Habitat structure and phenotypic variation in the invading butterfly *Coenonympha tullia*

Richard A. Lent

Updated 30 November 2016

Preamble

This is a mockup of a scientific paper (although the field data are real) to illustrate the use of R Notebooks as a means of creating reproducible research. The objective is to produce publication-quality output, in HTML, PDF, and Microsoft Word formats, with text, literature citations, a formatted bibliography, statistical analyses, tables, and graphics, all from one, plain-text R Notebook. It illustrates the basic components for creating reproducible research using the tools available in R and RStudio.

The R Notebook file sites.Rmd that produced the rendered document you are now reading contains all of the R code that produced the accompanying analysis and graphics. Within the R code chunks embedded in the text of the notebook, hash symbols (`#`) are used to denote comments. Comments are ignored by the R interpreter and are used to document the code. Thus the main text of the notebook describes the research, using the format of a traditional scientific paper, and the embedded R code and comments document how the analyses and graphics were produced. Additional comments in the main text are placed inside of HTML comment tags (`<!-- -->`). An additional file, sessioninfo.txt, was produced using the R `sessionInfo()` function to document the R environment (R version, computer platform and operating system, and installed R packages) used to produce this document, the analysis, and graphics. By storing the raw data file (sites.csv) along with a companion metadata file (sites.metadata.txt) together with the R Notebook and `sessioninfo.txt`, we can provide other researchers with all of the materials required to replicate our research.

A zip archive that packages all of the above-mentioned files, along with publication-quality documents rendered from `sites.Rmd` in PDF and Microsoft Word format, can be downloaded from here. When extracted, the zip archive will produce a folder named `sites` that contains all of the data, code, and documentation required to replicate the analyses and produce the graphics. It is thus a self-contained package of reproducible research.

An HTML notebook (sites.nb.html) is automatically created whenever the R Notebook `sites.Rmd` is modified and saved in RStudio, and is the usual way to share reproducible research documents created with RStudio. The HTML notebook contains the rendered HTML version of the R Markdown file sites.Rmd plus all R output, graphics, code, and a downloadable copy of `sites.Rmd`

What follows is a presentation of the research in standard scientific format, followed by some closing comments in the Epilogue.

Introduction

The satyrid butterfly *Coenonympha tullia*, the Common Ringlet, has recently expanded its range southward into New England.[1] Formerly restricted to the north and western United States and Canada, the Ringlet is now found throughout much of New England.[2] Its range continues to expand in the eastern United States.[3] A small butterfly, with maximum wingspan around 1.5 inches, Ringlets are quite phenotypically variable (Figure 1), particularly in background color and in the size and number of eyespots and other markings on the wing surfaces.[4–6] This phenotypic variation, in combination with the range expansion, suggests the possibility of segregation by habitat of different phenotypes, with resultant variation in the forces of natural selection.[1,7] Because the Ringlet is a weak flier and is specific in its habitat requirements, selecting early-successional fields and pastures containing certain species of grasses (e.g., *Festuca*, *Poa pratensis*, *Stipa*, *Agrostis*), it exists

as a metapopulation, consisting of many locally isolated colonies.[8] The Ringlet system is thus ideally suited for field studies in population ecology and evolutionary biology.

Methods

In summer 1994 we[9–11] visited 11 grassland sites located from north-central Massachusetts to the Canadian border (Figure 2). Habitat structure was measured at each site along a 50-meter transect running through the center of the site (Figure 3). Habitat structure variables are shown in Table 1. Values for the eight habitat variables are the percentage of 50 transect points at which each habitat type (grass, moss, sedge, forb, fern, bare ground, woody herbaceous plants, and shrub stems) was present. A habitat type was considered present at a transect point if it touched a thin metal rod passed down vertically through the vegetation at the point.[12]

Specimens of male Common Ringlets were collected from each site and their phenotypes measured (Table 2). The five phenotypic variables in Table 2 are length of right forewing measured from insertion on the thorax to the wing apex (fwlength, mm), length of thorax (thorax, mm), number of eyespots or ocelli on the undersides of the right fore- and hindwings (spotting, count), length of right forewing ray (rfray, ordinal 0 through 3, where 0 codes for absent and 1 - 3 code for small, medium, and large), and length of right hindwing ray (rhray, ordinal 0 through 3). The final variable, n, is the number of male Ringlets measured at each site. Values of the phenotypic variables in Table 2 are means for each site calculated across the n individuals. Ocelli, forewing rays (linear, light-colored patches), and hindwing rays are on the ventral wing surfaces and are shown in Figure 1.

Two Euclidean distance matrices were computed from the field data, based on habitat structure variables (Table 1) and butterfly phenotype variables (Table 2). A third Euclidean distance matrix contained geographic distances between pairs of sites, computed using the latitude and longitude of each site in UTM coordinates (meters). (When Euclidean distance is calculated using spatial coordinates it yields the approximate geographic distance between sites.) A distance matrix expresses dissimilaries between all pairs of objects (in this case, the 11 sites) using a numeric distance measure based on several variables. Euclidean distance is one of several such distance measures[13,14] and is the simple straight-line distance between two points in a Cartesian coordinate system. Euclidean distance can be extended to an n-dimensional space in which each of the n dimensions is a variable used to compute the distance matrix. Pairs of sites that have similar values of all vegetation or phenotypic variables have smaller values of Euclidean distance, and are thus more similar to each other in a multivariate sense, as if they were plotted in an n-dimensional space where n is the number of variables. Euclidean distance ranges from zero for identical sites (i.e., along the diagonal of the distance matrix, the distance of each site to itself) to an arbitrary upper bound.[14]

Hierarchical cluster analysis was performed on the habitat and phenotype distance matrices. Cluster analysis is a technique for finding groups in multivariate data. The technique produces a tree diagram, or dendrogram, that groups similar objects together and shows the distance relationships among objects.[13,15]

We examined between-site relationships in habitat structure, butterfly phenotype, and geographic distance using the Mantel test[14,16] for matrix correlation. Matrix correlation can be quantified by computing the Pearson correlation coefficient between the elements of two distance matrices. But because of the statistical dependence among the cells of a distance matrix, ordinary tests of significance cannot be used. Instead, the Mantel test uses random permutations of the rows and columns of the distance matrices. This is done a large number of times (10,000 in our analyses), and the correlation coefficient calculated each time, forming a sampling distribution. The observed matrix correlation is then compared with the sampling distribution to assess the significance of the observed correlation under the assumption of randomness.

We tested for matrix correlations between (1) habitat and phenotype, (2) habitat and geographic distance, and (3) phenotype and geographic distance. The questions[17] addressed with these tests were:

(1) Do sites that are similar in terms of their habitat structure variables also tend to be similar in terms of butterfly phenotype variables?

(2) Are sites that are closer together also similar in habitat structure?

(3) Are sites that are closer together also similar in butterfly phenotype?

All computations, graphics, and document preparation were done using R and RStudio.[18,19]

Results and Discussion

Because sites were initially chosen to be predominantly herbaceous, we dropped habitat variable `grass` from the analysis because it did not vary much among sites (Table 1). The 11 sites ranged from 46 to 50% grass with varying amounts of forbs.

Three site clusters were recognized based on habitat structure (Table 3 and Figure 4). Cluster 1 (Orleans, Bellows Falls, Springfield, Bernardston) contained sites that were shrubby and woody. Sites in cluster 2 (Mount Ascutney, Lyme, Fairlee, Lyndonville) had the lowest percentage occurrence of forbs. Moss, sedges, bare ground, and shrubs were absent at the sites in Cluster 3 (Hanover, Newport, St. Johnsbury). The habitat structure clustering was broadly related to geographic location of the sites (Figure 2). Cluster 1 sites were mostly southern, cluster 3 sites were mostly northern, and sites in cluster 2 were somewhat in the middle.

Three site clusters were also apparent based on butterfly phenotype (Table 4 and Figure 5). Cluster 1 (Bellows Falls, Lyme) had the smallest butterflies in terms of forewing and thorax length, with more ocelli but shorter fore- and hindwing rays (Table 4). Ringlets in cluster 3 (Bernardston, Newport) had the largest forewings. Butterflies in cluster 2 (St. Johnsbury, Springfield, Fairlee, Hanover, Mount Ascutney, Orleans, Lyndonville) had an intermediate forewing length and the longest rays on both forewings and hindwings. Clusters 1 and 3 each contained only two sites and differed from each other mainly in Ringlet body size. In contrast, sites in the large cluster 2 had Ringlets with characters intermediate to those of clusters 1 and 3. Clusters 1 and 3 can be viewed as outlying clusters that differ from each other in addition to being different from the central cluster 2 containing the majority of sites. In contrast to the habitat clustering results, no strong relationship was apparent with regard to butterfly clusters and site location (Figures 2 and 5).

While these patterns in the clustering of sites by habitat structure and butterfly phenotype are interesting, Mantel tests showed that the underlying matrix correlations were all nonsignificant (Figure 5). The observed matrix correlations of r = 0.199 for phenotype vs distance, r = -0.061 for habitat vs distance, and r = -0.25 for phenotype vs habitat were indistinguishable from randomly-generated values.

The most likely explanation for the lack of a relationship between site habitat structure and butterfly phenotype is that we measured the wrong variables, or that we measured habitat and butterfly variables at the wrong scale. For example, Dennis and Eales[20] classified 166 UK sites according to occupancy by *Coenonympha tullia* (simple presence-absence) and found that habitat quality, size of habitat patch, and patch isolation accounted for 61% of the variation in *C. tullia* occupancy of sites. Their 23 measures of habitat quality were much more detailed and species-specific than our eight simple measures of habitat structure.

Habitat variables thus may have a greater influence on whether or not Ringlets are present at a site, rather than being related to phenotypic characteristics of the local population. Morphological variation in Ringlets, for example thorax size, may be related to factors such as flight activity in differing thermal environments.[21] Wing size, ocellation, and size of rays and other wing markings have been shown to be influenced by opposing forces of natural selection favoring crypticity versus predator avoidance.[22]

Our data do suggest the presence of some heterogeneity in Ringlet phenotypes among the 11 sites that were sampled (Table 4, Figure 5). This agrees with the findings of Wiernasz,[1] who showed that range expansion in *C. tullia* was associated with significant increases in morphological variation and changes in other life history parameters. Such variation, coupled with small population sizes occupying isolated habitat patches, may result in increased genetic differentiation among local populations.[23]

Epilogue

In this example of using RStudio for creating reproducible research, everything was kept in a single R Notebook. Our intent was to show the basic pieces needed in order to produce reproducible research using the

tools available in R and RStudio (e.g., R Markdown, code chunks, BibTeX). A more complex project would require a more complex organizational structure. As discussed in detail by Gandrud,[24] this might entail a separate folder for the raw data and metadata files, one for R scripts that process the data, a folder that has R scripts for doing the analysis, one for graphics, one for production of the final documents, a plain-text `README` file to explain how everything fits together, and so on. Having all R code in a single R notebook, as we have done in this example, would probably only be suitable for a small project. Moreover, to control the length and complexity of this example we did not provide the original, raw field data collected on the habitat structure and morphology of individual butterflies that was processed and averaged to produce Tables 1 and 2. Truly reproducible research should provide all of the original data along with all of the computer code used to process and reduce the data for the final results. We also could have included some exploratory data analysis,[25] but in the interest of space we did not include this important step (see Introduction to R).

Creating reproducible research is good not only for the advancement of science, but is also good for the individual researcher. Gandrud[24] lists the following benefits of reproducible research:

- It creates a standard to judge scientific claims
- Avoids duplication of effort
- Encourages the development of cumulative knowledge
- Creates better work habits (and helps you remember what you did 10 years ago)
- Fosters teamwork and collaboration
- Facilitates changes in the analysis and document revisions
- Reproducible research is more useful to other researchers and has a greater impact

I hope that this document encourages you to explore the free tools provided by R and RStudio and to use them to generate your own brand of plain-text, cross-platform, future-proof reproducible research.

Table 1. Habitat structure variables.

locality	code	grass	moss	sedge	forb	fern	bare	woody	shrub
Bellows Falls	BF	46	4	0	34	0	4	10	0
Springfield	SP	49	0	0	24	1	0	14	0
Mount Ascutney	MA	49	0	4	17	5	0	0	0
Bernardston	BE	49	1	0	29	0	0	11	7
Lyme	LY	50	0	0	14	4	0	0	0
Fairlee	FA	50	5	0	12	0	2	0	0
Newport	NW	49	0	0	35	0	0	0	0
Orleans	OR	49	0	0	41	0	0	19	0
Lyndonville	LN	50	0	0	9	0	0	3	0
St. Johnsbury	SJ	49	0	0	42	0	0	0	0
Hanover	HA	49	0	0	38	5	0	5	0

Table 2. Butterfly phenotypic variables (means).

locality	code	fwlength	thorax	spotting	rfray	rhray	n
Bellows Falls	BF	15.8	3.6	1.6	1.6	1.4	62
Springfield	SP	16.0	3.7	1.1	2.0	1.5	51
Mount Ascutney	MA	16.2	3.8	1.2	1.9	1.6	58
Bernardston	BE	16.7	3.7	1.4	1.8	1.6	27
Lyme	LY	15.6	3.6	1.1	1.8	1.3	52
Fairlee	FA	15.9	3.8	1.0	1.8	1.5	25
Newport	NW	16.7	3.7	1.1	1.8	1.3	21
Orleans	OR	16.1	3.8	1.4	1.9	1.7	41
Lyndonville	LN	16.0	3.7	1.4	1.9	1.7	52
St. Johnsbury	SJ	15.7	3.6	1.2	1.9	1.7	56
Hanover	HA	16.0	3.6	1.2	1.8	1.8	65

Table 3. Means of site habitat variables by cluster; see Figure 4.

Cluster	moss	sedge	forb	fern	bare	woody	shrub
1	1.1	0	32.0	0.4	1.1	13.6	1.8
2	1.2	1	13.0	2.2	0.5	0.8	0.0
3	0.0	0	38.3	1.7	0.0	1.7	0.0

Table 4. Means of site phenotype variables by cluster; see Figure 5.

Cluster	fwlength	thorax	spotting	rfray	rhray
1	15.7	3.6	1.4	1.7	1.4
2	16.0	3.7	1.2	1.9	1.6
3	16.7	3.7	1.2	1.8	1.5

Literature Cited

1. Wiernasz, D. C. Ecological and genetic correlates of range expansion in Coenonympha tullia. *Biological Journal of the Linnean Society* **38,** 197–214 (1989).

2. Brown, F. M. Coenomympha tullia on islands in the St. Lawrence River. *Can. Entomol.* **93,** 107–117 (1961).

3. Butterflies and moths of north america: Attributes of coenonympha tullia. (2016). Available at: http://www.butterfliesandmoths.org/species/coenonympha-tullia. (Accessed: 29th November 2016)

4. Brown, F. M. Studies of Nearctic Coenonympha tullia (Rhopalocera, Satyridae): Coenonympha tullia inornata Edwards. *Bull. Amer. Mus. Nat. Hist.* **105,** 359–410 (1955).

5. Howe, W. H. *The butterflies of North America.* (Doubleday; Company, Inc., 1975).

6. Scott, J. A. *The butterflies of North America: A natural history and field guide.* (Stanford University

Figure 1. Ringlet specimens collected from a single site, showing ventral wing markings measured in the field, and also the range of phenotypic variation among individuals.

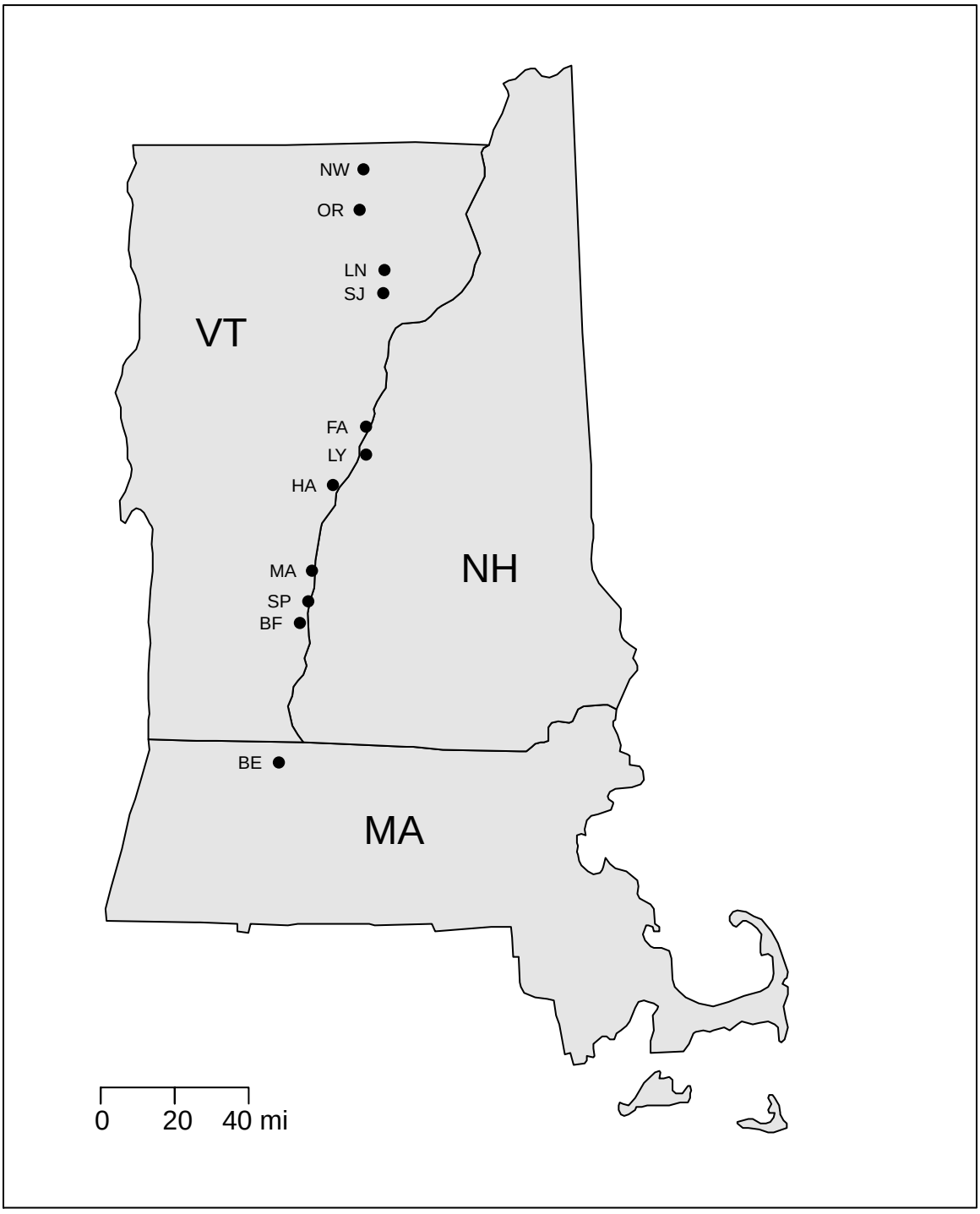

Figure 2. Locations of Study Sites. From south to north, sites are Bernardston, Bellows Falls, Springfield, Mount Ascutney, Hanover, Lyme, Fairlee, St. Johnsbury, Lyndonville, Orleans, and Newport.

Figure 3. Habitat structure was measured along a 50-meter transect running through the center of each site.

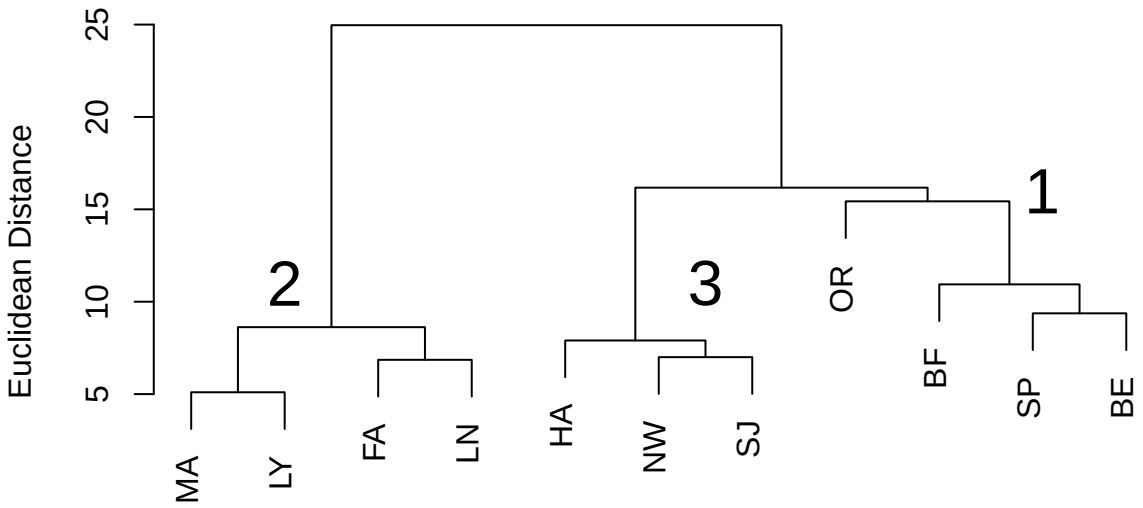

Figure 4. Clustering of sites by habitat structure. Site codes are in Table 1. Cluster numbers are referenced in Table 3.

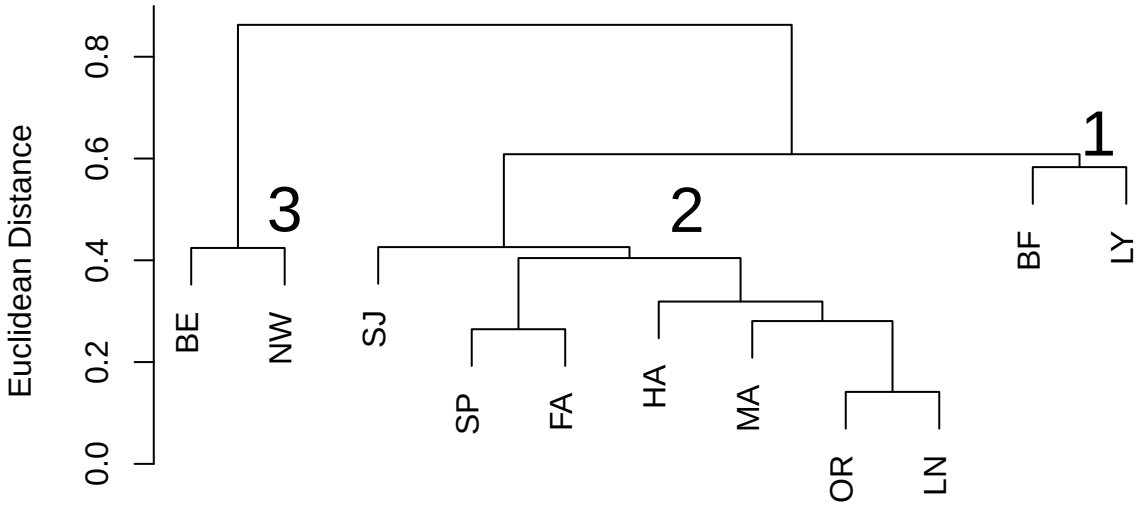

Figure 5. Clustering of sites by butterfly phenotype. Site codes are in Table 1. Cluster numbers are referenced in Table 4.

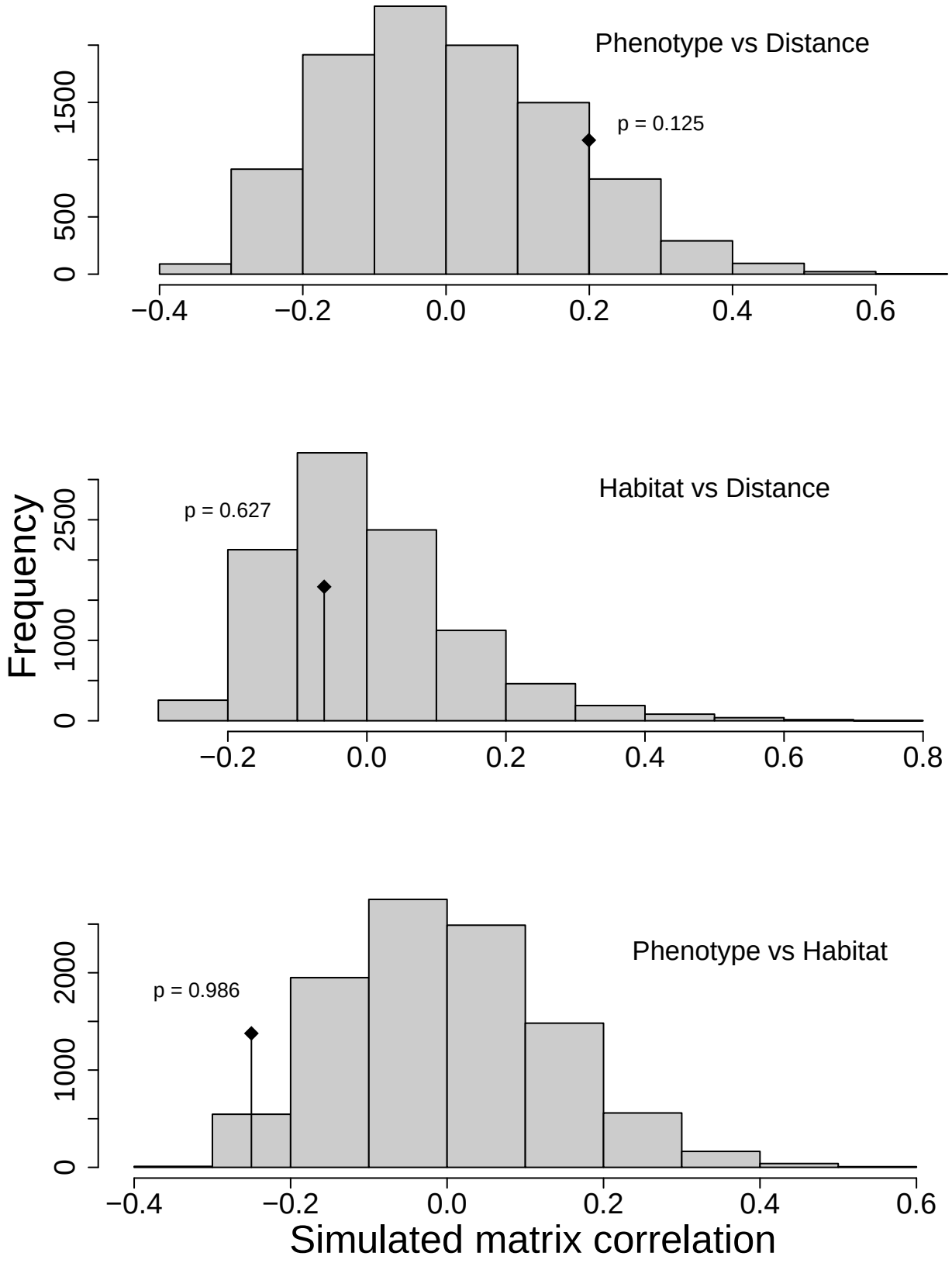

Figure 6. Mantel tests of distance matrix correlation. Diamond symbol indicates the observed matrix correlation, each with its associated p-value based on 10,000 random permutations.

Press, 1986).

7. Endler, J. A. *Natural selection in the wild.* (Princeton University Press, 1986).

8. Hanski, I. *Metapopulation Ecology.* (Oxford University Press, 1999).

9. Chen, J. C. An Analysis of Phenotypic Variation in New England Populations of Cercyonis pegala in Relation to Environmental Variables. (Amherst College, 1995).

10. Rob, C. M. Thermoregulatory Behavior and Phenotypic Variation in Two Massachusetts Satyrid Butterflies: Coenonympha inornata and Cercyonis pegala. (Amherst College, 1997).

11. Schumann, M. E. Geographic, Environmental, and Genetic Components of Phenotypic Variation in New England Populations of Coenonympha tullia (Lepidoptera: Satyridae). (Amherst College, 1995).

12. Greig-Smith, P. *Quantitative plant ecology.* (University of California Press, 1983).

13. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: An introduction to cluster analysis.* (Wiley, 2005).

14. Manly, B. F. J. *Multivariate statistical methods: A primer.* (Chapman & Hall/CRC Press, 2005).

15. Romesburg, H. C. *Cluster analysis for researchers.* (Lulu Press, 2004).

16. Manly, B. F. J. *Randomization and Monte Carlo methods in biology.* **1st,** (Chapman; Hall, 1991).

17. Urban, D. L. *Spatial Analysis in Ecology: Mantel's Test.* (Nat. Center for Ecological Analysis & Synthesis, 2003).

18. R Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2016).

19. RStudio Team. *RStudio: Integrated Development Environment for R.* (RStudio, Inc., 2016).

20. Dennis, R. L. H. & Eales, H. T. Patch occupancy in Coenonympha tullia (Muller, 1764) (Lepidoptera: Satyrinae): Habitat quality matters as much as patch size and isolation. *Journal of Insect Conservation* **1,** 10 (1997).

21. Heinrich, B. Thermoregulation and flight activity of a satyrine, Coenonympha inornata (Lepidoptera: Satyridae). *Ecology* **67,** 593–597 (1986).

22. Dennis, R. L. H., Porter, K. & Williams, W. R. Ocellation in Coenonympha tullia (Lepidoptera: Satyridae) II. Population differentiation and clinal variation in the context of climatically-induced antipredator defense strategies. *Entomol. Gaz.* **37,** 133–172 (1986).

23. Habel, J. C., Rödder, D., Lens, L. & Schmitt, T. The genetic signature of ecologically different grassland lepidopterans. *Biodiversity and Conservation* **22,** 2401–2411 (2013).

24. Gandrud, C. *Reproducible research with r and RStudio.* (CRC Press, Taylor & Francis Group, 2015).

25. Tukey, J. W. *Exploratory data analysis.* (Addison-Wesley, 1977).