Local adaptation of a native herbivore to a lethal invasive plant.

Nitin Ravikanthachari\textsuperscript{1}, Rachel Steward\textsuperscript{2}, and Carol Boggs\textsuperscript{1}

\textsuperscript{1}University of South Carolina  
\textsuperscript{2}Stockholm University

June 9, 2023

Abstract

Understanding the evolutionary processes that influence fitness is critical to predicting species’ responses to selection. Interactions among evolutionary processes including gene flow, drift and the strength of selection can lead to either local adaptation or maladaptation especially in heterogeneous landscapes. Populations experiencing novel environments or resources are ideal for understanding the mechanisms underlying adaptation or maladaptation, specifically in locally co-evolved interactions. We used the interaction between a native herbivore that oviposits on a patchily distributed introduced plant that in turn causes significant mortality to the larvae to test for signatures of local adaptation in areas where the two co-occurred. We used whole genome sequencing to explore population structure, patterns of gene flow and signatures of local adaptation. We found signatures of local adaptation in response to the introduced plant in the absence of strong population structure with no genetic differentiation and low genetic variation. Additionally, we found localized allele frequency differences within a single population between habitats with and without the lethal plant, highlighting the effects of strong selection. Our work highlights the potential for adaptation to occur in a fine-grained landscape in the presence of gene flow and low genetic variation.
Local adaptation of a native herbivore to a lethal invasive plant.

**Running title:** Local adaptation to a novel hostplant.

**Author list**

Nitin Ravikanthachari$^{1,2,5,*}$, Rachel A Steward$^{1,2,3}$, Carol L Boggs$^{1,2,4}$

**Author affiliations**

1 Dept. of Biological Sciences, University of South Carolina, Columbia, South Carolina 29208, USA
2 Rocky Mountain Biological Laboratory, Crested Butte, Colorado 81224, USA
3 Department of Biology, Lund University, Lund, Sweden
4 School of Earth, Ocean & Environment, University of South Carolina, Columbia, South Carolina 29208, USA
5 current address: Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA

*Corresponding author: Nitin Ravikanthachari

ravikanth.nitin@gmail.com
Abstract

Understanding the evolutionary processes that influence fitness is critical to predicting species’ responses to selection. Interactions among evolutionary processes including gene flow, drift and the strength of selection can lead to either local adaptation or maladaptation especially in heterogeneous landscapes. Populations experiencing novel environments or resources are ideal for understanding the mechanisms underlying adaptation or maladaptation, specifically in locally co-evolved interactions. We used the interaction between a native herbivore that oviposits on a patchily distributed introduced plant that in turn causes significant mortality to the larvae to test for signatures of local adaptation in areas where the two co-occurred. We used whole genome sequencing to explore population structure, patterns of gene flow and signatures of local adaptation. We found signatures of local adaptation in response to the introduced plant in the absence of strong population structure with no genetic differentiation and low genetic variation. Additionally, we found localized allele frequency differences within a single population between habitats with and without the lethal plant, highlighting the effects of strong selection. Our work highlights the potential for adaptation to occur in a fine-grained landscape in the presence of gene flow and low genetic variation.

Keywords
evolutionary traps, novel resources, maladaptation, *Pieris*, Brassicaceae, gene flow - selection
balance
Introduction

Strong spatially heterogenous selection can lead to locally adapted populations. The ability of populations to adapt to local conditions depends on substantial genetic variation, large effective population sizes and the changes in their habitat (Franch-Gras et al., 2018; Hedrick, 2011; Lande, 1976; Lenormand, 2002; Perrier et al., 2020; Simons, 2011). Additionally, local adaptation is maintained through the balance of strong selection and limited gene flow between divergently adapted populations (Blanquart et al., 2013; Hereford, 2009; Savolainen et al., 2013; Whitlock, 2015). Across a heterogeneous landscape, the interplay between gene flow and selection can lead to a geographic mosaic with varying levels of adapted and maladapted sub-populations (Gomulkiewicz et al., 2000; Thompson, 1999a, 2005).

Novel environments can derail adaptative evolution by decreasing the fitness of formerly locally adapted individuals (Brady, Bolnick, Angert, et al., 2019; Reed et al., 2003; Robertson & Hutto, 2006; Schlaepfer et al., 2002, 2005). The outcome of interactions between native populations and novel environments depends strongly on population connectivity among divergently adapted populations as well as the population’s potential to adapt to novel conditions (Bolnick & Nosil, 2007; Farkas et al., 2015, 2016; Nosil et al., 2019). This often results in persistent mismatches with or maladaptation to the new resources or conditions (Brady, Bolnick, Angert, et al., 2019; Chew, 1977b; Cotto & Ronce, 2014; Singer & Parmesan, 2018; Steward et al., 2022; Steward & Boggs, 2020). Such cases provide unique opportunities for understanding the process of local adaptation because the character, scale and timeline of the environmental change are often well known (Chew, 1977b; Singer & Parmesan, 2018, 2019; Steward et al., 2022; Steward & Boggs, 2020). Identifying and understanding these interactions are critical to
understanding species responses to novel environments, especially in the face of anthropogenic disturbance and climate change.

Herbivorous insects have a narrow range of hostplants that support development, and those insects form tightly co-evolved interactions with their hosts (Ehrlich & Raven, 1964; Hardy & Otto, 2014; Jaenike, 1990; Joshi & Thompson, 1995; Thompson, 1999b). Non-native or invasive species can break down long-standing interactions and lead to maladaptation in either or both of the interacting partners, either by outcompeting and replacing native species or through eco-evolutionary traps (LaForgia et al., 2020; Richard et al., 2019; Saul & Jeschke, 2015; Schlaepfer et al., 2005; Schweiger et al., 2010). Eco-evolutionary traps arise when previously adaptive traits become maladaptive in the face of novel resources or environments, such that the organisms prefer the low fitness-value novel resource instead of the high fitness-value native resources (Robertson et al., 2013; Robertson & Chalfoun, 2016; Robertson & Hutto, 2006; Schlaepfer et al., 2002).

Eco-evolutionary traps resulting from maladaptive oviposition choice have been documented in many insects (Horváth et al., 2007, 2010; Singer & Parmesan, 2018). Lepidoptera, specifically butterflies, are susceptible to laying eggs on low quality novel hosts plants (Nakajima & Boggs, 2015; Singer & Parmesan, 2010, 2019; Steward et al., 2022; Steward & Boggs, 2020; Yoon & Read, 2016). Hence, butterflies serve as an ideal system to test the impacts of novel resources on local adaptation in species interactions. We here use a landscape genetics approach to identify genomic signatures of local adaptation in the specialist herbivore Pieris macdunnoughii Remington, 1954 (Pieridae; formerly P. napi macdunnoughii) (Chew & Watt, 2006) in the presence of the toxic invasive plant Thlaspi arvense (Brassicaceae). Pieris
macdunnoughii females recognize T. arvense as a potential host plant due to chemical cue similarity with the native brassicaceous hosts and lay eggs on it (Chew, 1977b, 1977a, 1980; Nakajima, 2014; Steward et al., 2019, 2022). However, larvae from eggs laid on T. arvense have significant mortality due to the toxicity of the plant (Chew, 1977b; Steward et al., 2019). We sampled P. macdunnoughii in the East River Valley, Gunnison County, Colorado, where the butterfly oviposits on both native host plants and T. arvense.

Using genome-wide data from whole genome resequencing of individuals from areas where both the plants occur and from areas where T. arvense is absent, we identified candidate single nucleotide polymorphisms (SNPs) that are under selection in relation to the presence of T. arvense in the habitat. We specifically quantify: i) population structure of P. macdunnoughii in the East River Valley, ii) gene flow patterns of P. macdunnoughii between areas with and without T. arvense and iii) underlying signatures of selection in P. macdunnoughii in response to the presence of P. macdunnoughii in the landscape. Our work builds on decades of existing research on this system to address the genetic basis underlying maladaptive behavior in the face of novel resources.

Materials and methods

Study organisms

Pieris macdunnoughii is a montane butterfly distributed across the southern Rocky Mountains in North America. Pieris macdunnoughii is a specialist herbivore, and females oviposit on and the larvae feed on several native Brassicaceae (mustard) plants including Boechera spp., Cardamine cordifolia, Draba aurea and Descurainia incana (Chew, 1977b, 1977a, 1980). Pierine larvae have evolved resistance to the toxicity of glucosinolates (secondary metabolites) found in the Brassicaceae by modifying the glucosinolate hydrolysis pathway to
form thiocyanates and epithionitrilies that are less toxic, instead of forming isothiocyanates (Edger et al., 2015; Wheat et al., 2007). Although pierine butterflies are resistant to glucosinolate toxicity, they are susceptible to species whose glucosinolate concentration or composition differs from the ones with which they have co-evolved locally (Keeler & Chew, 2008).

_Thlaspi arvense_ (L.) (Brassicaceae) is native to Eurasia and was introduced to the Gunnison basin, Colorado in 1880s (Chew, 1977b). It is an early successional plant and is found in heavily disturbed areas. It has colonized elevations up to 2900m (Best & McIntyre, 1975; Warwick et al., 2011). *Pieris macdunnoughii* females recognize and lay eggs on *T. arvense* as a potential host plant in areas where they both co-occur due to similarity of glucosinolates, and other cues shared among the native and nonnative host plants. However, none of the larvae reared on *T. arvense* in the lab survive, thus leading to an evolutionary trap (Nakajima, 2014; Steward & Boggs, 2020). Previous research has indicated that the chemical profile of *T. arvense* is comprised mostly of the aliphatic glucosinolate sinigrin, whereas native mustards (e.g., *C. cordifolia*) that contain both aliphatic and aromatic glucosinolates (Rodman & Chew, 1980). Additionally, recent research has indicated that sinigrin acts as a pre-ingestive deterrent and larvae exhibit delayed feeding on *T. arvense* compared to *C. cordifolia* (Steward et al., 2019).

**Population sampling and study area**

The study area is situated in the East River valley, Gunnison County, Colorado spanning an altitude from ca. 2800 to 3400m a.s.l. (Figure 1). We sampled 100 female adult butterflies during June - August 2016 from 5 locations and 235 adult butterflies of both sexes during June - August 2019 from 7 locations. Our final sample size consisted of 335 individuals from 11 locations. The distance between sampling sites ranged from 1km to 28km. All butterflies were
caught using an aerial net. We collected whole body specimens of females and stored them in 98% ethanol in 2016. In 2019, 2mm x 2mm square of hindwings and a pair of midlegs were clipped and stored in 98% ethanol, after which the butterflies were marked and released. We also surveyed the presence/absence of T. arvense in all our sampling locations. We recorded T. arvense as present in 2 sites in 2016 (of 5 sites sampled) and 2019 (of 7 sites sampled). Our final sampling list consisted of 3 sites where T. arvense was present (of 11 sites sampled).

**DNA extraction and WGS library preparation**

We extracted genomic DNA from thorax (2016 samples) and from wing clips and a pair of midlegs (2019 samples) using the DNeasy Blood & Tissue Kit (Qiagen, USA). We followed the manufacturer's protocol, with the following modifications: We increased the Proteinase K incubation step to 16 hours, used ice-cold ethanol to precipitate DNA, and heated the elution buffer to 57 ºC before DNA elution. We quantified the DNA concentration using a NanoDrop™ 2000/2000c. We used an in-house Tn5 tagmentation protocol following Andolfatto et al. (2011) for our library preparation. 2-10ng/ul of genomic DNA was tagmented with 1:4 diluted Tn5 transposase and 40uM pre-annealed oligonucleotides. The tagmentation was carried out in a final volume of 20uL containing 4uL of 5x TAPS buffer. Samples were incubated at 55 ºC for 7 minutes followed by rapidly lowering the holding temperature to 10 ºC. The Tn5 transposase was deactivated using 2.5uL 0.2% SDS at 55 ºC for 7 minutes with a heated lid. PCR-based barcoding and enrichment was carried out in a reaction mixture of 20uL consisting of 2 uL of the tagmentation reaction product, 10uL of 2x OneTaq Hot Start DNA polymerase, 4uL of nuclease free water and 2uL (10uM) each of i7 and i5 Illumina primer. The thermocycling conditions included initial denaturation at 72°C for 3 minutes followed by denaturation at 94°C, followed by
10 cycles with 10s at 94°C, 15s at 62°C, 30s at 68°C and final extension for 5min at 68°C. We constructed 5 replicate libraries for each sample, then pooled all the replicates and samples (1920 libraries). Size selection of 300-500 bp was carried out using 0.8x volume and 0.6x volume AmpureXP beads. All the libraries were sequenced at MedGenome on an Illumina Novaseq S4 platform, generating 150bp paired end reads.

**SNP (Single nucleotide polymorphisms) calling**

Demultiplexed raw Illumina reads and adapter sequences were trimmed using fastp (Chen et al., 2018) and mapped against the reference genome (Steward et al., 2021) using bwa mem with default parameters (Li, 2013). The resulting SAM files were converted to BAM format, sorted and indexed using samtools (Li et al., 2009). Duplicates were marked and identifier groups were added using PICARD TOOLS with default parameters (“Picard Toolkit,” 2019). The genome dataset had a coverage of $6.84 \pm 2.42x$ (mean ± s.d) and was quantified using genozip (Lan et al., 2021). Single nucleotide polymorphisms (SNPs) were called across all samples using GATK HaplotypeCaller to generate individual intermediate gVCF files that were then imported using GATK GenomicsDBImport and were finally genotyped using GATK GenotypeGVCFs. We used GATK VariantFilteration to further hard filter SNPs and Indels separately using the best recommended workflow practices (Van der Auwera et al., 2013), and finally used GATK SelectVariants to include only those variants that met the criteria: $QD > 2$, base quality > 30, $SOR < 3.0$, $FS < 60$, $MQ > 40$, $MQ$ Rank Sum > 12.5, Read Pos Rank Sum > 8.0. We then used VCFtools to retain only high quality bi-allelic variants using the following parameters: minimum allele count = 4, max missing = 0.20, min q= 30, min-mean DP = 6, max mean DP = 60, minDP = 6, max DP = 60. We further used PLINK (Purcell et al., 2007) to prune those loci that were at
LD with the following parameters: $r^2 > 0.2$ in a window of 50bp. Our final sample size consisted of 335 individuals and 768,339 SNPs.

**Population structure**

We used a subset of 735,000 putatively neutral SNPs that were obtained after removing outlier loci (see below) to discern the population structure. Principal Component Analysis (PCA) clustering was carried out using PLINK (Purcell et al., 2007). We complemented our PCA analysis with archetypal analysis following Gimbernat-Mayol et al. (2022), to test biases in the PCA analysis due to irregular sample sizes and to identify latent factors. We removed multiallelic SNPs prior to running archetypal analysis and performed the analysis with $k$ ranging from 2 to 4. We picked the $k$ whose PC1 and PC2 axes together accounted for the most variance. We additionally used ADMIXTURE to corroborate the results from PC and archetypal analysis and to assess population structure and ancestry (Alexander et al., 2009). We performed ADMIXTURE analysis for ancestral clusters $K$ ranging from 1 to 6 and selected the $K$ value with the lowest cross-entropy as the best estimate of population admixture.

**Nucleotide summary statistics**

We used geoVar (Biddanda et al., 2020) to assess if alleles were shared among habitats/sites or if they were localized to each habitat/site. We estimated allele frequency distribution (site frequency spectrum) in a) all 11 sites in the East River Valley and b) in areas with and without *T. arvense*. We converted our VCF into a frequency table using geoVar ($n = 768339$ SNPs), then calculated the cumulative fraction of variants that contributed to the allele frequency pool for each site/habitat. We used the allele frequency distribution in geoVar in place of widely used SFS (Site Frequency Spectrum) methods (Gutenkunst et al., 2009), since geoVar
permits simultaneous comparisons of SFS for more than 2 populations. Furthermore, geoVar allows us to classify minor alleles into common (>5% frequency among all samples), low (1 - 5% frequency), rare (<1% frequency) and unobserved (allele not present).

Using VCFtools (Danecek et al., 2011) on our SNP dataset (consisting only of variant sites, n = 768339 sites), we calculated expected and observed heterozygosity of butterflies for sites where T. arvense is present and sites where T. arvense is absent. We used pixy (Korunes & Samuk, 2021) to estimate genome-wide nucleotide (π) diversity. For the input for pixy, we specifically used both the invariant and variant sites in our input VCF file (n = 69339609 sites) as recommended by the authors to prevent bias in our estimates. We used a 10kb sliding window with a 50bp step to calculate π estimates. We used Bartlett’s test to test for differences in heterozygosity and nucleotide diversity between the habitats. We used VCFtools on our SNP dataset (consisting only of variant sites, n = 768339 sites) to calculate pairwise Fst for all combinations of sites in the East River Valley. We used a sliding window approach with an interval window of 1kb with a 50bp step for estimating pairwise Fst.

Migration surface/gene flow

We used divMigrate from the diveRsity package (Keenan et al., 2013) in R (R Core Team, 2020) to estimate directional relative migration rates among sampling sites and PGDSpider (Lischer & Excoffier, 2012) to process the input for divMigrate. The relative migration network was scaled to the largest estimated magnitude based on Nm as a measure of genetic distance. We assessed the significance of the migration network by running 1000 bootstrap iterations.
We used EEMS (Petkova et al., 2015) to examine spatial variation in migration among populations (demes) and genetic diversity within populations. EEMS estimates genetic differentiation using an isolation-by-distance model of geo-referenced samples to visualize patterns of potential barriers and corridors of gene flow. We used PLINK to convert the VCF file to BED files and calculated a genetic dissimilarity matrix using bed2diff in EEMS. We ran EEMS using 400 and 800 demes and a MCMC run with 1.5 million burn-in iterations followed by 15 million sampling iterations. We ran multiple iterations and adjusted the proposal variance rates after each run until runs converged (Figure S1). The parameters used for the final run are provided in the supplementary file (Table S1). We then combined the final output of all the demes to produce a composite migration and diversity landscape. Migration and diversity rates were illustrated on a log10 scale relative to the overall migration and diversity across the entire landscape, such that a rate of one indicated tenfold higher migration and diversity rates relative to the average.

**Genotype-environment association (GEA) analysis**

Genome-wide scans for outlier loci were carried out in BayPass (Gautier, 2015). We converted our VCF file into an allele count matrix (n=768339 SNPs) using an in-house python script for BayPass input. We first used the standard covariate model with the Importance Sampling (IS) approximation (-covis) in BayPass to obtain the following parameters: a) the population scaled covariance matrix and b) the XTX (SNP-specific Fst that corrects for observed population covariance) score of overall differentiation among sites. We then used the auxiliary covariate model using the MCMC approximation along with the Bayesian auxiliary variable to identify loci that were associated with the presence/absence of *T. arvense*. We additionally
simulated 100,000 loci using the BayPass sim to calibrate the top 1% significance threshold for XTX. For environmental association, we used the Bayes Factor (BF) metric in deciban units (dB) as a measure of support for association with *T. arvense*. We used a cut-off of db>20 (decisive evidence) in favor of association based on Jefferey's rule (Jeffreys, 1961). Loci associated with *T. arvense* as well as those under adaptive differentiation were used as input in SNPEff (Cingolani et al., 2012) and BlastX (Altschul et al., 1990) to identify the underlying genes and their effects on protein structure.

**Results**

**Population structure and admixture**

Our Principal Component Analysis (PCA) and archetypal analysis using neutral loci failed to separate individuals into distinct clusters and hence did not reveal any population structure among all sites in the East River Valley (PCA: PC1=6.56% and PC2=5.65%; Figure 2a; archetypal analysis: PC1=52.92%, PC2=47.08%; Figure 2b). Our ADMIXTURE analysis corroborated the PCA and archetypal analysis. The lowest Cross Validation (CV) indicated K =1 (Table S1). Together, these results indicate that *P. macdunnoughii* sampled from sites throughout the East River valley comprise a single population.

**Genetic variation and differentiation**

We quantified the relative abundance of private and shared alleles in a) between all sites where *T. arvense* is absent or present (i.e. between habitats) and b) pairwise comparison among individual sites. The allele frequency distribution for *P. macdunnoughii* in habitats with and without *T. arvense* were mostly low frequency alleles that were localized to each habitat (Figure 3a; 26% of cumulative fraction of variants), followed by 22% of shared and common alleles. The
rest of the variants were either rare or low frequency alleles that were localized to each habitat (Figure 3a). In contrast to the habitat-based allele frequency distribution, pairwise comparison of allele frequencies among sites consisted of alleles that were shared/common amongst all the sites (Figure 3b; 14% of the cumulative fraction of variants).

Overall, we found higher genetic variation among *P. macdunnoughii* in areas without *T. arvense* compared to areas with *T. arvense* (Figure 4a; Bartlett's $K^2=6.95$; df=1; \( p=0.008 \)). The observed heterozygosity was less than expected in areas without *T. arvense* (Figure 4a; Bartlett's $K^2=39.63$; df=1; \( p<0.001 \)) and the observed vs expected heterozygosity was similar in areas with *T. arvense*. Our genome-wide average estimates of nucleotide diversity (\( \pi \)) diversity indicated a significant increase in \( \pi \) in sites without *T. arvense* compared to sites with *T. arvense* (Figure 4b; Bartlett's $k^2=9.611$; df=1; \( p=0.001 \)).

Our estimate of genome-wide mean pairwise Fst among all combinations of sites was less then \( <0.001 \) (Figure S2), indicating low genetic differentiation and high relatedness among sites. A correlation map based on the population co-variance matrix indicated that all sites were weakly positively correlated. Accordingly, the Mantel test did not reveal any signs of isolation by distance (\( p=0.14, r=0.05 \)).

**Migration**

We observed significant variation in the magnitude and direction of relative migration of *P. macdunnoughii* among all sites (Figure 5). The estimated relative migration rates ranged from 0 to 1 with an average of 0.56. We observed unidirectional and bidirectional migration from areas with *T. arvense* to areas without *T. arvense* and vice versa. The highest rates of gene flow (Nm>0.85) were observed from Gothic Townsite to Elko Park (Nm=0.89) and vice versa.
(Nm=1), and from Quigley Creek to Gothic Townsite (Nm=0.96). Additionally, we observed high rates of relative gene flow to and from Gothic Townsite compared to other sites (Figure 5).

Estimated effective migration surface (EEMS) contours revealed the low relative effective gene flow of *P. macdunnoughii* in the East River Valley (Figure 6a). Most of the potential barriers also showed high posterior probabilities in the Bayesian estimation of migration parameters (Figure S3). Posterior probabilities of migration parameters for sites in the Upper East River valley (Copper Creek, Copper Creek 1st Crossing, Gothic Townsite, Rustler’s Gulch, Quigley Creek, and Elko Park) were higher (>0.95) compared to the sites in the Lower East River valley (Lower and Upper Brush Creek and Lower and Upper Cement Creek), which had lower posterior probabilities <.90. We also estimated the relative effective genetic diversity of *P. macdunnoughii* in the East River valley. Our analysis highlighted low relative effective genetic diversity of *P. macdunnoughii* in all sampled sites (Figure 6b). All sites showed a posterior probability of >.90 (Figure S4).

**Genotype-environment association (GEA) analysis**

The BayPass core model (-covis) allowed us to estimate the scaled covariance matrix of population allele frequencies that quantify genetic relationship among pairs of sites. The results of Ω estimates agreed with our Fst estimates indicating that all sites are genetically similar (Figure S2).

To identify outlier loci, the XtX (SNP specific Fst) estimates were calibrated by analyzing a POD (pseudo-observed data set) of 100,000 SNPs. At the 1% threshold (XtX >26) for POD, we identified 8600 outlier SNPs (Figure S5). Our analysis that included presence/absence of *T. arvense* (as the environmental covariate) under the auxiliary covariate model identified 1008
SNPs that were associated with the presence of *T. arvense* (BF > 20) (Figure S6). Overall, we identified nine SNPs that were shared by both the XtX outlier loci analysis and environmental association analysis (Figure 7).

Our SNPeff analysis indicated that of the nine SNPs, one was identified as a low impact protein coding change in the exon (unlikely to change protein behavior) with the rest of the SNPs impacts classified as modifiers (effects on non-coding regions and/or effects of gene regulation). The BlastX analysis of these regions revealed that these genes were primarily involved in cytoskeletal organization, DNA damage repair, lipidation of chylomicrons in the intestines, eye development, epithelial development, and catalysis of phosphoric acid. Importantly, regions that were under selection and associated with the presence of *T. arvense* have previously been linked to larval development and metabolism (Table 1, S2).

**Discussion**

*Pieris macdunnoughii* butterflies in the East River Valley comprise a single admixed population. Allele frequency differences between habitats (i.e., presence or absence of *Thlaspi arvense*) were largely driven by localized alleles, whereas pairwise allele frequency differences among sites were driven by geographically widespread alleles. Additionally, heterozygosity of butterflies was lower in areas where *T. arvense* occurred compared to areas where the plant was absent. Examining relative and effective gene flow patterns revealed bidirectional, asymmetric relative gene flow and low effective gene flow across the East River valley. Finally, we identified signatures of selection in *P. macdunnoughii* in response to *T. arvense*. Based on our annotation of the assembly, it appears that selection is affecting loci involved in larval ability to feed on *T. arvense*, not the adult females’ recognition and discrimination of potential hosts. In sum, we
identified genetic signatures of local selection in a native herbivore in response to a novel toxic hostplant at a fine geographic scale despite absence of strong population structure and genome-wide differentiation.

Evidence for a single population among sites

Our results did not separate individuals into distinct clusters and indicated that all sites comprise a single population, which accords with previous dispersal estimates for *P. macdunnoughii* in this area. Mark-release-recapture surveys in the 1970s and early 2000s estimated *P. macdunnoughii* mean dispersal between 400m and 700m (among dispersants, with approximately 0.42% of recaptures being dispersants) (Nakajima, 2014) which generally exceeds the extent of *T. arvense* in invaded areas (Nakajima et al., 2013). *P. macdunnoughii* disperses along large elevational gradients (Nakajima, 2014), occurring up to 4500m.a.s.l. Several species of butterflies are known to disperse larger distances, thus increasing gene flow and blurring population boundaries across heterogeneous environments (Kitahara, 2016; Spieth & Cordes, 2012; Takami et al., 2004).

Distinct allele frequency distribution patterns between habitats and among sites

Recent research has highlighted the utility of allele or site frequency spectra (AFS/SFS) to identify outlier loci, understand population structure, infer demographic changes, and identify signatures of positive selection. In our study, AFS differences of *P. macdunnoughii* between areas with and without *T. arvense* was largely driven by alleles that were localized or private to each habitat, even in the absence of strong population structure. However, overall differences among sites in general were largely due to differences in shared, common alleles. This is in contrast with theoretical and empirical work in other systems, which showed that pairwise
differences among sites within a population were driven by shared, common alleles (Biddanda et al., 2020; Gutenkunst et al., 2009). For instance, analysis of large-scale human genomic data from the 1000 Genomes Project (1KGP) showed that allele differences between populations were due to localized and rare alleles and differences between pairs of individuals regardless of population origin were due to common variants found globally (Biddanda et al., 2020). Allele frequency differences between populations can result due to local adaptation and/or new mutations specific to the population (Günther & Coop, 2013). In our case, the abundance of localized, low frequency alleles between the habitats could be due to the selection imposed by $T. arvense$ leading to putative adaptive alleles.

**Mechanism underlying local adaptation with high gene flow and low genetic variation.**

Our results highlight important mechanisms that underly local adaptation. Population genetics theory suggests that local adaptation occurs in the absence of gene flow and that high gene flow often leads to homogenization and maladaptation (Bachmann et al., 2020; Farkas et al., 2016; Gandon et al., 1996; Garant et al., 2007; Kirkpatrick & Barton, 1997; Lenormand, 2002). However, recent studies showed that local adaptation occurs in the presence of strong gene flow in certain scenarios (Fitzpatrick et al., 2015; Tigano & Friesen, 2016). For example, in spatially and temporally varying habitats, gene flow augments standing genetic variation and thus increases local adaptation (Blanquart et al., 2012, 2013). Local adaptation can also occur in the presence of gene flow through adaptive introgression and strong selection against the immigrant alleles (Griffiths et al., 2021; Leroy et al., 2020; Rendón-Anaya et al., 2021).

Our results indicated that all sites in the East River Valley were panmictic, with high gene flow, low genetic differentiation, and low genetic variation within the population. Nonetheless,
we were able to identify strong signatures of local adaptation in *P. macdunnoughii* where *T. arvense* occurred. Eggs laid on *T. arvense* die before pupation, which introduces fitness costs for immigrant individuals that prefer *T. arvense* (Nakajima, 2014; Nakajima & Boggs, 2015).

Similarly, local adaptation in lava flow lizards (melanism) occurred in the presence of strong gene flow and low genetic variation (Krohn et al., 2019). This is in line with recent theoretical and empirical work that suggests that environmentally driven local adaptation does not lead to genome wide differences or require substantial standing genetic variation unless the underlying traits are linked to reproduction (Feder et al., 2012; Krohn et al., 2019; Shafer & Wolf, 2013).

**Selection on oviposition vs larval performance.**

We did not find that the fitness costs for naïve females ovipositing on *T. arvense* translated into signatures of selection at the genomic level. Rather, loci under selection were near or in genes annotated with functions that support larval feeding and performance. Recent studies have found that genomic bases of herbivorous insects’ response to novel hostplants are polygenic, involving genes underlying oviposition, larval feeding, larval metabolism, and detoxification (Egan et al., 2015; Gompert et al., 2022; Vertacnik & Linnen, 2017). Similarly, we identified a polygenic response to selection from *T. arvense*, including 9 loci distributed across five chromosomes, causing 25 variant effects. Identification using SNPeff and BlastX revealed that the majority of the genes were involved in larval development and metabolism and most of the variant changes affected non-coding regions. Non-coding regions in the genome contain regulatory elements that play a critical role in protein assembly, gene expression and regulation and are under purifying selection (Andolfatto, 2005; Bird et al., 2006; Loehlin et al., 2010).
Therefore, these variant changes might have significant effects in larval feeding and detoxification.

We did not identify any loci under selection that were associated with sensory (olfactory, gustatory, or visual) functions that underlie female oviposition choice (de Fouchier et al., 2017; Engsontia et al., 2014; Ramaswamy et al., 1987; K. Yang et al., 2020). Lepidopteran females use a combination of sensory receptors to identify a potential hostplant (Haverkamp et al., 2018; Renwick & Chew, 1994; Thompson & Pellmyr, 1991). Females use olfaction and visual cues for long range detection of hostplants and the final decision is made after gustatory tactile contact with the hostplant using the first pair of foretarsi in their legs (Ozaki et al., 2011; Ryuda et al., 2013). Differences in oviposition choice in females are driven by underlying differences in their chemosensory repertoire. Thus, our failure to identify chemosensory genes in our outlier analysis suggests that selection might be acting on larvae instead of the females. Existing evidence suggests that error prone oviposition in Lepidoptera females can drive hostplant range expansions and adaptation to new hosts, since repeated oviposition on less suitable/non hostplants imposes selection on the larvae to evolve to feed on the plant (Janz et al., 1994; Nylin et al., 2000; Nylin & Janz, 2009; Stefanescu et al., 2012). Furthermore, *P. macdunnoughii* larvae that survive to later instars can be rescued when they are provided with suitable host plants (*C. cordifolia* or *D. incana*) and individual based models have suggested that fine-grained distribution of *T. arvense* and native host plants can alter population dynamics of the butterfly (Nakajima & Boggs, 2015). Thus, we hypothesize that selection is acting on larvae for longer survival on *T. arvense* potentially enabling later instar larvae to find suitable native host plants.

**Evolutionary constraints on oviposition preference and larval performance**
*Pieris macdunnoughii* and its sister taxa in North America are derived from the Eurasian *P. napi* during the last Holarctic speciation event (Chew & Watt, 2006). *Pieris napi* larvae can develop successfully on *T. arvense* and other invasive Eurasian mustards found in the Eastern US including *Alliaria petiolata* (Forsberg, 1987; Friberg & Wiklund, 2019; Prasai & Karlsson, 2011). *Pieris oleracea*, another North American species in the *Pieris* species complex has reduced fitness when feeding on *A. petiolata* due to its toxicity to the larvae (Chew, 1977b; Haribal et al., 2001, 2001; Haribal & Renwick, 1998; Huang et al., 1994; Keeler & Chew, 2008). Thus, the North American *P. oleracea* larvae have lost the ability to develop or have reduced fitness on introduced mustards from Eurasia while females have retained the ancestral preference for hostplants. This mismatch in preference-performance is the underlying cause for maladaptation in the larvae. Our results suggest that any adaptation by *P. macdunnoughii* to be able to use the plant will likely involve evolution of larval ability to develop on *T. arvense* and not of the females' avoidance of oviposition on *T. arvense*. Adaptation in response to *T. arvense* in the larvae would involve longer survival on *T. arvense* followed by rescue in the later instars. Indeed, after decades of maladaptation of *P. oleracea* on *A. petiolata*, the larvae are now able to develop on the plant during its bolting stage but not the rosette stage (Keeler & Chew, 2008). This highlights the importance of understanding the interactions between plant phenology, plant chemical composition and larval performance. Our results suggest that a similar outcome may occur in *P. macdunnoughii* in the event of rapid evolution due to selection. Future research quantifying fine scale spatial and temporal patterns of larval performance coupled with quantifying *T. arvense* distribution and variation in plant chemical profiles might provide insights on the escape from or persistence of the evolutionary traps.
Conclusions:

Instances of interactions between native and non-native species are projected to increase across the globe due to range shifts, competition from invasive species and climate change. Understanding the effects of these interactions requires careful dissection of the ecological and molecular processes that mediate these interactions. Our results build upon decades of research aimed at understanding the causes and maintenance of evolutionary traps, to elucidate the molecular response and potential for adaptation to novel resources (Brady, Bolnick, Barrett, et al., 2019; Farkas et al., 2015; Gilroy & Sutherland, 2007; Robertson et al., 2013; Schlaepfer et al., 2002, 2005). As insects are currently experiencing a global decline (Hallmann et al., 2017; Nakajima & Boggs, 2015; Wagner et al., 2021), the Pieris macdunnoughii - Thlaspi arvense system can serve as a model to understand and predict the outcomes of these interactions in insects even at a fine geographical scale. This work identifies the molecular mechanisms that underlies the evolutionary trap, thus pioneering Pieris macdunnoughii - Thlaspi arvense as a model system to understand maladaptation and evolutionary traps in the face of climate change (Chew, 1977b; Nakajima, 2014; Steward et al., 2019, 2021, 2022; Steward & Boggs, 2020). Our work also highlights the potential of adaptation to occur in a fine-grained landscape in the absence of genetic variation and high gene flow.
Acknowledgements

We thank W. Watt, J. Quattro, P. Andolfatto, B. Hollis and C. Wessinger for providing feedback on the manuscript. We thank J. Nicklas, L. Tran, L. Fisher, L. Hubbard, T. Feemster, M. Brown, M. Olson, S. McDaniel, H. Walton, E. Wagner, and O. Shrestha for assisting with field work. We thank K. Deitz and A. Pinharanda for assistance and advice on Tn5 library preparation. We thank K. Correa Veléz, J. Rivera-Hernandez, B. Stokes, J. Stewart, and D. Reisman for advice on PCR and sequencing. M. Kuwalekar, S. Baral, R. Deshmukh, M. Stager, O. Linestad, D. Speiser and C. Wessinger provided feedback on bioinformatic data analysis. Field research permits from the US Forest Service were arranged by the Rocky Mountain Biological Laboratory (RMBL). Funding was provided by the University of South Carolina College of Arts & Sciences to CLB, an NSF Graduate Research Fellowship to RAS, and RMBL graduate fellowships to NR and RAS.

References


Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence...


Data accessibility statement

Raw sequence reads will be deposited in the NCBI SRA database upon acceptance and prior to publication. Metadata will also be stored in the NCBI SRA using the Invertebrate MIxS version 1.0 Package. The source code for data analysis will be made available on Github upon acceptance and prior to publication.

Benefits generated

Benefits from this research will ensue from the data and results shared from this study on public databases as described above.

Author contributions

NR and CLB designed the research. RAS collected specimens in 2016; NR collected specimens in 2019. NR did the molecular work and analyzed the data. NR wrote the draft manuscript. All authors contributed to manuscript revisions. CLB provided oversight for all stages of the work.
Tables

Table 1: Annotation of outlier loci identified by BayPass X\(\text{tX} \ (\text{XtX} > 26)\) and BF(dB > 20), associated isoforms and predicted effects. Annotation was carried out using SNPeff and BlastX.
<table>
<thead>
<tr>
<th>Loci (SNP) position</th>
<th>Chromosome</th>
<th>Codon affected</th>
<th>Intervals affected</th>
<th>BlastX identification</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>156011</td>
<td>1</td>
<td>STOP</td>
<td>153581 - 153583</td>
<td>cilia- and flagella-associated protein 410 isoform</td>
<td>Regulation of cell morphology and cytoskeletal organization (Bai et al., 2011); Involved in DNA damage repair (Lai et al., 2011).</td>
</tr>
<tr>
<td>3430127</td>
<td>7</td>
<td>STOP</td>
<td>3428853 - 3428855</td>
<td>uncharacterized protein</td>
<td></td>
</tr>
<tr>
<td>11523752</td>
<td>9</td>
<td>START</td>
<td>11545167 - 11545169</td>
<td>ARFRP 1</td>
<td>Lipidation of chylomicrons in the intestine and required for VLDL lipation in the liver (Jaschke et al., 2012).</td>
</tr>
<tr>
<td>4595201</td>
<td>9</td>
<td>START</td>
<td>4592359 - 4592361</td>
<td>homeobox protein Hox-A3-like</td>
<td>Part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior axis (Gaudet et al., 2011).</td>
</tr>
<tr>
<td>4595201</td>
<td>9</td>
<td>STOP</td>
<td>4595024 - 4595026</td>
<td>homeobox protein Hox-A3-like</td>
<td>Part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior axis (Gaudet et al., 2011).</td>
</tr>
<tr>
<td>4595201</td>
<td>9</td>
<td>START</td>
<td>4597927 - 4597929</td>
<td>retinal homeobox protein Rax-like</td>
<td>Plays a critical role in eye formation by regulating the initial specification of retinal cells and/or their subsequent proliferation (Kimura et al., 2000).</td>
</tr>
<tr>
<td>4595201</td>
<td>9</td>
<td>STOP</td>
<td>4599618 - 4599620</td>
<td>retinal homeobox protein Rax-like</td>
<td>Plays a critical role in eye formation by regulating the initial specification of retinal cells and/or their subsequent proliferation (Kimura et al., 2000).</td>
</tr>
<tr>
<td>11703864</td>
<td>14</td>
<td>START</td>
<td>11697968 - 11697970</td>
<td>DOCK 4</td>
<td>Functions as a guanine nucleotide exchange factor (GEF) that promotes the exchange of GDP to GTP, converting inactive GDP-</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Accession</td>
<td>Start</td>
<td>Stop</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>11703864</td>
<td>14</td>
<td>STOP</td>
<td>11701979 - 11701981</td>
<td>inx3 Structural components of the gap junctions. Essential for proper epithelial development of the epidermis (Lehmann et al., 2006).</td>
<td></td>
</tr>
<tr>
<td>5741973</td>
<td>14</td>
<td>START</td>
<td>5730927 - 5730929</td>
<td>mediator of RNA polymerase II transcription subunit 15 Required for activated transcription of the MtnA, MtnB and MtnD genes. Negatively regulates sex comb development (Boube et al., 2000); Required for cholesterol-dependent gene regulation. Positively regulates the Nodal signaling pathway (F. Yang et al., 2006).</td>
<td></td>
</tr>
<tr>
<td>5741973</td>
<td>14</td>
<td>STOP</td>
<td>5742219 - 5742221</td>
<td>tripartite motif-containing protein 45 May act as a transcriptional repressor in mitogen-activated protein kinase signaling pathway (Wang et al., 2004).</td>
<td></td>
</tr>
<tr>
<td>261573</td>
<td>xfSc0000009</td>
<td>STOP</td>
<td>261794 - 261796</td>
<td>protein ALP1-like Encodes an alkaline phosphatase. Alkaline phosphatases catalyze the hydrolysis of monoesters of phosphoric acid and a transphosphorylation reaction in the presence of large concentrations of phosphate acceptors (Harper &amp; Armstrong, 1972).</td>
<td></td>
</tr>
<tr>
<td>261573</td>
<td>xfSc0000009</td>
<td>STOP</td>
<td>263610 - 263612</td>
<td>Transposase</td>
<td></td>
</tr>
<tr>
<td>577356</td>
<td>xfSc0000009</td>
<td>START</td>
<td>573908 - 573910</td>
<td>uncharacterized protein</td>
<td></td>
</tr>
</tbody>
</table>
**Figure legends**

**Figure 1:** Map of sites where butterflies were collected in the East River valley. Sites in red represent areas where *Thlaspi arvense* does not occur and sites in brown represent areas where *T. arvense* is present. The size of the circle represents the magnitude of sample sizes from each location.

**Figure 2:** a) Genome-wide Principal Component Analysis (axes 1 and 2) using neutral loci indicating a mixed population without distinct clustering by site. b) Archetypal analysis (axes 1 and 2) with k=3 again showing a mixed population without site-specific clusters.

**Figure 3:** a) Relative abundance of allele frequency variants of *P. macdunnoughii* in areas with *T. arvense* and areas without *T. arvense*. The codes depicted in the figure represents the frequency of the minor allele in the habitat (U-undetected: No alleles present, R-rare: <1% minor allele frequency (MAF), L-low frequency: 1%-5% MAF, C-common: >5% MAF). The percentages and their corresponding numbers indicate the number of variants in each class and are grouped based on SNP identity. Grey rows represent alleles whose contribution was not significant to the differences between comparisons. b) Relative abundance of variants of *P. macdunnoughii* in our study area. Ucc: Upper Cement Creek, Ubc: Upper Brush Creek, Cc: Copper Creek, 1c: 1st Crossing, 401: 401 site, Rg: Rustler's Gulch, Qc: Quigley Creek, Ep: Elko Park, Lcc: Lower Cement Creek, Lbc: Lower Brush Creek, Gt: Gothic townsite.

**Figure 4:** a) Genome-wide variance in expected and observed heterozygosity of *P. macdunnoughii* in the East River valley. Bartlett's tests were used to compare observed heterzygosity between habitats where *T. arvense* was present or absent, and observed versus expected heterozygosity within each of these habitat types. b) Genome-wide variance in nucleotide diversity (π) of *P. macdunnoughii* in habitats without and with *T. arvense* (Bartlett’s test, ***P < 0.001***).  

**Figure 5:** Relative migration rates of *P. macdunnoughii* among the 11 sites in the East River valley estimated using divMigrate based on the number of migrants per generation (Nm)). Colors indicate migration levels (low = blue, high = red), and arrows indicate direction of migration. Solid lines represent unidirectional migration and dashed lines represent bi-directional migration. Colors in bidirectional migration correspond to the migration level of the closest arrow. Migration routes are only shown for Nm > 0.55 and migration routes from areas with *T. arvense* to areas without *T. arvense* and vice-versa.

**Figure 6:** a) Estimated Effective Migration Surfaces and b) Estimated Effective Diversity Surfaces for *P. macdunnoughii* in the East River Valley. The migration and diversity rates, log(m) and log(q) represent gene flow or genetic diversity barriers and corridors in the habitat, respectively. Each value corresponds to a 10-fold increase (blue) or decrease (orange/brown) in migration or genetic diversity compared to the null hypothesis of isolation by distance (white).

**Figure 7:** Pairwise comparison of median SNP XtX and BFmc values from three independent BayPass analyses for adaptive differentiation associated with *T. arvense* presence in the habitat. The vertical dashed line represents the 1% POD (Pseudo Observed Dataset) significance threshold (XtX=26), and the horizontal dashed line represents the 20-dB threshold for BFmc. Blue dots represent the 9 loci of interest (outliers for both XtX and BFmc).
Figure 3
Figure 4

![Box plots showing heterozygosity and average \( \pi \) for Thlaspi arvense.](image)

**Legend**

- **Absent**
- **Present**

- **Expected**
- **Observed**

- **Significance Levels**
  - ***p < 0.001***
  - **p < 0.01***
  - **n.s.** (not significant)