Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species

Supplemental Material

Sequence data collection and assembly

We isolated and purified DNA from each of the sampled butterflies from approximately 10 mg of thoracic tissue using Qiagen’s DNeasy 96 Blood and Tissue Kit (Cat. No. 69581; Qiagen Inc., Valencia, CA, USA). We generated reduced genomic complexity libraries for each individual using a restriction fragment based procedure (van Orsouw et al., 2007; Gompert et al., 2010a; Andolfatto et al., 2011). Briefly, we digested genomic DNA with two restriction endonucleases (EcoRI and MseI) and ligated double stranded adaptors oligonucleotides to the digested fragments. These adaptors contained the priming sites necessary for Illumina sequencing and included 10 bp individual identification sequences (i.e., barcodes). We amplified the barcode-adapted fragments using two separate, replicate PCR reactions for each individual, which were then pooled. Following PCR, amplicons for sets of 96 individuals (four sets total) with unique barcodes were pooled. We separated pooled amplicons on a 2% agarose gel and excised fragments between approximately 250 and 500 bp in length. We purified these fragments using Qiagen’s Qiaquick Gel Extraction Kit (Cat. No. 28706; Qiagen Inc., Valencia, CA, USA). A detailed description of the molecular protocol used will be published in a separate manuscript (Parchman et al., 2012). DNA sequencing of the four libraries was performed by the National Center for Genome Research (Santa Fe, NM, USA) using the Illumina GAII platform. Each of the four libraries was sequenced in a separate lane and contained a haphazard combination of individuals from each population.

We used SeqMan NGen 3.0.4 (DNASTAR) to perform a de novo assembly for a subset of the sequences (12 million). We used a match size of 71 base pairs (bp), a
minimum match percentage of 92%, a match score of 10, a mismatch penalty of 15, and a
gap penalty of 30. We ran the de novo assembly on a 3.2 GHz Intel Xeon processor
running CentOS 5 (Linux), with 48 GB of RAM. 7.1 million of the sequences were
assembled into 149,632 contigs (mean number of sequences per contig = 74; contigs N50 =
92 bp). After removing low-quality contigs, we generated a partial reference genome from
the consensus sequences of contigs from the de novo assembly. Specifically, we removed
contigs that included complemented reads (the library preparation procedure we used
should not generate complemented reads), and contigs with an excessively short (less than
88 bp) or long (greater than 96 bp) consensus sequence. The expected contig length was 92
bp (108 bp sequences, 10 bp barcode and 6 bp association with the restriction site). Short
and long contigs generally represented poor alignments of non-homologous sequences. The
final reference contained 139,509 contigs each 88 to 96 bp in length. We then assembled the
full sequence data set (110 million sequences) to the reference using SeqMan xng 1.0.3.3
(DNASTAR). We used a mer size of 51 bp and a minimum match percentage of 90% for
the reference-based assembly. We only assembled reads that could be uniquely placed and
required a minimum aligned length of 81 bp for assembly. We ran the reference-based
assembly on a 2.7 GHz Intel Xeon processor running Mac OS X, with 19 GB RAM and 2
TB of scratch hard disk space. Eighty million of the 110 million (73%) sequences were
assembled to the reference. We used custom Perl scripts in conjunction with samtools
and bcftools (Li et al., 2009) to identify variant sites in the assembled sequence data and
to determine the number of reads supporting each alternative nucleotide state for each
individual and locus. We only called variant sites if data were present for at least 50% of
the sampled individuals and if the probability of the data assuming all samples were
homozygous for the reference allele was less than 0.01. We used the default full prior in
bcftools for calling variants. We ignored insertion deletion variation. To further prune
potential sequence errors from the final data set, we removed all loci where allele counts for
apparent heterozygotes were unlikely given a binomial distribution. Specifically, we
discarded a locus if $P(X \leq x|p = 0.5, n) \leq 0.05$, where $x$ denotes the count of the less frequently observed allele. We identified 119,677 variable sites using these stringent criteria. The mean number of sequences per variable site per individual was 2.21.

**Statistical models**

**Population allele frequency model**

We define a simple Bayesian model to estimate population allele frequencies. This model provides a means to estimate allele frequencies from low-coverage DNA sequence data, but is otherwise similar to alternative likelihood or Bayesian models for population allele frequencies (e.g., Pritchard et al., 2000; Gillespie, 2004; Hedrick, 2005). We assume that the data ($X$) consist of DNA sequences for $I$ loci and $J$ individuals. We further assume that the sequences do not contain errors, but that they are sampled stochastically with limited coverage per nucleotide (as is common with next-generation sequence data sets). We believe that the assumption of no errors is nearly met for our data, as we use rigorous post-assembly processing methods to minimize sequence errors in the final data set. We treat the genotype of individual $j$ locus $i$ ($a_{ij}$) as an unknown variable. If we assume sequences are sampled randomly from the underlying genotype, the likelihood of sequence data (counts of the number of sequence reads with each of the two possible states) of individual $j$ at locus $i$ ($x_{ij}$) can be written as:

$$P(x_{ij}|a_{ij} = kk') = \begin{cases} 
(1 - \delta_k'(n) \binom{n}{x_{ijk}} \frac{0.5^{x_{ijk}}1_a(x_{ijk})0.5^{x_{ijk'}}1_a(x_{ijk'})}{1} & \text{if } u_{ij} = 2 \\
\delta_k'1_a(x_{ijk}) + (1 - \delta_k')(0.5^{x_{ijk}}1_a(x_{ijk}) + 0.5^{x_{ijk'}}1_a(x_{ijk'})) & \text{if } u_{ij} = 1 \\
1 & \text{if } u_{ij} = 0
\end{cases}$$

(S1)
$a_{ij} = kk'$ is an ordered genotype with alleles $k$ and $k'$, $u_{ij}$ is the number of unique alleles observed, $\delta_{k k'}^k$ is Kronecker’s delta (i.e., $\delta_{k k'}^k = 1$ if $k = k'$ and 0 otherwise), and $1_a = 1$ if $x_{ijk} > 0$ is an indicator function. The likelihood of the full data set $X$ is obtained by taking the product of Eqn. S1 across all $I$ loci and $J$ individuals.

Following standard procedure, we define the sample allele frequency for the population as $\psi_{ik} = \frac{1}{2J} \sum_j 2I_{pp}(a_{ij}) + I_p(a_{ij})$, where $I_{pp}(a_{ij}) = 1$ if $a_{ij}$ contains two copies of allele $k$ and $I_p(a_{ij}) = 1$ if $a_{ij}$ contains one copy of allele $k$. The probability of the sample allele frequency for locus $i$ is given by the multinomial distribution:

$$P(\psi_i|\pi_{ij}, n) = \frac{n!}{\psi_{i1}! \cdots \psi_{ik}! \cdots} \pi_{i1}^{\psi_{i1}} \cdots \pi_{ik}^{\psi_{ik}}$$ (S2)

where $\pi_{ik}$ is the frequency of allele $k$ in the population, and $n$ is the number of gene copies sampled from the population (twice the number of diploid individuals). This conditional probability assumes Hardy-Weinberg and linkage equilibrium within the population. We specify an uninformative Dirichlet prior for the population allele frequencies such that $\pi_i \sim \text{Dirichlet}(1, \cdots, 1)$. Thus, the full model is $P(\pi, a|X) \propto P(X|a)P(a|\pi)P(\pi)$. Note, $P(a|\pi)$ is equivalent to $P(\psi|\pi)$. Samples from this posterior probability distribution are obtained using Markov chain Monte Carlo (MCMC) using software written in C++ by ZG and CAB. The software utilizes the GNU Scientific Library (Galassi et al., 2009).

**F-model**

We use a hierarchical Bayesian F-model to estimate coefficients of genetic differentiation ($F_{ST}$) among populations. This model is related to a model originally developed by Balding and Nichols (1995) and extended by several others (Nicholson et al., 2002; Falush et al., 2003; Beaumont and Balding, 2004; Foll and Gaggiotti, 2008; Guo et al., 2009). The F-model does not treat $F_{ST}$ as a simple summary statistic of the allele frequencies, but rather models uncertainty in $F_{ST}$ due to evolutionary sampling and statistical sampling.
(Holsinger and Weir, 2009). Moreover, our implementation of the Bayesian F-model allows information sharing among loci to capture the fact that they likely have some shared evolutionary history. In other words, locus-specific estimates of $F_{ST}$ will be a compromise between their maximum likelihood estimates and the mean genome-level $F_{ST}$. Similar to the basic allele frequency model, we assume Hardy-Weinberg and linkage equilibrium within each population, that the sequences do not contain errors, and that the sequences are sampled stochastically with limited coverage per nucleotide. The data ($X$) consist of DNA sequences for $I$ loci, which are associated with $J$ individuals. We treat the genotype of individual $j$ locus $i$ ($a_{ij}$) as an unknown variable. The likelihood of sequence data of individual $j$ at locus $i$ ($x_{ij}$) is given by Eqn. S1 above, and the likelihood of the full data set $X$ is obtained by taking the product of Eqn. S1 across all $I$ loci and $J$ individuals (from all sampled populations).

We define the sample allele frequency for population $m$ as

$$\psi_{ikm} = \frac{1}{2J} \sum_j 2I_{pp}(a_{ij}) + I_p(a_{ij}),$$

where $I_{pp}(a_{ij}) = 1$ if $a_{ij}$ contains two copies of allele $k$ and $I_p(a_{ij}) = 1$ if $a_{ij}$ contains one copy of allele $k$. The probability of the sample allele frequency for locus $i$ and population $m$ is given by the multinomial distribution:

$$P(\psi_{im}, n_m) = \frac{n_m!}{\psi_{1im} \cdots x_{ikm} \cdots \psi_{ikm} n_{1im} \cdots \pi_{1km}}$$

where $\pi_{imk}$ is the frequency of allele $k$ in population $m$, and $n_m$ is the number of gene copies sampled in population $m$ (twice the number of diploid individuals). We then assume that the probability of population allele frequency is described by a Dirichlet distribution, as follows:

$$P(\pi_{ij}, n_m) = \frac{1}{B(\phi_i \theta_{ij})} \prod_k \pi_{ij}^{\phi_i \theta_{ij} - 1}$$

where $\phi_i$ is the allele frequency vector for locus $i$ in the hypothetical ancestral population and $\theta_{im} = 1/F_{ST_{im}} - 1$. When only bi-allelic loci are considered, Eqn. S3 simplifies to a
binomial distribution and Eqn. S4 simplifies to a Beta distribution. This model holds for many neutral population genetic models (Balding and Nichols, 1995; Rannala and Hartigan, 1996; Balding, 2003), most notably the infinite-island model proposed by Wright (1943), and approximately for a model of divergence from a common ancestral population (Nicholson et al., 2002).

We define a hierarchical linear model for $F_{ST}$. Specifically, we assume:

$$\log\left(\frac{1}{\theta_i}\right) = \log\left(\frac{F_{ST_i}}{1 - F_{ST_i}}\right) = \psi_i. \tag{S5}$$

$\psi_i$ denotes the locus effect for $F_{ST}$. We assume $\psi_i \sim \text{Normal} (\mu, \tau)$, and $\mu$ is proportional to a genome-level metric of $F_{ST}$. Moreover, we impose a sum-to-zero constraint on the $\psi_i$ to ensure identifiability of $\mu$ and $\psi$.

We assign an uninformative gamma prior to the precision parameter $\tau$, Gamma $(0.001, 0.001)$, and an uninformative Normal prior to $\mu$, N $(\mu = 0, \tau = 10^{-4})$.

Finally, we assign an uninformative Dirichlet prior to the hypothetical ancestral population allele frequencies: $\phi_i \sim \text{Dirichlet} (1, \cdots, 1)$. This gives the following complete Bayesian model:

$$P(a, \pi, \phi, \psi, \mu, \tau | X, n) \propto P(X | a), P(a | \pi)P(\pi | \phi, \psi)P(\psi | \mu)P(\tau)P(\mu)P(\phi). \tag{S6}$$

The function $P(a | \pi)$ is equivalent Eqn. S3 as $\psi$ (the sample allele frequency) is a simple function of the sample genotypes, and $P(\pi | \phi, \psi)$ is equivalent to Eqn. S4 because $\theta$ is a deterministic transformation of $\psi$. Samples from the posterior probability distribution of this model are obtained using Markov chain Monte Carlo (MCMC) using software written in C++ by ZG and CAB. The software utilizes the GNU Scientific Library (Galassi et al., 2009).
Similar to Gompert and Buerkle (2011a,b), we designate statistical outlier loci based on locus-specific estimates of $\psi = \log \frac{F_{ST}}{1-F_{ST}} - 1$ relative to the estimated distribution of $\psi$ across the genome. In other words, rather then assuming a specific neutral evolutionary model, we concurrently estimate locus-specific values of $\psi$ and parameters that describe the distribution of $\psi$ across the genome from the sequence data. We then use the expected genome-wide distribution of $\psi$, which is specified by these estimated genome-level parameters, to identify loci with an unlikely or extreme magnitude of genetic differentiation. The estimated genome-wide distribution of $\psi$ is $\int \int \text{Normal}(\hat{\mu}, \hat{\tau}) d\mu d\tau$, where $\hat{\mu}$ and $\hat{\tau}$ are the posterior probability distributions of $\mu$ and $\tau$. We call locus $i$ a statistical outlier if the posterior point estimate of $\psi_i$ is not contained in the interval $q_N$, which is the interval bounded by the $\frac{N}{2}$ and $\frac{1-N}{2}$ quantiles of the genome-wide distribution of $\psi$.

**Genomic cline model with genotype uncertainty**

Genomic clines are mathematical functions that describe the probability of locus-specific ancestry given genome-wide ancestry or hybrid index (Gompert and Buerkle, 2011a). Genomic cline models are used to quantify variation in the form and rate of genomic introgression across the genome, where genomic introgression is defined as the movement of alleles from one genomic background into another. Two parameters are used to quantify locus-specific genomic introgression: genomic cline parameter $\alpha$ specifies an increase (positive values of $\alpha$) or decrease (negative values of $\alpha$) in the probability of parental population one (e.g., *L. idas*) ancestry relative to the base expectation, whereas the genomic cline parameter $\beta$ specifies an increase (positive values) or decrease (negative values) in the rate of transition from low to high probability of parental species one (e.g., *L. idas*) ancestry as a function of hybrid index (Fig. S1). Genomic clines describe the probability of ancestry at a locus, or whether an allele was inherited from parental population zero or one. Ancestry at a locus is estimated from the allelic state and the
parental allele frequencies. Full details of a Bayesian genomic cline model are presented by Gompert and Buerkle (2011a), and similar models are described by others (e.g., Szymura and Barton, 1986; Rieseberg et al., 1999; Lexer et al., 2007; Tang et al., 2007; Gompert and Buerkle, 2009; Macholán et al., 2011). Here we simply describe an extension of the Bayesian genomic cline model proposed by Gompert and Buerkle (2011a) to incorporate uncertainty in the genotype of individuals.

The Bayesian genomic cline model assumes two parental populations or species are known a priori, and that hybridization between these populations has produced one or more admixed populations. The model assumes Hardy-Weinberg equilibrium and linkage equilibrium within the parental populations, but does not assume these populations are fixed for different alleles. A genetic map can be used to model correlated genomic introgression for linked loci in the admixed population(s) Gompert et al. (2011). Similar to the other models we have described the data (X) consist of DNA sequences for I loci and J individuals. The J individuals were sampled from the two parental populations and one or more admixed populations. We wish to estimate the joint posterior probability distribution of the genotypes (A), parental allele frequencies (π), locus-specific ancestry for each admixed individual (Z; where z_{ij} denotes whether an allele was inherited from parental population 0 or parental population 1), hybrid index (h), genomic cline parameters (α and β), and hyperparameters that describe variation in α and β (τ_α and τ_β).

We treat the genotype of individual j locus i (a_{ij}) as an unknown variable. For individuals sampled from either parental population, the likelihood of sequence data of individual j at locus i (x_{ij}) is given by Eqn. S1 above, and the likelihood of the full data set X is obtained by taking the product of Eqn. S1 across all I loci and J individuals. We assign an uninformative Dirichlet prior to the parental population allele frequencies. Thus, the posterior distribution for parental allele frequencies is given by the allele frequency model described in the Appendix (A1). For a putatively admixed individual, the likelihood of the sequence data x_{ij} is dependent on ancestry at locus i (z_{ij} \in \{00, 01, 10, 11\}):
\[ P(x_{ij}|a_{ij}, z_{ij}, \pi_i) = P(x_{ij}|a_{ij} = k k') P(a_{ij}|z_{ij}, \pi_i) \] (S7)

where

\[
P(a_{ij}|z_{ij}, \pi_i) = \begin{cases} 
\delta_{k}^{k'} \pi_{ik0} + (1 - \delta_{k}^{k'})2\pi_{ik0}\pi_{ik'0} & \text{if } z_{ij} = 00 \\
\delta_{k}^{k'} \pi_{ik0}\pi_{ik'1} + (1 - \delta_{k}^{k'})\left(\pi_{ik0}\pi_{ik'1} + \pi_{ik1}\pi_{ik'0}\right) & \text{if } z_{ij} = 01 \\
\delta_{k}^{k'} \pi_{ik0}\pi_{ik'1} + (1 - \delta_{k}^{k'})\left(\pi_{ik0}\pi_{ik'1} + \pi_{ik1}\pi_{ik'0}\right) & \text{if } z_{ij} = 10 \\
\delta_{k}^{k'} \pi_{ik1} + (1 - \delta_{k}^{k'})2\pi_{ik1}\pi_{ik'1} & \text{if } z_{ij} = 11 
\end{cases} \] (S8)

and \( P(x_{ij}|a_{ij} = k k') \) is given by Eqn. S1. Model variables are described above in this and previous sections. The probability of ancestry is a function of hybrid index and the genomic cline parameters as described by Gompert and Buerkle (2011a). We use truncated gamma priors, rather than full gamma priors for \( \tau_\alpha \) and \( \tau_\beta \) to avoid infinitely large variance in genomic introgression rates among loci and to stabilize the MCMC analysis. All other model details are as described by Gompert and Buerkle (2011a). Samples from the posterior probability distribution of this model are obtained using Markov chain Monte Carlo (MCMC) using software written in \texttt{C++} by the authors. The software utilizes the GNU Scientific Library (Galassi et al., 2009).

### Supplemental analyses

#### Population genetic structure, pair-wise F\textsubscript{ST}:

We used estimates of genome-level F\textsubscript{ST} between pairs of populations to characterize population genetic structure in \textit{Lycaenides}. We used the MCMC implementation of the \textit{F-model} to estimate the posterior probability distribution of genome-level F\textsubscript{ST} and F\textsubscript{ST} for each locus for each of 78 comparisons. We ran two independent chains for 25,000 steps.
each. Samples were recorded every 10\textsuperscript{th} step following a 1,000 step burn-in. We based posterior probability estimates on a single chain after inspecting the MCMC output of both chains to assess convergence to the stationary distribution. Genome-level estimates of \( F_{ST} \) for each pair of populations are given in Table S1. We used a numerical ordination technique, non-metric multidimensional scaling (NMDS), to visualize population genetic structure based on genome-level \( F_{ST} \). We performed NMDS using the \texttt{R} function \texttt{isoMDS}. We chose to summarize the data using two dimensions based on initial diagnostics (Fig. S2). The ordination results are consistent with the PCA results presented in the main paper and indicate the admixed populations are intermediate in genomic composition relative to \textit{L. idas} and \textit{L. melissa} (Fig. S3).

\textbf{Quantifying genetic differentiation, \( G_{ST} \):}

We quantified genetic differentiation between \textit{L. idas}, \textit{L. melissa}, and Jackson Hole \textit{Lycaeides} using Nei’s \( G_{ST} \) (Nei, 1973). Unlike \( F_{ST} \) estimated using the F-model approach, \( G_{ST} \) is modeled as a simple deterministic function of the allele frequencies and does not incorporate evolutionary sampling (Holsinger and Weir, 2009). Specifically, following Nei (1973) we calculate \( G_{ST} \) for a locus as: \( \frac{H_T - H_S}{H_T} \), where \( H_T \) is the expected heterozygosity the total sample and \( H_S \) is the mean expected heterozygosity within each population. We calculated expected heterozygosity based on estimates of the population allele frequencies for \textit{L. idas}, \textit{L. melissa}, and Jackson Hole \textit{Lycaeides} (see the main text). Allele frequency estimates for each taxon were obtained from a single chain iterated for 5,000 steps (samples were recorded every fifth step following a 1000 step burn-in). We calculated \( G_{ST} \) for each locus and comparison using a custom \texttt{perl} script.

Genetic differentiation was greatest between \textit{L. idas} and \textit{L. melissa} (mean \( G_{ST} \) = 0.0543, maximum = 0.9379; Fig. S6). We detected moderate genetic differentiation between \textit{L. melissa} and Jackson Hole \textit{Lycaeides} (mean \( G_{ST} \) = 0.0377, maximum = 0.9454) and minimal genetic differentiation between \textit{L. idas} and Jackson Hole \textit{Lycaeides}. These
results are consistent with patterns of genetic differentiation measured with $F_{ST}$.

**Admixture simulations**

We simulated admixed populations to determine whether correlations between $F_{ST}$ and $\alpha$ can arise without selection. We have described, validated and used this model to simulate admixture previously (Buerkle and Rieseberg, 2008; Gompet and Buerkle, 2009, 2011a; Gompert et al., 2011). The model assumes a single admixed population is formed $N_{gen}$ generations in the past by equal mixing of parental gametes from two parental populations. We modeled diploid, hermaphroditic individuals. We assumed 10 chromosome pairs each 1 Morgan in length. We assumed a constant population size of $N_{ind}$ individuals and discrete generations. We simulated reproduction and generated the admixed population each generation by sampling pairs of gametes to form new individuals. We sampled each gamete from either parental population with probability $\frac{m}{2}$ or the admixed population with probability $1 - m$. We refer to $m$ as the migration rate from the parental populations. We sampled gametes after meiosis. We modeled recombination as a Poisson process with an expectation of one recombination event per Morgan. We tracked ancestry blocks and junctions along chromosomes to retain the complete ancestry information for each simulated individual.

Initially, we simulated admixed populations under four sets of conditions: (i) $N_{ind} = 5000$, $N_{gen} = 50$, $m = 0.02$, (ii) $N_{ind} = 5000$, $N_{gen} = 50$, $m = 0.2$, (iii) $N_{ind} = 250$, $N_{gen} = 100$, $m = 0.001$, (iv) $N_{ind} = 100$, $N_{gen} = 50$, $m = 0.001$. We simulated 100 data sets for each set of conditions. We sampled 100 individuals from each simulated admixed population. We then determined genotypes at each of 204 bi-allelic loci spread across four of the chromosome pairs (51 loci each). For each of the 204 loci, we first randomly sampled the estimated $L. idas$ and $L. melissa$ allele frequencies from one of the 17,693 loci in the *Lycaeides* data set and used these as the parental allele frequencies. We then sampled genotypes for each of the 100 admixed individuals and 50 parents from each population.
using these allele frequencies and the locus-specific ancestry of each admixed individual. We estimated $F_{ST}$ and genomic cline parameters for each of the simulated data sets using the models described in the main paper, but with known genotypes. We ran 50,000 MCMC iterations for each $F_{ST}$ and genomic cline analysis and based inferences on sampled parameter values after a 10,000 iteration burnin and retaining one of every ten samples. We then calculated the Pearson correlation coefficient between locus-specific $F_{ST}$ and locus-specific $|\alpha|$ for each simulated data set.

The distribution of correlation coefficients we obtained for each of the four sets of demographic conditions is provided in Fig. S8. The mean correlation coefficient under each set of demographic conditions exceeded the correlation coefficient estimated for the *Lycaeides* data and ranged from 0.32 to 0.46. This demonstrates that a correlation of the magnitude estimated from the *Lycaeides* data could easily be generated in the absence of selection. This occurs because the absolute value of each estimated $\alpha_i$ is constrained by the amount of information on ancestry present in the sequence data, which is related to the difference in allele frequency between the parent populations and thus to $F_{ST}$.

Given these results, we wanted to know whether the intrinsic correlation between $F_{ST}$ and $|\alpha|$ disappeared when $F_{ST}$ was greater than a threshold value. This might be expected if loci with $F_{ST}$ greater than a threshold value contained sufficient information regarding ancestry to take on any value of $\alpha$ over the range observed for the data (i.e., if estimates of $\alpha$ were limited by true variability in the probability of ancestry across the genome rather than variation in the information content of loci). We simulated ten additional data sets to address this question. We used the fourth set of demographic conditions from the previous simulations: $N_{\text{ind}} = 100$, $N_{\text{gen}} = 50$, $m = 0.001$. We chose this set of conditions because they resulted in a range of estimated values for $\alpha$ that was similar to the range for the *Lycaeides* data. We simulated the data sets as described previously, but with 15,010 bi-allelic loci spaced evenly across the 10 chromosome pairs. We increased the number of loci to better mimic the distribution of parameter estimates for the *Lycaeides*
data set and to ensure the presence of more high $F_{ST}$ loci. We estimated locus-specific $F_{ST}$
and $\alpha$ using MCMC as previously described. We then calculated the Pearson correlation
between $F_{ST}$ and $\alpha$ for loci with $F_{ST} \geq \{0.0, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18\}$. The
results are described in the main text.

**Morphometric analyses**

We quantified morphological variation within and among the *Lycaeides* populations using
three measurements of the male genitalia and 23 measurements of ventral wing pattern
(the area and location of orange and black pattern elements). The data used for this
analysis were previously published and are described by Gompert et al. (2010b). Data were
available for 146 individuals from 11 of the 13 sampled populations (we did not have
morphological data from Sinclair or Lander, WY *L. melissa*). We summarized variation in
each character (i.e., male genitalia and wing pattern) using principal component analysis
(PCA). The analyses were conducted in R using the `prcomp` function. The first principal
component for male genitalia explained 81% of the variation in this character and captured
genitalia size. The first principal component for wing pattern explained 35% of the
variation in this character and captured the relative area (relative to wing size) of orange
and black spots. We plotted the mean ($\pm$ one standard deviation) of male genitalia
principal component one (PC1) by wing pattern PC1 (Fig. S4).

**References**

D. L. Stern, 2011. Multiplexed shotgun genotyping for rapid and efficient genetic


Table S1: Genome level $F_{ST}$ ($\mu$) for pairs of populations. The median (upper triangle) and 95% ETPI (lower triangle) are shown.

<table>
<thead>
<tr>
<th></th>
<th>lan</th>
<th>sin</th>
<th>vic</th>
<th>bcr</th>
<th>btb</th>
<th>mrf</th>
<th>tss</th>
</tr>
</thead>
<tbody>
<tr>
<td>lan</td>
<td>0.0438</td>
<td>0.0669</td>
<td>0.0696</td>
<td>0.0627</td>
<td>0.0714</td>
<td>0.0584</td>
<td></td>
</tr>
<tr>
<td>sin</td>
<td>0.0583</td>
<td>0.0785</td>
<td>0.0844</td>
<td>0.0801</td>
<td>0.0889</td>
<td>0.0716</td>
<td></td>
</tr>
<tr>
<td>vic</td>
<td>0.0554</td>
<td>0.0749</td>
<td>0.0836</td>
<td>0.0761</td>
<td>0.0882</td>
<td>0.0678</td>
<td></td>
</tr>
<tr>
<td>bcr</td>
<td>0.0044</td>
<td>0.0285</td>
<td>0.0395</td>
<td>0.0186</td>
<td>0.0619</td>
<td>0.0242</td>
<td></td>
</tr>
<tr>
<td>btb</td>
<td>0.0041</td>
<td>0.0267</td>
<td>0.0381</td>
<td>0.0189</td>
<td>0.0603</td>
<td>0.0229</td>
<td></td>
</tr>
<tr>
<td>mrf</td>
<td>0.0036</td>
<td>0.0266</td>
<td>0.0382</td>
<td>0.0178</td>
<td>0.0617</td>
<td>0.0226</td>
<td></td>
</tr>
<tr>
<td>tss</td>
<td>0.0042</td>
<td>0.0272</td>
<td>0.0390</td>
<td>0.0167</td>
<td>0.0612</td>
<td>0.0234</td>
<td></td>
</tr>
<tr>
<td>btp</td>
<td>0.0076</td>
<td>0.0378</td>
<td>0.0187</td>
<td>0.0598</td>
<td>0.0216</td>
<td>0.0245</td>
<td></td>
</tr>
<tr>
<td>mtp</td>
<td>0.0257</td>
<td>0.0290</td>
<td>0.0199</td>
<td>0.0155</td>
<td>0.0340</td>
<td>0.0180</td>
<td></td>
</tr>
<tr>
<td>bnp</td>
<td>0.0363</td>
<td>0.0394</td>
<td>0.0187</td>
<td>0.0214</td>
<td>0.0220</td>
<td>0.0282</td>
<td>0.0227</td>
</tr>
<tr>
<td>bts</td>
<td>0.0176</td>
<td>0.0196</td>
<td>0.0167</td>
<td>0.0167</td>
<td>0.0443</td>
<td>0.0158</td>
<td></td>
</tr>
<tr>
<td>btp</td>
<td>0.0051</td>
<td>0.0165</td>
<td>0.0127</td>
<td>0.0167</td>
<td>0.0423</td>
<td>0.0056</td>
<td></td>
</tr>
<tr>
<td>bts</td>
<td>0.0205</td>
<td>0.0232</td>
<td>0.0167</td>
<td>0.0190</td>
<td>0.0419</td>
<td>0.0172</td>
<td>0.0341</td>
</tr>
</tbody>
</table>
Table S2: Number of loci that were classified as statistical $F_{ST}$ outliers ($q \geq 0.95$) in at least $n$ (column names) out of $N$ pairwise population analyses. Each row provides data for a subset of population comparison: $I = L. idas$, $M = L. melissa$, $A = Jackson$ Hole $Lycaeides$. Comparisons with "*" give the count of outlier loci in the global $L. idas$ by $L. melissa$ comparison (80 total) that were also outliers in individual population comparisons. As an example, 11 loci were outliers in at least two $L. idas \times L. idas$ comparisons.

<table>
<thead>
<tr>
<th>Populations (N)</th>
<th>Number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$I \times I$ (10)</td>
<td>107</td>
</tr>
<tr>
<td>$M \times M$ (3)</td>
<td>164</td>
</tr>
<tr>
<td>$A \times A$ (10)</td>
<td>563</td>
</tr>
<tr>
<td>$I \times M$ (15)</td>
<td>475</td>
</tr>
<tr>
<td>$I \times A$ (25)</td>
<td>354</td>
</tr>
<tr>
<td>$M \times A$ (15)</td>
<td>455</td>
</tr>
<tr>
<td>$I^* \times M^*$ (15)</td>
<td>80</td>
</tr>
<tr>
<td>$I^* \times I^*$ (15)</td>
<td>80</td>
</tr>
<tr>
<td>$M^* \times M^*$ (3)</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure S1: Plots depict hypothetical genomic clines. For both panes the solid black line depicts a genomic cline when $\alpha = \beta = 0$, which is the genome-wide mean expectation with hybrid index equal to the probability of ancestry from parental population 1. In the left panel $\beta_i$ is set to 0 and $\alpha_i$ is varied from 0.1 to 1 (solid gray lines) and from -0.1 to -1 (dashed gray lines). Similarly, in the right panel $\alpha_i$ is set to 0 and $\beta_i$ is varied from 0.1 to 1 (solid gray lines) and from -0.1 to -1 (dashed gray lines).
Figure S2: Goodness of fit for NMDS. A scree plot showing stress and the number of dimensions used (A). A Shepard plot depicting the relationship between pair-wise $F_{ST}$ and the ordination distance in three dimensions. Predicted ordination distances are shown (solid line).
Figure S3: Ordination-based summary of genetic differentiation in *Lycaeides* based on non-metric multidimensional scaling (NMDS). The ordination used pair-wise $F_{ST}$ and assumed three dimensions. Colored symbols represent individual populations: *L. idas* (orange triangles, 5 localities), *L. melissa* (blue circles, 3 localities) and Jackson Hole *Lycaeides* (green crosses, 5 localities).
Figure S4: Statistical summary of morphological variation in *Lycaeides* based on male genitalic morphology PC1 and wing pattern PC1. Points denote population means for five *L. idas* populations (orange), one *L. melissa* population (blue) and five Jackson Hole (green). Bars denotes ± one standard deviation. The Jackson Hole populations possess intermediate male genitalic morphology, but *L. idas*-like wing patterns.
Figure S5: Estimated distribution of hybrid index in Jackson Hole Lycaeides ($0 = L. melissa$, $1 = L. idas$).

Figure S6: Scatterplot depicting genetic divergence (measured by Nei’s $G_{ST}$) between $L. idas$, $L. melissa$, and Jackson Hole Lycaeides. Each point represents a locus. The 99th (red line) and 95th (pink line) quantiles of the empirical distribution of $G_{ST}$ are shown. The mean value of $G_{ST}$ is given for each comparison.
Figure S7: Estimated distribution of genomic cline parameters $\alpha$ and $\beta$ in Jackson Hole Lycaenidae.
Figure S8: Correlation between $F_{ST}$ and the absolute value of $\alpha$ for simulated data sets. Each pane shows the distribution of correlation coefficients from 100 replicate conditions. Results in each pane are for a different set of demographic parameters: $N_{ind} = 5000$, $N_{gen} = 50$, $m = 0.02$ (A), $N_{ind} = 5000$, $N_{gen} = 50$, $m = 0.2$ (B), $N_{ind} = 250$, $N_{gen} = 100$, $m = 0.001$ (C), and $N_{ind} = 100$, $N_{gen} = 50$, $m = 0.001$ (D).
Figure S9: Scatterplot depicting the relationship between $G_{ST}$ between *L. idas* and *L. melissa* and genomic cline parameter $\alpha$. Each circle represents a single locus and contour lines denote the bivariate density. Dark gray circles denote loci with 95% CI for $\alpha$ that do not include 0. We detected an association between $G_{ST}$ and the absolute value of $\alpha$ ($r = 0.146$, $P < 2.2e^{-16}$).
Figure S10: Histogram of expected admixture linkage disequilibrium (LD). We calculated expected admixture LD for each locus pair as $D_0 = Q(1-Q)\delta_A\delta_B$ following Pfaff et al. (2001). This expectation assumes admixture occurs in a single generation and that Hardy-Weinberg and linkage equilibrium hold in the parental populations. $Q$ is the contribution of *L. idas* to the admixed populations, and $\delta_A$ and $\delta_B$ are the allele frequency differentials. We calculated allele frequency differentials from point estimates of allele frequencies in *L. idas* and *L. melissa*. Low levels of expected admixture disequilibrium ($q_{0.25} = 0.0009$, $q_{0.50} = 0.0026$, $q_{0.75} = 0.0064$, $q_{0.99} = 0.0340$) are consistent with the low levels of genetic differentiation between *L. idas* and *L. melissa*.