Ecology shapes epistasis in a genotype-phenotype-fitness map for stick insect color

Patrik Nosil[1,2], Romain Villoutreix[1], Clarissa F. de Carvalho[1], Jeffrey L. Feder[3], Thomas L. Parchman[4], and Zach Gompert[2]

[1]*Centre d'Ecologie Fonctionelle et Evolutive, Centre National de la Recherche Scientifique, Montpellier, 34293, France*
[2]*Department of Biology, Utah State University, Utah 84322, USA*
[3]*Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA*
[4]*Department of Biology, University of Nevada Reno, Nevada,* 89557, USA
Lead contact: patrik.nosil@cefe.cnrs.fr

Abstract

Genetic interactions such as epistasis are widespread in nature and can shape evolutionary dynamics. Epistasis occurs due to non-linearity in biological systems, which can arise via cellular processes that convert genotype to phenotype and via selective processes that connect phenotype to fitness. Few studies in nature have connected genotype to phenotype to fitness for multiple potentially interacting genetic variants. Thus, the causes of epistasis in the wild remain poorly understood. Here, we show that epistasis for fitness is an emergent and predictable property of non-linear selective processes. We do so by measuring the genetic basis of cryptic colouration and survival in a field experiment with stick insects. We find that colouration exhibits a largely additive genetic basis, but with some effects of epistasis that enhance differentiation between colour morphs. In terms of fitness, different combinations of loci affecting colouration confer high survival in one host-plant treatment. Specifically, non-linear correlational selection for specific combinations of colour traits in this treatment drives the emergence of pairwise and higher-order epistasis for fitness at loci underlying colour. In turn, this results in a rugged fitness landscape for genotypes. In contrast, fitness epistasis was dampened in another treatment, where selection was weaker. Patterns of epistasis that are shaped by ecologically based selection could be common, and central to understanding fitness landscapes, the dynamics of evolution, and potentially other complex systems.

Genes that control adaptive traits have now been identified in many organisms [1,2] and some pioneering work has even connected genotype to phenotype to fitness (i.e., a genotype-phenotype-fitness map) for individual genes [3–8]. However, adaptation may often involve multiple genes [9–11], with potential interactions

among them. Here, we focus on epistasis, defined as interactions between genes, where the effects of an allele at a locus depend on an allele (or alleles) at one or more other loci in the genome [12]. Such epistasis can make it difficult to predict evolution based on information from single genes alone [9,13,14], and has implications for adaptation [11,12,15] and speciation [16–18]. For example, epistasis can affect the evolution of complex traits, sex and recombination [19], parasite and antibiotic resistance [20,21], reproductive isolation [16,17], and missing heritability in human disease [9]. Epistasis is fundamental for understanding the structure of fitness landscapes [18,22,23], including their ruggedness and the number of adaptive peaks they contain, features that shape evolutionary dynamics.

Epistasis can arise at two fundamental levels of biological organisation (Figure 1)[10,24,25]. First, cellular and molecular processes can result in non-linearity in the conversion of genotypic to phenotypic variation [12,15,24], for example due to protein interactions and the complexity of metabolic and developmental networks [24] (a 'non-linear genotype-phenotype map' hypothesis, Fig. 1). Second, non-linear forms of phenotypic selection can cause epistasis for fitness at genes underlying trait variation [11,25,26] (a 'fitness epistasis' hypothesis). Critically, this can occur even if alleles contribute additively to trait variation, because when it comes to fitness *per se* the effect of an allele can still depend on the genetic background in which it occurs [25]. For example, under stabilizing or disruptive phenotypic selection (i.e., common forms of non-linear selection) [27], the fitness effect of a mutation that additively increases a trait value (e.g., body size) will depend on whether the mutation occurs in a genetic background where it moves the multi-locus genotype closer or further from the selective optimum. Likewise, correlational selection for combinations of trait values results in some underlying gene combinations having higher fitness than others, i.e., fitness epistasis rather than an additive relation between genotype and fitness [28]. Parsing these two main causes of epistasis is important because it determines whether interactions arise via inherent cellular features or through variation in ecological factors.

Empirically, epistasis is difficult to study due to the vastness of genotype space and the challenge of connecting genotype to phenotype to fitness for multiple genetic variants. For example, in terms of genotype space, with just five mutational steps separating two DNA sequences, there are 5! = 120 possible mutational paths between them [21]. Nonetheless, some progress has been made. A number of innovative combinatorial studies have engineered mutational steps that separate forms of single proteins and tested their fitness effects [29]. Some of these studies provide evidence for sign epistasis (i.e., alleles beneficial in one genetic background are deleterious in another) and rugged fitness landscapes [21,29], as do experimental evolution studies in microbes [13,30]. Most of this work has focused on simple pairwise genetic interactions such that higher-order interactions remain poorly understood, despite the potential importance of the dimen-

sionality of the fitness landscape for evolutionary dynamics [18,22]. In terms of genotype-phenotype-fitness maps, a general understanding of how genetic variation is converted into phenotypes is accumulating rapidly [1–9]. In contrast, much less is known about the genetic basis of organismal fitness in the wild [1,3–8], and very few studies have connected genotype to phenotype to fitness for combinations of loci (but see [11]). Until more such studies emerge, the prevalence, causes, and predictability of genetic interactions will remain unclear.

Connecting genotype to phenotype to fitness for multiple loci is not a trivial task. Moreover, in many systems where gene combinations presumably affect fitness, suppressed recombination (e.g., supergenes) prevents the existence of a range of genetic combinations whose fitness can be assayed [19,31]. Thus, a key to being able to test hypotheses concerning epistasis is the existence of recombination among adaptive genetic variants, and the ability to measure fitness of different gene combinations. Here, we use segregating genetic variation created by recombination to experimentally connect genotype to phenotype to fitness for loci controlling cryptic colouration in an insect. This enables us to elucidate the causes of epistasis and to infer features of the fitness landscape. In turn, our results inform why gene combinations are often packaged into distinct units of biodiversity via suppressed recombination.

We study wingless, herbivorous *Timema* stick insects, which rely on crypsis for protection against visual predators such as birds while resting on their host plants [32–36]. *Timema* body colouration has thus evolved to approximate the colours of the leaves and stems of their hosts (i.e., green versus brown / 'melanistic' morphs), and colour is a major axis of selection and adaptation in these insects [32,34,35] (Figures 2, 3). A recent study revealed that colour in *T. chumash* is controlled by a moderate number (~5-7) of linked but recombining genetic variants that reside in a ~1 mega-base region of linkage group eight (LG8 hereafter) [37]. Accordingly, although *T. chumash* exhibits statistically distinguishable green and melanistic morphs, it exhibits wide ranging colour variation overall, including individuals that are shades of yellow, pink, tan, beige, and blue [37]. Here, we leverage this segregating genetic variation for a range of colouration to test for selection on combinations of colouration loci (i.e., fitness epistasis). We integrate our findings with the fact that other *Timema* species exhibit more distinct colour morphs due to suppressed recombination among colour loci [37]. For example, morphs in *T. cristinae* are distinguished by a large (~10 mega-base) region of suppressed recombination on LG8 named the '*Mel-Stripe'* locus [35,37,38]. We thus focus here exclusively on the *Mel-Stripe* region (note that the majority of colour loci in *T. chumash* map to a ~1 mega-base subset of the *Mel-Stripe* region) [37].

Results and Discussion

Genotype-phenotype map for cryptic colouration. We first tested the hypothesis of a non-linear genotype-phenotype map for colouration in *T. chumash* (see Methods for details). We did so using standardized photos of 437 *T. chumash* for which we quantified body colour using red, green, and blue (R, G, B, respectively) pixel values. Following the approach of Endler [39], we calculated chromatic contrasts as the relative difference between: (1) red and green channels (RG = (R – G)/(R +G), a trait referred to as 'RG' hereafter) and (2) between green and blue channels (GB = (G – B)/(G + B), a trait referred to as 'GB' hereafter; Figure 2). This approach is based on the two most common differences between photoreceptor signals, resulting from the wiring of visual systems [40,41]. Such contrasts thus yield more biologically meaningful results for comparing colour patterns than do raw RGB values [39]. In addition, our methods capture the major axes of variation in *T. chumash* colour space, given this species does not reflect ultraviolet spectra (see Supplementary Information and Supplementary Figure 1 for results using reflectance data from *T. chumash*, modelled to avian photoreceptor sensitivities). These same photographed individuals were subsequently used in the mark-recapture experiment described below to estimate phenotypic selection and the genetic basis of survival. Thus, from each individual we took a tissue sample before release into the field that allowed us to collect genotyping-by-sequencing (GBS) data for all individuals (both those recaptured and not). Past work comparing individuals from whom a tissue sample was taken to unperturbed individuals has shown that tissue sampling does not affect survival in the laboratory or field [42].

The GBS data were first used for genetic mapping of colour (Figure 3). To this effect, we employed a purely additive Bayesian multi-locus genome-wide association (GWA) mapping approach in GEMMA that accounts for linkage disequilibrium (LD) among SNPs [43]. This revealed that ~80% of the variation in colour was explained by additive genetic variation, with narrow credible intervals on these estimates (Figure 3, Supplementary Figure 2). Consistent with past work, the SNPs associated with colour were concentrated in the *Mel-Stripe* locus region of the *T. cristinae* genome [37]. We return to these individual SNPs in more detail below when connecting genotype to fitness.

Given that most (i.e., ~80%) of the variation in colour was accounted for using an additive genetic model, epistasis could at most explain 20% of colour variation. Past work on colouration in *T. chumash* using a population from a different geographic site failed to detect evidence for epistasis [37]. However, explicit tests for marginal epistasis in the current data set using MAPIT [44] revealed statistical evidence of epistasis for five SNPs in the *Mel-Stripe* region (Figure 3).

Consequently, we re-ran GEMMA analyses, this time including in the model the 10 possible pairwise in-

teractions between these five SNPs (i.e., including epistatic interactions). This revealed a noticeable increase (~10%) in the percent variance explained (PVE) for RG, but not for GB where PVE was comparable to that in the additive model (Figure 3). Moreover, the interaction between two pairs of SNPs was consistently retained as colour-associated across Markov chain Monte Carlo (MCMC) model steps (posterior inclusion probability ~1.0 for one SNP pair affecting RG and ~0.6 for one SNP pair affecting GB). We also quantified genomic estimated breeding values (GEBVs; these quantify the total effect of genetic markers on phenotype) for GEMMA models with and without interactions. The GEBVs from the two analyses were highly correlated such that similar estimates emerge whether epistasis was included or not (RG, r = 0.95, $P < 0.001$ GB, r = 0.98, $P < 0.001$; both tests two-tailed). However, visual inspection of the results revealed that allowing for epistasis does slightly alter GEBVs in a manner that enhances differentiation between green versus darkly coloured individuals (Figure 3). Finally, genomic prediction of colour based only on genotype revealed modestly increased predictive power in models that included epistatic interactions. In sum, colour exhibits a largely additive basis, but with some moderate additional effects of epistasis.

Phenotype-fitness map for survival. Having established that colour has a largely additive basis, we designed a mark-transplant-and-sequence experiment to estimate phenotypic selection on colour and to map the genetic basis of survival, a core component of fitness (Figure 4). Our design reflects patterns of *Timema* evolution and host-plant use in Southern California, where *T. chumash* occurs. These considerations focus on two closely related species, *T. chumash* and *T. podura*, which are broadly and locally co-occurring (i.e., sympatric). *T. podura* exhibits highly distinct green and melanistic morphs and its core hosts, *Ceanothus* (C) and *Adenostoma* (A), select for green versus melanistic colouration, respectively [34,45]. These hosts are also used, albeit more secondarily, by *T. chumash.* The combination of these hosts (AC hereafter) is predicted to generate strong correlational selection for green versus melanistic colouration (i.e., high GB and low RG values, or low GB and high RG values, respectively), and against other colours. In contrast, a core host of *T. chumash* is *Cercocarpus*, i.e., mountain mahogany (MM), which exhibits wide colour variation, as do *T. chumash* individuals found on it [37]. We thus collected *T. chumash* from MM and transplanted them within the same area to adjacent, touching individuals of A and C (AC treatment) and to MM (control). Notably, such a scenario is biologically realistic given that these three hosts co-occur at small spatial scales throughout Southern California.

We applied individual and block-specific pen-marks on the abdomen of each *T. chumash*. On June 18th, these individuals were released onto bushes representing the two treatments, in three paired blocks. On June 21st, the survivors of the experiment were recaptured. Several past experiments have revealed that

this procedure results in minimal dispersal, with individuals that are not recaptured suffering mortality and those recaptured representing survivors [36,42,46]. We estimated selection coefficients and standardized selection gradients on colour by comparing survivors to non-survivors (Supplemental Table 1 for details including sample sizes for numbers recaptured).

We did not detect strong evidence for selection on MM (Figure 4, Supplementary Figure 3). In contrast, our prediction of correlational selection on AC was supported. Specifically, the combination of A and C generated correlational selection for either: (1) high GB and low RG values (i.e., green morphs), or (2) low GB and high RG values (i.e., melanistic morphs). Consequently, there was strong selection against individuals with colouration intermediate or otherwise mismatched from that mentioned above. This was the case for the experiment overall, and within two of three blocks individually. Standardized linear, quadratic, and correlational selection gradients on AC were generally in the range of 0.05 – 0.15 (Supplemental Table 2). Thus, selection was moderately strong, but within the range documented for other systems and traits [27]. Correlational selection such as documented here can result in some gene combinations having higher fitness than others (i.e., fitness epistasis). We thus conducted further analyses that integrate the results from genotype, phenotype, and fitness to test for the effect of interactions between SNPs on survival probabilities.

Integrating components of the phenotype-genotype-fitness map. Our next goal was to connect genotype to phenotype to fitness for the specific genetic regions (i.e., SNPs) associated with colour. The first step in doing so was to return to the results from the analyses reported above for mapping colour, this time focusing on the individual SNPs most strongly associated with colour. Specifically, we used the aforementioned results from the additive GEMMA model to: (1) quantify the weight of evidence that each individual SNP was associated with colour, and (2) estimate the number of genetic variants (i.e., quantitative trait nucleotides, QTN) controlling each colour trait [43]. This was done by considering how often SNPs were retained as trait-associated across different MCMC steps in the GWA. The proportion of such steps is termed the posterior inclusion probability, PIP hereafter, and reflects the weight of evidence that a SNP is associated with colouration. In the case of multi-genic control with recombination among loci, the one or few SNPs that best tag each causal variant are expected to consistently be trait-associated across MCMC steps (i.e., exhibit high PIP values). In turn, PIP values across SNPs sum to the number of total causal variants (i.e., even if the causal variants are not unambiguously identified, the number of such variants can be estimated) [43].

These analyses revealed that ~5 genetic variants control RG (posterior mean and s.d. = 4.63 +- 0.81), and

~4 control GB (posterior mean and s.d. = 3.94 +- 1.03). Overall, we estimated that ~6-7 genetic variants control colouration (posterior mean and s.d. = 6.51 +- 1.15), because some, but not all, colour-associated SNPs affected both traits. Notably, LD among the top colour-associated SNPs was generally modest, indicative of recombination between them (Supplementary Figure 4). Given these results, we focused our fitness analyses on the five most strongly colour-associated SNPs, with PIPs > 0.70 and minor allele frequencies (MAF) greater than 0.05. Notably, these five SNPs were 478-672 times more likely to be colour-associated than to have no effect on colour (Supplemental Table 3 for statistics). Finally, generation of a new *de novo* chromosome-level genome assembly for *T. chumash* confirmed that these SNPs are in synteny between *T. cristinae* and *T. chumash,* and exist as a single copy in the *T. chumash* genome (Supplementary Figure 10).

We used Bayesian multiple regression with variable selection and model averaging to connect genotype to phenotype to fitness. Our dependent variable was survival probability (i.e., expected fitness) inferred from the analyses of phenotypic selection (i.e., each individual was assigned a survival probability based on its colour score and the selection analyses). Our independent variables were the main (i.e., additive) effects of the five SNPs and their possible interactions (i.e., epistasis). In this context, two-way interactions represent pairwise epistasis and other interactions represent higher-order epistasis.

The full results are depicted in Figure 5. In the AC treatment, we found additive and epistatic effects on fitness, with the latter involving marked two- and three-way epistatic interactions. Thus, we estimated that the number of effects of epistasis on survival on AC was ~6, with ~2-3 stemming from pairwise interactions and ~2-3 stemming from three-way interactions. Notably, predictive power in the AC treatment from leave one-out cross-validation was ~64% better for a model with both additive and epistatic effects than for a model with only additive effects (both effects, predictive r = 0.60, 95% CI = 0.50-0.68, r^2 = 0.36, $P < 0.001$, additive effects only, predictive r = 0.47, 95% CI = 0.35-0.57, $r^2 = 0.22$, $P < 0.001$; both tests two-sided; Table 1).

To test if these results could arise from chance (i.e., from no true association between genotype and fitness), we ran null simulations that repeated the analyses 100 times using five randomly drawn SNPs that were matched for MAF with the colour-associated SNPs. The null model simulations revealed that our results from AC were unlikely to arise by chance ($P < 0.01$ for main effects and two- and three-way interactions; one-sided test). In contrast, additive and epistatic effects estimated in the MM treatment could be explained by chance ($P > 0.05$ for all model terms; one-sided test), consistent with the weaker phenotypic selection on MM than AC. Finally, these results were robust to other methods of analysis such as

Bayesian ridge and lasso regression, and could not be explained by dominance within SNPs (Supplementary Figure 5-8). We note that the genetic architecture of colour itself does not differ between treatments. Consequently, our results are consistent with non-linear selection (specifically correlational selection) driving the emergence of fitness epistasis on AC, and an absence of appreciable selection and fitness epistasis on MM. Thus, fitness epistasis was predictable based on patterns of ecologically based selection.

Moreover, these effects of epistasis can be understood at the level of underlying pairs and triplets of SNPs. One example in the AC treatment is shown in Figure 5, where alleles at two interacting SNPs have been coded by whether they cause colouration to become green (G) or more melanistic (M). It can be seen that high fitness is associated with having multiple copies of either G or M, with low fitness of genotypes that combine these alleles (e.g., to result in intermediate colouration). Such effects can be extended beyond pairs of SNPs to higher-order interactions. For example, Supplementary Figure 9 uses a three-way fitness interaction to illustrate how the survival effects of two SNPs that affect only GB depend on a third SNP that affects both GB and RG. In this case, the effects of the first two SNPs on lowering GB scores only improve survival if they are found in a genetic background at the third SNP that increases the RG score to result in more melanistic colouration. Thus, epistasis for fitness can be understood predictably via observed patterns of selection on colour in the transplant experiment.

Inference of fitness landscapes and their ruggedness. We next connected the detected effects of epistasis to the structure of the adaptive landscape. Specifically, we inferred the ruggedness of the fitness landscape (where peaks denote high fitness and valleys represent regions of low fitness) in our transplant experiment using the unique fitness expectation for each genotype provided by our Bayesian regression model. The multi-dimensional fitness landscape in each experimental treatment is depicted in Figure 5, where nodes denote genotypes, node size represents sample size (where the smallest nodes represent genotypes that were not observed in our sample), edges/lines connect genotypes that differ by one substitution, and colours denote relative fitness. Visual inspection of the landscape suggests greater ruggedness on AC than MM, consistent with the observed greater fitness epistasis on AC. Analyses executing random walks on the landscape from different starting points confirm that this visual intuition is correct, with greater ruggedness metrics on AC (Figure 5).

Supporting analyses and pleiotropic effects of colour-associated loci. Thus far, our fitness analyses have focused on colour-associated SNPs, and we have implicitly assumed that the main phenotypic effect of these SNPs is on colour, not other traits. In the supplementary materials and Table 1, we report additional analyses that explore and relax these assumptions, including a test for pleiotropic effects of colour-

associated SNPs on other (unmeasured) traits affecting survival (analogous to that introduced by Rennison and colleagues and applied to stickleback fish) [47]. These results show that our finding of fitness epistasis appears robust to different analytical approaches, but stronger results are obtained when considering selection acting through colour phenotype than when considering genotype alone. They also maintain clear evidence for selection on colour phenotype, but also suggest some pleiotropic effects of colour loci on other (unmeasured) traits influencing fitness.

Conclusions. Epistasis can arise via inherent cellular features or through variation in ecological factors. In our experiment, epistasis for fitness arose predictably as an emergent property of ecological variation in natural selection. This result informs three core issues in biology: (1) the predictability and repeatability of evolution, (2) exploration on adaptive landscapes, and (3) the packaging of multi-locus genetic variation into distinct units of biological diversity.

First, debate exists about the role of epistasis in the predictability and repeatability of evolution. Studies of proteins suggest that if evolution were repeated from the same starting point (i.e., genetic background), epistasis might increase the predictability of the mutational path taken to a given endpoint [21,29]. Specifically, deleterious gene combinations constrain the number of paths that are accessible to selection, causing predictable evolution. However, evolution will often proceed from different starting points, for example in variable genetic backgrounds of different populations and species. In such cases, genes with strong epistatic effects may only function well in a narrow range of genetic backgrounds, reducing their repeated use [12]. Our results add an important component to this debate. A major cause of epistasis in our study was selection itself. Thus, a key determinant of our ability to predict evolution might be our understanding of selection and its ecological causes, rather than only the cellular features that create epistasis in the genotype to phenotype map.

Second, our results shed light on the exploration of fitness landscapes. Specifically, how do populations traverse fitness valley to find global fitness peaks and avoid getting stuck on local optima [22,25]? A famous solution offered by Wright's 'Shifting Balance Theory' [23,48] invokes a delicate balance of migration, drift, and inter-demic selection, which may be difficult to achieve [49] (but see [50]). A potentially more general solution involves the stability of the landscape. If the environment is not static but rather fluctuates, then a peak at one point in time can become a valley at another, and vice-versa. Thus, valleys are temporary and crossable at certain points in time, i.e., the landscape is more a shifting 'seascape' [26]. The weaker selection on MM in our experiment could maintain standing genetic variation and help bridge peaks offered by AC, facilitating exploration of the fitness landscape. Other forms of fluctuating selection, such as negative fre-

quency-dependent selection, could further enable the exploration process [26], and indeed such selection has been documented in *Timema* [36]. Thus, the ecologically mediated epistasis documented here may enable the exploration of fitness landscapes in *Timema*. Given the ecological complexity of nature in time and space [14,51], similar processes likely apply to other organisms.

Third, our results increase understanding of the processes that package genetic variation into distinct units of biodiversity, such as morph or species. Divergence into such units is facilitated by reduced recombination [19,38] and by reproductive incompatibility [16,52]. Epistasis may play a role in both [16,17]. The alternative gene combinations favoured here could generate a seed of linkage disequilibrium that promotes the evolution of further reduced recombination between colour genes, and enhances the efficacy of selection for such reduced recombination. Indeed, this may have occurred in *T. cristinae*, a relative of *T. chumash* that feeds primarily on the hosts AC and exhibits reduced recombination between colour genes [36,38]. Moreover, our genome comparison here revealed that a chromosomal inversion on LG8 distinguishes these two species (Supplementary Figure 10). Finally, we note that specific combinations of traits and genes were selected against. Thus, natural selection itself may create ecological incompatibilities between forms that are analogous to the classical Dobzhansky-Muller genetic incompatibilities that cause hybrid dysfunction [17]. Although our focus was on morphs, similar processes could apply to ecotypes or species, as demonstrated under semi-natural conditions in stickleback fish [11]. Whether distinct units of diversity form could be mediated by the temporal and spatial scales at which selection (and thus fitness epistasis) fluctuates; fluctuations promote peak shifts but counteract consistent pressure for divergence.

Although experimental studies in nature akin to ours are few, genetic interactions have been studied extensively in several other contexts. For example, epistasis has been reported in experimental evolution studies in the lab [30,53], work on protein evolution [12,21], and in genome-wide scans of population genetic patterns [54,55]. Ideally, future work would combine these approaches to parse the causes of epistasis across different contexts, thus testing the generality of the patterns reported here. Nonetheless, given the prevalence of non-linear selection in nature [27], we expect fitness epistasis to be common. Thus, a collective body of emerging evidence suggests that genetic interactions may be central to understanding biological diversification, rather than being only complex second-order effects [56]. Even more generally, interacting components and non-linear functions are aspects of many biological, chemical, physical, and social systems [57,58]. Thus, evolutionary principles learnt from the study of genetic interactions may aid understanding of other complex systems.

ACKNOWLEDGEMENTS

We thank T. Reimchen, M. Joron, Luis-Miguel Chevin, and D. Ayala for discussion and comments on previous versions of the manuscript, M. Muschick for help with photography and reflectance measurements, and T. Oakley for lab space. 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 and ShARC HPC clusters, from the Corporate Information and Computing Services at the University of Sheffield. The work was funded by a grant from the European Research Council (EE-Dynamics 770826, https://erc.europa.eu/) and a grant from the National Science Foundation of the USA (NSF DEB 1638768). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions. All authors (PN, RV, CFC, JLF, TLP, and ZG) conceived the project. RV, CFC, TLP, and PN collected data. ZG led data analysis, aided by all authors. All authors (PN, RV, CFC, JLF, TLP, and ZG) contributed to writing.

Data Availability Statement. DNA sequences have been deposited in the NCBI SRA (BioProject PRJNA656892). Other data, including colour measurements and results from the experiment,have been archived in Dryad Digital Repository (doi:10.5061/dryad.2z34tmpjr). Correspondence and requests for materials should be sent to patrik.nosil@cefe.cnrs.fr.

Code Availability Statement. Core scripts have been archived in Dryad Digital Repository (doi:10.5061/dryad.2z34tmpjr).

Competing Interests Statement. The authors declare no competing interests.

Methods.

Insect sampling. The experimental *T. chumash* were collected from *Cercocarpus* sp. (Mountain Mahogany, MM) in the vicinity of the locality Horse Flats 5 (HF5, N 34 15.584, W 118 6.254). Over 700 individuals were collected between June 11th and June 13th, 2019. These were kept alive in plastic containers and moved to laboratory space where 465 healthy adults were chosen for use in the transplant experiment, including photography for phenotyping and molecular genotyping.

Photography and marking. All individuals from the transplant experiment were photographed with a digital Nikon 5600 camera equipped with a macro lens (Nikon AF-S VR Micro-Nikon 105mm f/2.8G IF-ED) and two external flashes (Yongnuo YN560-II speedlights). The images were taken with the camera set on manual, an aperture of f/14, a shutter speed of 1/200 s, a sensitivity of 100 ISO, and flashes adjusted to 1/4+0.5 env power in S1 mode in an output angle corresponding to 24-mm focal length on full frame (~84° diagonal). These settings gave the closest output possible from the picture obtained with the Canon EOS 70D body and 100mm f/2.8L Macro IS USM lens. To avoid shadows and reduce external luminosity interference, LumiQuest SoftBox LTp softboxes were attached to the flashes. In addition to the *Timema* specimens, the pictures included a ruler and a standard colour chip (Colorgauge Micro, Image Science Associates LLC, Williamson, NY, USA). Each specimen was photographed at least twice in different perpendicular positions to capture the body colour without gleam or shade. Once photographed, the individuals and labels were put back in 8-ounce deli cups (one insect and its associated label per cup), until random assignment to an experimental treatment and block. This process took a total of three days.

For tissue sampling, we took a portion of a leg for every individual included in the transplant experiment. Past work comparing individuals from whom a tissue sample was taken to unperturbed individuals has shown that this tissue sampling does not affect survival in the laboratory or the field [42]. In doing so, we cut the middle right leg (when facing the dorsal surface of the specimen) using nail scissors. Scissors and tweezers were sterilized before each leg cutting, wiping them with paper towel soaked in pure alcohol followed by heating the blades or tips with the flame of a lighter. The leg was placed in a two millilitre (mL) round bottom eppendorf tube filled with 1mL pure ethanol. We added the insect's identification label to the tube, and stored the tube for molecular work. Once we took the leg of an individual, we directly marked it. Marking consisted of four colour dots applied on the ventral side of the insect using a fine tip Sharpie pen. We used three different colours (red, black, and blue) to create unique colour sequences associated to a given individual. Once marked, we placed the individuals in an experimental unit box. We had a total of six unit boxes (one per treatment per block, given two treatments and three blocks per treatment). This process took us a total of two days.

Phenotyping. All photographs were corrected for white balance, adjusting the temperature and tint based on the values obtained from the neutral grey colour in a standard colour chip (target #10 in Colorgauge Micro, Image Science Associates LLC, Williamson, NY, USA), using ADOBE PHOTOSHOP LIGHTROOM 5.7 software (Adobe Systems Software Ireland Ltd). Due to the light standardization when the photographs were taken, the white balance did not change significantly between the pictures, where the temperature was around 6150, the tint -4, and a mean RGB balance of 95%. The images were then exported as high quality TIFF. The RGB colour channels (red, green and blue) were measured and processed following [39] to obtain the RG and GB estimates.

Sequencing, alignment, and variant calling. We generated reduced representation sequencing data for the 465 *T. chumash* using our standard genotyping-by-sequencing approach [60], which has been used in many past *Timema* studies [35,36,38,45]. We extracted genomic DNA for each individual from three to five legs using the DNeasy Blood and Tissue Kit (Qiagen). We then generated barcoded single-end DNA libraries for each individual following standard restriction-site digest protocols. These individual libraries were then distributed into pools (containing sets of different individuals). These pools were size selected for fragments of size 300-500 base pairs (including adaptors) and sequenced (one pool per lane) on an Illumina HiSeq2000 platform at the University of Texas Genome Sequencing and Analysis Facility (Austin, TX).

As in previous studies [35,36,38,45], we demultiplexed the data using custom Perl scripts that identify and remove the in-line barcodes, including those that were 1 bp away due to synthesizing or sequencing errors, and remove the subsequent six base pairs of the EcoRI cut site and the adapters at the 3' end when present. Sequences lacking barcodes (including PhiX sequences), or those shorter than 16 bp after parsing, were discarded. We then aligned the DNA sequences to the *T. cristinae* reference genome 1.3c2 (NCBI WGS PGFK01000000) with the *bwa mem* algorithm (version 0.7.17-r1188). For the alignments, we set the minimum seed length to 15, looked for internal seeds inside seeds longer than 19 bp, and only output alignments with mapping qualities of 30 or more. We then used *samtools* (version 1.5) to compress, sort and index the alignments.

We used the classic variant caller (-c) in *bcftools* for variant calling. This was done using the *mpileup* command in *samtools* (version 1.5) and the *call* command in *bcftools* (version 1.6). We used the recommended mapping quality adjustment for Illumina data (-C 50), skipped alignments with mapping quality less than 20 and bases with quality less than 20, we set the prior for variants to 0.001 and only called variants when the posterior probability that a locus was fixed for the reference base was less than 0.01.

This initial set of variants was filtered with a custom perl script that removed SNPs with mean coverage (per individual) of less than 2x, SNPs not supported by at least 10 reads of the non-reference allele, SNPs fixed for the non-reference allele in the sample of individuals, SNPs with a mapping quality less than 30, missing data for more than 20% of the individuals, or a minimum minor allele frequency of less than 0.005. We also excluded SNPs with more than two alleles and SNPs where forward and reverse reads were both observed at appreciable frequencies (> 0.01; this is not expected with our library preparation protocol). This left us with 11,990 SNPs for downstream analyses.

Genotype estimation. We used an empirical Bayesian approach to estimate genotypes. First, we used the expectation-maximization algorithm described in [61] and implemented in our own computer program (*estpEM* version 0.1; [62]) to obtain maximum likelihood estimates of allele frequencies at the 11,990 SNPs while accounting for uncertainty in genotypes as captured in the genotype likelihoods calculated by *samtools/bcftools*. We then obtained the posterior probabilities of the three genotypes at each SNP (i.e., homozygous reference, heterozygote, homozygous non-reference) based on the genotype likelihoods and using the allele frequency estimates to define the prior probability of sampling each allele (the two samples at each biallelic locus were assumed to be independent with probabilities given by the allele frequencies). We used the mean of the posterior distribution as a point estimate of the genotype. These estimates for the

number of non-reference alleles range from zero (homozygous reference) to two (homozygous non-reference), and are not constrained to be integer-valued.

Multi-locus genome-wide association mapping of colour in *T. chumash* with GEMMA. We fit Bayesian sparse linear mixed models (BSLMMs) with GEMMA (version 0.95) [43] to estimate the proportion of variance in the colour variables (RG and GB) explained by the additive effects of the SNPs, and to identify the individual SNPs most associated with colour. Unlike traditional genome-wide association (GWA) mapping methods, this polygenic GWA method fits a single model with all SNPs simultaneously and thus mostly avoids issues related to testing a large number of null hypotheses. In particular, trait values are modelled as a function of a polygenic term and a vector of the (possible) measurable effects (associations) of each SNP on the trait (**β**). Variable selection is used to estimate the SNP effects. SNPs can be assigned an effect of 0 (not in the model) or a non-zero effect (in the model) [43,63]. A Markov chain Monte Carlo (MCMC) algorithm is used to infer the posterior inclusion probability (PIP) for each SNP, that is, the probability that each SNP has a non-zero effect. The polygenic term defines an individual's expected deviation from the grand phenotypic mean based on all of the SNPs (this assumes all SNPs have near-infinitesimal effects on the trait). It accounts for phenotypic covariances among individuals caused by their relatedness or overall genetic similarity as captured by a genome-estimated kinship matrix. The kinship matrix also serves to control for population structure and relatedness when estimating the effects of individual SNPs (**β**) along with their PIPs. Likewise, SNPs in linkage disequilibrium with the same causal variant effectively account for each other, such that only one or the other is needed in the model, and this is captured by the PIPs.

We fit separate BSLMMs for RG and GB colour variables. We ran 10 MCMC chains for each colour trait, and each chain comprised 1 million MCMC steps and a 200,000 step burn-in. Samples were recorded on every 10^{th} step in each chain. Individuals with low sequence coverage (less than 2x averaged across SNPs) or missing phenotype data were excluded from this analysis, leaving 437 individuals. Model-averaged effect estimates of each SNP on each colour trait were obtained by first combining the 10 chains, and then multiplying the PIP for each SNP by its effect conditional on it being in the model and adding the small contribution of each SNP to trait variation via the polygenic term [43]. Five SNPs had PIPs > 0.5 (and greater than 0.7) as well as minor allele frequencies > 0.05. These were used for subsequent analyses connecting genotype to fitness.

We additionally summarized the posterior distribution for the proportion of variation explained (PVE) by the genetic data (via additive effects) and the proportion of the PVE that was attributable to SNPs with in-

dividually measurable (i.e., non-infinitesimal) effects (PGE). These quantities are captured by model hyper parameters that integrate over the effects of individual SNPs on the traits.

Testing for marginal epistasis among colour-associated SNPs using MAPIT. We used the program MAPIT to test for marginal epistasis for colour among the SNPs in *Mel-Stripe*. This method avoids the large combinatorial search space that must be considered when testing for pairwise or higher order epistasis, by instead testing for a combined pairwise interaction between a given SNP and all other loci [44]. We focused this analysis specifically on the 158 SNPs in the *Mel-Stripe* locus and fit models for RG and GB colour traits separately. P-values for tests of marginal epistasis (i.e., non-zero variance components for epistatic effects) were computed using the recommended hybrid method that first implements a z-test to compute a p-value, but then re-computes the p-value with the Davies method if the initial values is less than 0.05 (this enhances the precision of calculations for low p-values without adding a large computational burden). We focused further analyses on the set of five SNPs with p-values less than 0.05/158 = 0.0003 for RG or GB colour variables.

We ran two follow-up analyses in GEMMA to determine whether and to what extent allowing for epistatic effects on colour among the five SNPs showing evidence of epistasis improved our ability to explain variation in RG and GB colour. First, we re-fit the BSLMMs for colour described in the preceding section with 10 epistatic terms accounting for all of the possible pairwise interactions among these five SNPs. This was done by including the products of the centred genotypes for each pair of SNPs in the model. MCMC conditions were as described in the previous section. Our focus here was primarily on the extent to which adding epistasis increased the trait variance explained. Second, as a guard against model overfitting, we used genomic prediction with 10-fold cross-validation to compare the predictive power of models with and without the 10 epistatic terms. Specifically, we divided the data set into 10 (nearly equal) subsets of individuals. Then, for each subset, we fit the BSLMM model with all other subsets and used the fitted model to predict the colour phenotype of the dropped subset. We did this using the same MCMC conditions described in the previous section, but with five rather than 10 chains for each model fit. Genomic predictions were averaged over the five chains. We repeated these analyses with and without the epistatic terms included in the model, and in each case computed the Pearson correlation and the coefficient of determination (r^2) between the predicted and observed color values. This was done for RG and GB colour variables.

Release and recapture experiment. On June 18th, 2019 we transplanted the marked specimens back onto host plant individuals at the locality they were collected from. This was done in three blocks, where each

block contained each treatment (MM and AC), using a single plant individual of each host species. A total of 76 randomly chosen individuals were released in each treatment in each block. Experimental plants were chosen to be separated from other plants by 'bare ground' (sandy or gravelly regions not containing plants), forming an 'experimental island'. Past studies have shown that dispersal across such bare ground is near absent [42,46,64–66]. The geographic locations of each treatment and block, and number of recaptured individuals, are provided in Table 1 in the main text.

We were interested in rapid changes in cryptic colouration because past studies in *Timema* have documented adaptive divergence between experimental populations within days upon transplantation to new environments, and because adult and penultimate instar *Timema* tend to live for only one to three weeks in the field, with bird predation being a major source of selective mortality [42,46,64,65]. Thus, on June 21st, 2019 we recaptured the surviving individuals using visual surveys and sweep nets. Past mark-recapture work has shown this protocol is highly effective at recapturing the overwhelming majority of surviving individuals [42,46,64–66].

Estimating phenotypic selection gradients. We fit Bayesian hierarchical generalized linear models to estimate the strength and form of phenotypic selection on colour on each host plant treatment (AC = *Adenostoma*/*Ceanothus*, MM = mountain mahogany). We assumed that the probability of survival (*p*, i.e., recapture) for each individual (*i*) was,

$$log\left(\frac{p_i}{1-p_i}\right)=\beta_{0k}+\beta_{1k}\times RG_i+\beta_{2k}\times GB_i+\beta_{3k}\times RG_i^{2}+\beta_{4k}\times GB_i^{2}+\beta_{5k}\times RG_i\times GB_i.$$

Here, RG and GB denote standardized colour scores (mean of zero and standard deviation of one). Squared terms correspond with quadratic selection (i.e., disruptive of stabilizing) and the final term captures correlational selection. Distinct regression coefficients were inferred for each block (*k*). We placed hierarchical priors on each of the regression coefficients, such that $\beta_{jk}\sim Normal(\mu_j,\tau_j)$.Here, *j* denotes one of the five sets of coefficients. We then placed mostly uninformative hyperpriors on the unknown mean and precision terms, that is, $\mu_j\sim Normal(0,1e^{-5})$and $\tau_j\sim gamma(0.01,0.001)$.We assumed a bernoulli error distribution. We obtained samples from the posterior distribution of this model using MCMC via the *rjags* (version 4.8) interface with *JAGS* (version 4.3.0) [67]. We ran three chains each with a 40,000 step burn-in, 40,000 sampling steps, and a 20-step thinning interval. We summarized posterior distributions for the model parameters based on the median and 95% equal-tail probability intervals.

Because we fit a generalized linear model (Bernoulli error, logit link), our estimates of selection coefficients are not equivalent to classical selection gradients as defined by [68]. Specifically, the phenotype-fitness map is in terms of change in log odds survival rather than the probability of survival itself. We thus computed approximate selection gradients from these coefficients following [69].

Connecting genotype to expected fitness with Bayesian model averaging. We next estimated the effects of the five colour-associated SNPs on expected fitness (i.e., survival), including all two-way (10 terms), three-way (10 terms), four-way (five terms) and five-way (one term) interactions. We used the expected fitness from the generalized linear models connecting colour to survival (see the preceding section) for the dependent variable in this analysis. Thus, we were interested specifically in the effect of genotype on fitness due to the effect of genotype on colour, and of colour on fitness. We used multiple regression with Bayesian model averaging to for this analysis, as this allowed us to consider a large number of covariates (five additive and 26 epistatic terms) while accounting for uncertainty in the effects of each covariate including which covariates have non-zero effects. Specifically, the multiple regression models were of the form: $\square p_i = \mu_i + \sum_j \delta_j \times \beta_j \times g_{ij}$, where $\square p_i$ is the expected fitness from the phenotypic selection analysis for individual *i*, μ_i is a block-specific intercept for individual *i* (always in the model), δ_j is a binary indicator variable denoting whether a term is or is not in the model, β_j is the effect of term *j* when it is in the model, and g_{ij} is the genotype (additive) or product of genotypes (epistasis) for individual *i* associated with term *j*, and the summation is over the 31 genetic covariates (additive and epistatic terms). The multiple regression models were fit with the *bms R* package (version 0.3.4; [70]). Zellner's g-prior was used for the regression coefficients with $g = N$, where N is the number of individuals in each host treatment [71]. A uniform prior was used for the different models; that is for the sets of covariates with non-zero effects. We fit separate models for each host plant treatment, and allowed the three blocks to have distinct intercepts. Parameter estimates were obtained using MCMC with a 10,000 step burn-in followed by 200,000 sampling steps, and using the birth-death sampler for exploring model space. Five MCMC chains were run for each host plant treatment. We summarized the results in terms of the posterior inclusion probability (PIP) for each possible additive and epistatic effect and the model-averaged values of these effects (that is, averaged over the possibilities of the terms being zero versus non-zero). We also used the PIPs to compute the expected number of additive, two-way, three-way, four-way and five-way epistatic effects. This is simply given by the sum of the PIPs across the covariates comprising each of these types of effects.

We used leave-one-out cross validation to assess the performance of these models. Using predictive rather than explanatory power to assess model fit is expected to help guard against overfitting/over parameterization of models. Importantly, assessment of predictive power necessarily averages over uncertainty in covariate effects (including which covariates have non-zero effects). Our tests of predictive power involved 200 cross-validation runs, with one observation withheld from model fitting each time. The value of the missing observation was then estimated from the model. For each cross-validation run, we used a single MCMC chain (to avoid a very large computational burden) but otherwise used the same MCMC conditions described in the preceding paragraph. We then measured predictive power based on the Pearson correlation between the observed (or rather, the values predicted from the selection gradient analyses) and predicted expected fitness values. We repeated the cross-validation analysis with models allowing for only additive effects to assess the difference in predictive performance of additive only versus additive plus epistasis models for expected fitness.

We further evaluated the robustness of our genotype to fitness analyses by fitting similar models with Bayesian lasso and Bayesian ridge regression (see Supplementary Information). This yielded results similar to those presented in the main text. We also considered models with dominance effects (see Supplementary Information). Finally, we asked whether our estimates for the number of additive, two-way, three-way, four-way and five way epistatic effects differed from null expectations if genotype was independent of expected fitness. To do this, we repeatedly (100 times) selected five random SNPs with allele frequencies similar to the five colour-associated SNPs (+- 0.025 of the minor allele frequency for each). We then fit the same Bayesian multiple regression model used for the five colour-associated SNPs (and including the five additive and 26 epistatic terms based on the five random SNPs) using the same MCMC conditions described above. From this output, we derived a null distribution for the expected number of additive, two-way, three-way, four-way and five-way epistatic effects, which we compared to the numbers inferred based on the five colour-associated SNPs for both the AC and MM treatments.

Visualizing and quantifying the fitness landscape. We characterized the fitness landscape linking genotype at the five colour-associated SNPs to fitness in each host treatment based on the results from the Bayesian multiple regression with model-averaging described in the previous section. Specifically, we estimated the expected fitness of each five-locus genotype based on the model-averaged coefficient estimates. Note that this includes possible additive and epistatic effects of the SNPs on fitness. From this, we were constructed the fitness landscape on each host as a *n*-dimensional hypercube where each vertex if a five-locus genotype and edges connect all genotypes that differ by a single allele (here *n* is 10 as there are five SNPs each with three possible genotypes). We visualized the *n*-dimensional hypercube as a network

that retains the appropriate edges but distorts the length and organization of the edges relative to the projection of the hypercube in a 10 dimensional space. This was done with the *igraph* package in *R* (version 1.2.4) [72].

We then formally characterized the ruggedness of the fitness landscape in each treatment. Ruggedness has been defined in myriad ways (e.g., [73–77]). Here we used a slight variant/modification of an approach for defining and quantifying ruggedness described by [78]. Specifically, from either (i) a random genotype among those observed empirically in the sample of *T. chumash* individuals, or (ii) any random five-locus genotype (i.e., including genotypes that were not observed) we seeded a random walk across the fitness landscape. For each random walk, ten steps were taken along the fitness landscape. Each step involved a move to a neighbouring genotype (that is, one that differed by one allele) that was in no way dependent on the fitnesses of the genotypes. Then, for each walk, we calculated the net change in fitness between the first and last genotype, and the cumulative, absolute change in fitness values over the walk (i.e., the sum of the absolute values of the differences in fitnesses between each successive pair of genotypes). We repeated this procedure 10,000 times, and then measured the ruggedness of each landscape as the difference between the net change in fitness and the cumulative, absolute change in fitness.

Note, that if a walk involves a monotonic increase or decrease in fitness, these values will be the same and our metric will be zero. In contrast, if moves alternate between large fitness increases and decreases (as can occur when epistasis for fitness is ubiquitous), this value will be large. We report results starting from occupied (i.e., genotype empirically observed in the specific individuals we sequenced) regions of the landscape, and from any random genotype. The former provides a better characterization of the landscape in the region where *T. chumash* currently occur, whereas the latter provides a view across all possibilities (at least when ignoring loci in the genome other than the five colour-associated SNPs treated here).

Test for synteny between *T. cristiane* and *T. chumash* colour loci. We generated a chromosome-scale reference genome for *T. chumash* from information on proximity ligation of DNA in chromatin and reconstituted chromatin, which we then used to test for synteny of the colour loci between *T. cristinae* and *T. chumash*. The *T. chumash* specimen used was collected in 2017 from oak (*Quercus* sp.) from the population GR8.06Q (latitude and longitude 34.22, -117.71, respectively), flash frozen in liquid nitrogen, stored at -80C, and de-gutted before being shipped on dry ice for sequencing. The *T. chumash* genome was assembled from a combination of PacBio reads and DNA sequence data from Chicago and Hi-C li-

braries. Library construction, sequencing and assembly with the HiRise pipeline were outsourced to Dovetail Genomics. The assembled genome had a N50 of 176 Mb and a total size of 2.4 Gb.

We used the program CACTUS (version 1.0) to align the *T. chumash* genome to our existing, annotated *T. cristinae* genome [79] . We next extracted synteny blocks from the HAL-format graph-based comparative alignment produced by CACTUS with HALSYNTENY [80]. *T. cristinae* scaffold 128 from LG8 (the region spanning the colour loci) aligned with *T. chumash* scaffold 504 (a 140 Mb, chromosome-size scaffold). We visually examined the alignment between these two scaffolds by constructing a dot plot in R based on the synteny blocks.

Phenotype-free analyses and tests for pleiotropy of colour loci

We fit Bayesian regression models with model averaging as described above but with survival (binary) rather than expected fitness from the phenotypic selection analysis as the response variable. We considered two sets of models: (i) those with only additive effects for the five colour-associated SNPs, and (ii) those with additive and epistatic effects for the colour-associated SNPs. Leave-one-out cross-validation was used to assess predictive performance of each set of models (in terms of predicting survival) as described above. We further assessed the direct association of genotype with survival (independent of colour) by testing for marginal epistasis of each SNP within Mel-Stripe on survival with MAPIT [44]. We did this treating survival as a continuous rather than binary response variable, as we obtained inconsistent results using the latter approach. This is also consistent with the robustness of linear models to misspecification noted by [63]. The hybrid algorithm noted above was again used to determine *P*-values.

We fit additional Bayesian regression model with model averaging to test for pleiotropic effects of the colour-associated SNPs on other traits affecting survival. To do this, we used survival (binary) as the response variable and included colour (linear and quadratic terms for RG and GB, and the interaction between RG and GB). We again considered two sets of models: (i) those with only colour, and (ii) those with colour and with additive and epistatic effects for the five colour-associated SNPs. This allowed us to compare the predictive performance of colour alone, with colour and SNPs. Cross-validation analyses were used to compare the sets of models and conducted as described above. Increased predictive performance of colour and SNPs relative to colour is predicted if the SNPs affect fitness through traits other than just colour [47].

References

1. Barrett, R. D. H. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**, 767–780 (2011).
2. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution (N. Y).* **67**, 1235–1250 (2013).
3. Barrett, R. D. H., Rogers, S. M. & Schluter, D. Natural selection on a major armor gene in threespine stickleback. *Science (80-.).* **322**, 255–257 (2008).
4. Barrett, R. D. H. *et al.* Linking a mutation to survival in wild mice. *Science (80-.).* **363**, 499–504 (2019).
5. Gratten, J. *et al.* A localized negative genetic correlation constrains microevolution of coat color in wild sheep. *Science (80-.).* **319**, 318–320 (2008).
6. Lamichhaney, S. *et al.* A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science (80-.).* **352**, (2016).
7. Coberly, L. C. & Rausher, M. D. Pleiotropic effects of an allele producing white flowers in Ipomoea purpurea. *Evolution (N. Y).* **62**, 1076–1085 (2008).
8. Korves, T. M. & others. Fitness effects associated with the major flowering time gene FRIGIDA in Arabidopsis thaliana in the field. *Am. Nat.* **169**, 141–157 (2007).
9. Rockman, M. V. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1–17 (2012).
10. de Visser, J., C., F., T. & Elena, S. F. The causes of epistasis. *Proc. R. Soc. B Biol. Sci.* **278**, 3617–3624 (2011).
11. Arnegard, M. E. *et al.* Genetics of ecological divergence during speciation. *Nature* **511**, 307–311 (2014).
12. Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).
13. Kryazhimskiy, S., Rice, D. P., Jerison, E. R. & Desai, M. M. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522 (2014).
14. Marques, D. A. *et al.* Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nat. Ecol. Evol.* **2**, 1128–1138 (2018).
15. Natarajan, C. *et al.* Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. *Science (80-.).* **340**, 1324–1327 (2013).
16. Dettman, J. R., Sirjusingh, C., Kohn, L. M. & Anderson, J. B. Incipient speciation by divergent adaptation and antagonistic epistasis in yeast. *Nature* **447**, 585–588 (2007).
17. Orr, H. A. The population genetics of speciation - the evolution of hybrid incompatibilities. *Genet-*

ics **139**, 1805–1813 (1995).

18. Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).
19. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and Complex Phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
20. Wilfert, L. & Schmid-Hempel, P. The genetic architecture of susceptibility to parasites. *BMC Evol. Biol.* **8**, 187 (2008).
21. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science (80-.).* **312**, 111–114 (2006).
22. Gavrilets, S. *Fitness Landscapes and the Origin of Species*. (Princeton University Press, 2004). doi:10.2307/j.ctv39x541
23. Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. Sixth Int. Congr. Genet.* **1**, 356–366 (1932).
24. Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27**, 323–331 (2011).
25. Whitlock, M. C., Phillips, P. C., Moore, F. B. & Tonsor, S. J. Multiple Fitness Peaks and Epistasis. *Annu. Rev. Ecol. Syst.* **26**, 601–629 (1995).
26. Whitlock, M. C. Founder effects and peak shifts without genetic drift: Adaptive peak shifts occur easily when environments fluctuate slightly. *Evolution (N. Y).* **51**, 1044–1048 (1997).
27. Kingsolver, J. G. *et al.* The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245–261 (2001).
28. Sinervo, B. & Svensson, E. Correlational selection and the evolution of genomic architecture. *Heredity (Edinb).* **89**, 329–338 (2002).
29. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
30. Plucain, J. *et al.* Epistasis and Allele Specificity in the Emergence of a Stable Polymorphism in Escherichia coli. *Science (80-.).* **343**, 1366–1369 (2014).
31. Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biol.* **8**, (2010).
32. Sandoval, C. P. Differential visual predation on morphs of Timema cristinae (Phasmatodeae:Timemidae) and its consequences for host range. *Biol. J. Linn. Soc.* **52**, 341–356 (1994).
33. Sandoval, C. P. The effects of the relative geographic scales of gene flow and selection on morph frequencies in the walking-stick Timema cristinae. *Evolution (N. Y).* **48**, 1866–1879 (1994).
34. Sandoval, C. P. & Nosil, P. Counteracting selective regimes and host preference evolution in eco-

types of two species of walking-sticks. *Evolution* **59**, 2405–13 (2005).

35. Comeault, A. A. *et al.* Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. *Curr. Biol.* **25**, 1975–1981 (2015).
36. Nosil, P. *et al.* Natural selection and the predictability of evolution in Timema stick insects. *Science (80-.).* **359**, 765–770 (2018).
37. Villoutreix, R. *et al.* Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science* 369, 460-466 (2020).
38. Lindtke, D. *et al.* Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Mol. Ecol.* **26**, 6189–6205 (2017).
39. Endler, J. a. A framework for analysing colour pattern geometry: adjacent colours. *Biol. J. Linn. Soc.* **107**, 233–253 (2012).
40. Endler, J. A. On the measurement and classification of colour in studies of animal colour patterns. *Biol. J. Linn. Soc.* **41**, 315–352 (1990).
41. Hurvich, L. M. *Color vision*. (Sinauer Associates, Sunderland, MA, 1981).
42. Gompert, Z. *et al.* Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17**, 369–79 (2014).
43. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
44. Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.* **13**, (2017).
45. Comeault, A. A., Ferreira, C., Dennis, S., Soria-Carrasco, V. & Nosil, P. Color phenotypes are under similar genetic control in two distantly related species of Timima stick insect. *Evolution (N. Y).* 1283–1296 (2016). doi:10.1017/CBO9781107415324.004
46. Nosil, P. & Crespi, B. J. Experimental evidence that predation promotes divergence in adaptive radiation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9090–5 (2006).
47. Rennison, D. J., Heilbron, K., Barrett, R. D. H. & Schluter, D. Discriminating selection on lateral plate phenotype and its underlying gene, Ectodysplasin, In threespine stickleback. *Am. Nat.* **185**, 150–156 (2015).
48. Wright, S. The shifting balance theory and macroevolution. *Annu. Rev. Genet.* **16**, 1–19 (1982).
49. Coyne, J. A., Barton, N. H. & Turelli, M. Perspective: A Critique of Sewall Wright's Shifting Balance Theory of Evolution. *Evolution (N. Y).* **51**, 643–671 (1997).
50. Wade, M. J. & Goodnight, C. J. Perspective: The theories of Fisher and Wright in the context of metapopulations: When nature does many small experiments. *Evolution (N. Y).* **52**, 1537–1553 (1998).

51. Reimchen, T. E. Predator-induced cyclical changes in lateral plate frequencies of Gasterosteus. *Behaviour* **132**, 1079–1094 (1995).
52. Coyne, J. A. & Orr, H. A. *Speciation*. (Sinauer Associates, 2004).
53. Sackman, A. M. & Rokyta, D. R. Additive phenotypes underlie epistasis of fitness effects. *Genetics* **208**, 339–348 (2018).
54. Knief, U. *et al.* Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. *Nat. Ecol. Evol.* **3**, 570–576 (2019).
55. Hench, K., Vargas, M., Höppner, M. P., McMillan, W. O. & Puebla, O. Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nat. Ecol. Evol.* **3**, 657–667 (2019).
56. Lewontin, R. C. *The genetic basis of evolutionary change*. (Columbia University Press, 1974).
57. Scheffer, M. *Critical transitions in nature and society*. (Princeton University Press, 2009).
58. Scheffer, M. *et al.* Anticipating Critical Transitions. *Science (80-.).* **338**, 344–348 (2012).
59. Endler, J. A. & Mielke, P. W. Comparing entire colour patterns as birds see them. *Biol. J. Linn. Soc.* **86**, 405–431 (2005).
60. Parchman, T. L. *et al.* Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol. Ecol.* **21**, 2991–3005 (2012).
61. Li, H. *A statistical framework for {SNP} calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data*. (Bioinformatics 27, 2011).
62. Soria-Carrasco, V. *et al. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation*. *Science* **344**, (2014).
63. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
64. Nosil, P. Reproductive isolation caused by visual predation on migrants between divergent environments. *Proc. R. Soc. B Biol. Sci.* **271**, 1521–1528 (2004).
65. Nosil, P. *et al.* Genomic consequences of multiple speciation processes in a stick insect. *Proc. R. Soc. B Biol. Sci.* 5058–5065 (2012). doi:10.1098/rspb.2012.0813
66. Sandoval, C. P. Persistence of a walking-stick population (Phasmatoptera: Timematodea) after a wildfire. *Southwest. Assoc. Nat.* **45**, 123–127 (2000).
67. Plummer, M. rjags: Bayesian Graphical Models using MCMC. *R Packag. version 4-8.* (2018).
68. Lande, R. & Arnold, S. J. The Measurement of Selection on Correlated Characters. *Evolution (N. Y).* **37**, 1210–1226 (1983).
69. Janzen, F. J. & Stern, H. S. Logistic regression for empirical studies of multivariate selection. *Evolution (N. Y).* **52**, 1564–1571 (1998).

70. Zeugner, S. & Feldkircher, M. Bayesian Model Averaging Employing Fixed and Flexible Priors: TheBMSPackage forR. *J. Stat. Softw.* **68**, (2015).
71. Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference Decis. Tech.* (1986).
72. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex Syst.* **1695**, 1–9 (2006).
73. Weinberger, E. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.* **63**, 325–336 (1990).
74. Vassilev, V. K., Fogarty, T. C. & Miller, J. F. Information Characteristics and the Structure of Landscapes. *Evol. Comput.* **8**, 31–60 (2000).
75. Kouyos, R. D. *et al.* Exploring the complexity of the HIV-1 fitness landscape. *PLoS Genet.* **8**, e1002551–e1002551 (2012).
76. Malan, K. M. & Engelbrecht, A. P. A survey of techniques for characterising fitness landscapes and some possible ways forward. *Inf. Sci. (Ny).* **241**, 148–163 (2013).
77. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
78. Poursoltan, S. & Neumann, F. Ruggedness Quantifying for Constrained Continuous Fitness Landscapes. *Evolutionary constrained optimization* 29–50 (2015). doi:10.1007/978-81-322-2184-5_2
79. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).

80. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

Figure 1. Schematic of the hypotheses examined here. A) Epistasis can arise at two fundamental levels; due to non-linear cellular processes in the genotype to phenotype map and due to non-linear selection in the phenotype to fitness map.

Figure 2. Objectively quantifying colour variation from digital photographs. A) Spectral sensitivities of cones of a hypothetical tetrachromatic receiver (e.g., a bird [59]). The human visible spectrum represents cones that capture long wavelengths (L, red), medium wavelengths (M, green) and short wavelengths (S, blue). Differences between L-M and M-S activities are the most common responses in colour perception [40,41]. Relative differences between values of red and green (RG=(R – G)/(R +G)) and of green and blue (GB=(G – B)/(G + B)) [39] extracted from digital photographs can be used as an approximation to this physiological response. This approach was used considering *T. chumash* does not reflect UV (see Supplementary Information and Supplementary Figure 1). B) RG and GB are orthogonal measures and together capture the range of colour variation observed in *T. chumash*.

Figure 3. Genetics of cryptic colouration in *T. chumash*. A) Hyper-parameter estimates from GEMMA models without versus with interactions between pairs of the five SNPs shown to have epistatic effects on colouration in MAPIT (additive versus epistasis models hereafter). PVE = percent variance explained. Vertical bars denote 95% credible intervals. The darker shading within each bar represents PGE = percent of the PVE explained by SNPs with individually detectable effects. RG = red-green. GB = green-blue. B) Posterior inclusion probabilities (PIPs) for SNPs for RG and GB colour traits. The grey, shaded region represents the *Mel-Stripe* locus. C) Results from MAPIT analyses testing for epistatic effects on colouration. Dashed lines represent strict Bonferroni corrected thresholds for statistical significance. D) Genomic estimated breeding values (GEBVs) for additive versus epistasis models. E) and F) Genomic prediction of phenotype from genotype in the additive versus epistasis models, for RG and GB, respectively. Lines denote predicted values from linear regression.

Figure 4. Phenotypic selection on colouration in a field experiment. A) Photographs of representative *T. chumash* used to phenotype colouration and to measure survival in the field experiment. B) and C) show red-green (RG) and green-blue (GB) colour scores for the individuals used in each treatment. Filled circles represent recaptured individuals whereas empty circles represent those not recaptured (survivors versus dead, respectively). D) Coefficients (Coef.) of correlational selection in each treatment and block. Errors bars represent 95% credible intervals. E) A heat map of survival probabilities from the selection gradient analysis, here shown for block 1 in the AC treatment (see Supplementary Figure 3 for additional results). Darker colours reflect higher survival probability.

Figure 5. Evidence of epistasis for fitness and the fitness landscape. A) Model-averaged effects from Bayesian regression. Shown are main effects and interactions between SNPs, which represent additive versus epistatic effects, respectively. B) The number of estimated effects. Error bars represent +- posterior standard deviations (analogous to standard errors), and asterisks denote significant differences ($P < 0.05$, one-sided test) from null simulation results using randomly drawn SNPs. C) An example of pairwise epistasis between two loci (i.e., SNPs). Alleles at each locus have been coded in terms of whether they increase green or melanistic colouration (G and M, respectively). Darker colours represent higher survival probability. D) and E) show empirical fitness landscapes in each treatment. Nodes are genotypes, with those separated by a single substitution connected by a line. Darker colours represent higher survival probability. The smallest nodes represent genotypes not directly observed in our study whereas other nodes are scaled in size according to observed sample size. Values shown are metrics of ruggedness based on repeated 10-step random walks on the landscape, either from anywhere in the landscape or restricted to space occupied by observed genotypes.

Comparison	Response	Independent terms						
		Addi-tive	Epista-sis	Colour	CV r	Cis	*P*	r^2
Main	Exp. Fit-ness	Y	N	N/A	0.47	0.35-0.57	<0.001	0.22
Main	Exp. Fit-ness	Y	Y	N/A	0.60	0.50-0.68	<0.001	0.36
Phenotype-free	Survival	Y	N	N	0.19	0.05-0.32	0.008	0.04
Phenotype-free	Survival	Y	Y	N	0.32	0.19-0.44	<0.001	0.10
Pleiotropy	Survival	N	N	Y	0.48	0.36-0.58	<0.001	0.23
Pleiotropy	Survival	Y	Y	Y	0.52	0.42-0.62	<0.001	0.28

Table 1. Summary of cross-validation (CV) predictive performance from Bayesian model averaging. Comparison refers to the analyses considered, where 'Main' are the focal analyses first reported using the five strongly colour-associated SNPs, 'Phenotype-free' do not consider colour in the response variable directly, and 'Pleiotropy' consider genotype and colour as independent variables (see text for further details). Response = response variable, either expected fitness (Exp. Fitness) from the phenotypic selection analysis that includes colour, or survival directly. Additive, Epistasis, and Colour denote independent terms, specifically whether the model included additive effects for the five focal SNPs, epistatic interactions among these SNPs, and phenotypic colour scores (Y is yes, N is no, and N/A is not applicable). We report the leave-one-out cross-validation Pearson correlation between the observed and predicted response variable (CV r), the 95% confidence intervals on the correlation (CIs), the *P*-value for a two-tailed test of the null hypothesis that the correlation is 0, and the predictive r^2.