Historical refuges and recolonization routes in the Southern Appalachian Mountains, inferred through phylogeographic analysis of the spotted wintergreen (Chimaphila maculata)

John Daniel Banusiewicz
University of Mississippi
HISTORICAL REFUGES AND POSTGLACIAL RECOLONIZATION ROUTES IN THE SOUTHERN APPALACHIAN MOUNTAINS, INFERRED THROUGH PHYLOGEOGRAPHIC ANALYSIS OF THE SPOTTED WINTERGREEN (CHIMAPHILA MACULATA)

A Thesis
Presented in partial fulfillment of requirements for the degree of Master of Science in the Department of Biology
The University of Mississippi

by

JOHN BANUSIEWICZ

May 2017
ABSTRACT

Phylogeography has recently benefited from incorporation of coalescent modeling to test competing scenarios of population history of a species. Ecological niche modeling has also been useful in inferring areas of likely suitable habitat during past climate conditions. Several studies have examined the population history of biota in the Southern Appalachian Mountains and how they responded to climate change associated with the end of the Last Glacial Maximum (~18,000 years ago), though few studies have focused on understory plants. This study redressed that knowledge gap thorough examination of the phylogeographic history of the understory plant *Chimaphila maculata*, which is native to the Appalachian Mountains. Three microsatellite loci were developed and used in genetic analyses to assess how genetic diversity varied along a latitudinal gradient, and also to examine the number and geographic range of genetically distinct population clusters. An ecological niche model for *C. maculata* was generated to identify areas of past suitable habitat. Coalescent modeling was used to test four competing scenarios relating to past refugia and recolonization routes. Genetic variation was found to be higher in northern latitudes, though coalescent modeling was unable to discern one refugia and recolonization scenario for another. Four genetically distinct clusters were detected, and their geographic distribution showed similar patterns as other studies of the region. Results of this study were informative to the direction future studies could take to answer unresolved research questions regarding *C. maculata* in the Southern Appalachian Mountains.
DEDICATION

This thesis is dedicated to Mom and Dad, for everything.
LIST OF ABBREVIATIONS AND SYMBOLS

LGM  Last Glacial Maximum
kya  Thousand years ago
bp   Base pair
FAM  6-carboxy-fluorescine:
HEX  Hexachloro-6-carboxy-fluorescine
MCMC Markov chain Monte Carlo
PML  Potential microsatellite loci
PCR  Polymerase chain reaction
RFU  Relative fluorescence unit
SNP  Single nucleotide polymorphism
ACKNOWLEDGMENTS

This work was partly supported by funds from the University of Mississippi Graduate School Council. This project benefited from discussion with my committee, Dr. R. Garrick, Dr. R. Symula, and Dr. J. S. Brewer of the University of Mississippi Department of Biology. Thank you to Dr. R. Garrick, Dr. R. Symula, Dr. R. Dyer, C. Hyseni and E. Collier for much of the specimen collection. I would like to thank E. Collier for aiding in establishing C. maculata for the basis of research and aiding in DNA extraction, C. Hyseni, T. C. Glenn and T. Kieran for their assistance in microsatellite development, S. Burgess for assistance with ArcGis, and L. A. Passarella for helping with PCR.
# TABLE OF CONTENTS

ABSTRACT ............................................................................................................................................ ii  
DEDICATION....................................................................................................................................... iii  
LIST OF ABBREVIATIONS AND SYMBOLS....................................................................................... iv  
ACKNOWLEDGMENTS ........................................................................................................................... v  
LIST OF TABLES ................................................................................................................................ vii  
LIST OF FIGURES ................................................................................................................................ viii  
INTRODUCTION .................................................................................................................................... 1  
METHODOLOGY .................................................................................................................................. 6  
RESULTS ............................................................................................................................................... 21  
DISCUSSION ......................................................................................................................................... 26  
REFERENCES ....................................................................................................................................... 31  
APPENDIX ............................................................................................................................................ 44  
VITA...................................................................................................................................................... 46
# LIST OF TABLES

1. PCR cycling conditions used for testing potential microsatellite loci........................................9

2. Summary of non-redundant bioclimatic variables used in ecological niche modeling...............20

3. Summary of characteristics of three successful microsatellite loci...........................................21

4. Test of linkage disequilibrium within each genetic cluster.....................................................24

5. GPS coordinates and elevations of sample sites............................................................................45
LIST OF FIGURES

1. Photograph of *C. maculata*, showing its distinctive leaf color pattern................................. 5
2. Map of sampling sites from which *C. maculata* individuals were collected.......................... 7
3. The analytical approach used to determine number and location of genetic populations....... 13
4. Preliminary geographic range of clusters defined by individuals with no missing data......... 16
5. Four competing colonization scenarios.............................................................................. 18
6. Allelic richness compared to latitude of sampling sites....................................................... 22
7. Geographic range of four genetic clusters.......................................................................... 23
8. Approximate Bayesian computation test of four competing colonization scenarios......... 25
9. Ecological niche models of habitat suitability generated by MAXENT.............................. 25
10. Cluster convergence in the Great Smoky Mountains......................................................... 27
11. Comparison of past estimates of habitat suitability............................................................ 29
INTRODUCTION

Phylogeography combines molecular phylogenetics and population genetics with geographic data to study the evolutionary history of a species across its entire geographic range, over a broad temporal scale (Avise 2000, 2009). Microsatellite molecular markers have proven useful in understanding dispersal, and for reconstructing historical events, and are often incorporated into phylogeographic studies (Morgante et al. 2002; Varshney et al. 2005; Soltis et al. 2006; Ashley 2010). Phylogeography has also benefited from the growing wealth and accessibility of information on past climatic conditions through integration of ecological niche modeling and coalescent modeling (Carstens and Richards 2007). Ecological niche modeling is an algorithmic approach that is used to determine associations between a given species' geographic distribution and environmental conditions that are correlated with its occurrence. These associations provide information about current habitat suitability, and that information can be projected onto past climate conditions to estimate where populations were likely confined to in the past (Richards et al. 2007; Walker et al. 2009). This approach of understanding past responses to climate change provides a framework for anticipating future population responses (McLachlan et al. 2005; Cordellier and Pfenninger 2009; Razgour et al. 2013).

In addition to insights generated from ecological niche modeling using non-genetic data, coalescent modeling of genetic data allows one to test a set of competing hypotheses about long-
term population history that have been informed by knowledge of likely past distributions (Richards et al. 2007). Coalescent modeling scenarios attempt to explain the current genetic structure of a species by postulating the timing of genetically distinct populations coalescing ultimately into an ancestral population. These approaches have led to advances in phylogeographic inference by facilitating the testing of specific divergence and/or colonization scenarios that are formulated at the outset, rather than relying on post hoc interpretations based on phylogenetic trees overlaid on geographic maps (Knowles 2004; Hickerson et al. 2010).

The Southern Appalachian Mountains are a group of mountain ranges in the eastern United States dominated by mature hardwood forests. The southern boundary of the range is within Alabama and the southern edge of the Piedmont Plateau. The range expands northward into the Great Smoky Mountains that straddle Tennessee and North Carolina, and reach their northern limit at the Blue Ridge Mountains of Virginia. Elevation ranges from about 230 m in the valleys carved by rivers to peaks of nearly 1,300 m. The climate is temperate, and rainfall averages about 152 cm annually. Today, the Southern Appalachian Mountains have some of the highest levels of biodiversity in the world (Haskell 2000).

Glaciers covered northern regions of North America during the Last Glacial Maximum (LGM) of the Pleistocene epoch (Hewitt 2000). The most recent glaciation to spread across North America, the Wisconsinan, peaked ~18-23 thousand years ago (kya; Dyke and Prest 1987). Though glaciers did not cover the Southern Appalachians, the climate was much colder and drier than it is today (French and Millar 2014). Such conditions restricted suitable habitats to pockets of refugia in both terrestrial and aquatic biota (Bernatchez and Wilson 1998; Hewitt
2000; Brant and Ortí 2003; Church et al. 2003; Jaramillo-Correra et al. 2009). Recession of the Wisconsinan Glaciation ~18 kya (Dyke and Prest 1987) provided opportunities for recolonization throughout the region. This, in turn, lead researches to identify locations of refugia and recolonization pathways for several species. The results of several studies have shown that areas of suitable habitat differed among species, as did recolonization pathways (Gonzales et al. 2008; Potter et al. 2008; Jaramillo-Correra et al. 2009; Walker et al. 2009; Emerson et al. 2010; Li et al. 2013).

**Study species.**

Understory plants are underrepresented in Southern Appalachian phylogeographic studies, despite being the one of the most diverse components of the Appalachian forests (Huebner et al. 1995). Understory plants are usually short-lived and thus may have a generation time of a few years or less, and this provides opportunity to examine responses to a changing climate on a more incremental scale than larger, long-lived trees with longer generation times. To redress the knowledge gap, the present phylogeographic study focused on the spotted wintergreen (*Chimaphila maculata*; Figure 1). This species is a member of the Ericaceae family, a dioecious family in which most members have some degree of self-incompatibility (Sleumer 1960, Chavez and Lyerne 2009). *C. maculata* has 3–7 cm leaves, and a 10–25 cm flowering stalk that emerges during its annual pollination cycle. It is a perennial plant native to North America that is found throughout the Appalachian Mountains, and that seems to prefer the slightly acidic (pH 4–5) soil found in patches of the Eastern Boreal Forest (Anghinoni et al. 1996; Knoepp and Swank 1998). Wind is the primary seed dispersal agent, though its ~10 cm flowering stalk has a
conical shape and may harness raindrops to aid in dispersal (Amador et al. 2013). The species also has a conspicuous leaf morphology and color pattern that makes it easily recognizable, even when not in flower. Furthermore, C. maculata is diploid (2n = 26 chromosomes; Kondo 1972). This characteristic allows a variety of well-established population genetic analyses to be applied to understand population history, including inference of the number and location(s) of glacial refugia, and the timing and directionality of subsequent recolonization events.

Research questions.

In this study, sampling of C. maculata throughout the Southern Appalachian Mountains was coupled with analyses of microsatellite data, and ecological niche modeling, to reconstruct the phylogeographic history of this species. The following questions were addressed:

1. Does allelic richness of C. maculata vary over a latitudinal gradient?

2. How many genetically distinct populations of C. maculata currently exist in the Southern Appalachians, and where are they located?

3. Where were the areas of likely refugia for C. maculata during the LGM?

4. What recolonization pathways were likely used by C. maculata to emanate from its refuge or refugia to form its current distribution?
Figure 1. Photograph of *C. maculata*, showing its distinctive leaf color pattern.
METHODOLOGY

Sampling and DNA extraction.

One to two leaves from non-flowering *C. maculata* were collected by hand, preserved in silica gel and stored at -80° C. Collection was conducted in National Forests, National Parks, State Parks and other public lands from 2012–2015. During that time, leaf samples from 303 individuals across 81 sites were collected. One to six plants were collected at each location to examine the extent of local genetic diversity. DNA extraction for each plant was performed using a DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. GPS coordinates of each site were recorded in decimal degrees to four decimal places (Figure 2).
Figure 2. Map of sampling sites from which *C. maculata* individuals were collected. Each dot represents one sampling site (n = 81). A total of 303 individuals were collected.

Development of genetic markers.

In order to conduct analyses of population genetic structure in *C. maculata*, species-specific microsatellite markers were developed. Briefly, at the Georgia Genomics Facility T. C. Glenn performed low-coverage Illumina sequencing of pooled genomic DNA (Davey *et al.* 2011) from several *C. maculata* individuals, and T. Kieran conducted preliminary bioinformatics analyses. The latter involved scanning for microsatellite-containing sequence reads using the software MSATCOMMANDER v1.0.9 (Faircloth 2008). Thirteen loci containing either tri- or tetra-nucleotide microsatellite repeats were identified. The search was expanded by C. Hyseni to find dinucleotide microsatellite repeats. Based on this, a subset of 12 of these potential microsatellite loci (PLM) was included in a testing pool. Additionally, primer pairs for six
microsatellite loci designed for *C. japonica* by Liu *et al.* (2012) were tested, giving a total of 31 potential microsatellite loci to test.

*Selection of potential microsatellite loci (PML) suitable for downstream analysis.*

A panel of individuals was established to test if each PML was suitable for downstream analysis. The composition of individuals within the test panel differed for each PML, but every test panel contained DNA from seven to 10 individuals that represented the entire latitudinal range of the study region. The first goal was to identify PML that amplified successfully via polymerase chain reaction (PCR), and whose bands matched the anticipated size range (measured in base pairs; bp) based on MSATCOMMANDER analyses.

*Polymerase Chain Reaction (PCR) amplifications.*

Test panels for every PML were initially run using the following mix of reagents, henceforth referred to as PCR Mix 1: 3 µL 5x Buffer, 1.20 µL MgCl₂ (25 mM), 2.40 µL dNTPs (1.25 mM; Promega), 0.75 µL Bovine Serum Albumin (10 mg/mL; New England Biolabs), 4.50 µL dH₂O, 0.75 µL Forward Primer (10 µM), 0.75 µL Reverse Primer (10 µM), 0.15 µL GoTaq (5U/µL Promega), and 1.5 µL 1:19 diluted genomic DNA. Each PCR mixture was placed in a BioRad T100 thermal cycler and was run using a touchdown protocol shown in Table 1.
Table 1. PCR cycling conditions used for testing potential microsatellite loci. Each locus was assessed using a test panel of seven individuals.

<table>
<thead>
<tr>
<th>Step</th>
<th>Temperature (°C)</th>
<th>Duration (seconds)</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95</td>
<td>120</td>
<td>1x</td>
</tr>
<tr>
<td>2</td>
<td>95</td>
<td>30</td>
<td>10x*</td>
</tr>
<tr>
<td>3 (denature)</td>
<td>60, then -1° per cycle</td>
<td>20't0^2</td>
<td>10x*</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
<td>20</td>
<td>10x*</td>
</tr>
<tr>
<td>5</td>
<td>95</td>
<td>30</td>
<td>25x</td>
</tr>
<tr>
<td>6 (denature)</td>
<td>50</td>
<td>20't0^2</td>
<td>25x</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>20</td>
<td>25x</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>120</td>
<td>1x</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>∞</td>
<td>1x</td>
</tr>
</tbody>
</table>

* steps 2-4 repeated 9 times before moving on to step 5
1 tri and tetranucleotide PML
2 dinucleotide PML

Amplified products were run for two hours on a 1.5–2% 1x TBE agarose gel stained with GelRed (Phenix Research) at 90 volts and 380 milliamps alongside a 100-bp ladder. Amplified bands were compared to the 100-bp ladder to estimate fragment length, and photographed using a BioRad GelDoc XR gel documentation system with the program Quality One v4.6.5.

*PCR amplifications using a fluorescently labeled primer.*

PML that showed suitable amplification in the test panel were tested a second time, this time with the fluorescently labeled version of the forward primer. Forward primers of tri- and tetra-nucleotide PML were substituted in PCR mixtures for either a 6-carboxy-fluorescine (FAM), or hexachloro-6-carboxy-fluorescine (HEX) fluorescently labeled forward primer of equal concentration and quantity as previously described. PCR mixtures for dinucleotide PML were modified to incorporate a HEX dye attached to the 5' end of an abbreviated version of a non-fluorescent forward primer (Forward Primer Tail: M13Hex (5'-/5HEX/TCCCAGTCACGACGT-3', Schuelke 2000). PCR conditions when M13Hex were
incorporated were modified to the following: 3 µL 5x Buffer, 1.20 µL MgCl₂ (25 mM), 2.40 µL dNTPs (1.25 mM), 0.75 µL Bovine Serum Albumin (10 mg/mL), 3.70 µL dH₂O, 0.80 µL Forward Primer Tail (1 µM), 0.75 µL Reverse Primer (10 µM), 0.75 µL M13Hex primer (1 µM), 0.15 µL GoTaq (5U/µL), and 1.5 µL 1:19 diluted genomic DNA. Thermal cycler protocols for di-, tri-, and tetra-nucleotide PML were the same as in Table 1, and electrophoresis were the same as those listed above.

Genotype scoring.

PCR products that amplified with a fluorescent primer were prepared in a 1:9 dilution in HiDi Formamide (Applied Biosystems) and sent to Yale University’s DNA Analysis Facility on Science Hill for fragment analysis on an ABI 3730x capillary sequencer, with fragment sizes estimated using a Rox500 ladder. Allele size estimation and genotype scoring were performed using the software GENEIOUS v9.1.2 and the GENEIOUS Microsatellite Plugin v1.4.4 (Kearse et al. 2012). At this point, the following three criteria were used to identify a suitable locus:

1) An individual from the test panel showed either one or two clearly defined peaks of at least 500 relative fluorescence units (RFUs), with any immediately surrounding stutter peaks less than 100 RFUs, within size range expected for the microsatellite locus being evaluated, plus or minus 30-bp to account for the inherent variable length of microsatellite alleles. Since diploid individuals have a maximum of two different alleles per locus, a locus that showed any individual as having any more than two alleles was removed from the set. The above RFU values
were chosen to avoid ambiguity when distinguishing a true peak from smaller stutter peaks that generally accompany a true peak.

2) The majority of individuals from the test panel showed one or two clearly defined peaks within the parameters of criterion one. The locus needed to show it could reliably generate peaks in the test panel before testing could commence on all individuals.

3) Collectively, the pool of test panel individuals with clear peaks represented a mixture of heterozygous and homozygous genotypes. This would suggest the locus showed Mendelian inheritance patterns suitable for downstream analysis.

PML that fulfilled all three criteria were used to screen every individual at all sampling sites in the Southern Appalachian Mountains. Three loci fulfilled all criteria and were included in subsequent analyses (see Results).

*Latitudinal patterns of allelic richness.*

Allelic richness measures the number of different alleles at a given sampling location. It has been a useful measure in phylogeographic studies because areas of refugia are often characterized by high allelic richness, with a subsequent loss of allelic richness in more recently recolonized areas (Hewitt 2000; Takahashi *et al.* 2005; Maggs *et al.* 2008). Twenty-seven sampling sites with at least three individuals that amplified at loci "Cm79332" and "Cm81520" (the two most successful loci) were used in this analysis. The number of individuals per sample
site ranged from three to six, and so rarefaction correction was applied to weigh sampling sites with varying number of individuals (subsample of three per sampling location) using software HP-RARE v1.0 (Kalinowski 2004, 2005).

*Examination of genetic structure.*

Several steps were used to assess evidence for genetic structure. The goal was to assign all 242 individuals that were genotyped at any number of loci to a genetic population (Figure 3).
Figure 3. The analytical approach used to determine number and location of genetic populations. a) STRUCTURE runs determined four genetic clusters from 42 individuals with no missing data. Colored dots on the map are sampling sites of those individuals that were strongly assigned to one cluster ($Q > 0.80$) and the colored polygons represent geographic area of each cluster. Grey dots are sites with weakly assigned individuals ($Q < 0.80$). b) Expansion of cluster area through incorporation of all 264 individuals with some missing genotypic data. Black dots on the top map are sampling sites that were assigned to clusters after a STRUCTURE run. The bottom map shows all individuals assigned to a cluster and the geographic area of each cluster.
The program STRUCTURE v2.3.4 (Pritchard et al. 2000) was used to examine evidence for genetic structure within *C. maculata*. This software uses Bayesian computation and Markov chain Monte Carlo (MCMC) algorithms to identify the number of distinct genetic populations represented by the sampled individuals, and also to assign individuals into genetic populations that fit the expectations of Hardy-Weinberg and linkage equilibrium, based on the multi-locus diploid genotype data (Pritchard et al. 2000). A summary of relevant parameters that were used in STRUCTURE analyses are as follows: $K$: number of clusters (i.e. genetic populations), $\alpha$: relative admixture between populations, $\lambda$: similarity of allele frequencies among populations, $Q$: membership co-efficient (i.e., probability an individual belongs to a certain cluster), POPINFO: pre-specified genetic population information, in which each integer corresponds to a different genetic population.

Genotypic clustering was performed in STRUCTURE to establish a basic insight into the number and locations of genetic populations. Genotypes from the 42 individuals with no missing data were first used to identify putative populations. For this and all subsequent STRUCTURE runs, 100,000 MCMC iterations were used after discarding the first 10,000 as burn-in. For this first run, parameters were set to reflect the as-yet unknown number of populations by considering a range of $K$ values (i.e., $K=1–8$). Also, rather than assuming an $\alpha$-value, the software was set to infer $\alpha$ for each population. This was done to mitigate the possibility that each genetic population might not be equally represented in the sample set, as directed by the STRUCTURE manual. Additionally, $\lambda$ was set to 1.0 (default) for each putative genetic population because inheritance of alleles was assumed to be identical for the species regardless
of genetic population. Ten iterations under the listed conditions were run for each assumed $K$ value.

STRUCTURE HARVESTER Web v0.6.94 (Earl and vonHoldt 2012) is a web-based program that estimates the true $K$ value from a range of $K$ using the delta-$K$ method (Evanno et al. 2005). Results of the above STRUCTURE runs were processed using STRUCTURE HARVESTER, and $K=3$ was identified as the most likely scenario (see Results). All 42 individuals were assigned one of these $K=3$ clusters based on their $Q$ value (membership coefficient). Any individual with an average $Q$ value of 0.80 for a given population across the 10 replicate STRUCTURE runs was assigned to that population.

Following Evanno et al. (2005), evidence for subtle population structure within each of the three major clusters was assessed by reanalyzing each of them separately in STRUCTURE, using the same parameters as above. Output from the latter STRUCTURE runs were again analyzed in STRUCTURE HARVESTER. Two of the three clusters revealed a likely $K=1$, but one population cluster revealed a likely $K=2$, which suggested that this was actually two clusters. $Q$ values were used to assign individuals into one of the two new clusters using a simple majority vote criterion (i.e. $Q>0.50$).

Results of the prior analyses were used to reassign all 42 individuals to one of four populations. STRUCTURE was used again under the same conditions, but $K$ was fixed at four. Ten iterations were run, and the $Q$ values of each of the iterations were then averaged for each individual. Any individual with $Q>0.80$ for a single population was assigned to that population.
All 42 individuals were plotted on a map using their GPS coordinates. A polygon was drawn around each genetic population and any individual that fell within that polygon was assumed to belong to that population (Figure 4). These preliminary population assignments were incorporated into subsequent STRUCTURE runs.

![Figure 4](image_url)

Figure 4. Preliminary geographic range of clusters defined by individuals with no missing data. Individuals are plotted as dots. If two or more individuals had the same GPS coordinates, they were pooled together. Each color corresponds to a different genetically distinct population, as follows: red=North, green=Central, yellow=East, and blue=South. Shaded regions are the tentative geographic ranges of clusters. Grey dots are locations where individuals with Q<0.80 for any one cluster, and thus remained unassigned.

**Refinement of genetic populations thorough incorporation of individuals with missing data.**

GPS coordinates of all individuals with genotype data for at least one locus, including the 42 already used (n=264) were plotted over the preliminary cluster map. All individuals that fell within a cluster's polygon were assigned to that cluster. Individuals that fell outside the polygon that denoted a population were not assigned to a population. Overall, 84 of 264 individuals fell
outside the area of a polygon. STRUCTURE was run to include all 264 individuals using the same settings as the previous run. As before, $Q$ values were averaged across 10 replicate runs. An individual that had yet to be assigned to a cluster was assigned to the cluster for which $Q > 0.50$. Any individual with $Q < 0.50$ for any one cluster was assigned to the cluster of other individuals from its sampling location. GENEPOP v4.2 (Raymond and Rousett 1995, Raymond 2008), a population-genetic toolbox that estimates Weir and Cockerham's (1984) $F$ statistics and uses MCMC algorithms to calculate their significance, was used with default parameters to confirm consistent with Hardy-Weinberg Equilibrium and linkage equilibrium.

*Likely refugia and recolonization pathways.*

Initially, STRUCTURE HARVESTER misidentified as one cluster what was later revealed to be two clusters. To verify that individuals within these two clusters belonged to closely related populations, pairwise $F_{st}$ values (Weir and Cockerham 1984) were calculated in POPTREE2 (Takezaki *et al.* 2010), and this showed with high confidence that those individuals that belonged to two separate clusters (North and East, see Results) were closely related. Accordingly, those two clusters were combined into one cluster for the purposes of defining alternative recolonization scenarios, as fewer populations make subsequent analyses more tractable. To test four competing scenarios of refugia and recolonization (Figure 5), DIYABC v2.1 (Cornuet *et al.* 2014) was used.
Figure 5: Four competing colonization scenarios. Each "t" is a different range of times of a colonization event. More recent events (t1) are 5-20,000 generations ago and older events (tx2) 25-50,000 generations ago. Maps show timing of recolonization events and arrows show directionality. Each population tree corresponds to the scenario directly above it, and each time and color corresponds to a time and cluster on the map. Asterisks on population trees represent a bottleneck (10-100 generations) that occurred as the result of a found effect, each time a new lineage was being formed.

For each parameter that was used to define a scenario, a broad range of possible values (as defined by priors) were used. Effective population size for each present day cluster was set at 100-10,000 individuals. Every recolonization event was assumed to involve a bottleneck resulting from a founder event of 10-500 individuals that lasted for 10-100 generations. Effective population size of the ancestral cluster (or clusters in the case of the no-recolonization scenario) was set at 100-10,000 individuals. In both the southern refuge and northern refuge scenarios, the oldest recolonization event (t2 on Figure 5) was set at 15-25 kya and the more recent event (t1 on Figure 5) from 5-20 kya. Though these time ranges have some overlap, conditions were set such that t2 was always in the more distant past than t1. The central refuge scenario assumed two simultaneous recolonization events that occurred 15-25 kya (tx1). The three refugia model assumed any recolonization occurred beyond the scope of this study (tx2, 25-50 kya). Across all
scenarios, four million 3-locus microsatellite datasets for 42 individuals with no missing data were simulated in DIYABC. The summary statistics used to characterize the simulated (and empirical) datasets were mean number of alleles, mean genic diversity, and mean size variance for alleles within each cluster, and a pairwise $F_{st}$ between each combination of clusters. Posterior probabilities calculated in DIY-ABC were used to determine which of the four scenarios was most consistent with the empirical genetic data.

*Ecological niche modeling.*

The program MAXENT v3.4.0 (Phillips and Dudík 2008) was used to generate ecological niche models for *C. maculata* at two time periods: present day, and LGM. Occurrence records were compiled from personal collections (n = 81 non-redundant records), and also from occurrence data available via the United States Geological Survey (bison.usgs.gov). Occurrence data from the United States and Canada (with the exception of museum collections in Arizona, which were discarded) were trimmed to incorporate only individuals with GPS precision to four decimal places and duplicates were deleted (n=655). In total, 736 non-redundant records were used for MAXENT analysis. Current climate data and climate data from the LGM was compiled for 19 bioclimatic variables (Table 2) at 2.5-minute resolution (WorldClim.org, Hijmans *et al.* 2005). All occurrence data were contained within this trimmed area. Fifteen iterations of MAXENT were used, and each one pulled a random 25% subsample of occurrence records. Results of all iterations were averaged.
<table>
<thead>
<tr>
<th>Bioclimatic Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO1</td>
<td>Annual Mean Temperature</td>
</tr>
<tr>
<td>BIO2</td>
<td>Mean Diurnal Range (Mean of monthly (max temp - min temp))</td>
</tr>
<tr>
<td>BIO3</td>
<td>Isothermality (BIO2/BIO7) (* 100)</td>
</tr>
<tr>
<td>BIO4</td>
<td>Temperature Seasonality (standard deviation *100)</td>
</tr>
<tr>
<td>BIO5</td>
<td>Max Temperature of Warmest Month</td>
</tr>
<tr>
<td>BIO6</td>
<td>Min Temperature of Coldest Month</td>
</tr>
<tr>
<td>BIO7</td>
<td>Temperature Annual Range (BIO5-BIO6)</td>
</tr>
<tr>
<td>BIO8</td>
<td>Mean Temperature of Wettest Quarter</td>
</tr>
<tr>
<td>BIO9</td>
<td>Mean Temperature of Driest Quarter</td>
</tr>
<tr>
<td>BIO10</td>
<td>Mean Temperature of Warmest Quarter</td>
</tr>
<tr>
<td>BIO11</td>
<td>Mean Temperature of Coldest Quarter</td>
</tr>
<tr>
<td>BIO12</td>
<td>Annual Precipitation</td>
</tr>
<tr>
<td>BIO13</td>
<td>Precipitation of Wettest Month</td>
</tr>
<tr>
<td>BIO14</td>
<td>Precipitation of Driest Month</td>
</tr>
<tr>
<td>BIO15</td>
<td>Precipitation Seasonality (Coefficient of Variation)</td>
</tr>
<tr>
<td>BIO16</td>
<td>Precipitation of Wettest Quarter</td>
</tr>
<tr>
<td>BIO17</td>
<td>Precipitation of Driest Quarter</td>
</tr>
<tr>
<td>BIO18</td>
<td>Precipitation of Warmest Quarter</td>
</tr>
<tr>
<td>BIO19</td>
<td>Precipitation of Coldest Quarter</td>
</tr>
</tbody>
</table>

Table 2. Summary of non-redundant bioclimatic variables used in ecological niche modeling as shown on worldclim.org
RESULTS

Microsatellite discovery.

Three loci of the 31 tested (9.7%) fulfilled all criteria and were included in subsequent analyses (Table 3). All three of these loci were identified using Illumina sequences from C. maculata. Every C. japonica primer amplified, but failed to show evidence of polymorphism.

Table 3. Summary of characteristics of three successful microsatellite loci. Primer sequences for forward (F) and reverse (R) primers are included. Repeat motif is followed by the number of repetitions in parentheses. Size is the expected length of the amplified product.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primer Sequences (5'-3')</th>
<th>Repeat Motif</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cm79332</td>
<td>F: ACCCAATTGCTACTACATGGAC R: GAGTGTCAATCATGCGCA</td>
<td>ATC (11)</td>
<td>101</td>
</tr>
<tr>
<td>Cm81520</td>
<td>F: TGCGATCAAGAGAAGACGAC R: GCTGATAGGTGGCAGTCCTTG</td>
<td>ACT (8)</td>
<td>147</td>
</tr>
<tr>
<td>Cm114691</td>
<td>F: TATCGACAGGCGCATGGGAC R: CGGCGGTAGGGTTATATCGG</td>
<td>CT (20)</td>
<td>343</td>
</tr>
</tbody>
</table>

Of the 301 individuals sampled, at least one microsatellite locus was successfully scored in 264 individuals. For 165 individuals two microsatellite loci were scored successfully. For 42 individuals, all three loci were scored successfully.
Latitudinal patterns of allelic richness.

Linear regression of allelic richness vs. latitude showed allelic richness to be significantly higher (i.e. show greater genetic diversity) in more northern latitudes (Figure 6; p = 0.022).

![Figure 6. Allelic richness compared to latitude of sampling sites. Each dot is a unique sampling site with at least three individuals that amplified at both Cm79332 and Cm81520 (n=27). Allelic richness used rarefaction correction to normalize each sampling site, and latitude is expressed in decimal degrees (dd). The trend line shows linear regression, $r^2=0.192$ (p = 0.022).](image)

Examination of genetic structure.

*C. maculata* exists in four genetically distinct populations in the Southern Appalachian Mountains: North, East, Central and South (Figure 7). Three of the 264 individuals used to infer the spatial extent of each cluster failed to show $Q>0.80$ after the final round of STRUCTURE analysis and were omitted from subsequent analyses. Their exclusion did not alter the spatial
There was no significant departure from Hardy-Weinberg Equilibrium. All pairs of loci across all four clusters were in accordance with linkage equilibrium, except for one pair in one population (Table 4).

Figure 7. Geographic range of four genetic clusters. Top: each dot corresponds to GPS coordinates of a sampling site where an individual was scored at minimum one locus (n=81; 261 total individuals). Dot colors denote a distinct genetic cluster. A colored polygon was drawn around the outermost colored dots for each of the four colors. Each colored polygon represents the geographic range of a cluster. Bottom: STRUCTURE plot showing genetic clusters. Y-axis is Q value measuring likelihood an individual belongs to one of four clusters, which are denoted by colors that correspond to colored polygons on map above. Each bar on the X-axis is an individual (n=261). Individuals are grouped by Q value of their predominant cluster (i.e. the color most predominantly displayed for each individual).
Table 4. Results of linkage disequilibrium test as conducted in GENEPOP. A pairwise test of each locus combination in each of the four clusters as determined by STRUCTURE was performed. P-values were not significant (ns) unless otherwise noted.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Locus 1</th>
<th>Locus 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>Cm79332</td>
<td>Cm81520</td>
<td>ns</td>
</tr>
<tr>
<td>North</td>
<td>Cm79332</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>North</td>
<td>Cm81520</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>East</td>
<td>Cm79332</td>
<td>Cm81520</td>
<td>ns</td>
</tr>
<tr>
<td>East</td>
<td>Cm79332</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>East</td>
<td>Cm81520</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>Central</td>
<td>Cm79332</td>
<td>Cm81520</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Central</td>
<td>Cm79332</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>Central</td>
<td>Cm81520</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>South</td>
<td>Cm79332</td>
<td>Cm81520</td>
<td>ns</td>
</tr>
<tr>
<td>South</td>
<td>Cm79332</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
<tr>
<td>South</td>
<td>Cm81520</td>
<td>Cm114691</td>
<td>ns</td>
</tr>
</tbody>
</table>

Likely refugia and recolonization pathways.

Analyses aimed at inferring likely refugia and recolonization pathways were inconclusive. The posterior probability of any scenario could not be meaningfully distinguished from any other. No one scenario could be considered a likely representation of population nor dismissed (Figure 7).

Results of ecological niche modeling (Figure 8) showed present habitat suitability to match current known distribution of *C. maculata*, and suggested suitable habitat during the LGM
would have been found in the south and southeastern most parts of currently suitable habitats (Figure 9).

![Figure 8. Approximate Bayesian computation test of four competing colonization scenarios. Posterior probability of each of the four scenarios following four million simulations in DIY-ABC is displayed. Prior to the simulations, each scenario had an equal 0.25 posterior probability.]

![Figure 9. Ecological niche models of habitat suitability generated by MAXENT. Left are current climate conditions and right are climate conditions during the LGM. Suitability is displayed as a continuum where blue is an area of low habitat suitability and red is an area of high habitat suitability. Intermediate suitability areas are shown with middle colors on the heat map.]

25
DISCUSSION

Comparison of genetic population locations.

This study identified four distinct population clusters of *C. maculata* in the Southern Appalachians: North, South, Central, and East (Figure 7). Multiple clusters were expected, as population substructure is evident in nearly all phylogeographic studies conducted in the Southern Appalachians to date. Though no recolonization scenario separated itself from the other equally likely scenarios (Figure 8), the presence of multiple populations suggests that some event in the past, and/or environmental factors, have inhibited gene flow among *C. maculata* throughout its distribution.

Two neighboring populations of *C. maculata*, Central and East, met at the Great Smoky Mountains (Figure 10A). Analysis of mitochondrial and nuclear DNA in *Pseudotriton ruber* and *Gyrinophilus porphyriticus* salamanders (Figure 10B,C; Folt et al. 2016; Kuchta et al. 2016) and *Diadophis punctatus* snakes (Figure 10F; Fontanella et al. 2008) demonstrated the existence of a similar division, as did karyotype analysis of *Cryptocercus punctulatus* wood roaches (Figure 10G; Nalepa et al. 2002). Mitochondrial DNA alone revealed this same pattern in *Neotoma* woodrats (Figure 10E; Hays and Harrison 1992). Mitochondrial DNA, nuclear DNA, and ribosomal RNA analysis uncovered genetic structure of *Fumonta deprehendor* and *Sabacon*
cavicolens spiders within the Great Smoky Mountains on a much smaller geographic scale than the present study (Figure 10H, I; Thomas and Hedin 2008; Hedin and McCormack in press). Nuclear microsatellite markers were used in the present project and pointed toward some similarities in geographically neighboring yet genetically distinct populations in and around Great Smoky Mountains National Park.


Most interesting, however, are the parallels between results of the present study and those from Gonzales et al.’s (2008) study of the understory plant *Trillium cuneatum* (Figure 10D). The
latter study was previously the only examination of an understory plant in the Southern Appalachians. *C. maculata* was found to exist in several distinct population clusters in much the same way as haplotype clusters of *T. cuneatum* did. For both species, the Great Smoky Mountains were an area where edges of distinct populations met. Gonzales *et al.* (2008) pooled closely related chloroplast DNA haplotypes together into larger groups. Similar grouping was used in the present study when initial STRUCTURE and *F*<sub>st</sub> analyses independently suggested the North and East cluster were sister populations (see Methods). Though this spatial clustering has been observed in two different understory plants, it remains unclear as to why the Great Smoky Mountains are geographically coincident with the location of population differentiation.

*Past habitat suitability.*

Estimation of areas of past habitat suitability inferred through ecological niche modeling for *C. maculata* were consistent with models generated for two co-occurring species, the southern leopard frog *Rana sphenophalala* (Newman and Rissler 2007) and the millipede *Narceus americanus* (Walker *et al.* 2009; Figure 11). The three models all showed habitat suitability to the south and east of the Appalachian Mountains during the LGM. In all cases, past habitat suitability projections restricted the focal species to a small subset of its current distribution. Based on results from the other studies, it is possible that the area of suitable habitat during the LGM that overlapped with present day distribution acted as a refuge for *C. maculata* (Figure 11).
Figure 11: Comparison of past estimates of habitat suitability. A) Results from this study. Circled area is approximate area of past suitability that overlaps with current distribution of *C. maculata*. B) Model as it appeared in Walker et al. (2009). C) Model as it appeared in Newman and Rissler (2007). All black or colored area was deemed a suitable habitat.

*Future directions informed by results from present study.*

The results presented here suggest that *C. maculata* once persisted in a restricted habitat, and that it currently exists in four genetic populations in the Southern Appalachians. Surprisingly, *C. maculata* showed higher allelic richness in the north (Figure 6), in what would have been a much cooler climate than the south during the LGM. Results of ecological niche modeling also seem indicate an affiliation for cooler temperatures, as it seems the range of suitable habitat has shifted northward since the LGM (Figure 9). This study was unable, however, to statistically distinguish between alternative scenarios relating to where a past refuge or refuges were, and what recolonization routes, if any, were used. Spatial genetic clustering and past habitat suitability informed though ecological niche modeling provided a framework for follow-up studies to answer unresolved questions.
It is possible that a refuge area was not identified from the genetic data because it was not represented in the present sampling. Overlapping areas of past suitability and current distribution (Figure 11) were not sampled because those areas fall to the east of the Appalachian Mountains. The assumption that a refuge population must have existed in the Appalachians may have been incorrect. Present-day populations may all have originated from the interior coastal plains and Fall Line Hills (aka “Sandhills” in South Carolina). Expansion of sampling to that area could provide greater insight into refugia location(s) and subsequent recolonization pathways. Also, three microsatellite loci are few relative to the number of genetic markers used in other phylogeographic studies. They were successful in uncovering genetic structure, but may not have shown enough power to identify a recolonization scenario or association between allelic richness and latitude.

Based on information obtained during development of species-specific microsatellites, a high attrition rate (only 9.7% of primer pairs tried were successful) suggests that these other types of molecular markers should be considered. An alternative way to identify the most likely recolonization scenario could be the incorporation of single nucleotide polymorphism (SNP) markers. SNP data informed through next generation sequencing has emerged as a lower-cost, larger-output alternative to microsatellites in phylogeographic studies and can even be more successful at teasing apart complex population histories (Morin et al. 2004; Coates et al. 2009; Emerson et al. 2010; Bradbury et al. 2015). Microsatellites have a faster mutation rate, which makes them more viable candidates for determining recent changes cross a species' range (i.e. after the LGM) but SNPs could be used in conjunction with the existing microsatellite data presented here.


APPENDIX
Table 5. GPS coordinates and elevations of sample sites. Coordinates are recording in decimal degrees (dd) and elevation in meters (m).

<table>
<thead>
<tr>
<th>Location ID</th>
<th>Specimens Collected</th>
<th>Latitude (dd)</th>
<th>Longitude (dd)</th>
<th>Elevation (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A18</td>
<td>3</td>
<td>34.87866</td>
<td>-84.71337</td>
<td>354</td>
</tr>
<tr>
<td>A57</td>
<td>1</td>
<td>35.77140</td>
<td>-83.21343</td>
<td>575</td>
</tr>
<tr>
<td>A41</td>
<td>1</td>
<td>35.91838</td>
<td>-83.49964</td>
<td>257</td>
</tr>
<tr>
<td>A43</td>
<td>4</td>
<td>35.08652</td>
<td>-84.98366</td>
<td>472</td>
</tr>
<tr>
<td>A44</td>
<td>4</td>
<td>36.31100</td>
<td>-82.07211</td>
<td>648</td>
</tr>
<tr>
<td>A45</td>
<td>4</td>
<td>36.20811</td>
<td>-82.07604</td>
<td>1126</td>
</tr>
<tr>
<td>A46</td>
<td>4</td>
<td>36.22344</td>
<td>-82.10981</td>
<td>1123</td>
</tr>
<tr>
<td>A47</td>
<td>4</td>
<td>36.17875</td>
<td>-82.10726</td>
<td>1026</td>
</tr>
<tr>
<td>A48</td>
<td>2</td>
<td>36.15563</td>
<td>-82.25122</td>
<td>886</td>
</tr>
<tr>
<td>A50</td>
<td>4</td>
<td>36.64575</td>
<td>-81.74012</td>
<td>732</td>
</tr>
<tr>
<td>A51</td>
<td>4</td>
<td>37.73672</td>
<td>-81.43256</td>
<td>874</td>
</tr>
<tr>
<td>A53</td>
<td>4</td>
<td>37.40959</td>
<td>-79.80718</td>
<td>662</td>
</tr>
<tr>
<td>A54</td>
<td>4</td>
<td>37.66844</td>
<td>-79.69961</td>
<td>651</td>
</tr>
<tr>
<td>A55</td>
<td>3</td>
<td>37.52083</td>
<td>-79.72324</td>
<td>476</td>
</tr>
<tr>
<td>A56</td>
<td>4</td>
<td>38.12502</td>
<td>-78.78368</td>
<td>814</td>
</tr>
<tr>
<td>A57</td>
<td>4</td>
<td>38.17707</td>
<td>-78.66551</td>
<td>835</td>
</tr>
<tr>
<td>A58</td>
<td>2</td>
<td>38.76103</td>
<td>-78.28269</td>
<td>1051</td>
</tr>
<tr>
<td>A59</td>
<td>4</td>
<td>38.55696</td>
<td>-78.37990</td>
<td>1047</td>
</tr>
<tr>
<td>A60</td>
<td>4</td>
<td>38.32792</td>
<td>-79.27944</td>
<td>723</td>
</tr>
<tr>
<td>A61</td>
<td>4</td>
<td>38.22080</td>
<td>-79.41093</td>
<td>657</td>
</tr>
<tr>
<td>A62</td>
<td>4</td>
<td>38.07403</td>
<td>-79.44785</td>
<td>669</td>
</tr>
<tr>
<td>A63</td>
<td>4</td>
<td>38.11541</td>
<td>-79.34189</td>
<td>573</td>
</tr>
<tr>
<td>A64</td>
<td>4</td>
<td>38.04004</td>
<td>-79.34949</td>
<td>764</td>
</tr>
<tr>
<td>A65</td>
<td>1</td>
<td>37.27274</td>
<td>-80.10394</td>
<td>826</td>
</tr>
<tr>
<td>A66</td>
<td>2</td>
<td>36.16141</td>
<td>-80.91481</td>
<td>527</td>
</tr>
<tr>
<td>A67</td>
<td>4</td>
<td>36.12452</td>
<td>-80.74478</td>
<td>378</td>
</tr>
<tr>
<td>A68</td>
<td>4</td>
<td>36.01002</td>
<td>-80.63088</td>
<td>407</td>
</tr>
<tr>
<td>A69</td>
<td>4</td>
<td>35.65133</td>
<td>-83.43462</td>
<td>313</td>
</tr>
<tr>
<td>A70</td>
<td>1</td>
<td>36.05814</td>
<td>-83.67426</td>
<td>564</td>
</tr>
<tr>
<td>A71</td>
<td>4</td>
<td>36.03148</td>
<td>-83.67422</td>
<td>425</td>
</tr>
<tr>
<td>A72</td>
<td>4</td>
<td>36.03848</td>
<td>-83.63612</td>
<td>416</td>
</tr>
<tr>
<td>A73</td>
<td>4</td>
<td>35.23991</td>
<td>-84.54751</td>
<td>250</td>
</tr>
<tr>
<td>A74</td>
<td>4</td>
<td>35.34883</td>
<td>-84.24767</td>
<td>327</td>
</tr>
<tr>
<td>A75</td>
<td>4</td>
<td>35.34356</td>
<td>-84.19383</td>
<td>425</td>
</tr>
<tr>
<td>A76</td>
<td>3</td>
<td>35.32614</td>
<td>-84.07657</td>
<td>1139</td>
</tr>
<tr>
<td>A77</td>
<td>4</td>
<td>35.27122</td>
<td>-83.08566</td>
<td>683</td>
</tr>
<tr>
<td>A78</td>
<td>4</td>
<td>35.48535</td>
<td>-83.13284</td>
<td>1449</td>
</tr>
<tr>
<td>A79</td>
<td>4</td>
<td>35.59535</td>
<td>-82.48742</td>
<td>722</td>
</tr>
<tr>
<td>A80</td>
<td>4</td>
<td>37.91938</td>
<td>-80.27498</td>
<td>614</td>
</tr>
<tr>
<td>A81</td>
<td>4</td>
<td>38.82374</td>
<td>-79.38618</td>
<td>528</td>
</tr>
<tr>
<td>A82</td>
<td>2</td>
<td>38.82395</td>
<td>-79.38306</td>
<td>548</td>
</tr>
</tbody>
</table>
VITA

**Education**

2014  BS Biology, Towson University, Towson MD

**Honors and Awards**

2016  University of Mississippi Graduate Student Council Grant, University of Mississippi

**Posters**

2016  Society for the Study of Evolution, Austin TX
      John Banusiewicz and Dr. Ryan Garrick
      Historical Refuges and Postglacial Recolonization Routes in the Southern Appalachian Mountains, Inferred from Phylogeography of the Spotted Wintergreen (*Chimaphila maculata*)