Article:
Nosil, P. orcid.org/0000-0002-8271-9005, Villoutreix, R. orcid.org/0000-0002-1815-3844, de Carvalho, C.F. orcid.org/0000-0002-5780-5454 et al. (5 more authors) (2018) Natural selection and the predictability of evolution in Timemastick insects. Science, 359 (6377). pp. 765-770.

https://doi.org/10.1126/science.aap9125

Natural Selection And The Predictability Of Evolution In *Timema* Stick Insects

Patrik Nosil[1], Romain Villoutreix[1], Clarissa F. de Carvalho[1], Timothy E. Farkas[2], Víctor Soria-Carrasco[1], Jeffrey L. Feder[3], Bernard J. Crespi[4], Zach Gompert[5]

[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06369, USA
[3]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA
[4]Department of Biological Sciences, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[5]Department of Biology, Utah State University, Logan, Utah 84322, USA

Predicting evolution remains difficult. We study the evolution of cryptic body coloration and pattern in a stick insect using 25 years of field data, experiments, and genomics. We find that evolution is more difficult to predict when it involves a balance between multiple selective factors and uncertainty in environmental conditions than when it involves feedback loops that cause consistent back and forth fluctuations. Specifically, changes in color morph frequencies are modestly predictable through time (r^2 = 0.14), and driven by complex selective regimes and yearly fluctuations in climate. In contrast, temporal changes in pattern morph frequencies are highly predictable due to negative frequency-dependent selection (r^2 = 0.86). For both traits, however, natural selection drives evolution around a dynamic equilibrium, providing some predictability to the process.

Introduction:

Evolutionary biology is often portrayed as a descriptive rather than predictive science (*1, 2*). Nonetheless, the extent to which past evolution predicts future evolution can be quantified by testing how well early subsets of a time series predict subsequent changes. However, predictability in the form of such temporal autocorrelation does not consider the underlying mechanisms driving evolutionary change, and thus can be inherently low. Considering the mechanisms of evolution can lead to increased understanding of evolutionary change and its predictability, and we study such mechanisms here.

Evolutionary predictability is mediated by several factors. First, evolution can be unpredictable because of random processes, such as genetic drift (*3*). Second, even when evolution occurs by deterministic natural selection, predictive power can be diminished if multiple, complex forms of selection act simultaneously and by uncertainties in how the ecological conditions that affect selection change through time (*1, 2, 4-7*). For example, negative frequency-dependent selection (NFDS) favoring rare alleles can enhance predictability by causing increases in allele frequency to be followed predictably by decreases (and *vice-versa*)(*8, 9*). However, the extent of predictability will depend on how NFDS interacts with other evolutionary processes (e.g., directional selection stemming from climate change), and on whether the ecological conditions that affect these other processes can themselves be predicted. Third, the interaction of genes within their genomic context (i.e., dominance and epistasis) and with the environment (i.e., plasticity) may affect the anticipated trajectory of evolution (*10-12*). For example, directional selection is expected to produce a predictable evolutionary response only if the traits affected are reasonably heritable, and even then responses can be complex and nuanced (*4-7, 10*).

Studying the predictability of evolution across different timescales is particularly challenging (*1, 2*). For example, the immediate impact of natural selection can be readily measured in short-term field studies or experiments. Such studies suggest that strong selection is not uncommon at the scale of one or a few generations (*13, 14*), especially when new environments are colonized (*15-17*). However, short-term changes need not translate into long-term directional trends. Rather, evolution across geologic and phylogenetic time scales may be characterized by periods of relative stasis interspersed between occasional bursts of sustained directional change, consistent with Simpson's fossil-record inspired model of 'adaptive zones' (*18, 19*). Indeed, strong but fluctuating selection can generate a pattern of little change when averaged over longer time periods (*14*). Field studies that measure patterns of evolution over many years or decades are somewhat intermediate between immediate and geologic time scales and thus have the potential to illuminate how short-term selection relates to longer-term patterns of evolution (*1, 20-22*).

Such studies are rare because they require long-term temporal monitoring that cannot be sped up with more effort. We here analyze the predictability of evolution in a long-term study of the stick insect *Timema cristinae* (Fig. 1), and bolster our inferences using manipulative experiments and genomic analyses.

Polymorphism in stick insects

T. cristinae is a univoltine, wingless, plant-feeding stick insect that exhibits three morphs that are cryptic on different plant species or tissues (Figs. 1, 2)(*23-27*). A green morph bearing a white dorsal stripe is cryptic on the leaves of *Adenostoma fasciculatum*, a green and unstriped morph is cryptic on the leaves of *Ceanothus spinosus*, and a melanistic (i.e., brownish/grey and unstriped) morph is cryptic on the stems of both hosts (but is conspicuous on leaves). Accordingly, the striped morph is common on *Adenostoma*, the unstriped morph is common on *Ceanothus*, and the melanistic morph is found at ~10% frequency on both hosts. We refer to the variation between green (striped plus unstriped) and melanistic individuals as ‘color polymorphism’ and that between green-striped and green-unstriped individuals as ‘pattern polymorphism’ (i.e., color and pattern are different ‘traits’). These polymorphisms are highly heritable with strong genetic dominance (melanistic body coloration is recessive to green and stripe pattern is recessive to unstriped; details below)(*23, 24*).

Several processes maintain color and pattern polymorphism (*23-27*). A balance between divergent selection and gene flow between populations on *Adenostoma* versus *Ceanothus* helps maintain pattern polymorphism (*26, 27*). However, other factors likely contribute because even areas dominated by one host rarely fix for a single pattern morph (*26, 27*). The frequency of melanism does not vary markedly among populations such that variation in color is maintained by balancing selection within populations, potentially involving heterozygote advantage and selection that varies with microhabitat (stems versus leaves)(*23, 24*). Spatial and host-related aspects of evolution for these morphs are thus reasonably well understood. In contrast, whether and how morph frequencies change through time, and if they do so predictably, is unknown.

We studied temporal dynamics in *T. cristinae* using 25 years of field data from the mountains around Santa Barbara, California (545 locality-by-host-by-year estimates of morph frequency on the basis of 34,383 individuals collected from 1990 to 2017, mean n per locality-by-host-by-year = 63, s.d. = 141, Database S1). We focus our autocorrelation analyses of predictability (*28*) on the locality HV (an acronym for Hidden Valley), where we collected data over a continuous 18-year period, with no years of missing data (total n = 3470, mean yearly n = 193, s.d. = 268).

Results:

Temporal change in allele frequencies

We tested the hypothesis that temporal change in allele frequencies at the genetic region underlying the morphs is due, in part, to natural selection. Although past studies support selection on the morphs (*23-27*), these results do not mean that all temporal changes are due to selection, because selection and drift are not mutually exclusive (*29, 30*). Genomic data from different time points provide a means to test for selection, because strongly selected regions are expected to show greater change through time than the more neutral genomic background (*29, 31*). Genomic data are further required in our specific instance because genetic dominance and heterozygote excess complicate inference of allele frequency change using phenotypic data alone (*23, 24*).

We used *de novo* genome sequencing of a melanistic and green morph, with Dovetail hi-rise scaffolding of Illumina reads (N50 = ~16 and 8 megabases, respectively)(*32*), linkage mapping (*25*), and genome-wide association (GWA) mapping to explicitly delimit a single, contiguous genomic region (~10.5 megabases in size) associated with color and pattern variation (Figs. 2, S1-2). Consistent with a similar study with a more fragmented reference genome, this region exhibits three core haplotypes (i.e., alleles), one corresponding to each morph, designated *s*, *u*, and *m* for green-striped, green-unstriped, and melanistic, respectively (i.e., in terms of diploid genotypes and phenotypes: *uu, us* and *um* = green-unstriped; *ss* and *sm* = green-striped; *mm* = melanistic)(*23*). We refer to this region as the *Mel-Stripe* locus hereafter.

We quantified allele frequency changes at *Mel-Stripe* over time within three published data sets: (1) genotyping-by-sequencing (GBS) data collected in a natural population on *Adenostoma* (FHA, acronym for locality Far Hill on *Adenostoma*) in 2011 and 2013 (*30, 33*), (2) re-sequenced whole genomes from individuals collected in FHA and used in an eight-day (i.e., within-generation) release and recapture field experiment (*30*), and (3) GBS data in a between-year (i.e., between-generation) field transplant experiment (*25*).

In each case, we contrasted change through time at *Mel-Stripe* to that of the remainder of the genome (to all genomic scaffolds, i.e., loci, that harbored as many single nucleotide polymorphisms as *Mel-Stripe*, which was 40, 16, and 39 loci, respectively, for the data sets noted above). Due to our explicit interest in *Mel-Stripe* we did not attempt to delimit other loci under selection. If such loci exist they could upwardly bias our estimates of genome-wide change relative to a case of neutrality, making our results for *Mel-Stripe* conservative.

We found that *Mel-Stripe* showed the greatest temporal allele frequency change of all genomic regions, in all three data sets (FHA, change = 0.0273, $P = 0.024$; within-generation experiment, change = 0.0340, $P = 0.059$; between-generation experiment, change = 0.0988, $P = 0.025$; exact probabilities; Fisher's combined probability test across data sets: $X^2 = 20.50$, d.f. = 6, $P = 0.0023$, Fig. 2). Dispersal alone is unlikely to drive these observed patterns because FHA was sampled over an area that is larger (>10,000 m^2) than the dispersal capacity of *T. cristinae* (i.e., one to a few dozen meters per generation)(*30, 33, 34*). Furthermore, field surveys detected essentially no dispersal off experimental bushes in the recapture study (*30*), and selection on pattern has been previously observed in the presence, but not the absence of predation (with dispersal possible in both treatments)(*35, 36*). Thus, selection likely contributed to the genetic change we observed at the *Mel-Stripe* locus. We thus next turned to whether such selection was associated with weakly or strongly predictable patterns of evolution.

Predictability of the evolution of body color and complex selection regimes

We quantified the predictability of evolution using autoregressive moving average models, ARMA (*28*). This analysis revealed that that color morphs at Hidden Valley (HV) exhibited subtle and only moderately predictable changes through time (median predictive r^2 = 0.14, ARMA, Fig. 3, Table S1). This was associated with support for multiple, complex, and counteracting sources of selection (Fig. 4). Across the 25-year study period, the frequency of melanistic morphs increased in years where spring temperatures were warmer (overall effect of temperature = 0.187, 95% equal-tail probability intervals = 0.063-0.309; correlation between observed and predicted frequency from cross-validation = 0.16, 95% CI = 0.04-0.28, P = 0.0102, Bayesian hierarchical linear model, Fig. 4, Table S2). However, lab experiments indicate that melanistic individuals have weaker heat tolerance, relative to green individuals (B = 3.57, 95% CIs = 1.34-9.51, P = 0.0111, Cox proportional hazards regression model using exact likelihood). These results imply that selection for crypsis on dry, brownish plants in warmer years may favor dark colors, but that thermoregulatory selection acts in an opposing direction. However, further work is required to test this hypothesis directly, and to establish how well the laboratory experiments match field conditions. Notably, melanistic individuals exhibit fewer fungal infections and greater mating success than other morphs, further suggesting that selection is multi-faceted (*24*).

Given selection appears complex, we used our genomic data to estimate selection coefficients explicitly (for all six diploid genotypes underlying the morphs, with three alleles: *s*, *m*, and *u*). This analysis revealed that viability selection on *Adenostoma* during late life-history stages (i.e., in the within-generation experiment) favored the *s/s* homozygote (posterior probability that fitness of *s/s* > than the following genotypes: *m/u* = 0.93, *u/u* = 0.81, *m/m* = 0.82, *m/s* = 0.84, *u/s* = 0.91). In contrast, the most-fit genotype between years at the FHA locality (also on *Adenostoma*) was the *s/m* heterozygote (posterior probability that fitness of *s/m* > than the following genotypes: *m/u* = 0.90, *s/s* = 0.97, *u/u* = 0.81, *m/m* = 0.92, *u/s* = 0.94). Both *s/s* and *s/m* are green-striped in terms of phenotype, and cryptic on *Adenostoma*. Thus, a fluctuating balance between many factors, potentially including heterozygote advantage (*23*), may explain why evolution involving the deterministic process of selection was not more highly predictable.

Predictability of the evolution of pattern

Our findings for color raise the question of whether evolution is ever highly predictable? We address this issue by re-analyzing the data from Hidden Valley (HV) considering striped versus unstriped individuals (i.e., pattern, rather than color, polymorphism). This demonstrates that striped morph frequencies at HV exhibited consistent increases followed by decreases (i.e., up and down fluctuations) across 18 consecutive years (Binomial sampling probability < 0.0001, Fig. 3). Thus, the evolution of pattern was highly predictable (median predictive $r^2 = 0.86$, ARMA). In fact, predictive power at the scale of three to five years was near perfect (>0.95), and remained high even after a decade (>0.80). The observed pattern of predictable of up and down fluctuations could reflect a case where predators have specific search images for common prey, resulting in NFDS selection favoring rare prey phenotypes.

Morph frequency and selection

To test the NFDS hypothesis, we transplanted green-striped and green-unstriped *T. cristinae* to *Adenostoma* in either 1:4 or 4:1 ratios (n = 1000 individuals, 10 replicates per treatment, Fig. 5). The NFDS hypothesis predicts that the striped morph, cryptic on *Adenostoma*, would exhibit a stronger survival advantage when rare. Supporting this prediction, the striped morph experienced strong selection when initially rare (selection coefficient, $s = 0.70$), and increased in frequency in all 10 experimental replicates (posterior probability that change > 0 was >0.999). In contrast, the striped morph showed idiosyncratic changes when initially common ($s = -0.04$, posterior probability that change > 0 = 0.43). Although our results differ from NFDS where the sign of selection reverses very strongly, they are consistent with the strength of selection being dependent on frequency (i.e., directional selection that weakens with increasing allele frequency is akin to frequency-dependent selection). It is possible that selection against the striped morph would be more strongly negative if ratios were manipulated more extremely (e.g., to 10:1).

Conclusions:

We observed complex and fluctuating sources of selection. Together with gene flow (*26*), these selection pressures likely contribute to relative stability in the difference between hosts in morph frequency over the 25-year study period (Fig. 5). Complex and fluctuating selection may thus help maintain polymorphism, but prevent divergence of sufficient magnitude to strongly drive speciation (*24*). NFDS in particular may cause evolutionary systems to exhibit resilience, as reported in other complex ecological, social, and physical systems (*37-39*). Such resilience increases predictability of short-term evolution, because a system returns to its former state following perturbation. However, it can make long-term predictions difficult because substantial evolution only happens when the system reaches a tipping point that pushes it more permanently to a very different state.

Finally, our results suggest that the predictability of evolution can depend on the nature of selection, and our understanding of it. Thus, we predicted evolution more accurately for pattern, where selection appears to be strongly associated with frequency, than for color, where a myriad of factors, some poorly understood, affect fitness. As further illustration of this point, predictability in the form of temporal autocorrelation is modest to weak in other well-known studies of contemporary evolution: beak and body size changes in Darwin's finches and morph frequency changes in the scarlet tiger and peppered moth(*21, 40, 41*)(median predictive power per study system, r^2, in 3-10 year forecasting analyses 0.03-0.18, ARMA, Figs 6, S3-S6, Table S1). Adding climatic (i.e., rainfall) data to the *Geospiza* finch case, where climate is known to affect seed distributions, improves predictive power in *Geospiza fortis* (e.g., mean r^2 increase relative to a model without rainfall = 0.08, ARMA; Table S3). Nonetheless, predictive power even considering rainfall is modest and increased only in one of the two finch species examined (Table S3). It is possible that predictability remains limited because seed size itself was not modeled, because the relationship between evolution and rainfall is complex such that only extreme droughts have strong effects, and because some extreme climatic events precede the ten-year period that our forecasting is based upon.

In conclusion, our constrained understanding of selection and environmental variation (i.e., limits on data and analysis), rather than inherent randomness, can thus limit ability to predict evolution. In turn, these limitations may affect our understanding of ecological processes, because to the extent that evolution can be predicted, perhaps so can its ecological consequences for population dynamics, community structure, and ecosystem functioning (*42-44*).

Acknowledgements

The work was funded by a grant from the European Research Council (NatHisGen R/129639) to P. Nosil and a Natural Sciences and Engineering Research Council grant to B.J. Crespi (NSERC 06505). V. Soria-Carrasco was supported by a Leverhulme Trust Early Career Fellowship. We thank C. Sandoval for contributing data on morph frequencies. C. Sandoval, A. Comeault, M.

Muschick, R. Riesch, and T. Reimchen provided discussion and comments on previous versions of the manuscript. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged, as well as access to the High Performance Computing Facilities, particularly to the Iceberg HPC cluster, from the Corporate Information and Computing Services at the University of Sheffield. The genome drafts reported in this paper have been deposited in the Genomes database of the National Center for Biotechnology Information (NCBI) under BioProject Accession PRJNA417530. Other data, including custom code written for analyses, have been archived in Dryad Digital Repository XXXX (number can only be provided by Dryad if the manuscript is accepted). Rosa Ribas drew all the figures.

Author contributions

PN and ZG conceived the project. PN, RV, CC, TEF, VS, ZG contributed data and analyses. PN, JLF, and ZG wrote the manuscript, with feedback from all authors.

Figure 1. Drawings of the three morphs of *T. cristinae.*

Figure 2. Genomic change through time at the *Mel-Stripe* locus. (a) Manhattan plots showing results for genome-wide association mapping of color. The y-axes show *P*-values, with red denoting genome-wide significance. The left-land plot shows results genome wide (LG = linkage group). The right-hand plot is zoomed in on LG8, which shows the bulk of association, and here numbers below the x-axis delimit different genomic scaffolds. The *Mel-Stripe* locus is evident by the block of strong association spanning scaffolds 702.1 and 128. (b) Allele-frequency change through time in the natural population FHA (2011 versus 2013). (c) Allele-frequency change through time in the within-generation experiment. (d) Allele-frequency change through time in the between-generation experiment. In panels b-d the vertical red line shows change at the *Mel-Stripe* locus and the histogram shows the distribution of change across other similar-sized scaffolds in the genome (i.e., the genomic background).

Figure 3. Predicting evolution in *Timema cristinae* stick insects. (a) Schematic of analytical approach for predicting evolution using temporal autocorrelation. Black points represent

observed values in a time series. Some number of these observed points are removed (in this case the six right-hand most points) and the missing values for them are predicted using the remainder of observed points. Predictive power is the strength of association between observed and predicted values. (b) Color morph frequencies through time. (c) Pattern morph frequencies through time. (d) Change in color morph frequencies. (e) Change in pattern morph frequencies. (f) Predicting change in color morph frequencies (r^2). (g) Predicting change in pattern morph frequencies (r^2). (h) Predicting change in color morph frequencies (r). The difference from panel (f) is that r-values are not squared such that their sign is evident. Shaded areas are 95% confidence intervals. (i) Predicting change in pattern morph frequencies (r). The difference from panel (g) is that r-values are not squared such that their sign is evident. Shaded areas are 95% confidence intervals.

Figure 4. Complex patterns of natural selection. (a) Associations between the frequency of melanistic morphs and yearly spring temperature. Positive effects indicate increases in melanistic frequency with increased temperature. Significant effects are shown in red. The left-most data point represents the average effect across populations, and the remaining points are for individual populations. Note that among-population variation is high, but that all significant effects are positive. (b) Morph-specific survival time in thermoregulatory (i.e., heat tolerance) lab experiments. (c) Genotype specific fitness in the within-generation experiment on the host *Adenostoma* (*s/s* is most fit). Shown are the posterior probabilities from estimates of genotype-specific fitness. (d) Genotype specific fitness in the natural population FHA on *Adenostoma* (*s/m* is most fit). Shown are the posterior probabilities from estimates of genotype-specific fitness.

Figure 5. Evidence for negative-frequency dependent selection on pattern, and resulting stability in morph frequency differences between hosts. (a) Posterior probability estimates of the selection coefficient in each treatment. Positive values on the x-axis represent selection favoring striped individuals. (b) Changes in the frequency of striped morphs in releases at 20% of the population and 80% of the population during the course of the experiment. (c) Posterior probability distributions of change in the frequency of striped morph per treatment. (d) Yearly differences between host plant species in natural populations in the frequency of striped morphs (lines denote posterior medians and shaded regions given the 95% equal-tail probability intervals).

Figure 6. Predicting evolution in different systems. (a) Predicting evolution for evolutionary time series in (left to right) *Geospiza fortis* body size, *Geospiza fortis* beak morphology (principle components 1 and 2), *G. scandens* body size, *G. scandens* beak morphology (principle components 1 and 2), *Panaxia dominula* and *Biston betularia* morph frequencies, and *T. cristinae* color and pattern morph frequencies. Boxplots show the distribution of r^2 between true and predicted evolutionary change across 3 to 10 year model-based forecasts. (b) Predicting evolution in studies (r). The difference from panel (a) is that r-values are not squared such that their sign is evident.

References

1. P. R. Grant, B. R. Grant, Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**, 707-711 (2002).
2. M. Lässig, V. Mustonen, A. M. Walczak, Predicting evolution. *Nature Ecology & Evolution* **1**, 0077 doi:0010.1038/s41559-41017-40077 (2017).
3. D. L. Hartl, A. G. Clark, *Principles of population genetics, fourth edition*. (Sinauer, 2007).
4. A. Ozgul, S. Tuljapurkar, T. G. Benton, J. M. Pemberton, T. H. Clutton-Brock, T. Coulson, The Dynamics of Phenotypic Change and the Shrinking Sheep of St. Kilda. *Science* **325**, 464-467 (2009).
5. T. Coulson, D. R. MacNulty, D. R. Stahler, B. Vonholdt, R. K. Wayne, D. W. Smith, Modeling Effects of Environmental Change on Wolf Population Dynamics, Trait Evolution, and Life History. *Science* **334**, 1275-1278 (2011).
6. J. Merila, Evolution in response to climate change: In pursuit of the missing evidence. *Bioessays* **34**, 811-818 (2012).
7. M. Bosse, L. G. Spurgin, V. N. Laine, E. F. Cole, J. A. Firth, P. Gienapp, A. G. Gosler, K. McMahon, J. Poissant, I. Verhagen, M. A. M. Groenen, K. van Oers, B. C. Sheldon, M. E. Visser, J. Slate, Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science* **358**, 365-368 (2017).

8. M. Chouteau, V. Llaurens, F. Piron-Prunier, M. Joron, Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8325-8329 (2017).

9. D. I. Bolnick, W. E. Stutz, Frequency dependence limits divergent evolution by favouring rare immigrants over residents. *Nature* **546**, 285-+ (2017).

10. R. Lande, Quantitative genetic analysis of multivariate evolution, applied to brain-body size allometry. *Evolution* **33**, 402-416 (1979).

11. C. Rueffler, T. J. M. Van Dooren, O. Leimar, P. A. Abrams, Disruptive selection and then what? *Trends Ecol. Evol.* **21**, 238-245 (2006).

12. X. Thibert-Plante, A. P. Hendry, The consequences of phenotypic plasticity for ecological speciation. *J. Evol. Biol.* **24**, 326-342 (2011).

13. J. G. Kingsolver, H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, P. Beerli, The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245-261 (2001).

14. G. Bell, Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philos. Trans. R. Soc. B-Biol. Sci.* **365**, 87-97 (2010).

15. P. F. Colosimo, K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, D. M. Kingsley, Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928-1933 (2005).

16. S. M. Rogers, P. Tamkee, B. Summers, S. Balabahadra, M. Marks, D. M. Kingsley, D. Schluter, GENETIC SIGNATURE OF ADAPTIVE PEAK SHIFT IN THREESPINE STICKLEBACK. *Evolution* **66**, 2439-2450 (2012).

17. C. Heliconius Genome, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94-98 (2012); published online Epub2012-Jul-5 (

18. G. G. Simpson, *Tempo and mode in evolution*. (Columbia University Press, New York, 1944).

19. G. G. Simpson, *The major features of evolution*. (Columbia University Press, New York, 1953).

20. A. M. Siepielski, J. D. DiBattista, S. M. Carlson, It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecol. Lett.* **12**, 1261-1276 (2009).

21. P. R. Grant, B. R. Grant, *40 Years of Evolution: Darwin's Finches on Daphne Major Island*. (Princeton University Press, Princeton, 2014).

22. T. Dobzhansky, *Genetics and the Origin of Species*. (Columbia University Press, New York, NY, ed. 3rd, 1951), pp. 364.

23. D. Lindtke, K. Lucek, V. Soria-Carrasco, R. Villoutreix, T. E. Farkas, R. Riesch, S. R. Dennis, Z. Gompert, P. Nosil, Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Mol. Ecol.* **26**, 6189-6205 (2017).

24. A. A. Comeault, S. M. Flaxman, R. Riesch, E. Curran, V. Soria-Carrasco, Z. Gompert, T. E. Farkas, M. Muschick, T. L. Parchman, T. Schwander, J. Slate, P. Nosil, Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. *Current Biology* **25**, 1-7 (2015).

25. V. Soria-Carrasco, Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S. Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, S. P. Egan, B. J. Crespi, P. Nosil, Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science* **344**, 738-742 (2014).

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

27. P. Nosil, Divergent host plant adaptation and reproductive isolation between ecotypes of Timema cristinae walking sticks. *Am. Nat.* **169**, 151-162 (2007).

28. N. Cressie, C. K. Wikle, *Statistics for Spatio-Temporal Data*. (John Wiley and Sons, 2011).

29. Z. Gompert, Bayesian inference of selection in a heterogeneous environment from genetic time-series data. *Mol. Ecol.* **25**, 121-134 (2016).

30. Z. Gompert, A. A. Comeault, T. E. Farkas, J. L. Feder, T. L. Parchman, C. A. Buerkle, P. Nosil, Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17**, 369-379 (2014).

31. M. Foll, H. Shim, J. D. Jensen, WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular ecology resources* **15**, 87-98 (2015).

32. N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, R. E. Green, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* **26**, 342-350 (2016).

33. R. Riesch, M. Muschick, D. Lindtke, R. Villoutreix, A. A. Comeault, T. E. Farkas, K. Lucek, E. Hellen, V. Soria-Carrasco, S. R. Dennis, C. F. de Carvalho, R. J. Safran, C. P. Sandoval, J. L. Feder, R. Gries, B. J. Crespi, G. Gries, Z. Gompert, P. Nosil, Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology and Evolution* **1**, 0082 (2017).

34. C. Sandoval, Persistence of a walking-stick population (Phasmatoptera : Timematodea) after a wildfire. *Southwestern Naturalist* **45**, 123-127 (2000).

35. P. Nosil, Reproductive isolation caused by visual predation on migrants between divergent environments. *Proc. R. Soc. B-Biol. Sci.* **271**, 1521-1528 (2004).

36. P. Nosil, B. J. Crespi, Experimental evidence that predation promotes divergence in adaptive radiation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9090-9095 (2006).

37. M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel, I. A. van de Leemput, S. A. Levin, E. H. van Nes, M. Pascual, J. Vandermeer, Anticipating Critical Transitions. *Science* **338**, 344-348 (2012).

38. M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, B. Walker, Catastrophic shifts in ecosystems. *Nature* **413**, 591-596 (2001).

39. P. Nosil, J. L. Feder, S. M. Flaxman, Z. Gompert, Tipping points in the dynamics of speciation. *Nature Ecology and Evolution* **1**, 0001 (doi:0010.1038/s41559-41016-40001) (2017).

40. L. M. Cook, D. A. Jones, The medionigra gene in the moth Panaxia dominula: The case for selection. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* **351**, 1623-1634 (1996).

41. L. M. Cook, S. L. Sutton, T. J. Crawford, Melanic moth frequencies in Yorkshire, an old English industrial hot spot. *Journal of Heredity* **96**, 522-528 (2005).

42. A. P. Hendry, *Eco-evolutionary dynamics*. (Princeton University Press, Princeton, New Jersey, 2017).

43. T. W. Schoener, The Newest Synthesis: Understanding the Interplay of Evolutionary and Ecological Dynamics. *Science* **331**, 426-429 (2011).

44. T. E. Farkas, T. Mononen, A. A. Comeault, I. Hanski, P. Nosil, Evolution of Camouflage Drives Rapid Ecological Change in an Insect Community. *Current Biology* **23**, 1835-1843 (2013).

Supplementary References

45. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354 (2012).

46. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

47. Z. Gompert, L. K. Lucas, C. C. Nice, C. A. Buerkle, GENOME DIVERGENCE AND THE GENETIC ARCHITECTURE OF BARRIERS TO GENE FLOW BETWEEN LYCAEIDES IDAS AND L-MELISSA. *Evolution* **67**, 2498-2514 (2013).

48. Y. S. Aulchenko, S. Ripke, A. Isaacs, C. M. Van Duijn, GenABEL: an R library for genorne-wide association analysis. *Bioinformatics* **23**, 1294-1296 (2007).

49. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).

50. P. Rastas, F. C. F. Calboli, B. C. Guo, T. Shikano, J. Merila, Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F-2 Recombinant Crosses as an Example. *Genome Biology and Evolution* **8**, 78-93 (2016).

51. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

52. T. Schwander, B. J. Crespi, Multiple Direct Transitions from Sexual Reproduction to Apomictic Parthenogenesis in Timema Stick Insects. *Evolution* **63**, 84-103 (2009).

53. S. V. Angiuoli, S. L. Salzberg, Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334-342 (2011).

54. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis

Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).

55. Z. Gompert, J. P. Jahner, C. F. Scholl, J. S. Wilson, L. K. Lucas, V. Soria-Carrasco, J. A. Fordyce, C. C. Nice, C. A. Buerkle, M. L. Forister, The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Mol. Ecol.* **24**, 2777-2793 (2015).

56. O. J. T. Briet, P. H. Amerasinghe, P. Vounatsou, Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Numbers. *PLoS One* **8**, (2013).

57. M. C. Jones, Randomly choosing parameters from the stationary and invertibility region of autoregressive-moving average models. *Applied Statistics* **36**, 134–138 (1987).

58. C. P. Sandoval, Differential Visual Predation on Morphs of Timema-Cristinae (Phasmatodeae, Timemidae) and Its Consequences for Host-Range. *Biol. J. Linnean Soc.* **52**, 341-356 (1994).

59. Q. D. Team, **QGIS Geographic Information System. Open Source, Geospatial Foundation Project.** http://www.qgis.org/. (2016).

60. T. Therneau, P. Grambsch, *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag (2000).

61. T. Therneau. A Package for Survival Analysis in S. version, 2.38, <URL: http://CRAN.R-project.org/package=survival>. (2015).

62. R. D. C. Team. (Vienna, Austria, 2013).

63. Z. Gompert, F. J. Messina, Genomic evidence that resource-based trade-offs limit host-range expansion in a seed beetle. *Evolution* **70**, 1249-1264 (2016).

Supplementary Materials for

Natural Selection And The Predictability Of Evolution In *Timema* Stick Insects

Patrik Nosil, Romain Villoutreix, Clarissa F. de Carvalho, Timothy E. Farkas, Víctor Soria-Carrasco, Jeffrey L. Feder, Bernard J. Crespi, Zach Gompert

correspondence to: p.nosil@sheffield.ac.uk and zach.gompert@usu.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S6
Tables S1 to S3
Caption for database S1

Other Supplementary Materials for this manuscript includes the following:

Database S1 [Raw population data]

Materials and Methods

Approach for delimiting the genetic region affecting color and color-pattern

Previous work showed that each morph in *T. cristinae* is a chromosomal form underlain by a haplotype on a single linkage group (LG8), with restricted recombination between chromosomal forms (*23, 24*). However, it relied on a fragmented reference genome such that it could not delimit a single, contiguous region (i.e., locus) underlying each morph. We here delimit the locus underlying the morphs and quantify its change through time relative to the rest of the genome.

To do so, we generated higher-quality reference genomes for a melanistic and a green morph of *T. cristinae* using Dovetail hi-rise scaffolding of Illumina reads (N50 = ~16, 8 megabases, respectively)(*32*). Comparison of the reference genomes, linkage mapping (*25*), and genome-wide association (GWA) mapping allowed us to explicitly delimit a single, contiguous genomic region associated with color and pattern variation (Figure 2, S1, S2). Accordingly, this region exhibits three core haplotypes (i.e., alleles), one corresponding to each morph (with melanistic recessive to green body coloration and stripe recessive to unstriped pattern), and we refer to it as the *Mel-Stripe* locus hereafter. Details are contained below.

Reference genome with Dovetail

We generated reference genomes for a melanistic and a green morph using Dovetail technology (*32*). For the melanistic morph we used two sequencing runs. The first run (short reads + Chicago library) was done on a melanistic female from FHA caught in 2015 (id: 15_0190). The second run (Chicago library only) was done using another melanistic female caught in 2016 in FHA (id 16_0359). For the green reference (short reads + Chicago libraries), a green unstriped female from population PRC caught in 2015 was used (id 15_0802). The 2015 samples were flash frozen in liquid nitrogen, shipped to Sheffield and stored in a -80°C freezer. It was de-gutted prior to shipping to Dovetail. The 2016 sample was caught and degutted ‘fresh’ in California and sent directly to Dovetail.

The Dovetail assembly method relies on building a conventional reference assembly using Meraculous with paired-end Illumina reads and then using Chicago libraries for scaffolding using the HiRise pipeline (*32*). Chicago libraries are produced by reconstituting chromatin *in vitro* with chaperones and histones, followed by crosslinking (i.e. DNA stabilization by creating covalent bonds among the histones), digestion with restriction enzymes, and ligation. This process results in many chimeric fragments composed from physically distant regions, but ensures they come from the same stabilized large fragment. In theory, the read pairs produced can have separations up to the maximum fragment size of the DNA. A model of insert distribution derived from the distances among the original fragments is then used for scaffolding.

The assembly based on melanistic females (draft 1.3) had a 63.0x sequencing depth with a total length of 953.3 Mb (73.3% of the estimated genome size by flow cytometry)(*25*). It comprised 4068 scaffolds (N50=16.4 Mb, N90=1.1 Mb, L50=16, L90=135), a significant improvement relative to the previous draft 0.3 (14,221 scaffolds, N50=312.5 Kb, N90=52 Kb, L50=788, L90=3869; DDBJ/ENA/GenBank accession MSSY00000000.3)(*33*). We clustered scaffolds in major linkage groups as described in detail below in the section on delimitation of the *Mel-Stripe* locus, resulting in draft 1.3c2. This Whole Genome Shotgun project was deposited at DDBJ/ENA/GenBank under the accession PGFK00000000. The version described in this paper is version PGFK01000000. The assembly based on the green female had a 42.7x sequencing depth with a total length of 932.1Mb (71.1% of the estimated size). This assembly was poorer than the 1.3, but still significantly better than the previous 0.3 (5653 scaffolds, N50=8.2 Mb, N90=503.2 Kb, L50=22, L90=222). This assembly was labeled as draft 2.1. This Whole Genome Shotgun project was deposited at DDBJ/ENA/GenBank under the accession PGTA00000000. The version described in this paper is version PGTA01000000.

Genome-wide association (GWA) mapping

We mapped color and pattern variation using previously published GBS data (*33*) aligned to the new reference genome 1.3b2. We aligned 96.1% (789,388,267) of reads from 602 individuals using BOWTIE 2.2.9 (*45*) with the '--very-sensitive-local' preset. We used SAMTOOLS 1.3.1 (*46*) to sort and index the alignments. We used aligned reads with a mapping quality score of at least 20 to call single nucleotide polymorphisms (SNPs) with SAMTOOLS mpileup and BCFTOOLS 1.3.1 (*46*), using the original consensus caller (-c) with a P-value threshold of 0.05. From the 1,369,070 variants called, we excluded those with quality score of less than 20, sampling coverage of less than 50%, maximum depth more than 10 times the number of total, minor-allele frequency (MAF) equal or less than 0.01, and more than two alleles. The number of phenotyped individuals was different for color (590) and pattern (536) and we subsequently subset variants and applied filters relative to the respective number of samples. Thus, we retained 418,209 bi-allelic variants for color and 416,405 variants for pattern. Both datasets were very similar, showing the same mean coverage depth per SNP per individual of 5.1x (95%: 0-15; per SNP average: 5.1x, 95%: 1.0-9.5; per individual average: 5.1, 95%: 2.2-7.9). We used custom Perl scripts along with a custom C++ program (alleleEst 0.1b) to co-estimate allele frequencies and genotypes using a Bayesian model (*47*). Genotype estimates were stored in BIMBAM format as values ranging from 0 to 2 representing minor allele dosage.

Following past work (*24*) , we used GENABEL v1.8.0 (*48*) to perform single locus GWA mapping analyses. Briefly, we recoded genotype probabilities into genotype values accepted by GENABEL using a custom Perl script as follows: [0-0.5]=homozygote for major allele, [0.5-1.5]=heterozygote, [1.5-2]=homozygote for minor allele. Transformed genetic probabilities were filtered using GENABEL quality control function. SNPs with MAF inferior or equal to 1%, if any, were excluded from analysis. Individuals with extreme heterozygosity at a false discovery rate <1% and too high an identity by state (hereafter IBS >=0.95, calculated on a subset of 2000 SNPs), if any, were discarded from analysis.

Analyses were run controlling for population structure using the GENABEL egscore function (*49*). This function extracts principal components of a kinship matrix (here IBS indices) calculated using a subset of 2000 SNPs. The principal components are then used as covariates in the GWA linear models. The kinship matrix was computed excluding markers on linkage group 8 (to avoid over-correcting for genome-wide population structure by including causal variants), and excluding markers that were not assigned to linkage groups. We display results in the form of Manhattan plots. These graphics shows the association score (expressed as –$\log_{10}$(*pvalue*)) of every SNP tested along their physical position on the genome. Gaps between scaffolds are not represented on these graphics. SNP with a significant P-value after Bonferroni correction (calculated as 0.05/number of tested SNPs) are displayed in red in the Manhattan plots.

Defining the *Mel-Stripe* locus

We combined results from GWA mapping of color and pattern with whole genome comparative alignments and recombination rate estimates from crosses to define approximate boundaries for the main locus responsible for color and pattern variation in *T. cristinae* (Figures S1, S2). We focused on scaffolds 702 and 128 from the melanistic genome, which contained the overwhelming majority of SNPs significantly associated with color (96%) and pattern (73%)(numbers refer to significance following a strict Bonferroni correction, i.e., $P < 0.05$/(no. of tests)). Our approach included the following steps, which we detail below: (i) generate a linkage map with the genome scaffolds, (ii) split one key scaffold (702) based on inconsistencies in the linkage map, (iii) align the green and melanistic morph genomes to each other, (iv) delimit the *Mel-Stripe* locus based on the total evidence from the mapping results and comparative alignment. These boundaries are meant to serve as a working hypothesis for the region controlling color and pattern (which can then be usefully contrasted to the genomic background), and not as the precise boundaries of the functional variant(s).

Linkage map- We used the *LepMap2* software (*50*) and previously published data from three F1 crosses to construct a linkage map for the *T. cristinae* melanistic morph genome sequence scaffolds (the data, comprising 158 million ~100 base pair, bp, genotyping-by-sequencing reads, are fully described in (*25*)(NCBI BioProject PRJNA356911). Families consisted of 114 (female melanistic by male green), 48 (female green by male melanistic), and 24 (female green striped by male melanistic) full-sib offspring. However, note that the GWA described above used a draft (1.3c2) based on only the largest family. Sequence data for the parents and offspring were aligned to the melanistic morph genome using *bwa aln* and *samse* (version 0.7.10-r789)(*51*) with a maximum of 4 miss-matches, and not more than 2 miss-matches in a 20 bp seed. We then compressed, sorted and indexed the alignments using SAMTOOLS (1.2)(*46*), and identified variable nucleotides using the *call* variant caller in BCFTOOLS (version 1.3)(*46*). We only considered alignments with a mapping quality of 10 or more and bases with a base quality of 15 or more, and we applied a population prior with theta set to 0.001 when calling variants and only considered a SNP if the probability of the data assuming the locus was invariant was less than 0.01. We then applied a variant filter using vcfutils varFilter to retain only those SNPs with a total read depth of 464 and that were more than 5 bp from the nearest gaps (insertion-deletions).

We then generated the genotype input data for the mapping program, *LepMap2*. In doing so, we used custom Perl scripts to select the subset of SNPs that were recombination informative for each parent, and then estimated offspring genotype posterior probabilities using the genotype likelihood from BCFTOOLS (*46*)(from the vcf file) with a prior given by Mendelian inheritance expectation. We then only retained genotypes when the posterior probability of the most probable genotype was 0.95 or greater (in other cases the genotype estimate was converted to missing data). From this, we retained 17,478 SNPs (across all three families) for linkage map construction. As a first step with *LepMap2*, we further filtered the data for each family to retain only markers with missing data from fewer than 10 individuals, and with a P-value for segregation distortion greater than 0.005 (i.e., to remove loci with substantial deviations from Mendelian expectations). We allowed for a data error rate of 0.01. This resulted in a total of 4312 maternally informative SNPs and 5989 paternally informative SNPs.

We next used the *LepMap2 SeparateChromosomes* algorithm with a LOD minimum of 4 and with a minimum linkage group size of 50 SNPs for initial assignment of SNPs to LGs. This resulted in 6873 SNPs being assigned to 12 linkage groups (i.e., autosomes, *T. cristinae* has 13 chromosomes, see below for consideration of the sex chromosome). The *JoinSingles* algorithm was then used to assign additional SNPs to these linkage groups at the lower LOD threshold of 3, if the difference in support between their best and next best possible linkage group differed by 2 LOD units. Next we used a custom Perl script and approach to assign entire scaffolds to linkage groups based on the SNP assignments. Specifically, for a scaffold to be assigned to a linkage group (and thus all of its SNPs to be assigned to a linkage groups) required at least two SNPs (and 10% of all SNPs on a scaffold) to have been assigned to that linkage group, and for fewer than half as many SNPs to have been assigned to the next best linkage group. Based on this, we were able to assign 237 scaffolds (which accounted for 89% of all SNPs) to linkage groups. Finally, the *OrderMarkers* algorithm in *LepMap2* was used to estimate marker/SNP order on each linkage group. We took the median position in cM for all markers on a scaffold as the position for each scaffold in each cross.

As one of the filters applied with *LepMap2* was to remove markers with non-Mendelian patterns of inheritance, we expected to miss the sex (i.e., X) chromosome, and thus to find 12 of the 13 chromosomes, as we did. We thus employed a complementary approach to identify the X-linked scaffolds (in *T. cristinae* males are XO and females are XX)(*52*). Using SAMTOOLS DEPTH (version 1.2)(*51*) and custom Perl scripts, we extracted the coverage data from a previously published GBS data set that was used for genome-wide association mapping and comprised 395 female and 197 male *T. cristinae* (data from (*24*), but aligned to the current genome as described above; we lacked data on offspring sex in the mapping families so used this data instead). We then identified scaffolds where the ratio of read depth for males to females was less than the expected 1:1 ratio expected for autosomal markers (specifically less than 0.75). Twenty-nine scaffolds met this requirement, and also were not assigned to the 12 autosomal scaffolds described above. These included 380 recombination informative markers. Seventeen of these scaffolds were joined into a single linkage group (presumably the X chromosome) using the *SeparateChromosomes* algorithm in *LepMap2* with a LOD limit of 1.5 and a minimum size of 50 SNPs. The 17 scaffolds included 93.4% of the SNPs on the 29 scaffolds we identified as possibly being X-linked based on the coverage ratio. We used *OrderMarkers* to order these markers as described for the X-chromosome.

<u>Splitting and re-mapping scaffold 702</u>- Scaffold 702 from the melanistic morph genome showed a strong association with color and pattern in GWA analyses, but was not originally assigned to a linkage group. Upon examining this further we noted that one large chunk (SNPs up to position 14,171,514) of this scaffold was assigned to linkage group 8 (the linkage group where another scaffold, 128, showed a strong association with color and pattern and where we

had previously seen associations with these traits) whereas a second large chunk (SNPs after position 14,757,049) was assigned to linkage group 5 (preventing placement of this scaffold). The portion assigned to linkage group 8 showed an association with color and pattern, whereas the other half of the scaffold did not. Based on this evidence we inferred that this scaffold was over-assembled and thus we split scaffold 702 into three new scaffolds: 702.1 (positions 1-14,171,514), 702.2 (starting at position 14,757,049) and 702.3 (the middle ambiguous section lacking an informative SNP from 14,171,414-14,757,049). The new scaffolds 702.1 and 702.2 were added to their respective linkage groups and the *OrderMarkers* algorithm in *LepMap2* was re-run for these linkage groups.

Whole genome comparative alignment and defining *Mel-Stripe*- We aligned the melanistic and green morph genomes to each other using *Mugsy* (v1r2.3)(*53*). Our goal was twofold: (i) to refine the orientation of scaffolds 702.1 and 128 (the two scaffolds with the greatest association with color and pattern) based on overlap between these and scaffolds from the green morph genome, and (ii) to identify possible structural variants associated with the GWA color and pattern signal. Scaffold 702.1 (from the melanistic genome) partially aligned to green scaffold 1575; green scaffold 1575 also aligned to melanistic scaffold 2963 (which was 'left' of scaffold 702.1). Melanistic scaffold 2963 showed a negative correlation between SNP map position and physical position, suggesting it was in a reverse orientation. This combined with the overlap of both melanistic scaffolds 2963 and 702.1 with green scaffold 1575 allowed us to also specify (flip) the orientation of 702.1. Melanistic scaffold 128 was in the correct orientation based on the correlation (positive) between SNP physical and cM positions. Many small green morph scaffolds with uncertain orientations span the right side of the re-orientated melanistic scaffold 702.1 and melanistic scaffold 128 (> 15 small scaffolds). No green scaffold was found that aligned the portion of melanistic scaffold 128 from approximately 5 to 6.4 mbps. This region also exhibits lower sequence coverage in green individuals, suggesting it might be a large insertion-deletion polymorphism.

Given this information and the GWA mapping signal, we defined the bounds of a putative *Mel-Stripe* color and pattern locus as comprising melanistic scaffold 702.1 starting from the edge of the alignment with green scaffold 1575 (702.1 4,139,489 bp) to the edge of 702.1 (bp 1, given the reverse orientation) along with the neighboring melanistic scaffold 128 from bp 1 to right edge of the putative insertion-deletion polymorphism (bp 6,414,835). This specific region (that is, the *Mel-Stripe* locus) contains 70% and 31% of SNPs associated with color and pattern, respectively (59% of color or pattern-associated SNPs). As a comparison, this region only contains about 1% of the sequenced SNPs. Thus there is a 61 and 31-fold enrichment of color and pattern associated SNPs, respectively, in the *Mel-Stripe* locus. As emphasized above, our main goal was to delimit a *Mel-Stripe* locus that could be contrasted to the genomic background, and not to precisely identify causal functional variants affecting color and pattern. A schematic summary of the delimitation of *Mel-Stripe* can be found in Figure S1.

Genomic change at the *Mel-Stripe* locus

We quantified changes at *Mel-Stripe* between time periods using three published data sets: (1) genotyping-by-sequencing (GBS) data from 1102 individuals collected in a natural population on *Adenostoma* (FHA) in 2011 and 2013 (n = 500 and 602, respectively)(*30, 33*), (2) 491 re-sequenced whole genomes from an eight-day (i.e., within-generation) release and recapture field experiment (*30*), and (3) GBS data from 451 individuals in a between-year (i.e., between-generation) field transplant experiment (*25*). The within-generation experiment involved releasing 500 *T. cristinae* in a paired-block design and recapturing the survivors (*30*). We obtained whole genome sequence data from 491 of these individuals (*33*), allowing us to compare allele frequency changes between release and recapture. As described previously (*25*), the between-generation experiment involved transplanting 2000 stick insects from a single variable population (OGA) onto 10 host plant bushes in a block design (five blocks each with one *Adenostoma* bush and one *Ceanothus* bush per block; 200 *T. cristinae* were released on each bush). 421 F1 descendants of these individuals were then captured the following year (2011). We compared 30 individuals representative of the founders (collected in 2010) to the 421 F1s. Phenotypic change (proportion at time period two minus proportion at time period one) for each of these three data sets was as follows: FHA, stripe change = 0.06, unstriped change = -0.11, melanistic change = 0.05; within-generation experiment, stripe change = 0.05, unstriped change = -0.04, melanistic change = 0.01; between-generation experiment, stripe change = -0.24, unstriped change = 0.32, melanistic change = -0.07).

The GBS data were aligned to the *T. cristinae* reference genome with *bwa* (version 07.10-r789)(*51*) using the *aln* and *samse* algorithms. We allowed 5 miss-matches, 2 miss-matches in an initial 20 bp seed, trimmed bases with phred-scaled quality scores lower than 10, and only placed reads with a single best match. We then used SAMTOOLS (version 1.2)(*46*) and the BCFTOOLS call algorithm (version 1.3)(*46*) to identify SNPs and calculate genotype likelihoods. We used the recommended mapping quality adjustment (-C 50), skipped alignments with a mapping quality less than 20 and bases with a base quality less than 30, and used the multi-allelic SNP caller with θ set to 0.001 and a posterior probability of 0.01 or less for the homozygous reference genotype given the data to consider a SNP variable. We then filtered the initial set of SNPs to retain only those with a mean coverage of >= 2X (per individual), total coverage (across all individuals) less then three standard deviations above the mean across all loci, at least 10 reads of the non-reference allele, a mapping quality of 30, sequence data for at least 80% of the individuals, a minimum minor allele frequency of 0.01, less then 1% of reads in the reverse orientation (with our GBS method all reads should be in the same orientation), and separated by at least 5 bps. Filtering was done using custom Perl scripts. Following filtering, we retained 178,141 SNPs for the natural FHA population and 249,074 SNPs for the between-generation experiment.

We aligned the whole genome re-sequence data from the within-generation experiment to our reference genome using the *bwa* (version 07.10-r789)(*51*) mem algorithm with a band width of 100, a 20 bp seed length and a minimum score for output of 30. We then used SAMTOOLS (*46*) to compress, sort and index the alignments, and *Picard Tools* to mark and remove PCR duplicates (version 2.1.1)(https://broadinstitute.github.io/picard/). We then used the *GATK* HaplotypeCaller and GenotypeGVCFs modules (version 3.7)(*54*) to call variants and calculate genotype likelihoods. We used a minimum base quality score of 30 for consideration in calculations, a prior probability of heterozygosity of 0.001, and called variants with a minimum phred-scaled confidence of 50. The following filters were then applied using custom Perl scripts: minimum coverage of 1x per individual, a minimum value of the base quality rank sum test of -8, a minimum value of the mapping quality rank sum test of -12.5, a minimum value of the read position rank sum test of -8, a minimum ratio of variant confidence to non-reference read depth of 2, a minimum mapping quality of 40, a maximum phred-scaled P-value of Fisher's exact test for strand bias of 60, and a minimum minor allele frequency of 0.01. The resulting 6,175,495 SNPs were used for downstream analyses.

We obtained maximum likelihood estimates of allele frequencies for all populations / experimental samples using an expectation-maximization (EM) algorithm, as described in (*55*). For this, we used a convergence tolerance of 0.001 and allowed for a maximum of 20 EM iterations.

Population genomic parameters were then calculated based on the *Mel-Stripe* locus and additional reference loci based on the maximum likelihood allele frequency estimates. Additional loci were defined for all genome scaffolds placed on linkage groups that contained as many SNPs as *Mel-Stripe* and were defined by selecting (at random) a contiguous block of SNPs of the same number as *Mel-Stripe* (FHA: 780 SNPs, 40 reference loci; between-generation experiment: 1180 SNPs, 39 reference loci; within-generation experiment: 47,305 SNPs, 16 reference loci).

We analyzed genomic change based on raw allele frequency changes, allele frequency changes controlling for underlying genetic diversity (i.e., residual change), and using Wright's Fixation Index (F_{ST}). Specifically, we calculated nucleotide diversity (π) within the 2011 FHA sample or the founders of each experiment, allele frequency change between these samples and the 2013 FHA sample (natural FHA population) or recaptured stick insects (both experiments), the residuals from regressing change on diversity, and $F_{\mathrm{ST}} = \Sigma(\pi_{\mathrm{total}} - \pi_{\mathrm{subpop}})/\Sigma(\pi_{\mathrm{total}})$. In all cases, *Mel-Stripe* showed the most extreme change (more than any other locus). Detailed results are as follows. For FHA, raw change was = 0.0273, residual change was = 0.00516, and F_{ST} was = 0.0051 (P = 0.024, Exact probability). For the within-generation experiment, raw change was = 0.0340, residual change was = 0.00212, and F_{ST} was = 0.0030 (P = 0.059). For the between-generation experiment, raw change was = 0.0988, residual change was = 0.0595, and F_{ST} was = 0.0540 (P = 0.025; Fisher's combined probability test across data sets: X^2 = 20.50, d.f. = 6, P = 0.0023).

Autoregressive-moving-average models fit to different long-term evolutionary data sets

We fit Bayesian autoregressive-moving-average (ARMA) models to 10 evolutionary time-series data sets (details of each data set are given below; two are from *T. cristinae* and the others from published data in other systems). This approach uses past observations as covariates in a model. There are two specific types of terms in these models, autoregressive terms (AR) and moving-average terms (MA). AR terms use the data values from prior years as covariates whereas the MA terms use residuals from prior years as covariates. Different numbers of prior years (i.e., different order models) can be considered.

Specifics of the models are as follows. We first considered models with order 0, 1 or 2 for the auto-regressive and moving-average components of the model; a null model with a constant expectation was included for comparison. As an example, ARMA (1,2) denotes a model with order 1 for the autoregressive component and order 2 for the moving-average component, meaning that information from the last year is used for the autoregressive component and that information from the last two years is used for the moving-average component. The general form of the model is $y_t \sim \text{Normal}(\mu_t , \tau)$ and $\mu_t = c + \Sigma_i \, \theta_i \, y_{(t-i)} + \Sigma_j \, \varphi_j \, \varepsilon_{(t-1)}$, where $y_{(t-i)}$ is the data value *i* years in the past, $\varepsilon_{(t-1)}$ is the error term from *j* years, and the sums are over the order of the autoregressive and moving-average components of the model. We assumed a weakly stationary model and thus applied the re-parameterization and Beta prior scheme proposed by (*56, 57*). We placed a normal prior on the grand mean, $c \sim$ Normal (mean = 0, precision = 0.01), and gamma prior on the precision for the sampling distribution, $\tau \sim$ gamma (0.01, 0.001).

Each model was fit for each data set and the best model was selected based on deviance information criterion (DIC; the model with the lowest DIC was chosen). When the null model was best, the next best model was used for downstream analyses (the null model would not provide meaningful results for cross-validation or forecasting as the expectation would be the same for each year). Two estimates of DIC were obtained for each model (to verify consistency), each based on 10 Markov chain Monte Carlo (MCMC) chains each with 100,000 iterations, a 50,000 step burn-in and a thinning interval of 50. MCMC analyses were conducted using the *rjags JAGS* interface.

We then quantified the predictability of each evolutionary time series using the best ARMA model. We used two complementary approaches: leave-one-out cross-validation and forecasting. For leave-one-out cross-validation, we fit the relevant ARMA model for each data set, but with one year of the data set removed (this was done with each year in turn). The missing year's data value (evolutionary change) was then predicted from the ARMA model. We used these estimates to assess the relationship (based on a simple linear model) between the true and predicted evolutionary change.

For forecasting, we dropped the most recent *n* years of data, where *n* was (3, 4, ..., 9, 10), and fit the relevant ARMA model to predict the data values for the dropped data. We then calculated the Pearson correlation coefficient and coefficient of determination between the observed and predicted (forecast) change for the dropped years for each value of *n*. This is conceptually analogous to predicting/forecasting future (as of yet unobserved) evolutionary change. Cross-validation and forecasting results were also based on average of results from two independent MCMC model fits, each comprising 10 chains with 100,000 iterations, a 50,000 iteration burn-in and a thinning interval of 50.

The data analyzed include evolutionary time series for discrete trait frequencies, and in the case of Darwin’s finches, quantitative traits (mean value). In both cases, we first obtained point estimates of the value (mean or frequency) for each generation and then converted these into evolutionary change data sets (i.e., the data point for year *i* was the value [mean or frequency] in year *i+1* minus the value in year *i*). The nature and source of each data set are described below. Results are provided in the main text, Database S1, and Figures S3-S6.

Long-term field studies in *T. cristinae*

We compiled data on morph frequencies in *T. cristinae* using samples collected in the spring using sweep nets between 1990 and 2017. All individual were scored as ‘striped’, ‘unstriped’, or ‘melanistic’, or occasionally when it was difficult to distinguish between the first two categories as 'intermediate-striped'. These classifications have been found to be highly repeatable in past work (*26, 35, 36, 58*). Samples from 1990 to 1999 were taken and scored by Cristina Sandoval, who then trained PN in 2000. PN collected and scored most samples from 2000 to 2017.

GPS coordinates of all localities were taken at and then used to estimate elevation using 'point sampling tool' on QGIS 2.16.2 (*59*). The elevation values were extracted from 1/3 arc-sec Digital Elevation Models (DEM) at the location of the populations' coordinates. All DEMs were obtained from United States Geological Survey Dataset (USGS), available at National Map Viewer (https://viewer.nationalmap.gov/). Host-plant collected on (*Ceanothus* or *Adenostoma*) was recorded for all individuals. We estimated the proportion of individuals in a sample that were striped (% striped) using all striped and unstriped individuals (excluding melanistic individuals). We estimated the proportion of individuals in a sample that were melanistic (% melanistic) using all individuals. Detailed information on these localities (i.e., GPS coordinates and elevations), morph frequencies, sample sizes, etc. is provided in Database S1.

We observed consistent year-to-year increases and then decreases in the frequency of striped morphs at HV (see main text). We thus computed the binomial probability of the observed stripe time series alternating between an increase and decrease in stripe frequency every other year. Specifically, conditional on the first year, we calculated the probability that every other year showed a reversal in the direction of evolution as $0.5^{17} = 7.6e^{-6}$ (the full time series includes 18 years, the null probability that evolution reverses direction was assumed to be 0.5, and thus the probability of not changing direction was also 0.5).

Climatic data and analyses

We collated data on mean springtime statewide temperature in California using publicly available records (National Centers for Environmental Information, https://www.ncdc.noaa.gov/cag/time-series/us/4/0/tavg/3/4/1990-2016?base_prd=true&firstbaseyear=1901&lastbaseyear=2000). We focused on temperature averages across March, April, and May as these are the three months that *T. cristinae* is by far most active (most of the rest of the year is spent in egg diapause)(*26, 34, 58*). Nonetheless, we present results from different combinations of spring months below.

We fit a hierarchical Bayesian model for the full *T. cristinae* color data set, using data from all populations (i.e., not just HV) collected from 1990 to 2017. We did so to: (i) test for an association between climate and the melanistic morph frequency, and (ii) determine how well climate predicts color morph frequency across space and time.

We assumed a binomial sampling distribution for the observed number of melanistic morphs for a site and year (y_{ij}) given the number of *T. cristinae* sampled (n_{ij}) and the true melanistic morph frequency (p_{ij}). We connected this to a linear model with the logit link function, such that $\text{logit}(p_{ij}) = \alpha_i + \beta_i x_{\text{temperature}} + \theta \, \text{x}_{\text{year}} + \varepsilon_{ij}$, where α_i is a population (site) specific intercept, β_i denotes the effect of climate (temperature, see details below) on melanistic morph frequency for population i, θ is an overall effect of year (allowing for a general increase or decrease in melanistic morph frequency), and ε_{ij} is an error term that accounts for overdispersion relative to binomial sampling. We gave the ε values a normal prior with mean of 0 and precision parameter $\tau \sim$ gamma(0.1, 0.01) (we imposed a sum-to-zero constraint on the ε values). We then defined linear models at the next level of the hierarchy for the population specific α and β coefficients (for the intercept and effect of temperature, respectively), such that,

$$\alpha_i = a_1 + b_1\, x_{elevation} + c_1\, x_{host} + d_1\, x_{mountain}$$

$$\beta_i = a_2 + b_2\, x_{elevation} + c_2\, x_{host} + d_2\, x_{mountain}$$

Here, $x_{elevation}$ is the elevation at a location, x_{host} is a binary indicator variable for host plant (*Adenostoma* = 0, *Ceanothus* = 1), $x_{mountain}$ is a binary indicator variable denoting the mountain range (0 = Highway 154; 1 = Refugio), and a_1, a_2, b_1, b_2, c_1, c_2, d_1 and d_2 are regression coefficients (all given Normal priors with mean 0 and precision 0.0001).

We fit this model with three different temperature variables: (i) mean temperature for March, April and May (when *T. cristinae* are most active), (ii) mean temperature for February, March, April and May, and (iii) mean temperature for February, March and April. We used the *rjags* interface with JAGS to obtain Markov chain Monte Carlo (MCMC) parameter estimates for the model parameters. In each case, we ran three chains, each with a 10,000 iteration burn-in, 25,000 post burn-in iterations and a thinning interval of 10. We used four-fold cross-validation to determine the predictive power of the models. Specifically, we split the data set into four random subsets (only considering cases where the sample size was 25 or greater) and used three subsets to fit the model and validated the model by predicting morph frequencies for the other subset (MCMC options identical to those for the main models were used).

Temperature was generally associated with a higher frequency of melanistic *T. cristinae* (a_2 was positive), but less so at *Ceanothus* sites (c_2 was negative) (Table S2; estimates of the effect of temperature for each site and year are shown in the main text). Melanistic morphs were less common at higher elevations and on Refugio independent of temperature. Cross-validation results showed that the models had significant but modest predictive power. For example, with the March, April, May temperature model, the Pearson correlation between observed and predicted melanistic morph frequencies was r = 0.16 (95% CIs = 0.040-0.28, P = 0.0102, r^2 from a linear model = 0.027). The other temperature variables gave similar results: February, March, April, May temperature, r = 0.15 (95% CIs = 0.025-0.27, P = 0.0188, r^2 from a linear model = 0.022); February, March, April temperature, r = 0.19 (95% CIs = 0.069-0.31, P = 0.0024, r^2 from a linear model = 0.037).

Thermoregulatory experiments

We conducted lab thermoregulatory experiments testing the desiccation / heat tolerance of green versus melanistic *T. cristinae*. Heat stemmed from a desk lamp (K-mart model ksn: 0-02546202-9), raised 4.5 inches above two petri dishes that were stacked on top of each other and pushed to touch the base of the lamp. A third petri dish containing an individual *T. cristinae* was placed on top of the other two. The bulb used a Sylvania A19 halogen 100-watt replacement that used 72-watts. A total of four such lamp set-ups were used, allowing simultaneous assays of four *T. cristinae* (always two green and two melanistic, assigned randomly to one of the four lamps at the initiation of an assay, and then randomly re-assigned to one of the four after each weighing census, see below). Details of the procedure were as follows. Each individual was weighed. Each lamp was then turned on for ten minutes. Placing test animals underneath the lamps then started the trials. Every twenty minutes all four individuals were removed simultaneously and weighed in a random order, and scored as dead or alive. They were then assigned randomly back to one of the four test lamps. This procedure was repeated until 180 minutes had passed. A total of eight sets of such trials were run (total n = 32).

We fit a Cox proportional hazards model to the survival data to test for an effect of morph (green versus melanistic) on survival (*60*). For this, we used the *survival* package in R (*61*). We used the exact partial likelihood method, which is advantageous relative to the more common Efron method when time is measured in discrete intervals and tied times of death are thus more likely. We detected a significant effect of morph on survival time (exp(B) = 3.57, 95% CIs = 1.34-9.51, $P = 0.0111$). Note that exp(B) > 1 indicates melanistic morphs died from desiccation more rapidly than green morphs.

Estimating genotype-specific fitness using genomic data

We estimated selection coefficients/relative fitnesses for different genotypes at the *Mel-Stripe* locus based on the within-generation release-recapture experiment and based on patterns of evolutionary change between the 2013 and 2011 FHA samples. Similar to (*23*) we used PCA and k-means clustering to assign individuals one of six *Mel-Stripe* genotypes: homozygous for the stripe haplotype/allele (*s/s*), homozygous for the green unstriped haplotype (*u/u*), homozygous for the melanistic haplotype (*m/m*), or one of the three possible heterozygotes (*s/u, s/m or u/m*)(Fig. S2). We conducted a PCA on the individual genotype matrix for each of the two data sets. This was done for all individuals and the 780 SNPs comprising the *Mel-Stripe* locus. We then clustered *T. cristinae* based on the first two genetic PCs with k-means clustering; this was done with the R *kmeans* function with six centers, 100 starts and a maximum of 200 iterations. An initial round of clustering was performed to define cluster centers. For this round an equal number of green, striped and melanistic individuals were used (42 of each, which was the number of green individuals). We then used those centers to cluster all individuals with a second round of k-means clustering (this included individuals with no phenotypic data). Assignments from k-means clustering corresponded well with groups of individuals with the same color/pattern (i.e., stripe) phenotype, and were the basis for designating genotypes.

For the within-generation experimental data, we fit a Bayesian beta-binomial model to infer fitness values. Here, we inferred the survival probability of individuals with each genotype using a binomial sampling distribution for the number of recaptures given the probability of survival and recapture (p_{genotype}) and the number of individuals released with that genotype (n_{genotype}). We assigned an uninformative (Jeffery's) beta prior for each survival probability. Posterior samples (N = 5000 each) were obtained from the closed form solution for the posterior using *R* (*62*), and were then used to calculate the relative fitness of each genotype by dividing the survival probability by the survival probability with the highest fitness (based on the point estimate; s/s).

An alternative model was required for the FHA data, which was based on change over two generations (2011 versus 2013). During this time haplotype frequencies went from $m = 0.316$, *s*

= 0.602, and *u* = 0.082 to *m* = 0.360, *s* = 0.570, and *u* = 0.071. Perhaps more importantly, in both years, we detected an excess of the *s/m* heterozygotes (0.514 in 2011 and 0.502 in 2013) relative to Hardy-Weinberg expectations (0.380 and 0.410, respectively). For this analysis, we assumed the following relative fitness values*:* *s/m* = 1 (based on observed patterns of change this genotype appeared to have the highest fitness), *m/m* = 1 + *s1*, *s/s* = 1 + *s2*, *u/u* = 1 + *s2* + *s3*, *s/u* = 1 + *s2* + *s3* * *s4*, and *m/u* = 1 + *s1* + *s3* * *s4*. Thus, *s1* and *s2* define the fitness value of the *m/m* and *s/s* homozygote in a way that allows for any form of dominance. In turn, *s3* defines the fitness of *u/u* relative to *s/s* (i.e., after adding s2). The *s/u* heterozygote is 1 + *s2* + *s3* * *s4*, thus *s4* is the heterozygous effect. This is similar for *m/u*. We took an approximate Bayesian computation (ABC) approach to estimating the selection coefficients. We first sampled selection coefficients from their priors, U(-0.5, 0.5) for s1, s2, and s3, and U(0,1) for s4. We then simulated evolution forward in time for two generations according to a Wright-Fisher model with the observed starting genotype frequencies, and dynamics governed by drift and the sampled the selection coefficients (assuming viability selection). We assumed a variance effective population size of 110.3, which was inferred from patterns of change across 178,141 SNPs (following general procedures outlined in (*63*)). We ran 1,000,000 ABC simulations. We then used the ridge regression adjustment method in the R *abc* package to obtain samples form the posterior distribution from the simulation output. We retained the top 0.5% of simulations with the smallest distance between the simulated and observed genotype frequencies in the 2013 sample. We then converted the estimates of selection coefficients to relative fitnesses.

Field experiment testing for NFDS

We implemented a field transplant experiment testing for NFDS. A total of 1000 individuals were transplanted, collected from March 21-24, 2017 from populations PRNC (latitude 34.53, longitude -119.85), OUTA (latitude 34.53, longitude -119.84), HVC (latitude 34.49, longitude -119.79), and HVA (latitude 34.49, longitude -119.79). Numbers were as follows: green-unstriped morphs, PRCN 220, OUTA 50, HVC 140, HVA 90; green-striped morphs, PRCN 30, OUTA 100, HVA 280, HVC 90. Individuals were kept in groups of 10 and each group was randomly assigned to one of two treatments: striped individuals common (40 striped and 10 unstriped individuals) versus striped individuals rare (10 striped and 40 unstriped individuals). Each of these groups of 50 individuals was then randomly assigned to one of 20 experimental bushes (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 each day March 21-24. Past work demonstrates that this clears bushes of the overwhelming majority of *Timema (25, 26, 58).* Nonetheless, as an additional measure for ensuring accurate identification of experimental animals, each transplanted individual was marked with fine tip sharpie on the underbelly. This mark allowed us to distinguish experimental animals from any remaining residents, and the marks are not visible when *Timema* are resting on leaves. Individuals were released on March 26th between 9am and 3pm. Each individual was released with tweezers onto an experimental plant and checked to cling well to their transplanted host. Individuals were recaptured using visual surveys and sweep nets on March 31st, as in past work *(25, 26, 30, 35, 36, 58)*, and scored as striped or unstriped.

We fit a Bayesian beta-binomial model to assess the effect of initial stripe frequency on the recapture stripe frequency. We assumed that the recapture stripe count for bush *i* was y_i ~

binomial(p_i, n_i), where p_i is the true stripe recapture rate for a given initial release stripe frequency. We then placed independent, uninformative beta priors on p_i for each treatment. MCMC (via *rjags*) was then used to draw samples from posterior distribution. Stripe frequencies clearly increased when stripe was initially rare (recapture frequency = 0.46, 95% CIs - 0.37-0.55; change in stripe frequency = 0.26, 95% CIs = 0.17-0.35; posterior probability that stripe increased > 0.99). In contrast, we found no clear, consistent pattern of change when stripe was initially common (change in stripe frequency = -0.006, 95% CIs = -0.093-0.063; posterior probability that stripe increased > 0.43). We inferred selection coefficients for each treatment (20% vs. 80% initial stripe frequency) based on the estimated posterior distribution for the true stripe recapture rate. We defined relative fitnesses for striped and green stick insects as $w_{stripe} = 1$ and $w_{green} = 1 - s$, respectively. Here s is the selection coefficient. We then estimated w_{green} based on the difference between release and recapture frequencies of the striped morph, such that $p_i = (p_0\ w_{stripe})/(p_0 * w_{stripe} + (1\text{-}p_0) * w_{green})$, which can be rearranged as $w_{green} = (p_0 * p_i - p_0)/((p_0\text{-}1) * p_i)$. Here p_0 is the stripe release frequency (0.2 or 0.8).

Estimation of differences between hosts

We fit a hierarchical Bayesian model to quantify the overall difference in stripe frequency between hosts across years. A key aspect of this model was that it allowed us to account for the heterogeneity in sampling, including the fact that a subset of sites was sampled each year. We used all samples from the main mountain, Highway 154. This included 21,067 data points (*T. cristinae* scored as striped versus unstriped, we excluded melanistic morphs) from 274 collections (site by year combinations; 29 sites with a mean of 9.4 visits per site) spanning 27 years (1990 to 2017).

We specified generalized linear models for the stripe frequency at each location (site) for each year (nearby or inter-digitated samples from different hosts were considered different sites). We included effects for site and year, and modeled each of these hierarchically by placing a normal prior on them with parameter values estimated from the data (except the means for the year effects, which were fixed at 0 to ensure the model parameters were identifiable). We placed uninformative priors on the site means, normal with mean 0 and precision 1e^{-6}, and on the precision parameters, gamma(0.01, 0.001). We used Markov chain Monte Carlo to generate samples from the posterior distribution and used these samples to compute several key derived parameters: the yearly mean stripe frequency for each host and the yearly mean difference in stripe frequency between hosts. Inferences were based on three MCMC chains, each with a 10,000 iteration burn-in, 20,000 sampling iterations and a thinning interval of 5 (MCMC analyses were conducted with *rjags*). Point estimates (posterior medians) for the difference between hosts (stripe frequency on *Adenostoma* minus *Ceanothus*) ranged from 0.30 to 0.64 (mean = 0.56), and for all but one year (2011) the 95% CIs for the difference in stripe frequency excluded 0 (i.e., they were significantly positive).

Estimating predictability in finches and moths

The data analyzed were obtained as follows. We obtained data on *Geospiza fortis* and *Geospiza scandens* body size and beak size from (*21*). The data are from Daphne Major from 1973 to 2012. Three measurements were included: principal components (PC) 1 body size, PC1 beak size and PC2 beak size. We obtained data on *Panaxia dominula medionigra* allele frequency from (*40*). We used the data from 1940-1978, as there were no gaps in sampling during this time interval. We obtained data on *Biston betularia* peppered moth morph frequency from (*41*). We used the data from Leeds, which was most complete, and restricted analysis to years 1967 to 1995 because there were several years after 1995 with very low sample sizes. ARMA Models were fit to the data as described for *T. cristinae* above.

We then asked whether and to what extent including rainfall data on Daphne Major (also from 1973 to 2012) improved the fit of the *Geospiza* time series data sets. We focused on rainfall as it is thought to be a strong determinant of seed size, which is a key source of selection on these finches (*1, 21*). We obtained the rainfall data from (*21*). We fit Bayesian ARMA models of order 0, 1, or 2 with respect to the AR and MA components (as described previously) that also included rainfall (MCMC details were identical to those described above). We placed an uninformative prior, Normal(mean = 0, precision = 1e-5), on the coefficient for rainfall. We then used the best ARMA model that included rainfall (based on DIC) for predictive cross-validation and forecasting as described above for the pure ARMA models (without rainfall). We then compared the predictive performance of the best ARMA models with and without rainfall.

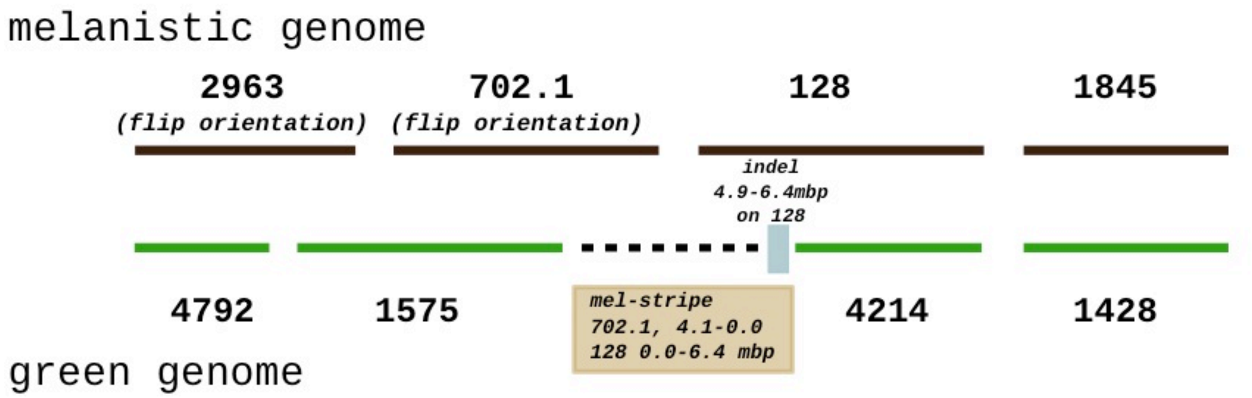


Fig. S1. Schematic illustrating the delimitation of the Mel-Stripe locus using two reference genomes. See text of supplementary materials for details.

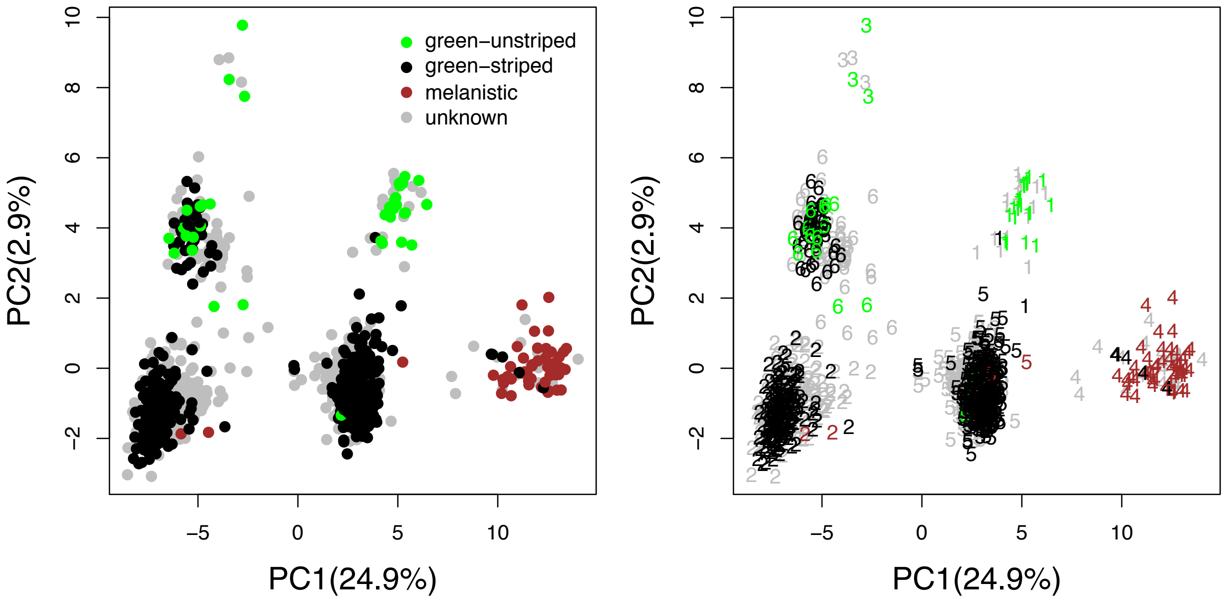


Fig. S2. Principal Components Analysis (PCA) ordination of 1102 *T. cristinae* from FHA based on genetic data from the Mel-Stripe locus. Points (left panel) and numbers (right panel) denote individuals, and are colored based on color and pattern phenotypes (we did not have phenotypic data for some individuals). In the right panel, numbers denote cluster/group assignments from k-means clustering with k=6. Cluster assignments were used to assign genotypes when estimating selection.

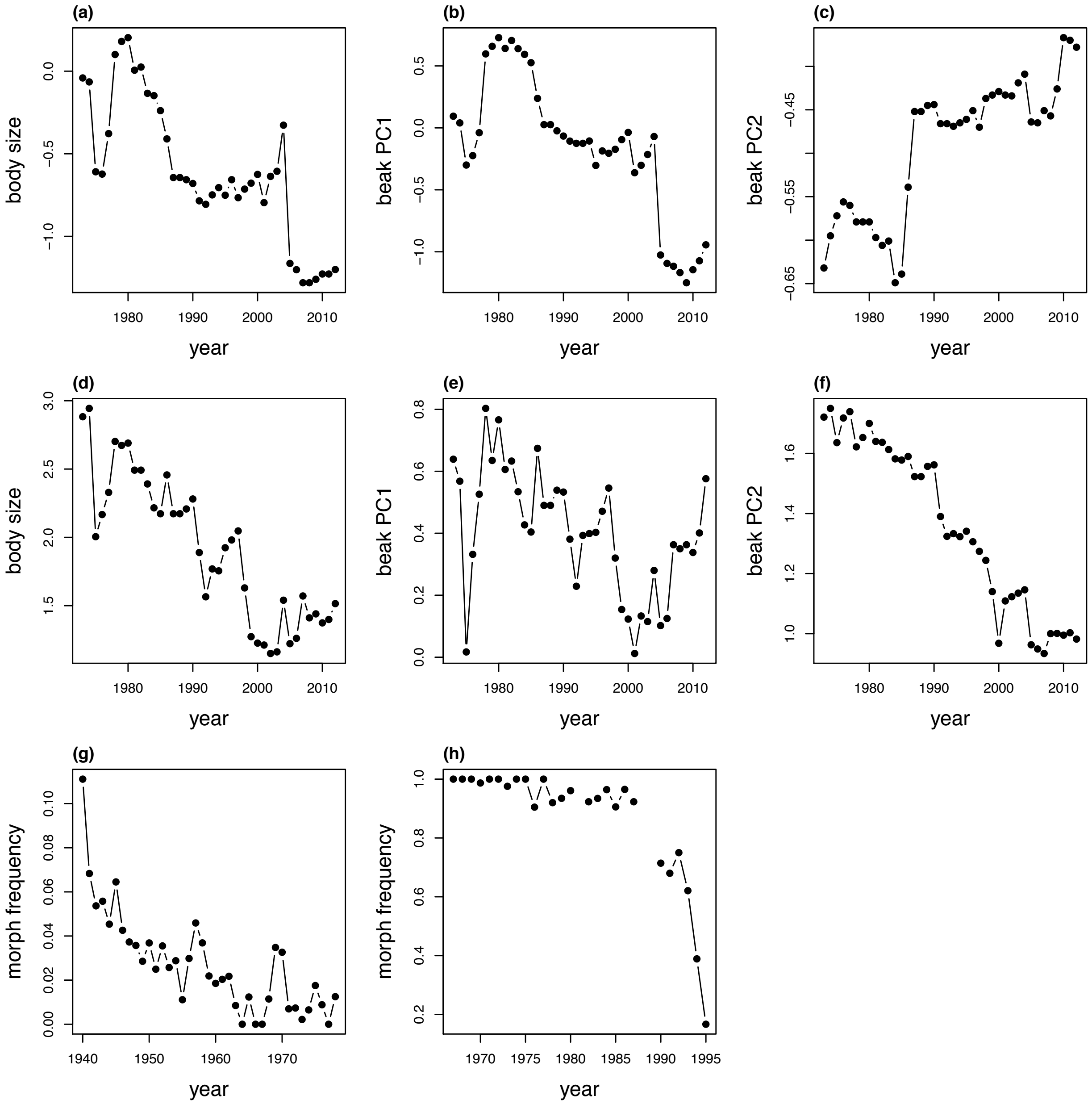


Fig. S3. Evolutionary time series for *Geospiza fortis* body size (a), beak PC1 (b), beak PC2 (c), *G. scandens* body size (d), beak PC1 (e), beak PC2 (f), *Panaxia dominula medionigra* frequency (g), and *Biston betularia* "peppered" frequency.

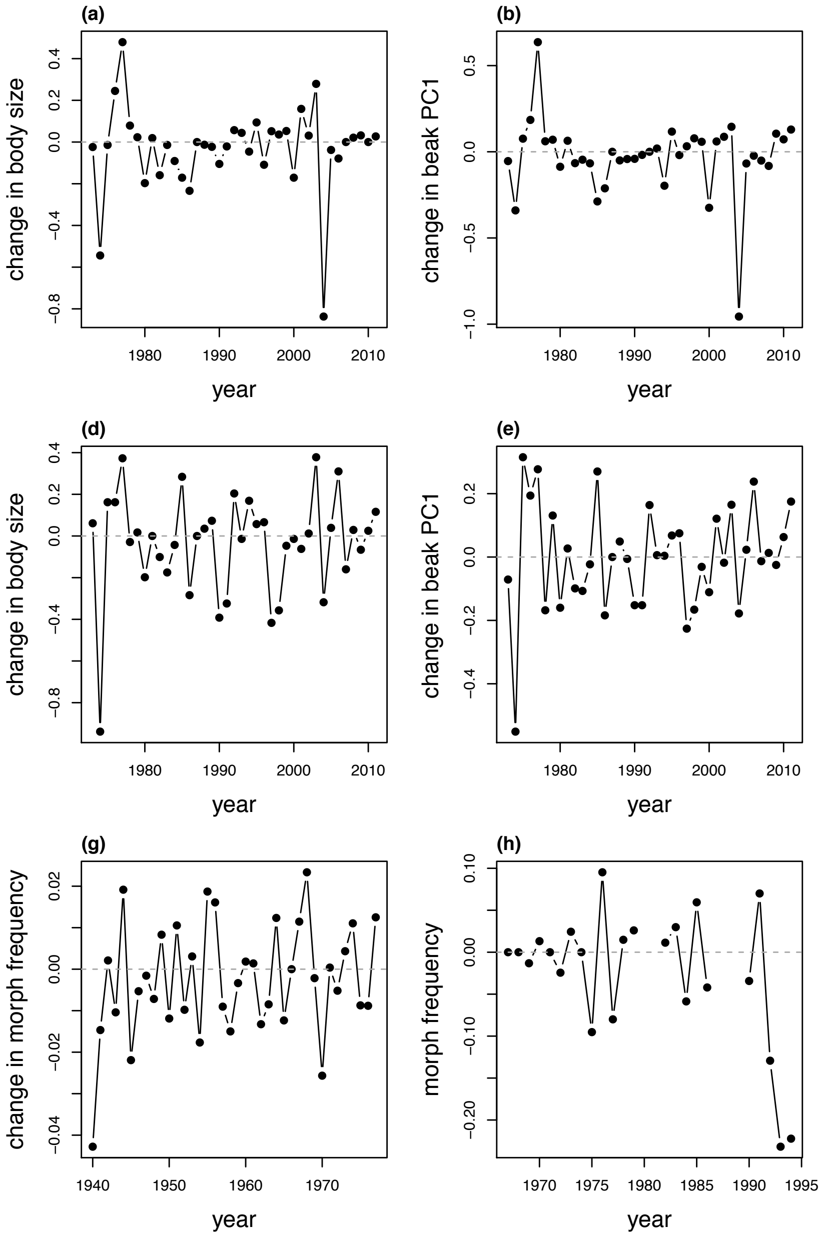


Fig. S4. Change in mean trait values or morph/allele frequency for *Geospiza fortis* body size (a), beak PC1 (b), beak PC2 (c), *G. scandens* body size (d), beak PC1 (e), beak PC2 (f), *Panaxia dominula medionigra* frequency (g), and *Biston betularia* "peppered" frequency. Data points for each year denote that change observed from that year to the next year.

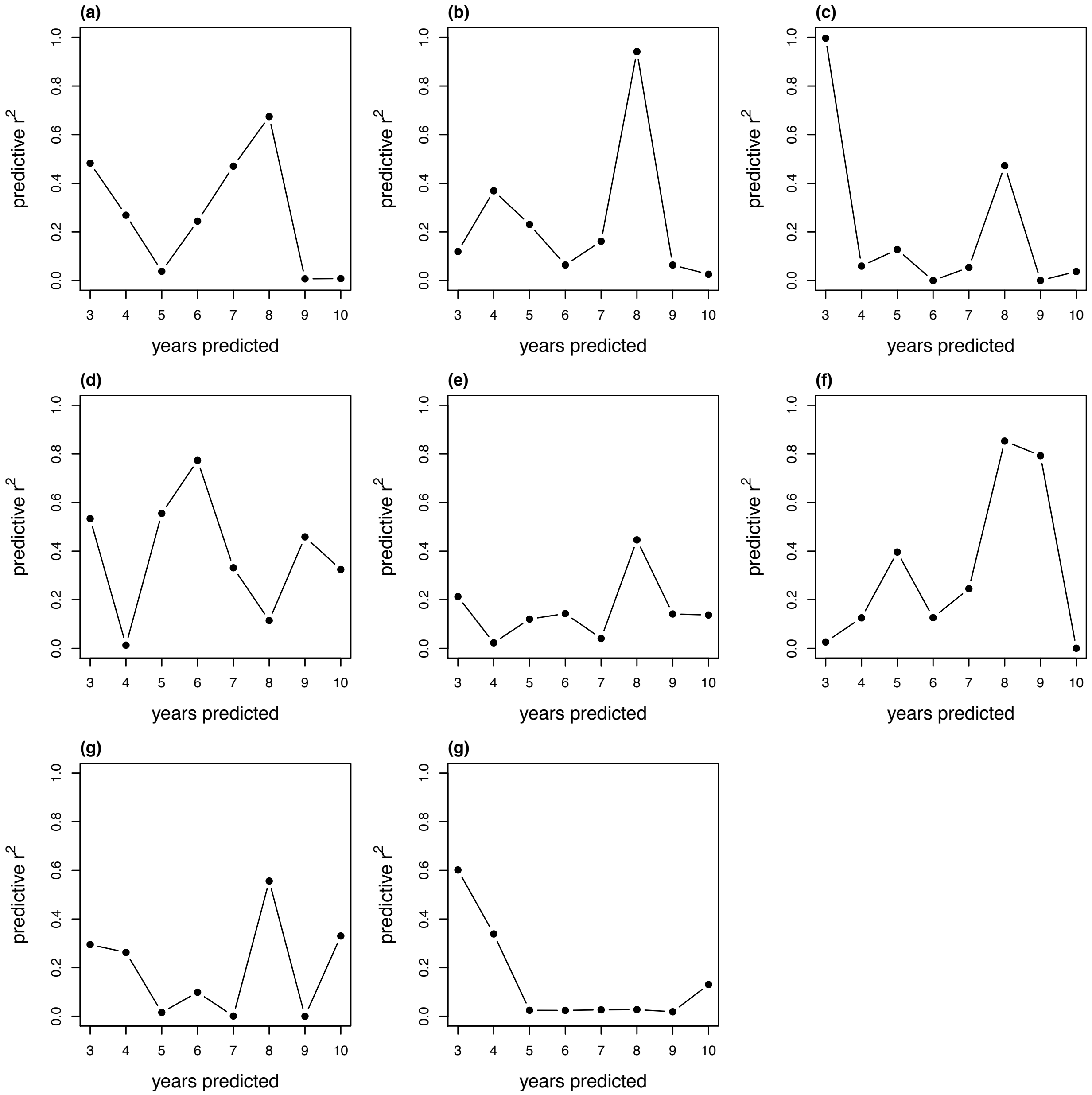


Fig. S5. Predictive r^2 from ARMA forecasting models for evolutionary time series in *Geospiza fortis* body size (a), beak PC1 (b), beak PC2 (c), *G. scandens* body size (d), beak PC1 (e), beak PC2 (f), *Panaxia dominula medionigra* frequency (g), and *Biston betularia* "peppered" frequency. r^2 between the observed and predicted values of change are shown from models dropping (and predicting) the last three to 10 years (r^2 was computed from a simple linear model).

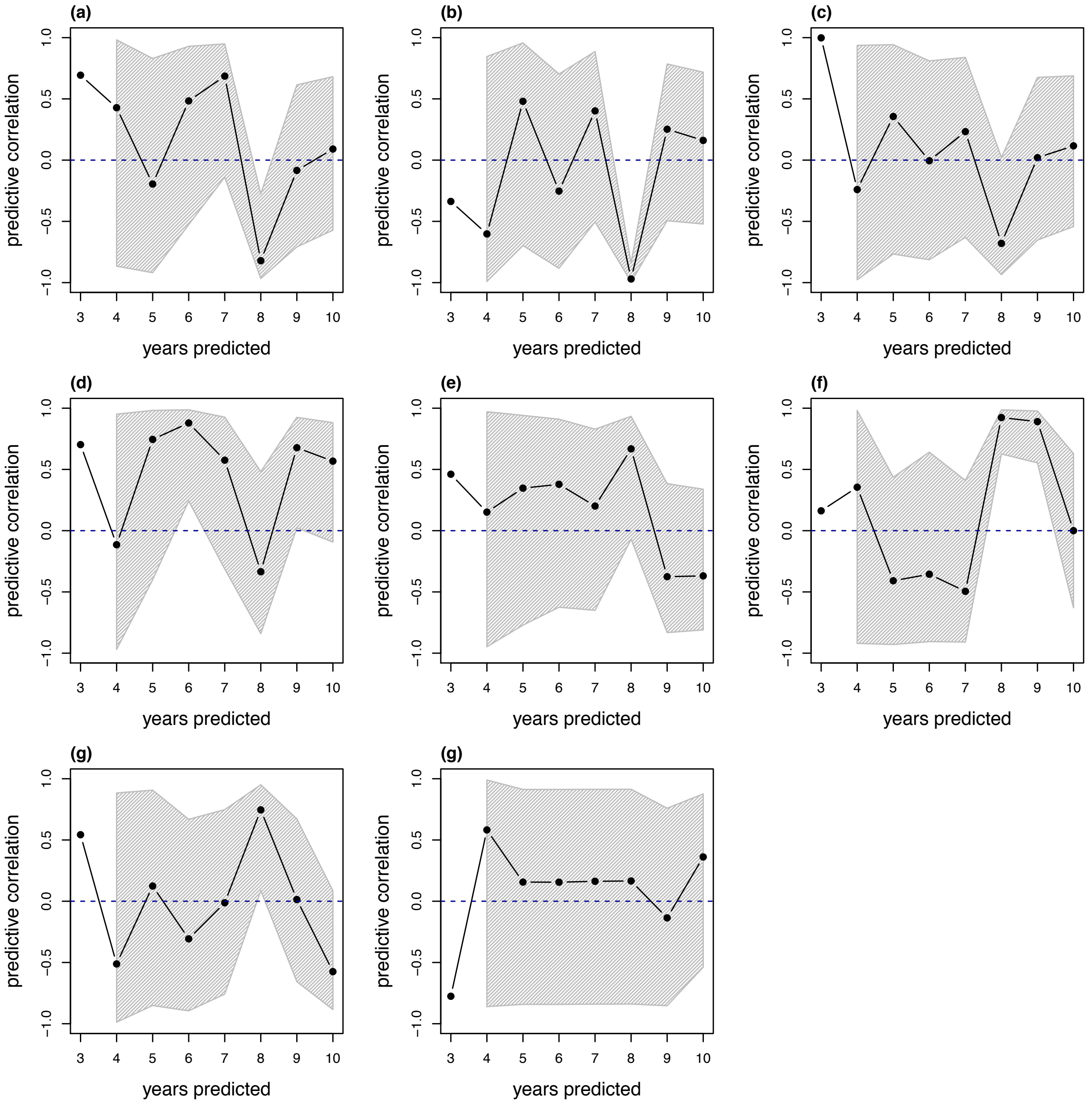


Fig. S6. Predictive correlations from ARMA forecasting models for evolutionary time series in *Geospiza fortis* body size (a), beak PC1 (b), beak PC2 (c), *G. scandens* body size (d), beak PC1 (e), beak PC2 (f), *Panaxia dominula medionigra* frequency (g), and *Biston betularia* "peppered" frequency. Pearson correlations (solid line and points) and 95% confidence intervals (shaded polygons) between the observed and predicted values of change are shown from models dropping (and predicting) the last three to 10 years.

Table S1. Summary of cross-validation and forecasting results (values for forecasting are medians from estimates based on 3 to 10 year forecasts). Bold font denotes cases where the ARMA model was preferred over a null model with a constant expectation.

Data set	Best model	cross-validation intercept	cross-validation slope	cross-validation r^2	forecasting r	forecasting r^2
Timema stripe	**ARMA(1,2)**	-0.005463	0.938310	0.6974	0.9282905	0.8618326
Timema color	**ARMA(1,2)**	0.03373	-1.24295	0.1019	-0.2959806	0.1388707
G. fortis body size	ARMA(2,2)	-0.07142	-1.89589	0.2581	0.2593157	0.2565920
G. fortis beak size (PC1)	ARMA(0,1)	-0.1872	-6.0475	0.2769	-0.0460291	0.1405218
G. fortis beak size (PC2)	**ARMA(0,1)**	0.002860	0.462132	0.03286	0.06829066	0.05675793
G. scandens body size	**ARMA(1,2)**	-0.005377	-0.159569	0.05488	0.6263869	0.3951535
G. scandens beak size (PC1)	**ARMA(1,2)**	0.02175	-0.22193	0.05206	0.2741161	0.1395220
G. scandens beak size (PC2)	**ARMA(1,2)**	-0.002938	0.546308	0.05978	0.08118622	0.18602100
P. dominula medionigra	**ARMA(1,1)**	-0.0005844	0.4419179	0.01698	0.000594893	0.180800958
B. betularia peppered	**ARMA(1,0)**	-0.05174	-1.80925	0.6584	0.15945025	0.02692756

Table S2. Posterior median and 95% credible intervals for key model parameters from the March, April, May melanistic morph model. All continuous covariates were standardized.

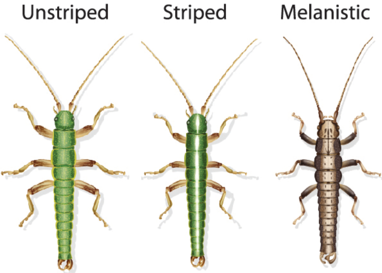
Parameter	Median	Lower bound 95% CI	Upper bound 95% CI
a_1	-2.31	-2.44	-2.20
a_2	0.187	0.063	0.309
b_1	-0.163	-0.260	-0.061
b_2	-0.0060	-0.0151	0.0274
c_1	0.164	-0.001	0.341
c_2	-0.197	-0.362	-0.249
d_1	-0.500	-0.749	-0.249
d_2	-0.050	-0.323	0.219

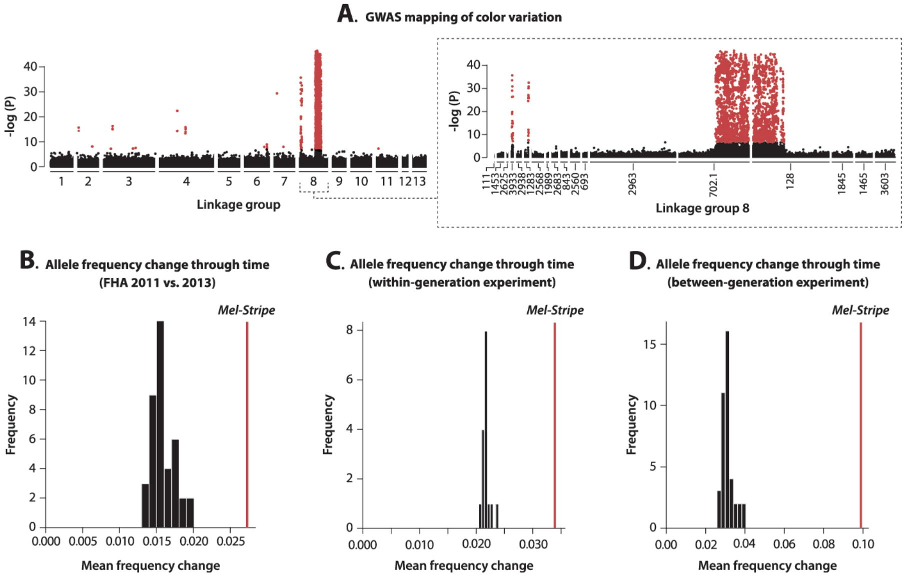
Table S3. Summary of model fit for the *Geospiza* data when rainfall is included in the model (based on rainfall and trait measurements from 1973-2012). We report the r^2 (mean across 3-10 years) for forecasting for the best ARMA model with rainfall, as well as the change in forecasting r^2, r (unsquared), and the lower and upper bounds on of the 95% confidence interval on r (lb and ub, respectively)(all of these values are averages across 3-10 year forecasts) obtained by including rainfall (positive values mean that rainfall improved the predictive forecast).

Data_set	Model	r^2	Change in r^2	Change in r	lb	ub
G. fortis						
body size	ARMA(2,2)	0.434	0.178	0.238	0.192	0.102
beak PC1	ARMA(0,1)	0.174	0.034	0.365	0.027	0.116
beak PC2	ARMA(0,1)	0.080	0.023	0.207	0.078	0.047
G. scandens						
body size	ARMA(2,1)	0.059	-0.337	-0.632	-0.421	-0.206
beak PC1	ARMA(1,2)	0.249	0.109	-0.476	0.014	-0.210
beak PC2	ARMA(1,2)	0.121	-0.065	-0.054	0.002	0.100

Database S1. Raw population data. See attached .csv sheet. Variable names are as follows: location = population/locality, year = year collected, latitude = latitude, longitude = longitude, elevation = elevation in meters, host = host plant collected on (A = *Adenostoma*, C = *Ceanothus*), melanistic = number of melanistic individuals collected, striped = number of striped individuals collected, unstriped = number of unstriped individuals collected, intermediate = number of intermediately striped individuals collected, total = total number of individuals collected, proportion_melanistic = proportion of the sample that was melanistic, proportion_striped_no_mel = proportion of the sample that was striped (excluding melanistics), mean_FebMarApr_temp = mean temperature in Fahrenheit for February, March, and April, mean_MarAprMay_temp = mean temperature in Fahrenheit for March, April, and May, mean_FebMarAprMay_temp = mean temperature in Fahrenheit for February, March, April, and May, refugio_yn = Mountain collected on (1 = Refugio, 0 = Highway 154).

The three morphs of *Timema cristinae*



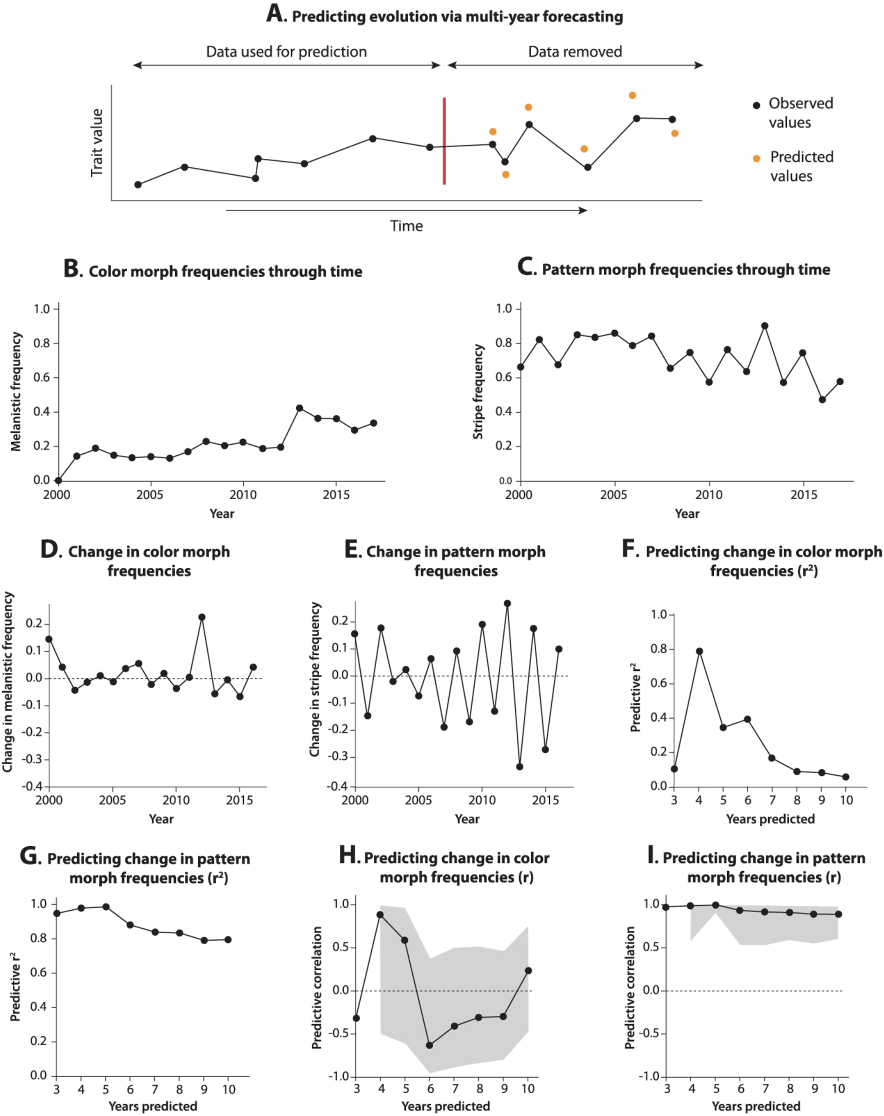


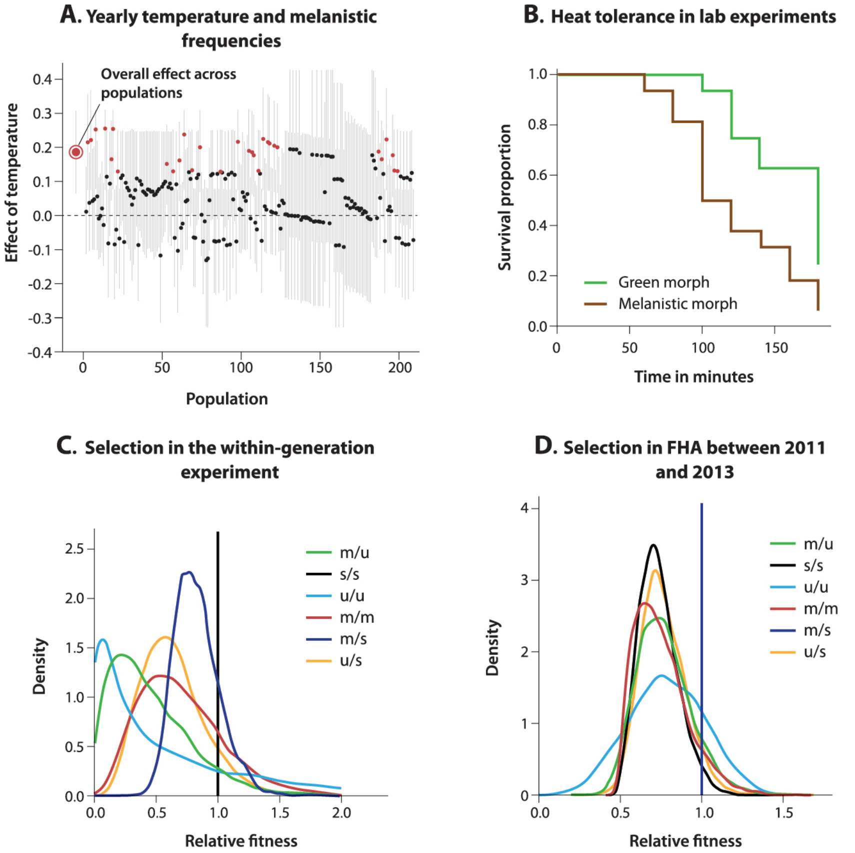
A. GWAS mapping of color variation

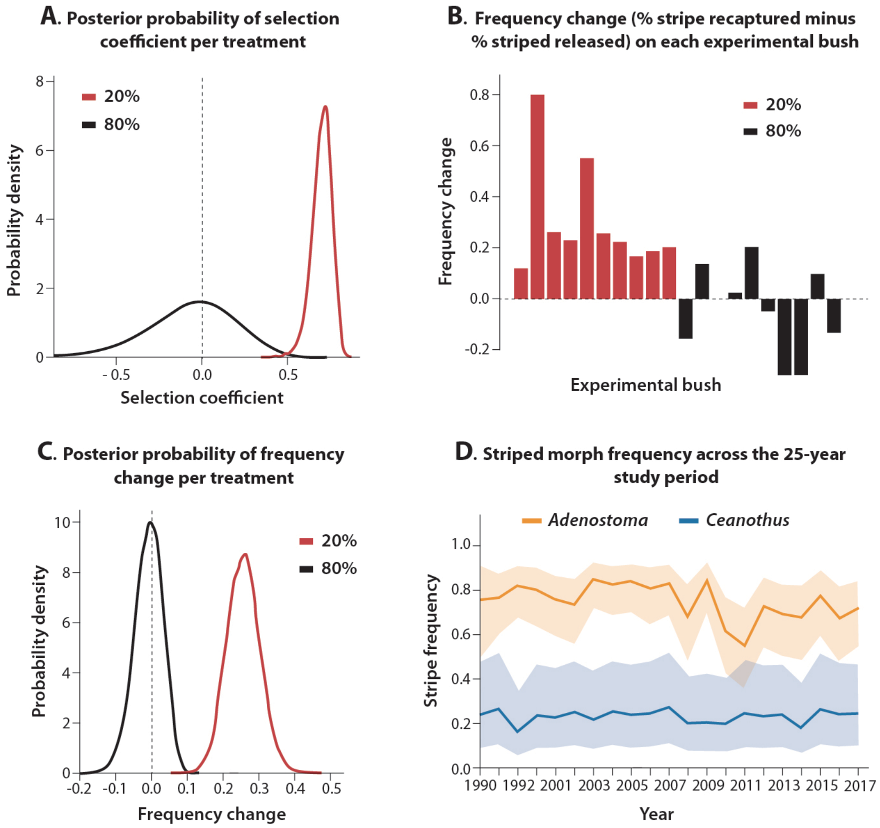
B. Allele frequency change through time (FHA 2011 vs. 2013)

C. Allele frequency change through time (within-generation experiment)

D. Allele frequency change through time (between-generation experiment)


A. Predicting evolution via multi-year forecasting
Data used for prediction
Data removed
Trait value
Time
Observed values
Predicted values
B. Color morph frequencies through time
Melanistic frequency
1.0
0.8
0.6
0.4
0.2
0.0
2000
2005
2010
2015
Year
C. Pattern morph frequencies through time
Stripe frequency
1.0
0.8
0.6
0.4
0.2
0.0
2000
2005
2010
2015
Year
D. Change in color morph frequencies
Change in melanistic frequency
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
2000
2005
2010
2015
Year
E. Change in pattern morph frequencies
Change in stripe frequency
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
2000
2005
2010
2015
Year
F. Predicting change in color morph frequencies (r²)
Predictive r²
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
10
Years predicted
G. Predicting change in pattern morph frequencies (r²)
Predictive r²
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
10
Years predicted
H. Predicting change in color morph frequencies (r)
Predictive correlation
1.0
0.5
0.0
-0.5
-1.0
3
4
5
6
7
8
9
10
Years predicted
I. Predicting change in pattern morph frequencies (r)
Predictive correlation
1.0
0.5
0.0
-0.5
-1.0
3
4
5
6
7
8
9
10
Years predicted


A. Yearly temperature and melanistic frequencies
Overall effect across populations
Effect of temperature
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
50
100
150
200
Population
B. Heat tolerance in lab experiments
Survival proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
Time in minutes
Green morph
Melanistic morph
C. Selection in the within-generation experiment
Density
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
Relative fitness
m/u
s/s
u/u
m/m
m/s
u/s
D. Selection in FHA between 2011 and 2013
Density
4
3
2
1
0
0.0
0.5
1.0
1.5
Relative fitness
m/u
s/s
u/u
m/m
m/s
u/s

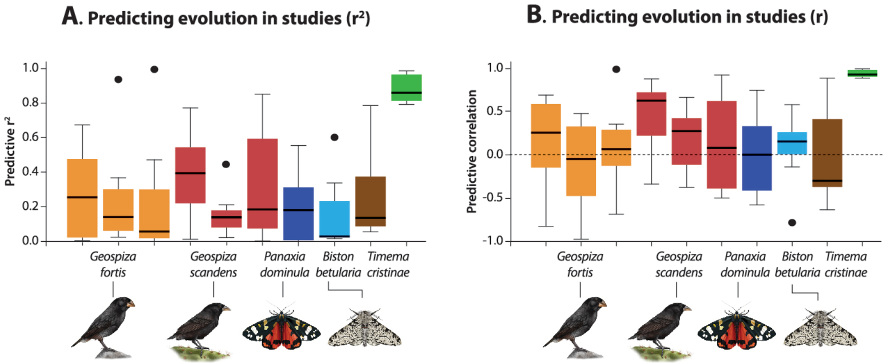


A. Posterior probability of selection coefficient per treatment

B. Frequency change (% stripe recaptured minus % striped released) on each experimental bush

C. Posterior probability of frequency change per treatment

D. Striped morph frequency across the 25-year study period



A. Predicting evolution in studies (r^2)

B. Predicting evolution in studies (r)