Received: 11 January 2023 | Accepted: 23 May 2023

DOI: 10.1002/aps3.11546

APPLICATION ARTICLE

Applications in Plant Sciences

A pipeline for the rapid collection of color data from photographs

Yvonne Luong[1] | Ariel Gasca-Herrera[1] | Tracy M. Misiewicz[2] | Benjamin E. Carter[1]

[1]Biological Sciences, San Jose State University, San Jose, California 95192, USA

[2]University and Jepson Herbaria, University of California, Berkeley, California 94720, USA

Correspondence

Benjamin E. Carter, Biological Sciences, San Jose State University, One Washington Square, San Jose, California 95192, USA.
Email: benjamin.carter@sjsu.edu

This article is part of the special issue "Advances in Plant Imaging across Scales."

Abstract

Premise: There are relatively few studies of flower color at landscape scales that can address the relative importance of competing mechanisms (e.g., biotic: pollinators; abiotic: ultraviolet radiation, drought stress) at landscape scales.

Methods: We developed an R shiny pipeline to sample color from images that were automatically downloaded using query results from a search using iNaturalist or the Global Biodiversity Information Facility (GBIF). The pipeline was used to sample ca. 4800 North American wallflower (*Erysimum*, Brassicaceae) images from iNaturalist. We tested whether flower color was distributed non-randomly across the landscape and whether spatial patterns were correlated with climate. We also used images including ColorCheckers to compare analyses of raw images to color-calibrated images.

Results: Flower color was strongly non-randomly distributed spatially, but did not correlate strongly with climate, with most of the variation explained instead by spatial autocorrelation. However, finer-scale patterns including local correlations between elevation and color were observed. Analyses using color-calibrated and raw images revealed similar results.

Discussion: This pipeline provides users the ability to rapidly capture color data from iNaturalist images and can be a useful tool in detecting spatial or temporal changes in color using citizen science data.

KEYWORDS

biogeography, citizen science, digital photographs, *Erysimum*, flower color, iNaturalist, R shiny

Flower color is among the most striking outcomes of angiosperm diversification, and the relationship between flower color and pollinators has been appreciated since the earliest days of formal biological inquiry. However, over the past few decades, relationships between flower color and other factors have been noted, suggesting that the role of flower color may extend far beyond pollinator interactions. For example, infraspecific flower color polymorphisms have been shown to be related to water stress (Warren and Mackenzie, 2001; Arista et al., 2013), irradiance (Winkel-Shirley, 2002; Koski and Ashman, 2015), and defense against pathogens or herbivores (Strauss and Whittall, 2006). Furthermore, it has long been understood that in some cases flower color may be a neutral trait, with drift explaining observed patterns of flower color polymorphism within populations (Epling and Dobzhansky, 1942). While pollinator attraction remains at the forefront of our understanding of flower color, especially in the context of within-species variation and the putative role of flower color in diversification (reviewed by Schiestl and Johnson, 2013; van der Kooi et al., 2019), it is clear that much work remains to be done in understanding the full array of patterns and processes in the ecology and evolution of flower color.

Although traditional studies of flower color variation have often been modest in spatial scale, recent biogeographic-scale studies have delivered important insights into the drivers of flower color within variable species. A strong latitudinal gradient, for example, was documented by Arista et al. (2013) in which orange and blue morphs of *Lysimachia arvensis* (L.) U. Manns & Anderb. were shown to inhabit overlapping but different latitudinal zones across Europe. Similarly, Koski and

Ashman (2015) demonstrated a global latitudinal gradient in flower pigmentation associated with ultraviolet (UV) irradiance in *Argentina anserina* (L.) Rydb. Warren and Mackenzie (2001) likewise uncovered striking differences in the distribution of white and purple morphs of several species across the British Isles that correspond to differences in water availability. Koski and Galloway (2020) demonstrated a stronger relationship in eastern North American *Campanula americana* L. of petal reflectance with local temperature than with post-glacial recolonization or pollinator communities. In each of these cases, analyses across large geographic areas revealed important spatial patterns in flower color that appeared to be driven by abiotic factors rather than pollinator preference.

Increasingly, large-scale biogeographical questions are being addressed by citizen scientist data sets that vastly increase the quantity and geographic scope of observations that can be leveraged to identify patterns and test hypotheses at broad scales. For decades, spatial and temporal analyses of avian diversity have been possible using data from the citizen science portal eBird (Wood et al., 2011). Similarly, studies of urban ecology (Weckel et al., 2010), invasion biology (Bois et al., 2011; Werenkraut et al., 2020; Tran et al., 2022), and phenology (Barve et al., 2020) have employed citizen science data to address questions that were intractable prior to the development of iNaturalist (https://www.inaturalist.org/) and other citizen science platforms. Similar to museum collections, citizen science observations are generally strongly biased spatially and taxonomically, and should therefore be used with caution. However, for some biogeographic questions the data possess tremendous potential for both generating and answering broad-scale questions.

In this study, we developed a pipeline to obtain flower color data from iNaturalist images to study flower color variation within the *Erysimum capitatum* (Douglas ex Hook.) Greene complex (wallflowers, Brassicaceae). The complex is native to North America and is considered to have radiated rapidly since first colonizing from the Old World less than 2 mya (Moazzeni et al., 2014; Züst et al., 2020). Estimates for the number of distinct taxa in the species complex have ranged from 17 to 33 since the mid 20th century, and the number of species recognized has ranged from 11 to 25 (Rossbach, 1958; Price, 1987; Al-Shehbaz, 2010). Flower color in the species complex ranges from outliers that are creamy white (isolated to a small number of distinctive taxa along the Pacific Coast) to lemon yellow, through orange to bright red and maroon, with a few outlying populations that are purple (Price, 1987). The source of color variation in *Erysimum* L. has not been studied; however, a similar range of colors (white, yellow, pink, and bronze morphs) have been studied in the genus *Raphanus* L., which is also in the Brassicaceae. In that genus, different combinations of anthocyanins and carotenoids produce color morphs, with white flowers expressing neither pigment group, yellow produced by carotenoids only, pink produced by anthocyanins only, and bronze expressing both pigment groups (Narbona et al., 2021).

Flower color variation in *Erysimum* appears to be continuous, with the most common colors being in the yellow to orange range. The flowers are relatively showy and have no remarkable structural modifications, and the pollinators include a wide array of bees, butterflies, and other generalists, with no apparent preferences documented in association with color differences (Price, 1987). There is currently no indication that selection by pollinators would be involved in driving shifts in floral color. To better understand the extent to which abiotic factors may be driving floral color differentiation in this complex, we set out to identify whether shifts in color are associated with geography or environmental factors.

To accomplish this, we developed an R shiny (Chang et al., 2020; https://shiny.rstudio.com/) pipeline to rapidly acquire quantitative color estimates from iNaturalist photographs. R shiny is a user interface built on the R language that allows users to harness R's power and flexibility through an interface with drop-down menus, buttons, and similar features without having to interact with command line prompts. Although similar color extraction procedures can be accomplished through ImageJ (Abramoff et al., 2004), the R package *colorZapper* (Valcu and Dale, 2023), and other image processing software, this pipeline was built with features that interface specifically with the export formats of query results from both iNaturalist queries and queries of iNaturalist records via the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/). Users perform a query in iNaturalist or GBIF (e.g., "*Erysimum* from western North America") and download the results of the query as a .csv file. The R shiny app then allows the user to click through images corresponding to the records in the query results to obtain color estimates for each image. Results of the color sampling are automatically saved along with the metadata (e.g., spatial coordinates, observation date) to facilitate downstream data visualization, spatial analysis, and integration with spatial data sets like climate layers, soil maps, or distributions of pollinators of interest. The tool works equally well in an offline mode to analyze, for example, herbarium records or separate folders of images representing color variation in different populations.

This paper has three primary aims: (A) to demonstrate the utility of the R shiny pipeline to both add value to existing iNaturalist collections and to serve as a tool for a broad range of users to investigate natural history (the tool has been tested by university undergraduates and is probably appropriate for high school students), (B) to understand spatial patterns in color variation in the *E. capitatum* complex. Our specific hypotheses related to wallflowers were that (1) flower color is spatially autocorrelated (i.e., geographic proximity is a good predictor of similarity in flower color); (2) any deviations from random spatial patterns are correlated with climate; and (3) based on anecdotal information, flower color varies predictably across elevational gradients. Finally, (C) we used a data set of images including flowers and ColorCheckers to quantify error from field photos and to determine the extent to which analyses using uncalibrated field images can capture true shifts in color.

METHODS

Color selector pipeline

We developed a pipeline consisting of R shiny apps, and associated R scripts, to extract color from iNaturalist images or other folders of images. A manual is provided (https://github.com/bencarter125/FlowerColor; see Data Availability Statement) that outlines in detail the operation of each component of the pipeline. Searches for iNaturalist records most commonly occur either through the iNaturalist website (https://www.inaturalist.org/) or by doing a search on GBIF constrained to only iNaturalist photos, as all research-grade iNaturalist records are made available through GBIF. For GBIF downloads, users have the option to select from multiple images associated with iNaturalist records if present (searches through iNaturalist only make the primary image accessible). Users can then either download all the images for later processing or run a web version that downloads and processes a single image at a time to avoid data storage issues. Both the iNaturalist and GBIF apps allow the user to then collect as many color samples from each image as desired by clicking on the image, with the color data retaining the important metadata (location, date) so that downstream analyses of color across landscapes can be performed; additional details are provided in the manual (see Data Availability Statement).

Data acquisition

All iNaturalist observations of *Erysimum* from the western United States up to May 2022 were downloaded from iNaturalist, including both research-grade and non-research-grade occurrences. Preliminary analysis indicated that iNaturalist users can usually accurately identify the genus *Erysimum*, but identification of putative species is more challenging. Indeed, expert opinion of species circumscriptions varies substantially (Rossbach, 1958; Price, 1987; Al-Shehbaz, 2010, 2012) within the North American *Erysimum* clade. Because the North American clade is known to be monophyletic and originated less than 2 mya (Moazzeni et al., 2014; Züst et al., 2020), we chose to treat the entire group as a complex and eliminated only the introduced Eurasian species from consideration (e.g., *E. cheiri* (L.) Crantz, *E. cheiranthoides* L.).

Occurrences were cleaned geographically to exclude any occurrences associated with anthropogenic landscapes. This was particularly important in this group because images of ornamental cultivars of *E. capitatum*, *E. cheiri*, and others are present in both iNaturalist and GBIF. A 2019 land use map of the United States with a spatial resolution of 30 m was obtained from the National Land Cover Database (https://www.mrlc.gov), and any records occurring on developed land (cover classes 21–24) or on planted/cultivated land (cover classes 81 and 82) were excluded from the data set. Following this, the 19 BIOCLIM variables were obtained from the WorldClim 2.0 data set (Fick and Hijmans, 2017) at the highest available spatial resolution (30 second), and values were extracted for each of the iNaturalist occurrences.

Images were then downloaded from iNaturalist and color was sampled by selecting approximately 15 pixels to represent the range of color in the inflorescence in each photo using our R shiny pipeline. Petals that were shaded, overexposed, or withered were avoided, and images that were entirely shaded or overexposed or included targeted species that were misidentified as *Erysimum* were removed. After cleaning, the data set included 4886 occurrences (81% of the 6031 downloaded occurrences). Color was converted from the red-green-blue (RGB) to hue-saturation-value (HSV) color space, and then the hue value was retained for each of the sampled pixel values for the image. Preliminary analyses indicated that the hue value nearly perfectly captured the range of color variation in *Erysimum*, which extends primarily from lemon-yellow through orange to red. Hereafter, "color" indicates the hue from the HSV values for images.

Spatial patterns of color

Two approaches were used to determine whether color was distributed non-randomly across the landscape. First, Moran's *I* was calculated to determine whether color was spatially autocorrelated across the occurrences, and second, spatial randomization tests were performed to determine whether any regions had more yellow or orange occurrences than expected by chance or whether any regions had more color heterogeneity or homogeneity (i.e., high or low color variance) than expected by chance.

Moran's *I* was calculated using the R package *spdep* (Bivand and Wong, 2018) after transforming to an Albers equal-area projection. A neighbor index with distances <100 km was created and resulted in all but 17 of the 4886 occurrences having at least one neighbor. After removal of those 17 occurrences, the neighbor index was constructed again and used to build a set of neighbor distances using binary weighting. This was then used to calculate Moran's *I*, and a permutation test with 999 iterations was used to test whether Moran's *I* differed significantly from zero.

Spatial randomization tests were then performed to identify which areas had more yellow or orange flowers than expected by chance and which had more homogeneous or heterogeneous flower color than expected. First, occurrences were aggregated to 0.5-degree grid cells (ca. 50 × 50 km), and then any grid cells with fewer than three occurrences were omitted, leaving a total of 256 grid cells. The three-occurrence threshold was used to balance retention of cells in order to derive meaningful geographic signal while also retaining cells with enough occurrences to calculate a meaningful mean and standard deviation. The observed mean and standard deviation of color were calculated for each grid cell. Then, for each of 999 iterations, occurrences were randomly reassigned to grid cells while keeping the number of occurrences per grid cell constant.

For each randomization, the mean and standard deviation of color were obtained for each grid cell. Significance for a two-tailed test with alpha of 0.05 was determined by finding grid cells for which the observed value was greater than 97.5% or less than 0.5% of the randomly generated values for each grid cell.

Color patterns with respect to climate

To determine the extent to which any non-random color patterns were explained by climate, an ordination was used to visually identify patterns and a spatial regression was employed to test for statistical associations of color and climate. For both analyses, a data set of six of the 19 BIOCLIM variables was used. The six variables used (temperature seasonality, mean temperature of the wettest quarter, mean temperature of the driest quarter, mean temperature of the coldest quarter, precipitation of the warmest quarter, and precipitation of the coldest quarter) were obtained by removing one or both of all variable pairs with correlation coefficients of greater than 0.70. To aid in the visualization of the ordination, biogeographic groups of occurrences were obtained by subjecting the climate data set to k-means clustering with $k = 5$, as determined by the total within-group sum of squares for k values of 1–10.

A principal component analysis (PCA) was conducted on the climate data set after scaling the variables. Observed flower color, biogeographic clusters, and occurrences associated with significantly orange or significantly yellow regions were mapped on the ordination axes.

A spatial regression was performed on the six (normalized) climate variables to test whether any climate variables were significant predictors of flower color after incorporating spatial autocorrelation. Specifically, a spatial error version of the spatial autoregressive model (SAR) was implemented in the R package *spatialreg* (Bivand et al., 2021) following Dormann et al. (2007) and Kissling and Carl (2008), and using the spatial weights index generated for the Moran's I test (see above).

Because flower color patterns in the Southern Rocky Mountains were anomalous relative to those in California and Colorado (all of which have a substantial number of orange flowers), we further explored regional differences by plotting histograms of flower color, plotting flower color against elevation, and obtaining correlation coefficients for flower color and elevation for these three regions.

Impacts of image color correction

iNaturalist images are inherently variable because there is no standardization for weather conditions, camera type, exposure, and other sources of error, with error here being the difference between true flower color and the color captured in images. To understand the impacts of this error, we compiled a data set of images taken with a 24-panel color checker (ColorChecker Classic Mini; Calibrite, Wilmington, Delaware, USA) in the frame. We used these data to answer two questions: (1) Are certain colors more variable in field photos than others? and (2) Do analyses of raw photos (similar to iNaturalist photos) yield similar results to analyses using color-corrected versions of the same photos?

Calibrated images were taken as part of an ongoing population genetic study of *Erysimum* and, for the purposes of this study, can be considered haphazard with respect to geography and climatic conditions. Images were taken of a total of 439 flowers representing 38 populations of *Erysimum* from across its distribution in western North America from 15 March 2021 through 9 July 2021. For each image, we used the ColorSelectorLocal R shiny app to capture color from flowers in the image (approximately 15 samples per image) and, separately, to capture color from each of the 24 ColorChecker panels (three samples per panel for each image). Images were calibrated (i.e., color-corrected) using equations provided in the R package *patternize* (Van Belleghem, 2018); however, the package itself was not used because our field photos of ColorCheckers varied too much in orientation and angle.

To determine whether particular colors are more variable than others in the field, we generated boxplots for each of the 24 ColorChecker panels of the distance between observed color and true color for all 439 images. We also plotted histograms to determine whether univariate errors differed across the hue, saturation, or value components of the HSV color values. To determine whether the calibration of images might substantively change interpretation of our results, we used a subset of six of the 39 populations distributed along strong elevational color gradients (two low- and two high-elevation populations in the Sierra Nevada, and one high- and one low-elevation population in the Southern Rocky Mountains). Boxplots of the observed color and corrected color for images in each of the six populations were compared qualitatively to known trends in the iNaturalist data set.

RESULTS

Spatial patterns of color

Color in the *E. capitatum* complex is significantly more clustered than expected by chance (Moran's $I = 0.087$, $P < 0.001$; Figure 1). Spatial randomization tests illustrated that this pattern is largely driven by the presence or absence of orange flowers. Yellow flowers are nearly ubiquitous, and many regions have significantly more yellow flowers than expected by chance (Figure 2). Regions with significantly more orange flowers than expected are more localized and occur primarily in

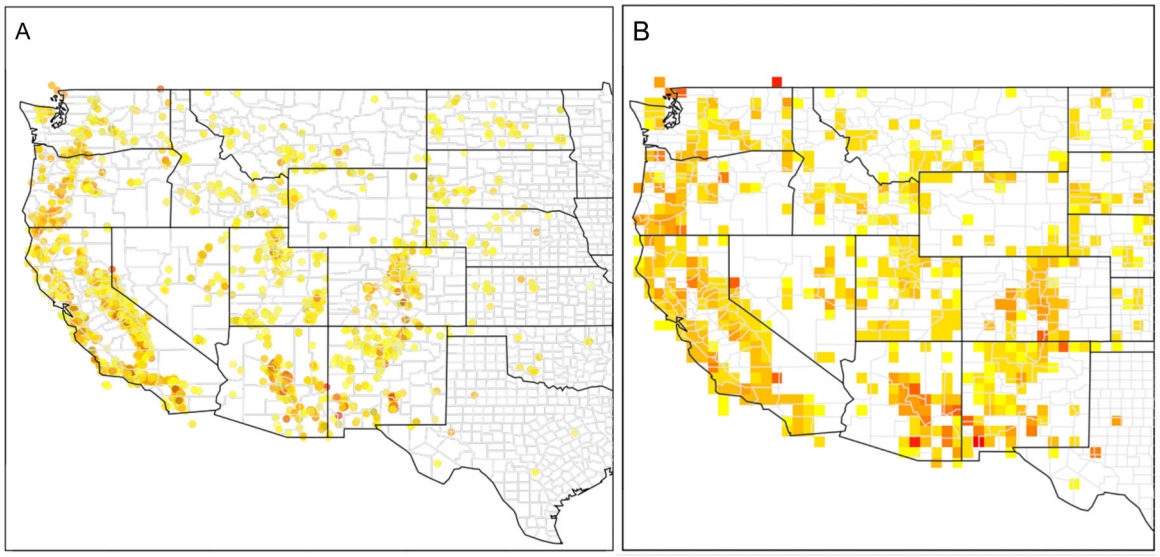

FIGURE 1 Geographical distribution of flower color in the *Erysimum capitatum* complex. (A) Occurrences colored using color obtained from iNaturalist records. (B) Occurrences aggregated into grid cells, colored using the mean color of occurrences in each grid cell.

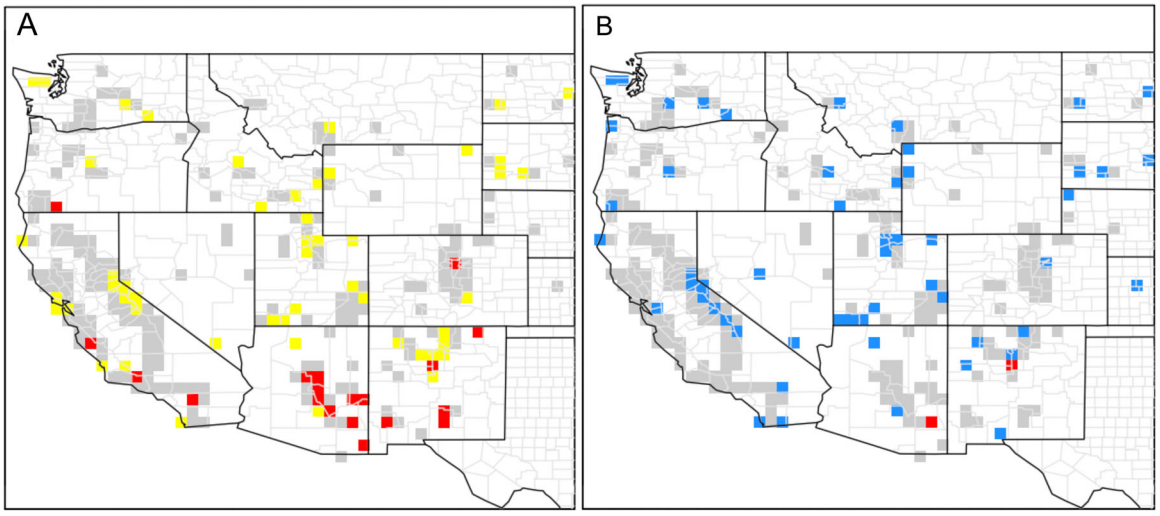

FIGURE 2 Geographical distribution of flower color in the *Erysimum capitatum* complex, showing regions with (A) flowers having significantly more yellow values (yellow cells) and significantly more orange values (red cells) than expected by chance and (B) with more homogeneous (blue cells) and more heterogeneous (red cells) flower color than expected by chance.

Arizona and New Mexico, with scattered regions in Colorado and California (Figure 2). There is a significant correspondence between pixels that are more yellow than expected and those that are more homogeneous than expected (Table 1); there are no pixels that are both more homogeneous and orange than expected or both more heterogeneous and more yellow than expected ($\chi^2 = 98.37$, $P < 0.001$; Table 1). Together, these indicate that orange flowers are relatively restricted and that when local flower color heterogeneity exists, it is driven by the co-occurrence of yellow and orange flowers, as compared to homogeneous areas that only have yellow flowers.

Color patterns with respect to climate

Overlaying biogeographic groups and regions with significantly more yellow or orange flowers than expected onto the climate PCA revealed that, while the relationship of color and climate is not random, there are not broadscale generalities to be made across the entire range of the species complex (Figure 3). Regions with more yellow flowers are aggregated in clusters across all five of the biogeographic regions, while regions with more orange flowers than expected are found in only two of the five regions and these two clusters are widely separated climatically. Together, these give a clear indication that

the distribution of orange and yellow flowers is not random, but that the association of flower color and climate depends greatly on the regional context.

Correlations between color and climate in the spatial regression were significant, but climate did not explain much of the variation in color (Table 2). The spatial regression demonstrated that four of the six climate predictors contributed significantly (at $\alpha = 0.01$) to the model, with precipitation of the warmest quarter (i.e., summer rainfall) contributing the most by far based on the change in Akaike information criterion (AIC; Table 2). The full regression model with no spatial term was significant as well, but with an adjusted R^2 of only 0.095. The spatial regressions do not produce R^2 values, but the low R^2 of the nonspatial model is consistent with the univariate correlation coefficients of each predictor against color, all of which had an absolute value less than 0.20.

Both the standard deviation of color and mean color indicated that the Southern Rocky Mountains differ from other regions. Further exploration confirmed this, with the Southern Rockies exhibiting both a bimodal distribution of flower color (red and yellow) and a generally high proportion of occurrences on the red end of the spectrum (Figure 4). A relationship with elevation was also present, with redder flowers tending to occur at higher elevations. In

TABLE 1 Correspondence of grid cells that were significantly different from random in mean color (orange/yellow) and standard deviation of color (homogeneous/heterogeneous). See also Figure 2.

Color	Homogeneous	Not significant	Heterogeneous
Orange	0	17	2
Not significant	21	165	0
Yellow	33	18	0

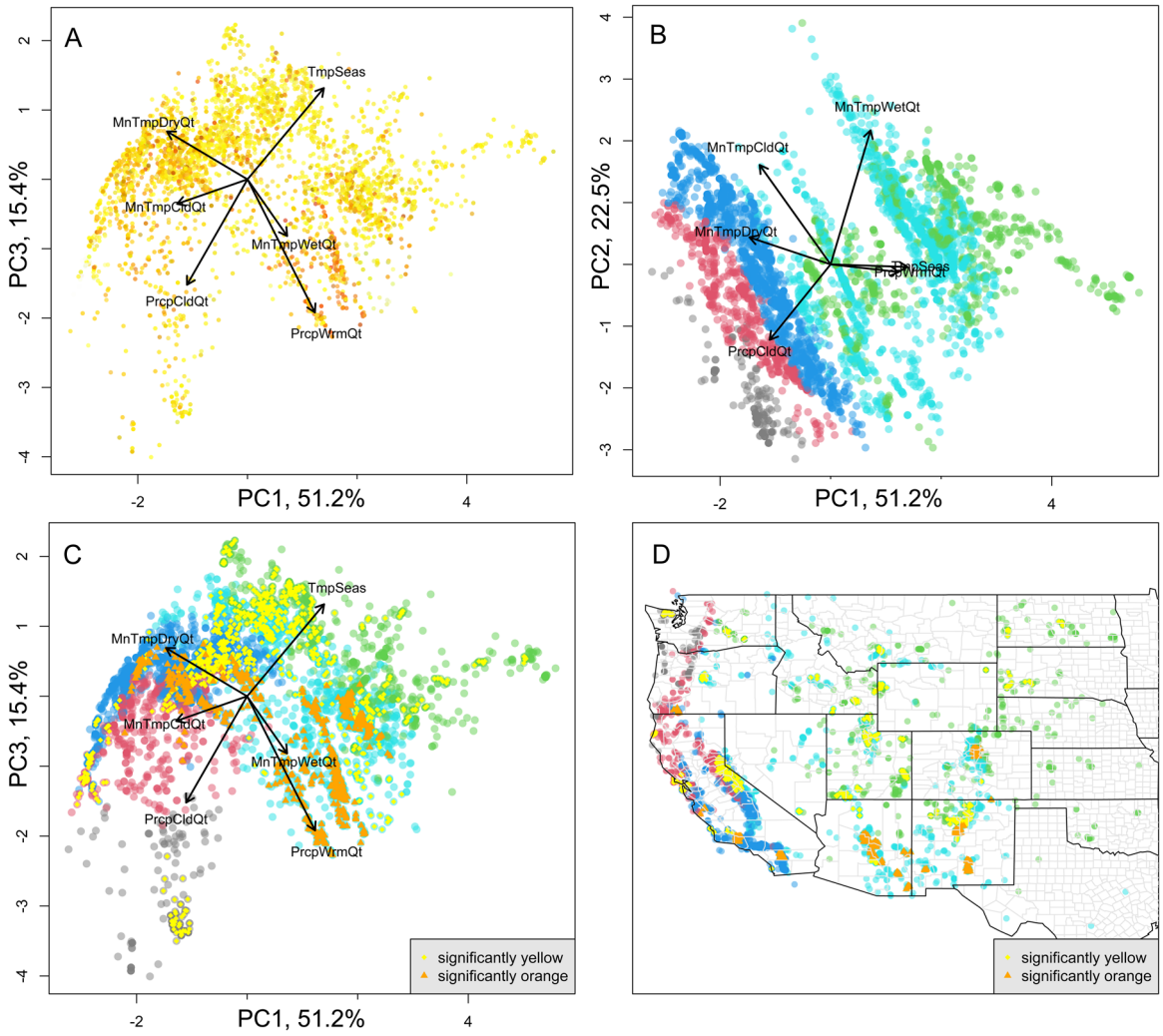

FIGURE 3 Relationships of flower color to climate. (A) Principal component analysis (PCA) axes 1 and 3 with occurrences across six BIOCLIM variables plotted with mean color of each occurrence. (B) PCA axes 1 and 2 plotted with bioregions assigned by *k*-means clustering. (C) PCA axes 1 and 3 plotted with bioregions assigned by *k*-means clustering, with occurrences from significantly yellow or orange regions in yellow and orange. (D) Map of occurrences displaying bioregions and significantly yellow and orange regions.

TABLE 2 Results from regression analyses. Standardized coefficients are provided for both the simple linear (non-spatial) regression and spatial regression, both of which had full models with significance of $P < 0.001$. The full spatial model had an Akaike information criterion (AIC) of 48835; ΔAICs are the change in AIC from a model that left out each predictor individually. Correlation coefficients are provided for each (unstandardized) variable against flower color, and the overall R^2 of the non-spatial model was 0.095.

Climate variable	Non-spatial standardized coefficients	Spatial standardized coefficients	Spatial *P* value	Spatial ΔAIC	Univariate correlation coefficient
Temperature seasonality	0.041	0.022	<0.001	−11	0.097
Mean temperature of wettest quarter	0.065	0.214	0.165	−11	−0.105
Mean temperature of driest quarter	−0.880	−1.332	<0.001	−21	−0.023
Mean temperature of coldest quarter	−0.341	−0.207	0.492	−12	−0.066
Precipation in warmest quarter	−0.194	−0.222	<0.001	269	−0.120
Precipation in coldest quarter	0.021	0.011	0.002	−4	0.001

contrast, the Sierra Nevada had a unimodal distribution with yellow at the mode and demonstrated the opposite elevational trend, with yellow flowers at higher elevations and orange at lower.

Impacts of image color correction

Error, quantified as the difference between observed ColorChecker color and true ColorChecker color, was generally low and was not substantially concentrated in one portion of the color spectrum (Figure 5). The three-dimensional distances (i.e., incorporating hue, saturation, and value distances together) ranged from 0.117 to 0.494 with a mean of 0.243 across the 24 ColorChecker colors. Grayscale colors tended to have higher error, but otherwise no strong trends were detected with respect to the distribution of error across the 24 colors.

In the analyses of differences between different-colored populations at different elevations (Figure 6), we found that we were able to recover the expected color shifts across elevation, and that analyses performed with calibrated images were qualitatively very similar to those performed on raw images.

DISCUSSION

Citizen science data sets have tremendous potential to increase the breadth of sampling for scientific research, and we demonstrate that here with our analysis of color variation in western wallflowers from thousands of sites across western North America. We found that color is strongly non-randomly distributed across the range of the *E. capitatum* complex as well as non-randomly distributed with respect to climate; however, patterns are localized and appear to be context specific. This study also demonstrates that error (the difference between the true color of an object and the color of the object as recorded in a photograph) was reasonably low for a data set of more than 400 images taken in the field, indicating that uncontrolled iNaturalist images, for at least some projects with reasonably high sample sizes, provide usable estimates of color. Finally, the R shiny pipeline introduced here provides a rapid method for acquiring color data from iNaturalist or other images.

Flower color variation in *Erysimum*

The analyses presented here indicated that both mean color and color variance were distributed non-randomly across the range of *E. capitatum*, and that the two were not independent. Yellow flowers were common throughout the range, and many areas had significantly homogeneously yellow flowers. In contrast, areas with more orange flowers were localized in the Southern Rocky Mountains and coastal California, and these areas contained both yellow and orange flowers. This non-randomness at the landscape scale is an important finding that was possible only through the large number of citizen science records.

The non-random pattern of color variation we observed could be driven by either natural selection or by non-selective processes. On one hand, divergent floral colors could experience higher fitness across a heterogeneous landscape, with selection driven by biotic factors such as pollinator preference or herbivore defense, or by abiotic factors such as the ability to tolerate environmental stressors including UV exposure, drought, or cold. Alternatively, shifts in flower color may arise through non-selective factors such as genetic drift or hybridization, with allelic recombination resulting in novel phenotypes. We found that at the scale of the entire range of the *E. capitatum*

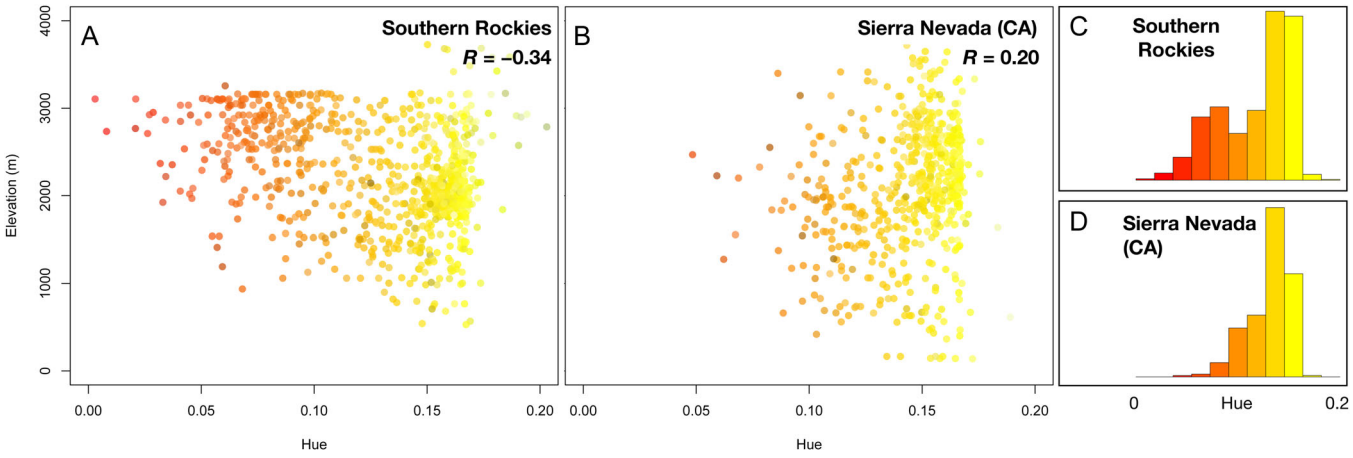

FIGURE 4 Distribution of flower color (hue) for two regions in western North America. (A, B) Relationship of flower color to elevation. (C, D) Histograms of flower color.

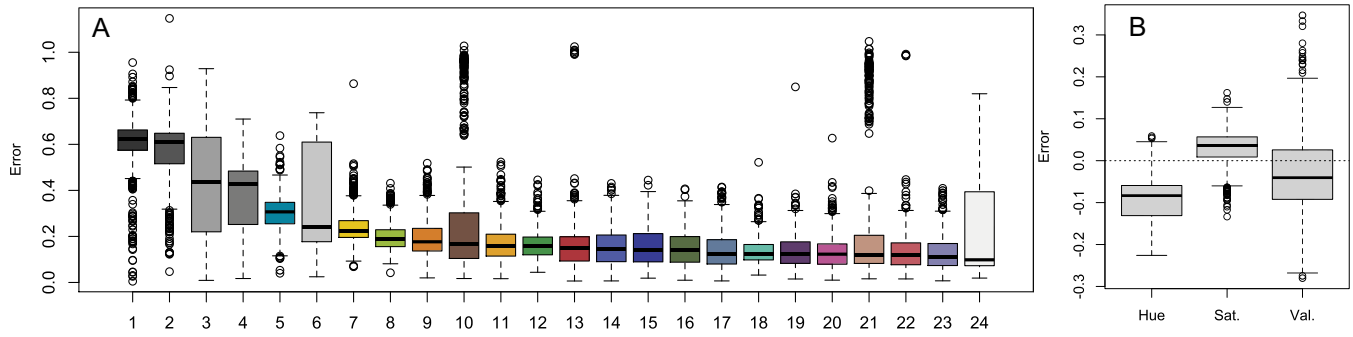

FIGURE 5 Error in color between observed ColorChecker and true ColorChecker values across photographs from the field ($n = 439$). (A) Euclidean distance, for each of the 24 ColorChecker panels, between true color and observed color. (B) Error (distance between true and observed values) in the hue, saturation, and value components across the 24 ColorChecker panels combined.

complex, floral color does not appear to be strongly correlated with climate. Although color is clearly not random with respect to climate (Figures 3 and 4), most of the clustering can be explained by spatial autocorrelation (Table 2).

To date, our understanding of flower color evolution in *Erysimum* comes from a single study of Old World taxa. Gómez et al. (2015) investigated floral color evolution across a clade of 40 species distributed in Western Europe and North Africa. Their results indicated little phylogenetic signal associated with flower color, suggesting that color is a highly labile trait. Ancestral state reconstructions demonstrated that there have been multiple independent transitions from yellow to lilac-colored flowers as well as secondary transitions back to yellow. Anthocyanins that underlie purple color in Brassicaceae are associated with stress tolerance (Chalker-Scott, 1999; Dick et al., 2011). Furthermore, because Old World *Erysimum* taxa with lilac-colored flowers are primarily relegated to the arid areas at the southern end of the genus' distribution, Gómez et al. (2015) hypothesized that floral color shifts may originally have arisen as an adaptation to stressful environments, in

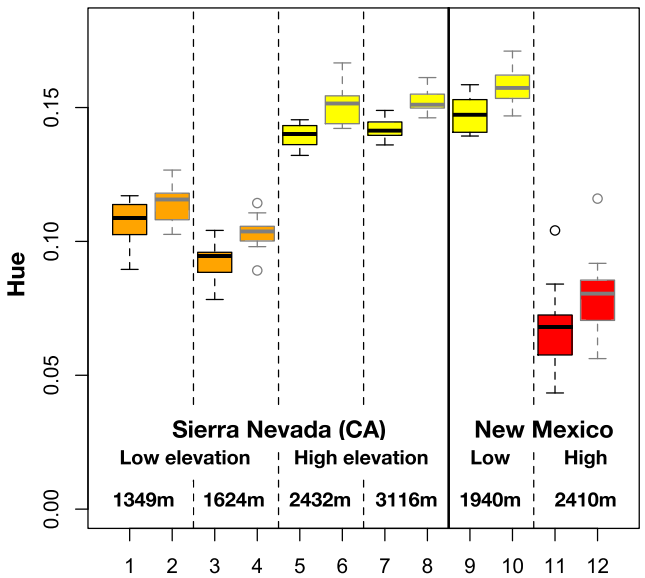

FIGURE 6 Comparison of color values (hue) for raw images (left, in black outline) and corresponding color-calibrated images (right, in gray outline) for six populations in California and New Mexico. Note that the relationship of color to elevation switches between the California and New Mexico populations.

alignment with a broader understanding across angiosperms that anthocyanin production is associated with stressful environments (Landi et al., 2015). While differences in pollinator niches were observed for yellow and lilac taxa, evolutionary reconstructions suggested that floral color shifts in this group likely preceded shifts in pollinator niche.

While to our knowledge there has been no investigation of floral color evolution in North American *Erysimum* taxa, Lay et al. (2013) quantified color across four populations of *E. capitatum* distributed along an elevational gradient in the Rocky Mountains. They found, similar to the results we presented here, that red flowers were associated with higher elevations and yellow flowers were associated with lower elevations. While pollinator observations suggested a shift from bee to fly floral visitation as elevation increased, pollinator communities still significantly overlapped across populations, suggesting that they are unlikely to be driving the observed floral color shifts. This is consistent with observations by Price (1987), who anecdotally found no differences among the diverse pollinators of orange and yellow flowers in California and observed no additional floral characteristics that might suggest selection by pollinators. Although we did find that redder flowers were associated with higher elevation in the Southern Rockies, we also found the opposite to be true in the Sierra Nevada. Together, these results highlight that floral color shifts in the *E. capitatum* complex are likely not being driven by one single factor (e.g., increased UV exposure or cooler temperatures at higher elevations); instead, it is much more likely that a variety of mechanisms (selective and non-selective) are occurring at local scales, resulting in the non-random but variable patterns we documented here. While further research is needed to tease apart the relative importance of ecological factors vs. non-ecological factors (e.g., drift or hybridization) in the establishment and maintenance of floral color variation, our work highlights the importance of thorough sampling when attempting to make broad-scale inferences about the role of environment in driving phenotypic change and identifies regions where selective pressure may be driving floral color shifts in the *E. capitatum* complex.

Color correction and color selection best practices

We encountered two primary challenges in compiling this data set. First, iNaturalist photographs were unstandardized and were taken under a wide range of lighting conditions in the field. Second, within an image, shading of flowers by other flowers in the same inflorescence creates a mosaic of lighting conditions at a finer scale that must be sampled consistently across images.

The uncontrolled nature of the iNaturalist images would be problematic if either the magnitude of the error was large relative to the magnitude of biologically interesting color variation (signal-to-noise ratio) or the distribution of error was not centered on zero (systematic directional bias), or if particular colors were substantially more error prone than other colors. Using our data set of images containing ColorCheckers, we found reasonably low levels of color error that were both symmetrically distributed and centered near an error of zero for the red, green, and blue components individually. This suggests that, while error was present, it was not systematically biased. In our data set, the color variation relevant to our question occurred along the green component between values of approximately 225 (corresponding to pure yellow) and 150 (corresponding to orange). In comparison, standard deviations of error for the red, green, and blue components were 27.2, 23.0, and 21.7, respectively. The error in the blue component was substantially greater in magnitude than in the red or green components, which we interpreted as being particularly influenced by shade/sun differences, but all three were reasonably small relative to the magnitude of color differences relevant to our research question.

Among the 24 colors in our ColorChecker, we similarly found consistent errors, but with the error relatively low compared to the magnitude of color differences of biological interest. This is particularly evident in the direct comparisons of calibrated and uncalibrated images across elevation gradients (Figure 6). While the distribution of within-population variation shifted slightly and unpredictably between calibrated and uncalibrated images, the same general patterns were evident across populations. Taking each of these results into account and considering the ability to easily generate large sample sizes using iNaturalist, we felt confident that using iNaturalist images provided results that were both unbiased and with a relatively small amount of error compared to the biological differences we aimed to quantify.

The second major challenge was consistency of data capture across images, as only approximately 15 pixels were sampled from each flower. We found, similar to other citizen science projects using iNaturalist images (Barve et al., 2020), that repeatability required iterative training and clear guidelines for data capture. For this study, these guidelines included avoiding portions of petals that were shaded by other petals, avoiding portions of flowers that had glare or were clearly overexposed, and eliminating images with poor quality (e.g., photos of flowers taken from a distance or in deep shade). We recommend, similar to Barve et al. (2020), iteratively developing a set of guidelines on a case-by-case basis for each study using a representative training set of images. Particularly with flower color, species that range from, for example, white to pink might be expected to present different sets of challenges than species that range from blue to purple. Additionally, pigmentation varies with age for some flowers, or can vary in relation to veins or other floral features, and it is important to incorporate these sources of variation into the protocol for answering a particular research question.

Comparison to similar recent approaches

Two recent studies use similar approaches to quantify color from citizen science images, and a description of the differences between these studies and the workflow presented here could help potential users decide on an appropriate approach for their particular question. Laitly et al. (2021) used a workflow employing the R package *colorZapper* (Valcu and Dale, 2023). Their workflow, in contrast to the workflow presented here and in the study discussed below, does not directly interface with the results of iNaturalist queries. However, their use of *colorZapper*, which can be used to extract color from user-defined pixels or user-defined polygons, is arguably the most flexible and can be used for any set of images. Laitly et al. (2021) also provide a head-to-head comparison of image analyses across HSV, RGB, and CIELAB color spaces, comparisons across cameras, and a comparison of museum specimen images taken under controlled conditions with citizen science images. A particularly relevant and broadly applicable finding of that study was that, for separation of species using these approaches, the use of more than 12–14 images per species provided diminishing returns.

In contrast, Perez-Udell et al. (2023) provided a Python-based workflow that directly interfaces with iNaturalist search results. Their workflow has an automated approach to color selection that first identifies clusters of pixels based on color using *k*-means and then provides summary statistics for the color clusters, with an option for automatically classifying images into groups (e.g., in cases of taxa with discrete color polymorphisms).

The approach introduced in this paper uses the R environment and interfaces directly with iNaturalist queries, similar to Perez-Udell et al. (2023). However, in this approach, users select pixels from images rather than automatically defining color clusters. While we are enthusiastic about the potential of automation and explored an automated approach in initial trials, we pursued user-defined pixel selection for two reasons. First, automated color extraction requires a strong contrast between the target segment (flowers) and background that is not always reliably present (e.g., white flowers on light-colored sand substrates). In our experience, automated approaches will reject many usable images that have similar background colors or in which the flowers define a relatively small portion of the overall image. Second, there is no automated way to control for shading of portions of petals, which often differ strongly from the primary petal color but also fade gradually between shade and unshaded, thus thwarting a clustering algorithm's ability to separate shaded from unshaded portions of the image. In the future, as image segmentation artificial intelligence becomes more accessible, both of these challenges are likely to be overcome. Until such time, we did find (unpublished data) that density-based spatial clustering, implemented in the R package *dbscan* (Hahsler et al., 2019), outperformed *k*-means clustering, and this could be considered in ongoing attempts to automate the color extraction process. Our approach also offers a comparison of iNaturalist images to color-calibrated images taken in the field, which complements Laitly et al.'s (2021) comparison to museum specimen images taken under controlled conditions.

Caveats of citizen science data

In addition to the error associated with images taken under uncontrolled lighting conditions, other factors should be considered when using iNaturalist data sets. Similar to museum specimens, iNaturalist records are notoriously poorly identified and are strongly biased in space, in time, and toward charismatic and/or common or rare species. This is a well-known and commonly discussed issue for museum collections (Meineke and Daru, 2021), and citizen science records are subject to the same challenges. In our study, we focused on a genus that is showy and is easily visually separated from closely related genera (by both humans and iNaturalist's suggestions based on artificial intelligence). Species-level identifications in the genus are often not confidently assigned even by taxonomic experts, and the clade in North America is known to be monophyletic and of recent origin, thereby justifying an analysis at the clade level rather than attempting to work with individual species. Other studies should similarly focus on discrete groups for which misidentifications will be low and/or allocate time to confidently identifying records used for color capture.

AUTHOR CONTRIBUTIONS

B.E.C. and T.M.M. conceived the research, and B.E.C. wrote the code for the workflows and analyzed the data. Y.L. and A.G.-H. assisted in writing code, collected the data, wrote the manual, and tested the workflows. All authors contributed to writing the manuscript and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank J. Whittall (Santa Clara University) and C. M. Guilliams (Santa Barbara Botanic Garden) for many valuable discussions on flower color, and the developers and users of iNaturalist for the images and metadata used in this study. We also thank the many San Jose State University undergraduates, especially J. Le, J. Trow, B. Tiburcio, and G. Dunayen, who helped early in the development of the project.

DATA AVAILABILITY STATEMENT

The data and code for this project are available on GitHub (https://github.com/bencarter125/FlowerColor).

ORCID

Tracy M. Misiewicz http://orcid.org/0000-0002-2683-8405

Benjamin E. Carter http://orcid.org/0000-0002-6473-3866

REFERENCES

Abramoff, M. D., P. J. Magalhaes, and S. J. Ram. 2004. Image processing with ImageJ. *Biophotonics International* 11(7): 36–42.

Al-Shehbaz, I. A. 2010. *Erysimum*. *In* Flora of North America Editorial Committee [eds.], Flora of North America North of Mexico, 7. Oxford University Press, New York, New York, USA.

Al-Shehbaz, I. A. 2012. *Erysimum*. *In* B. G. Baldwin, D. Goldman, D. J. Keil, R. Patterson, T. J. Rosatti, and D. Wilken [eds.], The Jepson Manual, 2nd ed. University of California Press, Berkeley, California, USA.

Arista, M., M. Talavera, R. Berjano, and P. L. Ortiz. 2013. Abiotic factors may explain the geographical distribution of flower colour morphs and the maintenance of colour polymorphism in the scarlet pimpernel. *Journal of Ecology* 101: 1613–1622.

Barve, V. V., L. Brenskelle, D. Li, B. J. Stucky, N. V. Barve, M. M. Hantak, B. S. McLean, et al. 2020. Methods for broad-scale plant phenology assessments using citizen scientists' photographs. *Applications in Plant Sciences* 8: e11315. https://doi.org/10.1002/aps3.11315

Bivand, R. S., and D. W. S. Wong. 2018. Comparing implementations of global and local indicators of spatial association. *TEST* 27: 716–748. https://doi.org/10.1007/s11749-018-0599-x

Bivand, R., G. Millo, and G. Piras. 2021. A review of software for spatial econometrics in R. *Mathematics* 9: 1276. https://doi.org/10.3390/math9111276.

Bois, S. T., J. A. Silander, and L. J. Mehrhoff. 2011. Invasive plant atlas of New England: The role of citizens in the science of invasive alien species detection. *BioScience* 61: 763–770. https://doi.org/10.1525/bio.2011.61.10.6

Chalker-Scott, L. 1999. Environmental significance of anthocyanins in plant stress responses. *Photochemistry and Photobiology* 70: 1–9.

Chang, W., J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. 2020. shiny: Web Application Framework for R. Website https://CRAN.R-project.org/package=shiny [accessed 17 August 2023].

Dick, C. A., J. Buenrostro, T. Butler, M. L. Carlson, D. J. Kliebenstein, and J. B. Whittall. 2011. Arctic mustard flower color polymorphism controlled by petal-specific downregulation at the threshold of the anthocyanin biosynthetic pathway. *PLoS ONE* 6: e18230. https://doi.org/10.1371/journal.pone.0018230

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. C. Davies, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* 30: 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Epling, C., and T. Dobzhansky. 1942. Genetics of natural populations. VI. Microgeographic races in *Linanthus parryae*. *Genetics* 27: 317–332. https://doi.org/10.1093/genetics/27.3.317

Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37: 4302–4315. https://doi.org/10.1002/joc.5086

Gómez, J. M., F. Perfectti, and J. Lorite. 2015. The role of pollinators in floral diversification in a clade of generalist flowers. *Evolution* 69: 863–878. https://doi.org/10.1111/evo.12632

Hahsler, M., M. Piekenbrock, and D. Doran. 2019. dbscan: Fast density-based clustering with R. *Journal of Statistical Software* 91(1): 1–30. https://doi.org/10.18637/jss.v091.i01

Kissling, W. D., and G. Carl. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* 17: 59–71. https://doi.org/10.1111/j.1466-8238.2007.00334.x

Koski, M. H., and T. Ashman. 2015. Floral pigmentation patterns provide an example of Gloger's rule in plants. *Nature Plants* 1: 14007. https://doi.org/10.1038/nplants.2014.7

Koski, M. H., and L. F. Galloway. 2020. Geographic variation in floral color and reflectance correlates with temperature and colonization history. *Frontiers in Plant Science* 11: 991. https://doi.org/10.3389/fpls.2020.00991

Laitly, A., C. T. Callaghan, K. Delhey, and W. K. Cornwell. 2021. Is color data from citizen science photographs reliable for biodiversity research? *Ecology and Evolution* 11: 4071–4083. https://doi.org/10.1002/ece3.7307

Landi, M., M. Tattini, and K. S. Gould. 2015. Multiple functional roles of anthocyanins in plant-environment interactions. *Environmental and Experimental Botany* 119: 4–17. https://doi.org/10.1016/j.envexpbot.2015.05.012

Lay, C. R., Y. B. Linhart, and P. K. Diggle. 2013. Variation among four populations of *Erysimum capitatum* in phenotype, pollination and herbivory over an elevational gradient. *American Midland Naturalist* 169: 259–273. https://doi.org/10.1674/0003-0031-169.2.259

Meineke, E. K., and B. H. Daru. 2021. Bias assessments to expand research harnessing biological collections. *Trends in Ecology and Evolution* 36: 1071–1082. https://doi.org/10.1016/j.tree.2021.08.003

Moazzeni, H., S. Zarre, B. E. Pfeil, Y. J. Bertrand, D. A. German, I. A. Al-Shehbaz, K. Mummenhoff, and B. Oxelman. 2014. Phylogenetic perspectives on diversification and character evolution in the species-rich genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Botanical Journal of the Linnean Society* 175: 497–522. https://doi.org/10.1111/boj.12184

Narbona, E., J. C. del Valle, and J. Whittall. 2021. Painting the green canvas: How pigments produce flower colours. *Biochemist* 43(3): 6–12. https://doi.org/10.1042/bio_2021_137

Perez-Udell, R. A., A. T. Udell, and S. Chang. 2023. An automated pipeline for supervised classification of petal color from citizen science photographs. *Applications in Plant Sciences* 11(1): e11505. https://doi.org/10.1002/aps3.11505

Price, R. A. 1987. Systematics of the *Erysimum capitatum* alliance (Brassicaceae) in North America. Ph.D. dissertation, University of California, Berkeley, California, USA.

Rossbach, G. B. 1958. The genus *Erysimum* (Cruciferae) in North America north of Mexico—A key to the species and varieties. *Madroño* 14: 261–267.

Schiestl, F. P., and S. D. Johnson. 2013. Pollinator-mediated evolution of floral signals. *Trends in Ecology and Evolution* 28: 307–315. https://doi.org/10.1016/j.tree.2013.01.019

Strauss, S. Y., and J. B. Whittall. 2006. Non-pollinator agents of selection on floral traits. *In* L. D. Harder and S. C. H. Barrett [eds.], Ecology and evolution of flowers, 120–138. Oxford University Press, Oxford, United Kingdom.

Tran, T., B. E. Carter, and J. Castillo-Vardaro. 2022. Predicted threats to a native squirrel from two invading species based on citizen science data. *Biological Invasions* 24: 3539–3553. https://doi.org/10.1007/s10530-022-02859-7

Valcu, M., and J. Dale. 2023. colorZapper: Color extraction utilities. R package version 1.4.8. Website https://github.com/mpio-be/colorZapper [accessed 17 August 2023].

Van Belleghem, S. 2018. Patternize: Quantification of color pattern variation. R package version 0.0.2. Website https://CRAN.R-project.org/package=patternize [accessed 17 August 2023].

van der Kooi, C. J., A. G. Dyer, P. G. Kevan, and K. Lunau. 2019. Functional significance of the optical properties of flowers for visual signalling. *Annals of Botany* 123: 263–276. https://doi.org/10.1093/aob/mcy119

Warren, J., and S. Mackenzie. 2001. Why are all colour combinations not equally represented as flower-colour polymorphisms? *New Phytologist* 151: 23–241. https://doi.org/10.1046/j.1469-8137.2001.00159.x

Weckel, M. E., D. Mack, C. Nagy, R. Christie, and A. Wincorn. 2010. Using citizen science to map human–coyote interaction in suburban New York, USA. *Wildlife Management* 74: 1163–1171. https://doi.org/10.2193/2008-512

Werenkraut, V., F. Baudino, and H. E. Roy. 2020. Citizen science reveals distribution of the invasive harlequin ladybird (*Harmonia axyridis* Pallis) in Argentina. *Biological Invasions* 22: 2915–2921. https://doi.org/10.1007/s10530-020-02312-7

Winkel-Shirley, B. 2002. Biosynthesis of flavonoids and effects of stress. *Current Opinion in Plant Biology* 5: 218–223.

Wood, C., B. Sullivan, M. Illiff, D. Fink, and S. Kelling. 2011. eBird: Engaging birders in science and conservation. *PLoS Biology* 9: e1001220. https://doi.org/10.1371/journal.pbio.1001220

Züst, T., S. R. Strickler, A. F. Powell, M. E. Mabry, H. An, M. Mirzaei, T. York, et al. 2020. Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *Elife* 9: e51712. https://doi.org/10.7554/eLife.51712

How to cite this article: Luong, Y., A. Gasca-Herrera, T. M. Misiewicz, and B. E. Carter. 2023. A pipeline for the rapid collection of color data from photographs. *Applications in Plant Sciences* 11(5): e11546. https://doi.org/10.1002/aps3.11546