ECOLOGY

Predicting patch occupancy reveals the complexity of host range expansion

M. L. Forister[1,2]*, C. S. Philbin[2,3], Z. H. Marion[4], C. A. Buerkle[5], C. D. Dodson[2,3], J. A. Fordyce[6], G. W. Forister[7], S. L. Lebeis[8], L. K. Lucas[9], C. C. Nice[10], Z. Gompert[9]

Specialized plant-insect interactions are a defining feature of life on earth, yet we are only beginning to understand the factors that set limits on host ranges in herbivorous insects. To better understand the recent adoption of alfalfa as a host plant by the Melissa blue butterfly, we quantified arthropod assemblages and plant metabolites across a wide geographic region while controlling for climate and dispersal inferred from population genomic variation. The presence of the butterfly is successfully predicted by direct and indirect effects of plant traits and interactions with other species. Results are consistent with the predictions of a theoretical model of parasite host range in which specialization is an epiphenomenon of the many barriers to be overcome rather than a consequence of trade-offs in developmental physiology.

INTRODUCTION

Emerging infectious diseases and crop pests are examples of host range expansion in which an organism with a parasitic life style colonizes and successfully uses a novel host (*1*). Many aspects of host range are poorly understood, including why most herbivorous insects and other parasites are specialized and the conditions under which new host-parasite interactions develop and persist. Mechanistic approaches in focal systems have revealed key aspects of host recognition (*2*) and other relevant biological processes (*3*) but, by design, do not encompass context dependence including interacting species and abiotic variation. Ecological studies of host range, in contrast, might quantify context dependence but have not included both modern genomic and metabolomic approaches (*4*). Here, we use the colonization of alfalfa, *Medicago sativa*, by the Melissa blue butterfly, *Lycaeides melissa* (Fig. 1), to present what is, to our knowledge, the most thorough picture of a recent (within the last 200 years) host range expansion in terms of number of populations studied and breadth of interacting species and host traits characterized.

Theoretical work in this area can be divided into two partially overlapping groups: those that emphasize developmental performance (including trade-offs in the ability to use different hosts) and those that stress opportunity and constraint imposed by exogenous factors, primarily natural enemies (*5*) and geography (*6*). Although developmental trade-offs in host use are rare (*7*), it is clear that plant defenses are a barrier to insect colonization, as performance is often reduced for herbivores in experiments with novel versus ancestral hosts (*8*). What we do not know is whether the magnitude of performance effects studied in the laboratory will be informative under field conditions. Predation pressure could, for example, remove all opportunity for successful development on a novel host that would otherwise be suitable. Equally unknown is whether variation within and among host populations might have compensatory effects, such that a direct negative effect of a particular toxin on an herbivore is balanced by similar effects on a competitor.

The Melissa blue is widespread in western North America, where it can be found in association with native legume (Fabaceae) host plants, and typically persists in isolated subpopulations connected by limited gene flow (*9*). The association with alfalfa is heterogenous, most often occurring in areas where the plant has escaped cultivation,

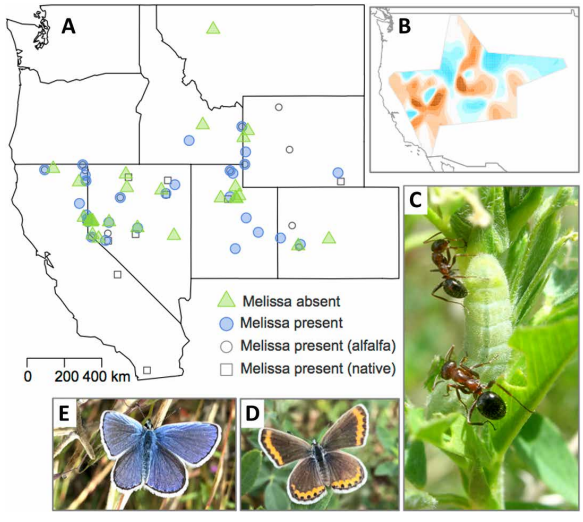

Fig. 1. Map of study locations, dispersal surface, and images of butterflies, ants and caterpillar. (**A**) Solid symbols (circles and triangles) are focal alfalfa locations from which arthropods and plants were collected: Blue sites are locations where the Melissa blue butterfly (*L. melissa*) has colonized the novel host; green sites are alfalfa locations not colonized by the butterfly. Open symbols (circles and squares) are locations used in the quantification of gene flow; in some cases (where an open circle appears within a blue circle), sites were represented in both datasets. (**B**) Effective migration surface used to generate covariates representing rates of effective dispersal (blue is faster than average; red is slower). (**C**) *L. melissa* caterpillar being tended by mutualist ants on alfalfa (photo credit: Chris Nice, Texas State University). (**D**) Female and (**E**) male Melissa blue butterflies (photo credit: Matthew Forister, University of Nevada).

[1]Department of Biology, University of Nevada, Reno, NV 89557, USA. [2]Hitchcock Center for Chemical Ecology, University of Nevada, Reno, NV 89557, USA. [3]Department of Chemistry, University of Nevada, Reno, NV 89557, USA. [4]Bio-protection Research Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand. [5]Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY 82071, USA. [6]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA. [7]Bohart Museum of Entomology, University of California, Davis, Davis, CA 95616, USA. [8]Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA. [9]Department of Biology, Utah State University, Logan, UT 84322, USA. [10]Population and Conservation Biology, Department of Biology, Texas State University, San Marcos, TX 78666, USA.
*Corresponding author. Email: forister@gmail.com

and is the result of at least two independent colonization events by the Melissa blue (*10*). Alfalfa was introduced to western North America in the mid-1800s (*9*) and is a poor food plant for Melissa blue caterpillars, which develop into adults that are, on average, half the size of individuals experimentally reared on a native host (*11*), with direct and indirect fitness consequences (*12*, *13*). The use of alfalfa does not appear to be constrained by genetic, developmental tradeoffs in the Melissa blue or a lack of genetic variation in ability to use that host (*10*, *14*). Nevertheless, unoccupied patches of alfalfa have remained unoccupied by the butterfly for years or even decades, even in close proximity to occupied patches (*15*). We took advantage of that landscape heterogeneity, as pictured in Fig. 1A, to quantify and model the factors controlling patch occupancy of the novel host by the Melissa blue butterfly.

RESULTS

To understand host plant and arthropod community variation associated with Melissa blue patch occupancy, samples and data were taken from more than 1600 individual plants from 56 alfalfa locations with and without the Melissa blue (Fig. 1A). Arthropod collections included 20,890 individuals that were sorted into 298 species (these were morphospecies in the vast majority of cases, identified to taxonomic family and assigned a unique morphospecies number based on phenotype) from 123 taxonomic families and 16 orders (spiders were not identified beyond order). Specimens were further parsed into functional groups, with proportional representation as follows: 56% of individuals were ant-tended herbivores (aphids, treehoppers, and relatives), 21% were predators, 8% were other herbivores (not ant-tended), 7% were ants, 2% were parasitoids (with the potential for attacking caterpillars), and the rest (6%) were parasitoids of other groups, flower visitors, or incidentals without direct ecological relevance for our focal species. Assignment to these functional categories was based largely on our knowledge of the natural history of the system. All aphids and treehoppers are not, of course, tended by ants but the majority of individual aphids and treehoppers that we see at our study sites do appear to engage in this facultative mutualism.

Because movement across the landscape is an essential component of novel host colonization and use, we assembled a population genomic dataset to estimate effective migration surfaces for the focal butterfly from 541 individuals for rare and common genetic variants, assuming 200 and 400 demes (surfaces are shown for both classes of variants in the 400 deme model in fig. S1). The migration surface models were more successful than simple isolation by distance at predicting genetic dissimilarity among spatial units, as can be seen in the comparison of fig. S1C versus fig. S1D for isolation by distance versus fitted values based on rare variants in the 400 deme model. Rates of effective migration from all four models (rare and common, 200 and 400 demes) were moved forward into a Bayesian ridge regression predicting Melissa blue presence and absence: the model of rare variants at the finer scale of 400 demes was the most influential in the ridge regression (more details below), which is why it is highlighted in fig. S1.

As a characterization of phytochemical variation among individual alfalfa plants, our analyses produced 849 metabolomic features, which were simplified through factor analysis. Similarly, factor analysis was used as a data reduction step for site-specific climate data. We retained two factors from the climate data and six from the metabolomic data (fig. S2) as optimal in the sense that we maximized the amount of variation captured with the constraint of having few enough factors to be both interpretable and tractable in downstream analyses. In the climate model, the two factors explained 72% of the variation among sites, with factor 1 describing a gradient of increasing temperatures and drier conditions, particularly maximum daily temperatures, while factor 2 is a gradient of increasing minimum temperatures (fig. S3). Analysis of the metabolomic data explained 33% of the variation with six factors, and we used relative mass defect for annotation of major compound classes (fig. S4). Given the large number of compounds involved, we focused on annotation of the higher-loading compounds (see Materials and Methods).

In preparation for structural equation modeling (SEM), we used Bayesian ridge regression to identify the most influential predictors for five response variables: presence and absence of the focal butterfly (our chief variable of interest), as well as abundance of ants, predators, ant-tended herbivores, and other (nontended) herbivores. Slightly different sets of biologically relevant predictors were examined for each response variable. All ridge regression models readily converged, and effective sample sizes tended to be in the thousands. Using 75% as the cutoff for confidence in our regression coefficients, we found between 4 and 10 variables that rose to the top as candidates for inclusion in structural equation models (table S2). Almost without exception, effects estimated with ridge regressions agreed with a priori expectations and previous results in this system. For example, ants had the strongest effect on butterfly presence and absence (*16*) and one of the strongest effects on the abundance of tended herbivores. Specific leaf area was highly ranked for the Melissa blue (table S2) with a negative effect as previously observed in an experimental context (*17*), and specific leaf area had a negative effect on other (largely chewing) herbivores (table S2) but not on other ant-tended (sucking) herbivores, which likely interact differently with physical leaf traits.

Spatial autocorrelation was investigated using Moran's *I* and comparisons against null simulations. Overall, we found that spatial autocorrelation was low: Only a few variables were significantly more clustered on the landscape than would be expected by chance, with (expectedly) dispersal having the strongest spatial autocorrelation (table S3). In addition to the tests with Moran's *I*, we generated MEMs (Moran's eigenvector maps) as covariates for spatial structure and included them in a Bayesian ridge regression for our variable of primary interest, the presence and absence of the Melissa blue. Two MEMs fell within the top variables following ridge regression (see last section of table S2), using the criterion of 75% confidence. However, the vast majority of other variables did not change in their importance while accounting for spatial autocorrelation. Thus, we concluded that spatial autocorrelation is present in the system but do not address it further because it does not alter our goal of understanding direct and indirect effects on the presence and absence of the butterfly.

Using the important variables from ridge regressions (table S2), we initially constructed a structural equation model that fits the data but had three unresolved paths (see the base model in table S4). For example, ants are a top variable (in ridge regressions) affecting tended herbivores, and tended herbivores are a top variable for ants, hence the unresolved (or "double-headed arrow") path in the base model. Subsequent model comparisons (detailed in table S4) resolved two of those paths, with other herbivores affecting ants and ants pointing to tended herbivores. The latter agrees with our observations: As

ecological generalists, ants can be present without tended herbivores, but tended herbivores are less likely to be successful without ants. The final model is shown in Fig. 2 (with both R^2 and leave-one-out correlations for individual endogenous variables), along with effects summarized by functional categories so that, for example, the direct and indirect importance of plant structure can be compared to the direct and indirect importance of phytochemistry (Fig. 2B). Full results from the structural equation model including path coefficients and associated *P* values are given in table S5. In our null model simulations, the variation explained by the real model was roughly threefold greater than the average variation explained in simulated datasets (fig. S5). Effects of a few of the most important variables are shown in Fig. 3, both individually (e.g., the influence of ants on Melissa blue presence and absence in Fig. 3A) and in conjunction with other factors (e.g., ants and phytochemical factor 4 in Fig. 3E). The importance of phytochemical variation is also visualized for individual metabolomic features and their direct (Fig. 3F) and indirect (Fig. 3G) associations with butterfly occupancy.

DISCUSSION

Like most butterflies in the family Lycaenidae, Melissa blue caterpillars engage in a facultative mutualism with ants (Fig. 1C), where caterpillars produce specialized secretions in exchange for protection from natural enemies (*18*). Previous experimental work in this system found that excluding ants from individual plants reduced caterpillar survival (*16*). We find here that ant abundance is the most influential variable or control on Melissa blue presence across the 56 sites (Figs. 2 and 3A). This is true even when considering the fact that ants facilitate many hemipterans (aphids, treehoppers, and other myrmecophiles), which, in turn, have a negative competitive effect on the Melissa blue (Fig. 3B). The balance of ant and hemipteran effects is such that the negative effect of the latter is most influential at intermediate ant densities (Fig. 3D). Similar complexity arises through direct and indirect effects of metabolomic variation. Phytochemical factor 4 has a direct negative association with Melissa blue presence (Fig. 3C) but an indirect positive effect mediated through other herbivores and their effect on ants (Fig. 2). That axis of plant variation is positively associated with a number of alkaloids, among other compounds, with potential herbivore toxicity (see fig. S4). In general, we find that roughly three-quarters of the variation in butterfly presence and absence at the landscape scale can be explained with the suite of variables that includes ants and plant metabolomic variation, as well as host patch area, natural enemies, and dispersal (relative rates of effective migration; Fig. 1B). The success of the model is apparent in cross-validation (Fig. 2) and null simulations of site-level properties (fig. S5).

Considering the summed totals of direct and indirect effects estimated through path analysis (Fig. 2B), we find that metabolomic variation is associated with the most pronounced, direct negative

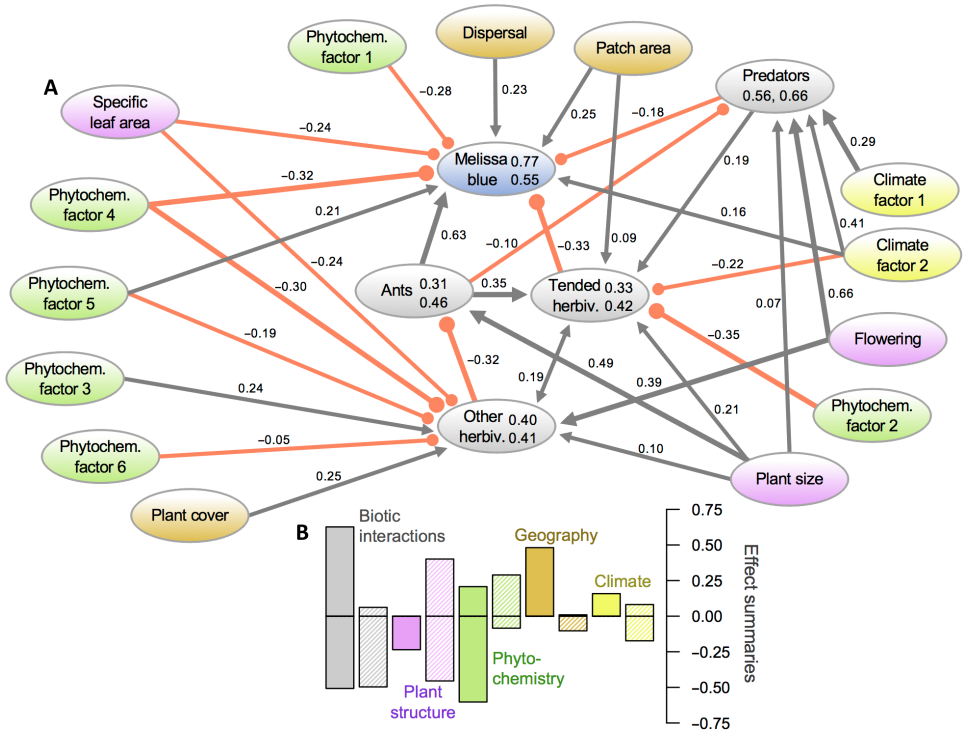

Fig. 2. Structural equation model and summary of direct and indirect effects. (**A**) Path diagram illustrates coefficients estimated in structural equation model predicting Melissa blue presence and absence across the landscape, as well as abundance of ants, tended herbivores, other herbivores and predators (model fit: Fisher's $C = 67.66$, $P = 0.995$). Negative effects are indicated by red lines, and positive effects are indicated by gray lines; width of lines is scaled to the magnitude of the coefficients. For the endogenous variables, two numbers are shown within ovals: R^2 values (on top) and observed-versus-predicted correlations (below) from leave-one-out cross-validation. Color coding of exogenous variables indicates plant metabolomics data (green), plant structural traits (violet), geographic variables (brown), and climate (yellow); color coding also corresponds to bar chart (**B**), which summarizes relative magnitude of direct and indirect effects (solid and hashed bars, respectively), both positive and negative. For example, climate has a modest positive direct effect, a smaller positive indirect effect (mediated through tended herbivores), and a larger negative indirect effect (through predators). Additional details from the structural equation model are in table S5.

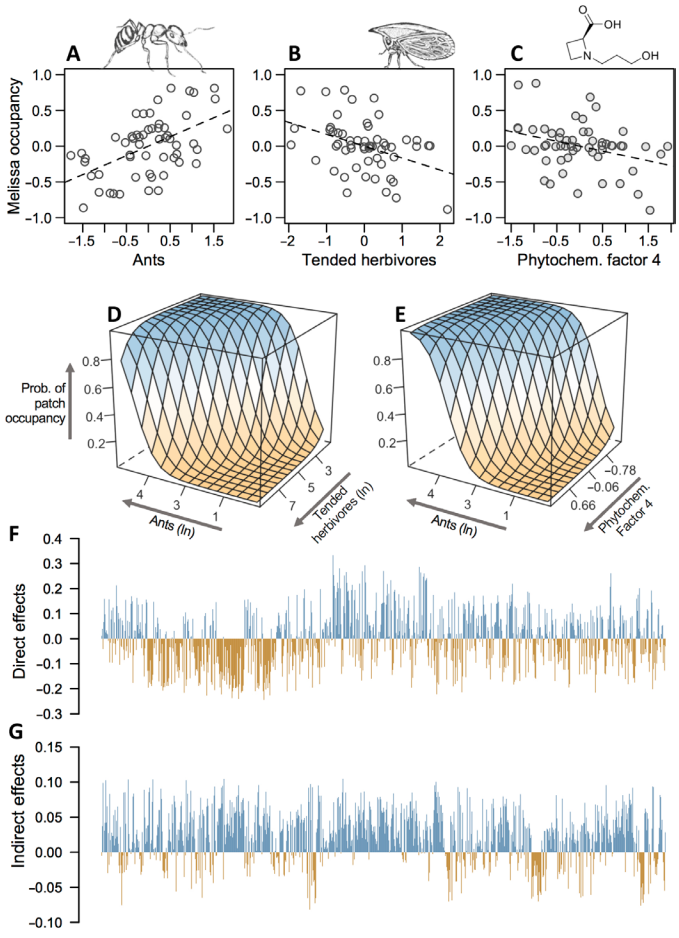

Fig. 3. Illustration of effects for a subset of variables predicting *L. melissa* presence and absence across the landscape. (**A** to **C**) Partial effects of ants, tended herbivores, and phytochemical factor 4 on Melissa blue occupancy; in other words, these are the effects of those individual factors while controlling for other factors predicting occupancy (see paths in Fig. 2). (**D** and **E**) Predicted probability of patch occupancy across a range of values for ant and tended herbivore abundance (D) and for ant abundance and phytochemical factor 4 (E) where it can be seen, for example, that at high values of phytochemical factor 4, a higher abundance of ants is needed before the probability of occupancy rises. (**F** and **G**) Direct and indirect effects of 849 metabolomic features, both positive (blue) and negative (orange); see text for more details on calculation of individual effects. *Formica* ant and *Campylenchia* treehopper (one of the more abundant ant-tended herbivores at our study sites) illustrations are by M.L.F.; the alkaloid shown above (C) is medicanine (see fig. S4 for additional examples).

effects, followed closely (among negative effects) by direct and indirect interactions with other arthropods and then indirect effects of plant structure. A negative effect of phytochemical variation on patch occupancy is consistent with the idea that plant defenses are, at least, a partial barrier to colonization but does not imply that trade-offs in host use (in the sense of host range being constrained by antagonistic pleiotropy among host-associated alleles) are present in this system. Previous experimental work and surveys of genetic variation in the field have suggested that antagonistic pleiotropy is not an important constraint on expansion of diet breadth for the Melissa blue (*10*, *14*). The effect of specific leaf area is also consistent with a previous experimental study (*17*) but is small compared to both positive and negative indirect effects associated with plant size and the density of flowers mediated through enemies and competitors (Fig. 2B). In terms of positive effects on the Melissa blue, the importance of ants is followed by geographic factors including patch area and dispersal (effective migration rates). These results demonstrate the value of studying plant variation in the context of geography and interacting species. While individual components of the results reported here are consistent with experimental work, other aspects are less accessible to manipulation. Individual metabolites, for example, have a mix of positive and negative direct effects on the Melissa blue (Fig. 3F), as observed in a previous rearing experiment (*17*), while the indirect effects of individual compounds are characterized more by positive effects mediated through numerous other species in the wild (Fig. 3G).

CONCLUSION

The theory of ecological fitting suggests that novel hosts are colonized if they are "close enough" to native hosts in key traits (*19*–*22*), but we have few cases in which that close enough distance has been quantified as we have done in this system. We find diverse factors or controls on colonization that are encountered in multifarious combinations (*23*). When all factors align, butterfly populations persist on the novel host, but the diversity of challenges (plants, enemies, and abiotic conditions) undoubtedly makes adaptation to the novel host difficult, especially when some or all of those factors likely shift in character from year to year. This possibility is consistent with only minimal local adaptation that has been observed in alfalfa-associated populations (*14*). It would, of course, be informative to have a similarly detailed picture of the factors associated with patch occupancy of native hosts with which the focal butterfly has a long evolutionary history, but this will have to await future studies. Given the results presented here, we can see the theory of host range evolution approaching maturity: genetic trade-offs are possible but rare (*24*, *25*); instead, it is likely that a balance of factors (both positive and negative) associated with novel host use exist in any system but are only infrequently encountered in combinations that allow host range expansion (*26*, *27*). Further complexity is added by the fact that changing abiotic conditions can affect the suitability of novel hosts (*28*). Thus, generalist herbivores or parasites (with many accumulated hosts) are predictably rare across geographic and phylogenetic scales (*29*, *30*). The complexity of barriers to novel host use and the ecological contingency of colonization challenge our ability to forecast new crop pests or emerging infectious diseases (*1*), but the multidisciplinary approach illustrated here does raise the promise, at least for herbivorous insects, that expansions of host range can be understood given current technologies and sufficient sampling effort.

MATERIALS AND METHODS

Site identification, plant collections, and associated data

Our goal was to sample a roughly equal number of alfalfa (*M. sativa*) locations that had and had not been colonized by the Melissa blue butterfly (*L. melissa*). Throughout the arid western United States, the plant is favored by the butterfly in places where it has escaped cultivation and exists in mostly discrete patches along roadsides and in invaded or partly degraded natural communities. Before the field sampling associated with this project in 2017 and 2018, we had accumulated a database of observations of alfalfa locations that support Melissa blue populations, as described in previous publications

(*9*, *31*, *32*), as well as (to a lesser extent) a database of alfalfa locations not associated with Melissa presence (*15*). Both presence and absence of the butterfly are relatively stable over time: We have observed presence to be unchanged at locations that we have visited annually for more than 20 years, and focal absences have been tracked for up to 12 years (*15*).

Throughout the summers of 2017 and 2018, we visited a total of 56 sites (Fig. 1A and table S1), mixing presence and absence sites haphazardly in space so as not to confound latitude or longitude with order (date) of sampling or with site status (presence and absence of the butterfly). In other words, sites were visited so that presence and absence sites were interdigitated both in space and time (note that we also include year as a categorical variable in analyses, as described below). In many cases, sites had been previously identified (as mentioned above), while in other cases, new sites were discovered during the summers of 2017 and 2018.

Because adults have high site fidelity (*33*) and are easily observed, the status of a location (butterfly presence or absence) can, in most cases, be determined in a single visit. However, the Melissa blue has multiple generations in a season, which makes it possible that a visit could coincide with low density between generations and thus produce a false-negative observation (which is more likely early in the summer, as the first and second generations are more discrete, while the second and third tend to overlap). Thus, all absence locations were revisited at least once or twice, approximately 2 to 6 weeks after the initial sampling date to confirm that a site does not support a Melissa blue population (the only exception to that was our most northern absence site, in Montana, which was not revisited for logistical reasons). Note that we did not attempt to quantify butterfly abundance, which (unlike presence or absence) would have required multiple visits to reliably estimate.

Following our earlier studies with arthropod communities on alfalfa (*34*, *35*), we used flowering as a guide to phenology and considered a site appropriate for sampling if at least half of the individual plants were flowering. Some of our previous work with alfalfa arthropods and plant traits has involved repeated sampling at individual locations (*35*), although arthropod communities in alfalfa were found to be relatively stable throughout the summer (*34*); thus, the present project used single visits to individual locations as a way to maximize effort spent sampling additional sites (Fig. 1A). Most sampling days were in July and August, with fewer in June and September, and an index of flowering was included as a covariate in models (see below).

Sampling at a site began with a marker placed in the center of the patch of alfalfa, and 30 individual alfalfa plants were then flagged at random compass directions and distances (up to 25 m) from the center of the patch (at two sites where a patch had fewer than 30 plants, they were all sampled). In some cases, when patches were essentially linear (for example, along a roadside), compass directions were converted to binary orientations forward or backward from the center. The total areal extent (length and width) of the patch was measured and percent cover of alfalfa was visually estimated. Each of the randomly selected focal plants was measured for size (as a box measurement of length, width, and height), and the number of flowering stems was counted. Three mature (but not senescent) leaves were selected haphazardly for leaf toughness measurements using a penetrometer (Chatillon 516 series) through the center of the middle leaflet (*17*).

For metabolomic work (and the measurement of leaf area and mass), three small clusters of leaves (three to five leaves in each cluster) were collected from the top, middle, and lower portions of the plant, thus encompassing as much whole plant phytochemical variation as possible. The three clusters were pooled into a single large coin envelope. Thus, the current study does not quantify intraindividual variation or attempt to separate induced from constitutive defenses (alfalfa is attacked by a wide range of vertebrate and invertebrate herbivores and the vast majority if not all of our sampled plants had been damaged to some extent before sampling). All of the envelopes from a single site were stored in an open paper bag for air drying before being delivered to a laboratory at the University of Nevada, Reno, where drying was completed in a vacuum. After vacuum drying, sample envelopes were kept in plastic bins with desiccating crystals until they were needed for either metabolomic work or area and mass measurements. For the latter (area and mass), five leaves were selected haphazardly from each envelope and weighed to the nearest tenth of a milligram on a microbalance and taped to a sheet of plain white paper that was then scanned. Leaf area (in cm^2) was taken using ImageJ software (version 1.52A) on the scanned images of individual leaves and used to calculate specific leaf area as area divided by mass (*17*). Because all of the analyses presented here focus on variation at the patch scale, plant measurements were averaged across plants within a location. Values for leaf toughness and plant volume, as well as patch area and cover, were log-transformed before analyses.

Climatic data for each of the 56 sites were generated as monthly averages for 2008 to 2018 using the get_prism:monthlys function from the prism package (*36*) for average daily minimum temperatures, average daily maximum temperatures, and precipitation totals (data from the PRISM Climate Group, Oregon State University). These analyses and all others described below (except where noted) were done using the R statistical computing language (*37*).

Arthropod collections and identification

Arthropods were collected using a sweep net from each of the randomly selected individual plants at each of the 56 field sites, as we have done previously (*16*, *34*, *35*). Unlike alfalfa in cultivation, alfalfa that has escaped along roadsides and into seminatural communities tends to grow as larger individuals with open spaces between plants, which facilitates sweep netting of individual plants. Each plant was swept four times in rapid succession, with the collector gradually moving around the plant during sweeping. The collected arthropods were then transferred from the net into a plastic vial with ethanol for storage (an aspirator was used to get most arthropods out of the net, with larger specimens transferred directly into alcohol). We attempted to collect all arthropods large enough to be seen with the naked eye, with the exception of thrips (Thysanoptera), which we have found to be both too small and too numerous on alfalfa for efficient collection. Vials were each labeled by location and plant.

Individual arthropods within samples (vials associated with individual plants) were counted and identified using a dissecting microscope (with ×90 maximum magnification) to the lowest possible taxonomic level, which was almost always family and, in some cases, genus and species, using standard taxonomic keys appropriate to different groups. If a specimen could only be identified to family, it was given a morphospecies number that identified unique morphospecies across all of our field sites. This work builds on a previous study of alfalfa insects in the Great Basin (*35*), and many of the morphospecies numbers used in the current study were established in that previous work. The only exception to that pipeline

involved spiders, which were only counted and not identified to any lower taxonomic level, such that total spider abundance was used in analyses (as part of the pool of predators at each site).

Following taxonomic identification, each species (or morphospecies) was given one of the following ecological assignments: ant-tended herbivores, other herbivores, parasitoids, predators, and ants. These assignments were partly based on our own observations of alfalfa-insect communities and on literature searches for specific taxa (for example, as a way to determine whether a particular chewing herbivore could have been feeding on alfalfa or might have more likely been there only as a flower visitor). Ants were treated as their own ecological group for the simple reason that they are considered separately in analyses as mutualists of our focal caterpillars and of other ant-tended herbivores. Parasitoids include wasps and flies with some history (based on literature searches) of potentially attacking caterpillars (we also identified a large number of hemipteran parasitoids, but those were not included in these analyses). For analyses reported here, arthropods were totaled within functional (ecological) groups at the plant patch level, and abundances were natural log–transformed before analyses.

Plant metabolomics

Individual plants across all sites in each collection year were randomized before extraction and analysis. After leaves were vacuum dried and finely ground (using a TissueLyser II, QIAGEN, Hilden, Germany), approximately 10 mg of dried tissue was extracted in 2.00 ml of aqueous ethanol (70%), vortexed briefly, and sonicated for 15 min. The resulting suspension was centrifuged at 500 rpm for 10 min. Aliquots of 1 ml from the supernatant were then passed through a 96-well filter (1 ml, 1-μm glass fiber; AcroPrep) into glass vials, covered with silicone mats and stored at −10°C. Chromatographic analyses were conducted on an Agilent 1200 analytical HPLC (high performance liquid chromatography) coupled to an Agilent 6230 time-of-flight mass spectrometer via an electrospray ionization source (gas temperature, 325°C; flow, 10 liter/m; nebulizer pressure, 35 psig; VCap, 3500 V; fragmentor, 165 V; skimmer, 65 V; and octopole, 750 V). A solution of digitoxin (0.50 μl at 0.200 mM methanol; Sigma-Aldrich), a commercially available cardenolide that has been used as an internal standard in other analyses of saponins (*38*), was coinjected with extracts (1.00 μl) and eluted at 0.500 ml/min through a Kinetex EVO C18 column (2.1 mm by 100 mm, 2.6 μ, 100 Å; Phenomenex) at 40°C. Buffers A (water containing 0.1% formic acid) and B (acetonitrile containing 0.1% formic) composed the linear binary gradient, changing over 30 min as follows: 0 to 1 min 5% B, ramp to 50% B at 4 min, ramp to 100% B at 21 min, 21 to 25 min 100% B ramping to 1.00 ml/min, before reequilibrating the column from 25 to 30 min at 5% B, 0.5 ml/min.

Raw data files were converted to mzML format using ProteoWizard msConvert 3.0 (*39*) before processing using the Bioconductor R package XCMS (*40*). Chromatographic features (retention time and m/z bins) were extracted from raw data files before retention time correction using the digitoxin internal standard, peak density grouping, and gap-filling. The Bioconductor R package CAMERA (*41*) was then used to identify groups of features (pseudospectra) with similar retention time that were highly correlated across chromatograms ($r > 0.8$), with similar ($r > 0.8$) peak shape and by characteristic isotopic patterns. The feature that was most highly represented across all individual chromatograms was then used as the representative feature from each pseudospectrum. Features were then normalized to plant mass and natural log–transformed.

To accommodate differences in ionization efficiency between individual phytochemicals, z transformation was applied to standardize means and variance of all features. Given the large number of plant specimens processed and the considerable amount of HPLC time involved, we applied this correction batchwise to correct for technical effects arising from changes in instrument response and unavoidable mechanical artifacts such as the changing of the column (the internal standard did not adequately correct for technical error across all compound classes and was not used for normalization). We applied z transformation across four different batches of samples that were identified through manual inspection of individual compound response across analysis time and corresponded to analysis batches. In addition, we implemented a "floor" correction across batches so that the highest minimum z score (comparing across batches) became the lowest value for all batches. The latter correction was suggested by the fact that the overall sensitivity of detection varied among batches, with some batches recovering a greater range of small values. Our primary variable of interest (the presence and absence of the butterfly across the landscape) was not confounded with batches (i.e., each batch included samples associated with presence and absence locations). Moreover, we repeated core analyses with different approaches to batch effect correction (e.g., fewer batches or without floor correction) and obtained results that were qualitatively similar to those reported in Fig. 2. In other words, the main result that phytochemical variation has both direct and indirect negative and positive effects that are comparable in magnitude to other factors (e.g., biotic interactions) is robust to the technicalities of mass spectra processing.

Putative annotations were attempted for all features having factor loadings (see Analyses: Factor analysis, below) with an absolute value greater than 0.3 based on previously described approaches (*17*). Briefly, initial classification of phenolics [200 to 400 parts per million (ppm)], saponins (400 to 650 ppm), lipids, and sterols (greater than 400 to 650 ppm) was done using the relative mass defect as a characteristic of each compound class. Annotations were further refined on the basis of expected retention time, amine-characteristic masses, and molecular ion mass. Features of interest were extracted from raw data and examined for characteristic fragments, adducts, and isotopes before cross-referencing against the METLIN mass spectrometric database (*42*) as a way to further categorize annotations. Because of the implications for bioactivity associated with small molecule alkaloids, nitrogenous compounds with ambiguous database hits were characterized as peptides so as to not overstate the presence of defensive alkaloids. One reason we could not annotate potential alkaloids is the lack of literature regarding the presence or characterization of alkaloids in *Medicago*, despite numerous studies of alkaloids in other legumes. Thus, more targeted investigation of *Medicago* alkaloids is warranted. Compounds that did not yield a chemically rational database match as a molecular ion or fragment were classified as unknown.

Dispersal (estimation of effective migration surface)

We analyzed genotyping-by-sequencing (GBS) data from 541 Melissa blue butterflies collected from 27 populations in western North America (see table S1). These DNA sequences were previously described by Chaturvedi *et al.* (*10*). For the current study, we used the bwa mem algorithm (version 0.7.17) (*43*) to align these data to a new version of the *L. melissa* reference genome (*44*). We ran bwa mem with a minimum seed length of 15, considered internal seeds of longer than

20 base pairs, and only output alignments with a quality score of >30. We then used samtools (version 1.5) to compress, sort, and index the alignments (*45*). We used samtools (version 1.5) and bcftools (version 1.6) for variant calling with the original consensus calling algorithm. We used the recommended mapping quality adjustment for Illumina data (-C 50), skipped alignments with mapping quality less than 20, skipped bases with base quality less than 30, and ignored insertion-deletion polymorphisms. We set the prior on single-nucleotide polymorphisms (SNPs) to 0.001 (-P) and called SNPs when the posterior probability that the nucleotide was invariant was ≤0.01 (-p). We filtered the initial set of variants to retain only SNPs with sequence data for at least 80% of the individuals, a mean sequence depth of 2× per individual, at least 10 reads of the alternative allele, a minimum quality score of 30, and no more than 1% of the reads in the reverse orientation (this is an expectation for our GBS method). We then split the SNP data into rare versus common SNPs, as these sets of SNPs can reveal different aspects of demographic history with rare variants being especially informative about recent gene flow and fine-scale population structure (*9*). We specifically delineated rare variants as those with an overall minor allele frequency of 1 to 5% (47,470 SNPs) and common variants as those with minor allele frequencies >5% (20,449) (variants with less than 1% frequency were discarded from downstream analyses).

We used entropy (version 1.2) to estimate genotypes. This program jointly infers genotypes and allele frequencies while accounting for uncertainty in each, as well as uncertainty in population assignment and ancestry (*9*). The latter is accomplished via an admixture model that assumes that the allele copies at each SNP locus are drawn from unknown, hypothetical source populations with each individual having a genome composed of some mixture of the source populations (*9*). Uncertainty in genotypes comes from limited coverage and sequencing error, as encoded in the genotype likelihoods estimated by samtools and bcftools. We estimated genotypes assuming two or three source populations. Estimates were obtained via Markov chain Monte Carlo (MCMC) with five chains, each with 5000 iterations as a burn-in followed by 8000 sampling iterations with a thinning interval of 5. Point estimates of genotypes were obtained as the posterior mean estimate of the number of nonreference alleles averaged across chains and numbers of source populations. Genetic distances were then calculated between all pairs of individuals on the basis of average identity by state.

Last, we estimated relative effective migration rates among populations on the basis of the genetic distances and sampling location; this was done separately for rare versus common variants. Relative effective migration rates were inferred using the program eems (version 0.0.0.9000) (*46*). This method does not infer absolute migration rates but rather identifies regions in space with low or high gene flow relative to a simple two-dimensional stepping-stone isolation-by-distance model (*46*). Thus, this method can identify regions in space receiving limited dispersal or gene flow, even if these regions do not harbor resident butterfly populations (for example, one of our sampled alfalfa locations that has not been colonized by the butterfly). In addition to estimating migration rates separately for rare and common variants, we fit the model (using MCMC with three chains, 4 million sampling iterations, 2 million burn-in iterations, and a thinning interval of 10,000) separately assuming 400 demes and 200 demes (to allow for more fine-scale and more coarsely estimated variation), evenly spaced on a triangular grid.

Analyses: Overview

Our goal was to use SEM to understand potentially complex direct and indirect effects on the presence and absence of our focal butterfly across the landscape. We implemented two levels of data reduction before SEM analysis. First, we used exploratory factor analysis on the climate and metabolomic data. Then, we used Bayesian ridge regression as a way to reduce the possible number of predictor (exogenous) variables that would need to be included for each response (endogenous) variable in our SEM. Last, we used the most successful SEM model to discuss the relative importance of different direct and indirect paths affecting the presence and absence of the focal butterfly. The success of the SEM model was judged by leave-one-out cross-validation and null simulations that permuted the site-level properties (predictor variables) for Melissa blue presence and absence.

Analyses: Factor analysis

Factor analysis is similar to other ordination techniques (such as principal components analysis) in dealing with suites of correlated variables, but the emphasis in factor analysis is on the identification of underlying structure (associated with factors that are not necessarily orthogonal) giving rise to the observed data (*47*). We used the same approach for both datasets (climate and metabolomic), specifically the factanal function with promax rotation and Thompson's regression scores. The number of factors calculated for each dataset was determined partly on the basis of inspection of scree plots but mainly through experimentation by fitting different numbers of factors and repeating downstream analyses (to learn, for example, if additional factors produced additional insight or meaningful effects in SEM models). For the climate dataset, the values going into the factor analysis were already summarized at the site level; thus, factor scores could be moved directly into the next analyses. For the metabolomic dataset, data from all 1651 individual plants were used in the factor analysis, and then, average scores at the site level for each factor were retained for ridge regressions and SEM models.

Analyses: Bayesian ridge regression

Before constructing SEMs, we used Bayesian ridge regression as a way to focus on a subset of important predictor variables for each of our endogenous variables (presence and absence of the butterfly and the abundance of interacting species). Ridge regression is a constrained regression in which coefficients are penalized in a way that reduces coefficients toward zero but does not exclude them (known as an l_2 penalty), thus allowing for the simultaneous estimation of effects of a large number of predictors (*48*). In a Bayesian context, ridge regression can be implemented by placing a hyperprior on the precision (equal to 1/variance) of regression coefficients, which lets the model learn how much the variance on coefficients should be constrained. Ridge regressions were either logistic (for presence and absence of the butterfly) or Gaussian (for logged abundance of ants and functional groups) and were run using JAGS (version 3.2.0) in R with the rjags package (*49*). Minimally influential priors on regression coefficients were modeled as normal distributions with a mean of zero and precision drawn from a (hyperprior) gamma distribution (rate, 0.1; shape, 0.1). After sampling with two Markov chains for 100,000 steps each (burn-in was not required), performance was evaluated by plotting chain histories, examining effective sample sizes, and calculating the Gelman and Rubin convergence diagnostic (*50*).

All variables were *z*-transformed before analyses, and different subsets of variables were examined for different response variables.

For example, for nonherbivorous groups (ants and predators), we allowed for the possibility that plant architecture might be important (as we have seen previously (*35*)), but not phytochemistry, specific leaf area, or leaf toughness. Year was included as a binary variable in all ridge regression models to allow for the possibility that variation among the two sampling years should be controlled for while estimating other variables of interest. After ridge regressions had been run (and diagnostics were checked) for each response variable, we calculated the fraction of the posterior distribution above or below zero for coefficients whose point estimate (median) was above or below zero, respectively. This value represented our confidence in the sign of the coefficients (positive or negative), and we selected variables (separately for each response variable) with confidence equal to or greater than 75% as variables of potential importance to move forward into SEMs. The cutoff of 75% was determined through experimentation: A higher cutoff missed some variables that were interesting in the downstream SEM, and a lower cutoff included variables that were unimportant (e.g., had very small effect sizes) in the subsequent SEM models.

Analyses: Spatial autocorrelation

Our previous work on alfalfa and the Melissa blue butterfly in the Great Basin has suggested to us that spatial autocorrelation might not be a major factor in this system, as spatially proximate patches of alfalfa are, in some cases, similar and, in other cases, very different with respect to female oviposition and larval performance (*15*). Nevertheless, the present project encompasses more space than our previous ecological work, which raises the importance of quantifying spatial autocorrelation, for which we have taken two approaches. First, we calculated Moran's *I* statistic of spatial autocorrelation for each of our variables and asked whether observations have greater or lesser autocorrelation than would be expected based on 1000 random permutations, using the moran.randtest function of the adespatial package (*51*). Moran's *I* is not, however, appropriate for binary presence and absence data, which is, of course, the variable of central interest (presence and absence of our focal butterfly). Thus, a complementary approach involved the generation of MEMs, which can be used in a multiple regression context to account for spatial autocorrelation at a range of scales (*51*). MEMs were generated using the dbmem function in adespatial, and significant MEMs (at $\alpha = 0.05$) were included in a Bayesian logistic ridge regression along with other predictors of Melissa blue presence and absence.

Analyses: Structural equation models

Following Bayesian ridge regressions, we had a suite of predictor variables for each of our five endogenous variables (butterfly presence and absence, ants, tended herbivores, other herbivores, and predators; note that caterpillar parasitoids could have been another endogenous variable in SEM models but were not included as such because they did not emerge as a top variable for the butterfly). Our initial SEM included three unresolved relationships: For example, ants are important for tended herbivores, and (expectedly) tended herbivores are important for ants. Unresolved causality can be retained in SEMs as correlations, but inferences are stronger if relationships (directions of influence or association) can be resolved (*52*). To that end, we compared the fit [using the Akaike information criterion (AIC)] of models with associations pointing in different directions; if the fit in one direction was a marked improvement, then the resolved relationship was retained in the final model. We used the piecewise package (*53*) to fit our SEM model, which allows for the inclusion of different error structures (binomial and Gaussian). To estimate standardized beta coefficients across the variables on different scales, we chose the "Menard.OE" option and judged overall model fit using Fisher's *C* and the associated significance test (*52*). In addition to the traditional R^2 reported by the piecewise package, we manually generated a cross-validation measure of model fit by repeating the SEM 56 times, leaving out one location with each iteration and calculating the correlation between observed and predicted values for each of our endogenous variables. Last, we compared the R^2 from the full model to the distribution of R^2 values from 1000 null simulations in which site-level attributes were shuffled among locations with each simulation.

For interpretation and visualization, we produced partial plots for certain relationships of interest by reproducing components of the full SEM as stand-alone generalized linear models from which residuals were saved and plotted as either bivariate plots or three-dimensional plots (exploring, for example, the probability of Melissa blue presence or absence as a function of combinations of ant abundance and tended herbivore abundance). In addition, we calculated indirect effects following standard procedures in path analysis involving, for example, the multiplication of coefficients leading from one exogenous variable through an intermediary endogenous variable to a final endogenous variable, as well as the addition of indirect effects to estimate, for example, the total indirect effect associated with a suite of predictor (exogenous) variables of a particular type (such as the indirect effect associated with all plant structural exogenous variables) (*52*). We also used direct and indirect path coefficients associated with phytochemical factors as a way to visualize the potential importance of individual compounds. For each compound, we multiplied its loading on a particular factor by the path coefficient associated with that factor and then summed across factors within indirect effects and direct effects separately.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/48/eabc6852/DC1

REFERENCES AND NOTES

1. S. Nylin, S. Agosta, S. Bensch, W. A. Boeger, M. P. Braga, D. R. Brooks, M. L. Forister, P. A. Hambäck, E. P. Hoberg, T. Nyman, A. Schäpers, A. L. Stigall, C. W. Wheat, M. Österling, N. Janz, Embracing colonizations: A new paradigm for species association dynamics. *Trends Ecol. Evol.* **33**, 4–14 (2017).
2. J. Fleischer, P. Pregitzer, H. Breer, J. Krieger, Access to the odor world: Olfactory receptors and their role for signal transduction in insects. *Cell. Mol. Life Sci.* **75**, 485–508 (2018).
3. D. G. Heckel, Insect detoxification and sequestration strategies. *Annu. Plant Rev.*, 77–114 (2018).
4. L. A. Dyer, C. S. Philbin, K. M. Ochsenrider, L. A. Richards, T. J. Massad, A. M. Smilanich, M. L. Forister, T. L. Parchman, L. M. Galland, P. J. Hurtado, A. E. Espeset, A. E. Glassmire, J. G. Harrison, C. Mo, S. A. Yoon, N. A. Pardikes, N. D. Muchoney, J. P. Jahner, H. L. Slinn, O. Shelef, C. D. Dodson, M. J. Kato, L. F. Yamaguchi, C. S. Jeffrey, Modern approaches to study plant–insect interactions in chemical ecology. *Nat. Rev. Chem.* **2**, 50–64 (2018).
5. M. S. Singer, J. O. Stireman, The tri-trophic niche concept and adaptive radiation of phytophagous insects. *Ecol. Lett.* **8**, 1247–1255 (2005).
6. E. P. Hoberg, D. R. Brooks, A macroevolutionary mosaic: Episodic host-switching, geographical colonization and diversification in complex host-parasite systems. *J. Biogeogr.* **35**, 1533–1550 (2008).
7. A. A. Agrawal, J. K. Conner, S. Rasmann, Tradeoffs and negative correlations in evolutionary ecology, in *Evolution After Darwin: The First 150 Years*, M. Bell, D. J. Futuyma, W. F. Eanes, J. Levinton Eds. (Sinauer Associates, 2010), pp. 243–268.
8. S. A. Yoon, Q. Read, Consequences of exotic host use: Impacts on Lepidoptera and a test of the ecological trap hypothesis. *Oecologia* **181**, 985–996 (2016).

9. Z. Gompert, L. K. Lucas, C. A. Buerkle, M. L. Forister, J. A. Fordyce, C. C. Nice, Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Mol. Ecol.* **23**, 4555–4573 (2014).
10. S. Chaturvedi, L. K. Lucas, C. C. Nice, J. A. Fordyce, M. L. Forister, Z. Gompert, The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies. *Mol. Ecol.* **27**, 2651–2666 (2018).
11. M. L. Forister, C. C. Nice, J. A. Fordyce, Z. Gompert, Host range evolution is not driven by the optimization of larval performance: The case of *Lycaeides melissa* (Lepidoptera: Lycaenidae) and the colonization of alfalfa. *Oecologia* **160**, 551–561 (2009).
12. M. L. Forister, C. F. Scholl, Use of an exotic host plant affects mate choice in an insect herbivore. *Am. Nat.* **179**, 805–810 (2012).
13. S. A. Yoon, J. G. Harrison, C. S. Philbin, C. D. Dodson, D. M. Jones, I. S. Wallace, M. L. Forister, A. M. Smilanich, Host plant-dependent effects of microbes and phytochemistry on the insect immune response. *Oecologia* **191**, 141–152 (2019).
14. Z. Gompert, J. P. Jahner, C. F. Scholl, J. S. Wilson, L. K. Lucas, V. Soria-Carrasaco, J. A. Fordyce, C. C. Nice, C. A. Buerkle, M. L. Forister, The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Mol. Ecol.* **24**, 2777–2793 (2015).
15. J. G. Harrison, Z. Gompert, J. A. Fordyce, C. A. Buerkle, R. Grinstead, J. P. Jahner, S. Mikel, C. C. Nice, A. Santamaria, M. L. Forister, The many dimensions of diet breadth: Phytochemical, genetic, behavioral, and physiological perspectives on the interaction between a native herbivore and an exotic host. *PLOS ONE* **11**, e0147971 (2016).
16. M. L. Forister, Z. Gompert, C. C. Nice, G. W. Forister, J. A. Fordyce, Ant association facilitates the evolution of diet breadth in a lycaenid butterfly. *Proc. Biol. Sci.* **278**, 1539–1547 (2011).
17. M. L. Forister, S. A. Yoon, C. S. Philbin, C. D. Dodson, B. Hart, J. G. Harrison, O. Shelef, J. A. Fordyce, Z. H. Marion, C. C. Nice, L. A. Richards, C. A. Buerkle, Z. Gompert, Caterpillars on a phytochemical landscape: The case of alfalfa and the Melissa blue butterfly. *Ecol. Evol.* **10**, 4362–4374 (2020).
18. N. E. Pierce, M. F. Braby, A. Heath, D. J. Lohman, J. Mathew, D. B. Rand, M. A. Travassos, The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). *Annu. Rev. Entomol.* **47**, 733–771 (2002).
19. S. J. Agosta, On ecological fitting, plant-insect associations, herbivore host shifts, and host plant selection. *Oikos* **114**, 556–565 (2006).
20. S. Nylin, N. Janz, Butterfly host plant range: An example of plasticity as a promoter of speciation? *Evol. Ecol.* **23**, 137–146 (2009).
21. S. Y. Strauss, J. A. Lau, S. P. Carroll, Evolutionary responses of natives to introduced species: What do introductions tell us about natural communities? *Ecol. Lett.* **9**, 357–374 (2006).
22. S. B. L. Araujo, M. P. Braga, D. R. Brooks, S. J. Agosta, E. P. Hoberg, F. W. von Hartenthal, W. A. Boeger, Understanding host-switching by ecological fitting. *PLOS ONE* **10**, e0139225 (2015).
23. M. C. Singer, C. S. McBride, Geographic mosaics of species' association: A definition and an example driven by plant-insect phenological synchrony. *Ecology* **93**, 2658–2673 (2012).
24. Z. Gompert, F. J. Messina, Genomic evidence that resource-based trade-offs limit host-range expansion in a seed beetle. *Evolution* **70**, 1249–1264 (2016).
25. N. B. Hardy, C. Kaczvinsky, G. Bird, B. B. Normark, What we don't know about diet-breadth evolution in herbivorous insects. *Annu. Rev. Ecol. Evol. Syst.* **51**, (2020).
26. L. M. Brown, G. A. Breed, P. M. Severns, E. E. Crone, Losing a battle but winning the war: Moving past preference–performance to understand native herbivore–novel host plant interactions. *Oecologia* **183**, 441–453 (2017).
27. C. W. Fox, J. A. Nilsson, T. A. Mousseau, The ecology of diet expansion in a seed-feeding beetle: Pre-existing variation, rapid adaptation and maternal effects? *Evol. Ecol.* **11**, 183–194 (1997).
28. R. M. Pateman, J. K. Hill, D. B. Roy, R. Fox, C. D. Thomas, Temperature-dependent alterations in host use drive rapid range expansion in a butterfly. *Science* **336**, 1028–1030 (2012).
29. M. L. Forister, V. Novotny, A. K. Panorska, L. Baje, Y. Basset, P. T. Butterill, L. Cizek, P. D. Coley, F. Dem, I. R. Diniz, P. Drozd, M. Fox, A. E. Glassmire, R. Hazen, J. Hrcek, J. P. Jahner, O. Kama, T. J. Kozubowski, T. A. Kursar, O. T. Lewis, J. Lill, R. J. Marquis, S. E. Miller, H. C. Morais, M. Murakami, H. Nickel, N. A. Pardikes, R. E. Ricklefs, M. S. Singer, A. M. Smilanich, J. O. Stireman, S. Villamarín-Cortez, S. Vodka, M. Volf, D. L. Wagner, T. Walla, G. D. Weiblen, L. A. Dyer, The global distribution of diet breadth in insect herbivores. *Proc. Natl. Acad. Sci.* **112**, 442–447 (2015).
30. J. C. Vamosi, W. S. Armbruster, S. S. Renner, Evolutionary ecology of specialization: Insights from phylogenetic analysis. *Proc. R. Soc. B* **281**, 20142004 (2014).
31. C. C. Nice, A. M. Shapiro, Molecular and morphological divergence in the butterfly genus *Lycaeides* (Lepidoptera : Lycaenidae) in North America: Evidence of recent speciation. *J. Evol. Biol.* **12**, 936–950 (1999).
32. M. L. Forister, C. F. Scholl, J. P. Jahner, J. S. Wilson, J. A. Fordyce, Z. Gompert, D. R. Narala, C. A. Buerkle, C. C. Nice, Specificity, rank preference, and the colonization of a non-native host plant by the Melissa blue butterfly. *Oecologia* **172**, 177–188 (2013).
33. M. D. Preston, M. L. Forister, J. W. Pitchford, P. R. Armsworth, Impact of individual movement and changing resource availability on male--female encounter rates in an herbivorous insect. *Ecol. Complex.* **24**, 1–13 (2015).
34. M. L. Forister, Anthropogenic islands in the Arid West: Comparing the richness and diversity of insect communities in cultivated fields and neighboring wildlands. *Environ. Entomol.* **38**, 1028–1037 (2009).
35. J. G. Harrison, C. S. Philbin, Z. Gompert, G. W. Forister, L. Hernandez-Espinoza, B. W. Sullivan, I. S. Wallace, L. Beltran, C. D. Dodson, J. S. Francis, A. Schlageter, O. Shelef, S. A. Yoon, M. L. Forister, Deconstruction of a plant-arthropod community reveals influential plant traits with nonlinear effects on arthropod assemblages. *Funct. Ecol.* **32**, 1317–1328 (2018).
36. E. Hart, K. Bell, A. Butler, prism: Download data from the Oregon prism project. R package version 0.0.6, 10.5281/zenodo.33663 (2015).
37. R Core Development Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2016).
38. J. J. Balsevich, G. G. Bishop, L. K. Deibert, Use of digitoxin and digoxin as internal standards in HPLC analysis of triterpene saponin-containing extracts. *Phytochem. Anal.* **20**, 38–49 (2009).
39. M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Schulman, B. Frewen, T. A. Baker, M. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Landridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
40. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
41. C. Kuhl, R. Tautenhahn, C. Bottcher, T. R. Larson, S. Neumann, CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/ mass spectrometry data sets. *Anal. Chem.* **84**, 283–289 (2011).
42. C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagayan, G. Siuzdak, METLIN. *Ther. Drug Monit.* **27**, 747–751 (2005).
43. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).
44. S. Chaturvedi, L. K. Lucas, C. A. Buerkle, J. A. Fordyce, M. L. Forister, C. C. Nice, Z. Gompert, Recent hybrids recapitulate ancient hybrid outcomes. *Nat. Commun.* **11**, 2179 (2019).
45. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
46. D. Petkova, J. Novembre, M. Stephens, Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
47. G. R. Norman, D. L. Streiner, *Biostatistics: The Bare Essentials* (PMPH USA, 2008).
48. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013), vol. 112.
49. M. Plummer, JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing.* **24**, 1–10 (2003).
50. S. P. Brooks, A. Gelman, General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434–455 (1998).
51. S. Dray, G. Blanchet, D. Borcard, G. Guenard, T. Jombart, G. Larocque, P. Legendre, N. Madi, H. H. Wagner, adespatial: Multivariate multiscale spatial analysis. R Package version 0.0. 3 (2016).
52. B. Shipley, *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations, and Causal Inference* (Cambridge Univ. Press, 2000).
53. J. Lefcheck, J. Byrnes, J. Grace, Package 'piecewiseSEM.' R Packag. version. 1 (2016).

Acknowledgments: We thank the Hitchcock Center for Chemical Ecology; S. Flanagan and K. Bell for help with collection; E. Perry for help with plant specimens; and J. Jahner, S. Yoon, and J. Harrison whose travels in the Great Basin discovered many important field sites. We also thank three anonymous reviewers who improved the manuscript. **Funding:** NSF grant DEB-1638793 to M.L.F. and C.D.D., DEB-1638768 to Z.G., DEB-1638773 to C.C.N., DEB-1638922 to J.A.F., and DEB-1638602 to C.A.B. M.L.F. was additionally supported by a Trevor James McMinn

professorship. **Author contributions:** Overall concept and approach by M.L.F., C.S.P, C.A.B, C.D.D., J.A.F., S.L.L., L.K.L., C.C.N., and Z.G. Data collection by Z.H.M., C.S.P, G.W.F., and M.L.F. Data analysis by M.L.F., C.S.P., Z.G., J.A.F., C.C.N., and Z.H.M. All authors read and edited the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data analyzed in this study are available through Dryad: https://doi.org/10.5061/dryad.cz8w9gj23. Additional data related to this paper may be requested from the authors.

Submitted 7 May 2020
Accepted 6 October 2020
Published 27 November 2020
10.1126/sciadv.abc6852

Citation: M. L. Forister, C. S. Philbin, Z. H. Marion, C. A. Buerkle, C. D. Dodson, J. A. Fordyce, G. W. Forister, S. L. Lebeis, L. K. Lucas, C. C. Nice, Z. Gompert, Predicting patch occupancy reveals the complexity of host range expansion. *Sci. Adv.* **6**, eabc6852 (2020).

Science Advances

Predicting patch occupancy reveals the complexity of host range expansion

M. L. Forister, C. S. Philbin, Z. H. Marion, C. A. Buerkle, C. D. Dodson, J. A. Fordyce, G. W. Forister, S. L. Lebeis, L. K. Lucas, C. C. Nice and Z. Gompert

Sci Adv **6** (48), eabc6852.
DOI: 10.1126/sciadv.abc6852