Caterpillars on a phytochemical landscape

Forister, M.L.[1,2], Yoon, S.[1,2], Philbin, C.S.[2,3], Dodson, C.D.[2,3], Hart, B.[4], Harrison, J.G.[5], Shelef, O.[6], Fordyce, J.A.[7], Marion, Z.H.[8], Nice, C. C.[9], Richards, L.A.[1,2], Buerkle, C.A.[5], Lebeis, S.[10], and Z. Gompert[11]

[1] Department of Biology, Program in Ecology, Evolution and Conservation Biology, University of Nevada, Reno, NV 89557, USA; Email: forister@gmail.com

[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] Department of Biochemistry, University of Nevada, Reno, NV 89557, USA

[5] Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY 82071, USA

[6] Department of Natural Resources, Institute of Plant Sciences, Volcani Center, Agricultural Research Organization, Rishon LeZion, Israel

[7] Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

[8] Department of Ecology, University of Canterbury, Christchurch, New Zealand

[9] Department of Biology, Population and Conservation Biology Program, Texas State University, San Marcos, TX 78666, USA

[10] Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA

[11] Department of Biology, Utah State University, Logan, UT 84322, USA

Abstract

Foundational theories in plant-animal interactions are being updated with input from modern metabolomic approaches that offer more comprehensive phytochemical profiles than were previously available. Here we use a recently-formed plant-insect interaction, the colonization of alfalfa (*Medicago sativa*) by the Melissa blue butterfly (*Lycaeides melissa*), to describe the landscape of primary and secondary plant metabolites and the performance of caterpillars as affected by both individual compounds and suites of covarying phytochemicals. We find that survival, development time and adult weight are all affected by a large number of compounds, including biomolecules associated with plant cell function and putative anti-herbivore action. The dimensionality of the plant-insect interface is high, with clusters of covarying compounds in many cases encompassing divergent effects on different aspects of caterpillar performance. The sapogenic glycosides are represented by more than 20 individual compounds with some of the strongest beneficial and detrimental effects on caterpillars, which highlights the value of metabolomic data as opposed to previous approaches that relied on total concentrations within defensive classes. Considering positive and negative effects of both secondary compounds and primary metabolites (possibly associated with nutritional imbalance), theories of the evolution of plant defense based on a simple dichotomy between investment in defense or primary metabolism appear to be overly simplistic. Results are also discussed in light of previous work on local adaptation to alfalfa by the focal herbivore.

Keywords

Lycaeides melissa, Medicago sativa, metabolomics, plant defense, specialization

1. Introduction

One of the conceptual pillars of trophic ecology is the idea that herbivores must overcome the barrier of plant secondary chemistry before extracting the nutrients necessary for growth and reproduction [1]. The success of this idea is reflected in areas of research that include coevolution [2], ecological specialization [3], and nutrient flow in ecosystems [4]. In most cases, progress has been made by chemical ecologists focusing on tiny subsets of the secondary metabolites produced by plants and consumed by herbivores. The focus on a few charismatic molecules or classes of compounds, such as furanocoumarins [5] or cardiac glycosides [6], was at least in part necessitated by early methods in natural products chemistry that were targeted and not easily optimized for the discovery of large suites of co-occurring primary and secondary metabolites [7,8]. As technological limitations have dissipated, the opportunity now exists for a deeper understanding of the challenges faced by herbivores, with the possibility of discovering, among other things, novel compounds and synergistic interactions among compounds [9–11]. More generally, an important task is to quantify the phytochemical dimensionality of the antagonistic interaction between plants and herbivores, with an eye towards understanding constraints on the evolution of both players [12,13] and predicting the formation of new plant-herbivore interactions [14]. Here we use the example of a specialized herbivore and a novel host plant to investigate the phytochemical landscape from the perspective of developing caterpillars.

The Melissa blue butterfly, *Lycaeides melissa*, specializes on larval host plants in the pea family (Fabaceae), primarily in the genera *Astragalus* and *Lupinus*. Within the last 200 years, *L. melissa* has colonized introduced alfalfa, *Medicago sativa* (Fabaceae), at least twice and probably multiple times [15], forming a heterogeneous patchwork of association throughout the range of the butterfly in western North America. *M. sativa* is a suboptimal host, relative to native hosts that have been examined, and populations of the butterfly that persist on *M. sativa* show evidence of loss of preference for native hosts [16], reduced caterpillar performance on native hosts, and a slight increase in ability to develop on the suboptimal novel host [17]. The genetic architecture of host use in this system is known to be polygenic and characterized by loci with conditionally neutral (host-specific) effects [17], but what is needed next is

an understanding of how many independent or covarying phytochemical compounds have consequential effects on caterpillars eating the novel host. For example, will the trajectory of further local adaptation by *L. melissa* to *M. sativa* be a matter of evolving the ability to detoxify one or a large number of compounds? We would also like to know how key compounds covary among individual plants, which should help us understand the puzzle faced by ovipositing females facing an array of co-occurring *M. sativa* chemotypes. Here we use a common garden approach (to minimize non-genetic phenotypic variation among plants) and caterpillars individually reared in a controlled environment to address these questions while describing the effects of metabolomic variation in *M. sativa* on *L. melissa*.

2. Methods

(a) Plants and caterpillars

Plants used in this project were grown at the University of Nevada, Reno, Main Station experimental farm. The common garden was planted in 2016 with seeds collected the previous year from 45 plants (previously studied by Harrison et al. [18]) growing in a fallow field in north-western Nevada on the western edge of the Great Basin Desert. The focal butterfly, *L. melissa*, was present in the source field but has not yet colonized the university farm where experimental plants were grown. The 45 maternal plants each contributed 15 offspring to a randomized grid design in the common garden, irrigated with broadcast sprayers in 2016 and drip in 2017, without supplemental fertilization. Out of each maternal family, a single plant was randomly selected for use in the rearing experiment reported here as a way to capture as much genetic and phenotypic variation as possible.

On 17 and 18 July 2017, a total of 45 *L. melissa* females were collected from an alfalfa-associated population near Verdi, NV, and confined to oviposition arenas (500 mL plastic cups) in groups of three with host plant leaves and supplied with Gatorade on mesh lids. After three days, eggs were removed from leaves, pooled, and kept at room temperature until hatching, at which time caterpillars were placed individually in Petri dishes (100 x 25 mm) with leaves of a particular *M. sativa* individual (which became the only plant from which they were fed throughout the experiment). Ten caterpillars were assigned to

each of the 45 experimental *M. sativa* plants (for a total of 450 independently-reared caterpillars) and kept in a growth chamber set to 25 C and a 12 hour light / 12 hour dark cycle. From each caterpillar we recorded survival to adult, date of eclosion (if successful) and adult weight to the nearest 0.01 mg on a Mettler Toledo XP26 microbalance.

(b) Phytochemistry and plant traits

Metabolomic variation among individual plants was characterized with liquid chromatography–mass spectrometry (LC-MS, [19]) using leaves collected on a single day at the start of the rearing experiment. Dried, ground leaves (10 mg) were extracted in 2 mL of 70% aqueous ethanol and injected into an Agilent 1200 analytical high performance liquid chromatograph paired with an Agilent 6230 Time-of-Flight mass spectrometer via an electrospray ionization source. Resulting chromatograms were analyzed using MassHunter Quantitative Analysis (v.B.06.00, Agilent, Santa Clara, CA); see electronic supplementary material for additional phytochemical protocols. Major classes of compounds were identified using characteristic relative mass defects [20]. Leaf protein content was quantified with three replicates (~2 mg each) per plant using the Bicinchoninic acid assay (Pierce Biotechnology, Waltham, MA). Before grinding, five dried leaflets from each sample were weighed to the nearest 0.1 mg, scanned, and area was measured using ImageJ (v.1.52a); specific leaf area (SLA) was calculated as leaf area divided by dry mass. Finally, leaf toughness was measured on fresh material in the common garden, at the start of the experiment (mid-July, when leaves were also sampled for chemistry and protein) and at the end of the experiment (mid-August), from three leaves per plant at each date, with a penetrometer (Chatillon 516 Series) through the center of the middle leaflet (as in [18]); the three leaves were selected haphazardly, avoiding the oldest and youngest leaves. The six leaf toughness measurements per plant were averaged for a single toughness measure used in analyses.

(c) Analyses of plant traits and caterpillar performance

(i) Overview

125 Our analytical strategy to understand the association between phytochemical variation and caterpillar
126 performance followed two complementary paths, one focusing on dimension reduction and feature
127 selection to produce relatively simple models, and the other on the estimation of effects of all individual
128 compounds. For the first path, involving dimension reduction and feature selection, we utilized an
129 approach developed for gene transcription studies that identifies groups or modules of correlated variables
130 with hierarchical clustering [21]; after clustering, we reduced the number of independent variables by
131 selecting among modules and other plant traits (specific leaf area, protein and leaf toughness) using lasso
132 regression [22]. Selected modules (and other plant traits) were analyzed in Bayesian linear models that
133 are useful in this context because they allowed us to quantify our confidence in the sign of effects
134 (positive or negative) as continuous probabilities (as opposed to relying on arbitrary significance cutoffs).
135 For the second analytical path, we utilized ridge regression [22] to estimate effects for all compounds
136 simultaneously, which allowed us to investigate the distribution of effects among compounds and classes
137 of compounds. Both analytical paths incorporated cross-validation during the lasso and ridge regressions,
138 and as a means of evaluating the predictive success of the Bayesian models. We also used randomization
139 tests to compare the performance of modules and individual compounds with randomly-chosen suites of
140 compounds.
141
142 **(ii) Dimension reduction and feature selection**
143 We chose an approach that reduces the number of independent variables while allowing us to learn
144 something about the correlational structure of the data, specifically unsupervised hierarchical clustering as
145 implemented in the blockwiseModules function of the WGCNA package [21] in R [23]. Among the
146 options in the pipeline, we used positive correlations among variables ("signed" network type), merge cut
147 height at 0.25, and correlations raised to the power of five (which is where the scale free topology index
148 reached a plateau). Through experimentation, we found that our results with LC-MS data were robust to
149 variation in these choices, including the choice of signed or unsigned networks. After an initial round of
150 clustering, we took a remaining 19 unassigned compounds and put them through a second round of

151 clustering (although the majority of consequential compounds were identified in the first round). One
152 output of the WGCNA procedure is the first eigenvector from each cluster of compounds, which reduced
153 our number of predictor variables by a factor of ten. The resulting eigenvectors plus protein, SLA
154 (specific leaf area) and leaf toughness were then put through the feature reduction step of lasso regression
155 [22], a penalized regression that allows beta coefficients to be constrained to zero (thus excluding
156 variables). We used the cv.glmnet function of the glmnet package [24] with cross-validation during error
157 reduction set to leave out one plant (and associated caterpillars) at each iteration. The variables selected
158 by the lasso were then put into a Bayesian linear model to estimate coefficients and associated credible
159 intervals using JAGS (version 3.2.0) run in R with the rjags package [25]. Two Markov chains were run
160 for 10,000 steps for each analysis (no burn in was required) and chain performance was assessed by
161 plotting chain histories, and calculating the Gelman and Rubin convergence diagnostic and effective
162 sample sizes [26,27]. For all models, uninformative priors for the regression coefficients were modeled
163 as a Normal distribution with a mean of zero and variance of 0.01. We quantified our confidence in the
164 sign of coefficients (positive or negative) as the fraction of the posterior samples that were less than zero
165 (for coefficients with a median negative value) or greater than zero (for coefficients with a median
166 positive value).

167 All analyses were done using the R statistical language [23] on scaled (z-transformed) predictor
168 variables, and both lasso and Bayesian models used binomial (for survival), Poisson (for development
169 time) and Gaussian (for adult weight) errors. The latter two analyses (development time and adult
170 weight) included sex as a factor. The analysis of development time also included adult weight as a
171 covariate; while (reciprocally) the analysis of adult weight included development time as a predictor.
172 These variables are negatively correlated (at -0.52), and they function as useful covariates of each other,
173 allowing us to investigate the possibility of unique plant effects on weight gain and development time,
174 which could not be discovered if, for example, these variables were combined into a single performance
175 index.

176 The success of models developed with the dimension reduction and feature selection pipeline was
177 judged in two ways. We used a cross-validation procedure in which we left out one plant (and associated
178 caterpillars) in each iteration of the Bayesian model and then used the estimated coefficients (for
179 phytochemical variables and other plant traits) to predict the performance of the unobserved caterpillars.
180 After 45 iterations (one for each plant), we calculated a simple correlation coefficient between the
181 observed and predicted performance of caterpillars across plants. In addition, we repeatedly resampled
182 the original LC-MS data to match the structure of the reduced set of predictor variables to ask to what
183 extent randomly assembled modules could outperform the empirically-derived modules.
184
185 **(iii) Individual compound effects**
186 The second path of our two-part analytical strategy involved simultaneous estimation of the effects of all
187 individual chemical compounds on caterpillar survival, development time and adult weight. For this
188 approach, we again used penalized regression (in the glmnet package [24]), but this time with ridge
189 regression (instead of lasso) which constrains beta coefficients to avoid variance inflation but does not
190 eliminate variables. As with the analyses above, ridge regression was done using error structures
191 appropriate to the specific response variables, and included additional covariates where possible (in
192 models of development time and adult weight). The resulting coefficients associated with all individual
193 compounds were examined as a second perspective on the modules examined in the first set of analyses,
194 and were used to ask to what extent individual compound effects could be predicted by the degree to
195 which they vary among individual plants as quantified with the simple coefficient of variation. To assess
196 confidence in the results of ridge regressions, we used a bootstrap approach, repeatedly resampling the
197 data and estimating coefficients 1000 times, noting the compounds whose bootstrap confidence intervals
198 did or did not overlap zero [28]. We also allowed for the discovery of interactions among compounds
199 using penalized regression on all individual compounds and all pairwise interactions between compounds.
200 For ease of interpretation, this final analysis of potential interactions used lasso (not ridge) regression,
201 letting the coefficients for many of the individual compounds and pairwise interactions go to zero.

202

203 **3. Results**

204 Of the 450 caterpillars that started the experiment, 261 were reared to eclosion as adults (a mortality rate
205 similar to previous work with this system [17]) on leaves from 45 individual alfalfa plants that were
206 characterized for protein, leaf toughness, specific leaf area and 163 individual metabolomic features (see
207 figure 1 for variation among plants in caterpillar performance and a subset of plant traits, and electronic
208 supplementary material table S1 for a list of compounds). Hierarchical clustering identified 14 subsets (or
209 modules) of compounds with generally low correlations among modules and high correlations within
210 modules (see electronic supplementary material figures S1 and S2 for correlations within and among
211 modules, and figure S3 for module variation among plants). The correlational structure of the
212 phytochemical data is illustrated as an adjacency network in figure 2, where it can be seen that some
213 modules (e.g., modules 1, 2, and 3) contain a great diversity of compound types, while other modules are
214 made up of more narrow classes (e.g. modules 7 and 8 which are mostly saponins, a class of defensive
215 secondary metabolites [29]). From the 14 eigenvectors summarizing variation in the modules, as well as
216 the other plant traits, lasso regression produced a reduced set of potential predictors which were then used
217 in Bayesian multiple regression models that included between seven and nine independent variables (table
218 1). The models had reasonably high performance in leave-one-out cross-validation (correlations between
219 observed and predicted values were between 0.50 and 0.59, table 1), and also in resampling analyses
220 (electronic supplementary material figure S4), where a small fraction (never more than 4%) of randomly-
221 generated models exceeded the variance explained of the models reported in table 1.

222 Variation among plants in the suites of covarying compounds had large effects on caterpillar
223 performance: for example, the beta coefficient of -2.33 (on the log-odds scale) associated with module 3
224 corresponds to a reduction in mean survival from 0.58 to 0.12 associated with a one unit change in
225 variation associated with that phytochemical module (table 1). The phytochemical predictor variables are
226 eigenvectors from clustering analysis, and thus are not entirely straightforward to interpret, especially
227 when the clustering analysis was itself based on z-transformed data. It is useful to note that our LC-MS

228 data consists of peak areas divided by the peak of an internal standard, and again divided by the dry
229 weight of the sample (thus, in total, referred to as "relative abundance per dry weight"). Variation in
230 these numbers reflects variation in concentrations within compounds (among plants), but care should be
231 used in comparing among compounds because of different ionization responses relative to the standard
232 (thus the use of z-transformation for among-compound analyses). Nevertheless, the effects reported in
233 table 1 reflect real variation in suites of compounds, as can be seen in correlations between eigenvectors
234 and individual compounds in electronic supplementary material figure S2, and in variation among plants
235 in average z-scores in figure S3.

236 Modules included in multiple regression models frequently had common effects across response
237 variables (e.g., the positive association of module 10 with both survival and adult weight), with the
238 exceptions of module 11 that had a solitary effect on survival and module 6 with an effect only on
239 development time (although the probability of the latter having a negative effect was only 0.75). Specific
240 leaf area had a negative effect on survival and adult weight, and the coefficients for specific leaf area (-
241 0.31 for survival and -0.42 for weight) were of smaller magnitude than most phytochemical effects.

242 Module-based analyses (as in table 1) focused on feature reduction with lasso regression; as a
243 complementary analytical approach, we used ridge regression on all of the compounds (which estimates
244 effects of individual compounds without excluding variables as in lasso regression). Analyses of
245 individual compounds by ridge regression (figure 3) were broadly consistent with the strongest module-
246 specific effects, as can be seen, for example, with module 10 having positive effects on survival and adult
247 weight in module analyses (table 1) and in compound-specific analyses (figure 3). Similarly, the
248 individual compounds in module 3 had negative compound-specific effects on survival (figure 3), and
249 that module had the strongest negative effect on survival in the eigenvector-based analyses in table 1. Not
250 surprisingly, the larger modules (with a greater number of covarying compounds, including many primary
251 metabolites) tended to have a more complex mix of positive and negative effects (for examples, modules
252 1 and 2, figure 3). For ease of interpretation, the coefficients from compound-specific regressions of
253 survival and development time (in figures 3 and 4) have been back-transformed to be on the scales of

254 probability and days (respectively), and displayed as changes relative to intercepts. For example, a
255 compound with a relatively large effect on survival in figure 3 could be associated with a one half percent
256 (0.005) reduction in the probability of survival relative to average survival and while holding other
257 compounds constant.

258 We also considered potential pairwise interactions among individual compounds, and found few
259 interactions that passed the filter of the penalized regression (electronic supplementary material table S2),
260 at least relative to the large number of potential interactions. We did not find evidence that more or less
261 variable compounds had differential effects on caterpillars, although there was a trend towards both
262 greater positive and greater negative effects being associated with less variable compounds (figure S5).
263 We did, however, see variation among classes of compounds in their effects on caterpillars (figure 4). All
264 classes included positive and negative effects, although phenolic effects on survival were more often
265 negative. The widest breadth of effects (including the most extreme positive and negative compound-
266 specific coefficients) tended to be found among peptides, saponins and phospholipids, with lipids and
267 terpenoids having a more narrow range (figure 4).

268

269 **4. Discussion**

270 The results reported here represent a dissection of the phytochemical landscape facing a specialized insect
271 herbivore attacking a novel host plant [30]. The phytochemical landscape is both physical, referring to
272 variation in compounds among individual plants in a common garden, and hypothetical to the extent that
273 effects of individual compounds on caterpillars are estimated, although compounds are of course not
274 encountered in isolation. Our exploration of the phytochemical landscape facing *L. melissa* on *M. sativa*
275 is necessarily a first draft based on a snapshot in time. Nevertheless, models including suites of covarying
276 compounds and other plant traits had predictive success and suggested different natural products affecting
277 survival, development time and adult weight (performance measured as adult weight is a proxy for
278 fecundity [31]). Previous work with *M. sativa* and insect herbivores has focused on sapogenic glycosides
279 [29], and a simple outcome from our study could have been that one or a small number of saponins have

anti-herbivore properties that reduce fitness of our focal insect. Instead, we find large numbers of compounds with potentially consequential effects on caterpillars (figure 3).

The precise identification of specific compounds is not as important as the more general result that the prominent class of defensive secondary chemistry, saponins, includes compounds with both positive and negative effects. Moreover, saponin effects tend slightly towards the positive with improved survival, faster development and increased adult size (figure 4). Positive effects are potentially associated with feeding stimulation, as has been observed (along with other positive effects) for other specialist herbivores and plant toxins [32,33]. Negative effects of saponins on insects potentially include disruption of hormone production [34], although exact modes of action on *L. melissa* will await further study. Perhaps more intriguing than the range of effects associated with saponins are the negative effects associated with variation in certain primary metabolites (figure 4). For example, some of the largest negative effects in module 3 (figure 2) are phospholipids, peptides and even sucrose. These could be direct effects if a compound is suboptimal for development, or they could be associated with nutritional imbalance [35], such that too much of one nutrient makes it difficult for caterpillars to consume a balanced diet. It has been suggested that the presentation of unbalanced nutrition can be a kind of anti-herbivore strategy for a plant that does not depend on secondary metabolites [36]. Although this possibility has not been thoroughly investigated in many systems with full metabolomic profiling, the idea that nutritional imbalance could be as important as direct toxicity suggests that we might update theories of the evolution of plant defense that were built on differential investment into simple categories of plant growth versus defense [37].

The finding that our specialist herbivore is affected by a wide range of metabolites, primary and secondary, that vary greatly even within a single host population has implications for our understanding of heterogeneity in the system and for the course of local adaptation of the herbivore to the novel host. *Lycaeides melissa* typically colonizes weedy or feral patches of *M. sativa* on roadsides or integrated into natural communities, and previous work has documented dramatic variation among individual alfalfa locations (often in close proximity) in the extent to which they can support caterpillar development [38].

306 Previous phytochemical data with a lower resolution was less successful in explaining that variation [38],
307 and the results reported here suggest that among patch variation could be explained by future studies
308 using metabolomic data as used here. The within-population complexity described in the current study
309 combined with previous evidence for dramatic among-population variation in *M. sativa* suitability for the
310 focal herbivore also raises the possibility that the novel host presents a multi-faceted and potentially ever-
311 shifting target from the perspective of evolving butterfly populations. In particular, it is possible that *M.*
312 *sativa* defense against a specialist herbivore might be realized through different combinations (within and
313 among populations) of individually-acting compounds, thus making it unlikely for butterflies in any one
314 population to possess an effective suite of alleles that improve fitness on the novel host. In this context, it
315 is interesting to note that a molecular genetic dissection of caterpillar performance in this system found a
316 large number (potentially hundreds) of individual loci associated with performance on *M. sativa*, yet
317 evolution in populations associated with the novel host is primarily associated with a loss (through
318 genetic drift) of the ability to eat a native host and only slight improvement in the ability to eat the novel
319 host [17].

320 The correlational structure of the phytochemical variation that we observed has implications for
321 the evolution of plant defense and the accumulation of insect herbivores on *M. sativa*. Specifically,
322 correlations among modules should make it possible to hypothesize directions of least resistance for
323 defense evolution. Compounds in module 9 had a negative effect on survival (table 1), and module 9
324 negatively covaried with module 10 (electronic supplementary material figure S1), which itself had a
325 positive association with caterpillar survival. Thus an increase in module 9 and an associated decrease in
326 10 would be beneficial for the plant, at least with respect to herbivory by our focal herbivore. Of course,
327 most plants do not have the luxury of optimizing defense against a single herbivore, and it is easy to
328 imagine that improvements in defense against one enemy could lead to increased attraction to another,
329 especially given the diversity of effects even within major classes studied here, including saponins and
330 phenols. Compounds in the latter class (phenolics) were found to have strong positive and negative
331 effects on assemblages of arthropods associated with the maternal plants from which seeds were collected

332 to start the common garden used in the present study [18]. We have not attempted to separate constitutive
333 and induced defenses [39] as the plants in our common garden were exposed to natural and continuous
334 levels of herbivory. We have also focused on simple effects rather than interactions among compounds,
335 although some were detected (electronic supplementary material table S2), and interactions could
336 certainly add to the complexity of effects on different herbivores. Future studies in this system involving
337 greater numbers of plants will have greater power to test for robust interactive effects. In the meantime, it
338 is clear that metabolomic data such as that analyzed here has the potential to both open up new avenues of
339 conceptual development in plant-insect interactions and to link micro-evolutionary trajectories across
340 hosts and herbivores.
341

342 **Acknowledgements**

343 This work was supported by National Science Foundation grant DEB-1638793 to MLF and CDD, DEB-
344 1638768 to ZG, DEB-1638773 to CCN, DEB-1638922 to SL and JAF, and DEB-1638602 to CAB; MLF
345 was additionally supported by a Trevor James McMinn professorship. Thanks to Ian Wallace, the
346 Hitchcock Center for Chemical Ecology and the P.I.G. group for discussion and expertise.
347

348 **References**

349 1. Feeny P, Rosenthal GA, Berenbaum MR. 1992 The evolution of chemical ecology: contributions
350 from the study of herbivorous insects. *Herbiv. their Interact. with Second. plant Metab.* **2**, 1–44.
351 2. Agrawal AA, Petschenka G, Bingham RA, Weber MG, Rasmann S. 2012 Toxic cardenolides:
352 chemical ecology and coevolution of specialized plant--herbivore interactions. *New Phytol.* **194**,
353 28–45.
354 3. Dyer LA. 1995 Tasty generalists and nasty specialists - antipredator mechanisms in tropical
355 lepidopteran larvae. *Ecology* **76**, 1483–1496.
356 4. Olson JS. 1963 Energy storage and the balance of producers and decomposers in ecological
357 systems. *Ecology* **44**, 322–331.

358 5. Berenbaum M. 1983 Coumarins and caterpillars: a case for coevolution. *Evolution (N. Y).* **37**, 163–
359 179.
360 6. Zalucki MP, Brower LP, Alonso A. 2001 Detrimental effects of latex and cardiac glycosides on
361 survival and growth of first-instar monarch butterfly larvae *Danaus plexippus* feeding on the
362 sandhill milkweed Asclepias humistrata. *Ecol. Entomol.* **26**, 212–224.
363 7. Dyer LA *et al.* 2018 Modern approaches to study plant–insect interactions in chemical ecology.
364 *Nat. Rev. Chem.* **1**. (doi:10.1038/s41570-018-0009-7)
365 8. Maag D, Erb M, Glauser G. 2015 Metabolomics in plant-herbivore interactions: challenges and
366 applications. *Entomol. Exp. Appl.* **157**, 18–29.
367 9. Richards LA, Dyer LA, Smilanich AM, Dodson CD. 2010 Synergistic effects of amides from two
368 Piper species on generalist and specialist herbivores. *J. Chem. Ecol.* **36**, 1105–1113.
369 10. Sardans J, Penuelas J, Rivas-Ubach A. 2011 Ecological metabolomics: overview of current
370 developments and future challenges. *Chemoecology* **21**, 191–225.
371 11. Prince EK, Pohnert G. 2010 Searching for signals in the noise: metabolomics in chemical ecology.
372 *Anal. Bioanal. Chem.* **396**, 193–197.
373 12. Fordyce JA, Nice CC. 2008 Antagonistic, stage-specific selection on defensive chemical
374 sequestration in a toxic butterfly. *Evolution (N. Y).* **62**, 1610–1617. (doi:10.1111/j.1558-
375 5646.2008.00388.x)
376 13. Macel M, van Dam NM, Keurentjes JJB. 2010 Metabolomics: the chemistry between ecology and
377 genetics. *Mol. Ecol. Resour.* **10**, 583–593.
378 14. Erbilgin N, Ma C, Whitehouse C, Shan B, Najar A, Evenden M. 2014 Chemical similarity
379 between historical and novel host plants promotes range and host expansion of the mountain pine
380 beetle in a native host ecosystem. *New Phytol.* **201**, 940–950.
381 15. Chaturvedi S, Lucas LK, Nice CC, Fordyce JA, Forister ML, Gompert Z. 2018 The predictability
382 of genomic changes underlying a recent host shift in Melissa blue butterflies. *Mol. Ecol.*
383 16. Forister ML, Scholl CF, Jahner JP, Wilson JS, Fordyce JA, Gompert Z, Narala D, Buerkle CA,

384 Nice CC. 2012 Specificity, rank preference and the colonization of a non-native host plant by the
385 Melissa blue butterfly. *Oecologia* **DOI: 10.10**.
386 17. Gompert Z *et al.* 2015 The evolution of novel host use is unlikely to be constrained by trade-offs
387 or a lack of genetic variation. *Mol. Ecol.* **24**, 2777–2793.
388 18. Harrison JG *et al.* 2018 Deconstruction of a plant-arthropod community reveals influential plant
389 traits with nonlinear effects on arthropod assemblages. *Funct. Ecol.* **32**, 1317–1328.
390 19. Jorge TF, Mata AT, António C. 2016 Mass spectrometry as a quantitative tool in plant
391 metabolomics. *Phil. Trans. R. Soc. A* **374**, 20150370.
392 20. Ekanayaka EAP, Celiz MD, Jones AD. 2015 Relative mass defect filtering of mass spectra: a path
393 to discovery of plant specialized metabolites. *Plant Physiol.* **167**, 1221–1232.
394 (doi:10.1104/pp.114.251165)
395 21. Langfelder P, Horvath S. 2008 WGCNA: an R package for weighted correlation network analysis.
396 *BMC Bioinformatics* **9**, 559.
397 22. Ogutu JO, Schulz-Streeck T, Piepho H-P. 2012 Genomic selection using regularized linear
398 regression models: ridge regression, lasso, elastic net and their extensions. In *BMC proceedings*, p.
399 S10.
400 23. Team RDC, R Development Core Team R. 2016 R: A Language and Environment for Statistical
401 Computing. *R Found. Stat. Comput.* (doi:10.1007/978-3-540-74686-7)
402 24. Friedman J, Hastie T, Simon N, Tibshirani R. 2016 Lasso and Elastic-Net Regularized
403 Generalized Linear Models. R-package version 2.0-5. 2016.
404 25. Plummer M, others. 2003 JAGS: A program for analysis of Bayesian graphical models using
405 Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical*
406 *computing*,
407 26. Gelman A, Rubin DB, others. 1992 Inference from iterative simulation using multiple sequences.
408 *Stat. Sci.* **7**, 457–472.
409 27. Brooks SP, Gelman A. 1998 General methods for monitoring convergence of iterative simulations.

410 *J. Comput. Graph. Stat.* **7**, 434–455.
411 28. Delaney NJ, Chatterjee S. 1986 Use of the bootstrap and cross-validation in ridge regression. *J.*
412 *Bus. Econ. Stat.* **4**, 255–262.
413 29. Levin DA. 1976 The chemical defenses of plants to pathogens and herbivores. *Annu. Rev. Ecol.*
414 *Syst.* **7**, 121–159.
415 30. Hunter MD. 2016 *The phytochemical landscape: linking trophic interactions and nutrient*
416 *dynamics*. Princeton University Press.
417 31. Forister ML, Nice CC, Fordyce JA, Gompert Z. 2009 Host range evolution is not driven by the
418 optimization of larval performance: the case of Lycaeides melissa (Lepidoptera: Lycaenidae) and
419 the colonization of alfalfa. *Oecologia* **160**, 551–561. (doi:10.1007/s00442-009-1310-4)
420 32. Seigler D, Price PW. 1976 Secondary compounds in plants: primary functions. *Am. Nat.* **110**, 101–
421 105.
422 33. Smilanich AM, Fincher RM, Dyer LA. 2016 Does plant apparency matter? Thirty years of data
423 provide limited support but reveal clear patterns of the effects of plant chemistry on herbivores.
424 *New Phytol.* **210**, 1044–1057.
425 34. Chaieb I. 2010 Saponins as insecticides: a review. *Tunis. J. Plant Prot.* **5**, 39–50.
426 35. Behmer ST. 2009 Insect herbivore nutrient regulation. *Annu. Rev. Entomol.* **54**.
427 36. Berenbaum MR. 1995 Turnabout is fair play - secondary roles for primary compounds. *J. Chem.*
428 *Ecol.* **21**, 925–940.
429 37. Stamp N. 2003 Out of the quagmire of plant defense hypotheses. *Q. Rev. Biol.* **78**, 23–55.
430 38. Harrison JG *et al.* 2016 The many dimensions of diet breadth: phytochemical, genetic, behavioral,
431 and physiological perspectives on the interaction between a native herbivore and an exotic host.
432 *PLoS One* **11**, e0147971.
433 39. Jansen JJ, Allwood JW, Marsden-Edwards E, van der Putten WH, Goodacre R, van Dam NM.
434 2009 Metabolomic analysis of the interaction between plants and herbivores. *Metabolomics* **5**,
435 150.

Author contributions:

MLF: designed experiment, conducted analyses, wrote first draft.

SY: conducted experiment and contributed to experimental design.

CSP, CDD, BH: generated and interpreted phytochemistry and protein data.

MLF, JGH, OS: developed and maintained common garden.

JAF, ZHM, CCN, LAR: contributed to analyses and experimental design.

CAB, JAF, ZG, SL, CCN: contributed to experimental design.

All authors: contributed to writing.

Figure legends (also reproduced below individual figures)

Figure 1. Variation among plants in caterpillar survival (a), development time (b) adult weight (c), three individual compounds (d-e) and two external plant traits, specific leaf area (g) and leaf toughness (h). The three example compounds shown here (out of the 163 assayed) were among the top five most influential compounds for survival, development time and adult weight: cpd. 9 is a peptide with a negative association with survival, cpd. 94 (another peptide) has a negative association with development time, and cpd. 160 is a phospholipid with a negative association with adult weight. Individual plants in all panels are organized from left to right by decreasing caterpillar survival in the top panel (a). Standard errors are shown for panels b, c, g and h. The units for d-e are compound relative abundance per dry weight of sample; the units for specific leaf area are cm^2/mg, and grams/newton for leaf toughness.

Figure 2. Illustration of correlational structure among compounds: each node in the network is a compound, and compounds are linked by a line if they are correlated among individual plants at 0.5 or above (links among compounds in modules 12-14 represent weaker correlations, greater than 0.1; see main text for details). Two letter codes within nodes indicate compound classes, as explained in the legend. Colors of nodes correspond to membership in modules as determined by hierarchical cluster analysis; the color key to the 14 modules is shown in the lower left. Not shown are a small number of compounds with weak connections to all other compounds, including two compounds that were not included in any module (shown as module zero in figure 3).

Figure 3. Effects of individual compounds on survival, development time and adult weight, as estimated by ridge regression (using binomial, Poisson and Gaussian models, respectively). The strength of effect for each compound is indicated by the horizontal extent of each bar, and compounds are grouped by modules (m1, m2, etc.); the order of compounds along the vertical axis is arbitrary within modules and fixed across columns. Orange colors indicate negative effects on survival, development and weight, while blue colors are positive effects. The darker shades of orange and blue mark coefficients whose 95% confidence intervals did not overlap zero in 1,000 bootstrap samples. Values for survival and development time have been back-transformed from units on the log-odds and log scales to units of probability and days to pupation, and are shown as changes from the mean or intercept values. For example, a negative (orange) survival coefficient of 0.005 means a one-half percent reduction in average probability of survival associated with variation in a particular compound. The ten compounds with the largest coefficients (by absolute value) and bootstrap intervals not overlapping zero are labelled by compound classes (see figure 2 for abbreviations) in each panel.

Figure 4. Violin plots of compound-specific effects (coefficients from ridge regressions) summarized by chemical classes. Plots show median (black dot), interquartile range (box) and 95% confidence intervals (whiskers) surrounded by kernel density envelopes. Sample sizes for each category as follows: l5 lipids, 17 terpenoids, 24 saponins, 48 peptides, 43 phospholipids, 7 phenolics and 9 other. Categories are as shown in electronic supplementary material table S1, with the exception of terpenoids (which is shown here as a pooled category of 5 sterols, 5 vitamins and 7 carotenoids) and "other" (which is 3 alkaloids, 2 amino acids, 1 halogenated compound and 3 sugars). Categories are arranged from left to right based on the gradient of median positive to negative effects on survival. Coefficients for survival (a) and development time (b) have been back-transformed from the units of log-odds and log to probability and days to pupation, respectively, and shown as deviations from the mean or intercept value (as in figure 3).

Table 1. Results from Bayesian regressions of module eigenvectors and covariates predicting caterpillar survival, development time and adult weight (as binomial, Poisson, and Gaussian regressions, respectively, with corresponding units in log-odds, log number of days, and milligrams). For each regression coefficient, numbers in parentheses are 95% credible intervals (the first two numbers) and the probability that the coefficient has the estimated sign (e.g., 0.63 for the m2 survival coefficient of 0.37 indicates a 63% probability that the m2 module has a positive effect on survival). Note that negative coefficients for development time indicate faster caterpillar development (fewer days) associated with variation in a particular compound. Modules (listed in the left column) are only shown if they were included in one of the three regressions following feature selection using lasso regression (see main text for additional details). Empty spaces in the table appear if a particular module was selected through lasso regression for one or two analyses but not all three (m3, for example, was not selected by lasso regression for development time). Slash marks (/) indicate variables not considered for a particular analysis (e.g., sex, adult weight [mg] and development time [days] were not possible for the survival analysis because they are not observed on dead individuals). Values for "validation" shown in the last row are the correlation between observed and predicted values in a cross-validation analysis (electronic supplementary material figure S4).

	Survival coefficient (CI; prob.)		Development time coefficient (CI, prob.)		Weight coefficient (CI, prob.)	
m2	0.37	(-1.87, 2.67; 0.63)	-0.07	(-0.26, 0.12; 0.78)		
m3	-2.33	(-3.94, -0.72; >0.99)			-2.33	(-4.87, 0.24; 0.96)
m6			-0.05	(-0.19, 0.09; 0.75)		
m9	-2.31	(-4.49, -0.15; 0.98)	0.064	(-0.11, 0.24, 0.78)		
m10	2.54	(0.81, 4.23; >0.99)			3.56	(1.22, 5.84; >0.99)
m11	2.01	(0.64, 3.42; >0.99)				
SLA	-0.31	(-0.55, -0.08; >0.99)			-0.42	(-0.78, -0.06; 0.99)
Protein	0.061	(-0.16, 0.28; 0.71)			0.08	(-0.26, 0.42; 0.68)
Tough.	0.036	(-0.17, 0.24; 0.36)	-0.002	(-0.02, 0.02; 0.57)		
Sex	/		0.06	(0.02, 0.10; >0.99)	1.10	(0.38, 1.83; >0.99)
mg	/		-0.03	(-0.05, -0.01; >0.99)	/	
Days	/		/		1.10	(0.38, 1.83; >0.99)
Intercept	0.34	(0.14, 0.53; >0.99)	3.48	(3.45, 3.52; >0.99)	10.36	(9.81, 10.92; >0.99)
Validation	0.53		0.59		0.50	

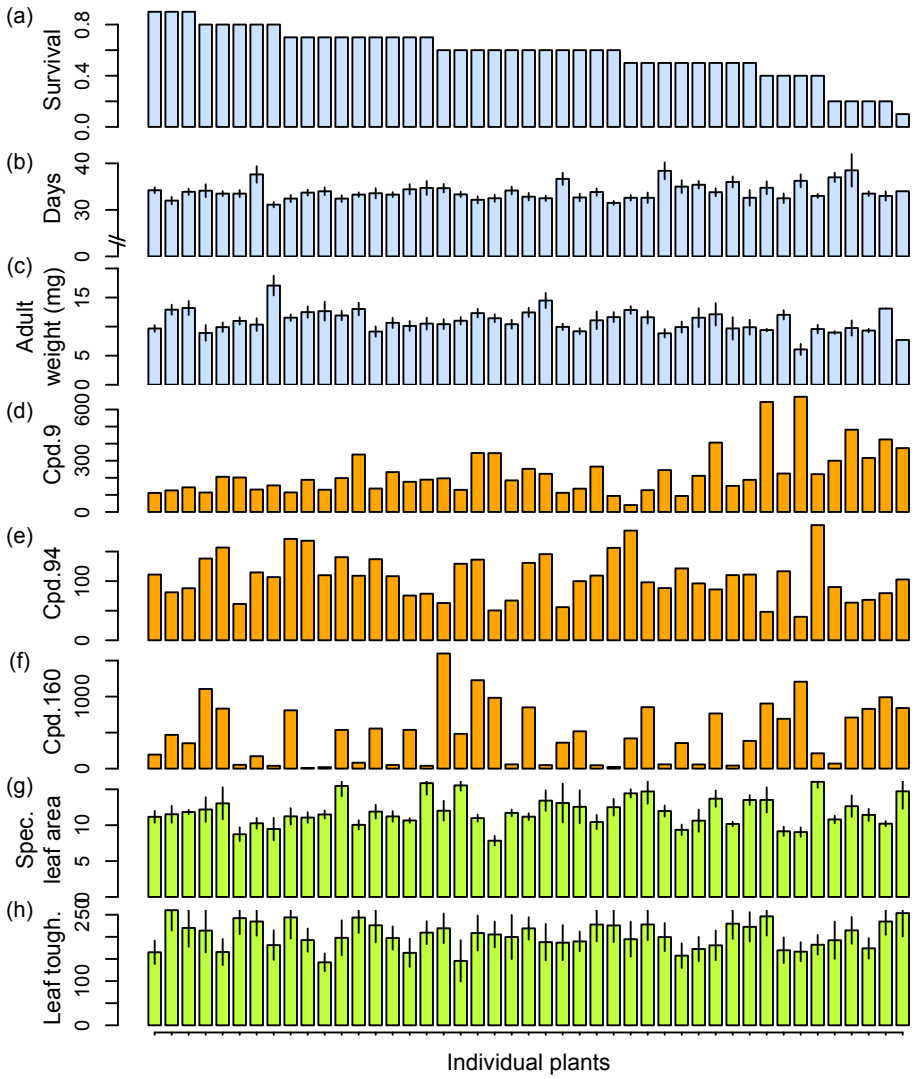

Figure 1. Variation among plants in caterpillar survival (a), development time (b) adult weight (c), three individual compounds (d-e) and two external plant traits, specific leaf area (g) and leaf toughness (h). The three example compounds shown here (out of the 163 assayed) were among the top five most influential compounds for survival, development time and adult weight: cpd. 9 is a peptide with a negative association with survival, cpd. 94 (another peptide) has a negative association with development time, and cpd. 160 is a phospholipid with a negative association with adult weight. Individual plants in all panels are organized from left to right by decreasing caterpillar survival in the top panel (a). Standard errors are shown for panels b, c, g and h. The units for d-e are compound relative abundance per dry weight of sample; the units for specific leaf area are cm^2/mg, and grams/newton for leaf toughness.

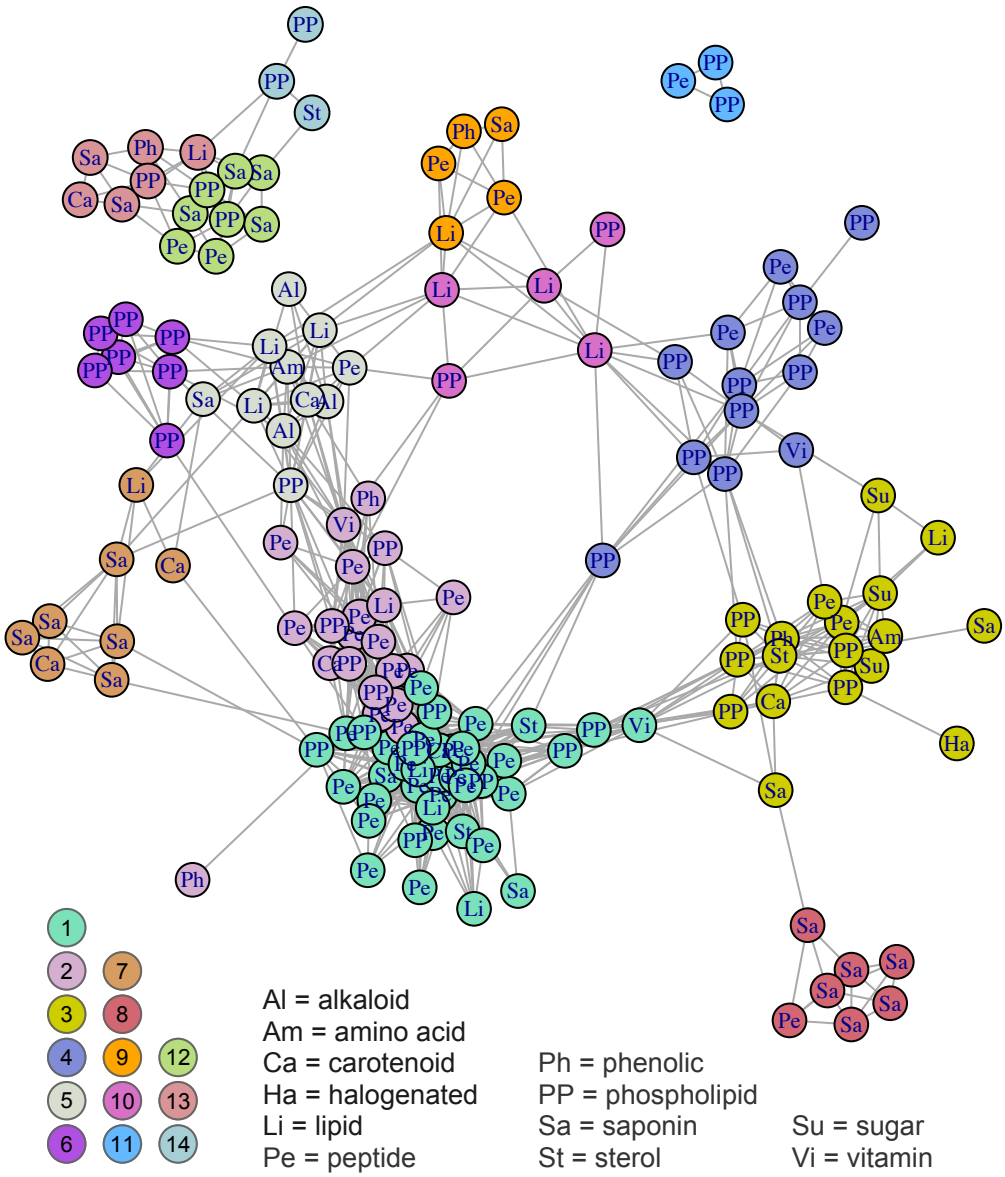

Figure 2. Illustration of correlational structure among compounds: each node in the network is a compound, and compounds are linked by a line if they are correlated among individual plants at 0.5 or above (links among compounds in modules 12-14 represent weaker correlations, greater than 0.1; see main text for details). Two letter codes within nodes indicate compound classes, as explained in the legend. Colors of nodes correspond to membership in modules as determined by hierarchical cluster analysis; the color key to the 14 modules is shown in the lower left. Not shown are a small number of compounds with weak connections to all other compounds, including two compounds that were not included in any module (shown as module zero in figure 3).

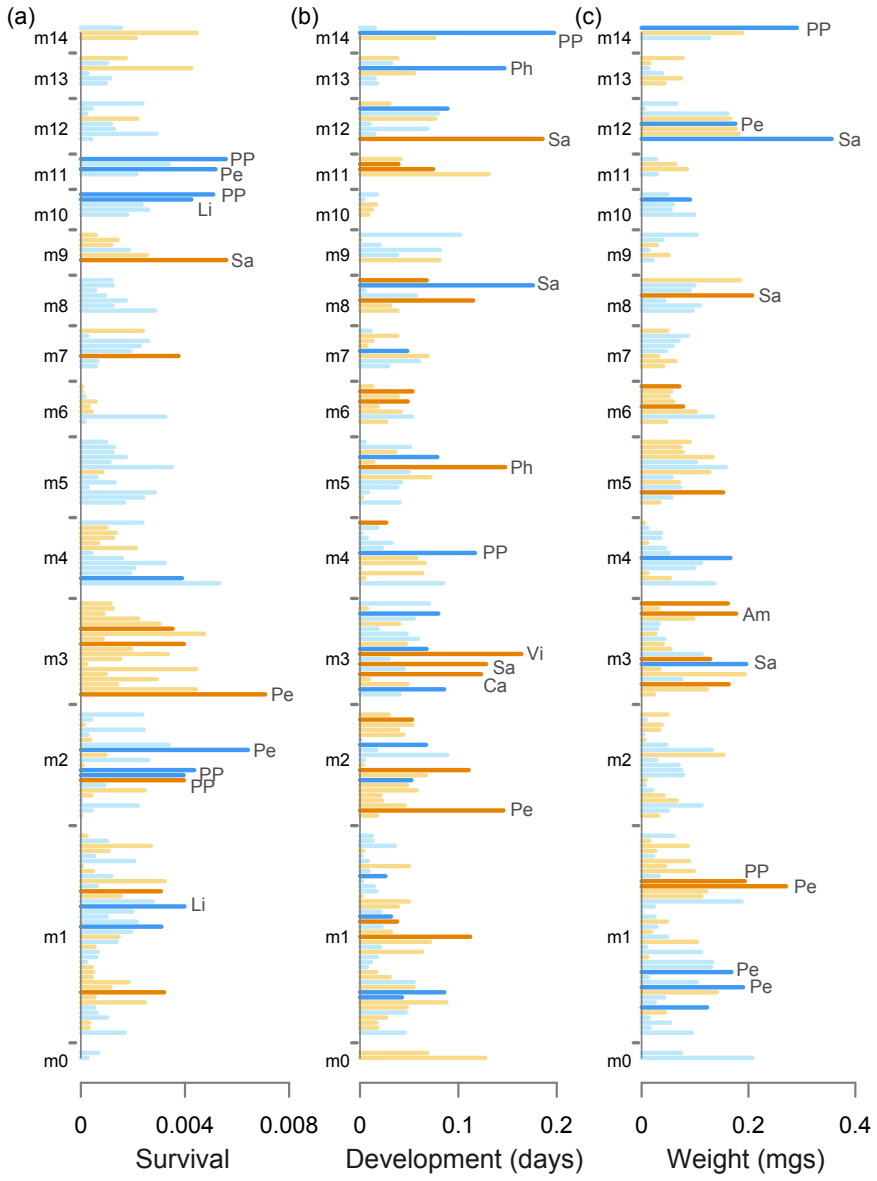

Figure 3. Effects of individual compounds on survival, development time and adult weight, as estimated by ridge regression (using binomial, Poisson and Gaussian models, respectively). The strength of effect for each compound is indicated by the horizontal extent of each bar, and compounds are grouped by modules (m1, m2, etc.); the order of compounds along the vertical axis is arbitrary within modules and fixed across columns. Orange colors indicate negative effects on survival, development and weight, while blue colors are positive effects. The darker shades of orange and blue mark coefficients whose 95% confidence intervals did not overlap zero in 1,000 bootstrap samples. Values for survival and development time have been back-transformed from units on the log-odds and log scales to units of probability and days to pupation, and are shown as changes from the mean or intercept values. For example, a negative (orange) survival coefficient of 0.005 means a one-half percent reduction in average probability of survival associated with variation in a particular compound. The ten compounds with the largest coefficients (by absolute value) and bootstrap intervals not overlapping zero are labelled by compound classes (see figure 2 for abbreviations) in each panel.

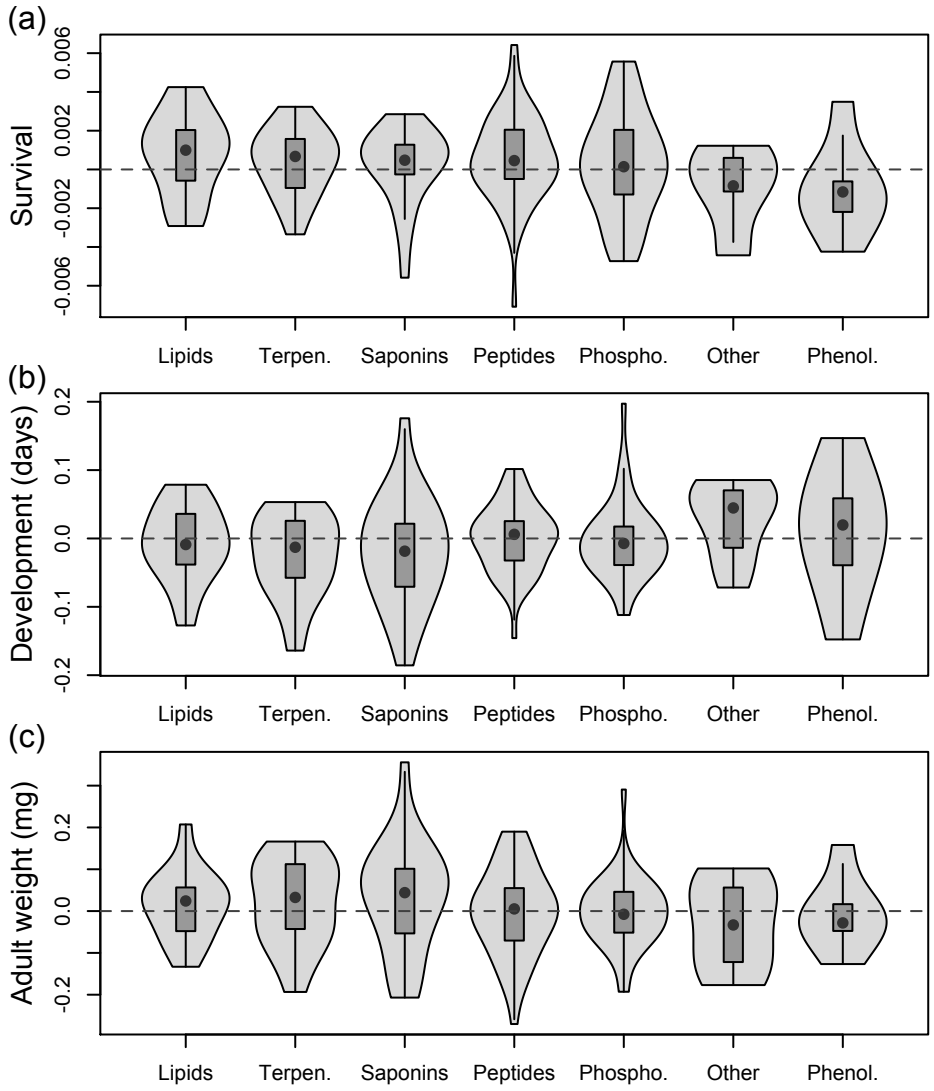

Figure 4. Violin plots of compound-specific effects (coefficients from ridge regressions) summarized by chemical classes. Plots show median (black dot), interquartile range (box) and 95% confidence intervals (whiskers) surrounded by kernel density envelopes. Sample sizes for each category as follows: l5 lipids, 17 terpenoids, 24 saponins, 48 peptides, 43 phospholipids, 7 phenolics and 9 other. Categories are as shown in electronic supplementary material table S1, with the exception of terpenoids (which is shown here as a pooled category of 5 sterols, 5 vitamins and 7 carotenoids) and "other" (which is 3 alkaloids, 2 amino acids, 1 halogenated compound and 3 sugars). Categories are arranged from left to right based on the gradient of median positive to negative effects on survival. Coefficients for survival (a) and development time (b) have been back-transformed from the units of log-odds and log to probability and days to pupation, respectively, and shown as deviations from the mean or intercept value (as in figure 3).