

GEOGRAPHIC VARIATION IN FLORAL PIGMENTATION OF US NATIVE HERB,
GERANIUM MACULATUM, AS EVALUATED BY TRADITIONAL AND CITIZEN
SCIENCE APPROACHES

by

RACHEL ANN PEREZ-UDELL

(Under the Direction of SHU-MEI CHANG)

ABSTRACT

Floral pigmentation is an ecologically critical trait frequently examined in context of plant's response to its environment (biotic and abiotic). Geographical variation is an important but difficult phenomenon to study. It requires dedicated field work at fine-tuned locations to establish a pattern. These patterns can provide insight on how selective and neutral evolutionary processes shape variation. This paper provides an overview of geographic variation in the native herbaceous species, *Geranium maculatum*. We have found an innovative method for classifying petal color using citizen science photographs from iNaturalist.org. We use this plus a traditional field study to describe the pattern of floral color in petal and pollen across the native range. Then, we use correlative studies to establish environmental influences that may select for pigmentation. Using linear regressions, random forests, niche model, mantel tests and PCA we determine that pollen and petal color may be influenced by precipitation, temperature, and UV. Finally, we used the RADcap approach to identify SNP loci to describe population differentiation with F_{st} . We investigate population structure and compare neutral genetic variation to

phenotypic variation previously found. In summary, our results suggest that floral pigmentation in this species is influenced by its environment.

INDEX WORDS: *Geranium maculatum*, geraniaceae, variation, computer vision, population genetics, citizen science, floral traits

GEOGRAPHIC VARIATION IN FLORAL PIGMENTATION OF US NATIVE HERB,
GERANIUM MACULATUM, AS EVALUATED BY TRADITIONAL AND CITIZEN
SCIENCE APPROACHES

by

RACHEL ANN PEREZ-UDELL

BS, University of North Georgia, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

© 2022

Rachel Ann Perez-Udell

All Rights Reserved

GEOGRAPHIC VARIATION IN FLORAL PIGMENTATION OF US NATIVE HERB,
GERANIUM MACULATUM, AS EVALUATED BY TRADITIONAL AND CITIZEN
SCIENCE APPROACHES

by

RACHEL ANN PEREZ-UDELL

Major Professor:	Shu-Mei Chang
Committee:	Dorset Trapnell
	Jill Anderson
	Jim Leebens-Mack
	John Wares

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2022

DEDICATION

This dissertation is dedicated to Antonio M. Perez, my paternal grandfather. Grandpa should have been the first Dr. Perez. Many decades ago, he was a veterinary student in Havana, Cuba. He had a wife, Lidia, and a young son, my father, Daniel. They were a happy family with a good future. Already, we know where this story takes its turn. Castro, his revolution, and my grandfather being a bit too loud with his oppositions left my family with little choice. They left, and in a hurry, as there was a warrant out for my grandfather's arrest. My paternal family settled down in Brooklyn, New York in the early 60's.

Unfortunately, I never got to meet my grandfather. He died before I was born. However, his vision of higher education lived through and here I am, proudly carrying on the Perez name. Over 60 years later, we've done it.

Here's to you, grandpa.

Yo sé los nombres extraños

De las yerbas y las flores,

Y de mortales engaños,

Y de sublimes dolores.

-from *Versos Sencillos* by Jose Martí.

ACKNOWLEDGEMENTS

Thank you to my parents, Ann, and Dan, for working hard to give Garrett and I a good life. Thank you for all the help along the way. Dad, you got me first interested in nature when I was young. You both always advocated for me when I could not. Without y'all I wouldn't be here, a first-generation college graduate and a PhD. I am blessed to have a supportive family, including my in-laws, who I adore.

Thank you to my partner, Andrew. You've been my other half since we began dating in 2014, after we took Introduction to Biology together in Fall 2013. To think our relationship started with a single paper for that course and we're still doing science together- and married! I couldn't have completed this without you. Your support and love help more than you'll ever know.

Thanks to UNG for even getting me on this path in the first place. If I had not have been required to take an honors course, I would not have found myself in a freshman biology course. That course is what gave me the confidence to tackle the degree (and it's where I met my husband. Win-Win). Throughout the program, I had wonderful mentors and gained a greater appreciation for the world around me.

Of course, I've had amazing mentors at UGA as well. Shu-Mei is not only a great advisor, but she is also a wonderful person. My entire committee has been good to me. I've taken courses with most of my committee members, and I've thoroughly enjoyed each one. Outside of my committee I want to thank Kim Martin, Sandi Lessl, and Brigitte Bruns, for their help outside the lab. You three have helped me grow as an educator. I

also want to thank the Greenhouse staff, especially Kevin, Mike, and Greg for their devoted love to all the plants they house. Their expertise not only helped keep my plants alive, but they helped keep a beautiful teaching collection flourishing. I also wish to extend a big thank you to the Travis Glenn lab for all the support and help with RADcap. Then of course, I wouldn't have completed this program without the aid of our fabulous office staff. During my time here, we've been blessed with some top-notch office help. Chelsea not only are you my go-to for book recs, but you've also helped answer every question I have. Gretchen and Amanda have always been there for the \$\$\$ questions.

Thank you to everyone who has helped me with field work, including Amy Lin, Lingyan Wang, Shu-Mei, my husband, my dad, my brother, my mom and my dog Violet. Of course, this field work would not be possible without the parks and organizations that allowed me to sample plants on their land, and all the contributors to *G. maculatum* records on iNaturalist.org. During my time here I was also fortunate enough to work with five amazing woman scientists while they worked on their undergraduate plant biology or biology degrees here. A big shout out to Amy Lin, Emma Horne, Mackenzi Costly, Deanna Negru and Bina Nechtman.

Then there are the friends I've met along the way. I recall meeting Patrick Smallwood, and his wife, Brittany at my interview weekend. When we meet people, we sometimes don't realize the impact they are going to have on our life. Brittany has become one of the people I'm closest to during my time at UGA. Patrick has been a constant source of aid with my dissertation. He has helped me with R and troubleshooting CTAB. Ashley has been my writing buddy. Kelly, Mandy, Marissa, Julia, and Natalia have been integral to getting sequencing data for Chapter 4. They aided in

troubleshooting benchwork and the program Stacks. There has been no shortage of good people I've met while I've been here. There's been my trivia gang; Patrick, Brittany, Andrew, Jacque, Prince, Jenn and Phil, my stitch'n'bitch group pre-pandemic, my office buddies, the friends I've made while TAing, and participating in activities such as the Athens Science Café and PBGSA.

It takes a village.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
 CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
References	19
2 AUTOMATED PIPELINE FOR SUPERVISED CLASSIFICATION OF PETAL COLOR FROM CITIZEN SCIENCE PHOTOGRAPHS	24
Abstract	25
Introduction	26
Methods	28
Results	34
Discussion	36
References	46
3 VARITATON IN FLORAL COLOR OF THE HARDY GERANIUM, <i>GERANIUM MACULATUM</i> , IS CORRELEATED WITH BIOCLIMATIC VARIABLES	49
Introduction	49
Materials and Methods	54

Results.....	63
Discussion	72
References.....	100
4 PILOT RADCAP STUDY IDENTIFIES LACK OF ISOLATION BY DISTANCE, BUT FINDS ISOLATION BY ENVIRONMENT IN GERANIUM MACULATUM.....	107
Introduction.....	107
Methods.....	111
Results.....	117
Discussion	119
References.....	143
5 Conclusion	151
References.....	150
APPENDICES	
A SUPPLEMENTAL MATERIALS FOR CHAPTER 2	151
B SUPPLEMENTAL MATERIALS FOR CHAPTER 3	157

LIST OF TABLES

	Page
Table 1.1: Patterns of variation across 30 plant studies.....	11
Table 3.1: Multiple linear regression of surveyed and iNaturalist petal data	84
Table 3.2: Linear regression of surveyed populations for pollen	84
Table 3.3: Logistic regression of iNaturalist data for pollen	86
Table 3.4: Confusion matrix for pollen color	88
Table 4.1: RADcap pilot study sample details.....	124
Table 4.2: Plate one pooling guide	125
Table 4.3: Plate two pooling guide	128
Table 4.4: Samples included for analysis	131
Table 4.5: Population metadata.....	134
Table 4.6: Population summary statistics	135
Table 4.7: Pairwise F_{st} values for all populations	136
Table 4.8: Pairwise P_{st} for petal and pollen.....	137
Table 4.9: AMOVA results.....	138

LIST OF FIGURES

	Page
Figure 1.1: Common patterns of geographic variation	18
Figure 2.1: Workflow.....	41
Figure 2.2: Cluster example.....	43
Figure 2.3: Petal morphs of <i>G. maculatum</i>	44
Figure 2.4: Petal morphs of <i>L. parryae</i>	45
Figure 3.1: Swatch and example of flower colors (petal and pollen)	89
Figure 3.2: Map of petal color	90
Figure 3.3: Map of pollen color	91
Figure 3.4: Maxent models for light, medium and dark iNaturalist petal occurrence records.....	92
Figure 3.5: Differences in habitat suitability between the two pollen morphs	93
Figure 3.6: PCA biplot.....	94
Figure 3.7: Shap summary violin plot for pollen color classification.....	95
Figure 3.8: One-way partial dependence plots	97
Figure 4.1: Maps of populations	139
Figure 4.2: STRUCTURE plot	140
Figure 4.3: Plot of isolation by distance	141
Figure 4.4: Plot of P_{st} vs F_{st}	142

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Geographic variation, geographically distributed population-level differences in a species throughout at least a section of its range, is part of a central topic in evolutionary biology: What are the consequences and mechanisms of evolution?

The geographic patterns displayed by an organism can provide insight into how evolutionary processes, both selective and neutral, have shaped variation among landscapes (Kawakami et al., 2011). Clinal variation is among the most common form of geographic variation, often being attributed to natural selection favoring different forms along an environmental gradient (Haldane, 1948; Slatkin, 1973; Endler, 1977; Storz, 2002, Figure 1). Common examples of clines include the famous Gloger's, Bergmann's, and Allen's rule, detailing adaptations along environmental gradients (Gloger, 1833; Bergmann 1847; Allen, 1877). Despite the association of clines with a consequence of natural selection, their presence alone does not provide sufficient evidence. Nonadaptive forces such as drift, gene flow, and migration from a gene origin are also contenders (Haldane, 1948; Endler, 1977; Storz, 2002; Vasemagi, 2006). The presence of a cline, however, from the same trait across different regions against the same environmental/geographic factor suggests evidence of selection (Reinhardt et al., 2014).

Patterns of Variation

A literature search (Table 1.1) in the spring/summer of 2019 was performed using Web of Knowledge, Google Scholar, and personal databases to identify papers in which geographic variation was present. Keywords used included “geographic variation”, “variation maintenance”, “spatial differentiation”, “floral variation”, “phenotypic variation”, “clinal variation”, and similar queries. Papers involving plants were further reviewed. A consensus was drawn on the common patterns of variation seen in plants and compared to common patterns found in two highly cited sources regarding clinal variation, Thorpe (1987) and Futuyma (1986). Neither source is plant-centric, therefore some patterns described are either elusive in nature or found to be uncommon in empirical studies on non-animal species. Figure 1.1 visualizes common patterns of geographic variation in plants found through the literature search. **A** shows a transition or hybrid zone between two morphologies or species with an intermediate contact zone in between. **B** displays a discrete break between two traits or morphologies without a transition zone. **C** demonstrates the traditional clinal pattern of gradual increase in trait. A concordant cline (**D**) is composed of multiple clines along the same axis. **E** shows a ratio/frequency cline, where the frequency of a certain trait, such as the frequency of a certain color, increases over a geographic scale. Finally, **F** displays a mosaic pattern where environmental features could be spread out in a nonlinear fashion, such as wetlands and uplands or xeric versus mesic.

Phenotypic variation

Variation can be found for phenotypic (Arista et al., 2013), physiological (Antonovics, 1971), genetic (Hamrick and Allard, 1972), and life history traits (Kawakami et al., 2011). A common suite of traits of interest to plant biologists is floral

variation. Highly variable floral traits, such as petal and pollen color, present dynamic relationships with biotic (e.g., pollinators, seed dispersals, and florivores) (Lau and Galloway, 2004; Amico et al., 2010; Sobral et al., 2021) and abiotic factors. In addition, they are relatively easy to identify and measure in the field, especially when the traits are separated by large magnitudes, such as shifts from one color category to another.

Polymorphisms in flower color are frequently maintained by direct pollinator selection (Rauscher, 2008). However, a growing body of evidence indicates the biotic environment can be just as important (Berardi et al., 2016; Koski and Galloway 2018, Van der Kooi., 2019; Peach et al., 2020). Understanding floral pigmentation variation presents challenges and can elude researchers; notably with the discordance between findings regarding *Linanthus parryae* and Sewall Wright's idea of isolation by distance (IBD) (Wright, 1943A; B). When variation in flower color was first thought to be caused by drift, as expected by IBD, it ended up being caused by selection due to patchy habitat and precipitation Schemske and Bierzychudek, 2001; 2007 The implication is that understanding whether natural selection, or non-adaptive forces influences floral pigmentation takes multiple approaches, including identifying a pattern, correlating that pattern to the environment (or lack of pattern), and conducting genetic work.

Measuring Phenotypic Variation Against Biotic Variables

The most common method for collecting phenotypic method in wild, native populations is a field survey in which a representative sampling of various populations over the range of a species is characterized for the trait of interest. Color can be measured with prelabeled color swatches. In addition, a growing number of studies utilized citizen science datasets to gather morphological data (i.e., Lehtinen, et al. 2020; Killen et al.,

2022; Rhodes, et al. 2022). These methods combined create a powerful approach to detect patterns in variation across a wide range and depth of sites.

Anthocyanins, a common floral pigment that creates pinks, reds, and purples can be induced by the environment either as a response to the environmental stressors or an adaptive response to provide protection against the stressors (Mierziak et al., 2014; Kim et al., 2017). These pigments can confer a selective advantage as they aid in protection from UV-B rays, (Gould, 2004; Rausher, 2008), heat stress (Rausher, 2008), drought stress (Kovinich et al., 2015), and thermoregulation (Berardi et al., 2016). The role of anthocyanins in these tissues present a potential target of selection against UV-B, temperature, and precipitation related stress.

UV-B (wavelength 280-325nm) is a common instigator of DNA damage and oxidative stress (Ahman et al., 1991; Brosché and Strid, 2002) and anthocyanins have generally been identified as pollen pigments that provide protection against UV-B induced damage (Koski and Galloway, 2018; Peach et al., 2020). Mechanistically, anthocyanins scavenge free radicals and absorb UV radiation (Woodall and Stewart 1998; Bieza and Lois, 2001). Increases in UV have been shown to increase anthocyanin concentration in numerous species and within various tissues, such as leaf (del valle et al., 2015), flower (Peach et al., 2020) and fruit (Zhao et al., 2017), suggesting that UV-B plays a selective role in pigmentation.

Thermoregulation and response to temperature stress can also influence floral pigmentation. Anthocyanins are found to increase in response to oxidative stress, occurring with temperature stress by absorbing visible light to increase temperature (Zhang et al., 2019). Cold temperatures have been shown to influence anthocyanins

concentration in many tissues: leaves (Zhang et al., 2019), seedlings (Christie et al., 1994; Chalker-Scott 1999), fruit (Lo Piero et al., 2005; Gaiotti et al., 2018; Takai et al., 2019) and calyx (Berardi et al 2016). Anthocyanin rich (darker) flowers have been shown to perform better in cold temperatures via warming the reproductive structures and improving reproductive success (Jewell et al., 1994; Lacey et al., 2010; Ida and Totland, 2012). Interestingly, reports have also found that spring ephemerals that regularly experience colder mornings also perform better when under warmer conditions (reviewed in Van der Kooi et al., 2019). However, at least one study showed the opposite: for the two pollen morphs in *Campanula americana*, the purple pollen outperforms white pollen in high temperature (Koski and Galloway, 2018). The mixed findings regarding the function of pollen pigmentation suggests more studies, particularly in species with color polymorphism, are needed before a generality can be made.

Anthocyanins also function as osmoregulators and various species with flower color polymorphism have showed that pigmented morphs perform better in lower water conditions than their non-pigmented counterparts (Grossenbacher et al., 2001; Schemske and Bierzychudek, 2001, Warren and Mackenzi, 2001; Tang and Huang, 2010). However, pollen appears to display a different reaction. For example, in *Nigella degenii* dark pollen was found in areas with reduced drought (Jorgensen and Andersson, 2005), but darker petals tend to be found in areas with low precipitation (Schemske and Bierzhudek, 2001; Warren and Mackenzi, 2001). Indeed, the role of anthocyanins in floral tissues is diverse and the role of pigmentation is important for continued reproductive success. In comparison to the environment, pigmentation, if under selection, should display a relationship.

When combined with environmental data, phenotypic data on floral pigmentation can provide a better understanding of the landscape of pigmentation variation. Correlative tests such as regressions, random forests, PCAs, and Mantel tests, in addition to modeling niche overlap, provides multiple lines of evidence suggesting a role of selection on floral pigmentation.

Neural genetic variation against phenotypic variation

Observing a geographic pattern of variation is not sufficient to elucidate the evolutionary process maintaining it, but it does take a reasonable first step (Endler, 1977). A multitude of ways measure the evolutionary mechanisms that influence populations. This dissertation will focus on methodology that is not constrained by pricier genetic solutions, such as McDonald Kreitman tests that require genes to be sequenced in the species of interest and a related species, as well as maintains affordability for most researchers, and focuses on in-field experiences.

A comparison of the divergence between phenotypic traits (P_{st}) and neutral marker differentiation (F_{st}) has become a popular and widely used metric to differentiate between natural selection and genetic drift across populations. A part of its popularity stems from its ease of use and flexibility. A summary of the potential results is given below (reviewed in Leinonen et al., 2013).

$P_{st} > F_{st}$: Divergent selection

$P_{st} < F_{st}$: Stabilizing selection

$P_{st} = F_{st}$: Drift

Isolation by distance (IBD) and isolation by environment (IBE) are metrics for population differentiation that can help indicate what evolutionary process shapes variation across populations. Isolation by distance was developed by American geneticist Sewall Wright and is described as population differentiation (F_{st}) correlated to geographic distance; the further populations, the more differentiated they should be. Isolation by distance finds its origin in limited dispersal or genetic drift (Wright, 1943A). Populations placed further than the average dispersal distance of an organism will experience random fluctuations in allele frequencies, causing them to diverge in different ways without the ability to homogenize from gene flow.

Isolation by environment is a similar metric, but correlates environmental factors to population differentiation, such that populations in similar environments should be more related than populations in different environments, though both mechanisms may influence this diversity. IBE can occur through various means, including selection against immigrant morphs (an alien enters a new environment and is selected against), reduced hybrid fitness (hybrids perform badly causing a reduction in gene flow), and biased dispersal (in regards to plants this can happen in a few ways; the first is selection against an environment where seeds or pollen disperse, the other way is a bias created by the agent of pollen movement) (Wang et al., 2013 and reviewed in Wang and Bradburd, 2014). These patterns are common within the field of geographic variation, as reviewed in Sexton et al., 2014 by an extensive literature search in which isolation by distance and isolation by environment studies were found and analyzed for the pattern that was most prevalent. Of 70 studies, 26 were on plants. Within plants, eight showed evidence of IBD, five with evidence of IBE, and 10 that supported evidence of both patterns (see Figure 1

from Sexton et al., 2014). It is critical to understand how both these metrics can influence patterns of variation, however, disentangling the effects of both is difficult. Proper sampling schemes necessarily involve breaking correlations between distance and environment, as well as using high-quality geospatial data. Differentiating between IBD and IBE is important because it implicates how genes move between populations, which translates to how adaptation may occur when environments change (Sexton et al., 2014).

Study Species

Geranium maculatum L. is a common understory perennial with a wide distribution spanning latitudinally from Georgia to southern Canada and longitudinally from the east coast to the Great Plains (Marin, 1965). Their flowers possess a great diversity in their pigmentation. Petals range from a pale pink, pink, magenta to deep purple and pollen ranges from various shades of yellow and purple (Martin, 1965, S-M, Chang, personal communication, 2017). The variation in pigmentation appears to be geographically sectioned (S-M, Chang, personal communication, 2017). Populations are composed of protandrous hermaphrodites and female individuals, as this species is gynodioecious. Flowers bloom from early March to August, with peak bloom between April-June, depending on locality (personal obs). Flowering individuals average six flowers per inflorescence with each flower producing up to five seeds that are dispersed a maximum of 4.5 meters by a dehiscing fruit in explosive fashion (Stamp and Lucas, 1983; Chang, 2006). *G. maculatum* is pollinated by an assortment of insects and is considered to suffer from either pollen limitation and/or resource limitation (Bertin and Sholes, 1993).

Dissertation Research

This dissertation will provide us with the first record of a detailed map on the distribution of floral variation in *G. maculatum*. Contrasting with the available reports in the literature indicating that *G. maculatum* generally presents pink petals or displays no geographic variation in flower color (Martin, 1965), data collected with our two-prong approach will outline a comprehensive pattern of floral color distribution. We use two methodologies for surveying floral pigmentation: first we use a conventional field survey and then we include sampling from iNaturalist.org, a frequently visited citizen science website. We develop a unique classification system using these citizen-recorded photographs of *G. maculatum* to classify petal color with a k-means clustering approach. In addition, with the aid of volunteers, we classify pollen color in these images as well.

The method we propose for classifying petal color in iNaturalist can extend to future studies in other species. Future studies can examine color distribution within one species, spatially and temporally, as well as aggregate data across multiple species to access color community in various geographic regions. We hope this tool can be used to gain a broader understanding of regional floral color.

We use this survey data to explore correlations of floral pigmentation to geographic and environmental factors. We use regressions and random forest classifiers to coorelate a latitude, longitude, and elevation, as well as a suite of bioclimatic variables to predict floral pigmentation in both our datasets (survey and iNaturalist). Niche modeling is used to determine niche overlap and habitat suitability for samples of different pigmentation. This project will help correlate phenotypic variation to genotypic variation in *G. maculatum* on a large scale. Phenotypic variation alone will not provide

the opportunity to detect if selection or drift influences floral variation across elevation. Therefore, genetic data will be used in tandem with our survey data to compare populations for signs of drift or selection by using population differentiation metrics isolation by distance (indicative of genetic drift) and isolation by environment (indicative of selection). This study will allow me to use different metrics of population differentiation to better understand how effects selection or drift could impact floral traits across a large area. This is important because plants may respond to the increasing effects of climate change. Understanding the processes that underly floral variation in a common, widespread species could help researchers predict how phenotypic ranges may change. The mechanisms that influence phenotypic maintenance in a population have vastly different consequences when the need for a population to adapt to a changing environment is applied (Sexton et al., 2014).

Table 1.1: Patterns of variation across 30 plant studies. Chart shows species, what trait was found to vary geographically and its pattern as well as what maintains that variation. Species names are taken from the paper and may not represent current nomenclature.

*Subsequent row is from the same study; blanks indicate same text as above.

Citation	Species	Pattern of Variation	Character Influenced	Character Description	Geographic Influence	Agent	Cause	Character Placement
Scouto-Vilaros et al 2018	<i>Ficus trichoncersa subspecies</i>	Discrete break Unknown if hybrid zone	Scent profile	Organic volatile compounds	Elevation	Pollinators	Reproductive isolation caused by pollinator (wasp) selection of their host	
Wang et al 2018	<i>Epimedium pubescens</i>	Ratio monomorph ic green to dimorphic green/yello w	Androecium (anther and pollen) color	Green or yellow	Light intensity, elevation	Abiotic factors: light intensity and elevation	Indirect selection on pollen amount and seed size	Green found in lower light intensity and lower elevation
Costa et al 2016*	<i>Lythrum salicaria</i>	Ratio Concordant with example below	Style morph	3 3 style sizes (L- Long, M- Medium, S- Small	Latitude	Morph incompatibility & pollinators	Negative frequency dependent selection & gene flow from agriculture (small pops.)	Ratios become more event (closer to 1:1:1) with an increase in latitude

		Cline	Population size					Population size increases with latitude
Arista et al 2013	<i>Lysimachia arrensis</i>	Transition zone	Flower color	Blue and red	Latitude	Flowering time, Potentially stress	Reproductive isolation caused by flowering time differences. Potential: indirect selection on floral color	Red in the north, blue in the south
Santos-Galley et al 2013*	<i>Narcissus papyraeus</i>	Ratio	Style morph	2 style sizes (L-Long, S-Small)	Distance from center of species range	Pollinators	Visitation rates of pollinators, possible selection	Long morph frequency increases with increasing distance from the center of the range
		Discrete break	Perianth size	Large or small	Latitude	Unknown	Unknown	Flowers larger in south of range, smaller in north
Arroyo and Dafni 1995	<i>Narcissus tazetta</i>	Mosaic	Style morph	2 style sizes (L-Long, S-Small)	Hill vs marsh habitat	Pollinator efficiency	Reproductive isolation caused by	Small found in marsh with

							Pollinator efficiency	hawkmoth pollinators, L found in hills with short tongued pollinators
Koski and Galloway 2018	<i>Campanula americana</i>	Ratio	Pollen color	White/tan to purple	Longitude	Temperature	Indirect selection on dark pollen color	Dark pollen found in west, light pollen found in east
Koski and Ashman 2015B	<i>Argentia anserina</i>	Cline	UV bullseye	Size of UV bullseyes in middle of flower	Altitude	Light intensity	Potentially indirect selection on UV bullseye size by pollinators' ability to see contrasting color in lower light intensity	Higher altitude has larger bullseyes, lower altitude had smaller
Koski and Ashman 2015A	<i>Argentia anserina</i>	Cline	UV bullseye	Size of UV bullseye in middle of flower	Proximity to equator	UV irradiance	Selection on bullseye size.	Follows Gloger's rule
Preneville et al 2013*	<i>Campanulastrum americanum</i>	Cline Concordant with below	Germination		Latitude	Temperature	Selection on phenology	Earlier in North
		Cline	Fruit maturation					

Berardi et al 2016*	<i>Silene vulgaris</i>	Cline Concordant with below	Calyx color	White or purple	Elevation	Temperature and seasonal rain variation	Selection on flavonoids	White is found at lower elevations, darkens to purple as elevation increases
		Cline	Non-Leaf anthocyanin flavonoid concentration					Increases with elevation
Galen 1989*	<i>Polemonium viscosum</i>	Cline Concordant with below	Flower size		Elevation	Pollinators	Selection by bumblebees and flies	Increases with elevation
		Cline	Height					
Kawakami et al 2011*	<i>Helianthus maximillani</i>	Cline Concordant with below	Days to flower		Latitude	Heterogeneous environment	Selection and gene flow	Days to flower increases as latitude decreases
		Cline	Growth Rate					Increases with latitude
		Cline	Plant Size	Includes height, stem diameter, dry biomass, and more. See paper				Larger in North, smaller in South

Zanetto and Kemer 1995	<i>Quercus petraea</i>	Cline	Allele frequency	13 loci	Longitude		Gene flow, post-glacial recolonization	
Streisfeld and Kohn 2005	<i>Mimulus aurantiacus</i>	Transition zone	Flower color	Red or yellow. Intermediate contact zone	Longitude	Potentially pollinator abundance	Divergent selection	Red in the West, yellow in the East
Ellis and Johnson 2009	<i>Gorteria diffusa</i>	Transition zone	Capitulum and insect mimic spots	14 non overlapping morphotypes	Unknown	Unknown	Reproductive isolation	
Ison et al 2018	<i>Campanula americana</i>	Ratio	Pollen color	White/tan to purple	Longitude	Pollinator preference maintains light morph in East	Direct selection	
Koski et al 2017	<i>Campanula americana</i>	Cline	Autonomy	Selfing fruit set	SE-NW		Post-glacial recolonization	Autonomy increases from South to North and East to West
Lau and Galloway 2004	<i>Campanula americana</i>	Ratio	Pollen color	Tan to purple	Mountain vs Riverside	Fitness consequences caused by low efficiency pollinators	Selection	Mountain population has purple pollen, riverside population has tan pollen
Zhang et al, 2008	<i>Glycine max, glycine soja</i>	Cline	GMCRY1a expression	Controlled in a circadian rhythm	Latitude	Photoperiod	Selection	

Hall et al 2007	<i>Populus tremula L</i>	Cline	Growth Cession	Traits associated with growth cession (i.e., bud flush, leaf abscission)	Latitude	Length of growing season	Selection and gene flow	
Streisfeld and Kohn 2007	<i>Mimulus aurantiacus</i>	Transition zone	Flower color	Red or yellow. Intermediate contact zone	Longitude	Pollinator	Direct selection on flower color	Coastal red, inland yellow
Tastard et al 2008	<i>Antirrhinum majus</i> spp <i>pseudomajus</i> and <i>striatum</i>	Hybrid zone	Flower color	Magenta - <i>pseudomajus</i> Yellow - <i>striatum</i>	Unspecified	Pollinators	Selection, hybrid zone stability	
Galen and Kevan 1980	<i>Polemonium viscosum</i>	Ratio	Flower color	Light blue to purple	Altitude	Pollinators	Visibility against background, direct selection	Teeline has dark background ; flowers are light blue. Apline has light background , flowers purple
Newman et al 2012	<i>Disa ferruginea</i>	Discrete break	Flower color	Red or orange	Longitude	Pollinators	Selection, flower mimicking rewarding plant	West is red, East is orange

Antonovics 1971*	<i>Anthoxanthum odoratum</i>	Discrete break	Zinc tolerance		Soil zinc conc.	Environment heterogeneity	Selection on heavy metal tolerance	
		Cline	Various morphological traits	Ex. flag leaf length, height See paper for more				
Sobral et al 2015	<i>Gentiana lutea</i>	Cline	Flower color	Orange to yellow	Longitude	Pollinators	Selection	West to East, orange to yellow
Hamrick and Allard 1972	<i>Avena barbata</i>	Mosaic	Allele frequency		Xeric or mesic	Environment heterogeneity	Selection	
Warwick and Briggs 1978	<i>Poa annua</i>	Discrete break (One growth form to both)	Growth form	Prostrate or erect	Lawn or flowerbed	Mowing	Disruptive selection	Prostrate are only in lawn, while erect and prostrate are found in flowerbed

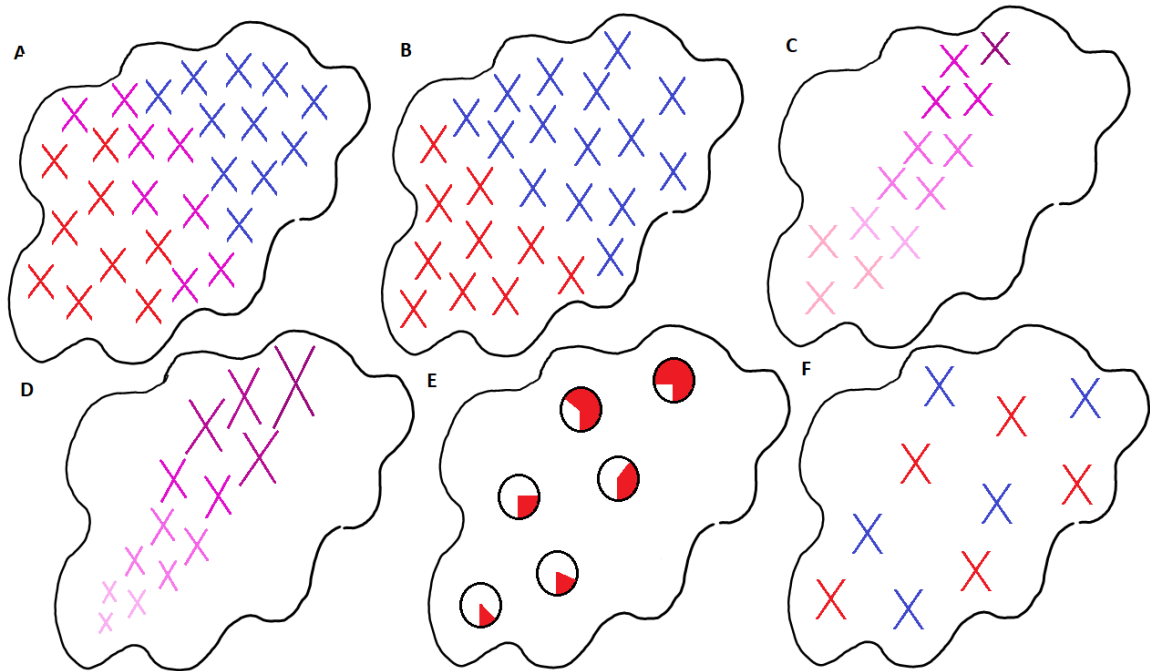


Figure 1.1: Common patterns of geographic variation. A. Transition or hybrid zone. B. Discrete break/Categorical C. Cline D. Concordant Cline E. Ratio/Frequency F. Mosaic

References

- Allen, J. A. 1877. The influence of physical conditions in the genesis of species. *Radical Review*. 1: 108–140
- Amico, G.C., M. A. Rodriguez-Cabal and M. A. Aizen. 2010. Geographic variation in fruit colour is associated with contrasting seed disperser assemblages in a South-Andean mistle. *Ecography*. 34(2):318-326
- Antonovics, J. 1971, The effects of a heterogeneous environment on the genetics of natural populations: The realization that environments differ has had a profound effect on our views of the origin and role of genetic variability in populations. *American Scientist*. 59 (5): 593-599
- 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. *Jour. of Eco*. 101: 1613-1622
- Arroyo, J., and A. Dafni. 1995. Variations in habitat, season, flower traits and pollinators in dimorphic *Narcissus tazetta* L. (Amaryllidaceae) in Israel. *New Phytol*. 129: 135-145
- Berardi, A., P. D. Fields, J. L. Abbate, and D. R. Taylor. 2016. Elevational divergence and clinal variation in floral color and leaf chemistry in *Silene vulgaris*. *Amer. Jour. of Botany*. 103 (8): 1508-1523
- Bergmann, C. 1847. Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Grösse. *Göttinger Studien*. 3 (1): 595–708.
- Costa, J., S. Castro, J. Loureiro, S. C. H. Barret. 2016. Variation in style morph frequencies in tristylous *Lythrum salicaria* in the Iberian Peninsula: the role of geographical and demographic factors. *Annals of Botany*. 117: 331-340
- Endler, J.A. 1977. Geographic variation, speciation, and clines. Princeton University Press. Princeton, New Jersey.
- Galen, C. 1989. Measuring pollinator-mediated selection on morphometric floral traits: bumblebees and the alpine sky pilot, *Polemonium viscosum*. *Evolution*. 43 (4); 882-890
- Galen, C., and P. G Kevan. 1980. Scent and color, floral polymorphisms and pollination biology in *Polemonium viscosum* Nutt. *The Am. Midl. Nat.* 104 (2): 281- 289
- Gloger, C. W. L. 1833. 5. Abänderungsweise der einzelnen, einer Veränderung durch das Klima unterworfenen Farben. Das Abändern der Vögel durch Einfluss des Klimas

- [The Evolution of Birds Through the Impact of Climate] (in German). Breslau: August Schulz. pp. 11–24.
- Gould, K. S. 2004. Nature's swiss army knife: the diverse protective roles of anthocyanins in leaves. *J. Biomed Biotechnol.* 5: 314-320
- Haldane, J.B.S. 1948. The theory of a cline. *J. of Genetics.* 48 (3): 277-284
- Hall, D., V. Luquez, V. M. Garcia, K. R. S. Onge, S. Jansoon, and P. K. Ingvarsson. 2007. Adaptive population differentiation in phenology across a latitudinal gradient in European Aspen (*Populus Tremula*, L.): a comparison of neutral markers, candidate genes and phenotypic traits. *Evolution.* 61 (12): 2849-2860
- Hamrick, J. L., and R. W. Allard. 1972. Mircogeographical variation in allozyme frequencies in *Avena barbata*. *Proc. Nat. Acad. Sci.* 69 (8): 2100-2104
- Ison, J. L., E. S. L. Tuan, M. H. Koski, J. S. Whalen, and L. F. Galloway. 2018. The role of pollinator preference in the maintenance of pollen colour variation. *Annals of Botany.* 1-10
- Kawakami, T., T. J. Morgan, J. B. Nippert, T. W. Ocheltree, R. Keith, P. Dhakal, and M. C. Ungerer. 2011. Natural selection drives clinal life history patterns in the perennial sunflower species, *Helianthus maximiliani*. *Molecular Ecology.* 20: 2318- 2328
- Killen, H., Chang, L., Soul, L. and Barclay, R., 2022. Combining Physical and Digital Data Collection for Citizen Science. Climate Research. *Citizen Science: Theory and Practice.* 7(1): 10
- Koski, M. K., and T. Ashman. 2015A. Floral pigmentation patterns provide an example of Huxley's rule in plants. *Nature Plants.* 1: 14007
- Koski, M. K., and T. Ashman. 2015B. An altitudinal cline in UV floral pattern corresponds with a behavioral change of a generalist pollinator assemblage. *Ecology.* 96 (12): 3343-3353
- Koski, M. K., D. L. Grossenbacher, J. W. Busch, and L. F. Galloway. 2017. A geographic cline in the ability to self-fertilize is unrelated to the pollination environment. *Ecology.* 98 (11): 2930-2939
- Koski, M. K., and L. F. Galloway. 2018. Geographic variation in pollen color is associated with temperature stress. *New Phytologist.* 218: 370-379
- Kovinich, N., G. Kayanja, A. Chanoca, M. S. Otegui, and E. Grotewold. 2015. Abiotic stresses induce different localization of anthocyanins in Arabidopsis. *Plant Signaling and Behavior.* 10 (7): e1027850

- Lau, J.A., and L. F. Galloway. 2004. Effects of low-efficiency pollinators on plant fitness and floral trait evolution in *Campanula americana* (Campanulaceae). *Oecologia*. 141 (4): 577-583
- Leinonen, T., R. J. S. McCairns, R. B. O'Hara, and J. Merila. 2013. Qst-Fst comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics*. 14:179-190
- Lehitnen, R. M., B. M. Carlson, A. R. Hamm, A. G. Riley, M. M. Mullin, and W. J. Gray. 2020 Dispatches from the neighborhood watch: Using citizen science and field survey data to document color morph frequency in space and time. *Ecology and Evolution*. 10(3):1526-1538
- Martin, C. M. 1965. An ecological life history of *Geranium maculatum*. *The Am. Mid. Nat.* 73(1): 111-149
- Newman, E., B. Anderson, and S. D. Johnson. 2012. Flower colour adaptation in a mimetic orchid. *Proc. of the Royal Soc. B*. 279: 2309-2313
- Petit, R. J., J. Duminil, S. Fineschi, A. Hampe, D. Salvini, and G. C. Venramin. 2005. Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol. Ecol.* 14:689-701
- Preneville, H. R., K. Barnard-Kubow, C. Dai, B. C. Barringer, and L. F. Galloway. 2013. Clinal variation for only some phenological traits across a species range. *Oecologia*. 173: 421-430
- Rausher, M.D. 2008. Evolutionary transitions in floral color. *Int. J. Plant Sci.* 169 (1): 7-21
- Reinhardt, J.A., B. Kolazkowski, C. D. Jones, D. J. Begun, and A. D. Kern. 2014. Parallel geographic variation in *Drosophila melanogaster*. *Genetics*. 197 (1): 361-373
- Rhodes, C. R., W. Haunfelder, B. E. Carlson. 2022. Citizen science reporting indicates geographic and phenotypic drivers of road use and mortality in a threatened rattlesnake. *Current Zoology*. XX, 1-13
- Santos-Galley, R., R. Perez-Barrales, V. I. Simon, Arroyo. 2013. The role of short-tongued insects in floral variation across the range of a style-dimorphic plant. *Annals of Botany*. 111: 317-328
- Schemske, D. W., and P. Bierzychudek. 2001. Perspective: evolution of flower color in the desert annual *Linanthus parryae*: Wright revisited. *Evolution*. 55 (7): 1269-1282

- Schemske, D. W., and P. Bierzychudek. 2007. Spatial differentiation for flower color in the desert annual *Linanthus parryae*: was wright right? *Evolution*. 61 (11): 2528-2543
- Scouto-Vilaros, D., M. Proffit, B. Buatois, M. Rindos, M. Sisol, T. Kuyaiva, B. Isua B, and et al. 2018. Pollination along an elevation gradient mediated both by floral scent and pollinator compatibility in the fig and fig-wasp mutualism. *J. of Ecology*. 106: 2256-2273
- Sexton, J. P., S. B. Hangartner, A. A. Hoffmann. 2014. Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*. 68 (1): 1-15
- Sholes, R. I. B., and O.D. V. Sholes. 1993. Weather, pollination and the phenology of *Geranium maculatum*. *The Am. Mid. Nat.* 129(1):52-66
- Slatkin, M. 1973. Gene flow and selection in a cline. *Genetics*. 75 (4); 733- 756
- Slatkin, M. 1985. Rare alleles as indicators of gene flow. *Evolution*. 39 (1): 53-65
- Sobral, M., T. Veiga, P. Dominguez, J. A. Guitian, P. Guitan, J. M. Guitian. 2015. Selective pressures explain differences in flower color among *Gentiana lutea* populations. *Plos One*. 10 (7): 0132522
- Stamp, N. E., and J. R. Lucas. 1983. Ecological correlates of explosive seed dispersal. *Oecologia*. 59: 272- 278
- Storz, J. F. 2002. Contrasting patterns of divergence in quantitative traits and neutral DNA markers: analysis of clinal variation. Jay F. Storz Publications. 16.
- Streisfeld, M. A, and J. R. Kohn. 2005. Contrasting patterns of floral and molecular variation across a cline in *Mimulus aurantiacus*. *Evolution*. 59 (12): 2548-2559
- Streisfeld, M. A., and J. R. Kohn. 2007. Environment and pollinator-mediated selection on parapatric floral races in *Mimulus aurantiacus*. *J. of Evol. Biol.* 20 (2007): 122-132
- Tastard, E., C. Andalo, M. Giura, M. Burrus, and C. Thebaud. 2008. Flower color variation across a hybrid zone in *Antirrhinum* as perceived by bumblebee pollinators. *Arthropod-Plant Interactions*. 2: 237-246
- Thorpe, R. S. 1987. Geographic variation: a synthesis of cause, data, pattern and congruence in relation to subspecies, multivariate analysis and phylogenesis. *Italian J. of Zoology*. 54 (1): 3- 11
- Vasemagi, A. 2006. The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*. 173: 2411-2414

- Wang, I. J., and G. S. Bradburd. 2014. Isolation by environment. *Molecular Ecology*. 23: 5649 - 5662
- Wang I. J., R. Glor, and J. B. Losos. 2013. Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology letters*. 16: 165- 182
- Wang, Z., Q. Quan, B. Wang, Y. Li, and S. Huang. 2018. Discovery of androecium color polymorphism in *Epimedium pubescens* with habitat preference or anther/pollen color in the genus. *J. of Plant Ecol.* 11 (4): 533- 541
- Warwick, S. I., and D. Briggs. 1978. The genecology of lawn weeds. *New. Phytol.* 81: 725- 737
- Wright, S. 1943A. Isolation by distance. *Genetics*. 28 (2): 114-138
- Wright, S. 1943B. An analysis of local variability of flower color in *Linanthus parryae*. *Genetics*. 28 (2): 139-156
- Zanetto, A., and A. Kremer. 1995. Geographical structure of gene diversity in *Quercus petraea* (Matt.) Liekl. I. Monolocus patterns of variation. *Heredity*. 75: 506-517
- Zhang, Q., H. Li, R. Li, R. Hu, C. Fan, F. Chen, Z. Wang, and et al. 2008. Association of the circadian rhythmic expression of GmCRY1a with a latitudinal cline in photoperiodic flowering of soybean. *PNAS*. 105 (52): 21028 - 21033

CHAPTER 2

AUTOMATED PIPELINE FOR SUPERVISED CLASSIFICATION OF PETAL COLOR FROM CITIZEN SCIENCE PHOTOGRAPHS¹

¹ Perez-Udell, R. A., Udell, A. T., and S-M. Chang. Submitted to *Applications in Plant Sciences*, June 2022

Abstract

- *Premise of the study:* Petal color is an ecologically important trait and uncovering color variation over a geographic range, particularly in species with large distributions and/or short bloom times, requires extensive fieldwork. We have developed a method to segment images from citizen science repositories using Python and *k*-means clustering in the Hue-Saturation-Value (HSV) color space.
- *Methods:* Our method uses *k*-means clustering to aggregate like-color-pixels in sample images to generate the HSV value space encapsulating the color range of petals. Using the HSV values, our method isolates photographs containing clusters in that range, and then bins them into a classification scheme based on user-defined categories.
- *Results:* We demonstrate the application of this method using two species: one with a continuous range of variation of pink-purple petals in *Geranium maculatum*, and second with a binary classification of white versus blue in *Linanthus parryae*. We demonstrate results that are repeatable and accurate.
- *Discussion:* This method provides a flexible, robust, and easily adjustable approach for classification of color images from citizen science repositories. By using color to classify images, this pipeline sidesteps many issues encountered using more traditional computer vision applications. This approach provides a tool for making use of large citizen scientist datasets.

Introduction

Variation in floral pigmentation is an ecologically important trait frequently studied in the context of plants' response to biotic and abiotic factors, demographic changes, population structure, and community interactions (Endler, 1977). It is particularly intriguing when a species shows nonrandom variation, such as clines, patches, and other gradational patterns, across a geographic range. A well-known case study of corolla color variation among populations of *Linanthus parryae* in western North America has been used as a prime example for the population genetics concept of isolation by distance described by the renowned geneticist, Sewall Wright (Wright, 1943, though see alternative conclusions by Schemske and Bierzychudek, 2007). Additional examples of flower pigmentation revealed natural selection following secondary contact hybrid zones (*Phlox*, Hopkins et al., 2012), and balancing selection imposed by pollinators (*Ipomoea*, Fry and Rausher, 1997; *Raphanus*, Irwin and Strauss, 2005; and *Clarkia*, Eckhart et al, 2006) among others. Despite the rich evolutionary and ecological insight provided by patterns of variation over geographical scales, there are relatively few cases of detailed description available – partly because characterizing natural variation at a large geographic level can be difficult. It not only requires a great amount of time and effort, but also includes other challenges inherent to field studies. Short bloom times in spring ephemerals, inclement weather during field seasons, long distances between populations, scarcity of funding for fieldwork, and travel restrictions all pose additional barriers to the conventional field surveys used to collect such data. As a result, researchers have sought alternative approaches to alleviate some of these challenges.

Recent studies using easily accessible, and often large, databases generated by ever-increasing digitization efforts present an exciting alternative to field surveys. Digitization of herbarium specimens has made collections spanning large temporal and spatial scales widely accessible to users through portals like the SouthEast Regional Network of Expertise and Collections (<https://sernecportal.org/portal/>). These digital images greatly improved our ability to take measurements on morphological traits and phenological shifts (Willis et al., 2017; Espinosa and Castro, 2018). Herbarium records have also been used in studies of floral pigmentation patterns, but many pigments degrade over time in pressed specimens (Koski et al., 2020), rendering these datasets less suitable for some topics.

Another form of digitized data is the image records found in the photograph repositories such as iNaturalist.org and the Global Biodiversity Information Facility (<https://www.gbif.org/>). With the popularity of smartphones and wide coverage of wireless networks, these virtual collections taken by citizen scientists contain floral pigmentation of fresh samples collected from a large number of georeferenced locations, sidestepping the problems of degrading pigments in the dried herbarium samples. Additionally, with the popularity of mobile apps like iNaturalist and Seek (iNaturalist.org), aided by community curation and artificial intelligence algorithm in species identification, the number and quality of these image records increases steadily each year. Indeed, several recent studies clearly showed the power of extracting the occurrence data from these citizen science collection (Li et al., 2019; Mancinelli et al., 2021; Mesaglio et al., 2021). However, obtaining morphologic data from these records, often requires manually sifting through the large number of photographs. While feasible,

the time and effort required can potentially delay research progress and may introduce biases or inconsistency due to subjectivity from the researchers on traits like pigmentation levels (Webster et al., 2002; Emery et al., 2017).

To improve the evaluation process on utilizing photographs taken by citizen scientists, we developed a simple computer vision methodology to automate the classification of petal pigmentation in large photograph datasets. This method couples the Anaconda distribution of Python, a programming language that is easily customizable for mathematics and image processing, and Jupyter Notebooks, an integrative development environment, to generate a pipeline with four scripts. It uses a relatively low amount of computer storage at ~5,000 records stored in just over half a gigabyte. It is well suited for large floral structures that display either a gradient in color variation or separate, discrete colors, and works for both *a priori* and *de novo* color classification. We include two examples, one with a continuous color gradient in petals and the other with discrete binary color morphs, to illustrate the range of usage.

Methods

This analysis pipeline we have developed downloads and analyzes photographic records using four Python scripts formatted as Jupyter notebooks (*Downloader*, *Color Cluster Visualizer*, *Data Collector* and *Classifier* (Figure 2.1)). The notebooks, a quick start guide, and example data sets are freely available at (<https://github.com/atudell/Color-Cluster-Kit>). The examples we include use two flowering plant species and their image records from iNaturalist, though our pipeline can also work with similar photographic data from other sources.

The pipeline described consists of four scripts. The first script, “*Downloader*,” downloads and names images onto the user’s computer. The second script, “*Color Cluster Visualizer*,” examines pixel colors of a subset of images and identifies major color clusters. The same clustering method is used in the third script, “*Data Collector*,” as a form of image segmentation to isolate the flower from the background for each image and to return summarized pixel values describing its colors. The fourth script, “*Classifier*,” puts images into color bins based on user-specified rules (Appendix S2.1-S2.8; see Supporting Information with this article)

To demonstrate how this pipeline can be used for both continuous and discrete color variation, we showcase implementations using *Geranium maculatum* and *Linanthus parryae*, respectively. We use the Hue-Saturation-Value (HSV) model to describe the color (H), the amount of gray in the color (S), and the brightness of the color (V) (Smith, 1978).

Data Acquisition

In this study, we use digitized photographs, such as those from iNaturalist.org, GBIF.org, or other similar archives of citizen scientist submitted data. Photograph retrieval is done with the *Downloader*, which automates the downloading and naming of photographs. As written, the *Downloader* works for iNaturalist but can be modified to use for any other database. Mechanically, the *Downloader* loads a specific web page for the displayed image and interacts with the HTML to download it. Minimal changes to the URL scheme would be needed to obtain similar results with other data archives, provided these archive uses a similar structure. Because the method determines color on a pixel-by-pixel basis, this pipeline is flexible in terms of the overall photograph quality. For

example, since neither shape nor prominence of the key feature is critical in our color clustering, our pipeline can be used on flower images that are blurry or off centered without severe impairing the analysis, Additional data qualities can be specified depending on requirements of individual research projects (see examples below).

Data Preparation

The second script, *Color Cluster Visualizer*, helps the user determine the number of clusters needed to distinguish the color(s) of the target feature from the background. It also derives an estimate for the typical ranges of HSV for these segments as part of the output. The script uses *k*-means clustering, a statistical method that groups like-data points, to explore the pixel properties within each image. Based on a user specified parameter *k* (the number of clusters), the pixel colors are placed into *k* clusters. The script then outputs *k* partial images, each highlighting only the pixels corresponding to one particular cluster and its average H-, S-, and V-values, as highlighted in Figure 2.2.

A main goal of the *Color Cluster Visualizer* script is to find the optimal value for *k*, defined as the smallest cluster number that allows clear separation of petal from the background. Too big a *k* will increase the computational time without adding additional benefit to the output, but too small a *k* will not properly segment the flower from the background. In this pipeline, the optimal *k* needs to be determined using a small representative subset of images that encapsulates the color variation of the target feature. The *Color Cluster Visualizer script* will be run iteratively with the *k* initially changing in increments of 5 as coarse adjustment and then fine tuning in smaller increments at the user's discretion. An optimal *k* results in all pixels of the target feature (a flower in this

case) falling entirely within one or a set of clusters. After iterating through the subsample of photos, a value of k should be estimated.

The other main goal of *Color Cluster Visualizer* script is to determine the range of HSV encapsulating color variation in the target feature that is captured by more than one cluster. For example, if a flower is captured in multiple clusters, one petal in its own cluster and the rest in a separate cluster, the HSV values in all clusters in which the feature appears should be noted and factored into the estimated HSV range used in the *Data Collector* script.

The third script, *Data Collector*, performs the same k -means clustering for every photograph in the data set and aggregates clusters whose averages fall within the HSV range determined in the *Color Cluster* step. Under ideal conditions, the HSV range will ensure that only the pixels associated with the feature of interest are captured. Consequently, images with no pixels captured are inferred as missing the feature of interest, and, in our case, are labeled as “non-flowering.” The output returns a set of summary statistics for each photograph that describes the color profile of the captured pixels. The decision of which summary statistics to output is species specific. Average H, S, and V values provide a good baseline description of the colors and have been set as the default in the script, but the script may be modified to report any other summary statistics to aid in classification of color. For example, the color continuity found in *G. maculatum* suggested that a good way to bin the petal color is by the average values of S, whereas the distinct blue and white flowers in *L. parryae* only required a count of pixels in the (blue) color range. We present this in more detail in the Results section.

Data Cataloging

The final script, *Classifier*, sorts the photographs into their respective clusters, or bins, for subsequent analysis. The details of this script are dictated by the user's wishes on how to classify their data for later analyses and will, therefore, need to be developed by individual users. Here, we illustrate how we customize the final step to specifically fit our two examples. When color bins are established *a priori* (such as light, medium, and dark in our *G. maculatum* example, or white versus blue in our *Linanthus* example), a classification scheme may be developed by a simple algorithm with a predefined set of rules. The example section provides two basic classification schemes. *Geranium maculatum* utilizes the saturation value as a percentile of the sample population, while *L. parryae* looks for a large number of blue pixels in the image.

Geranium maculatum specifics

We obtained the records for *G. maculatum* from iNaturalist with the designation of 'research-grade', defined as from a wild location, georeferenced, and with verified identification (Appendix S2.9). A total of 5,049 'medium' resolution photographs were downloaded using our *Downloader* script. Additional curation removed 47 records that were either posted outside of the possible flowering time (i.e., between September and February) or registered in locations missing environmental data from WorldClim (our subsequent analysis in other studies). The photographs took a total of 576MB, which is relatively small for the average consumer-grade laptop, enabling any machine to conduct this type of analysis.

Geranium maculatum is a species with continuous variation in floral color that ranges from light pink to magenta and deep purple. In previous studies, we developed a physical color-swatch that encompassed the range of observed petal colors to allow easy

scoring of this continuous variation during field surveys. To create a scoring system for the iNaturalist images comparable to the swatch method, we set the parameters for color clustering in the pipeline to replicate the swatch scales and preselected three color bins for classification (light, medium and dark) (Figure 2.3). Using a Samsung Galaxy S10 camera, we took images of greenhouse grown *G. maculatum* flowers (143) whose color had been designated using the swatch method. The greenhouse images were then processed using the *Color Cluster Visualizer script* and an estimated value of $k = 25$ was found to best separate flower(s) from their respective backgrounds. In addition, the HSV averages of the cluster(s) in which the flower(s) appeared were noted and used to estimate a HSV range (H: 123-157, S: 15-255, V: 0-255). This initial step revealed that the variety of colors found in *G. maculatum* fell within a small range of hue and the majority of variation in colors observed was from color saturation.

The estimates for k and HSV were entered into the *Data Collector*, the averages of H, S, and V were obtained, and the script was run using photographs downloaded from iNaturalist. Any photographs that did not return a cluster within the HSV range were labeled as “no flowers.” Given the number of observations and the value of k , the script took approximately 10 hours to complete and return a file detailing each observation and its summary statistics. Looking again at the images from the greenhouse to calibrate our results, we developed a simple binning system where any flower with an S-value below 70 were “Light”, above 92 were “Dark”, and anything in-between as “Medium.” We then applied this simple system in the *Classifier script* and made the final color classification for each observation in the iNaturalist dataset.

Linanthus parryae specifics-

Linanthus parryae displays two flower color morphs: blue and white corolla (Figure 2.4). The pipeline for this species includes first describing the HSV space of the blue seen in *L. parryae* and then determining a metric in which to classify a flower. Records from iNaturalist were obtained in the same manner as for *G. maculatum*. The *Downloader* collected 282 photographs, utilizing 38.9 MB of space (Appendix S2.10). All research grade occurrences of this species contain at least one flower though some contain both color morphs. We removed the multi-flower color images and retained 224 images in our final dataset, each having either white or blue corolla.

Results

In the continuous color example, we observed that *G. maculatum*'s petal pigmentation ranges from light pink to dark purple. Preliminary analysis showed that this petal color variation is largely explained by the change in S value, with the H and V values showing relatively less variation among individuals (Appendix S2.11). In our discrete color example, *L. parryae* presents either blue or white petals that differ in all three metrics of H, S, and V.

Application of the pipeline: Example 1 – petal pigmentation of Geranium maculatum

We found that the key metric that covaried with our *G. maculatum* swatch scales was the saturation values, with the H and V showing relatively minimal variation. From this pilot study, we chose S= 78 and S=92 (in the scale of 0-255) as the criteria to separate our Light, Medium, and Dark bins. These values were chosen arbitrarily to correspond to existing categories on the swatch used in our previous field studies.

Out of the 5,004 input photographs, 818 (16.3%) occurrences were classified as non-flowering. We validate this script by manually examining a random subsample of 500 non-flowering images and found that 87.0% were classified correctly (Appendix S12). Similarly, we also validated a random subset of 500 images that were classified as ‘flowering’ and yielded a 92.8% accuracy. Of the 36 records classified as having flowering incorrectly, 30 of them were classified as light due to leaf litter.

Within the 4,189 records labeled as flowering, 2,874 were classified as ‘light’, 832 as ‘medium’, and 480 as ‘dark’. While validation of this more nuanced classification can be difficult to do manually, we attempted to compare how manual evaluation of the color using the swatch differ from the classification of the same images using the reported pipeline. To do so, all three authors of this paper evaluated a subsample of 150 images independently by using the same swatch scales used for field surveys. We unanimously agreed on a classification for approximately 73% of the images, highlighting the inherent human variability in identifying the nuances in color variation, as well as the likely variation in computer monitor display. Of the observations for which there was unanimous agreement using manual scoring, the algorithm gave the same score 85.3% of the time. Among the observations that did not receive a unanimous vote, the algorithm always agreed with at least one of the scorers.

Application of the pipeline: Example 2 – petal pigmentation of Linanthus parryae

The algorithm required for a binary color classification in this species is simpler than the *G. maculatum* case; we only needed to accurately recognize one morph and assign the remainders to the other morph. We chose blue due to the rocky white background in most photographs. We used 50 images of blue flowers to derive an

estimation of its HSV color range for the *Data Collector* script. To accommodate the fact that a small portion of white flowers contain purple pigments in their corolla tubes (as shown in Fig 4), we modified the *Data Collector* to exclude photographs with $< 3,000$ blue pixels. The 3,000 figure was empirically determined based on the actual values observed in the images and only meant as a cut-off measure specific to this species. We chose the absolute number, rather than %, of pixels in the images, because the photos varied in resolution and the flowers varied in their prominence within the photos. This value successfully distinguished the white flowers with the blue corolla tube from blue flowers because the latter contain a significantly larger amount of blue pixels. Our pipeline predicted the presence of a blue flower with a precision of 91.5%. There were no images in this dataset without flowers.

Discussion

We present in this paper an efficient, reliable, and automated color classification system that can allow researchers to tap into the wealth of citizen-contributed images available in public databases that are currently under-utilized. The use of a computer algorithm greatly reduces the time and effort required to score large datasets and also provides a much more consistent method of classification than conventional manual-scoring. The color identification and classification methods used in our pipeline are flexible, can generate discrete and unambiguous classifications, and can be used with floral features that display continuous variation. Finally, this method also allows users to easily modify the analysis scripts to fit their needs for subsequent analyses. These qualities make the method reported here a great improvement from the existing method

of hand classifying for documenting and classifying floral pigmentation at large geographical scales.

Our method provides an easy interface with citizen science data and returns respectable results. While citizen scientists offer a trove of curated and georeferenced records, the pictures are often difficult for a computer to decipher, because they are taken from various angles, with varying amounts of prominence in the image, with variable number of species present, using a variety of camera types, and under many different lighting conditions. Employing image segmentation by color alone alleviates most of these concerns. Simply by searching the image for pixels within a target color range and ignoring other traditional hallmarks of computer vision, such as shape and texture (Nilsback and Zisserman, 2008; Guru et al., 2010; Hsu et al., 2011), we can sidestep the issue of images taken from various angles and distances. In addition, it is well recognized that humans have variable ability to perceive and distinguish saturation of colors (Cooper et al., 1991; Webster et al., 2002; Emery et al., 2017.). Breaking down an image into its pixel components and analyzing the color mathematically, our pipeline helps to avoid the problems of intrinsic biases or inconsistency introduced when manually scoring colors, as was demonstrated by the moderate level of agreement among the three authors scoring the same set of images.

The conversion of color data into the HSV space in our methodology addresses a major issue for color analysis using iNaturalist images, which were often taken under variable lighting. In traditional RGB analysis systems, changes in illumination alter the values of R, B, and G in a discontinuous and non-linear manner (Chavolla et al. 2018). Other existing floral classification systems avoid this problem by using curated datasets

that employ high-quality images taken with prominent blooms and/or fruit from consistent angles and reliable lighting conditions (Blasco et al., 2009; Singh et al., 201; Rosyani et al., 2018). Our approach enables images taken in the “wild”, such as those from iNaturalist to be used for analysis by using the HVS color space. In the HSV color space, changes in illumination primarily affect V-values and minimally impact H- and S-values (Chavolla et al. 2018). As a result, our approach allows images to be used even when taken under different light condition and at different angles.

Another advantage of our method is its simplicity and flexibility compared to other approaches commonly used in computer visualization. Many other machine learning approaches attempt to classify floral images use texture and shape in addition to color. Hence, these methods are often computationally expensive, difficult to understand and implement for non-experts, and unnecessary for simple color classifications (Nilsback and Zisserman, 2008; Guru et al., 2010). Cutting edge computer vision approaches, such as neural networks, generally also require manual drawing of bounding boxes for large training and testing datasets to properly calibrate deep learning (Venegas et al., 2021). Finally, for especially large datasets, specialized equipment, such as multiple graphics processing units GPUs, which are often expensive, may be required (Van Horn et al., 2018). Though these other approaches may produce complex measurements of target features, the stringent requirements and complexity of these methods make them less accessible to many researchers.

Our pipeline simplifies analysis by focusing only on the color range of the target feature, requires minimal manual labor, and can be performed quickly on consumer grade computers with large numbers of medium resolution images. Though we showed two

simple methods of predetermined classification in our examples, the data generated from the first three scripts can be readily fed into other more complex classification methods such as Support Vector Machines or Random Forests (Cortes and Vapnik, 1995; Breiman, 2001) with a predetermined number of color-bins. If there is no prior knowledge or preference for the number of bins, one can use unsupervised machine learning techniques such as OPTICS or DBSCAN (Ester et al., 1996; Ankerst et al., 1999) to create color bins using the HSV data collected in the “*Data Collector*” script. Discussion of these approaches is beyond the scope of this protocol, but information can be found in Kanagala and Krishnaiah (2016).

Even though our method provides many strengths for working with citizen scientist data, it does have some limitations. For example, this method works best for large, conspicuous blooms that display high contrast relative to the background. Small, inconspicuous blooms, or blooms that blend in with their background, would generally require additional image segmentation and machine learning. However, in some cases, creative workarounds can be established, as illustrated in our *L. parryae* example in which we used blue pixels, showing that it is possible to use the flexibility of our scripts against this limitation.

A second limitation is that, by ignoring shape, our algorithm cannot distinguish between separate bodies with the same color and will aggregate them together. Consequently, the presence of a second species within the same color range, other flowers of the same species but with different color profiles, or even another object in the photo with a similar color will skew the resulting data. For example, in the case of *G. maculatum*, if both dark and light morph flowers were present in the same image, the

algorithm would combine and average out their colors, giving a result that doesn't accurately convey the colors of the individual flowers. Similarly, our algorithm will average differences in coloration patterns present within the same flower, such as dark veins or a dark inner corolla, even if human judgement often places a higher degree of priority to the edges.

Lastly, our algorithm is not able to account for the other factors that can affect the quality of citizen scientist photographs such as additional objects in photographs (people, trash, keys fobs, etc.), a lack of standardization of cameras, and photomanipulation, which can potentially skew color perception. For example, discrepancies in how various cameras capture color can distort the results and many smartphone cameras, while incredibly accessible to citizen scientists, also allow the option of applying filters to an image which may disrupt the ability to capture the natural color of a plant. To this extent we provide recommendations for taking photographs for this type of usage in the appendix (Appendix S2.13). This can guide the creation of new datasets or broaden the viability of existing citizen scientist datasets.

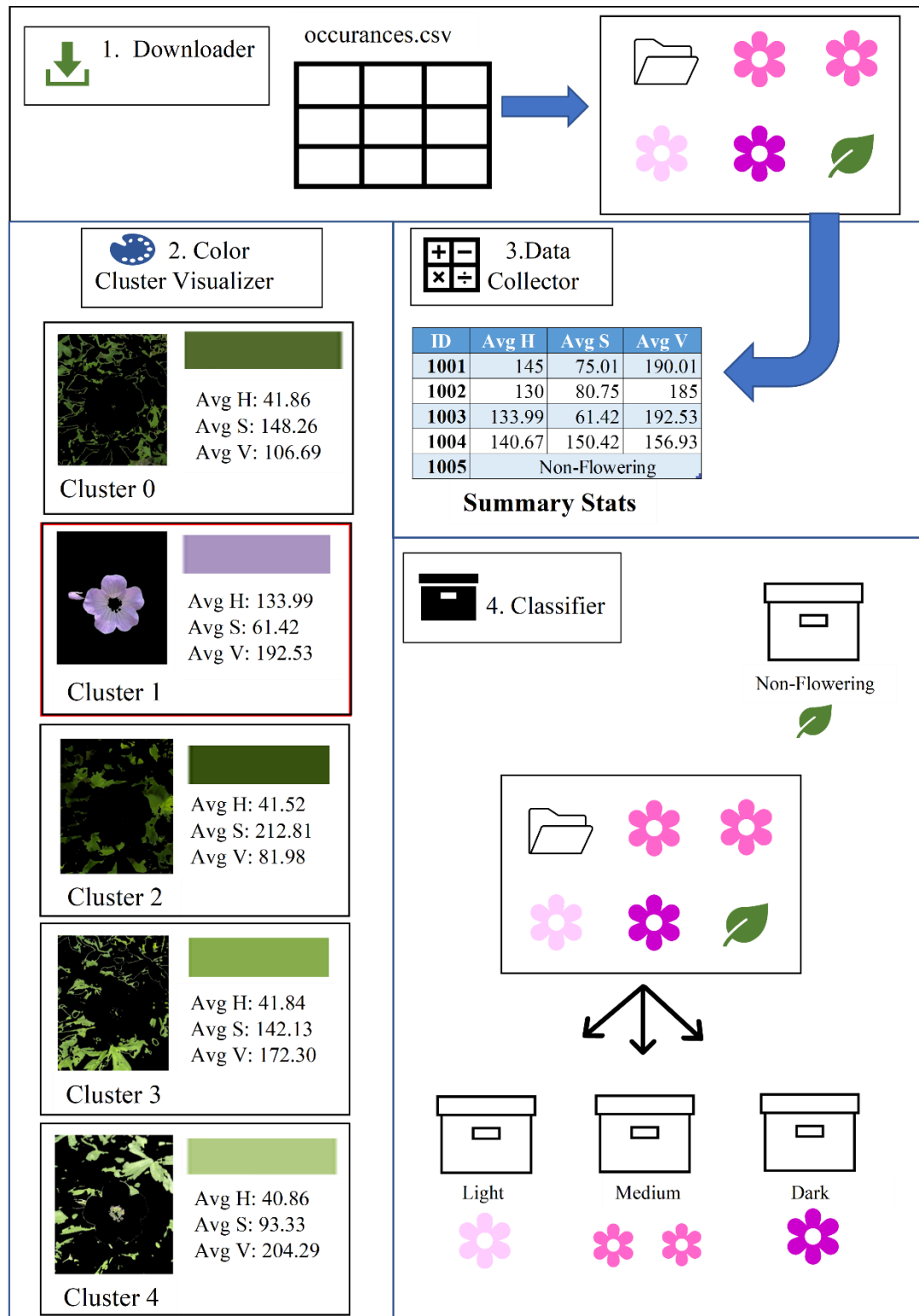


Figure 2.1: Workflow. Script 1 downloads jpegs from a CSV file. Their file names will match that of the sample ID. Script 2 creates images cluster within photographs for visualization. Replicating across variation gives an HSV range and k to consider. Script 3 repeats script 2 for all photographs and outputs a CSV file with the summary stats only for clusters that fall within the ranged defined with the help of Script 2. If no cluster in range exists for a photograph the output is 'non-flowering'. Script 4 takes the summary stats and bins them according to user input.



Figure 2.2: Cluster example. Original photos (top) vs isolated color cluster of interest (bottom) from white (left) and blue (right) *L. parryae*.



Figure 2.3: Petal morphs of *G. maculatum*. From left to right: light, medium, dark.



Figure 2.4: Petal morphs of *L. parryae*. From right to left, white morph with blue inner corolla, white morph, blue morph.

References

- Ankerst, M., M. M. Breunig, H. Kriegel, and J. Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD International Conference on management of Data*. ACM Press. 49-60
- Blasco, J., S. Cubero, J. Gomez-Sanchis, P. Mira, and E. Molto. 2009. Development of a machine for the automatic sorting of pomegranate (*Punica granatum*) arils based on computer vision. *Journal of Food Engineering*. 90(1):27-34
- Breiman, L. 2001. Random Forests. *Machine Learning*. 45:5-32
- Chavolla, E., D. Zaldivar, E. Cuevas, and M.A. Perez. 2018. Color Spaces Advantages and Disadvantages in Image Color Clustering Segmentation. In: A. Hassanien, D. Olivia [eds.] *Advances in Soft Computing and Machine Learning in Image Processing* 730, 3-22. Springer International Publishing.
- Cooper, B.A., M. Ward, C. A. Gowland, and J. M. McIntosh. 1991. The use of lanthony new color test in determining the effects of aging on color vision. *Journal. of Gerontology*. 46(6): 320-324
- Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning*. 20, 273-297
- Eckhart, V.M., N. S. Rushing, G. M. Hart, and J. D. Hansen. 2006. Frequency-dependent pollinator foraging in polymorphic *Clarkia xantiana* ssp. *xantiana* populations: implications for flower colour evolution and pollinator interactions. *Oikos*. 112(2):412-421
- Emery, K.J., V. J. Volbrecht, D. H. Peterzell, and M. A. Webster. 2017. Variations in normal color vision. VI. Factors underlying individual differences in hue scaling and their implications for models of color appearance. *Vision Research*. 141: 51-65
- Endler, J.A. 1977. Geographic variation, speciation and clines. *Monographs in Population Biology*. 10(3):1-246
- Espinosa, F., and M. P. Castro. 2018. On the use of herbarium specimens for morphological and anatomical research. *Botany Letters*. 165 (3-4): 361-367
- Ester, M., H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press. 226-231
- Fry, J., and M. D. Rausher. 1997. Selection on a floral color polymorphism in the tall morning glory (*Ipomoea purpurea*): transmission success of the alleles through pollen. *Evolution*. 51(1): 66-78
- Guru, D.S., Y. H. Sharath, and S. Manjunath. 2010. Texture Features and KNN in Classification of Flower Images. *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition"*. RTIPPR.

- Hopkins, R., D. A. Levin, and M. D. Rausher. 2012. Molecular signatures of selection reproductive character displacement of flower color in *Phlox drummondii*. *Evolution*. 66:469-485.
- iNaturalist. Website. from <https://www.inaturalist.org>.
- Irwin, R.E., and S. Y. Strauss. 2005. Flower color mircoevolution in Wild Radish: evolutionary response to pollinator-mediated selection. *The American Naturalist*. 165 (2): 225-237
- Kanagala, H.K., and V. V. J. R. Krishnaiah. 2016. A comparative study of k-means, DBSCAN and OPTICS. 2016 *International Conference on Computer Communication and Informatics (ICCCI)*. 1-6
- Koski, M.H., D. MacQueen, and T. Ashman. 2021. Floral Pigmentation Has Responded Rapidly to Global Change in Ozone and Temperature. *Current Biology*. 30 (15). 4425-4431.e3
- Li, E., S. S. Parker, G. S. Pauly, J. M. Randall, B. V. Brown, and B. S. Cohen. 2019. An urban biodiversity assessment framework that combines an urban habitat classification scheme and citizen science data. *Frontiers in Ecology and Evolution*. 7: 277
- Mancinelli, G., R. Bardelli, and A. Zenetos. 2021. A global occurrence database of the Atlantic blue crab *Callinectes sapidus*. *Scientific Data*. 8: 111
- Mata-Montero, E., and J. Carranza-Rojas. (2016) Automated Plant Species Identification: Challenges and Opportunities. In: Mata F., Pont A. (eds) ICT for Promoting Human Development and Protecting the Environment. WITFOR 2016. IFIP Advances in Information and Communication Technology, vol 481. Springer, Cham
- Mesaglio, T., A. Soh, S. Kurniawidjaja, and C. Sexton. 2021. 'First known photographs of living specimens': the power of iNaturalist for recording rare tropical butterflies. *Journal of Insect Conservation*. 25:905-911
- Nilsback, M., and A. Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. Sixth Indian Conference on Computer Vision, Graphics & Image Processing.
- Rosyani, P., M. Taufik, A. A. Waskita, and D. H. Apriyanti. 2018. Comparison of color model for flower recognition. 2018 *3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*.10-14
- Schemske, D.W., and P. Bierzychudek. 2007. Spatial differentiation for flower color in the desert annual *Linanthus parryae*: was Wright right? *Evolution*. 61 (11):2528-2543
- Singh, S., D. Dhyani, A. K. Yadav, and S. Rajkumar. 2011. Flower colour variations in gerbera (*Gerbera jamesonii*) population using image analysis. *Indian Journal of Agricultural Sciences*.81(12):1130-1136.

- Smith, A.R. 1978. Color gamut transform pairs. *Computer Graphics*. 12(3):12-19
- Van Horn, G., M. O. Aodha, Y. Song, Y. Cui, S. Sun, A. Shepard, H. Adam, et al. 2018. The iNaturalist Species Classification and Detection Dataset. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (IEEE)*. 8796-8778
- Venegas, P., F. Calderon, D. Riofrío, D. Benítez, G. Ramón, D. Cisneros-Heredia, M. Coimbra, et al. 2021. Automatic ladybird beetle detection using deep learning models. *PLoS ONE* 16(6): e0253027.
- Webster, M.A., S. M. Webster, S. Bharadway, R. Verma, K. Jaikumar, G. Madam, and E. Vaithilingham. 2002. Variations in normal color vision. III. Unique hues in Indian and United States Observers. *Journal of the Optical Society America*.19(19):1951-1962.
- Willis, C.G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost et al. 2017. Old plants, new tricks: phenological research using herbarium specimens. *Trends in Ecology & Evolution*. 32(7):531-546.
- Wright, S. 1943. Isolation by distance. *Genetics*. 28:114-138

CHAPTER 3

VARIATION IN FLORAL COLOR OF THE HARDY GERANIUM, *GERANIUM MACULATUM*, IS CORRELATED WITH BIOCLIMATIC VARIABLES

Introduction

Variation in floral color has played a critical role in our understanding of the causes for floral evolution. Studies focusing on the frequency differences of color morphs in natural or artificial populations have provided insights on how such variation could be the results of natural selection from biotic selecting agents, e.g. pollinator (Lau and Galloway, 2004; Streisfeld and Kohn, 2005, 2007; Schiestl and Johnson, 2013; Trunschke et al., 2021), seed dispersers (Amico et al., 2010) or floral enemy preferences (Sobral et al., 2021), or abiotic conditions, e.g., temperature, water availability, and solar radiation (Gould, 2004; Rausher, 2008; Lacey et al., 2010; Arista et al., 2013; Kovicich et al., 2015; Berardi et al., 2016; Koski and Galloway, 2018). Examples of species that display trait polymorphism across a wider geographic range provide particular interesting insight into how selective and/or non-selective mechanisms could be responsible for a larger-scale variation (Epling and Dobzhansky, 1942; Hamrick and Allard, 1972; Lau and Galloway, 2004; Newman et al., 2012), as well as how these mechanisms may vary across spatial scales. Unfortunately, studies that have investigated variation on a distribution-wide scale remain scarce to date and, as such, how floral trait variation is generated and maintained in a species' entire distribution range remains poorly understood.

The first step to fill this knowledge gap is to document trait variation across a species' range distribution. Our study system, *Geranium maculatum* is a common herbaceous species widely distributed in eastern North American, and an earlier survey had shown intriguing color polymorphisms in both petal and pollen colors, likely due to varying level of anthocyanin production (Chang, unpublished data). Pollinators in this species include a wide variety of insects (bees (including *Apis mellifera*, and members of the Halictidae, Andrenidae and Apidae families), flies, and butterflies (Bertin and Sholes, 1993; Van Etten and Chang, 2014). Pollinator preference studies in this species show mixed results but did indicate that members of Andrenidae prefer lighter pigmented version of this species. (Deen, unpublished data). Though the field survey approach provides data on the trait of interest and other important features at the population level, it tends to leave gaps in the coverage of the species range (Killen et al., 2022). In contrast, recently popularized data from citizen science platform, such as iNaturalist, provides good geographic coverage in data collection, but the information comes from the individual level without population delineation (Hochmair et al., 2020; Rochman et al., 2020) With the strengths each dataset brings, combining both in the geographic pattern analysis will allow a more rigorous investigation than using each of the dataset alone. Indeed, using this approach, we discovered a complex geographic pattern in petal and pollen color variation in *G. maculatum*, indicating a suitable system for studying the likely mechanisms underlying floral color variation, most likely due to anthocyanin levels, in this species.

When examined across the larger geographic regions, petal and pollen color has been found to correlate with latitude (Arista et al., 2013), longitude (Streisfeld and Kohn,

2005; Newman et al., 2012; Koski and Galloway, 2018), or elevation (Galen and Kevan, 1980; Berardi et al., 2016). While non-selective mechanisms can lead to geographic clinal variation including drift, gene flow and migration (Haldane, 1948; Endler, 1977; Vasemagi, 2006), it is often attributed to natural selection (Haldane 1948; Slatkin 1973; Endler 1977; Storz, 2002) due to abiotic factors such as temperature (Berardi et al., 2016; Koski and Galloway, 2018) or precipitation (Berardi et al., 2016)) or biotic reasons (namely pollinators (Galen and Kevan, 1980; Newman et al., 2012), or reproductive isolation resulting from flowering time differences (Arista et al., 2013) are responsible for such large scale variation. One problem with this approach is that several of these potential factors may naturally covary across the geographic range (Hoffman and Parson, 1991; Berardi et al., 2016). As such, collinearity of factors with geographic features makes it difficult to confirm any causal-effect relationships using correlative data between biological patterns and environmental variation. Studies focusing on the known environmental stressors can provide stronger inference for the causes of natural variation in floral traits.

Environmental stressors, such as UV-B (wavelength 280-325nm), temperature, and precipitation (Landi et al., 2015), have been identified as agents of selection in floral pigmentation. Pigmentation in plants, particularly the anthocyanins, have been shown to act as antioxidants, osmoregulators, and photo-protectants (Gould, 2004) and can be induced either as a direct physiological response to the environmental stressors or an adaptive response to provide protection to the plant tissues against the stressors (Mierziak et al., 2014; Kim et al., 2017). For example, increases in UV have been shown to darken pigmentation in numerous species and within various tissues; leaf (del Valle et al., 2015),

flower (Peach et al., 2020), pollen (Koski and Galloway, 2018; Peach et al., 2020) and fruit (Zhao et al., 2017), suggesting that UV-B plays a selective role in pigmentation. Pigmented morphs in several polymorphic species were shown to perform better in lower water conditions than their non-pigmented counterparts (Grossenbacher et al., 2001; Schemske and Bierzychudek, 2001, Warren and Mackenzi, 2001; Tang and Huang, 2010). Additionally, flowers which are anthocyanin rich (darker) have been shown to perform better in cold temperatures via warming the reproductive structures and improving reproductive success (Jewell et al., 1994; Lacey et al., 2010; Ida and Totland, 2012). Interestingly, reports have also found that spring ephemerals that regularly experience colder mornings also perform better when under warmer conditions (reviewed in Van der Kooi et al., 2019)

Pigmentation in pollen grains, though not studied as extensive as petal pigmentation, have also been shown to act as photo-protectant against the damaging effect of UV radiation on pollen germination and viability (Ahmad et al., 1991; Demhick and Day, 1996; Feng et al., 2000; Bohrerova et al., 2009; Llorens et al 2015,). Anthocyanins can also serve as osmoregulators to protect against the effect of desiccation on pollen viability (Gould, 2004). For example, in *Nigella degenii* dark pollen was found in areas with reduced drought (Jorgensen and Andersson, 2005). The association between pollen pigment and temperature shows species specific patterns. For example, Citrus showed that pigments may serve the antioxidant function to protect pollen against low temperatures (Mohammadrezakhani et al., 2017). However, an opposite pattern was observed in *Campanula americana* – purple pollen outperforms white pollen in high temperature (Koski and Galloway, 2018). The mixed findings thus far suggest that more

studies, particularly in species with color polymorphism, would be needed before a generality can be made.

This study examines petal and pollen color variation across the range *Geranium maculatum* using both traditional field surveys and citizen science records from iNaturalist, we characterize the comprehensive patterns of petal and pollen pigmentation and dissect the role of geographic and environmental factors (UV-B, temperature, and precipitation) may have played. We used multiple linear and logistic regressions to examine the relationship between floral pigmentation and these features. Niche overlap calculated through ecological niche model will determine if differently pigmented morphs have significantly different realized niches. Additionally, suggestive implications of selection of pigmentation will be accessed correlative tests such as random forest (powerful predictive models) and PCA. Lastly, we use mantel and partial mantel tests similarity in population phenotype to environmental similarity and population distance. Our aims are to: 1) document the geographic pattern of variation in both pollen and petal pigmentation, 2) determine if color morphs differ in their predicted suitable niche, and if so, to identify the factors that differentiate the niches between morphs, 3) correlate environmental factors to floral pigmentation, 4) test population pigmentation based on similarity in geographic and environmental space. Using a variety of analytical approaches, we aim to provide multiple lines of evidence to determine if the abiotic variables are important in shaping the distribution of floral pigmentation. Basing on results from studies of other systems, we expect to see a general pattern of darker pigmentation with high UV-B. Though we also generally predict that darker pigmentation will be correlated with harsher environmental conditions (colder and drier), conflicting

results in the literature makes it difficult to establish *a priori* predictions regarding temperature and precipitation. However, we expect that if selection from one or more of these variables influences petal or pollen color in *G. maculatum*, we would be able to detect significantly different niches between morphologies, populations in similar habitats to be similar in morph composition, and a correlative relationship between the morph and the environmental factors examined.

Materials and Methods

Study Species

Geranium maculatum is a native herbaceous perennial commonly found in central and eastern North America. Plants are found primarily in deciduous forests but can also be found in fields and mixed forests (Bertin and Sholes, 1993). Flowers are either protandrous hermaphrodites or females. Hermaphroditic flowers have 5 petals and 10 anthers arranged in two whorls that mature sequentially. Flowers are 1.5 to 3.0 cm in diameter, and their petal color varies from light pink to deep purple (Chang, 2006). Pollen is found to be yellow (anthocyanin-less) or purple (anthocyanin) with variation in the intensity of the pigment (personal observation) (Figure. 3.1). *G. maculatum* usually blooms late March to early June, though the flowering season shifts depending on the locality.

Documenting pigmentation variation

Field Survey: Field surveys of petal and pollen colors were conducted across 45 natural populations of *G. maculatum* throughout its distribution range in the springs of 2019 and 2021. We surveyed reproductive individuals along a transect line in each

population. Samples were spaced out to avoid recording different inflorescences of the same individual. The number of plants sampled per population varied between five and 142, determined by the availability of hermaphroditic flowers that were in their male phase and still contained fresh pollen (see Supplemental Table 3.1). The number of reproductive individuals of these populations were counted to represent their size, but this is always an underestimate since there are usually many non-reproductive individuals in the population. Though the petal color is a continuous trait, to enable efficient data collection during our survey trips, we used a color swatch developed from previous data for each trait (Figure 1). Swatch colors were arbitrarily coded from 11 (light pink) to 15 (dark purple), increasing in color saturation (Figure 3.1). Petal color was surveyed from all 45 populations. Because petal color fades as the flower ages (Martin, 1965), we limited our color scoring to flowers in the male phase (i.e., newly opened flowers). Petal color of all samples were averaged to get the mean petal color for each population.

Pollen of *G. maculatum* also shows variation in the level of purple coloration. Of the 40 populations surveyed (five populations were missing due to rainy weather), 24 populations have only yellow pollen grains, and the remainders contain both yellow and purple pollen grains (Figure 3.2). Though we observed different shades of yellow and purple in *G. maculatum* pollen, we chose to separate the individuals into just two discrete color categories in this study – yellow and purple. To analyze pollen color in finer categories of yellow and purple will require additional knowledge on the genetic inheritance of this trait which is currently unavailable. We calculated the frequency of yellow to purple pollen individuals per population using the survey data (Supplemental Table 3.1).

iNaturalist dataset: To broaden the geographic coverage of the *G. maculatum* records, we took advantage of the big dataset stored in the iNaturalist.org. There were over 15,000 observations of *G. maculatum* collected and submitted by citizen scientists to the website iNaturalist.org, as of summer 2020 when we downloaded the data for this study. We used the subset of the records that was deemed as research grade, a designation with the following qualifications: photographed in the wild, properly georeferenced, and have their identification verified by multiple users. A total of 5,007 records were retrained for analysis after we imposed additional metrics of filtering for further quality control: including adding a coordinate uncertainty threshold of 2,000 meters, removing records from unlikely flowering time (Sept-Feb), and images without any flowers (Chapter 2). To collect data from the photographs of *G. maculatum* that are comparable to our survey results, we digitized the color information of the photographs and collected the Hue-Saturation-Value (HSV) scores for the petal to be used as color categories for later analysis (See more details on the protocols in Chapter 2).

Research-grade records were downloaded from iNaturalist.org in 2021 to evaluate pollen color of citizen science records. We evaluated a total of 2,287 records that contained observable anthers and pollen morphology. A team of four evaluators manually scored the pollen color for each photograph with one of the authors (S-M Chang) reviewing and verifying the scores for all records. Photographs were scored as yellow or purple pollen and the nonscorable (small, blurry photographs, female phased, or non-dehiscent anthers) were removed from the analysis.

Data visualization: To investigate the geographic patterns of petal and pollen color, we plotted the surveyed records as pie charts representing the frequency of color

morphs in their corresponding GPS locations using ArcGIS Pro, a geospatial mapping software. For the citizen science occurrences data from iNaturalist, we plotted individual observations with their computer scored color onto the map. GPS coordinates of each population (survey) and individual record (iNaturalist) were used to mine environmental variables from two databases; the bioclimatic variable set at the spatial scale of 30s (~1km²) from WorldClim.org (19 variable, <https://worldclim.org/data/bioclim.html>, Fink and Hijmans, 2017) and the glUV UV-B dataset (6 variables, Supplemental Table 3.2) (Beckmann et al., 2014). We will refer to the analysis that uses survey data as the ‘population’ analysis and those that use the iNaturalist records as the ‘occurrence’ analysis hereafter.

Geographical distribution patterns: To test how the observed variation is distributed across the three key geographic features in the plotted map, we ran multiple regression analyses with latitude, longitude, and elevation as the predicting variables and the population and individual petal and pollen colors as the dependent variables in separate analyses. For the survey data, we used the average petal color for each population for analysis, so our sample size was 45. For the iNaturalist records, we used an automated pipeline (Chapter 2) to assign photographs into three categories: Light (recorded as 1; equivalent to swatch saturation scores for 11-12), Medium (recorded as 2; equivalent to swatch 13) and Dark (recorded as 3; equivalent to swatch 14-15) petal. This conversion from 5 (swatch) to 3 (HSV values) categories was done to help reduce the imbalance of category sample sizes due to very low numbers in the extreme categories (14 and 15). We similarly ran a multiple regression analysis for the pollen data by using the frequency of yellow pollen in each population as the dependent variable for survey

data and individual pollen records coded as 0 (purple) or 1 (yellow) in the occurrence data (logistic regression). An arcsine square root transformation was used on frequency data. Logistic regressions were utilized for the iNaturalist pollen data. Predictor variables were scaled using a Z transformation to account for variation in measurement units. We used R package ‘glmulti’ for this analysis and allowed all interaction terms in the model. The best model was selected using the AICc criterion (Calcagno, 2020).

Climatic factors: To determine whether the color variation can be explained by the climatic features, we ran a series of regressions on population and individual level pigmentation in petal and pollen using a reduced set of predicting variables (to avoid issues of multicollinearity among the full set of bioclimatic variables). The variance inflation factor (VIF) was used to measure the degree of multicollinearity with a value over 10 indicating multicollinearity (Vittinghott et al., 2012). The annual mean temperature (Bio 1), annual precipitation (Bio 12), and annual mean UV-B level (UV 1) were selected to meet these two criteria. Regressions were run with the same methodology as the geographic variables as in a Z transformation was used to scale predictor variables and ‘glmulti’ was used for model selection. To visualize interactions for better interpretation from the model, we graphed partial residuals using R Package ‘interactions’ (Long, 2019).

In addition to running the multiple linear regressions using all of the data, we repeated these analyses using only the eastern part of the distribution range. This was done because we noted that survey populations in the western range of *G. maculatum* consisted of solely yellow pollen and, similarly, the iNaturalist dataset contains almost exclusively yellow pollen. Since natural selection can only be detected when trait

variation is present, to better examine whether the abundance of the purple pollen is related to environmental factors, we repeated our MLR analysis using only data from the eastern regions where both morphs coexist. We removed six western populations monomorphic for yellow, including Morton Arboretum, Hubbard Valley, Mary Beth Doyle, Pearson Metroparks, Blue Creek, Eagle Creek (Supplemental Table 3.1). To trim the iNaturalist dataset correspondingly, we generated an arbitrarily imposed break based on the populations removed (Figure 3.2). It should be noted that we chose to run the break along the western edge of Appalachian Mountain foothills to retain the elevational gradient in the analysis because elevation has been shown to be an important predictor in previous studies.

Niche Differentiation

We tested for niche differentiation among the different morphologies using the ecological niche modeling software Maxent (Phillips et al., 2006). Significantly differentiated niches would provide evidence that morphologies utilize different niches and are thus potentially adapted to different environments (Suárez-Mota and Villaseñor, 2020). We also conducted principal component analyses of all the environmental factors used in the niche modeling to verify whether morphs cluster along the principal component axes of the environmental factors.

Input data for the model was the subsets of different color morphs from the iNaturalist records: light, medium, and dark for petal, and yellow and purple for pollen. After the Maxent models of habitat suitability for each morph was built, we calculated the pairwise niche overlap between morphs of the same trait using ENMtools (Warren et

al., 2008) and determined whether there is a significant difference between niches using the niche equivalency (identity) test.

Maxent uses a maximum entropy approach based on occurrence records to build models of habitat suitability (Phillips et al, 2006). We clipped environmental rasters to an extent, a 2x2 XY coordinate matrix, of -120, -60, 30, 50 to define the area. Using the package “raster” in R, we sampled 15,000 background points for each dataset, then thinned occurrence records with the package “spThin,” to reduce sampling bias (Aiello-Lammens et al., 2015). “raster” was used to turn the thinned occurrence records into an SWD file (sample with data). Maxent was run using all bioclimatic variables with only linear features, the simplest possible model, and was replicated 10 times. Highly correlated variables ($|r| > 0.7$) were removed. After the initial 10 replications, Maxent outputs a file containing various performance statistics. One statistic is the area under the curve (AUC) for each variable when the model is run with only that variable. The variable with the highest average AUC would have all highly correlated features removed. Working in descending order, this procedure was repeated. Maxent was re-ran with the remaining uncorrelated features. Variables were further removed using a stepwise procedure that compared model’s Akaike information criterion (AICc). The model with the best AICc was used as the final maxent parameters for that morph. After variable selection, regularization was tested in increments of 0.25 between one and 3. The final model is an average of 10 maxent runs with identical parameters.

To calculate the niche overlap between pairs of pollen or petal morphs, we calculated the Schoener’s D and the I statistic using ENMtools (Schoener, 1968; Warren et al., 2008). Values of D and I range from 0-1 and demonstrate how much niches

overlap. The pair-wise values of D and I were placed into a matrix format (dimension of 3x3 for petal, and 2x2 for pollen). A niche identity test was run to compare the real I and D values against 100 replicates of permuted randomly assigned values to determine significance (Warren et al 2010). The 95% confidence interval is used to compare against the real I and D values. If the real value is less than the 95% CI then the niches are significantly different between the two morphs. However, if the 95% CI is greater than the real value, the two niches are not significantly different.

To complement the niche modeling approach, we also conducted principal component analyses (PCAs) to reduce the highly correlated bioclimatic features into ranked major principle-components (PCs, i.e., the composite variables of the individual features) and check whether the color morphs separate along these major PCs. Separate PCAs were run for the petal and pollen data for survey population and iNaturalist occurrence datasets using `sk.learn` in Python (Van Rossum and Drake, 2009; Pedregosa et al., 2011). All 25 bioclimatic variables were loaded and scaled using `StandardScaler` from the `sk.learn` package. Visualization of data separation was done via cluster biplots, while the strongest loadings from PC1 and PC2 were used to interpret how each variable was influencing the separation of morphs.

Morph prediction based on environmental factors

Random Forest Classifiers, a machine learning approach, were employed to predict petal and pollen color using 25 bioclimatic variables from the occurrence records. Random Forest is a nonparametric, supervised, machine learning application valued for its high accuracy and ability to model complex interactions (Breiman, 2001; Cutler et al.,

2007). Accurate prediction of color by random forest with bioclimatic features imply floral color and the features are correlated.

To run the random forest we trained, trimmed features, and tuned parameters. Each occurrence dataset was split (70/30) into training and testing data so that 70% of the data would be used to train the model and remaining 30% to validate the results. The training data was fed into the random forest, an ensemble form of machine learning that uses a series of individual decision trees. 100 decision trees were trained on the data and each tree made its own prediction regarding color. A majority vote is taken among the predictions made by the individual decision trees to make a final, overall classification for each sample. To ensure the results were not skewed by data outliers, or the branching of the decision trees, K-Fold cross validation was performed with 10 folds, which splits the data into 10 equal partitions. Nine of the folds were used to train the random forest and the results were evaluated on the final fold. The folds were then replaced until each of the ten folds were tested. If one of the folds gave an aberrant outcome, it may indicate skew in the model. To validate the results and to prevent overfitting, the finished model made predictions on the remaining 30% of data (the test set). Because the model was not trained on this test set, we can assess performance on new observations. The input bioclimatic variables (features) were ranked based on feature importance (impurity-based). Features with a value below 0.05 were considered redundant and were removed from the model. The model was then rerun with the trimmed set of features. A randomized grid search was used to hyper-tune the model's parameters, essentially rerunning the model with a different, randomized combination of parameters within a predetermined range to determine which set of parameters return the best results

(Supplemental Tables 3.3-4). Parameters include the number of trees, and guides how the trees should branch. Three metrics were used to evaluate the classification model: R^2 , out of bag (OOB) score, a metric for prediction error, and Cohen's kappa, a metric designed to check how the model performs relative to chance (Landis and Koch, 1977). Partial dependence and Shap summary plots were used to interpret the model (Lundberg and Lee, 2017).

Population similarity by geographic and environmental distance

We conducted Mantel tests (using “vegan” package in R) to determine whether the trait differences between a pair of populations (survey data) is correlated with the geographic and/or environmental distances (Oksanen et al., 2019). Geographic distance was calculated using the American Museum of Natural History's Geographic Distance Matrix Calculator (Found at: https://biodiversityinformatics.amnh.org/open_source/gdmg/, Ersts.). Petal color distance was measured as the absolute difference in petal average between pairs of populations. Pollen color distance was measured as $|\text{population X1 frequency of yellow}| - |\text{population X2 frequency of yellow}|$ for each pair of populations. We also conducted partial mantel tests to determine the relationship between population color differences and the elevation and environmental differences while controlling for geographic distance. The environmental distance was calculated as the Euclidean distance of the three annual average bioclimatic variables between populations.

Results

Geographical patterns of floral pigmentation

Though the survey and iNaturalist datasets were collected using quite different approaches, they both showed that the lighter morphs (petal 11&12, pollen color “yellow”) are more common than the darker morphs (petal 14 & 15, pollen color “purple”). In the 45 populations surveyed, we scored petal of 2,027 flowers and found 46.3% to be light (13.6% for color 11, and 32.7% for color 12), 21.2% to be medium, and 32.5% to be dark (10.5% of color 14, and 22.0% of color 15). Most populations contained a mixture of at least two color-categories, with the exception of one having only light and another having only dark flowers. iNaturalist dataset contains 5,004 petal records, of which 57.4%, 16.4%, and 26.1% are classified as light, medium, and dark, respectively. Of the 40 populations surveyed for pollen color, 16 had both yellow and purple pollen with the former constituting 3-95% in the population, The remaining 24 had only yellow pollen. Overall, we scored 1,183 plants for pollen color and found 79.0% (934) having yellow pollen and 21.0% (249) having purple pollen. Similarly, of the 2,287 iNaturalist occurrence samples, 72.1% (1,650) were classified as yellow and 27.9% (637) as purple. In both datasets, yellow pollen is the majority morph, with purple being the minority, less geographically spread morph.

Our survey data showed that the average population petal color of *G. maculatum* displays a clear elevational cline (Figure 3.2) such that darker petals are more likely to locate in higher elevation and lighter petals in lower elevation. The best model includes the highly significant factor of elevation and its interaction with latitude (R^2 : 0.73, Table 1a). The darkest petals (code=15) are found in the southern Appalachian Mountains and predominantly in high elevation areas. The next darkest color (code=14) was found sparsely spread throughout the entire distribution range. Similarly, when analyzing the

individual iNaturalist data, though elevation and its interaction with longitude, and the interaction between longitude and latitude all have a significant effect on the petal color (Table 3.1c), the overall model had minimal power to explain the variation pattern of color occurrences (R^2 : 0.07). It is, however, clear from both the survey and iNaturalist maps that the southern Appalachians have a high percentage of the dark petal individuals (Figure 3.2).

The distribution of pollen color exhibits a more complex pattern (Figure 3.3). Specifically, there appears to be a contrast between eastern and western halves of the distribution range (See line in Figure 3.3b). The western half of the distribution range has almost exclusively yellow pollen, but in most areas/populations in the eastern half, including the Appalachian Mountains, both yellow and purple pollen are routinely found intermixed within a population (Figure 3.3a). Multiple linear regression (MLR) analysis on the survey data shows that geographic variables (elevation, latitude, and longitude) were able to explain ~ 57% of the pollen color variation among populations, and elevation has the most significant effect on the frequency of yellow pollen in the populations (Table 3.2a). Populations in high elevations tend to have higher frequencies of purple pollen individuals than those in lower elevations (Figure 3.3). There was also a significant interaction between elevation and longitude (Table 3.2a), which disappeared when we removed data from the western populations (areas) (Table 3.2c; R^2 0.48, elevation p : <0.001), indicating that the effect of elevation is most significant in the eastern regions, particularly around the Appalachian Mountains range. MLR analysis of the individual data revealed a similar pattern and all three geographical variables - elevation, longitude, and latitude, show a significant effect on pollen color (Table 3.3b).

When we removed the western regions, the overall trends remain the same (Table 3.3d), but the explanatory power of the model was much reduced ($R^2 = 0.19$).

Niche Differentiation

The suitable niche predicted by Maxent for dark petal morph was significantly different from light and medium morphs (Figure 3.4) but the latter two shared similar niches (see Figure 3.4 legend for the pairwise Schoener's D and I). Response curves indicate that niches with high suitability for dark morph contrast from that for light morph by their relatively higher precipitation during the highest and coldest quarters, the colder temperatures during the driest times, and the higher UV-B during months with peak UV radiation. Suitability for light morph is further characterized by low precipitation seasonality (Supplemental Figure 3.2)

Yellow pollen and purple pollen niches predicted by Maxent are also significantly different, shown by Schoener's D (0.571695, CI 5: 0.84) and I (0.770994, CI 5: 0.971) (Figure 3.5). Interestingly, yellow pollen is excluded from the highest elevation locations of the Appalachians. Suitable niche for purple pollen extends from the lowest latitude of the Appalachians, continuing north towards the highest points of the Appalachians, then veering east towards the coast up to the mid-Atlantic and New England regions (Figure 5b). The best-suited habitat for yellow pollen appears to be the northern part of the distribution range, from central to eastern US (Figure 3.5a). Response curves show that, in comparison to yellow pollen individuals, purple pollen individuals are found in locations with higher elevation, lower precipitation seasonality (<10%), and lower UV-B level in the summer (Supplemental Figure 3.3). In addition, yellow pollen individuals are

found in areas with moderate precipitation during the warmest and coldest quarter and moderate-high summer temperatures ($\sim 27^{\circ}\text{C}$) (Supplemental Figure 3.3a).

PCA of the bioclimatic variables associated with the surveyed petal populations revealed that the top two principle-components explain 66.28% and 18.76% of the environmental variation. Populations did not clearly separate along the PC1 axis but those with medium-dark petals separate from the light petal populations along the axis of PC2 with large loadings for the maximum temperature of warmest month and mean temperature of warmest quarter. Specifically, summers with colder temperatures separate dark and many medium-colored populations from the lighter populations (Figure 3.6a). In contrast, the PCA for iNaturalist records colored by petal color does not show separation of points, while the PCs explain 53.84% and 18.73% of the variation respectively (Figure 6b)

The two PCs associated with bioclimatic variables explain 86.49% of the variance of populations surveyed for pollen color frequency. The first PC separates locations with high UV-B, warm winters, low-temperature seasonality, high precipitation from those with inverse conditions. The second PC separates populations with lower warm-season temperature and low annual temperature from those with the inverse conditions (see full list of loadings in Supplemental Table 3.5). Populations with different frequency of yellow pollen show separation along the combination of the first two PC axes in which populations that with majority purple pollen locating in the top-right quadrant of the biplot (Figure 3.6c). Similarly, a PCA run that associates bioclimatic variables with pollen color occurrence resulted in 75.82% of the variance explained by the first two PCs (PC1: 59.51%, PC2: 16.31%). Populations showed separation along PC2 which is

described by lower temperatures during warm months, low variation in precipitation, high precipitation in ‘drier’ months, low UV-B seasonality, low UV-B in the highest month and low UV-B in the highest quarter (Figure 3.6c).

Morph prediction based on environmental factors

Random Forest does not accurately predict occurrence petal color in *G. maculatum* with bioclimatic features, indicating that environmental factors used in our analysis may not be critical in explain the petal color geographic variation. The model’s R^2 value were 0.74, 0.67 (train, test), and the OOB score was 0.70, suggesting accurate prediction. Light petals had a moderate precision and excellent recall (0.69, 0.97). However, the kappa value of the model is exceptionally low - 0.06, which suggests that the constructed model’s prediction is no better than just random chance. Light petals represented an uneven portion of the sample at 842 records of 1256. A simple assessment of classifying every record as ‘light’ would lead to a 67% accuracy. According to McHugh (2012) there is little data reliability in the output of this model. Our population survey data proved to have too small a sample size for Random Forest analysis, as indicated by its accuracy being highly dependent on the test-train split.

Random Forest analysis performs much better with the pollen color data. Pollen color is accurately predicted by Random Forest as shown by the following statistics and confusion matrix. As shown by the confusion matrix (Table 3.4a) the model had good R^2 (0.99, 0.86: train, test), and good recall (yellow: 0.73, purple: 0.91), which is the percentage of correctly identified classifications for that class. The total weighted average was 0.86 indicating overall good accuracy. The high OOB score (0.86) and substantial kappa value (0.64) indicates good model performance and data reliability. The key

features for pollen classification were precipitation seasonality, precipitation of the driest quarter, precipitation of coldest quarter, UV-B seasonality, and annual precipitation in descending order of Shap value. A Shap summary plot indicates that low values of precipitation and UV-B seasonality (Bio15, UV2) towards a prediction of purple, while high values of annual, driest quarter and coldest quarter precipitation push towards purple (Bio12, Bio17, Bio19) (Figure 3.7a).

The Random Forest algorithm allows the detection of complex interactions between variables, which can be interpreted using partial dependence plots. Partial dependence plots (Figure 3.8a) showed that, in general, increased precipitation seasonality and decreased overall precipitation would both lead to an increased prediction of yellow pollen. More specifically, when the precipitation goes beyond 150 mm, an increase in precipitation during the driest quarter will lead to a decreasing prediction for yellow pollen. However, the marginal effect of precipitation seasonality is stronger when the value of precipitation seasonality is low (<20%) than other precipitation values. A general trend of increasing UVB seasonality increases prediction towards yellow pollen (Figure 3.8a).

In addition, we also ran a Random Forest classifier for the occurrence dataset with west-wards populations removed. The model performed well, as indicated by the confusion matrix (Table 3.4b). The R^2 was respectable at a (train, test) of (0.96, 0.80), and a weighted accuracy of 0.80. This is paired with an OOB score of 0.75 and a moderate kappa value of 0.59, indicating moderate agreement. Features important in descending order are precipitation seasonality (Bio 15), min. temperature of coldest month and mean temperature of coldest quarter (Bio 6, Bio 11), annual mean temperature

(Bio 1), and temperature seasonality (Bio 4). Shap plots indicate that high values of these features increase the prediction of pollen color towards yellow (Figure 3.7b).

Multiple linear regressions indicate that lower annual temperatures with high precipitation and higher UV-B coorelate with darker petal populations (R^2 : 0.77) (Table 3.1b). The partial dependence plot shows that low temperatures, regardless of annual precipitation are already predicted to be above 13, while increasing the precipitation increases the population color to 15. Meanwhile higher temperature populations are predicted to be just under 12 and slowly decreases with increasing precipitation (Supplemental Figure 3.4). However, multiple linear regression of iNaturalist occurrences do not yield a strong relationship (R^2 : 0.06, Table 3.1d), which agrees to the lack of predictive power with Random Forest.

The best multiple linear regression model to explain population pollen color, was based on annual temperature, annual UV-B and an interaction of annual temperature and annual precipitation (R^2 0.60) (Table 3.2b). Pollen color can be partially explained in populations with higher yellow pollen frequency by increasing annual temperature, as well as decreasing annual UV-B (Table 3.2b). Furthermore, in populations at relatively lower precipitations, pollen color is influenced less by temperature than in populations with higher amounts of precipitation. At higher precipitation, populations with higher temperature have high yellow pollen frequency while populations with lower temperature have low yellow pollen frequency (Supplemental Figure 3.5). The same trend is seen when the monomorphic westward populations are removed (R^2 : 0.58, Table 3.2d). Among individual samples of *G. maculatum*, geographic location (elevation, latitude, and longitude) is able to explain pollen color better than the three environmental factors (R^2

0.41 vs R^2 0.30) (Table 3.3a-b). In addition, the interaction of annual precipitation and annual temperature partially explain pollen color; areas with both low temperatures and low precipitation are predicted to have yellow pollen while areas with low temperature but high precipitation have purple (Supplemental Figure 3.5) and vice versa. However, when we used only the eastern region data, the results changed to that low temperature areas tend to have more purple pollen, and warmer areas tend to have yellow, with less influence from lower precipitation areas by temperature.

Population similarity by geographic and environmental distance

All results presented in this section are from survey data because Mantel tests can only be done using population level data. The mantel test showed that populations that are closely together in geographic distance are not necessarily similar in petal color composition (r : -0.04, p : 0.80) (Supplemental Figure 3.6a). However, when controlling for geographic distance, partial mantel test shows that elevational difference has a relatively strong correlation with the population average colors (r : 0.61, p < 0.001), suggesting that elevation is highly influential for population petal color. Similarly, a partial mantel test using the Euclidean distance of annual temperature, precipitation and UV-B controlled by geographic distance shows a moderate correlation (r : 0.36, p : 1e-04).

Pollen color similarity in populations does not relate to geographic distance between populations according to the Mantel test (all populations, r : -0.12, p : 0.99, west removed r : -0.11, p : 0.99) (Supplemental Figure 3.6b). However, when geographic distance is controlled in a partial mantel test, we found that elevation and population pollen color are highly correlated (all populations, r : 0.58, p : 1e-04, west removed: r : 0.52, p : 1e-04), suggesting that population at similar elevations tend to have similar pollen color

frequencies. All three environmental factors showed relatively low correlation with pollen composition (Partial Mantel test: r : 0.13, p : 0.07) indicating that these three environmental factors do not provide additional explanation power for the population pollen colors. This result holds when we removed the monomorphic western populations (r : 0.10, p : 0.13).

Discussion

We examined geographic patterns of petal and pollen color variation in *G. maculatum* and found that populations vary over space and environment. Combined, our results indicate that colder temperatures, and high UV-B are related to populations with darker petal composition. Colder temperatures, high precipitation, and high UV-B favor populations with high purple pollen frequency. While low precipitation, higher temperatures, and higher UV-B seasonality, but lower UV-B favors populations with high yellow pollen frequency. Overall, our collection of tests paints a picture of how abiotic factors may be influencing pollen color.

Pigmentation in *G. maculatum* shows an impressive example of geographic variation in floral traits. Both petal and pollen color in *G. maculatum* are highly variable among populations. Our survey indicates that populations have a general trend of darker pigmentation in high elevational areas and lighter in lower elevation area. Pollen color variation displays a complex pattern - our survey indicates that populations with darker pollen are found along the southern Appalachians and along New England, while the western range is mostly yellow pollen.

Geographically close areas are expected to diverge based on genetic drift from reduced gene flow from increased population separation. But selection may result in

populations that have a similar agent and direction of selection to look similar. Mantel and partial mantel tests indicated that populations in similar environments are similar in morphology, indicating that populations may be under selection by their environment. Furthermore, pollen morphs have significantly different predicted niches, as do light and medium petal color in comparison to dark petals. Mantel tests furthered that these realized niches are different and correlated to population morphology, however once the monomorphic west populations are removed from the analysis the similarity in environment as defined by annual mean temperature, annual mean UV-B, and annual precipitation no longer correlates to similarity in population color. While elevational similarity holds, this may suggest that our environmental composition is not exploring the true variables, as similarity of populations along different mountain peaks (similar elevation=similar color) would necessitate gene flow down and up mountains, therefore there is an expectation that different environments would have different color composition.

Our regressions and PCA propose that population petal color separate based on environmental factors while tests examining the individual occurrence dataset struggled to identify relationships. However, the later may be due to bias withing citizen science sampling, or the data structure itself. Furthermore, the regressions and Random Forests complimented each other; the Random Forest allowed us to develop alternative hypothesis stemmed from the features the model decided were important. The linear regressions, however, are informed by previous studies; annual mean temperature, annual mean precipitation and annual mean UV-B were chosen based on *a priori* knowledge of these factors effect on production and expression of anthocyanins in plant tissues from

previous studies. Both Random Forests and regressions are used to test the relationships between climatic variables and pigmentation, the addition of both approaches increases our confidence in the output. Regressions and Random Forests both predicted that less UV-B and low precipitation are related to yellow pollen, while regressions indicated an interaction of temperature and precipitation. What is interesting is the Random Forest did not draw importance from temperature for the entire range. Our regressions of pollen morphology within populations with the westwards removed had the same trend, indicating these environmental factors may be playing a real selective role against pollen coloration. However, among occurrence records, the patterns were different. UV-B has an interaction with precipitation throughout the range and develops an interaction with temperature when we remove westwards populations. As UV-B increases with precipitation, pollen tends to be purple. In the eastern side of the range, UV-B has an interaction with temperature. Lower temperatures influence pollen color more strongly than high temperatures when interacting with UV-B. As UV-B increases, pollen tends to be purple. However, in higher temperatures, this interaction is weaker. This is likely due to the high prevalence of purple pollen on the peaks of the Appalachians, where the temperatures will be lower than the surrounding lowlands and a higher degree of UV-B radiation would be expected. However, when we run the Random Forest without the westward's occurrences, suddenly temperature becomes important in predicting pollen color. Precipitation could be speculated to lead to fixation in the west, while temperature influences color in the east.

Temperature

The literature indicates that colder temperatures can select for darker petal pigmentation to increase the temperature of reproductive structures (Jewell et al., 1994; Lacey and Herr, 2005; Stiles et al., 2007; Ida and Totland, 2012; Van der Kooi et al., 2019), which agrees with our survey results. Our multiple linear regression and PCA clearly demonstrate that colder temperatures correlate to darker petals, while our mantel tests and niche modeling as a whole indicate a separation of petal pigmentation based off the general environment. *G. maculatum* is a spring ephemeral that would experience cold spring mornings in the mountains of the Appalachians, which tracks with the elevational cline of pigmentation found. Biologically, elevated floral temperatures in spring ephemerals can aid reproduction success. For insect pollinators, warmer flowers can indicate better rewards. In fact, bees prefer warmed nectar, and warm flowers yield better quality and quantity of nectar. In addition to food rewards, insects raise their temperature via ‘basking’ in flowers, which would be critical in early spring (Van de Kooi et al., 2019).

In the polymorphic region of pollen color, low temperature correlated with purple pollen as indicated by our PCA, and regressions in both our survey and occurrence data. This was opposite than what was expected from the literature. In *Campanula americana*, another herbaceous species with exposed pollen with a purple (dark) and a tan (light) morph, it was found that higher temperatures partially explain purple pollen location due to temperature reducing germination of tan pollen grains, while not effecting germination in darker grains. It is important to note however, that the pattern of pollen color was opposite of ours with darker pollen on the west and lighter pollen on the east (Koski and Galloway, 2018). We see predominantly purple populations on the east coast

and yellow populations on the west of the range. This can be potentially explained by low temperatures that have been shown to increase flavonoid synthesis, while high temperature has been shown to decrease anthocyanin synthesis (Kim et al., 2017). In tomatoes, high-temperature stress increased reactive oxygen species (ROS) in pollen tubes. ROS damage cells and DNA. Flavonoids are ROS scavengers and increased with high temperatures in the study (Muhlemann et al., 2018). However, in 2 *Citrus* species, low temperatures lead to high flavonoid accumulation in pollen (Mohammadrezakhani et al., 2017). While in *Parrya nudicaulis* populations in higher temperature regions had a higher percentage of non-pigmented flowers, likely due to the loss of cold tolerance (Dick et al 2011). Anthocyanin increases owing to cold temperature is likely a consequence of oxidative stress (Zhang et al., 2019) caused by the antioxidant nature of anthocyanins. This is likely the case in pollen to preserve viability while in petals the underlying mechanism is likely to be that of thermoregulation.

Ultraviolet Radiation B

UV has an interesting relationship with floral pigmentation in many species. UV absorbing pigments aid in UV protection and pollinator attraction (Hirota et al., 2019, Koski et al., 2020). High UV does correlate with purple pollen and darker petals, so there is potential for this to be an agent of selection. In our Random Forest, it was UV seasonality that was a principal factor, while in our niche modeling UV-B during summer was important for establish species distrubtion and niche overlap. From that, the results indicate that moderate UV-B in the summer indicates yellow pollen, while low UV-B during summer indicates purple pollen while high values of UV-B are important in habitat suitability for darker petals. It is important to note that during summer, the plants

will have passed peak bloom and will be fruiting, which could indicate another trait correlated with pollen color that UV-B acts upon. As a potential explanation, sister taxa, *Geranium robertianum* has been characterized as having anthocyanins in leaf tissue. The anthocyanins concentration was higher in *G. robertianum* that grew in high light, indicating that anthocyanins do offer photo protectant qualities in at least this congener species, and perhaps offers the same protection in *G. maculatum* (Wierzbicka et al., 2013). If a correlation between leaf anthocyanin concentration and floral anthocyanin may exist and be driven by selection upon leaf tissue.

Additionally, while a human visible pattern in petal color is evident, UV markings visible to potential pollinators could additionally be present. UV bullseyes in *Argentina anserina* was found to have a positive correlation with UV-B, driven by distance to the equator (Koski and Ashman, 2015A). Furthermore, an elevational cline in bullseye pattern was responsible for shift in pollinator preference (Koski and Ashman, 2015B). While it is unknown if *G. maculatum* has an UV bullseye, anthocyanins are expected to be increased in areas with higher UV. Higher UV-B, according to our population regression does increase the prediction of darker populations' is an important factor to investigate in the context of ozone depletion, as flower pigmentation has been impacted by increasing UV (Koski et al., 2020).

Precipitation

Populations with purple pollen were found in well-watered areas as shown by high values of precipitation of driest quarter (Bio17), coldest quarter (Bio19) and annual precipitation (Bio12) pushing towards a classification of purple in our Random Forest. We can speculate based off the values of precipitation seasonality (Bio15), that low

precipitation seasonality indicates in our populations being well watered throughout the year. It was against our expectations that populations with less precipitation would lean towards yellow pollen. There is not a clear indication of the effect of precipitation on petal color, both regression and maxent indicate that precipitation interactions with temperature, but the interaction is inconsistent among tests. While anthocyanins have been shown to be osmoregulators in leaf tissue, it is unclear how this would benefit floral tissue, with the exception of desiccation in pollen (Chalker-Scott, 2002). As the *Geranium* genus does display stressed based anthocyanin accumulation, one could speculate that these darker morphed plants may display a heightened degree of anthocyanin accumulation in leaf tissue which is correlated to floral tissue pigmentation. (Wierzbicka et al, 2013). However, while this may make sense in terms of light/UV-B stress, the lack of darker pollen morphs in potentially water stressed environments indicates this would not be the cause.

Alternate Explanations

Determining whether selection is responsible for the variation using observational data showing the distribution pattern of a trait is a complex process. This is because environmental factors from available databases, such as worldclim.org (Fink and Hijmans, 2017, and glUV (<https://www.ufz.de/gluuv/>) (Beckman et al., 2014), are naturally correlated, which makes it difficult to isolate the key factor for selection, if existed, among the correlated ones. Regressions and Random Forests are predictive models, not explanative models. While Random Forests will measure feature importance, it cannot decide which of two highly correlated traits is the ‘correct’ trait. It merely moves forth with one, punting the other to a lower priority. However, it gives us reason to

believe that the environment does play a role in the selection of floral pigmentation, even if the exact variable may not be correct. It does provide an interesting alternative hypothesis from the regressions, however, in that precipitation and UV-B are responsible for pollen color.

The variation within populations, along with the general trend of populations as a whole correlating to the environment could indicate that selection is acting. However, traits can be constrained by correlated traits (Connor, 2002). Therefore, selection could be driving anthocyanin for other reasons and in other tissues that just pull petal color with it. The anthocyanin pathway has many paths and controls pigments in various tissues, and stress in one tissue can derive pigmentation in another (Armbruster, 2002). For instance, color of co-flowering species, pollen/petal contrast and pollinator community preference and efficiency could all drive floral pigmentation, while retaining a correlation to the environmental landscape (Lau and Galloway, 2004; Trunschke et al. 2021) We have noted in the field that dark pollen *G. maculatum* very often are against a backdrop of purple petals, while yellow pollen morphs are often against pink petals. We speculate this may be a form of ‘visual crypsis’ to hide pollen from folivores and pollen thieves (Ying-Ze et al., 2019) Ying-Ze et al found that many plants with exposed pollen match petal color. It has also been established that anthocyanins can provide protection against herbivory in flowers and leaves (Gould 2004; McCall et al., 2011), which may be a two-folded benefit of purple pollen against a purple background. On the other side, yellow against a pink background could signal for more pollinator attraction, if there enough spectral reflectance differentiation (Lunau and Maier, 1995; Lunau, 2000). Therefore, color could be under selection by pollinators and thieves alike, and pollen distribution

could reflect differing pollinator community compositions, or reproductive trade-offs.

iNaturalist bias

Field surveys are a popular and traditional method to explore the variation in traits of a plant. Typically, these surveys are of a higher quality than citizen science data, as records are confirmed by an expert with consistent quality. The process of conducting a field survey can require multiple seasons, depending on the range of the species. Often surveys are narrower than what an aggregation of citizen science records can accomplish in space and time. Limited money, time (including field season length), and resources are hurdles that prevent detailed, intensive field surveys. While citizen science records are free, encompass broad swatches of time they records suffer from a bad reputation for quality control and bias (Burgess et al., 2016).

Biased sampling from citizen scientists can influence color surveys. We cannot confirm in populations with variation in color, that each color was sampled equally to their true proportion when individual persons are sampling independently. Furthermore, clusters that are closer to trails and easily accessible areas are more likely to be sampled than plants deeper in the forest. A color that stands out against the majority of the population may be sampled more frequently, thus skewing the frequency. On the other hand, information often goes missing from citizen science data because not enough members in the community identify species, leaving good records left out of ‘research grade’ classification. Cell coverage also has an impact on citizen science records, as locations with poor coverage have less records (personal obs).

Next steps

Further experimentation is needed to assess the role of temperature, precipitation and UV-B on petal and pollen color and fitness, especially since these three variables are correlated with other features both ecological and geographic (Hoffman and Parsons, 1991; Berardi et al., 2016). Greenhouse reproductive studies coupled with common garden and transplant studies will help elucidate the selective role these environmental factors may have on color variation. However, the pattern presented in our field survey is a strong indicator that these environmental factors play a vital role in petal color (Endler 1977).

Further studies in *G. maculatum* should include experimental studies informed by this research, and genetic data that can tell us more about how populations are structured, phylogeography, and help test the contributions of drift and selection.

Table 3.1: Multiple linear regression of surveyed and iNaturalist petal data. All variables were Z-transformed. Significance values: 0 “****”, 0.001 “***”, 0.01 “**”, 0.05”.”

	Estimate	Std. Error	t value	pr(> t)	Significance	R ²
<i>A. Survey populations-Geographic</i>						0.73
Intercept	12.98	0.10	129.16	<2e-16	***	
Elevation	0.85	0.09	9.16	1.44E-11	***	
Elevation: Latitude	-0.31	0.12	-2.47	0.017	*	
<i>B. Survey populations-Environmental</i>						0.77
Intercept	12.90	0.10	130.80	<2e-16	***	
Annual mean temp.	-1.37	0.16	-8.86	5.60E-11	***	
Annual precip.	0.28	0.14	2.05	0.046	*	
Annual mean UV-B	1.04	0.21	4.99	1.24E-05	***	
Bio1:Bio12	-0.31	0.15	-2.14	0.039	*	
<i>C. iNaturalist occurrences-Geographic</i>						0.07
Intercept	1.41	0.01	124.89	<2e-16	***	
Elevation	0.14	0.02	8.14	533e-16	***	
Longitude	0.02	0.01	1.46	0.1446		
Elevation: Latitude	-0.03	0.01	-2.75	0.006	**	
Longitude: Latitude	-0.02	0.01	-2.06	0.0398	*	
<i>D. iNaturalist occurrences-Environmental</i>						0.06
Intercept	1.49	0.01	111.50	<2e-16	***	
Annual mean temp.	-0.19	0.03	-7.38	1.87e-13	***	
Annual mean UV-B	0.26	0.03	8.20	3.14e-16	***	
Annual mean temp: Annual precip.	-0.03	0.02	-1.51	0.131		

Annual mean temp: Annual mean UV-B	-0.09	0.01	-6.92	5.27e-12	***	
Annual mean UV-B: Annual precip.	0.06	0.02	3.79	0.0002	***	

Table 3.2: Linear regression of surveyed populations for pollen. Regression was run with the output being the frequency of yellow pollen in the population. All variables were Z-transformed. Significance values: 0 “***”, 0.001 “**”, 0.01 “*”, 0.05”.”

Model	Estimate	Std. Error	t value	pr(> t)	Significance	R²
<i>A. Geographic variables</i>						0.57
Intercept	4.84	0.17	28.35	<2e-16	***	
Elevation	-1.02	0.16	-6.62	1.04e-07	***	
Elevation: Longitude	0.50	0.22	2.30	0.028	*	
Elevation: Latitude	0.33	0.20	1.61	0.117		
<i>B. Bioclimatic variables</i>						0.60
Intercept	4.55	0.19	24.18	<2e-06	***	
Annual mean temp.	1.78	0.28	6.42	1.89e-07	***	
Annual mean UV-B	-1.97	0.27	-6.53	1.36e-07	***	
Annual mean temp: Annual precip.	0.85	0.27	3.12	0.004	**	
<i>C. Geographic variables - No Westwards</i>						0.48
Intercept	4.77	0.19	25.41	<2e-16	***	
Elevation	-1.03	0.19	-5.41	60.06e-06	***	
<i>D. Bioclimatic variables-No Westwards</i>						0.58
Intercept	4.4831	0.2027	22.117	<2e-16	***	
Annual mean	1.9553	0.3233	6.048	1.22e-06	***	

temp.						
Annual mean UV-B	1.66	0.31	-5.34	8.90e-06	***	
Annual mean temp: Annual precip.	0.87	0.32	2.74	0.010	*	

Table 3.3: Logistic regression of iNaturalist data for pollen. The response variable was coded as 0=yellow, 1=purple. All variables were Z-transformed. Significance values: 0 “****”, 0.001 “***”, 0.01 “**”, 0.05”.”

Model	Estimate	Std. Error	t value	pr(> t)	Significance	R²
A. Geographic variables						0.41
Intercept	2.11	0.11	18.55	<2e-16	****	
Elevation	-0.85	0.07	-11.50	<2e-16	****	
Longitude	-2.35	0.11	-20.20	<2e-16	****	
Latitude	0.65	0.08	7.90	1.36e-15	****	
B. Bioclimatic variables						0.30
Intercept	0.79	0.06	12.27	<2e-16	****	
Annual precip.	-1.77	0.10	-18.49	<2e-16	****	
Annual precip: Annual mean temp.	1.60	0.17	9.59	<2e-16	****	
Annual precip: Annual mean UV-B	-0.42	0.12	-3.41	0.001	**	
C. Geographic variables-No Westwards						0.19
Intercept	-0.27	0.07	-3.69	0.000	****	
Elevation	-1.43	0.13	-11.23	<2e-16	****	
Latitude	-1.98	0.17	-11.98	<2e-16	****	
Longitude	0.91	0.14	6.67	2.54e-11	****	

<i>D. Bioclimatic variables-No Westwards</i>						0.15
Intercept	-0.68	0.09	-7.60	3.05e-14	***	
Annual mean temp.	1.21	0.15	8.21	<2e-16	***	
Annual precip.	-0.35	0.09	-3.83	0.000	***	
Annual mean temp: Annual precip.	0.58	0.107	5.39	7.05e-08	***	
Annual mean temp: Annual mean UV-B	0.33	0.09	3.67	0.000	***	

Table 3.4: Confusion matrix for pollen color. (A) are all occurrence records, while (B) is with the west ward section of the range removed. Precision values measure what percentage of the positive predictions are actually positive. Recall is the true positive rate. F1 score is the harmonic mean of the two. Support is the number of total classifications.

A.	Precision	Recall	F1-score	Support
Purple	0.73	0.76	0.74	185
Yellow	0.91	0.89	0.90	502
Accuracy			0.86	687
Macro Avg	0.82	0.83	0.82	687
Weighted Avg	0.86	0.86	0.86	687
B.				
Purple	0.80	0.88	0.83	184
Yellow	0.81	0.71	0.76	142
Accuracy			0.80	326
Macro Avg	0.81	0.79	0.80	326
Weighted Avg	0.80	0.80	0.80	326



Figure 3.1: Swatch and examples of flower colors (petal and pollen). (Top) is the color swatch used for petal surveys of *G. maculatum*. (Bottom-left) is an example of a light (code=12), yellow pollen flower. (Bottom-right) is an example of a dark (code=15), purple pollen individual.

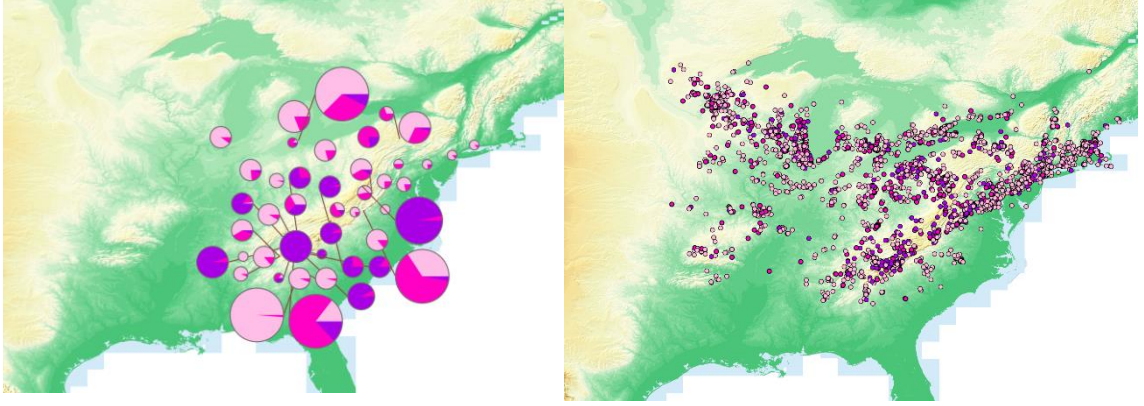


Figure 3.2: Map of petal color. Survey (left), iNaturalist (right). Pie charts size represent population size of reproductive individuals at time of sampling. The class bound represented are <50 and 1000+, see Supplemental Table 3.1 for more detail.

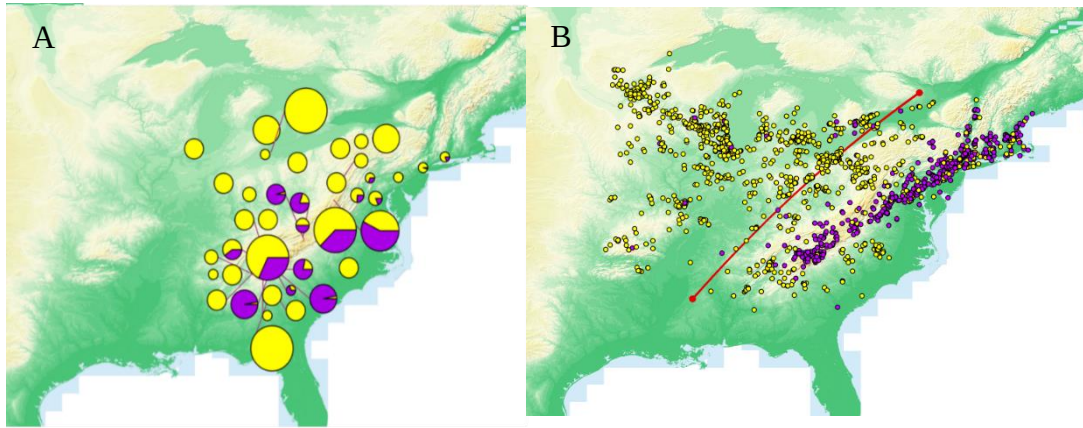


Figure 3.3: Map of pollen color. Pollen color frequency in populations (A). Pie charts represent estimated reproductive population size as indicated from Table S1. Individual occurrence records from iNaturalist colored by pollen color (B). Red line indicates the break where ‘westward’ records were removed in accordance with the monomorphic yellow populations of the west.

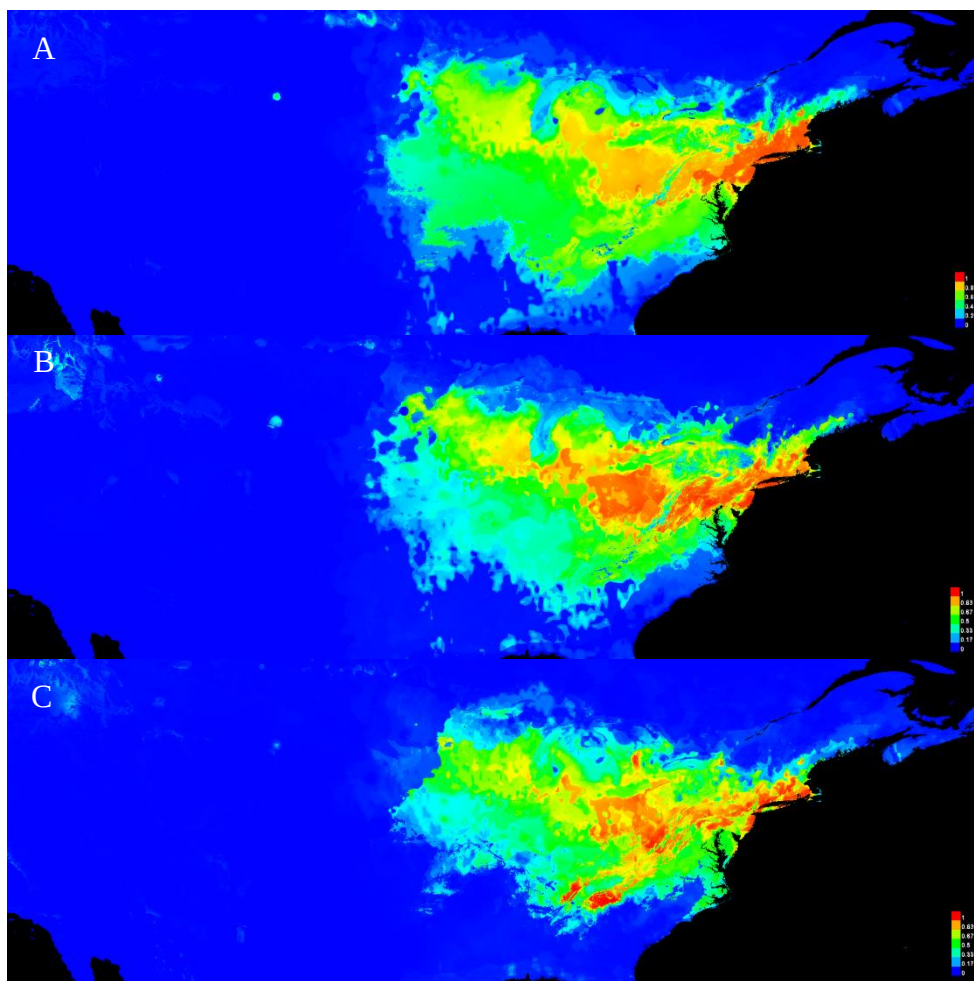


Figure 3.4: Maxent models for Light (A), Medium (B) and Dark (C) iNaturalist petal occurrence records. Light and medium model used linear, threshold and quadratic features and a regularization of 1.5. Dark model used linear and threshold. Schoender's D (dark x light: 0.807, CI 95: 0.845; dark x medium 0.826, CI 95: 0.852; light x medium: 0.872, CI 95: 0.868) and I (dark x light: 0.959, CI 95: 0.971; dark x medium: 0.958, CI 95: 0.971; light x medium: 0.977, CI 95: 0.977).

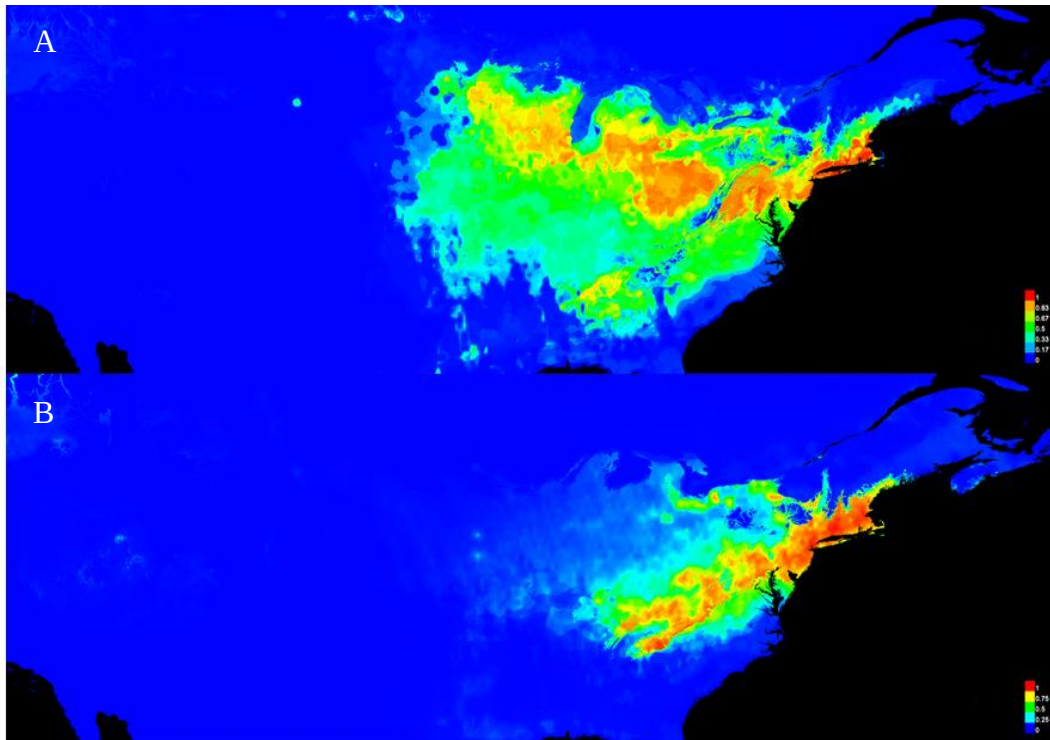


Figure 3.5: Differences in habitat suitability between the two pollen morphs. Yellow pollen habitat suitability (A) used features linear, product and threshold with a regularization of 2. Purple habitat suitability (B) uses features quadratic and threshold and regularization of 3.

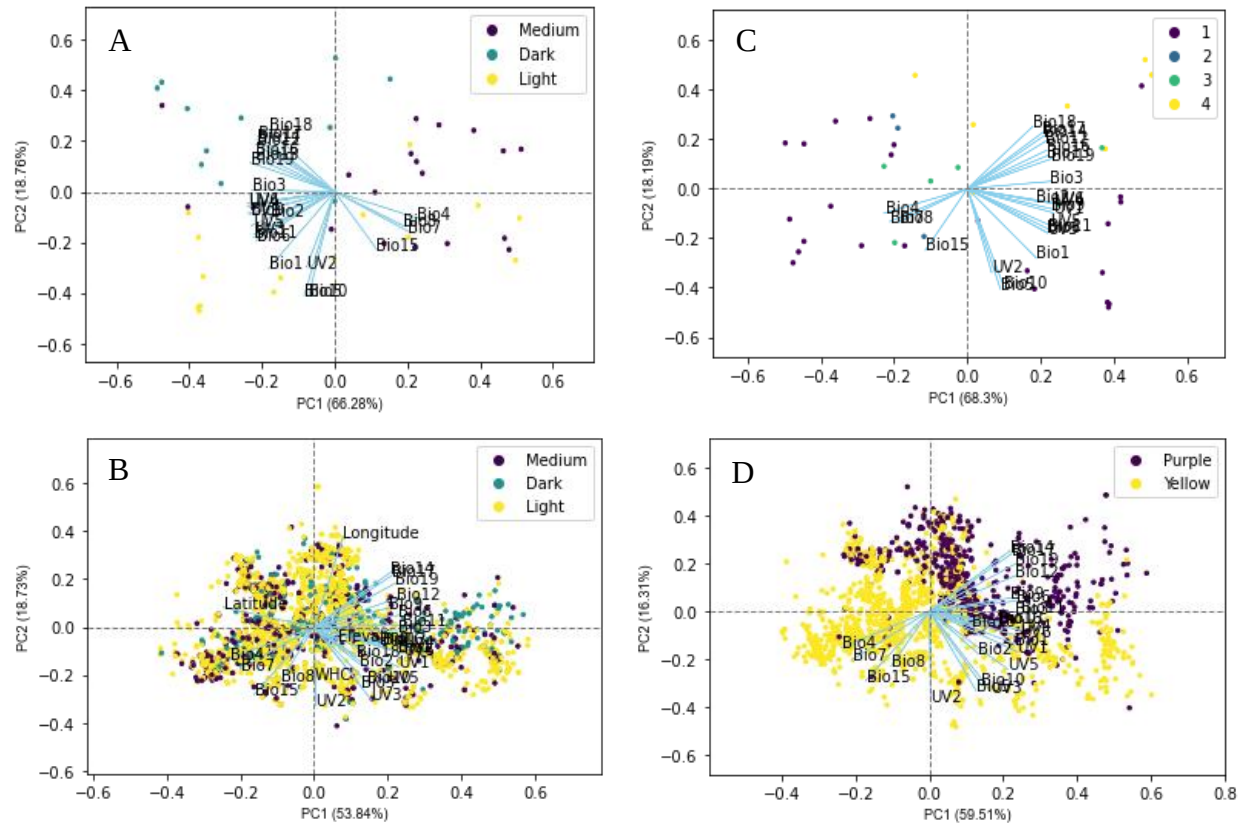


Figure 3.6: PCA biplot. (A) survey populations for petal color. Populations with color averages between 11-11.99 were colored as ‘light’, between 12-13.9 as ‘medium’ and >14 as ‘dark.’ (B) iNaturalist petal color occurrences. (C) survey pollen color as frequency of yellow pollen. Key categories: 1: 100% yellow pollen, 2: 75-99% yellow, 3: 50-75% yellow, 4 under 50% yellow. (D) iNaturalist pollen color occurrence.

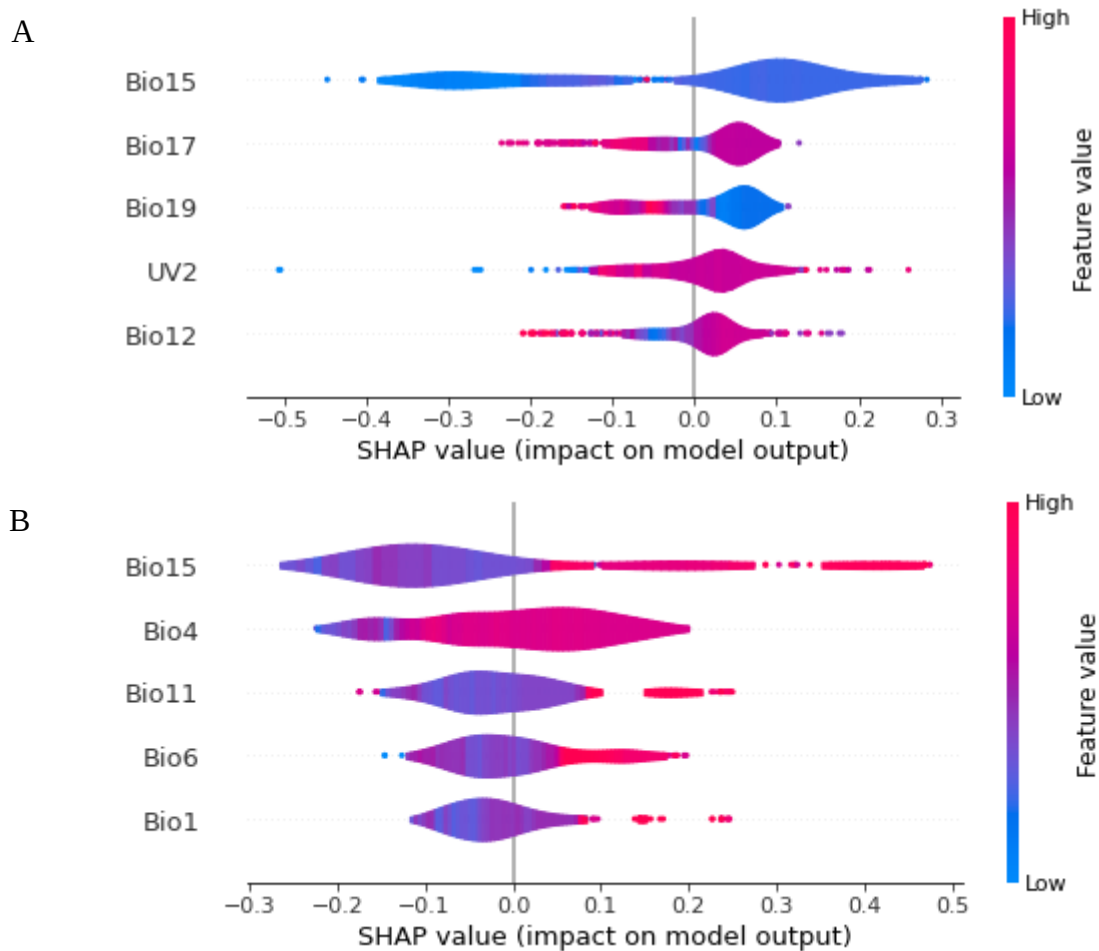
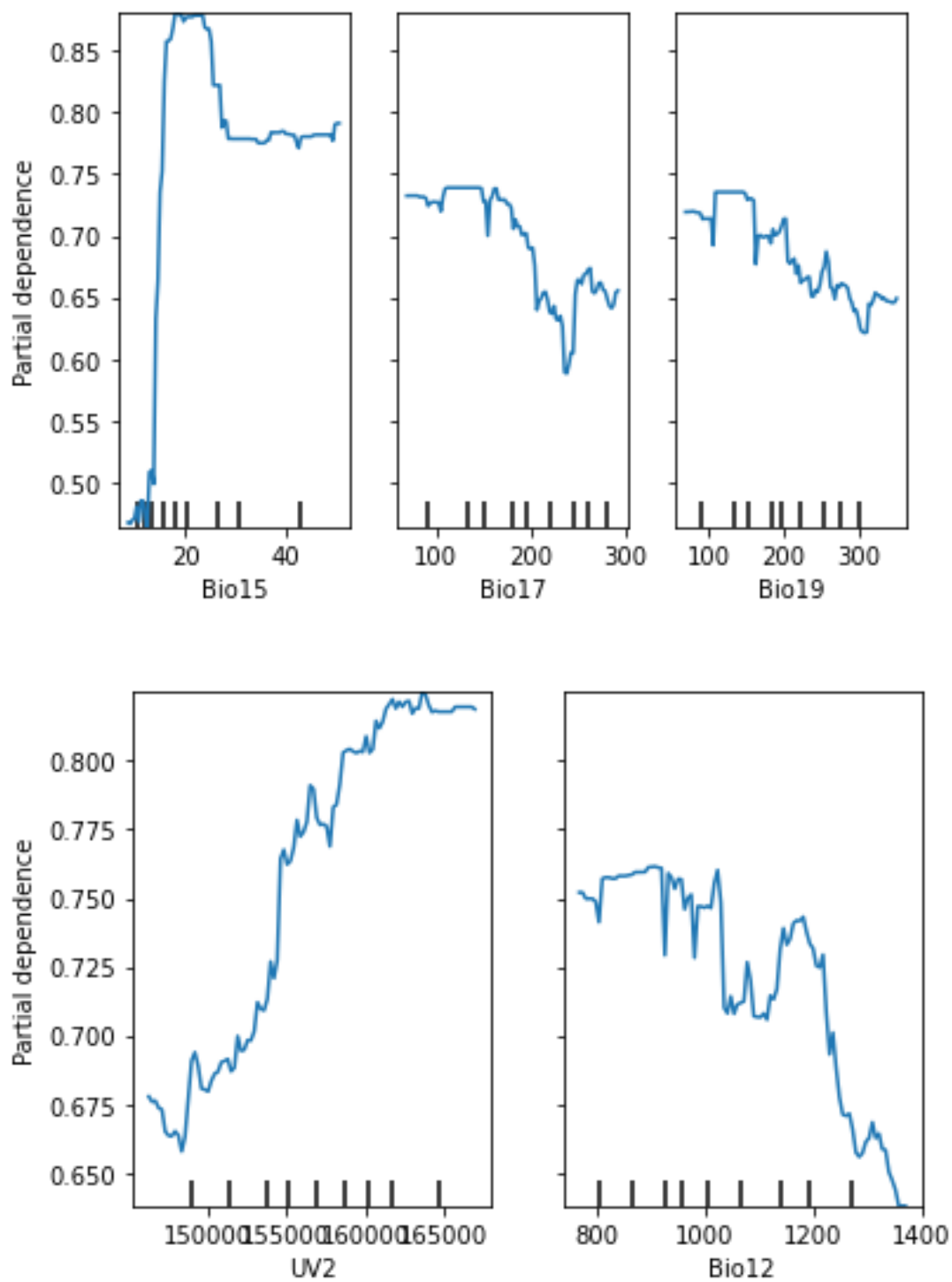


Figure 3.7: Shap summary violin plot for pollen color classification. Each feature contributes positively, towards a classification of yellow, or negatively towards a classification of purple. Shap summary violin plot, showing the impact each feature has on the prediction with warmer (pink) colors representing high feature values and cooler (blue) colors representing low feature values. Random forests do not require feature transformation. The (A) plot shows the plot for all occurrence records. The average values and standard deviation of each feature are 21.90/12.47, 192.83/70.27, 202.17/83.06, 156640.3/6173.29, 1030.31/190.97 and in descending order. B (bottom) shows the plot when westwards is removed. The average values and standard deviation of each feature are 13.00/3.46, 861.07/75.81, 0.47/2.86, -5.60/2.76, 11.32/2.15 in

descending order. Bio15= precipitation seasonality, Bio17= precipitation of driest quarter, Bio19= precipitation of coldest quarter, UV2= UV-B seasonality, Bio12= annual precipitation. Bio4=temperature seasonality, Bio11=mean temperature of coldest quarter, Bio6=min temperature of coldest month, Bio1=mean annual temperature.

A.



B.

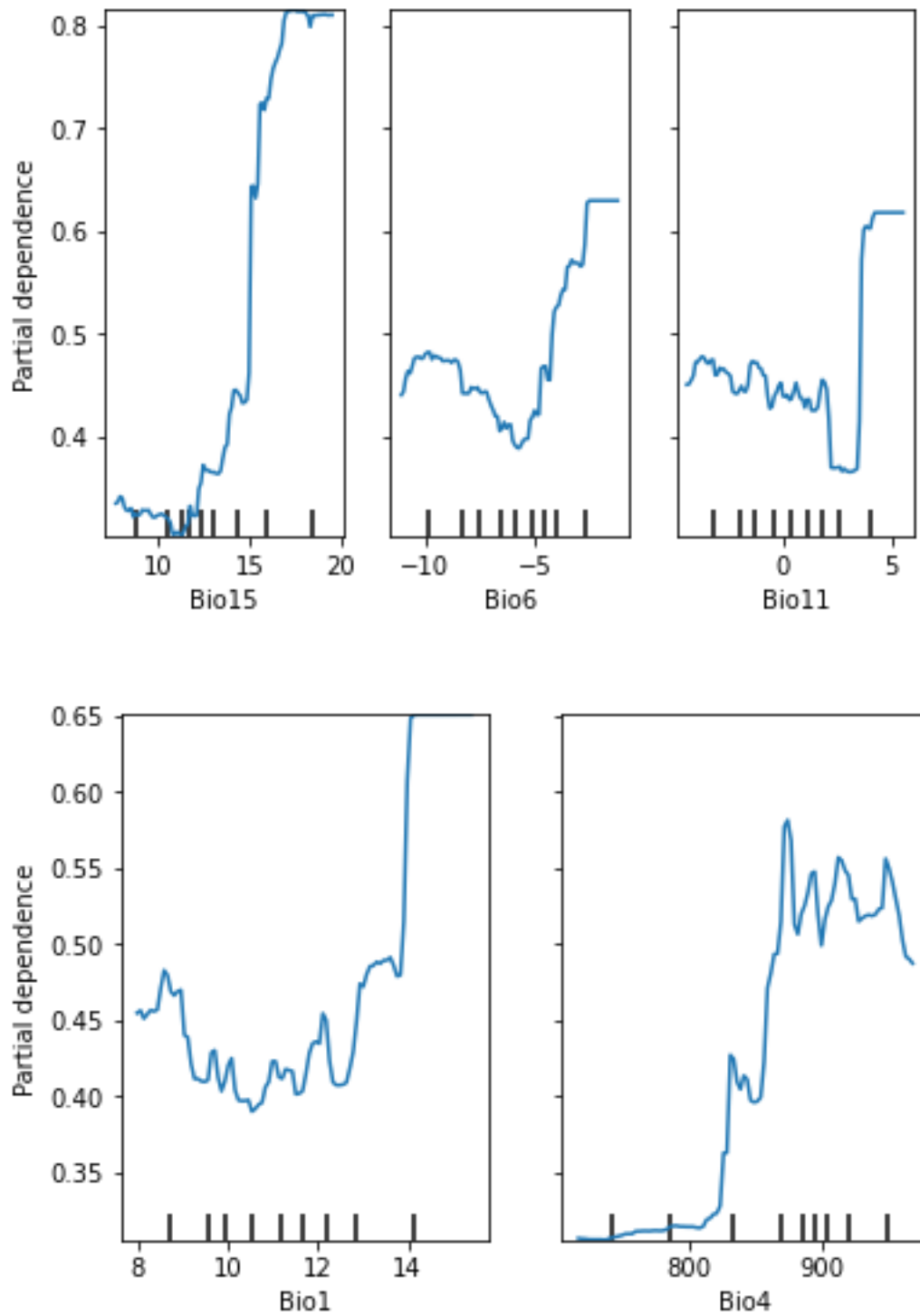


Figure 3.8: One-way partial dependence plots. Partial dependence, the marginal effect of

a single feature on the prediction class when all other features are controlled for, is shown on the y-axis as the probability of being classified as yellow. The average prediction is plotted in dark blue. (A) represents all pollen occurrence records, (B) represents when the west-wards records are removed.

References

- Ahmad, I., J. P. Day, M. V. MacDonald, and D. S. Ingram. 1991. Haploid culture and UV mutagenesis in rapid-cycling *Brassica napus* for the generation of resistance to chlorsulfuron and *Alternaria brassicicola*. *Annals of Botany*. 67(6):521
- Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38:541-545. URL: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01132>.
- Amico, G.C., M. A. Rodriguez-Cabal and M. A. Aizen. 2010. Geographic variation in fruit colour is associated with contrasting seed disperser assemblages in a South-Andean mistle. *Ecography*. 34(2):318-326
- 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. *Jour. of Eco.* 101: 1613-1622
- Armbruster, W. S. 2002. Can indirect selection and genetic context contribute to trait diversification? A transition-probability study of blossom-colour evolution in two genera. *J. of Evo. Biol.* 15:468–486
- Beckmann, M., T. Václavík, A. M. Manceur, L. Šprtová, H. von Wehrden, E. Welk, and A. F. Cord. 2014. glUV: a global UV-B radiation data set for macroecological studies. *Methods in Ecology and Evolution*. 5(4): 372-383.
- Berardi, A., P. D. Fields, J. L. Abbate, D. R. Taylor. 2016. Elevational divergence and clinal variation in floral color and leaf chemistry in *Silene vulgaris*. *Amer. Jour. of Botany*. 103(8): 1508-1523
- Bertin, R I., and O. D. V. Sholes. 1993. Weather, pollination and the phenology of *Geranium maculatum*. *Am. Midl. Nat.* 129(1):52-66
- Bohrerova, Z., G. Bohrer, K. D. Cho, M. A. Bolch, K. G. Linden. 2009. Determining the viability response of pine pollen to atmospheric conditions during long-distance dispersal. *Ecol. Appl.* 19:656-667.
- Breiman, L. 2001. Random Forests. *Machine Learning*. 45:5-32.
- Burgess, H. K., L. B. DeBey, H. E. Froehlich, N. Schmidt, E. J Theobald, A. K. Ettinger, J. HilleRisLambers, et al. 2016. The science of citizen science: Exploring barriers to use as primary research tool. *Biol. Conservation*. 208:113-120
- Calcagno, V. 2020. glmulti: Model Selection and Multimodel Inference Made Easy. R package version 1.0.8. <https://CRAN.R-project.org/package=glmulti>

- Chalker-Scott, L. 1999. Environmental significance of anthocyanins in plant stress responses. *Photochemistry and Photobiology*. 70:1-9
- Chang, S. 2006. Female compensation through the quantity and quality of progeny in a gynodioecious plant *Geranium maculatum* (Geraniaceae). *Am. J. Bot.* 93(2):263-70
- Conner, J. K. 2002. Genetic mechanisms of floral trait correlations in a natural population. *Nature*. 420:407-410
- Culter, D. R., T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, J. L. Lawler. 2007. Random forests for classification in ecology. *Ecology*. 88(11): 278-2792
- del Valle, J. C., M. L. Buide, I. Casimiro-Soriguer, J. B. Whittall, and E. Narbona. 2015. On flavonoid accumulation in different plant parts: variation patterns among individuals and populations in the shore campion (*Silene littorea*). *Front. Plant Sci.* 6: 939
- Demhick, S. M., and T. A. Day. 1996. Effect of enhanced UV-B radiation on pollen quantity, quality, and seed yield in *Brassica rapa* (Brassicaceae). *Am J. of Botany*. 83:573-579.
- Dick, C. A., J. Buenrostro, T. Butler, M. L. Carlson, D. J. Kliebenstein, and J. B. Whittall. 2011. Artic mustard flower color polymorphism controlled by petal-specific downregulation at the threshold of the anthocyanin biosynthetic pathway. *PLoS One*. 6(4):e18230
- Endler, J. A. 1977. Geographic variation, speciation and clines. Monographs in Population Biology. 10(3):1-246
- Epling, C., and T. Dobzhansky. 1942. Genetics of natural populations. VI. microgeographic races in *Linanthus parryae*. *Genetics*. 27 (3): 317- 332
- Ersts, P. J. Geographic distance matrix generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from https://biodiversityinformatics.amnh.org/open_source/gdmg/
- Feng, H., L. An, L. Tan, Z. Hou, and Z. Wang. 2000. Effect of enhanced ultraviolet-B radiation on pollen germination and tube growth of 19 taxa in vitro. *Environ. Exp. Bot.* 43:45-53
- Fink, S.E., and R.J. Hijmans, 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*. 37(12): 4302-4315.

- Galen, C., and P. G Kevan. 1980. Scent and color, floral polymorphisms and pollination biology in *Polemonium viscosum* Nutt. *The Am. Midl. Nat.* 104 (2): 281- 289
- Gould, K. S. 2004. Nature's Swiss Army Knife: The diverse protective roles of anthocyanins in leaves. *J. Biomed Biotechnol.* 5:314-320
- Grossenbacher, D., L Makler, M. McCarthy, and N. Fraga. 2021. Abiotic environment predicts micro- but not macroevolutionary patterns of flower color in monkeyflowers (Phrymaceae). *Front. Plant Sci.* 12:636133.
- Haldane, J.B.S. 1948. The theory of a cline. *J. of Genetics.* 48 (3): 277-284
- Hamrick, J. L., and R.W. Allard. 1972. Mircogeographical variation in allozyme frequencies in *Avena barbata*. *Proc. Nat. Acad. Sci.* 69 (8): 2100-2104
- Hochmair H.H., R. H. Scheffrahn, M. Basille, and M. Boone. 2020. Evaluating the data quality of iNaturalist termite records. *PLOS One.* 15(5): e0226534
- Hoffman, A.A., and P. A. Parson 199. Evolutionary genetics and environmental stress. Oxford University Press.
- Ida, T. Y., and Ø, Totland. 2014. Heating effect by perianth retention on developing achenes and implications for seed production in the alpine herb *Ranunculus glacialis*. *Alpine Botany.* 124: 37-47
- iNaturalist. Website. from <https://www.inaturalist.org>.
- Jewell, J., J, Mckee, and A, Richards. 1994. The keel colour polymorphism in *Lotus corniculatus* L.: differences in internal flower temperature. *New Phytologist.* 128(2): 363-368
- Jorgensen, T.H, and S. Andersson. 2005. Evolution and maintenance of pollen-colour dimorphisms in *Nigella degenii*: habitat-correlated variation and morph-by-environment interactions. *New Phytologist* 168: 487– 498.
- Killen, H., Chang, L., Soul, L. and Barclay, R., 2022. Combining Physical and Digital Data Collection for Citizen Science. Climate Research. *Citizen Science: Theory and Practice.* 7(1): 10
- Kim, S. G. Hwang, S. Lee, J. Zhu, I. Paik, T. T. Nguyen, J. Kim, and E. Oh. 2017. High ambient temperature represses anthocyanin biosynthesis through degradation of HY5. *Front. Plant Sci.* 8. 1787
- Koski, M. K., and T. Ashman. 2015A. Floral pigmentation patterns provide an example of glomer's rule in plants. *Nature Plants.* 1: 14007

- Koski, M. K., and T. Ashman. 2015B. An altitudinal cline in UV floral pattern corresponds with a behavioral change of a generalist pollinator assemblage. *Ecology*. 96 (12): 3343-3353
- Koski, M. H., and L.F. Galloway. 2018. Geographic variation in pollen color is associated with temperature stress. *New Phytologist*. 218(1):370-379
- Koski, M.H., D. MacQueen, T. Ashman. 2020. Floral pigmentation has responded rapidly to global change in ozone and temperature. *Current Biology*. 30:4425-4431
- Kovinich, N., G. Kayanja, A. Chanoca, M. S Otegui, and E. Grotewold. 2015. Abiotic stresses induce different localizations of anthocyanins in *Arabidopsis*. *Plant Signal Behav.* 10(7): e1027850
- Lacey, E., and D. Herr. 2005. Phenotypic plasticity, parental effects, and parental care in plants? I. An examination of spike reflectance in *Plantago lanceolata* (Plantaginaceae). *Am J Bot.* 92: 920–930
- Lacey, E., M. E. Lovin, S. J. Richter, and D. Herington. 2010. Floral reflectance, color, and thermoregulation: What really explains geographic variation in thermal acclimation ability of ectotherms? *The American Naturalist*. 175(3):335-349
- 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
- Landis, R. J., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*. 33(1):159-174
- Lau, J. A., and L.F. Galloway. 2004. Effects of low-efficiency pollinators on plant fitness and floral trait evolution in *Campanula americana* (Campanulaceae). *Oecologia*. 141(4): 577-583
- Llorens, L., F.R. Badenes-Pérez, R. Julkunen-Titto, C. Zidorn, A. Fereres, and M. A. K. Jansen. 2015. The role of UV-B radiation in plant sexual reproduction. *Perspectives in Plant Ecology, Evolution and Systematics*. 17(3):243-254
- Long, J.A. 2019. interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions. R package version 1.1.0, [URL:https://cran.rproject.org/package=interactions](https://cran.rproject.org/package=interactions).
- Luan, J., C. Zhang, B. Zu, Y. Xue, and Y. Ren. 2010. The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research*. 227: 105534
- Lunau, K., and J.E. Maier. 1995. Innate colour preferences of flower visitors. *Journal of Comparative Physiology*. 177: 1-19

- Lunau, K. 2000. Notes on the colour of pollen. *Plant Systematics and Evolution*. 198(3/4):235- 252
- Lundburg, S.M., and S. Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 30.
- Martin, C. M., 1965. An ecological life history of *Geranium maculatum*. *The Am. Mid. Nat.* 73(1):111-149
- McCall, A.C., S. J. Murphy, C. Venner, and M. Brown. 2013. Florivores prefer white versus pink petal color morphs in wild radish, *Raphanus sativus*. *Oecologia*. 172: 189-195
- McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem. Med.* 22(3): 276-282
- Mierziak, J., K. Kostyn, A. Kulma. 2014. Flavonoids as important molecules of plant interactions with the environment. *Molecules*. 19(10):16240-65
- Mohammadrezakhani, S., J. Hajilou, and F. Rezanejad. 2017. Evaluation of phenolic and flavonoid compounds in pollen grains of three Citrus species in response to low temperature. *Grana*. 57(3):1-9
- Muhlemann, J.K., T. L. B. Younts, G. K. Muday. 2018. Flavonols control pollen tube growth and integrity by regulating ROS homeostasis during high-temperature stress. *PNAS*. 115(47): e11188-e11197
- Newman, E., B. Anderson, and S. D. Johnson. 2012. Flower colour adaptation in a mimetic orchid. *Proc. of the Royal Soc. B*. 279: 2309-2313
- Oksanen, J.F., G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R.B. O'Hara, et al. 2019. vegan: Community ecology package. R package version 2.5-6. URL <https://CRAN.R-project.org/package=vegan>
- Peach, K., J. W. Liu, S. J. Mazer. 2020. Climate predicts UV floral pattern size, anthocyanin concentration, and pollen performance in *clarkia unguiculata*. *Front. Plant Sci.* 11:847
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12: 2825-2830.
- Phillips, S.J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modeling*. 190: 231-259

- R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rausher, M. D., 2008. Evolutionary transitions in floral color. *Int. J. Plant Sci.* 169 (1): 7-21
- Rohman M., L. B. Prasetyo, and M. D. Kusrini. 2020. Predicting spatial distribution of Asian Horned Frog (*Imegophrys montana* Kuhl & Van Hasselt 1882) in Java Island using citizen science's data. IOP Conf. Series: Earth and Environmental Sci. 771: 012027
- Schemske, D. W., and P. Bierzychudek. 2001. Perspective: evolution of flower color in the desert annual *Linanthus parryae*: Wright revisited. *Evolution*. 55 (7): 1269-1282
- Schiestl, F. P., and S. D. Johnson. 2013. Pollinator-mediated evolution of floral signals. *Trends in Ecology and Evolution*. 28(5):307-315
- Schoener, T. W. 1968. Anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology* 49: 704– 726.
- Sobral, M., I. P. Neylan, E. Narbona, and R. Dizro. 2021. Transgenerational plasticity in flower color induced by caterpillars. *Front. Plant Sci.* 12:617815
- Stiles, E.A., N. B. Cech, S. M. Dee, and E. P. Lacey. 2007. Temperature-sensitive anthocyanin production in flowers of *Plantago lanceolata*. *Physio Plant.* 129(4); 756-765
- Storz, J. F. 2002. Contrasting patterns of divergence in quantitative traits and neutral DNA markers: analysis of clinal variation. Jay F. Storz Publications. 16.
- Streisfeld, M.A., and J. R. Kohn. 2005. Contrasting patterns of floral and molecular variation across a cline in *Mimulus aurantiacus*. *Evolution*. 59 (12): 2548-2559
- Streisfeld, M.A., and J. R. Kohn. 2007. Environment and pollinator-mediated selection on parapatric floral races of *Mimulus aurantiacus*. *J. of Evol. Biol.* 20(1):122-132
- Suárez-Mota, M. E., and J. L. Villaseñor. 2020. Ecological niche overlap among species of the genus *Zaluzania* (Asteraceae) from the dry regions of Mexico. *Plant Ecology and Evolution*. 153(3): 337-347
- Tang, X-X., Huang, S-Q. 2010. Fluctuating selection by water level on gynoeceum colour polymorphism in an aquatic plant. *Annals of Botany*. 106(5):843-848

- Trunschke, J., K. Lunau, G. H. Pyke, Z-X, Ren, and H. Wang. 2021. Flower color evolution and the evidence of pollinator-mediated selection. *Front. Plant Sci.* 12:617851
- Van der Kooi C. J., A. G. Dyer, P. G. Kevan, K. Lunau. 2019. Functional significance of the optical properties of flowers for visual signaling. *Annals of Botany*. 123(2): 263-276
- Van Etten, M. L., and S-M. Chang. 2013. Context-dependent pollinator-discrimination acts against females in the gynodioecious *Geranium maculatum*. *Annals of Botany*. 114:1769-1778
- Van Rossum, G., and F. L Drake. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
- Vasemagi, A. 2006. The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*. 173: 2411-2414
- Vittinghoff, E., D. V. Glidden, S. C. Shiboski, C. E. McCulloch. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. 2nd ed. 2012 edition. Springer; 2011.
- Warren, D. L., Glor. R. E. and M. Turelli. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62: 2868–2883.
- Warren, D. L., R. E. Glor, M. Turelli. 2010. ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*. 33(3): 607-611
- Warren. J., and S. Mackenzie. 2001. Why are all colour combinations not equally represented as flower-colour polymorphisms? *New Phytologist*. 151(1):237-241
- Wierzbicka, M., H. Galera, B. Sudnik-Wójcikowska, and B. Wilkomirski. 2013. *Geranium robertianum* L., plant form adapted to the specific conditions along railway: "railway-wandering plant". *Plant Systematics and Evolution*. 300, 973-985
- Ying-Ze, X., L. Jia, C. Zhang, and S. Huang. 2019. Color-matching between pollen and corolla: hiding pollen via visual crypsis? *New Phytologist* 224(3):1142-1150
- Zhang, Q., J. Zhai, L. Shao, W. Lin, and C. Peng. 2019. Accumulation of anthocyanins: an adaptation strategy of *Mikani micrantha* to low temperature in winter. *Front. Plant Sci.* 10: 1049
- Zhao, Y., W. Dong, K. Wang, B. Zhang, A. C. Allan, K. Lin-Wang, K. Chen, and C. Xu. 2017. Differential sensitivity of fruit pigmentation to ultraviolet light between two peach cultivars. *Front. Plant. Sci.* 8:1552

CHAPTER 4

PILOT RADCAP STUDY IDENTIFIES LACK OF ISOLATION BY DISTANCE, BUT FINDS ISOLATION BY ENVIRONMENT IN *GERANIUM MACULATUM*

Introduction

Geographic variation in phenotypes provide a great opportunity to study one of the most important topics in evolutionary biology: what are the consequences and causes of evolution? Previous work has identified that floral pigmentation of *Geranium maculatum* shows a clear pattern related to both geographic and environmental factors (Chapter 3). However, these correlative results do not provide definitive answers to whether natural selection is responsible for the distribution patterns we observed. One alternative explanation is that the distribution patterns we observed are the results of non-selective mechanisms, including migration and genetic drift. To test this alternative hypothesis, we took a population genetic approach using single nucleotide polymorphisms (SNPs) markers to test whether the patterns we observed will allow for the evaluation of population genetic structure (F_{st}), and differentiation within the framework of phenotypic differentiation (P_{st}).

Phenotypic differentiation among population may reflect the combined effect from multiple processes, including natural selection, genetic drift, migration, and gene flow. Each of these processes could enhance or hinder population divergence, but they may also interact with each other in a spatially heterogeneous fashion to result in a more complex pattern of genetic and phenotypic variation across a larger geographic scale. For

example, the ability for populations to experience gene flow can be governed by geographic barriers (such as mountains, rivers, etc.), selective barriers (such as pollinator preference) and/or distance barriers (Sexton et al., 2014). Distance between populations can provide barrier for gene flow, especially for sessile organisms such as plant – a phenomenon referred to as isolation by distance (commonly termed IBD). It is, thus, expected that the further populations are geographically, the homogenizing effect of gene flow would decrease and the more differentiated these populations would be due to random allelic fluctuations (Wright 1943). However, other impediments, such as geographic barriers or selection against immigrants, can also restrict gene flow in the process of isolation by environment (IBE). IBE is a concept similar to IBD, but it emphasizes the similarity of environment (such as elevation or selecting agents) that allows greater gene flow than dissimilar environments (Wang and Bradburd, 2014). IBE also implies that selection plays a role in population differentiation. Comparing the relative contribution of IBE and IBD to the genetic and phenotypic differentiation across populations from a large geographic scale will provide insight into the evolutionary mechanisms likely to be responsible for floral pigmentation we observed in *G. maculatum*. It is also worth noting that IBD and IBE are not mutually exclusive and may work synergistically in nature.

We use clustering algorithm, STRUCTURE (Pritchard et al., 2000) and analysis of molecular variance (AMOVA) (Excoffier et al., 1992) to evaluate the role of gene flow between populations in *G. maculatum* and interpret the results in the context of population floral pigmentation. In our previous work, we identified that darker pollen and petals in *G. maculatum* are frequently found in higher elevation populations. We would

expect if selection were the underlying mechanism that darker populations would cluster separately than lighter populations, additionally, there would be evidence of population structure between populations in different elevational transects. We test for IBD and IBE using Mantel and Partial Mantel tests. Furthermore, genetic, and phenotypic differentiation can be compared to assess the role of drift and selection on quantitative traits (Spitze, 1993). This assumes that F_{st} reflects differentiation caused by genetic drift (Reynolds et al 1983), therefore, the metric Q_{st} , which is phenotypic variation, can be compared to assess if Q_{st} also falls under what would be expected under neutral processes. Comparisons of phenotypic differences (Q_{st}) against neutral marker differentiation (F_{st}) is a commonly used test to indicate if selection is influencing a population. If $Q_{st}=F_{st}$, then we assume that drift influences that phenotypic trait as neutral expectations are met. However, a Q_{st} greater than F_{st} indicates directional selection, while a Q_{st} less than F_{st} indicates stabilizing or uniform selection as the cause of phenotypic differentiation (Leinonen et al., 2013). However, due to the extensive breeding scheme required in estimating Q_{st} and the difficulties due to the long generation time in *G. maculatum*, we will use P_{st} as the approximation for Q_{st} . P_{st} does not consider any distinction between genetic and environmental variation that could cause floral pigmentation (Brommer, 2011; Leinonen et al., 2013). The advantage of using P_{st} is that it can come directly from our previous field survey, however, results should be taken with caution and approached conservatively (Brommer, 2011).

By combining genetic data with survey data, a comparison of phenotypic and genetic differentiation (P_{st}/F_{st}) can be tested, as we would expect patterns between neutral loci (F_{st}) and floral traits (P_{st}) to be equal under drift or dissimilar under selection. If

phenotypic differentiation is higher or lower than genetic differentiation, we can attribute selection to floral pigmentation. However, if phenotypic and genetic differentiation are the same, we can attribute floral pigmentation to genetic drift.

In this study, we chose to develop SNPs markers as a way to improve replicability and provide a universal set of baits that can be used across multiple projects on this species (Coates et al., 2009). To develop SNPs, we took the RADcap approach, a more economical alternative to the sequencing approach, that encompasses the pros of both 3RAD and sequence capture (Hoffberg et al., 2016). The strength of this sequence capture approach is the repeatable sequencing of specific loci that often obtain a high percentage of on target reads. This approach has been made even more desirable when paired with the much-reduced costs for the 3RAD library preparation (Hoffberg et al., 2016). This approach proved to be successful in our small pilot study using 24 samples total from 14 populations across the species distribution range. A population was defined as a deme of plants within a mile radius of one another. We identified putative loci and created custom baits to enrich libraries of the remaining project's 3RAD libraries. This will also us to measure population diversity and structure within *G. maculatum* samples at a greater number of loci than done previously, while increasing the confidence in clustering and population diversity results (Zimmerman et al., 2020).

In this study we describe the workflow it takes to go from DNA extraction to sequencing data, including the troubleshooting we had to purify extractions, and manipulating the library preparation protocol for our species. Finally, we analyze 93 individuals of *G. maculatum* that underwent sequence capture for population structure and diversity. Our goals for this study were to answer 1) Is RADcap a feasible method for

SNP loci discovery for our *G. maculatum* samples? 2) Can the phenotypic variation in pollen and petal colors observed among populations be explained by the neutral genetic differentiation (F_{st})? 3) Is there any evidence for selection on these color traits using the metric of the difference between P_{st} and F_{st} (i.e., $P_{st}-F_{st}$) to account for the populational and demographic history in our samples?

Methods

Pilot study on RADcap and SNPs discovery

The pilot study followed the protocol established by Hoffberg et al. (2016) to make 3RAD libraries of a small number of samples in order to design species specific baits for RADcap capture. For this pilot study, we used 24 individuals of *Geranium maculatum* from populations across the Appalachian Mountains (Table 4.1). DNA was extracted from silica dried leaf samples using a Qiagen DNeasy Plant Kit (QIAGEN) and samples were normalized to the concentration of 20ng/μl of DNA. The 3RAD library prep followed a modified protocol from Graham et al. (2015.) The samples underwent a one-hour digestion with the restriction enzymes BAMHi, MSP1, and CLA1. During digestion, a unique internal adapter was ligated onto the genomic fragments to enable barcoding. Unique barcoded primers for *itru5* and *itru7* (023 and 110) were added (Glenn et al., 2019) to each sample during a 14 cycle PCR. Samples were subsequently cleaned with 1.25x streptavidin coated magnetic beads (Rohland and Reich, 2012), plate read, pooled, and size selected for 550 bp fragments via Pippin Prep (Sage Science). These prepared samples were sequenced on a HiSeq X using a PE150 kit with a target of 24 million reads (1 million reads/sample).

We used the software pipeline Stacks (Catchen et al., 2013) to assemble the denovo loci and to identify the single nucleotide polymorphisms (SNPs) in our samples. Stacks has been commonly used for SNPs analysis and we will briefly describe the steps we took in this study. After the raw data with our sequences were demultiplexed and run through quality control, we used the *process_rad_tags* function in Stacks to trim all reads to 139 bp, as Stacks requires all reads to be the same length for subsequent loci assembly. Once all reads were cleaned up, Stacks took the reads to sort into multiple sequence stacks, each containing identical reads from the same sample.

Several parameters in Stacks require adjustment by users (See Paris et al., 2017 or Catchen et al., 2013 for a in depth review of the parameters). Parameter *m* specifies the number of reads required to form a stack. We retained the default value of *m*=3 since it has been shown to be suitable for multiple species (Paris et al., 2017), and a higher value does not appear to have a significant impact on polymorphism detection. Parameter *M* (note, distinct from *m*) specifies the number of nucleotide differences between stacks of the same sample that would be tolerated as allelic, and consequently merging these stacks into a single locus (stack). Parameter *n* specifies the threshold for the number of fixed nucleotide differences across samples that can be considered as alleles of the same locus. Both *M* and *n* were chosen by examining the resulting number of polymorphic loci. As the values of *M* and *n* increase, more stacks would be (correctly) collapsed and the number of polymorphic loci would gradually reduce. However, there exists a threshold where stacks of different loci start to be incorrectly collapsed which is signaled by a sudden big drop of the resulting polymorphic loci number. We used the parameter values before this drop (*M*=4, *n*=2) as the preferred value in our analysis (Paris et al., 2017). The

population script was run with the specification that at least 60% of the individuals within a population has the locus ($r = 0.6$) and at least two populations contain a particular locus ($p = 2$). We requested only one SNP per locus (*write single snp*) to avoid linkage between the SNP loci. The resulting *populations.snps.genepop* file was uploaded as a .gen file to be filtered for missing data using R package ‘Poppr’ (Kamvar, et al., 2014). Loci were filtered with the ‘missingno’ command (with threshold value = 30%) for removing loci and samples with missing data. A small number of samples were removed in this process (see results for more detail).

With the curation steps described above, we identified a total of 2,320 consensus sequences for our *G. maculatum* samples which were submitted to Arbor Biosciences for custom bait design and synthesis. Of those submitted, 1,814 were polymorphic (Arbor Biosciences, Ann Arbor, Michigan). A combined *myBaits Custom 20-40k* kit with *Pinus taeda* (as a benefit of RADcapture, baits can be designed with multiple species, thus further lowering the cost of buy-in) with 16 reactions was ordered with 1-2 baits per target. Arbor designed 80nt-overlapping baits. These baits were blasted against the organelle genome of both organisms and *P. taeda*’s nuclear genome. Baits were then blasted against themselves to ensure no bait-to-bait hybridization or target redundancy.

Full sample library preparation and capture

Our full sample list includes 25 populations and 500 *G. maculatum* individuals, but this pilot study only includes sequences from 1/5 of the samples, due to troubleshooting 3RAD library preparation and bait capture for this species (Figure 4.1). Dried leaf tissues were used for DNA extractions. Samples were sheered by glass beads and a tissulyzer, and a modified CTAB protocol (Cota-Sánchez, et al. 2006) was used to

extract their DNAs. The dried, ground samples were incubated with CTAB and (2-Mercaptoethanol) BME for 30 min and mixed periodically, followed with the chloroform/isoamyl extraction twice before DNA was precipitated by centrifuging in room temperature with isopropanol.

3RAD library prep used BadDNA's 3RAD with Molecular IDs protocol (found at: <https://baddna.uga.edu/protocols.html>) 3RAD Library prep was attempted for a plate's worth of samples but yielded poor results using the previous protocol. An RNase clean up step was added prior to library preparation to remove degraded RNA fragments that would act as an inhibitor. Each sample had 1µl of RNase added to 20µl of DNA extract. The reaction incubated for 1 hour at room temperature, followed by a 2x bead clean. Plates underwent either a three hour, or 40-minute digestion. Standard protocol ligation was followed, followed by 15 cycles of PCR and a 1.25x bead clean. 100 samples were pooled according to Table 4.2-4.3 for two, 100 sample sets for bait capture.

Arbor Biosciences' standard protocol for bait capture was used (Arbor Biosciences, 2022). Because the bait set used was custom, we expected target-bait divergence to be low, therefore the temperature for hybridization and wash was chosen to be 65°C, according to the protocol. The plant specific blockers to prevent random annealing were added. Then a wash that binds the biotin on the baits to magnetic beads to remove off target and poorly annealed samples was completed. Finally, workflow B was chosen for library recovery. The final enriched library was sequenced on a NovaSeq S4 with a PE150 kit at NC State's Genomic Lab. Stacks was used on returned sequences for loci recovery and SNP calling with the same population command as in the pilot.

Genetic diversity and population structure

We conducted our preliminary analysis of populations structure using the currently available 93 samples from the successful sequencing run (Table 4.4, see Table 4.5 for population geographic information). We estimated the summary statistics of heterozygosity, F_{is} , private alleles, and pairwise F_{st} using Stacks. Strength of differentiation was described as by Hartl and Clark (1997). In addition to F_{st} , we also used STRUCTURE to detect population genetic structure (Pritchard et al., 2000). STRUCTURE was run with the correlated allele frequencies model. Samples were assigned populations using PopData. A burn in period of 5,000 was followed by 5,000 MCMC iterations. Values of the cluster number K were evaluated from one to 12 with three replicates each. Optimal number of K clusters were chosen using the Evanno method (Evanno et al., 2005) via Structure Harvester (Earl and vonHoldt, 2011). Visualizations were created with CLUMPAK (Kopelman et al., 2015). AMOVA was conducted with hierarchical partitions using the Poppr ade4 AMOVA implementation in R. (Kamvar, et al., 2014; R Core Team, 2021). Populations were categorized into populations and two regions (east and west, and mountain (over 900m). Populations over 900m were all recorded to be ‘dark’ populations with an average petal color of over 13.5 and a high purple pollen frequency.

Isolation by distance (IBD) was tested by examining the relationship between pairwise $F_{st}/(1-F_{st})$ and geographic distance (ln transformed) with a Mantel test in R using the vegan package (Oksanen et al., 2019). Geographic distance was calculated as the ln transformed linear distance between populations using the American Museums of Natural History’s Geographic Distance Matrix Calculator (Found at: https://biodiversityinformatics.amnh.org/open_source/gdmg/, Ersts.). We also conducted

partial Mantel tests to assess the degree of Isolation by Environment (IBE), focusing on the elevation of populations, while controlling for the geographic distance between population pair (Wang and Bradburd, 2014). To investigate whether elevation difference and geographic distance are correlated, we also conducted a separate Mantel test between these two matrices.

P_{st} vs F_{st} comparisons were conducted to assess differentiation in floral pigmentation (petal and pollen color) against neutral genetic differentiation by a Mantel test. Separate pairwise P_{st} was calculated for petal, and pollen, color between populations using the R package Pstat (Silva and Silva, 2018). Only populations with > than five samples were included in this analysis. Color measurements were taken as described in in Chapter 3. Two parameters, c and h^2 , are used in calculating P_{st} . Parameter c is a scalar variable that states how much of the among population variance is due to additive genetic variance, and h^2 is narrow sense heritability (Seymour et al., 2019; Brommer, 2011) for the trait of interest. Because the heritability of these traits is unknown, we tested three possible c/h^2 ratios: 1 (null assumptions), 0.5, and 2 (Silva and Silva 2018). A positive correlation between P_{st} and F_{st} over the line of neutrality ($P_{st}=F_{st}$) would indicate that floral color is being driven by divergent selection, while a negative correlation would indicate stabilizing selection or uniform selection. In addition to analyzing F_{st} and P_{st} as separate identity, we also took their difference ($P_{st} - F_{st}$) to run additional Mantel tests against geographic, and elevational distances. Because F_{st} calculated using SNPs could be reflecting the differences in historical random and/or demographic differences between the population pair. By subtracting F_{st} from the corresponding P_{st} between the same population pair, we accounted for the historical factors and focus on the phenotypic

differences that are above and beyond the historical factors. This will allow for a Mantel test targeting more specific on divergent phenotypes that are likely due to selection (Berardi et al., 2016).

Results

Pilot RADcap establish SNP loci

In total, 32 million reads for the pilot pool were sequenced. Samples had an average of 1,664,000 reads each, and, on average, 97.8% of these reads were retained after quality control (Table 4.1). After parameter selection and filtering to remove missing data, there were 1,814 polymorphic loci. For bait creation, we included non-polymorphic loci with the intent of discovering polymorphism in the regions not sampled for the pilot. Although 2,320 putative *G. maculatum* loci for bait creation were considered, only 2,081 passed screening which removed target redundancy and bait-to-bait hybridization and ensured loci were nuclear. A total of 4,153 baits were created for our *G. maculatum* samples to be used in the subsequent bait-capture step.

Sequencing

Capture of plate one failed, as most samples did not receive any reads. Plate two performed better as 93/100 samples had reads. This plate had an average of 211,985 reads per sample with an average of 93.4% reads retained. 88 of the 93 samples saw over 90% of reads retained. (See Table 4.4 for the samples included in analysis and Table 4.5 for population location) From the successful plate, those with the ID beginning in “3” underwent a three-hour digestion, those with an ID beginning with a “6” had a 40-minute digestion plus a 1x bead clean prior to PCR. 6/7 samples that dropped out and the 5

samples with retrained reads less than 90% were from the former. Post Stacks there were a total of 2,250 loci.

Genetic diversity and population structure

For polymorphic loci, observed vs. expected heterozygosity differed significantly between all six populations with a sample size of over five. Two populations; BW and M have an excess of heterozygotes while CJ EC, GS and HM have a deficit. The F_{is} values were all small, with some positive, indicating a small degree of inbreeding within populations, and some negative, indicating an excess of heterozygotes. The number of private alleles ranged from 49-294 (Table 4.6). Southern populations have the highest number of private alleles, but also have small sample sizes. Nucleotide diversity (π) ranges from 0.06-0.14. The largest values are found in the SE of the range (Table 4.6). Pairwise genetic differentiation ranged from 0.05 to 0.22 (Table 4.7). Pairwise phenotypic differentiation ranged from 0.0006 to 0.988 (Table 4.8a) for petal and 0 to 1 for pollen under $c/h^2=1$ (Table 4.8b).

STRUCTURE Harvester identified a large number of clusters ($k=8$), where membership sharing across samples do not agree with a steppingstone model of gene flow. Populations to the west of the Appalachians cluster separately (Figure 4.2). Additionally, populations that are geographically close do not necessarily share cluster membership, as would be expected under a steppingstone model of gene flow, yet far apart populations can share greater membership than those that are nearby (Figure 4.2). Mantel tests further confirm that there is a lack of isolation by distance ($r: 0.09$, $p: 0.32$) (Figure 4.3). Populations do show a pattern of moderate level of isolation by elevation ($r: 0.22$, $p: 0.06$). Elevation and geographic distance between populations are not correlated

in this study (r : -0.15, p : 0.88). In all three AMOVA (Table 4.9) models, a significant amount of variation (~28%~35%) is observed among populations, indicating significant population structure (p : 0.01). When we split our populations into east and west, similar to what was done in Chapter 3, we did not detect a significant region effect (5.52% p : 0.17). However, when the populations were split based on their elevations (above or below 900m), we found a significant regional effect, though the amount of variance explained was modest (9.25%, p <0.05).

Pairwise P_{st} and F_{st} for petal and pollen did not significantly correlate with one another, regardless of the c/h^2 ratio. (Petal: r : 0.13, p : 0.24. Pollen: r : 0.03, p : 0.53. $c=h$) (Figure 4.4a-b). Mantel tests between P_{st} - F_{st} and geographic distance also did not show a significant relationship for either petal or pollen, providing additional support for a lack of isolation by distance at the phenotypic level in this species (petal: r : 0.02, p : 0.40. pollen: r : -0.23, p : 0.81), at least among the populations included in this study. However, (P_{st} - F_{st}) for petal and pollen color both showed a strong relationship with elevation (petal, r =0.77, p <0.01, pollen r =0.8241, p : 0.02, indicating that populations diverge according to elevation.

Discussion

We present a small, but informative, study to assess population diversity and structure in *G. maculatum* using SNP loci. Trouble shooting with RADcap protocol proved to be difficult for this species. The DNA recovered from Qiagen DNeasy Plant Kit were able to be processed for library preparation without any additional tweaking of the library preparation protocol, unlike samples processed with CTAB. Our assessment is this is due to the quality of the DNA. Product extracted using CTAB needed undergo the

addition of an RNase step to provide DNA to remove degraded RNA contaminant. Digestion times needed to be adjusted from the original protocol from one hour to three hours. Although, some samples needed to be redone at 40 minutes of digestion. Once these changes were implemented, RADcap became a feasible approach for identifying SNPs in *G. maculatum* with ample data returned.

Our genetic diversity summary statistics indicate low inbreeding. While a negative F_{is} value indicates an excess of heterozygotes, and a positive value indicates a deficit, due to small sample size, the values of F_{is} being so close to zero, likely reflect $F_{is}=0$. Otherwise, an excess of heterozygotes can reflect population bottlenecks, overdominance (heterozygote advantage), or disassortative mating, while deficits can reflect inbreeding, a Wahlund effect, or underdominance (heterozygote disadvantage). F_{is} values reflect what has been already found in this species. While this species is self-compatible, it very rarely selfs in the wild and suffers from high inbreeding depression (Chang, 2007; Van Etten, et al., 2014). Overall, values of F_{is} values are very small, and there is not a clear pattern of floral color and F_{is} that could indicate a role of heterozygote advantage or disadvantage as a function of color. However, all populations with negative F_{is} , except one, are located in parks, which may experience bottlenecks due to ground maintenance, trail/building construction and human disruption. Nucleotide diversity shows populations in the south-east of the range have the highest genetic diversity, which may be related to potential glacial refugium (Petit et al., 2003; Walker et al., 2009). While private alleles can be used as an estimate of gene flow, many of our populations are small in sample size (Slatkin, 1985). The counts of private alleles reflect clustering in

the STRUCTURE. Populations with high counts of private alleles share less membership with other populations.

Our STRUCTURE analysis provide some interesting results. STRUCTURE indicates that west populations (5, 10) cluster separately from east populations. This may reflect the effects of recolonization after the last glacial maximum due to the Appalachian Mountains or Mississippi River (Soltis et al., 2006). The STRUCTURE plots do not indicate clustering based upon locality, as some populations are close geographically but have different membership coefficients. Our AMOVA infers that populations are connected with high gene flow, with some limited structure between high elevation (dark) and low elevation (light) populations. Although variation between regions may be stronger if we additionally partitioned by longitude, which saw the greater between population variance in Christopher (2017). The inference of high gene flow, but different phenotype between these two regions suggests diversifying selection. While there is no evidence for structure between east and west populations, this may likely be due to sampling only two populations on the western region. A lack of isolation by distance was found in our study. This does not agree with what has previously been found in this species. Christopher (2017) did find evidence of isolation by distance in this species using ten microsat markers. A model of isolation by distance would be expected in this species due to its limited dispersal abilities (Stamp and Lucas, 1983) and the pattern is ubiquitous. However, *G. maculatum* has the Appalachians within its range, which may act as a barrier to gene flow, or an environmental gradient for diversifying selection (Streisfeld and Kohn, 2005; Leimu and Fischer, 2008; Bresowar and Walker, 2011; Li et

al., 2019) Instead of isolation by distance, this species has a pattern of isolation by environment

Our key finding is that populations experience isolation by environment by elevation. This result is suggestive of populations experiencing limited gene flow across elevational gradients. Our results of P_{st} - F_{st} against elevation indicate a similar result, indicating that populations diverge against environment based on their population morphology. This agrees with our previous work (Chapter 3), suggesting that populations experience selection on petal and pollen color due selection from environmental factors (UV-B, temperature, and precipitation) that vary with elevation. However, F_{st} differentiation based on elevation reflect a barrier to gene flow, perhaps either by reproductive isolation such as pollinator selection, bloom time differences, or selection against immigrant morphs (Wang and Bradburd, 2014). Furthermore, the majority of populations pairs have $P_{st} > F_{st}$ which is indicative of divergent selection. Three population pairs (M vs GS, CJ vs BW, and EC vs BW) show $P_{st} < F_{st}$ which is indicative of uniform selection, which can be explained by these populations being similar morphologically and geographically close (see Chapter 3). However, using P_{st} requires knowledge on the relative values of between- and among-population variance that is due to additive genetic effects. It is, therefore, important to treat P_{st} estimates with caution (Brommer, 2011). Furthermore, small populations size, under sampling, and a lack of comparisons across the range may skew the results.

Our results represent a pilot study on population diversity and population structure in *G. maculatum* using SNPS. While more loci are typically considered more important than increasing the number of individuals (Languth et al 2012), increasing the

sampling size increases certainty in the results. While measurements of F_{st} by SNPS are being found to not need large sample size, 20 is often given as a good representation of populations (Willing et al., 2012) Given our results of isolation by distance being different than what was previous found, we suggest further sampling in this species to reevaluate the processes influencing population structure, despite SNPS being considered more powerful than mircosats (Brumfield et al., 2003; Väli et al., 2008; Zimmerman et al, 2020). The results we present offer exciting interpretations of how the environment interacts with this species should have further evaluation with a more complete set of data that spans the range of *G. maculatum* among various populations. What we offer is a good foundation to begin making hypothesis for which to test.

Table 4.1: RADcap pilot study sample details including population location, total recovered sequences and percentage of sequences retained.

Sample ID	Population Name	Latitude	Longitude	Total Sequences	% Retained
A1	HM	36.87253	-81.5106	1664732	97.40552834
B3	HM	36.87253	-81.5106	3556876	98.60900408
C4	BT	35.3579	-83.7185	4024386	96.55433152
D5	IS	35.31417	-83.8408	3367004	98.3849737
E6	BH	33.91756	-83.3872	2009432	98.85992659
F7	LG	35.68303	-83.646	3988878	96.54546968
G8	LG	35.68303	-83.646	3104080	98.42922863
H9	FTT	37.46735	-79.5813	2842108	98.96611951
I10	LHi	40.00655	-79.246	3201700	98.75844083
J11	LHi	40.00655	-79.246	3484182	98.66496641
K13	SK	38.92528	-79.6644	3638268	98.61576992
L14	BW	39.20256	-78.1265	3265416	98.5059484
M15	CD	34.75056	-84.5564	3233926	98.95161485
N19	BR	34.73979	-83.9244	3867050	96.90285877
O18	BT	35.3579	-83.7185	4232706	96.62379102
P16	CD	34.75056	-84.5564	3899590	98.13029062
Q21	HWY	35.11003	-83.2676	3193094	98.76264213
R22	MK	37.38022	-80.0897	3559216	97.70727598
S23	BR	34.73979	-83.9244	3206144	94.79873643
T24	FTT	37.46735	-79.5813	2544394	98.8185792
IT2	LG	35.68303	-83.646	2857488	97.18927954
JT12	HR	36.71408	-81.4712	2157348	97.17324233
KT17	CD	34.75056	-84.5564	3006216	97.72980385
LT20	SK	38.92528	-79.6644	3214680	97.72950962

Table 4.2: Plate 1 pooling guide. Samples were split into pool A and pool B due to concentration and volume needed to avoid pipetting less than 2µL, but not more than the total sample volume. Pools A and B were combined. Post pipette prep, dry down and rehydration the final pool concentration was 47.5ng/ µL

Pool	Sample Name	Conc (ng/µL)	Volume to Add (µL)
A	F12	2.199	20.32
	H6	2.5525	17.51
	B11	2.9185	15.31
	MB2	3.21	13.92
	MA1	3.31	13.5
	A1	3.76	11.88
	H3	3.821	11.7
	G9	4.6255	9.66
	D5	4.7925	9.32
	C3	4.7935	9.32
	MG1	4.82	9.27
	D3	5.75	7.77
	E3	5.765	7.75
	B3	6.215	7.19
	C8	6.55	6.82
	E2	6.7	6.67
	MF1	6.85	6.52
	MB3	6.85	6.52
	MC2	7.02	6.37
	MA3	7.07	6.32
	ME1	7.44	6.01
	MC1	7.47	5.98
	MD1	7.95	5.62
	G12	8.185	5.46
	MC3	8.23	5.43
	C4	8.71	5.13
	F1	9.6	4.65
	B4	9.86	4.53
	E4	10.75	4.16

	H12	11.165	4
	D10	11.475	3.89
	E12	12.41	3.6
	F3	12.96	3.45
	G3	13.035	3.43
	D9	13.295	3.36
	F8	13.365	3.34
	A10	13.81	3.24
	F2	14.115	3.17
	H5	14.245	3.14
	D12	14.935	2.99
	G1	15.195	2.94
	H9	15.435	2.9
	G8	15.445	2.89
	A4	15.64	2.86
	A9	16.14	2.77
	E8	16.275	2.75
	D4	16.895	2.65
B	A5	18.67	4.93
	F9	19.15	4.8
	C1	19.69	4.8
	F11	19.88	4.75
	F4	19.89	4.75
	D1	20.61	4.58
	A11	20.69	4.57
	H10	20.74	4.55
	B8	20.85	4.53
	A2	21.14	4.47
	F6	21.42	4.41
	D8	21.92	4.31
	F7	21.99	4.3
	A7	22.09	4.27
	F10	22.18	4.26
	C6	22.24	4.25
	B6	22.42	4.21
	C11	22.94	4.12
	H2	23.32	4.05
	C5	23.50	4.02
	G10	24.09	3.92
	G11	24.56	3.85

	G6	24.59	3.84
	E10	24.73	3.82
	B10	26.01	3.63
	E1	26.79	3.52
	D6	26.92	3.51
	C12	27.04	3.49
	B7	27.47	3.44
	E9	27.55	3.43
	E11	28.05	3.37
	C10	29.13	3.24
	D7	30.28	3.12
	A6	31.58	2.99
	G5	31.95	2.96
	E6	32.51	2.91
	C7	33.96	2.78
	D11	34.21	2.76
	G2	34.87	2.71
	E7	35.69	2.65
	H7	36.13	2.61
	G7	36.74	2.57
	F5	36.76	2.57
	A8	36.97	2.55
	E5	38.63	2.44
	B9	38.65	2.44
	H4	38.70	2.44
	C9	39.10	2.42
	B2	39.26	2.41
	G4	39.49	2.39
	B5	39.65	2.38
	D2	39.78	2.37
	C2	41.13	2.3

Table 4.3: Pooling guide for Plate 2

Pool	Sample ID	location	Conc (ng/μL)	Volume to Add
A	PM2	3-B1	9.915	2.88
A	PM1	3-C1	8.89	3.22
A	CJ8	3-D1	8.77	3.26
A	CJ7	3-E1	7.765	3.68
A	CJ29	3-F1	9.445	3.03
A	CJ22	3-G1	7.66	3.73
A	CJ1	3-H1	8.74	3.27
A	CJ17	3-A2	8.18	3.5
A	CJ13	3-B2	9.415	3.04
A	CJ18	3-D2	4.5675	6.26
A	CJ2	3-F2	7.85	3.64
A	CJ25	3-H2	8.845	3.23
A	EC37	3-F3	9.78	2.92
A	EC33	3-G3	7.495	3.82
A	EC10	3-H3	8.035	3.56
A	EC45	3-H4	8.655	3.3
A	EC31	3-A5	7.745	3.69
A	EC5	3-B5	6.085	4.7
A	GS21	3-C5	4.549	6.29
A	GS11	3-H5	4.535	6.31
A	GS2	3-G6	4.9875	5.73
A	GS27	3-H6	4.524	6.32
A	GS45	3-A7	5.075	5.64
A	3BW13	3-C7	11.305	2.53
A	3BW8	3-F7	8.665	3.3
A	3BW18	3-H7	5.105	5.6
A	3BW5	3-B8	8.87	3.22
A	3BW31	3-D8	10.69	2.68
A	3BW4	3-H8	10.27	2.78
A	HM11	3-C9	7.845	3.65
A	HM6	3-H9	10.85	2.64
A	HM41	3-A10	7.365	3.88
A	HM19	3-H10	10.52	2.72
A	HR8	3-A11	7.805	3.66
A	HR29	3-C11	10.3	2.78
A	HR10	3-D11	7.515	3.81

A	BF86	3-H11	7.315	3.91
A	M36	3-A12	11.14	2.57
A	M13-19	3-B12	11.375	2.51
A	M30	3-E12	6.96	4.11
A	M21-19	3-G12	8.83	3.24
A	SM5	6-A3	3.329	8.59
A	PM48	6-G3	8.465	3.38
A	BH19	6-B4	2.2115	12.93
A	BH10	6-C4	4.667	6.13
A	SM19	6-D5	4.091	6.99
A	SM15	6-E5	8.945	3.2
A	BH32	6-B6	37.8	0.76
A	BH37	6-C6	5.42	5.28
A	PM65	6-A7	7.02	4.07
B	CJ15	3-C2	16.935	2.53
B	CJ5	3-G2	15.115	2.84
B	CJ4	3-A3	12.925	3.32
B	CJ30	3-B3	19.81	2.17
B	CJ12	3-C3	11.955	3.59
B	EC7	3-D3	12.565	3.41
B	EC11	3-E3	18.405	2.33
B	EC23	3-A4	14.02	3.06
B	EC6	3-B4	19.62	2.19
B	EC38	3-C4	15.91	2.7
B	EC18	3-E4	16.265	2.64
B	EC9	3-F4	20.085	2.14
B	EC49	3-G4	16.84	2.55
B	GS1	3-E5	19.665	2.18
B	GS5	3-F5	18.69	2.3
B	GS18	3-G5	15.645	2.74
B	GS53	3-A6	16.775	2.56
B	GS6	3-B6	15.77	2.72
B	GS46	3-C6	15.705	2.73
B	GS20	3-D6	15.75	2.72
B	GS13	3-E6	27.53	1.56
B	GS55	3-F6	13.745	3.12
B	GS34	3-B7	15.445	2.78
B	3BW23	3-D7	19.4	2.21
B	3BW16	3-E7	14.18	3.03
B	3BW12	3-G7	11.73	3.66

B	3BW6	3-A8	15.71	2.73
B	3BW29	3-E8	18.265	2.35
B	3BW2	3-F8	21.345	2.01
B	HM15	3-A9	26.14	1.64
B	HM14	3-B9	21.715	1.98
B	HM24	3-D9	19.93	2.15
B	HM7	3-E9	16.43	2.61
B	HM47	3-F9	14.945	2.87
B	HM1	3-B10	15.24	2.81
B	HM23	3-C10	23.195	1.85
B	HM35	3-D10	14.025	3.06
B	HM12	3-E10	23.425	1.83
B	HM4	3-F10	12.7	3.38
B	HM16	3-G10	13.275	3.23
B	BF21	3-E11	19.28	2.23
B	BF54	3-F11	19.565	2.19
B	BF53	3-G11	18.92	2.27
B	M42-19	3-C12	13.255	3.24
B	M31	3-D12	13.8	3.11
B	PM78	6-H3	13.4	3.2
B	SC43	6-F5	16.92	2.54
B	SC12	6-G5	19.355	2.22
B	SC2	6-H5	13.71	3.13
B	BH24	6-D6	13.85	3.1

Table 4.4: Samples included for analysis including morphological measurements and sequencing reads retained. Pollen color is coded as a binary, 0=purple, 1=yellow. Petals range in color from 11-15 (pink-purple).

Sample ID	Population	Pollen Color	Petal Color	Total Sequences	% Retained
3BW12	BW	1	13	144416	95.1%
3BW13	BW	1	12	160332	96.2%
3BW16	BW	0	12	132695	94.5%
3BW18	BW	1	12	278549	91.3%
3BW2	BW	1	11	174689	88.8%
3BW23	BW	1	12	144807	94.3%
3BW29	BW	1	12	189288	95.1%
3BW31	BW	0	12	126627	91.6%
3BW4	BW	0	12	161541	90.6%
3BW5	BW	0	12	120041	94.2%
3BW6	BW	1	12	163971	94.3%
3BW8	BW	1	12	184804	90.0%
BF53	BF	1	12	169874	94.3%
BF54	BF	1	12	239566	93.8%
BF86	BF	1	13	75389	94.7%
BH10	BH	1	11	180455	95.5%
BH19	BH	1	12	621315	94.2%
BH24	BH	1	11	159612	95.2%
BH32	BH	1	11	81030	94.7%
BH37	BH	1	12.5	513687	94.8%
CJ1	CJ	1	12	281424	92.2%
CJ12	CJ	1	12	131516	94.3%
CJ15	CJ	1	12	138632	91.6%
CJ17	CJ	1	12	302250	92.0%
CJ18	CJ	0	13.5	95101	92.1%
CJ2	CJ	0	12	172081	89.9%
CJ22	CJ	1	12	88288	92.8%
CJ25	CJ	1	12	147414	93.1%
CJ29	CJ	1	12	118785	85.7%
CJ30	CJ	1	13	167182	90.4%
CJ5	CJ	1	12	203114	94.0%

CJ7	CJ	1	13	124155	94.3%
CJ8	CJ	1	13	167473	92.4%
EC10	EC	1	13	122315	93.7%
EC11	EC	1	12	145396	91.5%
EC18	EC	1	12	256279	92.4%
EC23	EC	1	12.5	187505	94.4%
EC31	EC	1	12	228485	93.6%
EC33	EC	1	11	84618	93.3%
EC37	EC	1	11	98652	90.4%
EC38	EC	1	11	292145	94.0%
EC45	EC	1	12	274507	91.7%
EC49	EC	1	12	177035	93.7%
EC5	EC	1	13	170675	91.5%
EC6	EC	1	13	211258	94.4%
EC7	EC	1	12	323853	93.2%
EC9	EC	1	12	133098	89.5%
GS1	GS	0	15	199728	93.6%
GS11	GS	0	15	213884	93.6%
GS13	GS	0	15	191206	93.4%
GS18	GS	0	15	205377	93.0%
GS2	GS	0	15	60221	94.0%
GS20	GS	0	13.5	229557	93.8%
GS21	GS	0	14	128544	94.6%
GS27	GS	0	15	177965	92.2%
GS34	GS	0	15	257560	93.5%
GS45	GS	0	14	148004	93.2%
GS46	GS	0	14	424324	94.3%
GS5	GS	0	15	336336	91.3%
GS53	GS	0	15	170640	91.8%
GS55	GS	0	14	344909	88.3%
GS6	GS	0	14	138637	93.1%
HM1	HM	0	15	132022	94.6%
HM14	HM	0	15	232904	94.9%
HM15	HM	1	15	76477	95.0%
HM16	HM	0	14	214441	94.3%
HM19	HM	1	15	107620	93.0%
HM23	HM	0	14	153501	95.1%
HM35	HM	0	15	189443	92.8%
HM4	HM	0	15	213228	91.8%

HM41	HM	0	14	193773	94.5%
HM47	HM	0	15	154366	92.4%
HM6	HM	0	15	99551	93.1%
HM7	HM	0	14	182930	94.0%
HR10	HR	0	15	120358	96.0%
HR29	HR	0	13	91365	93.8%
HR8	HR	0	13	158732	94.9%
M13-19	M	female	14	242624	92.9%
M21-19	M	female	14	195145	95.2%
M30	M	0	15	98210	94.1%
M31	M	1	14	249073	93.8%
M36	M	0	15	170572	93.6%
M42-19	M	female	15	466547	95.3%
PM1	PM	1	12	284609	93.7%
PM2	PM	1	14	178334	94.0%
PM48	PM	1	12	104493	94.8%
PM65	PM	1	13	719566	96.4%
PM78	PM	1	12.5	130377	95.6%
SC12	SC	1	12	258135	94.2%
SC2	SC	1	11	336767	96.1%
SC43	SC	female	11	448945	95.0%
SM15	SM	1	12	289215	95.8%
SM5	SM	1	12	1128433	95.4%

Table 4.5: Population metadata. Population number, name, and geographic location (latitude, longitude, and elevation). Average petal color and frequency of yellow pollen (in M and SC calculated without females in final count)

Population #	Population Name	Latitude	Longitude	Elevation (m)	Average Petal Color	Freq of Yellow Pollen
1	BW	39.20256	-78.1265	201	12	0.67
2	BF	42.39696	-76.5117	324	12.33	1
3	BH	33.91756	-83.3872	219	11.5	1
4	CJ	39.02525	-77.1566	91	12.35	0.85
5	EC	39.87415	-86.295	260	12.04	1
6	GS	35.58152	-83.2401	929	14.57	0
7	HM	36.87253	-81.5106	711	14.67	0.13
8	HR	36.71408	-81.4712	1072	13.67	0
9	M	35.02322	-83.4117	860	14.5	0.33
10	PM	41.64264	-83.4371	186	12.7	1
11	SC	35.71884	-78.7589	105	11.33	1
12	SM	34.71054	-85.3775	325	12	1

Table 4.6: Population summary statistics.

Population Name	# Samples	Observed Heterozygosity	Expected Heterozygosity	F_{is}	Private Alleles
BW	12	0.06234	0.05847	-0.0294	57
BF	3	0.08001	0.076	-0.0882	64
BH	5	0.10184	0.09909	-0.0503	294
CJ	13	0.06438	0.06522	0.0219	85
EC	14	0.06044	0.06674	0.08767	163
GS	15	0.06367	0.06634	0.0565	176
HM	12	0.06127	0.06245	0.0262	63
HR	3	0.07453	0.06435	-0.1733	49
M	6	0.08289	0.07637	-0.0705	113
PM	5	0.08283	0.08277	-0.0240	157
SC	3	0.11669	0.1119	-0.0864	275
SM	2	0.10109	0.09532	-0.1062	142

Table 4.7: Pairwise F_{st} between all populations

1/BW	2/BF	3/BH	4/CJ	5/EC	6/GS	7/HM	8/HR	9/M	10/PM	11/SC	12/SM
0	0.129904	0.10056	0.050424	0.077957	0.082405	0.103523	0.165136	0.137902	0.113345	0.108057	0.117549
		0.186115	0.085887	0.061528	0.099512	0.11214	0.218365	0.179157	0.137005	0.160306	0.191133
			0.086649	0.077075	0.102062	0.109849	0.193849	0.158107	0.151446	0.14608	0.172439
				0.060324	0.073599	0.086871	0.118496	0.099965	0.097418	0.091843	0.085193
					0.068081	0.075655	0.094763	0.087742	0.054423	0.0721	0.069325
						0.079585	0.099921	0.080546	0.103114	0.090241	0.082907
							0.084096	0.111924	0.110486	0.071459	0.09148
								0.17131	0.182784	0.178115	0.216555
									0.162037	0.152127	0.156032
										0.158561	0.158179
											0.159942

Table 4.8: P_{st} for petal (a) and pollen (b) when $c=h$

A.

1/BW	2/BF	3/BH	4/CJ	5/EC	6/GS	7/HM	9/M
0	0.393939	0.623269	0.604679	0.011871	0.988417	0.990155	0.932801
0.393939	0	0.594295	0.000643	0.192186	0.951362	0.962236	0.938144
0.623269	0.594295	0	0.784187	0.521928	0.980134	0.982723	0.969286
0.604679	0.000643	0.784187	0	0.450009	0.982123	0.98381	0.968922
0.011871	0.192186	0.521928	0.450009	0	0.983253	0.98369	0.967414
0.988417	0.951362	0.980134	0.982123	0.983253	0	0.105042	0.029595
0.990155	0.962236	0.982723	0.98381	0.98369	0.105042	0	0.175824
0.932801	0.938144	0.969286	0.968922	0.967414	0.029595	0.175824	0

B.

1/BW	2/BF	3/BH	4/CJ	5/EC	6/GS	7/HM
0	0.393939	0.524476	0.346556	0.762626	0.932826	0.792
0.393939	0	0	0.192661	0	1	0.866667
0.524476	0	0	0.28777	0	1	0.91687
0.346556	0.192661	0.28777	0	0.540958	0.974557	0.90795
0.762626	0	0	0.540958	0	1	0.969977
0.932826	1	1	0.974557	1	0	0.581395
0.792	0.866667	0.91687	0.90795	0.969977	0.581395	0

Table 4.9: AMOVA Results. Significance tested as described by Excoffier et al. (1992).

Regions of ‘side’ are broken up by west (Populations EC (5) and PM (10)) and east.

Regions of “elevation” are broken up by mountain (populations at or over 900m elevation: GS, HM, HR, and M (6-9)). Negative variation values should be treated as a zero. Significance was accessed using the ‘Randtest’ function in Poppr with 999 permutations.

	% Variation	p value
Between Population	34.88	0.01
Within Populations	62.11	0.01
Between Side	5.52	0.17
Between Populations within Side	31.47	0.01
Within Populations	63.01	0.01
Between Elevation	9.25	0.019
Between Populations within Elevation	28.44	0.01
Within Populations	62.31	0.01

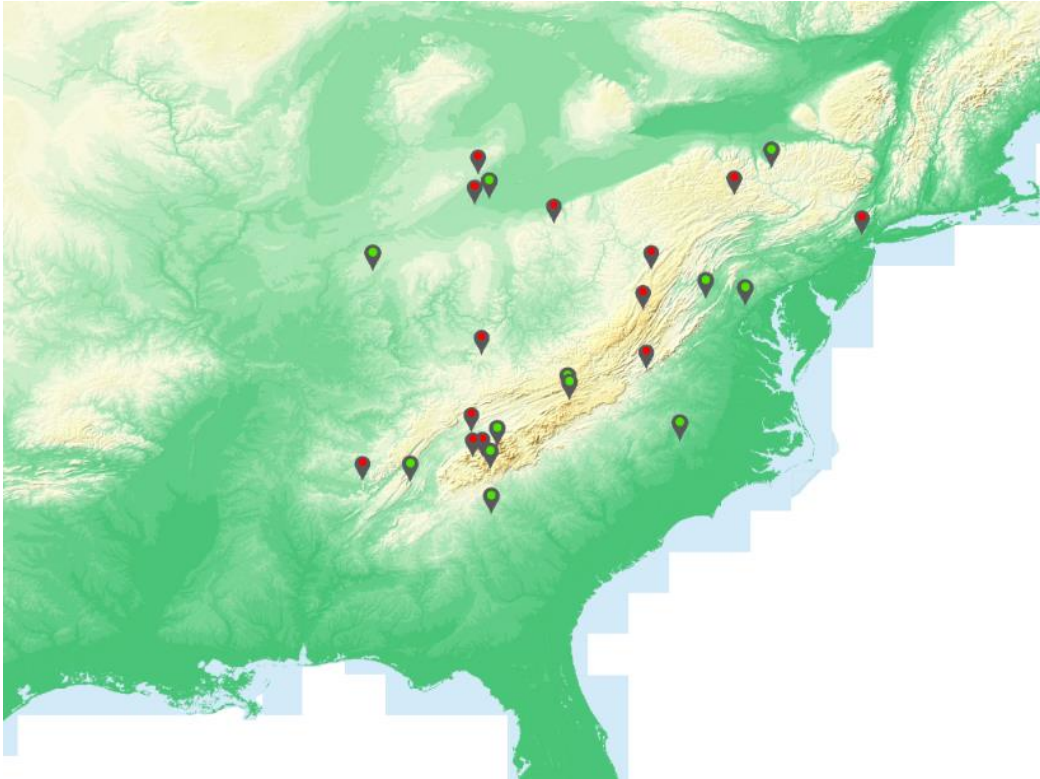


Figure 4.1: Map of populations for sequencing. Green pins represent populations included in this study.

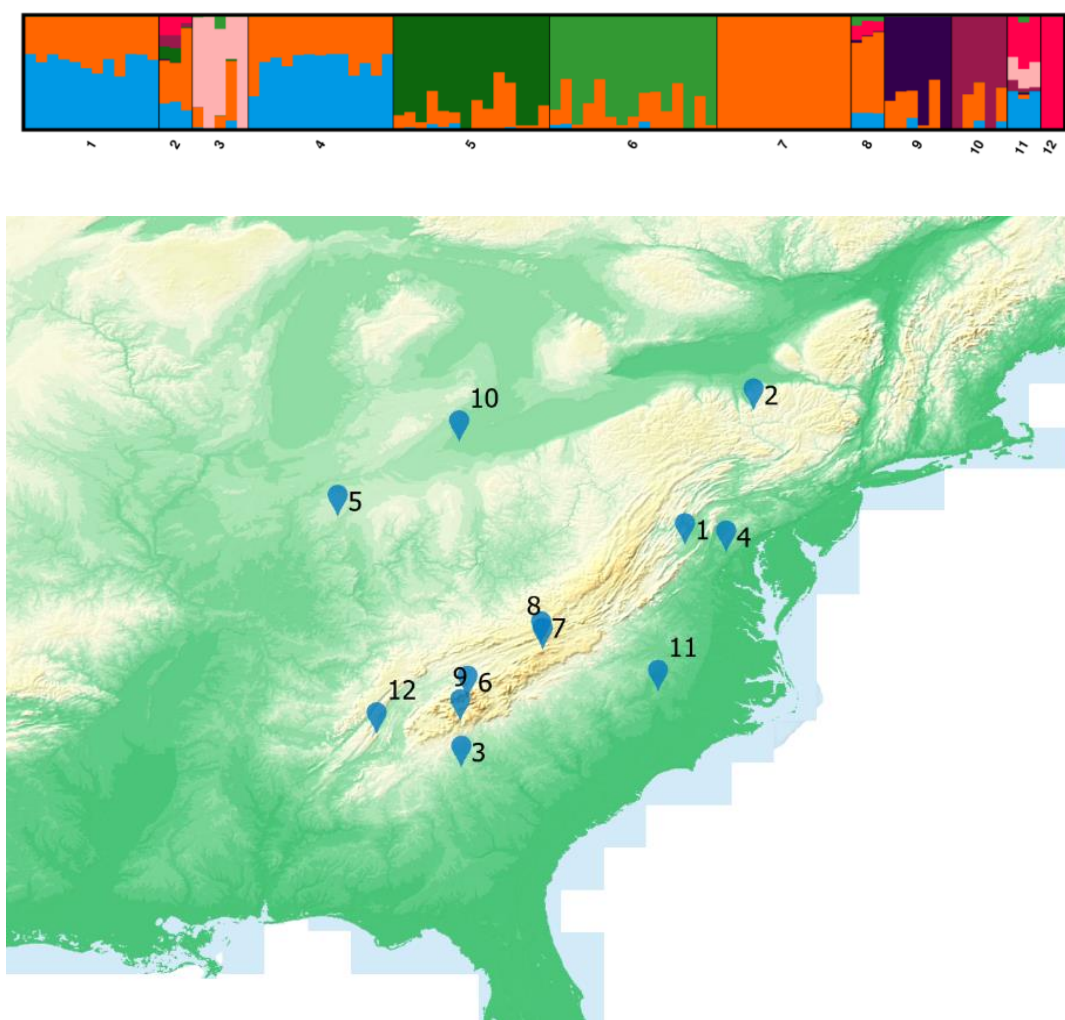


Figure 4.2: STRUCTURE plot of $K=8$. Colors indicate percent assignment to one of 8 clusters. Numbers show where each population is geographically located.

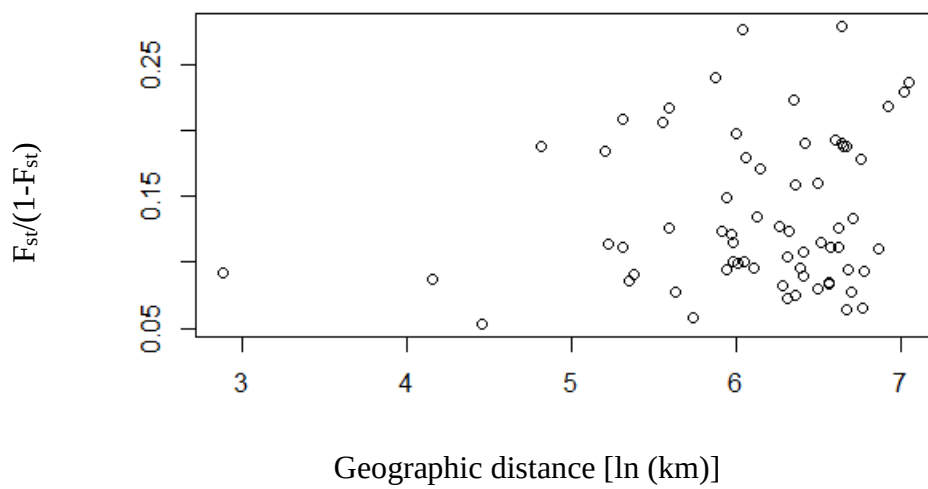
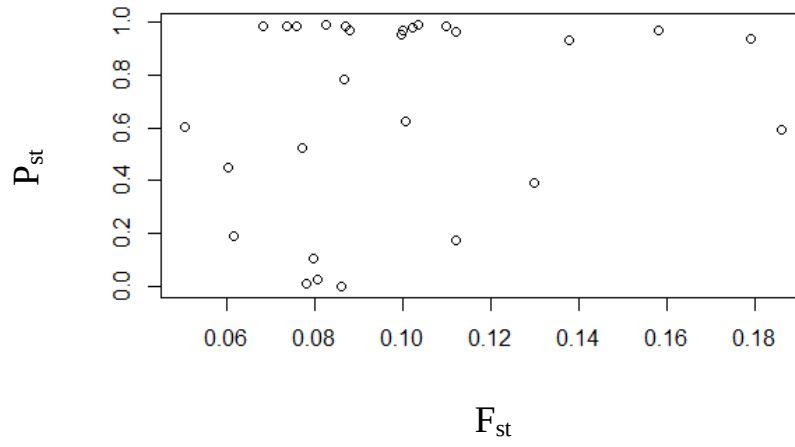


Figure 4.3: Plot of Isolation by Distance across all 12 populations. (r : 0.09, p : 0.32)

A.



B.

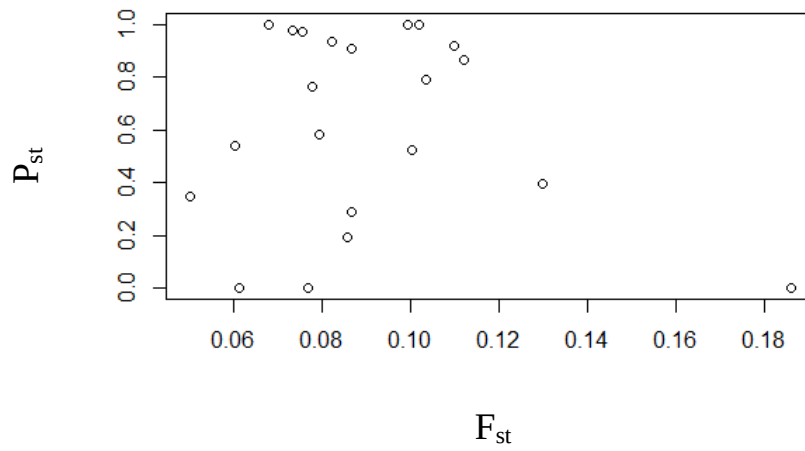


Figure 4.4: Plot of P_{st} vs F_{st} . No significant relationship was found between P_{st} and F_{st} for either population average (a) petal (r : 0.13, p : 0.24), or (b) pollen colors (r : 0.03, p : 0.53).

c/h^2 ratios was assumed to be 1 in the calculation of P_{st} .

References

- Arbor Biosciences. 2022. myBaits: Hybridization Capture for Targeted NGS User Manual. Ann Arbor: Arbor Biosciences
- Berardi, A., P. D. Fields, J. L. Abbate, and D. R. Taylor. 2016. Elevational divergence and clinal variation in floral color and leaf chemistry in *Silene vulgaris*. *Amer. Jour. of Botany*. 103 (8): 1508-1523
- Bresowar G.E., and G. L Walker. 2011. Patterns of genetic distribution of *Sibbaldiosis tridentata* (Rosaceae) in the Eastern United States. *Southern Appalachian Botanical Society*. 76(2):210-221
- Brommer, J.E. 2011. Whither P_{st} ? The approximation of Q_{st} by P_{st} in evolutionary and conservation biology. *J. of Evol. Biol.* 24(6):1160-1168
- Brumfield, R.T., P. Beerli, D. A. Nickerson, and S. V. Edwards. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends of Ecol. Evol.* 18(5):249-256
- Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22(11):3124-3140
- Chang S-M, 2007. Gender -specific inbreeding depression in a gynodioecious plant, *Geranium maculatum* (Geraniaceae). *American journal of Botany*. 94: 1193-1204
- Christopher, D.A 2017. Sex ratio variation and gene flow among populations of the gynodioecious *Geranium maculatum*. University of Georgia. PhD Dissertation.
- Coates, B. S., D. V. Sumerford, N. J. Miller, K. S. Kim. T. W. Sappington, B. D. Siegfried, L. C. Lewis. 2009. Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *J. of Heredity*. 100 (5): 556-564
- Cota-Sánchez, J. H., K. Remarchuk, and K. Ubayasena. 2006. Ready-to-Use DNA extracted with a CTAB method adapted for herbarium specimens and mucilaginous plant tissue. *Plant Mol. Biol. Reporter*. 24:161-167
- Earl, D.A, and B. M. vonHoldt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*. 4(2):359-361

- Ersts, P. J. Geographic distance matrix generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from https://biodiversityinformatics.amnh.org/open_source/gdmg/
- Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. 131(2):479-491
- Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14(8):2611-262
- Glenn, T. C., T. W. Pierson, N. J. Bayona-Vasquez, T. J. Kieran, S. L. Hoffberg, J. C. Thomas IV, D. E. Lefever, J. W., et al. 2019. Adapterama II: Universal amplicon sequencing on Illumina platforms (TaggiMatrix). *PeerJ* 7: e7786.
- Graham, C. F., T. C. Glenn, A. McArthur, D. R. Boreham, T. Kieran, S. Lance, R. G. Manzon, J. A. Martino, et al. 2015. Impacts of degraded DNA on Restriction enzyme Associated DNA Sequencing (RADSeq). *Molecular Ecology Resources*. 15:1304-1315.
- Hartl, D.L., and G. C. Clark. 1997. Principles of Population Genetics. Sinauer Associates, ... Sunderland.
- Hoffberg, S. L., T. J. Kieran, J. M. Catchen, A. Devault, B. C. Faircloth, R. Mauricio, and T. C. Glenn. 2016. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Mol. Ecol. Res.* 16(5): 1264-1278
- Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, I. Mayrose. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resources*. 15(5):1179-1191
- Leinonen, T., R. J. S. McCairns, R. B. O'Hara, and J. Merilä. 2013. Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nat. Rev. Genet.* 14(3):179-190
- Leimu, R., and M. Fischer 2008. 2008. A meta-analysis of local adaptation in plants. *PLOS ONE*. 3(12): e4010
- Lenguth, E. L., C. F. Bradley, S. J. Oyler-McCance, A. L. Garey, S. L. Emel, M. Mumma, H. H. Wagner, et al. 2012. Effect of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Mol. Ecol. Resources*. 12: 276-284
- Li, Y., K-M, Shih, C-T. Chang, J-D. Chung, and S-Y. Hwang. 2019. Testing the effect of

- mountain ranges as physical barrier to current gene flow and environmentally dependent adaptive divergence in *Cunninghamia konishii* (Cupressaceae). *Frontiers in Genetics*. 10:742
- Oksanen, J.F., G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R.B. O'Hara, et al. 2019. vegan: Community ecology package. R package version 2.5-6. <https://CRAN.R-project.org/package=vegan>
- Paris, J. R., J. R. Stevens, and J. M. Catchen. 2017. Lost in parameter space: a road map for STACKS. *Methods in Ecol. and Evol.* 10: 1360-1373
- Petit, J.R., I. Aguinagalde, J. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, et al. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*. 300(5625):1563-1565
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155(2): 945-959
- R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Reynolds, J., B. S. Weir, and C. C. Cockerham. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*. 105:767–779, 1983.
- Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*. 22, 939– 946.
- Sexton, J. P., S. B. Hangartner, A. A. Hoffmann. 2014. Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*. 68 (1): 1-15
- Seymour, M., K. Räsänen, and B. K. Kristjánsson. Drift versus selection as drivers of phenotypic divergence at small spatial scales: the case of Belgjarskógur threespine stickleback. *Evol. Evol.* 9(14):8133-8145.
- Silva SBD and Silva AD. 2018. Pstat: an R package to assess population differentiation in phenotypic traits. *The R Journal*. 10(1): 447-454
- Slatkin. M., 1985. Rare alleles as indicators of gene flow. *Society for the Study of Evolution*. 39(1):53-65
- Spitze K. 1993. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics*. 135: 367-374
- Stamp NE and Lucas JR. 1983. Ecological correlates of explosive seed dispersal. *Inter. Assoc. For Ecol.* 59 (2-3): 272- 278

- Streisfeld, M.A, ad J. R. Kohn. 2005. Contrasting patterns of floral and molecular variation across a cline in *Mimulus aurantiacus*. *Evolution*. 59 (12): 2548-2559
- Väli Ü, A. Einarsson, L. Waits, H. Ellegren. 2008. To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Mol Ecol*. 17:3808–3817.
- Van Etten., M. L. 2014. Mating system contributes only slightly to female maintenance in gynodioecious *Geranium maculatum* (Geraniaceae). 2014. *Heredity*. 113(5):464-470
- Walker, M. J., A. K. Stockman, P. E, Marek, and J. E. Bond. 2009. Pleistocene glacial refugia across the Appalachian Mountains and coastal plain in the millipede genus, *Narceus*: Evidence from population genetic, phylogeographic, and paleoclimatic data. *BMC Evolutionary Biology*. 9:25
- Wang, I. J., and G. S. Bradburd. 2014. Isolation by environment. *Molecular Ecology*. 23: 5649 - 5662
- Willing, E., C. Dreyer, and C. van Oosterhout. 2012. Estimates of genetic differentiation measured by F_{st} do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE*. 7(8): e42649
- Wright, S. 1943A. Isolation by distance. *Genetics*. 28 (2): 114-138
- Zimmerman, S. J., C. L. Aldridge, and S. J. Oyler-McCance. 2020. An empirical comparison of population genetic analyses using microsatellite and SNP data for species of conservation concern. *BMC Genomics*. 21:382

CHAPTER 5

CONCLUSION

Morphological traits that influence plant fitness often vary across a geographic region. Specifically, floral traits, most notably pigmentation have a long history of being studied for the effects of selection and/or genetic drift due to their significance in reproductive fitness (Waser and Price, 1985; Gigord et al, 2001; Jones and Reithel, 2001; Lau and Galloway, 2004; Eckhart et al., 2006) Mechanisms of evolution such as drift (movement of genes due to random chance) and natural selection (survival of the fittest) are often the driving force behind morphological variation in a species. Variation in floral color has shown to impact fitness through pollinator shifts (Yan et al., 2016) and protective anthocyanin changes (Berardi et al., 2016), making floral color an interesting and important trait to understand. I examined the herb *Geranium maculatum*, which has an interesting distribution in petal and pollen color recently documented by our lab.

In Chapter 2, we developed a method for classifying petal color in various species. We used python to separate pixels in photographs by color. Our method worked well for classifying *G. maculatum* photographs for petal color. And our accuracy with *Linanthus parryae* shows that this protocol has potential for other species. This sets the foundation for being able to process and utilize large-scale citizen science records for spatial and temporal studies in angiosperms.

In Chapter 3, we classified petal color from GPS associated photographs of *G. maculatum* from iNaturalist (inaturalist.org), an online citizen science project that

currently (summer 2022) has over 30k occurrence records for *G. maculatum*. We used our set of custom python scripts described in Chapter 2 to build a detailed distribution map of petal phenotypes. Additionally, we hand curated photographs to identify pollen color. This was paired with a traditional field survey approach to categorize populations by petal and pollen color. We found that dark (purple) pollen correlates with colder temperature, high precipitation, and high UV-B as well as a correlation with colder summers, higher UV-B and darker petal populations. However, we struggled to find correlations with individual level petal samples. This may indicate that selection within populations for petal color is not strong, sampling bias from citizen scientists that swamps out the true signal, a temporal effect, or skew from the unequal sampling of each color category. Overall, we do see clear patterns of floral variation among populations on a geographic scale.

In Chapter 4, we paired genetic data (F_{st}) from SNPS with our discovered phenotypic data from Chapter 3, which we used to calculate P_{st} for pollen and petal. We discover that populations show a pattern of isolation by environment (elevation) but find a lack of isolation by distance. P_{st} indicates a role of selection on floral pigmentation. Our Chapter 4 is a pilot study as we had many issues troubleshooting DNA extractions and 3RAD library preparation for *G. maculatum*. Therefore, the current data reflects only a portion of the range with less individuals sampled per population than originally desired.

Overall, this dissertation suggests that floral pigmentation in *G. maculatum* is due to selection caused by heterogeneous environment. Populations pigmentation appears to result in patterning due in part by elevation and the environmental space from precipitation, temperature and/or UV-B. With further examination, it is possible to test

further for support of selection from these abiotic agents underlying pigmentation in floral tissue in *G. maculatum*.

References

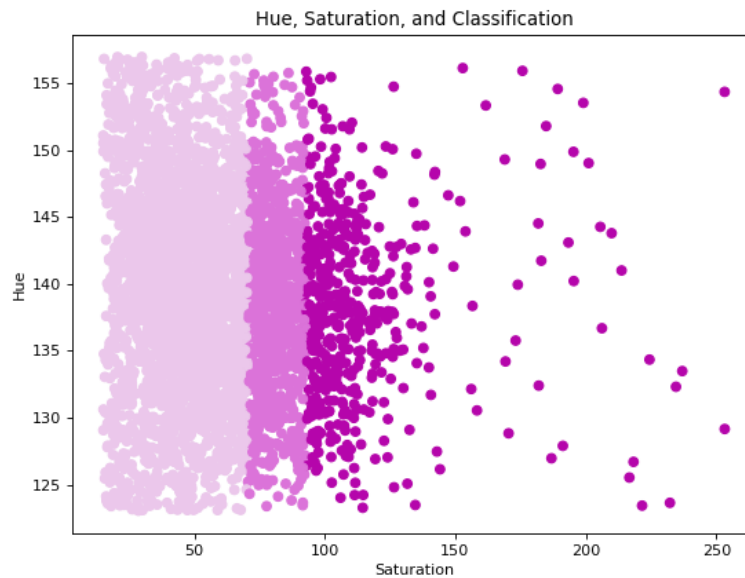
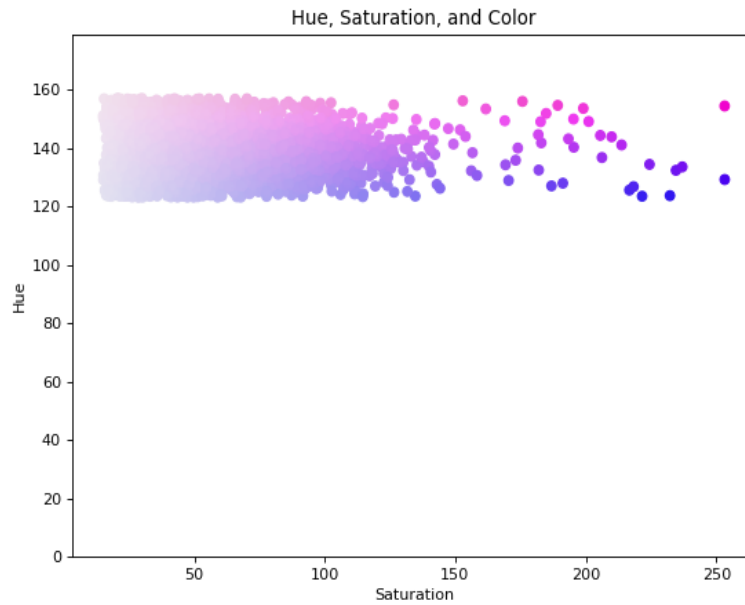
- Berardi, A., P. D. Fields, J. L. Abbate, and D. R. Taylor. 2016. Elevational divergence and clinal variation in floral color and leaf chemistry in *Silene vulgaris*. *Amer. Jour. of Botany*. 103 (8): 1508-1523
- Eckhart, V.M., N. S. Rushing, G. M. Hart, and J. D. Hansen. 2006. Frequency-dependent pollinator foraging in polymorphic *Clarkia xantiana* ssp. *xantiana* populations: implications for flower colour evolution and pollinator interactions. *Oikos*. 112(2):412-421
- Gigord, L. D., M. R. Macnair, and A. Smithson. 2001. Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid *Dactylorhiza sambucina* (L.) Soo. *PNAS*. 98(11):6253-6255
- Jones, K. N., and J. S. Reithel. 2001. Pollinator-mediated selection on a flower color polymorphism in experimental populations of *Antirrhinum* (Scrophulariaceae). *Am. J. of Botany*. 88(3): 447-454
- Lau, J.A., and L. F. Galloway. 2004. Effects of low-efficiency pollinators on plant fitness and floral trait evolution in *Campanula americana* (Campanulaceae). *Oecologia*. 141 (4): 577-583
- Waser, N. M., and M. V. Price. 1985. The effect of nectar guides on pollinator preference: experimental studies with a montane herb. *Oecologia*. 67:121-126
- Yan, J., G. Wang, Y. Sui, M. Wang and L. Zhang. 2016. Pollinator responses to floral colour change, nectar and scent promote reproductive fitness in *Quisqualis indica* (Combretaceae). *Scientific Reports*. 6: 24408

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTER 2

All scripts and raw data can be accessed from <https://github.com/atudell/Color-Cluster-Kit>

These include Supplemental S2.1-S2.8, S2.10, S2.12



Supplemental S2.9. Hue-Saturation of *G. maculatum* and resulting classification. **(Top)**

Each dot is colored in the HS (V was held stable at 94%) of the average HS value of the flower cluster. **(Bottom)** Resulting classification of light, medium or dark. Hue and

Saturation values are the converted values used in Python library Open CV due to

memory efficiency. To convert these values into true color space use $\text{Hue} \times 2$ and $\text{Saturation}/255$.

Appendix S2.11. Accuracy of flowering vs non-flowering in *G. maculatum*.

A random sample of ‘non-flowering’ classifications (500 samples) was taken and hand-scored for the presence or absence of flowers. The classification ‘non-flowering’ was correct 368/500 times. However, within the incorrectly scored instances, 67/132 were considered ‘not useful’ images of flowers. The ‘not useful’ designation was used for examples of images that had small, barely discernible flowers, overexposed photographs, partially closed flowers, among other issues explored in (Appendix). Therefore correctly identified images, as well as not useful, but incorrectly scored images, leaves us with an accuracy of removing bad photos for flower classification 87% of the time.

A random sample of ‘flowering’ classifications was taken (500 samples) and hand scored for the presence or absence of flowers. The classification of having flowers was correct 464/500 times (92.8%) Of the remaining 36 incorrectly classified, 30 were classified as being ‘light.’

Appendix S2.13. Photographing for floral structures for submission to iNaturalist and other citizen science repositories

Here we list good practices to remediate common issues we have identified in citizen scientist recorded photographs. If creating a project, we would suggest sharing this bulleted list, or creating your own for your project's interests, to share with volunteers.

- The specimen should be the object of focus.
 - Photographs should center on the flower/species of interest. Other objects can influence color segmenting.
 - We suggest taking a close up of the specimen, along with a 'body' shot.
 - Avoid taking photos of a person holding the specimen, including selfies.
 - Avoid body parts, extraneous objects such as trash, water bottles, bags, key fobs, etc.
- The specimen should be in the natural habitat.
 - Do not remove the specimen from its environment.
 - With the exception of deceased specimens that 'wash' ashore or have migrated to non-native environments, such as roadkill.
- Scales are often useful
 - Including a scale can be useful for researchers looking for measurements of traits.
 - Color scales can be a helpful addition, such as the color scales herbariums use

Scales with known color measurements can be useful for researchers to scale color properly.

- Document phenology, sex, and other descriptions on the plant
 - Phenology and sex can be selected in certain upload portals, such as iNaturalist, these traits allow for easy filtering of records
 - iNaturalist also offers sections for users to input notes about specimens.
- Ensure GPS coordinates are taken
 - When GPS coordinates are not easily established, such as when cell service is poor, take note of landmarks such as trail markers, parks, roads, etc to give a relative location.
- Do not edit photographs
 - Leave nature as nature is. Photo editing changes traits of interest and can skew data collection.

APPENDIX B

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Supplemental Table 3.1: Survey populations

Population Name	Latitude	Longitude	Population size	# flowers surveyed	Avg Pop. petal color	# androecium surveyed	Freq yellow pollen
McAfee Knob	37.38022	-80.0897	<50	14	12	-	-
Flat Top Trail	37.46735	-79.5813	1000+	54	12.69	47	0.63
Bear Creek	37.53227	-78.2745	<50	5	11.5	-	-
HWY 64	35.11003	-83.2676	<50	16	14.22	10	0.3
Macon	35.02322	-83.4117	100-300	83	14.40	23	0.22
Great Smokey	35.58152	-83.2401	300-500	80	14.71	40	0.03
Deep Creek	35.46479	-83.4337	100-300	34	14.26	-	-
Birding Trail	35.3579	-83.7185	1000+	63	14.75	-	-
Indian Store	35.31417	-83.8408	100-300	60	14.43	10	0.6
Little Greenbriar	35.68303	-83.646	1000+	62	13	53	0.68
Carl Dover	34.75056	-84.5564	50-100	21	12	15	1
Ellijay	34.77008	-84.59	100-300	71	12.47	53	1

Bear Hallow	33.91756	-83.3872	1000+	142	11.30	123	1
Bryon Reece	34.73979	-83.9244	<50	14	13.29	9	1
Homestead Road	36.71408	-81.4712	100-300	70	14.16	20	0.04
Mount Rogers	36.64565	-81.5834	100-300	25	14.74	-	-
Hungry Mother	36.87253	-81.5106	100-300	82	14.50	32	0.02
Lake Herrick	33.93015	-83.3699	100-300	73	11.30	73	1
National Forge	41.83957	-79.2738	100-300	13	13.23	10	1
R362	41.73037	-77.4253	50-100	31	12.71	5	1
Michaux State	39.8933	-77.4848	50-100	46	12.52	23	0.7
Laurel Hill	40.00655	-79.246	50-100	30	12.47	13	1
Bear Run	39.90364	-79.462	100-300	73	12.49	47	1
Spruce Knob	38.92528	-79.6644	500-1000	122	14.72	73	0.42
Miami Whitewater	39.42935	-84.747	50-100	24	11.92	13	1

Mortum Arboretum	41.82137	-88.0771	100-300	20	11.95	17	1
3rd Battle of Winchester	39.20256	-78.1265	50-100	43	12.14	37	0.74
Camp Creek	37.50239	-81.1369	50-100	16	12.88	11	0.5
French Creek	40.21192	-75.7821	<50	9	12.06	6	1
Orange Trail	33.90284	-83.3807	100-300	17	11.65	13	1
Wildflower Trail	34.73838	-86.549	<50	20	11.65	17	1
Swift Creek	35.71884	-78.7589	100-300	43	11.81	38	1
Shirley Miller	34.71054	-85.3775	100-300	29	12	18	1
William Hastie	35.93285	-83.877	100-300	12	11.67	9	1
Red River	37.82542	-83.6254	100-300	15	13.07	10	1
NOC	35.3318	-83.6033	300-500	40	14.95	30	0.03
South Mountain	40.75779	-74.2938	<50	22	11.95	21	0.95
West Rock	41.36179	-72.9735	<50	21	12.12	16	0.81

Buttermilk Falls	42.39696	-76.5117	300-500	100	12.27	38	1
Hubbard Valley	41.04251	-81.85130	100-300	65	11.99	28	1
Mary Beth Doyle	42.2423	-83.7094	300-500	53	11.99	29	1
Pearson Metropark	41.64264	-83.4371	1000+	100	12.35	96	1
Blue Creek	41.50357	-83.7984	<50	9	13	5	1
Eagle Creek	39.87415	-86.295	100-300	52	12.09	26	1
Cabin John	39.02525	-77.1566	50-100	33	12.17	27	0.82

Supplemental Table 3.2: Climatic variables. List of climatic variables, shorthand reference name and measurement units.

Variable	Short hand	Unit
Annual Mean Temperature	Bio1	°C
Mean Diurnal Range	Bio2	°C
Isothermality	Bio3	%
Temperature Seasonality	Bio4	%
Max Temperature of Warmest Month	Bio5	°C
Min Temperature of Coldest Month	Bio6	°C
Temperature Annual Range	Bio7	°C
Mean Temperature of Wettest Quarter	Bio8	°C
Mean Temperature of Driest Quarter	Bio9	°C
Mean Temperature of Warmest Quarter	Bio10	°C
Mean Temperature of Coldest Quarter	Bio11	°C
Annual Precipitation	Bio12	mm
Precipitation of Wettest Month	Bio13	mm
Precipitation of Driest Month	Bio14	mm
Precipitation Seasonality	Bio15	%
Precipitation of Wettest Quarter	Bio16	mm
Precipitation of Driest Quarter	Bio17	mm
Precipitation of Warmest Quarter	Bio18	mm
Precipitation of Coldest Quarter	Bio19	mm
Annual Mean UV-B	UV1	J/m ² /day
UV-B Seasonality	UV2	J/m ² /day

Mean UV-B of Highest Month	UV3	J/m ² /day
Mean UVB of Lowest Month	UV4	J/m ² /day
Sum of Monthly Mean UV-B During Highest Quarter	UV5	J/m ² /day
Sum of Monthly Mean UV-B During Lowest Quarter	UV6	J/m ² /day

Parameter	Value
n_estimators	100
min_samples_split	2
min_samples_leaf	1
max_features	auto
max_depth	None
bootstrap	True

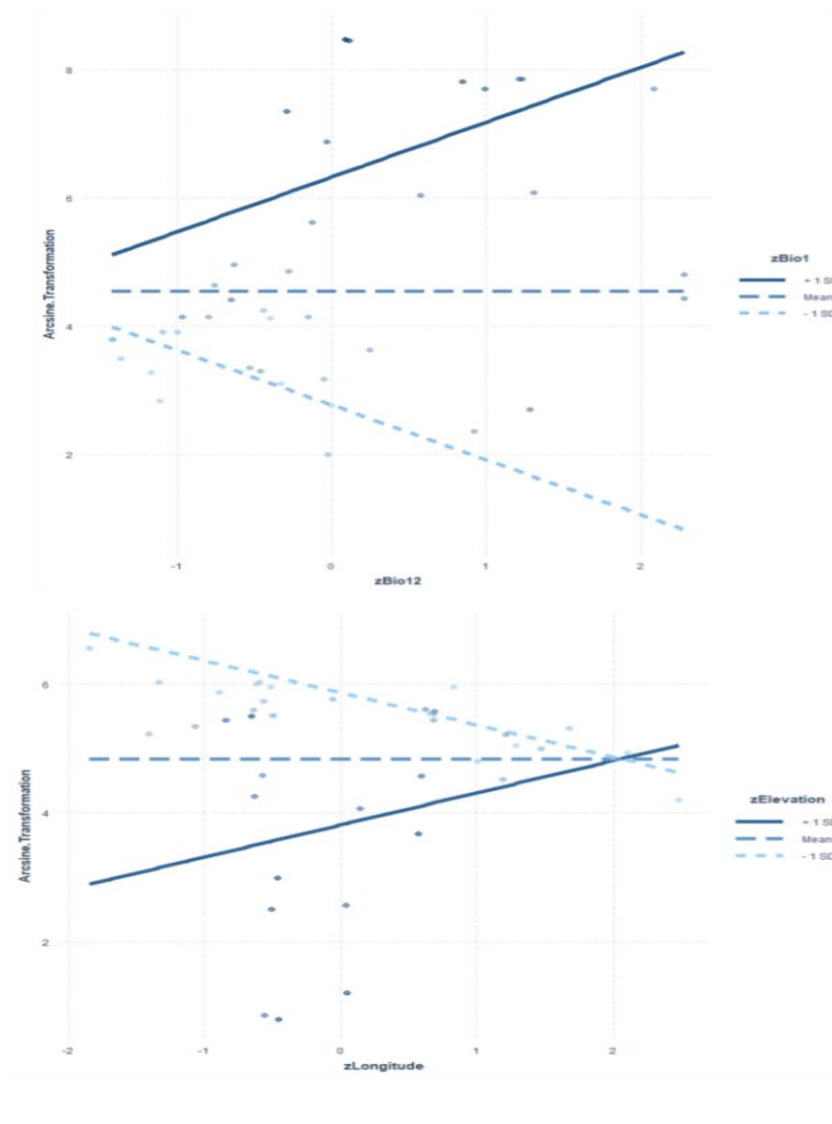
Supplemental Table 3.3: Random Forest default parameters. Initial parameters used.

Parameter	Value
n_estimators	1200
min_samples_split	5
min_samples_leaf	2
max_features	sqrt
max_depth	20
bootstrap	True

Supplemental Table 3.4: Random Forest classifier final parameters. Parameters chosen after hypertuning.

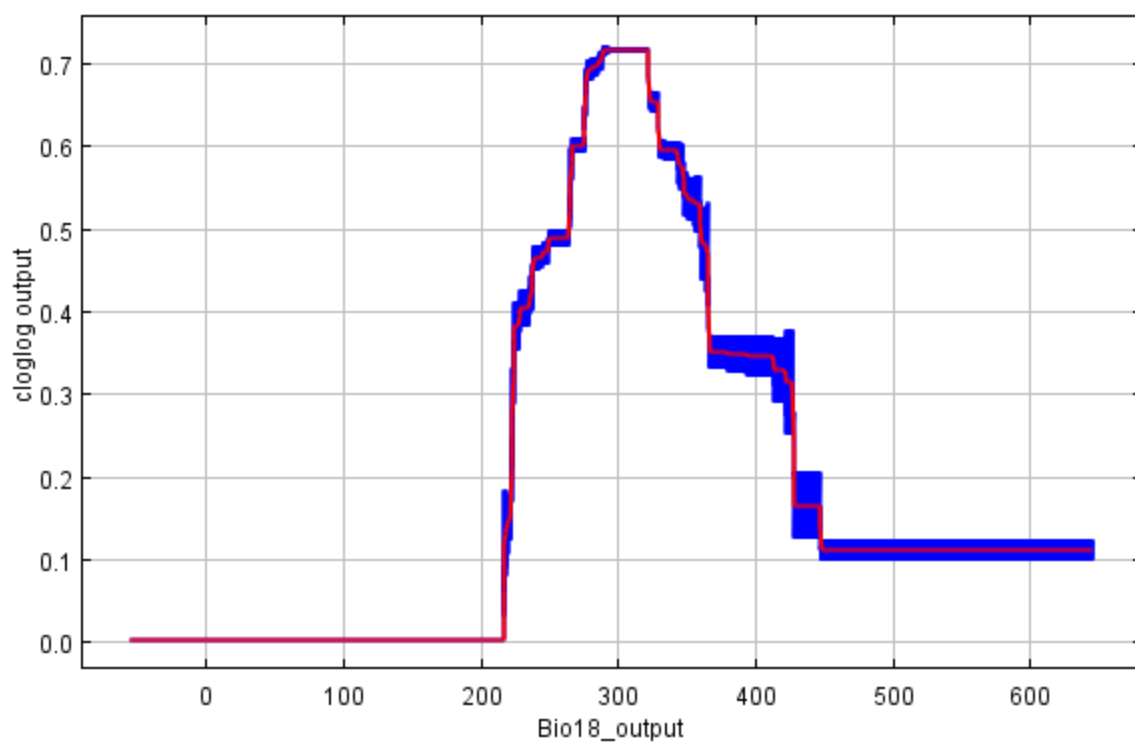
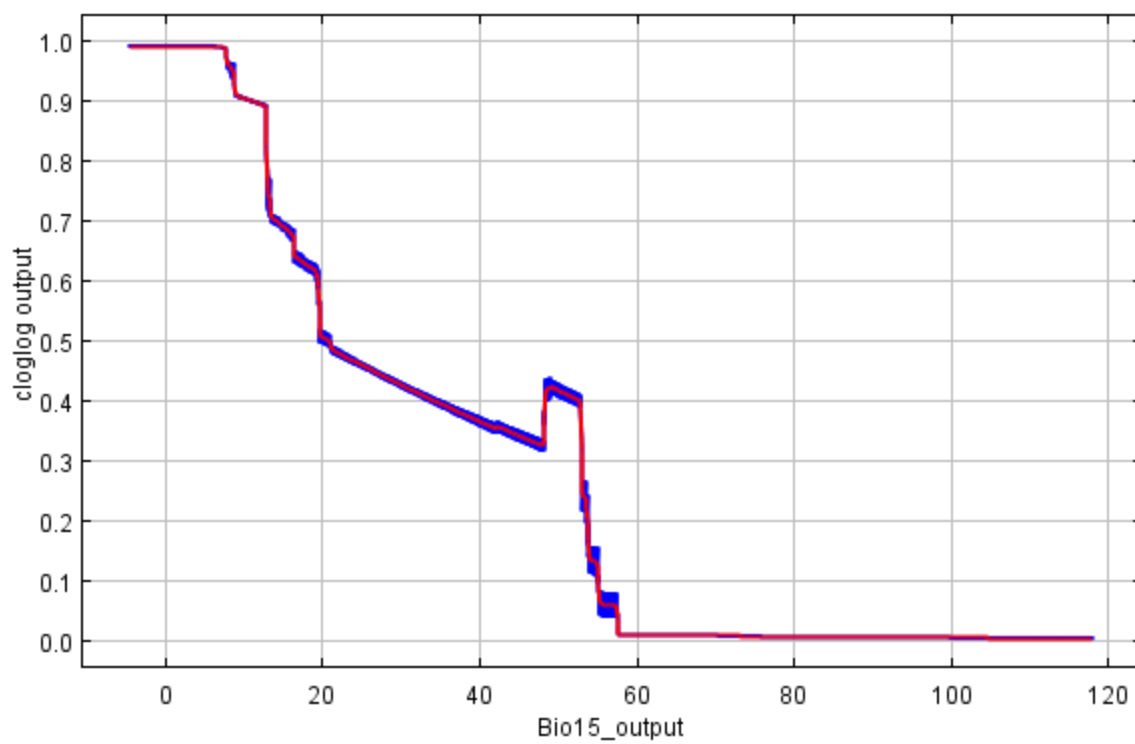
Supplemental Table 3.5: PCA loadings. Loadings for PC1 and PC2 for petal survey, pollen survey and pollen occurrence data.

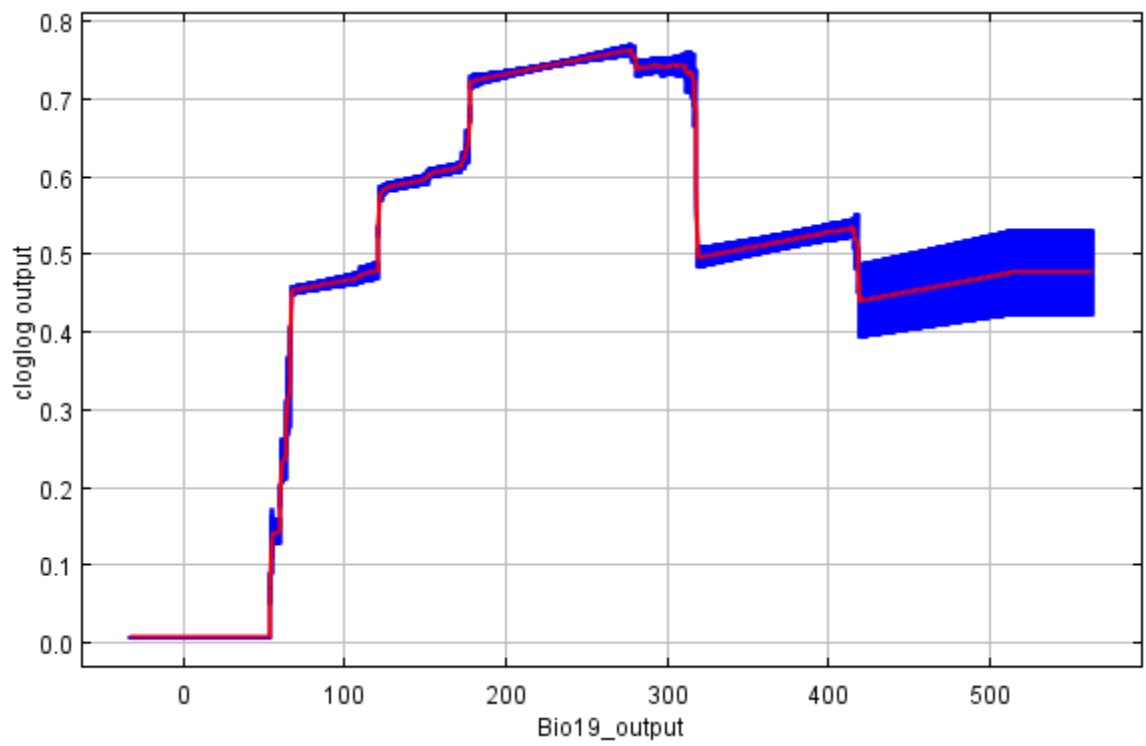
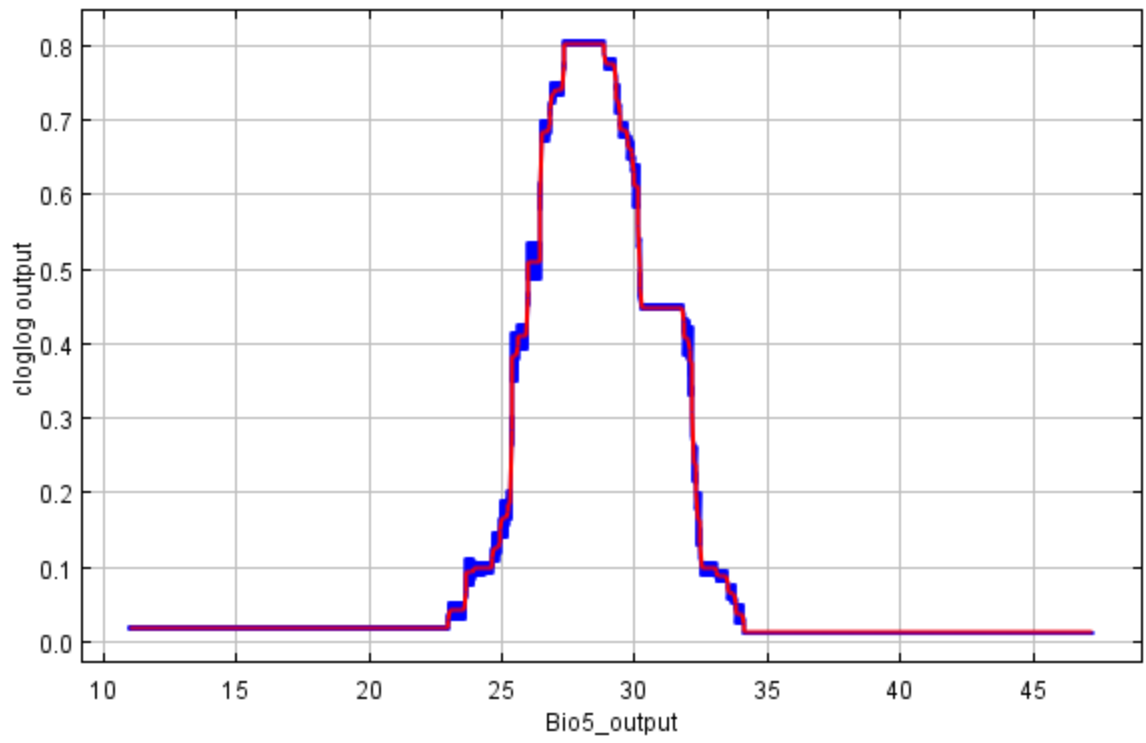
	Petal Survey Loadings		Pollen Survey Loadings		Pollen iNaturalist Loadings	
Variable	PC1	PC2	PC1	PC2	PC1	PC2
Bio1	-0.17906	-0.29985	0.18699	-0.2780544	0.231213	-0.13654714
Bio2	-0.17558	-0.09558	0.187272	-0.0504586	0.131765	-0.177158837
Bio3	-0.22477	0.006362	0.224691	0.0290516	0.23139	-0.006237664
Bio4	0.225866	-0.10404	-0.22485	-0.0979358	-0.2351	-0.150124988
Bio5	-0.07156	-0.41171	0.089643	-0.4024958	0.125369	-0.329554108
Bio6	-0.2127	-0.19022	0.214496	-0.1755007	0.235729	0.041655847
Bio7	0.201407	-0.15668	-0.20898	-0.1312992	-0.20871	-0.202479517
Bio8	0.188342	-0.13419	-0.18481	-0.1328861	-0.10348	-0.228060484
Bio9	-0.23018	-0.0877	0.229116	-0.081367	0.217601	0.053831556
Bio10	-0.0829	-0.40873	0.099374	-0.3956813	0.14095	-0.307244767
Bio11	-0.21977	-0.17976	0.220852	-0.1630914	0.247022	-0.00474553
Bio12	-0.2153	0.185872	0.212375	0.1899792	0.232918	0.146644491
Bio13	-0.2179	0.130921	0.216555	0.1319081	0.190895	-0.047165211
Bio14	-0.21124	0.20349	0.207932	0.2119224	0.21781	0.250128551
Bio15	0.115019	-0.22679	-0.11334	-0.2461315	-0.1719	-0.288391778
Bio16	-0.21545	0.152215	0.214449	0.153083	0.184749	-0.047134257
Bio17	-0.20931	0.217033	0.205285	0.2264464	0.221159	0.2370773
Bio18	-0.18102	0.24483	0.17752	0.2494715	0.115564	-0.058010704
Bio19	-0.23098	0.11381	0.228002	0.1198323	0.232535	0.196386028
UV1	-0.23478	-0.08556	0.232256	-0.0964302	0.236518	-0.168707358
UV2	-0.07682	-0.29775	0.065749	-0.3362459	0.003042	-0.376530906
UV3	-0.21975	-0.16811	0.217084	-0.1766918	0.165528	-0.333503042
UV4	-0.23696	-0.05606	0.235419	-0.0606686	0.246025	-0.079536027
UV5	-0.22655	-0.1324	0.223613	-0.1465837	0.210013	-0.240501531
UV6	-0.23706	-0.04657	0.235327	-0.0563936	0.24584	-0.106163221

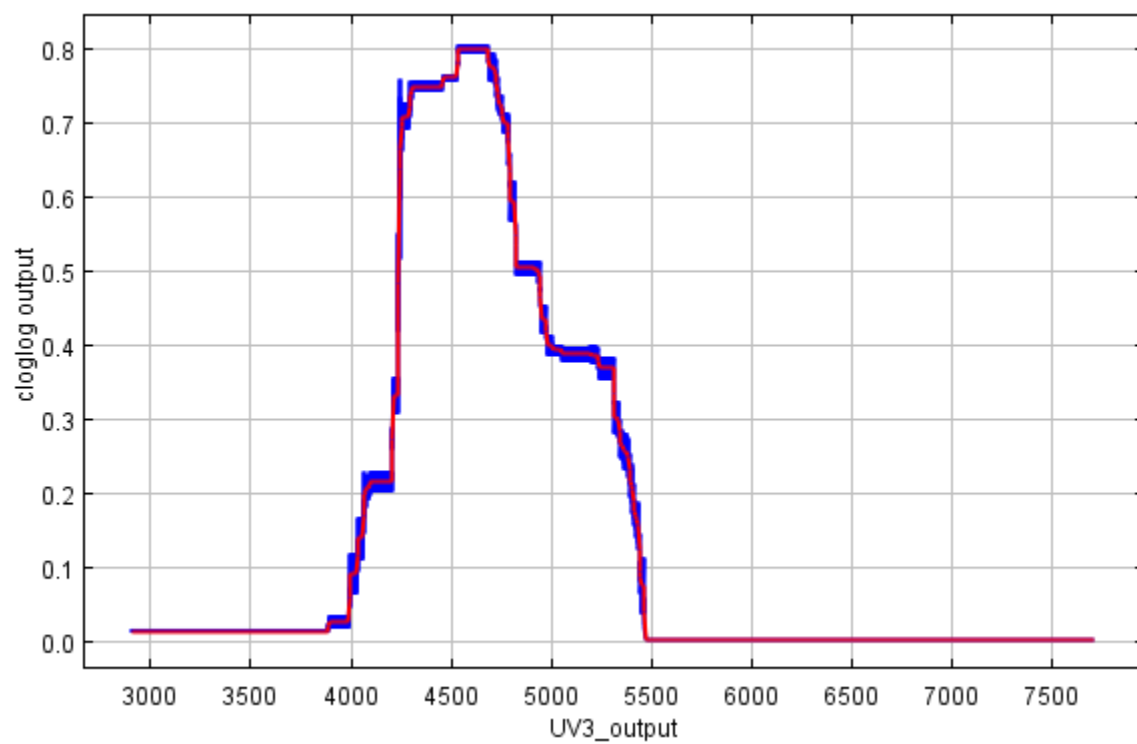


Supplemental Figure 3.1: Partial residual plot. Interaction of Bio 1 and Bio 12 (top) and Elevation and Longitude (bottom) from the survey based linear regression. Bio1 mean is 11.61817 with a SD of 2.405815. Bio 12 mean is 1230.027 with a SD of 261.7918.

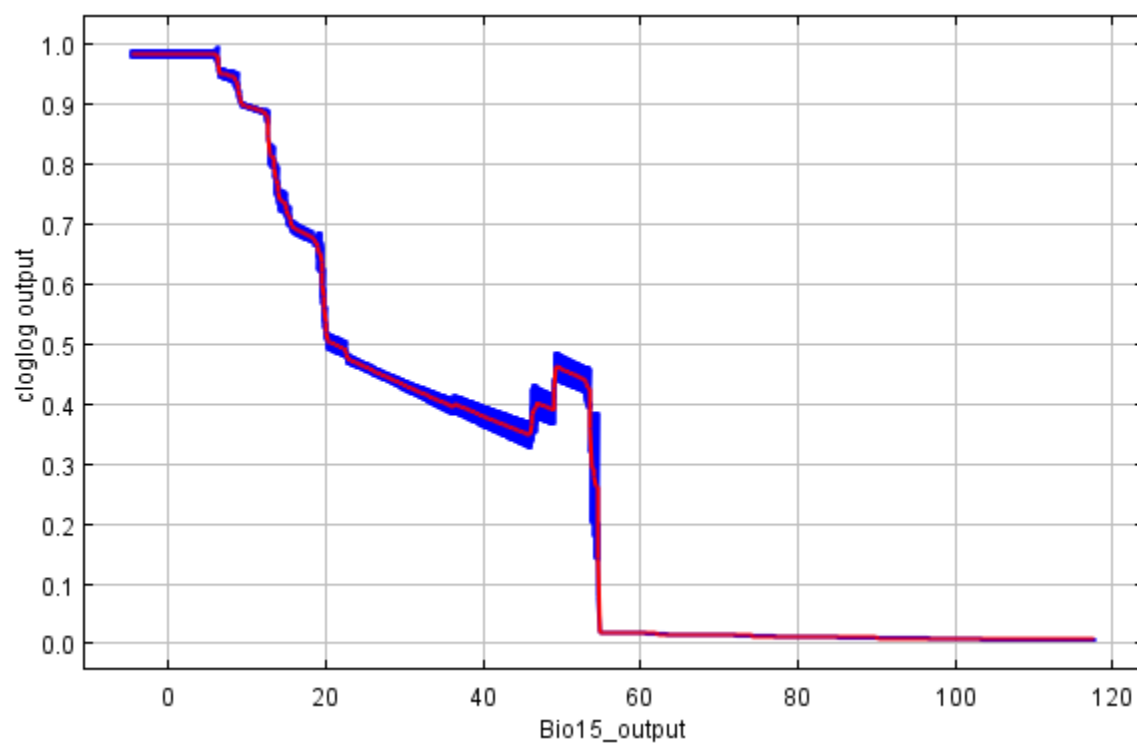
A.

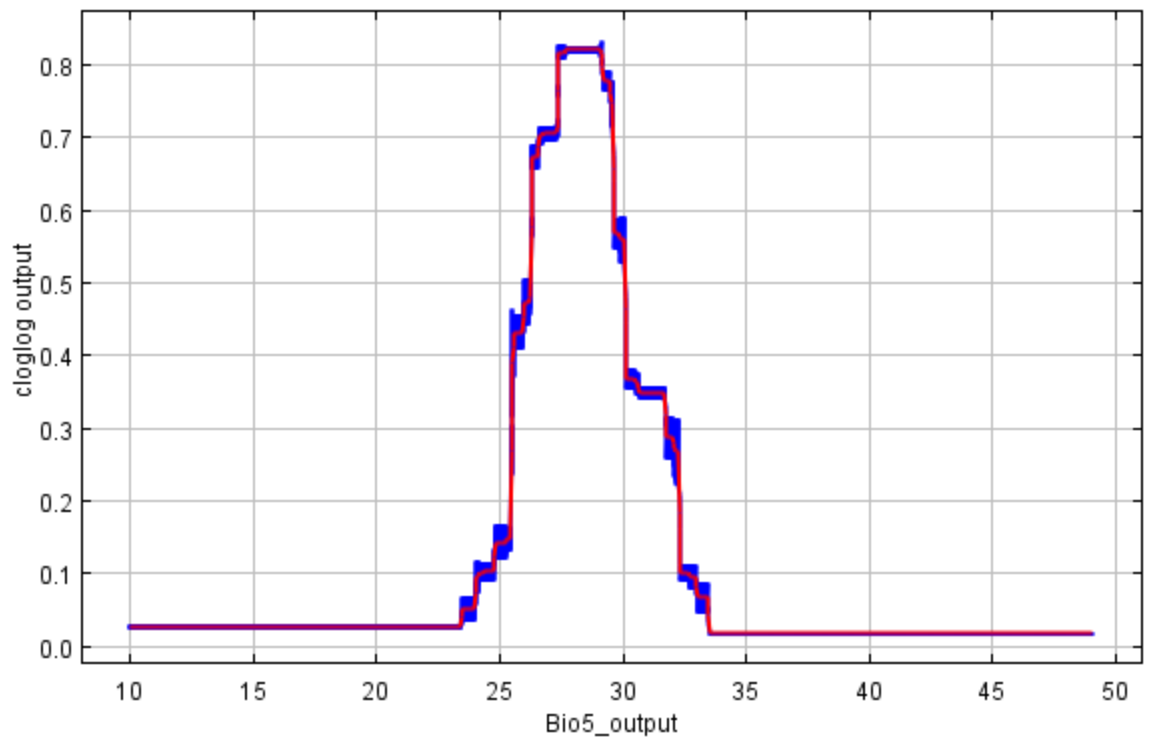
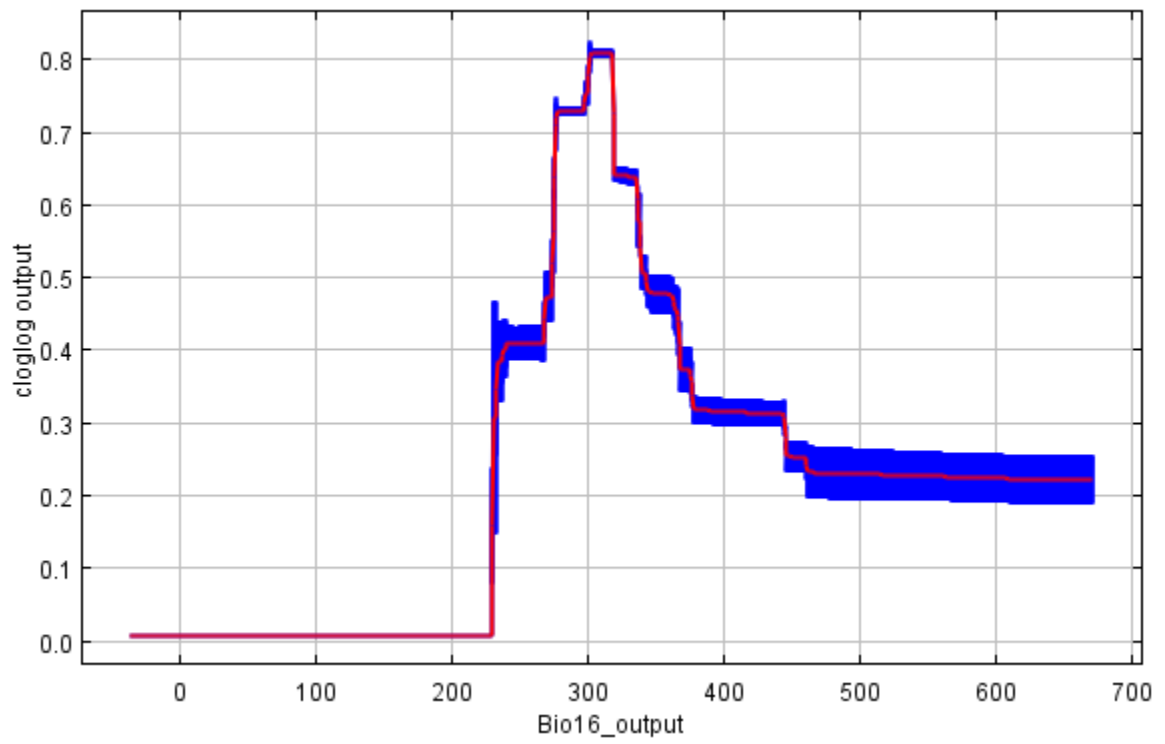


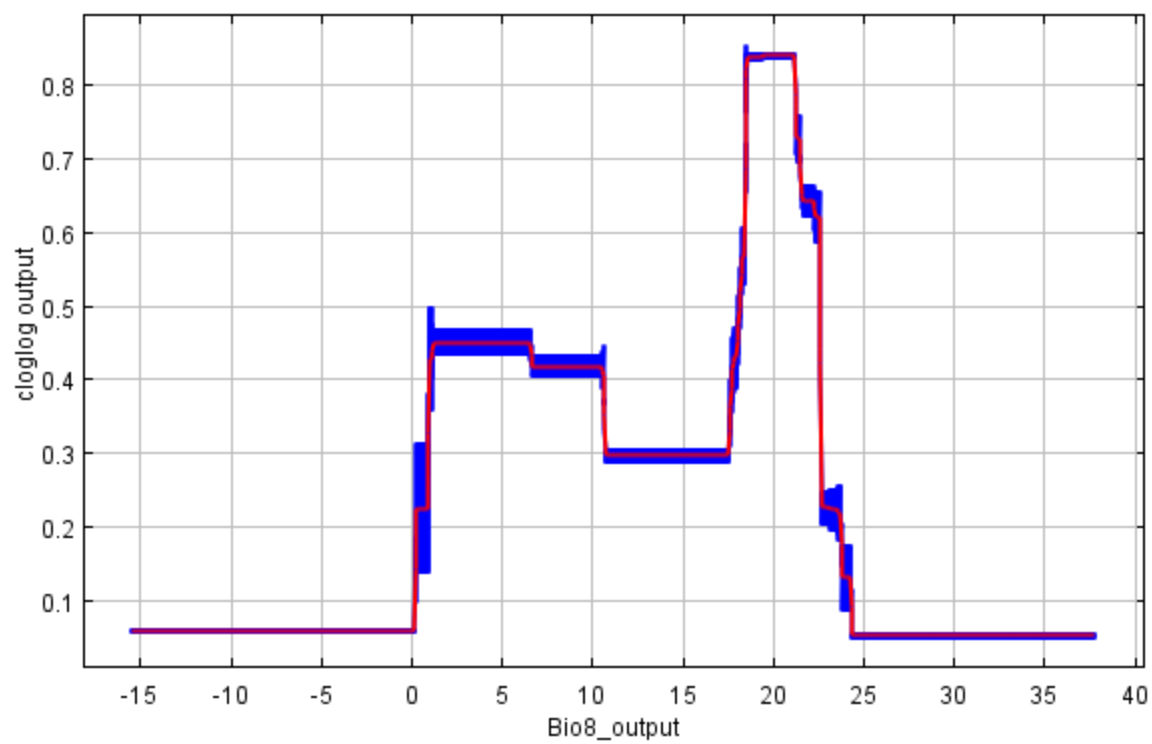
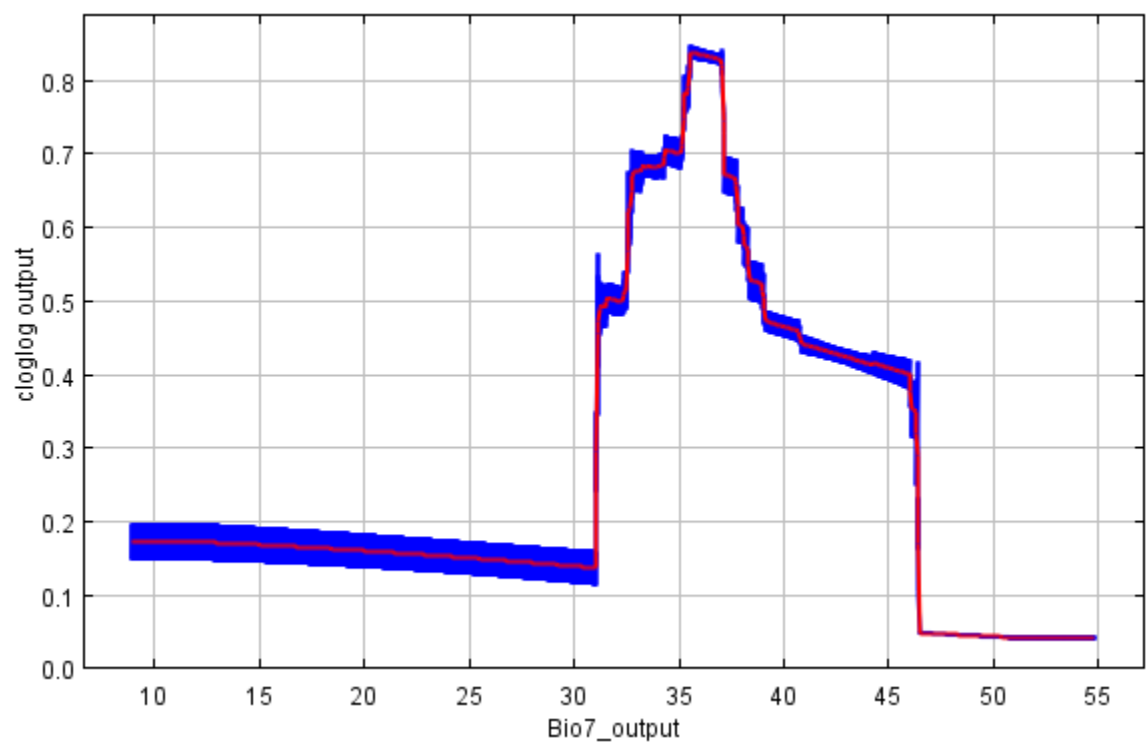


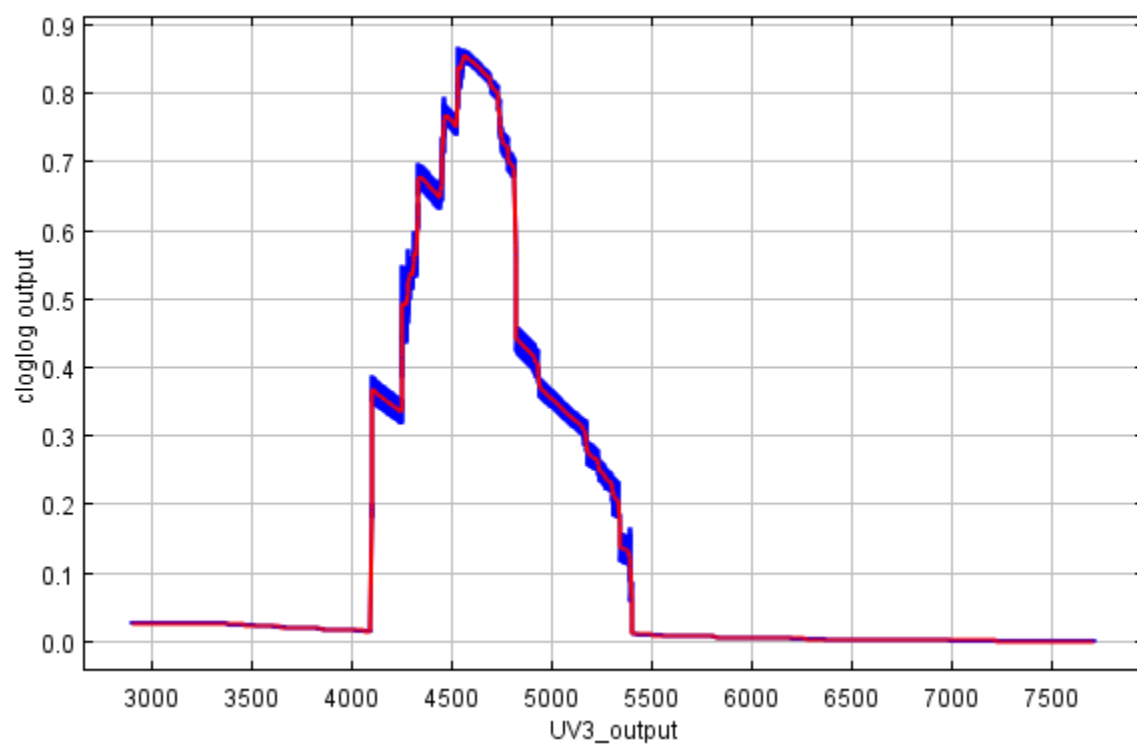


B.

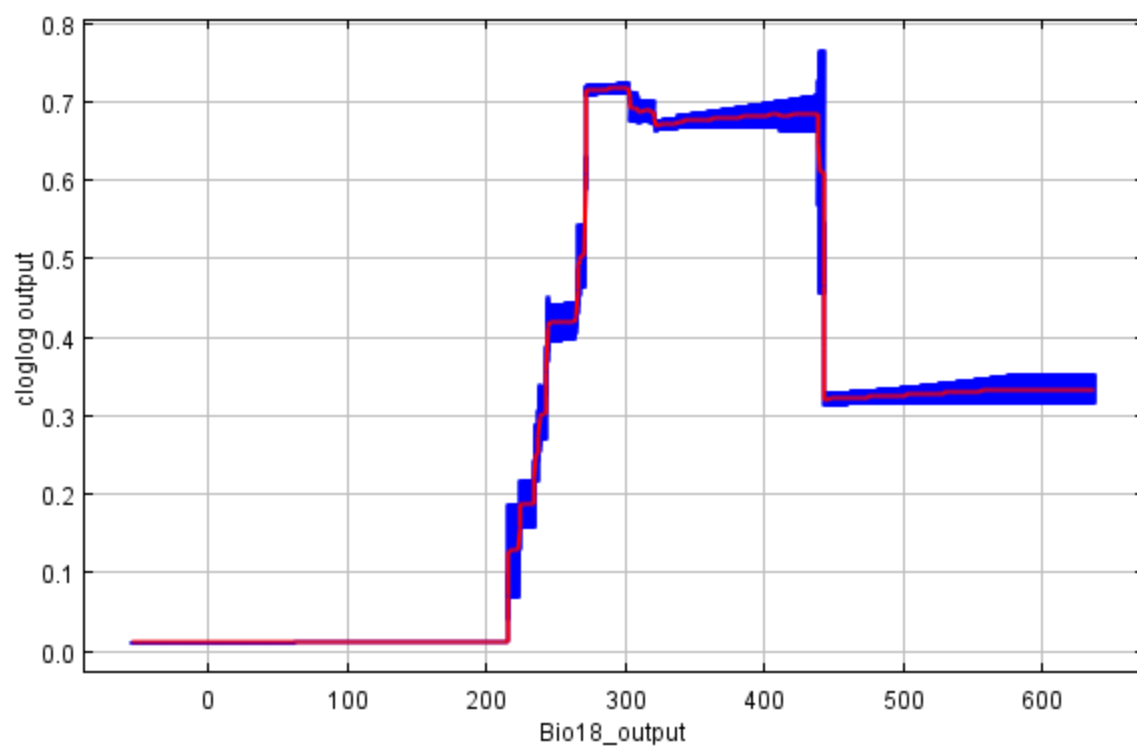


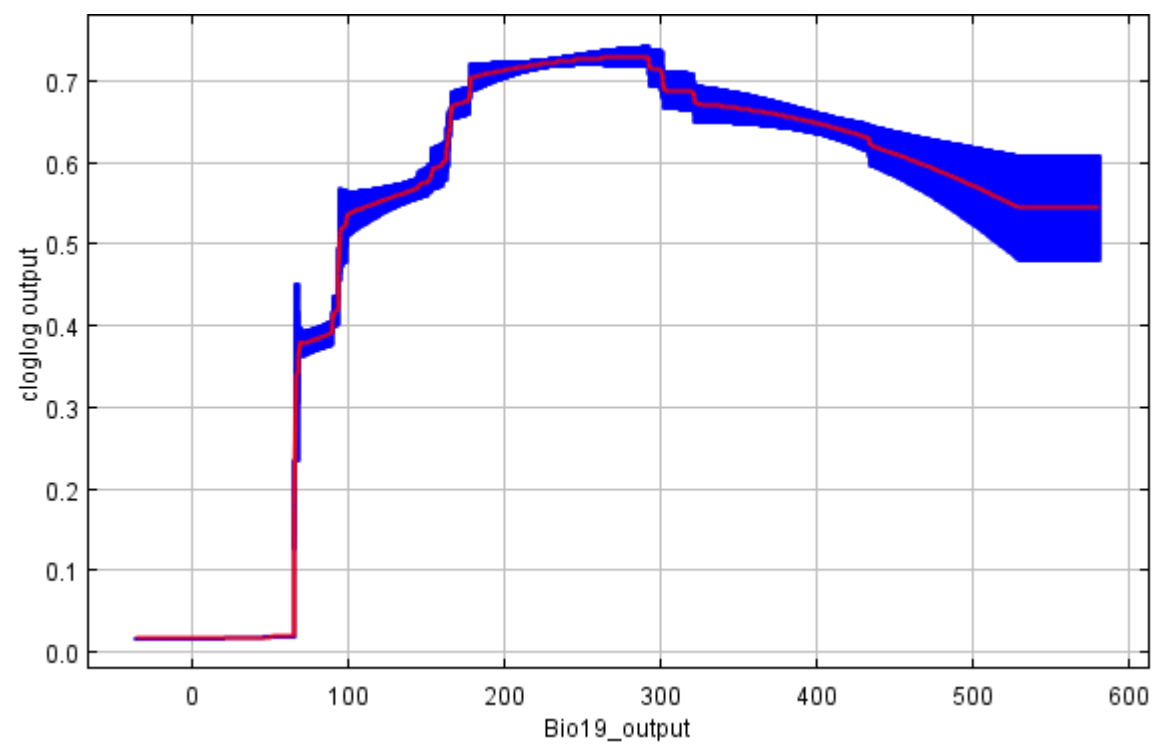
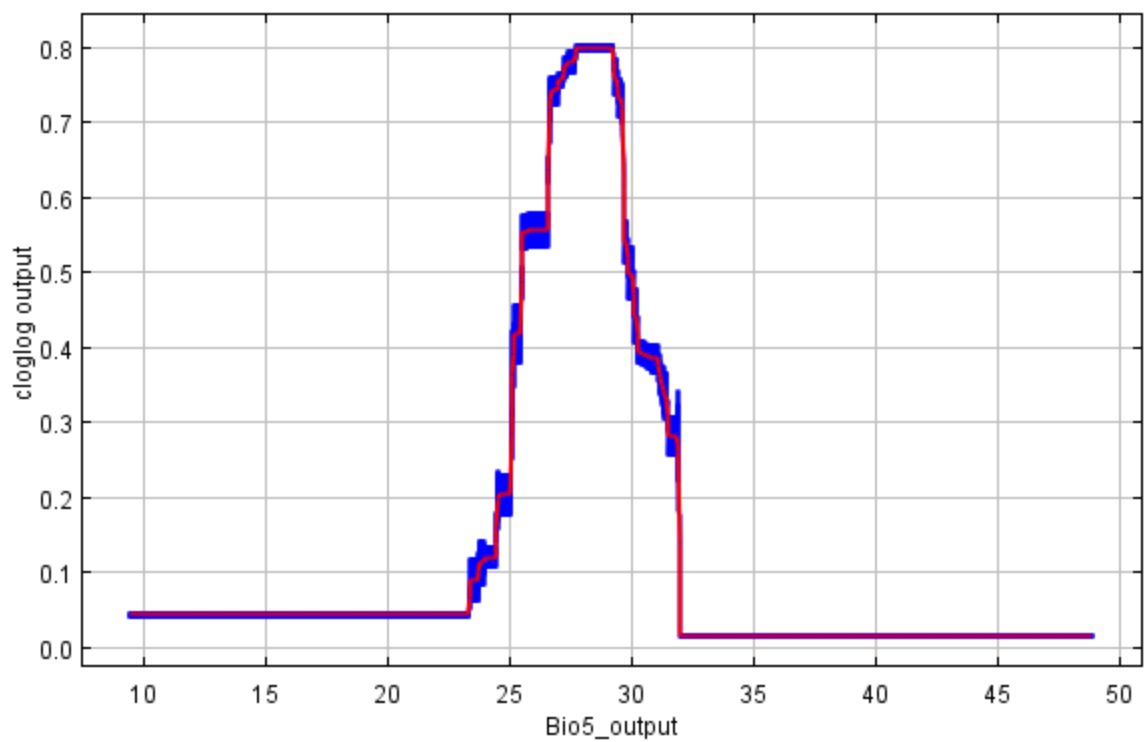


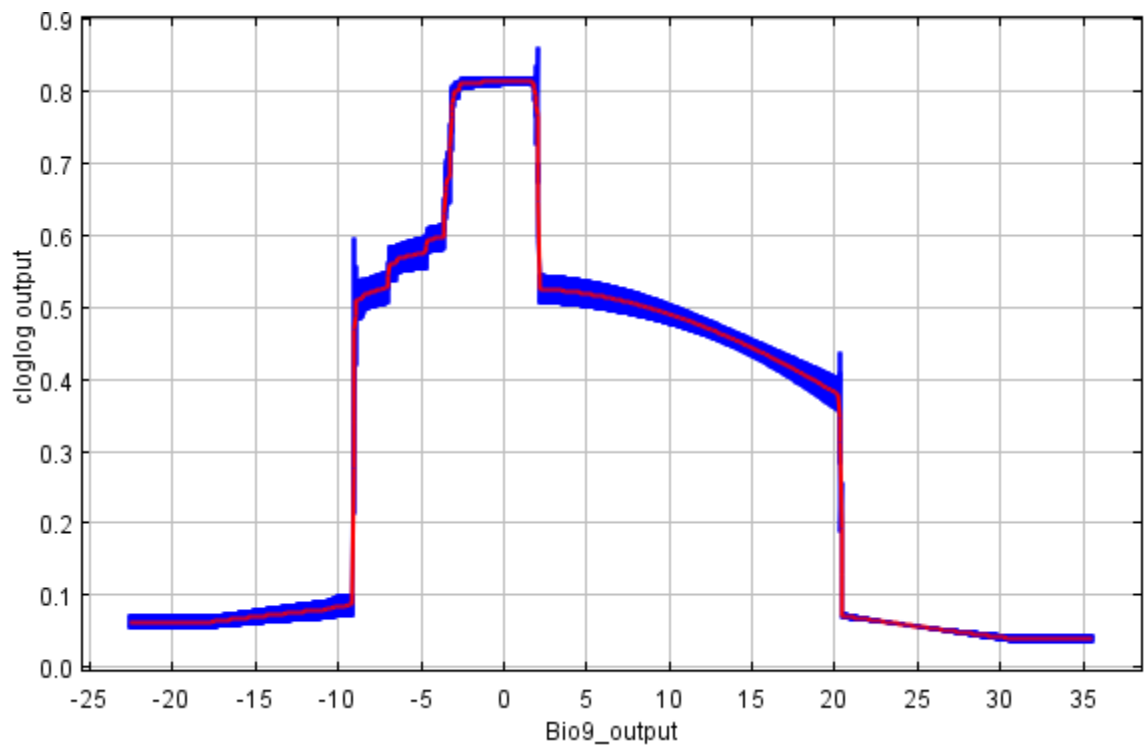
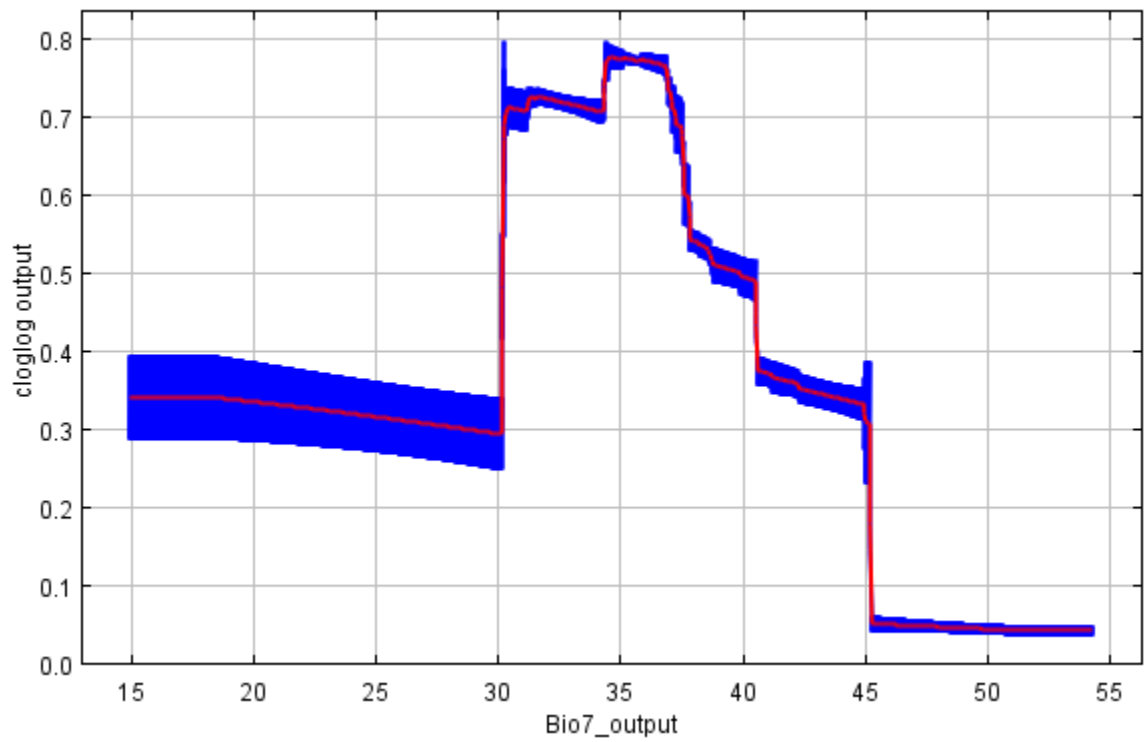


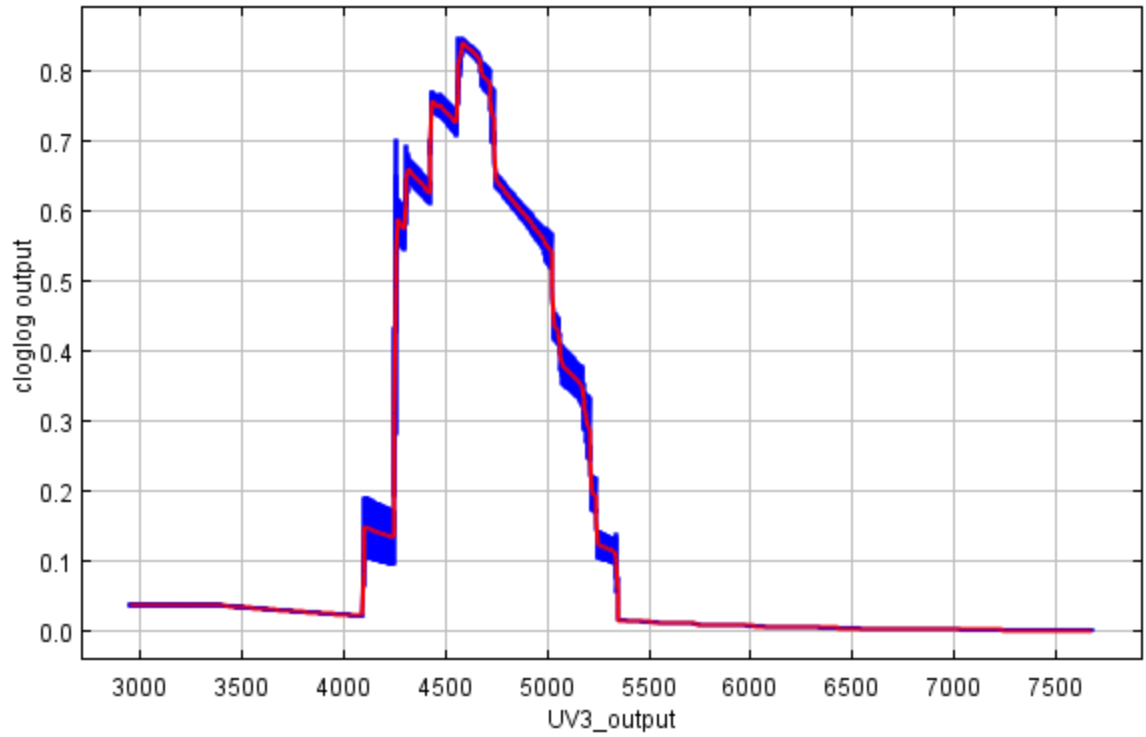


C.



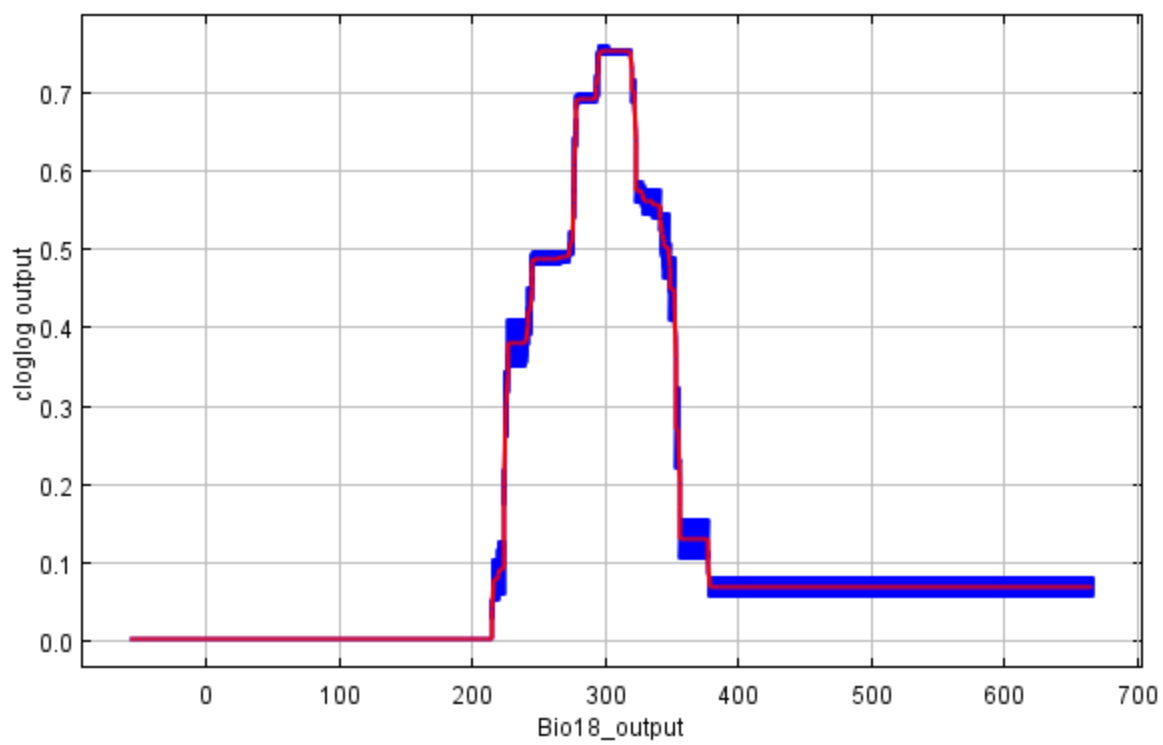
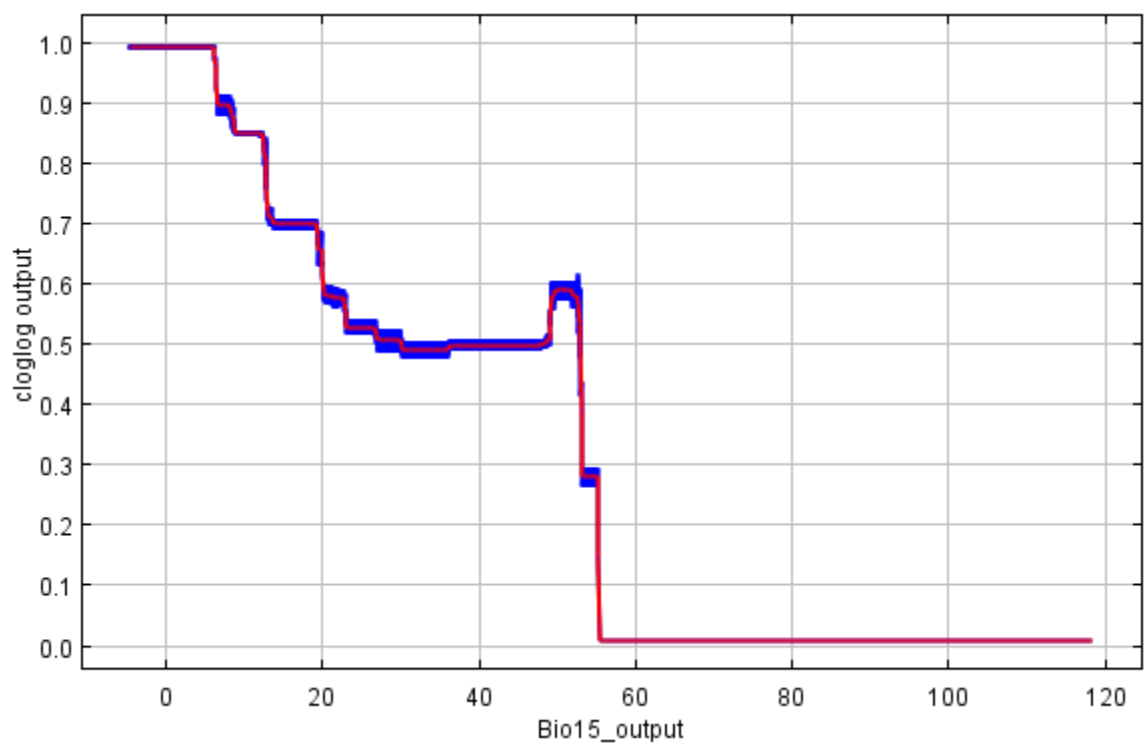


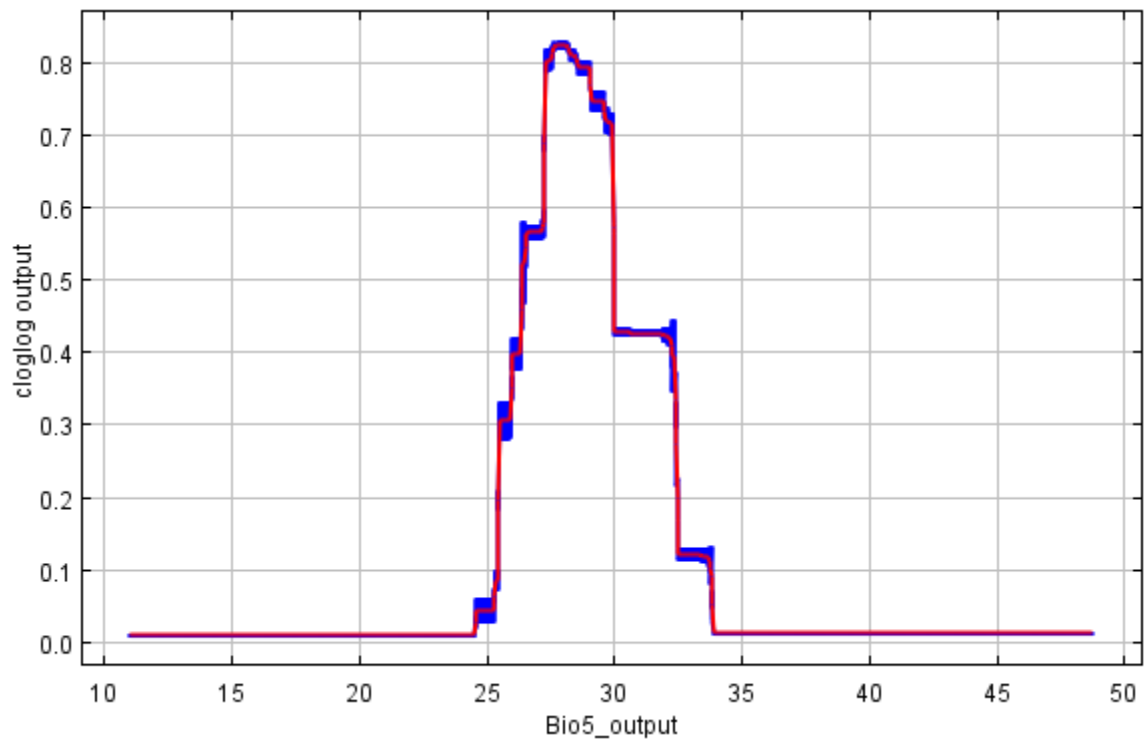
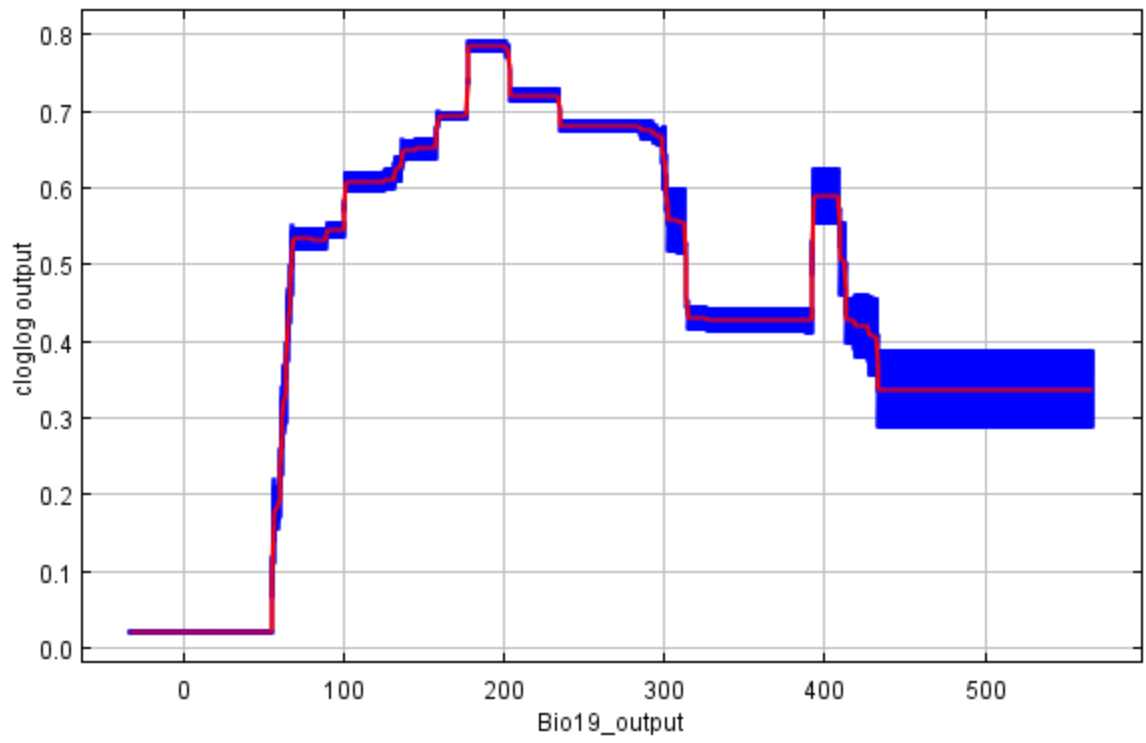


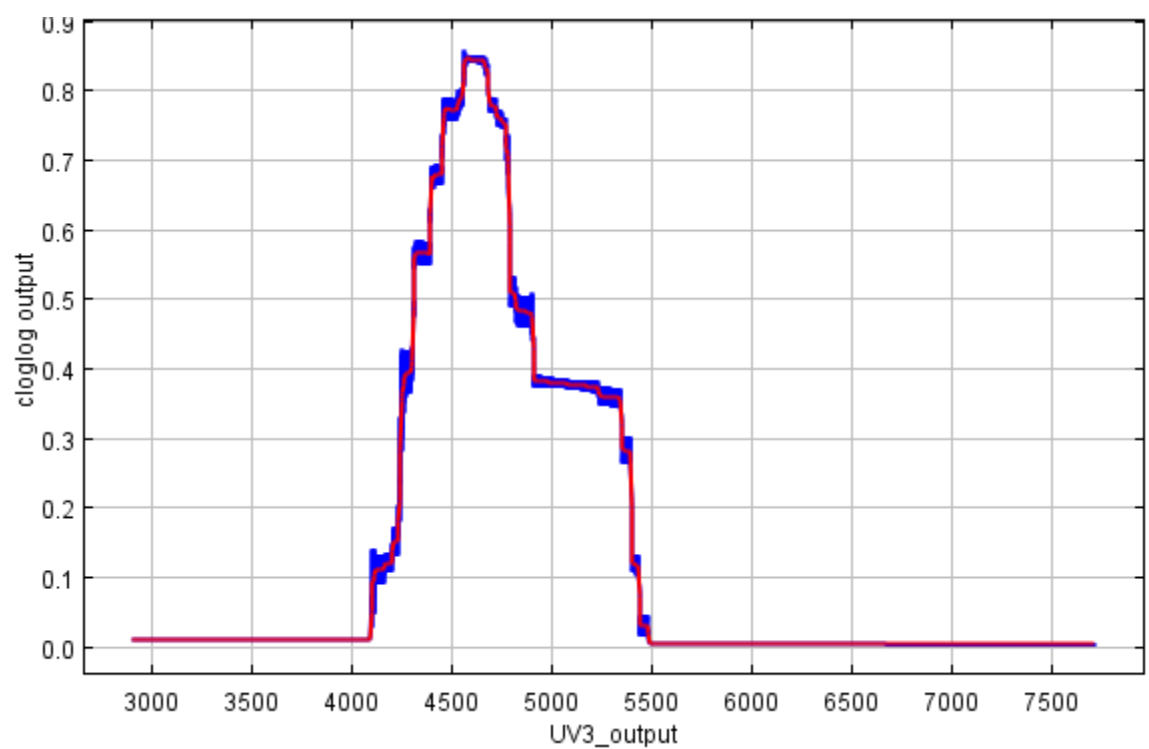
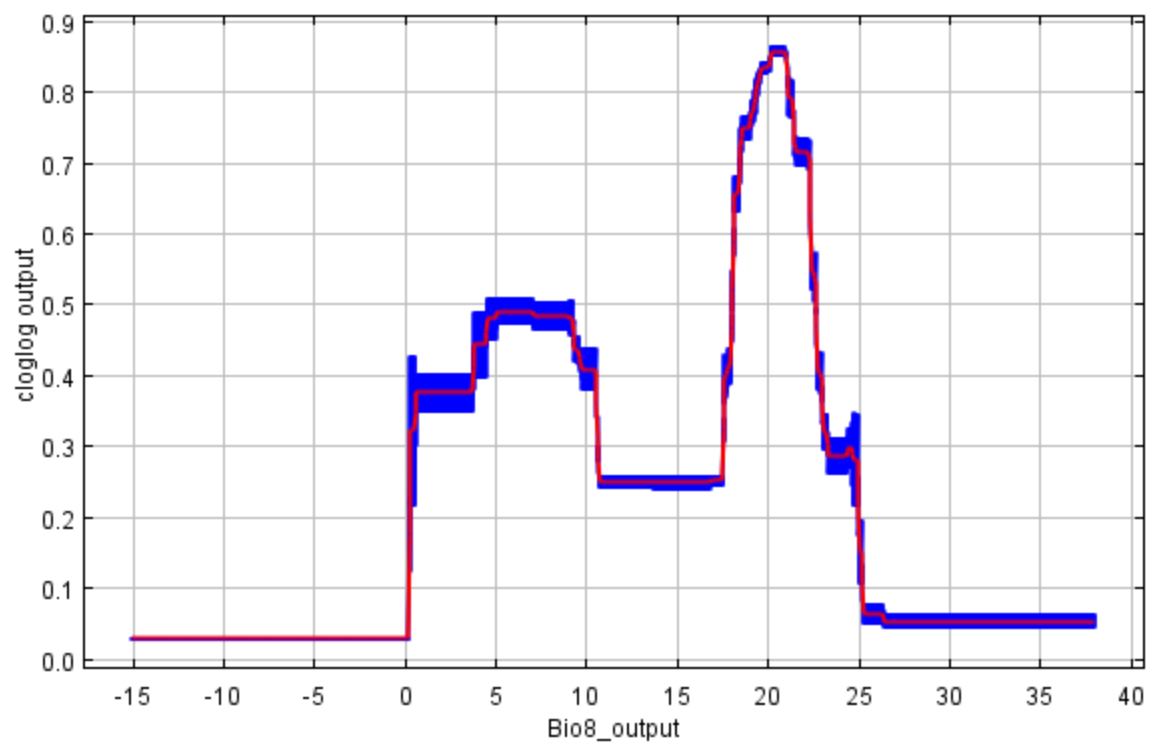


Supplemental Figure 3.2. Response curves for petals. **A.** is light petals. Each graph indicates a model made with only the indicated variable. **B.** is medium petals. **C.** is for dark petals. The red line is the mean for the Maxent runs, blue is one standard deviation.

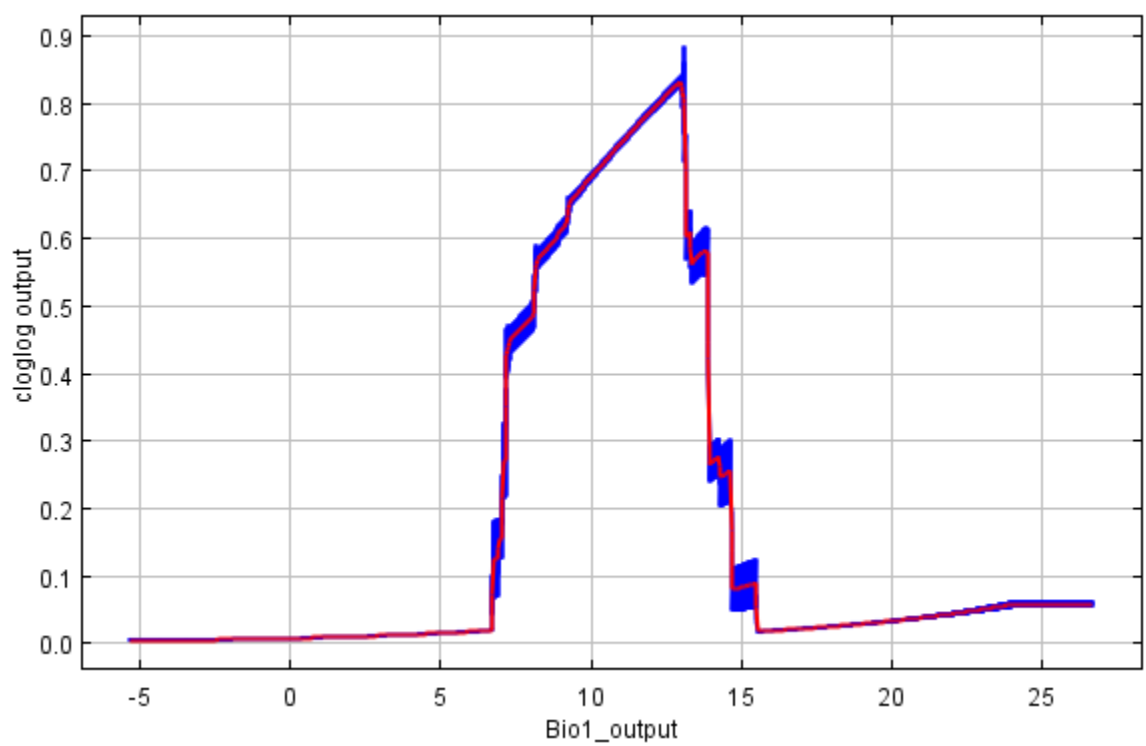
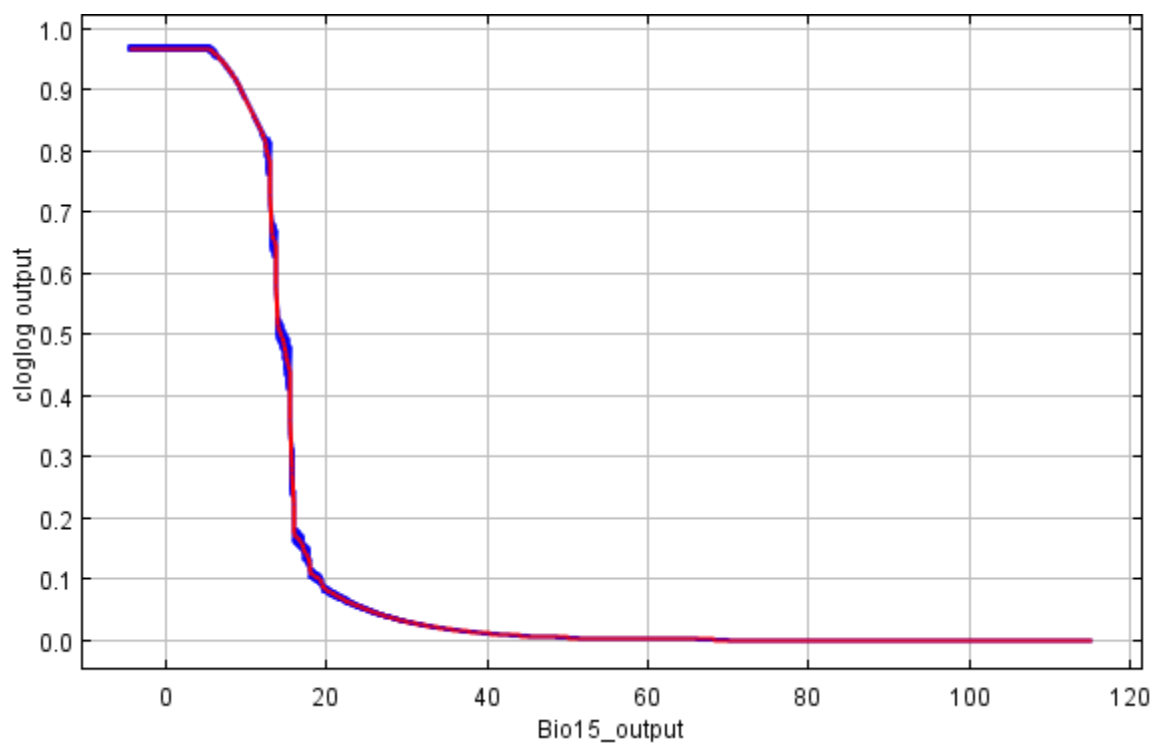
A.

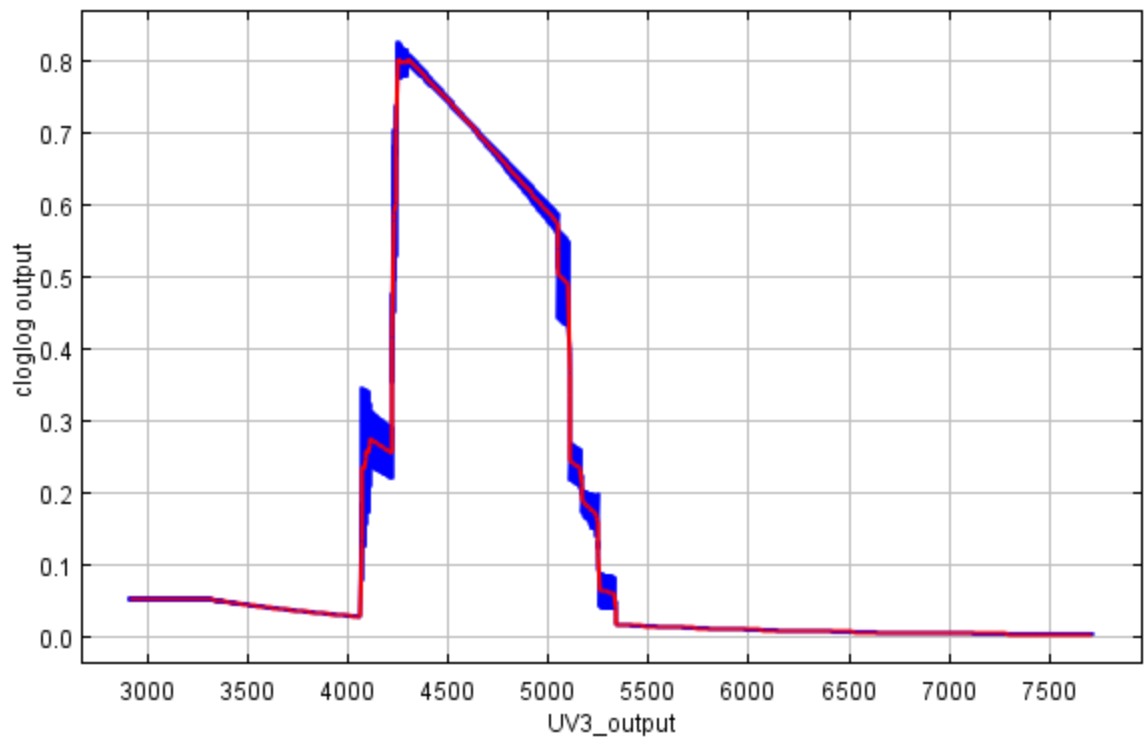
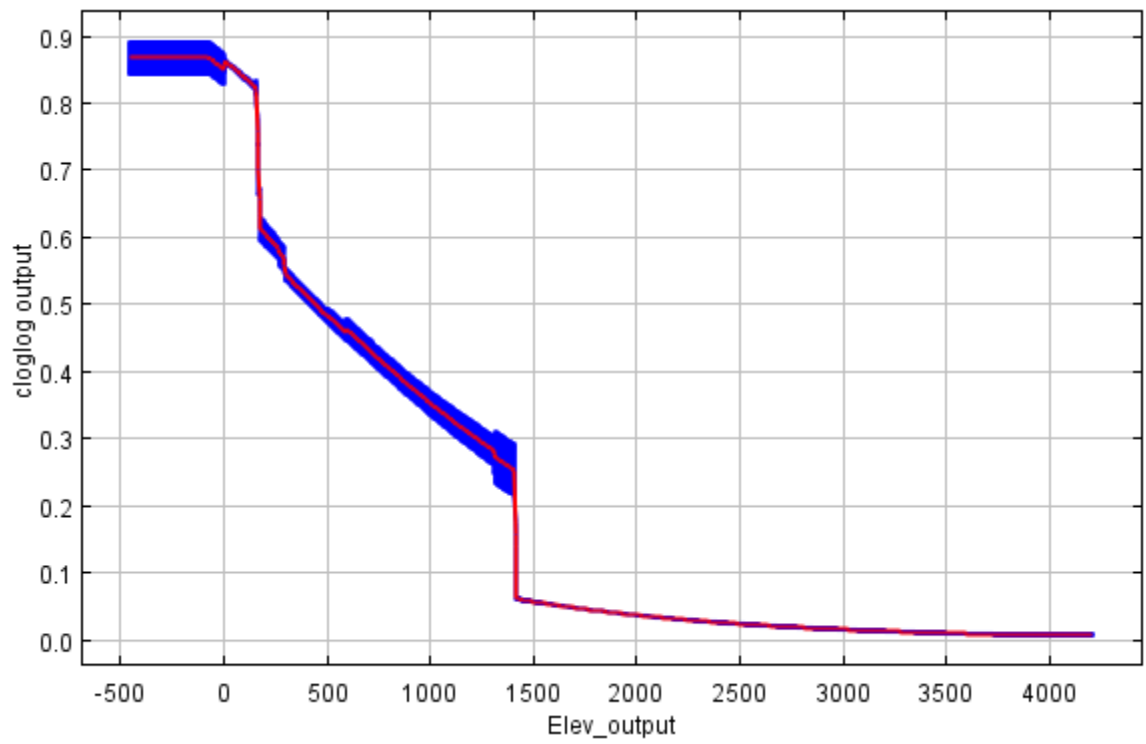




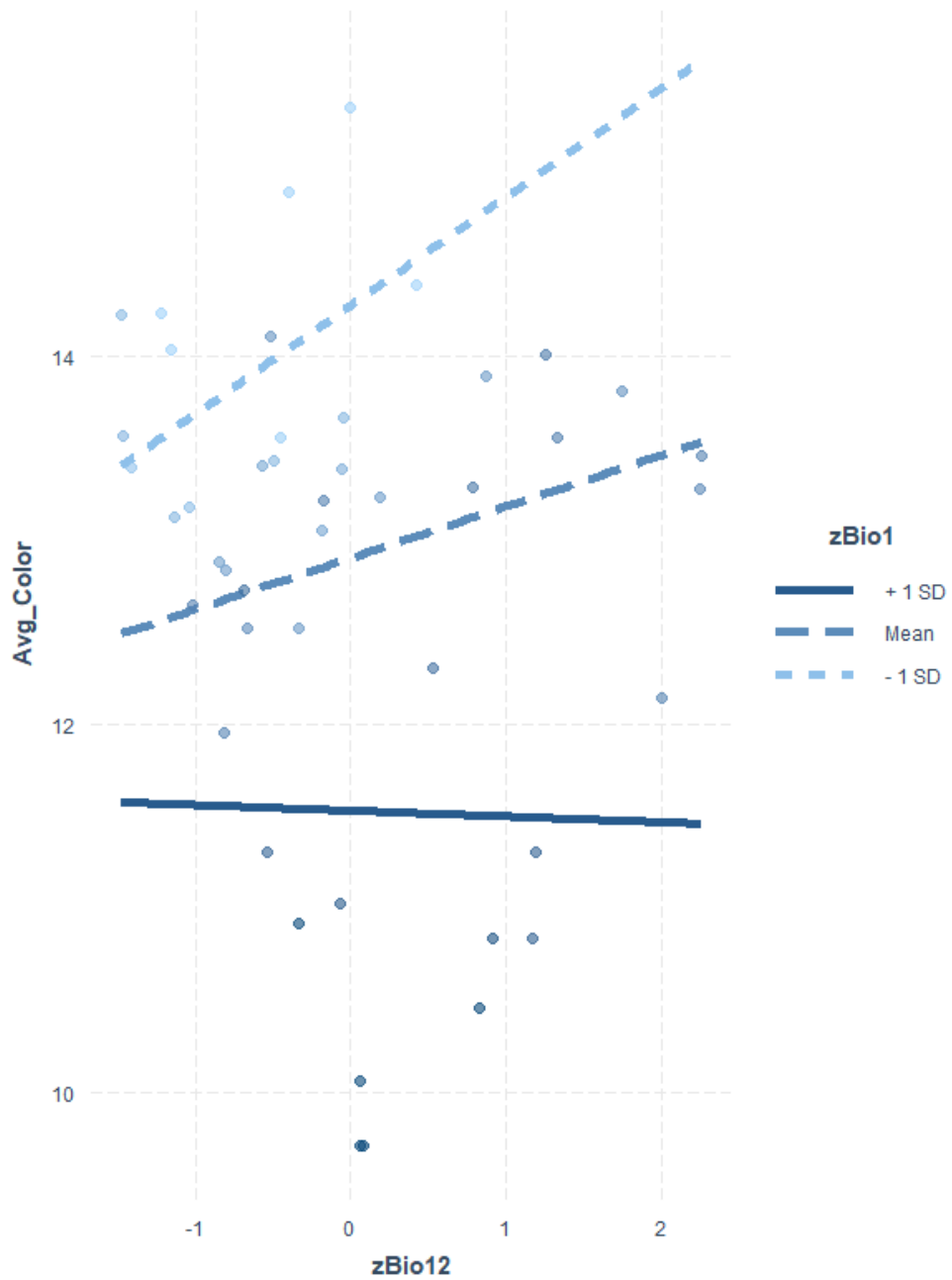


B.



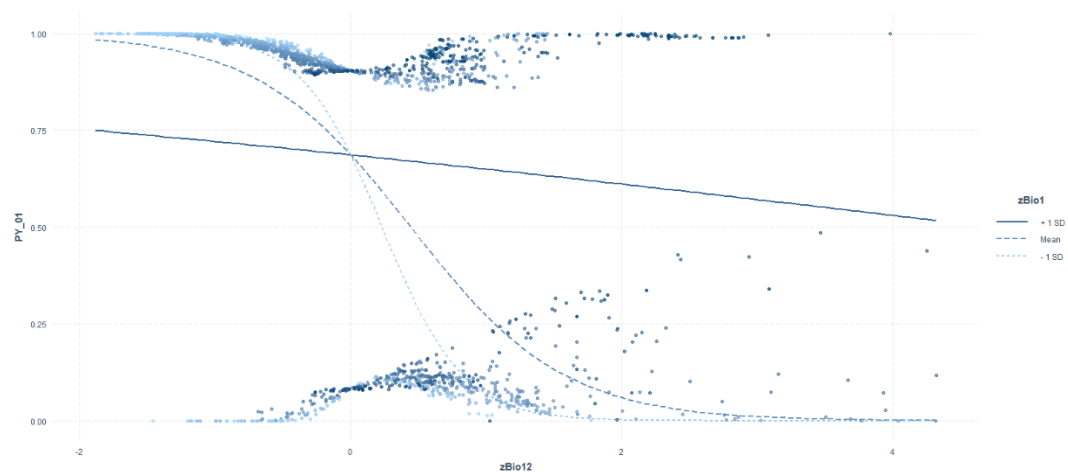


Supplemental Figure 3.3. Response curves for pollen. **A.** is yellow pollen. **B.** is purple pollen occurrences. The red line is the mean for the Maxent runs, blue is one standard deviation.

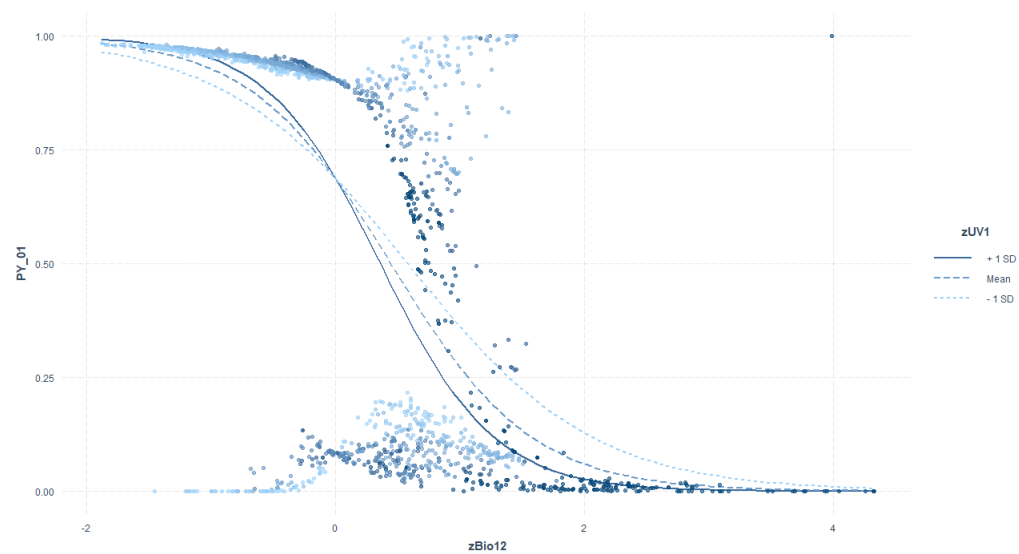


Supplemental Figure 3.4: Partial dependence plot of survey petal records. PDP showing relationship of Bio12, Bio1 and average population color. Climatic variables are Z transformed

A.



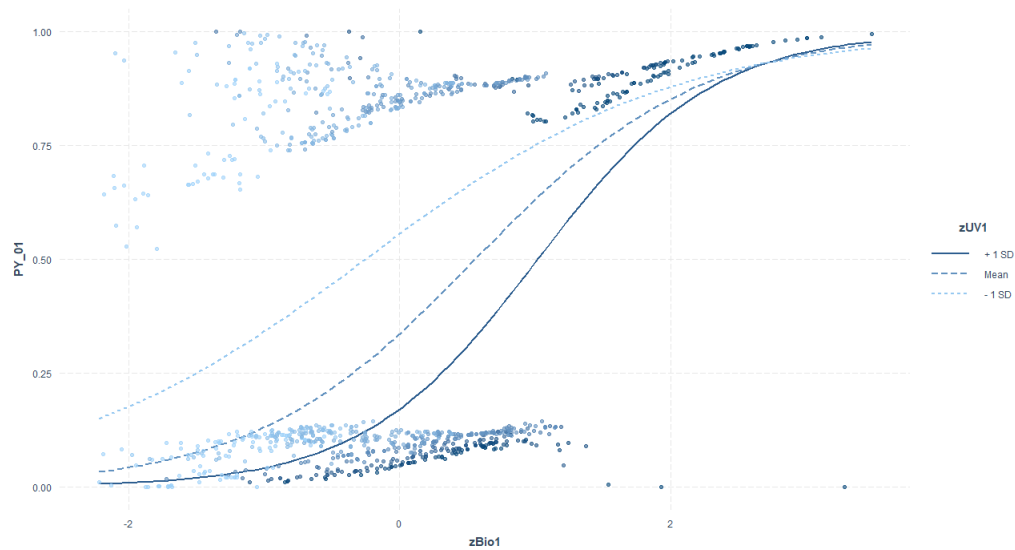
B.



C.

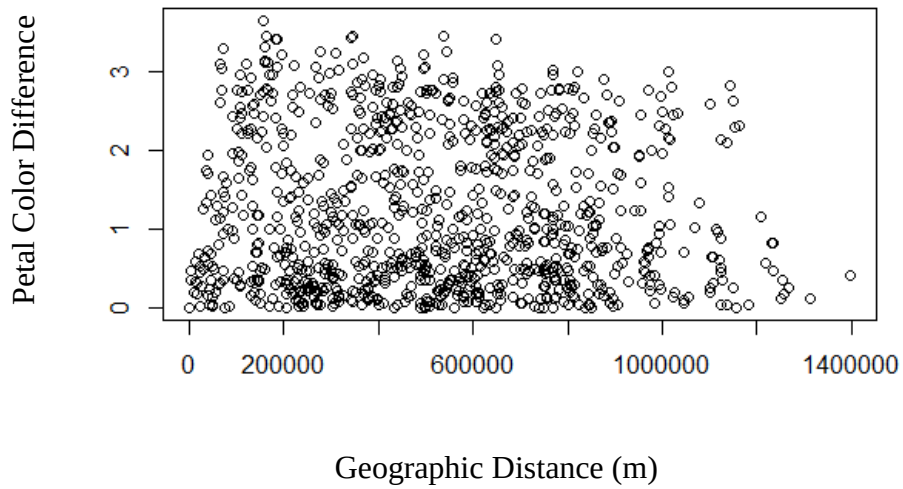


D.

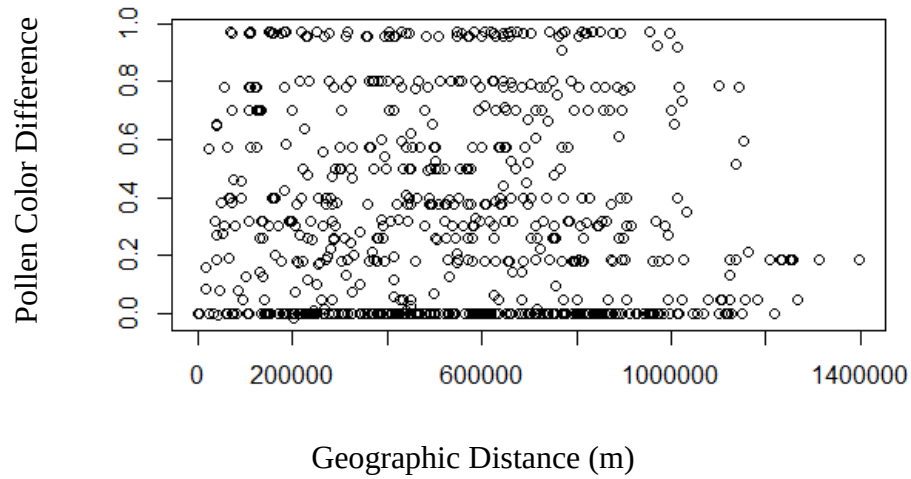


Supplemental Figure 3.5: Partial residual plots for iNaturalist pollen records. **A.** Full dataset, interaction of Bio 1 (Annual mean temperature) and Bio 12 (Annual precipitation). **B.** Full dataset, interaction of UV 1 (Annual mean UV-B) and Bio 1. **C.** Westward removed, interaction of Bio 1 and Bio 12. **D.** Westward removed, interaction of UV 1 and Bio 1. Y axis indicates 0=purple, 1= yellow. Full data set means and standard deviations (Bio1: 10.22, 2.34, Bio12: 1030.31, 190.97, UV1: 2452.387, 266.55). For westwards removed (Bio1: 11.32, 2.15, Bio12: 1168.185, 160.15, UV1: 2554.25, 29

A.



B.



Supplemental Figure 3.6: Difference in petal color (a) versus difference in geographic distance. Pollen (b) versus difference in geographic distance.