Host constraints on the post-glacial migration history of the parasitic plant, *Epifagus virginiana*

by Yi-Hsin Erica Tsai

Department of Biology
Duke University

Date:_____

Approved:

_____
Paul S. Manos, Supervisor

_____
James S. Clark

_____
François M. Lutzoni

_____
William F. Morris

_____
William Wilson

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biology in the Graduate School
of Duke University

2009

Abstract

Host constraints on the post-glacial migration history of the parasitic plant, *Epifagus virginiana*

by Yi-Hsin Erica Tsai

Department of Biology
Duke University

Date:_____

Approved:

_____
Paul S. Manos, Supervisor

_____
James S. Clark

_____
François M. Lutzoni

_____
William F. Morris

_____
William Wilson

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Department of Biology in
the Graduate School
of Duke University

2009

Copyright by
Yi-Hsin Erica Tsai
2009

Abstract

Because species respond individually to climate change, understanding community assembly requires examination of multiple species from a diversity of forest niches. I present the post-glacial phylogeographic history of an understory, parasitic herb (*Epifagus virginiana*, beechdrop) that has an obligate and host specific relationship with a common eastern North American hardwood tree (*Fagus grandifolia*, American beech). The migration histories of the host and parasite are compared to elucidate potential limits on the parasite's range and to understand their responses to shared climate change. Two chloroplast DNA regions were sequenced and 9 microsatellite loci genotyped from parasite specimens collected throughout the host's range. These data were compared with available cpDNA sequences from the host (McLachlan et al. 2005) and host fossil pollen records from the last 21,000 years (Williams et al. 2004). Analyses of genetic diversity reveal high population differentiation in the parasite's southern range, a possible result of long term isolation within multiple southern glacial refuges. Estimates of migration rates and divergence times using Bayesian coalescent methods show the parasite initiating its post-glacial range expansion by migrating northward into the northeast from southern areas, then westward into the midwest, a pattern consistent with the development of high density beech forests. This result is strongly confirmed through spatial linear regression models, which show host density plays a significant role in structuring parasite populations, while the initial migration routes of the host are irrelevant to parasite colonization patterns. Host density is then used as a proxy for the parasite's habitat quality in an effort to identify the geographic locations of its migration corridors. Habitat cost models are parameterized through use of the parasite's genetic data, and

landscape path analyses based on the habitat map show a major migration corridor south of the Great Lakes connecting the northeast and midwest. Host density was the major determinant controlling the parasite's range expansion, suggesting a lag time between host and parasite colonization of new territory. Parasites and other highly specialized species may generally migrate slower due to their complex landscape requirements, resulting in disassociation of forest assemblages during these times. From these results, the low migration capacities of highly specialized species may be insufficient to outrun extirpation from their current ranges.

Dedication

To Ma, who made my journeys possible,

and to Da, who taught me to love math.

感謝

Table of Contents

Abstract vii

List of Tables xi

List of Figures xii

Acknowledgments xiii

1. Comparative phylogeography, meet spatial statistics 1

    Introduction 1

    Current phylogeographic methods for genetic data 2

    Current comparative phylogeographic methods 8

    Introducing a GIS framework for comparative phylogeography 12

    Case study: Comparative phylogeographies in host-parasite systems 14

    Next steps in comparative phylogeography 24

2. Host density drives the post-glacial dispersal of the tree parasite, *Epifagus virginiana* 26

    Materials and methods 34

3. Using spatial statistics and landscape genetic tools to identify the migration corridors of *Epifagus virginiana* 50

    Introduction 50

    Materials and Methods 54

    Results 60

    Discussion 65

    Conclusions and future directions 69

References 86

Biography 101

List of Tables

Table 2.1: Genetic diversity statistics for *Epifagus virginiana*. 42

Table 2.2: Population size, migration, and divergence time estimates between geographic regions of *Epifagus virginiana* 43

Table 2.3: Genetic diversity statistics for *Fagus grandifolia* 44

Table 2.4: Analyses of Molecular Variance (AMOVA) in *Epifagus virginiana* 45

Table 2.5: Results of linear regression models. 45

Table 3.1: *Epifagus virginiana* genetic data and global diversity statistics by locus 71

Table 3.2: Population demographic parameters relating to geographic regions of *Epifagus virginiana* 72

Table 3.3: Probabilities that a divergence time between regions of *Epifagus virginiana* is greater than another 72

Table 3.4: Probabilities that a migration rate from one region of *Epifagus virginiana* is greater than another 73

Table 3.5: Genetic diversity statistics for different geographic regions of *Epifagus virginiana* 73

Table 3.6: Univariate host pollen density time-slice models predicting genetic distances of *Epifagus virginiana* 74

Table 3.7: Best spatial regression models of host variables on *Epifagus virginiana* genetic distances 75

Table 3.8: Performance of habitat cost models based on *Fagus grandifolia* pollen densities at 500 ybp 76

List of Figures

Figure 2.1: *Epifagus virginiana.* 46

Figure 2.2: Areas of high genetic differentiation of *Epifagus virginiana.* 48

Figure 2.3: Linear models predicting the genetic distances of *Epifagus virginiana.* 49

Figure 3.1: Regional definitions used in the study of the population structure of *Epifagus virginiana* 78

Figure 3.2: Cluster phenogram of *Epifagus virginiana* genotypes from 9 microsatellite loci. ...79

Figure 3.3: Geographic distribution of *Epifagus virginiana* microsatellite clusters....80

Figure 3.4: Data layers used in the spatial linear models of the relationship between *Epifagus virginiana* and different aspects of its host, *Fagus grandifolia*, biology 81

Figure 3.5: Changes through time in regression coefficients of *Fagus grandifolia* pollen layers in linear models predicting genetic distances of *Epifagus virginiana* 83

Figure 3.6: Performance of various Fagus grandifolia habitat cost models measured by Mantel test correlation coefficients....84

Figure 3.7: Migration corridors of *Epifagus virginiana* based on the best habitat cost models of Fagus grandifolia pollen densities at 500 ybp....85

Acknowledgments

The Ph.D. is an impossible task if you go it alone. As I pause and reflect, I realize how very fortunate I am to have had so many people show me kindness, generosity, and friendship throughout this process. At the top of this list is my advisor, Paul Manos. Simply put, Paul is a great advisor. He's read 95% of all the paper I've pushed in front of him, an astonishing rate that I would dare any other faculty member to beat. He's been steadfastly encouraging and has inspired a sense of confidence in myself that can't be topped. I very much appreciate how he's let me explore my interests, pushed me when I needed it, and helped to craft a project that I am passionate about.

I would like to thank Paul and the rest of my committee, Jim Clark, Bill Morris, Will Wilson, and François Lutzoni for their many contributions to my thesis and career. I've been amazed by how their most offhand comments can then lead to 6 months of fruitful work. In addition to their helpful comments on manuscripts, proposals, and ideas, I would like to thank them for the many letters of recommendation that they have written for me. Without which, I am sure, I would never have received the funding for my thesis or for my upcoming post-doc. I would also like to thank Cliff Cunningham for our discussions on phylogeography, which contributed much to my intellectual development, and left me pumped up about my project. I must also thank Jennifer Swenson for introducing me to GIS and spatial ecological tools, which played an enormous role in my thesis. Also, thank-you to Jon Shaw for the helpful comments and unique insights he brought to my manuscripts.

I would also like to thank the generous organizations that funded my doctoral work: Duke University Graduate School, Duke University Dept. of Biology, Mellon Foundation (through a training grant in Plant/Fungal Systematics), Sigma Xi, American Society of Plant Taxonomists, Association for Women in Science, Sigma Delta Epsilon/Graduate Women in Science, Google, Deep Time, Botanical Society of America, and the National Science Foundation.

Thanks must also go to my collaborators. Dylan Burge graciously introduced me to *Ceanothus*, shared his data, and got me thinking about how techniques developed in my thesis could be applied more broadly. David Kidd mentored me throughout the PhyloGeoViz project and shared many helpful discussions about phylogeographic techniques and visualizations. Jackson Fox's debugging abilities were invaluable in creating PhyloGeoViz and in many of the programming aspects of my thesis. I also owe thanks to the undergraduates who have passed through the lab, Kaila Davis, Claire Landgren, Billy Schnackel, and Amos Little, who have contributed many hours and sequences to this project.

I would like to thank the many specimen collectors of *Epifagus*: Robyn Burnham, Alison Colwell, Mark Fishbein, Jackson Fox, John Freudenstein, Fritz Gerhardt, Sally Gerhardt, Philip Hyatt, Gelyn Kline, Ross McCauley, Jason McLachlan, Jordan Metzgar, Dan Nickrent, Heather Nickrent, Paul Manos, Sasa Stefanovic, Erin Tripp, Yi-Show Tsai, Maggie Whitson, and Chuck Williams. In particular, I'd like to highlight the contribution of my mom, Yi-Show. She traveled with me 15,000+ miles around the continent, braved snow, rain, and fall-out from Hurricane Katrina, was a far better field biologist than me, and was a great travel companion. Furthermore, thanks to Brad Steinberg, Kimberley Hollier, Mary

"Emma" Debenport, Jim Robinson, Richard Broadwell, and other staff at Algonquin Provincial Park, the Louisiana State Arboretum, Umstead State Park, and the Duke Forest for granting me permits, helping me scope out field sites, and for making these enjoyable places to work.

Duke University has been a great place to work in part due to the fabulous staff. I'd like to thank Jim Tunney and Jill Danforth for their excellence in accounting, and Randy Smith and Anne Lacey for keeping the department running smoothly. Thanks also to Lisa Bukovnik and Graham Alexander of the IGSP Sequencing Facility for producing much of the raw data for the project, facilitating the robotic work that saved my hands and back, and expediting my submissions at the end. I also owe big thanks to Sandy Boles, who thought through many laboratory challenges with me, helped design more efficient projects, and shared so many useful tips and tricks of the lab. In the department's IT group, Andrew Turnier, Kevin Campbell, and Hunter Matthews helped solve many of my computing issues, and the OIT help desk always provided prompt and courteous service. I am indebted to Tom Milledge and the staff of the Scalable Computing Support Center who provided over-the-top service for the Duke Shared Cluster Resource, which was used extensively in this thesis. Also, the statistical consulting service offered by the Department of Statistics was immensely useful, particularly in devising some of the simple statistical tests performed in this thesis. Thanks also go to the staff of the Plant Teaching and Research Facility, especially Beverly Calhoun and Marcia Kirinus, who kept my plants alive though that project never fully matured.

Perhaps the greatest influences on my thesis work came from graduate students and post-docs that I was fortunate enough to overlap with. Foremost is Jason McLachlan, who

from day one took me under his wing and introduced me to graduate school. In many ways, it was due to his work on beech that I was able to turn a small species specific phylogeography into something much more interesting. Lab-mates Dylan Burge, Norm Douglas, Jay Horn, Dudu Meireles, Sang-Hun Oh, Taina Price, Tanya Rehse, Erin Tripp, and Kyle Williams helped in my project's development and kept the fun alive in the lab. I'll never forget the day we discovered a chicken in the lab freezer. Thanks also to Norm, Dylan, James Beck, Kyle Dexter, Christy Henzler, Nathalie Nagalingum, Mike Nowak, and Lisa Pokorny for their many stimulating conversations about phylogeographic methods, interesting biogeographic distributions, and phylogenetics. Many thanks to Sara Chun for explaining concepts in landscape ecology and GIS, and to Michelle Hersh for help in modeling and understanding ecology.

I must also thank my great friends that have made this a most enjoyable journey: James Beck, Moriah Beck, Dylan Burge, Audrey Chang, Sara Chun, Arielle Cooley, Julie DeMeester, Stephanie Diezmann, Norm Douglas, Paul Gong, Nat Grier, Michelle Hersh, Jerry Hsu, Suzanne Joneson, Melissa Kenney, Heidi Koschwanez, Phoebe Lee, Young Wha Lee, Nathalie Nagalingum, Sang-Hun Oh, Tong Ren, Carl Salk, Ariana Sutton-Grier, and Alexandra Tobler. Together, we completed many Department of Biology Triathlons (running, pushups, and gossip). And a special thanks to my best buds Michelle and Sara for our many rants and gyro Thursdays!

Thanks also to the Tsai and Fox families: Li-Ching, Yi-Show, and Janet Tsai, John, Jane, Jackson, and Mitch Fox. Everyone contributed something to my thesis. My mom could spot *Epifagus* from a mile away; my dad and I discussed the applications of circuit theory to ecology; and my sister read several drafts and was my biggest cheerleader. The

Foxes provided so much care and affection; the occasional surprise care package was very much needed and appreciated. And of course, thank you to my biggest supporter, Jack. Jack has sat through hours of practice talks, read through hundreds of pages of drafts, and loved me. I couldn't ask for anything more.

Last, I would like to thank the many teachers I've had throughout my life. Teachers have such power to inspire students, and they often have no idea they've had such a positive impact. The following have encouraged me to dream big, believed in my capabilities, and generally made the world a better place: Elaine Schuhrke (math, 6th grade), Marty Klein (math, 7th), Liz McGrew (English/German, 7th-8th), Steve Sayers (math, 9th-11th), Judy Halvorson (biology, 9th-12th), Robert Gaines (math, 12th), Angela Johnson (School of Education, Univ. of Colorado), Veronica Vaida (Dept. of Chemistry, Univ. of Colorado), Jeff Mitton (Dept. of EPO Biology, Univ. of Colorado), and, most of all, Paul Manos.

1. Comparative phylogeography, meet spatial statistics

Introduction

To understand what processes have shaped community assembly in the face of climate change, ecologists and biogeographers have looked to the post-glacial histories of the modern temperate forests as examples of how species respond to global warming (Davis and Zabinski 1992; Williams et al. 2004; Williams and Jackson 2007; Provan and Bennett 2008). The migration histories of a wide range of taxa have been investigated from multiple perspectives using independent tools such as fossil pollen data and genetic data. While each of these datasets is compelling and reveals an aspect of the migration history of a species, alone each is incomplete. For instance, fossil pollen data can precisely locate high density populations at multiple time steps but can miss the presence of small populations (McLachlan and Clark 2004). In contrast, genetic data can identify glacial refuges and colonization routes even from low density source populations, but timing these events in plants can be difficult due to slow mutation rates and the recent timescale (Schaal and Olsen 2000). Combining these datasets provides a comprehensive view of a species migration history: where the initial migration routes were, and how the abundance of the species developed. These data are increasingly available, enabling the construction of fuller migration histories. However, syntheses of these datasets have been limited due to the multidisciplinary nature of the data and the lack of accessible analytical tools.

The challenges facing multispecies phylogeography are broader than how to combine different types of data. Methods must also be developed to compare patterns across different species as well as datatypes. The field of phylogeography is rapidly expanding with 4600+ papers published this century using genetic tools (Web of KnowledgeISI Web of

Knowledge, topic search for 'phylogeography'). While data on individual species are increasing rapidly, these data are far more valuable scientifically when they can be synthesized and leveraged to address key ecological and evolutionary questions. While species respond individually to climate change (Davis 1983) and communities are not cohesive through time (Williams et al. 2004), they often are constrained by shared climate, geographical, or ecological barriers so that similar species associations may reassemble in new spaces. Commonalities between the species' histories could denote important biogeographical boundaries or point to shared glacial refuges. These possible long-term associations tell of the stability of the community and can reveal the time scale of interactions between species.

This paper's goals are to broadly review the current methods in phylogeography, to summarize previous comparative or regional studies, and to introduce spatial methods developed to synthesize patterns across different datasets and species. Following the methodological review, I explore applications to host-parasite comparative phylogeographies, describe a new host tree and parasitic plant model system, and discuss its utility in understanding post-glacial community assembly.

Current phylogeographic methods for genetic data

Summary statistics of genetic diversity

The field of phylogeography grew out of population genetics and has a long history of relying on various summary statistics to describe population structure, diversity, and differentiation. For instance, measures of population structure based on F_{ST} (Weir 1996), analysis of molecular variance (AMOVA) (Excoffier et al. 1992), or genetic distances (Nei

and Li 1979; Reynolds et al. 1983) have been used to quantify population differentiation, to calculate migration rates, or to estimate divergence times. Other measures of genetic diversity, such as H_T and V_T, can be used to identify refuges and suture zones. Refuges are assumed to be relatively stable populations that would allow accumulation of high allelic richness, though those alleles would be closely related (high H_T but low V_T) (Pons and Petit 1996; Petit et al. 2002). These summary statistics often form the basis of other analyses examining broader geographic patterns. For example, Dyer and Nason (2004) used a graph theoretic framework based on pairwise population distances to identify populations important in maintaining the genetic connectivity of the landscape. Conversely, Monmonier's analysis spatially locates genetic breaks by considering which adjacent populations have the highest pairwise genetic distances (Manni et al. 2004). Lastly, genetic distances can be interpolated to form map layers that easily visualize spatial patterns in population differentiation (Miller 2005; Miller et al. 2006). While some information is certainly lost by summarizing genotypes and their frequencies into a few descriptive measures, the summary statistics are a simple and straightforward way to search for broader geographic patterns.

In addition to some data loss by using summary statistics, a drawback of their use is that oftentimes the geographic regions must be defined a priori for comparison. In two or three dimensional systems, it can be difficult to divide the landscape into objective areas, and some results may be sensitive to regional definitions. For instance, in many temperate post-glacial studies, a decline in genetic diversity is expected to accompany increases in latitude due to bottlenecks and founder events that may have occurred during recolonization (Hewitt 1996; Ibrahim et al. 1996; Hewitt 2000). This signal may be lost if populations are binned

into faulty regional classifications. For instance, if many northern populations are misincluded within the southern region, then measures of haplotype diversity may be artificially low and less different from values in the "northern" region. Analyses that require these types of a priori definitions could benefit from the development of sensitivity tests and better ways to assess robustness to regional boundaries. This recommendation also applies to analyses that more explicitly account for space and require no a priori definitions. For example, even though Monmonier's method finds barriers anywhere on the landscape without regard to regional definitions, confidence in its results must still be assessed. A cross-validation method could explore the robustness of the genetic barriers to the sampling method (chapter 2). In short, the resulting barriers are validated by subsampling the genetic data by locality. The importance of particular sample sites can be understood, and it results in a more accurate measure of confidence.

Nested clade analysis

Many non-summary statistic approaches have been developed to more fully utilize the genetic datasets in phylogeographic analyses. One of the first methods to explicitly incorporate phylogenetic relationships among genotypes into phylogeographic inferences was nested clade analysis (NCA) (Templeton 1998). It was reasoned that certain patterns of allelic relationships and their geographic distributions were the result of specific demographic histories such as range expansion, allopatric fragmentation, long distance dispersal, etc., and an inference key was provided that led the user from phylogenetic and geographic patterns to an associated demographic history (Templeton 1998). The inference key was somewhat intuitive for certain patterns; for example, if closely related alleles at the tips of the haplotype network were distributed across a wide geographic range, an inference

of range expansion was returned (Templeton 1998). However, this method, especially the inference key, has not survived rigorous testing; several studies using simulated data have shown a high rate of false positives, where incorrect inferences were returned in over 75% of datasets (Knowles and Maddison 2002; Panchal and Beaumont 2007). Several authors have called on the community to abandon this method altogether (Beaumont and Panchal 2008; Petit 2008; Knowles 2009). Even so, proponents of the method have continued to endorse its use claiming that the rate of false positives will decrease sharply with the incorporation of more loci, that the simulations were conducted on biologically improbable datasets that had implemented old versions of the inference key (which the authors dispute (Beaumont and Panchal 2008)), and that tests with empirical data indicate a much lower fraction of failures (Templeton 2004; Templeton 2008; Templeton 2009).

While more loci will likely improve the accuracy of NCA as well as many other phylogeographic methods, the tests performed by Templeton (2004) that show high agreement between NCA results and a priori expectations do not adequately refute NCA criticisms. First, this result does not address the concern of false positives or error rates (Knowles 2009), but rather shows that NCA can produce some "correct" results. Second, even the error rates estimated using Templeton's approach are very troubling at 23% (Templeton 2004). Improvements have been recently suggested (Templeton 2009), but not yet tested by other authors. A more moderate view is that while the conclusions of NCA cannot be trusted, they are still valuable as a method of hypothesis generation (Garrick et al. 2008). Furthermore, proponents decry the lack of single method alternatives and accept NCA's use until one is developed (Garrick et al. 2008). However, if users already have a priori expectations, then what additionally can NCA provide as a hypothesis generating

mechanism? Also, while there is no silver bullet method that stands as an alternative to NCA that claims to reach as far ranging conclusions, there are a plethora of methodologies available that test specific hypotheses of population history (ex. mismatch distributions, assignment tests, coalescent estimates of migration rates, etc.). Due to these concerns, the plurality view is to drop the use of NCA and instead embrace more targeted phylogeographic methods.

Phylogeographic methods that utilize genealogies

In addition to NCA, many other methods have been developed that incorporate allelic relationships in estimating phylogeographic histories. They can be divided into two main categories: methods that explicitly account for space, and methods that do not. The case that maximally utilizes phylogenetic and geographic information was developed by Lemmon and Lemmon (2008). Their approach estimates the geographic coordinates of ancestral nodes within the genealogy of sampled individuals using a maximum likelihood framework (Lemmon and Lemmon 2008). While a powerful analysis technique, this method has high requirements for its input data that many plant systems will find challenging to meet. For instance, a fully resolved intraspecific phylogeny based on a single locus is necessary for estimations, and the strength of the results largely depends on the support for the phylogeny, which can be very low in most plant systems. More generally, genealogies have been plotted on maps (without specific inferences on ancestral distributions) to visualize the relationships between clades and geography (Kidd and Ritchie 2006). Once again, this approach is usually more fruitful in animal systems where better supported, more highly resolved intraspecific phylogenies exist.

While not explicitly accounting for space, several methods utilizing gene genealogies and the coalescent model can still be informative about migration histories. These methods estimate a variety of demographic parameters such as effective population sizes, migration rates, and divergence times (Beerli and Felsenstein 1999; Beerli and Felsenstein 2001; Hey and Nielsen 2004; Beerli 2006; Hey and Nielsen 2007). In these analyses, the data are divided into groups ("populations") from which parameters are estimated. If the groups are geographically defined (versus by host race, species, etc.), then the migration rates, divergence times, etc. refer to geographic structure and processes. For example, Carstens et al. (2005) used this approach to test colonization scenarios between the northern Rocky Mountains, northern Cascades, and southern Cascades in a suite of species from the Pacific Northwest. By estimating divergence times between geographic areas, pre- and post-Pleistocene migration scenarios were distinguished (Carstens et al. 2005). Similarly, migration rates were calculated among human populations in Pakistan to understand the directionality of movement patterns and to assess the effects of altitude on dispersal (Faubet and Gaggiotti 2008). These population demographic parameters can reveal directionality of gene flow on the landscape, age of divergences between geographic regions, and changes in effective population size between regions. Once again, however, the parameter estimates are dependent on how the groups are defined, and where geographic boundaries are drawn. Studies need to carefully justify regional boundaries or show the robustness of conclusions to different regional definitions.

The a priori definition of populations, groups, or regions can limit the ability to discern population structure at the appropriate scale. Assignment test or cluster based methods forgo this requirement and ignore any metadata on geographic location,

morphological variety, etc. as it sorts genotypes into groups (ex. Pritchard et al. 2000; Corander et al. 2003; Corander and Marttinen 2006; Corander and Tang 2007). These groups (sometimes called "populations") represent the most cohesive genetic entities in the dataset. The population assignments are then compared to the actual metadata (geographic locations, taxonomic variety, etc.) in the search for biological meaning in the delimitations of these clusters. If the clusters correspond to geographic regions, then genetically connected and disconnected areas are identifiable. Furthermore, the hierarchical relationships among clusters can be used to inform relationships among geographic areas. For example, in a study of 1056 humans from across the world, division of individuals into 4 groups recognized clusters primarily composed of Africans, Europeans + Middle Easterners + Central/South Asians, East Asians, and Americans (Rosenberg et al. 2002). When restricting assignments to only 2 clusters, East Asians and Americans were grouped together versus all other individuals, indicating a closer relationship between East Asia and America (Rosenberg et al. 2002). This method is very flexible and benefits from additional loci. Spatial information can also be used to further inform the cluster assignments (Guillot et al. 2005; Chen et al. 2007; Corander et al. 2008).

Current comparative phylogeographic methods

Methods that synthesize single species phylogeographies

After identifying patterns in a species of interest, the next step is to look for commonalities among multiple species. The simplest comparisons involve consistent regional definitions that hold across species. For instance, in the example of ecosystem phylogeography in the Pacific Northwest described above, Carstens et al. (2005) examined

patterns of dispersal and vicariance in 6 species with similar distributions broken into 3 broad regions. The possible migration scenarios between regions for each species were the same, and each species analysis was treated separately. Because the number of scenarios was limited (3 mutually exclusive hypotheses), comparisons of results were very manageable and common patterns were straightforward to identify (Carstens et al. 2005). When many more scenarios are possible, it becomes increasingly difficult to rely on a priori hypothesis discrimination to identify shared patterns. For instance in southeastern North American phylogeographies, many more migration scenarios are possible based on rivers, mountains, and other discontinuities (Soltis et al. 2006). At this larger scale, finding taxa with similar distributions also becomes more challenging, and unless an extreme geographic boundary exists, species tend to have unique phylogeographic patterns (Davis 1983). Because of the broader scope of the southeastern North American study, Soltis et al. (2006) identified the major genetic boundaries for each taxon, then employed a Geographic Information Systems (GIS) framework to check for similarities in boundary locations. No boundaries withstood this test, and it was concluded that no one barrier appears to have structured most southeastern North American taxa (Soltis et al. 2006). However, certain geographic discontinuities were common among a subset of taxa. These were identified generally through visual comparisons of single species datasets (Soltis et al. 2006). A concern for this approach is that it is difficult to discriminate between true congruence of patterns and pseudocongruence, or similar patterns arising from different migration processes, without more closely examining each taxon's migration history (Cunningham and Collins 1994).

One method of addressing the issue of pseudocongruence in phylogeographic patterns is to examine the timing of divergences across shared geographic barriers. A

hierarchical Bayesian method can be used to test for simultaneous divergences in many taxa across the same geographic boundary (Hickerson et al. 2006; Hickerson et al. 2007). For example, 8 pairs of echinoid sister species are split in geographic distribution by the isthmus of Panama, and show similar patterns of differentiation across this boundary (Hickerson et al. 2006). However, is this a case of pseudocongruence? Is the timing of divergence the same for all pairs? Estimates of simultaneous vicariance of all sister pairs indicated that not all divergences occurred at the same time; in particular, one sister pair diverged later than the other 7, which did diverge simultaneously at 3.1 million years ago, corresponding with the rise of the Panamanian isthmus (Hickerson et al. 2006). This is a useful method for testing the timing of a hypothesized barrier's effects on multiple taxa. However, to use this method one must have an a priori barrier of interest to investigate.

In contrast to the approach of combining multiple individual species analyses, Petit et al. (2003) first averaged data across species for each region, then analyzed the conglomerate data for geographic patterns. For example, allelic richness was calculated per species per region, then averaged by species to form a measure of allelic richness "per forest". These values were visualized geographically and interpreted to show whole community patterns. From this analysis, the highest levels of allelic richness per forest were obtained at high latitudes, north of the hypothesized refugial peninsulas (Petit et al. 2003). This result was interpreted to show that the overwhelming pattern for many forest species was one of suture zones north of the mountains where lineages met from multiple refugia (Petit et al. 2003). This study had a major advantage over other comparative phylogeographies -- the specimens for all species were collected within the exact same localities. This approach is quite remarkable and allows for average forest diversity indices

to be calculated without need for false combinations of nearby populations or spatial interpolations. On the other hand, by limiting the analysis to only shared populations, many species specific data points may have been discarded that could resolve finer scale patterns. Also, methods should be developed to extend this approach to less carefully collected comparative studies, where it is of interest to combine data with different collecting schema and with different distributions. A possible solution is to use spatial interpolations to bring differently developed datasets into comparable formats (see below).

Combining multiple datatypes

While a multitude of phylogeographic methods for single and multi-species analyses exist, a vacancy remains for a quantitative method to compare genetic and non-genetic datatypes. Most commonly, phylogeographers will use fossil datasets to inform hypotheses that are then tested using genetic methods (Cruzan and Templeton 2000; Palme et al. 2003; Tiffney 2008). Assessing agreement between fossil and genetic datasets commonly relies on visual analyses, especially concerning the general location of glacial refuges. The genetic data often show different refuges than the fossil data, perhaps located closer to the ice margin (McLachlan et al. 2005; Sommer and Zachos 2009), and migration from these "cryptic" refuges could drastically alter estimates of a species' migration capacity (Provan and Bennett 2008). In addition to hypothesis formation, fossils can contribute by calibrating divergence time estimates within the phylogeography (Morris et al. 2008). Moreover, fossils can inform the timing of range expansions of genotypes. For instance, studies that combine molecular and fossil evidence often do so through assigning fossil deposits likely genetic haplotypes (McLachlan et al. 2005; Magri et al. 2006). The assumption is that post-glacial range contractions are few, and founder events were responsible for the present patterns of genetic

diversity (Magri et al. 2006). Through these assignments, the fossil and genetic data are linked, and the spread of genotypes can now be tracked through time.

Introducing a GIS framework for comparative phylogeography

Most of the phylogeographic methods described thus far have relied on a priori divisions of space and have focused on understanding genetic patterns across those spatial boundaries. The opposite approach should also be explored: Given genetic patterns, is it possible to identify where spatial divisions occur? And where are areas of connectivity or migration corridors? This approach requires an explicit accounting for space in models of demographic history. The first step is to translate the phylogeographic data into a geographic map layer. A straightforward method is to construct a landscape surface based on a genetic diversity statistic, such as F_{ST} or Nei and Li distances (Nei and Li 1979; Weir 1996). In this visualization, the X and Y are the geographic coordinates for the range of the species of interest. The Z axis reflects the genetic data at a particular XY coordinate; in this case, the genetic data are summarized by genetic distances. Pairwise genetic distances are assigned to the midpoints between localities. The rest of the surface is constructed by interpolation (Miller 2005; Storfer et al. 2007). This same procedure can be repeated for multiple datasets and multiple species. In this way, even datasets that were sampled differently can be converted to the same cell size and extent, resulting in comparable maps. Also, different datatypes can be easily accommodated as long as they too can be converted into a geographic map layer. Now fossils can be integrated into analyses directly. Additionally, these layers can be more informative than in previous treatments, because not only can they contain information on range limits at different time slices but also on density.

Once map layers have been constructed for each of the datasets and species of interest, a suite of map functions and map algebra are at our disposal. Explicit spatial comparisons among datasets can now be made on a grid cell by grid cell basis. Linear models are a natural outgrowth of map algebra and can be used to quantify relationships between datasets. In this way, we are less reliant on visual pattern recognition, and instead can use regression, correlation coefficients, or other spatial statistical methodologies to check for agreement between datasets. In addition to providing a framework for comparing multi-species and multi-type datasets, the map layers can be used as inputs in landscape path analyses, where the goal is to spatially locate migration corridors. The genetic distances can be used as a proxy for habitat resistances (inverse of habitat porosity, or how easy it is to disperse through the habitat), and the resulting least-cost-paths between populations show the most likely routes of gene flow. This GIS framework is a gateway to landscape ecological tools commonly used in conservation biology.

In implementing these map based methods, one must acknowledge the non-independence of the map data. In addition to possible underlying spatial processes, the maps are created through interpolations, which necessarily create non-independent data points, which could result in inflated confidence in misleading results. However, this effect can be ameliorated by explicitly including spatial dependence in the models. For instance, in linear models a spatial lag term should be included (Ward and Gleditsch 2008), and autocorrelations among residuals should be checked to ensure that spatial variation has been accounted for.

Case study: Comparative phylogeographies in host-parasite systems

The habitable areas of many plant species will shift dramatically during the twenty-first century (Davis and Zabinski 1992; Iverson and Prasad 2002). As their distributions change and plants migrate to these new places, do forest communities act as cohesive units, or will species move independently and set up new associations? Much research on tree species have shown that they generally respond independently to climate change (Davis 1983; Comes and Kadereit 1998), and this suggests that communities are transient. However, little attention has been given to understory species that are dependent on trees for habitat or resources. Will species that have dependent associations with trees, such as parasites, have dependent migration routes? By studying the migration patterns of hosts and parasites, we can understand how the range of a species changes relative to that of a close associate, which is central to understanding the cohesiveness of communities and how assemblages of organisms move.

Parasites and other symbionts are increasingly used to track the migration histories of their hosts, because they often provide higher resolution patterns courtesy of shorter generation times and faster evolving genomes (Wirth et al. 2005; Nieberding and Olivieri 2007). Successful studies have mined the genomes of gut bacteria (Wirth et al. 2005), lice (Nadler et al. 1990), and other symbionts (Criscione and Blouin 2007) to trace complex human and other animal migratory patterns. But this strategy of using parasites as proxies has not yet caught on in the botanical world, where parasitism has evolved at least 11 times across many diverse angiosperm families (Barkman et al. 2007). Plant phylogeography has much to gain from this approach due to the challenges of slower mutation rates and subsequent low resolution of migration routes.

Population structure in parasitic plant systems

Previous studies of parasitic plant systems have focused on identifying host race formation and any major phylogeographical breaks between regions. Results have been largely mixed, with many studies showing genetic differentiation by host, fewer showing geographic patterns to the variation, and some showing no patterns discernable by host or geography. For instance, several species of mistletoes, *Arceuthobium vaginatum* and *A. americanum* (Santalaceae), showed differentiation at several allozyme loci dependent on what host the specimens were found infesting, *Pinus contorta* or *P. ponderosa* (Linhart et al. 2003). Similarly, another mistletoe, *Viscum album* (Santalaceae), was found with differentiated cpDNA based on its host, *Abies*, *Pinus*, or angiosperms, and within the host races further cryptic species were suggested within geographically distinct Turkish populations (Zuber and Widmer 2009). Other parasites with broad host ranges were not found to be structured along host lines. For example, taxa as divergent as *Pedicularis chamissonis* (Scrophulariaceae) in New Zealand and *Dactylanthus taylorii* (Balanophoraceae) from Japan instead showed substantial population structuring by rough north-south or east-west divisions (Fujii et al. 1997; Holzapfel et al. 2002). Some finer scale geographic patterns were found within the southern clade of *P. chamissonis* (Fujii et al. 1997), but *D. taylorii* did not show any relationship between geographic and genetic distances within its eastern or western ranges (Holzapfel et al. 2002). Most studies of parasitic species found some population differentiation that was either structured into host races or geographic entities.

A few comparative studies have been conducted on the population structure between parasitic plants and their hosts. Once again, a wide range of results were recovered, though geographic structuring of populations was unusual. For instance, the parasite *Cuscuta*

europaea (Convolvulaceae) had very high levels of gene flow, and its populations were less differentiated than its host, *Urtica dioica* (Urticaceae) (Mutikainen and Koskela 2002). No host race formation was recovered, and no correlation was found between host and parasite genetic distances (Mutikainen and Koskela 2002). On the other hand, a study of *Arceuthobium americanum* and its *Pinus* host species revealed varieties of the parasite structured by host (host race formation) (Jerome and Ford 2002; Jerome and Ford 2002). The parasite was more finely structured into three races versus the host into two groups, perhaps due to its more restrictive pollination and dispersal biology, but once again there was no relationship between geographic and genetic distances (Jerome and Ford 2002). Similarly, no isolation by distance was found in *Cytinus* (Cytinaceae), a Mediterranean holoparasite that parasitizes members of the Cistaceae, the rock-rose family (de Vega et al. 2008). However, extreme cases of host specialization and race formation were found in *Cytinus*. High levels of genetic differentiation were found based not only on host species, but also were related to the relationships between the hosts (de Vega et al. 2008). Parasite races were structured based on the infrageneric relationships of the host species, likely showing a true case of cospeciation or at least co-differentiation (de Vega et al. 2008).

Introduction to the Fagus-Epifagus system: The parasite, Epifagus virginiana

Epifagus virginiana (beechdrop) is a highly specialized parasitic plant that is commonly found on the roots of *Fagus grandifolia* (American beech). The herbaceous plant is a highly derived member of the Orobanchaceae (the broomrapes) (Bennett and Mathews 2006; Park et al. 2008). The Orobanchaceae is comprised of 89 genera and 2061 species with a worldwide distribution, though it is most commonly found in the temperate northern hemisphere and dry tropics of Africa (Bennett and Mathews 2006). Members of the

Orobanchaceae are exclusively root parasites (except for the non-parasitic *Lindenbergia*) and form a highly supported monophyletic group within the Scrophulariaceae sensu lato (Olmstead et al. 2001). *Epifagus* is closely related to other North American parasites, specifically *Conopholis* and *Boschniakia*, though *Epifagus* is very distinct with its annual habit, reduced haustorium (the parasitic organ), and slender, non-fleshy, stems (Wolfe et al. 2005; Park et al. 2008). Also, *Epifagus* is unique in that it is monotypic and highly host specific; it obligately parasitizes only the single species of *F. grandifolia*, except in a few isolated populations in Tamaulipas and Hidalgo, Mexico where it instead parasitizes the closely related *F. mexicana* (Thieret 1969). It is distributed across the entire range of *F. grandifolia* from eastern Wisconsin to Nova Scotia, south to Louisiana and northern Florida (Thieret 1969; USDA and NRCS 2009); it is usually very abundant when present but is patchily distributed.

In angiosperms, holoparasitism, or a complete heterotrophic lifestyle, is commonly accompanied by a suite of many morphological and genetic traits. *Epifagus* exemplifies many of these phenotypes. It is completely non-photosynthetic, possesses extremely reduced leaves and roots, and has chloroplasts with a drastically reduced thylakoid system (Walsh et al. 1980). Even so, there is some color variation in the stems (from cream to purplish-brown) that seems to be dependent on light availability and different chloroplast types (Schrenk 1894; Walsh et al. 1980), so there may be some remaining chloroplast functionality. Genetically, *Epifagus* has increased mutation rates, perhaps due to a lack of requirements for photosynthesis (Wolfe et al. 1992). Its chloroplast DNA retains only a fraction of the genes found in photosynthetic organisms, having lost most genes involved in photosynthesis, chlororespiration, and chloroplast specific RNA polymerases (dePamphilis and Palmer 1990;

Wolfe et al. 1992). Interestingly, however, the chloroplast does not appear entirely defunct with evidence for the selective preservation of certain open reading frames and tRNA genes (Wolfe et al. 1992; Wolfe et al. 1992; Lohan and Wolfe 1998). The high chloroplast mutation rate of *Epifagus* is reproduced in other parasitic lineages (Nickrent et al. 1997), where high mutation rates are not restricted to the chloroplast compartment but are also present in mitochondrial (Duff and Nickrent 1997) and nuclear DNA (Nickrent and Starr 1994).

The reproductive ecology of *Epifagus* is largely a story of selfing. *Epifagus* produces reddish-purple colored flowers and capsules along most of its stems and branches from July to November. While some flowers are chasmogamous, or open, in the upper portions of the stems, most seed set is a result of cleistogamous, or closed, flowers (Cooke and Schively 1904). With the high rate of cleistogamy, the necessity of pollinators seems negligible, and some plants never pierce the duff to even be exposed to pollination vectors (Leavitt 1902; Kuijt 1969). Nuclear microsatellite data confirm these observations with very few heterozygotes found (see Chapter 3). As a corollary, life below the duff likely ensures very limited dispersal of its small, dust-like, seeds. The seeds of *Epifagus* generally appear to be passively dispersed, with no elaisome to attract animal vectors, and being "slightly sticky" which discourages wind dispersal (Thieret 1969). The capsules open upward to form a "splash cup" morphology that may facilitate dispersal by water (Thieret 1969). The fleshy bodied sister genus to *Epifagus*, *Conopholis*, is thought to be an important source of food and nutrition for bears and other mammals (Seibert and Pelton 1994; Inman and Pelton 2002), and deer have been known to browse *Epifagus* in the autumn months (D Whigham, pers. comm.), so some unintentional mammal dispersal is not impossible. Seed germination in

Epifagus depends upon a thermolabile compound that is likely a host root exudate (Williams and Zuck 1986). A connection between parasite and host is formed early on, and the haustorium is likely constructed of recruited *host* tissue (Cooke and Schively 1904). *Epifagus* seems to parasitize fine, young host roots close to the soil surface, unlike *Conopholis* which forms thick haustoria on larger roots (Thieret 1969). The virulence of the parasitic infection is incredibly low with no known detrimental effects to the host, and the parasitism having no measureable effect on seedling survival (Musser 1979).

The study system: the host, Fagus grandifolia

Fagus grandifolia is a secondary succession species of the beech/maple climax community. It is fairly shade tolerant, so it can often invade under a thick forest canopy and sometimes spreads extensively through root suckering (Kitamura et al. 2001; Kitamura and Kawano 2001). *F. grandifolia* is most common in the Great Lakes region, the northeastern part of its range, and in the southern Appalachians. In the southern range, its abundance has been decreasing for the last 5000 years (Williams et al. 2004). Beech is limited by cold temperatures at its northern range and by lack of moisture at its western limits (Huntley et al. 1989), though in the upper peninsula of Michigan beech may be continuing to expand westwards (Woods and Davis 1989).

Fagus is a basal genus in the family Fagaceae (Manos and Steele 1997), comprised of beeches and oaks found mainly in the northern hemisphere with centers of diversity in southeast Asia and North America (Manos and Stanford 2001). Biogeographic studies suggest an Asian origin of this genus with subsequent dispersal to North America in the late Oligocene and early Miocene, followed by vicariance between continents (Manos and Stanford 2001). North America harbors two species of *Fagus*, *F. grandifolia* and *F. mexicana*,

with disjunct geographic distributions that diverged about 7 million years ago (Little 1965; Manos and Stanford 2001). Within *F. grandifolia* three varieties have been described (white, red, and gray beech) based on morphology, ecology, and geography (Camp 1950). These distinctions have not been thoroughly tested using genetic tools, but analyses based on cpDNA show many shared haplotypes across the range indicating high levels of gene flow and fertility across the geographic boundaries of the races described by Camp (1950), and many populations contain a mix of races (McLachlan et al. 2005). There is no evidence of discrimination of *Epifagus* based on host variety, although it is patchily distributed.

Fagus grandifolia is monoecious and flowers in late April to May (Fowells 1965). It primarily outcrosses and is wind pollinated, producing beechnuts that are dispersed by mammals or birds. Germination occurs in early spring to summer, first requiring cold stratification, and seedlings can establish in heavily shaded conditions (Fowells 1965). Beeches generally begin to be reproductive at age 40 and produce seed crops every 2-3 years (Fowells 1965). Roots develop fairly shallowly, and roots near the surface can be parasitized by *Epifagus*.

The post-glacial phylogeographic history of *F. grandifolia* has been extensively studied using both fossils and genetic data. At the last glacial maximum (~18 thousand years ago (kybp)), *F. grandifolia* was restricted to unglaciated territory south of the ice margin. Fossil pollen records suggest higher concentrations of beech in the southwestern part of the range in Mississippi (Davis 1983), and possibly in a refuge in northern Florida (Delcourt and Delcourt 1987). In addition to these suggested refuges, small deposits of *F. grandifolia* pollen were widespread during this period, suggesting diffuse, low density populations covering most of the range (Bennett 1985). Studies using cpDNA markers have identified refuges at

mid-latitudes in Indiana/Kentucky/Ohio that are closer to the ice margin and do not correspond to either of the refuges identified by high density pollen fossils (McLachlan et al. 2005), though see Morris et al. (2008). Following the retreat of the glaciers, beech fossil pollen became common first in the southeast (15 kybp), then across the entire south, peaking around 12 kybp (Williams et al. 2004). Beech density then increased dramatically in the northeast (9-7 kybp) and from there high density populations developed westwards into the midwestern and northern range (Williams et al. 2004). The genetic data show a different pattern, where a high diversity of northern haplotypes are found in the putative mid-latitude refuges. From there, the geographic distributions of the haplotypes trace northerly and northeasterly migration routes into the formerly glaciated territory (McLachlan et al. 2005; Morris et al. 2008). The two datasets can be reconciled by understanding that the paleo- and molecular data address different aspects of the migration history of beech. The fossil data are most informative in detecting high density populations, while the genetics are sensitive to population bottlenecks and genetic drift effects in small populations. Since the initial range expansion of a species likely occurs from low density populations at the range limits, the genetic data reveal the initial migration routes involved, while the fossil data more accurately describe the changes in abundance through space and time.

Estimates of migration rates for *F. grandifolia* differ dramatically based on the genetic or fossil pollen datasets. Based on fossils, rates as high as 300 m/yr were estimated (Davis 1981; Huntley et al. 1989), while rates as slow as 80 m/yr were recovered based on accounting for the genetic data (McLachlan et al. 2005). The accuracy of migration estimates is especially important when predicting the future migration capacity of this species. Numerous elevated CO_2 models predict the range of beech will shift radically northward into

Canada with its distribution in the USA shrinking by more than 90% due to moisture limitation in the south (Davis and Zabinski 1992; Iverson and Prasad 2002). Given these potential threats to beech and its importance in the northern hardwood communities, the study of beech and associated plants is especially timely.

What makes this system ideal?

The *Fagus-Epifagus* system provides a simple and contained example of a two species community with tightly associated members at two trophic levels, an idealized case for studying the factors controlling community assembly. Both the host and parasite are straightforward taxonomic entities with no records of hybridization with sympatric species or any other phylogenetic issues. Furthermore, the extensive data describing the migration history of *F. grandifolia* is an advantage that can be used to construct testable hypotheses concerning the range expansion of *Epifagus*. By comparing the population genetic structure of *Epifagus* against the host's initial migration routes from the genetic data and the temporal and spatial changes in host density through the fossil pollen record, the effect of host density on parasite range expansion can be evaluated. Furthermore, the host data provide an objective method to creating a priori regional boundaries, an issue that plagues most phylogeographies. For instance, instead of assigning populations to regions subjectively by eye, regions can be defined by host range distributions at different time slices in the host pollen data. So instead of "south" and "north" areas with unjustified distinctions, areas are labeled 13 kybp, 9 kybp, and 6 kybp based on the arrival time of host populations to those areas.

Critical to studying the phylogeography of *Epifagus* is the ability to identify geographically structured genetic variation. Several of the life history characters of *Epifagus*

strongly suggest that parasite populations should be highly differentiated. For instance, because *Epifagus* largely relies on selfing as a reproductive mechanism, pollen flow between populations is negligible and is unlikely to muddy patterns structured by seed dispersal. In addition, seed dispersal distances are likely low, because *Epifagus* is passively dispersed, so gene flow is very restricted, and populations of *Epifagus* should be highly differentiated. Selfing also results in a smaller effective population size (Pollak 1987), making populations more sensitive to the effects of genetic drift, and increasing the chances of homogeneous and highly differentiated populations. In other species, the advantage of selfing creating differentiated populations is balanced by a normal mutation rate and a low diversity of alleles, so that resolution of geographic patterns may be inhibited by isolated populations becoming fixed for similar alleles by chance. Fortunately, genetic variation is also easily uncoverable in *Epifagus*. Due to its heightened mutation rate, fewer base pairs need to be sequenced to reveal many polymorphisms, making collecting informative genetic data highly efficient. Also, its chloroplast genome has already been sequenced and annotated (Wolfe and dePamphilis 1998), making primer design straightforward. Lastly, the power to differentiate between migration scenarios largely rests on extensive data collection from many localities evenly spaced across the entire range of *Epifagus*. Collection of specimens of *Epifagus* is relatively straightforward because it is fairly abundant, easy to identify, and highly charismatic, so many local natural history societies and herbaria have records specifying its widespread distribution.

Next steps in comparative phylogeography

Comparative phylogeographies strive to find common processes that limit range expansion in a wide variety of taxa. Towards this goal, many single-species phylogeographic techniques have been developed. The challenge is how to overlap these species specific patterns into a multi-species history. In this review, I've presented several current comparative phylogeographic approaches and introduced a GIS framework for relating genetic population structure to other datatypes such as fossil pollen information. In the future, other data sources should also be included in these analyses; for instance, ecological niche models at different time periods could inform changes in habitat availability. Multi-datatype and multi-species comparisons could elucidate aspects of life history, ecology, or population history that explain what most limits a species' range expansion.

This work also has significant implications for conservation. Post-glacial migration is approximately analogous to present day range expansion due to climate change. Through identifying the factors that controlled past migration and the life history characters that are associated with these limitations, we are better able to predict future migration events and to devise management strategies. Design of corridors for conservation should be informed by current patterns in landscape connectivity and gene flow, which can be a direct result of phylogeographic studies. Genetic data on population structure should be used to parameterize habitat cost models to locate migration corridors, which can also be used in forecasting future patterns. For instance, Epps et al. (2007) discussed using a habitat cost model parameterized based on current genetic data from bighorn sheep to predict patterns of landscape connectivity if sheep populations were translocated to new areas.

Finally, understanding the migration patterns of parasites and pathogens is becoming more important as current climatic shifts open up new habitats for pests and as non-native pests are transported to previously inaccessible areas. Global climate change and increased habitat disturbance and fragmentation are likely to result in high likelihoods of invasions by non-native species (Dukes and Mooney 1999). For example, many species of *Striga* (Orobanchaceae) that parasitize grasses and important crop species across their native range in the Old World tropics and Africa are predicted to be of "great invasive potential" based on ecological niche models of current climatic conditions (Mohamed et al. 2007). Several *Striga* species have already been introduced to North America in the 1950s and were subject to an intense eradication regimen that largely controlled their impact (Shaw et al. 1962), but new invasions into Florida and other areas are still occurring (Botanga and Timko 2007). Studies of the constraints on parasite migration will be essential to controlling these outbreaks in the future.

2. Host density drives the post-glacial dispersal of the tree parasite, *Epifagus virginiana*

As climate changes and species strive to survive, they are forced to shift their geographic ranges in order to remain in suitable habitats. For parasites and other highly specialized species, their distributions are dictated not only by climate but by the geographic range of the host, variation in host density, and many other environmental factors engineered by associated species. Due to the often obligate relationship between host and parasite, parasite range expansions are thought to track the migration routes of their hosts, so they make attractive proxies for the host with their shorter generation times and more quickly evolving genomes (Wirth et al. 2005; Nieberding and Olivieri 2007). However, symbionts are affected by more than just their hosts' migration routes. Parasite and host phylogeographic histories may be incongruent, and that conflict reflects other processes that have limited the spread of the parasite relative to the host. In particular, the ability of the parasite to colonize the host landscape may be limited by its dispersal abilities or by the host's density, which is highly sensitive to changes in climate. A historical approach was taken, and I investigated the post-glacial migration and assembly of a tree-herb host-parasite system. The influence of host density was examined by comparing the migration history of a parasitic plant against both the host phylogeography and host abundance data from the last 21,000 years. Post-glacial migration is approximately analogous to present day range shifts due to climate change, so the results of this study are pertinent to understanding current factors controlling community assembly at different trophic levels (Petit et al. 2008).

The eastern North American study system consists of the parasitic plant, *Epifagus virginiana* (beechdrop; Orobanchaceae), which is an annual, self-fertilizing root parasite that

grows obligately on one tree species, *Fagus grandifolia* (American beech; Fagaceae). The migration history of *F. grandifolia* has been extensively studied with both molecular (McLachlan et al. 2005) and paleopollen data (Davis 1983; Bennett 1985). Along with other eastern North American species, *F. grandifolia* has contributed to a long-held paradigm concerning the capacity of temperate forests to successfully track climate change. Based on fossil pollen data of these species, rapid migration rates have been estimated for trees confined in southern refugia during glacial periods that expanded their ranges quickly northward to their current distributional limits (Davis 1983; Delcourt and Delcourt 1987). These migration rates are orders of magnitude faster than estimates based on life history and seed trap data (Clark et al. 1998). Recently, fossil pollen datasets have been re-interpreted as reflecting changes in abundance rather than documenting changes in range limits (Clark and McLachlan 2003). Molecular datasets, because they are sensitive to small population sizes and founder events likely to occur at the range margins, are more likely to show the initial migration pathways (Petit et al. 1997). Molecular phylogeographies have since uncovered more northerly refuges at mid-latitudes closer to the ice margin (Stewart and Lister 2001; McLachlan et al. 2005; Gonzales et al. 2008), that produce lower estimates of migration rates (Provan and Bennett 2008). Together, molecular and paleopollen datasets provide insights into the two host factors most likely to influence broad-scale parasite migration patterns: changes in host range and host density.

The availability of such detailed host information provides hypotheses against which to test parasite migration scenarios. If host susceptibility rates are high and density independent, then the appearance of new host habitat and the parasite dispersal rate are the major constraints on parasite range expansion. However, many disease systems are

additionally sensitive to fluctuations in host density, and successful establishment of a parasite may be limited by these effects (Kermack and McKendrick 1927; van Hulst et al. 1987; Arneberg et al. 1998). Because *E. virginiana* is obligate and host specific to this single species of *Fagus*, this two-species system is straightforward and ideal to address these potential host effects. Furthermore, due to the extensive sampling of parasite and host data, I have the opportunity to use powerful analytical tools to answer the question, did the parasite, *E. virginiana*, primarily respond to changes in the host's range or to changes in host density as it colonized the post-glacial landscape? To answer this question, comparisons between the host and parasite migration histories were implemented using tools from (1) comparative phylogeography, such as genetic diversity and differentiation statistics, Monmonier's algorithm, and Bayesian coalescent methods, and (2) from landscape ecology, where I present a newly developed method utilizing linear models to test relationships between the genetic and fossil density datasets.

The initial migration of *Fagus grandifolia* into newly unglaciated territory was largely accompanied by rapid increases in abundance except in the colonization of the midwest. The fossil pollen data show beech to be broadly distributed in the south at 13 thousand years before present (kybp), followed by density increases in the northeast at 9 kybp, then a steadily westward increase in density into the midwest by 6 kybp (H_p; Fig 2.1A) (Williams et al. 2004). In contrast, the host molecular data suggest a refuge near the glacial boundary from which colonists dispersed directly northwards into the midwest (H_m; Fig 2.1A) (McLachlan et al. 2005). The initial dispersal into the midwest must have occurred at low density, so that the pattern would have been missed in the pollen record (McLachlan and Clark 2004). If the parasite shows similar migration routes to H_m, then it was able to

colonize low density host populations at the range margins, and its migration was largely limited by dispersal rate. Alternatively, if the parasite's range tracks the development of high density beech forests (H_P), this suggests that the parasite's range expansion was limited by host density. The results have implications for the parasite's migration capacity. Even if the fossil pollen estimates were correct and tree species could rapidly equilibrate with global change, additional factors may cause parasite and other herbaceous species to lag behind. This would result in the disassembly of modern temperate forests and creation of communities with no past analogs (Williams and Jackson 2007), making forecasting future range expansions problematic.

To elucidate the migration history of *Epifagus virginiana* two chloroplast DNA (cpDNA) markers were sequenced from 88 localities across the range with ~5 samples/locality (see Materials and Methods). Forty-one haplotypes are observed forming 4 groups separated by multiple substitutions with high support (>65 parsimony/likelihood bootstrap, >97 posterior probability) (Fig 2.1B). The yellow, orange, and green haplotype groups are distributed almost exclusively south of the last glacial maximum ice margin, while the blue haplotype group spans the midwest, northeast, and south (Fig 2.1C). Not surprisingly, the hypothesized host refuge close to the ice margin contains all four major lineages. However, unlike for the host, this region contains roughly the northernmost extents of the ranges of the yellow, orange, and green haplotype groups, which indicate that this region was an unlikely source for north-bound colonists into the midwest.

While the northern range of the parasite is dominated by only the blue haplotype group, the population structure in the south is complex with the highest levels of total genetic diversity (V_T), allelic richness (A_R), and population differentiation (Gst = 0.782) (Tab

2.1). The southern populations have low within (17.7%) and high among (82.3%) population variation according to an analysis of molecular variance (AMOVA) (Tab 2.4). The opposite is true for the northern regions, where parasite variation is increased within localities (71.2% and 46.2% respectively), indicating low population differentiation (Tab 2.4). While these results suggest more isolation between the southern populations and gene flow in the north, they stop short of providing the geographic locations of the boundaries segregating the genetic variation. To locate these barriers I opted to use Monmonier's algorithm, a method that traverses the geographic network of populations and identifies the spatial boundaries that maximize genetic differences between areas (Manni et al. 2004). To explore the robustness of the inferred patterns, the data were cross-validated by reconstructing genetic boundaries in 1000 subsampled datasets. Subsampling occurred at the locality level with 50% of localities sampled with replacement.

The results of the cross-validated Monmonier analyses show genetic breaks concentrated in the region 13, with almost no genetic breaks north of the ice margin, and none between the northeast and midwest (regions 9 and 6) (Fig 2.2). The lack of boundaries in the north further supports the east-to-west migration route of the parasite into the midwest (H_p; Fig 2.1A), but the pattern of genetic breaks in the south generally do not correspond to previously identified biogeographic boundaries such as along the Apalachicola, Tombigbee, or Mississippi Rivers (Soltis et al. 2006). This is consistent with the interpretation that *E. virginiana* survived the glaciation in many small isolated host populations across the southern range. In the north, the increase in host density following the last glacial maximum appears to have allowed the rapid range expansion of the parasite

which resulted in a lack of population differentiation and diversity (Ibrahim et al. 1996; Petit et al. 2002).

The lack of northward migration into the midwest is further supported by migration rates among regions estimated using the Bayesian coalescent-based program IMa (Fig 2.1D, Tab 2.2) (Hey and Nielsen 2007). Geographic regions 13, 9, and 6 were defined by the geographic range of host fossil pollen at 13 kybp, 9 kybp, and 6 kybp (Fig 2.1A). Migration rates along south-to-north corridors into region 6, which encompasses the midwest, were significantly lower than those along east-to-west routes ($M_{13->6}$ = 0.40, $M_{9->6}$ = 3.32, P = 0.0072, Tab 2.2, Fig 2.1D), reflecting the extensive sharing of blue haplotypes between the midwest and the northeast versus the restricted southern distributions of the other haplotype groups. These results were robust to other regional definitions. For instance, even if the area south of the ice margin in region 6 is included within region 13 (forming region 13+), migration rates from the expanded region were still significantly less than from the northeast ($M_{13+->6}$ = 0.155, P = 0.003, Tab 2.2). A westerly migration into the midwest is consistent with the parasite following increases in host density (H_P).

While the results of the AMOVA, Monmonier, and coalescent analyses suggest that *Epifagus virginiana* may be broadly tracking changes in host density, these findings are consistent with many hypothetical migration routes within regions 9 and 6 that may not agree with more detailed host patterns. To better assess the relationships between parasite genetic patterns and a) geographic location, b) host genetic patterns, and c) host density, a method was developed to analyze these spatial data within a Geographic Information System (GIS) framework and using linear regression models (see Materials and Methods). Parasite genetic patterns were developed into a raster map layer through interpolations of pairwise

Nei and Li genetic distances assigned to midpoints between neighboring localities (Nei and Li 1979) (Fig 2.3A). The map layer was consistent with the Monmonier and AMOVA results that indicated high levels of differentiation among localities in the southern part of the range for *E. virginiana* (Fig 2.3A). Linear models were constructed to predict the parasite genetic map layer based on a) latitude and longitude information, b) a similarly interpolated genetic distance map for the host, and c) an interpolated host pollen density layer averaged from 0-21 kybp (Fig 2.3B-D; see Tab 2.5 for best models). The locations of strongest genetic differentiation for the host are in the midwestern and central parts of the range (Fig 2.3C), while host density is highest in the north (Fig 2.3D).

Parasite genetic patterns were significantly better predicted by the model that incorporated average host pollen density ($R^2 = 0.81$, $\Delta AIC = 0$, wAIC = 0.62, Fig 2.3G) than the model including host genetic distance as the explanatory variable ($R^2 = 0.71$, $\Delta AIC = 521$, wAIC = 0, Fig 2.3F) (see models 2 and 3 in Tab 2.5). Furthermore, a model that incorporated both host density and host genetic distance performed significantly better than those with host genetics alone ($R^2 = 0.81$, $\Delta AIC = 1$, wAIC = 0.38, model 4 in Tab 2.5), though it was marginally worse than a simpler model with just host density. The coefficient for the host genetic data was only marginally different from zero (0.04243, s.e. = 0.02158, P = 0.0495; model 2, Tab 2.5). In contrast, the model incorporating host density information showed visible trends (Fig 2.3G) and had a larger coefficient significantly different from zero (-0.8237, s.e. = 0.03235, $P < 0.001$; model 4, Tab 2.5). These results strongly suggest that host density was the primary determinant in the parasite's range expansion and population genetics -- not the host's own genetic patterns and migration routes.

Comparing the demographic histories of this parasite and its host indicates that the parasite's movements broadly track not host presence but rather host density, potentially highlighting a process underlying community assembly. Because of this dependence on host density, this parasite should not be used as a direct indicator of host range expansion even though its life history (obligate, host-specific parasitism, and mostly vertical transmission) suggests its utility as a proxy (Nieberding and Olivieri 2007). In this case, the parasite was unable to keep up with the range margin of its host, perhaps indicating that the host's initial range expansion occurred at low density and thus limited the establishment of the parasite in new populations. Meanwhile, the parasite differentiated within isolated refuges in the south. Once the threshold host density level was reached in contiguous northern populations, parasites from the opportune source population (likely with blue haplotypes) then quickly colonized. More generally, it seems likely that herbs and other highly specialized species may lag behind because the habitats they require are distributed in a more complex fashion across the landscape (Pakeman 2001). For instance, other parasitic plants have shown high levels of population differentiation typical of long periods of isolation, small population sizes, and patchy habitats (Schmidt and Jensen 2000; Zuber and Widmer 2009), and patterns of genetic diversity in the mistletoe, *Arceuthobium americanum*, and its host, *Pinus contorta* var. *latifolia*, have supported a lag between host and parasite colonization of formerly glaciated territory (Jerome and Ford 2002; Jerome and Ford 2002). The slower movement of herbs and parasites suggest that forest assemblages may partially disassociate while species react separately to climate change, and it is unclear if the migration capacity of highly specialized species is sufficient to outrun extirpation from their current ranges.

In the North American parasite, the inferred existence of southern refuges and the decline of genetic diversity with increasing latitude are similar to the patterns of genetic diversity seen in European oaks, maples, beeches, and other trees (Petit et al. 2002; Petit et al. 2003). Both groups show evidence for southern refuges from which colonists originated. In Europe, this compaction into refugia in Iberian, Italian, or Balkan peninsulas was presumably due to topographic barriers (e.g. the Alps), and differentiation between refuges was encouraged by extensive geographic boundaries (e.g. the Mediterranean). In eastern North America, such extreme east-west geographic boundaries were nonexistent. Hence, the host tree was widely distributed south of the ice margin, and its genetic diversity patterns do not show a signal of being isolated in multiple southern refugia (McLachlan et al. 2005). In contrast with the host, the eastern North American parasite does show this signal. Instead of *geographic* boundaries, its isolation and subsequent differentiation were likely the result of patchy, high density host populations. It seems likely that parasites and other highly specialized species that are associated with high density forests may lag behind trees during community assembly and would show the signature of southern refugia.

Materials and methods

Data collection

Epifagus virginiana: A total of 467 samples from 88 localities across the entire range of *E. virginiana* (Fig 2.1A) were sampled during 2003-2007. Plants from the same locality were collected at least 1 m apart, ensuring sampling of different individuals. DNA extractions were performed on fresh, frozen, or silica gel dried corm or stem material following a standard CTAB protocol (Doyle and Doyle 1990) or using DNeasy Plant Mini Kits (Qiagen

Inc., Valencia, CA.). PCR was performed to amplify two cpDNA regions: pseudogene *rbcL* with primers rbclf (GACTGAAAATCCTAGTGCCATCA) and rbclr (ACTAAACCGCCATCTTTCCA) and the first intron of *clpP* with primers clpP1f (AATGGTTTGCCTGTCCTTTG) and clpP1r (ACGTTTAGCATTCCCTCACG). The primers were specific to *E. virginiana* and designed based on its published cpDNA genome (Wolfe et al. 1992). PCR products were produced by *Taq* DNA polymerase or Phusion High-Fidelity DNA polymerase (New England BioLabs, Ipswich, MA). PCR products were cleaned using an enzymatic protocol (Werle et al. 1994) with (2 U) Exonuclease I and (0.1 U) Shrimp Alkaline Phosphatase or (0.1 U) Antarctic Phosphatase (New England BioLabs, Ipswich, MA). Cycle sequencing was performed following standard procedures using BigDye Terminator v3.1 (Applied Biosystems Inc., Carlsbad, CA). Dye-labeled amplicons were analyzed on an ABI 3700 or Applied Biosystems 3730xL DNA Analyzer (Applied Biosystems Inc., Carlsbad, CA). Sequences were edited and aligned using Sequencher v4.8 (Gene Codes Corporation, Ann Arbor, MI). Indel regions were excluded from analyses, leaving 1016 bp analyzed.

Fagus grandifolia: 1901 bp of cpDNA sequence data for 121 samples from 97 localities of *F. grandifolia* were analyzed. These data were from a previous study (McLachlan et al. (2005), and are publicly available in the Ecological Society of America's Ecological Archives, E086-110-S1 (http://www.esapubs.org/archive/ecol/E086/110/).

Fagus grandifolia fossil pollen data: Interpolated fossil pollen data were taken from Williams et al. (2004). These data offer pollen percentages at 1000 year time slices between the present and 21 kybp, and are publicly available via the Ecological Society of America's Ecological Archives, M074-007-S1 (http://www.ncdc.noaa.gov/paleo/pubs/williams2004/).

Phylogenetic analyses of E. virginiana

A subset of the cpDNA sequences consisting of only the 41 unique haplotypes were analyzed under maximum parsimony, maximum likelihood, and Bayesian frameworks using PAUP* v. 4.0b10 (Swofford 2003), GARLI v. 0.951 (Zwickl 2006), and Mr. Bayes v. 3.1.2 (Ronquist and Huelsenbeck 2003) respectively. A statistical parsimony haplotype network was constructed using TCS v1.21 (Clement et al. 2000) (Fig 2.1B). The haplotype network was consistent with trees constructed using any of the above approaches. Haplotype groups were defined by clades with >65% bootstrap support (out of 1000 replicates) in the MP and ML analyses and >95% posterior probability in the Bayesian analysis.

Data visualization and figure creation

Phylogeographic visualizations of haplotypes on maps were created using PhyloGeoViz v. 2.4.4 (http://phylogeoviz.org, Tsai and Kidd 2007) and plotted in Google Earth (Google Inc., Mountain View, CA). Base layers of state boundaries were acquired from GPS Visualizer (http://www.gpsvisualizer.com/kml_overlay, Schneider 2003) and modified using Adobe Photoshop and Illustrator CS3 (Adobe Systems Inc., San Jose, CA). The ice margin at the last glacial maximum was plotted from Dyke et al. (2003).

Monmonier analysis with cross-validation

The entire *Epifagus virginiana* dataset, except samples from the Mexican locality (MX), was analyzed using Monmonier's algorithm. Monmonier analyses were carried out using the Adegenet package (Jombart 2008) for the R software environment v. 2.7.2 (http://www.r-project.org, The R Foundation for Statistical Computing). Barriers were computed with nrun = 1, scanthres = 0, and threshold = NULL. To minimize barriers found at the edges

of the range due to concavity issues, nine virtual points were added at (W 90.38, N 45.77), (W 90.55, N 44.11), (W 90.38, N 42.34), (W 89.05, N 40.82), (W 89.44, N 39.52), (W 91.01, N 38.34), (W 92.61, N 37.40), (W 91.83, N 40.96), and (W 93.69, N 39.41). Cross-validation analyses were performed by randomly subsampling the original data by 50% of the localities at a time. Monmonier barriers were then recomputed for each replicate. Paths in both directions were recorded. 1000 randomly subsampled replicates were performed. To summarize results, paths from all replicates were analyzed together and binned into a 2-D 25 cell x 25 cell histogram. Data were visualized using a color ramp based on 0.05 quantiles. C++ and R scripts of the analyses are available from the authors.

Regional definitions

The geographic range map for *Fagus grandifolia* (Fowells 1965) was assumed to be the same for the parasite. The range was divided into 3 geographic regions based on fossil pollen evidence at different time slices. Region 13 was defined by the location of *F. grandifolia* fossil pollen at 13 kybp that was >2% of all pollen recovered at a site. Region 9 was similarly defined based on the fossil pollen range at 9kybp, but it was the subset of the distribution not already within region 13. The remaining area was named region 6, which also corresponds with the distribution fossil pollen at 6 kybp that excludes regions 13, and 9. To investigate the sensitivity of analyses to the regional definitions, analyses were rerun with an expanded region 13 (region 13^{+}) and smaller region 6 (region 6^{-}), where all localities south of the ice margin within region 6 were transferred into region 13^{+}. The Mexican population (MX) was treated separately and largely excluded from the subsequent analyses.

Genetic diversity statistics

Diversity indices such as G_{ST}, N_{ST}, H_T, and V_T were calculated based on formulae in Pons and Petit (1996) and Petit et al. (2002). These were calculated for each geographic region for both the parasite (Tab 2.1) and host data (Tab 2.2). C++ source code performing the calculations is available from the authors. Allelic richness and rarefaction analyses were performed using CONTRIB (http://www.pierroton.inra.fr/genetics/labo/Software/Contrib/, Petit et al. 1998).

Analyses of molecular variance (AMOVA) were performed on the *Epifagus virginiana* dataset both over the entire range and within regions using Arlequin v. 2.000 (Schneider et al. 2000).

Estimating migration rates and other population parameters using IMa

Migration rates, population sizes, and time since divergence were estimated between all regional pairs (excluding the Mexican population) of the parasite using the program IMa (Hey and Nielsen 2007). The program was run in 'M' mode with the following set parameters: maximum population sizes ($4N\mu$) = 10-35, maximum migration rates = 10-40, maximum divergence time = 10, 30 chains with a geometric heat mode, heating parameters g1 = 0.95 and g2 = 0.8, and burn in = 2500000 steps. All analyses were run for at least 1.5 million steps (and up to 15 million steps) following burn in and appeared to have converged according to trendline plots. To assess the probability that the migration rate from one region was larger than that from another region, migration values were randomly drawn from each marginal distribution. If the value from the first distribution was greater than the second, a value of 1 was assigned. If smaller, a value of 0 was assigned. This was repeated 1000 times. The average value of the 0 or 1 assignments was equivalent to the probability

that the first migration parameter was larger than the second. This method of assessing significance was repeated for each pair of migration parameters.

Interpolations and linear models

Nei and Li genetic distances (1979) were calculated between each pair of localities for the parasite and host separately. These data were interpolated to create raster maps of genetic distance. The interpolations were done to the exact same grid cell size and extent as the already interpolated fossil pollen data (Williams et al. 2004). In order to determine what populations were to be considered 'neighbors', a geographic network of all localities was constructed based on a Delaunay triangulation method implemented in the Adegenet R-package (Jombart 2008). Nei and Li genetic distances between these neighboring localities were then assigned to the midpoint coordinates between them. These formed the basis of the interpolations. Interpolations proceeded according to standard inverse weighting methods, where closer points were weighted more favorably than points farther away. The fossil pollen record data was summarized by averaging host pollen percentages from 0-21 kybp, not including zero or NA values. Maps were visualized in R using a 0.1 quantile color ramp.

The goal of the linear models was to assess similarity in the geographic patterns of genetic differentiation and density. Before the regressions were performed, the maps were transformed into comparable scales. This was done by relating each value to its percentile in the total distribution of the dataset, so that each map had values ranging from 0 to 1. The row and column number of each grid cell were used as proxies for latitude and longitude. Linear regression models were constructed using the stats R-package, and each model was of the form: Parasite genetic distance ~ Host data + Row + Column + Row:Column. The host

data included one or more of the following: host genetic distances or average host pollen density as described above. For each analysis with just one host element, multiple models were constructed with different numbers of polynomial terms (up to x^{10}). F-tests were performed to select the best polynomial model based on goodness of fit while biasing towards less complicated models with fewer terms. Models were also constructed with all combinations of one or two host explanatory variables. Models were compared using AIC values. The best models are reported in Tab 2.5. The absolute values of the residuals were mapped to highlight geographic areas where the models poorly explained the data. An equal interval color ramp (excluding the top 5%) was chosen to allow easy comparison between models.

Table 2.1: Genetic diversity statistics for *Epifagus virginiana*. L = No. of localities, N = No. of samples, G = No. of haplotype groups, A = No. of alleles, A_R = Allelic richness after rarefaction to 5. A_E = No. of endemic alleles. See Fig 2.1 for definition of regions. Diversity statistics were not calculated for the MX population due to low sample size.

Region	L	N	G	A	A_R	A_E	H_S (se)	V_S (se)	H_T (se)	V_T (se)	G_{ST} (se)	N_{ST} (se)
13	35	203	4	22	2.48	17	0.177 (0.047)	0.152 (0.048)	0.813 (0.045)	0.931 (0.117)	0.782 (0.059)	0.837 (0.056)
9	18	78	1	6	1.29	4	0.324 (0.074)	0.072 (0.017)	0.585 (0.034)	0.125 (0.012)	0.446 (0.131)	0.428 (0.150)
6	32	182	4	19	2.24	13	0.355 (0.051)	0.213 (0.058)	0.759 (0.054)	0.515 (0.104)	0.532 (0.075)	0.587 (0.140)
MX	1	5	1	1	0	1	~	~	~	~	~	~
All	86	468	4	41	2.82	~	0.272 (0.032)	0.156 (0.030)	0.861 (0.018)	0.862 (0.058)	0.684 (0.038)	0.819 (0.037)

Table 2.2: Population size, migration, and divergence time estimates between geographic regions of *Epifagus virginiana* defined in Fig 2.1(A). The highest point in the posterior probability of the marginal distribution followed by the 95% credible interval is reported for each parameter. *R1*, *R2* = Regions 1 and 2. θ_1, θ_2, and θ_A = Population size of region 1, 2, and ancestor respectively. $M_{X->Y}$ = Migration rate from region X into region Y. t = Divergence time. Estimates scaled by mutation rate. a > b (P > 0.99).

R1	*R2*	θ_1	θ_2	θ_A	$M_{2->1}$	$M_{1->2}$	t
13	6	20.20 (12.97-33.55)	17.22 (10.13-30.12)	0.07 (1.35-135.70)	0.28a (0.04-1.12)	0.40a (0.10-1.36)	4.64 (0.90-9.74)
9	6	4.61 (1.37-12.55)	26.11 (14.57-66.68)	12.24 (2.28-102.81)	0.02 (0.02-8.20)	3.32b (1.36-13.12)	4.04 (0.50-9.76)
13	9	24.67 (16.29-39.80)	4.06 (1.68-12.04)	5.18 (1.68-119.87)	0.10a (0.01-0.52)	0.02a (0.02-1.16)	3.64 (1.22-9.80)
13^+	6^-	25.98 (17.64-41.80)	11.03 (5.58-21.86)	8.43 (1.88-189.00)	0.12a (0.00-0.72)	0.16a (0.02-1.18)	2.90 (0.92-9.70)
9	6^-	4.08 (1.76-12.44)	17.09 (7.57-71.72)	6.87 (2.06-97.38)	0.05 (0.07-14.25)	7.78b (2.92-29.58)	3.40 (0.70-9.62)

Table 2.3: Genetic diversity statistics for *Fagus grandifolia* (calculated from McLachlan et al. 2005). L = No. of localities, N = No. of samples, A = No. of alleles, A_R = Allelic richness after rarefaction to 5. See Fig 2.1 for definition of regions. Because of low sample size, *F. grandifolia* samples were grouped into "localities" roughly by state boundaries.

Region	L	N	G	A	A_R	A_E	H_S (se)	V_S (se)	H_T (se)	V_T (se)	G_{ST} (se)	N_{ST} (se)
13	9	50	5	8	1.14	6	0.305 (0.126)	0.268 (0.111)	0.446 (0.127)	0.416 (0.114)	0.315 (0.343)	0.355 (0.320)
9	4	22	3	6	1.32	3	0.477 (0.160)	0.207 (0.155)	0.475 (0.117)	0.280 (0.148)	-0.004 (0.206)	0.261 (0.314)
6	9	49	4	8	2.44	4	0.563 (0.120)	0.489 (0.161)	0.815 (0.067)	0.893 (0.078)	0.309 (0.158)	0.452 (0.186)
All	22	121	5	17	1.85	~	0.442 (0.078)	0.347 (0.085)	0.636 (0.079)	0.638 (0.098)	0.305 (0.150)	0.456 (0.157)

Table 2.4: Analyses of Molecular Variance (AMOVA) in *Epifagus virginiana*. Regions are defined in Fig 2.1A.

	% of Variation		
Region	Among regions	Among populations	Within populations
13	~	82.28	17.72
9	~	28.83	71.17
6	~	57.36	42.64
Entire range	34.25	47.65	18.10

Table 2.5: Results of linear regression models. The genetic distance of *Epifagus virginiana* (*P*) is the response variable in all cases. Best model is in bold. *R* = row (latitude), *C* = column (longitude), *G* = genetic distance of *Fagus grandifolia* (Nei and Li distances), *D* = average density of host through time (% of overall pollen). All models have P-values <0.0001.

Model	Variables	Polynomial model (*P* =)	R^2	AIC	ΔAIC	wAIC
1	*R, C*	$1.5503 - 0.0308R - 0.0046C + 0.0001RC$	0.71	-1105	523	0.00
2	*G, R, C*	$1.675 + 0.0424G - 0.0359R - 0.006C + 0.0002RC$	0.71	-1107	521	0.00
3	***D, R, C***	$\mathbf{2.0679 - 0.8237D - 0.0299R - 0.0037C + 0.0002RC}$	**0.81**	**-1628**	**0**	**0.62**
4	*G, D, R, C*	$2.0951 + 0.0095G - 0.8225D - 0.031R - 0.0041C + 0.0002RC$	0.81	-1627	1	0.38

Figure 2.1: *Epifagus virginiana.* **(A)** Range map with competing migration route hypotheses shown. Regions shaded according to the distribution of host (*Fagus grandifolia*) pollen at 13 kybp, 9 kybp, and 6 kybp (Williams et al. 2004). H_p is the hypothesis based on host fossil pollen data; H_m is based on host cpDNA molecular evidence. **(B)** cpDNA haplotype network. Colors show haplotype group assignments. Numbers are the parsimony and likelihood bootstraps and Bayesian posterior probability values, respectively, for each group. Unfilled circles are unsampled haplotypes, and circle size reflects number of samples (min = 1, max = 128). **(C)** Geographic distribution of cpDNA haplotypes. Pie size is scaled to the number of samples per locality (min = 1, max = 11). Colors reflect haplotype group assignment as in (B). The Mexican population (MX) is >1250 km disjunct and is within a *F. mexicana* stand. LGM = last glacial maximum. **(D)** Migration rates. Rates were estimated in pairwise analyses between the three regions (A) using the program IMa (Hey and Nielsen 2007). Migration rates that are significantly faster than others into the same region are indicated by * for $P < 0.01$.

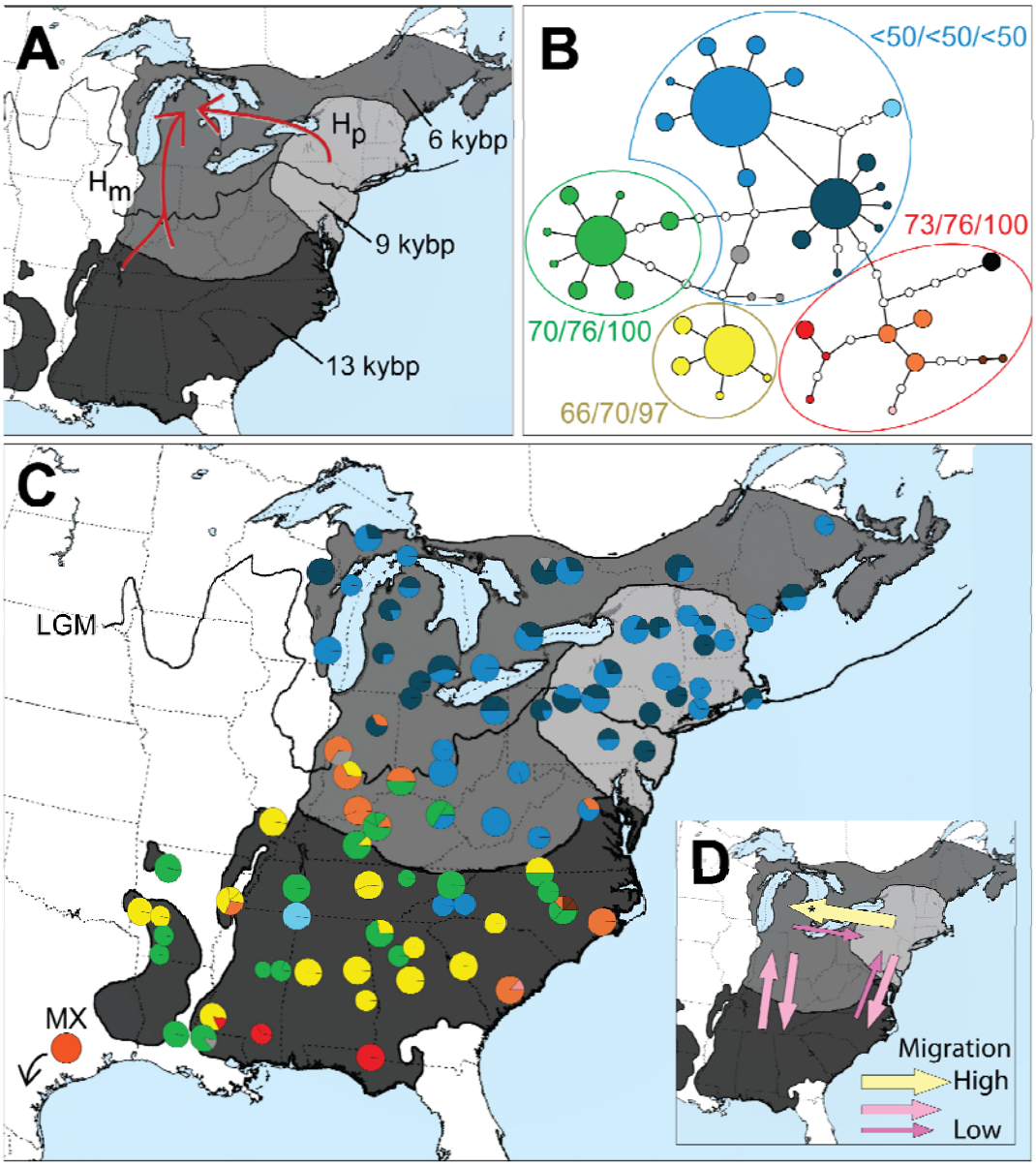

Figure 2.1: ***Epifagus virginiana.***

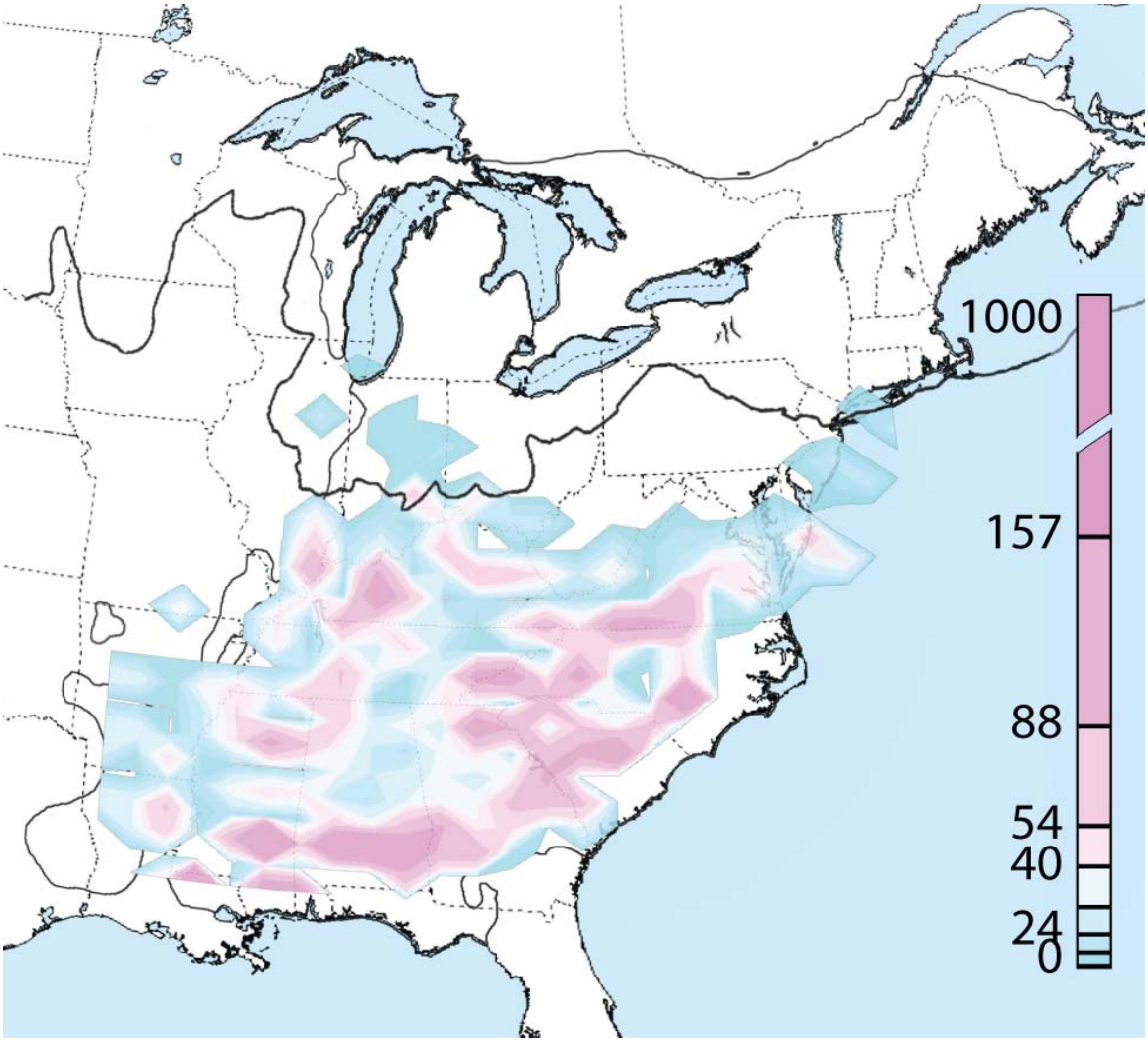

Figure 2.2: Areas of high genetic differentiation of *Epifagus virginiana*.
Discontinuities in the genetic structure were identified using Monmonier's algorithm. Colors correspond to the number of times a genetic break was found in that area over the 1000 cross-validation subsamples. Warmer colors have higher numbers of breaks; cooler or uncolored areas had few genetic breaks.

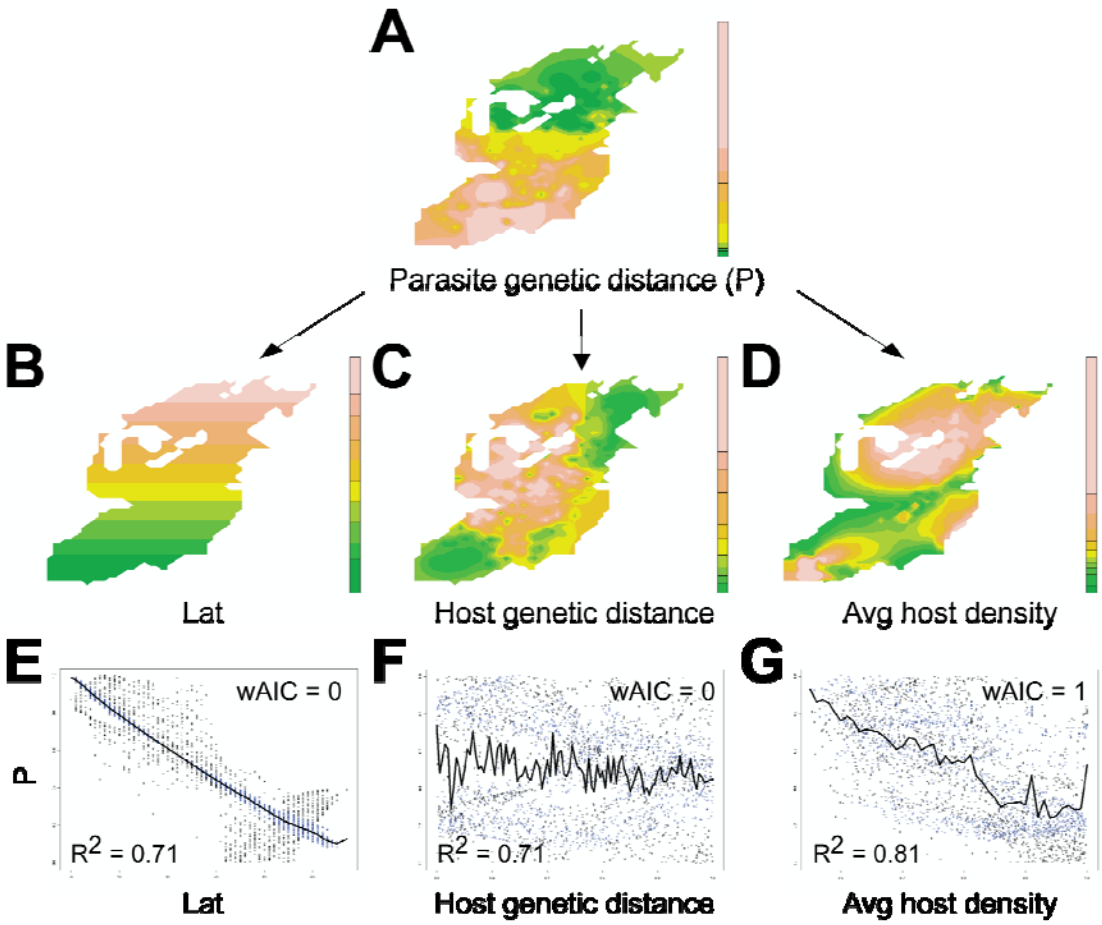

Figure 2.3: Linear models predicting the genetic distances of *Epifagus virginiana*. **(A)** Map of the genetic distances of *E. virginiana* (the response variable). **(B-D)** Maps of the major explanatory variables: latitude (B), host genetic distance (C), and average host density (D). **(E-G)** Graphs of explanatory variables against parasite genetic distance. Black circles are original data; blue circles are fitted parasite genetic distances based on the model. Black lines are average fitted data.

3. Using spatial statistics and landscape genetic tools to identify the migration corridors of *Epifagus virginiana*

Introduction

Phylogeographers and paleoecologists are interested in many of the same biogeographic questions regarding community assembly. For instance, in eastern North America, the migration routes of tree species and range changes of biomes have been intensively studied both with genetic and fossil data. However, these research endeavors have progressed largely independently. Standard methods of synthesis have yet to be developed that compare both genetic and paleo data in order to understand the relative colonization patterns of multiple members of the same community. I am in a unique position to develop such methods due to the availability of extensive molecular and fossil datasets for my two focal species, a host-parasite system. Previous phylogeographic work on the parasitic plant, *Epifagus virginiana* (beechdrop, Orobanchaceae), showed broad agreement between the parasite's population genetic structure and trends in the density of the host tree, *Fagus grandifolia* (American beech, Fagaceae) (see chapter 2). Here, this work is expanded with additional parasite molecular data and improved linear models to better understand how different host variables affect the parasite's post-glacial colonization in key areas of eastern North America. The knowledge of the relevant host constraints is then applied to build a habitat cost model for the parasite, from which the parasite's migration corridors can be explicitly located.

The aims of this study are four-fold: (1) to identify which host variables (genetic structure, density, variance in density, or arrival age) influenced the parasite's range expansion, (2) to understand the effects of host variables in the unglaciated versus the

formerly glaciated parts of the range, (3) to geographically locate the migration corridors for the parasite, and (4) to confirm the directionality of the post-glacial migration routes of the parasite.

Previous phylogeographic work based on cpDNA loci in *E. virginiana* revealed extensive population differentiation in the southern part of range, while the north is dominated by extensive sharing of closely related haplotypes that are rare in the south (see chapter 2). This pattern is broadly consistent with host fossil pollen data that showed development of high density beech forests first in the south of the range, then in the northeast, and finally westward into the midwest (Davis 1983; Davis and Zabinski 1992; Williams et al. 2004). However, the parasite migration patterns conflicted with the host's genetic population structure which suggests refuges close to the ice margin and a northward (instead of westward) range expansion into the midwest (McLachlan et al. 2005). These comparisons are interpreted to reflect the important role of host density on the parasite's range expansion, whereas changes in the host's own range distribution had nonsignificant effects on the parasite's phylogeography (see chapter 2). While the cpDNA was enough to elucidate these broad trends between datasets, alone it was insufficient to specifically locate the migration routes of the parasite. For instance, though it was established that the migration rate into the midwest was higher from the northeast than the south, migration rates into other regions were equivocal between potential source populations, and the credible intervals for regional divergence times were highly overlapping. Furthermore, the directionality of the migration route connecting the north remains in question, with the cpDNA data producing similar levels of migration between the northeast and midwest. Additional genetic loci should provide more accurate estimates of migration rates and

divergence times between geographic regions, thus increasing the resolution of the parasite's population genetic patterns and the ability to discern directionality in those patterns.

In addition to increasing the power to infer phylogeographic patterns, the new loci provide independent tests of the relationships between parasite population history and host variables. To quantify these relationships linear models are used. Previously in chapter 2, the models relating host and parasite variables included latitude and longitude parameters, but did not explicitly account for spatial non-independence of the data. This error could be non-negligible considering the parasite genetic map layer was created by spatial interpolation of data points. In fact, reanalysis of those models show strong spatial autocorrelations in the model residuals, suggesting that the resulting significance of model coefficients may be inflated. This bias is corrected by building spatial autoregressive linear lag models (Ward and Gleditsch 2008), an approach that includes spatial dependencies among the parasite genetic data.

In addition to data and model improvements, how the relationships among host and parasite variables may change across space is investigated. Are the host variables shaping parasite population structure in the north also responsible for parasite patterns in the south? To investigate the possibilities of spatially dependent parasite histories, parasite population structure is modeled against multiple host variables utilizing data from the southern and northern ranges separately in addition to building models based on the full geographic distribution.

From the parasite phylogeography and linear models, the host constraints that face the parasite on the landscape are better understood. These relevant host variables can be interpreted as indicators of the quality of the parasite's habitat. Given a habitat map,

landscape ecological tools such as landscape connectivity and path analysis can be applied to elucidate common migration corridors. Since the host fossil pollen data layers have higher resolution than the parasite genetic data, these data can be used to obtain with higher specificity the locations of parasite dispersal routes.

The landscape path analyses require a habitat layer that consists of resistance costs assigned to various habitat types. In the ideal case, the resistance value should be linearly correlated with the dispersal rate through the particular habitat (Beier et al. 2008). Accurately parameterizing the habitat cost model is difficult and can be highly controversial. Resistance values are often estimated based on expert opinion and/or empirical ecological studies of dispersal rates, which usually exhibit high species and locality specificity that may not be broadly applicable to other studies, so appropriate estimates may simply be unavailable in the literature (Beier et al. 2008). Also, it is not uncommon for habitats to be ranked and then arbitrarily assigned weights. The *Fagus-Epifagus* study system is uniquely suited to this analysis because the necessary empirical estimates of migration rates can be obtained through the genetic datasets. Population genetic theory provides a useful model selection framework: a positive relationship is expected between genetic distances and effective geographic distances (Wright 1943; McRae and Nürnberger 2009), i.e. the geographic distance accounting for resistances encountered in the landscape. The better the habitat cost model, the more accurate the estimates of effective geographic distances, and the higher correlation between the geographic distances and genetic distances (Cushman et al. 2006; Epps et al. 2007). The genetic data provide a method to discern between the various habitat cost models, which are then used to locate the most likely migration corridors.

Materials and Methods

Genetic data

Epifagus virginiana: Specimens were collected from 87 localities spread across the range of *E. virginiana* excluding the single Mexican population (see chapter 2 for more information). In addition to the two cpDNA loci analyzed previously, genotypic data were added from 9 presumably nuclear microsatellite loci of 113-362 bp: MS10 (177-187 bp), MS34 (218-259 bp), MS43 (221-303 bp), MS63 (113-136 bp), MS76 (173-182 bp), MS92 (345-362 bp), MS105 (143-161 bp), MS130A (260-266 bp), and MS135 (164-184 bp) (Tab 3.1). The microsatellites were developed following Zane et al.'s FIASCO method (2002) that utilizes AFLP's and a microsatellite enrichment step. PCR products were produced by *Taq* DNA polymerase or Phusion High-Fidelity DNA polymerase (New England BioLabs, Ipswich, MA). Samples were pooled into sets of 2-3 loci and run on an Applied Biosystems 3730xL DNA Analyzer with a ROX 500 size standard (Applied Biosystems Inc., Carlsbad, CA). Genotypes were scored using GeneMarker v.1.80 (SoftGenetics LLC, State College, PA). Because alleles were not equally spaced (e.g. every 2 bp), a stepwise, repeat-based mutation rate between alleles was not assumed in analyses. Instead, alleles of a locus were ranked by their size and assigned an index value. The locus MS92 was more complex than the others with several alleles (1-4) recovered for each sample; for instance, a sample might have bands at 345, 351, and 362 bp. *E. virginiana* is a diploid with 2n=38 (Hanna Weiss-Schneeweiss, unpublished data) and very few heterozygotes were recovered at the other loci (see results below), so this pattern appears to be the case of single alleles resulting in multiple bands. Each set of bands was scored as a single allele with 3 possible alleles in total; all

samples were assumed to be homozygous at this locus. In total, 450 samples were genotyped, 350 (78%) of which at all 9 loci, 403 (90%) at 6 or more.

Fagus grandifolia: Chloroplast DNA data from 121 samples from 97 localities were obtained from McLachlan et al. (2005) (see chapter 2 for more information).

Construction of genetic map layers: The genetic map layers, EV and F.gen, were based on interpolated genetic distances between localities of *E. virginiana* and *F. grandifolia* respectively. Pairwise genetic distances were estimated using Reynolds' method (Reynolds et al. 1983), a measure similar to Fst but linearized to divergence time, implemented in the Adegenet v.1.2-2 R-package (Jombart 2008). Genetic distances were assigned to the coordinates midway between each locality pair. The distances were interpolated on a 50 km grid, with grid cells corresponding to those in the fossil pollen data (Williams et al. 2004). In 17 instances, multiple parasite localities fell within the same grid cell. In these cases, the data were pooled to form a single population per grid cell (8 pooled populations in total). Interpolations proceeded according to a modified quadratic Shepard's inverse weighting method (Franke 1982), where closer points were weighted more favorably than points farther away. Only points within the present-day distribution of *F. grandifolia* were assigned values. All other cells were left empty (assigned zero). Maps were visualized in R using a 0.1 quantile color ramp.

Fossil pollen data

Fossil pollen data for the host, *Fagus grandifolia*, were obtained from Williams et al. (2004). The data are interpolated pollen percentages at 1000 yr time slices since 21 thousand years before present (kybp), and a layer at 500 ybp (see chapter 2 for more information). Absent data were assigned zero values. Three summary-pollen layers were constructed

based on the 23 time-slice layers: (1) an average pollen layer based on the arithmetic mean of all time-slices with all zeroes discarded (F.avgP), (2) a layer on the variance in pollen density across all time slices (including zero values) (F.varP), and (3) a colonization age layer which tracks the earliest date >2% pollen was recorded in that grid cell (F.age). The threshold of 2% is consistent with values adopted by other authors in studies of European *Fagus* (Magri et al. 2006), and it produced colonization ages consistent with prior interpretations of the data (Davis 1983; Davis and Zabinski 1992). Ten of the 1261 grid cells within the present-day distribution of *F. grandifolia* had <2% *F. grandifolia* pollen at all time-slices. For those cells, we estimated the arrival time by averaging values in the adjacent 8 cells. All layers were masked to include only points within the present-day distribution; all other cell values were set to zero. To facilitate comparisons among pollen density layers, the time-slice and avgP layers were visualized using a common 0.1 quantile color ramp based on their combined pollen frequency distributions. The varP and Fage layers were plotted on a 0.1 quantile color ramp based on their separate distributions.

Phylogeographic analyses

The population genetic structure of the parasite was investigated using standard F statistics and a Bayesian spatial clustering method. The F statistics, F_{IS}, F_{IT}, and F_{ST}, based on H_I, H_S, and H_T, were computed for each locus separately and then averaged over all loci. F statistics were estimated over the entire geographic distribution, and within geographic regions shown in Fig 3.1. Regions were defined by the distribution of >2 % pollen at 13, 9, and 6 kybp (see chapter 2 for more information). The Bayesian spatial clustering method was implemented in BAPS v.5.2 (Corander et al. 2008). This method utilizes only biallelic genotypic data (so only the microsatellite genetic data were used, and the cpDNA data

excluded) and assigns individuals to clusters that have restricted gene flow between them. In addition to the genetic data, this method incorporates spatial information within the prior distribution for the clusters (Corander et al. 2008). It systematically tries a range of numbers of clusters (up to 50), and it returns the most probable number of clusters and the cluster assignments. The geographic distributions of the cluster assignments were visualized using PhyloGeoViz v.2.4.4 (http://phylogeoviz.org) and plotted in Google Earth (Google Inc., Mountain View, CA). Colors were chosen based on the relationships between clusters as seen in the cluster phenogram.

The migration routes of *Epifagus virginiana* hypothesized in chapter 2 were further explored by using the full cpDNA and microsatellite datasets in estimating population demographic parameters between geographic regions (Fig 3.1). Migration rates, effective population sizes, and divergence times between regions were estimated using IMa (Hey and Nielsen 2007) (see chapter 2 for more information), a Bayesian, coalescent, MCMC method that incorporates both migration and isolation in its population divergence model. Differences between estimates were statistically quantified by randomly drawing values from each distribution, comparing the values, and calculating the average frequency that the value drawn from one curve was greater than the other (see chapter 2 for more information).

Linear models

To determine which *F. grandifolia* layer(s) best predicted the parasite's genetic population structure, many spatial autoregressive lag models were compared. These models are enhanced linear regressions that include an additional spatial lag term as an explanatory variable, thereby accounting for spatial non-independence in the *E. virginiana* genetic data

layer (Ward and Gleditsch 2008). Before considering models incorporating multiple host data layers, single variable models built around the host time-slices were evaluated for best fit of the parasite data. Only one host time-slice layer would be used in the multivariate analyses because of the non-independence of these temporally related layers. Single variable spatial regression models were constructed for each of the 23 fossil pollen layers and then competed by their Akaike Information Criterion (AIC) values. The best layer with smallest AIC value was chosen for the multivariate analyses (Wagenmakers and Farrell 2004). Multivariate spatial models were then constructed with combinations of one or more of the four host data layers: F.avgP, F.age, F.varP, and the best time-slice layer. Due to issues of non-independence, models contained either F.avgP or the best time-slice layer, never both. Models without significant regression coefficients were discarded. To evaluate the remaining multivariate models an AIC criterion was also used where Akaike weights represent the probability that a model is the best of the set (Wagenmakers and Farrell 2004).

The processes affecting parasite population structure may have differed in the unglaciated and formerly glaciated territories. To study the possible differences, the landscape was divided into southern and northern regions (corresponding to region 13 versus everywhere else, Fig 3.1). Linear models were constructed as described above for the southern and northern data subsets as well as for the entire range.

Since linear models rely on assumptions of relationship linearity and homoscedasticity, several data transformations were tried to minimize assumption violations. Specifically, all data layers were normalized by performing Box-Cox transformations (Box and Cox 1964), centering the data, and scaling the data by their standard deviations using the

qAnalyst v.0.5.1 R-package (www.quantide.com). Analyses were performed on both the original and transformed datasets.

Migration corridor analysis

After determining what host variable(s) best predict the parasite's population structure, the best host map(s) were used as "habitat" inputs in a landscape linkage analysis. The most likely paths connecting population localities reflect common migration corridors. Circuitscape v.3.4.2 (McRae et al. 2008) was used to estimate effective geographic distances and paths between populations. This program's methodology is similar to least-cost-path and other landscape corridor analyses except that it utilizes circuit theory to consider many possible paths instead of returning only a single route (McRae et al. 2008). Results were visualized in R using the base graphics R-package (www.r-project.org) and a 0.1 quantile color ramp.

Ideally, the habitat maps should be associated with resistance costs that are related to how difficult or easy it is to move through that habitat type. Measures of gene flow directly relate to the underlying habitat porosity, and studies of isolation by distance suggest that the log of the effective geographic distance between populations should be linearly correlated with genetic distance (Epps et al. 2007; McRae and Nürnberger 2009). Effective geographic distances were estimated as described above in Circuitscape v.3.4.2 (McRae et al. 2008). Genetic distances were estimated using F_{ST} or Reynolds' method through the R-packages Geneland v.3.1.4 and Adegenet v.1.2-2 (Guillot et al. 2005; Jombart 2008). The models were then assessed using Mantel tests (Vegan v.1.15-0 R-Package) of the log of effective geographic distance versus genetic distance; better habitat models would result in higher correlation coefficients between the two distance measures.

Habitat cost models were constructed based on the assumption that a threshold host density exists where habitats above that threshold are much easier to invade than localities below the threshold. Host density thresholds are common in many disease systems (Kermack and McKendrick 1927), though there are no empirical estimates of this value in *E. virginiana*. Since there were no starting estimates, models were constructed with thresholds of 0.23%, 2.85%, and 5.4% *F. grandifolia* pollen densities which correspond with the 25, 50, and 75th quantiles of the pollen distribution. Following the initial analyses, models were also run with 1.77% and 3.8% (the 40 and 60% quantiles) pollen density thresholds. Grid cells above the threshold were assigned resistances equal to the pollen percentage. Cells below the threshold were assigned the pollen percentage multiplied by a weighting factor. Multiple weighting schemes were tried ranging from 0.0001 (low host density cells were 10,000 times worse) to 1 (no difference in weights between low and high densities). Two nonlinear models were also considered that were not based on a set pollen threshold. A logarithm model was constructed based on the log of all pollen densities, and an exponential model was constructed with resistances based on exponentiation.

Results

Phylogeographic structure of Epifagus virginiana

The microsatellite dataset produced 161 unique genotypes, with 69% of genotypes occurring only once in the sample set, and the maximum times a genotype was recovered was 7. Each genotype was restricted to a single locality except in 4 cases where they were shared between geographically close populations (ex. between 3 localities in AR, between 4 localities in CT/VT/PA, between 3 localities in GA/KY, and between 3 localities in

VT/NH). The large number of private alleles led to high estimates of F_{ST} at all loci (Tab 3.1), indicating high levels of population differentiation. Interestingly, only 4 instances of heterozygosity were recorded in 3 individuals (1 individual was heterozygous at 2 loci). This was reflected in very high F_{IS} and F_{IT} values (>0.987), meaning almost all variation was outside of the individual, and *E. virginiana* is highly inbred.

Cluster analyses based on the microsatellite loci identified 20 population partitions as the most probable number of clusters in the dataset ($p = 0.99931$). The cluster phenogram (Fig 3.1) was divided into two geographically distinct clades: a northern one (blues and greens) and a southern one (reds and purples) (Fig 3.2 and 3). This pattern is very similar to the geographically structured cpDNA haplotype groups, where the north contained only members of the blue haplotype group (see chapter 1). While the cpDNA showed some extension of the predominately northern haplotypes into the south, particularly along the Appalachians, the northern microsatellite genotypes are more restricted to near or above the ice margin. The northeast is the center of diversity in the north by containing all the northern cluster types, while the midwest only harbors a subset of the northern clusters. This confirms the pattern seen in the cpDNA that there's a migration corridor connecting the northeast and midwest (rather than one from south), and is suggestive of an east-to-west directionality in gene flow.

To identify the colonization sources for each of the three geographic regions (regions 13, 9, and 6, Fig 3.1), migration rates and divergences times between regions were estimated. By increasing the number of loci from 1 (cpDNA only) to 10 (cpDNA + microsatellites), estimates were greatly improved showing smaller credible intervals that increased the ability to distinguish the population demographic parameters from different

regions (Tab 3.2). For instance, not only did the analyses confirm that immigration to the midwest was higher from the northeast than from the south ($P > 0.99$, Tab 3.4), the midwest's divergence time from the south was older than from the northeast ($P = 0.84$) (Tab 3.3). These results are consistent with the colonization route suggested by the development of host density. Furthermore, the directionality of migration in the north is answered by the migration rate estimates. Migration from the northeast to the midwest is greater than in the opposite direction ($P = 0.72$, Tab 3.4).

The genetic diversity statistics of *E. virginiana* estimated from both microsatellite and cpDNA data are largely in agreement with past cpDNA only estimates. Slightly lower levels of F_{ST} were found in the northern regions 9 and 6 than in the south (Tab 3.5). This increased genetic differentiation was also seen in the interpolated map of genetic distances between localities (EV, Fig 3.4). The distance map constructed from the full genetic dataset was qualitatively similar to the one constructed on only the cpDNA (EV, EV.c, EV.m in Fig 3.4). The south contained the highest levels of pairwise differentiation between localities, while the north was better connected with smaller genetic distances. However, the microsatellite data identified more areas of high differentiation in the northeast and at midlatitudes than the cpDNA (Fig 3.4). The upper midwest remained undifferentiated in both datasets, and low genetic distances connected it with areas slightly to the south in Indiana. The increased genetic differentiation in the south could be due to smaller population sizes. This is supported by estimates of effective population size from the IM analyses that show the southern population was much smaller than in the midwest ($P > 0.99$), though the northeast was even smaller ($P > 0.92$). The large population size in the

midwest appears to have grown from a much smaller ancestral population shared with the northeast (Tab 3.1) indicating a rapid westerly population expansion.

Spatial linear regressions of host variables against parasite genetic distances

The map layers used in the spatial linear regression analyses are presented in Figure 3.4. As described above, all the *E. virginiana* genetic distances maps were similar to one another, showing high levels of differentiation in the south, and low levels in the north (EV, EV.c, EV.m, Fig 3.4). In contrast, the genetic map layer for the host showed high levels of differentiation in the midwest and at midlatitudes (F.gen, Fig 3.4). The host pollen density time-slice layers show host pollen first restricted to the south at low density, then expanding its range northward with pockets of high density forest in the southeast, then across the south (F21-F11, Fig 3.4) At ~10 kybp, higher host densities occur in the northeast, which rapidly expand to cover most of the north (F10-F1, Fig 3.4). At the most recent time slices, host density shrinks overall (F0.5-F0, Fig 3.4). The colonization age summary pollen layer generally follows the description of the time-slice layers; the oldest inhabited region is in the southeast, which expanded westward, then jumped to the northeast, with the most recently colonized areas at the northern and western boundaries (F.age, Fig 3.4). The average density and variance in density pollen layers show similar patterns of high values in the southwest and northeast (F.avgP, F.varP, Fig 3.4).

The 23 host pollen density time-slice layers (F0-F21, Fig 3.4) were each in turn incorporated in a univariate spatial linear model predicting the parasite's genetic distances (EV). The original and normalized data produced similar results, so only the original data analyses are presented here. Most of the time-slice pollen layers produced models with significant coefficients (Fig 3.5, Tab 3.6); however, the directionality of the coefficients was

not constant. In older time-slices (>10 kybp), a positive relationship between pollen density and parasite distance was reported, while younger layers had significantly negative values (Fig 3.5). In general, though, the younger layers outperformed the older ones in weighted AIC values, where models of layers younger than 2 kybp were 91% more probable to be the best model than models based on older time-slices (Tab 3.6). The negative relationships were expected based on the results from the cpDNA only analyses (chapter 2) that showed increasing levels of parasite genetic distance related to lower average host pollen density. The most probable model was based on the 500 ybp time slice (F0.5, wAIC = 0.52), and also reflected a highly significant ($z < 0.0002$) negative relationship with parasite differentiation (Tab 3.6). The 500 ybp time slice was then used in the following multivariate parasite models.

Multivariate models that included the host genetic data were universally poor. Few had significant coefficients, and all models were outcompeted. This strongly indicates that host genetic structure has no influence on parasite patterns. Four models had significant host coefficients: (1) EV ~ F0.5, (2) EV ~ F.age, (3) EV ~ F.avgP, and (4) EV ~ F.avgP + F.age (Tab 3.7). Host density (F0.5 and F.avgP) and host colonization (F.age) coefficients were similar across models. AIC values were also similar, but the model with only F0.5 was at least 1.7 times better than the others (Tab 3.7). Recent patterns in host density appear to have dictated parasite range expansion.

In the southern range, no models had significant host coefficients (Tab 3.7), so the null model of no relationship (with only the spatial lag term) could not be rejected. In the north, the pollen based models (F0.5, F.avgP) outcompeted models that included the host

genetic data, though only the pollen time-slice model (F0.5) produced a significant regression coefficient (Tab 3.7).

Migration corridor analysis

Because the host pollen time-slice layer at 500 ybp (F0.5) best predicted the parasite's genetic distance in the overall range, it was chosen to be the habitat map for the migration corridor analyses. Habitat cost models were constructed as described above by applying various thresholds and weighting schemes (Tab 3.8). The Mantel tests showed that the models performed very similarly, and produced correlation coefficients narrowly distributed from 0.30 to 0.36. Both measures of genetic distance, Reynolds and F_{ST}, produced similar results with the 50 and 60th quantile thresholds generating the best estimates of effective geographic distances in each of the weighting schemes. The best weighting scheme in both cases was a factor of 0.30 multiplied to the lower threshold values (Fig 3.6).

Plots of the migration corridors based on the two best habitat models were very similar (Fig 3.7). High connectivity was found south of the Great Lakes connecting the northeast and midwest, and a north-south corridor existed between the upper and lower midwest. Dispersal patterns in the south were generally diffuse with few common corridors. An exception is in the southern Appalachians where connectivity along the mountain range is high.

Discussion

Reproductive biology of E. virginiana

A byproduct of the genetic data collected for this study was the revelation of very few heterozygotes of *E. virginiana* with correspondingly high values of F_{IS} and F_{IT} that

indicate this parasite is highly inbred. This result confirms early studies of the morphology and life history of the parasite, which note that the flowers are primarily cleistogamous, some plants never emerge above the duff, and that seed set is likely a result of selfing (Schrenk 1894; Leavitt 1902; Cooke and Schively 1904). *E. virginiana* being a selfer greatly reduces the complexity of this system and helps explain the clarity and congruence of the multiple genetic loci. Because pollen flow is extremely reduced in this system, all loci reflect primarily seed dispersal effects, which is ideal for tracking range expansion.

Host density not host range expansion controls parasite range expansion overall

The parasite's population genetic structure was heavily influenced by spatial changes in host density. This is seen in the significant relationships between the parasite genetic distance map (EV) and multiple map layers based on aspects of host density (e.g. F0.5, F.age, F.avgP; Tab 3.7). The most influential time period on the genetics of *E. virginiana* was very recent, at 500 ybp (F0.5, Tab 3.6). This suggests that the parasite is highly sensitive to host density changes and was able to react in a short amount of time. On the other hand, because many of the recent host density time-slice layers are very similar (Fig 3.4), the overall genetic patterns may have been established earlier, though still amenable to small changes. For instance, host density layers as early as 7 kybp show marginally significant negative relationships to the genetic structure of the parasite (Tab 3.6, Fig 3.5). Fluctuations in host density did not have as large an effect as host density itself, as seen in the better fit of models with F0.5 than F.varP.

While the parasite map layer was correlated with many of the host density related layers, no relationship was found between the parasite and host genetic layers. None of the models produced significant regression coefficients between them. This indicates that the

migration history of *F. grandifolia*, which produced the genetic geospatial patterns, was very different from the migration routes taken by the parasite. This result combined with the parasite's dependency on host density suggests that it significantly lagged behind the migration front of the host. From the pollen data, the initial range expansion of the host proceeded at low densities (Bennett 1985; Williams et al. 2004). If the parasite dispersed more slowly than the host, migration corridors of the host into new areas would not have been accessible regardless of host density there. By the time the parasite was capable of utilizing the host's migration corridor, the host would have colonized large parts of the landscape, perhaps with different host densities, making available alternate migration routes to the parasite.

No host variables relate to parasite population structure in the south

From the overall and northern linear models, the parasite is sensitive to host density variation as it expands its range. However, in the south this pattern has been erased, perhaps due to proximity to many potential source refuges clouding the pattern, or due to the more stable environment of the south. The southern range has remained unglaciated over the past 120,000 years (Dyke et al. 2003), so parasite patterns there could be complex and the result of multiple glacial/interglacial cycles. In contrast, the northern distribution was glaciated 18 kybp (Dyke et al. 2003), so northern populations are new arrivals and have a shorter and more interpretable colonization history. In support of this, only the most recent host pollen time-slice layers had a significant negative relationship with the parasite (Tab 3.6). The southern parasite patterns represent many accumulated patterns over a much longer time period that overlay and obscure relationships to the host's biology and history.

Location and directionality of migration corridors of E. virginiana

The major migration corridor for *E. virginiana* connects the northeast and midwest near the southern border of the Great Lakes (Fig 3.7). This route is primarily a westward one, as indicated by higher migration rates from east to west (Tab 3.2). Interestingly, a path cutting between lakes Ontario and Erie is less well traveled than this more southerly route, which provides a hypothesis to test in future studies of migration corridors near the Great Lakes. Studies in other systems, such as the herb *Boechera* and butterfly *Lycaeides* have documented northern refuges in this area (Nice et al. 2005; Kiefer et al. 2009).

Another migration corridor is seen connecting the southern and northern parts of the midwest (Fig 3.7). While the estimates of migration rates indicate low levels of migrants between the south and midwest in both directions (Tab 3.2), this area in Indiana and northern Kentucky may represent a suture zone between the north and south. Slight northward dispersal is evident here in the geographic distributions of several southern haplotypes and genotypes that extend across the ice margin into this area (Fig 3.3). In parallel, the northern haplotypes and genotypes that dominate here likely were the result of southward dispersal into this area.

A migration corridor was also identified in the southern Appalachians. This suggests that relatively high elevation was not a boundary for parasite gene flow. This mirrors the topographic distribution of the host. In the south, *F. grandifolia* is commonly found at relatively higher elevations (Fowells 1965), so mountain ranges are not expected to serve as genetic barriers.

Conclusions and future directions

Recent spatial patterns in host density best predict *E. virginiana*'s population structure over other host variables such as host migration routes, average host density, variance in host density, and host arrival date. These broad patterns held true in the northern part of the range, but in the south no host variables could predict the parasite's patterns, perhaps resulting from complex glacial processes older than the 21 kybp considered in the study. The parasite's population differentiation was highest in the south corresponding with small effective population sizes and low host densities. Migration proceeded first from the south to the northeast and then followed an east-west corridor into the midwest.

While the relationship between parasite and host distributions is now better understood, there is still a missing connection between a parasite's range and changes in climate. Ecological niche models could be produced and extrapolated back in time to more fully explore the lag time between habitat availability, host movements, and parasite colonization. The spatial linear model framework for phylogeography is flexible and an additional habitat availability variable could be incorporated. This approach could give more meaningful estimates of migration capacity. Also, the predictions of a threshold host density limiting *E. virginiana*'s establishment should be further investigated at different spatial scales. Parasite germination likely requires exudates from the growing tip of a host root (Williams and Zuck 1986), and the parasite must establish physical contact with a host root soon after germination for survival (Musselman and Press 1995). Based on these requirements, it is expected that a certain host root mass (and thus host density) is necessary to ensure a high probability of yearly reinfection by this annual parasite. By enumerating the threshold host

population density the amount of establishment limitation in the system could be directly estimated.

The highly structured populations of *E. virginiana* suggest that the parasite could be locally adapted to its host. Coevolutionary theory predicts parasite local adaptation in cases where its gene flow is restricted relative to the host's (Price 1980; Thompson 1994). Conversely, if the parasite's gene flow is much greater than its host's, locally adapted genotypes are expected to be inundated with genotypes from the rest of the range, so local adaptation is not expected. From the broad scale information on the genetic diversity of *E. virginiana* and *F. grandifolia* genetically distinct populations can be selected for local adaptation experiments.

Understanding what controls the spread of parasites is especially relevant today considering that future climate change may force plants to move at their full migration capacities (Davis and Zabinski 1992). The results of this study are especially applicable because the *F. grandifolia* community is predicted to be greatly affected by climate change(Iverson and Prasad 2002). In addition to this specific application, the *E. virginiana*/*F. grandifolia* system brings understanding to the general situation of associated species colonizing new environments. How the range of a species changes relative to that of a close associate is central to understanding the cohesiveness of communities and how larger assemblages of organisms move.

Table 3.1: *Epifagus virginiana* genetic data and global diversity statistics by locus.

Data type	Locus	Size (bp)	# alleles	H_I	H_S	H_T	F_{IS}	F_{IT}	F_{ST}
microsat	MS10	177-187	5	0	0.251	0.518	1	1	0.516
microsat	MS34	218-259	17	0.003	0.389	0.848	0.993	0.997	0.542
microsat	MS43	221-303	27	0	0.449	0.925	1	1	0.514
microsat	MS53	113-136	14	0	0.395	0.825	1	1	0.521
microsat	MS76	173-182	6	0	0.282	0.462	1	1	0.390
microsat	MS92	345-362	3	0	0.244	0.418	1	1	0.416
microsat	MS105	143-161	8	0.003	0.290	0.703	0.991	0.996	0.587
microsat	MS130A	260-266	4	0.004	0.336	0.702	0.987	0.994	0.520
microsat	MS135	164-184	4	0	0.156	0.193	1	1	0.191
cpDNA	rbcL + clpP1	1016	40	NA	0.297	0.852	NA	NA	0.652
Average				0.001	0.309	0.645	0.997	0.998	0.521

Table 3.2: Population demographic parameters relating to geographic regions of *Epifagus virginiana*. Regions are defined in Fig 3.1 and in text. θ_1, θ_2, and θ_A refer to the effective population sizes of region 1, region 2, and the ancestral population respectively. $m_{2->1}$ and $m_{1->2}$ are the migration rates from region 2 to region 1 and from region 1 to region 2 respectively. t is the divergence time between the regions. The 95% credible interval is given below each estimate.

region1	region2	θ_1	θ_2	θ_A	$m_{2->1}$	$m_{1->2}$	t
13	9	0.73 (0.2-1.7)	0.25 (0.08-0.59)	9.76 (4.95-26.72)	0.59 (0.11-5.90)	5.05 (0.62-13.75)	0.06 (0.02-0.37)
13	6	1.94 (1.27-3.02)	12.93 (10.03-21.15)	5.04 (3.42-21.15)	1.03 (0.38-2.23)	0.04 (0.01-2.02)	0.36 (0.30-3.90)
9	6	0.42 (0.08-0.59)	239.49 (47.80-236.38)	8.59 (5.39-46.12)	6.34 (3.11-11.20)	8.17 (6.28-11.63)	0.28 (0.18-1.28)

Table 3.3: Probabilities that a divergence time between regions of *Epifagus virginiana* is greater than another. Each value is the probability that the divergence time of the row heading is greater than that of the column. Probabilities were estimated by randomly sampling and comparing posterior distributions of estimates shown in Tab 3.2. Regions used in comparisons are as defined in Fig 3.1 and in text.

	13 vs 9	13 vs 6	9 vs 6
13 vs 9	0.00	0.01	0.04
13 vs 6	0.99	0.00	0.84
9 vs 6	0.96	0.16	0.00

Table 3.4: Probabilities that a migration rate from one region of *Epifagus virginiana* is greater than another. Each value is the probability that the migration rate of the row heading is greater than that of the column. Headings refer to migration rates from a region into another. Regions used in comparisons are as defined in Fig 3.1 and in text. Probabilities were estimated by randomly sampling and comparing posterior distributions of estimates shown in Tab 3.2.

	9->13	6->13	13->9	6->9	13->6	9->6
9->13	0.00	0.53	0.10	0.03	0.86	0.01
6->13	0.46	0.00	0.06	0.00	0.89	0.00
13->9	0.90	0.94	0.00	0.47	0.98	0.28
6->9	0.97	1.00	0.56	0.00	1.00	0.18
13->6	0.15	0.11	0.02	0.00	0.00	0.00
9->6	0.99	1.00	0.72	0.81	1.00	0.00

Table 3.5: Genetic diversity statistics for different geographic regions of *Epifagus virginiana*. For regional definitions see Fig 3.1.

Region	H_I	H_S	H_T	F_{IS}	F_{IT}	F_{ST}
13	0.000	0.277	0.603	1.000	1.000	0.540
6	0.001	0.360	0.640	0.997	0.998	0.437
9	0.003	0.265	0.519	0.989	0.994	0.489
Entire	0.001	0.309	0.645	0.997	0.998	0.521

Table 3.6: Univariate host pollen density time-slice models predicting genetic distances of *Epifagus virginiana*. Coefficients and their corresponding Z-values refer to host data layer coefficients, while ϱ is the spatial lag coefficient. All values of ϱ were significant ($p < 0.01$), and spatial lag models were significantly superior to standard models with no spatial components. The remaining residuals in any of the models were not spatially correlated as indicated by Lagrange multipliers. AIC, ΔAIC and weighted AIC values are reported following Wagenmakers and Farell (2004). The three models with >0.10 in probability are in bold; the best model, F0.5, is highlighted in red.

Time Slice	Intercept	Coef	Z-value	ϱ	AIC	ΔAIC	wAIC
F00	0.04281	-0.00189	0.002	0.949	-5355.3	5.1	0.04
F0.5	**0.04669**	**-0.00109**	**0.000**	**0.945**	**-5360.4**	**0.0**	**0.52**
F01	**0.04536**	**-0.00083**	**0.000**	**0.946**	**-5358.6**	**1.7**	**0.22**
F02	**0.04518**	**-0.00073**	**0.001**	**0.946**	**-5357.8**	**2.6**	**0.14**
F03	0.04173	-0.00062	0.003	0.950	-5354.9	5.4	0.03
F04	0.03783	-0.00048	0.019	0.954	-5351.7	8.7	0.01
F05	0.03715	-0.00044	0.024	0.955	-5351.2	9.1	0.01
F06	0.03628	-0.00047	0.029	0.956	-5350.9	9.5	0.00
F07	0.03434	-0.00044	0.090	0.958	-5349.0	11.4	0.00
F08	0.03324	-0.00060	0.204	0.959	-5347.7	12.6	0.00
F09	0.03048	0.00001	0.993	0.961	-5346.1	14.3	0.00
F10	0.03339	0.00207	0.041	0.956	-5350.3	10.1	0.00
F11	0.03402	0.00103	0.037	0.955	-5350.5	9.9	0.00
F12	0.03479	0.00034	0.031	0.955	-5350.8	9.6	0.00
F13	0.03246	0.00015	0.183	0.958	-5347.9	12.5	0.00
F14	0.03417	0.00055	0.047	0.955	-5350.1	10.3	0.00
F15	0.03489	0.00170	0.012	0.954	-5352.5	7.9	0.01
F16	0.03258	0.00337	0.064	0.958	-5349.6	10.8	0.00
F17	0.03135	0.00659	0.188	0.960	-5347.9	12.5	0.00
F18	0.03027	-0.00300	0.804	0.962	-5346.2	14.2	0.00
F19	0.03288	0.03297	0.074	0.957	-5349.3	11.1	0.00
F20	0.03067	0.00379	0.743	0.961	-5346.2	14.1	0.00
F21	0.03066	0.00844	0.813	0.961	-5346.2	14.2	0.00

Table 3.7: Best spatial regression models of host variables on *Epifagus virginiana* genetic distances. Models were based on regional subsets of the data as defined in Fig 3.1, or on the complete dataset. *Fagus grandifolia* data layers are from Fig 3.4, and the "Null" model refers to a model with only the spatial lag explanatory variable (EV ~ 1). Column headings are explained in Tab 3.6. Best models in each category are highlighted in red. Only models with marginally significant or significant regression coefficients are presented except in the case of the southern models. For the south, the best 3 (and the null) models are shown.

Region	Model (EV ~)	Intercept	Coef	Z-value	ϱ	AIC	ΔAIC	wAIC
South	1 (Null)	0.05048	NA	NA	0.940	-2260.5	1.1	0.18
	F.age	0.06784	-0.00136	0.113	0.942	-2261.0	0.6	0.24
	F.age + F.gen	**0.07443**	**-0.00178**	**0.047**	**0.938**	**-2261.7**	**0.0**	**0.32**
			0.00601	**0.104**				
	F0.5 + F.age + F.gen	0.07590	-0.00712	0.217	0.935	-2261.2	0.5	0.26
			-0.00166	0.064				
			0.00884	0.042				
North	**F0.5**	**0.05031**	**-0.00088**	**0.047**	**0.939**	**-3035.8**	**0.0**	**0.65**
	F.avgP	0.04885	-0.00118	0.099	0.939	-3034.6	1.2	0.35
Total	**F0.5**	**0.04669**	**-0.00109**	**0.000**	**0.945**	**-5360.4**	**0.0**	**0.59**
	F.age	0.03417	0.00061	0.002	0.949	-5355.9	4.5	0.06
	F.avgP	0.03670	-0.00103	0.034	0.957	-5350.6	9.8	0.00
	F.avgP + F.age	0.04195	-0.00116	0.020	0.943	-5359.3	1.1	0.35
			0.00065	0.001				

Table 3.8: Performance of habitat cost models based on *Fagus grandifolia* pollen densities at 500 ybp. Threshold values relate to the 25, 40, 50, 60, and 75th percentiles of the pollen density distribution, and weights were multiplied to densities lower than the set threshold to obtain habitat costs. The resulting log of effective geographic distances between populations were compared with two separate genetic distance calculations (Reynolds' and F_{ST}). Resulting Mantel statistics (Mantel) are reported here. All correlations are significantly nonzero. Best two models from the Reynolds' and F_{ST} analyses are in red.

Threshold quantile	Threshold pollen %	Weight	Mantel (Rey)	Mantel (F_{ST})
25	0.23	1.0000	0.329	0.327
25	0.23	0.1000	0.334	0.338
25	0.23	0.0100	0.335	0.340
25	0.23	0.0100	0.335	0.340
25	0.23	0.0010	0.334	0.341
25	0.23	0.0001	0.332	0.340
40	1.77	0.0900	0.351	0.341
40	1.77	0.0700	0.351	0.340
40	1.77	0.0500	0.352	0.340
40	1.77	0.0300	0.352	0.339
40	1.77	0.0100	0.352	0.337
40	1.77	0.0100	0.352	0.337
50	2.85	1.0000	0.329	0.327
50	2.85	0.1000	0.351	0.343
50	2.85	0.0900	0.352	0.344
50	2.85	0.0700	0.352	0.344
50	2.85	0.0500	0.353	0.344
50	**2.85**	**0.0300**	**0.354**	**0.345**
50	2.85	0.0100	0.354	0.344
50	2.85	0.0100	0.354	0.344
50	2.85	0.0010	0.353	0.341
50	2.85	0.0001	0.350	0.338
60	3.81	0.0900	0.353	0.343
60	3.81	0.0700	0.354	0.343
60	3.81	0.0500	0.355	0.343
60	**3.81**	**0.0300**	**0.356**	**0.342**

60	3.81	0.0100	0.355	0.339
60	3.81	0.0100	0.355	0.339
75	5.39	1.0000	0.329	0.327
75	5.39	0.1000	0.337	0.327
75	5.39	0.0100	0.327	0.310
75	5.39	0.0100	0.327	0.310
75	5.39	0.0010	0.314	0.293
75	5.39	0.0001	0.303	0.279
NA	NA	exponential	0.350	0.334
NA	NA	logarithmic	0.305	0.305

Table 3.8, continued.

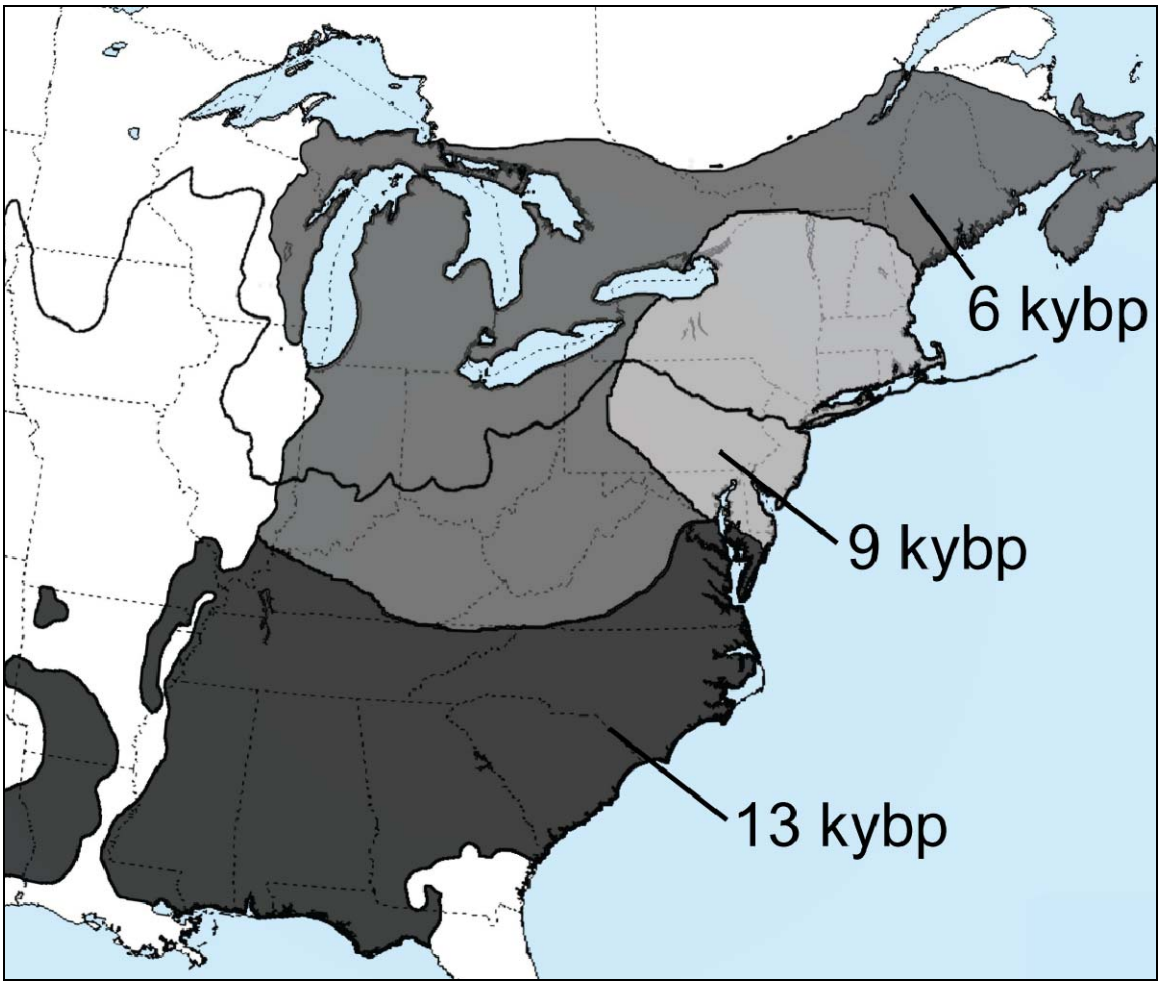

Figure 3.1: Regional definitions used in the study of the population structure of *Epifagus virginiana*. Regions were defined by the geographic distributions of >2% host pollen at 13, 9, and 6 kybp.

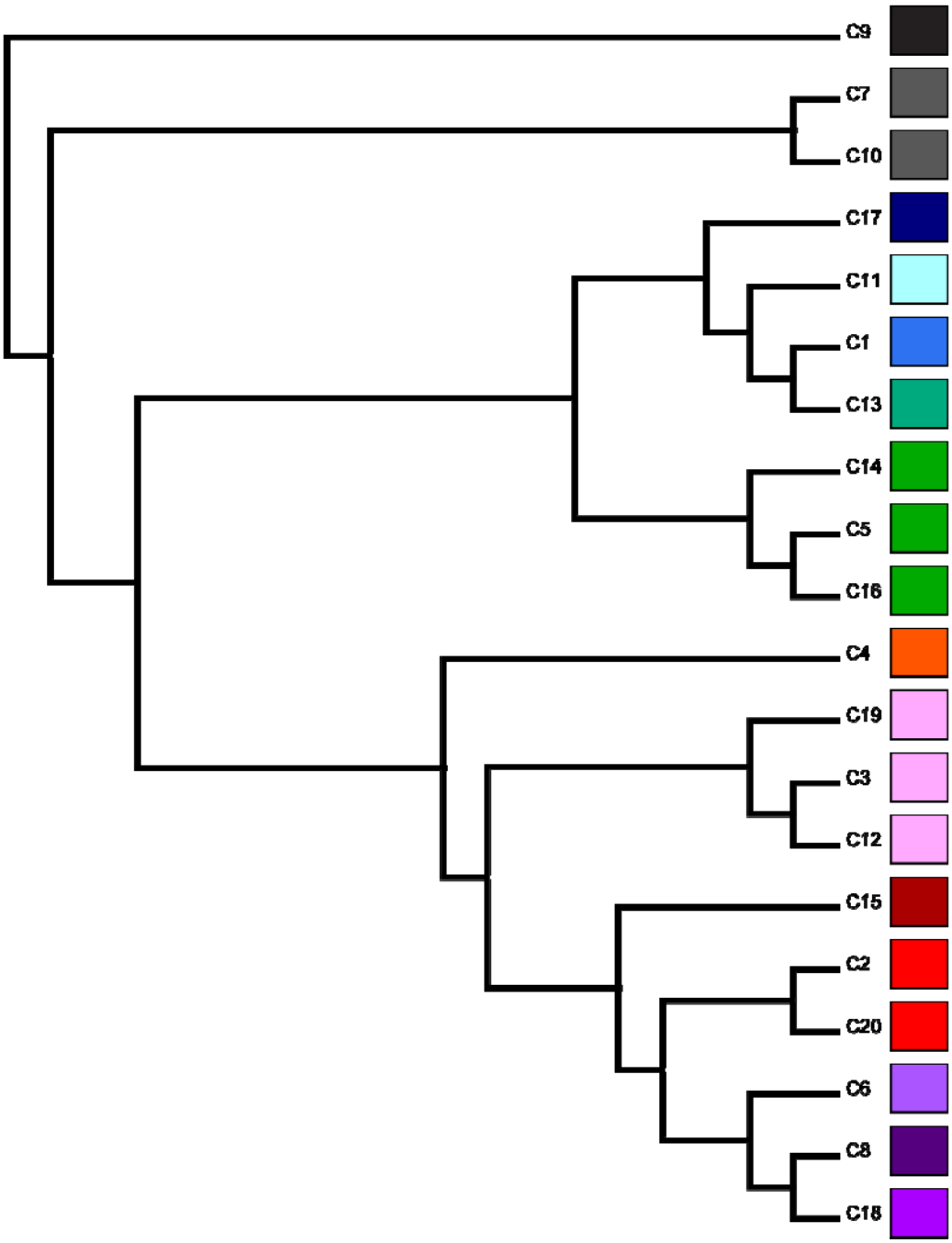

Figure 3.2: Cluster phenogram of *Epifagus virginiana* genotypes from 9 microsatellite loci. Population assignments were made using a spatial Bayesian clustering methodology in BAPS (Corander et al. 2008). Colors were chosen based on the relationships in the dendrogram. Cluster C9 (black) was only found in the Mexican population, and not included in any further analyses.

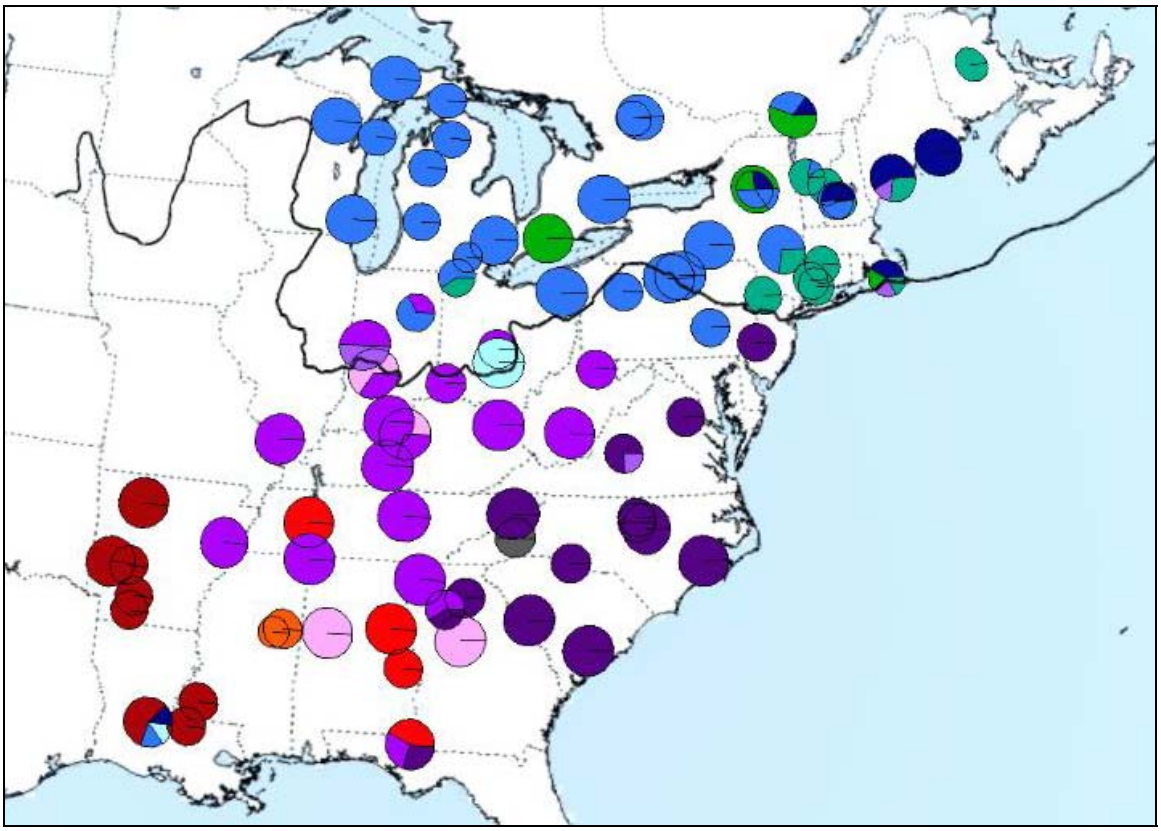

Figure 3.3: Geographic distribution of *Epifagus virginiana* microsatellite clusters. Colors correspond to clusters in Fig 3.2.

Figure 3.4: Data layers used in the spatial linear models of the relationship between *Epifagus virginiana* and different aspects of its host, *Fagus grandifolia*, biology. Data layers of genetic distances (row 1): The response variable, *E. virginiana*, based on the full dataset (EV), only the cpDNA (EV.c), only the microsatellite data (EV.m), and an explanatory variable of host genetic distance based on cpDNA (F.gen). Host pollen time-slices (rows 2-7): Host pollen layers from the present (F0), 500 ybp (F0.5), then in 1000 year increments to 21 kybp (F1-F21). Pollen summary layers (row 8): colonization age (F.age), average pollen density (F.avgP), and variance in pollen density through time (F.varP). All maps are shown in terrain colors where cooler greens represent smaller values and warmer pinks and whites correspond to high values.

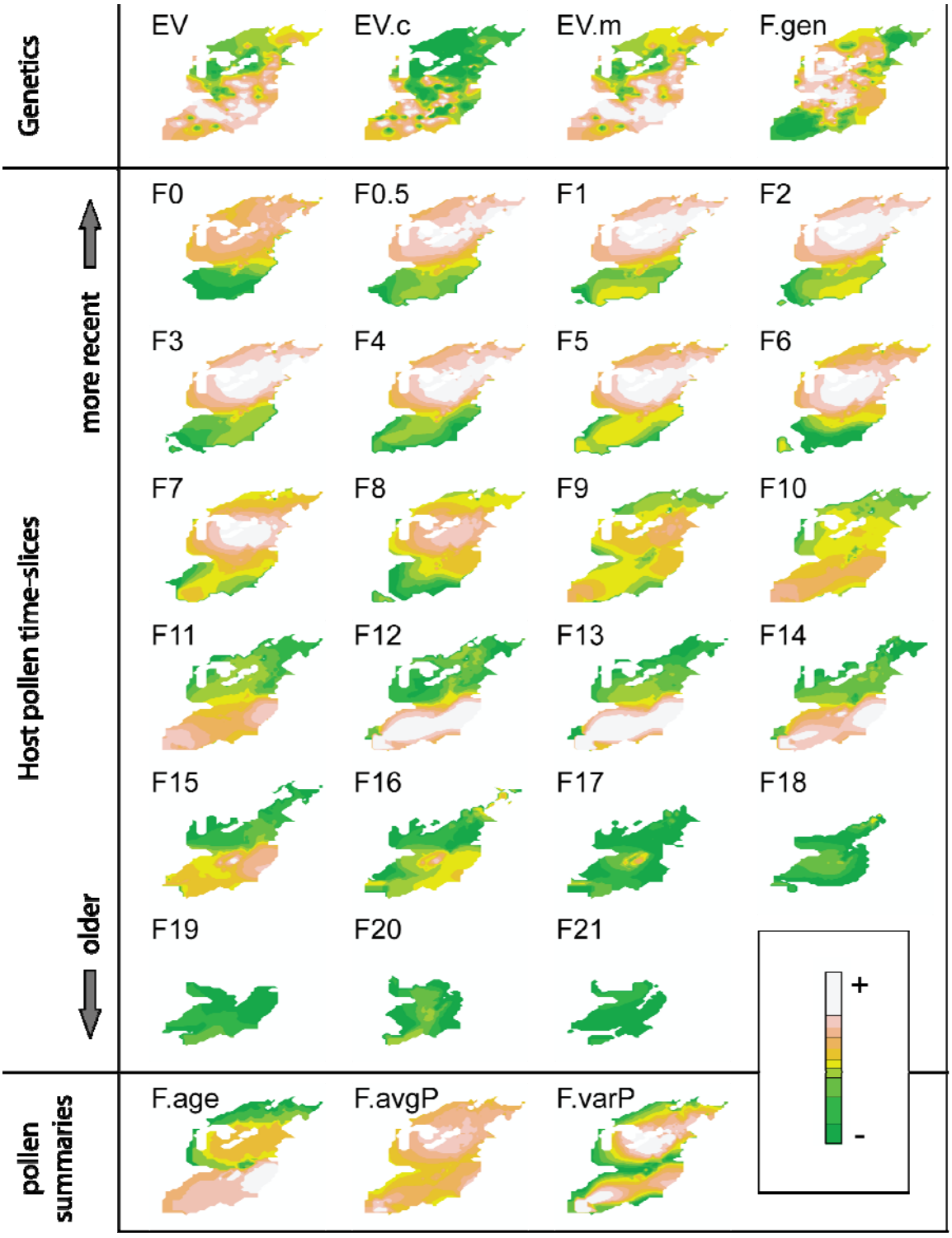

Figure 3.4: Data layers used in the spatial linear models of the relationship between *Epifagus virginiana* and different aspects of its host, *Fagus grandifolia*, biology.

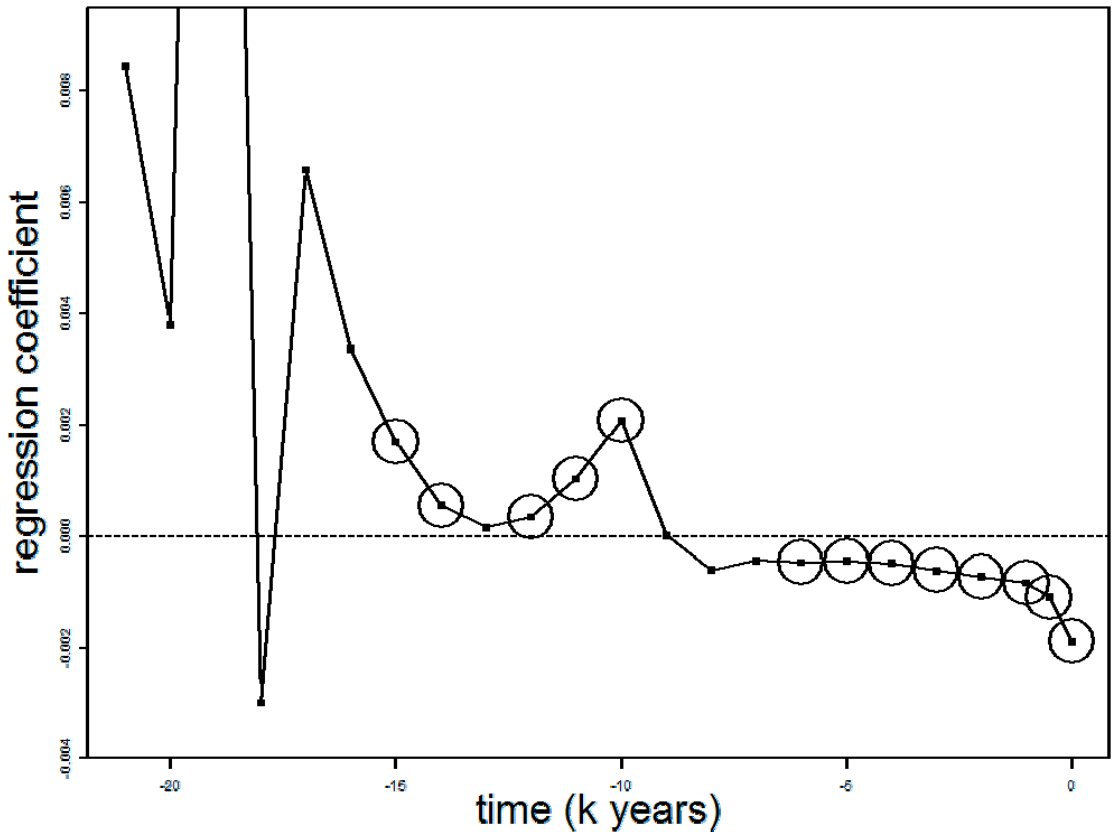

Figure 3.5: Changes through time in regression coefficients of *Fagus grandifolia* pollen layers in linear models predicting genetic distances of *Epifagus virginiana*. Results are shown from spatial linear models of untransformed, original data. Significant coefficients are circled, and the graph proceeds from the oldest time-slice (F21) to the present (F0). The dotted line is at zero for reference.

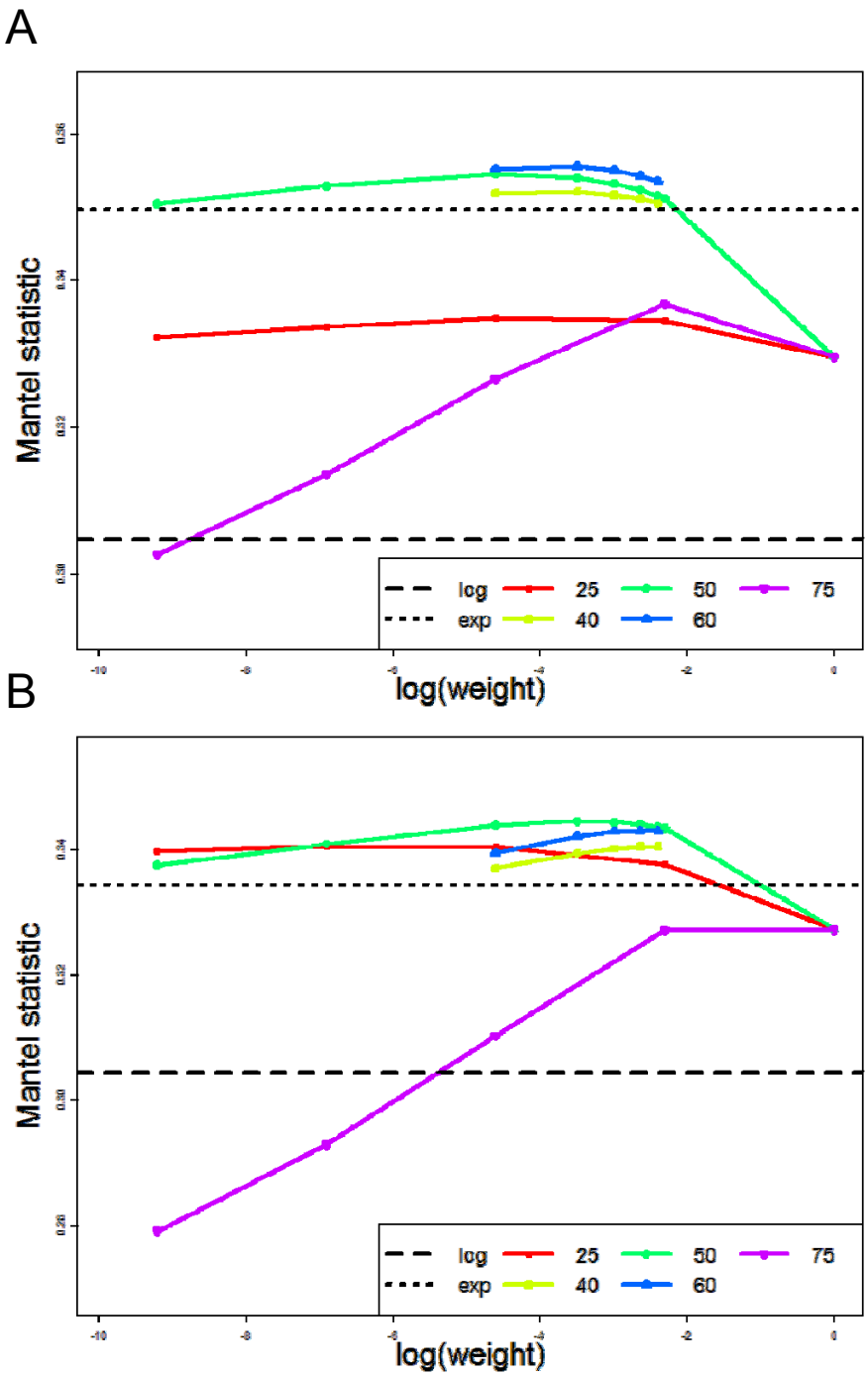

Figure 3.6: Performance of various Fagus grandifolia habitat cost models measured by Mantel test correlation coefficients. Mantel tests compared the log of the resulting effective geographic distances versus the (A) Reynolds' distances or (B) F_{ST}'s between populations of *Epifagus virginiana*. Lines connect models of the same threshold value with differing weighting schemes along the x-axis. Dotted horizontal lines correspond with the exponential and logarithmic weight models.

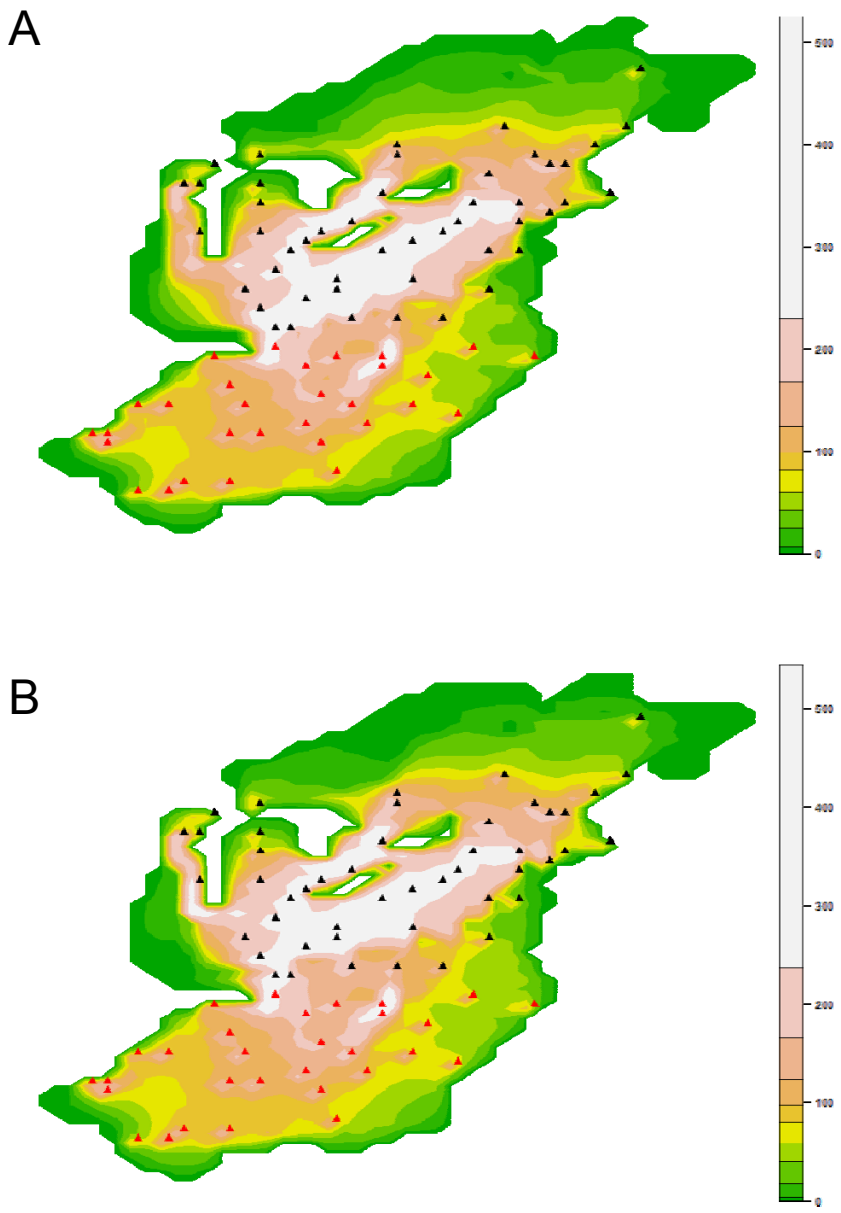

Figure 3.7: Migration corridors of *Epifagus virginiana* based on the best habitat cost models of Fagus grandifolia pollen densities at 500 ybp. (A) Cost model based on a 3.81% pollen threshold (60th quantile) and 0.03 weighting scheme indicated as the best model when tested against the parasite's Reynolds' genetic distances. (B) Cost model based on a 2.85% pollen threshold (50th quantile) and 0.03 weighting scheme indicated as the best model when tested against the parasite's pairwise F_{ST}s. Triangles are *E. virginiana* collection localities. Black and red triangle colors distinguish the northern (regions 9 & 6) and southern (region 13) ranges (see Fig 3.1 for region definitions).

References

Arneberg, P., A. Skorping, B. Grenfell and A. F. Read (1998). "Host densities as determinants of abundance in parasite communities." Proceedings: Biological Sciences **265**(1403): 1283-1289.

Barkman, T., J. McNeal, S.-H. Lim, G. Coat, H. Croom, N. Young and C. dePamphilis (2007). "Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants." BMC Evolutionary Biology **7**(1): 248.

Beaumont, M. A. and M. Panchal (2008). "On the validity of nested clade phylogeographical analysis." Molecular Ecology **17**(11): 2563-2565.

Beerli, P. (2006). "Comparison of Bayesian and maximum-likelihood inference of population genetic parameters." Bioinformatics **22**(3): 341-345.

Beerli, P. and J. Felsenstein (1999). "Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach." Genetics **152**(2): 763-773.

Beerli, P. and J. Felsenstein (2001). "Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach." Proceedings of the National Academy of Sciences **98**(8): 4563-4568.

Beier, P., D. R. Majka and W. D. Spencer (2008). "Forks in the road: Choices in procedures for designing wildland linkages." Conservation Biology **22**(4): 836-851.

Bennett, J. R. and S. Mathews (2006). "Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A." American Journal of Botany **93**(7): 1039-1051.

Bennett, K. D. (1985). "The spread of *Fagus grandifolia* across eastern North America during the last 18000 years." Journal of Biogeography **12**(2): 147-164.

Botanga, C. J. and M. P. Timko (2007). "Genetic structure and analysis of host and nonhost interactions of *Striga gesnerioides* (Witchweed) from central Florida." Phytopathology **95**(10): 1166-1173.

Box, G. E. P. and D. R. Cox (1964). "An analysis of transformations." Journal of the Royal Statistical Society. Series B (Methodological) **26**(2): 211-252.

Camp, W. H. (1950). "A biogeographic and paragenetic analysis of the American beech (*Fagus*)." American Philosophical Society Yearbook **1950**: 166-169.

Carstens, B. C., S. J. Brunsfeld, J. R. Demboski, J. M. Good and J. Sullivan (2005). "Investigating the evolutionary history of the pacific northwest mesic forest ecosystem: Hypothesis testing within a comparative phylogeographic framework." Evolution **59**(8): 1639-1652.

Chen, C., E. Durand, F. Forbes and O. Francois (2007). "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study." Molecular Ecology Notes **7**(5): 747-756.

Clark, J. S., C. Fastie, G. Hurtt, S. T. Jackson, C. Johnson, G. A. King, M. Lewis, L. Jason, S. Pacala, C. Prentice, E. W. Schupp, Thompson Webb, III and P. Wyckoff (1998). "Reid's paradox of rapid plant migration." BioScience **48**(1): 13-24.

Clark, J. S. and J. S. McLachlan (2003). "Stability of forest biodiversity." Nature **423**(6940): 635-638.

Clement, M., D. Posada and K. Crandall (2000). "TCS: a computer program to estimate gene genealogies." Molecular Ecology **9**(10): 1657-1660.

Comes, H. P. and J. W. Kadereit (1998). "The effect of Quaternary climatic changes on plant distribution and evolution." Trends in Plant Science **3**(11): 432-438.

Cooke, E. and A. F. Schively (1904). "Observations on the structure and development of Epiphegus Virginiana." Contributions from the Botanical Laboratory of the University of Pennsylvania **2**: 352-398 + plates.

Corander, J. and P. Marttinen (2006). "Bayesian identification of admixture events using multilocus molecular markers." Molecular Ecology **15**(10): 2833-2843.

Corander, J., J. Siren and E. Arjas (2008). "Bayesian spatial modeling of genetic population structure." Computational Statistics **23**(1): 111-129.

Corander, J. and J. Tang (2007). "Bayesian analysis of population structure based on linked molecular information." Mathematical Biosciences **205**(1): 19-31.

Corander, J., P. Waldmann and M. J. Sillanpaa (2003). "Bayesian analysis of genetic differentiation between populations." Genetics **163**(1): 367-374.

Criscione, C. and M. Blouin (2007). "Parasite phylogeographical congruence with salmon host evolutionarily significant units: implications for salmon conservation." Molecular Ecology **16**: 993-1005.

Cruzan, M. B. and A. R. Templeton (2000). "Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record." Trends in Ecology & Evolution **15**(12): 491-496.

Cunningham, C. W. and T. M. Collins (1994). Developing model systems for molecular biogeography: Vicariance and interchange in marine invertebrates. Molecular Ecology and Evolution: Approaches and Applications. B. Schierwater, B. Streit, P. Wagner and R. DeSalle. Basel, Switzerland, Birkhauser Verlag**:** 405-433.

Cushman, S. A., Kevin S. McKelvey, J. Hayden and Michael K. Schwartz (2006). "Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling." The American Naturalist **168**(4): 486-499.

Davis, M. B. (1981). Quaternary history and the stability of forest communities. Forest succession: concepts and application. D. C. West, H. H. Shugart and D. B. Botkin. New York, Springer**:** 134-153.

Davis, M. B. (1983). "Quaternary history of deciduous forests of eastern North-America and Europe." Annals of the Missouri Botanical Garden **70**(3): 550-563.

Davis, M. B. and C. Zabinski (1992). Changes in geographical range resulting from greenhouse warming: Effects on biodiversity in forests. Global Warming and Biodiversity. R. L. Peters and T. E. Lovejoy. New Haven, CT, Yale University Press**:** 297-308.

de Vega, C., R. Berjano, M. Arista, P. L. Ortiz, S. Talavera and T. F. Stuessy (2008). "Genetic races associated with the genera and sections of host species in the holoparasitic plant *Cytinus* (Cytinaceae) in the Western Mediterranean basin." New Phytologist **178**(4): 875-887.

Delcourt, P. A. and H. R. Delcourt (1987). Long term forest dynamics of the temperate zone: a case study of late-Quaternary forests in eastern North America. New York, New York, USA, Springer-Verlag.

dePamphilis, C. and J. Palmer (1990). "Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant." Nature **348**: 337 - 339.

Doyle, J. J. and J. L. Doyle (1990). "Isolation of plant DNA from fresh tissue." Focus **12**: 13-15.

Duff, R. J. and D. L. Nickrent (1997). "Characterization of mitochondrial small-subunit ribosomal RNAs from holoparasitic plants." Journal of Molecular Evolution **45**(6): 631-639.

Dukes, J. S. and H. A. Mooney (1999). "Does global change increase the success of biological invaders?" Trends in Ecology & Evolution **14**(4): 135-139.

Dyer, R. J. and J. D. Nason (2004). "Population graphs: The graph theoretic shape of genetic structure." Molecular Ecology **13**(7): 1713-1727.

Dyke, A. S., A. Moore and L. Robertson (2003). Deglaciation of North America. Geological Survey of Canada. **Open File 1574:** 2 sheets.

Epps, C. W., J. D. Wehausen, V. C. Bleich, S. G. Torres and J. S. Brashares (2007). "Optimizing dispersal and corridor models using landscape genetics." Journal of Applied Ecology **44**(4): 714-724.

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.

Faubet, P. and O. E. Gaggiotti (2008). "A new Bayesian method to identify the environmental factors that influence recent migration." Genetics **178**(3): 1491-1504.

Fowells, H. A. (1965). American beech (*Fagus grandifolia* Ehrh.). Silvics of Forest Trees of the United States, U.S. Dept. Agric. Forest Service**:** 172-180.

Franke, R. (1982). "Scattered data interpolation: Tests of some methods." Mathematics of Computation **38**(157): 181-200.

Fujii, N., K. Ueda, Y. Watano and T. Shimizu (1997). "Intraspecific sequence variation of chloroplast DNA in *Pedicularis chamissonis* Steven (Scrophulariaceae) and geographic structuring of the Japanese "Alpine" plants." Journal of Plant Research **110**(2): 195-207.

Garrick, R. C., R. J. Dyer, L. B. Beheregaray and P. Sunnucks (2008). "Babies and bathwater: a comment on the premature obituary for nested clade phylogeographical analysis." Molecular Ecology **17**(6): 1401-1403.

Gonzales, E., J. L. Hamrick and S.-M. Chang (2008). "Identification of glacial refugia in south-eastern North America by phylogeographical analyses of a forest understory plant, *Trillium cuneatum*." Journal of Biogeography **35**(5): 844-852.

Guillot, G., A. Estoup, F. Mortier and J. F. Cosson (2005). "A spatial statistical model for landscape genetics." Genetics **170**(3): 1261-1280.

Guillot, G., F. Mortier and A. Estoup (2005). "Geneland: a computer package for landscape genetics." Molecular Ecology Notes **5**(3): 712-715.

Hewitt, G. (2000). "The genetic legacy of the Quaternary ice ages." Nature **405**(6789): 907-913.

Hewitt, G. M. (1996). "Some genetic consequences of ice ages, and their role, in divergence and speciation." Biological Journal of the Linnean Society **58**(3): 247-276.

Hey, J. and R. Nielsen (2004). "Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*." Genetics **167**: 747-760.

Hey, J. and R. Nielsen (2007). "Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics." Proceedings of the National Academy of Sciences **104**(8): 2785-2790.

Hickerson, M., E. Stahl and N. Takebayashi (2007). "msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation." BMC Bioinformatics **8**(1): 268.

Hickerson, M. J., E. A. Stahl and H. A. Lessios (2006). "Test for simultaneous divergence using approximate Bayesian computation." Evolution **60**(12): 2435-2453.

Holzapfel, S., M. Z. Faville and C. E. C. Gemmill (2002). "Genetic variation of the endangered holoparasite *Dactylanthus taylorii* (Balanophoraceae) in New Zealand." Journal of Biogeography **29**(5-6): 663-676.

Huntley, B., P. J. Bartlein and I. C. Prentice (1989). "Climatic control of the distribution and abundance of beech (Fagus L.) in Europe and North America." Journal of Biogeography **16**(6): 551-560.

Ibrahim, K. M., R. A. Nichols and G. M. Hewitt (1996). "Spatial patterns of genetic variation generated by different forms of dispersal during range expansion." Heredity **77**(3): 282-291.

Inman, R. M. and M. R. Pelton (2002). "Energetic production by soft and hard mast foods of American black bears in the Smoky Mountains." Ursus **13**: 57-68.

ISI Web of Knowledge, Thomson Reuters.

Iverson, L. R. and A. M. Prasad (2002). "Potential redistribution of tree species habitat under five climate change scenarios in the eastern US." Forest Ecology and Management **155**: 205-222.

Jerome, C. A. and B. A. Ford (2002). "Comparative population structure and genetic diversity of *Arceuthobium americanum* (Viscaceae) and its *Pinus* host species: insight into host-parasite evolution in parasitic angiosperms." Molecular Ecology **11**(3): 407-420.

Jerome, C. A. and B. A. Ford (2002). "The discovery of three genetic races of the dwarf mistletoe *Arceuthobium americanum* (Viscaceae) provides insight into the evolution of parasitic angiosperms." Molecular Ecology **11**(3): 387-405.

Jombart, T. (2008). "adegenet: a R package for the multivariate analysis of genetic markers." Bioinformatics **24**(11): 1403-1405.

Kermack, W. O. and A. G. McKendrick (1927). "A contribution to the mathematical theory of epidemics." Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **115**(772): 700-721.

Kidd, D. M. and M. G. Ritchie (2006). "Phylogeographic information systems: putting the geography into phylogeography." Journal of Biogeography **33**(11): 1851-1865.

Kiefer, C., C. Dobes, T. F. Sharbel and M. A. Koch (2009). "Phylogeographic structure of the chloroplast DNA gene pool in North American *Boechera* - A genus and continental-wide perspective." Molecular Phylogenetics and Evolution **52**(2): 303-311.

Kitamura, K., K. Homma, H. Takasu, S. Hagiwara, F. H. Utech, D. F. Whigham and S. Kawano (2001). "Demographic genetics of the American beech, *Fagus grandifolia*. II. Genet substructure of populations for the Blue Ridge, Piedmont and the Great Smoky Mountains." Plant Species Biology **16**(3): 219-230.

Kitamura, K. and S. Kawano (2001). "Regional differentiation in genetic components for the American beech, *Fagus grandifolia* Ehrh., in relation to geological history and mode of reproduction." Journal of Plant Research **114**(1115): 353-368.

Knowles, L. L. (2009). "Why does a method that fails continue to be used?" Evolution **62**(11): 2713-2717.

Knowles, L. L. and W. P. Maddison (2002). "Statistical phylogeography." Molecular Ecology **11**(12): 2623-2635.

Kuijt, J. (1969). Biology of Parasitic Flowering Plants, Berkeley and Los Angeles, University of California Press.

Leavitt, R. G. (1902). "Subterranean plants of *Epiphegus*." Plant World **5**: 114.

Lemmon, A. R. and E. M. Lemmon (2008). "A likelihood framework for estimating phylogeographic history on a continuous landscape." Systematic Biology **57**(4): 544 - 561.

Linhart, Yan B., Leslie M. Ellwood, Jeffrey D. Karron and Janet L. Gehring (2003). "Genetic differentiation in the dwarf mistletoes *Arceuthobium vaginatum* and *Arceuthobium*

americanum on their principal and secondary hosts." International Journal of Plant Sciences **164**(1): 61-69.

Little, E. L., Jr. (1965). "Mexican beech, a variety of *Fagus grandifolia*." Castanea **30**(3): 167-170.

Lohan, A. J. and K. H. Wolfe (1998). "A subset of conserved tRNA genes in plastid DNA of nongreen plants." Genetics **150**: 425 - 433.

Magri, D., G. G. Vendramin, B. Comps, I. Dupanloup, T. Geburek, D. Gömöry, M. Latalowa, T. Litt, L. Paule, J. M. Roure, I. Tantau, W. O. van der Knaap, R. J. Petit and J.-L. de Beaulieu (2006). "A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences." New Phytologist **171**(1): 199-221.

Manni, F., E. Guerard and E. Heyer (2004). "Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm." Human Biology **76**(2): 173-190.

Manos, P. and K. Steele (1997). "Phylogenetic analyses of "higher" Hamamelididae based on plastid sequence data." Am. J. Bot. **84**(10): 1407-.

Manos, P. S. and A. M. Stanford (2001). "The historical biogeography of Fagaceae: Tracking the tertiary history of temperate and subtropical forests of the northern hemisphere." International Journal of Plant Sciences **162**(s6): S77-S93.

McLachlan, J. S. and J. S. Clark (2004). "Reconstructing historical ranges with fossil data at continental scales." Forest Ecology and Management **197**(1-3): 139-147.

McLachlan, J. S., J. S. Clark and P. S. Manos (2005). "Molecular indicators of tree migration capacity under rapid climate change." Ecology **86**(8): 2088-2098.

McRae, B. H., B. G. Dickson, T. H. Keitt and V. B. Shah (2008). "Using circuit theory to model connectivity in ecology, evolution, and conservation." Ecology **89**(10): 2712-2724.

McRae, B. H. and B. Nürnberger (2009). "Isolation by resistance." Evolution **60**(8): 1551-1561.

Miller, M. P. (2005). "Alleles In Space (AIS): Computer software for the joint analysis of interindividual spatial and genetic information." Journal of Heredity **96**(6): 722-724.

Miller, M. P., M. R. Bellinger, E. D. Forsman and S. M. Haig (2006). "Effects of historical climate change, habitat connectivity, and vicariance on genetic structure and diversity across the range of the red tree vole (*Phenacomys longicaudus*) in the Pacific Northwestern United States." Molecular Ecology **15**(1): 145-159.

Mohamed, K. I., J. F. Bolin, L. J. Musselman and A. T. Peterson (2007). Genetic diversity of *Striga* and implications for control and modeling future distributions. Integrating New Technologies for Striga Control: Towards Ending the Witch-hunt. G. Ejeta and J. Gresse. Hackensack, NJ, USA, World Scientific Publishing Co. Pte. Ltd.**:** 71-84.

Morris, A. B., S. M. Ickert-Bond, D. B. Brunson, D. E. Soltis and P. S. Soltis (2008). "Phylogeographical structure and temporal complexity in American sweetgum (*Liquidambar styraciflua*; Altingiaceae)." Molecular Ecology **17**(17): 3889-3900.

Morris, A. B., D. E. Soltis and P. S. Soltis (2008). Evidence for northern refugia? Reconsidering phylogeographic patterns in American beech (*Fagus grandifolia*; Fagaceae). Botany Conference, Vancouver, British Columbia, Canada.

Musser, R. L. (1979). The *Fagus-Epifagus* parasitic relationship: Field studies and modeling of beech seedlings and parasite carbon dioxide exchange. Department of Botany. Durham, NC, Duke University. **Ph.D. Dissertation**.

Mutikainen, P. and T. Koskela (2002). "Population structure of a parasitic plant and its perennial host." Heredity **89**(4): 318-324.

Nadler, S. A., M. S. Hafner, J. C. Hafner and D. J. Hafner (1990). "Genetic differentiation among chewing louse populations (Mallophaga: Trichodectidae) in a pocket gopher contact zone (Rodentia: Geomyidae)." Evolution **44**(4): 942-951.

Nei, M. and W. Li (1979). "Mathematical model for studying variation in terms of restriction endonucleases." Proc Natl Acad Sci USA **76**: 5269-5273.

Nei, M. and W. H. Li (1979). "Mathematical model for studying genetic variation in terms of restriction endonucleases." Proceedings of the National Academy of Sciences of the United States of America **76**(10): 5269-5273.

Nice, C. C., N. Anthony, G. Gelembiuk, D. Raterman and R. Ffrench-Constant (2005). "The history and geography of diversification within the butterfly genus *Lycaeides* in North America." Molecular Ecology **14**(6): 1741-1754.

Nickrent, D., Y. Ouyang, R. Duff and C. dePamphilis (1997). "Do nonasterid holoparasitic flowering plants have plastid genomes?" Plant Molecular Biology **34**: 717 - 729.

Nickrent, D. L. and E. M. Starr (1994). "High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants." Journal of Molecular Evolution **39**: 62 - 70.

Nieberding, C. M. and I. Olivieri (2007). "Parasites: proxies for host genealogy and ecology?" Trends in Ecology & Evolution **22**(3): 156-165.

Olmstead, R. G., C. W. dePamphilis, A. D. Wolfe, N. D. Young, W. J. Elisons and P. A. Reeves (2001). "Disintegration of the Scrophulariaceae." American Journal of Botany **88**(2): 348-361.

Pakeman, R. J. (2001). "Plant migration rates and seed dispersal mechanisms." Journal of Biogeography **28**(6): 795-800.

Palme, A. E., Q. Su, A. Rautenberg, F. Manni and M. Lascoux (2003). "Postglacial recolonization and cpDNA variation of silver birch, Betula pendula." Molecular Ecology **12**(1): 201-212.

Panchal, M. and M. A. Beaumont (2007). "The automation and evaluation of nested clade phylogeographic analysis." Evolution **61**(6): 1466-1480.

Park, J.-M., J.-F. Manen, A. Colwell and G. Schneeweiss (2008). "A plastid gene phylogeny of the non-photosynthetic parasitic Orobanche (Orobanchaceae) and related genera." Journal of Plant Research **121**(4): 365-376.

Petit, R. J. (2008). "The coup de grâce for the nested clade phylogeographic analysis?" Molecular Ecology **17**(2): 516-518.

Petit, R. J., I. Aguinagalde, J.-L. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineschi, D. Grivet, M. Lascoux, A. Mohanty, G. Muller-Starck, B. Demesure-Musch, A. Palme, J. P. Martin, S. Rendell and G. G. Vendramin (2003). "Glacial

refugia: Hotspots but not melting pots of genetic diversity." Science **300**(5625): 1563-1565.

Petit, R. J., U. M. Csaikl, S. Bordacs, K. Burg, E. Coart, J. Cottrell, B. van Dam, J. D. Deans, S. Dumolin-Lapegue, S. Fineschi, R. Finkeldey, A. Gillies, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. Konig, A. J. Lowe, S. F. Madsen, G. Matyas, R. C. Munro, M. Olalde, M.-H. Pemonge, F. Popescu, D. Slade, H. Tabbener, D. Taurchini, S. G. M. de Vries, B. Ziegenhagen and A. Kremer (2002). "Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations." Forest Ecology and Management **156**(1-3): 5-26.

Petit, R. J., F. S. Hu and C. W. Dick (2008). "Forests of the past: A window to future changes." Science **320**(5882): 1450-1452.

Petit, R. J., A. e. Mousadik and O. Pons (1998). "Identifying populations for conservation on the basis of genetic markers." Conservation Biology **12**(4): 844-855.

Petit, R. m. J., E. Pineau, B. Demesure, R. Bacilieri, A. Ducousso and A. Kremer (1997). "Chloroplast DNA footprints of postglacial recolonization by oaks." Proceedings of the National Academy of Sciences of the United States of America **94**(18): 9996-10001.

Pollak, E. (1987). "On the theory of partially inbreeding finite populations. I. Partial selfing." Genetics **117**(2): 353-360.

Pons, O. and R. J. Petit (1996). "Measuring and testing genetic differentiation with ordered versus unordered alleles." Genetics **144**(3): 1237-1245.

Price, P. W. (1980). Evolutionary Biology of Parasites. Princeton, NJ., Princeton University Press.

Pritchard, J. K., M. Stephens and P. Donnelly (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.

Provan, J. and K. D. Bennett (2008). "Phylogeographic insights into cryptic glacial refugia." Trends in Ecology & Evolution **23**(10): 564-571.

Reynolds, J., B. S. Weir and C. C. Cockerham (1983). "Estimation of the coancestry coefficient: Basis for a short-term genetic distance." Genetics **105**(3): 767-779.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky and M. W. Feldman (2002). "Genetic structure of human populations." Science **298**(5602): 2381-2385.

Schaal, B. A. and K. M. Olsen (2000). "Gene genealogies and population variation in plants." Proceedings of the National Academy of Sciences of the United States of America **97**(13): 7024-7029.

Schmidt, K. and K. Jensen (2000). "Genetic structure and AFLP variation of remnant populations in the rare plant *Pedicularis palustris* (Scrophulariaceae) and its relation to population size and reproductive components." American Journal of Botany **87**(5): 678-689.

Schneider, A. (2003). "GPS Visualizer." from http://www.gpsvisualizer.com/kml_overlay.

Schneider, S., D. Roessli and L. Excoffier (2000). Arlequin ver 2.000: A software for population genetics data analysis, Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Schrenk, H. (1894). "Parasitism of *Epiphegus Virginiana* (Broom Rape, Cancer Root.)." Proceedings of the American Microscopical Society **15**(2): 91-127.

Seibert, S. G. and M. R. Pelton (1994). "Nutrient content of squawroot, *Conopholis americana*, and its importance to southern Appalachian black bears, *Ursus americanus* (Carnivora: Ursidae)." Brimleyana **21**: 151-156.

Shaw, W. C., D. R. Shepherd, E. L. Robinson and P. F. Sand (1962). "Advances in witchweed control." Weeds **10**(3): 182-192.

Soltis, D. E., A. B. Morris, J. S. McLachlan, P. S. Manos and P. S. Soltis (2006). "Comparative phylogeography of unglaciated eastern North America." Molecular Ecology **15**(14): 4261-4293.

Sommer, R. S. and F. E. Zachos (2009). "Fossil evidence and phylogeography of temperate species: 'glacial refugia' and post-glacial recolonization." Journal of Biogeography **36**(11): 2013-2020.

Stewart, J. R. and A. M. Lister (2001). "Cryptic northern refugia and the origins of the modern biota." Trends in Ecology & Evolution **16**(11): 608-613.

Storfer, A., M. A. Murphy, J. S. Evans, C. S. Goldberg, S. Robinson, S. F. Spear, R. Dezzani, E. Delmelle, L. Vierling and L. P. Waits (2007). "Putting the 'landscape' in landscape genetics." Heredity **98**(3): 128-142.

Templeton, A. R. (1998). "Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history." Molecular Ecology **7**(4): 381-397.

Templeton, A. R. (2004). "Statistical phylogeography: methods of evaluating and minimizing inference errors." Molecular Ecology **13**(4): 789-809.

Templeton, A. R. (2008). "Nested clade analysis: an extensively validated method for strong phylogeographic inference." Molecular Ecology **17**(8): 1877-1880.

Templeton, A. R. (2009). "Why does a method that fails continue to be used? The answer." Evolution **63**(4): 807-812.

Thieret, J. W. (1969). "Notes on *Epifagus*." Castanea **34**(4): 397-402.

Thompson, J. N. (1994). The Coevolutionary Process. Chicago and London, University of Chicago Press.

Tiffney, B. H. (2008). "Phylogeography, fossils, and Northern Hemisphere biogeography: The role of physiological uniformitarianism." Annals of the Missouri Botanical Garden **95**(1): 135-143.

Tsai, Y. E. and D. Kidd. (2007). "PhyloGeoViz: A phylogeographic visualization tool." from www.phylogeoviz.org.

USDA and NRCS (2009). The PLANTS Database (http://plants.usda.gov), National Plant Data Center, Baton Rouge, LA USA.

van Hulst, R., B. Shipley and A. Theriault (1987). "Why is *Rhinanthus minor* (Scrophulariaceae) such a good invader?" Canadian Journal of Botany **65**: 2373-2379.

Wagenmakers, E.-J. and S. Farrell (2004). "AIC model selection using Akaike weights." Psychonomic Bulletin & Review **11**(1): 192-196.

Walsh, M. A., E. A. Rechel and T. M. Popovich (1980). "Observations on plastid fine-structure in the holoparasitic angiosperm *Epifagus virginiana*." American Journal of Botany **67**(5): 833-837.

Ward, M. D. and K. S. Gleditsch (2008). Spatial regression models. Thousand Oaks, CA, Sage Publications, Inc.

Weir, B. S. (1996). Genetic Data Analysis II. Sunderland, MA, Sinauer Associates, Inc.

Werle, E., C. Schneider, M. Renner, M. Volker and W. Fiehn (1994). "Convenient single-step, one tube purification of PCR products for direct sequencing." Nucleic Acids Research **22**(20): 4354-4355.

Williams, C. E. and R. K. Zuck (1986). "Germination of seeds of *Epifagus virginiana* (Orobanchaceae)." The Michigan Botanist **25**: 103-106.

Williams, J. W. and S. T. Jackson (2007). "Novel climates, no-analog communities, and ecological surprises." Frontiers in Ecology and the Environment **5**(9): 475-482.

Williams, J. W., B. N. Shuman, T. Webb, P. J. Bartlein and P. L. Leduc (2004). "Late-quaternary vegetation dynamics in North America: Scaling from taxa to biomes." Ecological Monographs **74**(2): 309-334.

Wirth, T., A. Meyer and M. Achtman (2005). "Deciphering host migrations and origins by means of their microbes." Molecular Ecology **14**(11): 3289-3306.

Wolfe, A., C. Randle, L. Liu and K. Steiner (2005). "Phylogeny and biogeography of Orobanchaceae." Folia Geobotanica **40**(2): 115-134.

Wolfe, A. D. and C. W. dePamphilis (1998). "The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants." **15**: 1243 - 1258.

Wolfe, K. H., C. W. Morden and J. D. Palmer (1992). "Function and evolution of a minimal plastic genome from a nonphotosynthetic parasitic plant." Proceedings of the National Academy of Sciences USA **89**: 10648-10652.

Wolfe, K. H., C. W. Mordent, S. C. Ems and J. D. Palmer (1992). "Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: Loss or accelerated sequence evolution of tRNA and ribosomal protein genes." Journal of Molecular Evolution **35**(4): 304-317.

Woods, K. D. and M. B. Davis (1989). "Paleoecology of range limits: Beech in the Upper Peninsula of Michigan." Ecology **70**(3): 681-696.

Wright, S. (1943). "Isolation by distance." Genetics **28**(2): 114-138.

Zane, L., L. Bargelloni and T. Patarnello (2002). "Strategies for microsatellite isolation: a review." Molecular Ecology **11**(1): 1-16.

Zuber, D. and A. Widmer (2009). "Phylogeography and host race differentiation in the European mistletoe (*Viscum album* L.)." Molecular Ecology **18**(9): 1946-1962.

Zwickl, D. J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion, The University of Texas at Austin. **Ph.D. dissertation**.

Biography

Yi-Hsin Erica Tsai was born in Taipei, Taiwan in 1980. Following her father's graduate studies in Chapel Hill, North Carolina, the family emigrated to the United States, and finally settled in Fort Collins, Colorado. In 2002, Erica received a Bachelor of Arts from the University of Colorado at Boulder with double majors in Biochemistry and Environmental, Population, and Organismic Biology and a minor in Chemistry. She graduated Summa Cum Laude and with Distinction.

While at the University of Colorado, Erica had the great opportunity to work with Dr. Jeff Mitton on a project studying population structure in quaking aspen. Through this work she became interested in genetics, evolution, and phylogeography. These interests led her to work on the comparative phylogeography of a host-parasite system with Dr. Paul Manos at Duke University, where she began her graduate studies in 2002. While a graduate student, Erica participated in the 2007 Google Summer of Code program at the National Evolutionary Synthesis Center. Through this experience, she gained valuable programming skills and produced PhyloGeoViz, a web-based phylogeographic visualization program. While at Duke she received a National Science Foundation Graduate Research Fellowship, James B. Duke Fellowship, and Department of Biology Fellowship. Her research was supported by grants from the National Science Foundation, Sigma Delta Epsilon/Graduate Women in Science, Duke University Graduate School, Duke University Department of Biology, Mellon Foundation, Association for Women in Science, American Society of Plant Taxonomists, Botanical Society of America, Deep Time, and Sigma Xi.

Erica's teaching experiences include being a co-instructor of a general chemistry co-seminar (Minority Arts and Sciences Program, University of Colorado), a teaching assistant

for an introductory computer science class (Dept. of Computer Science, University of Colorado), and a teaching assistant in introductory genetics and organismal diversity classes (Dept. of Biology, Duke University). During her doctoral studies, Erica has served as a mentor for several students through the National Science Foundation Research Experience for Undergraduates Program and as an assistant coordinator for the Howard Hughes Program for Precollege Students. She also volunteered and worked with the Duke Women in Science and Engineering student group as their webmaster, planning committee member, and planning committee chair.

Erica is a member of the following professional societies: Society for the Study of Evolution, American Society of Plant Taxonomists, Botanical Society of America, Sigma Xi, Sigma Delta Epsilon/Graduate Women in Science, and the Association for Women in Science.