Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climatic gradient

Ryan, Sean; Fontaine, Michael; Scriber, Mark; Pfrender, Michael; O'Neil, Shawn; Hellmann, Jessica

Published in:
Molecular Ecology

DOI:
10.1111/mec.14236

Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climatic gradient

Sean F. Ryan,[1,2] Michael C. Fontaine,[3] J. Mark Scriber,[4,5] Michael E. Pfrender,[2,6] Shawn T. O'Neil,[2,7] Jessica J. Hellmann[2,8]

[1]USDA ARS, 1600/1700 SW 23rd Drive, Gainesville, FL 32608 USA
[2]Department of Biological Sciences, University of Notre Dame, South Bend, IN 46556 USA
[3]Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen Nijenborgh 7, 9747 AG Groningen, Netherlands.
[4]Department of Entomology, Michigan State University, East Lansing, MI 48824 USA
[5]McGuire Center for Lepidoptera and Diversity, University of Florida, Gainesville FL 32611 USA
[6]Environmental Change Initiative, University of Notre Dame, South Bend, IN 46556 USA
[7] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, 97333 USA
[8]Institute on the Environment & Department of Ecology, Evolution and Behavior, University of Minnesota, 1954 Buford Ave., St. Paul, MN 55108 USA

Corresponding author: Sean F. Ryan; citscisean@gmail.com

Abstract

Hybrid zones are a valuable tool for studying the process of speciation and for identifying the genomic regions undergoing divergence and the ecological (extrinsic) and non-ecological (intrinsic) factors involved. Here, we explored the genomic and geographic landscape of divergence in a hybrid zone between *Papilio glaucus* and *Papilio canadensis*. Using a genome scan of 28,417 ddRAD SNPs, we identified genomic regions under possible selection and examined their distribution in the context of previously identified candidate genes for ecological adaptations. We showed that differentiation was genome-wide, including multiple candidate genes for ecological adaptations, particularly those involved in seasonal adaptation and host plant detoxification. The Z-chromosome and four autosomes showed a disproportionate amount of differentiation, suggesting genes on these chromosomes play a potential role in reproductive isolation. Cline analyses of significantly differentiated genomic SNPs, and of species diagnostic genetic markers, showed a high degree of geographic coincidence (81%) and concordance (80%) and were associated with the geographic distribution of a climate-mediated developmental threshold (length of the growing season). A relatively large proportion (1.3%) of the outliers for divergent selection were not associated with candidate genes for ecological adaptations and may reflect the presence of previously unrecognized intrinsic barriers between these species. These results suggest that exogenous (climate-mediated) and endogenous (unknown) clines may have become coupled and act together to reinforce reproductive isolation. This approach of assessing divergence across both the genomic and geographic landscape can provide insight about the interplay between the genetic architecture of reproductive isolation and endogenous and exogenous selection.

Introduction

Hybrid zones are ideal systems for exploring the interplay between exogenous (ecological factors, such as climate) and endogenous (genetic incompatibilities) selective pressures that drive genomic divergence. Although climate has been implicated in the maintenance of a number of hybrid zones across a broad range of taxa (Taylor *et al.* 2015), our understanding of how climate drives patterns of divergence and gene flow between the genomes of hybridizing species remains limited to a handful of systems (Taylor *et al.* 2014; de Villemereuil *et al.* 2014; Hamilton *et al.* 2014).

It has long been recognized that an advantage to studying hybrid zones is their utility in understanding the genetic basis of species divergence, i.e., identifying those regions of the genome under selection and potentially involved in reproductive isolation (Barton & Hewitt 1985). A classical and still widely used approach is to look at how genetic differences between species vary across geography and species boundaries, through a geographic cline analysis. Complementary to cline analysis, there are a number of population genomic approaches to identify genomic regions under selection (i.e., via "genome scans") that do not require spatial sampling along a cline. Genome scans can be used to identify outlier loci from a null expectation based on genome-wide ancestry that are potential indicators of divergent selection (Foll & Gaggiotti 2008). Given that geographic cline analyses and genome scans estimate selection dynamics operating on different spatiotemporal scales, from recent to more long-term time scale respectively, combing these approaches increases the power to detect regions of the genome involved in local adaptation and reproductive isolation (de Lafontaine *et al.* 2015).

Placing patterns of genetic differentiation within the genomic landscape can also provide insight as to how genetic architecture itself influences introgression across the genome—the

67 "genomic architecture of divergence" (Teeter *et al.* 2010; Gompert *et al.* 2012). A number of
68 studies suggest that gene flow may be heterogeneous across the genome of hybridizing species
69 (Mallet 2005; Teeter *et al.* 2008), with large amounts of the genome diffusing between species
70 while a few regions maintain exceptionally high levels of divergence (Fontaine *et al.* 2015).
71 Specifically, regions of low recombination such as inversions, centromeres and sex chromosomes
72 are often found to have elevated levels of divergence and often harbor a disproportionate amount
73 of the factors responsible for reproductive isolation (Coyne & Orr 1989; Presgraves 2008).

74 Both exogenous and endogenous factors can act to inhibit introgression and maintain
75 species boundaries. Endogenous factors have often been found to play a major role in driving
76 clinal patterns of genetic variation across many hybrid zones (Bierne *et al.* 2011). For example,
77 most tension zones, hybrid zones maintained by a balance between selection and migration, are
78 believed to be the result of intrinsic incompatibles. The geographic selection gradient model
79 (GSGM), a mathematically similar model to the tension zone model, can also produce similar
80 clinal patterns, with the main difference being selection is primarily exogenous (e.g., related to
81 abiotic factors) (Endler 1977). However, endogenous and exogenous barriers are not mutually
82 exclusive (Barton & Hewitt 1985; Hewitt 1988; Bierne *et al.* 2011). In fact, these barriers may
83 become "coupled" and act synergistically to increase reproductive isolation and maintain the
84 location of the hybrid zone, often resulting in tension zones becoming associated (co-occurring)
85 within an ecological gradient (Barton & Hewitt 1985; Bierne *et al.* 2011).

86 The hybrid zone between the eastern tiger swallowtail (*Papilio glaucus*) and the Canadian
87 tiger swallowtail (*P. canadensis*) butterfly is an excellent system for exploring how climate
88 influences genomic patterns of divergence. The ecology of this system has been studied
89 extensively for over thirty years and climate is implicated as a prominent factor in maintaining

reproductive isolation between the two species (Scriber 2011). However, an understanding of how endogenous and exogenous selective forces maintain divergence across the genome remains unresolved. The two species are estimated to have diverged 0.5-0.6 million years ago (Cong *et al.* 2015) and previous genetic and morphological analysis of individuals suggest they form a narrow (< 50 km) east-west hybrid zone in the north eastern United States, extending from Minnesota to New England (Fig 1A) (Scriber 1990). A number of phenotypic traits, including those that are physiological (Scriber *et al.* 1989; Kukal *et al.* 1991; Mercader & Scriber 2008), behavioral (Scriber *et al.* 1991; Deering & Scriber 2002; Mercader *et al.* 2009), and morphological (Luebke *et al.* 1988; Lehnert *et al.* 2012), appear to coincide (exhibit similar clines) with latitudinal and altitudinal variation in climate (Hagen & Scriber 1989; Scriber 2002; Scriber *et al.* 2003). A better understanding of how genes putatively involved in ecological adaptations in this system covary with environmental factors (the geographic landscape) and how they are distributed across the genome (the genomic landscape) will help to elucidate how ecological selection pressures drive reproductive isolation within the genomes of these hybridizing species.

The hybrid zone between *P. glaucus* and *P. canadensis* is believed to be maintained primarily by a climate-mediated developmental threshold, produced by geographic variation in the length of the growing season (Scriber & Lederhouse 1992). One line of evidence implicating the growing season is that the hybrid zone coincides with a change in a number of ecologically adaptive traits related to seasonality including the transition of facultative diapause induction of *Papilio glaucus* to the obligate diapause induction of *P. canadensis*. It is presumed that *P. glaucus* alleles involved in facultative diapause would be disadvantageous north of the hybrid zone where on average the growing season is too short to complete two generations (Ritland & Scriber 1985). Conversely, alleles underlying obligate diapause phenotype may be

113 disadvantageous south of the hybrid zone, as they would likely lead to phenotypes with lower
114 fitness than those that emerged for a second generation.

115 In the *P. glaucus* - *P. canadensis* system, diapause is largely Z-linked (Hagen & Scriber
116 1989); like birds, Lepidoptera has a ZW system (Z=X, W=Y) of sex determination and females
117 are the heterogametic sex. Divergent selection acting on genes involved in diapause has been
118 invoked as a possible explanation for the observed lack of introgression of mtDNA and Z-linked
119 markers (Scriber *et al.* 2008; Scriber 2011; Scriber *et al.* 2014). Previous studies have
120 documented a dramatic change in allele frequencies of Z-linked allozymes (LDH and PGD), an
121 mtDNA gene (cytochrome oxidase I) and physiological traits (geographic distribution of host-
122 specific detoxification abilities and host plant preferences) that coincide with the environmental
123 gradient that is the length of the growing season (reviewed in Scriber 2011). Given that diapause
124 appears to be an important ecological trait that differs between these species and may be involved
125 in maintaining the location of the hybrid zone, Scriber (2011) suggested that the Z-chromosome
126 may harbor a disproportionate amount of divergence.

127 In addition to climatic variation, other extrinsic factors may be involved in the
128 maintenance of this hybrid zone, including the distribution of host plants (Lindroth *et al.* 1988;
129 Mercader *et al.* 2009). While both species are highly phytophagous, they differ in their
130 preferences for, and ability to metabolize, linear and angular furanocoumarins, resulting in
131 differential detoxification abilities for some host species (e.g., *Populus tremuloides* can be
132 detoxified by *P. canadensis*, but not *P. glaucus*, with the opposite true for *Liriodendron*
133 *tulipifera*) (Hung *et al.* 1997). Candidate genes involved in, or linked to, traits associated with
134 various ecological adaptations, such as host plant detoxification, larval chemical defenses, and

diapause have been identified (Hagen & Scriber 1989; Li *et al.* 2002, 2003; Cong *et al.* 2015). Yet, how these candidate genes are distributed within the genome and coincide with patterns of genomic divergence, or how they vary with geography and ecological gradients is not well understood. Further, given that extrinsic factors vary in their geographic distribution, they may produce different underlying selective landscapes that could lead to different patterns of introgression across the hybrid zone.

In this study, we characterized genetic differentiation across the genomic and geographic landscape of the ecologically well-studied *P. glaucus* and *P. canadensis* hybrid zone. Specifically, we screened a reduced representation of genetic diversity across the genome using a ddRADseq protocol and characterized genome-wide patterns of differentiation between *P. glaucus* and *P. canadensis* along a 650 km latitudinal transect in the Wisconsin portion of the hybrid zone. First, we set out to characterize genetic structure between the two species and identify regions of the genome potentially impacted by natural selection. We then placed these patterns of differentiation within the genomic landscape to explore the genomic architecture of differentiation and assess whether differentiation is genome-wide or variable among chromosomes (particularly the Z-chromosome). Further, we evaluated the prediction that the distribution of genes previously identified as being putatively involved in ecological adaptations would be associated with regions of the genome under selection and exhibit steep clinal variation across the hybrid zone (i.e., provide further evidence for their being involved in ecological adaptations). Finally, we used a geographic cline approach to explore how significantly differentiated genomic regions (F_{ST} outliers in allopatry) vary across geography (and exhibit steep clinal variation in sympatry—hybrid zone) and evaluate the relative roles endogenous and exogenous factors may be contributing to patterns of differentiation observed within the genome.

Methods and materials

Sample collection and preservation

Three hundred and fourteen specimens were collected between 2007 and 2013 from across a latitudinal transect that spans the *P. canadensis* and *P. glaucus* hybrid zone, from the Upper Peninsula of Michigan through Wisconsin to central Illinois (Table 1; Fig 1A). Of these specimens, all 314 were used to genotype individuals for species diagnostic mitochondrial haplotypes in the cytochrome oxidase subunit1 (*COI*) gene (Sperling 1993), 192 for a species diagnostic Single Nucleotide Polymorphism (SNP) in the lactate dehydrogenase (*Ldh*) gene, and 101 (80 after filtering) used for the generation of reduced representation of the genome-wide diversity using the ddRADseq protocol. Only males were used in all analyses because females cannot be heterozygous for loci on the Z chromosome (females are hemizygous).

DNA extraction and genotyping/phenotyping of diagnostic markers

Briefly, gDNA was isolated from leg tissue using a phenol:chloroform protocol (Sambrook *et al.* 1989). The two species diagnostic mtDNA haplotypes (*COI* gene) were scored using a Restriction Fragment Length Polymorphism (RFLP) protocol adapted from Putnam *et al.* (2007). A TaqMan SNP genotyping procedure was used to genotype individuals for the Z-linked SNP located in the coding region of *Ldh* (Table S1). This mutation leads to a non-synonymous substitution that is fixed between the two species and is believed to lead to different LDH isozymes (Putnam *et al.* 2007) that may result in thermophysiological differences between the species (see Fig S1). More details on DNA extraction genotyping of diagnostic markers are provided in supplementary material; all data can be found in Data S1.

ddRADseq library preparation, sequencing, and bioinformatic post-processing

Two reduced-complexity libraries were generated for 101 individuals using a variation of the Restriction-Site Associated DNA fragment (RAD) procedure (adapted from Nosil *et al.* 2012) involving a double digest (ddRADseq; Parchman *et al.* 2012). Briefly, approximately 600 ng of genomic DNA was digested using the restriction enzymes EcoRI and MseI. For each individual, the EcoRI adapter contained a unique 8 to 10 base pairs (bp) index for barcoding. PCR product was then pooled (into a five and 96 individual library), purified with QIAquick Purification Kit (Qiagen), and size selected (400-600 bp range) on a BluePippin System (Sage Science). One library (containing five individuals) was run on an Illumina MiSeq (Notre Dame Genomics Core) and one (containing 96 individuals) was run on an Illumina HiSeq 2000 (BGI, UC Davis) using paired-end sequencing (150 bp reads on MiSeq and 100 bp reads on the HiSeq)

Raw reads were demultiplexed using a custom Perl script and trimmed and cleaned with the program Trimmomatic (v.0.32, Bolger et al. 2014) using default settings. The first 20 bp from all reads was clipped to remove barcodes and cutsite and only reads with a minimum read length of 30 bp were kept. The 3' end of forward unpaired reads were of lower quality and were trimmed to 80 and 60 bp for MiSeq and HiSeq reads respectively. Reads were then aligned to the *Papilio glaucus* genome v1 (Cong *et al.* 2015) using BWA-*aln* (v.0.7.8, Li and Durbin 2009). Read groups were added using custom Perl scripts, soft-clipped using Picard (v.1.33; http://picard.sourceforge.net) and realigned using GATK v.3.3 IndelRealigner (DePristo *et al.* 2011). Variant calling was performed using GATK HaplotypeCaller (McKenna *et al.* 2010; DePristo *et al.* 2011) with the default settings. Filters were applied to all SNPs in the following order. Variants were filtered for bialellic SNPs and indels removed using VCFtools v0.1.12a (Danecek *et al.* 2011). Individuals with excessive missing data (> 90%) were removed (n=3).

Rare alleles with a frequency >95% and <5% (pooling all individuals) were removed using VCFtools. GATK's default "hard filtering" was applied to keep only high quality variants (QD > 2.0, MQ > 30, ReadPosRankSum > -8.0, HaplotypeScore > 13, MappingQualityRankSum >-12.5). Only genotypes with a ~99% likelihood of being correct were kept by removing SNPs with a genotype quality (GQ) < 20. Following these filters, a minimum of 6X coverage for at least 75% of individuals in the "pure" species populations (*P. glaucus*: locations 1 and 2; *P. canadensis*: locations 7 and 8) was then applied (referred to as RAD1 dataset; 28,417 SNPs). For clinal analyses, we used a subset of the RAD1 dataset that had the additional requirement that a SNP had a minimum of 6X coverage in at least 75% of individuals from *all* eight locations. Following this filter, individuals with less than an average of 14X coverage (for all SNPs) were removed (referred to as RAD2 dataset; 18,316 SNPs) because the estimates of heterozygosity for individuals below this cut-off could be biased by low coverage (Fig S2).

Putative chromosome assignment of genetic markers

To determine the chromosome and orientation of each *P. glaucus* scaffold (while maintaining the orientation of SNPs within the *P. glaucus* scaffold) we used blastx (ncbi-blast-2.2.30+) (Altschul *et al.* 1990; Shiryev *et al.* 2007) to blast all peptide sequences within each scaffold of the *P. glaucus* assembly against the *Bombyx mori* genome using SilkDB v.2.0 (Wang *et al.* 2005). The *B. mori* scaffold with the most significant blast hits (based on Bit scores) was retained and used to determine the chromosome of each *P. glaucus* scaffold and their orientation within a chromosome. Given the high level of synteny between *P. glaucus* and *B. mori* (Cong *et al.* 2015), we expect this approach to reflect a moderately accurate representation of the genomic distribution of our ddRADseq loci.

227 *Genetic diversity and structure*

228 Genetic structure was assessed using the Bayesian clustering approach of STRUCTURE
229 v2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003) using an admixture model with correlated allele
230 frequencies, a burn-in of 100,000 Monte Carlo Markov chains (MCMCs) and a run length of
231 500,000 iterations, testing clusters (*K*) from 1 to 5; 15 replicates were run for each value of K.
232 We used the RAD2 dataset pruned for linkage disequilibrium (LD) by keeping only1 SNP per 1
233 kb (n=6,769 SNPs). The results were post-processed using CLUMPAK (Kopelman *et al.* 2015)
234 and the best *K* was chosen following the Evanno's method (Evanno *et al.* 2005) as implemented
235 in Structure Harvester v.0.6.94 (Dent & vonHoldt 2012).

236 A Principal Component Analysis (PCA) was also performed in R v 3.2.2 (R Core Team
237 2013) using the LD-pruned RAD1 dataset and the package *adegenet* v.1.4-2 (Jombart & Ahmed
238 2011) to provide a complementary visualization of the genetic structure without relying on any
239 model-based assumptions.

240 We estimated the levels of genetic diversity within each locality from the RAD2 dataset
241 using the per-site nucleotide diversity (π) (Nei & Li 1979), as well as the inbreeding index (F_{IS})
242 using VCFtools v0.1.12b. Linkage disequilibrium (LD) was calculated as mean r^2 for all pairwise
243 SNPs within each location using a modified version of plink-1.07 that allows for our large
244 number of "linkage groups" (scaffolds) (Malenfant *et al.* 2015). Confidence intervals were
245 generated by randomly subsampling 100 times seven individuals (smallest n of all locations)
246 from each location. Using these estimates of r^2 we then calculated the distance (in kb) at which r^2
247 is predicted to fall below 0.2 for each chromosomes (1-28) by fitting the relationship between r^2
248 and distance between SNPs within each scaffold to a non-linear model (Hill & Weir 1988)
249 describing the rate of LD decay similar to Marroni *et al.* (2011). A subset of the RAD2 dataset,

using SNPs with $F_{ST} > 0.9$ and filtered to 1 SNP per 1 kb (i.e. if more than one SNP were within < 1 kb within a scaffold, only the first SNP was retained; 228 SNPs) were used to calculate the hybrid index and heterospecific heterozygosity of each individual using the hi.index and int.het functions in the introgress package (Gompert & Buerkle 2010) in the R computing environment (R Core Team 2013).

Detecting (outlier) loci potentially under selection

We investigated the pattern of interspecific differentiation using two genome scan approaches conducted on the allopatric populations at the extremities of the transect, location 1 and 2 for *P. glaucus* (n = 18) and locations 7 and 8 for *P. canadensis* (n = 18). After removal of monomorphic loci, we first quantified the level of differentiation in allelic frequencies for each SNP using the dataset filtered using just the allopatric populations (RAD1; 28,417 SNPs) with the F_{ST} estimator of Weir & Cockerham (1984) using VCFtools. We identified SNPs potentially evolving under natural selection using the F_{ST}-outliers approach originally described in Beaumont and Nichols (1996) and implemented in ARLEQUIN v.3.5 (Excoffier & Lischer 2010). We estimated the 99.9% neutrality envelope of F_{ST} as a function of the heterozygosity using 500,000 coalescent simulations under a neutral model of population evolution and a Hierarchical Island Model consisting of 10 groups and 100 demes. SNPs above the upper 99.9% quantile were considered under potential divergent selection while those under the lower 0.001% quantile to be under potential balancing selection.

In addition, we used a Bayesian method implemented in BayeScan v.2.1 to estimate the probability that each locus is subject to divergent selection (Foll & Gaggiotti 2008). To account for multiple comparisons and reduce the amount of false positives, we set the prior odds for the

neutral model (pr_odds) to 10,000 and used a false discovery rate of 0.01. Twenty pilot runs of 5,000 iterations were used with a thinning interval of 10 and a burn-in of 50,000 iterations followed by 5,000 iterations for output.

Loci were considered as outliers for potential divergent selection if they were identified as significant with both methods (BayeScan and ARLEQUIN). To test whether the outlier loci were non-randomly distributed across chromosomes, a Fisher's Exact Test was used to compare the distribution of outlier loci for divergent selection and balancing selection to a null distribution, using all linkage groups as well as a separate analysis with just autosomes; a Bonferroni correction was applied to control for multiple testing ($\alpha = 0.01$/number of linkage groups; alpha = 0.00034 and 0.0038 when using all linkage groups and just autosomes, respectively). This analysis was also done using the "divergence hotspots" (genes that always show larger inter-species variation on both the protein and DNA sequence level) identified in Cong et al. (2015) to determine whether these highly divergent regions were non-randomly distributed across linkage groups and compare their distribution to outlier regions from our study.

Geographic cline analysis

We evaluated coincidence (amount of congruence among estimates for cline center for each SNP) and concordance (amount of congruence among estimates for cline width for each SNP) using the RAD2 dataset for ddRADseq loci identified as outliers (see above) that were in HWE in each location and filtered to 1 SNP per 1 kb (228 SNPs), as well as mtDNA (*COI*) and Z-linked (*Ldh*) loci. We fitted a combination of equations that describe the sigmoidal shape at the center of each cline and two exponential decay curves (Szymura & Barton 1986, 1991). A total of 15 different models that varied in shape (5 possibilities: having no tails, symmetrical tails,

296 asymmetrical tails, left tail only, or right tail only) and scaling of tails (3 possibilities: using no
297 scale, empirical estimates or best fit) were compared using a Metropolis-Hasting algorithm
298 implemented in the HZAR package in R (Derryberry *et al.* 2014). Posterior distribution of each
299 model parameter was estimated using three independent chains run using 500,000 MCMC steps
300 after a burn-in of 100,000 steps. This process was repeated three times for each model and each
301 MCMC trace was visually inspected to ensure convergence of the results. Model selection was
302 based on the lowest Akaike Information Criterion (AIC) value and parameter estimates of these
303 selected models were used (Akaike 1974). The cline center and width of any two loci were
304 considered non-coincident or discordant if the 2-log-likelihood unit (comparable to 95%
305 confidence intervals) for the upper and lower bound estimate of cline centers or width did not
306 overlap, respectively.

307 To determine whether the clinal patterns of loci coincide with a climatic gradient, the
308 estimated center of clinal loci were compared to geographic variation in length of the growing
309 season. Growing degree days (GDD) for each county along a transect spanning the hybrid zone
310 were calculated with PRISMs time series datasets modeled using climatologically-aided
311 interpolation at a 4km resolution (PRISM Climate Group 2014). For each year from 1980 to
312 2012, GDDs were calculated as the sum of the mean daily temperature ($(T_{min} - T_{min})/2 - T_{base}$),
313 between March 1st – Oct 31st. A T_{base} of 10° C was chosen based on the lower activity threshold
314 of *P. glaucus* estimated by Ritland & Scriber (1985) at 4km pixels in ArcGIS® (ESRI v10.2) and
315 then averaged to the county level. The 32-year mean and standard deviation among years for each
316 county were used to examine the relationship between the total GDDs and distance across this
317 transect. It has been estimated that the GDD requirement for *P. glaucus* to complete two
318 generations is at least 1300 GDDs (a conservative estimate) (Ritland & Scriber 1985). A linear

319 model was fit to the GDD data by location (distance along the transect). The fit from this model
320 was then used to predict the location where 1200, 1300, 1400 GDDs occurs; 1200 and 1400 GDD
321 were used to get an idea of the geographic variance of this climate-mediated threshold. The
322 distribution of other ecological factors—i.e., hosts associated with differential detoxification,
323 *Liriodendron tulipifera*, *Populus tremuloides*, and *Ptelea trifoliata* (Little 1971, 1976) and the
324 Batesian model *Battus philenor* for which *P. glaucus* is a mimic—were also visually compared to
325 the genetic clines (estimated center of the hybrid zone).
326

Results

Summary of ddRADseq data

One lane of Illumina HiSeq 2000 and one MiSeq run produced 346,321,096 and 12,818,490 paired reads, respectively, derived from 101 individuals. After mapping the reads to the *P. glaucus* reference genome, we called 1,835,135 SNPs, of which 28,417 loci showed a minimum of 6X coverage for 75% of individuals from "pure" parental populations and were used for F_{ST} outlier analysis. Out of these, 18,316 SNPs showed a minimum of 6X for at least 75% individuals for all populations, with an average coverage of 31.1X ± 8sd and average of 0.07% ± 0.6sd missing data across individuals. This second set was used for all other analysis. A detailed summary of sequencing and SNP filtering statistics is provided in supplementary results (Table S2).

Putative chromosome assignment of genetic markers

We used the close synteny of *Bombyx mori* and *P. glaucus* to infer the linkage group of our ddRADseq SNPs. Synteny is relatively high within Lepidoptera (>80%), and *P. glaucus* and *B. mori* share >85% of their genes in micro-syntenic blocks (Cong *et al.* 2015). For the 28,417 ddRADseq loci used for outlier analysis, 97% mapped to a *B. mori* scaffold, with 92% mapping to a linkage group (Fig S3; Data S2), spanning all 28 *B. mori* chromosomes. As expected if the chromosome assignments were correct, we found a significantly positive relationship ($r^2 = 0.69$; *P*-value < 0.0001) between *B. mori* chromosome size and number of mapped loci (Fig. S4). Thus, our ddRADseq SNPs appear to be well distributed across the genomes of *P. glaucus* and *P. canadensis.*

Population structure and diversity

The Bayesian clustering analysis of *STRUCTURE* showed the highest likelihood values and highest rate of likelihood change as the number of clusters (*K*) increased from 1 to 5 was obtained for *K*=2 (Fig. 1B and S5). Most individuals (18 *P. glaucus*, 55 *P. canadensis*) clustered into two distinct groups corresponding to the two species with more than 95% assignment probability to a species cluster (Fig 1B and Data S3). The Principal Component Analysis (PCA) revealed a similar clustering (Fig 1D).

Seven individuals showed an admixed ancestry with less than 95% of their genome coming from either species based on STRUCTURE analysis (Fig 1B, and Fig S5). Of the seven putative hybrids, two appeared to have mixed ancestry resembling early generation hybrids (<60% assignment to either cluster and nearly all diagnostic loci between the two species are heterozygous for both individuals) (see below and Fig. 1C) and five appear to be later-generation hybrids (Fig. 1B, D, and Fig. S6). Five of these putative hybrids were found at sampling location 3, previously identified as the location of the hybrid zone (Scriber *et al.* 2003) where hybrids made up 31% of the individuals from this location. However, using the species diagnostic marker, *Ldh*, for which we have a larger sample size and assuming individuals heterozygous for this marker are early generation hybrids (a reasonable assumption), we get a lower estimate of 15% (6/41) of individuals being putative hybrids in location 3, and only 3% (6/192) across the entire transect.

Genetic differentiation and detecting outlier loci

Genetic differentiation, measured by mean (weighted) (Weir & Cockerham 1984) F_{ST} between the two species (RAD1 dataset) was moderate (0.17). However, the distribution of F_{ST}

was slightly bimodal, made up largely of loci with F_{ST} value <0.25 (89%) and a small subset (3%) with estimates >0.75 (Fig 2C).

For the 28,417 SNPs we identified, of those that were at least 1 kb apart, 368 (1.3%) were significantly more differentiated in both the F_{ST}-outlier (Fig. S7A) and BayeScan (Fig. S7B) analysis and could potentially be under divergent selection (Fig 2B). As for those SNPs that were less differentiated than expected under a neutral expectation, we identified 148 (0.5%) falling outside the lower 99.9% envelope (Fig. 2A and S7A). Genetic differentiation between the two species was highly heterogeneous across the genome (Fig. 2A, S8). While most SNPs showed low levels of differentiation (F_{ST}), a small subset displayed very high F_{ST} values, with 239 fixed differences (of 28,417 loci), 301 SNPs with F_{ST} between 0.90 and 0.99 and 234 SNPs with F_{ST} between 0.80 and 0.89 (Fig 2C). These highly differentiated (F_{ST} > 0.90) SNPs that were also identified as outliers were distributed across all but five chromosomes (2, 3, 18, 20, and 26) of the *Bombyx* reference genome (Fig 2B, Fig. S8). Some of these chromosomes contained highly divergent regions spanning multiple scaffolds within the chromosome, most notably on the Z-chromosome where regions with high F_{ST} span nearly the entire chromosome. The Z-chromosome had an average F_{ST} ~4.4 times greater than the genomic background of differentiation compared to autosomes—Z-linked SNPs had an average F_{ST} of 0.36 ± 0.37 (median= 0.19), and autosomes an average F_{ST} of 0.08 ± 0.18 (median=0.0006).

Outlier SNPs were not randomly distributed across linkage groups (Fig S3). A significantly greater number of SNPs putatively under divergent selection as well as previously identified divergence hotspots (Cong et al., 2015) mapped to the Z-chromosome (*P*-value < 0.0001) (Fig. 2A, B and S3A). Among the autosomes, six chromosomes (4, 9, 10, 11, 12 and 19) were significantly enriched for outlier loci under divergent selection (*P*-value < 0.001) (Fig S3B).

Of these, chromosomes 4, 10, 12 and 19 were also significantly enriched for divergence hotspots. Contrastingly, there were twelve chromosomes (5, 7, 8, 16, 17, 18, 20, 21, 25, 26, 27, and 28) that had significantly fewer outliers than expected (*P*-value < 0.001) (Fig S3A).

Geographic cline analysis

The prediction that there will be a high degree of coincidence (similar cline center) among genetic and phenotypic clines was highly supported by our data. A majority of the outlier SNPs that were in HWE in each location and filtered to 1 SNP per 1 kb were coincident (81%; 187/230) and concordant (80%; 184/230) with the estimated median cline center of 387 km and 57 km median width (Fig 3B). This estimate places the center of the hybrid zone at ~ 43.8° N (Juneau Co., Wisconsin) (Fig 3B). The estimated center of the hybrid zone (location 3) also coincided with a steep cline in hybrid index and a peak in LD (r^2) and π (Fig 3C). Estimates of LD blocks (distance in bp between SNPs with $r^2 > 0.2$) were also largest in the hybrid zone (locations 3-5 averaging 75 kb), ranging from 36-78 kb across the transect and averaging 61 kb (se ± 6.5 kb) in size.

While there was an overall high level of coincidence and concordance across SNPs, 77% of chromosomes (17 of the 22 chromosomes containing SNPs for this analysis) had at least one locus that was discordant or non-coincident (based on non-overlapping confidence intervals with the average estimate) (Data S4). Of the 44 SNPS with non-coincident clines, the majority (40/44; 90%) were shifted in the direction of *P. glaucus*, while only 10% (4/44) were in the direction of *P. canadensis* (Fig 4). SNPs with clines shifting towards *P. glaucus* were found on thirteen different chromosomes. Of the four markers shifting towards *P. canadensis*, two were found on chromosomes 23 while one was on chromosome 1 (Z) and all of these markers were also

419 significantly wider (along with another locus on chromosome 23, they ranked among the top four
420 widest of all clines) (Fig 4).

421 Of the only 20% (47/230) of loci that were discordant, 98% (46/47) of these were
422 significantly wider and 2% (1/47) narrower than the median estimate (Data S4). The majority of
423 the significantly wider clines that were also noncoincident (24/28), were shifted to the south of
424 the transect. The only loci found to be significantly narrower than the median cline was a
425 RADseq locus located in a scaffold that did not map to a *B. mori* linkage group. However, it
426 should be noted that given hybrids are rare and make up a small portion of our samples within the
427 hybrid zone, we are only able to detect modest to large differences in cline width among clinal
428 loci.

429

430 *Evaluating ecological factors maintaining divergence*

431 The estimated center of clinal loci was compared to geographic variation in length of the
432 growing season. As expected, we found that the boundary of the growing degree day (GDD)
433 requirement to complete two generations coincided with the estimated center of the hybrid zone.
434 The cline center of the genetic loci was 387 km compared with the cline center for the GDD
435 threshold which was at 405 km, putting this climate-mediated developmental threshold ~18 km
436 north of the genetic cline center and similar in cline width (to the genetic loci)—GDD "width"
437 (variance in estimate based on ± 100 GDD) was 64 km, genetic cline width was ~ 57 km (Fig 3B
438 and 4A). Contrastingly, the range boundaries of three host plants (*Liriodendron tulipifera,*
439 *Populus tremuloides*, *Ptelea trifoliata*), for which differences in detoxification abilities exist
440 between *P. glaucus* and *P. Canadensis*, did not coincide with the hybrid zone. The east-west
441 boundary where *Ptelea trifoliata* reaches its northern limit and *Populus tremuloides* reaches its

southern limit is ~100+ kms from the hybrid zone in Wisconsin (Fig 5C, D, E). Similarly, the northern range boundary of the Batesian-mimicry model *Battus philenor* (for the female black morph of *P. glaucus*) does not coincide with the hybrid zone, as its most northern extent ends ~100 kms south of the estimated center of the hybrid zone (Fig 5B).

Discussion

The genomic landscape of divergence

Placing patterns of divergence within a genomic and geographic context can provide a better understanding of how natural selection and neutral processes drive and maintain species boundaries. Our results, based on a reduced representation of genomic variation, showed that genetic divergence and differentiation between *P. glaucus* and *P. canadensis* is widespread across the genome and is highly heterogeneous in distribution and size within and across chromosomes. A subset of chromosomes (Z-chromosome and the autosomes 4, 10, 12 and 19) showed a disproportionate amount of these differentiated and divergent regions. This pattern of loci putatively under divergent selection distributed across multiple linkage groups, with a sex chromosome harboring what appears to be a disproportionate amount of these divergent regions, is similar to what has been reported in a number of other systems including other butterflies such as *Heliconius* (Martin *et al.* 2013), mosquitoes (Fontaine *et al.* 2015), flies (Tao *et al.* 2003), mice (Teeter *et al.* 2008), and birds (Ellegren *et al.* 2012). This provides further evidence that regions of low recombination, and especially sex chromosomes, may include features that could play a disproportionate role in maintaining or facilitating divergence in the speciation process, as previously suggested (Fontaine *et al.* 2015; Mitchell *et al.* 2015).

Multiple factors can produce variable patterns of differentiation across the genome and interpreting the causes of these patterns depends in part on whether this hybrid zone is a primary contact zone (divergence-with-gene flow) or result of secondary contact (divergence in allopatry), which is not known in this system. However, there are a number of reasons why this hybrid zone is more likely the result of secondary contact. As with many hybrid zones in North America, it is unlikely that this hybrid zone survived climatic fluctuations over the last 0.6 million years of divergence between these two species (Cong *et al.* 2015). More likely, this hybrid zone formed as a result of secondary contact following the last glaciation (Barton and Hewitt, 1985). While not conclusive, the skewed ("F-shape") F_{ST} distribution we observe is similar to the pattern observed by Nosil *et al.* (2012) for allopatric *Timema cristinae* stick insect populations, in contrast to the "L-distribution" observed in parapatric *T. cristinae* populations experiencing gene flow. Some models predict that allopatric isolation should result in a more genome-wide distribution of genome divergence (Rieseberg 2001; Noor *et al.* 2001; Nosil *et al.* 2009; Yeaman & Whitlock 2011), in contrast to divergence-with-gene-flow which is predicted to result in a more clustered distribution of divergent loci. Our finding that multiple outlier loci are distributed across all but five chromosomes is thus more consistent with a process of allopatric divergence. However, discordance in the estimated divergence times of Z-linked markers suggests that gene flow has been ongoing during the speciation process (Putnam *et al.* 2007).

Regardless of the geographic origin of divergence, both selection and neutral processes can produce the heterogeneity in differentiation we observe across these genomes. In the case of the Z-chromosome, the nearly chromosome-wide high level of differentiation is likely, in part, the result of the increased genetic drift resulting from the reduced recombination rate and lower effective population size of this chromosome (Oyler-McCance *et al.* 2015). However, there are a

few regions that appear to be largely undifferentiated on the Z-chromosome. Whether these regions reflect further variation in recombination within this chromosome, variable gene flow, or are regions that are no longer in synteny (i.e., are actually located on a different chromosomes) is not known. In the case of autosomes, it seems unlikely that variation in recombination alone would explain a pattern that consists of a large number of differentiated regions, distributed widely both within and across nearly all chromosomes. However, there are a number of relatively large differentiated regions that may be the result of suppressed recombination, illustrated by the introgression of what appear to be large genomic blocks (Fig1 C, Fig S8). The importance and potential of inversions in facilitating divergence during speciation has been well documented in a number of systems (Trickett & Butlin 1994; Noor *et al.* 2001; Kirkpatrick & Barton 2006; Hoffmann & Rieseberg 2008; Love *et al.* 2016), however their role in facilitating divergence in this system remains unknown. A recombination and linkage map for *P. glaucus* would help in identifying possible inversions and more generally the role recombination is playing in producing variable patterns of differentiation in this system.

Alternatively, these patterns could reflect variable gene flow. There is now some controversy over whether patterns of heterogeneous differentiation across the genome, assessed by relative measures of differentiation such as F_{ST}, are caused by variable gene flow or processes that reduce variation in one population (e.g., selective sweeps or variation in recombination rates) (Noor & Bennett 2009; Cruickshank & Hahn 2014). Thus caution should be taken when using these measures alone. However, Cong et al. (2015) previously identified “divergence hotspots” (genes with sequence divergence at both the DNA and protein level that is greater between than within species; regions with elevated absolute divergence), as well as regions under positive selection and found little overlap (11%) between these regions. Similarly, only 10% (36/368) of

our ddRADseq SNPs (1 SNP per 1 kb) under divergent selection were closest to a gene identified as being under positive selection by Cong et al. (2015). Although these results are based on a reduced representation of the genetic variation across the genome, this lack of overlap between loci under divergent and positive selection suggest that positive selection is likely only playing a minor role in the variation in differentiation we observe. For variable introgression to be driving this pattern there would have to be ongoing gene flow. Given that all (F_{ST}) outliers in allopatric populations had relatively narrow clines in the hybrid zone and were largely coincident and concordant suggests the presence of strong reproductive barriers and very little (later generation) hybridization and gene flow. Thus, the observed variation in differentiation across the genome may reflect heterogeneous selection more than heterogeneous gene flow; although both may be playing a factor.

Evidence of ecological divergence throughout the genome

A major question remaining in speciation biology is how genomic architecture influences divergence (Nosil & Feder 2012). Resolving this question requires an understanding of how genes involved in ecological adaptation or reproductive isolation are arrayed within the genome. We find that candidate genes for divergent ecological traits (seasonal adaptations, host plant detoxification, and predator defense gland) in *P. glaucus – P. canadensis* system are distributed throughout the genome (across at least seven chromosomes), somewhat clumped and almost all associated with regions of elevated differentiation. Loci found to be significantly differentiated in allopatric populations (F_{ST} -outliers) that also exhibit steep clines across the hybrid zone are strong candidates for reproductive isolation (Harrison & Larson 2016). Interestingly, we find that nearly all ddRADseq outlier loci, spanning multiple chromosomes, exhibit steep clines across the

534 hybrid zone. That the majority of these loci were also highly coincident and concordant, suggests
535 that reproductive isolating factors are spread throughout the genome and/or selection against
536 hybrids is very strong.

537 While loci involved in reproductive isolation may be distributed across the genome, five
538 chromosomes seem to play a disproportionate role. Chromosomes 1 (Z), 4, 10, 12 and 19 harbor
539 a significantly greater number of loci potentially evolving under divergent selection as well as
540 containing divergence hotspots, and all exhibited steep clinal variation and (with the exception of
541 chromosome 12) contained candidate genes for ecologically divergent traits. The Z-chromosome
542 contains near chromosome-wide elevated differentiation (nearly 4 times the background of
543 autosomes) and is enriched with genes putatively involved in all traits that are known to be
544 ecologically divergent between these species, such as diapause (*cycle* and *period*), thermal
545 tolerance (*Ldh*) and a suppressor gene involved in female mimetic dimorphism (linked to *6-Pgd*),
546 all of which are in, or are near (< 30 kb from) highly differentiated and divergent regions.
547 Reduced recombination of this chromosome is likely facilitating some of this divergence, as a
548 region as large as ~700 kb has been estimated to be in linkage on the Z-chromosome (Winter &
549 Porter 2010). Regions of the genome experiencing low recombination such as near centromeres
550 (Ellegren *et al.* 2012), sites of inversions (Twyford & Friedman 2015) and often sex
551 chromosomes (Sæther *et al.* 2007; Teeter *et al.* 2008) have been shown to accumulate genes
552 involved in adaptive divergence. Both *period* and *cycle* of the Z-chromosome are central to the
553 circadian clock system that regulates the timing of eclosion (Blanchardon *et al.* 2001; Zhu *et al.*
554 2008) for which Cong et al. (2015) identified clusters of mutations on the surface of these
555 proteins that likely modulate differences in diapause between these species (obligate vs.
556 facultative diapause). Interestingly, while *period* was identified as a divergence hotspot by Cong

et al. (2015), we did not identify it as being under divergent selection (based on a number of RADseq loci within 1 kb of this gene). While only 11% of genes under positive selection were also divergence hotspots (as determined in Cong et al., 2015), of the few that were, four were found on the Z-chromosome, and include the gene *period*, suggesting that positive selection, not divergent selection, may be driving divergence in *period*. Thus, positive selection on traits involved in seasonal adaptation(s), including *period*, may be facilitating divergence on this chromosome (Dean *et al.* 2015).

Four autosomes (4, 10, 12 and 19) also stand-out with their enrichment of loci under potential divergent selection (with steep clinal variation) and divergence hotspot genes. Chromosome 10 contains one candidate gene for host plant detoxification (NADPH cytochrome P450 reductase; 95 kb from a RADseq locus under divergent selection) and one associated with development (ecdysone-inducible protein) that clustered closely in the center of this chromosome within what appears to be a large region of differentiation. Surprisingly, while chromosome four contains the gene *timeless*, a gene identified as a divergence hotspot (Cong et al., 2015) involved in seasonal adaptations (diapause induction) (Tauber *et al.* 2007), it was not identified as an F_{ST}-outlier in our study. In the fruit fly (*Drosophila melanogaster)* and the monarch butterfly (*Danaus plexippus*) this gene is involved in post-translational modifications of the core clock proteins and regulates differences in the incidence of diapause in response to changes in light and temperature (Reppert 2007; Tauber *et al.* 2007). Given that *P. canadensis* has an obligatory diapause and no longer uses photoperiodic cues, the lower than expected (moderate) differentiation we observed for a SNP in this gene may reflect relaxed selection on this trait in *P. canadensis* (Ryan *et al.* 2016).

579 Interestingly, a number of candidate genes for farnesyl pyrophosphate synthase (FPPS)
580 homologs hypothesized to synthesize terpenes for a predator defense gland (osmeterium) unique
581 to Papilioinidae (Cong et al., 2015) were found on chromosome 19. While the chemistry of this
582 gland is known to be affected by genetic background and varies substantially through larval
583 development, it is not known to what extent the chemical composition of this gland differs
584 between *P. glaucus* and *P. canadensis* (Frankfater *et al.* 2009). We identified at least two FPPS
585 genes in close proximity (< 2 kb) to loci under divergent selection as well as another FPPS gene
586 identified as a divergence hotspot by Cong et al. (2015). Given the large geographic distribution
587 of these species, it is possible that variation in predator communities could be driving divergence
588 in this trait, but this remains to be tested.
589 Of the few RADseq loci identified as being under balancing selection, three were in close
590 proximity to genes involved in traits that have been implicated as being under divergent or
591 balancing selection in other systems. Two of these SNPs were located on chromosome 16. This
592 includes a SNP within 2 kb of an opsin gene (opsin-1). A recent study of a North American
593 brush-footed butterfly (*Limenitis arthemis*) found clinal variation in a majority of opsin genes,
594 with many also under balancing selection (Frentiu *et al.* 2015). Interestingly the variation in the
595 long-wavelength opsin genes of *Limenitis arthemis* did not affect spectral sensitivity and instead
596 are speculated to play a role in unknown adaptive functions such as mediating responses to
597 temperature or photoperiod. The other SNP on chromosome 16 identified as a candidate for
598 balancing selection was closets (~10 kb) to a cytochrome P450 gene (CYP6k1). Hybrids of *P.*
599 *glaucus* and *P. canadensis* exhibit detoxification abilities of both parents (Mercader *et al.* 2009).
600 Whether overdominance is driving an increase in diversity of this cytochrome P450 gene (or
601 others) by increasing fitness, for example by expanding an individual's host breadth, is not

602 known and warrants further investigation. Chromosome 19 contained a SNP identified as a
603 candidate for balancing selection within 1 kb of a gene that's closest ortholog is the gene EbpIII
604 that is specifically expressed in the ejaculatory bulb and seminal fluid of *D. melanogaster*.
605 Balancing selection has been proposed as an explanation for high allelic diversity in a number of
606 seminal fluid accessory proteins (Acps), proteins that influence reproductive traits such as sperm
607 transfer, sperm storage, female receptivity, ovulation and oogenesis (Chapman 2001). While
608 these are an interesting set of candidate genes for balancing selection in this system, future
609 studies involving more fine scale genomic resolution and functional assays will help to evaluate
610 these candidates more thoroughly.
611 It should be noted that most of the distances (in bp) between these candidate genes and our
612 RADseq loci are generally less than one cM based on our estimate of a physical-to-map
613 recombination distance of ~322 kb/cM (with the genome being 376 Mb and the total map length
614 1167 cM). This value is similar to the 180 kb/cM estimated for *H. melpomene* (Jiggins *et al.*
615 2005). More importantly, these distances were also usually far less than our estimate of the
616 average size of regions in linkage disequilibrium (i.e., those with $r^2 > 0.2$), particularly those in
617 the hybrid zone that were estimated to be ~70-80 kb. Moreover, given that regions under
618 selection would likely be at the high end of such estimates, these regions may even be greater
619 than 100 kb in size.
620
621 *Climate a salient factor in hybrid zone maintenance*
622 While there are a number of ecologically divergent traits between the two species, climate
623 seems to be a salient factor in maintaining genomic divergence. The first and most prominent
624 reason is that nearly all divergent regions of the genome show a cline pattern that covaries with a

625 climatic gradient—the estimated center of the hybrid zone fell just 20 km south the estimated
626 location of the growing degree day (1300 GDD) threshold to complete two generations; all of the
627 narrowest clines (clines < 5 km in width; those putatively under the strongest selection) are even
628 closer to (~1km south of) this GDD threshold. This estimate is in agreement with previous
629 analyses based on a few species-diagnostic allozymes (Hagen 1990; Scriber *et al.* 2003).
630 Furthermore, while the genetic basis of diapause induction and termination remains elusive
631 (Denlinger 2002), some of the genes most consistently implicated as playing a role in regulating
632 diapause or eclosion phenology—*cycle, period, timeless, clock*—all mapped to chromosomes
633 identified as harboring a disproportionate amount of divergence. Third, candidate genes for other
634 ecologically divergent traits (i.e., host plant detoxification, and Batesian mimicry) (Zandt Brower
635 1957; Scriber *et al.* 1996; Mercader *et al.* 2009) do not covary with the ecological gradients that
636 would be associated with these traits, but instead largely covary with a climatic gradient. In the
637 case of three of the host plants associated with the greatest differences in detoxification abilities
638 between *P. glaucus* and *P. canadensis*—*Populus tremuloides, Ptelea trifoliata and Liriodendron*
639 *tulipifera*—the range boundary of these plants fall hundreds of kilometers south of the hybrid
640 zone. Similarly, the northern range boundary of the model *Battus philenor* is located hundreds of
641 kilometers south of the hybrid zone (at least in Wisconsin). Thus, it seems unlikely that the
642 distribution of host plants or a model for Batesian mimicry are playing a major role in
643 determining or maintain the location of this hybrid zone (at least in Wisconsin). Last, there is
644 evidence of steep clinal variation in genetic markers and a number of physiological, phenotypic,
645 and behavioral traits largely coincide with this same climate-mediated developmental threshold in
646 other portions of this hybrid zone (e.g., in Michigan and Vermont) (Scriber 2011).

Examples of climatic gradients driving clines in genetic and phenotypic variation have been well documented across a breadth of taxa, particularly in plants and insects (Knibb *et al.* 1981; Franks & Hoffmann 2012). Specifically, geographic gradients in the length of the growing season that result in changes in voltinism (number of generations in a growing season), so called "voltinism-transition-zones," often coincide with genetic and phenotypic changes (Bradford & Roff 1995; Ikten *et al.* 2011; Mallet *et al.* 2011). Other examples include adaptation to variation in the length of the growing season that has resulted in clinal variation in the body size of *Pteronemobius spp.* of cricket (Roff 1980), allele frequencies of a circadian gene in the European corn borer (*Ostrinia nubilalis*) (Levy *et al.* 2015), and the critical photoperiod of the pitcher plant mosquito (*Wyeomyia smithii*) (Bradshaw 1976). These examples demonstrate the powerful selective force that variation in the length of the growing season can impose, even in the face of gene flow, given that many of these clines exist within these species' ranges. The *P. glaucus* and *P. canadensis* hybrid zone also coincides with a voltinism-transition-zone, from bivoltine (facultative diapause of *P. glaucus*) to univoltine (obligate diapause of *P. canadensis*) and it has been hypothesized that strong climate-mediated divergent selection is acting to restrict gene flow, as genes associated with obligate and facultative diapause would be selected against south and north of the hybrid zone respectively (Scriber 2011). Our results lend further support for this hypothesis. Whether, the obligate diapause phenotype (of *P. canadensis*) is an adaptation that predates divergence of these species, or came about post divergence is still not known.

While most clines were largely coincident and concordant, the majority of discordant clines exhibited a pattern of widening ("leaning") in the direction of *P. glaucus*. This pattern can be an indication of hybrid zone movement (Moran 1981) (i.e., reflect a recent northward shift) or reflect biased gene flow towards *P. glaucus* that could be due to asymmetrical selective

pressures or assortative mating (Buggs 2007). Biased introgression into *P. glaucus* would agree with previous estimates of greater gene flow into *P. glaucus* (Zhang *et al.* 2013 and Cong *et al.* 2015). Although introgression appears biased in the direction of *P. glaucus* (southward), the most discordant of all clines were those located on chromosome 23 that were shifted north of the hybrid zone. One explanation of such a pattern could be adaptive introgression (Harrison & Larson 2016). That is, *P. glaucus* alleles are introgressing into *P. canadensis* because they confer some fitness advantage in both genetic backgrounds. Interestingly, two of the SNPs identified as being under divergent selection on this chromosome were within 1 kb of the gene *couch potato*; an amino acid polymorphism in the *couch potato* gene has been implicated in climatic adaptation in *Drosophila melanogaster* by altering reproductive diapause (Schmidt *et al.* 2008). Another SNP was closest to a juvenile hormone expoxide hydrolase, that is implicated in insect development by regulating juvenile hormone (Touhara & Prestwich 1993). Whether these genes are acting to restrict the introgression of favorable alleles on this chromosome is not known.

Potential for "coupling"

Although climate appears to be a salient factor in the maintenance of this hybrid zone, there are multiple lines of evidence that suggest intrinsic barriers might also be present in this system. One possibility is that the *P. glaucus – P. canadensis* hybrid zone is a tension zone that has become associated with a climatic gradient. In support of this idea, we observe a relatively high level of coincidence (81%) and concordance (80%) among both clinal ddRADseq loci that also coincide with a peak in *LD*. These patterns, along with what is estimated as a relatively moderate portion of the genome (1.3%) being under divergent selection have been proposed as an indication of the existence of intrinsic barriers (Bierne *et al.* 2011). Thus it is possible that

693 endogenous barriers have become coupled with an exogenous (climate-mediated) cline to impede
694 hybrid zone movement, due to either lowered population density across this ecotone (Barton &
695 Hewitt 1985) or strong extrinsic selection. Unfortunately, as is the case with many hybrid zones,
696 we do not have robust estimates of population density across the hybrid zone due to the difficulty
697 of obtaining such data. Thus, we cannot completely rule out this possibility. Such patterns do not
698 require invoking endogenous barriers, however, as a geographic-selection-gradient-model can
699 also produce clinal patterns indistinguishable from a tension zone model. Further, the large
700 amount of divergence across the genome may be explained by "genomic hitchhiking" as
701 (extrinsic) selection on many loci, such as that on the candidate genes for ecological adaptations
702 we observe, can result in an increase in genome-wide divergence, even in regions unlinked to
703 known adaptive loci (Feder *et al.* 2012).

704 Isolation during allopatry can lead to divergence in allelic variation, that when admixed,
705 such as during secondary contact, results in genotypes that are less fit due to the creation of
706 combinations of alleles that were previously hidden from selection (e.g., Dobzhansky-Muller
707 incompatibilities) (Coyne & Orr 2004). If the *P. glaucus* and *P. canadensis* hybrid zone is a
708 secondary contact zone as has been proposed (Scriber *et al.* 1991) and is partly supported by
709 results from this study, it is plausible that neutral process and selection have resulted in the build-
710 up of at least some intrinsic incompatibilities within these genomes. Intriguingly, 40 years of
711 research in this system have produced limited evidence of intrinsic incompatibilities, with
712 reductions in fitness often negligible. Tens-of-thousands of lab pairings from numerous
713 experiments have only observed hybrid offspring exhibiting slightly reduced egg viability (~9%),
714 low-to-moderate pupal mortality (~28%) and a reduction or even increase in fecundity,
715 depending on direction of cross (Scriber *et al.* 1995); there is even some evidence of heterosis,

716 translating into faster developmental rates and larger pupal size (Scriber *et al.* 2003). It is possible
717 that the fitness consequences of hybridization do not manifest until after multiple bouts of
718 hybridization—i.e., in later generation hybrids, such as has been observed in a number of
719 hybridizing species (Bierne *et al.* 2006; Pritchard & Edmands 2013). However, given that we
720 found few early generation hybrids, this suggests that most hybrids are not surviving long enough
721 for multiple bouts of hybridization to occur. Alternatively, it may be that we have been focusing
722 on the wrong phenotype. It has been observed that hybrids of these species exhibit a dramatically
723 altered phenology (emerge weeks to months prior to, or after, parentals) as well as aberrant
724 diapause phenotypes (i.e., failure to diapause even under short day conditions) (Ryan *et al.* 2016).
725 Presumably, these aberrant phenotypes could be less fit regardless of environmental conditions
726 and explain strong selection against admixed genomic backgrounds. Future studies that evaluate
727 the fitness of hybrids and parentals in the field, reared in sympatric (hybrid zone) and allopatric
728 (parental) populations, would help to identify possible intrinsic incompatibilities in this system.
729 Other non-environmental barriers, such as prezygotic barriers (i.e., mate preferences)
730 could also be preventing hybridization and aiding in the maintenance of this hybrid zone. There
731 is some evidence for mate preference, with what appears to be a preference by males of both
732 species for *P. glaucus* females (Deering & Scriber 2002). However, these preference assays have
733 assumed that males are making the choice, which has never been confirmed and there is actually
734 evidence that females in this system are making the choice when mating (Krebs 1988). Further,
735 hybrids based on our mtDNA and Z-linked species diagnostic markers have a relatively even mix
736 of both *P. glaucus* and *P. canadensis* mothers, suggesting that if there is a preference, it does not
737 appear to be substantial.

738 The finding that candidate genes putatively associated with host plant detoxification
739 overlap more with a climate-mediated developmental threshold than the geographic distribution
740 of differentially detoxified host plants is intriguing. The simplest explanation for this is that the
741 candidate genes associated with these outlier loci do not actually result in phenotypic divergence
742 in these traits (e.g., differences in detoxification ability). Another possibility is that they are
743 merely linked with other genes under exogenous (climate-mediated) or endogenous selection.
744 This may very well be the case, as the specific CYP450 genes we identified as being under
745 divergent selection are not those implicated in producing the greatest difference in
746 furanocoumarin metabolism. In fact, the RADseq loci closest to CYP6B4 genes (those shown to
747 have the greatest level of furanocoumarin metabolism) exhibited very low levels of
748 differentiation. It is also possible that the CYP6 and CYP4 genes that were implicated as being
749 under divergent selection are associated with detoxification abilities for host plants not explored
750 in this study (*P. glaucus* can feed on plants from > 34 families), or may only play a minor role.
751 An another explanation is that clines for ecological traits not involved in adaptations to climate
752 (e.g., CYP6B and CYP64 genes) have become associated with a climatic gradient as a result of
753 historical cline movement. There is some evidence that tension zones can become “attracted” and
754 move together after contact (Mallet & Barton 1989; Mallet & Turner 1998). Recently Rosser *et*
755 *al.* (2014) have shown that moving clines can also become trapped when they encounter
756 stationary clines maintained by exogenous selection along an ecotone due to an increase in per
757 locus selection resulting from disequilibrium among clines. Future studies that explore how clinal
758 patterns for a set of genome-wide SNPs vary across the multiple independent portions of this
759 hybrid zone where the distribution of host plants (as well as other ecological gradients) varies

760 should shed light on this question as well as the relative role of endogenous and exogenous
761 barriers in this system (Harrison & Larson 2016).

762 This study demonstrates the power of assessing how genomic differentiation varies along
763 a gradient of allopatry (pure parental populations) to sympatry (hybrid zone) to reveal insights
764 into the genetic architecture of reproductive isolation. With this approach we show that clinal
765 variation in a climate-mediated developmental threshold can act to maintaining divergence
766 between the genomes of hybridizing species. We also find evidence that previously unrecognized
767 endogenous barriers may be contributing to reproductive isolation in this system, providing some
768 support for the coupling hypothesis that hybrid zones with steep clinal variation associated with
769 an ecotone are the result of a coupling of endogenous and exogenous barriers. Last, similar to
770 what is now a growing number of systems, we find evidence that the Z-chromosome may be
771 playing a disproportionate role in the process of speciation.

Acknowledgements

Butterfly collections were undertaken by Sean F. Ryan, Mark Scriber, Jason Dzurisin, and Sarah Richman. Many thanks to Benjamin Clifford, Meredith Doellman and Jacqueline Lopez and Melissa Stephens in the Notre Dame Genomics & Bioinformatics Core Facility for their advice in various aspects of experimental design, NGS data collection and analysis. This work was sponsored by NSF grants DEB-0918879 to JMS and DEB-0919147 to JJH and a grant to SFR from the Environmental Change Initiative at the University of Notre Dame.

References

Aboagye-Antwi F, Alhafez N, Weedall GD *et al.* (2015) Experimental Swap of Anopheles gambiae 's Assortative Mating Preferences Demonstrates Key Role of X-Chromosome Divergence Island in Incipient Sympatric Speciation. *PLOS Genet*, **11**, e1005141.

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Barton NH, Hewitt and GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619–1626.

Bierne N, Bonhomme F, Boudry P, Szulkin M, David P (2006) Fitness landscapes support the dominance theory of post-zygotic isolation in the mussels Mytilus edulis and M. galloprovincialis. *Proceedings of the Royal Society of London B: Biological Sciences*, **273**, 1253–1260.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Blanchardon E, Grima B, Klarsfeld A *et al.* (2001) Defining the role of Drosophila lateral neurons in the control of circadian rhythms in motor activity and eclosion by targeted

genetic ablation and PERIOD protein overexpression. *European Journal of Neuroscience*, **13**, 871–888.

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

Bradford MJ, Roff DA (1995) Genetic and Phenotypic Sources of Life History Variation along a Cline in Voltinism in the Cricket Allonemobius socius. *Oecologia*, **103**, 319–326.

Bradshaw WE (1976) Geography of photoperiodic response in diapausing mosquito. *Nature*, **262**, 384–386.

Buggs RJA (2007) Empirical study of hybrid zone movement. *Heredity*, **99**, 301–312.

Chapman T (2001) Seminal fluid-mediated fitness traits in Drosophila. *Heredity*, **87**, 511–521.

Cong Q, Borek D, Otwinowski Z, Grishin NV (2015) Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. *Cell Reports*, **10**, 910–919.

Coyne JA, Orr HA (1989) Patterns of Speciation in Drosophila. *Evolution*, **43**, 362–381.

Coyne JA, Orr AH (2004) *Speciation*. Sinauer Associates.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dean R, Harrison PW, Wright AE, Zimmer F, Mank JE (2015) Positive Selection Underlies Faster-Z Evolution of Gene Expression in Birds. *Molecular Biology and Evolution*, **32**, 2646–2656.

822 Deering MD, Scriber MJ (2002) Field bioassays show heterospecific mating preference
823 asymmetry between hybridizing North American Papilio butterfly species (Lepidoptera:
824 Papilionidae). *Journal of Ethology*, **20**, 25–33.
825 Denlinger DL (2002) Regulation of Diapause. *Annual Review of Entomology*, **47**, 93–122.
826 Dent EA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for
827 visualizing STRUCTURE output and implementing the Evanno method. *Conservation*
828 *Genetics Resources*, **4**, 359–361.
829 DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and
830 genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.
831 Derryberry EP, Derryberry GE, Maley JM, Brumfield RT (2014) hzar: hybrid zone analysis using
832 an R software package. *Molecular Ecology Resources*, **14**, 652–663.
833 Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in
834 Ficedula flycatchers. *Nature*, **491**, 756–760.
835 Endler JA (1977) *Geographic variation, speciation, and clines*. Princeton Univ. Press., Princeton,
836 NJ.
837 Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the
838 software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
839 Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform
840 population genetics analyses under Linux and Windows. *Molecular Ecology Resources*,
841 **10**, 564–567.
842 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus
843 genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

844 Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence
845 and genome hitchhiking. *Philosophical Transactions of the Royal Society B: Biological*
846 *Sciences*, **367**, 461–474.
847 Flaxman SM, Feder JL, Nosil P (2013) Genetic Hitchhiking and the Dynamic Buildup of
848 Genomic Divergence During Speciation with Gene Flow. *Evolution*, **67**, 2577–2591.
849 Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for
850 Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–
851 993.
852 Fontaine MC, Pease JB, Steele A *et al.* (2015) Mosquito genomics. Extensive introgression in a
853 malaria vector species complex revealed by phylogenomics. *Science (New York, N.Y.)*,
854 **347**, 1258524.
855 Frankfater C, Tellez MR, Slattery M (2009) The scent of alarm: ontogenetic and genetic variation
856 in the osmeterial gland chemistry of Papilio glaucus (Papilionidae) caterpillars.
857 *Chemoecology*, **19**, 81–96.
858 Franks SJ, Hoffmann AA (2012) Genetics of Climate Change Adaptation. *Annual Review of*
859 *Genetics*, **46**, 185–208.
860 Frentiu FD, Yuan F, Savage WK *et al.* (2015) Opsin clines in butterflies suggest novel roles for
861 insect photopigments. *Molecular Biology and Evolution*, **32**, 368–379.
862 Gompert Z, Buerkle CA (2010) introgress: a software package for mapping components of
863 isolation in hybrids. *Molecular Ecology Resources*, **10**, 378–384.
864 Gompert Z, Lucas LK, Nice CC *et al.* (2012) Genomic Regions with a History of Divergent
865 Selection Affect Fitness of Hybrids Between Two Butterfly Species. *Evolution*, **66**, 2167–
866 2181.

867 Hagen RH (1990) Population Structure and Host Use in Hybridizing Subspecies of Papilio glacus
868 (Lepidoptera: Papilionidae). *Evolution*, **44**, 1914–1930.
869 Hagen RH, Scriber JM (1989) Sex-Linked Diapause, Color, and Allozyme Loci in Papilio
870 glaucus: Linkage Analysis and Significance in a Hybrid Zone. *Journal of Heredity*, **80**,
871 179–185.
872 Hamilton JA, Lexer C, Aitken SN (2013) Genomic and phenotypic architecture of a spruce
873 hybrid zone (Picea sitchensis × P. glauca). *Molecular Ecology*, **22**, 827–841.
874 Hamilton JA, Torre ARD la, Aitken SN (2014) Fine-scale environmental variation contributes to
875 introgression in a three-species spruce hybrid complex. *Tree Genetics & Genomes*, **11**,
876 817.
877 Harrison RG, Larson EL (2016) Heterogeneous genome divergence, differential introgression,
878 and the origin and structure of hybrid zones. *Molecular Ecology*, **25**, 2454–2466.
879 Hewitt GM (1988) Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology*
880 *& Evolution*, **3**, 158–167.
881 Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite
882 populations. *Theoretical Population Biology*, **33**, 54–78.
883 Hoffmann AA, Rieseberg LH (2008) Revisiting the Impact of Inversions in Evolution: From
884 Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual review*
885 *of ecology, evolution, and systematics*, **39**, 21–42.
886 Hung CF, Berenbaum MR, Schuler MA (1997) Isolation and characterization of CYP6B4, a
887 furanocoumarin-inducible cytochrome P450 from a polyphagous caterpillar
888 (Lepidoptera:papilionidae). *Insect Biochemistry and Molecular Biology*, **27**, 377–385.

889 Ikten C, Skoda SR, Hunt TE, Molina-Ochoa J, Foster JE (2011) Genetic Variation and
890 Inheritance of Diapause Induction in Two Distinct Voltine Ecotypes of Ostrinia nubilalis
891 (Lepidoptera: Crambidae). *Annals of the Entomological Society of America*, **104**, 567–
892 575.
893 Jiggins CD, Mavarez J, Beltrán M *et al.* (2005) A Genetic Linkage Map of the Mimetic Butterfly
894 Heliconius melpomene. *Genetics*, **171**, 557–570.
895 Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.
896 *Bioinformatics (Oxford, England)*, **27**, 3070–3071.
897 Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in
898 threespine sticklebacks. *Nature*, **484**, 55–61.
899 Kirkpatrick M, Barton N (2006) Chromosome Inversions, Local Adaptation and Speciation.
900 *Genetics*, **173**, 419–434.
901 Knibb WR, Oakeshott JG, Gibson JB (1981) Chromosome Inversion Polymorphisms in
902 DROSOPHILA MELANOGASTER. I. Latitudinal Clines and Associations between
903 Inversions in Australasian Populations. *Genetics*, **98**, 833–847.
904 Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program
905 for identifying clustering modes and packaging population structure inferences across K.
906 *Molecular Ecology Resources*, **15**, 1179–1191.
907 Krebs RA (1988) The mating behavior of Papilio glaucus (Papilionidae). *Journal of Research on*
908 *the Lepidoptera*, **26**, 27–31.
909 Kukal O, Ayres MP, Scriber JM (1991) Cold tolerance of the pupae in relation to the distribution
910 of swallowtail butterflies. *Canadian Journal of Zoology*, **69**, 3028–3037.

911 Kunte K, Shea C, Aardema ML *et al.* (2011) Sex Chromosome Mosaicism and Hybrid Speciation
912 among Tiger Swallowtail Butterflies. *PLoS Genet*, **7**, e1002274.
913 de Lafontaine G, Prunier J, Gérardi S, Bousquet J (2015) Tracking the progression of speciation:
914 variable patterns of introgression across the genome provide insights on the species
915 delimitation between progenitor–derivative spruces (Picea mariana × P. rubens).
916 *Molecular Ecology*, **24**, 5229–5247.
917 Lehnert MS, Scriber JM, Gerard PD, Emmel TC (2012) The "Converse to Bergmann's Rule" in
918 Tiger Swallowtail Butterflies: Boundaries of Species and Subspecies Wing Traits are
919 Independent of Thermal and Host-Plant Induction. *American Entomologist*, **58**, 156.
920 Levy RC, Kozak GM, Wadsworth CB, Coates BS, Dopman EB (2015) Explaining the sawtooth:
921 latitudinal periodicity in a circadian gene correlates with shifts in generation number.
922 *Journal of Evolutionary Biology*, **28**, 40–53.
923 Lexer C, Joseph JA, Loo M van *et al.* (2010) Genomic Admixture Analysis in European Populus
924 spp. Reveals Unexpected Patterns of Reproductive Isolation and Mating. *Genetics*, **186**,
925 699–712.
926 Li X, Baudry J, Berenbaum MR, Schuler MA (2004) Structural and functional divergence of
927 insect CYP6B proteins: From specialist to generalist cytochrome P450. *Proceedings of*
928 *the National Academy of Sciences of the United States of America*, **101**, 2939–2944.
929 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
930 *Bioinformatics (Oxford, England)*, **25**, 1754–1760.
931 Li W, Petersen RA, Schuler MA, Berenbaum MR (2002) CYP6B cytochrome p450
932 monooxygenases from Papilio canadensis and Papilio glaucus: potential contributions of
933 sequence divergence to host plant associations. *Insect Molecular Biology*, **11**, 543–551.

934 Li W, Schuler MA, Berenbaum MR (2003) Diversification of furanocoumarin-metabolizing
935 cytochrome P450 monooxygenases in two papilionids: Specificity and substrate
936 encounter rate. *Proceedings of the National Academy of Sciences of the United States of*
937 *America*, **100 Suppl 2**, 14593–14598.
938 Lima TG (2014) Higher levels of sex chromosome heteromorphism are associated with markedly
939 stronger reproductive isolation. *Nature Communications*, **5**.
940 Lindroth RL, Scriber JM, Hsia MTS (1988) Chemical Ecology of the Tiger Swallowtail:
941 Mediation of Host Use by Phenolic Glycosides. *Ecology*, **69**, 814–822.
942 Little ELJ (1971) Atlas of United States trees, volume 1, conifers and important hardwoods: U.S.
943 Department of Agriculture Miscellaneous Publication. , 9 p., 200 maps.
944 Little ELJ (1976) Atlas of United States trees, volume 3, minor Western hardwoods: U.S.
945 Department of Agriculture Miscellaneous Publication. , 13 p., 290 maps.
946 Love RR, Steele AM, Coulibaly MB *et al.* (2016) Chromosomal inversions and ecotypic
947 differentiation in Anopheles gambiae: the perspective from whole-genome sequencing.
948 *Molecular Ecology*.
949 Luebke HJ, Scriber JM, Yandell BS (1988) Use of multivariant discriminant analysis of male
950 wing morphometrics to delineate a hybrid zone for Papilio glaucus glaucus and P. g.
951 canadensis in Wisconsin. *American Midland Naturalist*, 366–379.
952 Malenfant RM, Coltman DW, Davis CS (2015) Design of a 9K illumina BeadChip for polar
953 bears (Ursus maritimus) from RAD and transcriptome sequencing. *Molecular Ecology*
954 *Resources*, **15**, 587–600.
955 Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**,
956 229–237.

957 Mallet J, Barton N (1989) Inference from Clines Stabilized by Frequency-Dependent Selection.
958 *Genetics*, **122**, 967–976.
959 Mallet J, Turner JRG (1998) Biotic drift or the shifting balance: did forest islands drive the
960 diversity of warningly coloured butterflies? In:, pp. 262–280. Oxford Univ. Press, Oxford,
961 U.K.
962 Mallet J, Wynne IR, Thomas CD (2011) Hybridisation and climate change: brown argus
963 butterflies in Britain (Polyommatus subgenus Aricia). *Insect Conservation and Diversity*,
964 **4**, 192–199.
965 Marroni F, Pinosio S, Zaina G *et al.* (2011) Nucleotide diversity and linkage disequilibrium in
966 Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics & Genomes*,
967 **7**, 1011–1023.
968 Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation
969 with gene flow in Heliconius butterflies. *Genome Research*, **23**, 1817–1828.
970 McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce
971 framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**,
972 1297–1303.
973 Mercader RJ, Aardema ML, Scriber JM (2009) Hybridization leads to host-use divergence in a
974 polyphagous butterfly sibling species pair. *Oecologia*, **158**, 651–662.
975 Mercader RJ, Scriber JM (2008) Asymmetrical thermal constraints on the parapatric species
976 boundaries of two widespread generalist butterflies. *Ecological Entomology*, **33**, 537–545.
977 Mitchell SN, Kakani EG, South A *et al.* (2015) Evolution of sexual traits influencing vectorial
978 capacity in anopheline mosquitoes. *Science*, **347**, 985–988.
979 Moran C (1981) Genetic demarcation of geographical distribution by hybrid zones. , 67–73.

980 Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction
981 endonucleases. *Proceedings of the National Academy of Sciences of the United States of*
982 *America*, **76**, 5269–5273.
983 Noor M a. F, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the
984 role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.
985 Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the
986 reproductive isolation of species. *Proceedings of the National Academy of Sciences of the*
987 *United States of America*, **98**, 12084–12088.
988 Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Phil.*
989 *Trans. R. Soc. B*, **367**, 332–342.
990 Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic
991 divergence. *Molecular Ecology*, **18**, 375–402.
992 Nosil P, Gompert Z, Farkas TE *et al.* (2012) Genomic consequences of multiple speciation
993 processes in a stick insect. *Proc. R. Soc. B*, rspb20120813.
994 Oyler-McCance SJ, Cornman RS, Jones KL, Fike JA (2015) Z chromosome divergence,
995 polymorphism and relative effective population size in a genus of lekking birds. *Heredity*,
996 **115**, 452–459.
997 Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an
998 adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.
999 Presgraves DC (2008) Sex chromosomes and speciation in Drosophila. *Trends in Genetics*, **24**,
1000 336–343.
1001 Prism Climate Group, Oregon State University, http://prism.oregonstate.edu, created 4 Feb 2014.

1002 Pritchard VL, Edmands S (2013) The Genomic Trajectory of Hybrid Swarms: Outcomes of
1003 Repeated Crosses Between Populations of Tigriopus Californicus. *Evolution*, **67**, 774–
1004 791.
1005 Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus
1006 Genotype Data. *Genetics*, **155**, 945–959.
1007 Putnam AS, Scriber JM, Andolfatto P (2007) Discordant Divergence Times Among Z-
1008 Chromosome Regions Between Two Ecologically Distinct Swallowtail Butterfly Species.
1009 *Evolution*, **61**, 912–927.
1010 R Core Team (2013) R: A language and environment for statistical computing. *R Foundation for*
1011 *Statistical Computing*.
1012 Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by
1013 geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.
1014 Reppert SM (2007) The Ancestral Circadian Clock of Monarch Butterflies: Role in Time-
1015 compensated Sun Compass Orientation. *Cold Spring Harbor Symposia on Quantitative*
1016 *Biology*, **72**, 113–118.
1017 Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology &*
1018 *Evolution*, **16**, 351–358.
1019 Ritland DB, Scriber JM (1985) Larval developmental rates of three putative subspecies of tiger
1020 swallowtail butterflies, Papilio glaucus, and their hybrids in relation to temperature.
1021 *Oecologia*, **65**, 185–193.
1022 Roff D (1980) Optimizing development time in a seasonal environment: The “ups and downs” of
1023 clinal variation. *Oecologia*, **45**, 202–208.

1024 Rosser N, Dasmahapatra KK, Mallet J (2014) Stable Heliconius butterfly hybrid zones are
1025 correlated with a local rainfall peak at the edge of the Amazon basin. *Evolution;*
1026 *International Journal of Organic Evolution*, **68**, 3470–3484.
1027 Ryan SF, Valella P, Thivierge G, Aardema ML, Scriber JM (2016) The role of latitudinal, genetic
1028 and temperature variation in the induction of diapause of Papilio glaucus (Lepidoptera:
1029 Papilionidae). *Insect Science*.
1030 Sæther SA, Sætre G-P, Borge T *et al.* (2007) Sex Chromosome-Linked Species Recognition and
1031 Evolution of Reproductive Isolation in Flycatchers. *Science*, **318**, 95–97.
1032 Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*. Cold
1033 Spring Harbor Laboratory Press, Cold Springs Harbor, New York.
1034 Schmidt PS, Zhu C-T, Das J *et al.* (2008) An amino acid polymorphism in the couch potato gene
1035 forms the basis for climatic adaptation in Drosophila melanogaster. *Proceedings of the*
1036 *National Academy of Sciences*, **105**, 16207–16211.
1037 Scriber JM (1990) Interaction of introgression from Papilio glaucus canadensis and diapause in
1038 producing "spring form" Eastern tiger swallowtail butterflies, P. glaucus. *Great Lakes*
1039 *Entomologist*, 127–138.
1040 Scriber JM (2002) Evolution of insect-plant relationships: chemical constraints, coadaptation, and
1041 concordance of insect/plant traits. *Entomologia Experimentalis et Applicata*, **104**, 217–
1042 235.
1043 Scriber JM (2011) Impacts of climate warming on hybrid zone movement: Geographically
1044 diffuse and biologically porous "species borders." *Insect Science*, **18**, 121–159.

1045 Scriber JM, Elliot B, Maher E, McGuire M, Niblack M (2014) Adaptations to "Thermal Time"
1046 Constraints in Papilio: Latitudinal and Local Size Clines Differ in Response to Regional
1047 Climate Change. *Insects*, **5**, 199–226.
1048 Scriber JM, Giebink BL, Snider D (1991) Reciprocal Latitudinal Clines in Oviposition Behavior
1049 of Papilio glaucus and P. canadensis across the Great Lakes Hybrid Zone: Possible Sex-
1050 Linkage of Oviposition Preferences. *Oecologia*, **87**, 360–368.
1051 Scriber JM, Hagen RH, Lederhouse RC (1996) Genetics of Mimicry in the Tiger Swallowtail
1052 Butterflies, Papilio glaucus and P. canadensis (Lepidoptera: Papilionidae). *Evolution*, **50**,
1053 222–236.
1054 Scriber JM, Lederhouse RC (1992) The thermal environment as a resource dictating geographic
1055 patterns of feeding specialization of insect herbivores. In: *Resource Distribution and and*
1056 *Animal-Plant Interactions*, pp. 429–466. Academic Press, New York.
1057 Scriber JM, Lederhouse RC, Dowell RV (1995) Hybridization studies with North American
1058 swallowtails. In: *Swallowtail Butterflies: Their Ecology and Evolutionary Biology*, pp.
1059 367–391. University of Chicago Press, Chicago, Illinois, USA.
1060 Scriber JM, Lindroth RL, Nitao J (1989) Differential toxicity of a phenolic glycoside from
1061 quaking aspen to Papilio glaucus butterfly subspecies, hybrids and backcrosses.
1062 *Oecologia*, **81**, 186–191.
1063 Scriber JM, Ording GJ, Mercader RJ (2008) Hybrid introgression and parapatric speciation in a
1064 hybrid zone. In: *Specialization, Speciation, and Radiation: the Evolutionary Biology of*
1065 *Herbivorous Insects*, pp. 68–87. Univ.California Press, Berkeley.
1066 Scriber JM, Stump A, Deering M (2003) *Ecology and Evolution Taking Flight: Butterflies as*
1067 *Model Study Systems*. University of Chicago Press, Chicago, Illinois, USA.

1068 Shiryev SA, Papadopoulos JS, Schäffer AA, Agarwala R (2007) Improved BLAST searches
1069 using longer words for protein seeding. *Bioinformatics (Oxford, England)*, **23**, 2949–
1070 2951.
1071 Sperling FA (1993) Heredity - Abstract of article: Mitochondrial DNA variation and Haldane's
1072 rule in the Papilio glaucus and P. troilus species groups. *Heredity*, **71**, 227–233.
1073 Szymura JM, Barton NH (1986) Genetic Analysis of a Hybrid Zone Between the Fire-Bellied
1074 Toads, Bombina bombina and B. variegata, Near Cracow in Southern Poland. *Evolution*,
1075 **40**, 1141–1159.
1076 Szymura JM, Barton NH (1991) The Genetic Structure of the Hybrid Zone between the Fire-
1077 Bellied Toads Bombina bombina and B. variegata: Comparisons between Transects and
1078 between Loci. *Evolution*, **45**, 237–261.
1079 Tao Y, Chen S, Hartl DL, Laurie CC (2003) Genetic dissection of hybrid incompatibilities
1080 between Drosophila simulans and D. mauritiana. I. Differential accumulation of hybrid
1081 male sterility effects on the X and autosomes. *Genetics*, **164**, 1383–1397.
1082 Tauber E, Zordan M, Sandrelli F *et al.* (2007) Natural Selection Favors a Newly Derived timeless
1083 Allele in Drosophila melanogaster. *Science*, **316**, 1895–1898.
1084 Taylor SA, Larson EL, Harrison RG (2015) Hybrid Zones: Windows on Climate Change. *Trends*
1085 *in ecology & evolution*, **30**, 398–406.
1086 Taylor SA, White TA, Hochachka WM *et al.* (2014) Climate-Mediated Movement of an Avian
1087 Hybrid Zone. *Current Biology*, **24**, 671–676.
1088 Teeter KC, Payseur BA, Harris LW *et al.* (2008) Genome-wide patterns of gene flow across a
1089 house mouse hybrid zone. *Genome Research*, **18**, 67–76.

1090 Teeter KC, Thibodeau LM, Gompert Z *et al.* (2010) The variable genomic architecture of
1091 isolation between hybridizing species of house mice. *Evolution; International Journal of*
1092 *Organic Evolution*, **64**, 472–485.
1093 Touhara K, Prestwich GD (1993) Juvenile hormone epoxide hydrolase. Photoaffinity labeling,
1094 purification, and characterization from tobacco hornworm eggs. *Journal of Biological*
1095 *Chemistry*, **268**, 19604–19609.
1096 Trickett AJ, Butlin RK (1994) Recombination suppressors and the evolution of new species.
1097 *Heredity*, **73 (Pt 4)**, 339–345.
1098 Twyford AD, Friedman J (2015) Adaptive divergence in the monkey flower Mimulus guttatus is
1099 maintained by a chromosomal inversion. *Evolution*, **69**, 1476–1486.
1100 de Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods
1101 against more complex models: when and how much should we trust them? *Molecular*
1102 *Ecology*, **23**, 2006–2019.
1103 Wang J, Xia Q, He X *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics.
1104 *Nucleic Acids Research*, **33**, D399–D402.
1105 Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure.
1106 *Evolution*, **38**, 1358–1370.
1107 Whitlock MC (1992) Temporal Fluctuations in Demographic Parameters and the Genetic
1108 Variance among Populations. *Evolution*, **46**, 608–615.
1109 Winter CB, Porter AH (2010) AFLP linkage map of hybridizing swallowtail butterflies, Papilio
1110 glaucus and Papilio canadensis. *The Journal of Heredity*, **101**, 83–90.
1111 Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection
1112 balance. *Evolution; International Journal of Organic Evolution*, **65**, 1897–1911.

1113 Zandt Brower JV (1957) Experimental Studies of Mimicry in some North American Butterflies.
1114 *Nature*, **180**, 444–444.
1115 Zhang W, Kunte K, Kronforst MR (2013) Genome-wide characterization of adaptation and
1116 speciation in tiger swallowtail butterflies using de novo transcriptome assemblies.
1117 *Genome Biology and Evolution*, **5**, 1233–1245.
1118 Zhu H, Sauman I, Yuan Q *et al.* (2008) Cryptochromes Define a Novel Circadian Clock
1119 Mechanism in Monarch Butterflies That May Underlie Sun Compass Navigation. *PLOS*
1120 *Biol*, **6**, e4.
1121
1122

Data accessibility:

Demultiplexed ddRAD sequence data were deposited in NCBI's SRA archives under BioProject ID PRJNA388284. Scripts and associated files for processing and analyzing all data in this manuscript were deposited in DRYAD (DOI: http://dx.doi.org/10.5061/dryad.t9221).

Author Contributions

SFR, JMS, MCF, MEP and JJH designed this research. SFR did field collections and all lab work. SFR performed all analysis, with help from MCF and SO. SO and MCF developed software and scripts for a portion of the genomic analyses. SFR, MCF, JJH, JMS and MEP were involved in writing and editing this manuscript.

Tables:

Table 1. Sampling locations along the transect. The transect is measured as the distance from Mason Co. using the 'Haversine equation' and 90° N as longitude and 6,371 km for the radius of the earth.

Location	County	Latitude	Distance (km)	n_{COI}	n_{Ldh}	n_{ddRAD}
1	Mason	40.38352	0	24	10	11
	Putnam	41.22890	94	1	-	-
	Rock Island	41.528373	127	-	-	-
2	Green	42.64672	252	10	10	2
	Iowa	43.05780	297	39	15	12
	Dane	43.076369	299	-	-	-
	Richland	43.37849	333	4	-	-
	Sauk	43.42960	339	11	8	2
3	Columbia	43.60436	358	7	7	2
	Green Lake	43.838402	384	-	-	-
	Marquette	43.89078	390	5	5	3
	Monroe	43.94569	396	1	-	-
	Adams	44.03663	406	14	13	9
	Juneau	44.16089	420	19	16	5
4	Jackson	44.23612	428	16	15	9
5	Wood	44.44880	452	16	15	6
	Portage	44.47409	455	1	1	1
	Clark	44.69498	479	21	15	8
	Marathon	44.89816	502	7	-	-
6	Taylor	45.16705	532	30	10	6
	Lincoln	45.36137	554	19	11	3
7	Oneida	45.57978	578	9	8	1
	Price	45.64626	585	14	14	10
	Vilas	45.95206	619	32	10	1
8	Iron	46.26355	654	5	-	-
	Ontonagon	46.51145	681	9	9	10
Total				314	192	101

n_{COI}: # of individuals genotyped for species-diagnostic SNP in *COI* gene
n_{Ldh}: # of individuals genotyped for species-diagnostic SNP in *Ldh* gene
n_{RADseq}: # of individuals sequenced with ddRADseq

Figures:

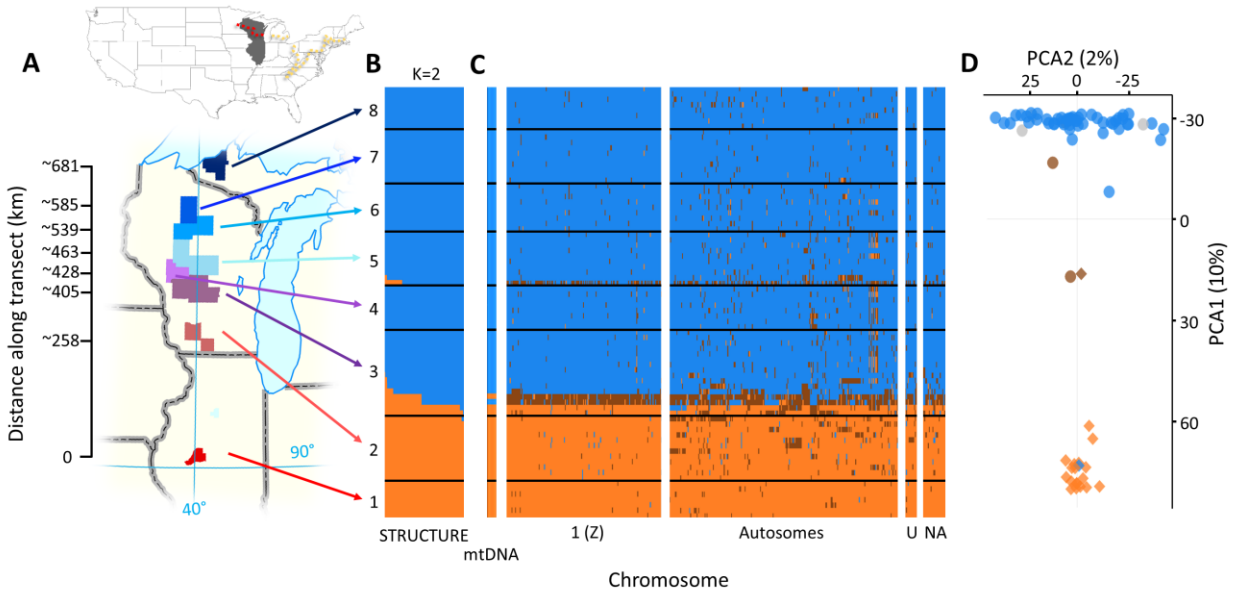


Figure 1: The genomic landscape of divergence: evidence of ecological selection across the genome. (A) Map showing sampling locations for each population; counties with the same color were pooled for analysis. (B) Bayesian clustering of ancestry proportion to K=2. (C) Maximum likelihood estimates of hybrid index of one mtDNA marker and 471 SNPs. Each row is an individual and column a genetic marker. Color of blocks indicates genotype or mtDNA haplotype (for the mtDNA) —orange and blue blocks represent homozygous genotype for *P. glaucus* and *P. canadensis* respectively and brown blocks indicate heterozygote genotypes. Black horizontal lines separate individuals by location. (D) Principal Components Analysis showing the individual scores for the first two PCs. The proportion of variance explained by each axis is indicated between brackets. Color indicates whether individuals had the *P. glaucus* (orange), *P. canadensis* (blue), or were heterozygous (brown) for a species diagnostic SNP in *Ldh* and shape indicates an individual's mtDNA haplotype (*P. glaucus*: diamonds, *P. canadensis*: circles); grey circles were individuals not genotyped for *Ldh*.

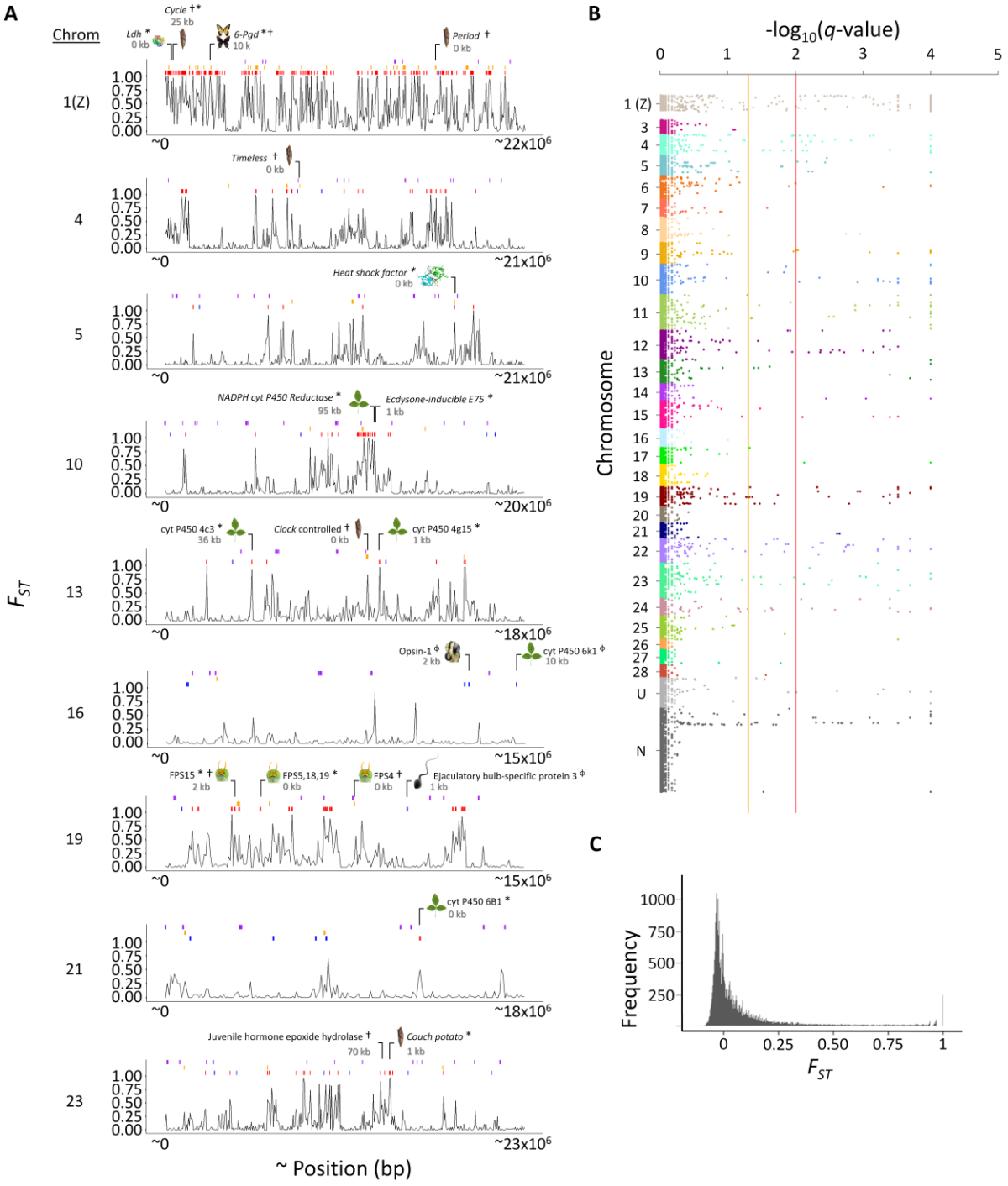
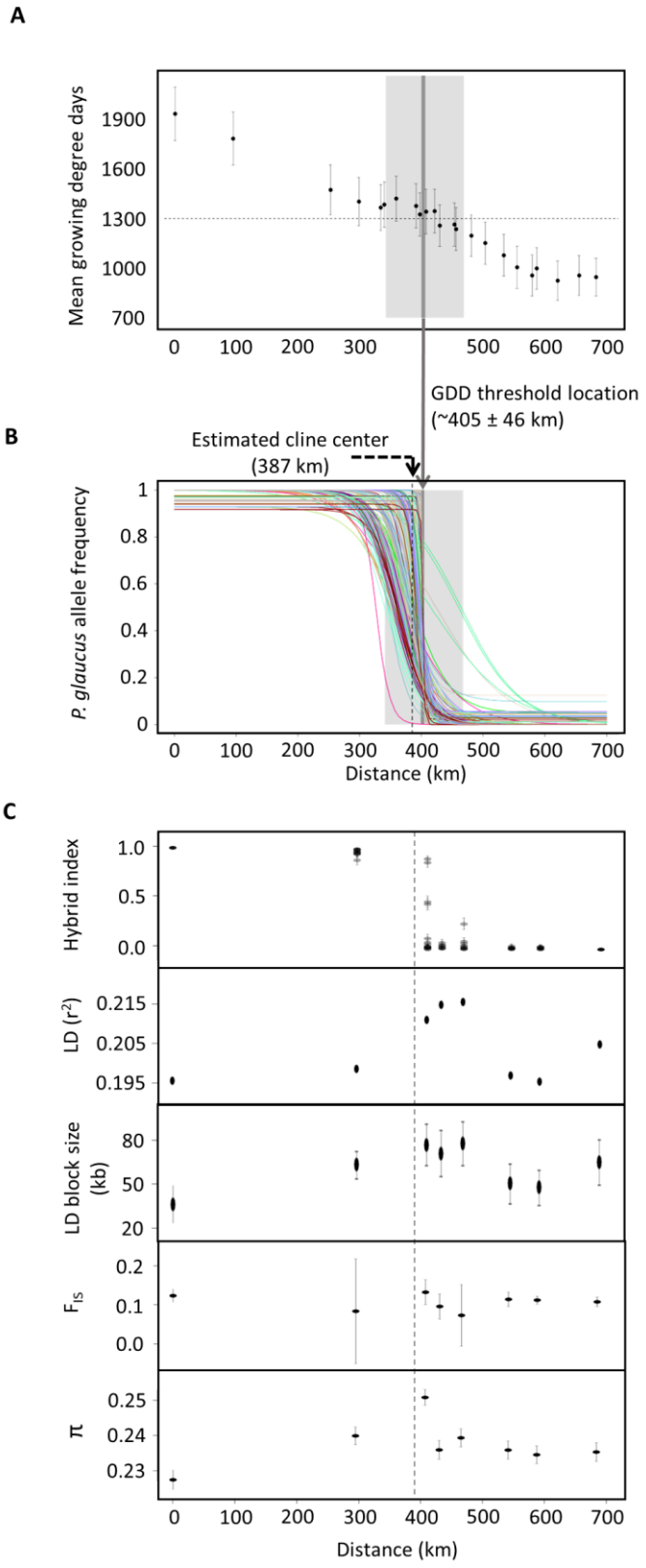


Figure 2: Genomic landscape of divergence between *P. glaucus* and *P. canadensis*. (A) Differentiation (F_{ST} averaged in non-overlapping 1 kb windows) within linkage groups containing candidate genes for ecologically divergent traits (position of candidate genes shown at the top of each linkage group along with distance from closest ddRADseq locus) ; note that these distances are less than one cM and the average LD decay of ~80 kb for $r^2 > 0.2$. *, ϕ, indicate

genes associated with loci under divergent and balancing selection, respectively. † indicate genes that were identified as divergence hotspots by Cong et al. (2015). Red and blue bars indicate loci under divergent and balancing selection, respectively (F_{ST} -outlier); orange and purple bars represent loci within 1 kb of genes identified as being divergence hotspots or under positive selection by Cong et al. (2015), respectively. (B) Manhattan plot of $\log_{10}$(*Q-values*) from BayeScan analysis by linkage group. Colored lines indicate threshold for False Discovery Rate (orange = 0.05; red = 0.01). Chromosome "U" and "N" represent loci that did not have linkage group information or did not map to a *B. mori* scaffold, respectively. (C) Histogram of F_{ST} values for all 28,417 ddRADseq loci.



1176

Figure 3: The geographic landscape of divergence. (A) 32 year average length of the growing season (mean growing degree days) across the *P. glaucus* / *P. canadensis* hybrid zone. Error bars are standard deviation. The dark grey shaded region illustrates the (1300) minimum growing degree day requirement (light grey regions ± 100 GDD) for *P. glaucus* to complete two generations (Ritland and Scriber 1985). (B) Maximum likelihood cline for 230 outlier SNPs (in HWE in each location and filtered to 1 SNP per 1 kb). Allele frequency clines that share the same color, map to the same (*B. mori*) chromosome (See Fig 2B). The shaded region illustrates the threshold required for two generations (panel A). (C) Hybrid index, linkage disequilibrium (LD), LD block size (average distance in bp between SNPs with $r^2 > 0.2$), fixation index (F_{IS}) and nucleotide diversity (π) across the *P. glaucus* / *P. canadensis* hybrid zone. Hybrid index ranges from 0 (*P. glaucus*-like) to 1 (*P. canadensis*-like). Error bars represent 95% confidence intervals. Estimates for LD were created by randomly subsampling seven individuals (smallest n of all locations) from each location 100 times. Vertical dashed lines represents the estimated center of the hybrid zone based on genetic clines.

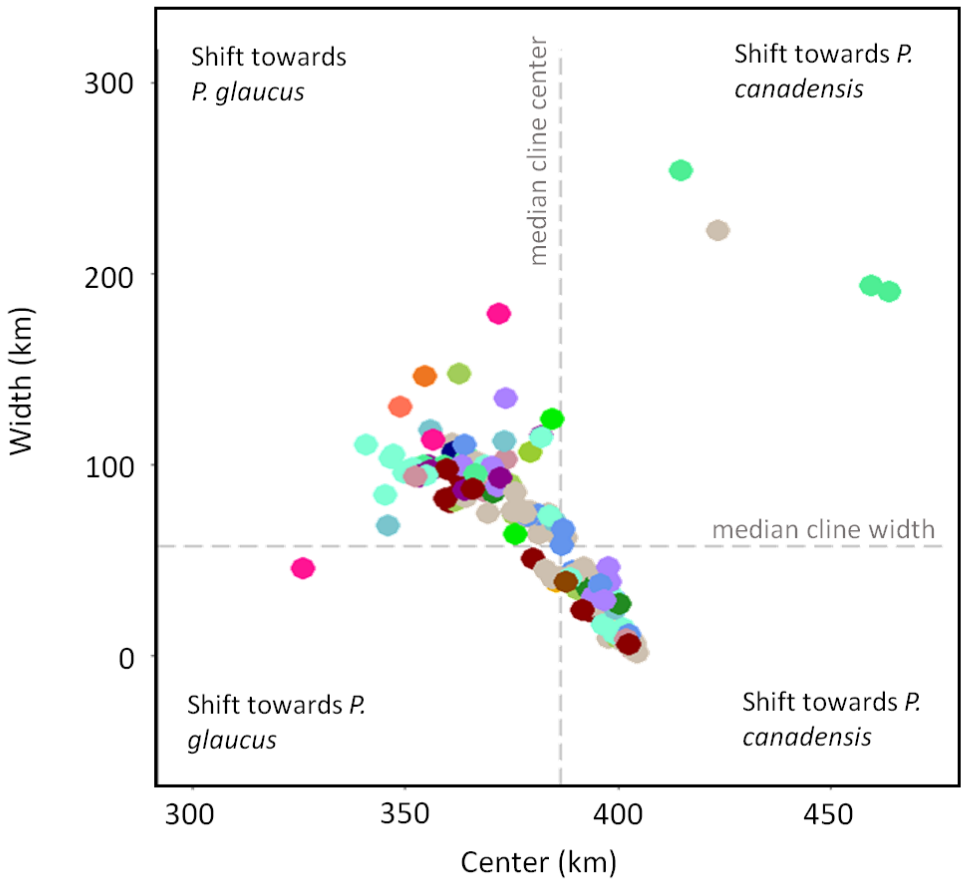
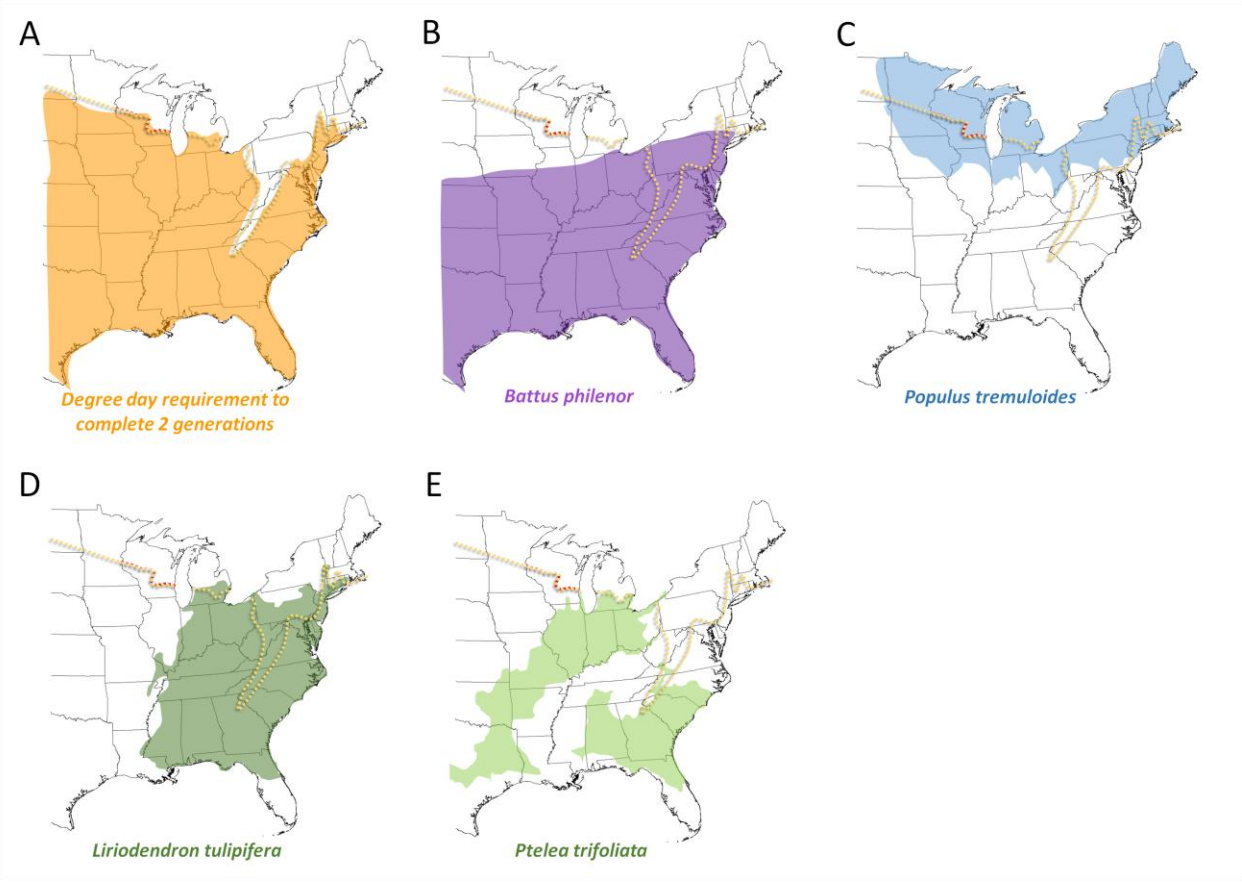


Figure 4: Maximum likelihood estimates of cline center and width for 230 (outlier) SNPs (circles). Dashed lines indicates estimated median cline center (387 km) and width (57 km) based on all genetic markers. Color indicates putative chromosome assignment (see Fig 2B).

1196



1197

1198 **Figure 5:** Approximate distributions of ecological factors associated with traits known to be
1199 ecological divergent between *P. glaucus* and *P. canadensis* in relation to the estimated location of
1200 the hybrid zone. The yellow dotted line indicates the estimated location of hybrid zone and the
1201 red portion highlights the region of the hybrid zone sampled in this study. (A) The estimated
1202 spatial distribution of growing degree days (GDD) required to complete two generations
1203 associated with facultative diapause; adapted from Scriber 2011 a threshold of >1300 GDD. (B)
1204 The distribution of the Batesian model (*Battus Philenor*) associated with the female black morph
1205 in *P. glaucus*; adapted from previously published distributions of *Battus Philenor* (Fordyce and
1206 Nice 2003). (C,D) Estimated distribution of two hosts plants (*Populus tremuloides* and
1207 *Liriodendron tulipifera*) associated with differential detoxification abilities between *P. glaucus*
1208 and *P. canadensis* (Lindroth *et al.* 1988; Li *et al.* 2004); adapted from range maps published in
1209 Little (1971, 1976).