bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Christopher M. Ward[1] and Simon W. Baxter[1]*

[1]School of Biological Sciences, University of Adelaide, Adelaide, Australia.

*Author for Correspondence: Simon Baxter, School of Biological Sciences, University of Adelaide, Adelaide, Australia 5005. P: +618 83132205 F: +61 8 8313 4362. E: **simon.baxter@adelaide.edu.au**

Data deposition: Sequence data has been deposited in GenBank SRR6023624, SRR6505218-SRR6505233, SRR6505268-SRR6505279 and MG787473.1. Nucleotide alignments will be provided upon request to the corresponding author.

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

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Cryptic species are genetically distinct taxa without obvious variation in morphology and are occasionally discovered using molecular or sequence datasets 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 MYA. 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 datasets including i) the f3 statistic at the level of the population, ii) pairwise comparisons of Nei's absolute genetic divergence (d_{XY}) between populations and iii) 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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (Bickford, et al. 2007; Pfenninger and Schwenk 2007; Stuart, et al. 2006). 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; Voda, et al. 2015).

Reproductive barriers can maintain boundaries between sympatric congeneric animal species (cryptic or non-cryptic) using a range of isolating mechanisms such as olfaction, pheromone cues and mating calls (Andersson, et al. 2007; Jones and Hamilton 1998), host plant preference or mating timing (Hänniger, et al. 2017), and endosymbiont infection (Bordenstein, et al. 2001; Shoemaker, et al. 1999). 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, inter-lineage sorting or homoplasic genetic drift. Patterson, et al. (2012) formalized statistical approaches to estimate admixture based on allele frequencies across multiple populations, namely the f3 and f4 statistics (*D*-statistic), which assess the likelihood of hybridization. The f4 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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hybridization and introgression of genetic variation from a donor species into a recipient can have adaptive advantages. The transfer of advantageous pre-adapted alleles from one species into another removes the reliance of new traits arising though mutation in the recipient. Examples include the transfer of rodenticide resistance between mice (Song, et al. 2011), coat colour alleles among jackrabbits and hares (Jones, et al. 2018), aposematic wing patterns in *Heliconius* butterflies (Mavárez, et al. 2006; Pardo-Diaz, et al. 2012; Wallbank, et al. 2016) and insecticide resistance genes in *Anopholes* 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 1880's (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 (Delgado and Cook 2009; Endersby, et al. 2006).

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* (Perry, et al. 2018; Ward and Baxter 2017), which are known to cause reproductive incompatibility in some cases (Duplouy, et al. 2013; Sasaki and Ishikawa 2000). *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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (*Diplotaxis 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 [*--gatkcigarworkaround, --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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 non-reference) 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 5 reads per base with no greater than 40% of samples missing genotypes at any one site.

Nei's mean intra-population 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.

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 mid-point 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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (15962 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 My^{-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. 2014) 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 × 4N generations ago and the root to 1.5 × 4N generations ago, providing the topology ((I_1, I_2), O). Simulated divergence was determined using mean d_{XY} values from *Plutella* samples (see Figure 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 × 4N generations ago with admixture frequencies (*f*) of $f = 0, 0.05, 0.1, 0.2$ 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 x 4N generations ago) and $f = 0, 0.05, 0.1, 0.2$ 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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Admixture

The F3-statistic

A formal test for admixture was calculated using the three population test, the f-statistic (f3) (Patterson, et al. 2012; Reich, et al. 2009). 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).

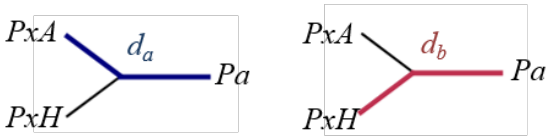
Absolute divergence (d_{XY})

Nei's absolute divergence, d_{XY}, was used to calculate the mean number of nucleotide differences between two populations across non-overlapping 50kb windows with *egglib_sliding_windows.py* (https://github.com/johnomics). Comparisons of d_{XY} were made first with simulated datasets, 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 i) *P. australiana* (O) and Australian *P. xylostella* (I_2) individuals, ii) *P. australiana* (O) and Hawaiian *P. xylostella* (I_1) and iii) 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 non-overlapping 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,

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



bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 approximately 0.5 if the phylogeny was concordant with the species tree. Values much greater than 0.5 indicate genomic windows more similar between *P. australiana* and Hawaiian *P. xylostella*, while values much less than 0.5 indicate genomic windows more similar between *P. australiana* and Australian *P. xylostella* and are candidate admixed regions. Only genomic windows with more than 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 above 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 non-redundant 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 molecular function, InterPro and UniProt annotations were obtained for each BLAST hit.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Table S1**). 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 approximately 70% of these sites were genotyped in *P. australiana* samples compared to ~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). *P. 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±0.003), as has been previously reported with microsatellite data (Endersby, et al. 2006). However, much higher levels of differentiation were observed when compared to Hawaiian *P. xylostella*, supporting the expectation of genetic isolation (average of F_{ST} =0.108±0.01). The average pairwise F_{ST} values were slightly lower between *P. australiana* and Hawaiian *P. xylostella* (F_{ST} =0.501±0.002) than *P. australiana* and Australian *P. xylostella* (F_{ST} =0.532±0.013) (**Table 2**).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure 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.

***P. 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, et al. 2016b). 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, et al. 2016a). 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 Mya (95% confidence interval ± 0.175 My **Figure 2**, **Table S2**). *Prays oleae* and *Leicoptera malifoliella* were used as the outgroups. The topology of the 13 mitochondrial genes used to date the split (**Table S3**) 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 F3-statistic

A formal test for genomic admixture was calculated using the three-population f-statistic (f3), first with simulated datasets 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 f3 statistic making negativity difficult to achieve, even with high levels of

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

introgression (Peter 2016). Consequently, this approach failed to indicate shared ancestry through a negative f3-statistic (**Figure 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 ≥20% (f=0.2), indicating high rates of recent hybridization are required to detect introgression using the f3-statistic (**Figure 3B**). Interestingly, spreading the total proportion of introgression to five equidistant time-points along the branch did not increase the detectability of admixture (**Figure S1**). This suggests the f3 is more dependent on the admixture frequency than the divergence between discordant and concordant regions.

Applying the f3-statistic to empirical data failed to identify negativity in any tip order between Australian *P. xylostella* and *P. australiana* (**Figure 3C**). Results for the f3-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*. A lower f3(*Pxyl* 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 f3 is decreased proportional to the frequency of minor alleles in the target population. As f3 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 datasets. In all cases, population wide comparisons of d_{XY} were performed between; i) two ingroup populations (I_1 and I_2), ii) ingroup 1 and the outgroup (I_1 and O) and iii) ingroup 2 and the outgroup (I_2 and O). Comparions returing 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 outgorup.

Absolute divergence in simulated populations was calculated for each 50 kb window (n=500), again for admixture occuring 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 normalised for density by rescaling such that the maxima of the distribution is 1 (**Figure 4, S2, S3**). Introgression either

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure 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 (**Figure 4C**), as values comparing the *P. australiana* outgroup with either *P. xylsotella* 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 f3-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 above. 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 phylogenic 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, normalising values within the range 0-1. Tree-tip distance proportions are presented as histograms to graph the distribution (*x axis*), and normalised density (*y axis*). Although introgression from I_2 into O (**Figure 5A, S4**) and from O into I_2 (**Figure 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 (**Table S4**), 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 equaly spaced events was effective at detecting admixture using tree-tip distance proportions (**Figure S5**).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This method was then applied to empirical data to identify genomic windows that were discordant with the species tree. Similar to the simulated datasets, genomic alignments were divided into non-overlapping 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-symetric 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 (**Figure 5C, Table S5**). Mean tree-tip distance proportions for each comparison were consistant 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 above 0.5 showing the Hawaiian *P. xylostella* had on average shorter branch lengths to *P. australiana*, consistant with the nuclear genome phylogeny (**Figure 1**). Under a widespread admixture hypothesis windows with distance prorortions 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 (**Table S6**) suggesting no clear evidence to support widespread admixture within *P. xylostella* and *P. australiana* individuals form three sympatric populations.

Despite lack of support for widespread hybridization and genome-wide introgression, we further investigated the tails 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 proprtions between *P. australiana* and Hawaiian *P. xylostella* (**Table S7**). Sliding window d_{XY} was performed on each of these scaffolds with 10 kb windows (sliding by 2kb), 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*. Hostorical admixture between Australian *Plutella* species is one possible explanation for this result, although the region does not contain any protein coding genes (**Figure S6A**). A region on scaffold, KB207380.1, was identfied in the tree-tip distribition tail in 12/32 comparisons however d_{XY} indicated admixture across this region was unlikely (**Figure S6B**).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 admixutre 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 (**Table S8**). 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 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. Non- redundant 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) (**Table S9**). 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 (**Figure S7, S8**).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

The discovery of cryptic species can often be inadvertent and arise from sequencing mitochondrial or nuclear amplicons (Landry and Hebert 2013; Stuart, et al. 2006), 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. *P. 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, et al. 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 shortend towards the reference, underestimating the diveregence between *P. australiana* and *P. xylostella* in the whole genome phylogeny. However, the introduction of this bias is unavoidable as the only reference geneome within the superfamily Yponomeutoidea is currenlty *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 f3-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 (Juric, et al. 2016; Perry, et al. 2018; Saw, et al. 2006), 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. 2016).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 million years. 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 two million years ago 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 deviatie from an expected consensus topology or species tree and can be used to identify genomic regions that may be of biological interst. 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 detecteting 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 behavious may provide further insight into these divergent loci.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 can't 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.

Acknowledgements

This work was supported by the Australian Research Council (grant numbers DP120100047, FT140101303) and Grains Research and Development Corporation (UOA1711-004RSX). CMW 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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

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: 964-970.

FastQC: a quality control tool for high throughput sequence data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baxter SW, et al. 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: 148-155.

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

Bordenstein SR, O'Hara FP, Werren JH 2001. Wolbachia-induced incompatibility precedes other hybrid incompatibilities in *Nasonia*. Nature 409: 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: 136-142.

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

Dai LS, et al. 2016b. The complete mitochondrial genome of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). Mitochondrial DNA A DNA Mapp Seq Anal 27: 1512-1513.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156-2158. doi: 10.1093/bioinformatics/btr330

Darriba D, Taboada GL, Doallo R, Posada D 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9: 772-772.

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.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

Dierckxsens N, Mardulyn P, Smits G 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45: e18-e18.

Drummond AJ, Suchard MA, Xie D, Rambaut A 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969-1973.

Duplouy 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: 1-13.

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: 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: 68.

Hasegawa M, Kishino H, Yano T-a 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22: 160-174. doi: 10.1007/bf02101694

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: 8313-8318.

Jones MR, et al. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360: 1355-1358. doi: 10.1126/science.aar5273

Jones T, Hamilton J 1998. A role for pheromones in mate choice in a lekking sandfly. Anim. Behav. 56: 891-898.

Juric I, Salzburger W, Balmer O 2016. 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. Bulletin of Entomological Research 107: 155-164. doi: 10.1017/S0007485316000766

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: 3059-3066.

Kozlov AM, Aberer AJ, Stamatakis A 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31: 2577-2579.

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

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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: 19854-19859.

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

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Lunter G, Goodson M 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 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: 28-28.

Martin SH, et al. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23: 1817-1828.

Martin SH, Davey JW, Jiggins CD 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol Biol Evol 32: 244-257.

Mavárez J, et al. 2006. Speciation by hybridization in *Heliconius* butterflies. Nature 441: 868-871.

Moran NA, Jarvik T 2010. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. Science 328: 624-627.

Nei M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press.

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: 815-820.

Papadopoulou A, Anastasiou I, Vogler AP 2010. Revisiting the Insect Mitochondrial Molecular Clock: The Mid-Aegean Trench Calibration. Mol Biol Evol 27: 1659-1672.

Paradis E, Claude J, Strimmer K 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290.

Pardo-Diaz C, et al. 2012. Adaptive Introgression across Species Boundaries in *Heliconius* Butterflies. PLoS Genetics 8. doi: 10.1371/journal.pgen.1002752

Patterson N, et al. 2012. Ancient Admixture in Human History. Genetics 192: 1065-1093.

Patterson N, Price AL, Reich D 2006. Population Structure and Eigenanalysis. PLoS Genetics 2: e190. doi: 10.1371/journal.pgen.0020190

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Perry KD, et al. 2018. Cryptic *Plutella* species show deep divergence despite the capacity to hybridize. BMC Evolutionary Biology 18: 77. doi: 10.1186/s12862-018-1183-4

Peter BM 2016. Admixture, Population Structure and *F*-Statistics. Genetics 202: 1485-1501.

Pfenninger 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: 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: 235-238.

Rambaut A, Suchard M, Xie D, Drummond A. 2014. Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer/.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L 2009. Reconstructing Indian Population History. Nature 461: 489-494.

Sasaki T, Ishikawa H 2000. Transinfection of Wolbachia in the Mediterranean Flour Moth, *Ephestia kuehniella*, by Embryonic Microinjection. Heredity 85: 130-135.

Saw J, N.M. E, McKechnie SW 2006. Low mtDNA diversity among widespread Australian diamondback moth *Plutella xylostella* (L.) suggests isolation and a founder effect. Insect Science 13: 365-373. doi: doi:10.1111/j.1744-7917.2006.00105.x

Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 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: 2805-2811.

Shoemaker DD, Katju V, Jaenike J 1999. Wolbachia and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria*. Evolution 53: 1157-1164.

Song Y, et al. 2011. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. Curr. Biol. 21: 1296-1301.

Stamatakis A 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 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: 470-474. doi: 10.1098/rsbl.2006.0505

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tang W, et al. 2014. DBM-DB: the diamondback moth genome database. Database 2014: bat087. doi: 10.1093/database/bat087

Tyron H. 1889. In. Report on Insect and Fungus Pest No.1. Brisbane, Australia: Department of Agriculture, Queensland, James C. Beal. Government Printer.

Voda R, Dapporto L, Dinca V, Vila R 2015. Cryptic matters: overlooked species generate most butterfly beta-diversity. Ecography 38: 405-409.

Walker BJ, et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9: e112963. doi: 10.1371/journal.pone.0112963

Wallbank RWR, et al. 2016. Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling. PLOS Biology 14: e1002353. doi: 10.1371/journal.pbio.1002353

Ward CM, Baxter SW 2017. Draft Genome Assembly of a Wolbachia Endosymbiont of *Plutella australiana*. Genome Announcements 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. In. Ggplot2: Elegant Graphics for Data Analysis: Springer, New York, NY. p. 1-212.

You M, et al. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet 45: 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: 1115-1129.

Zhang W, Dasmahapatra KK, Mallet J, Moreira GRP, Kronforst MR 2016. Genome-wide introgression among distantly related *Heliconius* butterfly species. Genome Biology 17: 25. doi: 10.1186/s13059-016-0889-0

Fig. 1: Maximum likelihood phylogeny of *P. 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*).

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Fig. 3. The three population f-statistic (f3). **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 grey boxes for f3(I_1,O;I_2), which did not reach negative values. For comparison, the f3-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 produced a significant f3 statistic at a mixing frequency >0.2, as indicated by the values below zero f3(I_1,O;I_2). **C** The f3-statistic was then applied to empirical data, testing for admixture in three possible scenarios between *P. xylostella* from Hawaii (*P.x. H,* I_1), *P. xylostella* from Australia (*P.x. A,* I_2) and *P. australiana* (*P.a.,* O). This suggests that, as no f3 values were below zero, if admixture was occurring it could not be detected using this method.

Fig. 4. Absolute genetic divergence (d_{XY}) between populations. Each plot is a histogram summarising 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 **Figure S2** and **S3** 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 abseence 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 *P. 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 2014, SA 2014). Each panel summarizes 7276 (±293) branch distance ratio calculations. The branch leading to ingoup 1 was standardized using the same two Hawaiian *P. xylostella* individuals for each of these comparisons.

bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Summary 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	*P. xylostella*	2	17	91.5	0.0151	(0.0042)
		P. australiana	4	12.5	71.8	0.0170	(0.0052)
Australian Capital Territory (ACT)	2015	*P. xylostella*	2	18	92.2	0.0150	(0.0045)
		P. australiana	4	13.5	72.3	0.0168	(0.0048)
South Australia (SA)	2014	*P. xylostella*	4	23.5	91.8	0.0157	(0.0040)
		P. australiana	4	17.5	65.9	0.0174	(0.0051)
Hawaii	2013	*P. xylostella*	8	13.125	91.7	0.0200	(0.0044)

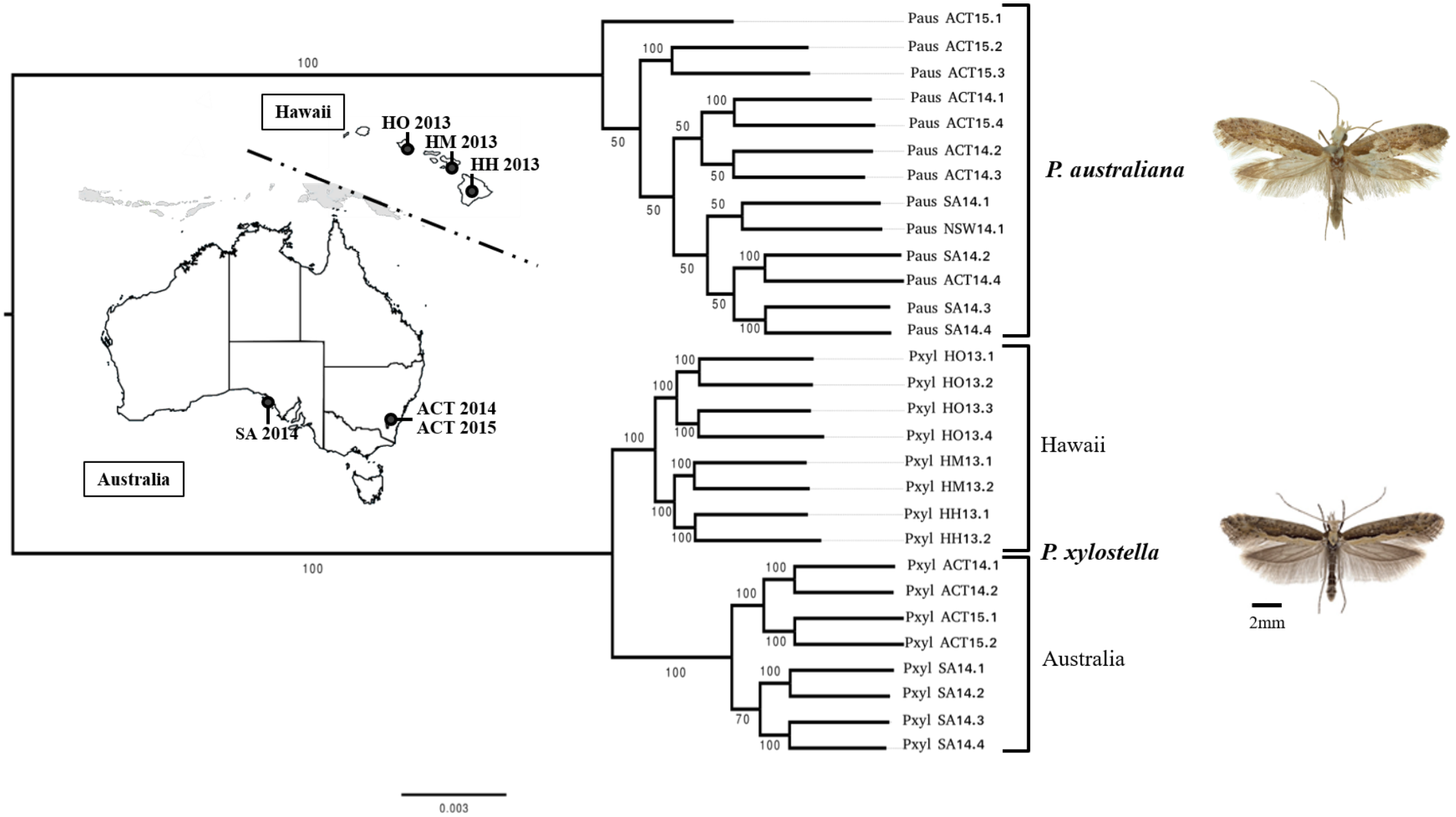
Table 2. Matrix of the minimum distance between collection sites (km's, above diagonal) and pairwise F_{ST} values of each *Plutella* population (below diagonal).

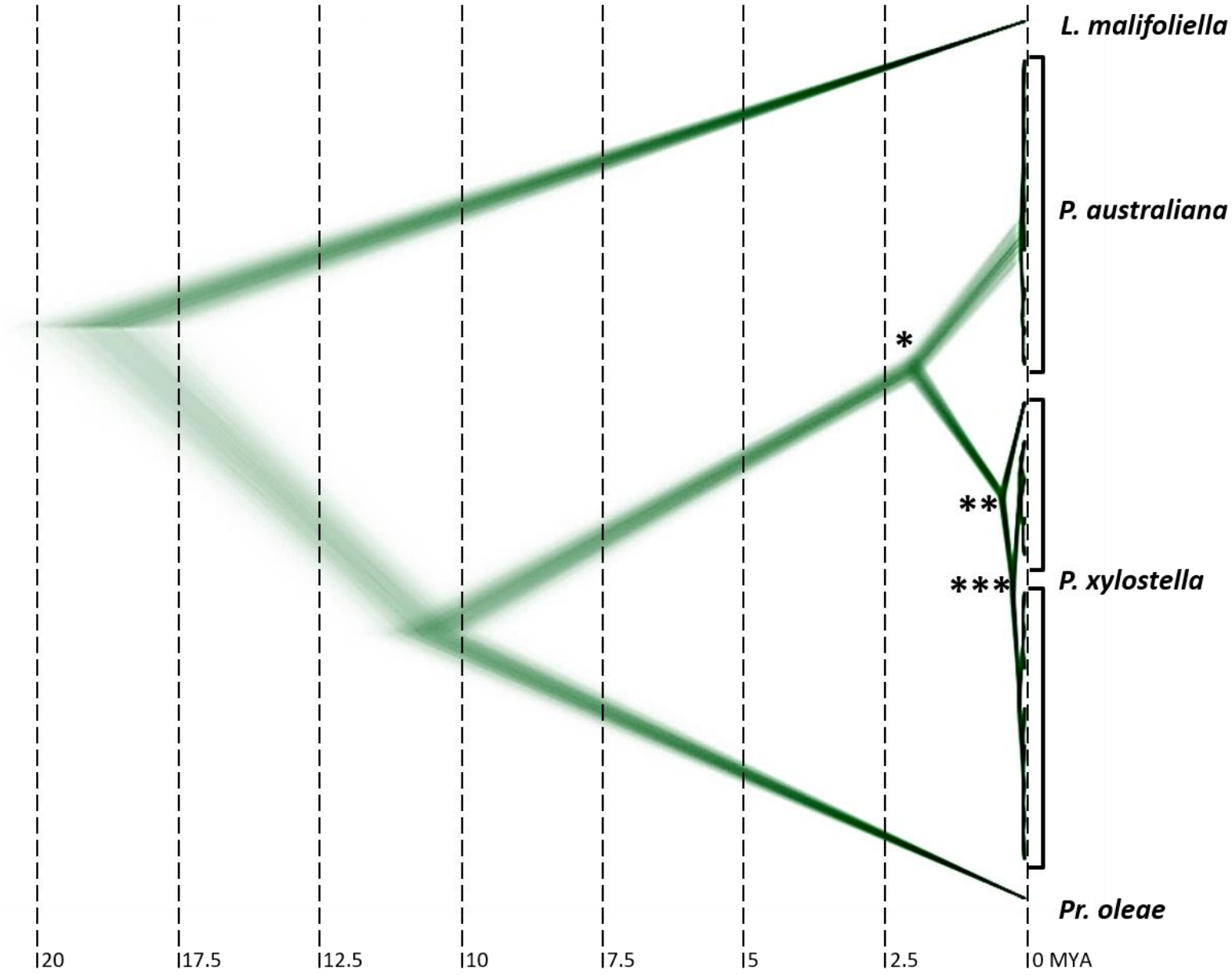
Species and population	*Pxyl*.ACT.2014	*Pxyl*.ACT.2015	*Pxyl*.SA.2014	*Pxyl*.Hawaii.2013	*Paus*.ACT.2014	*Paus*.ACT.2015	*Paus*.SA.2014
Pxyl.ACT.2014		8.3	1372	8470	0	8.3	1372
Pxyl.ACT.2015	0.000		1371	8466	8.3	0	1371
Pxyl.SA.2014	0.007	0.006		9480	1372	1371	0
Pxyl.Hawaii.2013	0.102	0.103	0.119		8470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

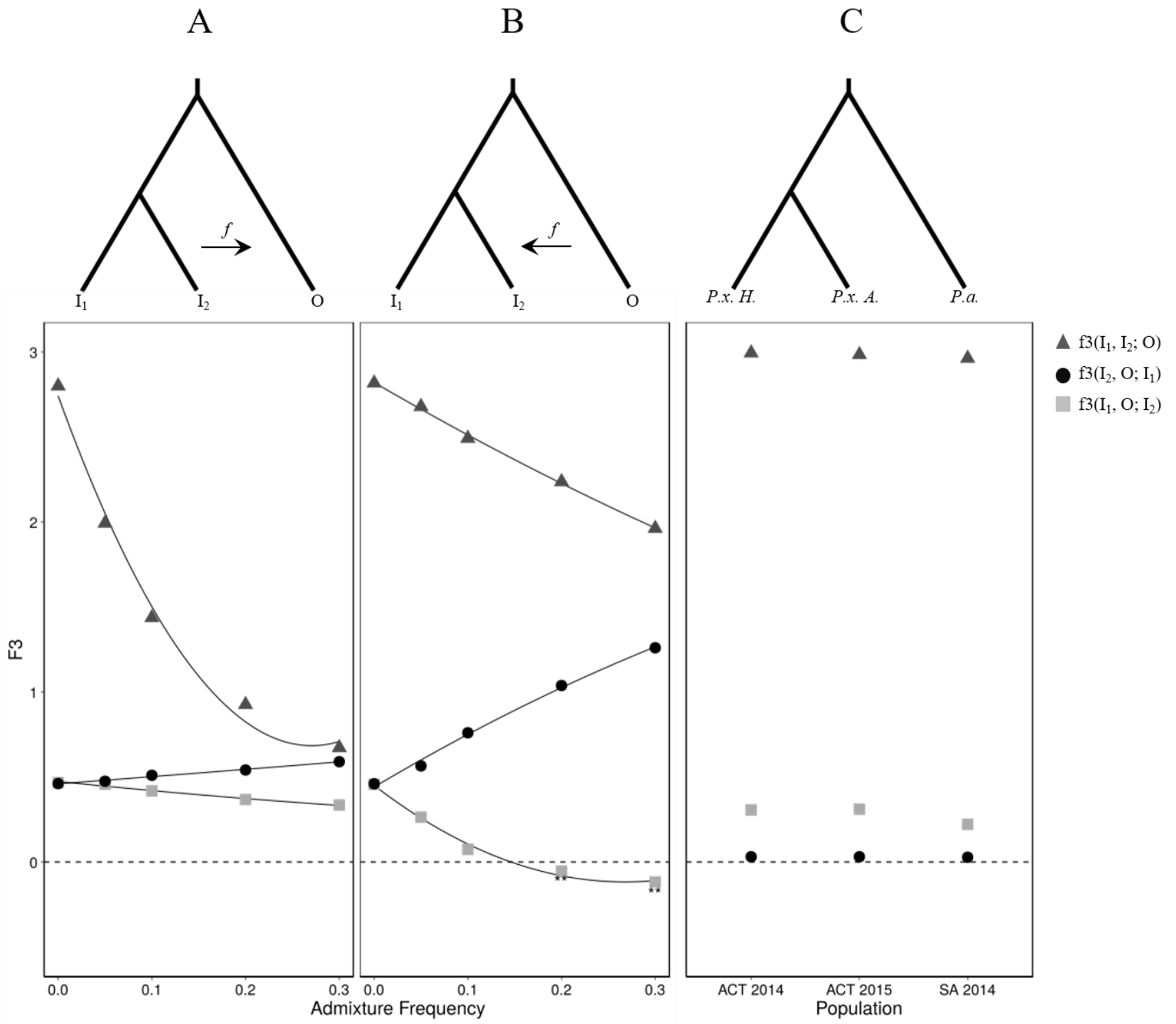
bioRxiv preprint doi: https://doi.org/10.1101/405498; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

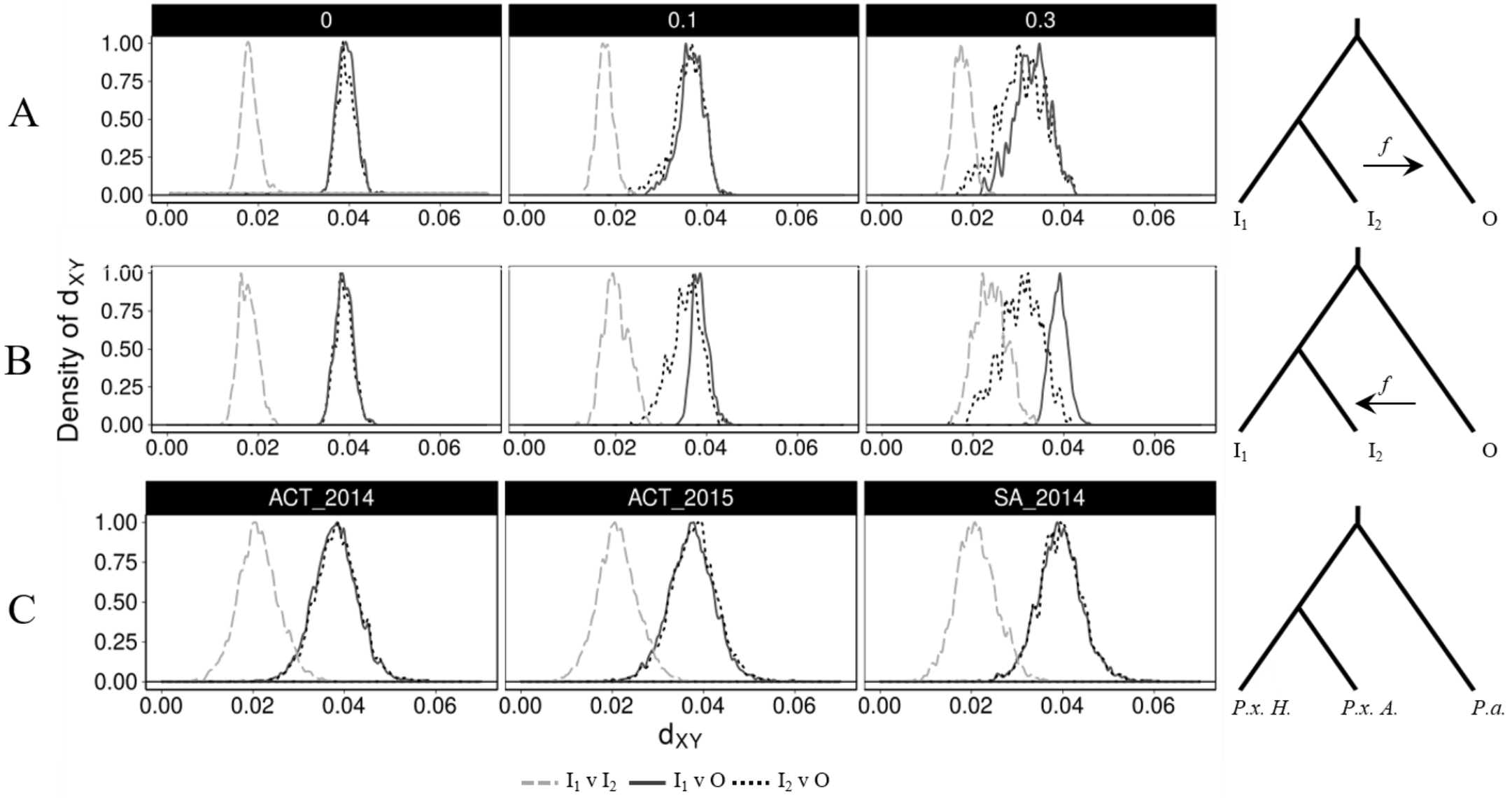
Table 3. Confidence intervals (95%) for d_{XY} comparisons of populations

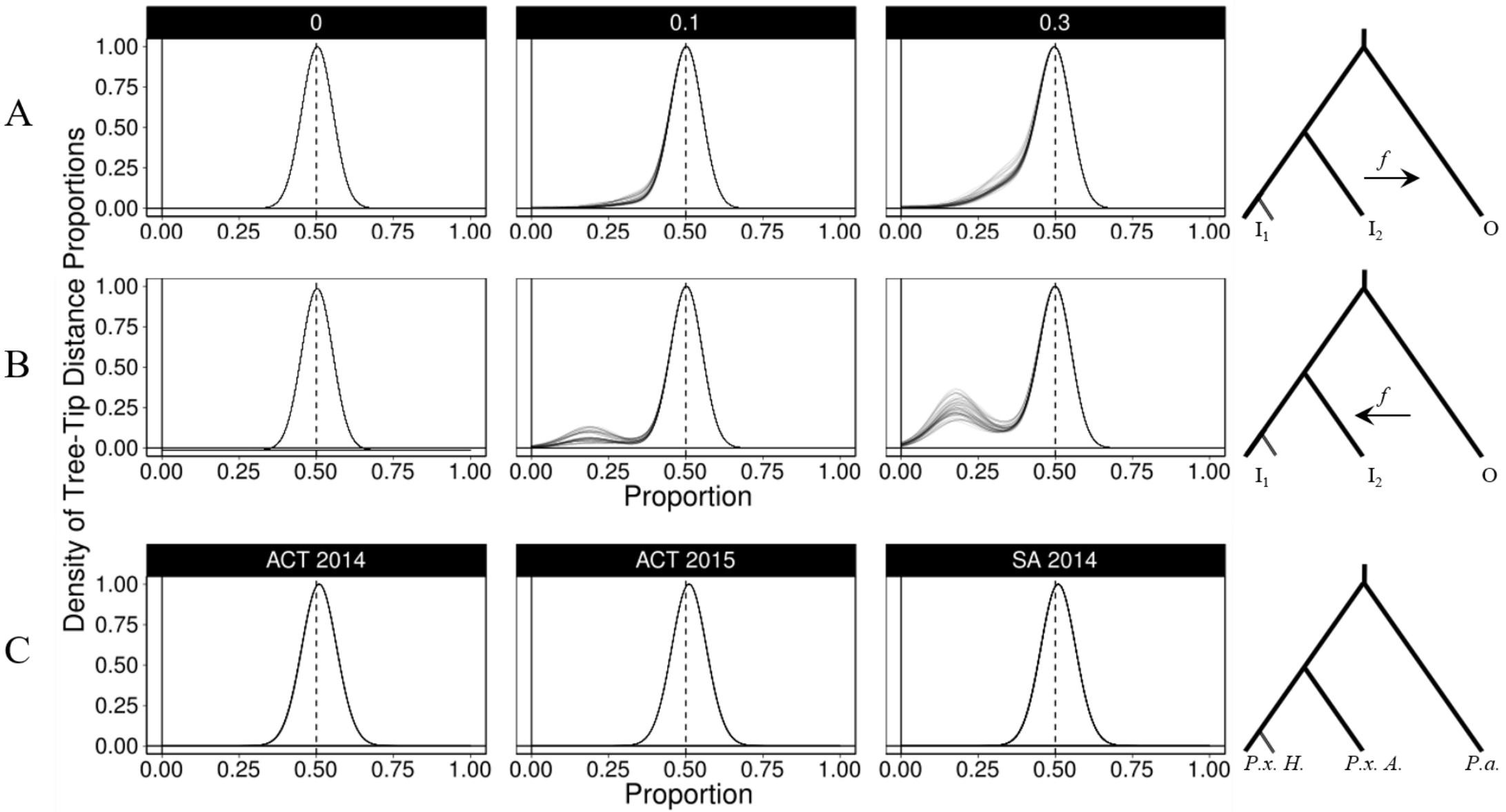
	Pxyl. Hawaii *vs* *Pxyl.* Australia	*Pxyl.* Hawaii *vs* *Paus.* Australia	*Pxyl.* Australia *vs* *Paus.* Australia
ACT 2014	0.02101 - 0.02122	0.03788 - 0.03812	0.03828 - 0.03852
ACT 2015	0.02102 - 0.02123	0.03721 - 0.03744	0.03771 - 0.03795
SA 2014	0.02087 - 0.02111	0.03898 - 0.03926	0.03926 - 0.03955


Hawaii
HO 2013
HM 2013
HH 2013
ACT 2014
ACT 2015
SA 2014
Australia
100
100
50
50
100
50
50
50
50
100
50
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
70
100
Paus ACT15.1
Paus ACT15.2
Paus ACT15.3
Paus ACT14.1
Paus ACT15.4
Paus ACT14.2
Paus ACT14.3
Paus SA14.1
Paus NSW14.1
Paus SA14.2
Paus ACT14.4
Paus SA14.3
Paus SA14.4
Pxyl HO13.1
Pxyl HO13.2
Pxyl HO13.3
Pxyl HO13.4
Pxyl HM13.1
Pxyl HM13.2
Pxyl HH13.1
Pxyl HH13.2
Pxyl ACT14.1
Pxyl ACT14.2
Pxyl ACT15.1
Pxyl ACT15.2
Pxyl SA14.1
Pxyl SA14.2
Pxyl SA14.3
Pxyl SA14.4
P. australiana
Hawaii
P. xylostella
Australia
2mm
0.003


L. malifoliella
P. australiana
*
**
***
P. xylostella
Pr. oleae
20
17.5
12.5
10
7.5
5
2.5
0 MYA


A
B
C
f
f
I1
I2
O
I1
I2
O
P.x. H.
P.x. A.
P.a.
f3(I1, I2; O)
f3(I2, O; I1)
f3(I1, O; I2)
F3
3
2
1
0
0.0
0.1
0.2
0.3
0.0
0.1
0.2
0.3
**
**
Admixture Frequency
ACT 2014
ACT 2015
SA 2014
Population


A
B
C
0
0.1
0.3
ACT_2014
ACT_2015
SA_2014
Density of d_{XY}
d_{XY}
1.00
0.75
0.50
0.25
0.00
0.00
0.02
0.04
0.06
I_1 v I_2
I_1 v O
I_2 v O
f
I_1
I_2
O
P.x. H.
P.x. A.
P.a.


A
B
C
0
0.1
0.3
ACT 2014
ACT 2015
SA 2014
Density of Tree-Tip Distance Proportions
Proportion
0.00
0.25
0.50
0.75
1.00
f
I1
I2
O
P.x. H.
P.x. A.
P.a.