

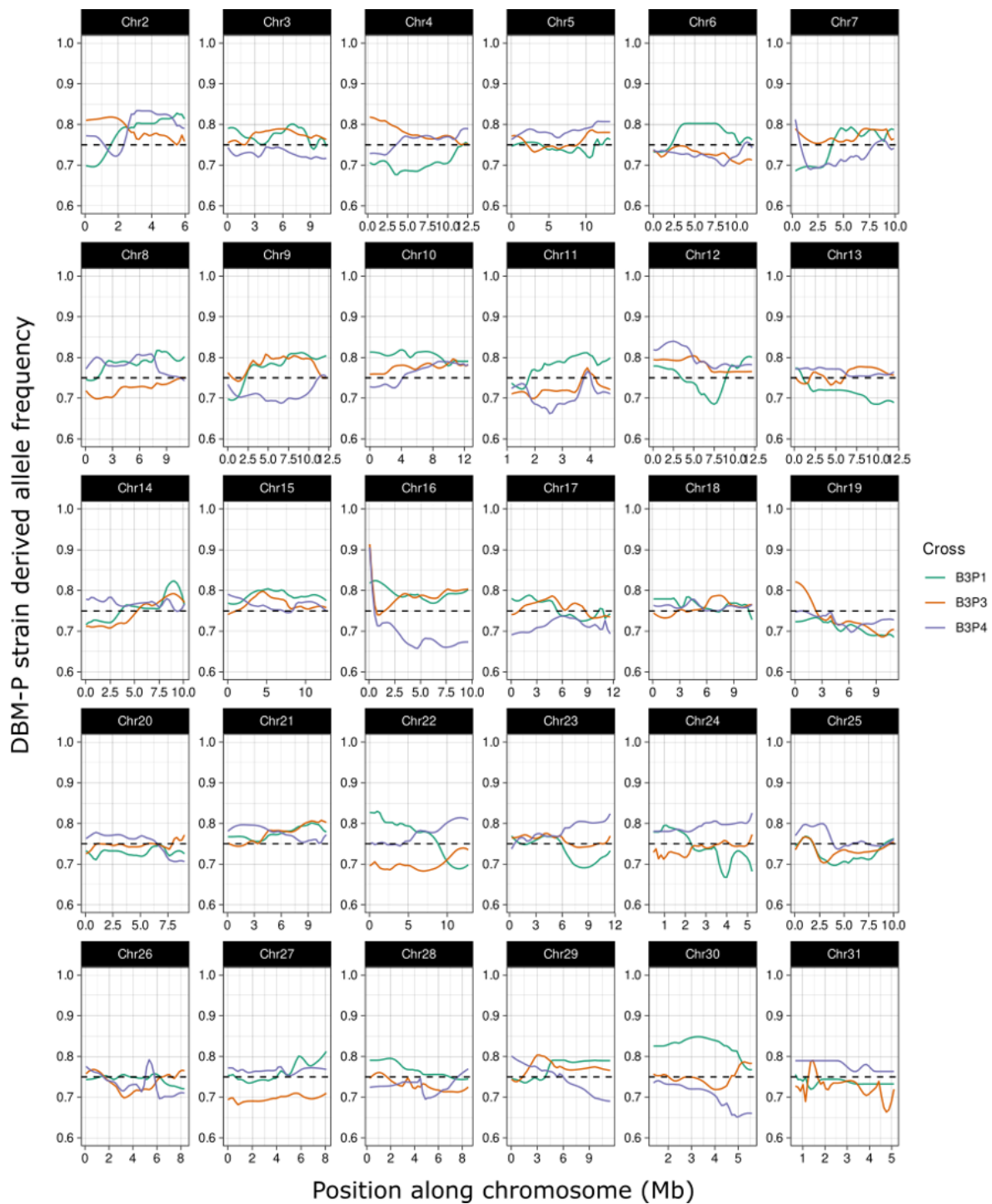
367 **Figure 6:** The frequency of DBM-P derived alleles among F₂ progeny across seven chromosomes
 368 with QTL recording LOD scores >1.5. The F₂ progeny from four separate crosses (A1, A2, B1, B2)

369 were reared either on the *Pisum* or *Brassica* host plant. The expected DBM-P derived allele frequency
370 under Hardy-Weinberg equilibrium is $f=0.5$. A) Allele frequency of two $F_1 \times F_1$ crosses with a female
371 DBM-P in the F_0 (grandparental) generation. Larvae reared on *Pisum* showed elevated DBM-P
372 derived alleles in both crosses for Chromosome 4 (pink and orange), however chromosome 10 only
373 showed increased allele frequency in cross A1 B) Allele frequency of the two $F_1 \times F_1$ crosses with a
374 male DBM-P individual in the F_0 (grandparental) generation. Chromosome 12 and chromosome 22
375 showed an increased abundance in moth crosses whereas chromosome 16 and 24 only showed an
376 increase in one cross.

377 ***Overrepresentation of DBM-P derived alleles in backcrosses***

378 Progeny from backcrosses into the DBM-P strain are expected to proportionately have 0.75
379 allele frequency derived from the DBM-P strain. Therefore, backcross progeny reared on *P.*
380 *sativum* were investigated for regions of the genome with DBM-P derived allele frequency
381 deviating from the $f=0.75$ expected under HWE. Although there were many regions of the
382 genome that deviated from 0.75, most loci showed $\sim 0.7-0.8$ DBM-P derived allele frequency
383 making it difficult to identify clear segregation patterns. Only one locus differed significantly
384 (Goodness of fit $p < 0.05$) from the expected 0.75 at the start of chromosome 16 (Figure 7).
385 The region with allelic bias spanned 0-280 kb in two out of the three sequenced backcrosses
386 providing further support for the start of chromosome 16 as a putative QTL.

387



388

389 **Figure 7:** The frequency of DBM-P derived alleles among backcross progeny for all autosomes. The
 390 backcross progeny from three separate crosses (B1P1, B1P3, B1P4) were reared on *Pisum sativum*.
 391 The expected DBM-P derived allele frequency under Hardy-Weinberg equilibrium is $f=0.75$ and is
 392 marked by a dashed line. Chromosome 16 shows a 280 Kb region with strong bias towards a DBM-P
 393 origin.

394

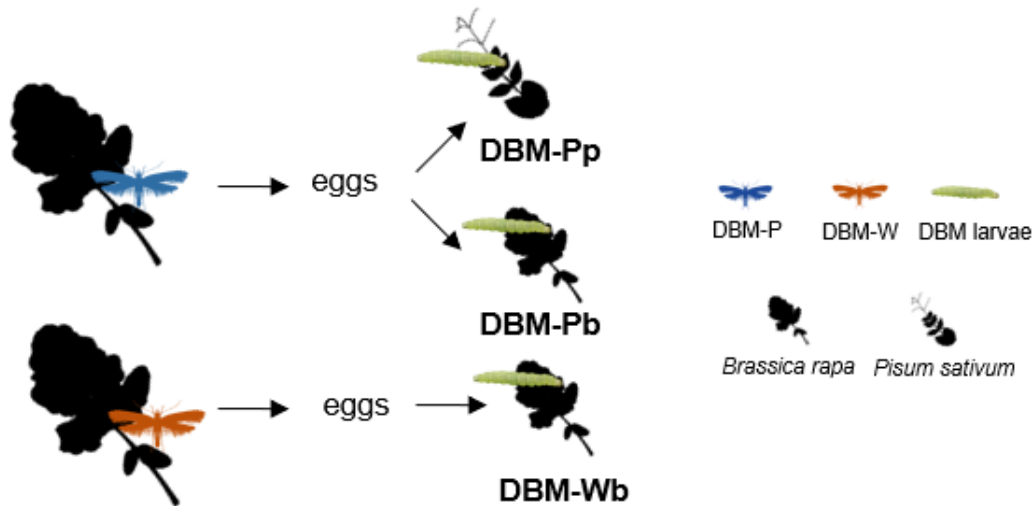
395

396 **Transcriptomic response in host plant adaptation**

397 Larval midgut and head capsule tissue was collected from the DBM-P strain after being
 398 reared on *Pisum sativum* (Pp) or on *Brassica napus* (Pb), and from the DBM-W strain after
 399 being reared on *B. napus* (Wb) (Figure 8). Larval midgut and head capsule tissue were
 400 selected due to the importance of chemosensory and behavioural genes in host plant
 401 selection (Poudel and Lee 2016, Yang et al. 2020c) and metabolism gene importance in
 402 insect detoxification (David et al. 2006, Nardini et al. 2012, Calla et al. 2017, Hu et al. 2019,
 403 Huang et al. 2019). Transcriptome sequencing was performed on three biological replicates
 404 of each condition, generating 20M (\pm 4.3) reads per library. Differentially expressed genes
 405 were sorted into three classes based on their pattern of differential expression 1) strain
 406 differences, 2) primed response to host plant and 3) dynamic response to host plant (Table
 407 1). Strain differences define genes that are differentially expressed between DBM-P and
 408 DBM-W strains, regardless of host plant. A primed response describes genes that are
 409 differentially expressed between DBM-W_bra and DBM-P_bra along with differential
 410 expression, in the same direction, between DBM-P_bra and DBM-P_pis (e.g. DBM-W_bra
 411 logCPM = 5, DBM-P_bra logCPM = 10, DBM-P_pis logCPM = 15). Finally, dynamic
 412 expression encompasses genes that show equivalent expression between DBM-W_bra and
 413 DBM-P_bra, and differential expression between DBM-P_bra and DBM-P_pis.

414 **Table 1:** Examples of differentially expressed gene patterns of interest in diamondback moth
 415 midgut and head tissue.

Class	Expression in DBM-P on <i>Pisum</i>	Expression in DBM-P on <i>Brassica</i>	Expression in DBM-W on <i>Brassica</i>
Strain	-	-	↑
Primed	↑↑	↑	-
Dynamic	↑	-	-



418
419

420 **Figure 8:** No choice feeding assays were performed using DBM-P larvae (blue) reared on either
 421 *Brassica napus* (DBM-Pb) or *Pisum sativum* (DBM-Pp) and wild type DBM-W larvae (orange, Waite
 422 strain) reared on *B. napus* (DBM-W_bra). Midgut and heads were dissected from 4th instar larvae,
 423 RNA isolated and transcriptomes sequenced to assess differential gene expression.

424 Of the ~19,000 protein coding genes in the PxLV.1 genome, differential midgut expression
 425 (FDR<0.05) occurred in 2968 genes were strain specific, 1869 had primed effect, and 1812
 426 had dynamic patterns. Gene Ontology (GO) terms significantly enriched for primed midgut
 427 responses revealed macromolecule compound biosynthetic and metabolic processes, along
 428 with gene expression and transcription (Table 2). Dynamic differentially expressed genes
 429 showed GO term enrichment for cell redox homeostasis, glutathione metabolic process as
 430 well as protein deubiquitylation and sulfur metabolism related genes, suggesting *P. sativum*
 431 secondary metabolites illicit high levels of oxidative stress (Table 2). The most significant
 432 differentially expressed genes with a dynamic response (top 1%) were further investigated
 433 for function in the response to oxidative stress, revealing differential expression of a
 434 Rhodanese-domain containing gene (FUN_006990) which respond to cyanide stress.

435 **Table 2:** Gene Ontology (GO) terms for the top five significantly enriched terms among midgut tissue,
 436 highlighting Primed, Strain, and Dynamic responses. The number of genes annotated with the GO
 437 term (Annotated), those significantly differentially expressed (Significant), and the number that would
 438 be differentially expressed by chance alone (Expected) are shown for reference, along with their p-
 439 value from a classic fisher test.

GO.ID	Term	Annotated	Significant	Expected	p-value	Class
-------	------	-----------	-------------	----------	---------	-------

GO:0006412	translation	171	65	23.01	1.50E-16	Primed
GO:0043043	peptide biosynthetic process	173	65	23.28	3.00E-16	Primed
GO:0043604	amide biosynthetic process	180	66	24.22	7.40E-16	Primed
GO:0006518	peptide metabolic process	179	65	24.09	2.20E-15	Primed
GO:0043603	cellular amide metabolic process	190	66	25.57	1.60E-14	Primed
--	--	--	--	--	--	--
GO:0005975	carbohydrate metabolic process	171	52	33.43	0.00036	Strain
GO:0009057	macromolecule catabolic process	73	26	14.27	0.00088	Strain
GO:1901575	organic substance catabolic process	135	41	26.39	0.00151	Strain
GO:0019318	hexose metabolic process	17	9	3.32	0.00214	Strain
GO:0009056	catabolic process	143	42	27.95	0.0027	Strain
--	--	--	--	--	--	--
GO:0006790	sulfur compound metabolic process	22	11	2.79	2.30E-05	Dynamic
GO:0006749	glutathione metabolic process	5	4	0.63	0.0012	Dynamic
GO:0016579	protein deubiquitination	23	9	2.92	0.0012	Dynamic
GO:0070646	protein modification by small protein re...	23	9	2.92	0.0012	Dynamic
GO:0034470	ncRNA processing	47	14	5.96	0.0015	Dynamic

440

441 Head tissue showed 3499 differentially expressed genes according to strain, highlighting the
442 impressive diversity between DBM-P and DBM-W, regardless of host plant. Enrichment of
443 genes related to GO terms *cellular response to stimulus* and *chemical synaptic transmission*
444 was observed in strain differentially expressed genes (Table 3) suggesting a difference in
445 brain function and response to external stimuli in DBM-P individuals. Significant expression

446 was relatively modest in head tissue, relative to midgut responses, with 1063 primed, and
 447 896 dynamic genes identified. Primed genes were enriched in GO terms *ATP metabolic*
 448 *process* and *oxidation-reduction process*, whereas the dynamic response to host plant was
 449 enriched in *oxidation-reduction process*, *IMP biosynthetic process* and *protein catabolic*
 450 *process* (Table 3).

451 **Table 3:** The top five most significantly enriched Gene Ontology (GO) terms for head capsule Primed,
 452 Strain, and Dynamic differentially expressed genes. The number of genes annotated with the GO
 453 term (Annotated), those significantly differentially expressed (Significant), and the number that would
 454 be differentially expressed by chance alone (Expected) are shown for reference, along with their p-
 455 value from a classic Fisher test.

GO.ID	Term	Annotated	Significant	Expected	p-value	Class
GO:0006091	generation of precursor metabolites	41	30	3.33	8.40E-25	Primed
GO:0015980	energy derivation by oxidation	22	19	1.79	1.50E-18	Primed
GO:0046034	ATP metabolic process	36	24	2.92	1.60E-18	Primed
GO:0045333	cellular respiration	21	18	1.71	1.70E-17	Primed
GO:0055114	oxidation-reduction process	419	84	34.02	1.70E-16	Primed
--	--	--	--	--	--	--
GO:0023052	signaling	397	112	84.58	0.00038	Strain
GO:0051716	cellular response to stimulus	447	124	95.23	0.00039	Strain
GO:0007268	chemical synaptic transmission	5	5	1.07	0.00044	Strain
GO:0098916	anterograde trans-synaptic signaling	5	5	1.07	0.00044	Strain
GO:0099536	synaptic signaling	5	5	1.07	0.00044	Strain
--	--	--	--	--	--	--

GO:0055114	oxidation-reduction process	419	51	28.29	1.40E-05	Dynamic
GO:1901565	organonitrogen compound catabolic process	81	17	5.47	2.00E-05	Dynamic
GO:0006188	IMP biosynthetic process	4	4	0.27	2.00E-05	Dynamic
GO:0046040	IMP metabolic process	4	4	0.27	2.00E-05	Dynamic
GO:0030163	protein catabolic process	51	13	3.44	2.10E-05	Dynamic

456

457

458

459 ***Differential expression of detoxification genes in the midgut***

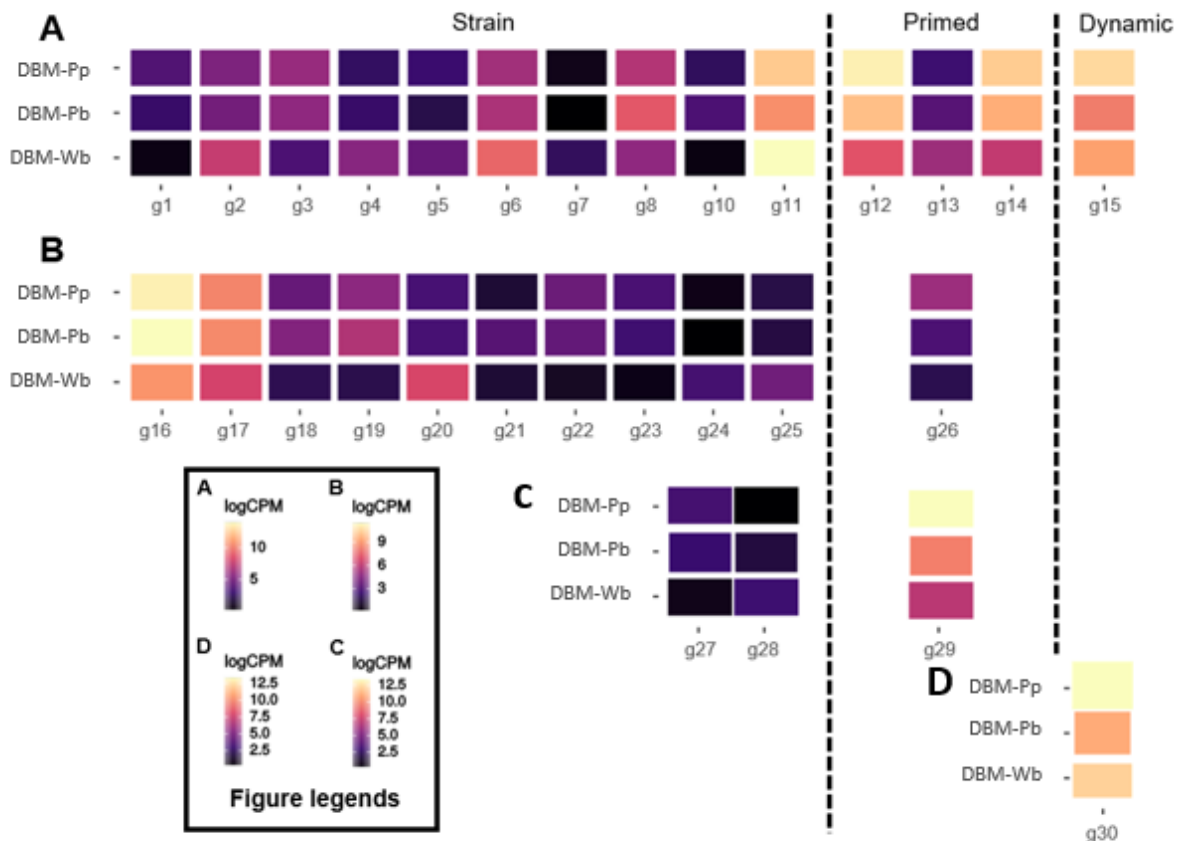
460 Cytochrome P450s, carboxylesterases, UDP-glucosyltransferases, glutathione-s-
461 transferases and Sulfurotransferases all play important roles in detoxification mechanisms of
462 plant defense compounds in the insect midgut (Li et al. 2007, Chahine and Donnell 2011,
463 Ketterman et al. 2011, Nardini et al. 2012, Halon et al. 2015). Conditional or constitutive
464 changes in expression of some members of these gene families may therefore provide a
465 strategy for adaptation to *P. sativum*. Identification of candidate genes was carried out by
466 searching the functional annotation of the PxLV.1 genome annotation for GO terms,
467 interproscan or pfam domains belonging to any of these five gene families (Supplementary
468 Table 1). This identified 245 annotations with putative detoxification function, 88 p450s, 78
469 Carboxylesterases, 27 UDP-glucosyltransferases, 32 glutathione-S-transferases, and 20
470 sulfurotransferases.

471 *Phase I metabolism in midgut tissue*

472 The majority of differentially expressed genes were categorized according to strain,
473 indicating transcript abundance in DBM-P and DBM-W larval midguts was, for the most part,
474 irrespective of diet. Strain specific expression was observed in the Phase I metabolism with
475 10 cytochrome P450s and 10 carboxylesterases differentially expressed (Figure 9A, B).

476 Cytochrome P540s with highest expression in DBM-P were identified among CYP4, CYP6,
 477 and CYP9 families, which have been shown to be differentially expressed in response to
 478 plant secondary metabolites (David et al. 2006, Huang et al. 2019, Li et al. 2019).
 479 Interestingly, differential expression was observed at the gene cluster level, with three
 480 CYP9s and two CYP4s within cytochrome P450 gene clusters on chromosome 17 and 26
 481 respectively.

482 Three Carboxylesterases and one Cytochrome P450 showed primed differential expression.
 483 Carboxylesterases FUN_003725 and FUN_020741 along with Cytochrome P450 (family
 484 CYP6) FUN_005580 showed highest expression in the DBM-P strain, suggesting these
 485 genes may be directly responding to novel secondary metabolites encountered by the DBM-
 486 P strain when reared on *P. sativum*. In contrast, carboxylesterase FUN_021236 showed
 487 lowest expression in the DBM-P strain suggesting this may detoxify secondary metabolites
 488 specific to *B. napus*. No cytochrome p450 genes exhibited dynamic response to host plant
 489 and only one carboxylesterase (FUN_005322) was dynamically differentially expressed.
 490



491
 492 **Figure 9:** Differential gene expression analysis of gene families involved with metabolism of
 493 compounds in the insect midgut. Significantly different expression patterns are categorized according

494 to strain, primed and dynamic responses for DBM-P_pis (DBM-Pp), DBM-P_bra (DBM-Pb), DBM-
495 W_bra (DBM-Pb). Colour legends are labeled according to their panel in the lower left corner. Data is
496 presented as normalized expression count data (log Counts Per Million reads, logCPM) ranging from
497 low values in black to high in yellow. A) Carboxylesterases, B) Cytochrome P450 genes, C) UDP-
498 glucosyltransferases, and D) Glutathione-S-Transferases. Sulfurotransferases did not exhibit strain,
499 primed or dynamic differential expression. The full list of gene names can be found in Supplementary
500 Table 3.

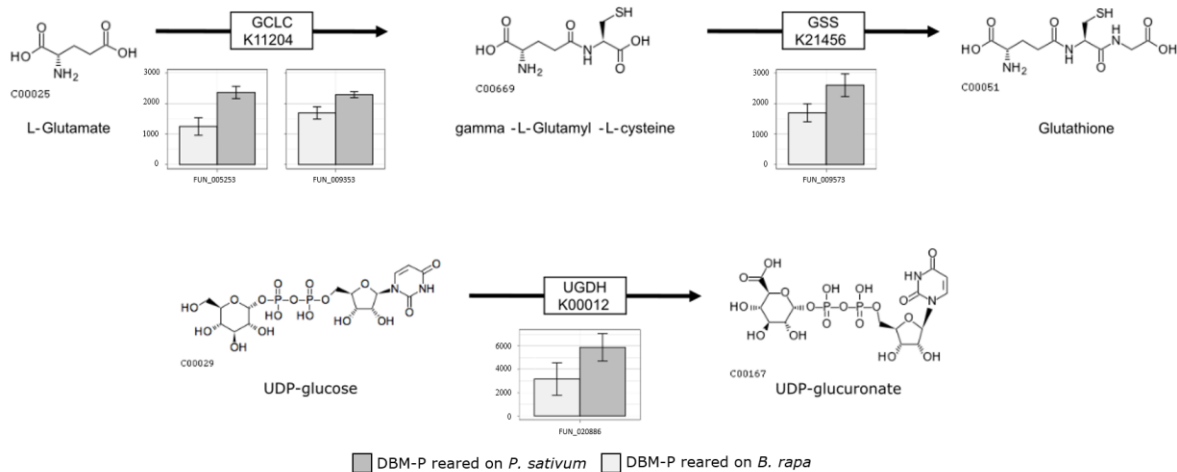
501

502 *Phase II metabolism in midgut tissue*

503 In the phase II metabolism, differential expression was observed in UDP-
504 glucosyltransferases and glutathione-S-transferases, yet no difference was observed for
505 sulfurotransferases, despite an increase in sulfur metabolism when the DBM-P strain is
506 reared on *P. sativum* (Table 1). This may suggest that sulfur conjugation is increased without
507 an increase in abundance of sulfurotransferases. Strain level differential expression was
508 observed in two UDP- glucosyltransferases belonging to family 302-E (FUN_020551) and
509 49-B (FUN_019637). Primed response was also observed in a single UDP-
510 glucosyltransferases 49-B gene (FUN_020553) suggesting the 49-B subfamily may play a
511 role in detoxification of *P. sativum* secondary metabolites in the diamondback moth. A single
512 family D1 glutathione-S-transferases gene (FUN_019678) showed dynamic expression, with
513 highest abundance in the DBM-P strain when reared on *P. sativum*. D1 glutathione-S-
514 transferases have been implicated in the adaptation of *Drosophila mojavensis* to feed on a
515 novel host plant group (Matzkin 2008).

516

517 Dynamic response to secondary metabolites in DBM-P individuals is further supported by
518 differential expression of key genes in the glutathione and glucuronate synthesis pathway.
519 Increased expression was observed in both glutamate-cysteine ligase catalytic subunits
520 (FUN_005253 and FUN_009353) along with the rate limiting step of glutathione synthesis
521 (glutathione synthase, FUN_009573) (Figure 10). The enzyme catalyzing UDP-Glucose to
522 UDP-Glucuronate conversion (UDP-glucose dehydrogenase, FUN_020886), the substrate
523 used in UDP- glucosyltransferase detoxification, was also differentially expressed (Figure
524 10). Collectively, this suggests a putative role for glutathione-S-transferases and UDP-
525 glucuronosyltransferases in a dynamic response to *P. sativum* secondary metabolites in the
526 diamondback moth.



527

528 **Figure 10:** Differential expression in the Glutathione (upper) and UDP-glucuronate (lower)
 529 synthesis pathway is observed in DBM-P when reared on a *Pisum sativum* host plant. The y
 530 axis of the barplots is measured in counts per million. Structures of compounds along with
 531 their KEGG IDs are displayed. KEGG numbers and names of enzymes that catalyze the
 532 reaction are shown within boxes for glutamate-cysteine ligase catalytic subunit (GCLC),
 533 glutathione synthase (GSS), UDP-glucose dehydrogenase (UGDH).

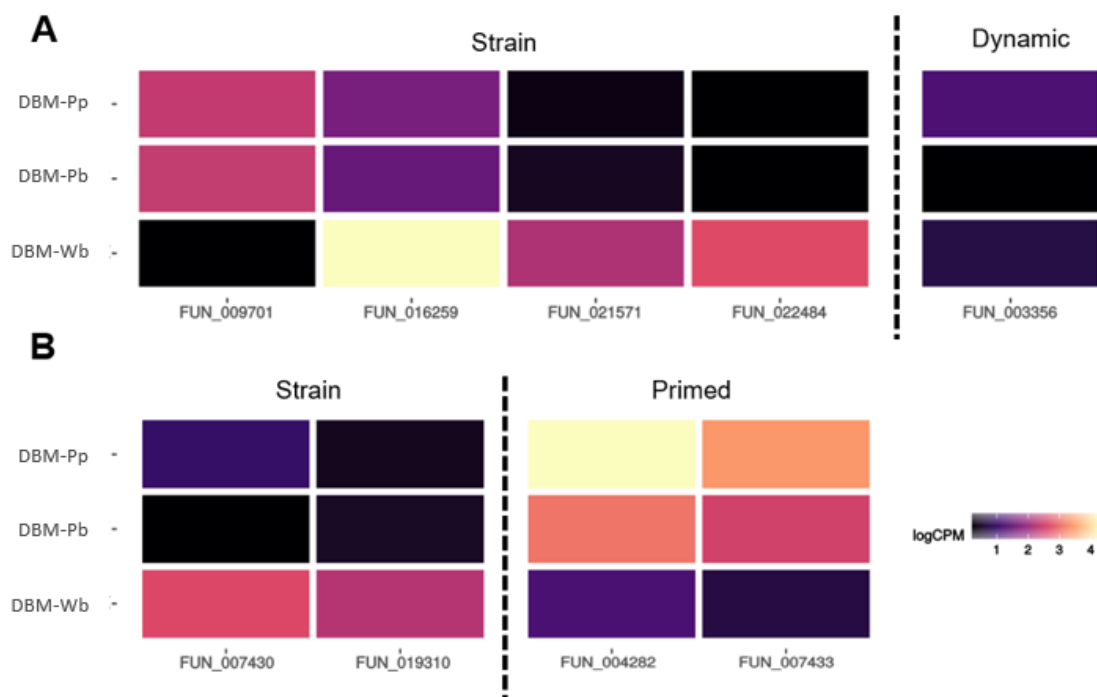
534

535 **Differential gene expression of chemosensory genes in head tissue**

536 Chemosensation plays an important role in host plant identification and to stimulate larval
 537 feeding. Head capsule tissue was investigated for differential expression in predicted
 538 gustatory receptors, odorant binding proteins and olfactory receptors. Predicted
 539 chemosensory genes were identified in a similar way to detoxification genes, however
 540 gustation and olfaction related GO terms, IPR or PFAM domains (Supplementary Table 2)
 541 were used, which identified 170 candidate annotations. Strain specific gene expression was
 542 common in both gustatory and olfactory receptors (Figure 11). Putative gustatory receptor
 543 FUN_016259, FUN_021571 and FUN_022484 all showed lower expression in DBM-P
 544 individuals regardless of host plant (Figure 11A). Gustatory receptor FUN_009701 showed
 545 almost no expression in DBM-W strain compared to the DBM-P strain (Figure 11A).
 546 Differentially expressed gustatory receptors were then BLAST against the *D. melanogaster*
 547 proteins revealing FUN_016259 had highest homology to the bitter taste receptor GR93a,
 548 FUN_021571 to GR21a which mediates acceptance or avoidance behavior, and
 549 FUN_022484 a sweet taste receptor GR64a. Primed response was identified in odorant
 550 binding proteins FUN_004282 and FUN_007433 with lower expression in the DBM-W strain
 551 (Figure 11B). This suggests FUN_004282 and FUN_007433 may be reacting preferentially
 552 to compounds present in *P. sativum* that are absent, or present at lower concentrations in *B.*

553 *napus*. Primed odorant binding proteins were most similar to *D. melanogaster* Odorant-
 554 binding protein 19d (FUN_004282) and *D. melanogaster* Odorant-binding protein 28a
 555 (FUN_007433). Furthermore, expression was higher in the DBM-W strain for odorant binding
 556 protein FUN_007430 (*D. melanogaster* Odorant-binding protein 19a) and olfactory receptor
 557 FUN_019310 (*D. melanogaster* Olfactory receptor 65b) (Figure 11B). Transcriptional
 558 differences in gustation and olfaction at both strain and primed levels suggests a complex
 559 breakdown in antifeedant behavior coupled with pro-feeding stimulus may have driven the
 560 acceptance of *Pisum sativum* as a new host plant.

561



562

563 **Figure 11:** Primed, Strain and Dynamic differential expressed chemosensory genes in head
 564 capsule tissue for DBM-P_pis (DBM-Pp), DBM-P_bra (DBM-Pb), DBM-W_bra (DBM-Pb). Data is
 565 presented as normalized expression count data (log Counts Per Million reads, logCPM)
 566 ranging from low values in black to high values in yellow. A) Gustatory receptors. B)
 567 Olfactory receptors and Odorant binding proteins.

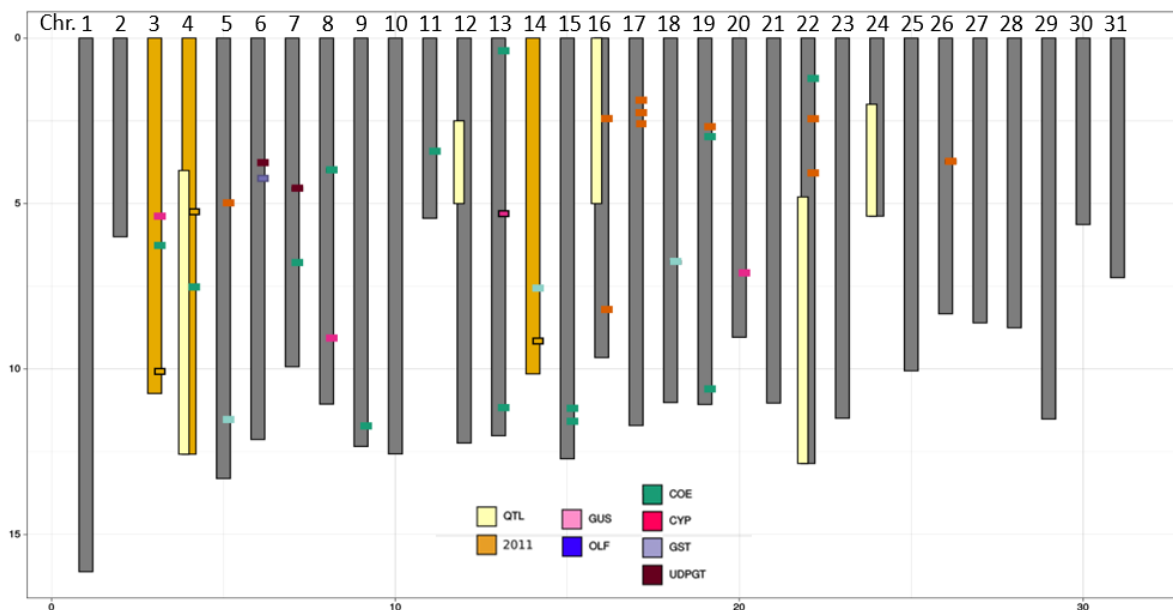
568

569 **Do QTL contain differentially expressed genes?**

570 Linkage group markers (n = 4) previously associated with DBM-P survival on *Pisum sativum*
 571 (Henniges-Janssen et al. (2011) were BLAST against the PxLV.1 genome. Three DNA
 572 markers returned single BLAST hits on Chromosome 3 (GU594729), 4 (GU594731), and 14

573 (GU594730), and a fourth (GU594732) marker produced >10 high identity BLAST hits
 574 across the genome and could not be definitively assigned. Despite this, only chromosome 4
 575 had a QTL from both this study and Henniges-Janssen et al. (2011) (Figure 12), suggesting
 576 the genetics of *Pisum* survivorship is highly complex. Furthermore, this provides added
 577 support for chromosome 4 as a major effect locus.

578 We then sought to determine if any differentially expressed genes were on the same
 579 chromosome as markers reported from Henniges-Janssen et al. (2011) (chromosomes 3, 4,
 580 14) or within QTLs identified in this study. Surprisingly, no differentially expressed gustatory
 581 receptors, olfactory receptors or odorant binding proteins were identified within QTLs (Figure
 582 12) suggesting the genetic mechanism underlying this *P. sativum* feeding trait may have a
 583 regulatory component. Two chromosomes (3 and 14) with markers from Henniges-Jansen et
 584 al (2011) contained differentially expressed chemosensory genes (Figure 12), gustatory
 585 receptor FUN_016259 and odorant binding protein FUN_004282 respectively. Detoxification
 586 genes were identified within two QTLs on different chromosomes (Figure 12), a single
 587 carboxylesterase (FUN_018127) on chromosome 4 and a cytochrome P450 on chromosome
 588 16 (FUN_006043).



589

590 **Figure 12:** Schematic representation of each chromosome with the PxLV.1 genome
 591 assembly. Allele frequency confirmed QTLs associated with the ability to survive on *Pisum*
 592 *sativum* from the multiple QTL model analysis are highlighted in yellow on the left-hand side
 593 of chromosomes 4, 12, 16, 22, and 24. The position of the best BLAST hit for each of the
 594 markers along chromosomes that were significantly associated with the ability to survive on
 595 *Pisum sativum* from Henniges-Jansen et al (2011) are highlighted in orange on the right-hand

596 side of chromosomes, along with the chromosome itself. Positional information for
597 differentially expressed chemosensory genes: Gustatory receptors (GUS, pink) and
598 Olfactory receptors/Odorant binding proteins (OLF, blue) along with differentially expressed
599 detoxification genes: Carboxylesterases (COE, green), Cytochrome P450s (CYP, red),
600 Glutathione-S-Transferases (GST, purple), and UDP- glucosyltransferases (UDPGT, brown)
601 are highlighted on the right side of chromosomes.

602

603

604 **Discussion:**

605 Holometabolous insects face a unique set of challenges when expanding their host plant
606 range (Bernays and Graham 1988, Jaenike 1990, Fry 1996). Larvae must first recognize a
607 plant as a suitable host, initiate and maintain feeding to support development and overcome
608 any novel secondary metabolites. Mated adult females must also accept the new host for
609 oviposition in order for the trait to propagate. In this study, we used reduced representation
610 genome sequencing and differential gene expression to uncover insights into host plant
611 adaptation in the diamondback moth. This revealed the genetic basis of adaptation to *P.*
612 *sativum* is highly complex, with multiple quantitative trait loci segregating among the *P.*
613 *sativum* adapted strain and diverse transcriptional responses contributed to the phenotype.
614 Differential expression was observed in multiple gustatory receptors and odorant binding
615 proteins, which may play a role in a breakdown in antifeedant behavior and host plant
616 acceptance. Furthermore, we showed that gene expression in midgut tissue, specifically
617 glutathione-s-transferases, UDP- glucosyltransferases and their respective substrates, in the
618 phase II metabolism play an important role in mediation of oxidative stress likely induced by
619 novel *P. sativum* secondary metabolites.

620 ***Phenotypic plasticity in the DBM-W strain***

621 Diamondback moth strains have been reported with low frequencies of survival on *P.*
622 *sativum* (Gupta and Thorsteinson 1960, Löhr and Gathu 2002, Yang et al. 2020b), however,
623 complete mortality is generally observed. Despite this, populations collected from *Brassica*
624 crops in Kenya, that were in close proximity to *P. sativum* feeding populations also show low
625 levels of survivorship (Löhr and Gathu 2002). In contrast to previous observations
626 (Henniges-Janssen et al. 2011), we observed low levels of survivorship among the *Brassica*
627 reared laboratory strain DBM-W when provided with *P. sativum*. This indicates diamondback
628 moth strain DBM-W may i) show some level of phenotypic plasticity for survivorship on *P.*
629 *sativum* host plants, ii) undergone a level of host acceptance through cultured exposure to *P.*

630 *sativum* volatiles (but not leaves) or iii) unexpected gene flow of DBM-P alleles into DBM-W
631 population has occurred in cultures. Future effort should determine whether independent
632 cultures of DBM-W, and indeed unrelated strains, can survive on *P. sativum* that have not
633 been reared in the same location as the DBM-P strain.

634

635 ***Patterns of recombination in linkage maps***

636 Heterogametic lepidopteran females do not undergo crossing over between non-sister
637 chromatids during meiosis, leading to progeny inheriting one complete chromosome from
638 each pair. Males however, do undergo crossing over during gamete formation which
639 provides opportunities for fine scale linkage mapping of traits (Turner and Sheppard 1975,
640 Rastas et al. 2013, Merrill et al. 2019). Therefore, traditional intercrossing experiments result
641 in both male and female linkage patterns in the F₂ generation. To remove one type of
642 segregation pattern, researchers sometimes employ either male or female informative
643 backcrosses rather than intercrosses to identify QTL (Henniges-Janssen et al. 2011, Groot
644 et al. 2013, Merrill et al. 2019).

645 Previous work identifying QTL associated with the DBM-P adaptation utilized female
646 informative backcrosses (Henniges-Janssen et al. 2011), whereby female F₁ progeny were
647 backcrossed into male DBM-P individuals. As females lack recombination, Henniges-
648 Janssen et al. (2011) was able to identify chromosome level association with the phenotype.
649 To investigate genomic loci associated with the phenotype we used both intercrosses, which
650 contain a mix of female and male informative markers, alongside male informative
651 backcrosses. Furthermore, we utilized male informative markers wherever possible to
652 construct linkage maps, maximizing the ability to detect changes in allele frequency within
653 crosses. This allowed us to resolve high density linkage groups for 30 autosomes and the Z
654 sex chromosome that were largely co-linear with the reference genome (PxLV.1, Chapter 6).
655 However, some inconsistencies remained between genetic and physical position which may
656 be due to the presence of female informative intercross markers or due to previously
657 reported structural variation between *P. xylostella* populations (You et al. 2013).

658 ***Adaptation to P. sativum is controlled by multiple combinations of unfixed loci***

659 Loci associated with host plant adaptation and specialization in insects are generally
660 complex and polygenic (Sheck and Gould 1996, Henniges-Janssen et al. 2011, Oppenheim
661 et al. 2012, Karpinski et al. 2014, Nouhaud et al. 2018) with multiple genes from both
662 chemosensation and detoxification pathways often contributing to the phenotype
663 (Breeschoten et al. 2019, Singh et al. 2020, Allio et al. 2021). Despite host plant adaptation

664 being polygenic in nature, large effect mutations have arisen in natural populations that play
665 a necessary role in their ability to survive on specialist host plants (Ratzka et al. 2002, Singh
666 et al. 2020). In this study, we identified multiple genomic regions associated with the ability
667 to survive on *P. sativum*. However, little consistency between crosses was observed, despite
668 F₂ progeny being derived from the same grandparents.

669 This was counter to our expectations that although the phenotype would remain polygenic,
670 the most fit major effect alleles would rise to fixation in the population over the 100
671 generations since Henniges-Janssen et al. (2011). Trait loci overlap was also poor between
672 Henniges-Janssen et al. (2011) and this study, with only one overlapping QTL. Although
673 many other loci were identified to segregate with the adaptation, we are unable to determine
674 if these are new mutations that have arisen in the population or simply due to sampling in the
675 original study. Presence of multiple different loci combinations contributing to complex
676 phenotypes, such as host plant specialization, are supported by theoretical models while the
677 fitness landscape has many local maxima (Kauffman and Weinberger 1989, Macken et al.
678 1991, Cooper and Podlich 2002). Especially when both large and small effect mutations
679 contribute to the phenotype, as has been observed in multiple other host plant adaptations
680 (Karpinski et al. 2014, Nouhaud et al. 2018). Furthermore, empirical examples of adaptive
681 novelty after a host plant range expansion are present in the literature (Drès and Mallet
682 2002, Bernal and Medina 2018, Singh et al. 2020). With continuous selection pressures
683 imposed by the novel host leading to an increase in novel adaptive loci contributing to their
684 survival (Bernal and Medina 2018, Singh et al. 2020). To explain the observed complexity of
685 the genetic mechanism, we propose the presence of minor and major effect loci has led to
686 multiple 'fit-enough' loci combinations arising in the population. However, further studies
687 utilizing large pedigrees of single paired crosses, preferably using multiple wild type strains,
688 will be necessary to definitively investigate this.

689 ***Trade-offs after prolonged rearing on the *P. sativum* host plant***

690 There has been a long-standing debate on the effect genetic antagonism and trade-offs
691 have during host plant specialization (Jaenike 1990, Joshi and Thompson 1995, Fry 1996,
692 Roff and Fairbairn 2007, Ali and Agrawal 2017), with few studies identifying genetic
693 mechanisms (Hawthorne and Via 2001, Via and Hawthorne 2002, Nouhaud et al. 2018). In
694 this study trait loci associated with survival on *P. sativum* were identified through an increase
695 in abundance of DBM-P derived alleles in progeny reared on *P. sativum* compared to *B.*
696 *napus*. However, loci with significantly lower DBM-P derived allele frequency on *B. napus*
697 represent regions of adaptive trade-off. Using this approach, we observed multiple loci
698 associated with survival on *P. sativum* that may also be deleterious on the native *B. napus*

699 host. Presence of genetic antagonistic loci is consistent with lowered survivorship in the
700 DBM-P strain, compared to wild type strains, when reared on *B. napus* reported elsewhere
701 (Löhr and Gathu 2002, Henniges-Janssen et al. 2011). However, we were unable to
702 statistically disentangle trade-off loci from positive effect loci. Scoring survivors allele
703 frequency on the novel host is not sufficient to identify regions negatively associated with the
704 phenotype if the organism is viable on both hosts (Joshi and Thompson 1995), as is the
705 case in the DBM-P strain (Löhr and Gathu 2002). Future studies should consider collecting
706 F₂ individuals that failed to develop on *B. napus* to confirm the antagonistic effect of the QTL
707 we identified. Though this is likely not feasible due to the size of neonate larvae and their
708 behavior to feed within leaves (Talekar and Shelton 1993).

709

710 ***Host plant specific differential expression of chemosensory proteins in head capsules***

711 Gene expression can provide the missing link between genotype and phenotype (Huestis
712 and Marshall 2009), especially in complex traits (Pavey et al. 2010) such as host plant
713 preference (Miller and Strickler 1984, Matsubayashi et al. 2011). Transcriptomic profiling of
714 head capsule tissue revealed differential expression of genes related to synaptic signaling
715 and response to stimulus between the DBM-P and DBM-W regardless of host plants
716 suggesting differences in brain function between the two strains. Differential expression of
717 genes involved in olfaction and gustation have been observed in insect host plant races
718 reared on their non-native host plant (Kopp et al. 2008, Shiao et al. 2015, Eyres et al. 2016).
719 Furthermore, feeding assays in both hemi (Eyres et al. 2016, Xiao et al. 2021) and
720 holometabolous (Shiao et al. 2015, Orsucci et al. 2018b) insects revealed differential
721 expression of chemosensory genes which may impact host plant acceptance.

722 In this study we identified both bitter and sweet taste receptors with lowered expression in
723 the DBM-P strain regardless of host plant. One such gene, an ortholog of GR93a, has been
724 shown to be involved in avoidance behavior of phenylpropanoids, a plant secondary
725 metabolite, in *Drosophila melanogaster* (Poudel and Lee 2016). This may indicate a
726 breakdown in antifeedant behavior drives the DBM-P strains acceptance of *P. sativum* as a
727 viable host plant. Interestingly, primed expression with highest expression in the DBM-P
728 strain was observed in multiple odorant binding proteins. Although odorant binding proteins
729 have been shown to be involved in host plant preference (Matsuo et al. 2007, Poudel and
730 Lee 2016, Chang et al. 2017) and the DBM-P strain shows a primed response, we were
731 unable to determine if this is involved in host plant preference.

732

733 ***Transcriptomic response to host plant in the phase I metabolism***

734 Gene differential expression has been used to identify detoxification plasticity associated
735 with host plant specialization in both hemi (Eyres et al. 2016, Singh et al. 2020) and
736 holometabolous insects (Pearce et al. 2017, Breeschoten et al. 2019). We investigated
737 expression levels of cytochrome p450 and carboxylesterase genes in the PxLV.1 genome to
738 provide insights into how the phase I metabolism responds to a novel host plant. This
739 revealed a large number of strain specific differences that were irrespective of host plant.
740 Three families of cytochrome P450s (CYP4, CYP6, CYP9) were differentially expressed
741 between the DBM-P and Waite strains. Differential expression of insect CYP4 genes have
742 been shown to occur due to xenobiotics (such as insecticides) (David et al. 2006, David et
743 al. 2013, Hu et al. 2019) and host plant usage (Huang et al. 2019). Furthermore, family
744 CYP6 and CYP9 genes have roles implicated in host plant utilization across Lepidoptera
745 (Calla et al. 2017); response to phytochemicals in insects (Li et al. 2019, Vandenhole et al.
746 2021) along with insecticide resistance (David et al. 2013, Liang et al. 2015). Strain specific
747 detoxification was also observed in type B carboxylesterases, which have been shown to be
748 differentially expressed during host plant transfers in *Sitobion avenae* (Wang et al. 2020) and
749 *Chilo suppressalis* when feeding on different host plants (Zhong et al. 2017). This provides
750 an attractive hypothesis: that differential expression of these carboxylesterases along with
751 CYP4, CYP6 and CYP9 genes is due to adaptive evolution of the DBM-P strain to detoxify
752 novel secondary metabolites present in *P. sativum*. However, we are unable to confidently
753 determine if these are adaptive changes or just due to divergence between the two strains.
754 Future experiments should investigate if other DBM strains show similar expression of these
755 genes compared to the Waite strain.

756 ***Mediation of oxidative stress by the phase II metabolism plays a dynamic role in***
757 ***adaptation to host plant***

758 The insect phase II metabolism facilitates detoxification of plant secondary metabolites
759 through conjugation of functional groups making them more easily excreted. We observed a
760 combination of strain and dynamic differential expression in glutathione-S-transferases along
761 with strain and primed expression of UDP- glucosyltransferases. This provides strong
762 evidence that plasticity in glutathione-S-transferases and UDP- glucosyltransferases
763 expression enables DBM-P larvae to survive novel secondary metabolic pressures exerted
764 by *P. sativum*. Furthermore, both of these genes have been shown to take part in plant
765 secondary metabolite detoxification (Li et al. 2007, Ketterman et al. 2011) and response to
766 oxidative stress (Yu et al. 2008, Yamamoto et al. 2009, Cui et al. 2020) in insects. With
767 members of the UDP- glucosyltransferases family have increased expression in both

768 *Spodoptera* (Breeschoten et al. 2019) and *Helicoverpa* (Pearce et al. 2017) moths while
769 feeding on different host plants.

770 Dynamic differential expression of midgut genes was significantly associated with synthesis
771 of substrates for Phase II enzymes, glutathione-s-transferases and sulfotransferases, along
772 with key genes that respond to oxidative stress such as a heat shock proteins (Kang et al.
773 2017, Farahani et al. 2020) and deubiquitinases (Cotto-Rios et al. 2012). Glutathione has
774 been shown to accumulate in insect tissue undergoing oxidative stress (Barbehenn et al.
775 2001, Barbehenn 2003). Metabolite profiling of a different strain selected for survivorship on
776 *Pisum sativum* for 17 generations showed indications of oxidative stress (Yang et al. 2020a)
777 supporting our observations and suggesting the DBM-P strain has not adapted to sufficiently
778 protect against oxidative stress induced by *P. sativum* in over 130 generations.

779

780

781

782 **References**

- 783 Ali, J. G., and A. A. Agrawal. 2017. Trade-offs and tritrophic consequences of host shifts in specialized
784 root herbivores. *Functional Ecology* **31**:153-160.
- 785 Allio, R., B. Nabholz, S. Wanke, G. Chomicki, O. A. Pérez-Escobar, A. M. Cotton, A.-L. Clamens, G. J.
786 Kergoat, F. A. Sperling, and F. L. Condamine. 2021. Genome-wide macroevolutionary
787 signatures of key innovations in butterflies colonizing new host plants. *Nature*
788 *communications* **12**:1-15.
- 789 Anholt, R. R. 2020. Chemosensation and Evolution of *Drosophila* Host Plant Selection. *Iscience*
790 **23**:100799.
- 791 Atijegbe, S. R., S. Mansfield, C. M. Ferguson, S. P. Worner, and M. Rostás. 2020. Host Range
792 Expansion of an Endemic Insect Herbivore is Associated With High Nitrogen and Low Fibre
793 Content in Exotic Pasture Plants. *Journal of Chemical ecology* **46**:544-556.
- 794 Barbehenn, R. V. 2003. Antioxidants in grasshoppers: higher levels defend the midgut tissues of a
795 polyphagous species than a graminivorous species. *Journal of Chemical ecology* **29**:683-702.
- 796 Barbehenn, R. V., S. L. Bumgarner, E. F. Roosen, and M. M. Martin. 2001. Antioxidant defenses in
797 caterpillars: role of the ascorbate-recycling system in the midgut lumen. *Journal of Insect*
798 *Physiology* **47**:349-357.
- 799 Becerra, J. X. 2007. The impact of herbivore–plant coevolution on plant community structure.
800 *Proceedings of the National Academy of Sciences* **104**:7483-7488.
- 801 Becerra, J. X. 2015. On the factors that promote the diversity of herbivorous insects and plants in
802 tropical forests. *Proceedings of the National Academy of Sciences* **112**:6098-6103.
- 803 Bernal, J. S., and R. F. Medina. 2018. Agriculture sows pests: How crop domestication, host shifts,
804 and agricultural intensification can create insect pests from herbivores. *Current opinion in*
805 *insect science* **26**:76-81.
- 806 Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
807 data. *Bioinformatics (Oxford, England)* **30**:2114-2120.

808 Breeschoten, T., V. I. D. Ros, M. E. Schranz, and S. Simon. 2019. An influential meal: host plant
809 dependent transcriptional variation in the beet armyworm, *Spodoptera exigua* (Lepidoptera:
810 Noctuidae). *BMC Genomics* **20**:845.

811 Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses.
812 *Bioinformatics* **19**:889-890.

813 Calla, B., K. Noble, R. M. Johnson, K. K. O. Walden, M. A. Schuler, H. M. Robertson, and M. R.
814 Berenbaum. 2017. Cytochrome P450 diversification and hostplant utilization patterns in
815 specialist and generalist moths: Birth, death and adaptation. *Molecular Ecology* **26**:6021-
816 6035.

817 Capinera, J. 2001. Handbook of vegetable pests. Academic press.

818 Chahine, S., and M. J. Donnell. 2011. Interactions between detoxification mechanisms and excretion
819 in Malpighian tubules of *Drosophila melanogaster*. *The Journal of*
820 *Experimental Biology* **214**:462.

821 Chang, H., D. Ai, J. Zhang, S. Dong, Y. Liu, and G. Wang. 2017. Candidate odorant binding proteins
822 and chemosensory proteins in the larval chemosensory tissues of two closely related
823 noctuidae moths, *Helicoverpa armigera* and *H. assulta*. *PLOS ONE* **12**:e0179243.

824 Cooper, M., and D. W. Podlich. 2002. The E(NK) model: Extending the NK model to incorporate gene-
825 by-environment interactions and epistasis for diploid genomes. *Complexity* **7**:31-47.

826 Cotto-Rios, X. M., M. Békés, J. Chapman, B. Ueberheide, and T. T. Huang. 2012. Deubiquitinases as a
827 signaling target of oxidative stress. *Cell reports* **2**:1475-1484.

828 Cui, X., C. Wang, X. Wang, G. Li, Z. Liu, H. Wang, X. Guo, and B. Xu. 2020. Molecular Mechanism of
829 the UDP-Glucuronosyltransferase 2B20-like Gene (AccUGT2B20-like) in Pesticide Resistance
830 of *Apis cerana cerana*. *Frontiers in Genetics* **11**.

831 David, J.-P., H. M. Ismail, A. Chandor-Proust, and M. J. I. Paine. 2013. Role of cytochrome P450s in
832 insecticide resistance: impact on the control of mosquito-borne diseases and use of
833 insecticides on Earth. *Philosophical transactions of the Royal Society of London. Series B,*
834 *Biological sciences* **368**:20120429-20120429.

835 David, J. P., S. Boyer, A. Mesneau, A. Ball, H. Ranson, and C. Dauphin-Villemant. 2006. Involvement of
836 cytochrome P450 monooxygenases in the response of mosquito larvae to dietary plant
837 xenobiotics. *Insect biochemistry and molecular biology* **36**:410-420.

838 Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R.
839 Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*
840 **29**:15-21.

841 Drès, M., and J. Mallet. 2002. Host races in plant-feeding insects and their importance in sympatric
842 speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological*
843 *Sciences* **357**:471-492.

844 Ehrlich, P. R., and P. H. Raven. 1964. Butterflies and plants: a study in coevolution. *Evolution*:586-
845 608.

846 Emelianov, I., J. Mallet, and W. Baltensweiler. 1995. Genetic differentiation in *Zeiraphera diniana*
847 (Lepidoptera: Tortricidae, the larch budmoth): polymorphism, host races or sibling species?
848 *Heredity* **75**:416-424.

849 Eyres, I., J. Jaquiéry, A. Sugio, L. Duvaux, K. Gharbi, J.-J. Zhou, F. Legeai, M. Nelson, J.-C. Simon, C. M.
850 Smadja, R. Butlin, and J. Ferrari. 2016. Differential gene expression according to race and
851 host plant in the pea aphid. *Molecular Ecology* **25**:4197-4215.

852 Farahani, S., A. R. Bandani, H. Alizadeh, S. H. Goldansaz, and S. Whyard. 2020. Differential expression
853 of heat shock proteins and antioxidant enzymes in response to temperature, starvation, and
854 parasitism in the Carob moth larvae, *Ectomyelois ceratoniae* (Lepidoptera: Pyralidae). *PLOS*
855 *ONE* **15**:e0228104.

856 Fry, J. D. 1996. The Evolution of Host Specialization: Are Trade-Offs Overrated? *The American*
857 *Naturalist* **148**:S84-S107.

858 Furlong, M. J., D. J. Wright, and L. M. Dosdall. 2013. Diamondback Moth Ecology and Management:
859 Problems, Progress, and Prospects. *Annual Review of Entomology* **58**:517-541.

860 Groot, A. T., H. Staudacher, A. Barthel, O. Inglis, G. Schöfl, R. G. Santangelo, S. Gebauer-Jung, H.
861 Vogel, J. Emerson, C. Schal, D. G. Heckel, and F. Gould. 2013. One quantitative trait locus for
862 intra- and interspecific variation in a sex pheromone. *Molecular Ecology* **22**:1065-1080.

863 Gupta, P., and A. Thorsteinson. 1960. Food plant relationships of the diamond-back moth (*Plutella*
864 *maculipennis* (curt.)): I. Gustation and Olfaction in Relation to Botanical Specificity of the
865 Larva. *Entomologia experimentalis et applicata* **3**:241-250.

866 Halon, E., G. Eakteman, P. Moshitzky, M. Elbaz, M. Alon, N. Pavlidi, J. Vontas, and S. Morin. 2015.
867 Only a minority of broad-range detoxification genes respond to a variety of phytotoxins in
868 generalist *Bemisia tabaci* species. *Scientific Reports* **5**:17975.

869 Hawthorne, D. J., and S. Via. 2001. Genetic linkage of ecological specialization and reproductive
870 isolation in pea aphids. *Nature* **412**:904-907.

871 Heidel-Fischer, H. M., and H. Vogel. 2015. Molecular mechanisms of insect adaptation to plant
872 secondary compounds. *Current opinion in insect science* **8**:8-14.

873 Henniges-Janssen, K., A. Reineke, D. G. Heckel, and A. T. Groot. 2011. Complex inheritance of larval
874 adaptation in *Plutella xylostella* to a novel host plant. *Heredity* **107**:421-432.

875 Henniges-Janssen, K., G. Schöfl, A. Reineke, D. G. Heckel, and A. T. Groot. 2010. Oviposition of
876 diamondback moth in the presence and absence of a novel host plant. *Bulletin of*
877 *Entomological Research* **101**:99-105.

878 Houseman, J. G., F. Campos, N. Thie, B. Philogene, J. Atkinson, P. Morand, and J. Arnason. 1992.
879 Effect of the maize-derived compounds DIMBOA and MBOA on growth and digestive
880 processes of European corn borer (Lepidoptera: Pyralidae). *Journal of Economic Entomology*
881 **85**:669-674.

882 Hu, B., S.-H. Zhang, M.-M. Ren, X.-R. Tian, Q. Wei, D. K. Mburu, and J.-Y. Su. 2019. The expression of
883 *Spodoptera exigua* P450 and UGT genes: tissue specificity and response to insecticides.
884 *Insect Science* **26**:199-216.

885 Huang, X., D. Liu, R. Zhang, and X. Shi. 2019. Transcriptional Responses in Defense-Related Genes of
886 *Sitobion avenae* (Hemiptera: Aphididae) Feeding on Wheat and Barley. *Journal of Economic*
887 *Entomology* **112**:382-395.

888 Huestis, D. L., and J. L. Marshall. 2009. From gene expression to phenotype in insects: non-
889 microarray approaches for transcriptome analysis. *BioScience* **59**:373-384.

890 Jaenike, J. 1990. Host Specialization in Phytophagous Insects. *Annual Review of Ecology and*
891 *Systematics* **21**:243-273.

892 Joshi, A., and J. N. Thompson. 1995. Trade-offs and the evolution of host specialization. *Evolutionary*
893 *Ecology* **9**:82-92.

894 Kang, Z.-W., F.-H. Liu, X. Liu, W.-B. Yu, X.-L. Tan, S.-Z. Zhang, H.-G. Tian, and T.-X. Liu. 2017. The
895 Potential Coordination of the Heat-Shock Proteins and Antioxidant Enzyme Genes of
896 *Aphidius gifuensis* in Response to Thermal Stress. *Frontiers in Physiology* **8**.

897 Kant, M., W. Jonckheere, B. Knecht, F. Lemos, J. Liu, B. Schimmel, C. Villarreal, L. Ataíde, W. Dermauw,
898 and J. Glas. 2015. Mechanisms and ecological consequences of plant defence induction and
899 suppression in herbivore communities. *Annals of Botany* **115**:1015-1051.

900 Karpinski, A., S. Haenniger, G. Schöfl, D. G. Heckel, and A. T. Groot. 2014. Host plant specialization in
901 the generalist moth *Heliothis virescens* and the role of egg imprinting. *Evolutionary Ecology*
902 **28**:1075-1093.

903 Kauffman, S. A., and E. D. Weinberger. 1989. The NK model of rugged fitness landscapes and its
904 application to maturation of the immune response. *Journal of Theoretical Biology* **141**:211-
905 245.

906 Ketterman, A. J., C. Saisawang, and J. Wongsantichon. 2011. Insect glutathione transferases. *Drug*
907 *metabolism reviews* **43**:253-265.

908 Komazaki, S. 1990. Variation in the Hatch Timing of the Overwintering Egg among Populations of
909 *Aphis spiraecola* PATCH (Homoptera: Aphididae) Collected from Different Host Plants and
910 Localities in Japan. *Applied Entomology and Zoology* **25**:27-34.

911 Kopp, A., O. Barmina, A. M. Hamilton, L. Higgins, L. M. McIntyre, and C. D. Jones. 2008. Evolution of
912 gene expression in the *Drosophila* olfactory system. *Molecular biology and evolution*
913 **25**:1081-1092.

914 Kortbeek, R. W. J., M. van der Gragt, and P. M. Bleeker. 2019. Endogenous plant metabolites against
915 insects. *European Journal of Plant Pathology* **154**:67-90.

916 Li, F., K. Ma, X. Chen, J.-J. Zhou, and X. Gao. 2019. The regulation of three new members of the
917 cytochrome P450 CYP6 family and their promoters in the cotton aphid *Aphis gossypii* by
918 plant allelochemicals. *Pest Management Science* **75**:152-159.

919 Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and
920 population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford,*
921 *England)* **27**:2987-2993.

922 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and
923 S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and
924 SAMtools. *Bioinformatics (Oxford, England)* **25**:2078-2079.

925 Li, X., M. A. Schuler, and M. R. Berenbaum. 2007. Molecular mechanisms of metabolic resistance to
926 synthetic and natural xenobiotics. *Annu. Rev. Entomol.* **52**:231-253.

927 Liang, X., D. Xiao, Y. He, J. Yao, G. Zhu, and K. Y. Zhu. 2015. Insecticide-mediated up-regulation of
928 cytochrome P450 genes in the red flour beetle (*Tribolium castaneum*). *International journal*
929 *of molecular sciences* **16**:2078-2098.

930 Liao, Y., G. K. Smyth, and W. Shi. 2013. The Subread aligner: fast, accurate and scalable read mapping
931 by seed-and-vote. *Nucleic acids research* **41**:e108-e108.

932 Löhr, B., and R. Gathu. 2002. Evidence of Adaptation of Diamondback Moth, *Plutella xylostella* (L.),
933 to Pea, *Pisum sativum* L. *International Journal of Tropical Insect Science* **22**:161-173.

934 Macken, C. A., P. S. Hagan, and A. S. Perelson. 1991. Evolutionary walks on rugged landscapes. *SIAM*
935 *Journal on Applied Mathematics* **51**:799-827.

936 Matsubayashi, K. W., S. Kahono, and H. Katakura. 2011. Divergent host plant specialization as the
937 critical driving force in speciation between populations of a phytophagous ladybird beetle.
938 *Journal of evolutionary biology* **24**:1421-1432.

939 Matsuo, T., S. Sugaya, J. Yasukawa, T. Aigaki, and Y. Fuyama. 2007. Odorant-Binding Proteins OBP57d
940 and OBP57e Affect Taste Perception and Host-Plant Preference in *Drosophila sechellia*. *PLOS*
941 *Biology* **5**:e118.

942 Matzkin, L. M. 2008. The molecular basis of host adaptation in cactophilic *Drosophila*: molecular
943 evolution of a glutathione S-transferase gene (*GstD1*) in *Drosophila mojavensis*. *Genetics*
944 **178**:1073-1083.

945 Mello, M. O., and M. C. Silva-Filho. 2002. Plant-insect interactions: an evolutionary arms race
946 between two distinct defense mechanisms. *Brazilian Journal of Plant Physiology* **14**:71-81.

947 Merrill, R. M., P. Rastas, S. H. Martin, M. C. Melo, S. Barker, J. Davey, W. O. McMillan, and C. D.
948 Jiggins. 2019. Genetic dissection of assortative mating behavior. *PLOS Biology* **17**:e2005902.

949 Miller, J. R., and K. L. Strickler. 1984. Finding and Accepting Host Plants. Pages 127-157 in W. J. Bell
950 and R. T. Cardé, editors. *Chemical Ecology of Insects*. Springer US, Boston, MA.

951 Müller, C., H. Vogel, and D. G. Heckel. 2017. Transcriptional responses to short-term and long-term
952 host plant experience and parasite load in an oligophagous beetle. *Molecular Ecology*
953 **26**:6370-6383.

954 Nardini, L., R. N. Christian, N. Coetzer, H. Ranson, M. Coetzee, and L. L. Koekemoer. 2012.
955 Detoxification enzymes associated with insecticide resistance in laboratory strains of
956 *Anopheles arabiensis* of different geographic origin. *Parasites & Vectors* **5**:113.

957 Nouhaud, P., M. Gautier, A. Gouin, J. Jaquiéry, J. Peccoud, F. Legeai, L. Mieuze, C. M. Smadja, C.
958 Lemaitre, R. Vitalis, and J.-C. Simon. 2018. Identifying genomic hotspots of differentiation

959 and candidate genes involved in the adaptive divergence of pea aphid host races. *Molecular*
960 *Ecology* **27**:3287-3300.

961 Oppenheim, S. J., F. Gould, and K. R. Hopper. 2012. The genetic architecture of a complex ecological
962 trait: host plant use in the specialist moth, *Heliothis subflexa*. *Evolution: International*
963 *Journal of Organic Evolution* **66**:3336-3351.

964 Orsucci, M., P. Audiot, F. Dorkeld, A. Pommier, M. Vabre, B. Gschloessl, S. Rialle, D. Severac, D.
965 Bourguet, and R. Streiff. 2018a. Larval transcriptomic response to host plants in two related
966 phytophagous lepidopteran species: implications for host specialization and species
967 divergence. *BMC Genomics* **19**:1-14.

968 Orsucci, M., P. Audiot, S. Nidelet, F. Dorkeld, A. Pommier, M. Vabre, D. Severac, M. Rohmer, B.
969 Gschloessl, and R. Streiff. 2018b. Transcriptomic response of female adult moths to host and
970 non-host plants in two closely related species. *BMC Evolutionary Biology* **18**:145.

971 Pashley, D. P., and J. A. Martin. 1987. Reproductive incompatibility between host strains of the fall
972 armyworm (Lepidoptera: Noctuidae). *Annals of the Entomological Society of America*
973 **80**:731-733.

974 Pavey, S. A., H. Collin, P. Nosil, and S. M. Rogers. 2010. The role of gene expression in ecological
975 speciation. *Annals of the New York Academy of Sciences* **1206**:110.

976 Pearce, S. L., D. F. Clarke, P. D. East, S. Elfekih, K. Gordon, L. S. Jermiin, A. McGaughran, J. G.
977 Oakeshott, A. Papanikolaou, and O. P. Perera. 2017. Genomic innovations, transcriptional
978 plasticity and gene loss underlying the evolution and divergence of two highly polyphagous
979 and invasive *Helicoverpa* pest species. *BMC biology* **15**:1-30.

980 Philips, C. R., Z. Fu, T. P. Kuhar, A. M. Shelton, and R. J. Cordero. 2014. Natural History, Ecology, and
981 Management of Diamondback Moth (Lepidoptera: Plutellidae), With Emphasis on the United
982 States. *Journal of Integrated Pest Management* **5**:D1-D11.

983 Poudel, S., and Y. Lee. 2016. Gustatory Receptors Required for Avoiding the Toxic Compound
984 Coumarin in *Drosophila melanogaster*. *Molecules and cells* **39**:310-315.

985 Rastas, P. 2017. Lep-MAP3: robust linkage mapping even for low-coverage whole genome
986 sequencing data. *Bioinformatics* **33**:3726-3732.

987 Rastas, P., L. Paulin, I. Hanski, R. Lehtonen, and P. Auvinen. 2013. Lep-MAP: fast and accurate linkage
988 map construction for large SNP datasets. *Bioinformatics* **29**:3128-3134.

989 Ratzka, A., H. Vogel, D. J. Kliebenstein, T. Mitchell-Olds, and J. Kroymann. 2002. Disarming the
990 mustard oil bomb. *Proceedings of the National Academy of Sciences* **99**:11223.

991 Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for
992 differential expression analysis of digital gene expression data. *Bioinformatics (Oxford,*
993 *England)* **26**:139-140.

994 Roff, D. A., and D. Fairbairn. 2007. The evolution of trade-offs: where are we? *Journal of*
995 *evolutionary biology* **20**:433-447.

996 Schweizer, F., H. Heidel-Fischer, H. Vogel, and P. Reymond. 2017. *Arabidopsis* glucosinolates trigger a
997 contrasting transcriptomic response in a generalist and a specialist herbivore. *Insect*
998 *biochemistry and molecular biology* **85**:21-31.

999 Sheck, A., and F. Gould. 1996. The genetic basis of differences in growth and behavior of specialist
1000 and generalist herbivore species: selection on hybrids of *Heliothis virescens* and *Heliothis*
1001 *subflexa* (Lepidoptera). *Evolution* **50**:831-841.

1002 Shiao, M.-S., J.-M. Chang, W.-L. Fan, M.-Y. J. Lu, C. Notredame, S. Fang, R. Kondo, and W.-H. Li. 2015.
1003 Expression divergence of chemosensory genes between *Drosophila sechellia* and its sibling
1004 species and its implications for host shift. *Genome biology and evolution* **7**:2843-2858.

1005 Singh, K. S., B. J. Troczka, A. Duarte, V. Balabanidou, N. Trissi, L. Z. C. Paladino, P. Nguyen, C. T.
1006 Zimmer, K. M. Papapostolou, and E. Randall. 2020. The genetic architecture of a host shift:
1007 An adaptive walk protected an aphid and its endosymbiont from plant chemical defenses.
1008 *Science Advances* **6**:eaba1070.

1009 Talekar, N. S., and A. M. Shelton. 1993. Biology, Ecology, and Management of the Diamondback
1010 Moth. *Annual Review of Entomology* **38**:275-301.

1011 Turner, J. R. G., and P. M. Sheppard. 1975. Absence of crossing-over in female butterflies
1012 (*Heliconius*). *Heredity* **34**:265-269.

1013 Vandenhole, M., W. Dermauw, and T. Van Leeuwen. 2021. Short term transcriptional responses of
1014 P450s to phytochemicals in insects and mites. *Current opinion in insect science* **43**:117-127.

1015 Via, S., and D. J. Hawthorne. 2002. The Genetic Architecture of Ecological Specialization: Correlated
1016 Gene Effects on Host Use and Habitat Choice in Pea Aphids. *The American Naturalist*
1017 **159**:S76-S88.

1018 Wang, D., D. Liu, X. Shi, Y. Yang, N. Zhang, and Z. Shang. 2020. Transcriptome profiling revealed
1019 potentially important roles of defensive gene expression in the divergence of insect
1020 biotypes: a case study with the cereal aphid *Sitobion avenae*. *BMC Genomics* **21**:546.

1021 Ward, C. M., T.-H. To, and S. M. Pederson. 2020. ngsReports: a Bioconductor package for managing
1022 FastQC reports and other NGS related log files. *Bioinformatics* **36**:2587-2588.

1023 Weinhold, L. C., S. Ahmad, and R. S. Pardini. 1990. Insect glutathione-S-transferase: A predictor of
1024 allelochemical and oxidative stress. *Comparative Biochemistry and Physiology Part B:
1025 Comparative Biochemistry* **95**:355-363.

1026 Wink, M., and O. Schimmer. 2018. Molecular modes of action of defensive secondary metabolites.
1027 *Annual Plant Reviews online*:21-161.

1028 Wittstock, U., N. Agerbirk, E. J. Stauber, C. E. Olsen, M. Hippler, T. Mitchell-Olds, J. Gershenzon, and
1029 H. Vogel. 2004. Successful herbivore attack due to metabolic diversion of a plant chemical
1030 defense. *Proceedings of the National Academy of Sciences* **101**:4859-4864.

1031 Xiao, Y., L. Sun, Q. Wang, X.-K. An, X.-Z. Huang, A. Khashaveh, Z.-Y. Li, and Y.-J. Zhang. 2021. Host
1032 plants transfer induced regulation of the chemosensory genes repertoire in the alfalfa plant
1033 bug *Adelphocoris lineolatus* (Goeze). *Comparative Biochemistry and Physiology Part D:
1034 Genomics and Proteomics* **38**:100798.

1035 Yamamoto, K., S. Nagaoka, Y. Banno, and Y. Aso. 2009. Biochemical properties of an omega-class
1036 glutathione S-transferase of the silkworm, *Bombyx mori*. *Comparative Biochemistry and
1037 Physiology Part C: Toxicology & Pharmacology* **149**:461-467.

1038 Yang, F.-Y., H. S. Saqib, J.-H. Chen, Q.-Q. Ruan, L. Vasseur, W.-Y. He, and M.-S. You. 2020a.
1039 Differential Profiles of Gut Microbiota and Metabolites Associated with Host Shift of *Plutella*
1040 *xylostella*. *International journal of molecular sciences* **21**.

1041 Yang, F.-Y., H. S. A. Saqib, J.-H. Chen, Q.-Q. Ruan, L. Vasseur, W.-Y. He, and M.-S. You. 2020b.
1042 Differential Profiles of Gut Microbiota and Metabolites Associated with Host Shift of *Plutella*
1043 *xylostella*. *International journal of molecular sciences* **21**:6283.

1044 Yang, K., X.-L. Gong, G.-C. Li, L.-Q. Huang, C. Ning, and C.-Z. Wang. 2020c. A gustatory receptor tuned
1045 to the steroid plant hormone brassinolide in *Plutella xylostella* (Lepidoptera: Plutellidae).
1046 *eLife* **9**:e64114.

1047 Yang, R. S. H., and F. E. Guthrie. 1969. Physiological Responses of Insects to Nicotine¹. *Annals of the
1048 Entomological Society of America* **62**:141-146.

1049 You, M., F. Ke, S. You, Z. Wu, Q. Liu, W. He, S. W. Baxter, Z. Yuchi, L. Vasseur, G. M. Gurr, C. M. Ward,
1050 H. Cerda, G. Yang, L. Peng, Y. Jin, M. Xie, L. Cai, C. J. Douglas, M. B. Isman, M. S. Goettel, Q.
1051 Song, Q. Fan, G. Wang-Pruski, D. C. Lees, Z. Yue, J. Bai, T. Liu, L. Lin, Y. Zheng, Z. Zeng, S. Lin,
1052 Y. Wang, Q. Zhao, X. Xia, W. Chen, L. Chen, M. Zou, J. Liao, Q. Gao, X. Fang, Y. Yin, H. Yang, J.
1053 Wang, L. Han, Y. Lin, Y. Lu, and M. Zhuang. 2020. Variation among 532 genomes unveils the
1054 origin and evolutionary history of a global insect herbivore. *Nature communications*
1055 **11**:2321.

1056 You, M., Z. Yue, W. He, X. Yang, G. Yang, M. Xie, D. Zhan, S. W. Baxter, L. Vasseur, G. M. Gurr, C. J.
1057 Douglas, J. Bai, P. Wang, K. Cui, S. Huang, X. Li, Q. Zhou, Z. Wu, Q. Chen, C. Liu, B. Wang, X. Li,
1058 X. Xu, C. Lu, M. Hu, J. W. Davey, S. M. Smith, M. Chen, X. Xia, W. Tang, F. Ke, D. Zheng, Y. Hu,
1059 F. Song, Y. You, X. Ma, L. Peng, Y. Zheng, Y. Liang, Y. Chen, L. Yu, Y. Zhang, Y. Liu, G. Li, L.

1060 Fang, J. Li, X. Zhou, Y. Luo, C. Gou, J. Wang, J. Wang, H. Yang, and J. Wang. 2013. A
1061 heterozygous moth genome provides insights into herbivory and detoxification. *Nature*
1062 *Genetics* **45**:220-225.

1063 Yu, Q., C. Lu, B. Li, S. Fang, W. Zuo, F. Dai, Z. Zhang, and Z. Xiang. 2008. Identification, genomic
1064 organization and expression pattern of glutathione S-transferase in the silkworm, *Bombyx*
1065 *mori*. *Insect biochemistry and molecular biology* **38**:1158-1164.

1066 Zalucki, M. P., A. Shabbir, R. Silva, D. Adamson, L. Shu-Sheng, and M. J. Furlong. 2012. Estimating the
1067 Economic Cost of One of the World's Major Insect Pests, *Plutella xylostella* (Lepidoptera:
1068 Plutellidae): Just How Long Is a Piece of String? *Journal of Economic Entomology* **105**:1115-
1069 1129.

1070 Zhong, H., F. Li, J. Chen, J. Zhang, and F. Li. 2017. Comparative transcriptome analysis reveals host-
1071 associated differentiation in *Chilo suppressalis* (Lepidoptera: Crambidae). *Scientific Reports*
1072 **7**:13778.

1073

Chapter 8

White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter.

Ward, C. M., Aumann, R. A., Whitehead, M. A., Nikolouli, K., Leveque, G., Gouvi, G., Fung, E., Reiling, S. J., Djambazian, H., Hughes, M. A., Whiteford, S., Caceres-Barrios, C., Nguyen, T. N. M., Choo, A., Crisp, P., Sim, S. B., Geib, S. M., Marec, F., Häcker, I., Ragoussis, J., Darby, A. C., Bourtzis, K., Baxter, S. W. & Schetelig, M. F. (2021). **Nature Communications**, 12(1), 491.

Statement of Authorship

Title of Paper	White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	<p>Christopher M. Ward^{1,12}, Roswitha A. Aumann^{2,12}, Mark A. Whitehead³, Katerina Nikolouli⁴, Gary Leveque^{5,6}, Georgia Gouvi^{4,7}, Elisabeth Fung¹, Sarah J. Reiling⁵, Haig Djambazian⁵, Margaret A. Hughes³, Sam Whiteford³, Carlos Caceres-Barrios⁴, Thu N. M. Nguyen^{1,8}, Amanda Choo¹, Peter Crisp^{1,9}, Sheina B. Sim¹⁰, Scott M. Geib¹⁰, František Marec¹¹, Irina Häcker², Jiannis Ragoussis⁵, Alistair C. Darby³, Kostas Bourtzis^{4*}, Simon W. Baxter^{8*} & Marc F. Schetelig^{2*}. White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter. <i>Nature communications</i>, 12(1), 1-12</p> <p>¹ School of Biological Sciences, University of Adelaide, 5005 Adelaide, Australia. ²Department of Insect Biotechnology in Plant Protection, Justus-LiebigUniversity Gießen, Institute for Insect Biotechnology, Winchesterstr. 2, 35394 Gießen, Germany. ³ Centre for Genomic Research, Institute of Integrative Biology, The Biosciences Building, Crown Street, L69 7ZB Liverpool, United Kingdom. ⁴ Insect Pest Control Laboratory, Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture, Seibersdorf, 1400 Vienna, Austria. ⁵ McGill University Genome Centre, McGill University, Montreal, QC, Canada. ⁶ Canadian Centre for Computational Genomics (C3G), McGill University, Montreal, QC, Canada. ⁷ Department of Environmental Engineering, University of Patras, 2 Seferi str., 30100 Agrinio, Greece. ⁸ Bio21 Molecular Science and Biotechnology Institute, School of BioSciences, University of Melbourne, Melbourne 3010, Australia. ⁹ South Australian Research and Development Institute, Waite Road, Urrbrae 5064, South Australia. ¹⁰ USDA-ARS Daniel K. Inouye US Pacific Basin Agricultural Research Center, 64 Nowelo Street, Hilo, HI 96720, USA. ¹¹ Biology Centre, Czech Academy of Sciences, Institute of Entomology, Branišovská 31, 370 05 České Budějovice, Czech Republic ¹² These authors contributed equally * Corresponding authors</p>

Principal Author

Name of Principal Author (Candidate)	Christopher Ward
Contribution to the Paper	<p>Co-first Author with R.A.A.</p> <p>Designed research. Performed research. Contributed new reagents/ analytic tools. Analysed the data. Wrote the paper.</p> <p>Specifically,</p> <ol style="list-style-type: none"> carried out methods for bioinformatic analysis in the results section "Resolving the <i>B. dorsalis</i> wp locus by introgression experiments". This included producing Figure 1. Performed analysis and wrote the text relevant to <i>B. dorsalis/B.tryoni</i> species in the section "Genome and transcriptome sequencing reveal a single candidate wp gene".

	<p>3. Created data and images for supplementary files Figure S1, Figure S2, Figure S5.</p> <p>4. Wrote or co-wrote Materials and Methods sections "Introgression and identification of <i>wp</i> in <i>B. dorsalis</i>", "Molecular analyses of <i>wp</i> mutants and mosaics", "Gene editing and generation of homozygous <i>wp</i> - strains".</p> <p>5. Provided text and editing for the introduction and discussion sections.</p> <p>A summary of author contributions as they appear in the manuscript are provided below.</p>
Overall percentage (%)	N/A
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper with R.A.A.
Signature	
Date	18/5/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- the candidate's stated contribution to the publication is accurate (as detailed above);
- permission is granted for the candidate to include the publication in the thesis; and
- the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Principal Author (Co-First)	Roswitha A. Aumann
Contribution to the Paper	Co-first Author with C.M.W. Designed research. Performed research. Contributed new reagents/ analytic tools. Analysed the data. Wrote the paper. A summary of author contributions as they appear in the manuscript are provided below.
Signature	
Date	17.05.2021

Name of Co-Author	Kostas Bourtzis
Contribution to the Paper	Designed research. Performed research. Contributed new reagents/ analytic tools. Analysed the data. Wrote the paper. A summary of author contributions as they appear in the manuscript are provided below.
Signature	
Date	17.05.2021

Name of Co-Author	Simon W. Baxter
Contribution to the Paper	Designed research. Performed research. Contributed new reagents/ analytic tools. Analysed the data. Wrote the paper. As a corresponding author S.W.B. is also signing for the contribution of: K.N., G.L., G.G., S.J.R., H.G., C.C.-B., S.B.S., S.M.G. J.R., E.F., T.N.M.N., A.C., P.C., M.A.W., M.A.H., S.W., F.M., I.H., A.C.D. A summary of author contributions as they appear in the manuscript are provided below.



Signature

Date 20/5/2021

Name of Co-Author	Marc F. Schetelig
Contribution to the Paper	Designed research. Performed research. Contributed new reagents/ analytic tools. Analysed the data. Wrote the paper. A summary of author contributions as they appear in the manuscript are provided below.
Signature	→
Date	11. May 2021

Contribution statement: R.A.A., C.M.W., C.C., P.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. designed the research; C.M.W., R.A.A., M.A.W., K.N., G.G., E.F., S.J.R., M.A.H., C.C., T.N.M.N., A.C., S.B.S., S.M.G., A.C.D., K.B., S.W.B., and M.F.S. performed the research; R.A.A., C.M.W., H.D., G.L., F.M., J.R., K.B., S.W.B., and M.F.S. contributed new reagents/ analytic tools; C.M.W., R.A.A., M.A.W., K.N., G.L., G.G., H.D., S.W., T.N.M.N., A.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. analyzed the data; R.A.A., C.M.W., K.N., G.L., G.G., S.J.R., S.W., A.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. wrote the paper

White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter

Christopher M. Ward ^{1,12}, Roswitha A. Aumann^{2,12}, Mark A. Whitehead³, Katerina Nikolouli ⁴, Gary Leveque^{5,6}, Georgia Gouvi ^{4,7}, Elisabeth Fung¹, Sarah J. Reiling⁵, Haig Djambazian⁵, Margaret A. Hughes³, Sam Whiteford³, Carlos Caceres-Barrios⁴, Thu N. M. Nguyen ^{1,8}, Amanda Choo ¹, Peter Crisp^{1,9}, Sheina B. Sim ¹⁰, Scott M. Geib¹⁰, František Marec ¹¹, Irina Häcker ², Jiannis Ragoussis ⁵, Alistair C. Darby³, Kostas Bourtzis^{4✉}, Simon W. Baxter ^{8✉} & Marc F. Schetelig ^{2✉}

Mass releases of sterilized male insects, in the frame of sterile insect technique programs, have helped suppress insect pest populations since the 1950s. In the major horticultural pests *Bactrocera dorsalis*, *Ceratitis capitata*, and *Zeugodacus cucurbitae*, a key phenotype white pupae (wp) has been used for decades to selectively remove females before releases, yet the gene responsible remained unknown. Here, we use classical and modern genetic approaches to identify and functionally characterize causal wp^- mutations in these distantly related fruit fly species. We find that the wp phenotype is produced by parallel mutations in a single, conserved gene. CRISPR/Cas9-mediated knockout of the wp gene leads to the rapid generation of white pupae strains in *C. capitata* and *B. tryoni*. The conserved phenotype and independent nature of wp^- mutations suggest this technique can provide a generic approach to produce sexing strains in other major medical and agricultural insect pests.

¹School of Biological Sciences, University of Adelaide, 5005 Adelaide, Australia. ²Department of Insect Biotechnology in Plant Protection, Justus-Liebig-University Gießen, Institute for Insect Biotechnology, Winchesterstr. 2, 35394 Gießen, Germany. ³Centre for Genomic Research, Institute of Integrative Biology, The Biosciences Building, Crown Street, L69 7ZB Liverpool, United Kingdom. ⁴Insect Pest Control Laboratory, Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture, Seibersdorf, 1400 Vienna, Austria. ⁵McGill University Genome Centre, McGill University, Montreal, QC, Canada. ⁶Canadian Centre for Computational Genomics (C3G), McGill University, Montreal, QC, Canada. ⁷Department of Environmental Engineering, University of Patras, 2 Seferi str., 30100 Agrinio, Greece. ⁸Bio21 Molecular Science and Biotechnology Institute, School of BioSciences, University of Melbourne, Melbourne 3010, Australia. ⁹South Australian Research and Development Institute, Waite Road, Urrbrae 5064, South Australia. ¹⁰USDA-ARS Daniel K. Inouye US Pacific Basin Agricultural Research Center, 64 Nowelo Street, Hilo, HI 96720, USA. ¹¹Biology Centre, Czech Academy of Sciences, Institute of Entomology, Branišovská 31, 370 05 České Budějovice, Czech Republic. ¹²Authors contributed equally to the study: Christopher M. Ward, Roswitha A. Aumann. ✉email: k.bourtzis@iaea.org; simon.baxter@unimelb.edu.au; marc.schetelig@agrar.uni-giessen.de

Tephritid species, including the Mediterranean fruit fly (medfly) *Ceratitis capitata*, the oriental fruit fly *Bactrocera dorsalis*, the melon fly *Zeugodacus cucurbitae*, and the Queensland fruit fly *Bactrocera tryoni*, are major agricultural pests worldwide¹. The sterile insect technique (SIT) is a species-specific and environment-friendly approach to control their populations, which has been successfully applied as a component of area-wide integrated pest management programs^{2–4}. The efficacy and cost-effectiveness of these large-scale operational SIT applications has been significantly enhanced by the development and use of genetic sexing strains (GSS) for medfly, *B. dorsalis* and *Z. cucurbitae*^{5,6}.

A GSS requires two principal components: a selectable marker, which could be phenotypic or conditionally lethal, and the linkage of the wild-type allele of this marker to the male sex, ideally as close as possible to the male determining region. In a GSS, males are heterozygous and phenotypically wild type, whilst females are homozygous for the mutant allele thus facilitating sex separation^{6–8}. Puparium color was one of the first phenotypic traits exploited as a selectable marker for the construction of GSS. In all three species, brown is the typical puparium color. However, naturally occurring color mutants such as white pupae (wp)⁹ and dark pupae (dp)¹⁰ have occurred in the field or laboratory stocks. The wp locus was successfully used as a selectable marker to develop GSS for *C. capitata*, *B. dorsalis*, and *Z. cucurbitae*^{6,11,12}; however, its genetic basis has never been resolved.

Biochemical studies provided evidence that the white pupae phenotype in medfly is due to a defect in the mechanism responsible for the transfer of catecholamines from the hemolymph to the pupal cuticle¹³. In addition, classical genetic studies showed that the wp phenotype is due to a recessive mutation in an autosomal gene located on chromosome 5 of the medfly genome^{9,14}. The development of translocation lines combined with deletion and transposition mapping and advanced cytogenetic studies allowed the localization of the gene responsible for the wp phenotype on the right arm of chromosome 5, at position 59B of the trichogen polytene chromosome map¹⁵. In the same series of experiments, the wp locus was shown to be tightly linked to a *temperature-sensitive lethal* (*tsl*) gene (position 59B–61C), which is the second selectable marker of the VIENNA 7 and VIENNA 8 GSS currently used in all medfly SIT operational programs worldwide^{7,15}.

The genetic stability of a GSS is a major challenge, mainly due to recombination phenomena taking place between the selectable marker and the translocation breakpoint. To address this risk, a chromosomal inversion called D53 was induced and integrated into the medfly VIENNA 8 GSS (VIENNA 8^{D53+})^{6,8}. Cytogenetic analysis indicated that the D53 inversion spans a large region of chromosome 5 (50B–59C on trichogen polytene chromosome map) with the wp locus being inside the inversion, close to its right breakpoint⁶.

Extensive genetic and cytogenetic studies facilitated the development of a physical map of the medfly genome^{8,16}. The annotated gene set provided opportunities for the identification of genes or loci-associated mutant phenotypes, such as the wp and *tsl*, used for the construction of GSS^{16,17}. Salivary gland polytene chromosome maps developed for *C. capitata*, *B. dorsalis*, *Z. cucurbitae*, and *B. tryoni* show that their homologous chromosomes exhibit similar banding patterns. In addition, *in situ* hybridization analysis of several genes confirmed that there is extensive shared synteny, including the right arm of chromosome 5 where the *C. capitata* wp gene is localized⁸. Interestingly, two recent studies identified SNPs associated with the wp phenotype in *C. capitata* and *Z. cucurbitae* that were also on chromosome 5^{18,19}.

In this work, we employ different strategies involving genetics, cytogenetics, genomics, transcriptomics, gene editing, and

bioinformatics to identify independent natural mutations in a gene responsible for puparium coloration in three tephritid species of major agricultural importance, *C. capitata*, *B. dorsalis*, and *Z. cucurbitae*. We then functionally characterize causal mutations within this gene in *C. capitata* and *B. tryoni* resulting in development of new white pupae strains. Due to its conserved nature²⁰ and widespread occurrence in many insect species of agricultural and medical importance, we also discuss the potential use of this gene as a generic selectable marker for the construction of GSS for SIT applications.

Results

Resolving the *B. dorsalis* wp locus by introgression experiments. The *B. dorsalis* white pupae phenotype was introgressed into *B. tryoni* to generate a strain referred to as the *Bactrocera* introgressed line (*BIL*, Supplementary Fig. 1). To determine the proportion of *B. dorsalis* genome introgressed into *BIL*, whole-genome sequence data from male and female *B. dorsalis*, *B. tryoni*, and *BIL* individuals were analyzed. Paired-end Illumina short read data from single *B. oleae* males (SRR826808) and females (SRR826807) were used as an outgroup. Single copy orthologs across the genome ($n = 1,846$) were used to reconstruct the species topology revealing a species-specific monophyly (Fig. 1a) consistent with published phylogenies^{21,22}. Reconstruction also showed monophyly between *B. tryoni* and *BIL* across 99.2% of gene trees suggesting the majority of loci originally introgressed from *B. dorsalis* have been removed during backcrosses.

Genomes were partitioned into 100 kb windows and pairwise absolute genetic distance (d_{XY}) calculated between each species and *BIL* to estimate admixture. *Bactrocera dorsalis* was found to be highly similar to a small proportion of the *BIL* genome (Fig. 1b; purple), as indicated by d_{XY} values approaching the median value of *B. dorsalis* vs *B. tryoni* (Fig. 1b; yellow).

Two formal tests for introgression were also carried out, the f_d estimator (Fig. 1c) and topology weighting (Fig. 1d). Three distinct local evolutionary histories (Fig. 1d) were tested using d_{XY} and topology weighting across the *B. dorsalis* wp Quantitative Trait Locus (QTL) i) *BIL* is closest to *B. tryoni* (Fig. 1d; purple, expected across most of the genome), ii) *BIL* is closest to *B. dorsalis* (Fig. 1d; orange, expected at the wp locus), and iii) *BIL* is closest to *B. oleae* (Fig. 1d; green, a negative control). Across the nuclear genome the species topology was supported in 98.82% of windows. Both f_d and topology weighting confirmed a lack of widespread introgression from *B. dorsalis* into *BIL* with few ($n = 42$) discordant outlier windows. Genomic windows discordant across all three tests were considered candidate regions for the wp mutation. Four scaffolds accounting for 1.18% of the *B. dorsalis* genome met these criteria and only two, NW_011876372.1 and NW_011876398.1, showed homozygous introgression consistent with a recessive white pupae phenotype (Supplementary Fig. 2).

To resolve breakpoints within the *B. dorsalis* wp QTL, a windowed analysis across NW_011876398.1 and NW_011876372.1 was performed using d_{XY} (Fig. 1e), topology weighting (Fig. 1f) and f_d (Fig. 1g). The maximum range of the introgressed locus was 4.49 Mb (NW_011876398.1 was 2.9–5.94 Mb and NW_011876372.1 was 0–1.55 Mb) (Fig. 1e–g). The wp QTL was further reduced to a 2.71 Mb region containing 113 annotated protein coding genes through analyzing nucleotide diversity (π) among eight pooled *BIL* genomes (3.8 Mb on NW_011876398.1 to 0.73 Mb on scaffold NW_011876372.1, Supplementary Fig. 2).

Resolving the *C. capitata* wp by genome sequencing and *in situ* hybridization. Cytogenetic studies have determined the gene responsible for the white pupae phenotype to be localized on the right arm of chromosome 5, at position 59B of the trichogen polytene

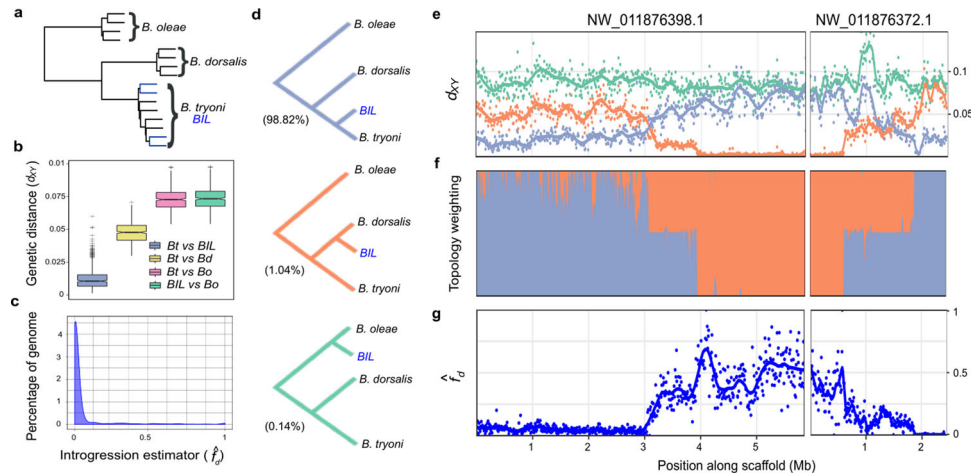


Fig. 1 Characterization of total introgression from *B. dorsalis* into the *Bactrocera* introgressed line and identification of the white pupae locus. **a** Species tree constructed from 1846 single copy ortholog gene trees for four haplotypes of *B. oleae*, *B. dorsalis*, *B. tryoni*, and *BIL*. Branches corresponding to *BIL* individuals are shown in blue. All nodes were well supported with posterior probabilities >0.97. **b** Nei's absolute genetic distance (d_{xy}) calculated for tiled 100 kb windows across the genome between *B. tryoni* vs *BIL* (*Bt* vs *BIL*); *B. tryoni* vs *B. dorsalis* (*Bt* vs *Bd*); *B. tryoni* vs *B. oleae* (*Bt* vs *Bo*); and *BIL* vs *B. oleae* (*BIL* vs *Bo*). Box and whisker graphs (including outliers) represent a summary of 2294 genomic windows. Boxes show the first and third inter quartile range (IQR) while whiskers extend to a maximum of $1.5 \times$ IQR. All values outside $1.5 \times$ IQR are shown as plus signs. **c** The introgression estimator (f_d) calculated across tiled 100 kb windows to identify regions of disproportionately shared alleles between *BIL* and *B. dorsalis*, f_d (*Bt*, *BIL*, *Bd*; *Bo*). **d** The three evolutionary hypothesis/topologies of interest to identify introgressed regions and their representation across the genome: species (purple, 98.82%), introgression (orange, 1.04%) and a negative control tree (green, 0.14%). **e** Nei's absolute genetic distance (d_{xy}) calculated for tiled 10 kb windows across the candidate *wp* locus for *B. tryoni* vs *BIL* (purple), *B. dorsalis* vs *BIL* (orange), *B. oleae* vs *BIL* (green). **f** Topology weighting for each topology shown in **d**, calculated for 1 kb tiled local trees across the candidate *wp* locus. **g** The introgression estimator (f_d) calculated across tiled 10 kb windows for the comparison f_d (*Bt*, *BIL*, *Bd*; *Bo*) to identify the start and end of the introgressed locus. Source data are provided in a Source Data file.

chromosome map¹⁵. The equivalent of position 59B is position 76B of the salivary gland polytene chromosome map, inside but close to the right breakpoint of the D53 inversion (69C–76B on the salivary gland polytene chromosome map). Long read sequencing data were generated of the wild-type strain Egypt II (EgII, WT), the inversion line D53 and the genetic sexing strain VIENNA 8 (without the inversion; VIENNA 8^{D53-|}) (Supplementary Table 1) to enable a comparison of the genomes and locate the breakpoints of the D53 inversion, to subsequently narrow down the target region, and to identify *wp* candidate genes.

Chromosome 5-specific markers¹⁶ were used to identify the EgII_Ccap3.2.1 scaffold_5 as complete chromosome 5. Candidate D53 breakpoints in EgII scaffold_5 were identified using the alignment of three genome datasets EgII, VIENNA 8^{D53-|}, and D53 (see material and methods). The position of the D53 inversion breakpoints was located between 25,455,334 and 25,455,433 within a scaffold gap (left breakpoint), and at 61,880,224 bp in a scaffolded contig (right breakpoint) on EgII chromosome 5 (Ccap3.2.1; accession GCA_905071925) (Fig. 2a). The region containing the causal *wp* gene was known to be just next to the right breakpoint of the D53 inversion. Cytogenetic analysis and in situ hybridization using the WT EgII strain and the D53 inversion line confirmed the overall structure of the inversion, covering the area of 69C–76B on the salivary gland polytene chromosomes (Fig. 2), as well as the relative position of markers residing inside and outside the breakpoints (Fig. 2 and Supplementary Fig. 3). PCRs using two primer pairs flanking the predicted breakpoints (Supplementary Fig. 4) and subsequent sequencing confirmed the exact sequence of the breakpoints. Thereby, the wild-type status was confirmed for EgII flies and VIENNA 7^{D53+|} GSS males, which are heterozygous for the inversion. Correspondingly, these amplicons were not present in D53 males and females or in VIENNA 7^{D53+|} GSS females (all homozygous for the inversion)

(Supplementary Fig. 4). Positive signals for the inversion were detected in D53 and VIENNA 7^{D53+} GSS males and females, but not in WT flies using an inversion-specific primer pair (Supplementary Fig. 4).

Genome and transcriptome sequencing reveal a single candidate *wp* gene. Orthologs within the QTL of *B. dorsalis*, *C. capitata*, and scaffolds known to segregate with the *wp* phenotype in *Z. cucurbitae* (NW_011863770.1 and NW_011863674.1)¹⁸ were investigated for null mutations under the assumption that errors within a conserved gene result in white pupae. A single ortholog containing fixed indels absent from wild-type strains was identified in each species. White pupae *B. dorsalis* and *BIL* strains showed a 37 bp frame-shift deletion in the first coding exon of LOC105232189 introducing a premature stop codon 210 bp from the transcription start site (Fig. 3a). Presence of the deletion was confirmed in silico using whole genome resequencing from the *wp* and wildtype mapped to the reference, and by de novo assembly of Illumina RNAseq data transcripts (Fig. 3a).

In *C. capitata*, a D53 Nanopore read alignment on EgII showed an independent approximate 8150 bp insertion into the third exon of LOC101451947 disrupting proper gene transcription 822 bp from the transcription start site (Fig. 3b). The insertion sequence is flanked by identical repeats, suggesting that it may originate from a transposable element insertion. The *C. capitata* mutation was confirmed in silico, as in *B. dorsalis*, using whole genome sequencing and RNAseq data (Fig. 3b).

Transcriptome data from the white pupae-based genetic sexing strain of *Z. cucurbitae* revealed a 13 bp deletion in the third exon of LOC105216239 on scaffold NW_011863770.1 introducing a premature stop codon (Fig. 3c).

The candidate white pupae gene in all three species had a reciprocal best BLAST hit to the putative metabolite transport

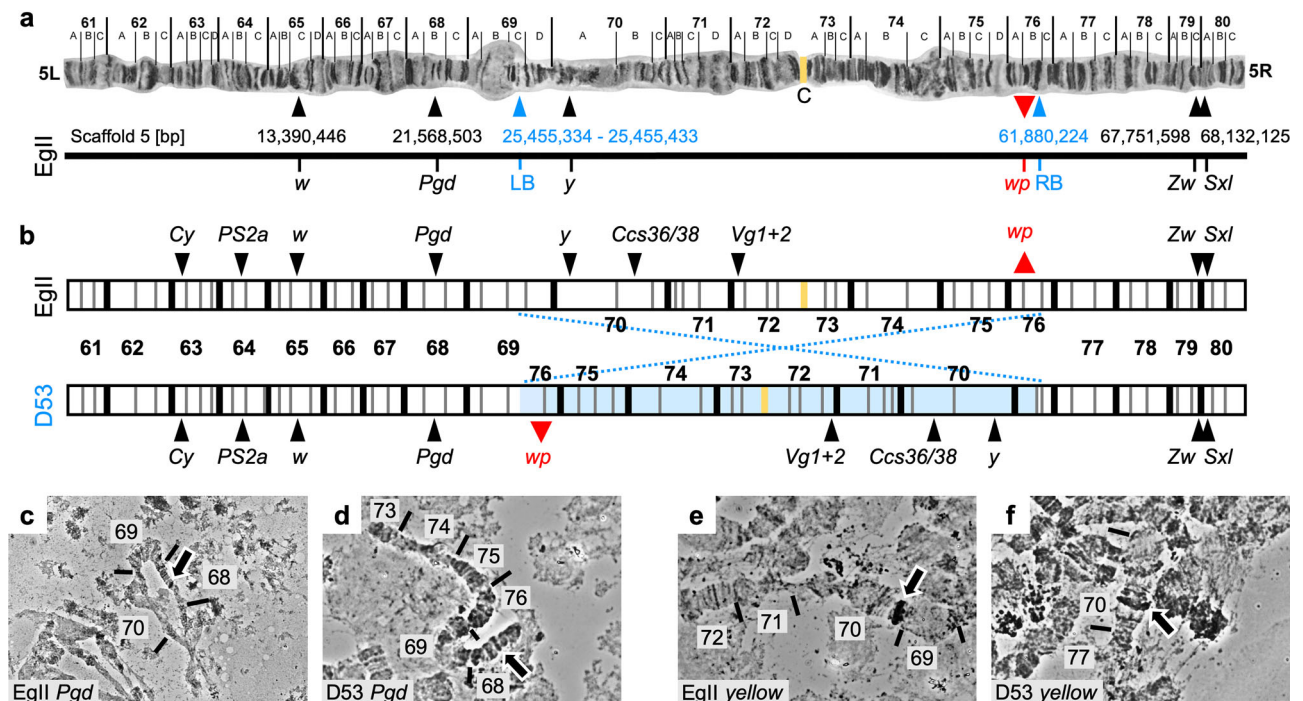


Fig. 2 Genomic positioning of the D53 inversion on chromosome 5 of *C. capitata*. **a** Chromosome scale assembly of *C. capitata* EglII chromosome 5. Shown are the positions of in situ mapped genes *white* (*w*), *6-phosphogluconate dehydrogenase* (*Pgd*), *glucose-6-phosphate 1-dehydrogenase* (*Zw*), and *sex lethal* (*Sxl*), the position of the D53 inversion breakpoints (blue; LB = left breakpoint, RB = right breakpoint), and the relative position of *white pupae* (*wp*) on the polytene chromosome map of chromosome 5⁷¹ (left (L) and right (R) chromosome arm, linked at the centromeric region (C)) and the PacBio-Hi-C EglII scaffold_5 (bp = base pairs), representing the complete chromosome 5 (Ccap3.2.1, accession GCA_905071925). The position of the *yellow* gene (*y*, LOC101455502) was confirmed on chromosome 5 70A by in situ hybridization, despite its sequence not been found in the scaffold assembly. **b** Schematic illustration of chromosome 5 without (EglII, WT) and with (D53) D53 inversion, with additional marker genes *Curly* (*Cy*), *integrin- α PS2* (*PS2a*), *white* (*w*), *chorion S36/38* (*Ccs36/38*), *vitellogenin-1/2-like* (*Vg1+2*). The inverted part of chromosome 5 is shown in light blue, the centromere in yellow. Two probes, one inside (*y*, 70A) and one outside (*Pgd*, 68B) of the left inversion breakpoint were used to verify the D53 inversion breakpoints by in situ hybridization. WT EglII is shown in **c** and **e**, D53 in **d** and **f**. Chromosomal segments are numbered, arrows in micrographs indicate in situ hybridization signal. In situ hybridizations were done at least in duplicates and at least ten nuclei were analyzed per sample, scale bar = 10 μ m. All replicates led to similar results. The source data underlying Fig. 2c-f are provided as a Source Data file.

protein CG14439 in *Drosophila melanogaster* and contains a Major Facilitator-like superfamily domain (MFS_1, pfam07690), suggesting a general function as a metabolite transport protein. In situ hybridization on polytene chromosomes of *B. dorsalis*, *C. capitata* and *Z. cucurbitae* was used to confirm the presence of the *wp* locus in the same syntenic position on the right arm of chromosome 5 (Fig. 3d–f). Therefore, all three species show a mutation in the same positional orthologous gene likely to be responsible for the phenotype in all three genera.

Knockout of the MFS gene causes white pupae phenotypes. An analogous *B. dorsalis wp*⁻ mutation was developed in *B. tryoni* by functional knockouts of the putative *Bt_wp* using the CRISPR/Cas9 system. A total of 591 embryos from the Ourimbah laboratory strain were injected using two guides with recognition sites in the first coding exon of this gene (Fig. 4a). Injected embryos surviving to adulthood ($n = 19$, 3.2%) developed with either wild-type brown ($n = 12$) or somatically mosaic white-brown puparia ($n = 7$, Supplementary Fig. 5). Surviving G₀ adults were individually backcrossed into the Ourimbah strain, resulting in potentially *wp*^{+|-(CRISPR)} heterozygous brown pupae (Fig. 4c). Five independent G₀ crosses were fertile (three mosaic white-brown and two brown pupae phenotypes). G₁ offspring were sibling mated and visual inspection of G₂ progeny revealed that three families contained white pupae individuals. Four distinct

frameshift mutations were observed in screened G₂ progeny (Fig. 4a) suggesting functional KO of putative *Bt_wp* is sufficient to produce the white pupae phenotype in *B. tryoni*. Capillary sequencing of cloned *Bt_MFS* amplicons revealed deletions ranging from a total of 4–155 bp, summed across the two guide recognition sites, introducing premature stop codons.

In *C. capitata*, CRISPR/Cas9 gene editing was used to knockout the orthologous gene and putative *Cc_wp*, LOC101451947, to confirm that it causes a white puparium phenotype. A mix of recombinant Cas9 protein and the gRNA_MFS, targeting the third exon and thereby the MFS domain of the presumed *Cc_wp* CDS (Fig. 4b), was injected into 588 EglII WT embryos of which 96 developed to larvae and 67 pupated. All injected G₀ pupae showed brown pupal color. In total, 29 G₀ males and 34 females survived to adulthood (9.3%) and were backcrossed individually or in groups (see material and methods) to a strain carrying the naturally occurring *white pupae* mutation (*wp*^{-(nat)}; strain #1402_22m1B)²³ (Fig. 4d). As *white pupae* is known to be monogenic and recessive in *C. capitata*, this complementation assay was used to reveal whether the targeted gene is responsible for the naturally occurring white pupae phenotype or if the mutation is located in a different gene. G₁ offspring would only show white pupae phenotypes if *Cc_wp* was indeed the *white pupae* gene, knocked-out by the CRISPR approach, and complemented by the natural mutation through the backcross (*wp*^{-(nat)|-(CRISPR)}). In the case that the *Cc_wp* is not the gene

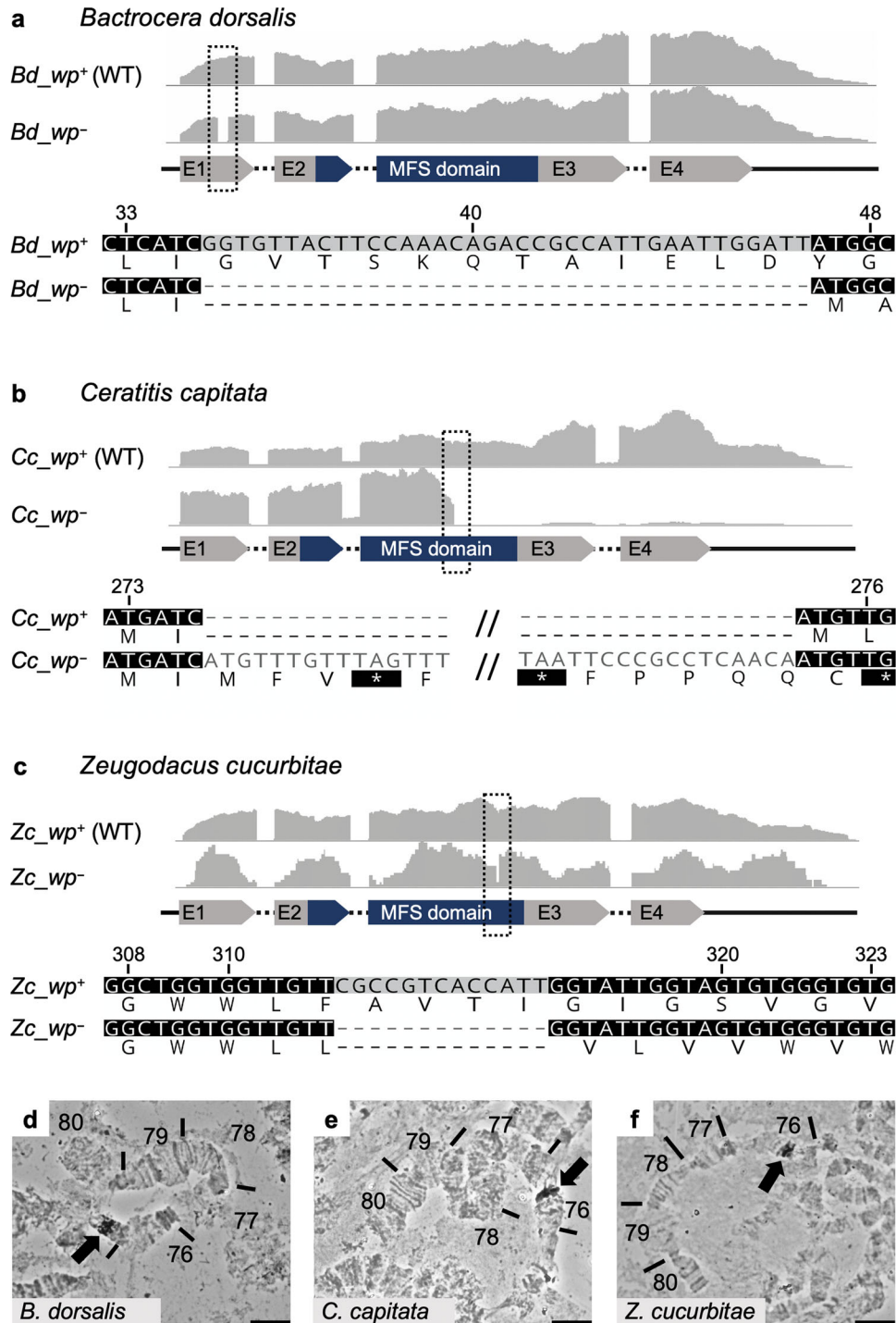
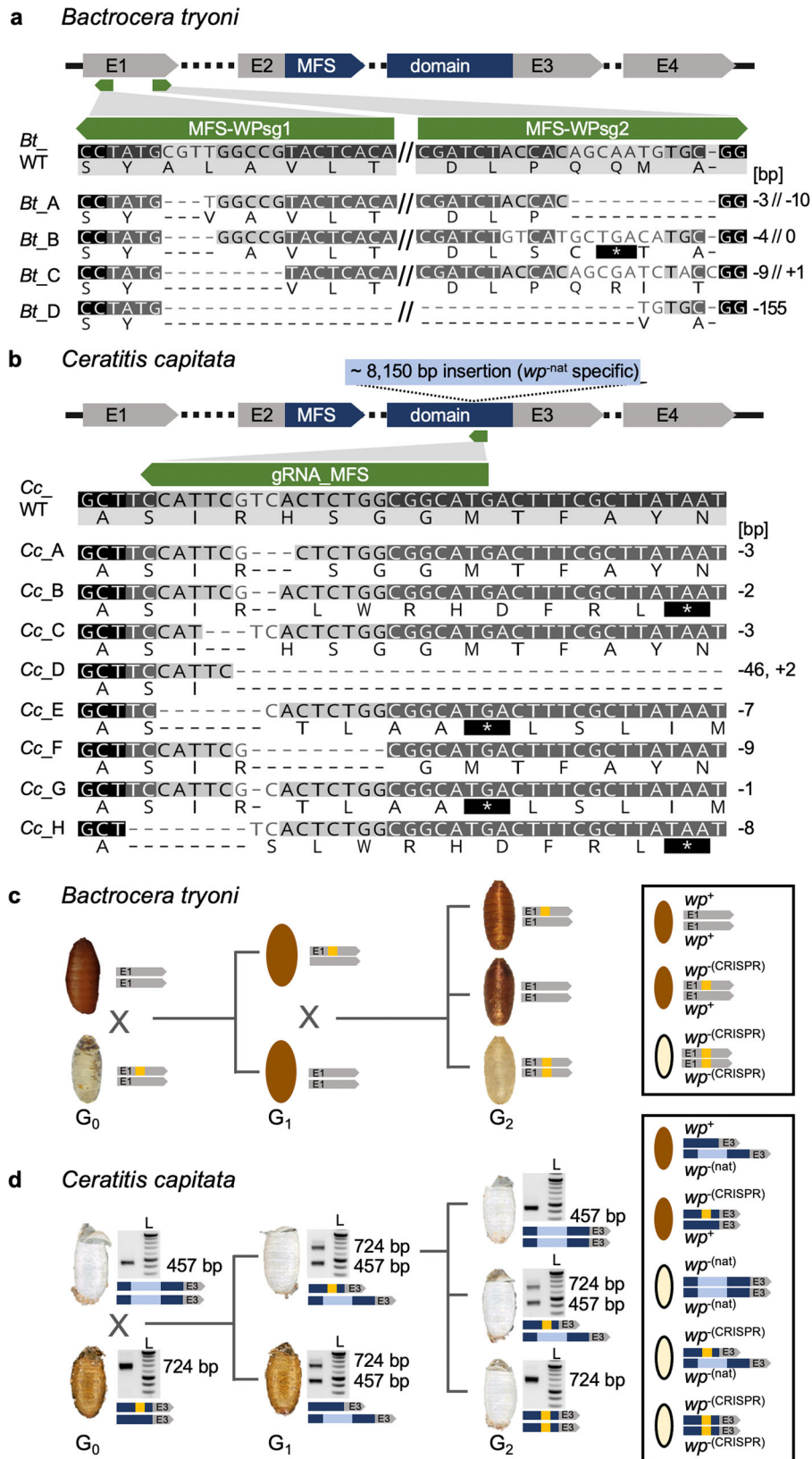


Fig. 3 Identification of the *wp* mutation in the transcriptomes of *B. dorsalis*, *C. capitata*, and *Z. cucurbitae*. The gray graphs show expression profiles from the candidate *wp* loci in WT (*wp*⁺) and mutant (*wp*⁻) flies at the immobile pupae stages of **a** *B. dorsalis*, **b** *C. capitata*, and **c** *Z. cucurbitae*. The gene structure (not drawn to scale) is indicated below as exons (arrows labeled E1–E4) and introns (dashed lines), the Major Facilitator Superfamily (MFS) domain is shown in blue. The positions of independent *wp* mutations (*Bd*: 37 bp deletion, *Cc*: approximate 8150 bp insertion, *Zc*: 13 bp deletion) are marked with black dashed boxes in the expression profiles and are shown in detail below the gene models based on de novo assembly of RNAseq data from WT and white pupae phenotype individuals (nucleotide and amino acid sequences). Deletions are shown as dashes, alterations on protein level leading to premature stop codons are depicted as asterisks highlighted in black. In situ hybridization on polytene chromosomes for **d** *B. dorsalis*, **e** *C. capitata*, and **f** *Z. cucurbitae* confirmed the presence of the *wp* locus on the right arm of chromosome 5 in all three species (arrows in micrographs). In situ hybridizations were done at least in duplicates and at least ten nuclei were analyzed per sample, scale bar = 10 μm. The source data underlying Fig. 3d–f are provided as a Source Data file.



carrying the natural wp^- mutation, a brown phenotype would be observed for all offspring. Here, five out of 13 crosses, namely M1, M3, F2, F3, and F4, produced white pupae phenotype offspring. The crosses generated 221, 159, 70, 40, and 52 G_1 pupae, of which 10, 30, 16, 1, and 1 pupa respectively, were white. Fifty-seven flies

emerged from white puparia were analyzed via non-lethal genotyping, and all of them showed mutation events within the target region. Overall, eight different mutation events were seen, including deletions ranging from 1 to 9 bp and a 46 bp deletion combined with a 2 bp insertion (Fig. 4b). Five mutation events

Fig. 4 CRISPR/Cas9-based generation of homozygous $wp^{-(\text{CRISPR})}$ lines in *B. tryoni* and *C. capitata*. A schematic structure of the wp CDS exons (E1, E2, E3, E4) including the MFS domain in *B. tryoni* (a) and *C. capitata* (b) are shown. Positions of gRNAs targeting the first and third exon in *B. tryoni* and *C. capitata*, respectively, are indicated by green arrows. Nucleotide and amino acid sequences of mutant wp alleles identified in G_1 individuals are compared to the WT reference sequence in *B. tryoni* (a) and *C. capitata* (b). Deletions are shown as dashes, alterations on protein level leading to premature stop codons are depicted as asterisks highlighted in black. Numbers on the right side represent InDel sizes (bp = base pairs). Crossing schemes to generate homozygous $wp^{-(\text{CRISPR})}$ lines in *B. tryoni* (c) and *C. capitata* (d) show different strategies to generate wp strains. Bright-field images of empty puparia are depicted for both species. Genotype schematics and corresponding PCR analysis (for *C. capitata*) validating the presence of CRISPR-induced (orange) and natural (blue, for *C. capitata*) wp mutations are shown next to the images of the puparia. c Injected G_0 *B. tryoni* were backcrossed to the Ourimbah laboratory strain resulting in uniformly brown G_1 offspring (depicted as illustration because no images were acquired during G_1). G_1 inbreeding led to G_2 individuals homozygous for the white pupae phenotype. d Injected WT G_0 flies were crossed to flies homozygous for the naturally occurring wp^- allele ($wp^{-(\text{nat})}$). $wp^{-(\text{nat})}$ (457 bp amplicon) and $wp^{-(\text{CRISPR})}$ or WT (724 bp amplicon) alleles were identified by multiplex PCR (left lane; L = NEB 2 log ladder). White pupae phenotypes in G_1 indicated positive CRISPR events. G_2 flies with a white pupae phenotype that were homozygous for the $wp^{-(\text{CRISPR})}$ allele were used to establish lines. PCR was done once for each individual, $wp^{-(\text{CRISPR})}$ alleles were verified and further analyzed via sequencing. The source data underlying Fig. 4d are provided as a Source Data file.

(B, D, E, G, H) caused frameshifts and premature stop codons. The remaining three (A, C, F), however, produced deletions of only one to three amino acids. Mutants were either inbred (mutation C) (Fig. 4d) or outcrossed to WT EgII (mutation A–H), both in groups according to their genotype. This demonstrated that Cc_wp is the gene carrying the $wp^{-(\text{nat})}$, and that even the loss of a single amino acid without a frameshift at this position can cause the white pupae phenotype. Offspring from outcrosses of mutation A, D, and H, as well as offspring of the inbreeding (mutation C), were genotyped via PCR, and $wp^{+|-(\text{CRISPR})}$ and $wp^{-(\text{CRISPR})|-(\text{CRISPR})}$ positive flies were inbred to establish homozygous $wp^{-(\text{CRISPR})}$ lines.

Discussion

White pupae (wp) was first identified in *C. capitata* as a spontaneous mutation and was subsequently adopted as a phenotypic marker of fundamental importance for the construction of GSS for SIT^{6,9}. Full penetrance expressivity and recessive inheritance rendered wp the marker of choice for GSS construction in two additional tephritid species, *B. dorsalis* and *Z. cucurbitae*^{11,12}, allowing automated sex sorting based on pupal color. This was only possible because spontaneous wp mutations occur at relatively high rates either in the field or in mass rearing facilities and can easily be detected^{6,9}. Despite the easy detection and establishment of wp mutants in these three species, similar mutations have not been detected in other closely or distantly related species such as *B. tryoni*, *B. oleae*, or *Anastrepha ludens*, despite large screens being conducted. In addition to being a visible GSS marker used to separate males and females, the wp phenotype is also important for detecting and removing recombinants in cases where sex separation is based on a conditional lethal gene such as the *tsl* gene in the medfly VIENNA 7 or VIENNA 8 GSS^{6,7}. However, it took more than 20 years from the discovery and establishment of the wp mutants to the large-scale operational use of the medfly VIENNA 8 GSS for SIT applications^{6,9} and the genetic nature of the wp mutation remained unknown. The discovery of the underlying wp mutations and the availability of CRISPR/Cas genome editing would allow the fast recreation of such phenotypes and sexing strains in other insect pests. Isolation of the wp gene would also facilitate future efforts towards the identification of the closely linked *tsl* gene.

Using an integrated approach consisting of genetics, cytogenetics, genomics, transcriptomics, and bioinformatics, we identified the white pupae genetic locus in three major tephritid agricultural pest species, *B. dorsalis*, *C. capitata*, and *Z. cucurbitae*. Our study clearly shows the power of employing different strategies for gene discovery, one of which was species hybridization. In *Drosophila*, hybridization of different species has played a

catalytic role in the deep understanding of species boundaries and the speciation processes, including the evolution of mating behavior and gene regulation^{24–28}. In our study, we took advantage of two closely related species, *B. dorsalis* and *B. tryoni*, which can produce fertile hybrids and be backcrossed for consecutive generations. This allowed the introgression of the wp mutant locus of *B. dorsalis* into *B. tryoni*, resulting in the identification of the introgressed region, including the causal wp mutation via whole-genome resequencing and advanced bioinformatic analysis.

In *C. capitata*, we exploited two essential pieces of evidence originating from previous genetic and cytogenetic studies: the localization of wp to region 59B and 76B on chromosome 5 in the trichogen cells and salivary gland polytene chromosome map, respectively^{15,29}, and its position close to the right breakpoint of the large inversion D53⁶. This data prompted us to undertake a comparative genomic approach to identify the exact position of the right breakpoint of the D53 inversion, which would bring us in the vicinity of the wp gene. Coupled with comparative transcriptomic analysis, this strategy ensured that the analysis indeed tracked the specific wp locus on the right arm of chromosome 5, instead of any mutation in another, random locus which may participate in the pigmentation pathway and therefore result in the same phenotype. Functional characterization via CRISPR/Cas9-mediated knockout resulted in the establishment of new white pupae strains in *C. capitata* and *B. tryoni* and confirmed that this gene is responsible for the puparium's coloration in these tephritid species. Interestingly, the wp phenotype is based on three independent and very different natural mutations of this gene, a rather large and transposon-like insertion in *C. capitata*, but only small deletions in the two other tephritids, *B. dorsalis* and *Z. cucurbitae*. In medfly, however, CRISPR-induced in-frame deletions of one or three amino acids in the MFS domain were sufficient to induce the wp phenotype, underlining the importance of this domain for correct coloration of the puparium.

It is worth noting that in the first stages of this study, we employed two additional approaches, which did not allow us to successfully narrow down the wp genomic region to the desired level. The first was based on Illumina sequencing of libraries produced from laser micro-dissected (Y;5) mitotic chromosomes that carry the wild-type allele of the wp gene through a translocation from the fifth chromosome to the Y. This dataset from the medfly VIENNA 7 GSS was comparatively analyzed to wild-type (Egypt II) Y and X chromosomes, and the complete genomes of Egypt II, VIENNA 7^{D53-} GSS, and a D53 inversion line in an attempt to identify the chromosomal breakpoints of the translocation and/or inversion, which are close to the wp locus (Supplementary Table 2). However, this effort was not successful due to the short Illumina reads and the lack of a high-quality reference genome. The second approach was based on individual scale whole-genome

resequencing/genotyping, and identifying fixed loci associated with pupal color phenotypes, which complemented the QTL analysis¹⁹. Seven loci associated with SNPs and larger deletions linked to the white pupae phenotype were analyzed based on their respective mutations and literature searches for their potential involvement in pigmentation pathways. However, we could not identify a clear link to the pupal coloration as shown by *in silico*, molecular, and *in situ* hybridization analysis (Supplementary Figs. 6 and 7, Supplementary Table 3).

The *wp* gene is a member of a Major Facilitator Superfamily (MFS). Orthologs of *white pupae* are present in 146 of 148 insect species aggregated in OrthoDB²⁰ v9 and are single copy in 133 species. Furthermore, *wp* is included in the benchmarking universal single copy ortholog (BUSCO) gene set for Insecta and according to OrthoDB³⁰ v10 has a below average evolutionary rate (0.87, OrthoDB group 42284at50557) suggesting an important and evolutionarily conserved function (Supplementary Fig. 8). Its ortholog in *Bombyx mori*, *mucK*, was shown to participate in the pigmentation at the larval stage³¹ whereas in *D. melanogaster* peak expression is during the prepupal stage after the larva has committed to pupation³², which is the stage where pupal cuticle sclerotization and melanization occurs. It is known that the insect cuticle consists of chitin, proteins, lipids, and catecholamines, which act as cross-linking agents thus contributing to polymerization and the formation of the integument³³. Interestingly, the sclerotization and melanization pathways are connected and this explains the different mechanical properties observed in different medfly pupal color strains with the dark color cuticles being harder than the brown ones and the latter harder than the white color ones³⁴. The fact that the white pupae mutants are unable to transfer catecholamines from the hemolymph to the cuticle is perhaps an explanation for the lack of the brown pigmentation¹³.

The discovery of the long-sought *wp* gene in this study and the recent discovery of the *Maleness-on-the-Y* (*MoY*) gene, which determines the male sex in several tephritids³⁵, opens the way for the development of a generic approach for the construction of GSS for other species. Using CRISPR/Cas-based genome editing approaches, we can: (a) induce mutations in the *wp* orthologues of SIT target species and establish lines with *wp* phenotype and (b) link the rescue alleles as closely as possible to the *MoY* region. Given that the *wp* gene is present in diverse insect species including agricultural insect pests and mosquito disease vectors, this approach would allow more rapid development of GSS in SIT target species, members of diverse families, such as the agricultural pest species *A. ludens*, *A. fraterculus*, *B. dorsalis*, *B. correcta*, *B. oleae*, *Drosophila suzukii*, *Cydia pomonella*, *Pectinophora gossypiella*, *Lobesia botrana*; the livestock pests *Glossina morsitans*, *G. pallidipes*, *G. palpalis gambiensis*, *G. austeni*; and the mosquito disease vectors *Aedes aegypti*, *Aedes albopictus*, and *Anopheles arabiensis*. However, the biological quality of any new strain which is considered for SIT application should be first thoroughly tested in respect to their fitness and male mating competitiveness. In principle, these GSS will have higher fertility compared to the semi-sterile translocation lines⁶. In addition, these new generation GSS will be more stable since the rescue allele will be tightly linked to the male determining region thus eliminating recombination which can jeopardize the genetic integrity of any GSS. The concept of the generic approach can also be applied in species which lack a typical Y chromosome such as *Ae. aegypti* and *Ae. albopictus*. In these species, the rescue allele should be transferred close to the male determining gene (*Nix*) and the M locus^{36,37}. It is hence important for this generic approach to identify regions close enough to the male determining loci to ensure the genetic stability of the GSS and to allow the proper expression of the rescue alleles. In the present study,

we have already shown that CRISPR/Cas9-induced mutations resulting in the white pupae phenotype can be developed in SIT target species and the resulting strains provide already new opportunities for GSS based on visible markers.

Methods

Insect rearing. *Ceratitis capitata*, *B. dorsalis*, and *Z. cucurbitae* fly strains were maintained at 25 ± 1 °C, 48% RH and 14/10 h light/dark cycle, and fed with a mixture of sugar and yeast extract (3 v:1 v) and water. Larvae were reared on a gel diet, containing carrot powder (120 g/L), agar (3 g/L), yeast extract (42 g/L), benzoic acid (4 g/L), HCl (25%, 5.75 mL/L), and ethyl-4-hydroxybenzoate (2.86 g/L). Flies were anesthetized with N₂ or CO₂ for screening, sexing, and the setup of crosses. To slow down the development during the non-lethal genotyping process (*C. capitata*), adult flies were kept at 19 °C, 60% RH, and 24 h light for this period (1–4 days).

Bactrocera tryoni flies were obtained from New South Wales Department of Primary Industries (NSW DPI), Ourimbah, Australia and reared at 25 ± 2 °C, 65 ± 10% RH and 14/10 h light/dark cycle. Flies were fed with sugar, Brewer's yeast and water and larvae were reared on a gel diet, containing Brewer's yeast (204 g/L), sugar (121 g/L), methyl p-hydroxy benzoate (2 g/L), citric acid (23 g/L), wheat germ oil (2 g/L), sodium benzoate (2 g/L), and agar (10 g/L).

Introgression and identification of *wp* in *B. dorsalis*. Interspecific crosses between male *B. tryoni* (*wp*^{+/+}) and female *B. dorsalis* (*wp*^{-/-}) were carried out. The F₁ *wp*^{+/-} hybrids developed with brown puparia and were mass crossed. F₂ *wp*^{-/-} females were backcrossed into *B. tryoni* *wp*^{+/+} males. Backcrossing was then repeated five additional times to produce the white pupae *Bactrocera* introgressed line (*BIL*, Supplementary Fig. 1).

Genome sequencing using Illumina NovaSeq (2 × 150 bp, Deakin University) was performed on a single male and female from the *B. dorsalis wp* strain, *B. tryoni*, and the *BIL* (~26X) and two pools of five *BIL* individuals (~32X). Quality control of each sequenced library was carried out using FastQC v0.11.6 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and aggregated using ngsReports³⁸ v1.3. Adapter trimming was carried out using Trimmomatic³⁹ v0.38 and paired reads were mapped to the *B. dorsalis* reference genome (GCF_000789215.1) using NextGenMap⁴⁰ v0.5.5 under default settings. Mapped data were sorted and indexed using SAMtools, and deduplication was carried out using Picard MarkDuplicates v2.2.4 (<https://github.com/broadinstitute/picard>). Genotypes were called on single and pooled libraries separately with ploidy set to two and ten, respectively, using FreeBayes⁴¹ v1.0.2. Each strain was set as a different population in FreeBayes. Genotypes with less than five genotype depth were set to missing and sites with greater than 20% missing genotypes or indels filtered out using BCFtools⁴² v1.9. Conversion to the genomic data structure (GDS) format was carried out using SeqArray⁴³ v1.26.2 and imported into the R package gear⁴⁴ v0.1 for population genetic analysis.

Single copy orthologs were identified in the *B. dorsalis* reference annotated proteins (NCBI *Bactrocera dorsalis* Annotation Release 100) with BUSCO^{45,46} v3 using the dipteran gene set⁴⁵. Nucleotide alignments of each complete single copy ortholog were extracted from the called genotype set using gear v0.1 and gene trees built using RAXML⁴⁷ v8.2.10 with a GTR + G model. Gene trees were then imported into Astral III⁴⁸ v5.1.1 for species tree estimation. Genome scans of absolute genetic divergence (d_{XY}), nucleotide diversity (π), and the f_d estimator f_d were carried out using gear v0.1. Two levels of analysis were carried out: i) genome wide scans of non-overlapping 100 kb windows and ii) locus scans of 10 kb tiled windows. Local phylogenies were built for nucleotide alignments of non-overlapping 1 kb windows using RAXML v8.2.10 with a GTR + G model and topology weighting was calculated using TWISST⁴⁹.

Introgressed regions (i.e., candidate *wp* loci) were identified by extracting windows in the genome wide scan with topology weighting and f_d greater than 0.75 and visually inspecting the 'locus scan' data set for d_{XY} , f_d , and topology weighting patterns indicative of introgression. Nucleotide alignments of all genes within candidate *B. dorsalis* introgressed regions were extracted from the GDS using gear and visually inspected for fixed mutations in *B. dorsalis wp*, *BIL* individuals, and the two *BIL* pools. Candidate genes were then searched by tBLASTn against the *D. melanogaster* annotated protein set to identify putative functions and functional domains were annotated using HMMer⁵⁰. Mapped read depth was calculated around candidate regions using SAMtools⁵¹ depth v1.9 and each sample's read depth was normalized to the sample maximum to inspect putative deletions. Called genotypes were confirmed by de novo genome assembly of the *B. dorsalis wp* genome using MaSuRCA⁵² v3.3 under default settings. The de novo scaffold containing LOC105232189 was identified using the BLASTn algorithm. *In silico* exon–intron boundaries were then manually annotated in Geneious⁵³ v11.

Identification of the D53 inversion and *wp* in *C. capitata*. Multiple *C. capitata* strains were used for this study. Egypt II (EgII) is a wild-type laboratory strain. D53 is a homozygous strain with an irradiation-induced inversion covering the area 69C–76B on the salivary gland polytene chromosome map (50B–59C on the trichogen cells polytene chromosome map). VIENNA 7 and VIENNA 8 are two GSS in which two (Y;5) translocations, in the region 58B and 52B of the trichogen cells polytene chromosome map, respectively, have resulted in the linkage of the wild-type allele of the *wp* and *tsl* genes to the male determining region of the

Y chromosome. Thus, VIENNA 7 and VIENNA 8 males are heterozygous in the *wp* and *tsl* loci but phenotypically wild type while VIENNA 7 and VIENNA 8 females are homozygous for the mutant alleles and phenotypically white pupae, and they die when exposed to elevated temperatures. The VIENNA 7 and VIENNA 8 GSS can be constructed with and without the D53 inversion (VIENNA 7/8^{D53+} or ^{D53-}). When the GSS have the inversion, females are homozygous (^{D53++}) for D53 while males are heterozygous (^{D53+-})^{6,8,16}.

To perform whole genome sequencing of *C. capitata* strains, high-molecular-weight (HMW) DNA was extracted from *C. capitata* lines (males and females of the WT EgII strain, the VIENNA 7^{D53-} and VIENNA 8^{D53-} GSS and the inversion line D53) and sequenced. Freshly emerged, virgin and unfed males and females were collected from all strains. For 10X Genomics linked read and Nanopore sequencing, the HMW was prepared as follows: twenty individuals of each sex and strain were pooled, ground in liquid nitrogen, and HMW DNA was extracted using the QIAGEN Genomic tip 100/G kit (Qiagen, Germany). For PacBio Sequel an EgII line was created with single pair crossing and subsequent sibling-mating for six generations. In all generations adult and larval diet contained 100 µg/mL tetracycline. HMW DNA from G₆ individuals was prepared as follows: five males from this EgII line were pooled and ground in liquid nitrogen, and HMW DNA was extracted using the phenol/chloroform Phase Lock Gel tubes (QuantaBio)⁵⁴. DNA for Illumina applications was extracted from individual flies (Supplementary Table 1). PacBio de novo sequencing: samples were purified with AMPure beads (Beckman Coulter, UK) (0.6 volumes) and QC checked for concentration, size, integrity, and purity using Qubit (Qiagen, UK), Fragment Analyser (Agilent Technologies) and Nanodrop (Thermo Fisher) machines. The samples were then processed without shearing using the PacBio Express kit 1 for library construction and an input of 4 µg DNA following the manufacturer's protocol. The final library was size-selected using the Sage Blue Pippin (Sage Sciences) 0.75% cassette U1 marker in the range of 25–80 kb. The final library size and concentrations were obtained on the Fragment Analyser before being sequenced using the Sequel 1 2.1 chemistry with V4 primers at a loading on plate concentration of 6 pM and 10 h movie times. For Nanopore sequencing, the ligation sequencing kits SQK-LSK109 or SQK-RAD004 were used as recommended by the manufacturer (Oxford Nanopore Technologies, Oxford, United Kingdom). Starting material for the ligation library preparation were 1–1.5 µg HMW gDNA for the ligation libraries and 400 ng for the rapid libraries. The prepared libraries were loaded onto FLO-PRO002 (R9.4) flow cells. Data collection was carried out using a PromethION Beta with live high accuracy base calling for up to 72 h and with mux scan intervals of 1.5 h. Each sample was sequenced at least twice. Data generated were 7.7 Gb for EgII male, 31.09 Gb for D53 male, 26.72 Gb for VIENNA 7^{D53-} male, and 24.83 Gb for VIENNA 8^{D53-} male. Run metrics are shown in Supplementary Table 4. The PacBio data were assembled using CANU⁵⁵ v1.8 with two parameter settings: the first to avoid haplotype collapsing (genomeSize = 500 m corOutCoverage=200 batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50) and the second to merge haplotypes together (genomeSize = 500 m corOutCoverage=200 correctedErrorRate=0.15). The genome completeness was assessed with BUSCO^{45,46} v3 using the dipteran gene set⁴⁵. The two assemblies were found to be duplicated due to alternative haplotypes. To improve the contiguity and reduce duplication, haploMerger2 v20161205 was used⁵⁶ and the assembly was assessed with BUSCO v3. Phase Genomics Hi-C libraries were made by Phase genomics from males (*n* = 2) of the same family used for PacBio sequencing. Initial scaffolding was completed by Phase Genomics, but edited using the Salsa⁵⁷ v2.2 and 3D-DNA (3D de novo assembly pipeline v180419; <https://github.com/theaidenlab/3d-dna>) software. The resulting scaffolds were allocated a chromosome number using chromosome specific markers¹⁶. Specific attention was made to the assembly and scaffolding of chromosome 5. Two contig misassemblies were detected by the Hi-C data and fitted manually. The new assembly (EgII_Ccap3.2.1) was then validated using the Hi-C data. Genes were called using the Funannotate v1.6.0-24f34f6 software making use of the Illumina RNAseq data generated by this project; mRNA mapping to the genome is described below.

To identify possible breakpoint positions, the Nanopore D53 fly assembly contig_531 was mapped onto the EgII scaffold_5 (from the EgII_CCAP3.2_CANU_Hi-C_scaffolds.fasta assembly) using MashMap v2.0 (<https://github.com/marbl/MashMap>). This helped to visualize the local alignment boundaries (Supplementary Fig. 10). MashMap parameters were set to kmer size = 16; window size = 100; segment length = 500; alphabet = DNA; percentage identity threshold = 95%; filter mode = one-to-one. Subsequent to this, and to help confirm the exact location of the identified breakpoints, minimap2 (v2.17, <https://github.com/lh3/minimap2>) was used to align D53 as well as VIENNA 8^{D53-} and VIENNA 7^{D53-} Nanopore reads onto the EgII scaffold_5 reference (Supplementary Fig. 10). Minimap2 parameters for Nanopore reads were: minimap2 -x map-ont -A 1 -a --MD -L -t 40. Samtools (v1.9, <https://github.com/samtools/samtools>) was used to convert the alignment.sam to.bam and prepare the alignment file to be viewed in the Integrative Genomics Viewer (IGV, <http://software.broadinstitute.org/software/igv/>). The expectation was to see a leftmost breakpoint in D53 read set alignments but not in VIENNA 8^{D53-} and VIENNA 7^{D53-} when compared to the EgII reference (Supplementary Fig. 9). Due to an assembly gap in the EgII scaffold_5 sequence, the exact location of the leftmost inversion breakpoint was not conclusive using this approach. A complementary approach was then used to facilitate detection of the leftmost inversion breakpoint in the D53 inversion line. Minimap2 was again used, but here D53 contig_531 was used as reference for the mapping of EgII male PacBio

reads as well as VIENNA 8^{D53-} male and VIENNA 7^{D53-} male Nanopore reads (Supplementary Fig. 10). Minimap2 parameters for PacBio reads were: minimap2 -x map-pb -A 1 -a --MD -L -t 40. Minimap2 parameters for Nanopore reads were: minimap2 -x map-ont -A 1 -a --MD -L -t 40. Samtools (v1.9, <https://github.com/samtools/samtools>) was used to convert the alignment.sam to.bam and prepare the alignment file to be viewed in the Integrative Genomics Viewer (IGV, <http://software.broadinstitute.org/software/igv/>). The expectation was to see a common breakpoint for all three of the above read set alignments when compared to the D53 genome in the area of the inversion. Position ~3,055,294 was identified in the D53 contig_531 as the most likely leftmost breakpoint. To determine the rightmost breakpoint, D53, VIENNA 8^{D53-} and VIENNA 7^{D53-} male nanopore reads were aligned on the EgII_scaffold_5 sequence. The expectation was to see a breakpoint in D53 read set alignments but not in VIENNA 7^{D53-} and VIENNA 8^{D53-}. This is the case here, since read alignments coming from both sides of the inversion are truncated at one position (Supplementary Fig. 9). Findings from genome version EgII_Ccap3.2 were extrapolated to the manually revised genome version EgII_Ccap3.2.1.

Predicted D53 inversion breakpoints were verified via PCRs on EgII, D53, and VIENNA 7^{D53+} GSS male and female genomic DNA, using PhusionFlash Polymerase in a 10 µL reaction volume [98 °C, 10 s; 30 cycles of (98 °C, 1 s; 56 °C, 5 s; 72 °C, 35 s); 72 °C, 1 min] (Supplementary Fig. 4). Sequences of all oligonucleotides used in this study are listed in Supplementary Table 5. The primer pair for the right breakpoint was designed based on EgII sequence information, primers for the left breakpoint were designed based on D53 sequence information. The wild-type status of chromosome 5 (EgII male and female, VIENNA 7^{D53+-} male) was amplified using primer pairs P_1794 and P_1798 (1950 bp) and P_1795 and P_1777 (690 bp). Chromosome 5 with the inversion (D53 male and female, VIENNA 7^{D53+-} male and VIENNA 7^{D53++} female) was verified using primer pairs P_1777 and P_1798 (1188 bp) and P_1794 and P_1795 (1152 bp) and amplicon sequencing (Macrogen Europe, Amsterdam).

Transcriptomic analysis of *C. capitata*, *B. dorsalis*, and *Z. cucurbitae* species were then conducted for RNA samples from 3rd instar larval and pre-pupal stages (Supplementary Table 1). Total RNA was extracted by homogenizing three larvae of *C. capitata* and *B. dorsalis* and a single larvae of *Z. cucurbitae* in liquid nitrogen, and then using the RNeasy Mini kit (Qiagen). Three replicates per strain and time point were performed. mRNA was isolated using the NEBNext polyA selection and the Ultra II directional RNA library preparation protocols from NEB and sequenced on the Illumina NovaSeq 6000 using dual indexes as 150 bp paired end reads (library insert 500 bp). Individual libraries were sequenced to provide >1 million paired end reads per sample. Each replicate was then assembled separately using Trinity⁵⁸ v2.8.5. The assembled transcripts from Trinity were mapped to the Ccap3.2 genome using minimap⁵⁹ (parameters -ax splice:hq -uf). The Illumina reads were mapped with STAR⁶⁰ v2.5.2.a. IGV⁶¹ v2.6 was used to view all data at a genomic and gene level. Given that the white pupae GSS^{12,62} was used to collect samples for RNA extraction from single larvae of *Z. cucurbitae*, larval sex was confirmed by a maleness-specific PCR on the *MoY* gene of *Z. cucurbitae*³⁵ using cDNA synthesized with the OneStep RT-PCR Kit (Qiagen) and the primer pair ZcMoY1F and ZcMoY1R amplifying a 214 bp fragment. Conditions for a 25 µL PCR reaction using the 1× Taq PCR Master Mix kit (Qiagen) were: [95 °C, 5 min; 30 cycles of (95 °C, 1 min; 51 °C, 1 min; 72 °C, 1 min); 72 °C, 10 min]. Presence of a PCR product indicated a male sample. Each, male and female sample was a pool of three individuals. Three replicates per strain and time point were collected.

Cytogenetic verification of D53 inversion and *wp* genes. Polytene chromosomes for in situ hybridization were prepared from third-instar larvae salivary glands⁶³. In brief, the glands were dissected in 45% acetic acid and placed on a coverslip in a drop of 3:2:1 solution (3 parts glacial acetic acid: 2 parts water: 1 part lactic acid) until been transparent (approximately 5 min). The coverslip was picked up with a clean slide. After squashing, the quality of the preparation was checked by phase contrast microscope. Satisfactory preparations were left to flatten overnight at -20 °C and dipped into liquid nitrogen until the bubbling stopped. The coverslip was immediately removed with razor blade and the slides were dehydrated in absolute ethanol, air dried, and kept at room temperature.

Probes were prepared by PCR. Single adult flies were used to extract DNA with the Extract me kit (Blirt SA), following the manufacturer's protocol. NanoDrop spectrometer was used to assess the quantity and quality of the extracted DNA which was then stored at -20 °C until used. Primers (P_1790/P_1791, P_1821/P_1822, Pgd_probe_F/R, vgl1_probe_F/R, Sxl_probe_F/R, y_probe_F/R, zw_probe_F/R, P_1633/P_1634, Zc_F/R, Bd_F/R, P_1395/P_1396, P_1415/P_1416) were designed for each targeted gene using the Geneious Prime software. PCR was performed in a 25 µL reaction volume using 12.5 µL PCR Master mix 2x Kit (Thermo Fisher Scientific), 60–80 ng DNA, and the following PCR settings [94 °C, 5 min; 35 cycles of (94 °C, 45 s; 56 °C, 30 s; 72 °C, 90 s); 72 °C, 1 min].

Probe labeling was carried out according to the Dig DNA Labelling Kit manual (Roche). Prior to in situ hybridization⁶⁴, stored chromosome preparations were hydrated by placing them for 2 min at each of the following ethanol solutions: 70%, 50%, and 30%. Then they were placed in 2× SSC at room temperature for 2 min. The stabilization of the chromosomes was done by placing them in 2× SSC at 65 °C for 30 min, denaturing in 0.07 M NaOH 2 min, washing in 2× SSC for 30 s, dehydrating (2 min in 30%, 50%, 70%, and 95% ethanol), and air drying. Hybridization was performed on the same day by adding 15 µL of denatured probe

(boiled for 10 min and ice-chilled). Slides were covered with a siliconized coverslip, sealed with rubber cement, and incubated at 45 °C overnight in a humid box. At the end of incubation, the coverslip was floated off in 2× SSC and the slide washed in 2× SSC for 3 × 20 min at 53 °C.

After 5 min wash in Buffer 1 (100 mM tris-HCl pH 7.5/ 1.5 M NaCl), the preparations were in Blocking solution (Blocking reagent 0.5% in Buffer 1) for 30 min, and then washed for 1 min in Buffer 1. The antibody mix was added to each slide and a coverslip was added. Then the slides were incubated in a humid box for 45 min at room temperature, following 2 × 15 min washes in Buffer 1, and a 2 min wash in detection buffer (100 mM Tris-HCl pH 9.5/ 100 mM NaCl). The color was developed with 1 mL of NBT/BCIP solution during a 40 min incubation in the dark at room temperature. The removal of the NBT/BCIP solution was done by rinsing in water twice. Hybridization sites were identified using 40× or 100× oil objectives (phase or bright field) and a Leica DM 2000 LED microscope, with reference to the salivary gland chromosome maps⁶⁵. Well-spread nuclei or isolated chromosomes were photographed using a digital camera (Leica DMC 5400) and the LAS X software 3.7.0. All in situ hybridizations were performed at least in duplicates and at least ten nuclei were analyzed per sample.

Gene editing and generation of homozygous *w⁺* strains. For CRISPR/Cas9 gene editing in *B. tryoni*, purified Cas9 protein (Alt-R Sp. Cas9 Nuclease V3, #1081058, 10 μg/μL) and guide RNAs (customized Alt-R CRISPR/Cas9 crRNA, 2 nmol and Alt-R CRISPR/Cas9 tracrRNA, #1072532, 5 nmol) were obtained from Integrated DNA Technologies (IDT). The guide RNAs were individually resuspended to a 100 μM stock solution with nuclease-free duplex buffer before use. The two customized 20 bp crRNA sequences (Bt_MFS-1 and Bt_MFS-2) were designed using CRISPOR⁶⁶. Injection mixes for microinjection of *B. tryoni* embryos comprise of 300 ng/μL Cas9 protein, 59 ng/μL of each individual crRNA, 222 ng/μL tracrRNA, and 1× injection buffer (0.1 mM sodium phosphate buffer pH 6.8, 5 mM KCl) in a final volume of 10 μL. The guide RNA complex containing the two crRNAs and tracrRNA was prepared by heating at 95 °C for 5 min before cooling to room temperature. The Cas9 enzyme along with the other injection mix components were then added to the guide RNA complex and incubated at room temperature for 5 min to assemble the ribonucleoprotein (RNP) complexes. Microinjections were performed in *B. tryoni* Ourimbah laboratory strain embryos that were collected over a 1 h time period. Injections were performed under paraffin oil using borosilicate capillary needles (#30-0038, Harvard Apparatus) drawn out on a Sutter P-87 flaming/brown micropipette puller and connected to an air-filled 20 mL syringe, a manual MM-3 micromanipulator (Narishige) and a CKX31-inverted microscope (Olympus). Microscope slides with the injected embryos were placed on agar in a Petri dish inside a vented container containing moist paper towels at 25 °C (± 2 °C). Hatched first instar larvae were removed from the oil and transferred to larval food. Individual *G₀* flies were crossed to six virgin flies from the Ourimbah laboratory strain and eggs were collected overnight for two consecutive weeks. *G₁* flies were then allowed to mate inter se and eggs were collected in the same manner. *G₂* pupae were then analyzed phenotypically and separated according to color of pupae (brown, mosaic, or white).

For *C. capitata* CRISPR/Cas9 gene editing, a guide RNA (gRNA_MFS), targeting the third CDS exon of *CcMFS* was designed and tested for potential off target effects using Geneious Prime⁵³ and the *C. capitata* genome annotation Ccap2.1¹⁶. In silico target site analysis predicted an on-target activity score of 0.615 (scores are between 0 and 1; higher score corresponds to higher expected activity⁶⁷) and zero off-targets sites in the medfly genome. gRNA_MFS was synthesized by in vitro transcription of linear double-stranded DNA template. Therefore, a linear DNA template was amplified in a 100 μL PCR reaction using primers P_1753 and P_369 and Q5 HF polymerase (NEB) according to the manufacturer's protocol (Bio-Rad C1000 Touch thermal cycler [98 °C, 30 s; 35 cycles of (98 °C, 10 s; 58 °C, 20 s; 72 °C, 20 s); 72 °C, 2 min]). The PCR reaction was purified using the Clean and Concentrator-25 kit. Subsequently, 500 ng were transcribed using the HiScribe T7 High Yield RNA Synthesis kit (NEB), followed by an DNase treatment (Invitrogen) and a final purification of the RNA using the Mega Clear Kit (Invitrogen). Injection mix for microinjection of embryos contained 360 ng/μL Cas9 protein (1 μg/μL, dissolved in its formulation buffer (PNA Bio Inc, CP01)), 200 ng/μL gRNA_MFS, and an end-concentration of 300 mM KCl^{68,69}. The mix was freshly prepared on ice followed by an incubation step for 10 min at 37 °C to allow pre-assembly of gRNA-Cas9 RNP complexes and stored on ice until use. Microinjections were conducted in WT EgII *C. capitata* embryos, collected over a 30–40 min period, chemically dechorionized (sodium hypochlorite, 3 min), fixed on double-sided sticky tape (Scotch 3 M), and covered with halocarbon oil 700 (Sigma-Aldrich). For injections, siliconized quartz glass needles (Q100-70-7.5; LOT171381; Science Products, Germany), drawn out on a laser-based micropipette puller (Sutter P-2000), were used with a manual micromanipulator (MN-151, Narishige), an Eppendorf FemtoJet 4i microinjector, and an Olympus SZX16 microscope (SDF PLAPO 1xPF objective). Injected embryos were placed in an oxygen chamber (max. 2 psi), first instar larvae were transferred from the oil to larval food.

As complementation assay, reciprocal crosses between surviving *G₀* adults and virgin adults of the *white pupae* strain #1402_22m1B (pBac_fa_attP-TREhs43-Ccra-I-hid^{Ala5}-SV40_a_Pub-nls-EGFP-SV40) (*w⁻(nat)*)²³ were set up either single paired (six cages) or in groups of seven to ten flies (seven cages). Eggs were

collected three times every 1–2 days. Progeny (*G₁*) exhibiting the white pupae phenotype (*w⁻(nat)*-(CRISPR)) were assayed via non-lethal genotyping and sorted according to mutation genotype (see Fig. 4). Genotypes 'A-H' were group-backcrossed to WT EgII (*w⁺*), genotype 'C' siblings mass-crossed. Eggs were collected four times every 1–2 days. Generation *G₂* flies were analyzed via multiplex PCR using three primers, specific for *w⁺* and *w⁻(CRISPR)* or *w⁻(nat)* allele size, respectively (see molecular analyses of *w⁺* mutants and mosaics, *C. capitata* non-lethal genotyping). Offspring of outcross cages showed brown pupae phenotype and either *w⁺*-(*w⁻(nat)*) or *w⁺*-(CRISPR) genotype. In order to make mutations A, D, and H homozygous, 40 flies (25 females, 15 males) were genotyped each, and *w⁺*-(CRISPR) positive flies were inbred (mutation A: 15 females, 7 males, mutation D: 12 females, 7 males, mutation H: 11 females, 8 males). *G₃* showing white pupae phenotype was homozygous for *w⁻(CRISPR)* mutations A, D, or H, respectively, and was used to establish lines. Inbreeding of mutation C *w⁻(nat)*-(CRISPR) flies produced only white pupae offspring, based on either the *w⁻(nat)*-(*w⁻(nat)*), *w⁻(nat)*-(CRISPR), or *w⁻(CRISPR)*-(CRISPR) genotype. 94 flies (46 females, 48 males) were genotyped, homozygous *w⁻(CRISPR)* were inbred to establish a line (13 females, 8 males).

Molecular analyses of *w⁺* mutants and mosaics. In *B. tryoni*, genomic DNA was isolated for genotyping from *G₂* pupae using the DNeasy Blood and Tissue Kit (Qiagen). PCR amplicons spanning both BtMFS guide recognition sites were generated using Q5 polymerase (NEB) with primers BtMFS_5primeF and BtMFS_exon2R. Products were purified using MinElute PCR Purification Kit (Qiagen), ligated into pGEM-t-easy vector (Promega) and transformed into DH5α cells. Plasmids were purified with Wizard Plus SV Miniprep (Promega) and sequenced.

In *C. capitata*, non-lethal genotyping was performed to identify parental genotypes before setting up crosses. Therefore, genomic DNA was extracted from single legs of *G₁* and *G₂* flies following an adapted version of an established protocol⁷⁰. Single legs of anesthetized flies were cut at the proximal femur, placed in vials containing ceramic beads and 50 μL buffer (10 mM Tris-Cl, pH 8.2, 1 mM EDTA, 25 mM NaCl), and homogenized for 15 s (6 m/s) using a FastPrep-24 5 G homogenizer. Then, 28.3 μL buffer and 1.7 μL proteinase-K (2.5 U/mg) were added. The reaction mix was incubated for 1 h at 37 °C, followed by 4 min at 98 °C, and subsequently cooled down on ice and used for PCR. For *G₁* flies, PCR on *w⁺* was performed in a 25 μL reaction volume using the DreamTaq polymerase, primers P_1643 and P_1644, and 3.75 μL reaction mix, whereby different amplicon sizes were expected for different alleles (*w⁺* and *w⁻(CRISPR)*: 724 bp, *w⁻(nat)*: 8872 bp). The *w⁻(nat)* amplicon was excluded via PCR settings [95 °C, 3 min; 35 cycles of (95 °C, 30 s; 56 °C, 30 s; 72 °C, 1 min); 72 °C, 5 min]. The 724 bp PCR product was verified by gel electrophoresis and purified from the PCR reaction using the DNA Clean & Concentrator-5 kit. PCR products were sequenced (P_1644) and analyzed using Geneious Prime⁵³. In generation *G₂*, flies were analyzed using multiplex PCR with primers P_1657, P_1643, and P_1644, to distinguish between the *w⁻(nat)* (457 bp; P_1643/P_1657), and *w⁻(CRISPR)* alleles (724 bp; P_1643/P_1644) using the above-described PCR protocol.

Image acquisition. Images of *B. tryoni* pupae were taken with an Olympus SZX16 microscope, Olympus DP74 camera, and Olympus LF-PS2 light source using the Olympus stream basic 2.3.3 software. Images of *C. capitata* pupae were taken with a Keyence digital microscope VHX-5000. Image processing was conducted with Adobe Photoshop CS5.1 software to apply moderate changes to image brightness and contrast. Changes were applied across the entire image.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this article is available as a Supplementary Information file. The datasets and insect strains generated and analyzed during the current study are available from the corresponding authors upon request. All sequences generated in this study from *B. dorsalis*, *B. tryoni*, *Bactrocera* introgressed line (*BIL*), *C. capitata*, and *Z. cucurbitae* samples are publicly available on NCBI within the ENA BioProject PRJEB36344 (for Ccap genome assembly EgII-3.2.1, WGS, PacBio, chromosome dissections, Illumina MiSeq, Illumina HiSeq 4000, RNaseq, Illumina NovaSeq 6000, Hi-C, and Nanopore data; see Supplementary Table 1 for detailed sample designation), BioProject PRJNA629430 (for WGS and Illumina DNaseq 2 × 250 PE data; see Supplementary Fig. 6 for detailed sample designation), and BioProject PRJNA682907 (for WGS and Illumina NovaSeq 6000 data; see Supplementary Table 1 for detailed sample designation). The source data underlying Figs. 1, 2c–f, 3d–f, and 4d, as well as Supplementary Figs. 3a–b, 4a–b, 4d–e, and 7 are provided as a Source Data file. Source data are provided with this paper.

Received: 16 June 2020; Accepted: 14 December 2020;

Published online: 21 January 2021

References

- Robinson, A. S. & Hooper, G. *Fruit Flies: Their Biology, Natural Enemies, and Control* Vol. 1 (Elsevier, 1989).
- Suckling, D. M. et al. Eradication of tephritid fruit fly pest populations: outcomes and prospects. *Pest Manag. Sci.* **72**, 456–465 (2016).
- Dyck, V. A. et al. *Sterile Insect Technique – Principles and Practice in Area-Wide Integrated Pest Management* (eds Dyck, V. A., Hendrichs, J. & Robinson, A. S.) (Springer, 2005).
- Vreysen, M., Robinson, A. S. & Hendrichs, J. *Area-Wide Control of Insect Pests: from Research to Field Implementation* (Springer, 2007).
- Rendon, P., McInnis, D., Lance, D. & Stewart, J. Medfly (Diptera: Tephritidae) genetic sexing: large-scale field comparison of males-only and bisexual sterile fly releases in Guatemala. *J. Econ. Entomol.* **97**, 1547–1553 (2004).
- Franz, G. *Sterile Insect Technique – Principles and Practice in Area-Wide Integrated Pest Management* (eds Dyck, V. A., Hendrichs, J. & Robinson, A. S.) (Springer, 2005).
- Augustinos, A. A. et al. *Ceratitis capitata* genetic sexing strains: laboratory evaluation of strains from mass-rearing facilities worldwide. *Entomol. Exp. Appl.* **164**, 305–317 (2017).
- Zacharopoulou, A. et al. A review of more than 30 years of cytogenetic studies of Tephritidae in support of sterile insect technique and global trade. *Entomol. Exp. Appl.* **164**, 204–225 (2017).
- Rössler, Y. The genetics of the Mediterranean fruit fly: a “white pupae” mutant. *Ann. Entomol. Soc. Am.* **72**, 583–585 (1979).
- Rössler, Y. & Koltin, Y. The genetics of the Mediterranean fruit fly, *Ceratitis capitata*: three morphological mutations. *Ann. Entomol. Soc. Am.* **69**, 604–608 (1976).
- McCombs, S. D. & Saul, S. H. Linkage analysis of five new genetic markers of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae). *J. Hered.* **83**, 199–203 (1992).
- McInnis, D. O. et al. Development of a pupal color-based genetic sexing strain of the melon fly, *Bactrocera cucurbitae* (Coquillett) (Diptera: Tephritidae). *Ann. Entomol. Soc. Am.* **97**, 1026–1033 (2004).
- Wappner, P. et al. White pupa: a *Ceratitis capitata* mutant lacking catecholamines for tanning the puparium. *Insect Biochem. Molec. Biol.* **25**, 365–373 (1995).
- Rössler, Y. & Rosenthal, H. Genetics of the mediterranean fruit fly (Diptera: Tephritidae): morphological mutants on chromosome five. *Ann. Entomol. Soc. Am.* **85**, 525–531 (1992).
- Kerremans, P. & Franz, G. Cytogenetic analysis of chromosome 5 from the Mediterranean fruit fly, *Ceratitis capitata*. *Chromosoma* **103**, 142–146 (1994).
- Papanicolaou, A. et al. The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biol.* **17**, 192 (2016).
- Papathanos, P. A. et al. A perspective on the need and current status of efficient sex separation methods for mosquito genetic control. *Parasites Vectors* **11**, 654 (2018).
- Sim, S. B. & Geib, S. M. A chromosome-scale assembly of the *Bactrocera cucurbitae* genome provides insight to the genetic basis of white pupae. *G3* **7**, 1927–1940 (2017).
- Sim, S. B., Ruiz-Arce, R., Barr, N. B. & Geib, S. M. A new diagnostic resource for *Ceratitis capitata* strain identification based on QTL mapping. *G3* **7**, 3637–3647 (2017).
- Zdobnov, E. M. et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45**, D744–D749 (2017).
- San Jose, M. et al. Incongruence between molecules and morphology: a seven-gene phylogeny of *Dacini* fruit flies paves the way for reclassification (Diptera: Tephritidae). *Mol. Phylog. Evol.* **121**, 139–149 (2018).
- Choo, A., Crisp, P., Saint, R., O’Keefe, L. V. & Baxter, S. W. CRISPR/Cas9-mediated mutagenesis of the white gene in the tephritid pest *Bactrocera tryoni*. *J. Appl. Entomol.* **142**, 52–58 (2018).
- Ogaugwu, C. E., Schetelig, M. F. & Wimmer, E. A. Transgenic sexing system for *Ceratitis capitata* (Diptera: Tephritidae) based on female-specific embryonic lethality. *Insect Biochem. Mol. Biol.* **43**, 1–8 (2013).
- Davis, A. W. et al. Rescue of hybrid sterility in crosses between *D. melanogaster* and *D. simulans*. *Nature* **380**, 157–159 (1996).
- Araripe, L. O., Montenegro, H., Lemos, B. & Hartl, D. L. Fine-scale genetic mapping of a hybrid sterility factor between *Drosophila simulans* and *D. mauritiana*: the varied and elusive functions of “speciation genes”. *BMC Evol. Biol.* **10**, 385 (2010).
- Brideau, N. J. & Barbash, D. A. Functional conservation of the *Drosophila* hybrid incompatibility gene Lhr. *BMC Evol. Biol.* **11**, 57 (2011).
- Kotov, A. A. et al. piRNA silencing contributes to interspecies hybrid sterility and reproductive isolation in *Drosophila melanogaster*. *Nucleic Acids Res.* **47**, 4255–4271 (2019).
- Barbash, D. A. Ninety years of *Drosophila melanogaster* hybrids. *Genetics* **186**, 1–8 (2010).
- Bedo, D. G. & Zacharopoulou, A. Inter-tissue variability of polytene chromosome banding patterns. *Trends Genet.* **4**, 90–91 (1988).
- Kriventseva, E. V. et al. 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 (2019).
- Zhao, Y. et al. A major facilitator superfamily protein participates in the reddish brown pigmentation in *Bombyx mori*. *J. Insect Physiol.* **58**, 1397–1405 (2012).
- The modEncode Consortium et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
- Wright, T. R. The genetics of biogenic amine metabolism, sclerotization, and melanization in *Drosophila melanogaster*. *Adv. Genet.* **24**, 127–222 (1987).
- Bourtzis, K., Psachoulia, C. & Marmaras, V. J. Evidence that different integumental phosphatases exist during development in the Mediterranean fruit fly *Ceratitis capitata*: possible involvement in pupariation. *Comp. Biochem. Physiol. Part B* **98**, 411–416 (1991).
- Meccariello, A. et al. *Maleness-on-the-Y (MoY)* orchestrates male sex determination in major agricultural fruit fly pests. *Science* **365**, 1457–1460 (2019).
- Hall, A. B. et al. Sex determination. A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**, 1268–1270 (2015).
- Liu, P. et al. *Nix* is a male-determining factor in the Asian tiger mosquito *Aedes albopictus*. *Insect Biochem. Mol. Biol.* **118**, 103311 (2019).
- Ward, C. M. & Pederson, T. H. S. M. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. *Bioinformatics* **36**, 2587–2588 (2020).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
- Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907* (2012).
- Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
- Zheng, X. et al. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).
- Ward, C. M., Ludington, A. J., Breen, J. & Baxter, S. W. Genomic evolutionary analysis in R with gear. <https://doi.org/10.1101/2020.08.06.240754> (2020).
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
- Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
- Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438 (2017).
- Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
- Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
- Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
- Green, M. R. & Sambrook, J. Isolation of High-Molecular-Weight DNA using organic solvents. *Cold Spring Harb. Protoc.* **2017**, pdb.prot093450 (2017).
- Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
- Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
- Ghuray, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).
- Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
- Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

61. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
62. Zacharopoulou, A. & Franz, G. Genetic and cytogenetic characterization of genetic sexing strains of *Bactrocera dorsalis* and *Bactrocera cucurbitae* (Diptera: Tephritidae). *J. Econ. Entomol.* **106**, 995–1003 (2013).
63. Zacharopoulou, A. et al. The genome of the Mediterranean fruit fly *Ceratitidis capitata*: localization of molecular markers by in situ hybridization to salivary gland polytene chromosomes. *Chromosoma* **101**, 448–455 (1992).
64. Mavragani-Tsipidou, P. et al. *Protocols for Cytogenetic Mapping of Arthropod Genomes* (ed Sakharov, I.) (CRC Press, Taylor and Francis Group, LLC, 2014).
65. Zacharopoulou, A. Polytene chromosome maps in the medfly *Ceratitidis capitata*. *Genome* **33**, 184–197 (1990).
66. Concordet, J. P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
67. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
68. Aumann, R. A., Schetelig, M. F. & Häcker, I. Highly efficient genome editing by homology-directed repair using Cas9 protein in *Ceratitidis capitata*. *Insect Biochem. Mol. Biol.* **101**, 85–93 (2018).
69. Burger, A. et al. Maximizing mutagenesis with solubilized CRISPR-Cas9 ribonucleoprotein complexes. *Development* **143**, 2025–2037 (2016).
70. Carvalho, G. B., Ja, W. W. & Benzer, S. Non-lethal PCR genotyping of single *Drosophila*. *Biotechniques* **46**, 312–314 (2009).
71. Gariou-Papalexiou, A. et al. Polytene chromosomes as tools in the genetic analysis of the Mediterranean fruit fly, *Ceratitidis capitata*. *Genetica* **116**, 59–71 (2002).

Acknowledgements

This study was financially supported by the Joint FAO/IAEA Insect Pest Control Sub-programme of the Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture, the Emmy Noether program of the German Research Foundation (SCHE 1833/1-1; to MFS), the LOEWE Center for Insect Biotechnology and Bioresources of the Hessian State Ministry for Higher Education, Research and the Arts (HMWK; to MFS), and the SITplus collaborative fruit fly program funded by the Hort Frontiers Fruit Fly Fund, part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from Macquarie University, USDA, and JLU Gießen and contributions from the Australian Government (FF17000 to MFS). The project was further supported by Hort Innovation, using the Apple & Pear, Strawberry, Citrus, Cherry, Summerfruit, Table Grape and Vegetable research and development levies (MT13059 to PC, AC, EF), with co-investment from South Australian Research and Development Institute (SARDI) and Primary Industries and Regions South Australia (PIRSA) and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture. SWB was supported by the Australian Research Council (FT140101303) and the Hermon Slade Foundation grant HSF 18/6. Furthermore, this work was supported by the Canadian Foundation for Innovation (33408, to JR) and Genome Canada Genome Technology Platform awards (JR), as well as the International Atomic Energy Agency research contracts no. 23358 (JR) and no. 23379 (FM) as part of the Coordinated Research Project “Generic approach for the development of genetic sexing strains for SIT applications”. The project benefitted from discussions at this CRP. Furthermore, resources were provided by the SCINet project of the USDA-ARS, the Detection, Control, and Area-wide Management of Fruit Flies and Other Quarantine Pests of Tropical/Subtropical Crops, and a USDA-NIFA grant (0500-00093-001-00-D, 2040-22430-026-00-D, and 2017-67012-26087 to SBS and SMG). The USDA-ARS is an equal opportunity/affirmative action employer, and all agency services are available without discrimination. Mention of commercial products and organizations in this manuscript is

solely to provide specific information. It does not constitute endorsement by USDA-ARS over other products and organizations not mentioned. The authors also wish to thank Tanja Rehling, Jakob Martin, and Johanna Rühl for technical assistance and Germano Sollazzo for input on injections and primers design (Justus-Liebig University Gießen and Insect Pest Control Laboratory); Elena Isabel Cancio Martinez, Thilakasiri Dammalage, Sohel Ahmad, and Gülizar Pillwax for insect rearing (Insect Pest Control Laboratory), Shu-Huang Chen (McGill University) for technical assistance with Nanopore library preparations; and Arjen van ’t Hof (University of Liverpool) for constructing libraries from micro-dissected chromosomes.

Author contributions

R.A.A., C.M.W., C.C., P.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. designed the research; C.M.W., R.A.A., M.A.W., K.N., G.G., E.F., S.J.R., M.A.H., C.C., T.N.M.N., A.C., S.B.S., S.M.G., A.C.D., K.B., S.W.B., and M.F.S. performed the research; R.A.A., C.M.W., H.D., G.L., F.M., J.R., K.B., S.W.B., and M.F.S. contributed new reagents/analytic tools; C.M.W., R.A.A., M.A.W., K.N., G.L., G.G., H.D., S.W., T.N.M.N., A.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. analyzed the data; R.A.A., C.M.W., K.N., G.L., G.G., S.J.R., S.W., A.C., S.B.S., S.M.G., I.H., J.R., A.C.D., K.B., S.W.B., and M.F.S. wrote the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-20680-5>.

Correspondence and requests for materials should be addressed to K.B., S.W.B. or M.F.S.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

Chapter 9

Discussion and closing words

1 In this thesis I sought to identify and understand the evolution of agriculturally important trait
2 loci in two insect pest species, the diamondback moth (*Plutella xylostella*, L.) and the
3 Queensland fruit-fly (*Bactrocera tryoni*, Froggatt). This involved developing two software
4 packages for the analysis of next generation sequence data (Chapters 2 & 3). These packages
5 then aided me to probe whole genome resequencing data for introgressed loci (Chapters 4
6 and 8), model the evolutionary history of a major insect pest (Chapter 4 & 5), assemble a
7 chromosome level reference genome for the diamondback moth (Chapter 6) and finally use a
8 multi-omic approach to investigate the mechanisms underlying host range expansion in a
9 major insect pest species (Chapter 7). Although these projects appear disparate in their
10 conclusions, they weave together the narrative of my PhD through their methodological
11 resemblance and shared research areas.

12 This final Chapter narrates the course of my PhD research in the context of each aim
13 presented in Chapter 1 and seeks to outline the broader narrative of my PhD research by
14 summarizing its contributions to the scientific community and literature as a whole.

15

16 **Aim 1: Develop computational packages to interpret high-throughput sequence**
17 **libraries metrics and perform memory efficient evolutionary analysis among**
18 **large genomic datasets.**

19 Sequencing hundreds of genomes or transcriptomes has become routinely feasible for
20 research projects, however, the ability to rapidly identify bias or contamination within raw data
21 of individual libraries has remained a laborious process. Initial assessment of quality control
22 logs generated from fastq sequence output files by programs such as fastqc
23 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) provides insight into systematic
24 bias and provides useful metrics to determine overall success, yet, the pipeline is
25 unmanageable when sample sizes are very large. This drove the development of ngsReports
26 whereby researchers are able to create customized quality control summaries according to
27 their needs in the R programming language. This was incredibly powerful for my PhD research
28 which, in some cases, relied on sequencing datasets with 100s of samples and was able to
29 identify if poor mapping rates were due to divergence or sequence quality when assessing
30 outgroup species mapped to an ingroup reference. Furthermore, ngsReports is being utilized
31 by the research community and has, at the time of writing, been installed ~3000 times
32 (<http://www.bioconductor.org/packages/release/bioc/html/ngsReports.html>), placing it in the
33 top ~40% of all R packages in the Bioconductor repository. Current and future work on the
34 ngsReports package aims to implement log files from other bioinformatics software platforms.

35 During the course of my PhD, I carried out computational analyses on multiple genetic
36 datasets that generally required file format conversion. However, this process was not lossless
37 and required custom scripts to be written for each analysis. This prompted me to develop
38 general purpose functions to carry out population genetic statistics on the VCF file format. In
39 their first iteration, functions were combined into an R package, **windowed evolutionary**
40 **analysis and visualization in R (weavR)**, <https://github.com/CMWbio/weavR>). However, due to
41 the large file size associated with VCF files the analysis was memory intensive compared to
42 other available tools (Pfeifer et al. 2014). Around this time, the SeqArray package (Zheng et
43 al. 2017) proposed a reworked structure for genotype data that was disk size efficient and had
44 a lowered memory footprint than traditional VCF files. This provided the opportunity to
45 implement the Genomic Data Structure (GDS) format into the R package I was developing.
46 The GDS format was highly applicable to population genetic tasks as it enabled direct querying
47 from GenomicRanges objects (Lawrence et al. 2013) and C++ scripting allowed for lightning-
48 fast querying of genotype arrays from the GDS file on disk, without the need to read the entire
49 locus into memory first. GenomicRange querying also provided a simple solution to another
50 shortcoming of many other population genetic tools - a constricted analysis space. Many tools
51 only allow genome-wide windowed analysis to be performed, making it difficult to select a
52 particular subset of sites to be analyzed (such as a specific set of genes or codon positions).
53 This prompted me to write a set of helper functions to select specific regions of the genome
54 for analysis from user provided tab-separated or GFF files and to combine multiple
55 GenomicRange objects into highly customizable ranges for analysis.

56 Functions in weavR were reworked or completely replaced to utilize the GDS format which
57 facilitated the incorporation of object-oriented programming concepts I learned through
58 development on the ngsReports package (Chapter 2). This converted the previously function
59 based weavR package into a modular, object-oriented R package that allowed users to carry
60 out a diverse array of functionality with a single object type – the GEAR. From weavR the R
61 package **genomic evolutionary analysis in R (gear)** was born and with each new analytical
62 challenge I encountered a new functionality was implemented (Chapter 3). This resulted in
63 gear being used in some form across the majority of the Chapters of this thesis. The current
64 and future goals of gear are to implement more population genetic metrics such as neutrality
65 statistics and allele frequency spectrum calculations.

66

67 **Aim 2: Investigate the extent of hybridization, admixture and introgression**
68 **among sympatric pest species and interspecific crosses.**

69 Introgression of insecticide resistance and other pre-adapted alleles from invasive pest
70 species into native non-pest species is an underestimated threat to agricultural and is an
71 expanding research front (Mesgaran et al. 2016, Bay et al. 2019, Valencia-Montoya et al.
72 2020). A potential introgression risk was identified with the discovery that invasive
73 diamondback moth, a major crop pest, is able to hybridize with its native ally, *Plutella*
74 *australiana*, under laboratory conditions (Perry et al. 2018). Genomic scans for introgression
75 between the two species using the F_3 statistic revealed no evidence for widespread gene flow
76 (Chapter 4). Yet, when applying F_3 to simulated datasets under different admixture scenarios
77 it failed to identify introgression between similarly divergent simulated species unless
78 admixture was recent and substantial. This prompted me to develop a novel metric that utilizes
79 phylogenetic branch lengths to identify introgressed loci. Applying this new metric to simulated
80 datasets showed that it was far more sensitive to low levels of introgression than the F_3 .
81 However, after re-investigating the genomes of diamondback moth and *P. australiana*, I found
82 no support for introgression between the two species, suggesting they display high levels of
83 assortative mating in the wild. This provided an interesting contrast to a growing body of
84 literature that suggests hybridization is common in insects (Harrison and Larson 2014) and
85 provided evidence that introgression of beneficial (eg. insecticide resistance) alleles from
86 diamondback moth is unlikely to occur in the field.

87 Utilizing the skills I developed to search for introgressed loci in the diamondback moth, I then
88 investigated resequencing datasets of an interspecific cross between the Queensland fruit-fly
89 and the white pupa strain of the Oriental fruit-fly (Chapter 8). The white pupa phenotype has
90 naturally arisen in three different fruit fly species, allowing it to be used as marker to sort males
91 from females without the need to know its genetic mechanism (Rendón et al. 2004,
92 Aketarawong et al. 2020). Yet, it remained undetected among Queensland fruit-fly cultures,
93 limiting efficiency of sterile insect technique applications in Australia. Although transference of
94 the white pupa phenotype present in the Oriental fruit-fly into a Queensland fruit-fly
95 background posed an interesting solution to the lack of a naturally occurring white pupa
96 mutant, import restrictions prevented introduction of the hybrid strain for use in Australia.

97 Using *gear*, I carried out a genome wide scans for introgressed loci, identifying a *Major*
98 *facilitator superfamily* gene that was known to be involved in *B. mori* larvae pigmentation (Zhao
99 et al. 2012) and was highly conserved among insects (Chapter 8) making it a powerful
100 candidate for further research. An international effort then resulted in independent discovery
101 of the white pupa gene in both Mediterranean fruit-fly and Melon fruit-fly, identifying the same
102 causal gene. CRISPR/cas9 mediated knock out of the MFS *white pupa* gene confirmed its
103 role as sufficient to cause the white pupa phenotype in Queensland fruit-fly. Current work is

104 ongoing to generate a *white pupa* based genetic sexing strain for the Queensland fruit-fly, and
105 other dipteran pest species, for sterile insect technique applications.

106

107 **Aim 3: Improve evolutionary and genomic resources for the major Brassica pest**
108 ***Plutella xylostella* L.**

109 Although the diamondback moth is the major insect pest of *Brassica* crop plants,
110 comparatively little is known about its closest relatives *Plutella australiana*, *P. porrectella* and
111 *P. armoraciae* or their evolutionary history. Expansion of the evolutionary history
112 reconstructed in Chapter 4 was carried out by sequencing the whole genome of *P. porrectella*,
113 *P. armoraciae* and an outgroup *Acrolepiosis assectella*. I then carried out assembly of the
114 mitochondrial genome for each newly sequenced moth providing a valuable resource to future
115 studies of these poorly understood *Plutella* species.

116 Fitting a molecular clock to the *Plutella* mitochondrial phylogeny raised an interesting problem
117 (Chapter 4 & 5): How did *Plutella* come to inhabit a large geographic range over such a short
118 time span (~4 million years)? Recent work has suggested a South American origin for
119 diamondback moth (Appendix A: paper 2), yet *P. australiana* appears to be native to Australia
120 (Landry and Hebert 2013), *P. armoraciae* to North America (Landry and Hebert 2013) and *P.*
121 *porrectella* to Europe (Smith and Sears 1984). Yet, phylogenetic dating of the *Plutella* crown
122 node suggested the worldwide dispersal of the *Plutella* genus is much more recent than
123 previously thought. Species descriptions for all known *Plutella* species (>1758) occur after
124 colonization of and subsequent trade with colonial America (>1492). Although this may simply
125 be due to formalization of binominal nomenclature in the late 18th century by Linnaeus (whom
126 first described diamondback moth), a more controversial reading presents an out of America
127 hypothesis for the genus as a whole. Further phylogenetic reconstruction employing more
128 species sampled from across the globe will be required in future studies to disentangle *Plutella*
129 moth's complex evolutionary history.

130 The lack of field introgression between diamondback moth and *P. australiana* identified in
131 Chapter 4, coupled with their ability to hybridize under laboratory conditions (Appendix A:
132 paper 4) provided a unique opportunity to carry out genome assembly through a relatively new
133 approach - trio-binning. By hybridizing diamondback moth and *P. australiana* under laboratory
134 conditions and sequencing a single pupa, I successfully assembled a chromosome level
135 reference genome from a single haplotype of the diamondback moth. Though this method has
136 been used on mammals (Koren et al. 2018), this represented the second such example in
137 invertebrates (Yen et al. 2020) and to my knowledge the first example employing two different

138 invertebrate species. This new reference genome constitutes one of the most high-quality
139 Lepidopteran assemblies to date and greatly improves upon the previous diamondback moth
140 reference genome (You et al. 2013) in both completeness and overall contiguity (Chapter 6).

141 Whole genome sequencing of a diamide resistant diamondback moth, originating from the
142 Lockyer Valley in Queensland, enabled identification of a ryanodine receptor mutation that is
143 known to cause chlorantraniliprole resistance (Jouraku et al. 2020). Discovery of this variant
144 in Australia is expected to have an impact on the way in which group 28 diamide chemistries
145 are managed in agricultural settings, and likely to create significant interest among agronomic
146 sectors. Future work will focus on assembly and annotation of the *P. australiana* haplotype
147 along with comparative analysis against the diamondback moth reference presented in
148 Chapter 6. This has the potential to reveal gene expansion events and genomic evolution
149 specific to the diamondback moth lineage which may have been involved in its evolution into
150 a major insect pest.

151

152 **Aim 4. Analysis of a complex host plant range expansion in *P. xylostella***

153 The diamondback moth's host plant range expansion to include legumes posed a troubling
154 problem for agriculture. Although initial investigation into the genetic mechanism of the DBM-
155 P strain revealed chromosomes that segregated with survival on *Pisum* (Henniges-Janssen
156 et al. 2011), important questions still remained. How many mutations control the phenotype?
157 Have they become fixed in the population in the >100 generations since Henniges-Janssen et
158 al. (2011)? What effect does *Pisum* feeding have on gene expression? Chapter 7 sought to
159 answer these questions using a multi-omic approach, sequencing both reduced
160 representation genome and poly-A selected RNA libraries.

161 By sequencing both backcross and intercross pedigrees I was able to resolve high density
162 linkage maps for the 31 chromosomes in the PxLV.1 reference genome (Chapter 6). It is worth
163 noting that initial analysis and linkage map construction was carried out on the You et al.
164 (2013) genome, however, little concordance was identified between physical genomic position
165 and genetic position. The observed discordance was the main motivator to assemble a new
166 reference genome for *P. xylostella*, in order to assure robustness in the DBM-P analysis. In
167 contrast, co-linearity between the linkage maps and PxLV.1 genomic position was high,
168 providing further support for the robustness of the reference genome assembled in Chapter 6
169 and that pedigrees could be used for QTL analysis.

170 I then carried out LOD analysis across the genome, revealing multiple regions of the genome
171 linked with the 'survival on *Pisum*' phenotype. However, QTL varied between crosses and only

172 one was overlapping with a chromosome identified by Henniges-Janssen et al. (2011). This
173 suggested the adaptation is highly complex with multiple allele combinations conferring the
174 phenotype. One limitation of this work was the relatively small sample size. Larger sample
175 sizes could potentially increasing signal for QTL against the genomic background, improving
176 identification of putative QTL. In Chapter 7, this may have been achieved by utilizing a single
177 cross, however multiple crosses were employed due to the adaptations unfixed nature
178 described by Henniges-Janssen et al. (2011). Another limitation of the study was the use of
179 the diamondback moth Waite strain, which unexpectedly showed low levels of survivorship on
180 a *Pisum* host plant. This probably enabled DBM-W alleles to persist among cross progeny
181 challenged with pea feeding, potentially preventing DBM-P allele abundance from reaching
182 significance at adaptive loci. The DBM-W stain was initially selected due to its inability to
183 survive on *Pisum*, however I observed low levels of survivorship suggesting environmental
184 conditions may impact host plant plasticity in the diamondback moth. Future work should aim
185 to utilize multiple different diamondback moth strains, preferably with low innate survivorship
186 on *Pisum*, to decrease the effect of phenotypic plasticity. Though this may not be simple due
187 to the seemingly widespread nature of host plant phenotypic plasticity in many diamondback
188 moth strains (Gupta and Thorsteinson 1960, Yang et al. 2020).

189 Due to the complex genetic nature of the phenotype, I then investigated transcriptomic
190 response due to host plant. No choice feeding assays utilizing the DBM-P strain identified a
191 broad expressional response that spanned both the phase I and phase II metabolisms in larval
192 midgut tissue, coupled with olfactory and gustatory transcriptomic changes in head capsules
193 (Chapter 8). One of the main strengths of this work was its utilization of specific organs for
194 analysis. This allowed identification of differential expression in lowly expressed genes, such
195 as gustatory receptors (Park and Kwon 2011), which may have been masked if whole larvae
196 were sequenced. Future work may wish to employ further dissection of tissues, especially to
197 understand if differential expression of brain specific signaling pathways play a role in the
198 adaptation separate from larval taste reception.

199

200 **Final words**

201 Overall, these manuscripts describe major advancements in our understanding of the *Plutella*
202 genus and the resources available for future *Plutella* genomic studies. The results outlined in
203 Chapters 4-7 further our understanding of how one of the foremost insect pest species has
204 evolved, whereas Chapter 8 provides a simple genetic mechanism for improvement of Sterile
205 Insect Technique programs in fruit-flies and beyond. In addition, bioinformatics tools
206 developed in Chapter 2-3 represent a significant advancement in the techniques available of

207 quality control and evolutionary analysis of next generation sequencing data. Taken together,
208 the entirety of this work constitutes a major contribution towards understanding the complexity
209 of host plant adaptation in diamondback moth and how it has diverged from its non-pest
210 relatives.

211

212 **References**

- 213 Aketarawong, N., S. Isasawin, K. Laohakieat, and S. Thanaphum. 2020. Genetic stability,
214 genetic variation, and fitness performance of the genetic sexing Salaya1 strain for
215 *Bactrocera dorsalis*, under long-term mass rearing conditions. *BMC Genetics* **21**:131.
- 216 Bay, R. A., E. B. Taylor, and D. Schluter. 2019. Parallel introgression and selection on
217 introduced alleles in a native species. *Molecular Ecology* **28**:2802-2813.
- 218 Gupta, P., and A. Thorsteinson. 1960. Food plant relationships of the diamond-back moth
219 (*Plutella maculipennis* (curt.)): I. Gustation and Olfaction in Relation to Botanical
220 Specificity of the Larva. *Entomologia experimentalis et applicata* **3**:241-250.
- 221 Harrison, R. G., and E. L. Larson. 2014. Hybridization, Introgression, and the Nature of
222 Species Boundaries. *Journal of Heredity* **105**:795-809.
- 223 Henniges-Janssen, K., A. Reineke, D. G. Heckel, and A. T. Groot. 2011. Complex inheritance
224 of larval adaptation in *Plutella xylostella* to a novel host plant. *Heredity* **107**:421-432.
- 225 Jouraku, A., S. Kuwazaki, K. Miyamoto, M. Uchiyama, T. Kurokawa, E. Mori, M. X. Mori, Y.
226 Mori, and S. Sonoda. 2020. Ryanodine receptor mutations (G4946E and I4790K)
227 differentially responsible for diamide insecticide resistance in diamondback moth,
228 *Plutella xylostella* L. *Insect Biochemistry and Molecular Biology* **118**:103308.
- 229 Koren, S., A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J.
230 L. Williams, T. P. L. Smith, and A. M. Phillippy. 2018. De novo assembly of haplotype-
231 resolved genomes with trio binning. *Nature Biotechnology* **36**:1174-1182.
- 232 Landry, J.-F., and P. D. Hebert. 2013. *Plutella australiana* (Lepidoptera, Plutellidae), an
233 overlooked diamondback moth revealed by DNA barcodes. *ZooKeys*:43.
- 234 Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan,
235 and V. J. Carey. 2013. Software for Computing and Annotating Genomic Ranges.
236 *PLOS Computational Biology* **9**:e1003118.
- 237 Mesgaran, M. B., M. A. Lewis, P. K. Ades, K. Donohue, S. Ohadi, C. Li, and R. D. Cousens.
238 2016. Hybridization can facilitate species invasions, even without enhancing local
239 adaptation. *Proceedings of the National Academy of Sciences* **113**:10210-10214.

240 Park, J.-H., and J. Y. Kwon. 2011. A systematic analysis of *Drosophila* gustatory receptor
241 gene expression in abdominal neurons which project to the central nervous system.
242 *Molecules and cells* **32**:375.

243 Perry, K. D., G. J. Baker, K. J. Powis, J. K. Kent, C. M. Ward, and S. W. Baxter. 2018. Cryptic
244 *Plutella* species show deep divergence despite the capacity to hybridize. *BMC*
245 *evolutionary biology* **18**:1-17.

246 Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher. 2014. PopGenome: An
247 Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology*
248 *and Evolution* **31**:1929-1936.

249 Rendón, P., D. McInnis, D. Lance, and J. Stewart. 2004. Medfly (Diptera: Tephritidae) genetic
250 sexing: large-scale field comparison of males-only and bisexual sterile fly releases in
251 Guatemala. *Journal of economic entomology* **97**:1547-1553.

252 Smith, D., and M. Sears. 1984. Life history of *Plutella porrectella*, a relative of the
253 Diamondback Moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *The Canadian*
254 *Entomologist* **116**:913-917.

255 Valencia-Montoya, W. A., S. Elfekih, H. L. North, J. I. Meier, I. A. Warren, W. T. Tay, K. H. J.
256 Gordon, A. Specht, S. V. Paula-Moraes, R. Rane, T. K. Walsh, and C. D. Jiggins. 2020.
257 Adaptive Introgression across Semipermeable Species Boundaries between Local
258 *Helicoverpa zea* and Invasive *Helicoverpa armigera* Moths. *Molecular Biology and*
259 *Evolution* **37**:2568-2583.

260 Yang, F.-Y., H. S. A. Saqib, J.-H. Chen, Q.-Q. Ruan, L. Vasseur, W.-Y. He, and M.-S. You.
261 2020. Differential Profiles of Gut Microbiota and Metabolites Associated with Host Shift
262 of *Plutella xylostella*. *International journal of molecular sciences* **21**:6283.

263 Yen, E. C., S. A. McCarthy, J. A. Galarza, T. N. Generalovic, S. Pelan, P. Nguyen, J. I. Meier,
264 I. A. Warren, J. Mappes, R. Durbin, and C. D. Jiggins. 2020. A haplotype-resolved, de
265 novo genome assembly for the wood tiger moth (*Arctia plantaginis*) through trio
266 binning. *GigaScience* **9**.

267 You, M., Z. Yue, W. He, X. Yang, G. Yang, M. Xie, D. Zhan, S. W. Baxter, L. Vasseur, G. M.
268 Gurr, C. J. Douglas, J. Bai, P. Wang, K. Cui, S. Huang, X. Li, Q. Zhou, Z. Wu, Q. Chen,
269 C. Liu, B. Wang, X. Li, X. Xu, C. Lu, M. Hu, J. W. Davey, S. M. Smith, M. Chen, X. Xia,
270 W. Tang, F. Ke, D. Zheng, Y. Hu, F. Song, Y. You, X. Ma, L. Peng, Y. Zheng, Y. Liang,
271 Y. Chen, L. Yu, Y. Zhang, Y. Liu, G. Li, L. Fang, J. Li, X. Zhou, Y. Luo, C. Gou, J.
272 Wang, J. Wang, H. Yang, and J. Wang. 2013. A heterozygous moth genome provides
273 insights into herbivory and detoxification. *Nature Genetics* **45**:220-225.

274 Zhao, Y., H. Zhang, Z. Li, J. Duan, J. Jiang, Y. Wang, S. Zhan, R. O. Akinkulore, A. Xu, and
275 H. Qian. 2012. A major facilitator superfamily protein participates in the reddish brown
276 pigmentation in *Bombyx mori*. *Journal of insect physiology* **58**:1397-1405.



277 Zheng, X., S. M. Gogarten, M. Lawrence, A. Stilp, M. P. Conomos, B. S. Weir, C. Laurie, and
278 D. Levine. 2017. SeqArray—a storage-efficient high-performance data format for WGS
279 variant calls. *Bioinformatics* **33**:2251-2257.

280

Appendix A

Co-authored publications with relevance to this work.

Identification of Y-chromosome scaffolds of the Queensland fruit fly reveals a duplicated *gyf* gene paralogue common to many *Bactrocera* pest species

Amanda Choo*,[§] Thu N.M. Nguyen*,[§] 
Christopher M. Ward*, Isabel Y. Chen*[†], John Sved[‡],
Deborah Shearman[‡], Anthony S. Gilchrist[‡],
Peter Crisp[†] and Simon W. Baxter* 

*School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia; [†]South Australian Research and Development Institute, Adelaide, South Australia, Australia; and [‡]Evolution and Ecology Research Centre, University of New South Wales, Sydney, New South Wales, Australia

Abstract

Bactrocera tryoni (Queensland fruit fly) are polyphagous horticultural pests of eastern Australia. Heterogametic males contain a sex-determining Y-chromosome thought to be gene poor and repetitive. Here, we report 39 Y-chromosome scaffolds (~700 kb) from *B. tryoni* identified using genotype-by-sequencing data and whole-genome resequencing. Male diagnostic PCR assays validated eight Y-scaffolds, and one (Btry4096) contained a novel gene with five exons that encode a predicted 575 amino acid protein. The Y-gene, referred to as *typo-gyf*, is a truncated Y-chromosome paralogue of X-chromosome gene *gyf* (1773 aa). The Y-chromosome contained ~41 copies of *typo-gyf*, and expression occurred in male flies and embryos. Analysis of 13 tephritid transcriptomes confirmed *typo-gyf* expression in six additional *Bactrocera* species, including *Bactrocera latifrons*, *Bactrocera dorsalis* and *Bactrocera zonata*. Molecular dating estimated *typo-gyf* evolved within the past 8.02 million years (95% highest posterior density 10.56–5.52 million years), after the split with *Bactrocera oleae*. Phylogenetic analysis also highlighted complex evolutionary histories among

several *Bactrocera* species, as discordant nuclear (116 genes) and mitochondrial (13 genes) topologies were observed. *B. tryoni* Y-sequences may provide useful sites for future transgene insertions, and *typo-gyf* could act as a Y-chromosome diagnostic marker for many *Bactrocera* species, although its function is unknown.

Keywords: Y-chromosome, sex determination, *Bactrocera*, GYF protein, tephritid, mito-nuclear discordance.

Introduction

Y-chromosomes are generally gene poor, highly repetitive and can lack any substantial interspecific similarity between related species, making them difficult to identify and validate from sequenced genomes (Carvalho et al., 2009; Sanchez, 2008; Tobler et al., 2017; Vicoso and Bachtrog, 2015). Much of the empirical work on Y-chromosome evolution has involved *Drosophila* species, which do share some features with mammals, including X–Y pairing during mitosis, low gene content and high levels of repetitive elements (Carvalho et al., 2009). Identifying active genes amongst repetitive sequences on Y-chromosomes remains challenging (Tsoumani et al., 2015), but their identification can help understand factors contributing to sex determination and male fertility (Hall et al., 2015).

The Y-chromosome is not essential for sex determination in *Drosophila melanogaster*, yet Y-chromosome do encode a number of male-specific genes required for fertility (Carvalho et al., 2009). The *Drosophila* Y can acquire genes from other chromosomes, and different rates of gene gain occur within the genus (Tobler et al., 2017), which complicates comparative genomic studies attempting to identify Y-chromosome orthologues (Koerich et al., 2008). A range of methods have been used to identify Y-chromosome genes and sequences in nonmodel dipteran insects, including suppression subtractive hybridization (Salvemini et al., 2014), representational difference analysis (Gabrieli et al., 2011), Y-chromosome isolation with pulse field gel electrophoresis and sequencing (Tsoumani et al., 2015) and developing sex-specific genomic libraries (Hall et al., 2013; Koerich et al., 2016). However,

First published online 21 June 2019.

Correspondence: Simon W. Baxter, Molecular Life Sciences Building, University of Adelaide, North Terrace, Adelaide 5005, Australia. Tel: +61 (0)8 8313 2205; e-mail: simon.baxter@adelaide.edu.au

[§]These authors contributed equally.

expressed Y-chromosome genes remain largely unknown among tephritid fruit flies.

A dominant system for determining maleness (*M* factor) was hypothesized to be conserved on the Y-chromosome among many non-*Drosophila* dipterans (Shearman, 2002) and act early in embryo development (Morrow et al., 2014a; Sanchez, 2008; Vicoso and Bachtrög, 2015). Unrelated male sex determination factors have now been reported for *Aedes aegypti* (Nix; Hall et al. (2015)), *Anopheles gambiae* (YoB; Krzywinska et al. (2016)) and, although not limited to the Y-chromosome, *Musca domestica* (*Mdmd*; Sharma et al. (2017)). Sex-determining *M* factors have been also been mapped to the Y-chromosome centromere in *Lucilia cuprina* (Bedo and Foster, 1985) and on the long arm of the Y-chromosome in *Ceratitis capitata* (Mediterranean fruit fly) (Willhoeft and Franz, 1996). Despite the Y being essential for determining maleness, deletion mapping demonstrated *C. capitata* lacked other genes required for male fertility outside this region (Willhoeft and Franz, 1996). The *M* factor regulates a pathway that produces male splice variants of sex-determination genes, including *transformer*, although suppression of female splice variants with RNA interference can produce sex-reversed XX males. Sex-reversed XX males of *L. cuprina* (Concha and Scott, 2009), *Bactrocera oleae* (Lagos et al., 2007), *Bactrocera tryoni* (Raphael et al., 2014) and *C. capitata* (Pane et al., 2002; Salvemini et al., 2009) are fertile, which indicates the Y-chromosome is not necessary for male fertility in these cases.

B. tryoni, the Queensland fruit fly, is a generalist of more than 20 different plant families (Clarke, 2017) and the most important horticultural pest of eastern Australia (Dominiak and Ekman, 2013). Sporadic outbreaks require significant intervention to achieve localized eradication (Dominiak and Daniels, 2012). Female flies puncture fleshy fruit with their ovipositor and lay eggs, causing susceptibility to bacterial infection at the puncture site, and larvae damage fruit through feeding (Hancock et al., 2000). A genome reference is available for *B. tryoni* (>31 960 scaffolds; Gilchrist et al., 2014) and was assembled using XY male flies of the *bent wings* strain. Genotype-by-sequencing (GBS) of single-pair insect crosses assigned and placed 1776 scaffolds to five autosomes (Sved et al., 2016), covering around 50% of the estimated genome length. However, scaffolds have not previously been assigned the small, highly heterochromatic Y-chromosome (Zhao et al., 1998).

A better understanding of insect pest genomes and their sex chromosomes has the potential to improve control strategies, eg through identifying Y-chromosome regions suitable for male-specific transgene insertions. Here, we use both GBS data from single-pair insect crosses and whole-genome resequencing to identify Y-chromosome scaffolds from the *B. tryoni* draft genome and develop molecular diagnostic assays for males. One scaffold was found to contain a novel, repetitive gene that evolved after

the split with *B. oleae* around 8 million years ago (MYA). This work contributes to the understanding of Y-chromosome evolution among *Bactrocera* fruit flies and provides male-specific sequences that could potentially be used for transgene insertions in *B. tryoni*.

Results

Identification of Y-chromosome scaffolds using GBS markers

GBS data from three single-pair *B. tryoni* crosses previously aligned to the reference genome (Sved et al., 2016) were re-analysed for extreme mapping bias between male and female progeny. Fifty-five scaffolds were identified with >99% of GBS reads originating from male progeny and <1% of females. We considered these candidate Y-chromosome sequences, and they had a combined length 1.306 Mb, including sequence gaps (Table S1). GBS read depth was highly variable (Cross 56 = 719.2 ± 547.9 per sequenced base, Cross 171 = 481.8 ± 403.1, Cross 271 = 504.1 ± 396.0), indicating some GBS sequences could be repetitive elements, but this may also be attributable to variation in numbers of reads given by the GBS procedure (Beissinger et al., 2013).

Whole-genome shotgun sequencing of the bent wings strain

Whole-genome sequence read data from pooled males or pooled females of the *bent wings* strain were normalized to approximately 128.655 million 93 base paired-end reads and then mapped to the *B. tryoni* reference genome (Gilchrist et al., 2014). After filtering for mapping quality (>MQ20), median read depth across the genome was 32-fold for both males and females. Coverage across the 55 candidate Y-scaffolds identified with GBS data was expected to be predominantly male, whereas coverage across autosomes was expected to be similar between males and females. Of 55 candidate Y-scaffolds, whole-genome sequence data supported 39 as Y-chromosomal, with male-only coverage accounting for 28.6% to 100% of sites across these scaffolds (mean 47.29%, SD ±22.17%) (Table S2, Fig. S1). The GC content of these 39 Y-scaffolds was 36.64%, which did not differ greatly from autosomes (35.17% for Btry176, Btry280, Btry315); however, the Y-scaffolds do contain large regions of missing data. Read depth was variable across these 39 scaffolds, indicating some scaffolds may contain misassembled repetitive or autosomal sequences (Fig. S1).

PCR validation of male-specific markers

Male-specific diagnostic PCR assays were designed within eight Y-chromosome scaffolds, at 11 sites with low or zero female read depth (Table 1, Fig. S2). Y-chromosome PCR

assays were successful using genomic DNA isolated from *B. tryoni* males, but not females from the Ourimbah laboratory culture. Amplicons for an autosomal control gene, *white*, were obtained from both male and female DNA (Fig. 1). As the Ourimbah culture is likely to be inbred with limited variation among Y-chromosomes, genomic DNA was also isolated from individual wings of wild *B. tryoni* flies. Although DNA yield was low, amplification of Y-chromosome assays was successful in almost all field-collected males; however, a product for *Btry5456(1)* also amplified in 8/27 females (Table 2). Amplification was inconsistent for two assays with wild flies (*Btry3330(1)* and *Btry13701*), and we found the most robust male Y-chromosome diagnostic assays were *Btry1523 (1 and 2)*, *Btry3150*, *Btry4688*, *Btry9748* and *Btry4096*.

Identification of a truncated Y-chromosome paralogue of GYF (typo-gyf)

Y-chromosome scaffold Btry4096 contains an annotated ~16 kb gene with five exons, predicted to produce a 575 amino acid protein (Gilchrist et al., 2014). BLAST analysis against the *D. melanogaster* database of annotated proteins (flybase.org/blast) showed highest homology with *GRB10 interacting GYF protein* (expect value $E = 7.75e^{-39}$, FlyBase FBgn0039936, *gyf*), which is proposed to maintain neuromuscular homeostasis through regulating autophagy (Kim et al., 2015). The

Y-chromosome paralogue was considerably shorter than *gyf* and we refer to this *B. tryoni* gene as *typo-gyf*, a truncated Y-chromosome paralogue of *gyf*.

Protein alignments of full-length *B. tryoni gyf* (1773 aa) and *typo-gyf* (575 aa) show 27.4% identity, and 71.65% identity when all gaps are removed. Intron boundaries are largely conserved between *B. tryoni gyf* and *typo-gyf* gene models, although, the first *gyf* exon (411 bp) is spliced into two shorter *typo-gyf* exons (exon 1: 88 bp; exon 2: 188 bp) plus an 82 bp intron. Conceptual read through of this intron would introduce a stop codon in *typo-gyf*. Large deletions have occurred relative to *gyf*, including loss of the GYF domain (Kofler and Freund, 2006) in exon 4, and removal of the last three exons (Fig. 2A). PCR primers designed in *typo-gyf* exon 1 and exon 2 amplified products using male *B. tryoni* DNA, but not female DNA template (Fig. 2B).

Median read depth across scaffold Btry4096 (JHQJ01002098.1), encoding *typo-gyf*, was high in males (677-fold, excluding sequence gaps) compared with other Y-scaffolds (Table S2). This indicated *typo-gyf* and surrounding sequence was repetitive and probably assembled into a single consensus scaffold. Gene *gyf* was located on scaffold Btry2907, and male read depth was consistently lower than female across the gene, suggesting it was X-chromosomal (Fig. 3). Copy number variation of *gyf* and *typo-gyf* were determined with quantitative PCR using genomic DNA from four male and four female *B. tryoni* flies.

Table 1. PCR primers for Y-chromosome diagnostic assays

Primer name	Sequence (5'–3')	Size (bp)	Annealing (°C)	Scaffold(Position from ... to)
Btry1523(1)_F	GGAGCAATCGCTTTCATC	260	60 ^a	JHQJ01001220.1
Btry1523(1)_R	CATCGAAGCGAAGGTAAC			(58 946...59 205)
Btry1523(2)_F	GGCCAAATAGCCAGGTCAAC	391	60 ^a	JHQJ01001220.1
Btry1523(2)_R	GCGTTGGGGTACACAAGATG			(75 311...75 701)
Btry3150_F	CTGGTTGTCTGTGATACTCC	234	60 ^a	JHQJ01003004.1
Btry3150_R	CGAATTACGTGCCTGTTTGC			(8454...8687)
Btry3330(1)_F	GGCCAGTTGATTGT GGTAG	350	60 ^a	JHQJ01003140.1
Btry3330(1)_R	GAT GGT CATGT GACCTACC			(19 359...19 708)
Btry3330(2)_F	GTTGAGTTAATCAACGTCG	285	55 ^a	JHQJ01003140.1
Btry3330(2)_R	GTACTCTTCAAACATTGCC		58 ^b	(20 436...20 720)
Btry4688_F	CT CCAACGCCAGT GAACT G	427	60 ^a	Table S1
Btry4688_R	CAAGGCGGCTATTACCACC			(3463...3889)
Btry5456(1)_F ^d	GACGAT CAGTCTGAGCAC	201	62 ^a	JHQJ01004522.1
Btry5456(1)_R	GACCACACGACAGAAGTG			(6802...7002)
Btry5456(2)_F	GAAGAAATGGCAAGCGAG	208	60 ^a	JHQJ01004522.1
Btry5456(2)_R	GTTTCGATTACACCCGAAC			(24 997...25 204)
Btry9748_F	GGTCTAGCTGCTGGATATG	463	60 ^a	JHQJ01006679.1
Btry9748_R	GTCCCATTACTTTCCCGAC			(5209...5671)
Btry13701_F	CATAGAT GCCTGGAATAGC	567	62 ^b	JHQJ01008474.1
Btry13701_R	GACTTACAAGCTTCGGTCTG		67 ^c	(7540...8106)
Btry4096_F	GTGGAT GTAATACTGGT GGAGA	314	60 ^a	JHQJ01002098.1
Btry4096_R	ACCT GTAGCAGTAGTTCATCTCT			(27 865...28 178)
<i>White</i> _F ^e	GCTAGCGAT CAT GGGCAG	600	50 ^a	JHQJ01000419.1
<i>White</i> _R ^e	GGT GTACCCAGAAAGCGC			(142 623...143 222)

Taq polymerase.

^aPhire polymerase.

^bKAPA Robust polymerase.

^cKAPA HiFi polymerase.

^dAmplification observed in some females.

^eAutosomal positive control gene.

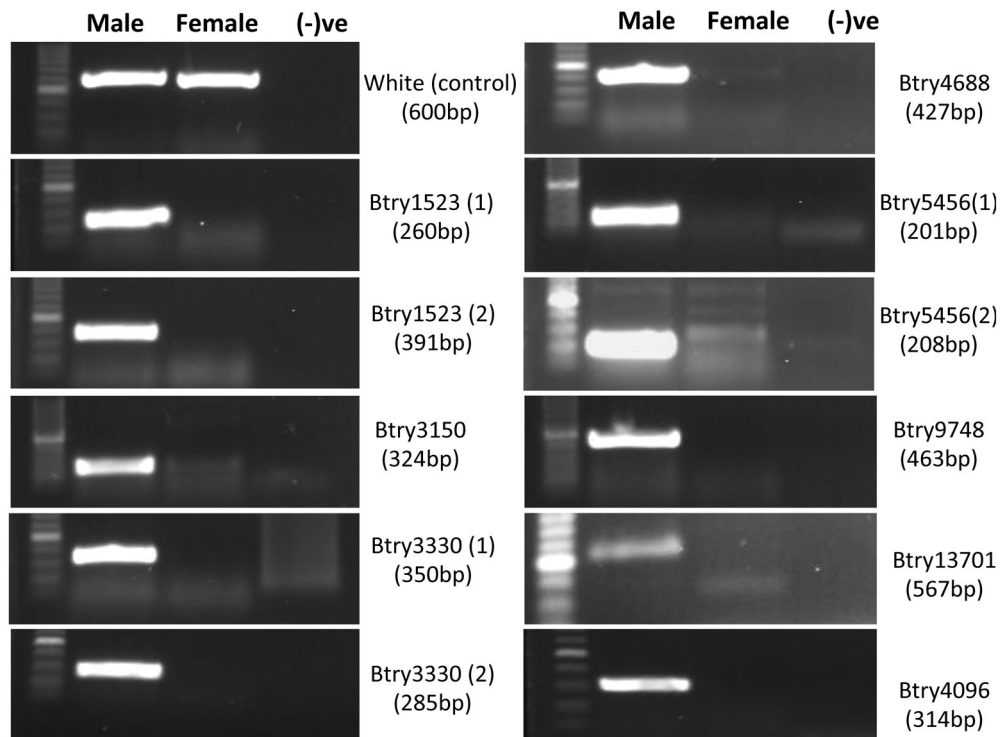


Figure 1. Agarose gel electrophoresis images of Y-linked amplicons. Each panel shows a DNA ladder (Promega, 100 bp), *Bactrocera tryoni* male and female PCR products plus a no-template negative control (-). The gene *white* is an autosomal control and amplicon sizes are indicated.

Relative to autosomal control gene *actin3*, *typo-gyf* had a relative target quantity (RQ) of 41.40 (SD \pm 4.87) and *gyf* was confirmed to have two copies in females (XX karyotype, 2.07 ± 0.156), and one copy in males (XY karyotype, 0.916 ± 0.081) (Fig. S3).

typo-gyf expression and diversity in *B. tryoni*

Expression of *typo-gyf* was assessed in teneral flies and pooled embryos collected 1, 5 or 24 h after egg laying (mixed sex, $n = 20$). DNase-treated total RNA was reverse transcribed

Table 2. PCR amplification results for 10 *Bactrocera tryoni* Y-chromosome assays using laboratory strain (Ourimbah) and wild samples (Auburn, NSW)

PCR assay	<i>Bactrocera tryoni</i>			
	Lab strain		Wild samples	
	Males	Females	Males	Females
Btry1523 (1)	17/17 (100%)	0/19 (0%)	30/30 (100%)	0/25 (0%)
Btry1523 (2)	16/16 (100%)	0/19 (0%)	30/30 (100%)	0/25 (0%)
Btry3150	16/16 (100%)	0/19 (0%)	30/30 (100%)	0/25 (0%)
Btry3330 (1)	16/16 (100%)	0/19 (0%)	NA	NA
Btry3330 (2)	16/16 (100%)	0/18 (0%)	30/30 (100%)	0/27 (0%)
Btry4688	16/16 (100%)	0/19 (0%)	30/30 (100%)	0/25 (0%)
Btry5456 (1)	16/16 (100%)	0/19 (0%)	30/30 (100%)	8/27 (30%)
Btry5456 (2)	14/16 (88%)	0/18 (0%)	29/30 (97%)	0/27 (0%)
Btry9748	16/16 (100%)	0/19 (0%)	30/30 (100%)	0/25 (0%)
Btry13701	13/16 (81%)	0/17 (0%)	NA	NA
Btry4096	16/16 (100%)	0/18 (0%)	29/30 (97%)	0/26 (0%)

NA, not available.

to cDNA, and expression of *typo-gyf* was confirmed in male flies and 24-h-old embryos with four PCR assays (Fig. 4, Table S3). A primer pair designed in exons 1 and 2 produced multiple amplicons from 24 h embryo cDNA, including a 243 bp amplicon with the correct reading frame, and a 325 bp unspliced amplicon containing an intron and premature stop codon. Weak amplification of unspliced transcripts was also observed in 5-h-old embryos (Fig. 4A, arrow). The X-chromosome *gyf* gene was expressed in all samples (Fig. 4E).

The *B. tryoni* Y-chromosome contains ~41 copies of *typo-gyf*. To assess their similarity, we generated cDNA template from a single male fly and PCR amplified *typo-gyf* transcripts with primers in the 5' and 3' untranslated regions (Table S3). Twenty-five cloned transcripts were sequenced, and most ($n = 19$) shared >97% nucleotide identity (Fig. S4), although conceptual translations showed that only 10 were predicted to produce a full-length 575 amino acid protein. The remaining 15 cloned transcripts encoded shorter proteins (62–567 aa) due to alternative splicing of messenger RNAs, deletions or nucleotide substitutions creating premature stop codons. However, some sequence errors may have been caused during PCR amplification or cloning procedures (Fig. S5). Nucleotide variability between sequenced transcripts indicates that many different *typo-gyf* copies are expressed, and their predicted proteins can vary in length.

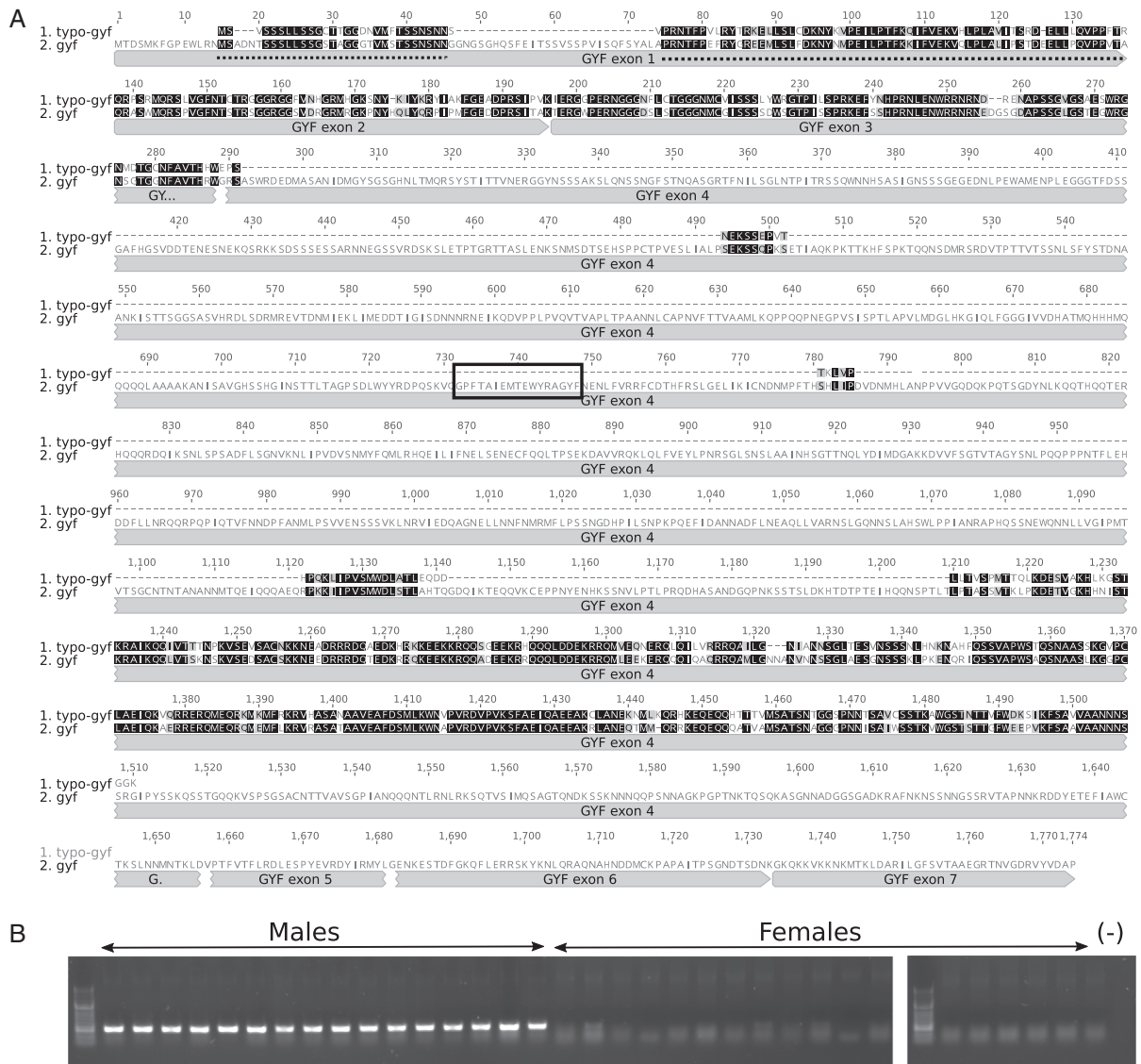


Figure 2. (A) Alignment of *Bactrocera tryoni* proteins TYPO-GYF (575 aa) and GYF (1773 aa). These proteins have conserved exon boundaries, aside from GYF exon 1, which is spliced into two *typo-gyf* exons (dotted lines). The final *typo-gyf* exon has multiple deletions relative to GYF, including removal of the core GYF motif (boxed). Conserved amino acids are shaded black; similar amino acids in grey and gaps are indicated with dashes. (B) Amplification of *typo-gyf* (325 bp) from *B. tryoni* male and female genomic DNA (laboratory culture).

Multiple *Bactrocera* species express *typo-gyf* orthologues

The National Centre for Biotechnology Information (NCBI) Transcriptome Shotgun Assembly database contains male-only or mixed-sex transcriptome assemblies for multiple tephritid species (Table S4). BLAST analysis using the *B. tryoni* TYPO-GYF protein confirmed partial matches in *Bactrocera bronyiae* (322 aa), *Bactrocera latifrons* (382 aa), *Bactrocera jarvisi* (310 aa), *Bactrocera dorsalis* (575 aa) and *Bactrocera kraussi* (808 aa, probable *gyf/typo-gyf* chimeric sequence). Transcriptome assemblies of *Bactrocera zonata* (SRR4024787), *Bactrocera correcta* (SRR4020110) and *B. oleae* (SRR5559327) only identified *typo-gyf* transcripts in *B. zonata* (Table S5). TYPO-GYF proteins shared between 67% and

86% amino acid identity with *B. tryoni* (Table S6). We also re-analysed transcriptomes from carefully staged male and female *B. jarvisi* embryos (Morrow et al., 2014b) and found that *typo-gyf* was the most significant differentially expressed transcript in 3- to 5-h-old samples: $\log(\text{fold change}) = 11.828$, $P < 1.310 \times 10^{-7}$, false discovery rate 0.00158 (Fig. S6). Transcriptomes of all insects analysed here also contained partial or full-length GYF transcripts (Table S5), and amino acid identity was >85% among the *Bactrocera* species.

Evolutionary origin of *typo-gyf*

A Bayesian phylogeny of 16 dipteran species, including the seven identified with *typo-gyf*, was produced to determine

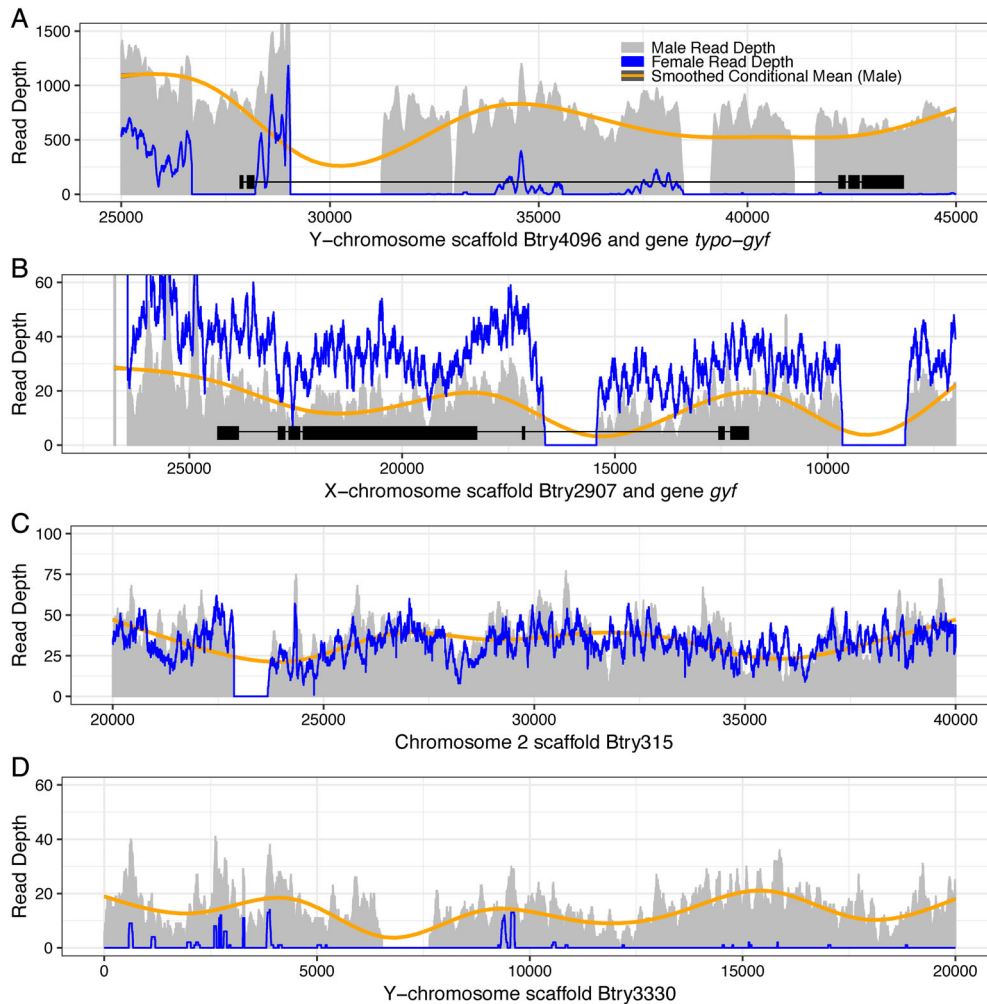


Figure 3. Read depth of male (grey bar) and female (blue line) resequenced genomes across 20 kb regions of four scaffolds. (A) Y-chromosome scaffold Btry4096 containing gene *typo-gyf* with five exons annotated in black. Male read depth is consistently higher than the genome-wide average, indicating the region represents a series of conserved duplications that assembled into a single scaffold. (B) X-chromosome scaffold Btry2907 and gene *gyf*. Male read depth is consistently lower than female depth due to X chromosome copy number. The scaffold was reverse complemented to orient *gyf* in the sense direction, and the seven exons are shaded black. (C) Autosomal scaffold Btry315 shows similar male and female read depth. (D) Scaffold Btry3330 represents male bias observed for a Y-chromosome region. Each plot also shows the smoothed conditional mean of male read depth (orange line with 95% confidence interval). [Colour figure can be viewed at wileyonlinelibrary.com].

the likely evolutionary origin of *typo-gyf* (Table S4). Phylogenetic reconstruction was performed using 116 single copy orthologues (155.7 kb), and all 16 species nodes had posterior probabilities of 1.0 (Fig. 5, Table S7). Molecular dating was performed on the phylogeny using a log-normal distribution of the *Drosophila-Rhagoletis* most recent common ancestor estimated by Misof et al. (2014) to inform the root prior (median 81.3037 MYA, 95% highest posterior density (HPD) 55.806–111.679 MYA). Tephritid species were monophyletic with a common ancestor 36.04 MYA (95% HPD 47.46–24.86 MYA), which predates the earliest known fossil for this group from the Miocene (~25 MYA) (Norbom, 1994; Poinar, 1992). Within the *Bactrocera* clade, Australian endemic species (*B. tryoni*, *B. jarvisi* and *B. kraussi*) were

monophyletic with a most recent common ancestor between 2.69 and 1.4 MYA.

The Bayesian phylogeny indicated *typo-gyf* evolved in *Bactrocera* after the split with *B. oleae* between 10.56 and 5.52 MYA (Fig. 5, dagger symbols). *Typo-gyf* was not present in four available *B. correcta* transcriptomes (Table S5), indicating it may have been lost from the Y-chromosome within the last 1.97 MYA, the transcripts are not expressed in this species, or their Y-chromosome has evolved independently.

The topology of the Bayesian nuclear phylogeny (Fig. 5) showed several differences to previous studies that largely used mitochondrial data (Kearse et al., 2012; Zhang et al., 2010), particularly the placement of *B. jarvisi* and *B. dorsalis*.

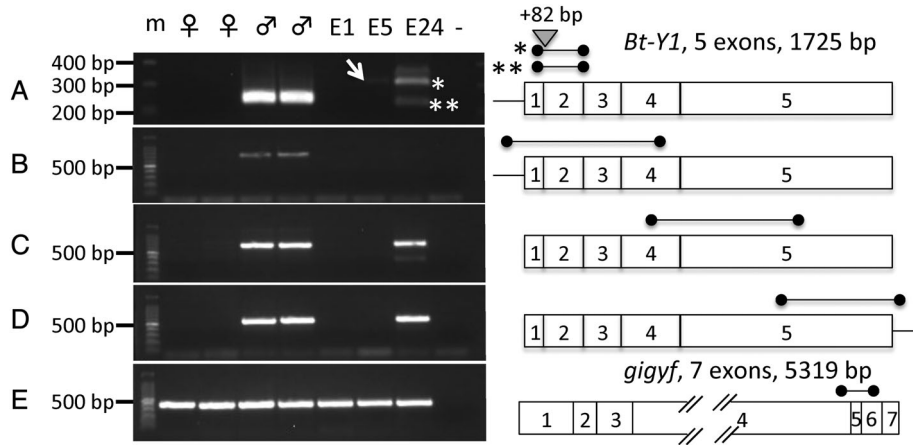


Figure 4. *typo-gyf* is male specific and expressed in embryos and flies. (A–E) PCR amplification was performed using cDNA from 3-day-old virgin females (♀), males (♂), 1-h embryos (E1), 5-h embryos (E5), 24-h embryos (E24) and a no-template control (–). A schematic of each amplicon size and location is shown. (A) A 243 bp *typo-gyf* amplicon using primers in exon 1 and exon 2 is produced in male flies and 24-h embryos (**). An unspliced product is also expressed in 24-h embryos (*) and weakly in 5-h embryos (arrow). (B) A 747 bp *typo-gyf* amplicon spanning the 5' untranslated region (UTR) to exon 4. (C) A 713 bp *typo-gyf* amplicon spanning exon 4 to exon 5. (D) A 579 bp *typo-gyf* amplicon spanning exon 5 to the 3' UTR. The DNA marker 'm' is a 100 bp DNA ladder (Promega). (E) The 426 bp *gyf* product was amplified from all samples. See Table S3 for primer sequences.

To examine this further, separate maximum likelihood (ML) trees of the 116 BUSCO nuclear gene set and 13 mitochondrial genes (8357 bp) obtained from transcriptomes or

genomes were compared (Table S4). The ML nuclear tree had similar topology to our Bayesian phylogeny; however, the mitochondrial tree was discordant across the *Bactrocera*

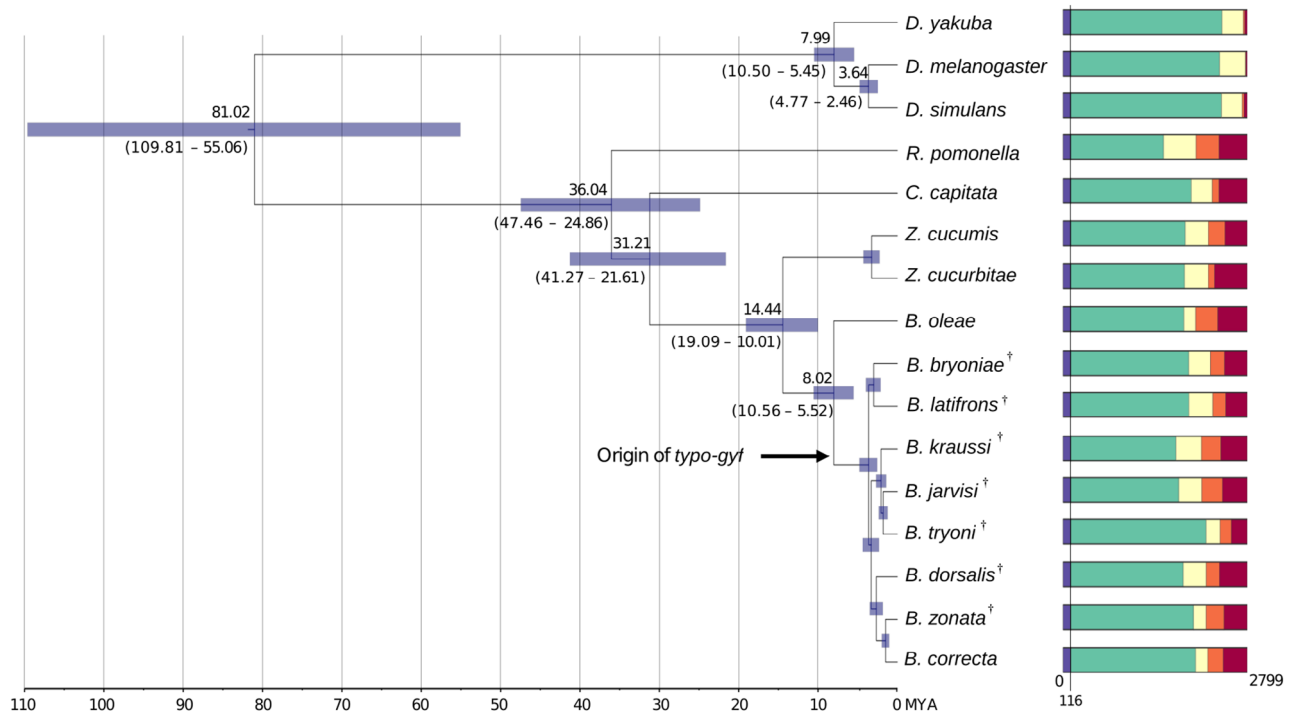


Figure 5. Bayesian phylogeny of three Drosophilidae and 13 Tephritidae species, generated using 116 single-copy gene orthologues. Median split times are shown, with 95% highest posterior density ranges indicated in brackets. Rectangular plots represent the proportion of 2799 BUSCO genes identified from genomes or transcriptomes of each species (purple, full-length gene in all species; green, full-length single-copy gene; yellow, full-length multicopy gene; orange, fragmented partial gene; red; absent). *Typo-gyf* transcripts were recovered from seven species, indicated with the dagger symbol. The gene arose after a common ancestor split with *Bactrocera oleae* 5.52–10.56 million years ago (MYA). Genus abbreviations: *D.*, *Drosophila*; *R.*, *Rhagoletis*; *C.*, *Capitata*; *Z.*, *Zeugodacus*; *B.*, *Bactrocera*. [Colour figure can be viewed at wileyonlinelibrary.com].

clade. Australian *B. jarvisi* was no longer monophyletic with *B. tryoni* and *B. kraussi*, but grouped with *B. zonata* and *B. correcta*, whereas *B. dorsalis* grouped with *B. tryoni* and *B. kraussi* (Fig. S7). Bootstrap support for taxon placement was only modest for *B. jarvisi* and *B. dorsalis* in the mitochondrial tree (70% and 77% respectively); nevertheless, mitochondrial and nuclear genomes of these *Bactrocera* species are discordant and appear to have alternate evolutionary histories.

Discussion

Sex chromosomes arise from autosomes (Charlesworth, 1996), and the *B. tryoni* X evolved from the small dot chromosome (Muller element F) (Vicoso and Bachtrog, 2015). Genes on the *Drosophila* dot and *B. tryoni* X are relatively conserved (Sved et al., 2016), despite extensive rearrangements, yet gene content and structure of dipteran Y-chromosomes can vary extensively, and identification of Y-chromosome genes remains challenging (Carvalho et al., 2009; Gabrieli et al., 2011). A primary goal of this study was to develop diagnostic Y-chromosome markers for *B. tryoni* and to identify potential sites for transgene insertions with CRISPR/Cas9 (Buchman and Akbari, 2019). Reduced representation genome sequencing is one strategy for developing genetic markers on autosomes (Davey et al., 2011) or sex chromosomes (Baxter et al., 2011; Gamble and Zarkower, 2014), and identifying Y-sequence from the fragmented *B. tryoni* draft genome with GBS was a relatively robust strategy. However, it has probably limited our overall findings to Y-chromosome scaffolds that contain GBS restriction enzyme sites.

Identification of *B. tryoni* Y-scaffolds, and the development of diagnostic Y-chromosome assays, could benefit several research applications. First, it may be possible to use Y-specific primers to genotype females collected from the field or pheromone traps to determine whether they have mated and carry sperm. Second, determining the sex of embryos for gene expression analysis could be performed, as shown by Morrow et al. (2014b) in *B. jarvisi*. Third, single-copy Y-chromosome sequences could be targeted for male-specific transgene insertion. Fourth, the assays could be used to help confirm karyotypes of individuals when performing experimental sex reversal with RNA-interference knock down, or CRISPR/Cas9 modification of sex-determination pathway genes (Li and Handler, 2017; Peng et al., 2015). Finally, the sequences identified here should assist with future *B. tryoni* Y-chromosome assemblies using long-read sequencing technologies (Jain et al., 2018). All of these future research topics may not be limited just to *B. tryoni*. Developing Y-chromosome markers in many other *Bactrocera* species will be possible using *typo-gyf* homologues.

Evolution of typo-gyf

Transfer of genes to the *Drosophila* Y-chromosome was recently found to occur 10-fold more frequently than expected (Tobler et al., 2017), although most suffer rapid degeneration and pseudonization. A copy of the X-chromosome gene *gyf* may have transferred to an ancestral *Bactrocera* Y-chromosome, where it subsequently degenerated, forming *typo-gyf* that lacks the GYF domain. The persistence of *typo-gyf* on the Y-chromosome for millions of years may indicate neofunctionalization, although it is unlikely to be involved with male fertility, as *B. tryoni* sex-reversed XX males that lack a Y-chromosome are fully fertile (Raphael et al., 2014).

Approximately 41 copies of *typo-gyf*-like transcripts are present in the *B. tryoni* genome. Expressed copies of *typo-gyf* are relatively similar at the DNA level, suggesting copy number expansion occurred a considerable time after diverging from the ancestral *gyf* paralogue. The *typo-gyf* genomic scaffold (Btry4096) contains extensive sequencing gaps, and long-read genome sequencing should help resolve the distribution and order of these *typo-gyf* copies, and enable neighbouring repetitive elements to be characterized.

Gene duplication and high rates of intrachromosomal gene conversion are likely to have important roles in the evolution of Y-chromosome genes in *Drosophila* (Chang and Larracunte, 2019). Gene conversion is known to maintain palindromic repeated sequences on the Y-chromosome of humans and apes (Rozen et al., 2003) to prevent their decay and can act to maintain the ancestral state (Skov et al., 2017). Although polymorphic variation was observed among some expressed *typo-gyf* transcripts, gene conversion could be acting to maintain many of these sequence repeats on the *B. tryoni* Y. Expression of *typo-gyf* transcripts was found among seven *Bactrocera* species we tested, and determining whether copy number expansions have occurred in these cases will provide insight into the origin of this duplication. There are documented examples of duplicated *B. oleae* Y-chromosome sequences, including *importin-4* gene fragments (Gabrieli et al., 2011), and although not exclusively Y-chromosomal, the long terminal repeat retrotransposon *Achilles* (Tsoumani et al., 2015). Further sequencing of Y-chromosomes from a range of tephritids, determining copy number variation using genomic DNA or performing chromosomal *in situ* hybridization experiments, will help determine the extent to which *typo-gyf* has been duplicated.

Phylogenetic analysis of tephritids and the molecular clock

We focused on establishing phylogenetic relationships from a limited number of Diptera with available transcriptome or genomic datasets. Despite the availability of an extensive dipteran BUSCO gene set (Waterhouse et al.,

2017), phylogenetic reconstruction of the nuclear genome was limited to just 116 of 2799 single-copy orthologues. The gene number was limited because of variation in library tissue type, the *Rhagoletis pomonella* transcriptome assembly was highly fragmented and we were limited to those with transcriptome or genome data available from males, as our primary aim was to resolve the origin of *typo-gyf*. Nevertheless, this Bayesian nuclear phylogeny provides a robust and well-supported topology of these *Bactrocera* species (Fig. 5). However, the topology of a species phylogeny can vary considerably depending upon genes selected (Degnan and Rosenberg, 2009; Pollard et al., 2006). Evolutionary relationships between *B. tryoni* and *B. jarvisi*, plus *B. dorsalis* to that of *B. zonata* and *B. correcta*, were notably different between the nuclear and mitochondrial phylogenies, suggesting alternate evolutionary histories of these two genomes.

Informative morphological characters for establishing evolutionary relationships within the Dacini tribe are relatively limited, prompting several recent efforts to reconstruct mitochondrial trees with amplicon sequence data (Krosch et al., 2012; San Jose et al., 2018; Zhang et al., 2010). Consistent topologies were produced with mitochondrial data among these studies, and data presented here (Fig. S7) where *B. jarvisi* was in monophyly to *B. zonata* and *B. correcta*, while nuclear phylogenies show *B. jarvisi* in monophyly with *B. tryoni* (Dupuis et al., 2018; San Jose et al., 2018). Phylogenies produced with mitochondrial genomes can therefore have alternative topologies to nuclear phylogenies among *Bactrocera*, and future work should investigate if historic introgression has occurred between *B. jarvisi* and *B. tryoni*, which may explain discordant relationships.

Molecular clocks use DNA sequence variation between species to estimate their point of evolutionary separation (Bromham and Penny, 2003). The reliability of molecular clocks is dependent upon factors including calibration rates that estimate substitutions in DNA or protein sequence over time, fossil records to introduce minimum divergence estimates, generation time and population size and model choice. However, factors other than clock parameters can affect evolutionary rates. Published estimates of when *C. capitata* last shared a common ancestor with *Bactrocera* include 24.9 MYA (Yaakop et al., 2015; no confidence intervals reported), 83 MYA (95% HPD: 64–103 MYA; Nardi et al., 2010) and 110.9 MYA [95% confidence interval (CI): 131.4–91.2 MYA] (Krosch et al., 2012). Our Bayesian phylogeny estimate of 31.21 MYA (95% CI: 41.27–21.16) is earlier than many others due to root prior calibration using the *Rhagoletis–Drosophila* split estimated by Misof et al. (2014), as opposed to calibration using maximum fossil ages. Therefore, assumptions made by Misof et al. (2014) will be carried into the age estimates presented here (Kjer et al., 2015; Tong et al., 2015).

The origin of *typo-gyf* was dated within the last 10.56 million years, after the *B. oleae* split; however, the timing and extent of gene duplication remain unclear. We found no evidence of a *typo-gyf* gene in *B. correcta*, indicating it may have been lost after splitting with sister species *B. zonata*. Major differences in Y-chromosome evolution between *B. correcta* and related *Bactrocera* species would not be unprecedented, as the Y-chromosome of *D. melanogaster* and related *Drosophila pseudoobscura* are known to be completely different, without any common genes (Carvalho and Clark, 2005).

Conclusion

Here, we identified *B. tryoni* Y-chromosome regions that may be suitable for transgene insertion with CRISPR/Cas9, as recently shown in *D. melanogaster* (Buchman and Akbari, 2019). Inserting markers or traits onto the Y-chromosome could potentially generate transgenic male-selecting strains used for the sterile insect technique, or sites to introduce X-chromosome-shredding gene-drives (Galizi et al., 2016). The *typo-gyf* genes could be used as Y-chromosome molecular markers in many other *Bactrocera* species. Although *typo-gyf* is not expected to play a role in fertility, this gene has been maintained on the Y-chromosomes of at least seven *Bactrocera* species for millions of years and may have a novel function.

Experimental procedures

Insect samples

A *B. tryoni* laboratory strain (Ourimbah) was obtained from the New South Wales (NSW) Department of Primary Industries, Ourimbah, Australia. The strain was maintained at $25 \pm 2^\circ\text{C}$, $65 \pm 10\%$ relative humidity and a 14 h : 10 h light : dark cycle. Adult flies were reared on sugar, brewer's yeast, water and additional yeast paste and the larvae on a gel diet (Moadeli et al., 2017). Field-collected *B. tryoni* samples were obtained from loquat fruits (*Eriobotrya japonica*) in 2017 from Auburn, NSW, and then reared for one generation on capsicum fruits (*Capsicum annuum*). The *B. tryoni* strain *bent wings* was used for GBS and whole-genome sequencing, as described previously (Gilchrist et al., 2014; Sved et al., 2016).

Nucleic acid isolation

DNA was extracted from individuals of the *B. tryoni* Ourimbah laboratory strain collected immediately upon eclosion to ensure females were virgins and would not be carrying male sperm that could potentially bias molecular analyses. Individual flies were homogenized using a TissueLyser (Qiagen, Hilden, Germany) with stainless-steel ball bearings, and DNA was extracted using the Wizard Genomic DNA Purification kit (Promega, Madison, WI, USA). DNA was isolated from single wings of individual field-collected flies using the Phire Animal Tissue Direct PCR kit (ThermoFisher Scientific, Waltham, MA, USA).

Embryos were collected for 1 h, removed from cages and then snap frozen in liquid nitrogen (1-h-old embryos) or left to develop for a further 4 h (5-h-old embryos) or 23 h (24 h embryos) before freezing. Total RNA was isolated using the RNeasy Lipid Tissue Mini Kit (Qiagen), treated with Ambion DNaseI (ThermoFisher Scientific) and reverse transcribed using Invitrogen SuperscriptIII (ThermoFisher Scientific) primed by a hexamer and oligo-dT mixture (Promega).

Analysis of GBS data

Three single-pair crosses were previously prepared for GBS using restriction enzyme *Eco*T221 and mapped to *B. tryoni* scaffolds with *bwa-mem*, outlined in Sved et al. (2016). Genomic scaffolds were considered candidate Y-chromosome sequences if >99% of mapped GBS reads (~85 bp) were from male progeny.

Analysis of whole-genome sequence data

Illumina sequence reads were generated from pooled males and pooled females (SRA accession: PRJNA531345) of the *bent wings* strain. TRIMMOMATIC (version 0.36) was used to quality filter reads using the parameters SLIDINGWINDOW:4:15 CROP:93 MINLEN:93. This produced 128 655 388 paired-end 93 bp female reads, and male sequence data was limited to a similar number of 128 654 791 paired-end 93 bp reads using HTSEQ (v.0.6.1) (Pyl et al., 2014). Reads were mapped to the *B. tryoni* reference genome, available from NCBI (JHQJ01). Twenty-one additional scaffolds identified as putative Y-chromosome sequences using GBS data were added to the NCBI reference using the Unix function *cat*. Equivalent numbers of male and female reads were then aligned to the reference genome using the aligner *BWA*, *bwa mem* with default parameters. The resulting SAM files were converted to BAM using SAMTOOLS (version 1.8) and then sorted and indexed. Mapping quality was filtered (MAPQ>20) and read depth calculated for each site with SAMTOOLS (version 1.8) using the parameters [depth -Q 20 -a].

PCR analysis of Y-chromosome scaffolds

Y-chromosome PCR primers were designed using PRIMER3PLUS (v. 2.4.2) (Untergasser et al., 2012). Amplification was performed using DNA isolated from individual males and females of the Ourimbah laboratory strain and samples collected from Auburn. PCR experiments were performed using the KAPA2G Robust HotStart ReadyMix (KAPA Biosystems) with the following cycling conditions: 95°C for 3 min, 35 cycles of 95°C for 15 s, 58–62°C for 15 s and 72°C for 15 s, followed by a final extension of 72°C for 1 min. Conditions for the Phire Hot Start II DNA polymerase kit (ThermoFisher Scientific) were 98°C for 5 min, 35 cycles of 98°C for 5 s, 55–62°C for 5 s and 72°C for 15 s, followed by a final extension of 72°C for 1 min. The KAPA HiFi HotStart PCR kit (KAPA Biosystems) required the following conditions: 95°C for 3 min, 35 cycles of 98°C for 20s, 57°C for 15 s and 72°C for 1 min, followed by a final extension of 72°C for 1 min (Table 1). cDNA was amplified using MyTaq polymerase (Bioline). Amplicons were purified using MinElute (Qiagen) kits, subcloned into pGEM-t-Easy vector (Promega) and transformed into DH5-alpha cells. Y-chromosome PCR products and plasmid clones were sequenced by the Australian Genome Research Facility.

Copy number variation of typo-gyf

Copy number variation of *gyf* and *typo-gyf* was estimated with an ABI StepOnePlus Real-Time PCR system. Comparative C_T experiments were performed using genomic DNA from four females and four male flies, using *actin3* as the control. All reactions were performed in triplicate using the SensiFAST SYBR Hi-ROX kit (BioLine Cat. Bio-92005) according to the manufacturer's instructions. Primer sequences include *actin3* (F 5'-CAGTATTGCTCACCGAAGCA-3', R 5'-GTACGACCGGAAGCGTAGAG-3'), *typo-gyf* (F 5'-AACATGTGCGTCATCAGCTC-3', R 5'-AAGATT CAGCACTGCCCACT-3') and *gyf* (F 5'-GGGTATAGCCGGAAGTGGACA-3', R 5'-GTTGAAAGTTCGACCGGAAG-3').

Data source and de novo assembly

Assembled transcriptome datasets were obtained from the NCBI and FlyBase (<http://flybase.org>). Short read sequence data for species without available assemblies were downloaded from the NCBI Sequence Read Archive and *de novo* assembled with TRINITY v.2.5.1 (Grabherr et al., 2011) with parameter --no_normalize_reads (Table S4). Coding sequence was predicted using TRANSDCODER v.5.1.0 (Haas et al., 2013) with parameter -LongOrfs.

Identification of typo-gyf and gyf sequences among insect species

The protein sequence of *B. tryoni* TYPO-GYF and *D. melanogaster* GYF (FlyBase ID FBpp0290107) was queried against 13 *Bactrocera* transcriptomes and three *Drosophila* genomes (Table S4) using the BLAST function in GENEIOUS (v.10.2.6) (Kearse et al., 2012). Query hits were compared with *D. melanogaster* GYF and *B. tryoni* TYPO-GYF to identify possible misassemblies and then sequences were aligned with GENEIOUS (v.10.2.6) using MAFFT v.7.388 (Katoh and Standley, 2013).

Identification of gene orthologues, sequence alignments and phylogenetic analysis

Open reading frames were identified from transcriptomes or annotated genomes using TRANSDCODER (v.5.4) (Haas et al., 2013). As many of the transcriptome libraries were assembled using pooled individuals and developmental stages, CD-HIT v.4.6.1 (Godzik and Li, 2006) was used to select the longest transcript with a 98% sequence identity threshold (-c 0.98). Single-copy nuclear orthologues were then identified using the orthoDB Diptera gene set (*dip-tera_odb9*, 2799 genes) (Zdobnov et al., 2017) with BUSCO (v.3) (Waterhouse et al., 2017) and AUGUSTUS (Stanke and Morgenstern, 2005). Orthologous genes annotated in all samples ($n = 16$) were extracted and nucleotides aligned based on reading frame using GENEIOUS v.10.2.6 with the MAFFT FFT-NS-2 algorithm (Katoh and Standley, 2013). Alignment gaps in three or more samples were removed and any incomplete reading frames were trimmed to the nearest codon shared between all individuals. Over half the 2799 full-length BUSCO genes were identified in all species, yet only 116 complete single-copy genes were shared between all samples, producing a concatenated length 155.7 kb.

Mitochondrial genes were extracted from reference mitogenomes, including *B. correcta* (NC_018787), *Bactrocera cucurbitae* (NC_016056), *B. dorsalis* (NC_008748), *B. latifrons* (NC_029466),

B. oleae (NC_005333), *B. tryoni* (NC_014611), *B. zonata* (NC_027725), *C. capitata* (NC_000857), *D. melanogaster* (NC_024511), *Drosophila simulans* (NC_005781) and *Drosophila yakuba* (NC_001322). The mitochondrial genes of *B. bryoniae*, *Bactrocera cucumis*, *B. jarvisi*, *B. kraussi* and *R. pomonella* were obtained by homology BLAST against transcriptome data.

A concatenated set of 13 mitochondrial genes were aligned using MAFFT v.7.388 (Kato and Standley, 2013) and trimmed with GENEIOUS (v.10.2.6) to remove missing data. ML trees were constructed using the general time-reversible model and a gamma distribution with gamma rate heterogeneity using RAXML v.8 (Stamatakis, 2014) – alignment available upon request. Bootstrapping was carried out to provide node support with 1000 replications. A nuclear ML tree was constructed using these same parameters with 116 complete single-copy nuclear orthologues, except 100 bootstrapping replications were performed (Table S7).

A species tree using 116 BUSCO genes was reconstructed with BEAST 2 (Bouckaert et al., 2014). Substitution models for all genes were unlinked Hasegawa–Kishino–Yano (Hasegawa et al., 1985) models with a four-category discrete gamma rate heterogeneity. Genes were partitioned by codon position (1st + 2nd and 3rd) according to Shapiro et al. (2006). Molecular clocks were set to relaxed with a log-normal distribution to allow for variation between branches (Drummond et al., 2006). A random starting tree was generated and tree prior set to calibrated Yule (Heled and Drummond, 2011). We used a log-normal distribution of the *Drosophila–Rhagoletis* most recent common ancestor estimated by Misof et al. (2014) to inform the root prior (median 81.3037 MYA, 95% HPD 55.806–111.679 MYA). Markov chain Monte Carlo analysis was carried out with a total chain length of 1×10^9 sampled every 1000 chains. This analysis was then repeated to compare topology and date local maxima from the Markov chain Monte Carlo. The two runs were combined and annotated using TREEANNOTATOR (Bouckaert et al., 2014). TRACER v.1.7 (Rambaut et al., 2018) was then used to compare replicates and determine chain burn-in, based on visual inspection of parameter Effective Sample Sizes, convergence and mixing (17%).

Data visualization

Plots depicting sequence read depth (Figs 3 and S2), sequence read bias (Fig. S1) and copy number variation (Fig. S3) were produced with the R programming language (R-Core-Team, 2018) and package GGPlot2 (Wickham, 2016). Images of sequence alignments were produced with GENEIOUS (v.10.2.6) (Kearse et al., 2012), and phylogenetic images were produced with FIGTREE v.1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) and INKSCAPE (v.0.91) (<https://inkscape.org>).

Acknowledgements

We would like to thank Phil Taylor, Polychronis Rempoulakis, Jess Inskeep and Md. Jamil Hossain Biswas (Macquarie University) for providing field-collected *B. tryoni* samples, and three anonymous reviewers for their comments. Supercomputing resources were provided by the Phoenix HPC service at the University of Adelaide. The project is part funded by the National SITplus program

and Horticulture Innovation Australia using the research and development levy funds from the vegetable, apple and pear, citrus, strawberry, table grape, cherry and summerfruit industries, with co-investment from South Australian Research and Development Institute and Primary Industries and Regions South Australia. This work was supported by The Commonwealth Hill Trust and Grains Research and Development Corporation scholarships to CMW, The Hermon Slade Foundation (Grant 18/6 to SWB, TNMN, AC) and Australian Research Council (FT140101303 to SWB). The authors have no conflict of interest to declare.

References

- Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D. *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, **6**, e19315.
- Bedo, D.G. and Foster, G.G. (1985) Cytogenetic mapping of the male-determining region of *Lucilia cuprina* (Diptera: Calliphoridae). *Chromosoma*, **92**, 344–350.
- Beissinger, T.M., Hirsch, C.N., Sekhon, R.S., Foerster, J.M., Johnson, J.M., Muttoni, G. *et al.* (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, **193**, 1073–1081.
- Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.
- Bromham, L. and Penny, D. (2003) The modern molecular clock. *Nature Reviews Genetics*, **4**, 216–224.
- Buchman, A. and Akbari, O.S. (2019) Site-specific transgenesis of the *Drosophila melanogaster* Y-chromosome using CRISPR/Cas9. *Insect Molecular Biology*, **28**, 65–73.
- Carvalho, A.B. and Clark, A.G. (2005) Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. *Science*, **307**, 108–110.
- Carvalho, A.B., Koerich, L.B. and Clark, A.G. (2009) Origin and evolution of Y chromosomes: *Drosophila* tales. *Trends in Genetics*, **25**, 270–277.
- Chang, C.H. and Larracuente, A.M. (2019) Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics*, **211**, 333–348.
- Charlesworth, B. (1996) The evolution of chromosomal sex determination and dosage compensation. *Current Biology*, **6**, 149–162.
- Clarke, A.R. (2017) Why so many polyphagous fruit flies (Diptera: Tephritidae)? A further contribution to the 'generalism' debate. *Biological Journal of the Linnean Society*, **120**, 245–257.
- Concha, C. and Scott, M.J. (2009) Sexual development in *Lucilia cuprina* (Diptera, Calliphoridae) is controlled by the transformer gene. *Genetics*, **182**, 785–798.
- Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

- Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, **24**, 332–340.
- Dominiak, B.C. and Daniels, D. (2012) Review of the past and present distribution of Mediterranean fruit fly (*Ceratitis capitata* Wiedemann) and Queensland fruit fly (*Bactrocera tryoni* Froggatt) in Australia. *Australian Journal of Entomology*, **51**, 104–115.
- Dominiak, B.C. and Ekman, J.H. (2013) The rise and demise of control options for fruit fly in Australia. *Crop Protection*, **51**, 57–67.
- Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88–e88.
- Dupuis, J.R., Bremer, F.T., Kauwe, A., San Jose, M., Leblanc, L., Rubinoff, D. et al. (2018) HIMAP: robust phylogenomics from highly multiplexed amplicon sequencing. *Molecular Ecology Resources*, **18**, 1000–1019.
- Gabrieli, P., Gomulski, L.M., Bonomi, A., Siciliano, P., Scolari, F., Franz, G. et al. (2011) Interchromosomal duplications on the *Bactrocera oleae* Y chromosome imply a distinct evolutionary origin of the sex chromosomes compared to *Drosophila*. *Plos One*, **6**, e17747.
- Galizi, R., Hammond, A., Kyrou, K., Taxiarchi, C., Bernardini, F., O’Loughlin, S.M. et al. (2016) A CRISPR-Cas9 sex-ratio distortion system for genetic control. *Scientific Reports*, **6**, 31139.
- Gamble, T. and Zarkower, D. (2014) Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources*, **14**, 902–913.
- Gilchrist, A.S., Shearman, D.C.A., Frommer, M., Raphael, K.A., Deshpande, N.P., Wilkins, M.R. et al. (2014) The draft genome of the pest tephritid fruit fly *Bactrocera tryoni*: resources for the genomic analysis of hybridising species. *BMC Genomics*, **15**, 1153.
- Godzik, A. and Li, W. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the TRINITY platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.
- Hall, A.B., Basu, S., Jiang, X., Qi, Y., Timoshevskiy, V.A., Biedler, J.K. et al. (2015) A male-determining factor in the mosquito *Aedes aegypti*. *Science*, **348**, 1268–1270.
- Hall, A.B., Qi, Y., Timoshevskiy, V., Sharakhova, M.V., Sharakhov, I.V. and Tu, Z. (2013) Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics*, **14**, 273–273.
- Hancock, D.L., Hamacek, E.L., Lloyd, A.C. and Elson-Harris, M.M. (2000) In: *The Distribution and Host Plants of Fruit Flies (Diptera: Tephritidae) in Australia*. Brisbane, QLD: Queensland Department of Primary Industries.
- Hasegawa, M., Kishino, H. and Yano, T.-a. (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.
- Heled, J. and Drummond, A.J. (2011) Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology*, **61**, 138–149.
- Jain, M., Olsen, H.E., Turner, D.J., Stoddart, D., Bulazel, K.V., Paten, B. et al. (2018) Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, **36**, 321–323.
- Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S. et al. (2012) GENEIOUS BASIC: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.
- Kim, M., Semple, I., Kim, B., Kiers, A., Nam, S., Park, H.-W. et al. (2015) *Drosophila* Gyf/GRB10 interacting GYF protein is an autophagy regulator that controls neuron and muscle homeostasis. *Autophagy*, **11**, 1358–1372.
- Kjer, K.M., Ware, J.L., Rust, J., Wappler, T., Lanfear, R., Jermin, L. S. et al. (2015) Response to comment on ‘Phylogenomics resolves the timing and pattern of insect evolution’. *Science*, **349**, 487–487.
- Koerich, L., Wang, X., Clark, A. and Carvalho, A. (2008) Low conservation of gene content in the *Drosophila* Y chromosome. *Nature*, **456**, 949–951.
- Koerich, L.B., Dupim, E.G., Faria, L.L., Dias, F.A., Dias, A.F., Trindade, G.S. et al. (2016) First report of Y-linked genes in the kissing bug *Rhodnius prolixus*. *BMC Genomics*, **17**, 100.
- Kofler, M.M. and Freund, C. (2006) The GYF domain. *FEBS Journal*, **273**, 245–256.
- Krosch, M.N., Schutze, M.K., Armstrong, K.F., Graham, G.C., Yeates, D.K. and Clarke, A.R. (2012) A molecular phylogeny for the tribe Dacini (Diptera: Tephritidae): systematic and biogeographic implications. *Molecular Phylogenetics and Evolution*, **64**, 513–523.
- Krzywinska, E., Dennison, N.J., Lycett, G.J. and Krzywinski, J. (2016) A maleness gene in the malaria mosquito *Anopheles gambiae*. *Science*, **353**, 67–69.
- Lagos, D., Koukidou, M., Savakis, C. and Komitopoulou, K. (2007) The transformer gene in *Bactrocera oleae*: the genetic switch that determines its sex fate. *Insect Molecular Biology*, **16**, 221–230.
- Li, J.W. and Handler, A.M. (2017) Temperature-dependent sex-reversal by a *transformer-2* gene-edited mutation in the spotted wing drosophila, *Drosophila suzukii*. *Scientific Reports*, **7**, 12363.
- Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C. et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.
- Moadeli, T., Taylor, P.W. and Ponton, F. (2017) High productivity gel diets for rearing of Queensland fruit fly, *Bactrocera tryoni*. *Journal of Pest Science*, **90**, 507–520.
- Morrow, J.L., Riegler, M., Frommer, M. and Shearman, D.C.A. (2014a) Expression patterns of sex-determination genes in single male and female embryos of two *Bactrocera* fruit fly species during early development. *Insect Molecular Biology*, **23**, 754–767.
- Morrow, J.L., Riegler, M., Gilchrist, A.S., Shearman, D.C. and Frommer, M. (2014b) Comprehensive transcriptome analysis

- of early male and female *Bactrocera jarvisi* embryos. *BMC Genomics*, **15**(Suppl 2), S7.
- Nardi, F., Carapelli, A., Boore, J.L., Roderick, G.K., Dallai, R. and Frati, F. (2010) Domestication of olive fly through a multi-regional host shift to cultivated olives: comparative dating using complete mitochondrial genomes. *Molecular Phylogenetics and Evolution*, **57**, 678–686.
- Norrbom, A.L. (1994) New genera of Tephritidae (Diptera) from Brazil and Dominican amber, with phylogenetic analysis of the tribe Ortalotrypetini. *Insecta Mundi*, **8**, 1–15.
- Pane, A., Salvemini, M., Bovi, P.D., Polito, C. and Saccone, G. (2002) The *transformer* gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. *Development*, **129**, 3715–3725.
- Peng, W., Zheng, W., Handler, A.M. and Zhang, H. (2015) The role of the transformer gene in sex determination and reproduction in the tephritid fruit fly, *Bactrocera dorsalis* (Hendel). *Genetica*, **143**, 717–727.
- Poinar, G.O. (1992) In: *Life in Amber*. Stanford, CA: Stanford University Press, pp. 368.
- Pollard, D.A., Iyer, V.N., Moses, A.M. and Eisen, M.B. (2006) Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics*, **2**, e173–e173.
- Pyl, P.T., Anders, S. and Huber, W. (2014) HTSEQ – a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
- Rambaut, A., Drummond, A.J., Xie, D., Baele, G. and Suchard, M. A. (2018) Posterior summarization in Bayesian phylogenetics using TRACER 1.7. *Systematic Biology*, **67**, 901–904.
- Raphael, K.A., Shearman, D.C.A., Gilchrist, A.S., Sved, J.A., Morrow, J.L., Sherwin, W.B. et al. (2014) Australian endemic pest tephritids: genetic, molecular and microbial tools for improved sterile insect technique. *BMC Genetics*, **15**, S9.
- R-Core-Team (2018) In: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) EDGER: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H. S., Waterston, R.H. et al. (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, **423**, 873–876.
- Salvemini, M., D'Amato, R., Petrella, V., Ippolito, D., Ventre, G., Zhang, Y. et al. (2014) Subtractive and differential hybridization molecular analyses of *Ceratitis capitata* XX/XY versus XX embryos to search for male-specific early transcribed genes. *BMC Genetics*, **15**(Suppl. 2), S5.
- Salvemini, M., Robertson, M., Aronson, B., Atkinson, P., Polito, L. C. and Saccone, G. (2009) *Ceratitis capitata transformer-2* gene is required to establish and maintain the autoregulation of *Cctra*, the master gene for female sex determination. *The International Journal of Developmental Biology*, **53**, 109–120.
- San Jose, M., Doorenweerd, C., Leblanc, L., Barr, N., Geib, S. and Rubinoff, D. (2018) Incongruence between molecules and morphology: a seven-gene phylogeny of Dacini fruit flies paves the way for reclassification (Diptera: Tephritidae). *Molecular Phylogenetics and Evolution*, **121**, 139–149.
- Sanchez, L. (2008) Sex-determining mechanisms in insects. *The International Journal of Developmental Biology*, **52**, 837–856.
- Shapiro, B., Rambaut, A. and Drummond, A.J. (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, **23**, 7–9.
- Sharma, A., Heinze, S.D., Wu, Y., Kohlbrenner, T., Morilla, I., Brunner, C. et al. (2017) Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science*, **356**, 642–645.
- Shearman, D.C.A. (2002) The evolution of sex determination systems in dipteran insects other than *Drosophila*. *Genetica*, **116**, 25–43.
- Skov, L., Danish Pan Genome Consortium and Schierup, M.H. (2017) Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLoS Genetics*, **13**, e1006834.
- Stamatakis, A. (2014) RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, **30**, 1312–1313.
- Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, **33**, W465–W467.
- Sved, J.A., Chen, Y.Z., Shearman, D., Frommer, M., Gilchrist, A.S. and Sherwin, W.B. (2016) Extraordinary conservation of entire chromosomes in insects over long evolutionary periods. *Evolution*, **70**, 229–234.
- Tobler, R., Nolte, V. and Schlotterer, C. (2017) High rate of translocation-based gene birth on the *Drosophila* Y chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 11721–11726.
- Tong, K.J., Duchêne, S., Ho, S.Y.W. and Lo, N. (2015) Comment on 'Phylogenomics resolves the timing and pattern of insect evolution'. *Science*, **349**, 487–487.
- Tsoumani, K.T., Drosopoulou, E., Bourtzis, K., Gariou-Papalexidou, A., Mavragani-Tsipidou, P., Zacharopoulou, A. et al. (2015) *Achilles*, a new family of transcriptionally active retrotransposons from the olive fruit fly, with Y chromosome preferential distribution. *PLoS One*, **10**, e0137050.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. et al. (2012) PRIMER3 – new capabilities and interfaces. *Nucleic Acids Research*, **40**, e115–e115.
- Vicoso, B. and Bachtrog, D. (2015) Numerous transitions of sex chromosomes in Diptera. *PLoS Biology*, **13**, e1002078.
- Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Kliuchnikov, G. et al. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35**, 543–548.
- Wickham, H. (2016) In: *GGPLOT2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.
- Willhoeft, U. and Franz, G. (1996) Identification of the sex-determining region of the *Ceratitis capitata* Y chromosome by deletion mapping. *Genetics*, **144**, 737–745.
- Yaakop, S., Ibrahim, N.J., Shariff, S. and Zain, B.M.M. (2015) Molecular clock analysis on five *Bactrocera* species flies (Diptera: Tephritidae) based on combination of COI and NADH sequences. *Oriental Insects*, **49**, 150–164.

- Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simao, F.A., Ioannidis, P. *et al.* (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, **45**, D744–D749.
- Zhang, B., Liu, Y.H., Wu, W.X. and Le Wang, Z. (2010) Molecular phylogeny of the *Bactrocera* species (Diptera: Tephritidae: Dacini) inferred from mitochondrial sequences of 16S rDNA and COI sequences. *The Florida Entomologist*, **93**, 369–377.
- Zhao, J.T., Frommer, M., Sved, J.A. and Zacharopoulou, A. (1998) Mitotic and polytene chromosome analyses in the Queensland fruit fly, *Bactrocera tryoni* (Diptera: Tephritidae). *Genome*, **41**, 510–526.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Density plots displaying the proportion of WGS reads from males relative to females (MAPQ \geq 20). Values were normalized between zero and one, by taking the number of reads for each sex and dividing (male reads/(male+female reads)). Values of 1.0 indicate that only male reads align and values approaching zero indicate excess female read depth. Intervening white spaces are gaps in the reference genome. Smoothed conditional means of male read depth (orange line with 95% CI in grey), scaffold length, number of sites with coverage and median site depth are shown. Scaffolds Btry176, Btry280 and Btry315 are autosomal and indicate similar read depth between males and females.

Figure S2. Select Y-chromosome scaffold regions used for PCR primer design. Male (grey bars) and female (blue lines) sequence read depth from whole genome sequence datasets is shown. PCR primer locations are indicated with black boxes in each panel.

Figure S3. Copy number variation of *gyf* and *typo-gyf* in *B. tryoni*. Comparative CT quantitation was performed relative to the autosomal control gene *actin3*. Each female had two copies of *gyf* (2.07 ± 0.156) and lacked *typo-gyf* (data not show). Males had one copy of *gyf* (0.916 ± 0.081) on the X-chromosome and 41.40 (SD ± 4.87) copies of *typo-gyf* on the Y-chromosome.

Figure S4. Nucleotide alignment of 25 *typo-gyf* clones from a single *B. tryoni* male. Substitutions are highlighted in blue, green, yellow and red and deletions indicated with dashes. Each exon is annotated with a grey arrow and primers within predicted untranslated sequence are shown in green. See Figure S5 for protein alignment.

Figure S5 Nucleotide sequences of 25 *typo-gyf* clones were conceptually translated and 25 proteins aligned. Proteins ranged in length from 62–575 amino acids. See Figure S4 for nucleotide alignment.

Figure S6. Expression of a *typo-gyf* homolog in *Bactrocera jarvisi*. Morrow *et al.* (2014b) sequenced transcriptomes from pools of sexed *Bactrocera jarvisi* embryos that were 2–3 hours or 3–5 hours old. Eight paired end fastq libraries were downloaded from NCBI using the SRA-Toolkit (v2.8.2) command 'fastq-dump' with the option "--defline-seq '@\$sn[\$rn]/\$ri' ", and a reference transcriptome assembled with Trinity (v2.5.1) using the --trimmomatic option. Quality filtered reads from each sample was then quantified with the Trinity function align_and_estimate_abundance.pl, which uses the program Salmon (v0.8.2) to quantify transcripts. A total of 16813 genes were reportedly expressed in at least one sample using the Trinity function abundance_estimates_to_matrix.pl. Statistical analysis of transcripts was performed using edgeR (Robinson *et al.*,) using the Trinity function 'run_DE_analysis.pl'. A. An MA plot generated using edgeR showing differential expression between female and male 3–5 hour embryos. The transcript with the highest differential expression is TRINITY_DN12000_c0_g0 and represents unspliced isoforms of the *typo-gyf* homolog. Female 3–5 hour embryo replicates both had zero read counts; male 3–5 hour embryo replicates had 447 and 74 counts. The three red points are the only three genes with false discovery rates (FDR) below 0.05. B. Nucleotide alignment of *typo-gyf* transcripts from *B. tryoni* 24 hour embryos and *B. jarvisi* 3–5 hour embryos. The intron is not spliced in either transcript, introducing a premature stop codon if protein translation occurs. The intron begins with characteristic donor splice site 'GT' and ends with the acceptor splice site 'AG'. The two sequences show 94.7% identity. *Bactrocera tryoni* 24 hour embryos also express spliced transcripts that lack the intron.

Figure S7. Maximum Likelihood trees of 16 dipteran species. A. ML tree of thirteen mitochondrial genes including ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6 (8357 bp alignment). B. ML tree of 116 nuclear genes (155.7 kb). The tree topology differs for *B. jarvisi*, *B. dorsalis* and *B. kraussi*, as indicated with red lines.

Table S1. GBS data from three single pair *B. tryoni* crosses identified 55 candidate Y-chromosome scaffolds with >99% male read bias. Candidate Y-scaffolds supported by WGS sequencing in Fig. S1 are indicated.

Table S2. Whole genome re-sequencing statistics and analysis.

Table S3. Primer sequences for *typo-gyf* and *gyf*.

Table S4. Genomes and transcriptomes used for identification of 116 nuclear BUSCO genes and 13 mitochondrial genes.

Table S5. Evidence for *typo-gyf* and *gyf* transcripts.

Table S6. Percent pairwise amino acid identity between *Bactrocera* TYPO-GYF and *B. tryoni* GYF.

Table S7. Single copy nuclear genes used for Bayesian phylogeny (Fig.) and Maximum Likelihood trees (Fig. S7B).

Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore

Minsheng You ^{1,2,20}✉, Fushi Ke^{1,2,20}, Shijun You ^{1,2,3,20}, Zhangyan Wu^{4,20}, Qingfeng Liu ^{4,20}, Weiyi He ^{1,2,20}, Simon W. Baxter^{1,2,5}, Zhiguang Yuchi ^{1,6}, Liette Vasseur ^{1,2,7}✉, Geoff M. Gurr ^{1,2,8}✉, Christopher M. Ward ⁹, Hugo Cerda^{1,2,10}, Guang Yang^{1,2}, Lu Peng^{1,2}, Yuanchun Jin⁴, Miao Xie^{1,2}, Lijun Cai^{1,2}, Carl J. Douglas^{1,2,3,21}, Murray B. Isman ¹¹, Mark S. Goettel^{1,2,12}, Qisheng Song ^{1,2,13}, Qinghai Fan^{1,2,14}, Gefu Wang-Pruski^{1,2,15}, David C. Lees¹⁶, Zhen Yue ⁴✉, Jianlin Bai^{1,2}, Tiansheng Liu^{1,2}, Lianyun Lin^{1,5}, Yunkai Zheng^{1,2}, Zhaohua Zeng^{1,17}, Sheng Lin^{1,2}, Yue Wang^{1,2}, Qian Zhao^{1,2}, Xiaofeng Xia^{1,2}, Wenbin Chen^{1,2}, Lilin Chen^{1,2}, Mingmin Zou^{1,2}, Jinying Liao^{1,2}, Qiang Gao⁴, Xiaodong Fang⁴, Ye Yin⁴, Huanming Yang^{4,17,19}, Jian Wang^{4,18,19}, Liwei Han^{1,2}, Yingjun Lin^{1,2}, Yanping Lu^{1,2} & Mousheng Zhuang^{1,2}

The diamondback moth, *Plutella xylostella* is a cosmopolitan pest that has evolved resistance to all classes of insecticide, and costs the world economy an estimated US \$4-5 billion annually. We analyse patterns of variation among 532 *P. xylostella* genomes, representing a worldwide sample of 114 populations. We find evidence that suggests South America is the geographical area of origin of this species, challenging earlier hypotheses of an Old-World origin. Our analysis indicates that *Plutella xylostella* has experienced three major expansions across the world, mainly facilitated by European colonization and global trade. We identify genomic signatures of selection in genes related to metabolic and signaling pathways that could be evidence of environmental adaptation. This evolutionary history of *P. xylostella* provides insights into transoceanic movements that have enabled it to become a worldwide pest.

¹ State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Institute of Applied Ecology, Fujian Agriculture and Forestry University, Fuzhou 350002, China. ² Joint International Research Laboratory of Ecological Pest Control, Ministry of Education, Fuzhou 350002, China. ³ Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. ⁴ BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China. ⁵ School of BioSciences, The University of Melbourne, Melbourne, VIC 3010, Australia. ⁶ Tianjin Key Laboratory for Modern Drug Delivery & High-Efficiency, Collaborative Innovation Center of Chemical Science and Engineering, School of Pharmaceutical Science and Technology, Tianjin University, Tianjin 300072, China. ⁷ Department of Biological Sciences, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, ON L2S 3A1, Canada. ⁸ Graham Centre, Charles Sturt University, Orange, NSW 2800, Australia. ⁹ School of Biological Sciences, University of Adelaide, Adelaide, Australia. ¹⁰ Instituto Superior de Formación Docente Salomé Ureña (ISFODOSU), Recinto Lus Napoleón Núñez Molina, Carretera Duarte, Km 10 1/2, Municipio de Licey Al Medio, Provincia de Santiago, República Dominicana. ¹¹ Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. ¹² Agriculture and Agri-Food Canada, Lethbridge Research Centre, Lethbridge, AB, Canada. ¹³ Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA. ¹⁴ Plant Health & Environment Laboratory, Ministry for Primary Industries, Auckland, New Zealand. ¹⁵ Department of Plant, Food, and Environmental Sciences, Faculty of Agriculture, Dalhousie University, PO Box 550, Truro, NS B2N 5E3, Canada. ¹⁶ Natural History Museum, Cromwell Road, South Kensington, SW7 5BD London, UK. ¹⁷ Institute of Plant Protection, Fujian Academy of Agricultural Sciences, Fuzhou 350003, China. ¹⁸ BGI-Shenzhen, Shenzhen 518083, China. ¹⁹ James D. Watson Institute of Genome Sciences, Hangzhou 310058, China. ²⁰ These authors contributed equally: Minsheng You, Fushi Ke, Shijun You, Zhangyan Wu, Qingfeng Liu, Weiyi He. ²¹ Deceased: Carl J. Douglas. ✉email: msyou@fafu.edu.cn; lvasseur@brocku.ca; ggurr@csu.edu.au; yuezhen@bgi.com

The diamondback moth, *Plutella xylostella* (L.) (Lepidoptera: Plutellidae), is the most widely distributed species among all butterflies and moths on Earth¹. This major pest is an oligophagous herbivore of cultivated and wild cruciferous plants (Brassicaceae), including many economically important food crops such as cabbage, cauliflower, and rapeseed^{1,2}. The total economic cost of its damage and management worldwide is estimated at US \$4–5 billion per year^{1,3}. This is the first species for which its field populations have been documented to have evolved resistance to DDT, the iconic chemical insecticide of the 1950s⁴ and to toxins of *Bacillus thuringiensis* (Bt) developed as an insecticide for control of Lepidoptera in the 1990s⁵. *P. xylostella* has now developed resistance to all major classes of insecticide and is increasingly difficult to control in the field¹.

Although intensive research has been done on the biology, ecology and management of *P. xylostella* during recent decades^{1,2,6}, our knowledge of its geographical origin and how it has become such a highly successful pest in all continents except Antarctica remains surprisingly incomplete and highly controversial^{7–10}. To date, little is known about the global patterns of genomic variation in this species, which is essential for understanding the evolutionary history of *P. xylostella* together with the genetic basis of its rapid adaptation to insecticides.

In this study, we identify the origin of *P. xylostella* as South America using a global sample collection and nuclear/mitochondrial genome sequencing of all individuals, along with *COI* sequences for these and other specimens from BOLD. Further, we analyze the nuclear genomes of our specimens combined with geographical and historical information to reveal its dispersal routes and the progressive timing of global expansion. Based on the sequenced SNPs, we investigate the genomic signatures of selection to address the underlying mechanism associated with the local adaptation of this pest species.

Results

Global pattern of genomic variation. We first characterized the global pattern of variation among 532 genomes of *P. xylostella* using a worldwide sample of specimens collected from different locations (sites) in a stratified fashion reflecting a diverse range of biogeographical regions (Fig. 1a, Supplementary Fig. 1 and Supplementary Table 1) and covering an extensive scope of the eco-climatic index (Supplementary Fig. 2). Each of the individual genomes was sequenced with the Illumina sequencing system (HiSeq 2000) to produce 90-bp paired-end raw reads (Supplementary Fig. 1). A total of 1,797 Gb quality filtered reads (Supplementary Table 2) were mapped to the *P. xylostella* reference genome¹¹ using Stampy v1.0.27¹². Individuals with low mapping rate or coverage (<60%) were excluded, and a total of 532 individuals were retained for variant discovery (Supplementary Table 2). After calibrating and filtering of the low-quality variants (Supplementary Fig. 1), we generated a genomic dataset containing 40,107,925 SNPs and 22,736,441 indels (Supplementary Tables 3 and 4), representing one variant on average in every six bp of the reference genome¹¹. This is the densest variant map for any organism, including the recently released data of human¹³ and *Arabidopsis thaliana*¹⁴. The global pattern of genomic variation (Fig. 1b), regional diversity of individual-based SNPs (Fig. 1c), and low ratio of shared SNPs (7.20%) among different geographical populations (Supplementary Table 5) revealed a high level of polymorphism that provided the capacity for *P. xylostella* to readily expand and adapt to different habitats worldwide.

Geographical origin. An earlier study proposed that *P. xylostella* originated from Mediterranean Europe⁸ as many Brassicaceae crops were first domesticated in the this region. Other studies

predicted a South African⁹ or Chinese origin⁷ based on the regional diversity of indigenous Brassicaceae hosts or parasitoids of *P. xylostella*. An mtDNA-based analysis had supported claims of Africa as the possible area of origin of the species but used as few as 13 sampling sites worldwide without samples from South America¹⁰. A much larger and more representative collection of samples was required for accurate identification of the pest's geographical origin and better understanding of the evolutionary history of *P. xylostella*.

We extended these efforts to conduct a genomic study with high-quality nuclear SNPs of the worldwide sample (Fig. 1a, Supplementary Fig. 1 and Supplementary Table 1). Using the nuclear SNP data, a neighbor-joining (NJ) tree was constructed for global *P. xylostella* populations with the congeneric *P. australiana* as an outgroup according to our *COI*-based phylogenetic analysis (see Methods). Results revealed that multiple individuals collected from South America (SA) formed a distinct and basal clade (Fig. 2a and Supplementary Fig. 3). This was further confirmed by mitochondrial genomic data using the same specimens (Supplementary Fig. 4) and *COI* gene data from our specimens combined with additional published data (Supplementary Fig. 5). We also generated summary trees of different groups (see population genetic structure analysis below) based on 3,256 genome-wide local trees and revealed that the most prevalent topologies across the genome support *P. xylostella* populations of South America as the basal node closest to *P. australiana* (Supplementary Table 6). These results strongly suggest that *P. xylostella* originated in South America, where other endemic *Plutella* species are known¹⁵. Our study convincingly repolarizes the evolutionary history of *P. xylostella* from hypotheses of Old-World origin^{7–10} to the New World.

The Brassicaceae family contains >3,700 species found on all continents but Antarctica^{2,16}. In South America, with the richest cruciferous flora of the Southern Hemisphere^{16,17}, *P. xylostella* would have evolved on native Brassicales. Following European colonization of South America in the late 15th to early 16th century, the introduction and widespread use of domesticated Brassicaceae crops^{18–20} would have expanded host plant resources used by *P. xylostella* on this continent. Initially, the species would have been confined to South America, isolated by oceans and limited by habitat/eco-climatic constraints until human interference²¹.

Evolutionary and expansion history. Based on our phylogenetic analysis, in addition to the basal clade of South America (SA), we found four additional clades of North America (NA), Afro-Eurasia (A-E), South East Asia (SEA), and Oceania (OC) (Fig. 2a). These geographically clustered groups were supported by genetic structure analysis (Fig. 2b). A principal component analysis (PCA) (Fig. 2c) provided further evidence of the population structure of *P. xylostella* worldwide, with gene flow across the continent of Afro-Eurasia. The d_{XY} -based analysis²² showed the lowest genetic differentiation between SA *P. xylostella* and *P. australiana*, followed by NA *P. xylostella* and *P. australiana*, A-E *P. xylostella* and *P. australiana*, SE Asian *P. xylostella* and *P. australiana*, and OC *P. xylostella* and *P. australiana* (Fig. 2d), which outlined the global colonization process of *P. xylostella* populations.

To further investigate the demographic history of *P. xylostella*, we estimated the population sizes and divergence times using a pairwise sequentially Markovian coalescent (PSMC) model²³. It revealed a strikingly concordant history among geographical groups with a sharp decline following the last glacial maximum (LGM) in the early phase of evolution and a pattern of divergence in the recent past (with low resolution) among

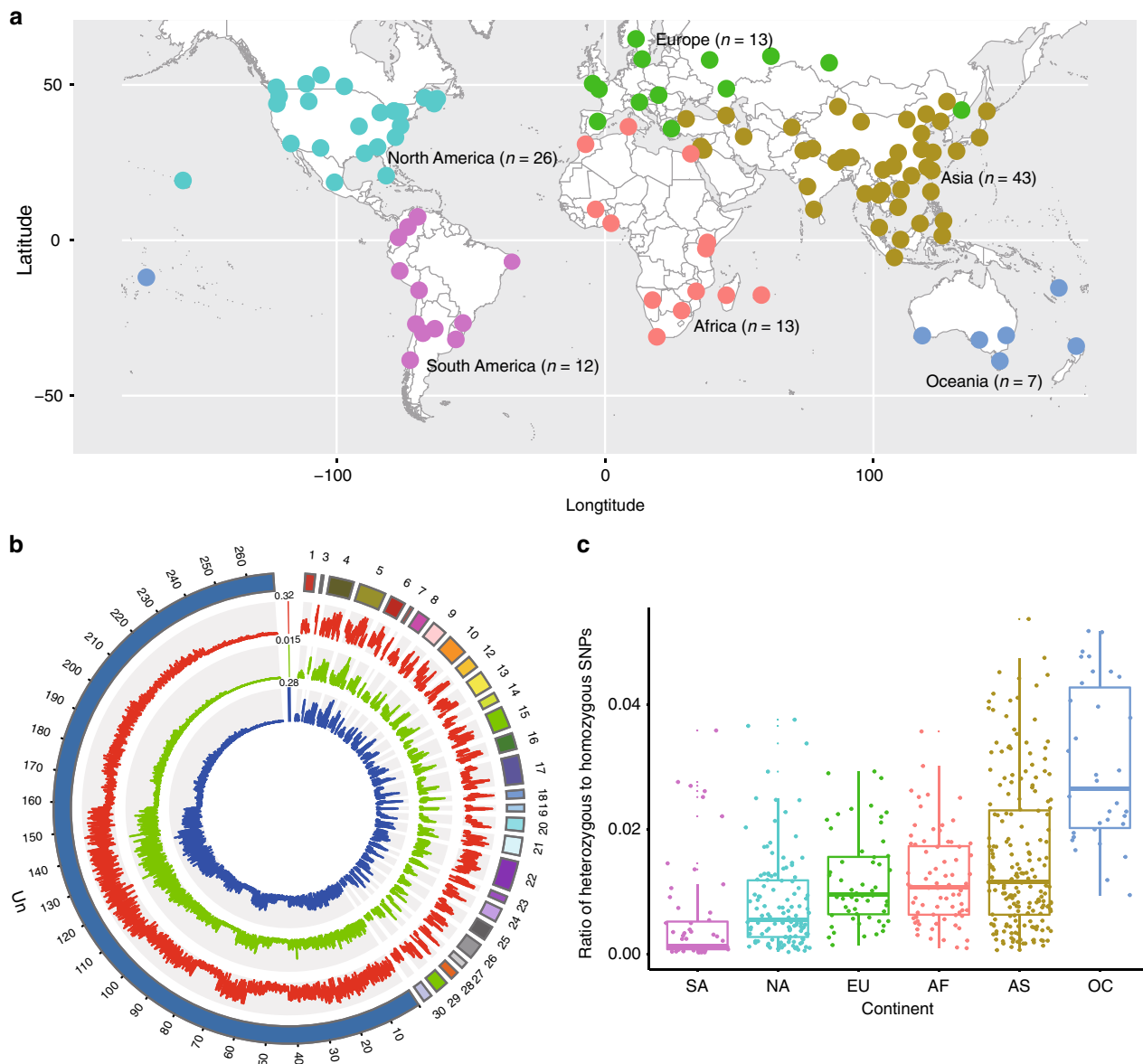


Fig. 1 Sampling locations and genomic variation of 532 *Plutella xylostella* individuals. **a** Illustration of 114 sampling locations showing that the *P. xylostella* specimens were collected from 55 countries across six of the seven continents in the world, i.e. excluding Antarctica. **b** Global patterns of the genomic variation. Circles from the outermost to innermost represent the reference genome of *P. xylostella* (including the partial sequences of 28 chromosomes and the scaffolds that were unable to be assigned (Un)), SNPs, nucleotide diversity and indels, respectively. **c** Ratio of individual-based heterozygous to homozygous SNPs in different continents of South America (SA), North America (NA), Europe (EU), Africa (AF), Asia (AS), and Oceania (OC). Boxes show the first and third quartile range (IQR) while whiskers extend to a maximum of 1.5 * IQR. Values for each of the individuals are shown as points surrounding boxplots. Source data are provided in the Source Data file. The map was generated with the *rworldmap* package v1.3-6⁷³.

different geographical groups (Supplementary Fig. 6). A recently published approach, SMC++²⁴, which estimates the historical population sizes with higher resolution in the recent past compared to other methods such as PSMC²⁵, was used to predict the historical population sizes and divergence times of different groups. We found that *P. xylostella* experienced three major expansions across the world, with both North American and Afro-Eurasian lineages splitting from the ancestral lineage approximately 500 years ago, followed by South East Asia and then Oceania (Fig. 3a).

P. xylostella has remarkable genetic plasticity²⁶ and high level of genomic variation¹¹ that enable it to rapidly adapt to local environments, potentially leading to the change in the levels of genetic diversity among geographical populations. After a new

founder event of *P. xylostella*, the sizes of derived populations tend to rapidly grow based on the SMC++ analysis (Fig. 3a). This may have led to accumulation of mutations that were not present in the ancestral population, especially reflecting the likelihood that the species would have been subject to new and diverse selection pressures from novel plant hosts, novel agonists (e.g., pathogens and parasitoids), habitats, and climates^{27,28}, as well as becoming established in extensive, heterogeneous geographical regions that limited mixing. Genetic admixture, which may generate novel genotypes^{27,29}, among the *P. xylostella* populations in both the Old-World and North America was frequent according to our phylogenetic reconstruction (Supplementary Fig. 3) and population structure analyses (Fig. 2b and c). Together, these effects help explain the pattern of increasing

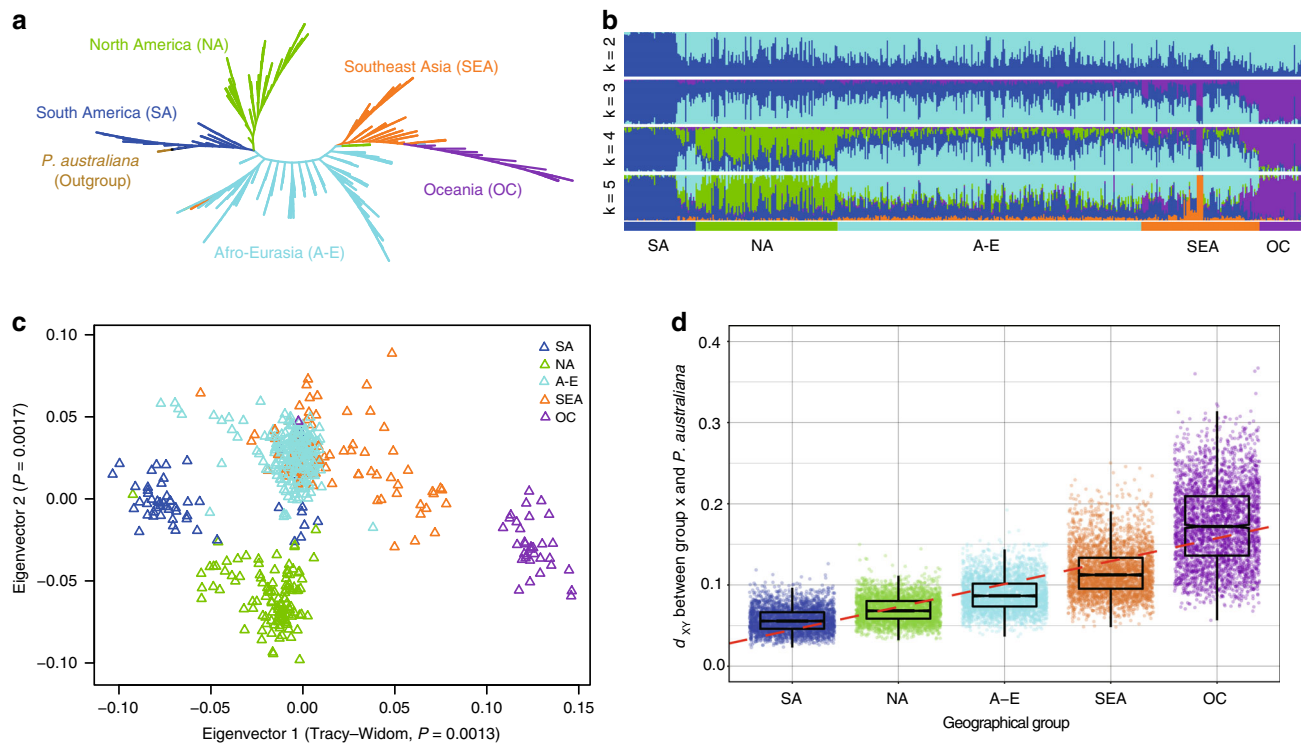


Fig. 2 Origin and global dispersal of *P. xylostella*. **a** Neighbor-joining phylogeny based on the nuclear genomes (SNPs) of 532 *P. xylostella* individuals collected worldwide, with two congeneric *P. australiana* individuals being used as an outgroup. The branch lengths are not scaled. See also Supplementary Fig. 3. **b** sNMF-based genetic structure and individual ancestry with colors in each column representing ancestry proportion over range of population sizes ($K=2-5$, with an optimal $K=5$). **c** PCA plot of the first two components generated by 2,839 SNPs. **d** d_{xy} calculated between each *P. xylostella* group and *P. australiana* based on 3,256 non-overlapping genome-wide windows. Boxes show the first and third quartile range (IQR) while whiskers extend to a maximum of 1.5 * IQR. Values for each window are shown as individual points surrounding boxplots. A simple linear model was fitted to the data (red dashed line), the line of best fit has an estimated R^2 value of 0.6149 and a slope of 0.02821. Source data are provided in the Source Data file.

genetic diversity in the range expansion process of *P. xylostella* populations (Fig. 1c).

Our phylogenetic and genetic analyses, PCA plot of the first two components, and d_{xy} -based analysis of differentiation (Fig. 2) supported the demographic evidence that the SA lineage was basal with the NA, A-E and SEA lineages diverging at progressively later stages and the OC lineage the most recent (Fig. 3a). These results were integrated with historical information^{18–20}, allowing us to propose a scenario of dispersal events for *P. xylostella* (Fig. 3b). We found that the major expansion events of *P. xylostella* were associated with human activities of agricultural production and trade. With European colonization, particularly the domestication of cruciferous crops with reduced glucosinolates and the introduction of Brassicaceae crops by European colonizers to South America^{18,30}, the original populations of *P. xylostella* in South America appear to have dispersed to and colonized various regions of the world (Fig. 3a and b). After colonizing the Mediterranean region, founder populations of *P. xylostella* likely dispersed across Europe, Western Asia, and Africa (Figs. 2 and 3)^{31,32}. Like the spreading trend of *A. thaliana*, the diversity of *P. xylostella* populations in Europe and Eurasia exhibits a latitudinal pattern along the east-west axis (Fig. 3b), which has been facilitated by the rapid expansion of agriculture^{14,33}. Around 200 years ago, independent dispersal events led the founder populations expanding eastwards into Asian countries first and then proceeding to Oceania (Figs. 2 and 3)^{34,35}, which corresponds to the most recent major region of colonization by Europeans. Records of Brassicaceae date from the “First Fleet” arrival in Australia in 1788 that carried produce and seeds of several *Brassica* species³⁵, and this was followed by

introduction and widespread cultivation of other brassicas by Chinese Australians³⁴. Relative to the predicted earliest time when a Plutellidae ancestor may have become a cruciferous specialist (~54–90 million years ago)^{11,36}, the recent expansion events of *P. xylostella* (~200–500 years ago) further indicate that it could have survived on the indigenous Brassicaceae plants in South America for a long time, possibly in the timeframe of millions of years, after its putative divergence from an ancestor shared with its closest known relative, *P. australiana*³⁷.

Genomic signatures of local adaptation. Based on our globally sampled genomic data, we found that *P. xylostella* populations across the world exhibited a dense map of variants (Supplementary Tables 3 and 4), high level of polymorphism (Fig. 1b, c and Supplementary Fig. 7), and rapid decay of linkage disequilibrium (LD; Supplementary Fig. 8). These findings suggest a large effective population size for this species. Considering its genetic heterozygosity and rapid insecticide-resistance evolution, this species is well suited for a study of evolutionary adaptation under strong environmental selection pressure^{38,39}. The intensive use of insecticides against *P. xylostella* has led to increased selection pressure for development of insecticide resistance^{1,2,4,40–45}. We identified a global pattern of adaptive variation shown by the frequency distribution of three reported SNPs associated with insecticide resistance^{46,47} (Supplementary Fig. 9). Such a global genotype distribution of three insecticide-resistance-related point mutation loci revealed that selection pressure resulting from insecticide applications had strong geographical dependence.

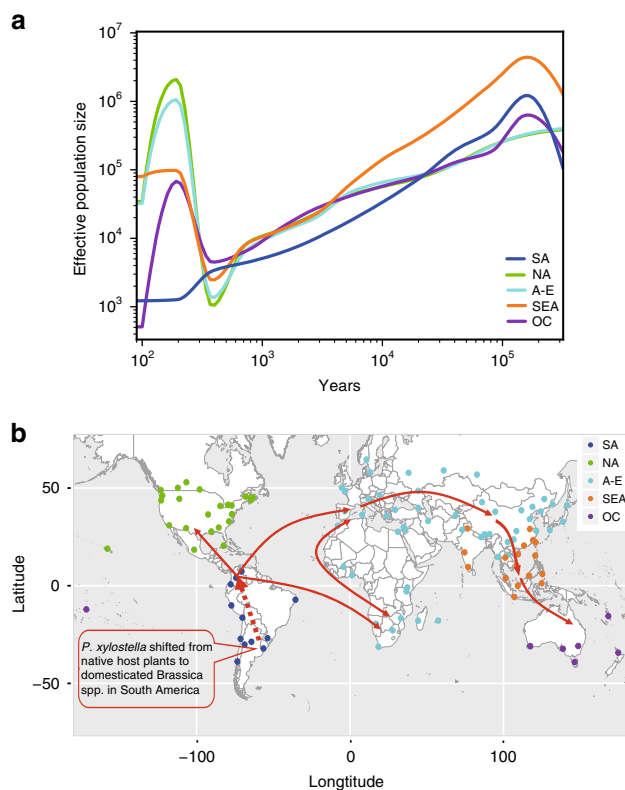


Fig. 3 Global expansion and demographic history of the *P. xylostella* populations. **a** Demographic history of *P. xylostella* illustrating the effective population sizes and divergence times based on multiple unphased individuals from different geographical groups and predicted by SMC++²⁴. Coordinates are logarithmically scaled. **b** A proposed scenario of global colonization of the *P. xylostella* populations. The red arrows denote the proposed dispersal events of *P. xylostella* from South America towards other continents based on phylogenetic result (Fig. 2a), population genetic analyses (Fig. 2b–d), and demographic history (Fig. 3a). Source data are provided in the Source Data file. The map was generated with the worldmap package v1.3-6⁷³.

To identify the genomic signatures of evolutionary adaptation for *P. xylostella*, we ran a genome-wide association study (GWAS) using the eigenvector1 of PCA as a “phenotype” (EigenGWAS)⁴⁸ to isolate a group of 75 individuals from Southeast Asia and Oceania (Supplementary Fig. 10). This reflects the fact that in these tropical and subtropical regions, cruciferous crops are massively and continuously grown year-round in a variety of cropping systems from backyard gardens to large-scale farms, resulting in favorable conditions for *P. xylostella* to develop and frequently outbreak throughout the year^{1,2}.

We identified 3,827 significantly differentiated SNPs ($P \leq 1e^{-8}$) (Supplementary Fig. 11a, right) with high level of genetic differentiation (F_{ST}) from 64,960 filtered SNPs (Supplementary Fig. 11a, left), which indicated that numerical distribution the significantly differentiated SNPs was proportionally similar to that of the filtered SNPs in each of the four genomic regions (Supplementary Fig. 11a). These outliers contained 1,179 candidate genes, being the most highly represented in metabolic and molecular signaling-related pathways according to the GO and KEGG analysis (shown with top 20 GO terms and KEGG pathways; Figs. 11b and 12). Among the 1,179 candidate genes under divergent selection we found 93 that were annotated in the published *P. xylostella* genome with known functions of detoxification of plant defense compounds and insecticide resistance¹¹. We then identified six genes with non-synonymous

SNPs in coding regions. Three of them, including carboxypeptidase A (Px005867), P450-CYP2 (Px002515), and juvenile hormone esterase (JHE, Px003448) have reliable structural templates available in the Protein Data Bank, which allowed us to create homology models for these three enzymes using Schrödinger software⁴⁹.

Signals of divergent selection for the three non-synonymous mutations were identified according to their global distribution of genotype frequency (Supplementary Fig. 14). Comparison of the predicted structures between wild-type (WT) and mutant (Mut) enzymes revealed the potential impacts of these three mutations on the structural changes of these enzymes (Supplementary Fig. 15), which provides a cue for further experimentally-based research to establish a functional relationship between mutations and insecticide resistance^{50,51}.

Discussion

The present study improves our understanding of the origin, evolution, and genetic bases of adaptation in *P. xylostella*, a species with worldwide importance for pest management and food safety. Using a global sample collection (532 individuals) covering all six continents where the species occurs and nuclear and mitochondrial genomes as well as *COI* sequencing of all individuals, we have identified the area of origin of *P. xylostella* as South America. The result contrasts with previous hypotheses that suggested the Mediterranean region⁸, South Africa^{9,10} or China⁷ as possible areas of origin of the species. Further, the phylogeographical profiling reveals that *P. xylostella* expansion events and timing have been facilitated by human socioeconomic activities. Genes in metabolic and molecular signaling-related pathways are putative candidates involved in evolutionary adaptation under the strong selective pressure of insecticides. Our results illustrate the utility of emerging genomic approaches to understand historical patterns of species expansion, and further address the underlying mechanisms associated with the worldwide dispersal of this notorious pest species.

Methods

Sample collection and DNA extraction. Based on the globally-distributed nature of *P. xylostella*, we developed a sampling plan with broad geographic scope (Fig. 1a and Supplementary Fig. 1). The global samples of *P. xylostella* were collected in 2012–2014 from 114 locations that cover broad regions throughout the world, with 13 samples from Africa including Madagascar, 43 samples from Asia, 13 samples from Europe, 26 samples from North America including Hawaii, 12 samples from South America, and 7 samples from Oceania (Fig. 1a and Supplementary Table 1). Our collection covers an extensive range of the eco-climatic index and areas that support differing numbers of annual generations, including those regions with year-round persistence of *P. xylostella* to others that are only seasonally suitable for growth and development of the species (Supplementary Fig. 2). Within each location, larvae, pupae, or adults were collected from cruciferous vegetable fields. Field-collected samples were morphologically inspected and genetically checked with *COI* sequences to confirm their identity.

The samples were preserved in 95% alcohol at -80°C prior to DNA extraction. At least five individuals from each sampling location were used for DNA extraction. For quality control, each individual was washed twice using double-distilled water, and then dissected to remove the midgut including its microbiome and parasitoids to eliminate potential DNA contamination (Supplementary Fig. 1). To avoid unintentional biases, the individuals were each allocated a code number (Supplementary Table 2) in a double-blind fashion to obscure the origin of the insect to all handlers and analysts who identified the insect, its DNA or any associated genomic data.

DNA was extracted from each individual using DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. DNA was eluted from the DNeasy Mini spin column in 200 μl TE buffer. Concentration and integrity of the total DNA for each individual was measured with a Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA) and agarose gel electrophoresis.

DNA barcoding and sequencing. A *Cytochrome Oxidase I (COI)* mitochondrial gene fragment of up to 658 bp was amplified and sequenced using Sanger sequencing, and then queried to the BOLD system⁵² to confirm the species identity for each individual. In the 468 sequences of this dataset [<https://doi.org/10.5883/DS-PLUT1>]^{37,52},

399 sequences belong to *P. xylostella* (BOLD:AAA1513 [http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:AAA1513]) and 58 to *P. australiana* (BOLD:AAC6876 [http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:AAC6876]) while the rest belongs to other potential outgroup taxa in the family Plutellidae.

Genomic sequencing was performed with Illumina HiSeq 2000 at BGI, Shenzhen, China, to produce 90 bp paired-end reads for every individual. Considering the cosmopolitan distribution of *P. xylostella*, we aimed to sequence a large number of individual genomes across various geographical locations with a 5–10× coverage for each individual, which is a strategy previously used for the 1000 human genomes project⁵³ and *Apis mellifera*⁵⁴. Two *P. australiana*³⁷ individuals were also sequenced with a 30× coverage and used as an outgroup for comparative analysis of genetic differentiation with the *P. xylostella* populations and construction of the phylogenetic tree. Sequencing libraries for each of the *P. xylostella* individuals were constructed according to the manufacturer's protocol. The quality and yield of the library were tested using the Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System.

Data filtering, mapping and SNP calling. Raw reads were processed to obtain clean reads using custom scripts. Poor reads with 10 ambiguous “N” bases, >40% low-quality bases, or identical sequences at the two ends were filtered out.

We artificially allocated scaffolds of the *P. xylostella* genome into 20 synthetic chromosomes, and the SNP calling, and subsequent analyses were all performed on these 20 “chromosomes”. Stampy (v1.0.27)¹² was employed to map the clean reads onto our *P. xylostella* reference genome (v2)¹¹ using default parameters. Subsequently, alignments for each of the individual samples were sorted with SortSam of Picard-tools (v-1.117, <https://sourceforge.net/projects/picard/>), and processed by removing duplicate reads with MarkDuplicates of Picard-tools. Reads with indels were realigned using RealignerTargetCreator and IndelRealigner in the Genome Analysis Toolkit (GATK v-3.2.2)⁵⁵ to avoid misalignment around indels. After realignment, base-quality scores were recalibrated using BaseRecalibrator based on the reference SNP set, which was generated using UnifiedGenotyper and SAMtools⁵⁶ from the 532 individuals. The sequencing and mapping statistics are summarized in Supplementary Table 2.

SNP calling was then performed using the GATK HaplotypeCaller with parameters `--emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter 128,000`. VariantRecalibrator was first used to create a Gaussian mixture model to examine the annotation values over a high-quality subset previously generated by UnifiedGenotyper and SAMtools and evaluate all input variants. ApplyRecalibration was used to designate the model parameters for each of the variants. Finally, VariantFiltration was used to filter the SNPs. The filtering settings were as follows: `QD < 2.0 | MQ < 40.0 | ReadPosRankSum < -8.0 | FS > 60.0 | HaplotypeScore > 13.0 | MQRankSum < -12.5`.

To resolve the origin of *P. xylostella*, two congeneric *P. australiana* individuals were used as an outgroup species. The sequencing reads of *P. australiana* were aligned to the *P. xylostella* genome using Stampy (v1.0.27)¹² with default parameters and reordered and sorted by Picard (<https://broadinstitute.github.io/picard/>). SOAPsnp (<http://soap.genomics.org.cn/soapsnp.html>) was then used to detect SNPs in each *P. australiana* individual with at least three supporting reads, and to assemble the consensus sequence for the *P. australiana* individuals based on the alignment of the sequencing reads with the *P. xylostella* genome. The genomic dataset of two *P. australiana* individuals and 532 *P. xylostella* individuals was used for phylogenetic tree construction.

To identify mitochondrial variants of *P. xylostella*, we also called SNPs using the mitochondrial genome of *P. xylostella* (GenBank KM023645 [<https://www.ncbi.nlm.nih.gov/search/all/?term=KM023645>]) as a reference. The same SNP calling procedure as done for nuclear SNP calling was employed, while a haplotype setting was used. The mitochondrial genome of *P. australiana* was reconstructed using MITObim⁵⁷ with a *P. australiana* COI barcode sequence as the seed and the *P. xylostella* mitochondrial genome as the reference.

Construction of the phylogenetic trees. Phylogenetic relationships of nuclear and mitochondrial genomes were analyzed among 532 individual samples of *P. xylostella* with two samples of *P. australiana* used as an outgroup. A phylogenetic tree based on the nuclear genomes was constructed (Fig. 2a and Supplementary Fig. 3), using the neighbor-joining (NJ) method⁵⁸, based on a genetic distance matrix (<https://github.com/BGI-shenzhen/VCF2Dis>), and calculated by the software PHYLP v3.695 (<http://evolution.genetics.washington.edu/phylip.html>). Mitochondrial genomes were also used for phylogenetic tree construction using the NJ method implemented in PHYLP, and a frequency tree (Supplementary Fig. 4) was generated using the consensus module with 1000 bootstraps.

To further confirm the origin and evolutionary relationships of *P. xylostella* populations based on nuclear and mitochondrial SNPs, a COI-based phylogenetic tree (Supplementary Fig. 5) was constructed based on NJ method with 1000 bootstraps using MEGA5⁵⁹. This tree included the sequences of 532 *P. xylostella* individuals collected worldwide and two *P. australiana* individuals collected in Australia, as well as individual sequences of five non-Australian *Plutella* species (with two individual sequences for each of *P. armoraciae*, *P. porrectella*, *P. geniatella*, *P. hyperboreella* and one of *P. notabilis*), *Eidophasia vanella*, *P. australiana*, and *P. xylostella* downloaded from BOLD [<https://doi.org/10.5883/DS-PLUTI>]^{37,52}, and

two undescribed taxa (‘kaloko’ and ‘napoopoo’) from Hawaii^{60,61} downloaded from GenBank, with accession codes AF019041 [<https://www.ncbi.nlm.nih.gov/nuccore/AF019041>] and AF019042 [<https://www.ncbi.nlm.nih.gov/nuccore/AF019042>].

Population genetic pattern analysis. Bi-allelic SNPs presenting >95% individuals with a minor allele frequency of over 0.2 in the dataset were kept using vcftools⁶² and used for population genetic structure analysis. We sampled one SNP from a 25 bp DNA window to generate loci independent of linkage disequilibrium. A total of 2,839 SNPs was retained for further analysis. The population genetic structure was analyzed using sNMF⁶³ with the pre-defined genetic clusters increased from $K = 2$ to $K = 8$ and illustrated with POPHELPER⁶⁴. Principal component analysis (PCA) was also conducted using PLINK⁶⁵ with the same dataset, and a Tracy-Widom test was used to determine the significant level of the eigenvectors. The results (Fig. 2b) further confirmed and supported the five geographically clustered groups of *P. xylostella* populations worldwide based on previous nuclear phylogenetic analysis.

We generated genome-wide summary trees of different groups based on local trees. Variants with a maximum missing rate of 70% were filtered, and then converted to Genomic Data Structure (GDS) using SeqArray⁶⁶. Local genetic distance matrix was calculated using R scripts (<https://github.com/CMWbio>) with a bin of 5,000 SNPs in a maximum interval of 100 kb. A total of 3,256 local trees were generated across the genome. TWISST⁶⁷ was then used to calculate topology weighting for each local tree with 1,000 iterations (Supplementary Table 6). We also calculated d_{XY} values²² in these 3,256 windows between five identified groups and the outgroup population to investigate genetic differentiation pattern during the global colonization of *P. xylostella* (Fig. 2d).

We presented the global genotype distribution of three previously reported SNPs associated with insecticide resistance^{46,47} to show the geographical dependence of these point mutation loci (G4946E, L1014F and T929I; Supplementary Fig. 9). G4946E in ryanodine receptor was involved in resistance to diamide⁴⁶, and L1014F and T929I in sodium channel were associated with resistance to pyrethroid⁴⁷.

Demographic history. We selected one individual with high sequencing depth from each group to estimate the demographic history of *P. xylostella* using Pairwise Sequentially Markovian Coalescence (PSMC)²³, with a generation time of 0.1 years and a mutation rate⁶⁸ of 8.4×10^{-9} (Supplementary Fig. 6). A recently published approach with higher resolution in the recent past compared to PSMC accuracy, SMC++^{24,25}, was used to predict the demographic history (or population sizes and divergence times) of *P. xylostella* based on multiple unphased individuals (Fig. 3a). Five previously defined groups (or clades) were used for the analysis. We used a mutation rate of 8.4×10^{-9} from *Drosophila*⁶⁸, and 10 generations of *P. xylostella* per year estimated with global observations and records^{1–3}. The short generation time of *P. xylostella* makes possible the reliable and precise estimation of effective population sizes in the recent past using the method of SMC++²⁵.

Identification of the loci under selection. An approach of genome-wide association study with the first eigenvector from the PCA as a “phenotype” (EigenGWAS)⁴⁸ was recently developed to identify single SNPs that contribute to the genetic differentiation (eigenvector) of two populations based on regression analysis. By using individual-level eigenvectors as phenotypes (Y in regression analysis) and single SNPs (X in regression analysis) in a linear regression, the resulting regression coefficients are equivalent to singular value decomposition (SVD) SNP effects and used to identify loci under selection along gradients of ancestry⁴⁸. EigenGWAS also used a correction parameter to filter out signals of population stratification (i.e. caused by geography/drift), which allows the loci under selection to be identified. This approach has been successfully used to identify the loci under divergent selection between the UK and Dutch populations of great tit (*Parus major*) for better understanding of how genetic signatures of selection translate into variation in fitness and phenotypes³⁸.

To identify the genomic signatures of selection for *P. xylostella*, 64,960 SNPs (Supplementary Fig. 11a) were obtained after filtering with a missing rate $\leq 5\%$ and a minor allele frequency ≥ 0.05 by vcftools⁶² from the genomic dataset of our global samples. Based on the filtered SNPs, we ran the approach of EigenGWAS using a stringent level of genome-wide significance threshold ($P \leq 1e^{-8}$)⁶⁹. A total of 3,827 loci (or outlier) under selection (Supplementary Fig. 11a) were identified for further functional annotation.

Based on the genomic database of *P. xylostella* (<http://iae.fafu.edu.cn/DBM/index.php>), we searched for candidate genes with the outliers. GO annotation and classification of the candidate genes were conducted using Blast2GO (version 2.5.0)⁷⁰ and WEGO⁷¹. Pathways of the candidate genes were identified (Supplementary Fig. 11b) using the KEGG database (<http://www.genome.jp/kegg/pathway.html>). The genes enriched in the first 20 GO terms and KEGG pathways, and the value of fixation index (F_{ST}) was calculated for each of the identified loci using vcftools⁶⁸ (Supplementary Figs. 11b; 12 and 13).

Homology modeling. The structural models for the wild-type carboxypeptidase A, P450-CYP2 and juvenile hormone esterase were created by Prime module of Schrödinger software using human carboxypeptidase structure (PDB ID: 1PCA), fish cytochrome P450 structure (PDB ID: 4R1Z) and human acetylcholinesterase

structure (PDB ID: 4BDT) as templates, respectively (Supplementary Fig. 15). The resistant-mutant models were developed by introducing mutations to the wild-type structural models, followed by further energy minimization using Chimera⁶⁹. All of the structural figures were also generated by Chimera⁶⁹.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw reads of all 532 sequenced individuals have been deposited in the CNSA (<https://db.cngb.org/cnsa/>) of CNGBdb with the accession code CNP0000018, and been synchronously deposited in the EMBL Nucleotide Sequence Database (ENA) (<https://www.ebi.ac.uk/ena/>) with the accession code PRJEB24034. Sequences of five non-Australian *Plutella* species (with two individual sequences for each of *P. armoraciae*, *P. porrectella*, *P. geniatella*, *P. hyperboreella* and one of *P. notabilis*), *Eidophasia vanella*, *P. australiana*, and *P. xylostella* were downloaded from BOLD [<https://doi.org/10.5883/DS-PLUT1>]^{37,52}. Sequences of two undescribed taxa ('kaloko' and 'napoopoo') from Hawaii^{60,61} were downloaded from the GenBank, with accession codes AF019041 [<https://www.ncbi.nlm.nih.gov/nuccore/AF019041>] and AF019042 [<https://www.ncbi.nlm.nih.gov/nuccore/AF019042>]. The source data underlying Figs. 1b, c, 2b–d, 3a as well as Supplementary Figs. 2, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15 are provided as a Source Data file.

Received: 1 June 2019; Accepted: 15 April 2020;

Published online: 08 May 2020

References

- Furlong, M. J., Wright, D. J. & Dossdall, L. M. Diamondback moth ecology and management: problems, progress, and prospects. *Annu. Rev. Entomol.* **58**, 517–541 (2013).
- Talekar, N. S. & Shelton, A. M. Biology, ecology, and management of the diamondback moth. *Annu. Rev. Entomol.* **38**, 275–301 (1993).
- Zalucki, M. P. & Furlong, M. J. Predicting outbreaks of a migratory pest: an analysis of DBM distribution and abundance revisited. In *Proceedings of The Sixth International Workshop on Management of Diamondback Moth and Other Crucifer Pests* (eds Srinivasan, R., Shelton, A. M. & Collins, H. L.) 8–14 (AVRDC, 2011).
- Ankersmit, G. W. DDT-resistance in *Plutella maculipennis* (Curt.) (Lep.) in Java. *Bull. Entomol. Res.* **44**, 421–425 (1953).
- Tabashnik, B. E., Cushing, N. L., Finson, N. & Johnson, M. W. Field development of resistance to *Bacillus thuringiensis* in diamondback moth (Lepidoptera: Plutellidae). *J. Econ. Entomol.* **83**, 1671–1676 (1990).
- Li, Z., Feng, X., Liu, S., You, M. & Furlong, M. J. Biology, ecology, and management of the diamondback moth in China. *Annu. Rev. Entomol.* **61**, 277–296 (2016).
- Liu, S., Wang, X., Guo, S., He, J. & Shi, Z. Seasonal abundance of the parasitoid complex associated with the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae) in Hangzhou, China. *Bull. Entomol. Res.* **90**, 221–231 (2000).
- Hardy, J. E. *Plutella maculipennis*, Curt., its natural and biological control in England. *Bull. Entomol. Res.* **29**, 343–372 (2009).
- Kfir, R. Origin of the diamondback moth (Lepidoptera: Plutellidae). *Ann. Entomol. Soc. Am.* **91**, 164–167 (1998).
- Juric, I., Salzburger, W. & Balmer, O. Spread and global population structure of the diamondback moth *Plutella xylostella* (Lepidoptera: Plutellidae) and its larval parasitoids *Diadegma semiclausum* and *Diadegma fenestrata* (Hymenoptera: Ichneumonidae) based on mtDNA. *Bull. Entomol. Res.* **107**, 155–164 (2017).
- You, M. S. et al. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225 (2013).
- Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
- 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Alonso-Blanco, C. et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
- Meyrick, E. *Exotic Microlepidoptera* (Taylor & Francis Group, 1931).
- Appel, O. & Al-Shehbaz, I. A. Cruciferae. In *The Families and Genera of Vascular Plants vol.5 Flowering Plants · Dicotyledons* (eds Kubitzki, K., Rohwer, J. G. & Bittrich, V.) 75–174 (Springer, 2003).
- Al-Shehbaz, I. A., Cano, A., Trinidad, H. & Navarro, E. New species of *Brayopsis*, *Descurainia*, *Draba*, *Neuontobotrys* and *Weberbaueria* (Brassicaceae) from Peru. *Kew Bull.* **68**, 219–231 (2013).
- Super, J. C. *Food, Conquest, and Colonization in the Sixteenth-century Spanish America* 192 (University of New Mexico Press, Albuquerque, 1988).
- de Oviedo y Valdés, G. F. *La Historia General de Las Indias* (1535).
- Gallagher, D. American plants in Sub-Saharan Africa: a review of the archaeological evidence. *Azania: Archaeol. Res. Afr.* **51**, 24–61 (2016).
- Frenot, Y. et al. Biological invasions in the Antarctic: extent, impacts and implications. *Biol. Rev.* **80**, 45–72 (2005).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, 1987).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
- Patton, A. et al. Contemporary demographic reconstruction methods are robust to genome assembly quality: a case study in Tasmanian devils. *Mol. Biol. Evol.* **36**, 2906–2921 (2019).
- Henniges-Janssen, K. et al. Complex inheritance of larval adaptation in *Plutella xylostella* to a novel host plant. *Heredity* **107**, 421–432 (2011).
- Estoup, A. et al. Is there a genetic paradox of biological invasion? *Annu. Rev. Ecol. Syst.* **47**, 51–72 (2016).
- Lavergne, S. & Molofsky, J. Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proc. Natl Acad. Sci. USA* **104**, 3883–3888 (2007).
- Facon, B., Pointier, J. P., Jarne, P., Sarda, V. & David, P. High genetic variance in life-history strategies within invasive populations by way of multiple introductions. *Curr. Biol.* **18**, 363–67 (2008).
- Fahey, J. W., Zalcmann, A. T. & Talalay, P. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* **59**, 5–51 (2002).
- Dixon, G. R. *Vegetable Brassicas and Related Crucifers* (CABI, Wallingford, 2006).
- Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A. & Mummenhoff, K. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* **16**, 108–116 (2011).
- François, O., Blum, M. G., Jakobsson, M. & Rosenberg, N. A. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet.* **4**, e1000075 (2008).
- Wahlqvist, M. L. Asian migration to Australia: food and health consequences. *Asia Pac. J. Clin. Nutr.* **11**, S562–S568 (2002).
- Frost, A. *The First Fleet: The Real Story*. (Black Inc., 2012).
- Wang, X. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035 (2011).
- Landry, J. F. & Hebert, P. D. N. *Plutella australiana* (Lepidoptera, Plutellidae), an overlooked diamondback moth revealed by DNA barcodes. *Zookeys* **327**, 43–63 (2013).
- Bosse, M. et al. Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science* **358**, 365–368 (2017).
- Hawkins, N. J., Bass, C., Dixon, A. & Neve, P. The evolutionary origins of pesticide resistance. *Biol. Rev.* **94**, 135–155 (2018).
- Atumurirava, F. & Furlong, M. J. Diamondback moth resistance to commonly used insecticides in Fiji. In *Proceedings of The Sixth International Workshop on Management of Diamondback Moth and Other Crucifer Pests* (eds Srinivasan, R., Shelton, A. M. & Collins, H. L.) 216–221 (AVRDC, 2011).
- Heisswolf, S., Houlding, B. J. & Deuter, P. L. A decade of integrated pest management (IPM) in Brassica vegetable crops—the role of farmer participation in its development in southern Queensland, Australia. In *Proceedings of The Third International Workshop on Management of Diamondback Moth and Other Crucifer Pests* (eds Sivapragasam, A., Loke, W., H. Hussan, A. K. & Lin, G. S.) 228–232 (MARDI, 1997).
- Walker, G. P., Cameron, P. J. & Berry, N. A. Implementing an IPM programme for vegetable brassicas in New Zealand. In *Proceedings of The Fourth International Workshop on Management of Diamondback Moth and Other Crucifer Pests* (eds Endersby, N. M. & Ridland, P. M.) 365–370 (The Regional Institute Ltd, 2001).
- Furlong, M. J. et al. Ecology of diamondback moth in Australian canola: landscape perspectives and the implications for management. *Aust. J. Exp. Agr.* **48**, 1494–1505 (2008).
- Sonoda, S., Tsukahara, Y. M. & Tsumuki, H. Genomic organization of the para-sodium channel alpha-subunit genes from the pyrethroid-resistant and -susceptible strains of the diamondback moth. *Arch. Insect Biochem. Physiol.* **69**, 1–12 (2008).
- He, W. et al. Developmental and insecticide-resistant insights from the de novo assembled transcriptome of the diamondback moth, *Plutella xylostella*. *Genomics* **99**, 169–177 (2012).
- Trocza, B. et al. Resistance to diamide insecticides in diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae) is associated with a mutation in the membrane-spanning domain of the ryanodine receptor. *Insect Biochem. Mol. Biol.* **42**, 873–880 (2012).

47. Endersby, N. M. et al. Widespread pyrethroid resistance in Australian diamondback moth, *Plutella xylostella* (L.), is related to multiple mutations in the para sodium channel gene. *Bull. Entomol. Res.* **101**, 393–405 (2011).
48. Chen, G., Lee, S. H., Zhu, Z., Benyamin, B. & Robinson, M. R. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity* **117**, 51–61 (2016).
49. Schrödinger Release 2017-1: Prime, Schrödinger, LLC, New York, NY. (2017).
50. Jackson, C. J. et al. Structure and function of an insect α -carboxylesterase (α Esterase7) associated with insecticide resistance. *Proc. Natl Acad. Sci. USA* **110**, 10177–10182 (2013).
51. Amichot, M. et al. Point mutations associated with insecticide resistance in the *Drosophila* cytochrome P450 Cyp6a2 enable DDT metabolism. *Eur. J. Biochem.* **271**, 1250–1257 (2004).
52. Ratnasingham, S. & Hebert, P. D. N. BOLD: the barcode of life data system (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* **7**, 355–364 (2007).
53. The 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
54. Wallberg, A. et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat. Genet.* **46**, 1081–1088 (2014).
55. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
57. Hahn, C., Bachmann, L. & Chevreaux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* **41**, e129 (2013).
58. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
59. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
60. Chang, W. X. Z. et al. Mitochondrial DNA sequence variation among geographic strains of diamondback moth (Lepidoptera: Plutellidae). *Ann. Entomol. Soc. Am.* **90**, 590–595 (1997).
61. Robinson, G. S. & Sattler, K. *Plutella* in the Hawaiian Islands: relatives and host-races of the diamondback moth (Lepidoptera: Plutellidae). *Bish. Mus. Occas. Pap.* **67**, 1–27 (2001).
62. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
63. Fricot, E., Mathieu, F., Trouillon, T., Bouchard, G. & Francois, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).
64. Francis, R. M. POPHELPER: an R package and web app to analyze and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).
65. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
66. Zheng, X. et al. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).
67. Simon, M. & Van Belleghem, S. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438 (2017).
68. Haag-Liautard, C. et al. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosoph.* *Nat.* **445**, 82–85 (2007).
69. Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).
70. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J. & Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
71. Ye, J. et al. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34**, W293–W297 (2006).
72. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
73. South, A. rworldmap: a new R package for mapping global data. *R J.* **3**, 35–43 (2011).

Acknowledgements

The authors are very grateful to many researchers and volunteers for their kind help with collection of the *P. xylostella* specimens worldwide. Our special thanks go to Sean Graham of the University of British Columbia (UBC) for his kind help with the use of a computer server in Canada, to Yuelin Zhang for his kind supervision of Shijun You's PhD thesis that is related to this work, to Angel L. Viloria for his comments on the potential origin of *P. xylostella* and the information about early exportation of agricultural products in Americas, to Ihsan Al-Shehbaz for his information about South America Brassicaceae diversity. This work was financially supported by the National Natural Science Foundation of China (No. 31320103922 and No. 31230061), Fujian-Taiwan Joint Innovation Centre for Ecological Control of Crop Pests, International science and technology cooperation and exchange program of FAFU (KXb16014A), the Thousand Talents Program and the “111” Program in China, Australian Research Council grant FT140101303, and the National Key Research and Development Program of China (No. 2017YFD0201403).

Author contributions

M.Y., F.K., S.Y., Z.W., Q.L., W.H., S.W.B., and Z. Yuchi contributed equally to this work. M.Y., L.V., G.M.G., C.J.D., Z. Yue, H.Y., and J.W. conceived, designed and/or managed the project. H.C., M.S.G., L.V., G.M.G., S.W.B., Q.S., Q.F., G.W.-P., D.C.L., J.B., T.L., L.P., M.X., L. Cai, Y.Z., Z.Z., S.L., Y.W., Q.Z., X.X., W.C., L. Chen, M. Zou, J.L., L.H., Y. Lin, Y. Lu, and M. Zhuang collected insects and/or prepared DNA samples for sequencing. F.K., S.Y., Z.W., Q.L., W.H., Z. Yuchi, Y.J., C. M.W., L.L., T.L., J.B., M.Z., Q.G., X.F., and Y.Y. performed experiments and/or data analyses. M.Y., F.K., S.Y., Z.W., W.H., Z. Yuchi, L.V., and G.M.G. co-wrote the manuscript. M.Y., F.K., S.Y., Z.W., W.H., Z. Yuchi, L.V., G.M.G., D.C.L., S.W.B., C.M.W., H.C., G.Y., M.B.I., M.S.G., Q.S., Q.F., G.W.-P. interpreted results and/or revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-16178-9>.

Correspondence and requests for materials should be addressed to M.Y., L.V., G.M.G. or Z.Y.

Peer review information *Nature Communications* thanks Nicola Nadeau, and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



Disruption of duplicated *yellow* genes in *Bactrocera tryoni* modifies pigmentation colouration and impacts behaviour

Thu N. M. Nguyen^{1,2} · Vivian Mendez³ · Christopher Ward² · Peter Crisp⁴ · Alexie Papanicolaou⁵ · Amanda Choo² · Phillip W. Taylor³ · Simon W. Baxter¹

Received: 24 June 2020 / Revised: 5 November 2020 / Accepted: 17 November 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Irradiated Queensland fruit flies (*Bactrocera tryoni*) used in Sterile Insect Technique (SIT) programmes are marked with fluorescent dyes to distinguish them from wild flies when recaptured in monitoring traps. However, coating sterile pupae with powdered dyes can reduce emergence rates and fly quality and can sometimes produce insufficiently certain discrimination through inadequate coating or because the dye is transferred to wild flies through contact. Here we created a phenotypically distinct *B. tryoni* strain that lacks typical melanisation patterns through CRISPR/Cas9-mediated mutagenesis of tandemly duplicated *yellow-y* genes and then assessed effects of this visible trait on fly performance. Recessive mutations are only required in one of these copies to restrict melanisation and generate a phenotype clearly distinguished from wild type. The yellow strain showed significant declines in eclosion rates and in the percentage of fliers directly after emergence. Locomotor activity was greater in the yellow strain, and these mutations did not generally affect mating probability, copula latency, or copula duration. The longevity of yellow flies was approximately 10 days shorter than wild-type flies in both sexes. Overall, replacing dyes with yellow body marker for SIT can simplify production, eliminate a step that is known to reduce fly quality, remove potentially hazardous dyes from production, enable accurate discrimination from wild flies, and improve cost-effectiveness; however, direct comparisons of the decrements in performance associated with dyes on mass-reared wild-type flies and disruption of *yellow-y* genes are now required to determine the relative suitability of these marking methods for *B. tryoni* SIT.

Keywords Queensland fruit fly · Yellow-y · Sterile insect technique · CRISPR/Cas9 · Melanisation

Key message

Communicated by Christian Stauffer.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10340-020-01304-9>) contains supplementary material, which is available to authorised users.

✉ Simon W. Baxter
simon.baxter@unimelb.edu.au

¹ School of BioSciences, Bio21 Institute, University of Melbourne, Melbourne, VIC, Australia

² School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia

³ Applied BioSciences, Macquarie University, Sydney, NSW, Australia

⁴ South Australian Research and Development Institute, Adelaide, SA, Australia

⁵ Hawkesbury Institute for the Environment, Western Sydney University, Sydney, Australia

- Pest control using the sterile insect technique requires mass releases of irradiated insects that are marked with dyes; however, these can affect insect quality
- CRISPR mutagenesis created a yellow Queensland fruit fly strain with a phenotype that differed to the wild type
- The yellow strain was assessed for locomotor activity, longevity, flight ability, and mating success
- The yellow phenotype shows potential for identification purposes in sterile insect release programmes

Introduction

The Queensland fruit fly, *Bactrocera tryoni* (Diptera: Tephritidae), is the most destructive horticultural insect pest in Australia and is able to thrive across a wide climatic and host plant range (Sultana et al. 2017). Managing *B. tryoni* has traditionally relied on pesticides; however, public health and environmental risks associated with chemical sprays now limit their use (Dominiak and Ekman 2013). The Sterile Insect Technique (SIT) is an alternative species-specific biocontrol strategy that has been used to suppress or eradicate various tephritid fruit fly pests (Bellini et al. 2013; Enkerlin et al. 2015; Vreysen et al. 2014; Wyss 2000). Pest control using the SIT involves releasing vast numbers of sterilised males that mate with wild females which then produce non-viable embryos resulting in a localised population collapse (Knipling 1955). SIT has been used intermittently to control *B. tryoni* outbreaks for more than 20 years (Fanson et al. 2014), and significant efforts are currently underway to improve efficiency of this approach (Adnan et al. 2020; Benelli et al. 2019a, 2019b; Mainali et al. 2019).

Safeguarding horticulture regions from *B. tryoni* outbreaks involves careful monitoring with traps. If population suppression with SIT is required, it is important to quickly and accurately discriminate between released sterile flies and wild flies collected in monitoring traps, to evaluate the efficiency of the SIT programme, to prevent mistaken declaration of outbreak, and to avoid failure to respond promptly to incipient outbreaks. False-positive detections in surveillance programmes could potentially impact market access for horticulture products, while false-negative assessments could result in containment failure. As is standard practise for most SIT programmes (FAO/IAEA/USDA 2019; Zavala-López and Enkerlin 2017), the current method for distinguishing sterile *B. tryoni* is by coating pupae with a fluorescent dye that is transferred to adult upon eclosion. However, this method has several drawbacks, including that the fluorescent dust (1) can fail to thoroughly coat flies, be washed off or transferred to wild flies upon contact, (2) can cause dehydration and death prior to release (Campbell et al. 2009; Dominiak et al. 2000, 2010b), (3) can reduce flight ability (Dominiak et al. 2010a), (4) may affect field performance due to the increased preening time, and (5) is considered as hazard for personnel working in the mass-rearing facility. Improved and safe visual markers to distinguish between sterile flies and wild flies would be beneficial to SIT programmes.

Identifying heritable visual phenotypes that are not found in wild flies is a potential alternative strategy to distinguish sterile flies collected in monitoring traps. Four

recessive visible phenotypic markers have been reported in *B. tryoni*, including three spontaneous mutations bent wings, white marks and orange eyes (Meats et al. 2002; Zhao et al. 2003), and a white eye mutant caused by targeted knockout of the *white* gene (Choo et al. 2018). However, these mutant strains were all likely to carry substantial fitness costs and/or affect behaviour. The *bent wings* mutation creates deformities of the wing and impedes flight, so it is not suitable for SIT (Meats et al. 2002). Although viability assays have not been performed on orange and white eyes *B. tryoni* mutants, eyes pigmentation mutations in other fruit fly species significantly affect performance (Saul and McCombs 1992), and mutations in the *Drosophila melanogaster white* ortholog severely impair biological functions including mobility, lifespan, and stress tolerance (Ferreiro et al. 2018). The *B. tryoni white marks* strain exhibits white rather than yellow marking on the thorax, and although it has been used in dispersal studies (Weldon and Meats 2007, 2010), the strain has been lost and its genetic base is unknown; therefore, it cannot be re-created. Developing a *B. tryoni* strain with a novel visual body marker that remains stable over the lifespan of the fly without substantially compromising performance of male flies could improve the efficiency of SIT programmes.

In *Drosophila melanogaster* (as well as many other insects), body pigmentation and colouration result from a combination of black and brown melanins as well as yellow-tan and colourless sclerotins which are produced by the melanin pathway (Massey and Wittkopp 2016; Wittkopp et al. 2003). The pathway begins with the conversion of tyrosine into DOPA (dihydroxyphenylalanine). Subsequently, DOPA can be used in two different pathways: to be polymerised into a black melanin or to be converted into dopamine. Dopamine can then have one of four fates: (1) to be converted into brown melanin, (2) into black melanin, (3) into NBAD (beta-alanyl dopamine) which is subsequently oxidised to produce a yellow-tan sclerotin, or (iv) into NADA (N-acetyl dopamine) which is the precursor of colourless sclerotin. The *yellow-y* gene plays essential role in modifying dopamine into black melanin (Massey and Wittkopp 2016). Disruption of the *yellow-y* gene causes a pale-yellow body colour which is clearly distinguishable in *D. melanogaster* (Wittkopp et al. 2002). Loss of activity of the corresponding yellow orthologs in other dipteran species such as *Ceratitis capitata* (Gourzi et al. 2000), *Musca domestica* (Heinze et al. 2017), *Lucilia cuprina* (Paulo et al. 2019) results in reduced black melanin synthesis, causing regions of the body that are normally black to display a lighter colouration.

In *Bactrocera dorsalis*, *yellow-y* has duplicated and each copy shows different expression profiles throughout development (Bai et al. 2019). Here we confirm that the *B. tryoni* genome also contains two *yellow-y* paralogs and assessed

the performance of a *yellow-y* knock-out strain developed using CRISPR/Cas9-mediated mutagenesis for potential use in SIT programmes.

Materials and methods

Fly rearing

Bactrocera tryoni were obtained from New South Wales Department of Primary Industries, Ourimbah, Australia, and were maintained at the biological control insectary of South Australian Research and Development Institute in Urrbrae, South Australia. Flies were reared in a controlled environment room (25 ± 2 °C and $65 \pm 10\%$ relative humidity (RH)) and under a photoperiod of 14:10 h light:dark and fed with sugar, either Brewer's yeast or yeast hydrolysate enzymatic (MP Biomedical, LLC), and water, while larvae were reared on 'Chang et al. (2006)' gel diet of Moadeli et al. (2017), based on liquid diet formulation of Chang et al. (2006). Eggs were collected using an 'oviposition device', which was a 30-mL semi-transparent white plastic cup covered by a green damp cloth and a pinpricked lid for the females to oviposit through. A piece of sponge saturated with 5 μ l of 100% tart apple juice was placed inside the oviposition device to stimulate oviposition.

Evolution of *yellow-y* gene in *Bactrocera* species

The *Drosophila melanogaster* *yellow-y* protein (FBpp0070070) was obtained from Flybase and queried against the *B. tryoni* genome (GCA_000695345.1) and the publicly available genomes of *Bactrocera latifrons* (MIMC00000000.1), *Bactrocera dorsalis* (JFBF00000000.1), *Bactrocera oleae* (LGAM00000000.2), *Zeugodacus curcubitae* (JRNW00000000.1), and *Ceratitis capitata* (AOHK00000000.2) using the tBLASTn algorithm in Geneious (v10.2.6). As full-length *yellow-y* sequence was absent from the *B. tryoni* reference genome, whole genome shotgun sequence reads of *B. tryoni* were obtained from Gilchrist et al. (2014) and a genome reassembly was performed using different parameters. Input data for genome reassembly included a 250-bp paired-end Illumina library (300-bp insert sizes) generated from a single individual (error-corrected with blue (Greenfield et al. 2014)), and two Illumina mate pair libraries of 3 kb and 10 kb. The mate pair libraries did not originate from a single fly, and therefore we focused on using an ALLPATHS-LG assembly ensuring the mate pair data did not contribute to the base sequence if filtered for paired-end contaminants (Gnerre et al. 2011). Paired-end contaminants in mate pair libraries were removed by creating a partial ALLPATHS-LG assembly and filtering out these

errors prior to proceeding with subsequent ALLPATHS-LG assemblies. We explored using the 250-bp paired-end data as assembled contigs via DISCOVAR denovo (Weisenfeld et al. 2014), and these were provided as possible inputs to the next stage of the ALLPATHS-LG process. Multiple ALLPATHS-LG assemblies explored the parameter space and combination of input data to optimise measures of contiguity (G50) and completeness (BUSCO and reference mRNA sequences identified). Correctness was not identified as prior experience has shown that ALLPATHS-LG has high sequence correctness due to internal processes based on the realignment of raw reads. An assembly was selected based on these metrics called 'refcon'. Subsequently, we removed haplotypes based on an all versus all alignments of repeatmasked contigs, raw sequence read coverage, and preservation of paired data. The assembly was finally repeatmasked using RepeatScout and RepeatMasker. A single contig containing two *B. tryoni* *yellow-y* paralogs was identified (Supplementary File 1). Full coding sequence of *B. tryoni* *yellow-y* genes was obtained from *B. tryoni* one-day-old male transcriptome (GenBank accession numbers MT656584, MT656584) and was used for downstream phylogeny analysis.

The NCBI Conserved Domain Search (CD-search) was used to identify signature motifs in the annotated *yellow-y1* and *yellow-y2* protein sequences. The second exons of *yellow-y1* and *yellow-y2* were PCR amplified and Sanger Sequenced (Australian Genome Research Facility) using genomic DNA from the *B. tryoni* Ourimbah laboratory strain to identify polymorphisms relative to the reference genome.

A multiple alignment of the *yellow-y* coding sequence was constructed using MAFFT v.7.388 (Katoh and Standley 2013) with Geneious (v10.2.6). Maximum likelihood trees of *yellow-y* ungapped coding sequence were constructed using RAXML v8 (Stamatakis 2014) using the general time-reversible model and a gamma distribution with gamma rate heterogeneity. Support values for each node were generated by 1000 bootstrap replicates. Phylogenetic trees were viewed and edited by FigTree v.1.4.4 (Rambaut 2018). The alignment and maximum likelihood phylogeny were then used as inputs into HyPhy (Pond et al. 2004) to test for diversifying selection on branches before and after the split of *yellow-y1* and *yellow-y2* using the aBSREL algorithm (Smith et al. 2015).

The *B. tryoni* *yellow* genes were identified by searching the re-assembled *B. tryoni* genome and one-day-old male transcriptome using *D. melanogaster* *yellow* proteins as queries. Full length of *B. tryoni*, *D. melanogaster*, and *Bombyx mori* *yellow* proteins were aligned using MAFFT v7.450 with L-INS-I algorithm, and maximum likelihood phylogeny was constructed using RAXML v8 (Stamatakis 2014) using PROTGAMEAUTO option. Support values for each node were generated by 1000 bootstrap replicates.

Expression profile of *Bactrocera tryoni* yellow-y genes

Three biological replicates representing different developmental stages were collected, which consisted of pools of approximately 100 embryos 24 h after laying, 1st instar ($n=22$) and 3rd instar ($n=10$) larvae, individual pools of five 1-, 3-, 5-, 7-, 9-, or 12-day pupae, and pools of two males and two females 1 or 7 days post-emergence. Total RNA was isolated using the TRIzol protocol (Ambion), and genomic DNA was removed using the TURBO DNA-free kit (Invitrogen) according to manufacturer's instructions. First-strand cDNA was synthesised using SuperScript IV reverse transcriptase (Invitrogen) with Oligo dT primer (Promega) and random hexamers (Promega) with 1 μ g total RNA template. Primers used for quantitative PCR (qPCR) were designed using Primer3Plus (<https://primer3plus.com/cgi-bin/dev/primer3plus.cgi>) (Table S1) and reactions performed with SensiFAST SYBR Hi-ROX Mix (BIOLINE) using the StepOnePlus Real-Time PCR system (Applied Biosystems). Quantification of gene expression was performed with the $2^{-\Delta\Delta C_t}$ method (Livak and Schmittgen 2001), using *B. tryoni actin3* as an endogenous control. Expression of *yellow-y1* and *yellow-y2* genes was normalised as one-day adults which enabled relative concentrations to be compared across other developmental stages.

CRISPR/Cas9 crRNAs design, preparation of the ribonucleoprotein complex and embryo microinjections

Two crRNAs were designed using CRISPOR (Haeussler et al. 2016) to target identical sequence regions in exon 2 of the *B. tryoni yellow-y1* and *yellow-y2* genes (crRNA#1-TTG TGCGAACAGCATCACCA and crRNA#2-GTCACTCTC ACCCATACGTT). The Alt-R® S.p. Cas9 Nuclease 3NLS (#1074181), two specific crRNAs and universal tracrRNA (#1072532), were obtained from Integrated DNA Technology (IDT), USA. The purified Cas9 protein was diluted to 1 μ g/ μ l in 20 mM HEPES (pH 7.5, 150 mM KCl) and stored at -20°C in small aliquots. The lyophilised crRNA and tracrRNA pellets were resuspended in Nuclease-Free Duplex Buffer (IDT) at 100 μ M concentration and stored at -20°C . Two guide RNA duplexes were individually annealed by mixing 1 μ l of each crRNA with 1 μ l of tracrRNA and 0.5 μ l of Nuclease-Free Duplex Buffer to create a final duplex concentration of ~ 40 μ M and incubated at 95°C for 5 min and then allowed to cool to room temperature (20 – 25°C). To generate the ribonucleoprotein (RNP) complex, 2.5 μ l of each annealed crRNA:tracrRNA, 3 μ l of diluted Cas9 protein (1 μ g/ μ l), 1 μ l of $10\times$ injection buffer (0.1 mM sodium phosphate buffer pH 6.8, 5 mM KCl) together with 1 μ l of Nuclease-Free Duplex Buffer were mixed. The mixture was

incubated at room temperature for 5 min to allow ribonucleoprotein complex formation and stored on ice until injecting. *Bactrocera tryoni* embryo microinjections were performed according to Choo et al. (2018). Embryos less than one hour after egg laying were injected at the posterior end and left until hatched.

Phenotypic screening and mutation analyses

Injected G_0 flies were mated inter-se, and G_1 adult progeny were assessed for visible change in pigmentation colour and photographed using an Olympus SZX16 microscope and an Olympus LC30 camera. The G_1 flies with reduced pigmentation were selected and reared as the 'yellow' strain.

Individual whole flies were homogenised using a TissueLyser II (Qiagen), and DNA was extracted with the DNeasy® Blood & Tissue Kit (Qiagen) for genotyping. A region in exon 2 of the *B. tryoni yellow-y1* and *yellow-y2* was PCR amplified using MyTaq® polymerase (Bioline) with two set of primer pairs *Qyellow_F3/Qyellow_R1* and *Qyellow_F3/Qyellow_R3*, respectively (Table S1). The PCR conditions were as follows: 95°C for 1 min, 35 cycles of 95°C for 15 s, 60°C for 15 s, and 72°C for 59 s, followed by a final extension at 72°C for 5 min. The PCR amplicons were purified using MinElute® PCR Purification Kit (Qiagen), cloned into pGEM®-T Easy Vector Systems (Promega), and then transformed into DH5 α cells. Plasmids were purified with Wizard Plus SV Minipreps kit (Promega) and Sanger sequenced (AGRF). Sequences were aligned using MAFFT v7.388 (Katoh and Standley 2013) in Geneious (v10.2.6).

Fitness comparison between wild-type and yellow fly strains

A culture of the yellow strain (generation six) and Ourimbah strain (wild-type strain) was maintained in a controlled environment room ($25 \pm 0.5^\circ\text{C}$, $65 \pm 5\%$ RH, and 12:0.5:11:0.5-h light:dusk:

dark:dawn photoperiod). Adult flies were fed with sugar, yeast hydrolysate enzymatic (MP Biomedical, LLC), and water, while larvae were reared on 'Chang et al. (2006)' gel diet (Moadeli et al. 2017).

Assessment of emergence rates, flight ability, and sex ratios

Three days before estimated emergence, replicates of 100 pupae were placed in 55-mm plastic Petri dish lids which were then centred on 90-mm Petri dish lids that were overlaid with a black paper. Black acrylic 'flight tubes' (100 mm tall) were placed over the 90-mm Petri dish, ensuring the only way for adults to escape was to fly upward. Flight tubes

were coated with unscented talcum powder on the interior surface to prevent flies from walking out. A 10-mm section at the base of the flight tubes remained powder free to provide newly emerged flies a vertical place to rest. An identical empty 'fly-back tube' without pupae was placed 50 mm away from the flight tube, to estimate the number of individuals that fly out from after emergence but then return and unable to escape again (fly-back). Both tubes were placed in a white mesh cage (325 × 325 × 325 mm Megaview BugDorm-43030F). The cages were kept in 14:10-h light:dark cycle, and light was emitted by a white 36-W fluorescent tube position 100 mm above the cage. To minimise fly-back, cages were checked daily and flies that had escaped from the tubes were collected. When all flies had emerged (6 days after the first flies emerged), the remaining contents of the tubes were counted. The flies were categorised as: (1) non-emerged (unopened pupal case), (2) partially emerged (portion of adult body stuck in puparium), (3) deformed (emerged but with deformed or damaged wings), (4) non-fliers (the number of morphological normal flies in the flight tube subtracting the flies in the fly-back tube).

- Percentage of adult emergence = $((N \text{ pupae} - (N \text{ non-emerged} + N \text{ partially emerged})) / N \text{ pupae}) \times 100$
- Percentage of partially emerged = $(N \text{ partially emerged} / N \text{ pupae}) \times 100$
- Percentage of deformed adult = $(N \text{ deformed} / N \text{ pupae}) \times 100$
- Percentage of fliers = $((N \text{ pupae} - (N \text{ non-emerged} + N \text{ partially emerge} + N \text{ deformed} + N \text{ non-fliers})) / N \text{ pupae}) \times 100$
- Rate of fliers (percentage of flies that are able to fly) = $(\text{percentage fliers} / \text{percentage emergence}) \times 100$
- Sex ratio = $N \text{ males emerged} / N \text{ total flies emerged}$

Locomotor activity

Flies were sorted by sex three days after emergence, before sexual maturation (Perez-Staples et al. 2007), and groups of 200 male or female virgin flies were transferred to 12-L ventilated plastic cages. When 12–17 days old, 84 flies from each strain and each sex were caught individually without anaesthesia in each glass tubes (0.8 cm internal diameter × 10 cm long). Each tube was plugged by a black soft plastic cap at one end and a loose cotton ball at the other end to allow ventilation and was then placed in a locomotor activity monitor (LAM10, TriKinetics, MA, USA) in which activity was measured by the number of times a fly crossed an infrared beam projected through the centre of each tube (following Fanson et al. (2013) and Dominiak et al. (2014)). Four monitor units (each containing 32 tubes) arranged in four rows of eight columns were used. We controlled for potential behavioural differences

due to tube location, by loading flies into activity monitors using a careful designed approach that systematically varied their positions based on sex and phenotype. Activity was assessed for 4 h during light period. Activity rate was calculated by total number of beam crosses throughout the duration of the experiment.

Non-competitive single pair mating assays

Virgin flies were sorted by sex and separated as described above. Non-competitive mating success experiments were conducted by pairing single virgin males with single virgin females and assessing mating at three different time points. The experiments were conducted simultaneously using wild-type flies 12, 14, and 16 days post-emergence, and yellow flies 13, 15, and 17 days post-emergence (as the yellow flies emerged a day earlier). On each experimental day, at least 4 h before onset of simulated dusk, each single pair of flies was placed in a transparent plastic 1.125-L container with a mesh cover window (ca. 28 cm²) for ventilation. *Bactrocera tryoni* usually mates at dusk in dim light (Smith 1979; Tychsen and Fletcher 1971), but in laboratory conditions it is common to have some matings start earlier. All four possible crossing combinations were set up with 30 replicates each: (1) wild-type male × wild-type female, (2) yellow male × wild-type female, (3) wild-type male × yellow female, and (4) yellow male × yellow female. Each pair was left to interact and was observed for mating from the time the flies were introduced into the container until all sexual activity had ended for the evening. Time when copulation began and ended was recorded.

- Percentage of mating = $(\text{number of successful mated pairs} / 30) \times 100$
- Latency until copulation = latency from introduction of flies until mating of each pair
- Copula duration = the mating duration of each pair

Assessment of longevity

For each strain, 10 single sex flies were sorted within 3 days after emerging and transferred to a 1.125-L container with a mesh-covered window (ca. 28 cm²) for ventilation. Each cage was provided with yeast hydrolysate enzymatic (MP Biomedical, LLC), sugar, and water-soaked cotton in separate 35-mm Petri dishes. Cages were checked, and dead flies were removed daily until day 217. The lifespan of each fly was calculated as the number of days it survived after emergence. Escaped flies or accidental deaths were recorded as censored.

Statistical analyses

Data analyses were performed in RStudio (v. 1.2.1335). Student's *t* test was used to test for statistical effect of pairwise comparisons. A two-way ANOVA followed by a Tukey post hoc test was used to compare data from multiple samples with two factors. Mating probability (binary outcome) was assessed using logistic regression followed by post hoc analysis.

Analyses of survival data were performed using the statistical software JMP 15.0.0 (SAS Institute Inc., Cary, NC, USA). The Akaike's information criterion (AICc) was used to identify the best fit parametric model for our survival data which was Weibull distribution with strain and sex as fixed effects.

Results

Duplication of the *yellow-y* gene

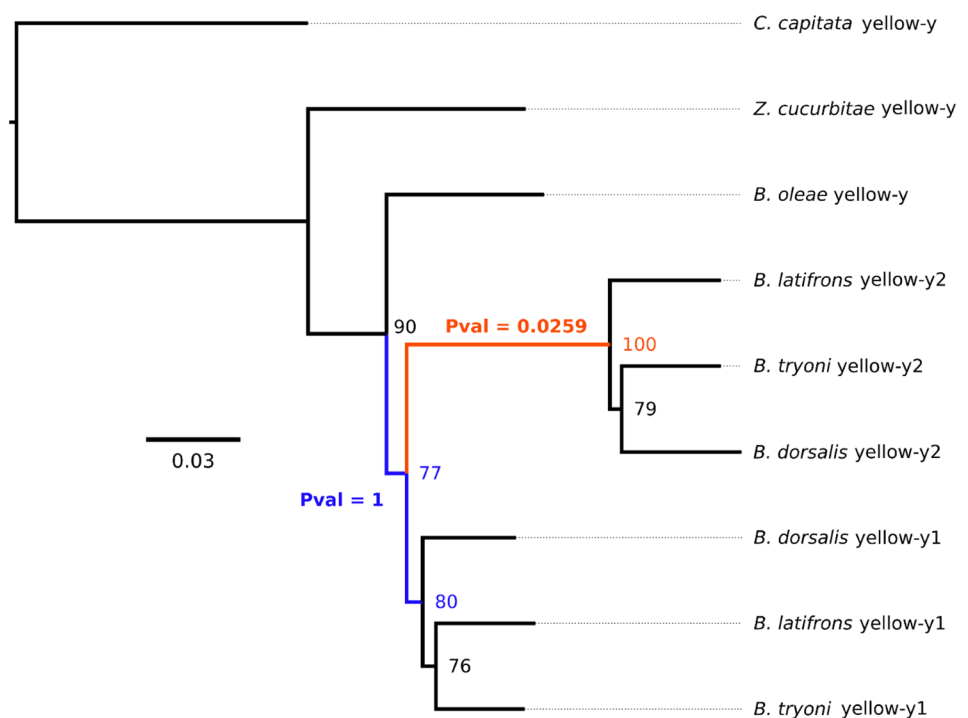
BLAST analysis of the *D. melanogaster* *yellow-y* protein against tephritid genomes identified a *yellow-y* gene duplication in *B. dorsalis* and *B. latifrons*, but not *Bactrocera oleae* or more distantly related species (Fig. 1, Figure S1). Only partial coding sequences of *yellow-y* genes were identified in the *B. tryoni* reference genome (contig JHQJ01009153.1, JHQJ01004462.1, and JHQJ01008064.1) (Figure S2), and consequently, we generated a revised genome assembly and recovered a *yellow-y* tandem duplication (Supplementary

File 1). *Yellow-y1* and *yellow-y2* paralogs contained a highly conserved Major Royal Jelly Protein (MRJP) domains (287 amino acids), although *yellow-y2* (493 amino acids) was considerably shorter than *yellow-y1* (553 amino acids) due to variation at the carboxy-terminal of the protein (Figure S3, Table S2). In *B. tryoni*, the *yellow-y* proteins shared 77.40% identity. Phylogenetic reconstruction of *yellow-y* homologs revealed the gene tree was consistent with previously published species trees (Choo et al. 2019), and although *yellow-y1* and *yellow-y2* were in reciprocal monophyly, there was some incongruence between paralog clade topologies (Fig. 1). To determine whether positive selection may have occurred after the duplication of *yellow-y*, three branches were individually tested, relative to background, using the aBSREL algorithm. Significant ($p < 0.05$) tests were found in the ancestral lineage of the *yellow-y2* paralog immediately after duplication suggesting neo-functionalisation may have occurred. However, no evidence for positive selection was identified in *yellow-y1* or prior to *yellow-y* duplication (Fig. 1).

Bactrocera tryoni *yellow-y* paralogs are differentially expressed during most developmental stages

Expression levels of *Bt-yellow-y1* and *Bt-yellow-y2* were compared across multiple developmental stages using quantitative polymerase chain reaction (qPCR) to assess transcript abundance. The transcripts of both genes were detected in all the stages tested. The *yellow-y1* gene is expressed approximately 100-fold higher than *yellow-y2*

Fig. 1 Maximum likelihood *yellow-y* tree from seven Tephritidae species generated using ungapped protein coding sequence. Duplication of *yellow-y* occurred in the common ancestor of *B. tryoni*, *B. dorsalis* and *B. latifrons*, after splitting from *B. oleae*. Bootstrap support was generated from 1000 replicates, and their values (percentage) are indicated on each node. Tests for positive selection following the split from *B. oleae* show that *yellow-y2* (red branch, $p = 0.0259$) was significant and *yellow-y1* was not significant (blue branch, $p = 1$). The tree was rooted using *Ceratitis capitata* as the outgroup



throughout embryo, larval, and pupae development, but despite major differences in transcript abundance, expression patterns of both genes had consistent profiles. Among developmental stages assessed, *Bt-yellow-y2* transcripts only exceeded *Bt-yellow-y1* in 7-day-old flies (Fig. 2). Differences in expression of *Bt-yellow-y1* and *Bt-yellow-y2* may have functional consequences.

CRISPR/Cas9-mediated mutagenesis of *B. tryoni* *yellow-y* paralogs

Two crRNA sequences were designed to recognise conserved regions of exon 2 in both *Bt-yellow-y1* and *Bt-yellow-y2*. Cas9-crRNA-tracrRNA complexes were co-injected into 800 *B. tryoni* embryos, and 19 G₀ adult flies with wild-type body colour were recovered (Fig. 3a). Mass mating of 12 surviving mature adults (eight males and four females) produced more than 800 G₁ adults, including 119 (54 males and 65 females) with distinct pale colouration that were mass-crossed to establish a stable yellow strain. Darker colour scales (black colour and dark brown) observed in wild-type flies were lighter in yellow mutants,

whereas other colour scales remained unchanged. Newly emerged yellow mutants have metallic yellow-coloured cuticle, which developed into pale brown a few hours after eclosion. The most distinguishable change in yellow mutants was a pale-yellow wing margin, which remains unchanged throughout their lifespan (Fig. 3c). Reciprocal crossing of yellow mutants to wild-type strain produced progeny with wild-type pigmentation colour, confirming that yellow was an autosomal recessive phenotype.

Mutations in *Bt-yellow-y1* and *Bt-yellow-y2* genes were characterised using cloned PCR amplicons from five individual G₁ yellow flies (three females and two males). We sequenced five cloned amplicons for each individual, for both *Bt-yellow-y* paralogs. Analysis of *Bt-yellow-y1* identified four different mutations and confirmed that all five individuals carried two mutant alleles, including an in-frame three base deletion (Threonine deletion) in *Bt-yellow-y1* (Fig. 3b). Five different mutations in *Bt-yellow-y2* were identified; however, three yellow individuals also carried wild-type alleles (Fig. 3b). Phenotypic screening of subsequent generations G₂–G₁₁ of the established yellow

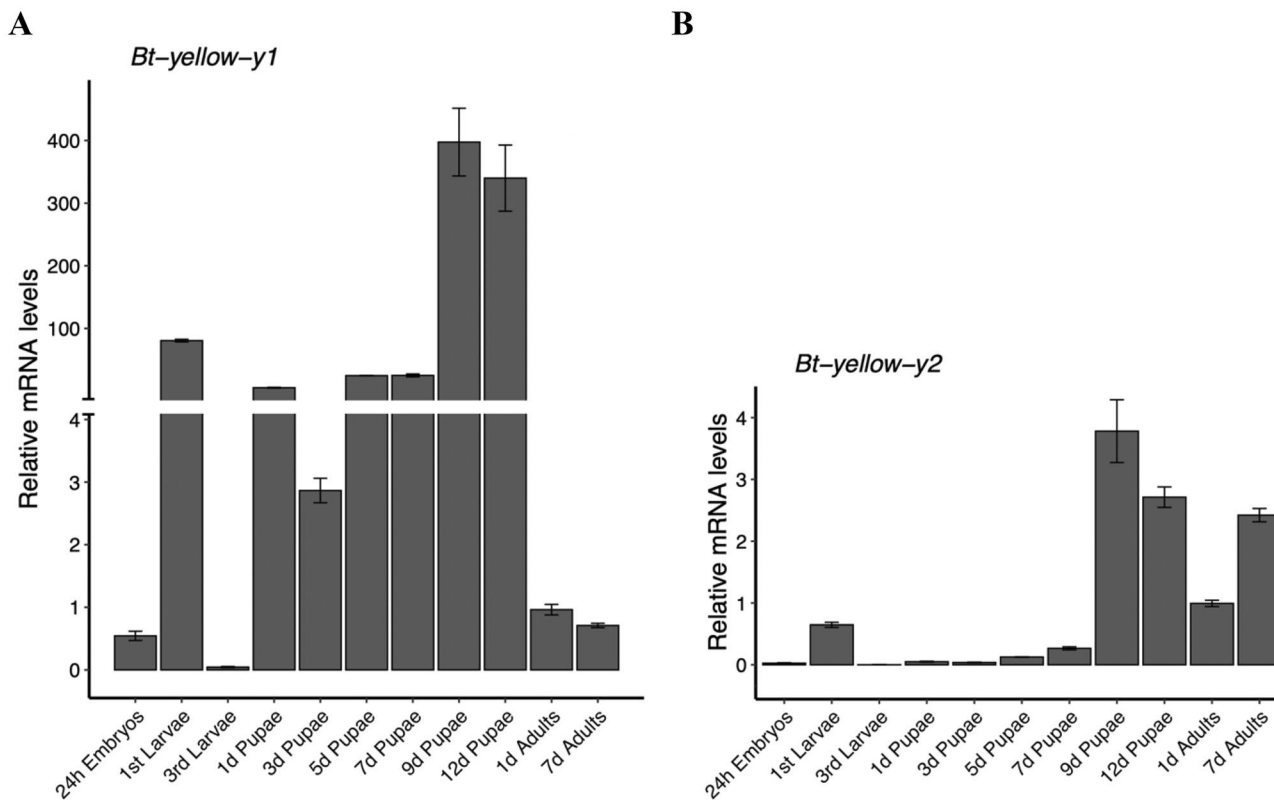


Fig. 2 Expression of *Bactrocera tryoni* genes (a) *yellow-y1* and b *yellow-y2* during development. Relative expression was calculated using the delta-delta Ct method ($2^{-\Delta\Delta Ct}$) with housekeeping gene *actin3* using tissue collected from 24-h embryos, first and third instar lar-

vae, six pupae ages (at 24–26 °C), and adults one and seven days after eclosion. Expression is plotted relative to 1-day-adult samples. Bar represents standard errors for three biological replicates

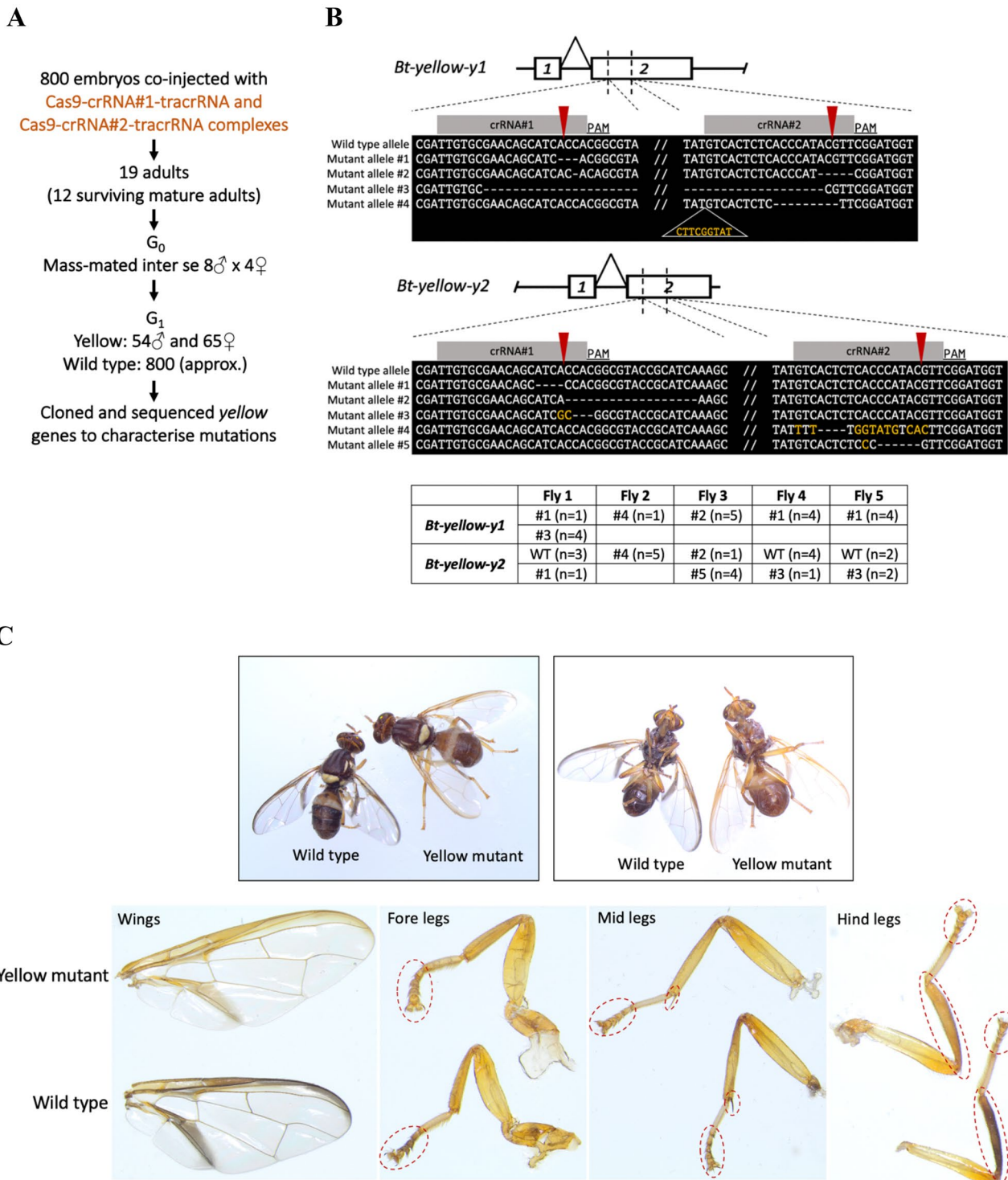


Fig. 3 CRISPR/Cas9-mediated mutagenesis of the *Bactrocera tryoni* *yellow-y1* and *yellow-y2* genes. **a** Flow diagram showing the process for the microinjections, mating of G₀ adults, and screening for yellow mutants in G₁ progeny. **b** A schematic of the genomic organisation of *Bt-yellow-y1* and *Bt-yellow-y2* genes and the two guide target sites in exon 2 of each gene. Exons 1 and 2 are shown as boxes, and the black dotted lines refer to the CRISPR/Cas9 target sites. Target sequences crRNA#1 and crRNA#2 (grey boxes), PAM motifs and expected Cas9 cut sites (red arrows) are shown above wild-type sequences.

Dashed lines in DNA sequence alignments represent CRISPR/Cas9 deletions, and insertions are indicated with yellow font. The table shows allelic mutations of *yellow-y1* (#1 to #4) and *yellow-y2* (#1 to #5) identified from sequencing cloned amplicons from five yellow individuals. WT is wild type, and “n” refers to the number of clones sequenced. **c** Comparison of *B. tryoni* 3-days-old male wild-type body colour, and the CRISPR/Cas9 induced yellow mutant phenotype. Knockdown *yellow-y* genes blocked melanin synthesis in specific tissues

strain did not identify any flies with wild-type body colour in the colony, indicating that the phenotype is stable.

Yellow-y mutations significantly reduce adult emergence rates and flight ability

Pupal emergence and subsequent flight ability were assessed for the wild type and yellow strains using 20 replicates of approximately 100 pupae each. To initiate the experiment, eggs were seeded from both strains simultaneously to establish large populations, yet we observed mostly yellow flies consistently emerged one day earlier

than the wild type. Adult emergence from the puparium was reduced in the yellow strain, but remained above 90% (Fig. 4a). Percentage of fliers (taking all fliers per total pupae) was 16% lower in yellow flies, compared with the wild type (Fig. 4b). Similarly, the rate of fliers (taking all fliers per total of emerged flies) was significantly lower in the yellow strain (Fig. 4c). There were no significant differences in the percentage of partial emerged and deformed flies between the wild type and yellow strain, with values being very low in both strains (Fig. 4d, e). Sex ratio did not vary significantly between the two strains (Fig. 4f).

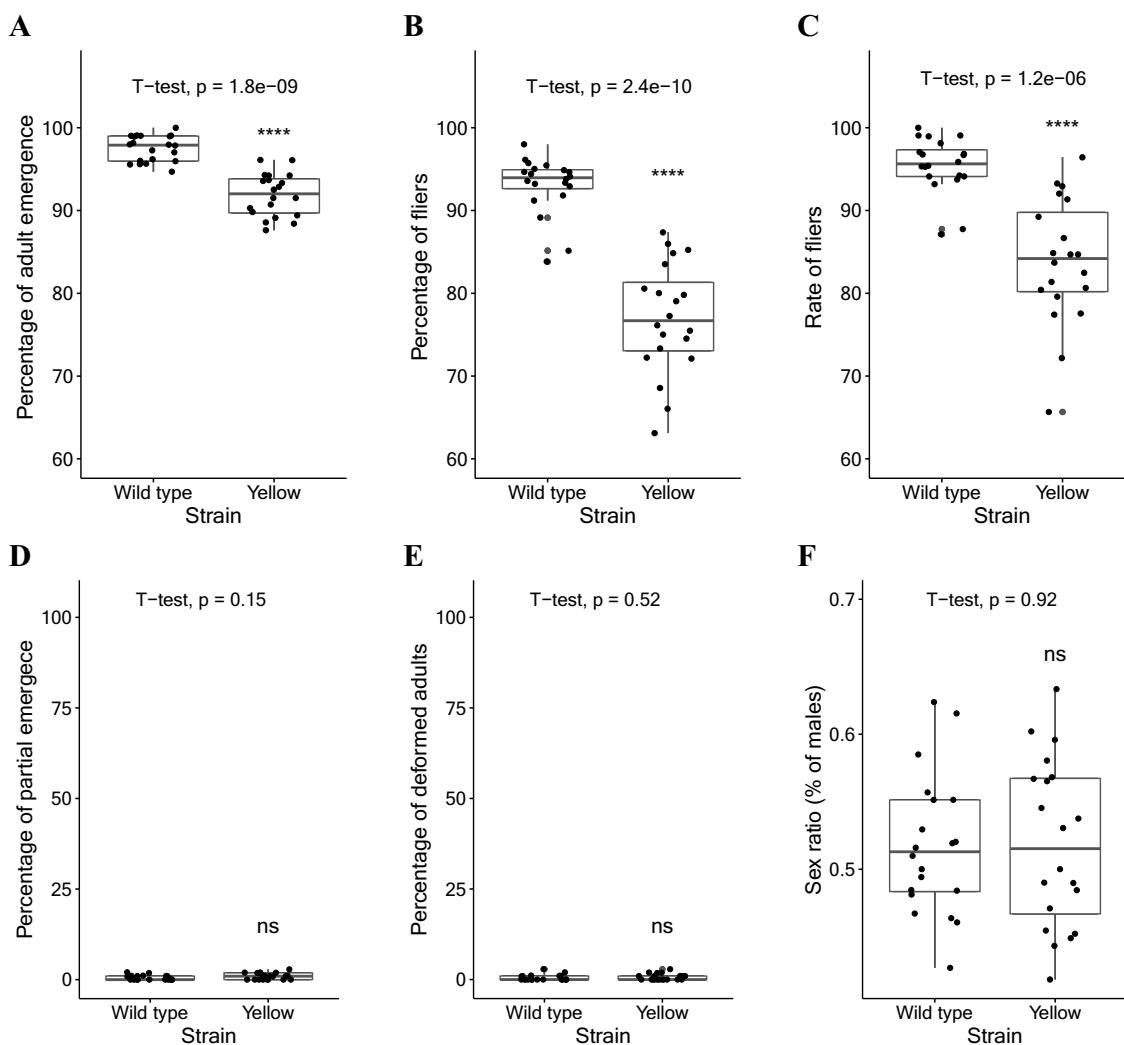


Fig. 4 Flight ability assay: **a** percentage of adult emergence, **b** percentage of fliers (pupae that eclosed as adults and could fly), **c** rate of fliers (comparing all flies that successfully emerged without malformation), **d** percentage of partially emerged, **e** percentage of deformed adults, and **f** sex ratio of *Bactrocera tryoni* wild-type and yellow flies. Twenty replicates were carried for each strain, with 100

pupae per replicate. Box plots and data points are shown for each replicate. Box plots represent the interquartile range, and the median value is indicated. Error bars represent 1.5 times the interquartile range. The statistical significance was calculated by Student's t test (**** $p < 0.00001$, ns not significant)

Yellow flies have a higher activity rate than wild-type flies

Eighty-four flies of each strain and each sex were individually assessed for activity levels during the four hours of light prior to dusk using a locomotor activity monitor. Flies were age matched and ranged between 12 and 17 days old. For both males and females, the total number of beam crosses during the testing period of yellow flies was significantly higher compared to wild type (Fig. 5).

Yellow flies are capable of normal courtship

There was no significant difference in percentage of mating among the four crossing combinations and across different

age time points, with more than 90% of pairs mating overall (Fig. 6a). Latency until copulation did not vary between pairings type or with age of the flies (Fig. 6b). When paired with wild-type females, yellow males 12–13 days old and 16–17 days old had shorter median copulation duration than wild-type males; however, these differences were not significant (Fig. 6c). There was no significant difference in copula duration between yellow and wild-type males when crossed with yellow females (Fig. 6c).

The lifespan of yellow flies is slightly shorter than wild-type males

Two-hundred flies of each strain and sex were assessed for longevity over a period of six months by recording mortality

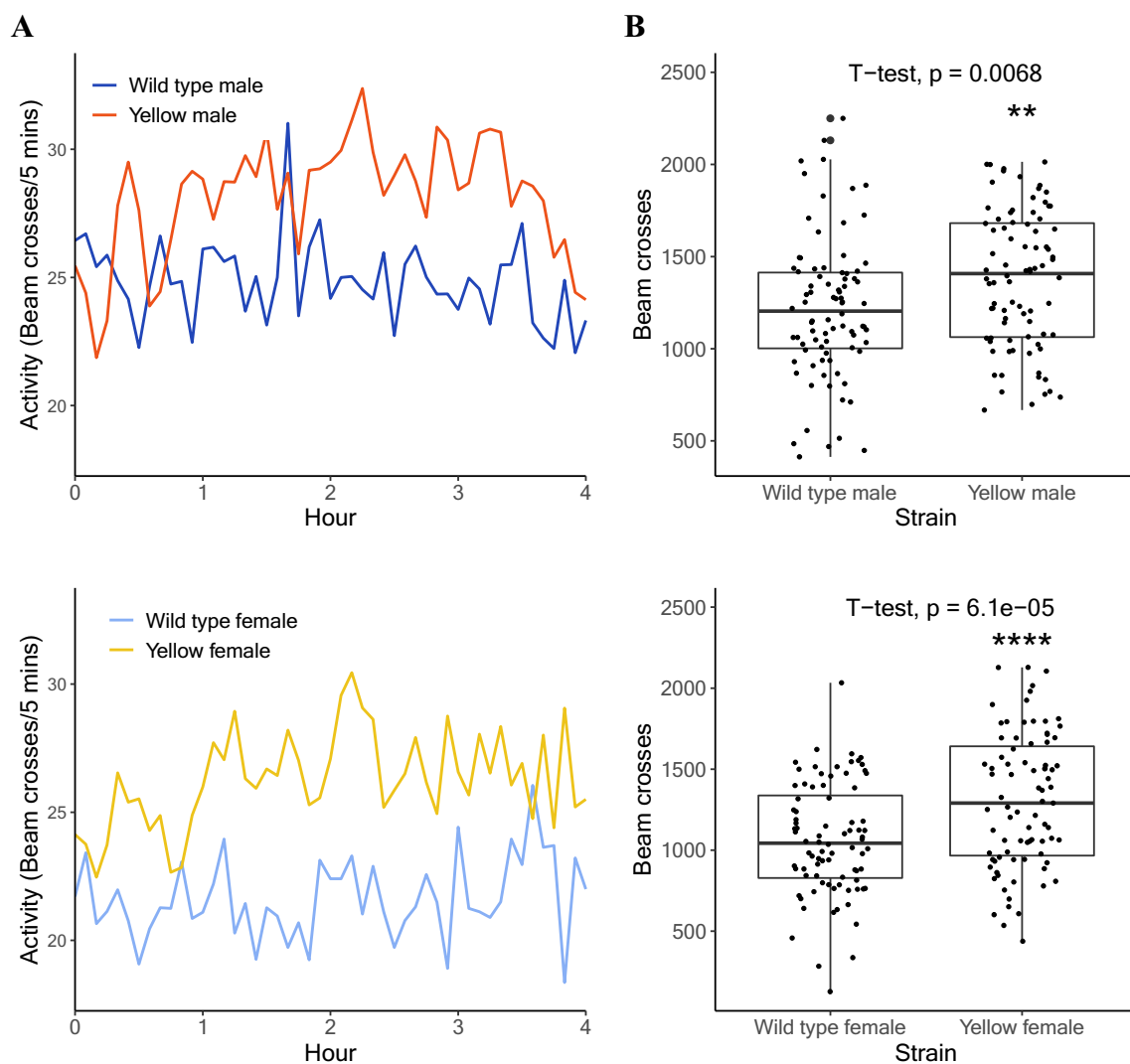


Fig. 5 Locomotor activity assay over four hours of day light: **a** average activity patterns (beam crossing per 5 min) for males (top row) and females (bottom row) of *Bactrocera tryoni* wild-type and yellow flies. **b** Total beam crosses of male flies (top row) and female flies

(bottom row). Box plots and raw data points are shown. Box plots represent the interquartile range, and the median value is indicated. Error bars represent 1.5 times the interquartile range. Significance was measured using Student's t test (** $p < 0.01$, **** $p < 0.001$)

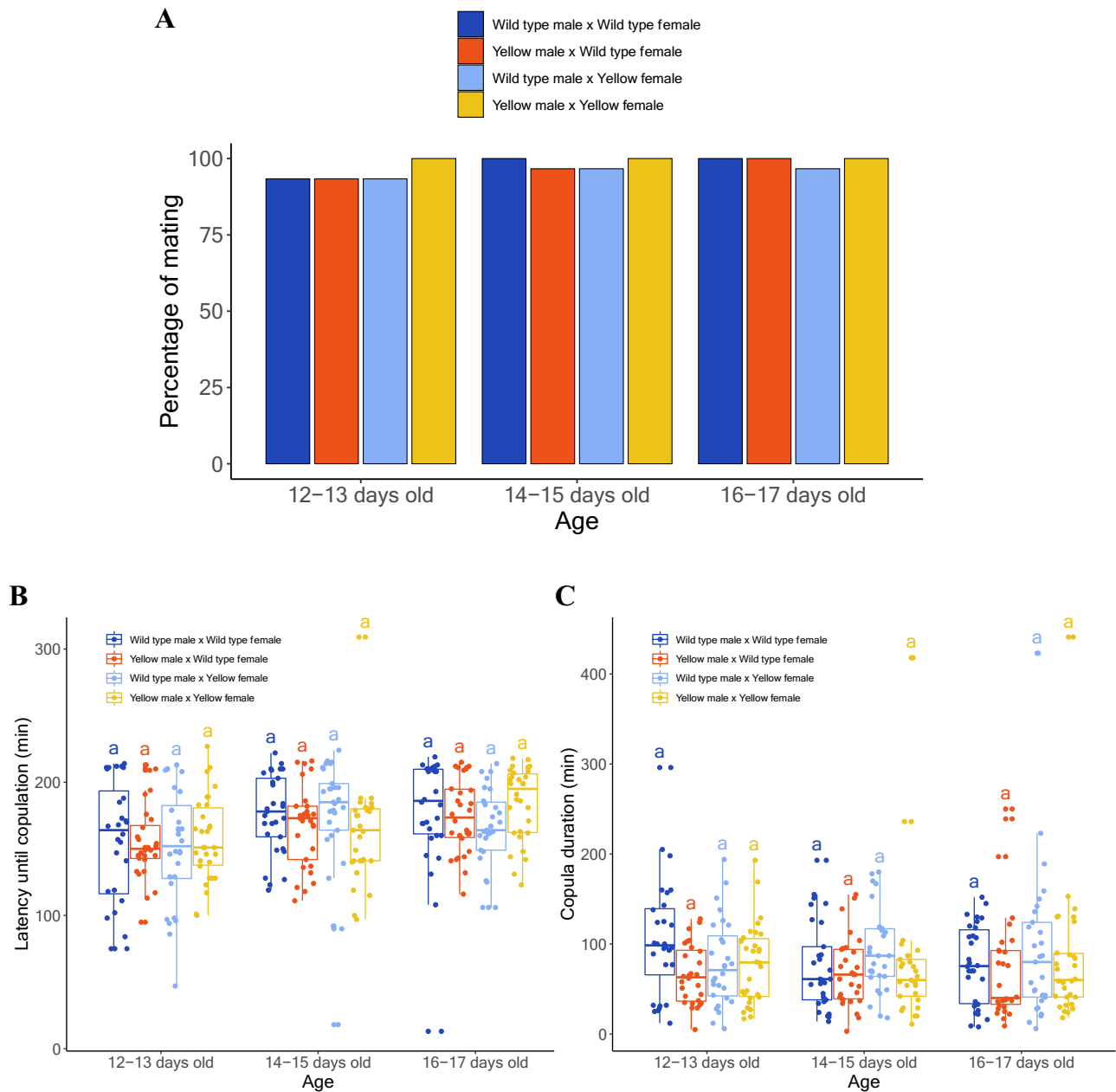


Fig. 6 Mating assays: yellow flies emerged a day earlier than wild-type flies, so mating assays were conducted with 12-, 14-, and 16-day-old wild-type flies and 13-, 15-, and 17-day-old yellow flies. **a** Percentage of pairs mated in each combination. **b** Latency until copulation. **c** Copulation duration. Box plots and raw data points are

shown. Box plots represent the interquartile range, and the median value is indicated. Error bars represent 1.5 times the interquartile range. Significance was assessed using Tukey's post hoc comparisons (no comparisons were significant)

daily (the lasted observed time is day 217). Under laboratory conditions, both strains maintained high survival rates for periods much longer than are relevant to SIT although there were modest differences between strains and sexes. The Weibull survival function was used to display the probability of survival at a given time according to sex and phenotype (Fig. 7a, Table S3). The survival curves of yellow males and females decrease faster than the curves for

wild-type male and female, respectively. The risk of death was quantified with the hazard function (Fig. 7b). The faster the survival function decreases, the higher the hazard. Corresponding to the survival curves (Fig. 7a), yellow males and females have higher hazard functions than wild-type males and females, respectively; however, a switch over between yellow and wild-type males is recorded at approximate day 150 (Fig. 7b). Figure 7c displays the estimates of the median

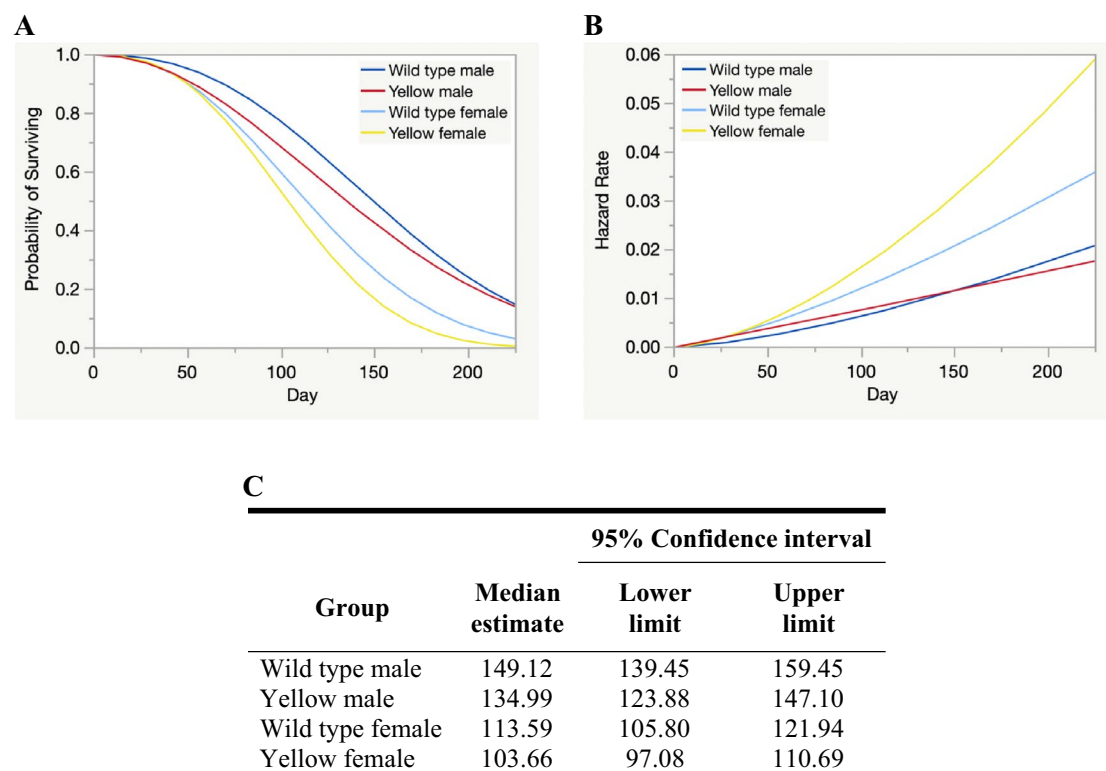


Fig. 7 Longevity assays. **a** Weibull survival functions showing the probability of survival over time. **b** The corresponding hazard functions in relation to age of yellow and wild-type males and females, and **c** estimates of median time to death for each group

and its 95% CI for each group. Weibull parametric regression with strain and sex as fixed effects detected significant differences in survivorship between the two strains and the two sexes, but no interaction (Wald test, strain: $\chi^2 = 5.8422$, $p = 0.0156$; sex: $\chi^2 = 67.5478$, $p < 0.0001$; interaction: $\chi^2 = 0.3019$, $p = 0.5827$).

Discussion

Evolution of *yellow-y* genes and their roles in body pigmentation

In 1911, Morgan described a *Drosophila melanogaster* strain with golden yellow wings, noting that the “entire fly is conspicuously yellow” (Morgan 1911). The phenotype was of notable importance, as it was subsequently used as one of six X-linked traits used to generate the first *D. melanogaster* linkage map, based on recombination rates (Sturtevant 1913). Two years later, Sturtevant (1915) compared mating competitiveness among several *D. melanogaster* mutant strains, including the yellow body flies. Although he argued that mating was an act of acceptance rather than choice, the yellow flies were “not so active as are the normal” and in terms of copulation success, “far behind their normal [wild

type] opponents”. Many independent mutations in the *D. melanogaster yellow-y* gene have been described (Drapeau 2006) and low mating success of yellow male mutants has been well documented (Bastock 1956; Drapeau 2006; Wilson et al. 1976). However, it was only recently confirmed that lack of melanisation of sex combs on the forelegs of *D. melanogaster* yellow males impeded mating success by limiting their ability to grasp and mount female (Massey et al. 2019).

Recessive mutations in the coding and regulatory sequence of the *yellow-y* gene cause the yellow phenotype through an overall reduction of melanin and produce a stable yellowish body colour throughout the flies’ lifespan (Wittkopp et al. 2002). However, *yellow-y* mutations in several other insect species can cause phenotypes markedly different to that of *D. melanogaster*. Knocking out *yellow-y* in the New World screw-worm fly *Cochliomyia hominivorax* and the Australian sheep blowfly *Lucilia cuprina* causes strong yellowish body colour upon eclosion but, unlike *D. melanogaster*, cuticle pigmentation largely recovers some hours later resulting in brownish body colour (Paulo et al. 2019). Loss of *yellow-y* expression also affects melanin production in the whole body in the housefly *Musca domestica* (Heinze et al. 2017), but only affects pigmentation of the hindwing in the coleopteran *Trinolium castaneum* (Arakane et al. 2010),

and specific body parts such as abdomen and hindwings in the hemipteran *Oncopeltus fasciatus* (Liu et al. 2016). In *B. tryoni*, the pale-yellow wing margin and brownish abdomens of our yellow mutants are readily distinguishable from wild type, although the mutation does not cause such extreme phenotypes as seen in *D. melanogaster*.

Duplication of *yellow-y* is likely to have occurred in a common ancestor of *B. tryoni*, *B. dorsalis* and *B. latifrons* and positive selection after the duplication suggest functional divergence of these paralogs. The yellow phenotype is recessive, and although wild-type *yellow-y2* alleles were detected in yellow G_1 *B. tryoni*, generating separate homozygous knockouts of *yellow-y1* or *yellow-y2* will be valuable in understanding the function each gene plays in pigmentation. *Bactrocera tryoni yellow-y1* showed the highest level of expression during pupal development, which is also the case in for *D. melanogaster*, where *yellow-y* is highly expressed in pupal stages and peak at the second half of pupal development (Walter et al. 1991). The highest expression levels of *B. tryoni yellow-y2* were detected in mature adults, when pigmentation has been completed and these expression profiles are consistent with those observed in *B. dorsalis* (Bai et al. 2019), which could contribute to the darkening of *B. tryoni* yellow flies in the hours after emergence.

Effect of disrupting *yellow-y* genes on the performance of *B. tryoni*

Decrements in performance following disruption of *yellow-y* genes were closely comparable to the decrements associated with the use of fluorescent dyes. In yellow mutant strain, the percentage of emergence was reduced by 5.80% compared with wild type (from 97.50% to 91.70%), whereas percentage of emergence of pupae coated with fluorescent dyes has been reported to be reduced by 8.30% (from 85.70% to 77.40%) (Dominiak et al. 2010a, b). Disruption of the *yellow-y* genes reduced rate of fliers by 11.58% (from 95.44% to 83.86%), while fluorescent dyes have been reported to reduce rate of fliers by 8.80% (from 92.1% to 83.3%) (Dominiak et al. 2010a).

Spontaneous locomotor activity is an integral component of most behavioural traits. In many insects, relations between activity levels and behaviours have been described, including finding essential resources at cold temperatures (Overgaard et al. 2010), adapting to changing temperatures (Loeschcke and Hoffmann 2007), as well as escaping from stressful conditions (Feder et al. 2010; Kjaersgaard et al. 2010). The dispersal of flies from release point is a key factor for the success of Sterile Insect Technique (SIT), and spontaneous activity monitor may provide an indication of dispersal tendency (Dominiak et al. 2014). Weldon et al. (2010) showed that activity level of wild *B. tryoni* is far higher than domesticated mass-reared flies. In the present

study, the activity rate of *B. tryoni* yellow flies was higher than wild-type flies over a four-hour period daylight that preceded mating activity at dusk, which may have benefits in mating performance or predator evasion. In *D. melanogaster*, disruption of DOPA-melanin synthesis in yellow mutants is predicted to lead to elevated dopamine levels, which in turn leads to increased levels of activity (Drapeau 2003; Rossi et al. 2015), which could explain the increased activity rate observed in *B. tryoni* yellow mutants.

In SIT, the ability of released sterile males to succeed in mating with wild females is an essential factor for programme success. Mutations in the *yellow-y* gene in *D. melanogaster* have been known to reduce mating success of the males, but not in all mutations (Drapeau 2006; Massey et al. 2019). In *D. melanogaster*, null *yellow-y* mutants caused by transversion of an A to C in the ATG initiation resulted in greatly reduced mating success than wild-type males due to structural change in foreleg bristles (sex combs) required for grasping females prior to copulation (Drapeau 2006; Massey et al. 2019). Here, our *B. tryoni* yellow flies were made by targeting exon 2. In non-competitive mating assays with wild-type females, yellow males and wild-type males both had comparable mating success, latency times prior to copulation as well as copula duration. *Bactrocera tryoni*'s forelegs and other legs do not have sex comb-like structures, but they do have foreleg hooks that are involved in contacting and grasping female during mating (Ekanayake et al. 2018). Reduced pigmentation was observed in fore-, mid-, and hindlegs of *B. tryoni* males, yet hook structures remain clear, and we did not observe significant differences in the mating ability of yellow flies. This finding indicates that disrupting *B. tryoni yellow-y* genes in exon 2 had very little effect on mating ability of male *B. tryoni*.

Under laboratory conditions, both the yellow and wild-type flies survived more than 217 days post-eclosion. When assessed over 217 days, lifespan of yellow flies was approximately 10 days shorter than wild-type flies in both sexes. To date, only one study has investigated the effect of fluorescent dyes on *B. tryoni* adult mortality, in which no adverse effect was found (Weldon 2005). It is likely that the decrement in longevity of yellow flies is of little consequence in practical terms. In SIT programmes, released flies rarely survive in the field for more than 3–4 weeks (Reynolds et al. 2012). Over this time frame the differences between yellow and wild-type strains were negligible, with both sexes of both strains exhibiting very high survivorship.

Our primary goal in this study was to develop a visible body marker which could potentially be used to distinguish released sterile *B. tryoni* flies from wild flies in nature. Disruption of *B. tryoni yellow-y* genes did cause some adverse effects on performance compared to the wild-type strain, but this generally appears to be less than or similar to effects of dye in other studies. Wild-type *yellow-y2* alleles

were present in yellow G_1 *B. tryoni*, which could affect the observed yellow strain's performance. And even though the off-target effects of CRISPR/Cas9 on yellow strain could have been minimised by inter-crossing 12 injected G_0 , off-target mutations could be in the background and affect the performance of the yellow flies. Therefore, backcrossing yellow strain to wild type, selecting homozygous strains for *yellow-y1* and *yellow-y2*, and then performing direct fitness comparisons between irradiated wild-type flies treated with fluorescent dye and irradiated yellow homozygous strains are now needed to more directly assess whether the yellow wing phenotype represents an improved strategy for identifying *B. tryoni* released in SIT programmes. Even if the yellow flies are slightly inferior in some performance measures, this may be more than compensated by reduced handling and cost, removal of a hazardous material from rearing facilities, and improved reliability of marking.

Authors' contributions SWB, PC, AC conceived the research. PT, AC, SWB designed experiments. AP generated a *B. tryoni* genome. TN performed experiments with assistance from AC, VM, CW. All authors played a role in data analysis. TN and SWB wrote the manuscript, and all authors edited and approved the final version.

Funding This work was funded by Hermon Slade Foundation Grant 18/06 and the National SITplus programme through Hort Innovation, using the research and development levy funds from the vegetable, apple and pear, citrus, strawberry, table grape, cherry, and summer fruit industries, with co-investment from South Australian Research and Development Institute (SARDI), the research arm of Primary Industries and Regions South Australia (PIRSA), and the Australian Government. Quality control assessment received additional support from the SITplus collaborative fruit fly programme. Project *Raising Q-fly Sterile Insect Technique to World Standard* (HG14033) is funded by the Hort Frontiers Fruit Fly Fund, part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from Macquarie University and contributions from the Australian Government. AP is funded by the Hawkesbury Institute for the Environment.

Availability of data and material Stuart Gilchrist and John Sved provided data for a *B. tryoni* genome assembly which was performed using the high performance computing resources at the Hawkesbury Institute for the Environment.

Compliance with ethical standards

Conflicts of interest The authors declare they have no conflicts of interest.

Ethical approval This research did not involve humans or vertebrates.

References

- Adnan SM, Farhana I, Inskoop J, Rempoulakis P, Taylor PW (2020) Dietary methoprene enhances sexual competitiveness of sterile male Queensland fruit flies in field cages. *J Pest Sci* 93:477–489. <https://doi.org/10.1007/s10340-019-01170-0>
- Arakane Y, Dittmer NT, Tomoyasu Y, Kramer KJ, Muthukrishnan S, Beeman RW, Kanost MR (2010) Identification, mRNA expression and functional analysis of several yellow family genes in *Tribolium castaneum*. *Insect Biochem Mol Biol* 40:259–266. <https://doi.org/10.1016/j.ibmb.2010.01.012>
- Bai X et al (2019) CRISPR/Cas9-mediated knockout of the eye pigmentation gene white leads to alterations in color of head spots in the oriental fruit fly, *Bactrocera dorsalis*. *Insect Mol Biol*. <https://doi.org/10.1111/imb.12592>
- Bastock M (1956) A gene mutation which changes a behaviour pattern. *Evolution* 10:421–439. <https://doi.org/10.1111/j.1558-5646.1956.tb02868.x>
- Bellini R, Medici A, Puggioli A, Balestrino F, Carrieri M (2013) Pilot field trials with *Aedes albopictus* irradiated sterile males in Italian urban areas. *J Med Entomol* 50:317–325. <https://doi.org/10.1603/ME12048>
- Benelli M, Ponton F, Lallu U, Mitchell KA, Taylor PW (2019) Cool storage of Queensland fruit fly pupae for improved management of mass production schedules. *Pest Manag Sci* 75:3184–3192. <https://doi.org/10.1002/ps.5436>
- Benelli M, Ponton F, Taylor PW (2019) Cool storage of Queensland fruit fly eggs for increased flexibility in rearing programs. *Pest Manag Sci* 75:1056–1064. <https://doi.org/10.1002/ps.5215>
- Campbell AJ, Lynch AJ, Dominiak BC, Nicol HI (2009) Effects of radiation, dye, day of larval hopping and vibration on eclosion of Queensland fruit fly, *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae). *Gen Appl Entomol J Entomol Soc NSW* 38:49–53
- Chang CL, Vargas RI, Caceres C, Jang E, Cho IK (2006) Development and assessment of a liquid larval diet for *Bactrocera dorsalis* (Diptera: Tephritidae). *Ann Entomol Soc Am* 99:1191–1198. [https://doi.org/10.1603/0013-8746\(2006\)99\[1191:DAAOAL\]2.0.CO;2](https://doi.org/10.1603/0013-8746(2006)99[1191:DAAOAL]2.0.CO;2)
- Choo A, Crisp P, Saint R, O'Keefe LV, Baxter SW (2018) CRISPR/Cas9-mediated mutagenesis of the white gene in the tephritid pest *Bactrocera tryoni*. *J Appl Entomol* 142:52–58. <https://doi.org/10.1111/jen.12411>
- Choo A et al (2019) Identification of Y-chromosome scaffolds of the Queensland fruit fly reveals a duplicated *gyf* gene paralogue common to many *Bactrocera* pest species. *Insect Mol Biol*. <https://doi.org/10.1111/imb.12602>
- Dominiak BC, Ekman JH (2013) The rise and demise of control options for fruit fly in Australia. *Crop Prot* 51:57–67. <https://doi.org/10.1016/j.cropro.2013.04.006>
- Dominiak BC, Schinagl L, Nicol H (2000) Impact of fluorescent marker dyes on emergence of sterile Queensland fruit fly, *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae). *Gen Appl Entomol J Entomol Soc NSW* 29:45–47
- Dominiak BC, Sundaralingam S, Jiang L, Jessup AJ, Barchia IM (2010) Impact of marker dye on adult eclosion and flight ability of mass produced Queensland fruit fly *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae). *Aust J Entomol* 49:166–169. <https://doi.org/10.1111/j.1440-6055.2010.00745.x>
- Dominiak BC, Sundaralingam S, Jiang L, Jessup AJ, Nicol HI (2010) Impact of marking dye, transport and irradiation on eclosion of mass produced Queensland Fruit Fly *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae). *Plant Prot Q* 25:141–143
- Dominiak BC, Fanson BG, Collins SR, Taylor PW (2014) Automated locomotor activity monitoring as a quality control assay for mass-reared tephritid flies. *Pest Manag Sci* 70:304–309. <https://doi.org/10.1002/ps.3559>
- Drapeau MD (2003) A novel hypothesis on the biochemical role of the *Drosophila* yellow protein. *Biochem Biophys Res Commun* 311:1–3. <https://doi.org/10.1016/j.bbrc.2003.09.106>
- Drapeau MD (2006) A cis-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating

- success. *Genetics* 172:1009–1030. <https://doi.org/10.1534/genetics.105.045666>
- Ekanayake WMTD, Clarke AR, Schutze MK (2018) Close-distance courtship of laboratory reared *Bactrocera tryoni* (Diptera: Tephritidae). *Austral Entomol* 58:578–588. <https://doi.org/10.1111/aen.12365>
- Enkerlin W et al (2015) Area freedom in Mexico from mediterranean fruit fly (Diptera: Tephritidae): a review of over 30 years of a successful containment program using an integrated area-wide SIT approach. *Fla Entomol* 98:665–681. <https://doi.org/10.1653/024.098.0242>
- Fanson BG, Petterson IE, Taylor PW (2013) Diet quality mediates activity patterns in adult Queensland fruit fly (*Bactrocera tryoni*). *J Insect Physiol* 59:676–681. <https://doi.org/10.1016/j.jinphys.2013.04.005>
- Fanson BG, Sundaralingam S, Jiang L, Dominiak BC, D'Arcy G (2014) A review of 16 years of quality control parameters at a mass-rearing facility producing Queensland fruit fly *Bactrocera tryoni*. *Entomol Exp Appl* 151:152–159. <https://doi.org/10.1111/eea.12180>
- FAO/IAEA/USDA (2019) Product quality control for sterile mass-reared and released tephritid fruit flies, Version 7.0. International Atomic Energy Agency, Vienna, Austria
- Feder ME, Garland T, Marden JH, Zera AJ (2010) Locomotion in response to shifting climate zones: not so fast. *Annu Rev Physiol* 72:167–190. <https://doi.org/10.1146/annurev-physiol-021909-135804>
- Ferreiro MJ et al (2018) *Drosophila melanogaster* white mutant w1118 undergo retinal degeneration. *Front Neurosci* 11:732. <https://doi.org/10.3389/fnins.2017.00732>
- Gilchrist AS et al (2014) The draft genome of the pest tephritid fruit fly *Bactrocera tryoni*: resources for the genomic analysis of hybridising species. *BMC Genom* 15:1153. <https://doi.org/10.1186/1471-2164-15-1153>
- Gnerre S et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* 108:1513. <https://doi.org/10.1073/pnas.1017351108>
- Gourzi P, Gubb D, Livadaras Y, Caceres C, Franz G, Savakis C, Zacharopoulou A (2000) The construction of the first balancer chromosome for the Mediterranean fruit fly, *Ceratitidis capitata*. *Mol Gen Genet* 264:127–136. <https://doi.org/10.1007/s004380000294>
- Greenfield P, Duesing K, Papanicolaou A, Bauer DC (2014) Blue: correcting sequencing errors using consensus and context. *Bioinformatics* 30:2723–2732. <https://doi.org/10.1093/bioinformatics/btu368>
- Haeussler M et al (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 17:148. <https://doi.org/10.1186/s13059-016-1012-2>
- Heinze SD et al (2017) CRISPR-Cas9 targeted disruption of the yellow ortholog in the housefly identifies the brown body locus. *Sci Rep* 7:4582. <https://doi.org/10.1038/s41598-017-04686-6>
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Bio Evol* 30:772–780. <https://doi.org/10.1093/molbev/mst010>
- Kjaersgaard A et al (2010) Locomotor activity of *Drosophila melanogaster* in high temperature environments: plastic and evolutionary responses. *Climate Res* 43:127–134. <https://doi.org/10.3354/cr00870>
- Knipling E (1955) Possibilities of insect control or eradication through use of sexually sterile males. *J Econ Entomol* 48:459–462
- Liu J, Lemonds TR, Marden JH, Popadić A (2016) A pathway analysis of melanin patterning in a Hemimetabolous insect. *Genetics* 203:403. <https://doi.org/10.1534/genetics.115.186684>
- Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25:402–408. <https://doi.org/10.1006/meth.2001.1262>
- Loeschcke V, Hoffmann AA (2007) Consequences of heat hardening on a field fitness component in *Drosophila* depend on environmental temperature. *Am Nat* 169:175–183. <https://doi.org/10.1086/510632>
- Mainali BP, Moadeli T, Ponton F, Taylor PW (2019) Comparison of gel larval diet with traditional lucerne chaff and carrot solid diets for rearing of Queensland fruit fly (Diptera: Tephritidae). *J Econ Entomol* 112:2278–2286. <https://doi.org/10.1093/jee/toz140>
- Massey JH, Wittkopp PJ (2016) The genetic basis of pigmentation differences within and between *Drosophila* species. *Curr Top Dev Biol* 119:27–61. <https://doi.org/10.1016/bs.ctdb.2016.03.004>
- Massey JH, Chung D, Siwanowicz I, Stern DL, Wittkopp PJ (2019) The yellow gene influences *Drosophila* male mating success through sex comb melanization. *Elife* 8:e49388. <https://doi.org/10.7554/eLife.49388>
- Meats A, Maheswaran P, Formmer M, Sved J (2002) Towards a male-only release system for SIT with the Queensland fruit fly, *Bactrocera tryoni*, using a genetic sexing strain with a temperature-sensitive lethal mutation. *Genetica* 116:97–106
- Moadeli T, Taylor PW, Ponton F (2017) High productivity gel diets for rearing of Queensland fruit fly, *Bactrocera tryoni*. *J Pest Sci* 90:507–520. <https://doi.org/10.1007/s10340-016-0813-0>
- Morgan TH (1911) The origin of nine wing mutations in *Drosophila*. *Science* 33:496. <https://doi.org/10.1126/science.33.848.496>
- Overgaard J, Sørensen JG, Jensen LT, Loeschcke V, Kristensen TN (2010) Field tests reveal genetic variation for performance at low temperatures in *Drosophila melanogaster*. *Funct Ecol* 24:186–195. <https://doi.org/10.1111/j.1365-2435.2009.01615.x>
- Paulo DF et al (2019) Specific gene disruption in the major livestock pests *Cochliomyia hominivorax* and *Lucilia cuprina* using CRISPR/Cas9. *G3 (Bethesda)* 9:3045–3055. <https://doi.org/10.1534/g3.119.400544>
- Perez-Staples D, Harmer AMT, Taylor PW (2007) Sperm storage and utilization in female Queensland fruit flies (*Bactrocera tryoni*). *Physiol Entomol* 32:127–135. <https://doi.org/10.1111/j.1365-3032.2006.00554.x>
- Pond SLK, Frost SDW, Muse SV (2004) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679. <https://doi.org/10.1093/bioinformatics/bti079>
- Rambaut A (2018) FigTree (version 1.4.4). <http://tree.bio.ed.ac.uk/software/figtree/>
- Reynolds OL, Smallridge CJ, Cockington VG, Penrose LD (2012) The effect of release method and trial site on recapture rates of adult sterile Queensland fruit fly, *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae) Australian. *J Entomol* 51:116–126. <https://doi.org/10.1111/j.1440-6055.2011.00845.x>
- Rossi FA, Boichicchio PA, Quesada-Allué LA, Pérez MM (2015) N- β -alanyldopamine metabolism, locomotor activity and sleep in *Drosophila melanogaster* ebony and tan mutants. *Physiol Entomol* 40:166–174. <https://doi.org/10.1111/phen.12100>
- Saul SH, McCombs SD (1992) Light eye color mutants as genetic markers for released populations of Hawaiian fruit flies (Diptera: Tephritidae). *J Econ Entomol* 85:1240–1245. <https://doi.org/10.1093/jee/85.4.1240>
- Smith PH (1979) Genetic manipulation of the circadian clock's timing of sexual behaviour in the Queensland fruit flies *Dacus tryoni* and *Dacus neohumeralis*. *Physiol Entomol* 4:71–78. <https://doi.org/10.1111/j.1365-3032.1979.tb00179.x>
- Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* 32:1342–1353. <https://doi.org/10.1093/molbev/msv022>

- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Sturtevant AH (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool* 14:43–59. <https://doi.org/10.1002/jez.1400140104>
- Sturtevant AH (1915) Experiments on sex recognition and the problem of sexual selection in *Drosophila*. *J Anim Behav* 5:351–366. <https://doi.org/10.1037/h0074109>
- Sultana S, Baumgartner JB, Dominiak BC, Royer JE, Beaumont LJ (2017) Potential impacts of climate change on habitat suitability for the Queensland fruit fly. *Sci Rep* 7:13025. <https://doi.org/10.1038/s41598-017-13307-1>
- Tychsen PH, Fletcher BS (1971) Studies on the rhythm of mating in the Queensland fruit fly *Dacus tryoni*. *J Insect Physiol* 17:2139–2156. [https://doi.org/10.1016/0022-1910\(71\)90174-0](https://doi.org/10.1016/0022-1910(71)90174-0)
- Vreysen MJB et al (2014) Sterile insects to enhance agricultural development: the case of sustainable tsetse eradication on Unguja Island Zanzibar, Using an Area-Wide Integrated Pest Management Approach. *PLOS Negl Trop Dis* 8:e2857. <https://doi.org/10.1371/journal.pntd.0002857>
- Walter MF, Black BC, Afshar G, Kermabon A-Y, Wright TRF, Biessmann H (1991) Temporal and spatial expression of the yellow gene in correlation with cuticle formation and DOPA decarboxylase activity in *drosophila* development. *Dev Biol* 147:32–45. [https://doi.org/10.1016/S0012-1606\(05\)80005-3](https://doi.org/10.1016/S0012-1606(05)80005-3)
- Weisenfeld NI et al (2014) Comprehensive variation discovery in single human genomes. *Nat Genet* 46:1350–1355. <https://doi.org/10.1038/ng.3121>
- Weldon C (2005) Marking Queensland fruit fly, *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae) with fluorescent pigments: pupal emergence, adult mortality, and visibility and persistence of marks. *Gen Appl Entomol* 34:7–13
- Weldon C, Meats A (2007) Short-range dispersal of recently emerged males and females of *Bactrocera tryoni*(Froggatt) (Diptera: Tephritidae) monitored by sticky sphere traps baited with protein and Lynfield traps baited with cue-lure Australian. *J Entomol* 46:160–166. <https://doi.org/10.1111/j.1440-6055.2007.00570.x>
- Weldon C, Meats A (2010) Dispersal of mass-reared sterile, laboratory-domesticated and wild male Queensland fruit flies. *J Appl Entomol* 134:16–25. <https://doi.org/10.1111/j.1439-0418.2009.01436.x>
- Weldon CW, Prenter J, Taylor PW (2010) Activity patterns of Queensland fruit flies (*Bactrocera tryoni*) are affected by both mass-rearing and sterilization. *Physiol Entomol* 35:148–153. <https://doi.org/10.1111/j.1365-3032.2010.00726.x>
- Wilson R, Burnet B, Eastwood L, Connolly K (1976) Behavioural pleiotropy of the yellow gene in *Drosophila melanogaster*. *Genet Res* 28(1):75–88. <https://doi.org/10.1017/s0016672300016748>
- Wittkopp PJ, True JR, Carroll SB (2002) Reciprocal functions of the *Drosophila* yellow and ebony proteins in the development and evolution of pigment patterns. *Development* 129:1849
- Wittkopp PJ, Carroll SB, Kopp A (2003) Evolution in black and white: genetic control of pigment patterns in *Drosophila*. *Trends Genet* 19:495–504. [https://doi.org/10.1016/S0168-9525\(03\)00194-X](https://doi.org/10.1016/S0168-9525(03)00194-X)
- Wyss JH (2000) Screwworm eradication in the Americas. *Ann N Y Acad Sci* 916:186–193. <https://doi.org/10.1111/j.1749-6632.2000.tb05289.x>
- Zavala-López JL, Enkerlin W (2017) Guideline for packing, shipping, holding and release of sterile flies in area-wide fruit fly control programmes, 2nd edn. Food and Agriculture Organization of the United Nations (FAO), Italy
- Zhao JT, Bennett CL, Stewart GJ, Frommer M, Raphael KA (2003) The scarlet eye colour gene of the tephritid fruit fly: *Bactrocera tryoni* and the nature of two eye colour mutations. *Insect Mol Biol* 12:263–269. <https://doi.org/10.1046/j.1365-2583.2003.00410.x>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix B

Supplementary figures and tables for Chapter 6.

Appendix B:

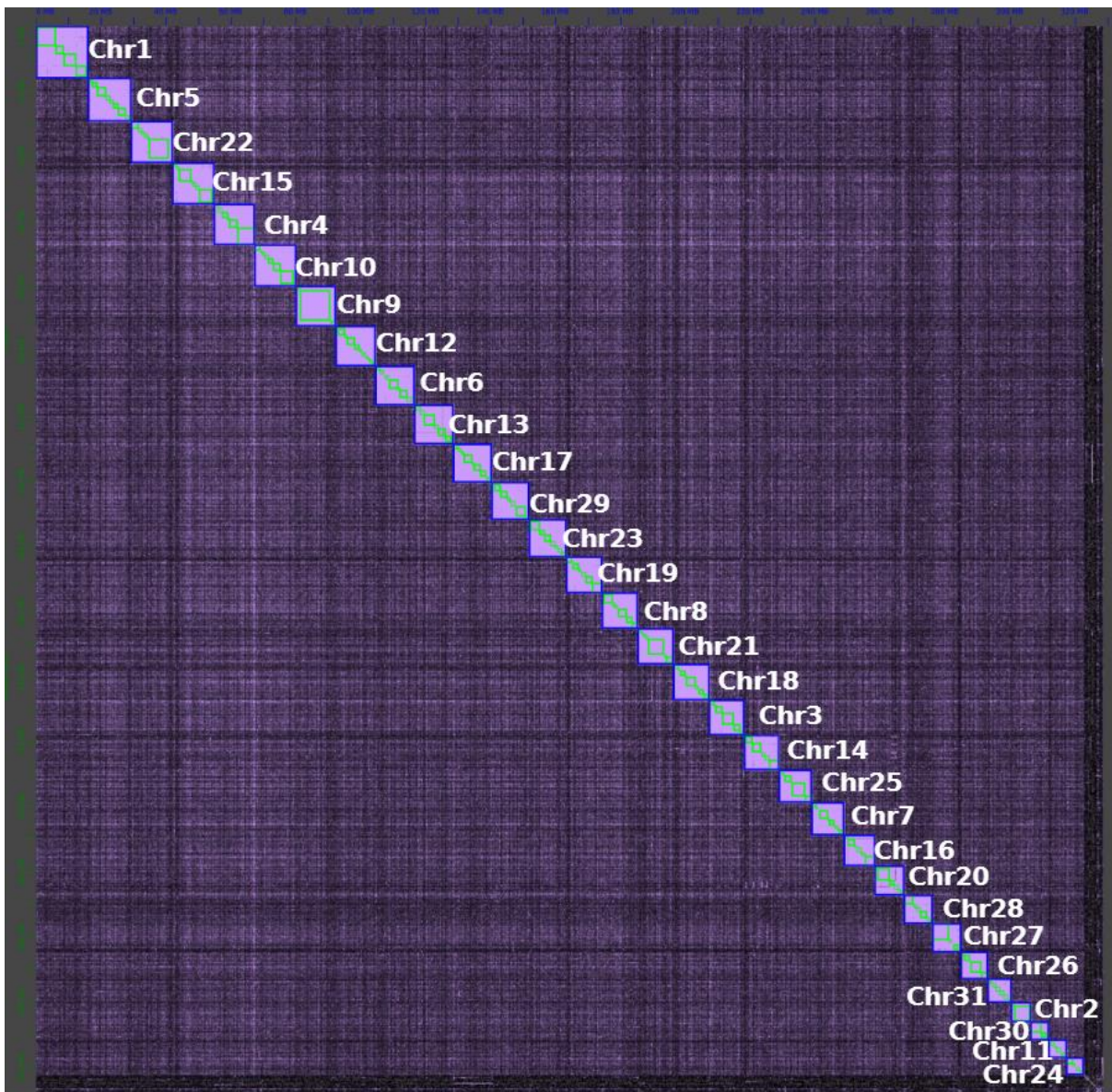


Fig. S1. Hi-C contact matrix output from JBAT (Durand et al., 2016) showing chromosome (blue squares) and contig (green squares) blocks. Chromosome blocks are ordered by size with chromosome numbers indicated. Differences in color intensity between Fig. S2. and the contact matrix of Fig. 2A. are due to the bin size, scaling and analysis programs. Higher resolution Hi-C interaction for Chr1, Chr5, Chr22 and Chr31 can be found in Fig S3.

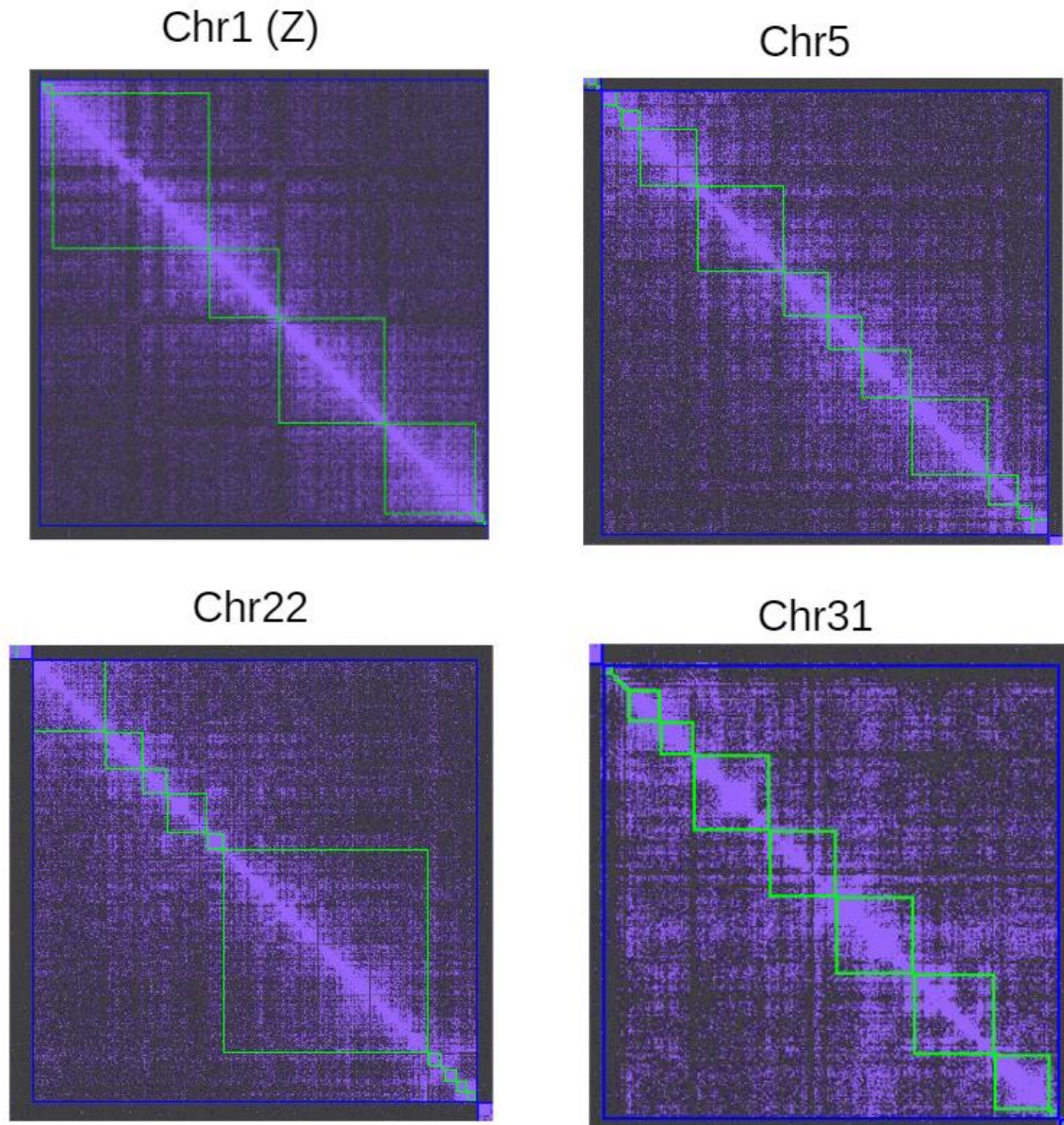


Fig. S2. Images generated using JuiceBox Tools showing Hi-C interactions between contigs for the three longest chromosomes (Chr1, Chr5, Chr22) and a chromosome containing the ryanodine receptor (Chr31). Green squares represent separate contigs. Hi-C interactions are highly contiguous between neighbouring contigs with no clear incongruence within contigs or chromosome level scaffolds, indicating that Hi-C scaffolding was robust.

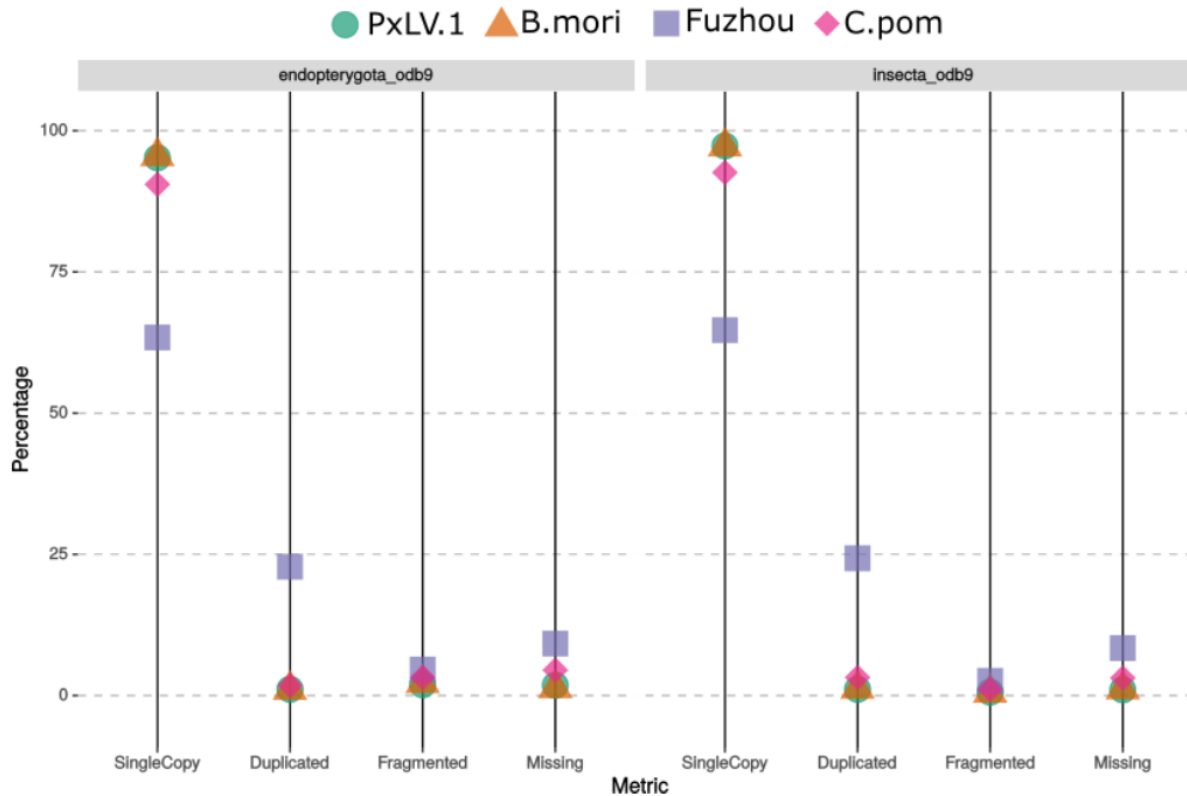


Fig. S3. Parallel coordinate plot presenting the percentage of Endopterygota and Insecta BUSCO gene sets identified in assembled genomes of *P. xylostella* from PxLV.1 (this study) and DBM_FJ_V1.1 (You et al., 2013), *Bombyx mori* from SilkBase-Nov.2016 (Kawamoto et al., 2019) and *Cydia pomonella* from GCA_003425675.2 (Wan et al., 2019). Metrics are shown for protein coding genes that are complete and single copy, complete and duplicated, fragmented, and missing.

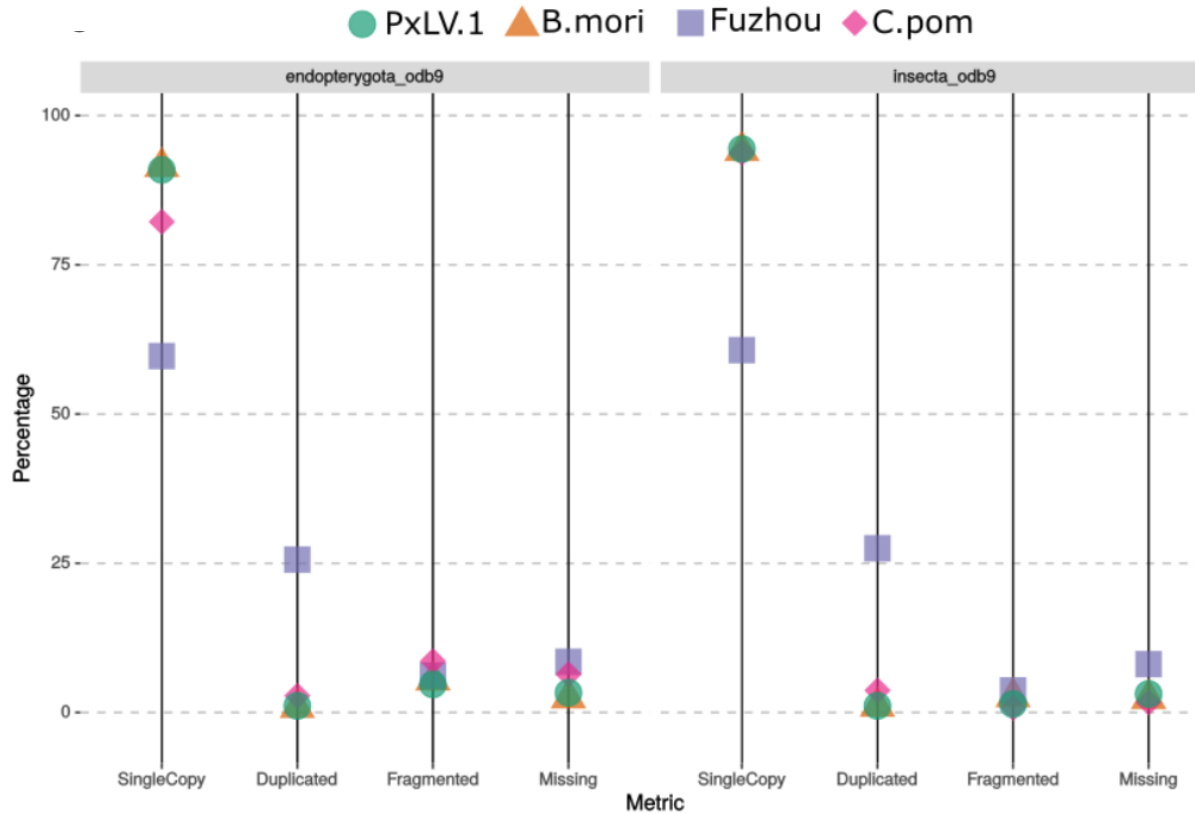


Fig. S4. Parallel coordinate plot presenting the percentage of Endopterygota and Insecta BUSCO gene sets identified in protein gene annotations from *P. xylostella* PxLV.1 (this study) and DBM_FJ_V1.1 (Tang et al., 2014; You et al., 2013), *Bombyx mori* SilkBase-Nov.2016 (Kawamoto et al., 2019) and *Cydia pomonella* GCA_003425675.2 (Wan et al., 2019). Metrics are shown for protein coding genes that are complete and single copy, complete and duplicated, fragmented, and missing.

A

KA17 CGGCGCCGCTGTCTCTCTCGGAGGAAGTTCTACACCTTGAAGTACGTGGCGCTGGTGCCTCTGCATCAACTTTGTGCTGTCTATAAGGTATCCACCCTGGAAGCAGCCTGGTGAAGGTCGGCATCGGAG
PxLV.1 CGGCGCCGCTGTCTCTCTCGGAGGAAGTTCTACACCTTGAAGTACGTGGCGCTGGTGCCTCTGCATCAACTTTGTGCTGTCTATAAGGTATCCACCCTGGAAGCAGCCTGGTGAAGGTCGGCATCGGAG
R R A V S F L A R K F Y T L K Y V A L V L A F C I N F V L L F Y K V S T L D A E P G E G S G I G

KA17 ACATCATCAGCGCTCCGGGTCAAGTCTGGTTCGGGCAGCGCGGAGGAGCGCGGGAAGAAGACGAGGACCCGATAGAGCTGGTCATATAGACGAGGATTACTTCTACATGGAACCGTCATCAAGATAGCTGCTTACT
PxLV.1 ACATCATCAGCGCTCCGGGTCAAGTCTGGTTCGGGCAGCGCGGAGGAGCGCGGGAAGAAGACGAGGACCCGATAGAGCTGGTCATATAGACGAGGATTACTTCTACATGGAACCGTCATCAAGATAGCTGCTTACT
D I I S G S G S G S G S G S G D G G S G E E D E D P I E L V H I D E D Y F Y M E H V I K I A A L L

KA17 GCATTCATAGTGTCTGGCTAAATGATTGGGTACTACCATTTGAAGTCCCGTAGCCATCTCAAGCTGAGAAAGGATAGCTGTAACCTGGAGTTCCAGGCTTACATAGCAGCAGCAAGAGAGACGACTG
PxLV.1 GCATTCATAGTGTCTGGCTAAATGATTGGGTACTACCATTTGAAGTCCCGTAGCCATCTCAAGCTGAGAAAGGATAGCTGTAACCTGGAGTTCCAGGCTTACATAGCAGCAGCAAGAGAGACGACTG
H S I V S L A K L I G Y Y H L K V P L A I F K R R E K E I A R K L E F D G L Y I A E Q P E D D D L

KA17 AAGAGCCACTGGGACAAGCTGCATATCTGCCAAGTCTCCAGTGAAGTCTGGGACAAAGTTCGTGAAGAAGAAGGTCGCGTCAAGTACTCGGAGACTTACGACTTCGACTCCATCTCCACATGCTGGCATGGAGAAGA
PxLV.1 AAGAGCCACTGGGACAAGCTGCATATCTGCCAAGTCTCCAGTGAAGTCTGGGACAAAGTTCGTGAAGAAGAAGGTCGCGTCAAGTACTCGGAGACTTACGACTTCGACTCCATCTCCACATGCTGGCATGGAGAAGA
K S H W D K L V I S A K S F P V N Y W D K F V K K K V R V K Y S E T Y D F D S I S N M L G M E K

KA17 GCGCTTCGGACGCAAGATGAGGGCAGAGGGTCTTCCACTACATCTTAAGCATAGACTGGCGTACCAAGTGTGGAAGCCGGGTACAGATCACAGACAACCTGCTCTTACTCTCTGGTACTTCTGCTCTGT
PxLV.1 GCGCTTCGGACGCAAGATGAGGGCAGAGGGTCTTCCACTACATCTTAAGCATAGACTGGCGTACCAAGTGTGGAAGCCGGGTACAGATCACAGACAACCTGCTCTTACTCTCTGGTACTTCTGCTCTGT
T A F A T Q E D E D E G R G F F H Y I L S I D W R Y Q V W R A G V T I T D N S F L Y S L W Y F S F S V

KA17 GATGGCAACTTCAACCACTTCTTTCGCGCCATCTGTTGGACGTGGCTGTGGTTCAAGACTGTGAGGACTAT
PxLV.1 GATGGCAACTTCAACCACTTCTTTCGCGCCATCTGTTGGACGTGGCTGTGGTTCAAGACTGTGAGGACTAT
M G N F N H F F F A A H L L D V A V G F K T L R T

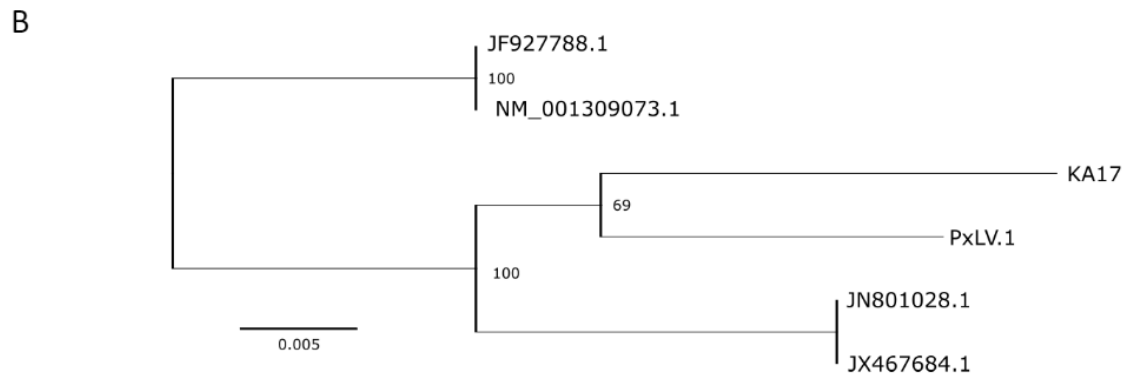


Fig. S5. Comparison of RyR partial coding sequence between the *P. xylostella* PxLV.1 genome, representing the LV-R strain from the Lockyer Valley, and the KA17 strain (Japan) assembled from RNAseq data. A) Sequence alignment of exons 107 – 119 of KA17 and PxLV.1 with amino acid translations. The 4790K mutation is outlined in black. Colored bases indicate 28 synonymous substitutions across 803 bp of coding sequence. B) Nucleotide alignment of six mRNA sequences produced an 803 bp maximum likelihood phylogeny of the RyR region shown in A). The mRNA accessions include the reference sequence NM_001309073.1, which was derived from a strain collected in Beijing, China (JF927788.1 (Sun et al., 2012)) and two sequences are from the *P. xylostella* strain Roth (JX467684.1 and JN801028.1).

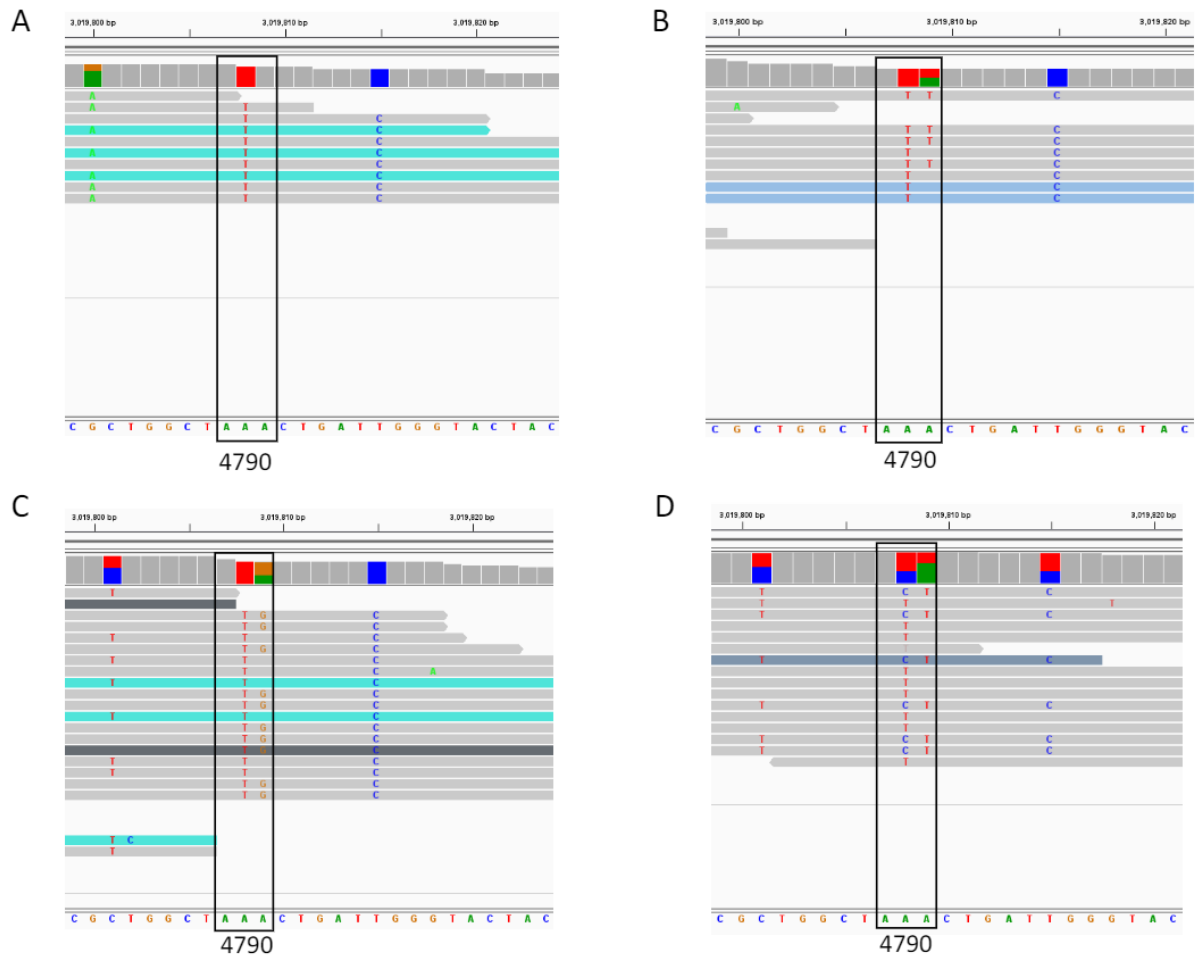


Fig. S6. Sequence reads from individual *P. xylostella* samples were aligned to the PxLV.1 genome and the RyR codon 4790 was genotyped. A) Example of wild type I4790 ATA codon (ERR2512145), B) example of a synonymous I4790 ATT codon (ERR2512250), C) I4790M mutation (ERR2512233), and D) I4790T mutation (ERR2520854). Screenshots were taken in IGV of the region immediately surrounding the 4790 codon beginning at base position 3,019,807 on Chromosome 31 of the PxLV.1 genome. To avoid potential genotyping bias caused by low sequence read depth, only one genotype was recorded for each individual, representing the best supported allele.

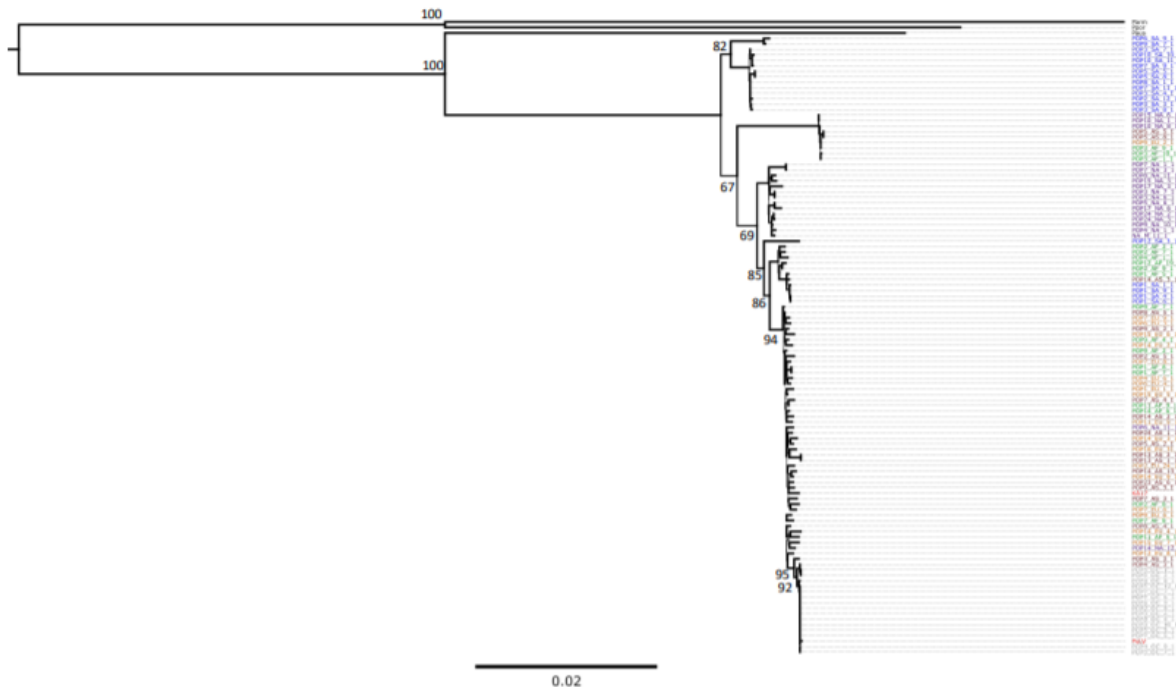


Fig. S7. Phylogeny of 12 mitochondrial genes (ND1, ND2, ND3, ND4, ND4L, ND5, ND6, CYTB, COX1, COX2, COX3, ATP6) using 113 samples from (You et al., 2020), KA17 and LV-R strains and three outgroup *Plutella* species *P. australiana* (*P. aus*), *P. porrectella* (*P. por*) and *P. armoraciae* (*P. arm*). Branch lengths between samples are scaled and bootstrap (n=10000) support values are provided for the same nodes as Fig. 5. *Plutella xylostella* samples are colored based on population of collection; South America (SA, blue), North America (NA), Asia (AS, maroon), Oceania (OC, grey), Africa (AF, green) and Europe (EU, orange). LV-R and KA17 are highlighted in red.

Table S1. *Plutella xylostella* genome sequences published by You et al. (2020) that were used to produce a mitochondrial phylogeny (Fig. 5 and Fig. S7) and genotype RyR codon 4790.

ENA accession	Sample Name	Sample in mitochondrial phylogeny	Location	Continent	Best supported genotype for RyR codon 4790	RyR 4790 amino acid
ERR2486112	NA_M_11_1	yes	Romita, Mexico	NA	ATA	Isoleucine
ERR2508315	POP9_SA_7_1	yes	Huaral, Peru	SA	ATA	Isoleucine
ERR2512131	POP9_NA_10_1	yes	Montana, USA	NA	ATA	Isoleucine
ERR2512133	POP9_EU_6_1	yes	Moscow, Russia	EU	ATA	Isoleucine
ERR2512134	POP9_EU_2_1	yes	Moscow, Russia	EU	ATT	Isoleucine
ERR2512135	POP9_AS_3_1	yes	Yamanashi-ken, Japan	AS	ATA	Isoleucine
ERR2512136	POP9_AS_21_1	no	Yamanashi-ken, Japan	AS	ATA	Isoleucine
ERR2512137	POP9_AS_2_1	yes	Yamanashi-ken, Japan	AS	ATA	Isoleucine
ERR2512140	POP9_AF_7_1	yes	Cairo, Egypt	AF	ATA	Isoleucine
ERR2512141	POP9_AF_4_1	yes	Cairo, Egypt	AF	ATA	Isoleucine
ERR2512143	POP9_AF_2_1	yes	Cairo, Egypt	AF	ATA	Isoleucine
ERR2512144	POP8_SA_1_1	yes	Osorno, Chile	SA	ATA	Isoleucine
ERR2512145	POP8_OC_4_1	no	Port Vila, Vanuatu	OC	ATA	Isoleucine
ERR2512153	POP8_NA_11_1	yes	North Carolina, USA	NA	ATA	Isoleucine
ERR2512161	POP8_AS_5_1	yes	Sapporo, Japan	AS	ATA	Isoleucine
ERR2512162	POP8_AS_4_1	yes	Sapporo, Japan	AS	ATA	Isoleucine
ERR2512169	POP7_SA_9_1	yes	La Serena, Chile	SA	ATA	Isoleucine
ERR2512173	POP7_SA_11_1	yes	La Serena, Chile	SA	ATA	Isoleucine
ERR2512174	POP7_OC_7_1	yes	Auckland, New Zealand	OC	NNN	-
ERR2512175	POP7_OC_6_1	yes	Auckland, New Zealand	OC	ATA	Isoleucine
ERR2512176	POP7_OC_5_1	yes	Auckland, New Zealand	OC	ATA	Isoleucine
ERR2512180	POP7_NA_3_1	yes	Hawaii, USA	NA	ATA	Isoleucine
ERR2512183	POP7_NA_1_1	yes	Hawaii, USA	NA	ATA	Isoleucine
ERR2512184	POP7_EU_9_1	yes	Marathon, Greece	EU	ATA	Isoleucine
ERR2512185	POP7_EU_8_1	yes	Marathon, Greece	EU	ATA	Isoleucine
ERR2512186	POP7_EU_6_1	yes	Marathon, Greece	EU	NNN	-
ERR2512189	POP7_AS_4_1	yes	Kagoshima, Japan	AS	ATA	Isoleucine
ERR2512190	POP7_AS_3_1	yes	Kagoshima, Japan	AS	ATA	Isoleucine
ERR2512195	POP7_AF_8_1	yes	Togo	AF	ATT	Isoleucine
ERR2512199	POP6_SA_9_1	yes	Arica, Chile	SA	ATA	Isoleucine
ERR2512207	POP6_NA_11_1	yes	Vauxhaul, Alberta, Canada	NA	ATA	Isoleucine
ERR2512210	POP6_EU_4_1	yes	Padova, Italy	EU	ATA	Isoleucine
ERR2512211	POP6_EU_3_1	yes	Padova, Italy	EU	ATA	Isoleucine
ERR2512213	POP5_SA_8_1	yes	Cordoba, Argentina	SA	NNN	-

ERR2512215	POP5_SA_6_1	no	Cordoba, Argentina	SA	NNN	-
ERR2512216	POP5_SA_5_1	yes	Cordoba, Argentina	SA	ATA	Isoleucine
ERR2512218	POP5_OC_9_1	yes	Upolu, Samoa	OC	ATA	Isoleucine
ERR2512222	POP5_OC_5_1	yes	Upolu, Samoa	OC	ATA	Isoleucine
ERR2512223	POP5_OC_2_1	yes	Upolu, Samoa	OC	ATA	Isoleucine
ERR2512224	POP5_OC_12_1	yes	Upolu, Samoa	OC	NNN	-
ERR2512225	POP5_OC_1_1	yes	Upolu, Samoa	OC	ATA	Isoleucine
ERR2512228	POP5_NA_6_1	yes	Saskatchewan, Canada	NA	ATA	Isoleucine
ERR2512232	POP5_AS_8_1	yes	North Sulawesi, Indonesia	AS	ATT	Isoleucine
ERR2512233	POP5_AS_5_1	yes	North Sulawesi, Indonesia	AS	ACT	Threonine
ERR2512235	POP5_AS_2_1	yes	North Sulawesi, Indonesia	AS	ATA	Isoleucine
ERR2512241	POP4_OC_9_1	yes	New South Wales, Australia	OC	ATA	Isoleucine
ERR2512250	POP4_NA_1_1	yes	Manitoba, Canada	NA	ATT	Isoleucine
ERR2512251	POP4_EU_6_1	yes	Bretagne, France	EU	ATA	Isoleucine
ERR2512259	POP4_AS_3_1	yes	West Java, Indonesia	AS	ATA	Isoleucine
ERR2512266	POP4_AF_1_1	yes	Namibia	AF	ATA	Isoleucine
ERR2519249	POP3_SA_8_1	yes	Montevideo, Uruguay	SA	ATA	Isoleucine
ERR2519250	POP3_SA_7_1	yes	Montevideo, Uruguay	SA	ATA	Isoleucine
ERR2519252	POP3_SA_4_1	yes	Montevideo, Uruguay	SA	ATA	Isoleucine
ERR2519253	POP3_SA_3_1	yes	Montevideo, Uruguay	SA	ATA	Isoleucine
ERR2519254	POP3_OC_8_1	yes	Western Australia, Australia	OC	ATA	Isoleucine
ERR2519255	POP3_OC_7_1	yes	Western Australia, Australia	OC	ATA	Isoleucine
ERR2519256	POP3_OC_4_1	yes	Western Australia, Australia	OC	ATA	Isoleucine
ERR2519257	POP3_OC_10_1	yes	Western Australia, Australia	OC	ATA	Isoleucine
ERR2519259	POP3_NA_7_1	yes	Quebec, Canada	NA	ATA	Isoleucine
ERR2519260	POP3_NA_6_1	yes	Quebec, Canada	NA	ATA	Isoleucine
ERR2519265	POP3_AS_3_1	yes	Kota Kinabalu, Malaysia	AS	ATT	Isoleucine
ERR2519269	POP3_AF_9_1	yes	Tanzania	AF	ATA	Isoleucine
ERR2519271	POP3_AF_10_1	yes	Tanzania	AF	ATA	Isoleucine
ERR2519273	POP3_AF_1_1	yes	Tanzania	AF	ATA	Isoleucine
ERR2519324	POP2_SA_12_1	yes	Santa Maria, Brazil	SA	ATA	Isoleucine
ERR2519325	POP2_OC_8_1	yes	Tasmania, Australia	OC	ATA	Isoleucine
ERR2519327	POP2_OC_5_1	yes	Tasmania, Australia	OC	ATA	Isoleucine
ERR2519348	POP10_NA_1_1	yes	Maine, USA	NA	ATA	Isoleucine
ERR2519351	POP10_NA_8_1	yes	Maine, USA	NA	ATA	Isoleucine
ERR2519358	POP10_SA_10_1	yes	Tulcan, Ecuador	SA	ATT	Isoleucine
ERR2519359	POP10_SA_11_1	yes	Tulcan, Ecuador	SA	ATA	Isoleucine
ERR2519365	POP11_AF_5_1	yes	Bobo-Dioulasso, Burkina Faso	AF	ATA	Isoleucine
ERR2519367	POP11_AF_9_1	yes	Bobo-Dioulasso, Burkina Faso	AF	ATA	Isoleucine
ERR2519376	POP11_EU_9_1	yes	Volgograd, Russia	EU	ATA	Isoleucine
ERR2519384	POP12_AF_19_1	yes	Manica, Mozambique	AF	ATA	Isoleucine
ERR2519395	POP12_EU_8_1	yes	Ekaterinburg, Russia	EU	ATA	Isoleucine
ERR2519396	POP12_NA_1_1	no	Michigan, USA	NA	ATA	Isoleucine

ERR2519403	POP12_SA_3_1	yes	Mucuchies, Venezuela	SA	ATA	Isoleucine
ERR2519410	POP13_AS_1_1	yes	Cameron highland, Malaysia	AS	ATA	Isoleucine
ERR2519473	POP13_AS_2_1	yes	Cameron highland, Malaysia	AS	ATA	Isoleucine
ERR2520498	POP14_AF_6_1	yes	Zewai, Ethiopia	AF	NNN	-
ERR2520835	POP14_AF_9_1	no	Zewai, Ethiopia	AF	ATA	Isoleucine
ERR2520836	POP14_AS_13_1	yes	Vientiane, Laos	AS	ATA	Isoleucine
ERR2520837	POP14_AS_2_1	yes	Vientiane, Laos	AS	ATA	Isoleucine
ERR2520838	POP14_AS_3_1	yes	Vientiane, Laos	AS	ATA	Isoleucine
ERR2520841	POP14_EU_2_1	yes	Vladivostok, Russia	EU	ATA	Isoleucine
ERR2520843	POP14_EU_4_1	yes	Vladivostok, Russia	EU	ATA	Isoleucine
ERR2520844	POP14_EU_9_1	yes	Vladivostok, Russia	EU	ATA	Isoleucine
ERR2520846	POP14_NA_12_1	yes	Maryland, USA	NA	ATA	Isoleucine
ERR2520854	POP14_EU_3_1	yes	Vladivostok, Russia	EU	ATG	Methionine
ERR2520862	POP15_EU_5_1	yes	Ocsa, Hungary	EU	ATA	Isoleucine
ERR2520863	POP15_EU_6_1	yes	Ocsa, Hungary	EU	ATA	Isoleucine
ERR2520864	POP15_EU_7_1	yes	Ocsa, Hungary	EU	ATA	Isoleucine
ERR2520866	POP15_NA_3_1	yes	Alabama, USA	NA	ATA	Isoleucine
ERR2520869	POP15_NA_6_1	yes	Alabama, USA	NA	ATA	Isoleucine
ERR2520881	POP16_EU_31_1	yes	Alnarp, Sweden	EU	ATA	Isoleucine
ERR2520886	POP17_NA_3_1	yes	Texas, USA	NA	ATA	Isoleucine
ERR2520889	POP17_NA_6_1	yes	Texas, USA	NA	ATA	Isoleucine
ERR2520898	POP1_AF_5_1	yes	Pretoria, South Africa	AF	ATA	Isoleucine
ERR2520900	POP1_AF_7_1	yes	Pretoria, South Africa	AF	ATA	Isoleucine
ERR2520901	POP1_AF_8_1	yes	Pretoria, South Africa	AF	ATA	Isoleucine
ERR2520909	POP1_EU_1_1	yes	West Cornwall, England	EU	ATA	Isoleucine
ERR2520916	POP1_SA_1_1	yes	Recife, Brazil	SA	ATA	Isoleucine
ERR2520917	POP1_SA_4_1	yes	Recife, Brazil	SA	ATA	Isoleucine
ERR2520918	POP1_SA_6_1	yes	Recife, Brazil	SA	ATA	Isoleucine
ERR2520920	POP1_SA_9_1	yes	Recife, Brazil	SA	ATA	Isoleucine
ERR2521228	POP23_AS_6_1	yes	Rahuri, India	AS	ATT	Isoleucine
ERR2521230	POP24_AS_1_1	yes	Fujian, China	AS	ATA	Isoleucine
ERR2521287	POP24_NA_2_1	yes	Seattle, USA	NA	ATA	Isoleucine
ERR2521288	POP24_NA_3_1	yes	Seattle, USA	NA	ATA	Isoleucine
ERR2521682	POP2_AF_2_1	yes	Cape Town, South Africa	AF	ATT	Isoleucine
ERR2521686	POP2_AF_7_1	yes	Cape Town, South Africa	AF	ATA	Isoleucine
ERR2521688	POP2_AF_8_1	yes	Cape Town, South Africa	AF	ATA	Isoleucine
ERR2521690	POP2_AF_9_1	yes	Cape Town, South Africa	AF	ATA	Isoleucine
ERR2521753	POP2_AS_5_1	yes	Phetchabun, Thailand	AS	ATA	Isoleucine
ERR2521759	POP2_EU_24_1	yes	Madrid, Spain	EU	ATA	Isoleucine

Table S2. Summary of PxLV.1 chromosome size and content.

Chromosome	Length (bp)	Number of genes	Number of exons	Total gene size (exons + introns)	Number of tRNAs	GC content (%)	Mean gene length	Mean exon length	Mean intron length	Mean exons per gene
1 (Z)	16,120,546	726	4695	9,525,684	118	38	13,120	340	1,998	6.5
2	6,001,377	384	2193	3,692,819	114	38.3	9,616	296	1,556	5.5
3	10,736,100	614	3158	5,214,616	122	38.2	8,492	338	1,630	5.1
4	12,569,824	798	4493	6,328,359	133	39	7,930	362	1,299	5.8
5	13,304,806	829	4897	7,172,082	133	38.4	8,651	302	1,399	5.9
6	12,125,418	721	3673	5,804,257	127	37.8	8,050	286	1,610	5.1
7	9,927,925	542	2811	5,089,143	136	37.8	9,389	315	1,853	5.2
8	11,057,709	649	4044	6,265,604	95	38.4	9,654	271	1,523	6.2
9	12,337,041	606	3420	5,723,042	153	37.9	9,443	320	1,645	5.6
10	12,562,405	764	4411	6,757,416	135	38.9	8,844	326	1,459	5.8
11	5,444,954	349	2124	3,143,508	120	38.9	9,007	309	1,401	6.1
12	12,234,112	691	3851	6,167,467	127	38.3	8,925	311	1,572	5.6
13	12,013,341	652	3623	5,476,352	143	38.3	8,399	324	1,448	5.6
14	10,144,654	509	2626	4,865,320	172	38.1	9,558	311	1,912	5.2
15	12,709,421	858	5357	7,330,769	103	38.9	8,544	286	1,289	6.2
16	9,650,281	590	3631	5,481,758	132	38.7	9,291	278	1,471	6.2
17	11,703,668	711	3975	6,040,875	147	38.4	8,496	291	1,496	5.6
18	11,009,671	602	2954	4,790,134	167	37.9	7,957	326	1,627	4.9
19	11,068,360	675	3543	5,413,292	203	38.4	8,019	325	1,486	5.2
20	9,037,267	566	3055	4,867,607	132	37.9	8,600	286	1,604	5.4
21	11,026,436	524	2898	5,004,539	134	38.3	9,550	341	1,691	5.5
22	12,850,310	728	3931	6,585,026	143	38.5	9,045	303	1,684	5.4
23	11,483,547	601	2844	4,357,323	171	38.1	7,250	343	1,508	4.7
24	5,378,391	402	2489	3,701,770	77	37.8	9,208	266	1,456	6.2
25	10,050,624	631	3712	5,468,182	126	37.7	8,665	283	1,434	5.9
26	8,328,171	415	2304	3,927,913	173	39.3	9,464	303	1,710	5.6
27	8,603,041	366	2269	4,349,594	133	38.8	11,884	326	1,897	6.2
28	8,750,093	369	2069	4,060,897	218	37.7	11,005	300	2,024	5.6
29	11,508,118	713	3351	6,041,497	126	38.3	8,473	361	1,831	4.7
30	5,632,285	286	2019	3,025,191	130	38.3	10,577	251	1,453	7.1
31	7,235,916	330	1662	3,432,925	153	37.7	10,402	340	2,153	5.0
Total	322,605,812	18,201	102,082	165,104,961	4,296	-	-	-	-	-
Average	10,406,639	587	3,293	5,325,966	139	38	9,210	310	1,617	6

Table S3. Comparison of Ryanodine Receptor protein sequences from *P. xylostella* diamide resistant strain LV-R (5164 aa) and laboratory colonies AET09964.1 (5164 aa) and XP_037970554 (5151 aa). Grey shading indicates amino acid substitutions that are absent in LV-R.

Table S3. Comparison of Ryanodine Receptor protein sequences from <i>P. xylostella</i> diamide resistant strain LV-R (5164 aa) and laboratory colonies AET09964.1 (5164 aa) and XP_037970554 (5151 aa). Grey shading indicates amino acid substitutions that are absent in LV-R.			
Accession (strain)	PxLV.1 (LV-R)	AET09964.1 (Roth)	XP_037970554 (Laboratory colony)
amino acid ^A			
1746	A	V	A
2596	E	G	E
2723	L	S	L
2842	Q	P	Q
2965-2977 ^B	EEEVQIEVSDTTT	EEEVQIEVSDTAT	-----
2976	T	A	-
3162	I	M	I
3166	T	M	T
3390	Q	Q	H
3533	D	G	D
3792	K	R	K
4790	K	I	I

^A Position based on AET09964.1

^B RyR XP_037970554 was predicted by automated computational analysis from genome assembly "Haplomerged_assembly (GCF_905116875.1)". The gene model lacks this exon, but it is present in the corresponding genome sequence.

Table S4. *Plutella xylostella* genotypes for the RyR 4790 codon among individuals from six Australian populations.

Australian State	Location	Latitude	Longitude	Year	Number	Gender	Codon 4790 genotype
Western Australia	Boyup Brook	-33.640168	116.401075	2014	1	male	ATA / ATA
					2	male	ATA / ATA
					3	male	ATA / ATA
					4	female	ATA / ATA
					5	male	ATA / ATA
					6	female	ATA / ATA
					7	male	ATA / ATA
					8	female	ATA / ATA
Victoria	Ouyen	-34.995278	142.314944	2014	9	male	ATA / ATA
					10	male	ATA / ATA
					11	female	ATA / ATA
					12	male	ATA / ATA
					13	male	ATA / ATA
					14	female	ATA / ATA
					15	male	ATA / ATA
					16	male	ATA / ATA
New South Wales	Henty	-35.59567	146.949933	2014	17	male	ATA / ATA
					18	female	ATA / ATA
					19	female	ATA / ATA
					20	male	ATA / ATA
					21	female	ATA / ATA
					22	male	ATA / ATA
					23	female	ATA / ATA
					24	male	ATA / ATA
South Australia	Moonaree	-31.99147	135.871023	2014	25	unknown	ATA / ATA
					26	unknown	ATA / ATA
					27	unknown	ATA / ATA
					28	unknown	ATA / ATA
					29	unknown	ATA / ATA
					30	unknown	ATA / ATA
					31	unknown	ATA / ATA
					32	unknown	ATA / ATA
Queensland	Bundaberg	-24.799891	152.262575	2014	33	male	ATA / ATA
					34	male	ATA / ATA
					35	male	ATA / ATA
					36	female	ATA / ATA
					37	female	ATA / ATA
					38	male	ATA / ATA

					39	female	ATA / ATA
Tasmania	Launceston	-41.469613	147.14097	2014	40	female	ATA / ATA
					41	male	ATA / ATA
					42	female	ATA / ATA
					43	male	ATA / ATA
					44	male	ATA / ATA
					45	female	ATA / ATA
					46	male	ATA / ATA
					47	male	ATA / ATA

Table S5. Backcross progeny from two families treated with or without 14 mg/L chlorantraniliprole then genotyped for the I4790K RyR mutation

Backcross	Group	I/K	K/K	Total	Parson's χ^2 test
F1 ♂ x LV-R ♀ (n=43)	untreated	7	5	12	χ^2 21.733, df = 1, p-value = 3.134e-06
-	treated	3	28	31	
F1 ♂ x LV-R ♀ (n=41)	untreated	5	5	10	
	treated	2	29	31	
Total (n=84)	untreated	12	10	22	
	treated	5	57	62	

Table S6. Mortality rates estimated from bioassay dose response curves (Fig. 6) for 14 mg/L chlorantraniliprole.

Chemical	Strain	Number	Dose (mg/L)	Estimated mortality	min	max
Coragen (Chlorantraniliprole)	WS	360	14	0.999448883	0.99857312	1.000324646
	LV-R	361	14	0.000832773	-0.000914261	0.002579806
	LV-Rf x WSm	360	14	0.714105876	0.629934272	0.79827748
	WSf x LV-Rm	361	14	0.787004383	0.706350519	0.867658248

Appendix C

Supplementary tables for Chapter 7.

Appendix C:

Supplementary Table 1: InterProScan (IPR) and PFAM domains used to select predicted detoxification genes from the PxLV.1 reference genome.

Gene Family	PFAM domains	IPR domains
P450		
	PF00067	IPR001128
		IPR036396
glutathione-S-transferases		
	PF02798	IPR004045
	PF00043	IPR004046
carboxylesterases		
	PF00135	IPR002018
sulfurotransferases		
	PF03567	IPR005331
	PF00685	IPR000863
	PF00685	IPR000863
UDP-glucuronyltransferase		
	PF00201	IPR002213

Supplementary Table 2: InterProScan (IPR) and PFAM domains along with Gene Ontology (GO) terms used to select predicted chemosensory genes from the PxLV.1 reference genome.

Gene Family	PFAM domains	IPR domains	GO terms
Gustatory receptors			
	PF06151	IPR009318	GO:0050909
	PF08395	IPR013604	GO:0050913
	PF05296	IPR007960	GO:0050916
			GO:0050915
			GO:0050917
Olfactory Receptor/Oderant binding protein			
	PF01395	IPR006170	GO:0007608
	PF02949	IPR004117	
	PF08395	IPR013604	

Supplementary Table 3: Gene names for detoxification genes shown in Figure 9.

g1	FUN_002060	COE	typeB
g2	FUN_003027	COE	typeB
g3	FUN_005302	COE	typeB
g4	FUN_007992	COE	typeB
g5	FUN_008558	COE	typeB
g6	FUN_010575	COE	typeB
g7	FUN_016333	COE	typeB
g8	FUN_018127	COE	typeB
g10	FUN_021242	COE	typeB
g11	FUN_022411	COE	typeB
g12	FUN_003725	COE	typeB
g13	FUN_021236	COE	typeB
g14	FUN_020741	COE	typeB
g15	FUN_005322	COE	typeB
g16	FUN_006043	CYP	Cyp6a
g17	FUN_006289	CYP	Cyp9f
g18	FUN_006297	CYP	Cyp9f
g19	FUN_006298	CYP	Cyp9f
g20	FUN_007972	CYP	Cyp6a
g21	FUN_010662	CYP	Cyp6a
g22	FUN_010798	CYP	Cyp4d
g23	FUN_013613	CYP	Cyp4d
g24	FUN_013615	CYP	Cyp4c
g25	FUN_018835	CYP	Cyp4c
g26	FUN_005580	CYP	Cyp6a
g27	FUN_020551	UDPGT	302-E
g28	FUN_019637	UDPGT	49-B
g29	FUN_020553	UDPGT	49-B
g30	FUN_019678	GST	D1