Evolution repeats itself in replicate long-term studies in the wild

Patrik Nosil, Clarissa F. de Carvalho, Romain Villoutreix,
Laura Zamarano, Marion Sinclair-Waters, Nick Planidin,
Tom Parchman, Jeffrey Feder, Zach Gompert

Supplementary materials

Materials and Methods

Long-term field studies in *Timema cristinae*

We compiled data on morph frequencies in *T. cristinae* using samples collected in the spring using sweep nets between 1990 and 2023. This builds on a previously published data set that spanned the years 1990-2017 (1). All individuals were scored as 'striped', 'green' (a green morph lacking a stripe), or 'melanistic'. These classifications have been found to be highly repeatable in past work (2,3). Samples from 1990 to 1999 were collected and scored by Cristina Sandoval, who then trained PN in 2000. PN collected and scored most samples from 2000 to 2023. The host-plant collected on (*Adenostoma* or *Ceanothus*) was recorded for all individuals. Detailed information on the collection localities (i.e., GPS coordinates and elevations), morph frequencies, sample sizes, etc. is provided in Database S1.

We used this data set to estimate the average frequency of striped morphs on *Adenostoma* and *Ceanothus* host plants each year of the time series. In this analysis, we analyzed the 32,867 stick insect observations from Santa Ynez mountain along highway 154 (i.e., we excluded a second mountain, Refugio mountain, from analysis but included these data in the data archive for completeness). This includes counts of stick insects from 213 population by year combinations on *Adenostoma* and 206 population by year combinations on *Ceanothus*. Here, we define a population as the stick insects collected from a locality on a specific host plant species. Treating the stick insects from different host plant species at the same locality as different populations is necessary to estimate morph frequencies by host. However, sympatric and parapatric host-associated populations, that is populations on different hosts but in the same or nearly the same spatial location, exhibit very low levels of genetic differentiation and differences in morph frequency (1, 2, 4). Thus, the analyses of frequency-dependent selection in subsequent sections focus on locality level data, rather than distinguishing between host-associated populations at

the same locality.

Our analysis was conducted using a hierarchical Bayesian generalized linear model. We assumed a binomial likelihood for the number of striped stick insects for each population and year, with a population (j) and year (i) specific probability of stripe (p_{ij}) and sample size (n_{ij}). Melanistic stick insects were excluded from this analysis. Thus, our estimate of stripe frequency is out of the total number of striped and green stick insects. We then defined a linear model for the logit frequency of stripe for each *Adenostoma* and *Ceanothus* population for each year as:

$$\text{logit}(p_{ij}) = \alpha_j + \beta_i$$

Here, α and β are population and year specific effects, where population refers to a specific location and host plant. We placed separate normal priors on the α parameters for *Adenostoma* and *Ceanothus* populations, with means (μ_A and μ_C) and precisions (i.e., reciprocal of the variance, τ_A and τ_C) estimated from the data. Separate normal priors were also placed on the year effects for *Adenostoma* and *Ceanothus* with means set to 0 and precisions estimated from the data. Sum-to-zero constraints were imposed on the year effects to ensure model identifiability. We then placed mostly uninformative priors on these higher level means and precisions, $\mu \sim \text{normal}(\text{mean} = 0, \text{precision} = 1e^{-6})$ and $\tau \sim \text{gamma}(\text{shape} = 0.01, \text{rate} = 0.001)$. We fit this model via Markov chain Monte Carlo (MCMC) using the `rjags` (version 4.14) interface with `JAGS` (5). We ran three chains each with a 10,000 iteration burn-in followed by 20,000 iterations for sampling with a thinning interval of 5. This analysis was done using `R` version 4.2.2.

Estimating D from the time-series data

The parameter D can be used to make inferences on selection, including negative frequency-dependent selection (NFDS). Specifically, theory shows that evolutionary dynamics and out-

comes (e.g., stable oscillations, dampening oscillations, etc.) can be characterized by D (6, 7), defined as

$$D = \left. \frac{\partial p \Delta_p}{\partial p} \right|_{p=\hat{p}}.$$

Here, D is the slope of the best fit line from the linear regression of Δp (the change in frequency) on p (frequency) evaluated near the equilibrium (i.e., where $\Delta p = 0$). Ignoring all other processes, gradual convergence to a stable equilibrium p occurs when $-1 < D < 0$, dampened oscillations occur for $-2 < D < -1$, stable oscillations occur for $D = -2$, and diverging oscillations occur for $D < -2$; $D > 0$ indicates positive frequency-dependent selection (6). Additional processes, such as genetic drift or fluctuating selection in a heterogeneous environment, can alter dynamics and outcomes relative to these expectations (7).

We thus sought to characterize NFDS dynamics and expected outcomes from our stripe-frequency time series by estimating D. We specifically estimated D for each of the 10 *T. cristinae* locations with 10 or more pairs of consecutive years (i.e., pairs of p and Δp) (see Table S1 for details about the 10 locations). We did this using a hierarchical Bayesian model. We accounted for uncertainty in stripe frequency in each location (j) and year (j) by assuming the observed stripe count was binomially distributed, such that $y_{ij} \sim \text{binomial}(p_{ij}, n_{ij})$, where p_{ij} is the true (unobserved) stripe frequency and n_{ij} is the sample size. We then defined the following linear model:

$$\Delta p_{ij} = p_{i+1j} - p_{ij} = \alpha_j + D_j * p_{ij} + \epsilon_{ij}$$

Here, α_j is a location-specific intercept, D_j is the location-specific slope parameter of interest (D), and ϵ_{ij} is an error turn. We placed hierarchical priors on intercept and slope parameters, $\alpha_j \sim \text{normal}(\mu_\alpha, \sigma_\alpha)$ and $D_j \sim \text{normal}(\mu_D, \sigma_D)$. The hyperparameters were given the following weakly informative priors and inferred from the data: $\mu_\alpha \sim \text{normal}(\text{mean} = 0, \text{SD} = 20)$,

$\mu_D \sim \text{normal}(\text{mean} = 0, \text{SD} = 20)$, $\sigma_\alpha \sim \text{gamma}(2, 0.1)$, and $\sigma_D \sim \text{gamma}(2, 0.1)$. We assumed a normal prior on the error terms with a mean of 0 and a standard deviation estimated from the data; the latter was given the following prior, $\sigma \sim \text{gamma}(2, 0.1)$. We fit this model using Hamiltonian Monte Carlo (HMC) (8) via the `rstan` (version 2.21.8) interface with `Stan` (9). We used the No-U-Turn (NUTS) sampler (10) and ran four HMC chains each comprising a 1000 iteration warmup and 1000 additional sampling iterations. We computed and examined the Gelman–Rubin convergence diagnostic to verify adequate HMC mixing and likely convergence of the HMC algorithm to the posterior distribution. We assessed repeatability of evolutionary dynamics and outcomes across the 10 locations in terms of whether estimates of D were all consistent with NFDS ($D < 0$) and suggestive of a similar outcome (e.g., all $-1 < D < 0$ indicative of gradual convergence to an equilibrium), and based on the variability (e.g., standard deviation) of D across locations (a smaller standard deviation implies greater repeatability, see, e.g., (11)).

We then estimated the expected equilibrium stripe frequency in each location as $\hat{p_j} = \frac{-\alpha_j}{D_j}$. We computed the Pearson correlation coefficient to measure the association between the predicted equilibrium stripe frequency ($\hat{p_j}$) and the mean stripe frequency in each location ($\bar{p_j}$). All analyses described in this section were conducted using `R` version 4.2.2.

Field experiment testing the form of the NFDS function

We collected hundreds of stick insects from several core localities to conduct field transplant experiment and thereby estimate the NFDS fitness function. A total of 513 individuals were collected from locations LA (latitude 34.41, longitude -119.80), PRC (latitude 34.53, longitude -119.86), PRNC (latitude 34.53, longitude -119.85), VPC (latitude 34.53, longitude -119.85), and WTA (latitude 34.52, longitude -119.78) between April 30th and May 4th, 2021. Numbers were as follows: green morphs, LA 17, PRC 64, PRNC 58, VPC 71, WTA 48; striped morphs,

LA 136, PRC 0, PRNC 3, VPC 26, WTA 90. Individuals were randomly assigned to one of 21 treatments (n = 20 stick insects per treatment), which varied the initial stripe frequency from 0% (0 striped, 20 green) to 100% (20 striped, 0 green) in 5% increments. Each of these treatments of 20 individuals was then randomly assigned to one of 21 experimental bushes (near locality N2, in the general area of latitude 34.51 and longitude -119.80). Each bush was cleared of existing *T. cristinae* (the only *Timema* species occurring in this area) by sampling it before release. Past work demonstrates that this clears bushes of the overwhelming majority of *Timema* (2, 12, 13). Individuals were released on the morning of May 5th by dumping them from containers; the insects cling to the branches when released in this way because of a strong clinging reflex. Individuals were recaptured using visual surveys and sweep nets on the evening of May 7th, as in past work (2, 12–14), and scored as striped or green. Past work suggests minimal dispersal occurs from the focal bush and that recapture is a good proxy for survival (3, 14).

We fit a Bayesian model to test for NFDS in this experiment and quantify the relationship between stripe frequency and fitness. For this, we assumed that the number of green and striped stick insects recaptured from an experimental treatment (i.e., a single bush with a unique initial stripe frequency) was binomially distributed with a morph and treatment-specific survival probability (i.e., absolute fitness), $y_i^{green} \sim \text{binomial}(w_i^{green}, n_i^{green})$ and $y_i^{striped} \sim \text{binomial}(w_i^{striped}, n_i^{striped})$. Here, i denotes the treatment index (i.e., the index for the initial stripe frequency) and w_i^{green} and $w_i^{striped}$ denote the absolute fitnesses of green and striped morphs in treatment i. We then defined a pair of linear equations that specify the absolute fitness and relative fitness of the stripe morph as a function of the initial stripe frequency,

$$\text{logit}(\bar{w}_i) = \mu_{\bar{w}} + \beta_{\bar{w}} * p_i^{stripe}$$
$$\frac{w_i^{stripe}}{\bar{w}_i} = e^{\mu_\omega + \beta_\omega * p_i^{stripe}}.$$

In these equations, the μ terms serve as intercepts and the β terms capture the effects of initial

stripe frequency on average survival probability (average of the two absolute fitnesses) and relative fitness of the stripe, here $\frac{w_i^{striped}}{\bar{w}_i}$. From these equations, it follows that the absolute fitness of the green morph is $w_i^{green} = 2\bar{w}_i - w_i^{striped}$. We placed weakly informative normal priors on the μ and β terms, all $\mathrm{normal}(\mathrm{mean} = 0, \mathrm{SD} = 10)$. We fit this model using HMC via the `rstan` (version 2.21.8) interface with `Stan` (9). For this, we used four HMC chains with the NUTS sampler, each comprising 4000 warmup iterations followed by 4000 sampling iterations, and set the target average proposal acceptance probability (`adapt_delta`) to 0.9 (the default value is 0.8). We computed and examined the Gelman–Rubin convergence diagnostic to verify adequate HMC mixing and likely convergence of the HMC algorithm to the posterior distribution.

We next compared this model to an alternative model allowing for a non-linear relationship between stripe frequency and fitness. The model was identical to the linear model described above, except a sigmoid function was used to model relative fitness,

$$\frac{w_i^{stripe}}{\bar{w}_i} = \frac{L}{1 + e^{-1 * k * (p_i^{stripe} - c)}}$$

Here, L denotes the maximum relative fitness, k specifies the steepness of the curve and c is the stripe frequency at which the relative fitness is halfway between its minimum and maximum value. Depending on the parameter values, this function can approximate a step function (very steep curves) or a near-linear function (very shallow curves). We placed weakly informative normal priors on L and k, each with a mean of 0 and SD of 10, and an uninformative beta prior on c (both shape parameters set to 1). We used approximate leave-one-out cross validation to compare this model to the linear model described above. This was done with the `loo` function from `rstan` (15). As shown in the supplemental results (Table S3, Figure S3), the two models performed similarly with the linear model marginally outperforming the sigmoid model. However, even more importantly, both models result in near linear relationships between stripe

frequency and fitness. In other words, the best sigmoidal model approximated a linear model not a step model.

The above models allowed for flexible specification for the affect of stripe frequency on fitness, however, they did not give a direct estimate of D, the key NFDS parameter. Thus, we also fit a second model to estimate D in a manner analogous to the time-series data. This second Bayesian model was identical to that described above for estimating D from the time-series data, except that a single D was inferred (rather than one per locality) and thus the model was not hierarchical. As such, we used weakly informative priors on α (the intercept) and D, both $\mathrm{normal}(\mathrm{mean} = 0, \mathrm{SD} = 1)$. We likewise placed a weakly informative prior on the residual standard deviation, $\sigma \sim \mathrm{gamma}(2, 0.1)$. Here, we again used four HMC chains with the NUTS sampler; in this case, each chain comprised a 3000 iteration warmup followed by 20,000 iterations for sampling and `adapt_delta` was set to 0.95. Once again, we computed and examined the Gelman–Rubin convergence diagnostic to verify adequate HMC mixing and likely convergence of the HMC algorithm to the posterior distribution. The equilibrium stripe frequency for the experimental location was predicted from these results as $\hat{p} = \frac{-\alpha}{D}$. All analyses described in this section were conducted using `R` version 4.2.2.

Testing for dampened oscillations and for the effect of drift on stripe frequency fluctuations

If constant NFDS was the only evolutionary process operating, the estimated values of D from the 10 *T. cristinae* localities would predict dampening oscillations in stripe frequency that gradually converge to the a constant at the locality-specific equilibrium frequency. Thus, we first asked whether there was evidence of such convergence. To do this, we fit a hierarchical Bayesian model for stripe frequency change (Δp_{ij}) as a function of time (year, which is equivalent to generation); we considered only the consistently sampled years from 2000 onward for

this analysis. The hypothesis of dampening oscillations would predict a negative effect of year on the absolute value of stripe frequency change. We assumed a normal likelihood for stripe frequency change, specifically $\Delta p_{ij} \sim \text{normal}(\mu_{ij}, \sigma)$, with a linear model for the expected change as a function of time and locality (j)

$$\mu_{ij} = \alpha_j + \beta_j x_i^{year}.$$

In this equation, α_j is a location-specific intercept and β_j denotes the location-specific effect of year on the absolute value of change. We placed hierarchical priors on the regression parameters, $\alpha_j \sim \text{normal}(\mu_\alpha, \sigma_\alpha)$ and $\beta_j \sim \text{normal}(\mu_\beta, \sigma_\beta)$. We then placed weakly informative priors on the hyperparameters and estimated these from the data: $\mu_\alpha \sim \text{normal}(\text{mean} = 0, \text{SD} = 1)$, $\mu_\beta \sim \text{normal}(\text{mean} = 0, \text{SD} = 1)$, $\sigma_\alpha \sim \text{gamma}(2, 0.1)$, and $\sigma_\beta \sim \text{gamma}(2, 0.1)$. We likewise placed a weakly informative prior on the residual standard deviation term, $\sigma \sim \text{gamma}(2, 0.1)$. We fit this model using HMC (8) via the `rstan` (version 2.21.8) interface with `Stan` (9). We used the NUTS sampler (10) and ran four HMC chains each comprising a 1000 iteration warmup and 1000 additional sampling iterations. We again computed and examined the Gelman–Rubin convergence diagnostic to verify adequate HMC mixing and likely convergence of the HMC algorithm to the posterior distribution.

As reported in the main text, we found no evidence of a decline in the magnitude of frequency change over time (posterior probability [pp] $\mu_\beta < 0 = 0.11$, two locations have credible evidence of increases in change over time with pp $\beta > 0$ greater than 0.98, whereas none exhibited credible evidence of decline, all pp $\beta < 0$ were less than 0.71, see Figure S2). Thus, some additional processes are required to explain the persistent fluctuations in stripe given the estimated values of D (see (7)).

One possibility is that NFDS combined with only genetic drift could explain the observed fluctuations. We used simulations of evolution by NFDS and drift to test this hypothesis.

Specifically, for each location we simulated replicate runs of evolution starting from the observed initial stripe frequency. We determined the expected stripe frequency each generation as $\Delta p = D(p - \hat{p})$, where D and $\hat{p}$ are the estimated values of the NFDS parameter D and the estimated equilibrium stripe frequency for the relevant location from the hierarchical Bayesian time-series model (as reported in Table S1). We incorporated genetic drift into each simulation by binomial sampling, such that the simulated stripe frequency after a generations of NFDS and drift was $p' \sim \text{binomial}(p + \Delta p, n = 2N_e)$. For this, we used a published estimate of contemporary N_e for a *T. cristinae* population (FHA, latitude 34.52, longitude -119.80) based on the magnitude of genome-wide allele frequency change, N_e = 110 (1,16). We conducted 100 replicate simulations for each of the 10 locations and summarized each simulation in terms of the mean absolute change in stripe frequency per generation. We then compared the observed mean absolute change from each of the 10 locations to the corresponding null distribution. As a sensitivity analysis, we repeated this entire procedure for values of N_e between 100 and 20 in increments of 10 (Figure S4). All analyses described in this section were conducted using `R` version 4.2.2.

Experimental test of the selective advantage of the striped morph on *Adenostoma*

We conducted a second field transplant experiment to ask whether striped *T. cristinae* morphs exhibit higher fitness on *Adenostoma* than melanic morphs. For this experiment, stick insects were collected from a single location, PRNC (latitude 34.53, longitude -119.85). Stick insects for each pair of eight experimental blocks were captured on a single day in 2023: May 9th for blocks 1 and 2, May 11th for blocks 3 and 4, May 13 for blocks 5 and 6, and May 15 for blocks 7 and 8. For each block, 10 striped and 10 melanic stick insects (n = 20 total stick insects per block) were released onto a single *Adenostoma* bush at the experimental transplant

locality (latitude 34.51, longitude = -119.80, as in the first experiment) the day after they were collected. Despite paste evidence for limited dispersal off experimental bushes (3, 14), we explicitly tested this assumption again by marking the underbelly of the experimental stick insects with fine-tip sharpie before release. Other than marking stick insects, the release procedures were as described for the NFDS experiment. Two days after release, surviving stick insects were recaptured using visual surveys and sweep nets (see, e.g., (2, 12–14)) and scored as striped or melanic. Nearby bushes ($\sim$ within 10 meters, which is roughly the per-generation dispersal rate, (17)) were searched for marked *T. cristinae* stick insects to detect dispersal from the focal experimental *Adenostoma* bush. No marked stick insects were found off the focal bush in any of the eight blocks.

We then fit a hierarchical Bayesian generalized linear model to determine whether striped *T. cristinae* were more fit than melanic morphs on *Adenostoma*. We assumed a binomial likelihood for the number of striped stick insects recaptured for each block (j), $y_j \sim \text{binomial}(p_j, n_j)$. Here, p_j and n_j denote the block-specific probability that a recaptured stick insect was striped and the block-specific number of stick insects recaptured. We assumed $\text{logit}(p_j) = \beta + \alpha_j$, where β denotes the mean effect of being striped (relative to melanic) and α is a replicate-specific deviation from this mean. We modeled the α hierarchically with a zero-centered normal prior, $\alpha_j \sim \text{normal}(\text{mean} = 0, \text{SD} = \sigma)$, and placed weakly informative priors on β and σ, $\beta \sim \text{normal}(\text{mean} = 0, \text{SD} = 10)$ and $\sigma \sim \text{normal}(\text{mean} = 0, \text{SD} = 1)$. We again fit this model with HMC and the NUTS sampled via the `rstan` interface with `Stan`. We used four HMC chains, each comprising a 1000 iteration warmup followed by an additional 1000 sampling iterations. We again computed and examined the Gelman–Rubin convergence diagnostic to verify adequate HMC mixing and likely convergence of the HMC algorithm to the posterior distribution. We then focused on the posterior distribution of β, the stripe effect, for inference. This analysis was conducted with `R` version 4.2.2.

Reference genome assembly for the green *Timema cristinae* morph

In total, this study used three, independently assembled chromosome-level reference genomes, including an existing reference genome for the striped *T. cristinae* morph (18) and two new reference genomes described in this paper (one for the green *T. cristinae* morph and one for *T. podura*). We created the *de novo* reference genome for the green *T. crisinae* morph using a combination of PacBio and Illumina reads from Chicago and Hi-C genomic libraries. DNA extraction, library preparation, DNA sequencing, and reference genome assembly were performed by Dovetail Genomics (completed June 2020). A single female stick insect from the locality Paradise Road North (PRCN, latitude 34.53, longitude -119.85, collected on *Ceanothus spinosus* in 2019) was used for the assembly, and the individual was chosen based on a preliminary analysis that suggested it was homozygous for the green (unstriped) allele.

The DNA sample was quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (20kb) for PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer recommended protocol. The library was bound to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II. Sequencing was performed on PacBio Sequel SMRT cells. An initial genome assembly was then performed using the `FALCON` 1.8.8 pipeline from Pacific Bioscience. The assembly was polished through PacBio's Arrow algorithm from SMRT Link 5.0.1, using the original raw-reads.

A Chicago library was prepared as described in (19). Briefly, $\sim$500ng of high molecular weight gDNA was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to $\sim$350 bp mean fragment size and sequenc-

ing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X to a target depth of 30x coverage.

A Dovetail Hi-C library was prepared in a similar manner as described previously (20). Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to $\sim$350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X to a target depth of 30x coverage.

The PacBio *de novo* assembly, Chicago library reads, and Dovetail Hi-C library reads were used as input data for `HiRise`, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (19). An iterative analysis was conducted. First, Chicago library sequences were aligned to the draft input assembly using a modified `SNAP` read mapper (`http://snap.cs.berkeley.edu`). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by `HiRise` to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding Chicago data, Dovetail Hi-C library sequences were aligned and scaffolded following the same method. The final assembly comprised 1,313,928,163 base pairs (bps) with an N50 of 63,125,338 bps and an L50 of 7. The assembly included 13 large scaffolds corresponding to the

12 *T. cristinae* autosomes and the X chromosome (as described in more detail below) (Figure S6). Based on BUSCO version 4.0.5 (linage dataset = eukaryota_odb10, 70 species, 255 BUSCOs), the assembly included 195 complete BUSCOs (188 single copy and seven duplicated), and 10 fragmented BUSCOs; 50 BUSCOs were missing.

We then aligned this new *T. cristinae* genome to our existing chromosome-level reference genome for the striped *T. cristinae* morph (18, 21). We did this using `Cactus` (version 1.0.0) (22, 23). First, we used `RepeatMasker` (version 4.0.7) to mask repetitive regions of the genome (24) based on a repeat library developed for *Timema* (18). We then performed a pairwise alignment between the genomes with `Cactus` and used `HalSynteny` to extract syntenic alignment blocks from the comparative alignment (25). This was done with the default lower bound for synteny blocks of 5000 bps. We then identified homologous chromosomes by summing the total length of syntenic segments between the two genomes; there was a one-to-one correspondence between chromosomes. We then constructed sequence alignment dot plots from the synteny blocks using `R` (version 4.0.2) to visualize inversions and other structural variation between the genomes of the two morphs. We focused on chromosome 8 because of past evidence that chromosome 8 affects color pattern (results from our current study further support this) (26).

Genome-wide association mapping of color-pattern within *T. cristinae*

We used genome-wide association (GWA) mapping to identify regions of the genome associated with color-pattern (striped versus green) variation in *T. cristinae*. For this we used a previously published DNA sequence data set comprising partial genome sequences (genotyping-by-sequencing or GBS data) for 602 *T. cristinae* from a single population (FHA, latitude 34.52, longitude -119.80) (27). We aligned these data to both the newly generated green and striped *T. cristinae* reference genomes. This was done using the `bwa aln` algorithm (version 0.7.17) (28)

with a maximum miss-match of 4 and allowing no more than 2 miss-matches in the first 20 bps of each sequence. Bases with quality score less than 10 were removed and only reads with a single best alignment (-n 1) were kept. We use `samtools` (version 1.5) to compress, sort and index the alignments. We then called SNPs using `samtools` (version 1.5) and `bcftools` (version 1.6) (29). We used the consensus caller (-c), applied the recommended mapping quality adjustment for Illumina data (-C 50), and only output SNPs when the probability of all individuals being homozygous for the reference allele conditional on the data was $<$0.01. We used a series of `Perl` scripts to filter each variant set (i.e., based on each reference genome). These filters retained SNPs that met the following criteria: 2X minimum coverage per individual, a minimum of 10 reads supporting the non-reference allele, Mann-Whitney P-values for base quality, mapping quality and read position rank-sum tests $>$ 0.005, a minimum ratio of variant confidence to non-reference read depth of 2, a minimum mapping quality of 30, no more than 20% of individuals with missing data, only two alleles observed, and coverage not exceeding 3 standard deviations of the mean coverage (at the SNP level). A total of 169,051 SNPs passed these filters and were used for GWA mapping based on the striped genome (121,635 for the green genome).

We used `gemma` to fit linear mixed models and thereby test for associations between each SNP and color-pattern, that is green (0) or striped (1) (the color-pattern data are described and reported in (27)). This was done with the 538 stick insects that were green or striped (melanic stick insects were excluded). This method uses a kinship matrix to control for population genetic structure when testing for genotype-phenotype associations (all samples are from a single population, and thus limited structure is expected) (30). This analysis was performed using `gemma` version 0.95a with SNPs based on alignment and variant calling with the striped and green genomes. Likelihood ratio tests were used to compute P-values for the null hypothesis test of no association between genotype and phenotype.

Determining the bounds of the color-pattern loci

The GWA mapping analysis identified two large blocks of SNP associations with color-pattern on chromosome 8 (see Figures 6 and S5). We hypothesized that these two large regions of association comprised regions of reduced recombination driven by structural genetic variation, as was previously suggested by past association mapping analyses of color and color-pattern based on more fragmented genome assemblies (1, 26). Regions of reduced recombination and structural variation can manifest as an increase in PCA eigenvalues for the first PCs and clustering of individuals in PC space based on structural variant alleles (e.g. (26, 31, 32)). We thus used local PCAs in combination with the GWA signal to delineate two large genetic regions (loci) associated with color pattern variation, denoted *Pattern* and *Mel-Stripe2*. We first performed separate (local) PCA ordinations of the genetic data (centered but not standardized genotype matrixes) for the 5890 SNPs on chromosome 8 in 100-SNP sliding windows (5791 windows total). We then summarized each PCA by the eigenvalue associated with the first eigenvector. Visual inspection of sliding window eigenvalues suggested two peaks of high eigenvalues (accentuated structure) on chromosome 8 coinciding with the GWA signal.

We formally delimited the color-pattern loci using a Hidden Markov model (HMM) approach. Specifically, we fit a HMM to the eigenvalues using the `HiddenMarkov` package (version 1.8.13) in `R` (33). We modeled two hidden states, which we initialized with expected values equal to the 25th and 75th percentiles of the empirical eigenvalue distribution for chromosome 8. We assumed a normal distribution for the observed eigenvalues with standard deviations initialized at half the empirical standard deviation. We then estimated the hidden state means, standard deviations, and transitions between hidden states using the Baum-Welch algorithm (i.e., we set initial values for means and standard deviations but these were then refined with the Baum-Welch algorithm) (34). For this, we allowed a maximum of 500 iterations and set the tolerance to $1e^{-4}$. Using this procedure, we identified states corresponding to high (mean

eigenvalue = 2.67, SD = 0.43) and low or background (mean eigenvalue = 1.87, SD = 0.14) population structure (i.e., high and low eigenvalues). We then used the Viterbi algorithm for decoding, that is for inferring the most likely hidden state for each 100 SNP window (35). This algorithm assigned 4072 SNP windows to the low (background) state and 1719 windows to the high state. The 1719 high state windows comprised 8 contiguous regions of the high state that broadly overlapped with the GWA signal. We combined neighboring high windows separated by narrow regions of the low HMM state to define the two color pattern loci corresponding to the GWA signals: *Pattern* = 1,096,929 to 12,418,814 bps, and *Mel-Stripe2* = 45,229,604 to 84,397,122 bps.

Finally, we examined the correspondence between these color-pattern loci and structural variation between the green and striped *T. cristinae* genomes as identified from the whole genome comparative alignment described above. This revealed a single large inversion corresponding approximately with *Pattern*, and several large inversions and insertions (in the striped genome) in *Mel-Stripe2*.

Reference genome assembly for *Timema podura* and comparative alignment with *T. cristinae*

We created a *de novo* reference genome for *T. podura*. This was done using a combination of PacBio and Illumina reads from a Omni-C genomic library. DNA extraction, library preparation, DNA sequencing, and reference genome assembly were performed by Dovetail Genomics (completed December 2021). A single female stick insect (green morph, ID = 17_0262) from Bay Springs, CA (BSC, latitude 33.82, longitude -116.70, collected on *Ceanothus* in 2017) was used for the assembly. The DNA sample was quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library ($\sim$20kb) for PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA,

USA) using the manufacturer recommended protocol. The library was bound to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II). Sequencing was performed on PacBio Sequel II 8M SMRT cells.

`wtdbg2` (36) was then run to generate a primary assembly. Next, `blobtools` v1.1.1 (37) was used to identify potential contamination in the assembly based on `blast` (v2.9) results of the assembly against the NT database. A fraction of the scaffolds was identified as contaminant and were removed from the assembly. The filtered assembly was then used as an input to `purge_dups` v1.1.2 (38), and potential haplotypic duplications were removed from the assembly, resulting in the final assembly.

For the Dovetail Omni-C library, chromatin was fixed in place with formaldehyde in the nucleus. Fixed chromatin was digested with DNaseI and then extracted, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce $\sim$30X sequence coverage.

The input PacBio *de novo* assembly and Dovetail Omni-C library reads were used as input data for `HiRise`, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (19). The Dovetail Omni-C library sequences were aligned to the draft input assembly using `bwa` (28). The separations of Dovetail Omni-C read pairs mapped within draft scaffolds were analyzed by `HiRise` to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. The final assembly comprised

1,202,903,989 bps with an N50 of 73,318,109 bps and an L50 of 7. The assembly included 14 large scaffolds corresponding to the 12 *T. cristinae* autosomes and the X chromosome (Figure S7). Based on BUSCO version 4.0.5 (linage dataset = eukaryota_odb10, 70 species, 255 BUSCOs), the assembly included 233 complete BUSCOs (232 single copy and seven duplicated), and 8 fragmented BUSCOs; 14 BUSCOs were missing.

Next, we aligned our new *T. podura* genome to the our new green morph *T. cristinae* genome (see above) and to our existing chromosome-level reference genome for the striped *T. cristinae* morph (18, 21). We did this separately for each pair of genomes using `Cactus` (version 1.0.0) precisely as described above for the *T. cristinae* comparative genome alignment (22,23). Briefly, this involved repeat masking with `RepeatMasker` (version 4.0.7), (24), pairwise genome alignment with `Cactus`, and the extraction of syntenic alignment blocks from the comparative alignment with `HalSynteny` (25). We then identified homologous chromosomes by summing the total length of syntenic segments between the two genomes. There was a one-to-one correspondence between most chromosomes, but *T. cristinae* chromosomes 1, 3 and 4 were each homologous with portions of *T. podura* scaffolds 3 and 12, and chromosome 3 and 4 also shared homology with *T. podura* scaffolds 2 (chromosome 3 only) and 8 (chromosome 4 only) (see Figure S7). We then constructed sequence alignment dot plots of *T. cristinae* chromosome 8 from the synteny blocks using `R` (version 4.0.2) to visualize inversions and other structural variation between morphs (this chromosome corresponded with a single scaffold, number 29, in *T. podura*).

Tables

Table S1: Characteristics of the 10 replicate locations studied over time. Years = number of consecutive year pairs analyzed, Observations = total number of observations (i.e., the sample size in terms of numbers of individuals), D = Bayesian point estimate (median) and 95% equal-tail probability intervals for D, $\hat{p}$ = predicted equilibrium stripe frequency.

Location	Latitude°	Longitude°	Years	Observations	D	$\hat{p}$
FH	34.5176	-119.8010	10	4531	-0.38 (-0.65, -0.13)	0.92
HV	34.4884	-119.7864	22	4580	-0.32 (-0.51, -0.11)	0.71
L	34.5089	-119.7956	15	4101	-0.48 (-0.81, -0.21)	0.89
M	34.5151	-119.7971	15	1543	-0.42 (-0.74, -0.14)	0.67
MBOX	34.5035	-119.8056	10	884	-0.38 (-.060, -0.12)	0.68
OG	34.5128	-119.7962	13	1659	-0.45 (-0.76, -0.18)	0.78
OUT	34.5318	-119.8435	14	1664	-0.70 (-0.96, -0.40)	0.43
PR	34.5331	-119.8574	17	2403	-0.61 (-1.04, -0.25)	0.02
SC	34.5226	-119.8318	12	776	-0.58 (-1.05, -0.24)	0.05
VP	34.5324	-119.8473	12	5255	-0.50 (-0.83, -0.22)	0.21

Table S2: Bayesian estimates of the slope parameters relating the change in stripe frequency to time. Point estimates and 95% equal-tail probability intervals (ETPIs) for each of the 10 focal locations are shown.

Location	Slope (95% ETPI)
FH	0.0001 (-0.0019, 0.0021)
HV	0.0004 (-0.0012, 0.0021)
L	-0.0006 (-0.0024, 0.0011)
M	0.0009 (-0.0009, 0.0026)
MBOX	0.0043 (0.0018, 0.0073)
OG	0.0011 (-0.0011, 0.0032)
OUT	0.0025 (0.0007, 0.0043)
PR	0.0003 (-0.0014, 0.0020)
SC	0.0006 (-0.0014, 0.0029)
VP	0.0009 (-0.0009, 0.0028)

Table S3: Comparison of linear and sigmoid functions for relative fitness in the negative frequency-dependent selection (NFDS) experiment based on approximate leave-one-out (LOO) cross validation. Estimates and standard errors (in parentheses) are shown for the Bayesian LOO estimate of the expected log pointwise predictive density ($\mathrm{ELPD}_{\mathrm{LOO}}$), the effective number of parameters ($\mathrm{P}_{\mathrm{LOO}}$), and the LOO information criterion (LOOIC). Smaller values of LOOIC denote better models in terms of expected predictive performance.

Model	$\mathrm{ELPD}_{\mathrm{LOO}}$	$\mathrm{P}_{\mathrm{LOO}}$	LOOIC
Linear	-70.9 (6.7)	5.6 (1.1)	141.8 (13.3)
Sigmoidal	-71.6 (7.0)	6.4 (1.3)	143.2 (14.1)

Table S4: Comparison of mean absolute stripe frequency change and null expectations with drift and negative frequency-dependent selection (NFDS) assuming an effective population size of 110. Results are shown for each of 10 locations. Mean absolute change, the ratio of mean change to the mean null expectation (X-fold change) and the P-value from the null hypothesis test that the observed change did not exceed expectations for drift and NFDS are given.

Location	Mean change	X-fold change	P-value
FH	0.027	1.32	0.03
HV	0.045	1.62	<0.01
L	0.032	1.60	<0.01
M	0.056	1.69	<0.01
MBOX	0.075	2.65	<0.01
OG	0.042	1.70	<0.01
OUT	0.080	2.33	<0.01
PR	0.021	2.48	<0.01
SC	0.022	1.61	<0.01
VP	0.058	2.25	<0.01

Figures

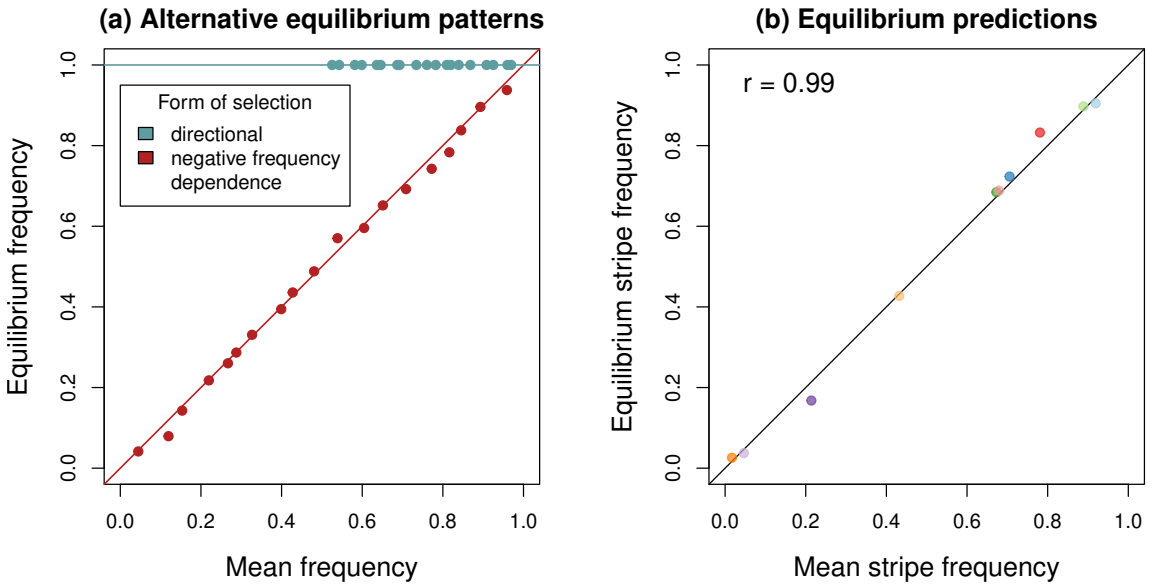

Figure S1: Replicate time series reveal a close match between observed and predicted equilibrium morph frequencies under negative frequency-dependent selection. Although the same data are used to estimate observed and predicted frequencies, the strong association detected in not a given and would not arise under all models of evolution (e.g., those with directional evolution rather than fluctuations due to NFDS). Panel (a) shows the general relationship between the mean frequency in a time series and the equilibrium predicted for directional selection favoring a morph or negative frequency-dependent selection. The relationship for the observed data is shown in (b) with points colored to denote locations as in Figure 1 of the main text (Pearson correlation = 0.99, $P < 0.0001$).

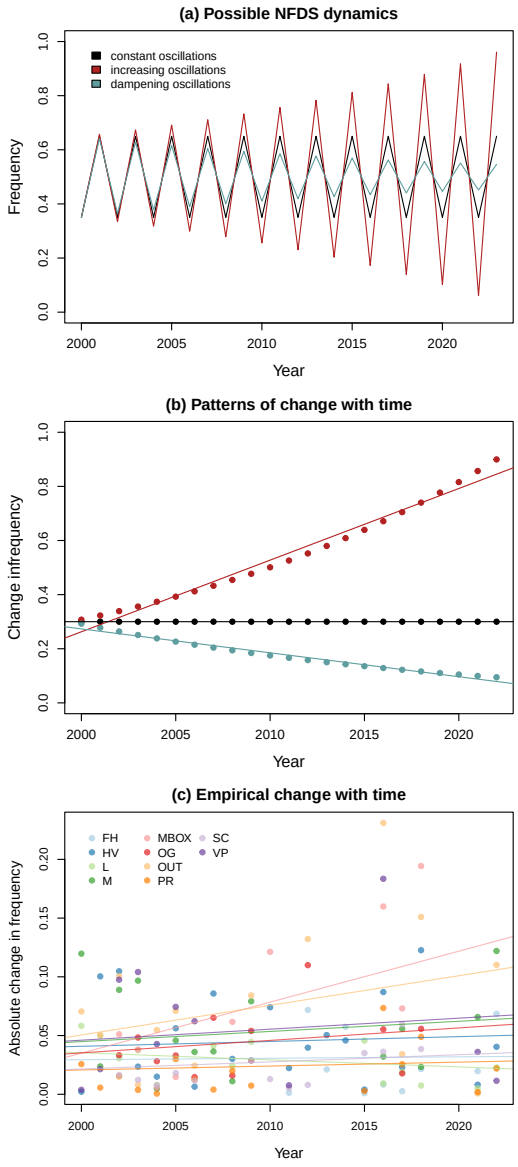

Figure S2: Alternative evolutionary dynamics with NFDS and in the empirical time-series data. Panel (a) shows possible NFDS dynamics with constant, increasing or dampening oscillations over time. These lead to different relationships of the absolute change in frequency (e.g., stripe frequency) as a function of time; points and a best-fit line are shown for these hypothetical examples (b). Panel (c) is scatterplot showing the absolute change in stripe frequency for the empirical stripe frequency data for the 10 focal locations as a function of year. Colored lines are best-fit lines for each location from a hierarchical Bayesian linear regression. The points and lines are colored to denote locations as in Figure 1. The analysis provides no evidence of a decline in the magnitude of stripe frequency change over time in any location.

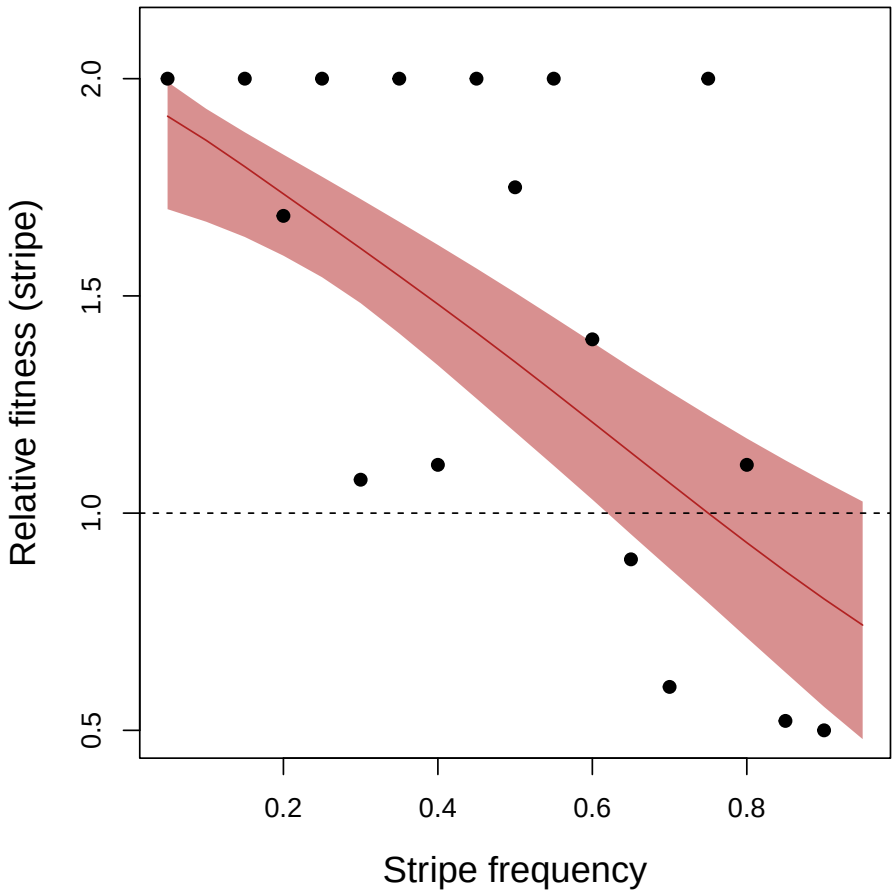

Figure S3: Relative fitness of striped versus green morphs as a function of stripe frequency. In contrast to the results shown in the main manuscript, this plot show results from fitting a sigmoid function for relative fitness (i.e., a non-linear model). Even with the assumed sigmoid model, the estimated parameters result in a relatively linear decline of relative fitness with frequency. Dots represent values for each experimental bush with lines and shading representing the estimated fitness function (medians and 95% equal-tailed probability interval, ETPI, respectively)

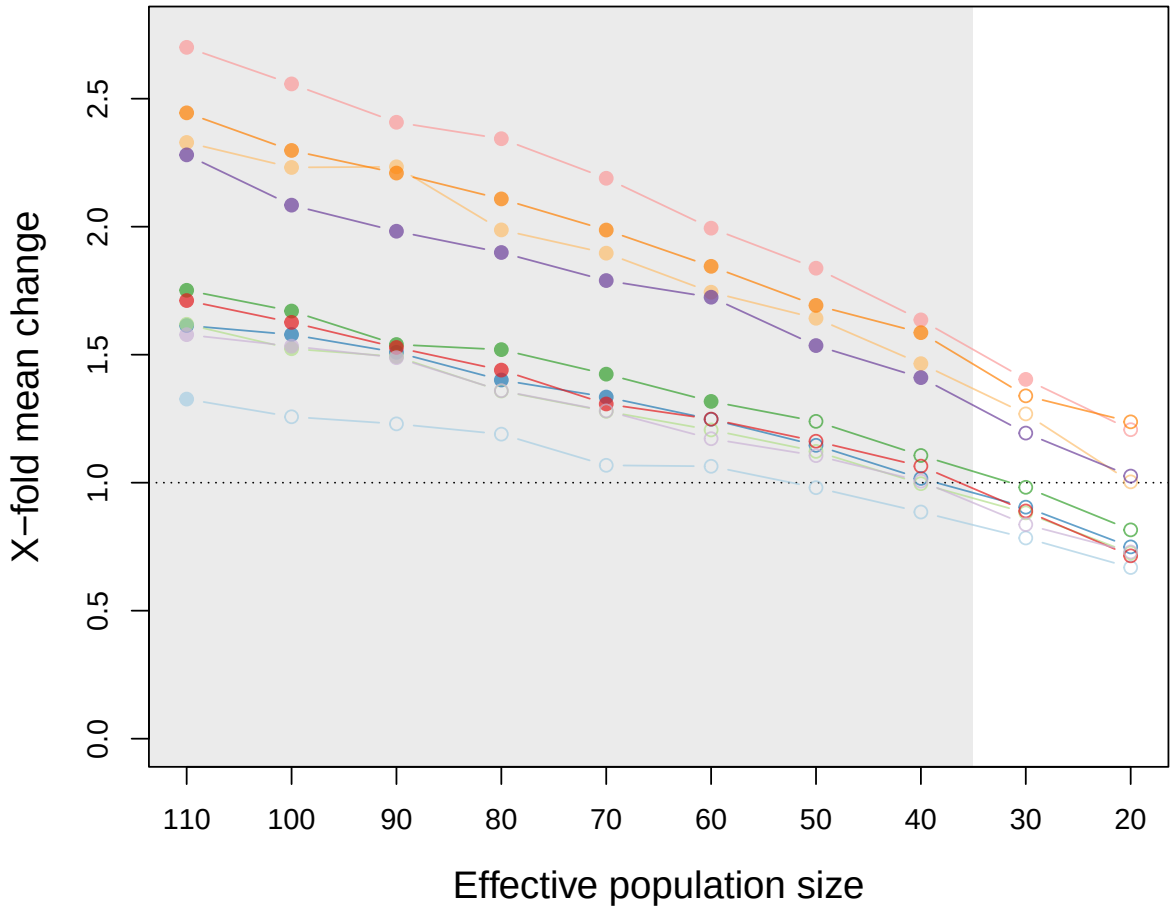

Figure S4: Graphical summary of the sensitivity analysis for explaining the observed fluctuations in stripe frequency by drift and negative frequency-dependent selection (NFDS). The plot shows the ratio of the mean absolute change in stripe frequency for each location relative to the expectation from simulations of evolution by drift (X-fold mean change). The dotted horizontal line indicates X-fold mean change of 1, which implies the observed mean change matches expectations for drift combined with NFDS. Results are shown for a range of effective population sizes, from 110 to 20. Colored points denote different locations (see the legend in Figure 1e). Filled circles denote conditions where the probability of the observed change by drift and NFDS was less than 0.05 (i.e., $P < 0.05$); open circles indicate $P \geq 0.05$. The shaded region of the plot encompasses effective population sizes where the Fisher combined P-value for the null hypothesis that drift can explain change for all 10 locations was less than 0.05.

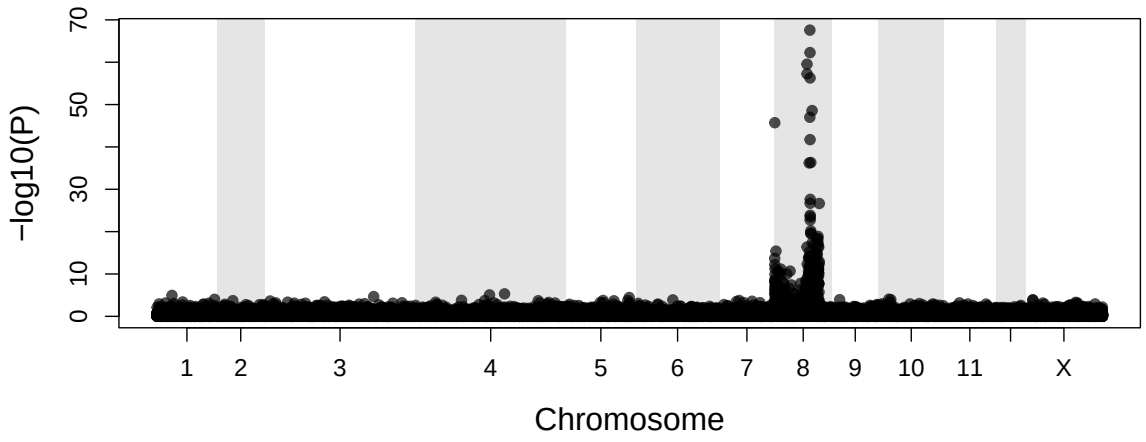

Figure S5: Genome-wide association mapping results for the green genome. This plot shows a Manhattan plot of the negative log10 P-value for the association between genome-wide SNPs and color pattern (green or striped). Points denote individual SNPs and shaded regions delimit chromosomes.

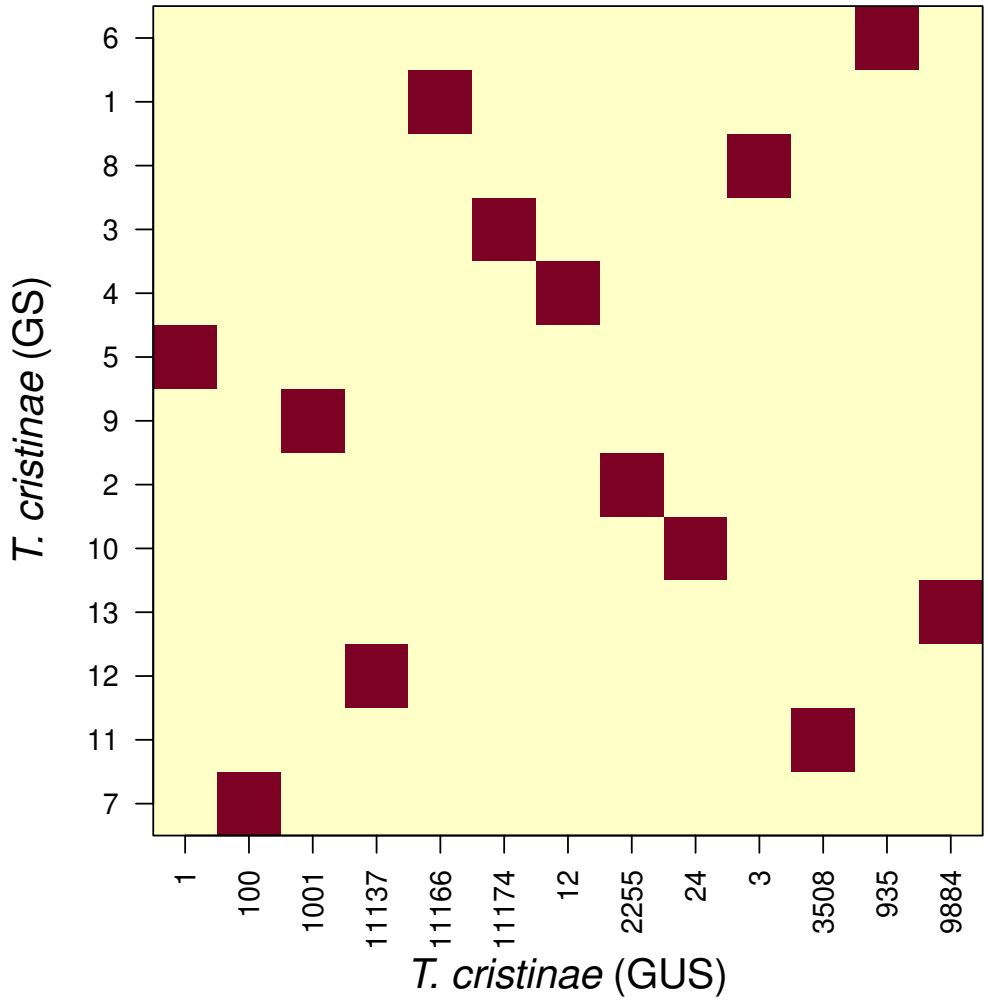

Figure S6: Correspondence between *T. cristinae* striped morph (GS) chromosome and scaffolds (chromosomes) for the green unstriped (GUS) morph. The heatmap shows the proportion of synteny blocks involving each striped morph chromosome that aligned to each of the green morph scaffolds.

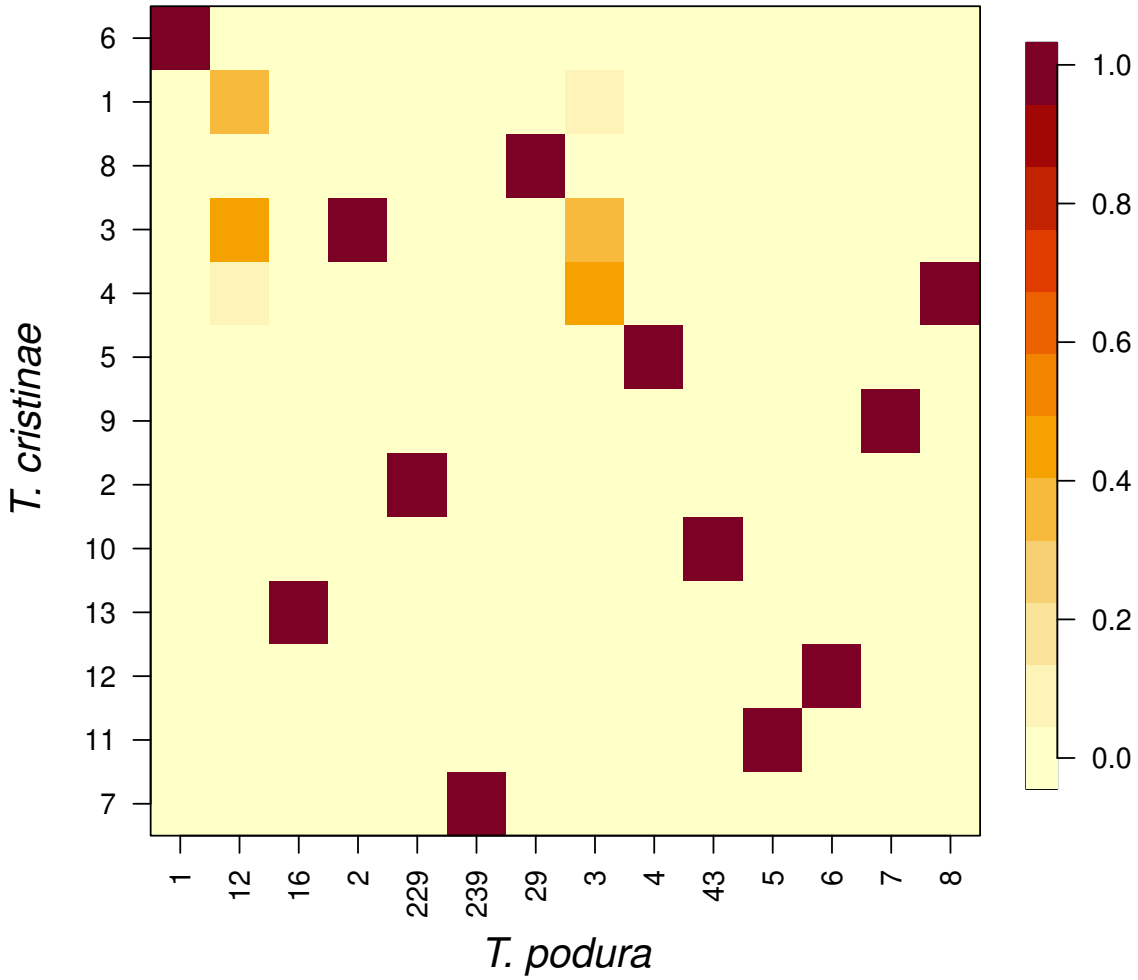

Figure S7: Heatmap shows the proportion synteny blocks on of each *T. podura* scaffold that aligned to each of the *T. cristinae* chromosomes. Results are shown for the striped morph.

References

1. Nosil, P. *et al.* Natural selection and the predictability of evolution in *Timema* stick insects. *Science* **359**, 765–770 (2018).

2. Sandoval, C. P. The effects of the relative geographic scales of gene flow and selection on morph frequencies in the walking-stick *Timema cristinae*. *Evolution* **48**, 1866–1879 (1994).

3. Nosil, P. Reproductive isolation caused by visual predation on migrants between divergent environments. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **271**, 1521–1528 (2004).

4. Nosil, P. *et al.* Genomic consequences of multiple speciation processes in a stick insect. *Proceedings of the Royal Society B: Biological Sciences* **279**, 5058–5065 (2012).

5. Plummer, M. *rjags: Bayesian Graphical Models using MCMC* (2023). URL `https://CRAN.R-project.org/package=rjags`. R package version 4-14.

6. Lewontin, R. C. A general method for investigating the equilibrium of gene frequency in a population. *Genetics* **43**, 419 (1958).

7. Chevin, L.-M., Gompert, Z. & Nosil, P. Frequency dependence and the predictability of evolution in a changing environment. *Evolution Letters* **6**, 21–33 (2022).

8. Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo* **2**, 2 (2011).

9. Stan Development Team. RStan: the R interface to Stan (2023). URL `https://mc-stan.org/`. R package version 2.21.8.

10. Hoffman, M. D. & Gelman, A. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15**, 1593–1623 (2014).

11. Gompert, Z., Flaxman, S. M., Feder, J. L., Chevin, L.-M. & Nosil, P. Laplace's demon in biology: Models of evolutionary prediction. *Evolution* **76**, 2794–2810 (2022).

12. Sandoval, C. P. Differential visual predation on morphs of *Timema cristinae* (Phasmatodeae: Timemidae) and its consequences for host range. *Biological Journal of the Linnean Society* **52**, 341–356 (1994).

13. Soria-Carrasco, V. *et al.* Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).

14. Gompert, Z. *et al.* Experimental evidence for ecological selection on genome variation in the wild. *Ecology Letters* **17**, 369–379 (2014).

15. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432 (2017).

16. Gompert, Z., Feder, J. L. & Nosil, P. The short-term, genome-wide effects of indirect selection deserve study: A response to Charlesworth and Jensen (2022). *Molecular Ecology* **31**, 4444–4450 (2022).

17. Sandoval, C. Persistence of a walking-stick population (Phasmatoptera: Timematodea) after a wildfire. *The Southwestern Naturalist* 123–127 (2000).

18. Villoutreix, R. *et al.* Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science* **369**, 460–466 (2020).

19. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* **26**, 342–350 (2016).

20. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

21. Nosil, P. *et al.* Complex evolutionary processes maintain an ancient chromosomal inversion. *Proceedings of the National Academy of Sciences* **120**, e2300673120 (2023).

22. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Research* **21**, 1512–1528 (2011).

23. Armstrong, J. *et al.* Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

24. Smit, A. F. Repeat-masker open-3.0. *http://www. repeatmasker. org* (2004).

25. Krasheninnikova, K. *et al.* halSynteny: a fast, easy-to-use conserved synteny block construction method for multiple whole-genome alignments. *GigaScience* **9**, giaa047 (2020).

26. Lindtke, D. *et al.* Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Molecular Ecology* **26**, 6189–6205 (2017).

27. Riesch, R. *et al.* Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology & Evolution* **1**, 0082 (2017).

28. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

29. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

30. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).

31. Li, H. & Ralph, P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211**, 289–304 (2019).

32. Todesco, M. *et al.* Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).

33. Harte, D. *HiddenMarkov: Hidden Markov Models*. Statistics Research Associates, Wellington (2021). URL `https://www.statsresearch.co.nz/dsh/sslib/`. R package version 1.8-13.

34. Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41**, 164–171 (1970).

35. Forney, G. D. The Viterbi algorithm. *Proceedings of the IEEE* **61**, 268–278 (1973).

36. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods* **17**, 155–158 (2020).

37. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).

38. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).