Natural selection drives genome-wide evolution via chance genetic associations

Zachariah Gompert (zach.gompert@usu.edu)
Utah State University

Jeffrey Feder
University of Notre Dame https://orcid.org/0000-0002-8797-980X

Patrik Nosil
University of Sheffield https://orcid.org/0000-0002-8271-9005

Article

Keywords: Functionally-neutral Genetic Regions, Statistical Associations, Indirect Selection

DOI: https://doi.org/10.21203/rs.3.rs-537420/v1

License: ☺️ ☀️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Natural selection drives genome-wide evolution via chance genetic associations

Zachariah Gompert*,1,2, Jeffrey L. Feder3,
Patrik Nosil4

1Department of Biology, Utah State University, Logan, UT 84322, USA
2Ecology Center, Utah State University, Logan, UT 84322, USA
3Department of Biological Sciences, University of Notre Dame,
Notre Dame, Indiana 46556, USA
4CEFE, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3,
Montpellier, 34293, France
*To whom correspondence should be addressed;
E-mail: zach.gompert@usu.edu.

May 2021
Understanding selection’s impact on the genome is a major theme in biology. Functionally-neutral genetic regions can be affected indirectly by natural selection, via their statistical association with genes under direct selection. The genomic extent of such indirect selection, particularly across loci not physically linked to those under direct selection, remains poorly understood, as does the time scale at which indirect selection occurs. Here we use field experiments and genomic data to show that widespread statistical associations with genes known to affect fitness in stick insects, deer mice and stickleback fish cause many genetic loci across the genome to be impacted indirectly by selection. We then show that statistical associations with other, unknown causal variants make aspects of evolution more predictable in stick insects. Thus, natural selection combines with chance genetic associations to affect genome-wide evolution across linked and unlinked loci and even in modest-sized populations.

Evolution is best understood for single traits and genes, such as the Ectodysplasin gene controlling lateral plates in stickleback fish [1], the Agouti gene controlling coat color in mice [2], and Cortex, which controls wing coloration in moths and butterflies [3]. The effects of selection on such traits and genes can, in theory, be transmitted to other (i.e., not directly selected) correlated genetic regions across the genome [4, 5, 6, 7]. Thus, even functionally neutral genetic regions can experience selection indirectly due to statistical association (i.e., linkage disequilibrium, LD) with selected regions (Fig. 1). There is increasing evidence that many traits are polygenic [8, 9] such that adaptation often involves selection on many regions of the genome, as documented in rabbits, cichlid fish, great-tits, Rhagoletis flies, and Arabidopsis [10, 11, 12, 13, 14]. In such cases, the indirect effects of selection could be even more pervasive.

Studies of molecular evolution in flies, humans and other organisms have shown that much of the genome can be impacted by indirect selection [15, 16, 17]. For example, correlations across the genome between genetic diversity and recombination provide evidence of widespread background selection [18, 19, 20], and frequent selective sweeps have been documented in Drosophila [21, 22]. However, such analyses of molecular evolution average over long periods of time, and thus it remains unclear whether indirect selection is pervasive on shorter time scales, such as generations or decades (but see, e.g., [23, 24]). Our best evidence for short-term genetic consequences of selection comes from studies of ecologically important traits determined by one or a few genes, such as in finches, stickleback, and mice [25, 26, 27, 28]. Strong selection on such traits can have important immediate evolutionary and ecological consequences [29, 30], but it is unclear if this affects few or many regions of the genome. Thus, the extent and genome-wide consequences of indirect selection on short to moderate time scales are poorly understood. This makes our understanding of evolution incomplete because the genomic distribution of selection...
affects the dynamics of adaptation and speciation, and the predictability of evolution. Indeed, indirect selection could have notable effects on the short-term predictability of evolution (e.g., the direction of allele frequency change), even if it does not increase the magnitude of per-generation change much beyond genetic drift alone. Finally, the genomic consequences of selection have implications for many evolutionary applications, such as demographic and phylogenetic inference, which often assume neutrality [31, 32].

Elucidating the consequences of selection on the genome is challenging because most genetic regions are affected by multiple factors, whose contributions are difficult to isolate without experimental approaches [33]. For example, the LD that mediates indirect selection is affected by past selection, recombination, chance in finite populations (i.e., genetic drift), and population structure [34]. Specifically, genetic drift and population structure can generate widespread LD, even among unlinked loci on different chromosomes. In terms of drift, LD arises when combinations of alleles are over-represented by chance. The effect of structure is commonly considered in genome-wide association mapping studies, where even structure within populations sometimes generates sufficient LD to confound genotype-phenotype associations if not properly controlled for [35, 36]. Thus, although facilitated by chromosomal linkage, we stress that LD (a statistical association) is distinct from physical linkage (a physical feature of genomes). Consequently, the effects of indirect selection may extend beyond loci physically-linked to those under direct selection.

Studies using whole genomes from many individuals and experimental approaches are thus now necessary to understand the potentially complex interplay of factors that shape genome-wide evolution. Here, we determine the genome-wide effects of selection by analyzing 491 whole-genome sequences from experimental populations of stick insects. We specifically test two competing hypotheses: (i) indirect selection caused by LD with causal variants is limited to a few regions of the genome, and (ii) indirect selection caused by LD with causal variants is widespread across the genome. The basis for the relatively untested widespread indirect selection hypothesis is that, especially in small or moderate-sized populations experiencing polygenic selection, sufficient stochastic LD arises by drift for many or even all single nucleotide polymorphisms (SNPs) to exhibit non-trivial LD with at least one causal variant under selection (Fig. 1). If this occurs, direct selection will be indirectly transmitted to many SNPs across the genome via LD. In reality, the hypotheses posed here form a continuum where a larger number of indirectly selected loci lend greater support to the widespread indirect selection hypothesis. We focus on the *Timema* stick insect system, but show the generality of our results by additional re-analysis of field transplant experiments in mice and stickleback fish [28, 37].
Results

Study system and measuring survival

Our focal organism is the wingless, herbivorous stick insect *Timema cristinae*. This species is endemic to southern California and exhibits green, green-striped, and melanistic body color morphs (Fig. 2A). This system has two key features that make it amenable to testing hypotheses concerning the genomic impact of selection. First, the three morphs are genetically-controlled and known to be subject to strong natural selection [38, 39]. This includes selection for crypsis on the leaves and stems of the host-plants *Adenostoma fasciculatum* and *Ceanothus spinosus*, and negative frequency-dependent selection. Moreover, the 10 megabase (Mbp) region on linkage group 8 (LG8) that controls color (*Mel-Stripe* locus hereafter) exhibits heterosis [40]. Thus, past evidence for selection sets up a clear expectation for selection on *Mel-Stripe*, predicted to impact the genetic loci in LD with this locus. There is potential at least for selection on other traits and genomic regions as well, given past evidence for population divergence in body size, host preference, cuticular hydrocarbons and subtle genome-wide differentiation [41, 42]. Second, controlled mark-recapture experiments are possible to estimate survival, and thus the effects of selection on *Mel-Stripe* and genome wide.

We leverage such an experiment here, in which past work reported evidence for selection on genetic regions associated with color and also other loci, with details that varied with elevation (i.e., microclimate) differences among the transplant sites [43]. This past work thus provides evidence for selection but considered only a small subset of the genome using genotyping-by-sequencing (GBS) data and lacked a well-assembled reference genome. Consequently, how much of the genome was indirectly affected by selection because of LD and the extent to which LD occurred between linked versus unlinked loci is unknown. Here, we use whole-genome sequences to estimate indirect selection in the experiment. Experimental details are as follows. In 2011, we collected 500 *T. cristinae* from a single population (4.51753°N, 119.80125°W, named FHA), where the dominant host is *Adenostoma* and all individuals were collected from this host [43]. A tissue sample was taken from each insect, with 491 used successfully for whole genome sequencing. These individuals were then deployed in a mark-recapture experiment, with 50 insects transplanted onto each of 10 experimental plants (five *Adenostoma* and five *Ceanothus*). Our experiment focused on short-term survival in adult and penultimate instar *Timema*, which tend to live for ~1-3 weeks in the field [43] and on which predation by birds is a major source of selective mortality [38]. Thus, after eight days, we recaptured surviving stick insects to quantify genome-wide allele frequency change between release and recapture, and thus test for selection. Notably, past work has shown that recapture is a good proxy for survival, due to the limited dispersal ability of these wingless insects [38].

The results presented in the main text focus on ~7 million single nucleotide polymorphism (SNPs) from whole-genome sequence data from 246 *T. cristinae* transplanted...
to *Adenostoma* [39]. This approach focuses on survival in a natural habitat, as exemplified by studies of Darwin’s finches and other birds [44, 45], and thus avoids effects of transplantation to a novel host. The results in the main text are thus based on allele-frequency change and LD in this experimental population of 246 individuals, which is representative of deme sizes of *T. cristinae* in nature [46], but also sufficiently finite to allow for stochastic LD to arise among unlinked loci via random sampling and pairing of gametes. Thus, our design realistically embodied how selection acts in this species, and others living in heterogeneous or patchy environments. We present complementary and largely parallel results based on 245 *T. cristinae* transplanted to *Ceanothus* in the Supplementary Material (Supplementary Text and Supplementary Figures S1-S3).

**Many genetic regions are affected by indirect selection via the *Mel-Stripe* locus**

We begin with an illustrative example of the potential for indirect selection focused on the *Mel-Stripe* locus, which exhibits multiple previous lines of evidence for selection including in the specific experiment analyzed here [40, 39]. This serves as a proof-of-concept of indirect selection using a single directly-selected locus, on which we subsequently expand to consider genome-wide selection. In this context, we focus on whether substantial LD with *Mel-Stripe* is limited to SNPs near the locus on LG8, or extends to distant SNPs including those on other chromosomes. We thus estimated LD as the coefficient of determination between *Mel-Stripe* genotype (determined from a principal components analysis of SNPs within the *Mel-Stripe* locus) and genotype estimates for each of ~7 million genome-wide SNPs (Supplementary Figure S4). As expected, we observed an excess of high LD between *Mel-Stripe* and individual SNPs in or near *Mel-Stripe* (Fig. 2B). However, numerous SNPs on other chromosomes also exhibited high LD with *Mel-Stripe*. For example, 64 SNPs on eight chromosomes other than LG8 showed LD > 0.1, and LD was small but decidedly non-zero (i.e., > 0.01) for 183,675 non-LG8 SNPs (~3% of all SNPs). Thus, direct selection on a single locus (i.e., *Mel-Stripe*) combined with long-range LD necessarily induces a degree of indirect selection across the 13 chromosomes in the *T. cristinae* genome; below, we show that this hypothesized indirect selection does indeed occur.

**Generality of results**

To verify that such patterns were not unique to our specific stick insect data set, we tested for genome-wide LD between SNPs and a putative causal variant in two additional, independent data sets from two other organisms (see the Supplementary Materials for details). First, we used genomic data from a field enclosure experiment involving cryptically colored populations of deer mice (*Peromyscus maniculatus*) transplanted to light or dark-soil enclosures [28]. In this experiment, 481 deer mice were collected and released
in light-soil ($N = 217$) or dark-soil ($N = 229$) field-enclosures where they were subject to natural selection. Previous analyses of this experiment combined with genetic manipulations showed that a deletion ($\Delta Ser$) within the Agouti gene affected the color of these mice and causally affected fitness (i.e., was subject to direct selection). We found evidence that SNPs across the genome were associated with the $\Delta Ser$ genotype in both experimental populations (Fig. 3A,B). LD exceeded $\sim0.04$ for the 1% of SNPs most associated with $\Delta Ser$ in each population, and these high-LD SNPs were distributed across all 23 autosomes and the X chromosome. Moreover, about 20% of SNPs were in LD with $\Delta Ser$ to a non-trivial extent (i.e., $r^2 > 0.01$ for 20.0% and 23.0% of SNPs in the light-soil and dark-soil experimental populations, respectively).

Second, we analyzed genetic data from an experimental field population of threespine stickleback (Gasterosteus aculeatus) [37]. The experiment uncovered a single QTL for fitness (offspring number), with the peak signal associated with a SNP on chromosome 4 near the Eda locus (position 12,815,024), a gene known to affect bony armor and fitness [1, 27, 47]. We treated this SNP as the putative causal locus, and quantified LD between this QTL marker and each of 387 additional genome-wide SNPs in the set of 224 F2 females. LD between SNPs and the QTL marker was exceptionally high for most SNPs on chromosome 4 (Fig. 3C). However, we also detected LD between the QTL marker and SNPs on other chromosomes (Fig. 3C,D). For example, LD with the QTL marker exceeded $\sim0.04$ for about 10% of SNPs and exceeded 0.01 for 29% of SNPs, representing 1% and 22% respectively, of such SNPs not on chromosome 4. Together, these results show that widespread indirect selection has the potential to occur on short times scales not just in stick insects but in these mice and fish populations as well. This finding was not a point noted by either of the original studies [28, 37]. Similar to the case with Mel-Stripe in T. cristinae, these results show that direct selection on a single locus may cause substantial indirect selection on other loci in the genome.

The genome is affected by indirect selection because of and independent from Mel-Stripe

Having documented patterns of LD that would lead to indirect selection, we next quantified the total extent of selection across the genome during the T. cristinae field experiment. To do this, we used a Bayesian model to estimate the total strength of selection (i.e., direct plus indirect) experienced by each SNP from its association with survival while accounting for the effects of genetic drift (this is akin to genetically mapping survival with effects equal to selection differentials). Because the overwhelming majority of SNPs are unlikely to causally affect fitness (but see [9]), these estimates of total selection should mostly reflect indirect selection. This analysis revealed that different alleles at loci were on average associated with a 11% difference in relative survival probabilities (i.e., a mean difference in marginal fitness or selection differential of 11%, i.e., $|\hat{s}| = 0.11$, 1st quartile = 0.049, 3rd quartile = 0.17) (Supplementary Figure S5). Notably, because of genetic
drift, there was substantial uncertainty in these estimates (i.e., 90% credible intervals on
$s$ excluded 0 for only 135,447 or 2% of SNPs). Still, for ∼2 million SNPs (i.e., ∼29%),
the odds-ratio for non-zero total selection, defined as $\frac{x_i}{1-x_i}$ where $x_i = 2|\Pr(s_i > 0) - 0.5|$, exceeded 1. Thus, largely because of indirect selection, many genome-wide SNPs were
likely impacted to some extent by selection. LD with Mel-Stripe explained about 0.6%
of the variation in the selection detected (linear regression, $n = 7,243,463$, $P < 1 \times 10^{-15}$,
$r^2 = 0.006$). Thus, LD with Mel-Stripe likely caused some indirect selection representing
the realization of a portion of the potential for deterministic genomic change at unlike
SNPs in response to selection at a specific target locus. However, other sources of direct
selection are necessary to explain the documented patterns of indirect selection.

**Genome-wide LD predicts indirect selection**

We thus hypothesized that the genomically widespread indirect selection documented
above was not solely caused by LD with Mel-Stripe, but rather also by LD with other, un-
known causal variants. To test this hypothesis, we asked whether LD between SNPs and
expected fitness (i.e., survival probability) predicted the total selection experienced by
each SNP. The analytical procedure we used to address this question involved three steps:
(1) estimate expected fitness of each individual stick insect using genomic prediction, (2)
quantify how strongly each of the ∼7 million SNPs was statistically associated with ex-
pected fitness at the organism level, and (3) test whether the strength of the expected
fitness relationship from (2) (an estimate of indirect selection) predicts independent esti-
mates of total selection on each SNP (as described in the previous paragraph)(Fig. 4A).
Such predictive power would be expected if our estimates total selection mostly reflect
indirect selection, and thus this procedure also serves as a semi-independent validation
of our estimates of total selection above.

Thus, we first estimated expected fitness (i.e., survival probabilities) based on the ∼7
million SNPs and an ABC model of direct selection on (unknown) causal variants. This
approach is similar to polygenic genomic prediction methods with fitness as the predicted
phenotype and direct selection being analogous to marker-associated phenotypic effects
[48]. Importantly, the ABC genomic prediction approach accounts for the effects and
patterns of LD for the (unknown) causal variants, while also modeling uncertainty in
the identity of causal variants and their effects on fitness. In other words, the genomic
predictions of expected fitness numerically integrate over uncertainty in the number of
locations of causal variants. The ABC model had modest predictive power. Specifically,
cross-validation showed that our model outperformed a null model, reducing errors for
predicting survival by 19% (error rate for null model = 37%, error rate for fit model =
30%, $\frac{0.37 - 0.30}{0.37} = 0.19$ or 19%, $P < 0.001$)(Fig. 4B).

Having obtained genomic predictions of expected fitness, we next estimated the sta-
tistical association between the genotype at each of the ∼7 million SNPs and these
fitness estimates. We summarized the strength of this association between SNP geno-
types and expected fitness using the coefficient of determination from linear regression \( (r^2) \) (Fig. 4A). Then for the aforementioned third step in our procedure, we found that the strength of association between SNP genotypes and expected fitness was strongly explained by our estimates of total selection (Pearson \( r \) for log-odds of selection = 0.447, 95% CIs = 0.447-0.448, \( P < 0.0001 \)) (Fig. 4C). Thus, LD among (unknown) causal variants and genome-wide SNPs, as captured by the association between SNPs and genomic predictions of fitness, predicts the selection experienced by SNPs during the experiment about 22× better than does LD with *Mel-Stripe* (i.e., \( r^2 = 0.22 \) versus 0.01). This result lends support to the widespread indirect selection hypothesis.

**Consequences of indirect selection compared to random mortality**

We next asked whether the documented patterns of indirect selection led to different patterns of short-term genomic change than would be expected with random mortality (pure drift). This was done to both assess the genomic consequences of indirect selection and to determine whether our results were expected under a null model of random mortality. We did this by generating and analyzing 10 data sets identical to our own, but with survival randomized across individuals within each experimental block (host plant) (we limited our analyses to 10 randomized data sets because of the very large computational burden of running these analyses). We found that the magnitude of genome-wide allele frequency change for the observed and randomized data sets were very similar, with a mean absolute change of \( \sim 0.009 \) and 95th percentile change of \( \sim 0.026 \) in all cases. Likewise, the distribution of inferred total selection differentials were nearly identical (i.e., \( |\hat{s}| = 0.11 \) for the observed and each randomized data set, with credible evidence of selection impacting \( \sim 130,000 \) SNPs in each data set).

However, the observed and randomized data sets differed in a notable way. Specifically, the extent to which estimates of total selection were correlated with the strength of statistical association between each SNP and genomic predictions of expected fitness was higher for the observed data (\( r = 0.447 \) and 0.466 for the *Adenostoma* and *Ceanothus* host treatments, respectively) than for any of the randomized data sets (\( \bar{r} = 0.249 \), minimum \( r = 0.214 \), maximum \( r = 0.316 \), \( P = \frac{2}{11} \times \frac{1}{10} = 0.018 \)). Consequently, these associations with genomic predictions of expected fitness explained 22% of the variation in total selection for the observed data, but only 4-9% of the variation in the random mortality data sets. This provides evidence that the documented change reflects a contribution from indirect selection, even if selection did not increase the magnitude of short-term genomic change compared to drift. Because selection adds a deterministic component to evolution, this necessarily increases the extent to which short-term evolutionary change is predictable. In contrast to random drift, this includes prediction of the direction of allele frequency change at each SNP (i.e., neutral alleles positively associated with beneficial alleles or expected fitness can be expected to increase in frequency).
We have shown that LD among genome-wide SNPs likely resulted in SNPs across the genome being impacted by selection, which necessarily makes the short-term direction of evolutionary change for these SNPs more predictable. However, long-range LD is readily broken down by the independent assortment of chromosomes and recombination (i.e., with independent assortment, LD decays by half each generation). Thus, long-range LD is not expected to predict evolution over longer time scales. In contrast, LD among physically-linked loci (local LD hereafter) can be more persistent and thus lead to predictable patterns of indirect selection over time.

Thus, we next quantified the contribution of “local” LD to indirect selection during the transplant experiment. Local LD pertains to sets of contiguous and physically-linked SNPs, rather than individual SNPs. We consequently turn our attention now to window-based analyses that better reflect the consequences of local LD than do analyses of individual SNPs. We used a Hidden-Markov Model (HMM) approach to delimit windows with elevated LD relative to the genomic background, and did the same for elevated total selection (Fig. 5A,B). As predicted by effects of local LD, genomic regions (windows) with higher local LD at the onset of the experiment exhibited greater evidence of selection (Pearson correlation between logit mean $r^2$ LD and the mean log-odds of selection for each window, $r = 0.23$, CI = 0.23-0.24, $P < 0.0001$, $n = 143,800$ windows) (Supplementary Figure S6). Notably, the strength of this correlation was greater than for any of the 10 data sets with randomized mortality (mean Pearson correlation $r = 0.18$, maximum = 0.21, $P = 1/11 = 0.09$). Thus, both local and long-range LD contributed to indirect selection during the *T. cristinae* experiment.

**Discussion**

Understanding the impact of selection on the genome is important for understanding adaptation, speciation, genome evolution, and eco-evolutionary dynamics. Our results, and related work on linked selection discussed in the introduction [18, 19, 49, 21, 15, 16, 20, 17, 22, 32], show that functionally neutral loci in many regions of the genome can be affected by selection. Specifically, many of the ~7 million SNPs we analyzed were likely affected by indirect selection during our relatively brief experiment. Indeed, our results demonstrate indirect selection affecting SNPs on all 13 *T. cristinae* chromosomes even when considering only selection on *Mel-Stripe*, a major locus known to affect color pattern and fitness [39, 50]. Such widespread indirect selection is not unique to the *T. cristinae* system, but is a near mathematical certainty in any finite population subject to selection. Indeed, we conducted additional new genomic analyses of field experiments involving field mice and stickleback fish that demonstrate widespread indirect selection similar to that seen in the stick insects (see Supplementary Material and [28, 37]). Thus, widespread indirect selection is not limited to linked loci and the long time scales of...
molecular evolution (e.g., [15, 16, 17]), but is also prevalent among unlinked loci and on very short time scales, operating even within a single generation.

Nonetheless, our results also show that genetic drift caused by random mortality can mirror many of the genomic patterns detected in our analyses. For example, simulated random mortality caused nearly identical average allele frequency change and associated estimates of total selection. Perhaps this is not surprising, as the intensity of genomic change with ∼70% mortality (as occurred in our experiment) should be similar regardless of the specific individuals that live or die, and genotypes at many loci will be statistically associated with realized fitness (survival) even when mortality is random. This conclusion assumes that overall mortality is not higher when selection occurs (i.e., that selection is “soft”) [51, 52]. If instead selection results in increased mortality (not just genotype-specific mortality), total genomic change and estimates of selection should be higher than they would be in the absence of selection [53]. Moreover, we have convincing evidence of phenotypic selection in the *T. cristinae* experiment, and consequently the genomic change we documented necessarily includes indirect selection (i.e., mortality was not random but depended on color pattern) [43], and our results did differ from expectations based on random mortality in a few ways, such as (i) an increased association between estimates of total selection and the extent to which SNPs were statistically associated with genomic predictions of expected fitness, and (ii) an enhanced association between local LD and genomic windows of elevated total selection. These findings suggest that it is easier to explain or predict the targets of indirect selection than to explain or predict which loci will change the most by drift alone. Still, fully realizing this potential to predict contemporary evolutionary change is not a trivial task, but rather would require a good understanding of the nature of selection, patterns of LD, and the genetic basis of relevant trait variation [54].

With the above caveats and considerations in mind, our results have implications for three general themes in biology: (1) the role of deterministic versus stochastic processes in evolution, (2) the predictability of evolution, and (3) the dynamics of adaptation and speciation. In terms of determinism and chance, indirect selection via long-range LD reflects a combination of each. There is a stochastic component to indirect selection that is similar to the chance role mutation plays in theories of hitchhiking and genetic draft [55]. However, conditional on patterns of LD, indirect selection is a deterministic process. Thus, our results support a role for conditional determinism in short-term evolution and provide a demonstration that evolution is simultaneously stochastic and deterministic. Concerning predictability, the direction and magnitude of change at functionally neutral alleles can be predicted from LD and direct selection, although local LD is more likely to generate longer-term predictions about genetic differentiation, as demonstrated in flycatchers, butterflies, and other taxa [56, 57]. This is because local LD decays much more slowly than long-range genome-wide LD. Still, even LD between loci on different chromosomes only decays by 50% each generation, making predictions across several generations possible provided selection pressures and gene interactions remain consistent.

In terms of adaptation, theory predicts that selection is more efficient and selective
sweeps more frequent in large populations. Our results do not contradict this, but do argue that chance long-range LD will cause more of the genome to be affected by indirect selection in smaller populations. Thus, selection in large populations is focused and efficient, whereas in small populations the effects of selection are more widespread but diffuse. Indeed, one way to view the effects of indirect selection for physically unliked genes in small populations is that across generations it generates a genome wide reduction in effective population size that can add to that of genetic drift alone. This is true even if the direction of selection fluctuates in space or time, and this effect should be enhanced if selection is accompanied by increased mortality or reduced fecundity (i.e., when selection is hard and maladaptation reduces population size). In sum, our collective results suggest that indirect selection is not simply a statistical nuisance that frustrates attempts to find individual genes under direct selection. Rather, indirect selection is a critical component of both short-term and longer-term processes that shape genome-wide evolution and generate biodiversity.

Methods

Field release-recapture experiment

In 2011, we collected 500 *T. cristinae* from a single population (4.51753°N, 119.80125°W, named FHA), where the dominant host is *Adenostoma* and all individuals were collected from this host [43]. A small tissue sample (i.e., a portion of one leg) was taken from each insect using sterile scissors. Past work has shown no effect of such tissue sampling on survival in the field or lab [43]. These individuals were then deployed in a mark-recapture experiment, with ~50 insects transplanted onto each of 10 experimental plants (five *Adenostoma* and five *Ceanothus*) (see [43] for details). After eight days, surviving stick insects were collected to quantify genome-wide allele frequency change between release and recapture, and thus test for selection.

DNA sequence alignment, variant calling and genotype estimation for the experiment

Whole genome DNA sequence data for individuals from this experiment were described in [42, 39], where data was successfully obtained for 491 of the 500 transplanted insects. For the current study, we aligned the whole genome DNA sequence data from each of these 491 *T. cristinae* to the *T. cristinae* reference genome (version 1.3) using the *bwa* (version 07.10-r789) *mem* algorithm, and then used *samtools* (version 1.5) to compress, sort and index the alignments, and to remove PCR duplicates [58]. We then used the *GATK*HaplotypeCaller and *GenotypeGVCFs* modules (version 3.5) to call
variants and calculate genotype likelihoods [59]. After filtering SNPs, we obtained maximum likelihood estimates of allele frequencies for all experimental samples using an expectation-maximization (EM) algorithm, as described in [60] and implemented in our own C++ program, estpEM [42]. We then used these allele frequency estimates and the genotype likelihoods from GATK to calculate empirical Bayesian genotype estimates. See the Methods section of the Supplementary Material for additional details.

**Testing for indirect selection associated with Mel-Stripe**

Following [39], we defined the Mel-Stripe color locus as spanning 4.1 Mbps on scaffold \(\sim 702.1\) and \(\sim 6.4\) Mbps on the adjacent scaffold 128, both on linkage group 8 (LG8). This is a large, complex structural variant that includes a \(\sim 10\) Mbp putative inversion and \(\sim 1\) Mbp deletion [61, 50]. As in past work [40], we inferred each individual’s Mel-Stripe genotype from patterns of clustering in principle components analysis (PCA) space [62]. We then measured pairwise LD between each of the \(\sim 7\) million SNPs and the Mel-Stripe locus as the square of the correlation between genotypes [63]. See the Methods section of the Supplementary Material for additional details.

**Estimating total selection**

We next inferred the total selection experienced by each of the \(\sim 7\) million SNPs. Here, total selection reflects the combined effects of direct and indirect selection as captured by differences in marginal fitness (expected fitness averaged across genetic backgrounds) of genotypes at each SNP locus. Because we assume most SNPs do not causally effect fitness in the short-term experiment (but see [9]), we equate total selection with indirect selection for simplicity. We used a Bayesian approach to estimate total selection on each SNP while accounting for genetic drift. Information for this inference comes from mortality in the experiment and patterns of allele frequency change. We describe the model and the procedure for fitting the model in the Methods section of the Supplementary Material.

**Quantifying LD between SNPs and expected fitness**

We asked whether the strength of the association (i.e., LD) between SNPs and genomic predictions of expected fitness could explain how strongly each SNP was impacted by indirect selection in the experiment. As noted in the main text, this involved three steps: (1) estimate expected fitness of each individual stick insect from our polygenic genomic prediction model, (2) quantify how strongly each of the \(\sim 7\) million SNPs was statistically associated with expected fitness at the organism level, (3) test whether the strength of the SNP-expected fitness relationship predicts independent estimates of total selection on each SNP.
We generated genomic predictions of expected fitness using an approximate Bayesian computation (ABC) framework based on survival and genotype data from the sequenced SNPs for the 246 individuals transplanted to the five *Adenostoma* bushes (see Supplementary Analyses and Results in the Supplementary Material for our complementary analysis of the *Ceanothus* treatment). The approach is similar to polygenic genome-wide association methods commonly used for genomic prediction of phenotypes [64, 48]. Our specific method is tailored to our data and thus has the added benefits of allowing for variation in absolute fitness among transplant sites, directly modeling the experimental design, and treating survival explicitly as a binary phenotype. We describe the model and the procedure for fitting the model in the Methods section of the Supplementary Material.

We then fit linear regression models with expected fitness as the response variable and genotype at each SNP (one SNP at a time) as the independent variable. Lastly, we computed the Pearson correlation between the $r^2$ association of each SNP with expected fitness and the evidence that each SNP experienced selection. See the Methods section of the Supplementary Material for additional details.

**Cross-validation to test performance of the genomic prediction model**

We used cross-validation to quantify the predictive power of the genomic prediction ABC model given our data set. To do this, we ran an additional 400 million simulations with the computer program described in the preceding section. However, in each simulation 80% of individuals were assigned to a training set, with the other 20% assigned to a test set. The training set was used to fit the model, and the test set allowed us to test the model on data not used for model fitting. Assignments to test versus training sets were random, and differed across simulations. See the Methods section of the Supplementary Material for additional details.

**Measuring LD and selection in windows**

To capture local (rather than long range) patterns of LD and selection, we calculated the average LD and total (mostly indirect) selection in contiguous regions (windows) of the genome. For LD, we calculated the mean LD (measured by the squared genotypic correlation, $r^2$) for all pairs of SNPs in 10 kilobase pair (kbp) windows along each genome scaffold (i.e., windows never spanned multiple scaffolds). We then fit discrete state, homogeneous hidden Markov models (HMMs) to delineate large-scale, contiguous regions of the genome with elevated LD or total selection (as in [65]). We fit the HMM for total selection similarly. See the Methods section of the Supplementary Material for additional details.
Analyses of randomized data sets

We repeated the main analyses described above with 10 randomized data sets. This was done to simulate and analyze cases of pure drift (random mortality) to compare to our observed data, and thereby to determine which of the documented patterns would be expected in the absence of selection. For each data set, we permuted survival among the 246 T. cristinae released on the five Adenostoma bushes. Permutations were only done within bushes (blocks) to retain bush-specific mortality rates. We then conducted analyses of selection during the experiment for each of the 10 random mortality data sets exactly as described for the observed data. Thus, these analyses used the same genotypic data, but simply differed in the specific individuals that lived and died.

References

[1] Colosimo, P. F. et al. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307, 1928–1933 (2005).

[2] Linnen, C. R. et al. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339, 1312–1316 (2013).

[3] Nadeau, N. J. et al. The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534, 106–110 (2016).

[4] Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. Genetics Research 23, 23–35 (1974).

[5] Barton, N. H. Multilocus clines. Evolution 454–471 (1983).

[6] Barton, N. H. Genetic hitchhiking. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 355, 1553–1562 (2000).

[7] Flaxman, S. M., Feder, J. L. & Nosil, P. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. Evolution 67, 2577–2591 (2013).

[8] 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).

[9] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177–1186 (2017).

[10] Carneiro, M. et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345, 1074–1079 (2014).
[11] Bosse, M. et al. Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science* 358, 365–368 (2017).

[12] Exposito-Alonso, M., Burbano, H. A., Bossdorf, O., Nielsen, R. & Weigel, D. Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature* 573, 126–129 (2019).

[13] Dowle, E. J. et al. Genome-wide variation and transcriptional changes in diverse developmental processes underlie the rapid evolution of seasonal adaptation. *Proceedings of the National Academy of Sciences* 117, 23960–23969 (2020).

[14] Kautt, A. F. et al. Contrasting signatures of genomic divergence during sympatric speciation. *Nature* 1–6 (2020).

[15] Cai, J. J., Macpherson, J. M., Sella, G. & Petrov, D. A. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genetics* 5, e1000336 (2009).

[16] Sella, G., Petrov, D. A., Przeworski, M. & Andolfatto, P. Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics* 5, e1000495 (2009).

[17] Elyashiv, E. et al. A genomic map of the effects of linked selection in *Drosophila*. *PLoS Genetics* 12, e1006130 (2016).

[18] Begun, D. J. & Aquadro, C. F. Levels of naturally occurring dna polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356, 519–520 (1992).

[19] Roselius, K., Stephan, W. & Stadler, T. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* 171, 753–763 (2005).

[20] McGaugh, S., Smukowski, C., Manzano-Winkler, B., Himmel, T. & Noor, M. Recombination modulates how selection affects linked sites in *Drosophila*. *PLoS Biology* 1, e1001422 (2012).

[21] Jensen, J. D., Thornton, K. R. & Andolfatto, P. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics* 4, e1000198 (2008).

[22] Campos, J. L., Zhao, L. & Charlesworth, B. Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. *Proceedings of the National Academy of Sciences* 114, E4762–E4771 (2017).

[23] Bergland, A. O., Behrman, E. L., O’Brien, K. R., Schmidt, P. S. & Petrov, D. A. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genetics* 10, e1004775 (2014).
Buffalo, V. & Coop, G. Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences* **117**, 20672–20680 (2020).

Grant, P. R. & Grant, B. R. Unpredictable evolution in a 30-year study of Darwin’s finches. *Science* **296**, 707–711 (2002).

Grant, P. R. & Grant, B. R. Evolution of character displacement in Darwin’s finches. *science* **313**, 224–226 (2006).

Barrett, R. D., Rogers, S. M. & Schluter, D. Natural selection on a major armor gene in threespine stickleback. *Science* **322**, 255–257 (2008).

Barrett, R. D. *et al.* Linking a mutation to survival in wild mice. *Science* **363**, 499–504 (2019).

Thompson, J. N. *Relentless Evolution* (University of Chicago Press, 2013).

Hendry, A. P. *Eco-evolutionary dynamics* (Princeton university press, 2020).

Schrider, D. R., Shanku, A. G. & Kern, A. D. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223 (2016).

Pouyet, F., Aeschbacher, S., Thiéry, A. & Excoffier, L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife* **7**, e36317 (2018).

Buerkle, C., Gompert, Z. & Parchman, T. The n= 1 constraint in population genomics. *Molecular Ecology* **20**, 1575–1581 (2011).

Barton, N. Estimating linkage disequilibria. *Heredity* **106**, 205 (2011).

Aranzana, M. J. *et al.* Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics* **1**, e60 (2005).

Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791 (2006).

Schluter, D. *et al.* Fitness maps to a large-effect locus in introduced stickleback populations. *Proceedings of the National Academy of Sciences* **118** (2021).

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

Nosil, P. *et al.* Natural selection and the predictability of evolution in *Timema stick* insects. *Science* **359**, 765–770 (2018).
[40] Lindtke, D. et al. Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Molecular Ecology* 26, 6189–6205 (2017).

[41] Nosil, P. Divergent host plant adaptation and reproductive isolation between ecotypes of *Timema cristinae* walking sticks. *The American Naturalist* 169, 151–162 (2007).

[42] Riesch, R. et al. Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology & Evolution* 1, 1–13 (2017).

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

[44] Sepil, I., Lachish, S. & Sheldon, B. C. MHC-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular Ecology* 22, 384–396 (2013).

[45] Grant, P. R. & Grant, B. R. 40 years of evolution: Darwin’s finches on Daphne Major Island (Princeton University Press, 2014).

[46] Farkas, T. E., Mononen, T., Comeault, A. A., Hanski, I. & Nosil, P. Evolution of camouflage drives rapid ecological change in an insect community. *Current Biology* 23, 1835–1843 (2013).

[47] O’Brown, N. M., Summers, B. R., Jones, F. C., Brady, S. D. & Kingsley, D. M. A recurrent regulatory change underlying altered expression and Wnt response of the stickleback armor plates gene EDA. *eLife* 4, e05290 (2015).

[48] Gompert, Z., Egan, S. P., Barrett, R. D., Feder, J. L. & Nosil, P. Multilocus approaches for the measurement of selection on correlated genetic loci. *Molecular Ecology* 26, 365–382 (2017).

[49] Hahn, M. W. Toward a selection theory of molecular evolution. *Evolution* 62, 255–265 (2008).

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

[51] Wallace, B. Hard and soft selection revisited. *Evolution* 465–473 (1975).

[52] Wade, M. J. Soft selection, hard selection, kin selection, and group selection. *The American Naturalist* 125, 61–73 (1985).

[53] Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10, 195–205 (2009).

[54] Nosil, P., Flaxman, S. M., Feder, J. L. & Gompert, Z. Increasing our ability to predict contemporary evolution. *Nature Communications* 11, 1–6 (2020).
[55] Gillespie, J. H. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155**, 909–919 (2000).

[56] Burri, R. *et al.* Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research* **25**, 1656–1665 (2015).

[57] Chaturvedi, S. *et al.* Recent hybrids recapitulate ancient hybrid outcomes. *Nature Communications* **11**, 1–15 (2020).

[58] Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[59] McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

[60] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

[61] Nosil, P. *et al.* Ecology shapes epistasis in a genotype–phenotype–fitness map for stick insect colour. *Nature Ecology & Evolution* 1–12 (2020).

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

[63] Sved, J. Correlation measures for linkage disequilibrium within and between populations. *Genetics Research* **91**, 183–192 (2009).

[64] Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics* **9**, e1003264 (2013).

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

**Acknowledgements**

We thank A. Buerkle, C. Nice, T. Parchman, R. Pineau, T. Reimchen and the Gompert and Nosil lab groups for discussion and comments on previous versions of the manuscript. R. Barrett and S. Laurent graciously provided access to genomic data for *Peromyscus*. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. The work was funded by a grant from the National Science Foundation of the United States (grant no. NSF DEB 1844941) and a grant from the European Research Council (grant no. EE-Dynamics 770826, https://erc.europa.eu/).
Author contributions

G, JF and PN conceived of the project. PN collectd the data. ZG analyzed the data. ZG, JF and PN wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

DNA sequence data have been archived on NCBI's SRA (PRJNAXXXX). Phenotypic data are available on Dryad (DOIXXX).

Code availability

Computer code and primary input files are available on github (https://github.com/zgompert/TimemaPolygenicSelection).

Figures
Figure 1: Schematic illustration of direct versus indirect selection. Panel (A) shows how sampling error (i.e., genetic drift) can generate long-range LD. Here, random sampling from a gene pool results in some multilocus genotype combinations being over-represented (i.e., statistically associated) in a population. Boxes denote the locations of three genetic loci, with filled-in boxes indicating the allele favored by selection (red) or in LD with the favored, causal variant (orange). Panel (B) illustrates how direct selection and LD cause indirect selection. A population is shown before and after selection (and, for simplicity, before any reproduction). The ‘X’s thus denote individuals that died. Selective mortality results in fixation of the favored allele, and also an increase in the frequency of the associated, functionally-neutral alleles. In this example, the effects of association include one allele on a different chromosome than the causal variant.
Figure 2: *Timema cristinae* morphs and empirical evidence of selection in the transplant experiment. Panel (A) depicts green unstriped, green striped and melanistic morphs, which are determined by the *Mel-Stripe* locus. Panel (B) shows linkage disequilibrium (LD) (measured as $r^2$) between genome-wide SNPs and the *Mel-Stripe* locus. Only the 1% of SNPs exhibiting the greatest LD are shown. Colors denote whether SNPs fall within (red) or outside of (dark gray) the *Mel-Stripe* locus.
Figure 3: Plots show linkage disequilibrium (LD) (measured as $r^2$) in deer mice (A,B) and threespine stickleback (C,D). Panels (A) and (B) show LD between genome-wide SNPs and the $\Delta$Ser mutation in Agouti in deer mice (Peromyscus maniculatus). Points denote each of $\sim$50 thousand SNPs; SNPs most strongly associated with $\Delta$Ser are shown in green (top 1%), other SNPs are shown in gray. The location of the $\Delta$Ser mutation is denoted by a vertical red line. Panels (A) and (B) show results for the light-soil ($N = 217$) and dark-soil enclosures ($N = 229$), respectively. Drawings by Rosa Ribas depict dark and light deer mice phenotypes. Panels (C) and (D) show LD between 387 SNPs and a QTL marker near the eda locus in threespine stickleback (Gasterosteus aculeatus). SNPs most strongly associated with the QTL markers are shown with green points (top 10%), and other SNPs are denoted by gray points. The location of the QTL marker is denoted by a vertical red line. Panel (C) is based on all SNPs, whereas panel (D) excludes SNPs on chromosome 4 (the vertical axis scale differs between these panels). Drawings by Rosa Ribas depict marine and freshwater stickleback fish.
Indirect selection is caused by statistical associations between SNPs and expected fitness caused by LD with causal variants.

Organismal level
- Expected fitness (organism)
- Linkage disequilibrium

Genetic level
- Causal variants
- Marginal fitness (locus)

Indirect selection

(A) LD drives indirect selection

(B) Model cross validation

(C) LD predicts selection

Figure 4: Evidence of widespread indirect selection. (A) Conceptual overview and illustration of our method for connecting LD with indirect/total selection. The schematic shows that causal variants determine each organism’s expected fitness (e.g., survival probabilities), and statistical association between SNP genotypes and organismal expected fitness determine indirect selection based on differences in the marginal fitness values of (potentially neutral) SNP genotypes. Example plots on the right illustrate our method. First, we estimate expected fitness from our polygenic, genomic prediction model. We then estimate the association between each SNP genotype on expected fitness using linear regression (bottom plot). The $r^2$ values for each SNP are then used as the independent variables in the next analysis where we ask whether they explain variation in the total selection experienced by each SNP. Panel (B) gives the distribution of error rates for genomic prediction of individual survival over a series of null models (gray region), which is compared to the error rate for the fitted model (orange vertical line). The fitted model outperformed the null models ($P < 0.001$). Panel (C) shows that SNPs more strongly associated with expected fitness have a higher posterior probability of having experienced (indirect) selection. The heatmap depicts the relationship across the $\sim7$ million SNPs with individual black points shown in the regions of lowest density. A best fit line is shown (Pearson correlation = 0.45, $r^2 = 0.2$, $P < 0.001$). The dashed line corresponds with an average posterior probability of 0.5 for selection affecting SNPs in a window.
Figure 5: Local LD predicts selection and evolution in natural populations. Panel (A) depicts estimates of selection (total selection, which we equate with indirect selection) across the genome. Points denote 50 SNP window averages of the log posterior odds that each SNP experienced selection. The dashed line corresponds to an odds ratio of 1:1. Windows of high indirect selection based on a hidden Markov model (HMM) are shown in red. Panel (B) gives mean pairwise LD (measured by $r^2$) in 10 kb windows along the genome. Windows of high LD based on HMM are shown in red.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IndirectSelectionSupplement.pdf