1-1-2020

# Evolutionary History Of Subterranean Termites In The Geographic And Ecological Context Of The Appalachian Mountains In The United States

Chaz Hyseni

# EVOLUTIONARY HISTORY OF SUBTERRANEAN TERMITES IN THE GEOGRAPHIC AND ECOLOGICAL CONTEXT OF THE APPALACHIAN MOUNTAINS IN THE UNITED STATES

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
in the Department of Biology
The University of Mississippi

by

Chaz Hyseni

May 2020

ABSTRACT

Termites in the genus *Reticulitermes* (Blattodea: Rhinotermitidae) are distributed across the eastern United States, including the southern Appalachian Mountains, a region incredibly rich in biodiversity. The eastern subterranean termite, *Reticulitermes flavipes*, has been uninentionally introduced to South America and Europe, and is predicted to further expand its geographic range. My goal was to determine how eco-evolutionary processes, operating at both long and short timescales, may have contributed to *R. flavipes* becoming an invasive species. I examined geographic and environmental influences at historical and contemporary timescales. To do this, I first determined the extent of niche divergence among three geographically overlapping *Reticulitermes* species, *R. flavipes*, *R. malletei*, and *R. virginicus*, and also identified the geographic areas and environmental conditions in which *R. flavipes* occurs to the exclusion of the other two species. Then, I assessed evidence for the influence of glacial-interglacial cycles on changes in the geographic distribution of *R. flavipes*, as well as potential genetic divergence within the species resulting from these past distributional shifts. In addition to historical eco-evolutionary processes, at the contemporary timescale I investigated how epigenetic mechanisms–specifically, DNA methylation–facilitate rapid responses to human-mediated disturbance of forest ecosystems. Finally, I developed a new landscape connectivity metric, $MS_{Conn}$, to help understand the effect spatial heterogeneity of environments plays on biological diversity at multiple levels of organization, from alleles to communities. In principle, $MS_{Conn}$ can be integrated into an eco-evolutionary framework, making it possible to quantify the effect of biotic and abiotic environments on gene flow between populations, and vice versa, the effect of gene flow on species interactions within and between communities.

DEDICATION

To Mom.

# ACKNOWLEDGEMENTS

It has been a long journey, a journey that started over 20 years ago. I survived a war, I worked 3 years as an interpreter for the United Nations, and started college thousands of miles away from home. Born in Germany, I spent my formative years in Kosovo, in the war-torn environment of the nineties in former Yugoslavia. I would not be where I am today, if not for my mom (Hateme) and dad's (Rexhep) unwavering support and selfless sacrifices. A piece of their heart left when I flew off to college in August 2002. My dad passed away in 2006 and I could not be there in his dying moments, or for his funeral. My sister (Merita) got married, but I could not go. She had her first son (Unik), and I was able to visit when he was 7 months old. He is now 7, and I have yet to visit him again. I have also missed the birth of my second nephew (Eden). I am grateful for my family's sacrifices. Sending me into the world meant, I had an unforgettable college experience at Yale University.

I have always been open-minded and eager to learn new things, but Yale is where my horizons were substantially broadened. What is more, I met my best friend, Michael Chen, and many other lifelong friends, including Jared Levant, and friends of the "round table" (Alex Millman, Ari Romney, Cerin Lindgrensavage, Chris Hagemann, Christina Meyer, Esme von Hoffman, Florence Wu, Jessica Feinstein, Laura Manville, Marta Herschkopf, Mary Matthews, Molissa Farber, Ryan Suplee, Sailaja Paidipaty, Shari Wiseman, Taylor Davis, Vlad Vainberg, Vivek Kasinath, and others). 18 years later, we are still in touch. I was best man at Michael and Christina's wedding in 2010. I will also never forget the road trip Michael and I took across the southwestern U.S. and California in 2007–a lot of beef jerky was consumed.

I got started on the academic path almost 15 years ago. On this path, I have

been extremely fortunate to come across people who have given me incredible support. In my third year in college, Gisella Caccone gave me an opportunity to do research in her lab, and the rest is history. As part of my research in Gisella's lab, I traveled for field work to the Galapagos islands–I will never forget those two weeks of witnessing astonishing wildlife, especially giant tortoises. After graduating in 2007, I spent an additional five years as a research assistant in Gisella's lab. In those years, she allowed me to spread my science wings, and I will forever be grateful for her leadership style and giving me so much creative freedom. She became more than just a mentor; she is, truly, my second mom. Thanks to her, New Haven, Connecticut became a home away from home. My best friend and lab manager in Gisella's lab, Carol Mariani, deserves a special thank you, for all the laughs we had in the lab/office. I am also immensely thankful for Jeff Powell's support throughout the years, and all the great people and friends in the Caccone and Powell labs (Beckie Symula, Ben Evans, Dan Edwards, Edgar Benavides, Julia Brown, Katy Richards-Hrdlicka, Kirstin Dion, Mark Sistrom, Ryan Garrick, to name a few).

Soccer has been a constant in my life, since my childhood days. The sport itself, and the friends I have met through it, especially in the last 15 years, have helped me maintain a work-life balance, positively affecting the quality of my work and bettering my professional experience. My soccer friends in New Haven, especially the bonds created within the Flower Power soccer team, made New Haven feel even more like home. Getting together for games and winning intramural titles with the team will always be some of my more cherished memories. Since Flower Power came to be, in 2009, I am still in touch with many former members, and best friends: Srinath Krishnan and Andy Wells (two members of the Campari Trio, C3O), Woosok Moon (member of the C3O expansion, C4O), Toshi Karato, Ronan Chalmin, Brad Foley, Anja Hafemann, Andy Robson, and others. I am also grateful for friendships that blossomed on the soccer field, outside of the Flower Power team, including David Wallmann, Eric Meridiano, Frank Limbrock, Fred Sowah, Philipp Weissert, and many others.

I left New Haven after 10 years, to go to Cornell University. It was almost as hard as leaving Kosovo. However, Cornell and Ithaca, New York will always have a special place in my heart. It is where my Ph.D. journey began. The professional situation I found myself in was not ideal and after just one year, I realized I had to leave. Things were very different on the friendship front. I may have only lived there from August 2012 to December 2013, but the friendships made at Cornell (Annise Dobson, Ben Marcy-Quay, Cat Sun, Jeremy Dietrich, Laura Eierman, Sarah Collins and Willie Fetzer, Suzanne Beyeler and Jason Martin, Wieteke Willemen, and many others) and through Ithaca soccer are timeless. Ibe Jonah was the first friend I made in Ithaca. Ibe's friendship, and the soccer league he has organized for so many years, made my Ithaca experience a fantastic one, despite a professional stumbling block. Playing on the Mystery Machine team was one of the highlights of my Ithaca experience.

Oxford, Mississippi would become the next destination. At Yale, in the Caccone lab, I became friends with Ryan Garrick and Beckie Symula, whom I have been fortunate to have as friends for over 10 years. Their tenure-track adventure at the University of Mississippi started a little before my Ph.D. journey started at Cornell. After things did not go as expected at Cornell, I listened to Ryan's suggestion to apply for admission to the Biological Science Ph.D. program at the University of Mississippi. I arrived in Oxford in 2014. Ryan has truly been a best friend, and an amazing advisor, always finding time to help with valuable feedback, guidance, and life and career advice. I am also very grateful for Beckie's loyal friendship, support, and guidance, both as a friend and a committee member. Ryan and Beckie have made my Ph.D. journey an incredibly rewarding experience. I am indebted to the other members of my committee (Brice Noonan, Erik Hom, Louis Zachos, Peter Zee, and Rodney Dyer) as well, for their feedback and support.

Aside from the memorable academic experience, I will remember my time in Oxford for the friendships that will last beyond my Ph.D. years, especially friends I shared a lab or office with (Amber Horning, Jason Payne, Jarrod Sackreiter, John Banusiewicz, Lauren Fuller, Stephanie Burgess, Reese Worthington, Zanethia Barnett, and many others), as well as a number of great friends (Amit Pillai, Andreas

Vortisch, Leti Wodajo, Saumil Jadhav, Shaheed Nazrul, to name a few) made on the soccer field. I would be remiss not to thank Cindy Rimoldi, Colin Jackson, Lance Sullivan, Linda Mota, Matt Ward, and especially Richard Buchholz, for all the laughs, and their support, of course.

I moved to Oxford in 2014, but it was in 2015 when Oxford became home. In January, Baxter decided that living with me was better than the dog shelter, and he has been my faithful companion ever since. In November, Estelle Blair came into my life. She is a strong, independent woman, and a future medical doctor. I am lucky to have found such a stalwart ally who has been there for me since our first date. She is the love of my life and deserves more than I was able to give at times, especially in the last 6 months of the dissertation writing process. She has been very supportive and patient throughout, and without her by my side, finishing the dissertation would have been much harder. Not only has she been there for me, but she also brought another joy in my life, Lucy, who was 3 years old when I first met her. She is 8 now, and has taught me so much already, about being a dad, about life, and stopping to smell the roses, or look for insects. As a matter of fact, in 2016, when I did most of my field work in the Appalachian Mountains, Estelle and Lucy came along for a few field trips, and helped me collect termites. It was an amazing summer adventure. Baxter came along on every trip. Nala came along, too, a few times. Nala is a beautiful German shepherd, who has completed our family of five. She is as loyal as it gets, and never leaves my side. Baxter and Nala "forcing" me to go on walks has helped me stay focused, especially in the last two months of writing the dissertation, and self-isolation during the coronavirus pandemic.

Words cannot express how grateful I am for all the people (many not mentioned here) who have made the United States feel like home. I became a U.S. citizen in April 2013, and the decision to do so was in large part because of the people I have known here. Thank you. Thank you, also, for making my Ph.D. journey a successful one. I will forever be grateful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# INTRODUCTION

Ecological and evolutionary processes are dynamically intertwined, not only historically, but also on contemporary timescales (e.g.,[3–7]). Species interactions in communities and ecosystems drive phenotypic changes within species, which reciprocally influence species interactions. Phenotypic plasticity is an important biological phenomenon that allows organisms to modulate their phenotypes in response to the local environmental conditions, including biotic interactions. Thus, phenotypic plasticity plays an important role in eco-evolutionary dynamics (e.g.,[8,9]).

Phenotypic plasticity is an important biological phenomenon that allows organisms to modulate their phenotypes in response to different biotic and abiotic environments. Epigenetic mechanisms can modulate phenotypes through changes in gene expression, without concomitant changes in DNA sequence. For instance, DNA methylation at the promoter of a gene may suppress expression of the gene. Epigenetic mechanisms have been associated with the phenotypic differences observed among castes (e.g., workers, soldiers, reproductives) in eusocial insects, including ants[10], bees[11] and wasps[12], as well as termites[13–15].

Caste differentiation and task specialization (e.g., workers provide food for the colony) have allowed eusocial insects–especially ants and termites–to become ecologically dominant. Indeed, in some tropical forests, ants and termites have been estimated to make up 30% of the animal biomass and 80% of the insect biomass[16]. Through their activities, ants and termites affect entire ecosystems, and are aptly described as ecosystem engineers. For instance, in West Africa and Uganda, termite activity increases the heterogeneity of savanna vegetation[17,18].

In the dead-wood microhabitats of forest ecosystems, the engineering activities of subterranean termites contribute to enhancing the internal heterogeneity of logs, making them habitable for a diverse array of dead-wood-dependent (saproxylic) arthropods. As ecosystem engineers, evolutionary change in subterranean termites is likely to affect local- and broad-scale ecological dynamics, including community structure and ecosystem processes, which, in turn, are likely to have an impact on evolutionary change in subterranean termites, at both long and short timescales.

Subterranean termites in the genus *Reticulitermes* (Blattodea: Rhinotermitidae) are broadly distributed across the eastern United States. Five *Reticulitermes* termite species are found in this part of the country (often sympatrically)[19], which includes the southern Appalachian Mountains, a region incredibly rich in biodiversity[20]. The eastern subterranean termite, *Reticulitermes flavipes* (Kollar), is predicted to expand its geographic range by 2050[21]. Native to the eastern United States, *R. flavipes* has been unintentionally introduced to other parts of the U.S. (e.g.[22]), as well as other countries (e.g.[23–26]).

Here, my goal was to determine how eco-evolutionary processes, operating at both long and short timescales, may have contributed to *R. flavipes* spreading to other parts of the world and becoming invasive. To gain insights into the success of this species, in Chapter 1 I examined whether *R. flavipes* evolved distinct niche requirements and identified geographic areas and environmental conditions in which *R. flavipes* occurs to the exclusion of two congeners (*R. malletei* and *R. virginicus*) that are also commonly found in the southern Appalachian Mountains, and from which *R. flavipes* is thought to have diverged over 10 million years ago[27].

In Chapter 2, I hypothesized that Pleistocene climatic fluctuations altered the geographic distribution of *R. flavipes*, repeatedly redistributing genetic diversity, and thus impacting the evolutionary history of the species. To determine whether glacial-interglacial climate change in the Pleistocene resulted in distributional shifts and genetic divergence within *R. flavipes*, I modeled contemporary and historical (up to 120,000 years ago) geographic distributions of *R. flavipes* in the eastern U.S., and also inferred the evolutionary and demographic history of the species using mitochondrial and nuclear DNA sequence data.

In addition to examining deep-time ecological niche divergence among *Reticulitermes* species, and genetic divergence within *R. flavipes*, in Chapter 3 I hypothesized that, at the contemporary timescale, human-mediated disturbance of forest ecosystems in the southern Appalachian Mountains has had effects on DNA methylation in *R. flavipes*, thus contributing to the species' phenotypic plasticity. This has the potential to impact interactions with closely related species, and could facilitate the invasiveness of *R. flavipes* in other parts of the world that are similarly altered by humans (e.g., in France[23,25,28,29]).

In Chapter 4, I developed a new metric, $MS_{Conn}$, which captures functional connectivity at multiple levels, from alleles to communities. $MS_{Conn}$ can be applied in different fields. For instance, in landscape genetics, it can be integrated into a framework for testing the effect of environmental features on gene flow. In community ecology, $MS_{Conn}$ can measure connectivity between species that are linked by dispersal in a network of communities. Furthermore, this metric can, in principle, be integrated into an eco-evolutionary framework, making it possible to quantify the effect of biotic and abiotic environments on gene flow between populations, as well as the effect of gene flow on species interactions within and between communities.

CHAPTER 1:

# ECOLOGICAL DRIVERS OF SPECIES DISTRIBUTIONS AND NICHE OVERLAP FOR THREE SUBTERRANEAN TERMITE SPECIES IN THE SOUTHERN APPALACHIAN MOUNTAINS, USA

Citation: Hyseni C, Garrick RC. Ecological drivers of species distributions and niche overlap for three subterranean termite species in the southern Appalachian Mountains, USA. **Insects** *2019*, 10. https://www.mdpi.com/2075-4450/10/1/33

Abstract: In both managed and unmanaged forests, termites are functionally important members of the dead-wood-associated (saproxylic) insect community. However, little is known about regional-scale environmental drivers of geographic distributions of termite species, and how these environmental factors impact co-occurrence among congeneric species. Here we focus on the southern Appalachian Mountains—a well-known center of endemism for forest biota—and use Ecological Niche Modeling (ENM) to examine the distributions of three species of *Reticulitermes* termites (i.e., *R. flavipes*, *R. virginicus*, and *R. malletei*). To overcome deficiencies in public databases, ENMs were underpinned by field-collected high-resolution occurrence records coupled with molecular taxonomic species identification. Spatial overlap among areas of predicted occurrence of each species was mapped, and aspects of niche similarity were quantified. We also identified environmental factors that most strongly contribute to among-species differences in occupancy. Overall, we found that *R. flavipes* and *R. virginicus* showed significant niche divergence, which was primarily driven by summer temperature. Also, all three species were most likely to co-occur in the mid-latitudes of the study area (i.e., northern Alabama and Georgia, eastern Tennessee and western North Carolina), which is an area of considerable topographic complexity. This work pro-

vides important baseline information for follow-up studies of local-scale drivers of these species' distributions. It also identifies specific geographic areas where future assessments of the frequency of true syntopy vs. micro-allopatry, and associated interspecific competitive interactions, should be focused.

## 1.1 Introduction

### 1.1.1 The Southern Appalachian Mountains: A Center of Endemism for Forest Biota

The southern Appalachian Mountains, extending latitudinally from northeast Alabama to northwest Virginia, are some of the oldest uplands in North America. These mountains have been exposed and unglaciated for over 100 million years[30]. Steep altitudinal precipitation gradients, a complex heavily dissected topography, and a humid, temperate climate, have shaped southern Appalachian forests into some of the most diverse environments in the eastern United States.[31]. While deciduous oak-hickory forests dominate much of the mid-elevation landscape[31], high elevations (above 1400 m) support spruce-fir forests[32], whereas mesic coves support hemlock, and pines are commonly found at xeric low- to mid-elevations[33].

The southern Appalachian Mountains are incredibly rich in biodiversity[20]. The region is thought to have served as a major Pleistocene refuge for numerous species. Past climatic cycles have affected distributions of forest biota, resulting in major range shifts or local extinction. Following the Last Glacial Maximum (ca. 21,000 years ago), recolonization is thought to have occurred relatively rapidly, from 7000–16,000 years ago[34–38]. The southern Appalachian Mountains are a well-known center of endemism for salamanders and other amphibians[39,40]. However, there is increasing evidence of short-range endemism in other groups, including dead wood-associated forest invertebrates (e.g., millipedes[41,42], cockroaches[43,44], and centipedes[45]).

### 1.1.2 Subterranean Termites: Functionally Important Ecosystem Service Providers in Temperate Forests

Dead-wood-dependent (saproxylic) arthropods play critical roles in maintaining healthy, productive forests by contributing to the decomposition of fallen trees and thus driving nutrient cycling that affects organisms at all trophic levels[46–50]. Indeed, rotting logs may be one of the most stable, thermally buffered, above-ground microhabitats that exist in forests, and the decomposition process has successional stages, facilitated by wood-feeding and wood-boring invertebrates[50,51]. Termites are some of the first to colonize a rotting log, and through feeding and tunneling activities of the worker caste, the dead-wood substrate is modified by the creation of galleries. Once established, these facilitate colonization by larger wood-feeding invertebrates[52]. Ultimately, the ecosystem engineering activities of termites contribute to enhancing the internal heterogeneity of logs, making them habitable by a diverse array of saproxylic species.

Termites in the genus *Reticulitermes* (Blattodea: Rhinotermitidae) are broadly distributed across the eastern United States. Morphological separation of species is notoriously difficult[53], particularly given that only the worker caste can usually be readily sampled. To address this, we developed an efficient molecular assay (i.e., polymerase chain reaction (PCR) amplification of a short region of mitochondrial cytochrome oxidase subunit II (COII) gene, followed by screening of restriction-fragment-length polymorphism (RFLP) banding profiles[54]) that can be used to distinguish each of the five eastern United States species. In the southern Appalachians, several *Reticulitermes* species can co-occur locally. However, true syntopy (i.e., two species co-inhabiting the same rotting log) appears to be very rare, but reported instances of fine-scale sampling have been limited.

### 1.1.3 Ecological Niche Models: Efficient Tools for Predicting Organismal Distributions

Ecological niche models (ENMs) are broadly useful spatially explicit analytical tools that relate species occurrence data with environmental variables, such as climatic temperature and precipitation data[55], or topographic and land cover data. Once constructed, ENMs generate maps of estimated habitat suitability, and can be used to describe the historical, current, and future climate space for a given species. For example, ENMs have been used to identify areas of high conservation importance[56–58], predict climate change effects on geographic ranges of species[59,60], as well as determine potential threats of invasive species[61,62]. These analytical tools are becoming widely used owing to the increasing accessibility of climatic data via public databases[63–65]. An important assumption when using ENMs to predict historical or future distributions is niche conservatism (i.e., the stability of ecological niches over time)[66]. However, evidence suggests that niche conservatism is common among closely related species[67–69], and the risks of erroneous inferences are further reduced when focusing only on contemporary climate and occurrence data (i.e., when reconstructing present-day ENMs).

### 1.1.4 The Current State of Knowledge about Subterranean Termite Distributions, and Goals of this Study

There is a general lack of data on the natural distributions of termites in temperate forests, given that most research has focused on damage that termites cause to man-made wooden structures. Accordingly, occurrence records mostly come from urban areas, and they are also of low resolution (e.g., presence/absence in a given county). Notwithstanding these limitations, Maynard et al.[70] recently provided valuable insights into the role of climatic (temperature and precipitation) variables in influencing distributions of termites in the eastern United States. Specifically, those authors performed ENM for two *Reticulitermes* species (*R. flavipes* and *R. virginicus*) and the invasive Formosan subterranean termite, *Coptotermes formosanus*. Furthermore, they synthesized pre-existing knowledge to identify the influence on termite distributions of biotic factors, such as tree species and wood traits, fungal preferences, phenology of predatory ants, and competitive asymmetries among coexisting termite species. While interspecific competition may result in spatial or temporal separation which could lead to niche divergence, to date, very little is known about niche partitioning in subterranean termites and the environmental factors that may lead to niche divergence.

In the present paper, we aimed to generate new insights into regional-scale environmental drivers of geographic distributions of termite species, and how these environmental factors impact co-occurrence among congeneric species. Focusing on the southern Appalachian Mountains and surrounding areas, we performed an ENM-based evaluation of niche divergence among the three most common *Reticulitermes* species in the eastern United States. In addition to identifying niche divergence, if present, we aimed to determine the environmental factors driving niche divergence among species.

## 1.2 Methods

### 1.2.1 Termite Sampling, Species Identification, and Ecological Niche Modeling

From 2012 to 2016, we collected *Reticulitermes* termites from 132 sites across the southern Appalachians Mountains and surrounding areas (Table A.1; Figure A.1). At most sites, termite workers were collected from a single rotting log at an intermediate to late stage of decay. However, at 10 sites, termites were also collected from additional logs within ~30 m of one another (i.e., samples came from a total of 2 logs at 8 sites, 3 logs at 1 site, and 4 logs at 1 site; Table A.1). Owing to the close proximity of these clustered logs (i.e., at or near the typical error associated with a handheld GPS unit), the same coordinates were assigned to them, but specimen collections were assigned log-specific identifiers. Molecular taxonomic identifications were based on a single termite per rotting log, using Garrick et al.'s[54] PCR-RFLP assay. Briefly, a short (376-bp) region of the mitochondrial COII gene was amplified (using PCR primers RetCo2-F and RetCo2-R), and products were then sequentially digested with three restriction enzymes (RsaI, TaqI, and MspI), which in combination generate diagnostic species-specific banding patterns. Ultimately, we identified 91 non-redundant occurrence points for *R. flavipes*, 30 for *R. virginicus*, and 17 for *R. malletei* (Table A.1). ENM was conducted with the 'biomod2' package[71,72] in R[73] using four modeling algorithms (e.g.,[74–76]). Distributions were reconstructed using mean climatological data for a period spanning 1960–1990, with all variables used at 1-km resolution. Nineteen bioclimatic variables[63] were obtained from the WorldClim database v.1.4 (`http://www.worldclim.org`), and then factor analysis was used to reduce the number of predictors, and the associated correlation among them (see Supplementary Material for full details of ENM methods). From the 19 bioclimatic variables, we generated four environmental factors (see Supplementary Material and Figures A.2 and A.3 for full details of factor analysis): dry-season precipitation, wet-season precipitation, summer temperature, and temperature range.

### 1.2.2 Niche Occupancy, Niche Identity, and Distributional Overlap

Predicted niche occupancy profiles were generated for each environmental factor following Evans et al.[77], implemented in the 'phyloclim' package[78]. Niche overlap for each environmental factor was summarized using both Schoener's D statistic[79], and the modified Hellinger statistic, I, as proposed by Warren et al.[80]. We also used the D and I statistics to determine pairwise niche equivalency/identity among the three *Reticulitermes* species. The niche equivalency test asks whether the ENMs of two species are more different than expected if they had been drawn from the same distribution. To perform the niche equivalency test, we generated a distribution using 999 pseudoreplicate datasets.

To assess distributional overlap based on ENMs, we used maps of binary presence/absence as well as continuous occurrence probabilities. We used binary predictions, because this allowed us to determine which species co-occurred in areas of distributional overlap. However, since the use of continuous predictions has been recommended when estimating species richness[81], we calculated the sum of *Reticulitermes* species' occurrence probabilities (Figure A.4), and calculated joint and exclusive occurrence probabilities for each of the three species (Figure A.5). For binary predictions, the approach of maximizing sensitivity and specificity has consistently performed better than other methods[82–84]. Thus, we used the True Skill Statistic (TSS = sensitivity + specificity − 1)[85] both as a model performance metric and to identify a threshold for converting continuous occurrence probabilities to binary classifications. The threshold was chosen based on maximizing the TSS, without risking under-prediction of presences (i.e., selecting the lowest threshold at which TSS is maximized). We used a threshold value of 0.2, where probability $> 0.2$ represented presence, and suitability $\leq 0.2$ represented absence. We merged the three species' binary maps by summing re-coded maps, where absence = 0, but presence was coded depending on species: *R. flavipes* = 4, *R. virginicus* = 2, and *R. malletei* = 1. This way, the sum of binary maps resulted in seven distinct categories: single-species areas (3 categories, with aforementioned scores); areas of two-species overlap (3 categories, scores of either 3, 5, or 6 depending on the identity of the species pair); and areas where all three species overlap (1 category, with a score of 7).

### 1.2.3 Environmental Factors and Niche Divergence

To determine the sources of variation in the *Reticulitermes* occurrence dataset, we included the effects of spatial structure and environmental factors, and performed variance partitioning using the 'varpart' function in 'vegan'[86]. To account for multiple predictors in the model, we used adjusted $R^2$. To determine which (if any) environmental factors have significantly contributed to niche divergence of

*Reticulitermes* species, we first removed the effect of spatial structure. We did this by performing distance-based redundancy analysis[87] using the 'capscale' function. To account for spatial structure, we transformed Euclidean geographic distances to a continuous rectangular vector by Principal Coordinates analysis of Neighbor Matrices (PCNM) using the 'pcnm' function in 'vegan'. Only significant PCNM axes were used in partialling out spatial structure. Significance of the environmental and spatial predictors was assessed using multivariate F-statistics with 9999 permutations.

## 1.3 Results

### 1.3.1 Niche Occupancy, Niche Identity, and Distributional Overlap

Predicted niche occupancy profiles for the three *Reticulitermes* species (Figure 1.1) showed differences in peak values across all four environmental factors. The two temperature factors, summer temperature and temperature range, showed differences in peaks between *R. flavipes* and *R. virginicus*, whereas *R. malletei* was intermediate. Similarly, the two precipitation factors, dry-season precipitation and wet-season precipitation, showed more marked differences between *R. flavipes* and *R. virginicus* than for any of the other pairwise species comparisons. The bimodality of wet-season precipitation is a result of occurrence of *Reticulitermes* species in two areas with pronounced differences in wet-season precipitation (see Figure A.3). Bimodality was also observed for summer temperature in *R. flavipes*, given that the species occurs in both low elevations and the cooler high-elevation areas of the Appalachians (see Figure A.3). Statistics that characterize the extent of niche overlap showed that *R. flavipes* and *R. virginicus* had the least amount of overlap (D = 0.582, I = 0.843; Table 1.1). Furthermore, the niche identity test between these two species showed significant differentiation ($p < 0.001$; Table 1.1). *R. malletei* was more similar to *R. flavipes* in terms of temperature range (D = 0.889) and summer temperature (D = 0.872), but showed more overlap with *R. virginicus* for dry- (D = 0.894) and wet-season precipitation (D = 0.848). *R. virginicus* showed the least overlap with *R. flavipes*, across all four environmental factors (Table 1.2).

The predicted distribution of *R. flavipes* spanned a larger area in the northern portion of the southern Appalachians than that of the other two species. *R. flavipes* overlapped with *R. malletei*, to the exclusion of *R. virginicus*, in an area including Kentucky, Virginia, and West Virginia (Figure 1.2; Figure A.5). The

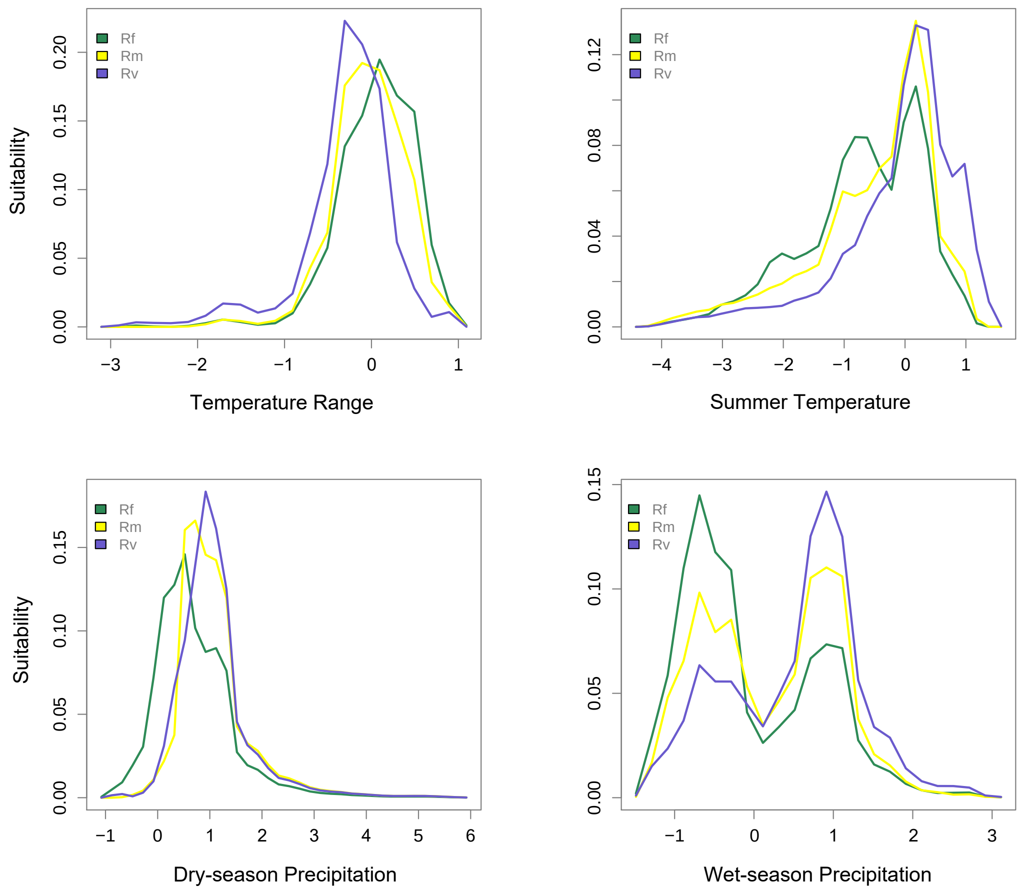



**Figure 1.1:** *Predicted niche occupancy*. Four environmental factors were used to estimate niche occupancy of *R. flavipes* (Rf), *R. malletei* (Rm), and *R. virginicus* (Rv): top two panels: temperature range and summer temperature; bottom two panels: dry- and wet-season precipitation. The y-axis represents niche occupancy, or suitability, and the area under the curves sums to 1, the total suitability.

**Table 1.1:** *Niche identity test*. The upper off-diagonal shows Schoener's D statistic, and the lower off-diagonals shows the modified Hellinger statistic, I. Significant niche divergence is reported in bold text with red highlighting. The more dissimilar of the other two niche comparisons is highlighted in pink. Abbreviations used for *R. flavipes*, *R. malletei*, and *R. virginicus* are Rf, Rm, and Rv, respectively.

| | Rf | Rm | Rv |
|---|---|---|---|
| Rf | - | $D = 0.744$<br>$p = 0.280$ | **$D = 0.582$**<br>**$p < 0.001$** |
| Rm | $I = 0.935$<br>$p = 0.239$ | - | $D = 0.788$<br>$p = 0.630$ |
| Rv | **$I = 0.843$**<br>**$p < 0.001$** | $I = 0.961$<br>$p = 0.750$ | - |

**Table 1.2:** *Pairwise niche overlap among* Reticulitermes *species for each of four environmental factors.* The top three rows show Schoener's D statistic, and the bottom three rows show the modified Hellinger statistic, I. The four environmental factors are: temperature range (TR), summer temperature (ST), dry-season precipitation (DP), and wet-season precipitation (WP). Niche overlap is highest in green and lowest in red. *R. flavipes*, *R. malletei*, and *R. virginicus* are abbreviated as Rf, Rm, and Rv, respectively.

| | | TR | ST | DP | WP |
|---|---|---|---|---|---|
| **D** | Rf/Rm | 0.889 | 0.872 | 0.693 | 0.820 |
| | Rf/Rv | 0.683 | 0.707 | 0.680 | 0.680 |
| | Rm/Rv | 0.791 | 0.809 | 0.894 | 0.848 |
| **I** | Rf/Rm | 0.991 | 0.990 | 0.919 | 0.982 |
| | Rf/Rv | 0.917 | 0.928 | 0.926 | 0.942 |
| | Rm/Rv | 0.952 | 0.961 | 0.990 | 0.984 |

overlap between *R. flavipes* and *R. virginicus*, excluding *R. malletei*, spanned a smaller area, with lower probability (Figure A.5). Predicted distributions of all three species overlapped in eastern Tennessee, western North Carolina, northern Alabama and Georgia (Figure 1.2; Figure A.4 and Figure A.5).

### 1.3.2 Environmental Factors and Niche Divergence

Distance-based redundancy analysis (Figure 1.3) showed that only the summer temperature factor contributed significantly ($F_{1, 127} = 8.673$, $p = 0.001$) to differences in occurrence among the three *Reticulitermes* species. After accounting for spatial structure by partialling out six significant spatial components (PCNM axes 1, 4, 6, 17, 43, and 58), summer temperature remained significant ($F_{1, 121} = 5.622$, $p = 0.003$). The six significant spatial components along with summer temperature accounted for 18.5% of the observed variation in the occurrence data. Spatial structure alone explained 9.6% of the variation, environmental factors accounted for 3.3%, and the interaction between the two explained an additional 5.6% of the variation.

Following the removal of spatial structure effects, the highest correlation coefficient between environmental factors and ordination axes of the distance-based redundancy analysis was observed for summer temperature (r = 0.730) and axis 1. This axis captured the divergence of *R. virginicus* from the other two species (Figure 1.3). Thus, summer temperature contributed significantly to *R. virginicus* divergence. While not significant, temperature range (r = −0.383) and wet-quarter precipitation (r = 0.376) were correlated with axis 2, which captured the divergence of *R. malletei* (Figure 1.3).

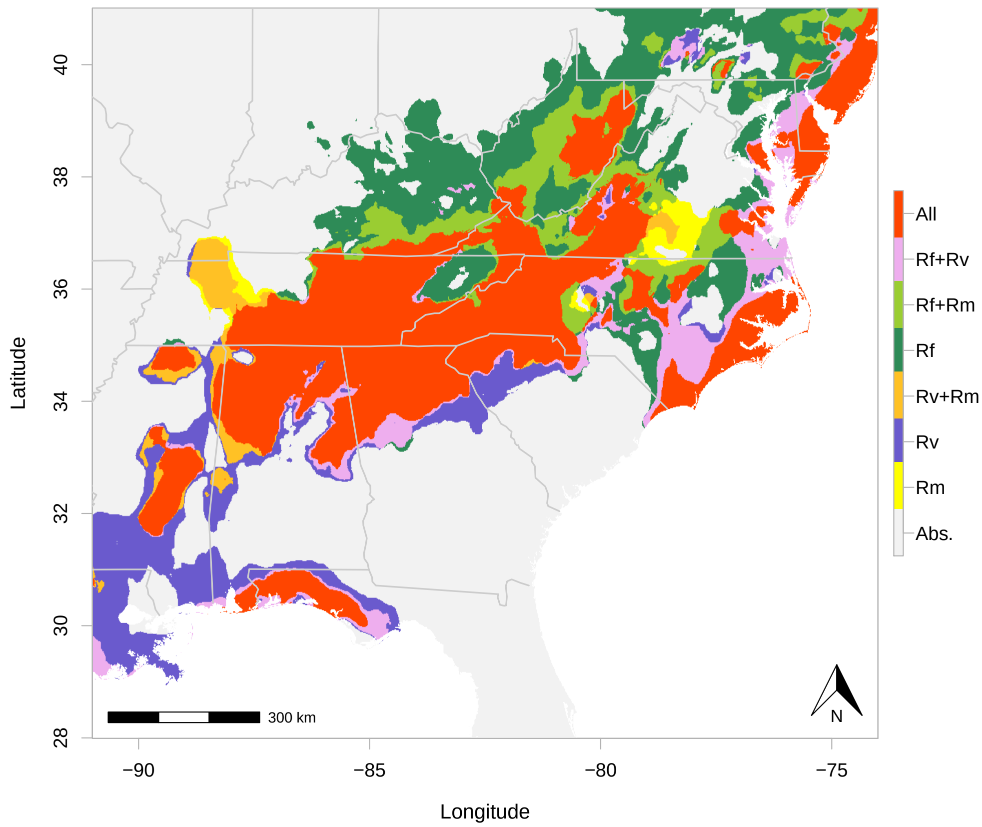



**Figure 1.2:** *Distributional overlap of* R. flavipes *(Rf),* R. malletei *(Rm), and* R. virginicus *(Rv).* Overlap is color coded based on the number of species. "All" is where occurrence of all three species is predicted. Areas of two-species overlap are shown in the legend as "Rf + Rv", "Rf + Rm", and "Rv + Rm". Absence of all three species is shown in grey and referred to in the legend as "Abs."

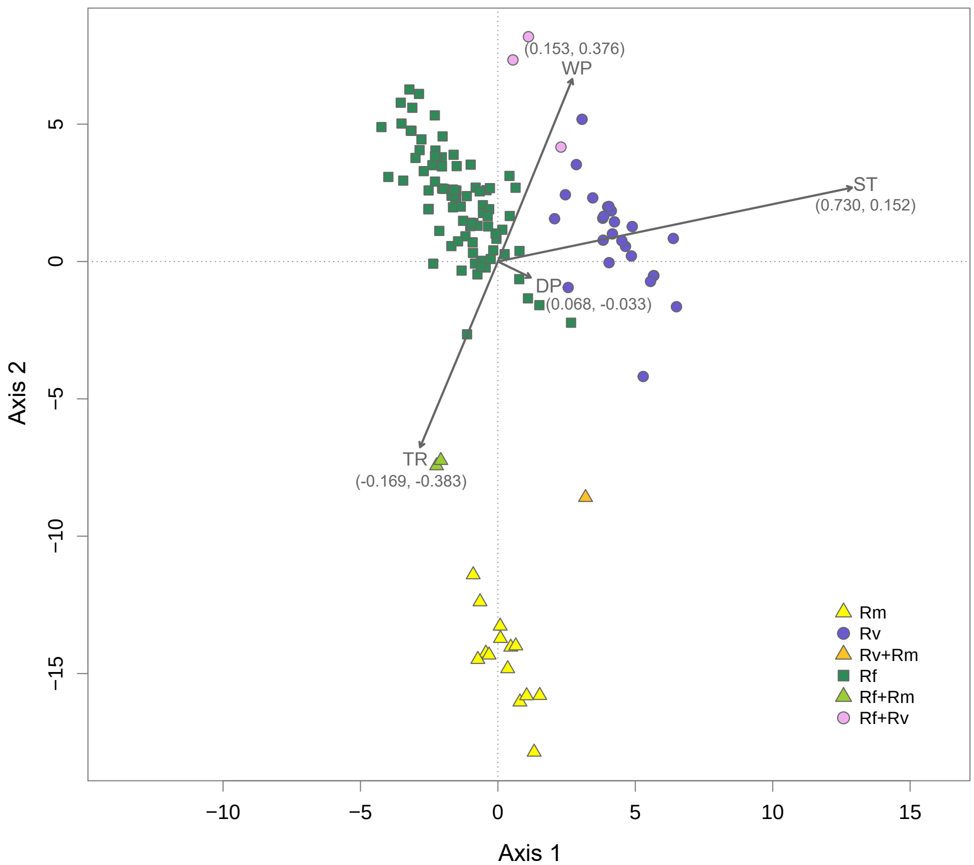



**Figure 1.3:** *Distance-based redundancy analysis.* The plot shows a constrained ordination of 132 sampling sites, color coded based on the number of species present. Sites where only *R. flavipes*, *R. virginicus*, or *R. malletei* were sampled are referred to in the legend as "Rf", "Rv", and "Rm", respectively. Two-species sites are shown in the legend as "Rf + Rv", "Rf + Rm", and "Rv + Rm". The ordination is conditional on six significant spatial components (PCNM axes 1, 4, 6, 17, 43, and 58) and constrained by four environmental factors: dry-season precipitation (DP); wet-season precipitation (WP); summer temperature (ST); temperature range (TR). Arrows show strength of correlation (coefficients in parentheses) of environmental factors with ordination axes 1 and 2.

## 1.4 Discussion

This study provides insights into the ecology of subterranean termites with regard to geographic distributions and niche partitioning among three broadly co-distributed *Reticulitermes* species in the southern Appalachian Mountains and surrounding areas. This region is a biogeographically significant center of endemism, yet the ecology of its resident invertebrate fauna–particularly saproxylic insects–is poorly known. Our ENMs suggest that an area in the mid-latitudes of the southern Appalachians, characterized by complex topography and multiple ecoregions, provides suitable habitat to support all three *Reticulitermes* species. Our study also highlights the roles that temperature and precipitation play in driving niche divergence among *Reticulitermes* species. To our knowledge, this work represents the first evidence of significant regional-scale niche divergence between *R. flavipes* and *R. virginicus*. Below, we consider the broader context of these findings, as well as caveats and future directions for follow-up studies that build on the information presented here.

### 1.4.1 *Reticulitermes* Distributions and Climatic Drivers of Niche Divergence among Species

Our analyses predicted extensive co-occurrence of all three *Reticulitermes* species in the mid-latitudes of the southern Appalachians (Figure 1.2; Figure A.4 and Figure A.5). Based on paleoclimatic[88], biogeographic[89] and comparative phylogeographic[90] data, the southern Appalachians remained free from Pleistocene ice sheets and served as a major refuge for many species during glacial periods, consequently maintaining higher levels of biodiversity. Indeed, the present-day complexity of this mid-latitude region harbors many different niches, which could facilitate long-term coexistence of closely related species. However, in addition to predicted co-occurrence of *Reticulitermes* species in the montane regions of the southern Appalachians, our ENMs also identified areas of two- and three-species co-occurrence along the Gulf coast of western Florida, and the Atlantic coast from North Carolina to New Jersey and New York. To empirically confirm the co-occurrence of subterranean termites in these coastal areas, future studies should include these regions in their sampling efforts. In the case of another forest-dependent invertebrate, the millipede *Narceus americanus*, the Florida Gulf coast has been identified as an important refuge during the Last Glacial Maximum[91]. Indeed, the paleoclimatic history of areas to the south and east of the southern Appalachian Mountains are increasingly being recognized as reservoirs of forest invertebrate biodiversity during past periods of environmental change. The incidence of high termite species diversity—even though only assessed here for one genus—is therefore not unexpected.

In addition to co-occurrence of *Reticulitermes* species, our study provides novel insights into climatic drivers of niche divergence. Consistent with the findings of Maynard et al.[70], we determined that *R. virginicus* is more restricted to the south, whereas *R. flavipes* has a broad latitudinal range. Furthermore, we determined that *R. flavipes* occurs farther north than the other two species, even excluding other *Reticulitermes* (Figure A.5), potentially because it tolerates lower amounts of precipitation (both dry- and wet-season; Figure 1.1). Maynard et al.'s[70] ENMs showed that temperature variables were the most important predictors of termite distributions. Based on our formal assessment of niche overlap between *R. flavipes* and *R. virginicus*, we determined that both temperature and precipitation seasonality (as represented by temperature range, summer temperature, and dry- and wet-season precipitation) play non-negligible roles in the significant niche divergence between *R. flavipes* and *R. virginicus*. Furthermore, using distance-based redundancy analysis, we identified summer temperature as a major driver of this divergence. In the mid-latitudes of the southern Appalachians, where dry-season precipitation is high (Figure A.3), all three *Reticulitermes* species co-occur (Figure 1.2; Figure A.4 and Figure A.5), but farther north, where dry- and wet-season precipitation is low (Figure A.3), *R. flavipes* is more competitive.

### 1.4.2 Potential Explanations for Lack of Empirical Evidence for Local-Scale Coexistence of *Reticulitermes* Species

Interestingly, despite the significant niche divergence between *R. flavipes* and *R. virginicus*, we collected both of these species from the same rotting log at one sampling site (i.e., #37 located near the Georgia/Southern Carolina state border; Table A.1). To our knowledge, this is the first record of true syntopy between *Reticulitermes* species. The apparent rarity of syntopy and general lack of coexistence of *Reticulitermes* species at local scales could be explained by competitive exclusion. Colony size and soldier number are important features for termite competitive ability. Termite species with small colonies have been observed to relinquish resources and be eliminated by dominant interspecific competitors with large colonies[92]. Through avoidance of dominant competitors, interspecific competition may result in spatial separation[93], but also temporal separation (i.e., phenological differences). Termites may be able to avoid other related species using vibrational cues. Indeed, vibrational cues are important for termite sensory perception and communication, as these signals can travel over long distances[94,95]. For instance, the drywood termite *Cryptotermes secundus* can distinguish conspecifics from the dominant competitor in the environment, the subterranean termite *Coptotermes acinaciformis*[94]. Furthermore, *Coptotermes acinaciformis* detects its major predator, the ant *Iridomyrmex purpureus*, using vibrational cues only[95]. Overall, given these highly tuned sensory capabilities, it stands to reason that competitive exclusion, or competitor avoidance, could be important factors in preventing local co-occurrence among *Reticulitermes* species. Alternatively, the dominant competitor may ultimately outcompete the other species. For instance, *R. flavipes* has a broad distribution and occurs farther north than the other two species, possibly due to a competitive advantage stemming from the fact that it tolerates conditions of lower dry- and wet-season precipitation. Furthermore, interspecific aggression coupled with low levels of intraspecific agonism (even colony fusion)[96,97], may make *R. flavipes* the dominant competitor.

### 1.4.3 Caveats and Future Directions

While our sampling suggests that true syntopy and local co-occurrence of different species at the same site is very rare, our detection of only one species in all but one rotting log, and at the majority of sampling sites (i.e., 126 out of 132), may actually be a consequence of the sampling strategy that was employed (see Section 1.2.1). Briefly, we simply aimed to collect termites from each site, rather than provide a complete assessment of termite diversity at each site. Indeed, variance partitioning reflects this, showing that most (81.5%) of the variance in the occurrence data did not stem from spatial structure (9.6%), or environmental differences (3.3%), or interaction between the two (5.6%). Accordingly, while competitive exclusion is a plausible explanation for apparent rare local-scale co-occurrence (i.e., micro-allopatry) among *Reticulitermes* species, a dedicated sampling approach would be required to formally test this idea. For example, exhaustively sampling multiple logs per site, at a series of sites arranged along a transect traversing a region where two or more species occur in close proximity would be a productive approach. Fortunately, the present study identified specific geographic areas where future assessments of the frequency of true syntopy vs. micro-allopatry, and associated interspecific competitive interactions, should be focused (Table A.1; Figure A.1).

Although we have shown separation in niche space between species, particularly *R. flavipes* and *R. virginicus*, these inferences were underpinned by regional-scale environmental variables, and so they do not take into account local-scale drivers of niche divergence such as differences in microhabitat preference, phenology, or diet. Indeed, Maynard et al.[70] highlighted that biotic and soil characteristics play a role in termite distribution and abundance. Thus, our assessment of niche divergence is necessarily incomplete. While it does provide important baseline information, follow-up studies of local-scale drivers of species' distributions could examine aspects of the microhabitat (e.g., humidity and temperature of soil and rotting logs), timing of nuptial flights along latitudinal and altitudinal clines, and/or use stable isotopes to determine decomposition stage of ingested wood and the importance of microbial biomass in termite diets at a given location[98].

Data Accessibility: The following are available online at `http://www.mdpi.com/2075-4450/10/1/33/s1`: File S1: Environmental variables and Ecological Niche Modeling methods; Table A.1: Sampling sites with number of species occurrences at each site and number of logs per site; Figure A.1: Map of *Reticulitermes* sampling depicting occurrences of one or more species at each site; Figure A.2: Factor analysis; Figure A.3: Environmental factors and bioclimatic variables; Figure A.4: Distributional overlap of *Reticulitermes* species; Figure A.5: Probability of joint and exclusive occurrence of *Reticulitermes* species.

CHAPTER 2:

# THE ROLE OF GLACIAL-INTERGLACIAL CLIMATE CHANGE IN SHAPING THE GENETIC STRUCTURE OF EASTERN SUBTERRANEAN TERMITES IN THE SOUTHERN APPALACHIAN MOUNTAINS, USA

CITATION: Hyseni C, Garrick RC. The role of glacial-interglacial climate change in shaping the genetic structure of eastern subterranean termites in the southern Appalachian Mountains, USA. **Ecology and Evolution** *2019*, 9(8). `https://doi.org/10.1002/ece3.5065`

ABSTRACT: The eastern subterranean termite, *Reticulitermes flavipes*, currently inhabits previously glaciated regions of the northeastern U.S., as well as the unglaciated southern Appalachian Mountains and surrounding areas. We hypothesized that Pleistocene climatic fluctuations have influenced the distribution of *R. flavipes*, and thus the evolutionary history of the species. We estimated contemporary and historical geographic distributions of *R. flavipes* by constructing Species Distribution Models (SDM). We also inferred the evolutionary and demographic history of the species using mitochondrial (cytochrome oxidase I and II) and nuclear (endo-beta-1,4-glucanase) DNA sequence data. To do this, genetic populations were delineated using Bayesian spatial genetic clustering, competing hypotheses about population divergence were assessed using approximate Bayesian computation (ABC), and changes in population size were estimated using Bayesian skyline plots. SDMs identified areas in the north with suitable habitat during the transition from the Last Interglacial to the Last Glacial Maximum, as well as an expanding distribution from the mid-Holocene to the present. Genetic analyses identified three geographically cohesive populations, corresponding with northern, central, and south-

ern portions of the study region. Based on ABC analyses, divergence between the Northern and Southern populations was the oldest, estimated to have occurred 64.80 thousand years ago (kya), which corresponds with the timing of available habitat in the north. The Central and Northern populations diverged in the mid-Holocene, 8.63 kya, after which the Central population continued to expand. Accordingly, phylogeographic patterns of *R. flavipes* in the southern Appalachians appear to have been strongly influenced by glacial-interglacial climate change.

## 2.1 Introduction

Geographic barriers to dispersal, such as mountains and rivers, are considered major drivers of genetic divergence within and among species. The influence of climate change (e.g., glacial-interglacial oscillations during the Pleistocene) in generating phylogeographic structure is also widely recognized ([99,100] and references therein). For example, in Europe, when ice sheets reached their maximum extent during glacials, this repeatedly resulted in range contraction into southern refugia, which subsequently served as key reservoirs for recolonization via northward expansion during interglacials[99,101]. In these regions at high latitudes, successive glacial-interglacial cycles were likely to reinforce the same genetic signatures of contraction and expansion (but see[102,103]).

In contrast to landscapes that were repeatedly covered by ice sheet advances throughout the Pleistocene, those in temperate or tropical regions that remained unglaciated potentially contained numerous refugia (e.g.,[104]). Indeed, in montane areas with deeply dissected topography, latitude alone may be a poor proxy for the locations of refugial areas, as the steep environmental gradients that occur locally can exert a strong influence on persistence of habitat patches that can support viable populations. In such regions—in contrast to the traditional view of refuges being continuously occupied long-term stable areas—successive glacial-interglacial cycles are less likely to have repeatedly played out in the same way. Owing to stochastic processes, they may have instead been somewhat ephemeral. For instance, a refugium may have been only periodically occupied, with the process of shifting between alternative refugia from one glacial cycle to the next involving extinction at the trailing edge and colonization at the leading edge. Herein, we refer to this particular case of contraction-expansion dynamics as "distributional shift" and consider it a plausible model for the focal landscape setting. Indeed, consideration of how major shifts in geographic distributions contributed to population differentiation during the Pleistocene is important for understanding speciation processes (e.g.,[105] and references therein).

The southern Appalachian Mountains represent some of the oldest uplands in North America ( 471–480 million years old;[106] and references therein) and harbor high levels of biodiversity[39,40,107,108]. This topographically complex temperate region is characterized by steep environmental gradients, which have promoted population divergence in many species, particularly those with poor dispersal abilities[109]. Paleoclimatic[88], biogeographic[89] and comparative phylogeographic[90] data indicate that the southern Appalachians remained free from Pleistocene ice sheet advances, and consequently, retained numerous refugial areas for forest-dependent biota during cool and dry glacial periods. Indeed, short-range endemism and high diversity have been well documented in plethodontid salamanders[39] and other amphibians[40]. Similar patterns have also been reported for invertebrate groups such as crayfish[107], arachnids[109,110], and millipedes[42]. While the role of the southern Appalachian Mountains as a major barrier driving an east-west divide among lowland taxa is widely recognized ([90] and references therein), there have been surprisingly few biogeographic and phylogeographic studies of upland species that occupy the mid- and high-elevation ridgelines, and research on invertebrates in particular is underrepresented.

The eastern subterranean termite, *Reticulitermes flavipes*, currently inhabits previously glaciated regions of the northeastern U.S., as well as the unglaciated southern Appalachian Mountains and surrounding areas. This species is a key ecosystem engineer that makes major contributions to dead wood decomposition and nutrient cycling in forests[48,111], and its distribution is influenced by humidity and temperature[112]. This diploid eusocial species lives in colonies that typically have a simple family structure, arising from an outbred primary reproductive pair that remains fertile for 6–11 years[19]. When the king or queen die, some full-sib workers differentiate into male and female secondary reproductives, at which point the colony becomes inbred[113]. However, in addition to temporal transitions from simple to extended families, there may also be spatial partitioning, whereby the initial reproductive center, with the primary reproductives, expands into satellite nests housing secondary reproductives[114]. Winged alates disperse away from the original colony and establish new colonies and then shed their wings. However, dispersal abilities are only moderate, with distances varying from a few meters to >1 km[19]. Such limited dispersal is conducive to strong historical inference[115].

Reconstructing long-term population history is often achieved via analyses of geo-referenced DNA sequence data, using spatially explicit phylogenetic and/or coalescent-based analytical approaches (see[116,117] and references therein). Increasingly, complementary non-genetic data are being employed to augment inferences or to generate hypotheses about past events and population processes. In particular, Species Distribution Models (SDM) are now widely used to locate glacial refugia (e.g.,[118]), or determine the influence of past climate change on current genetic structure (e.g.,[119]). In some cases, similar conclusions about phylogeographic history have been drawn from SDMs and genetic data[120]. Briefly, SDMs relate occurrence records for a given species with the environmental conditions in those same locations in order to estimate geographic areas in which the species is likely to be found[121]. Given that historical climatic fluctuations can trigger range contractions and expansions—including wholesale distributional shifts (e.g.,[122])—SDMs can form a framework for understanding the genetic consequences of glacial-interglacial climate change[123].

In this study, we investigated the genetic consequences of glacial-interglacial climate change on *R. flavipes* from the unglaciated southern Appalachian Mountains and surrounding areas, and considered distributional shifts as a plausible hypothesis (among others) to be assessed using SDMs and genetic data. Given the reliance of this species on dead-wood microhabitats, our expectation was that during the Pleistocene and earlier, *R. flavipes* closely tracked the changing distributions of forest habitats, and was strongly impacted by climatic fluctuations. Indeed, ecologically-specialized low-mobility forest insects may be particularly well-suited for reconstructing past climatic impacts on montane forest landscapes, in part owing to their short generation times and ability to persist in habitat patches too small to support more mobile vertebrates[124–126]. Furthermore, owing to the limited dispersal ability of *R. flavipes*, we expected that relatively fine-scale genetic structuring would be detectable. To test these expectations, we modeled present and past distributions and used contrasts between these SDMs to make inferences about distributional shifts and to identify areas of stability (i.e., potential refugia). Based on this, we generated competing hypotheses about drivers of genetic divergence, and then tested these via analyses of DNA sequence data using coalescent simulations. In addition to the effects of historical climatic conditions, we also considered the influence, if any, of contemporary climatic conditions and dispersal-based spatial structure on genetic variation in *R. flavipes*.

## 2.2 Methods

### 2.2.1 Phylogeographic framework

To address the aims of this study, we used the following workflow: Step 1 – Model present-day and historical climate-based distributions of *R. flavipes* in order to identify potential refugia and generate expectations about directionality of range contractions or expansions, including distributional shifts; Step 2 – Infer the number of distinct populations using spatial genetic clustering, and cross-validate via principal component analysis, and phylogenetic reconstruction; characterize genetic variation within and differentiation among populations, and; estimate the amount of genetic variation explained by dispersal (spatial structure) and environment (contemporary climatic conditions); Step 3 – Test alternative phylogeographic hypotheses to determine whether expansion out of refugia, distributional shifts, or vicariance was the underlying historical process generating the observed patterns of genetic variation within and among populations; estimate values of parameters included in the best-fit phylogeographic hypothesis; and assess evidence for changes in effective population size over time.

### 2.2.2 Genetic data collection

*Reticulitermes* termites were collected between 2012 and 2014 from locations in the southern Appalachian Mountains. Since it is not possible to reliably distinguish among several co-distributed species on the basis of morphology when only members of the worker caste are collected[127], termites were identified using a molecular assay[54]. Ultimately, *R. flavipes* were sampled from 50 rotting logs across 46 locations (Figure 2.1; also see Table B.1 in Supplementary Material). From each log, 1–3 individuals were used for phylogeographic analyses. For out-group taxa, we included specimens representing three close relatives (Table B.2): *R. virginicus* (n = 3 individuals), *R. malletei* (n = 1) and R. nelsonae (n = 1).

Extraction of genomic DNA was performed using a DNeasy tissue kit (Qiagen, Valencia, CA) following the manufacturer's recommendations. Portions of the mitochondrial cytochrome c oxidase subunit I (COI) and II (COII) genes, and an intronic portion of the nuclear endo-beta-1,4-glucanase (EB14G) gene, were amplified via Polymerase Chain Reaction using primers (Table B.3) and conditions reported in Section B.1.2 in Supplementary Material, and then sequenced at Yale University. Sequence alignments were performed using Geneious v.6.1.8[128], and

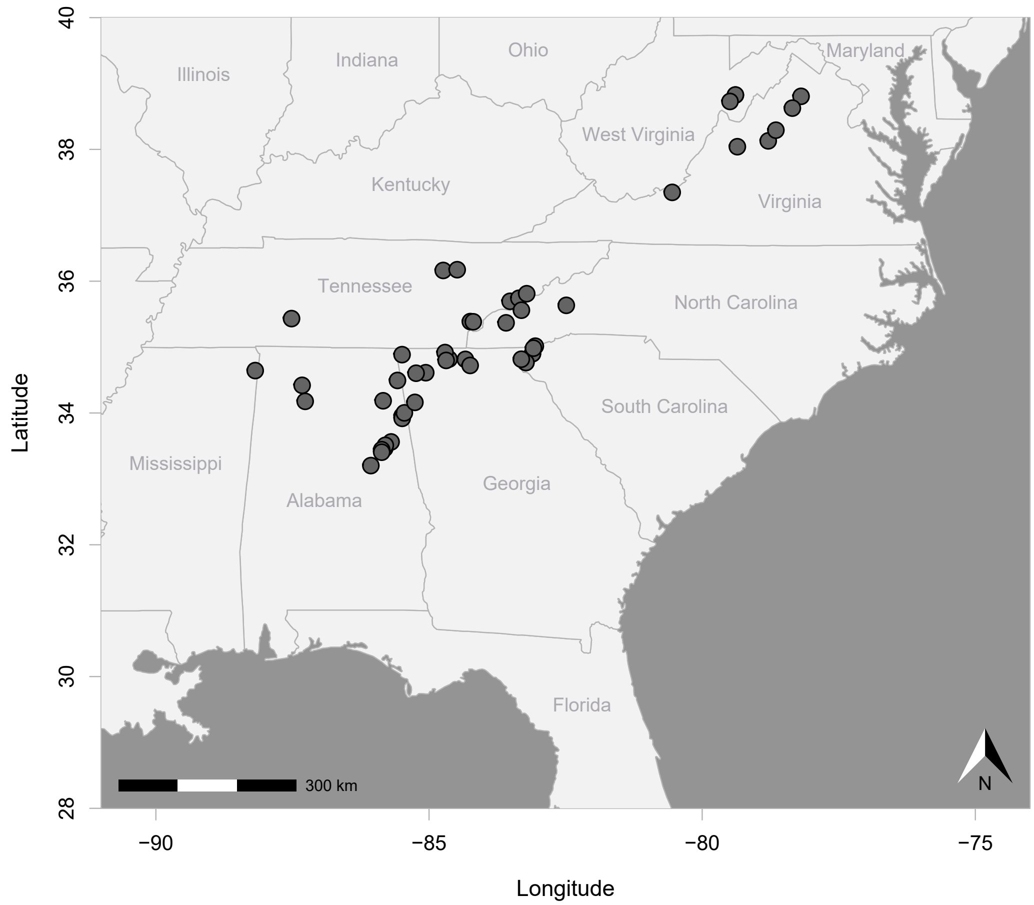



**Figure 2.1:** *Sites sampled for use in genetic analyses.* Geographic map showing sampling locations (gray dots, n = 46) from which *Reticulitermes flavipes* termites were collected in the southern Appalachian Mountains, southeastern USA.

manually edited as necessary. We concatenated COI and COII and refer to this sequence (COI+COII) as the mitochondrial DNA (mtDNA) locus; we refer to EB14G as the nuclear DNA (nDNA) locus. For the latter, heterozygous sites were scored using the "Find Heterozygotes" plugin in Geneious. For a site to be considered heterozygous, we required that height of the secondary peak was at least 50% of the primary peak (sites with quality scores < 20, were coded as 'N'). Allele haplotypes were inferred using PHASE v.2.1.1 [129], with the following settings: 90% phase certainty, 10,000 iterations, thinning interval = 10, burn-in = 1,000, and the default recombination model. PHASE was run three times to evaluate consistency of results.

### 2.2.3 Step 1: Present and past geographic distributions

There are few published occurrence records of forest populations of *R. flavipes* with confirmed species-level identifications and adequate geospatial precision for SDM Accordingly, in addition to the 46 sites that contributed to genetic analyses (above), the presence of *R. flavipes* at an additional 45 locations (surveyed from 2015 to 2016) was confirmed using Garrick et al.'s [54] molecular assay, resulting in a total of 91 occurrence points (Figure B.1). To construct SDMs, we used the 'biomod2' package [71,72] in R [73]. Full details about SDM construction are given in Section B.1.3 in Supplementary Material. Briefly, we used four machine learning algorithms to model distributions based on climatological data, presence records, and 20 independent sets of 100 pseudo-absence points (Figure 2.2). The latter choice was based on work by Barbet-Massin et al. [130], who showed that for machine learning methods it is better to use multiple replicates of pseudo-absence points, with the number of pseudo-absences in each replicate close to the number of occurrence points. We used environmental variables at a 1-km resolution for SDM construction. Present-day SDMs were based on mean climatological data spanning 1960–1990, and historical distributions were modeled for the Mid-Holocene (MH; 6 kya), the Last Glacial Maximum (LGM, 22 kya), and the Last Interglacial (LIG, 120–140 kya). For each period, 19 bioclimatic variables [63] were obtained from the WorldClim database v.1.4 (`http://www.worldclim.org`; Table B.4), and then factor analysis was used to retain maximum variation contained in the 19 variables while simultaneously: 1) reducing the number of predictors, to avoid overfitting, and 2) dealing with non-independence of predictors (i.e., collinearity), which represents a challenge to correlative modeling methods (e.g., [131]).

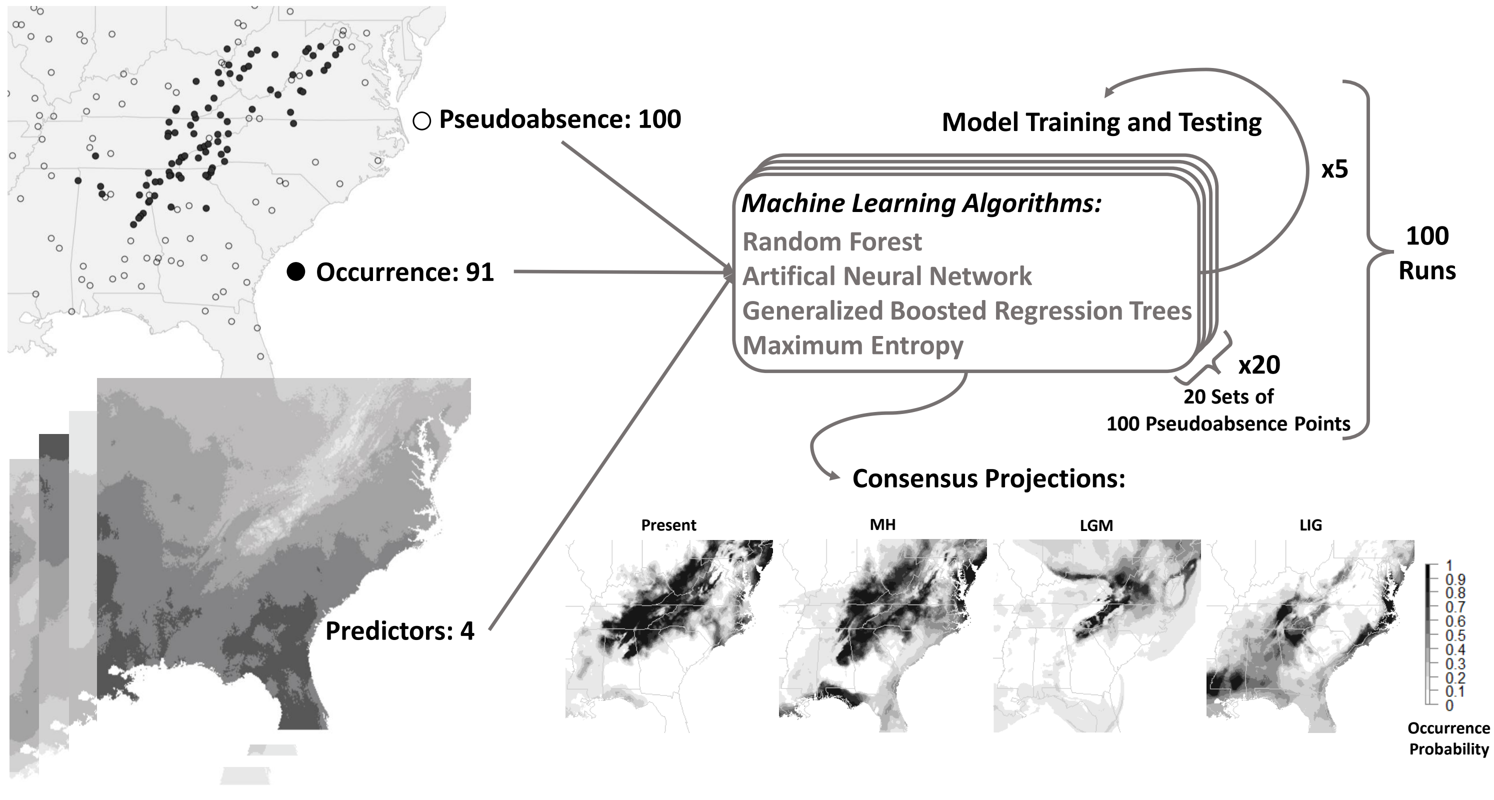



**Figure 2.2:** *Species Distribution Modeling.* Diagram showing the conceptual framework used to generate SDMs that enabled contrasts between successive time periods: "present" (1960–1990), Mid-Holocene (MH; 6 kya), Last Glacial Maximum (LGM; 22 kya), and Last Interglacial (LIG; 120–140 kya).

Distributional shifts and areas of stability. We used a threshold value to convert continuous occurrence probabilities to a binary classification of suitable (>0.2) vs. unsuitable (≤0.2). The occurrence probability threshold was chosen based on the True Skill Statistic (TSS;[85]). Specifically, we chose a threshold value that maximized the TSS, as this approach has consistently performed better than other thresholding methods[82–84]. However, since we used multiple pseudo-absence replicates, we had the opportunity to maximize TSS without risking under-prediction of presences, which results from choosing a high threshold value. Indeed, using distributions of TSS and threshold values, we were able to select the lowest threshold (0.2; Figure B.1), below which TSS had a steep slope. To calculate the distributional shift between two successive time periods (e.g., LIG to LGM, or LGM to MH), we took the difference of the two binary maps, after multiplying the more recent time period by two in order to ensure that we obtain four categories in the distributional shift calculation: colonization (difference = 2), stability (1), absence (0), and extinction (-1; see Figure B.2). Similarly, to estimate areas of stability (i.e., persistence in a location between successive time periods), we multiplied the binary occurrence maps (Figure B.2) of the corresponding periods: locations where the product is 1 were considered to harbor stable habitats across time periods (stability = 1).

### 2.2.4 Step 2: Genetic variation and the role of environment and space in genetic structuring

Bayesian clustering and Principal Components Analysis. To determine the number of geographically cohesive genetic groups of *R. flavipes*, we analyzed georeferenced mtDNA sequences in BAPS v.6.0[132]. We assessed values of K (i.e., the number of clusters) ranging from 2–20, with 10 replicate runs each. We also examined evidence for geographically cohesive genetic groups by representing the variance in mtDNA sequences using Principal Components Analysis (PCA), performed with the 'prcomp' function in R. Phylogenetic reconstruction and molecular dating. We reconstructed a mtDNA-based dated phylogeny to verify the existence of any genetic groups determined by BAPS, as well as to estimate divergence times. First, we used PartitionFinder 1.1.0[133] to determine the best partitioning scheme, and the Bayesian Information Criterion in jModelTest v.2.1.10[134] to identify the optimal model of sequence evolution. The best-fit model for all three codon positions was HKY + I[135]. Then, to estimate a dated phylogeny, we used BEAST v.2.4.5[136], with a relaxed log-normal molecular clock[137], and a coalescent tree prior. We used broad mutation rate priors. For the mtDNA locus, the range included Brower's[138] commonly used insect rate of 1.15% sequence divergence per lineage per million years, and Luchetti et al.'s[139] faster rate of up to 140% per million years, which was estimated from COII in European *Reticulitermes* taxa. Based on point estimates obtained using approximate Bayesian computation (ABC[140]) assessments of competing phylogeographic hypotheses (described in Methods – Step 3), we set the mean mutation rate at 12% per million years (see Results – Step 3) for the mtDNA locus. Since there was no mutation rate information available for the nDNA locus in *Reticulitermes*, we estimated the mean mutation rate in BEAST by conditioning on the mtDNA locus and setting the initial mean value at 0.6% with a range of 0.2–2% (obtained using ABC; see Results – Step 3). BEAST was run for 50 million Markov chain Monte Carlo generations, with samples saved every 2,500 generations, after discarding the first 5 million generations as burn-in. We used Tracer 1.6[141] to examine the stationarity of parameter estimates and to determine that effective sample sizes were greater than 500. BEAST was run with and without the out-group *Reticulitermes* taxa using the same settings. Results were summarized via a Maximum Clade Credibility tree in TreeAnnotator v.2.4.4[136], with the first 25% of trees discarded as burn-in. Diversity within and differentiation among genetic populations. To estimate levels of diversity within each genetic population, the following metrics were calculated separately for the mtDNA and nDNA loci using DnaSP v5.10.01[142]: number of segregating sites (S[143]), average number of nucleotide differences (K[144]), nucleotide diversity ($\pi$[143]), and the mutation-scaled effective population size ($\theta_W$[145]). To measure genetic divergence among genetic populations, the following statistics were also calculated: average number of nucleotide substitutions per site (Dxy[143]), net number of nucleotide substitutions per site (Da[143]), average number of pairwise nucleotide differences (Kxy[144]), and $F_{ST}$[2]. Genetic variation influenced by environment and dispersal. To estimate the amount of genetic variation explained by spatial structure versus the environment, we used distance-based redundancy analysis (dbRDA[87]). We computed the genetic distance matrix using the 'dist.dna' function of the 'ape' package[146], and performed dbRDA using the 'capscale' function of the 'vegan' package[86] in R. To compute the response variable, genetic distances (i.e., matrix of pairwise mutational differences between DNA sequences) were estimated using the TN93[147] model of sequence evolution, allowing for different rates for transitions and transversions. For environmental predictors, we used the contemporary environmental factors obtained via factor analysis (see Methods – Step 1). To obtain spatial structure predictors, we transformed Euclidean geographic distances to a continuous rectangular vector by Principal Coordinates analysis of Neighbor Matrices (PCNM) using the 'pcnm' function in 'vegan'. Significance of the predictors was assessed using multivariate F-statistics with 9999 permutations. We first analyzed the relationship between the genetic distance matrix and each environmental factor separately, and then performed a partial dbRDA for each variable while controlling for the influence of spatial structure, using only significant PCNM eigenvectors. Similarly, we analyzed the relationship between genetic distances and PCNM eigenvectors, retained the significant eigenvectors, and then removed interactions with the environment to obtain the contribution of spatial structure alone.

### 2.2.5 Step 3: Phylogeographic hypothesis testing and population size

Competing scenarios. We used ABC, as implemented in the software DIYABC v.2.1.0[148], to assess alternative hypotheses designed to determine whether expansion out of long-term stable refugia, distributional shifts, or vicariance was the major underlying process generating the present-day spatial distribution of genetic variation. MtDNA plus (phased) nDNA sequence data were used, and we conditioned these analyses on a posteriori knowledge of the existence of three distinct genetic clusters of *R. flavipes* (see Results – Step 2). Because ABC analyses can suffer when a large number of candidate models are simultaneously considered[149], we employed a two-tiered approach, where best-fit scenarios from separate analyses in the first tier are subsequently compared against each other in the second tier. This hierarchical or tournament-style approach has also been applied in other study systems (e.g.,[150,151]). All scenarios in both tiers incorporated bottleneck events, because they all involved divergence of new populations from an existing population, and thus founder effects. Indeed, our inclusion of bottleneck events enabled specification of progenitor-descendant relationships between pairs of diverging populations (as in[152]). Furthermore, the non-negligible role of bottlenecks during climatically-driven population divergence has been established. In one set of analyses in the first tier of ABC comparisons, we assessed scenarios in which *R. flavipes* persisted in a single major refugium (Figure B.3), such that the other areas were colonized via successive expansions out of that refugium. We considered three different refugial locations (i.e., the north, south, or central portion of the study region; see Results – Step 2). In a second set of analyses within the first tier, we assessed scenarios that involved distributional shifts (Figure B.4), whereby populations diverged in a stepping-stone fashion (i.e., one population gave rise to a descendant population, which later became the progenitor of the third population). Here, we considered all possible stepping-stone configurations (i.e., there was no assumption that only nearest neighbors can exhibit a progenitor-descendant relationship). In the second tier of ABC comparisons, the best-fit hypotheses from the refugial and distributional shift scenarios were directly compared, along with an additional hypothesis that incorporated vicariance (Figure B.5). The reason for including this third hypothesis was to test the possibility that the original ancestral population no longer exists, having split into two new populations, one of them giving rise to a third population. While there are other vicariance hypotheses that could have been compared in the first tier, we chose not to do this based on the sequence of divergence events best-fit refugial and distributional shift hypotheses had in common. This reduced the number of plausible vicariance hypotheses to one. ABC model specification, and model choice. Within the ABC framework, two classes of model parameters were used to characterize the phylogeographic hypotheses described above: effective population sizes (Ne), and divergence times (T). We performed two rounds of modeling: 1) a preliminary round with broad priors, and 2) the final round with narrower priors (Table B.5). Briefly, all competing scenarios had two divergence events: any two of $T_N$, $T_C$ or $T_S$, (where the subscript is the first letter abbreviation of the new cluster, i.e., Northern, Central, or Southern), the prior range for the more recent event encompassed the MH and the LGM whereas priors for the older event ranged from the LGM to the LIG assuming a 1-year generation time for *R. flavipes*. Full details of ABC priors on Ne and T parameters are given in Section B.1.4 in Supplementary Material. We set the mtDNA mutation rate priors from $5.0 \times 10^{-9}$ to $5.0 \times 10^{-7}$, a broad range encompassing the Brower[138] and Luchetti[139] rates (see Methods – Step 2). Similarly, since no rates were available for the nDNA locus in *Reticulitermes*, we used broad priors for this locus, from $5.0 \times 10^{-10}$ to $2.5 \times 10^{-8}$. Thus, the mean nDNA rate was an order of magnitude slower than the mean mtDNA rate, despite some overlap at the upper end of nDNA and lower end of mtDNA prior ranges. To characterize the empirical two-locus DNA sequence dataset, we used the following summary statistics: number of segregating sites (one- and two-sample) and private segregating sites (one-sample), mean (one- and two-sample) and variance of pairwise differences (one-sample), mean and variance of numbers of the rarest nucleotide at segregating sites (one-sample), Tajima's D[153] (one-sample), and $F_{ST}$[2] between two samples. ABC runs consisted of $1 \times 10^6$ simulated genetic datasets per competing phylogeographic hypothesis. We then compared the values of summary statistics calculated from simulated datasets to those from the empirical dataset. Following Cornuet et al.[148], model checking was performed via principal components analysis, and then posterior probabilities were calculated via logistic regression[154] on 1% of simulated data most similar to the empirical data, to identify the best-fit model![155]. We evaluated model performance (i.e., the ability to discriminate between the best-supported and alternative scenarios), by estimating type I and type II error rates. To do this, we simulated 500 data sets and estimated the most likely model using a polychotomous logistic regression[155,156]. The type I error rate was the proportion of data sets that were simulated under an alternative scenario but were incorrectly categorized under the best-supported scenario. The type II error rate was the proportion of instances in which the best-supported scenario was incorrectly selected as the most likely scenario. To calculate point estimates and confidence intervals for the values of parameters included in the best-fit model, we selected 1% of the simulated data closest to the observed data. Additionally, for the best-fit scenario, we estimated precision in parameter estimation[156] by computing the relative median of the absolute error for 500 simulated data sets with values drawn from posterior distributions. Population size changes over time. For each of the three *R. flavipes* genetic groups, we assessed evidence for population size changes vs. stability by calculating Tajima's D, and Fu and Li's D* and F*[157] from the mtDNA data, in DnaSP. To identify cases of departure from the null hypothesis of constant size, p-values for these statistics were obtained by computing 10,000 coalescent simulations based on θ from the observed data and assuming no recombination. We also calculated Ramos-Onsins and Rozas'[158] R2 statistic for which significantly small values indicate population growth, whereas significantly large R2 values indicate size reduction. Statistical significance of deviation from the null hypothesis of constant population size was assessed by performing 10,000 coalescent simulations in DnaSP. To complement the above analyses, we also estimated mismatch distributions, where a unimodal distribution indicates growth, whereas a multimodal distribution is indicative of size constancy[159]. Given that signatures of selection can mimic those of population size changes and therefore complicate interpretation of the above summary statistics, we examined evidence for non-neutrality using compound tests[160]. We performed the compound tests using the program DH (`http://zeng-lab.group.shef.ac.uk/wordpress`). The significance ($\alpha$ = 0.05) of each test was determined using 100,000 simulations. We also examined evidence for changes in Ne over time in each cluster by analyzing the combined mtDNA plus (unphased) nDNA sequence data using Extended Bayesian Skyline Plots (EBSP[161]) in BEAST. The same mutation rate parameters for phylogenetic tree estimation were used here, and EBSP searches were run for 50 million Markov chain Monte Carlo generations, with a burn-in of 5 million generations. Samples were saved every 2,500 generations and ESS and the stationarity of likelihood values were examined in order to make sure all ESS values were greater than 500.

## 2.3 Results

### 2.3.1 Genetic data collection

MtDNA sequences were obtained from 122 *R. flavipes* individuals, and the nDNA locus was sequenced from 124 individuals. The mtDNA alignment had 86 polymorphic sites and 32 haplotypes, while the nDNA locus had 5 polymorphic sites and 5 haplotypes (Table 2.1). All sampled logs contained individuals with the same mtDNA haplotype, with the exception of a rotting log sampled at site A41 (see Table B.1), which contained two different haplotypes from the same genetic population, suggesting a rare instance of colony fusion (see DeHeer and Vargo 2004).

### 2.3.2 Step 1: Present and past geographic distributions

When constructing SDMs, a strong correlation was observed among some of the 19 bioclimatic variables (Figure B.6). Three iterations of eliminating variables and factors with low contributions to the total variation were required until all retention criteria were met. Ultimately, four factors (MR1–4, $\alpha > 0.7$; Figure B.7) explained 100% of the variation in eight retained variables, and 84% of the variation in all 19 bioclimatic variables. Correlation among the four factors was lower than among the original variables in all four time periods considered (i.e., present, MH, LGM, and LIG; Table B.6). For convenience, we named the four factors according to the original variables with which they were strongly correlated (r > 0.9; Figure B.7; also see Figures B.8 and B.9). Distributional shift and stability maps (Figure 2.3) showed that: 1) from the LIG to the LGM, most of the suitable habitat shifted northward from the East Coast and the Gulf Coast toward the location of the southern edge of the Laurentide ice sheet, above 40° latitude; 2) from the LIG to the present, the southern edge of *R. flavipes*' distribution underwent a extinction-colonization (or contraction-expansion) cycle; 3) the eastern portion of West Virginia and areas around western North Carolina had suitable habitat from the LIG to the present; and 4) the amount of suitable habitat increased since the beginning of the Holocene.

### 2.3.3 Step 2: Genetic variation and the role of environment and space in genetic structuring

The BAPS analysis identified three genetic clusters, each with largely separate geographic distributions (Figure 2.4a). Herein, we refer to them as the Northern, Central, and Southern clusters. We used the first three principal components (PCs) to represent these clusters in three dimensions (Figure 2.4b). The three PCs accounted for 53% of the variance at the mtDNA locus; they showed that the Northern cluster is most similar to the Central cluster. Phylogenetic reconstruction using BEAST produced a Bayesian tree (Figure 2.4c) that corroborated the three clusters identified using BAPS and PCA, albeit with the Northern cluster as paraphyletic. Molecular dating using the mtDNA locus in BEAST estimated the Southern-Northern divergence at a median of 131.9 kya (95% CI: 83.6–195.0 kya; Figure B.10), and the Northern-Central divergence at a median of 35.8 kya (95% CI: 21.5–56.7 kya; Figure B.10).

**Table 2.1:** *Genetic diversity and tests of neutrality*. K: average number of nucleotide differences; S: segregating sites; $\theta_W$ = Ne$\mu$ for the mtDNA locus and 4Ne$\mu$ for the nDNA locus, where Ne is the effective population size, and $\mu$ is the mutation rate per nucleotide ($\theta_{Wnuc}$) and per generation ($\theta_{Wgen}$); $\pi$: nucleotide diversity. Significance: *****0.01***, ***0.05**, *#0.10*.

| | **Data** | | | **Neutrality** | | |
|---|---|---|---|---|---|---|
| | **Population** | **Locus** | **Individuals** | **TajimaD** | **FuLiD*** | **FuLiF*** |
| mtDNA | Southern | COI | 16 | 0.926 | 0.926 | 0.944 |
| | | COII | | -0.678 | -0.678 | -0.7 |
| | | COI+COII | | 0.027 | 0.027 | 0.028 |
| | Northern | COI | 24 | 0.483 | 0.303 | 0.388 |
| | | COII | | -1.182 | -1.431 | -1.535 |
| | | COI+COII | | -0.289 | -0.511 | -0.513 |
| | Central | COI | 82 | *****-1.957*** | ***-2.270** | ***-2.527** |
| | | COII | | *#-1.476* | *#-1.667* | *#-1.865* |
| | | COI+COII | | *****-1.900*** | ***-2.240** | ***-2.486** |
| | All | COI | 122 | -1.212 | -0.754 | -1.072 |
| | | COII | | ***-1.482** | *#-1.807* | ***-2.008** |
| | | COI+COII | | -1.377 | -1.316 | -1.583 |
| nDNA | All | EB14G | 124 | -0.562 | -0.562 | -0.578 |

| | **Diversity** | | | | | |
|---|---|---|---|---|---|---|
| | **No. of Haplotypes** | **S** | **$\pi$** | **$\theta_{Wnuc}$** | **K** | **$\theta_{Wgen}$** |
| mtDNA | 4 | 14 | 0.015 | 0.014 | 8.333 | 7.636 |
| | 4 | 18 | 0.017 | 0.018 | 9.167 | 9.818 |
| | 4 | 32 | 0.016 | 0.016 | 17.5 | 17.455 |
| | 8 | 18 | 0.013 | 0.012 | 7.278 | 6.623 |
| | 6 | 15 | 0.008 | 0.01 | 4.167 | 5.519 |
| | 9 | 33 | 0.01 | 0.011 | 11.444 | 12.142 |
| | 16 | 15 | 0.004 | 0.008 | 2.199 | 4.578 |
| | 9 | 9 | 0.003 | 0.005 | 1.485 | 2.575 |
| | 19 | 24 | 0.003 | 0.006 | 3.684 | 7.153 |
| | 28 | 46 | 0.014 | 0.021 | 7.823 | 11.671 |
| | 18 | 40 | 0.011 | 0.018 | 6.046 | 10.181 |
| | 32 | 86 | 0.012 | 0.02 | 13.869 | 21.851 |
| nDNA | 5 | 5 | 0.009 | 0.01 | 2.2 | 2.4 |

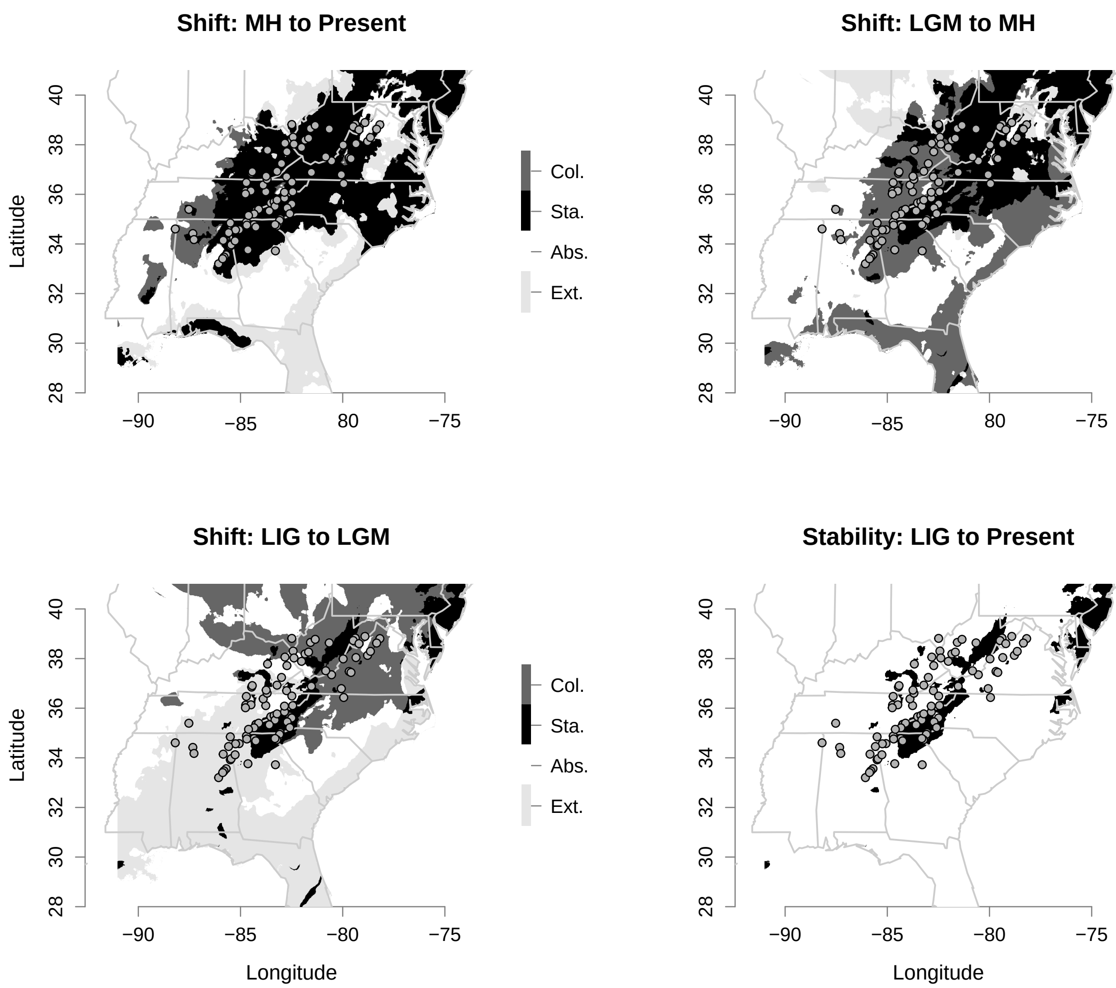



**Figure 2.3:** *Distributional shifts and stability*. Maps showing inferred distributional shifts and long-term stability for successive time periods: MH to present, LGM to MH, and LIG to LGM. Each panel depicts four occurrence categories: colonization (Col.), stability (Sta.), absence (Abs.), and extinction (Ext.). The superimposed gray dots represent the 91 occurrence points used for distribution modeling.

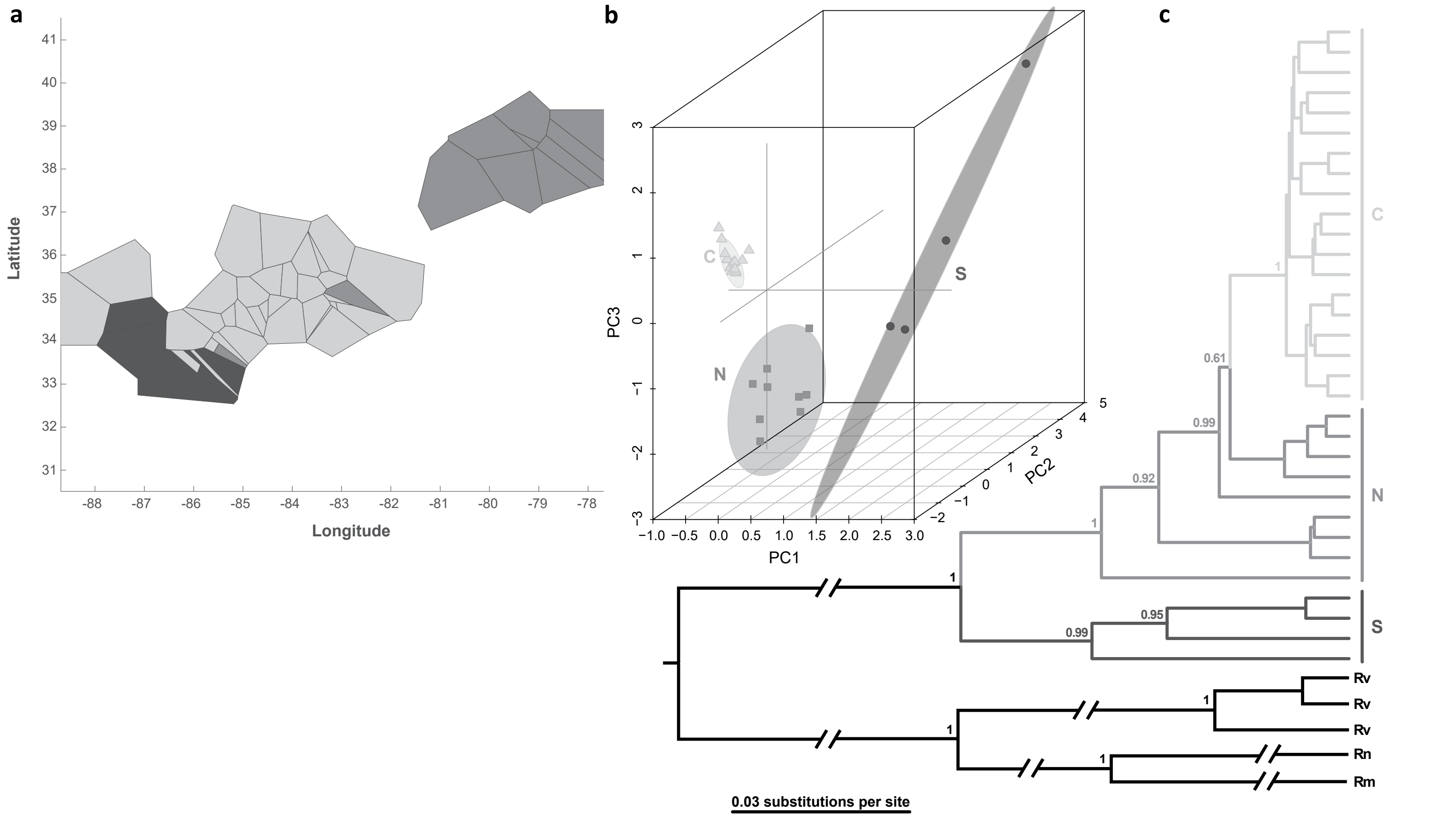



**Figure 2.4:** *Identification of natural genetic populations based on mtDNA sequences.* (a) Bayesian spatial-genetic clustering. The map shows the inferred locations of three genetic clusters recovered using BAPS: Northern (gray), Central (light gray) and Southern (dark gray). (b) Principal Components Analysis. Principal component scores are shown in three dimensions with grouping of individuals according to the BAPS clusters. (c) Bayesian Maximum Clade Credibility tree. For the in-group (*R. flavipes*), nodes and branches are shaded according to the BAPS clusters, and labels with abbreviations as follows: Northern (N), Central (C), and Southern (S). Only those node support values (posterior probabilities) > 0.50 are shown. Abbreviations for out-group taxa are: *R. virginicus* (Rv), *R. malletei* (Rm) and *R. nelsonae* (Rn).

Although the Southern cluster comprised only four mtDNA haplotypes, this group had the most genetic variation (nucleotide diversity, $\pi = 0.016$; mean number of nucleotide differences, K = 17.50; Table 2.1). Nine mtDNA haplotypes in the Northern cluster resulted in values of $\pi = 0.010$, and K = 11.44, and, although there were 19 haplotypes in the Central cluster, these diversity values were lowest (i.e., $\pi = 0.003$ and K = 3.68; Table 2.1). Genetic differentiation was highest between Southern vs. Central clusters ($F_{ST} = 0.659$) whereas Northern vs. Central differentiation was lowest (Table B.7). Genetic structure was influenced by environment and geography. The full model of environmental and spatial structure predictors accounted for 58.7% of the observed genetic variation at the mtDNA locus. Spatial structure alone explained 41.1% ($p < 0.001$) of the genetic variation. Environmental factors accounted for 5.2% ($p = 0.012$) of the variation. The interaction between the two explained an additional 12.4% of the genetic variation. After removing the effect of spatial structure, the factors with significant contribution to genetic variation were "temperature range" and "wet-season precipitation" (Figure B.11).

### 2.3.4 Step 3: Phylogeographic hypothesis testing and population size

In the two sets of first-tier ABC comparisons: 1) the refuge-based scenario with the highest posterior probability was the hypothesis that postulated the Northern region was the source from which the Southern cluster diverged first, followed by the Central cluster (scenario R3; Table 2.2; Figure B.3); and 2) the distributional shift scenario that provided the best fit to the empirical data was the hypothesis that represented a case of Southern-to-Northern-to-Central stepping-stone colonization (scenario DS1; Table 2.2; Figure B.4). In the second tier of ABC comparisons, the best-fit scenario was DS1 (Table 2.2; Figure B.5). The DS1 scenario had a posterior probability of 0.932 when compared against other DS scenarios in the first tier, but its posterior probability in the second tier was 0.495 compared to 0.332 for the second-best R3 scenario. Both of these scenarios had high type I and II error rates in the second-tier comparisons (Table B.8). Based on examination of estimated parameter values from the best-fit model, divergence between the Northern and Southern populations was the oldest, estimated to have occurred 64.80 kya (95% CI: 26.40–115.00 kya; Figure 2.5a;Table B.9), while the Northern and Central populations diverged 8.63 kya (95% CI: 2.75–22.50 kya; Figure 2.5a;Table B.9).

**Table 2.2:** *Two-tiered ABC hypothesis testing.* Best-fit scenarios are highlighted in bold font. ABC hypothesis testing was performed in two tiers. In the first tier, refugial and distributional shift scenarios were evaluated separately. In the second tier, these two scenarios, as well as a vicariance scenario (V; Figure B.5), were compared.

| ***Refugial Scenarios*** | | |
|---|---|---|
| *Scenario* | *Posterior Probability* | *95% CI* |
| R1: S-N;S-C | 0.103 | (0.087–0.120) |
| R2: S-C;S-N | 0.014 | (0.010–0.018) |
| **R3: N-S;N-C** | **0.861** | **(0.843–0.879)** |
| R4: N-C;N-S | 0.013 | (0.009–0.016) |
| R5: C-S;C-N | 0.006 | (0.003–0.009) |
| R6: C-N;C-S | 0.003 | (0.001–0.005) |

| ***Distributional Shift Scenarios*** | | |
|---|---|---|
| *Scenario* | *Posterior Probability* | *95% CI* |
| **DS1: S-N;N-C** | **0.932** | **(0.918–0.946)** |
| DS2: S-C;C-N | 0.002 | (0.001–0.003) |
| DS3: N-S;S-C | 0.064 | (0.050–0.078) |
| DS4: N-C;C-S | 0.002 | (0.001–0.003) |
| DS5: C-S;S-N | 0.000 | (0.000–0.001) |
| DS6: C-N;N-S | 0.000 | (0.000–0.001) |

| ***Refugium vs. Distributional Shift vs. Vicariance*** | | |
|---|---|---|
| *Scenario* | *Posterior Probability* | *95% CI* |
| R3: N-S;N-C | 0.332 | (0.313–0.351) |
| **DS1: S-N;N-C** | **0.495** | **(0.481–0.510)** |
| V: N/S;N-C | 0.173 | (0.159–0.187) |

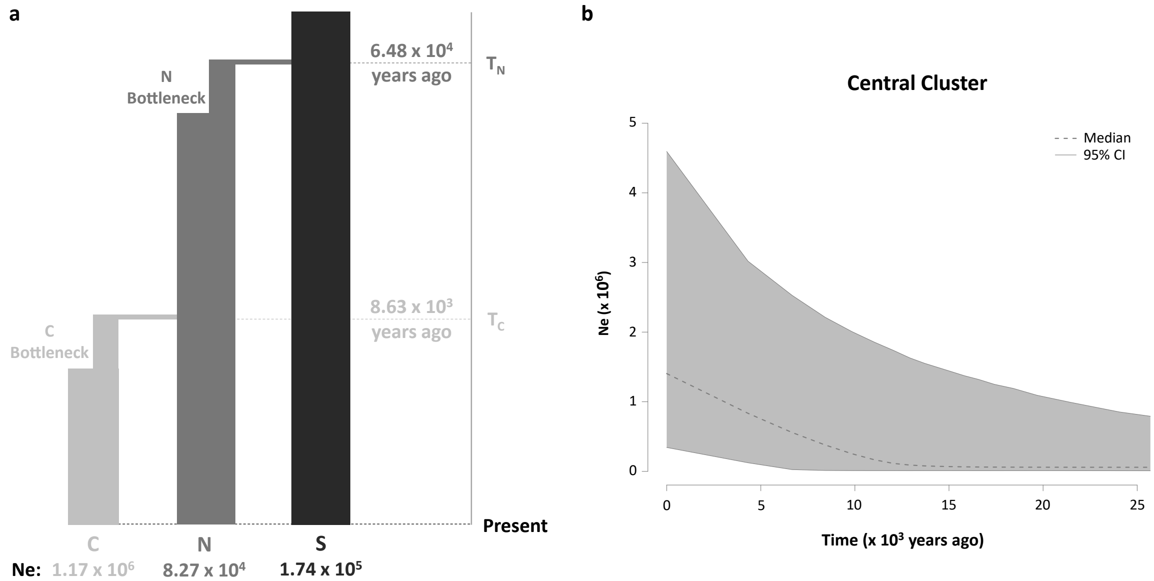



**Figure 2.5:** (a) Best-fit phylogeographic scenario inferred using ABC. The distributional shift hypothesis represents a case where the Northern (N) cluster first diverged from the Southern (S) cluster, and the Central (C) cluster subsequently diverged from the Northern cluster, in a stepping-stone fashion. Branch widths of the population tree represent effective population sizes (Ne), and the model includes brief bottlenecks associated with each founder event (see Section 2—Step 3). (b) Extended Bayesian skyline plot. The plot shows changes in effective population size (Ne) over time in the Central cluster, jointly estimated from mtDNA and nDNA data.

The Central population was the only cluster that showed a signature of population growth, based on significant results for Tajima's D (D = -1.90 for the mtDNA locus; Table 2.1), as well as Fu and Li's statistics (D = -2.24, F = -2.49; Table 2.1). Likewise, mismatch distribution analyses revealed evidence of population growth in the Central cluster only. This population experienced significant growth ($R2 = 0.047$; $p < 0.001$), whereas no size changes were detected in the Northern ($R2 = 0.166$; $p = 0.479$), or the Southern ($R2 = 0.154$; $p = 0.116$) clusters. The EBSP assessments of changes in Ne over time also showed evidence of growth of the Central cluster, initiated in the last 10,000 years (Figure 2.5b). Furthermore, non-significant outcomes from compound neutrality tests for the mtDNA locus suggested that the aforementioned inferences were not obscured by selection (Table B.10).

## 2.4 Discussion

This study provides new insights into how Pleistocene climatic fluctuations impacted the geographic distribution of *R. flavipes* in the southern Appalachian

Mountains and surrounding areas. The interplay between past climate change and complex montane topography, and its impact on the spatial distribution of intraspecific genetic diversity has been reported for other taxa from temperate regions[101]. While there has been extensive work on salamanders from the southern Appalachians (e.g.,[40,162–166]), relatively few studies have focused on reconstructing the long-term population history of forest-dependent arthropods in this region (but see[91,109,110,167,168]). Indeed, the predominant focus on vertebrates and vascular plants in conservation research and planning is likely to result in management strategies that fail to cater to a large proportion of biodiversity ([45] and references therein). To understand drivers of phylogeographic patterns in *R. flavipes*, we examined evidence for distributional shifts using SDMs, and reconstructed the evolutionary and demographic history of *R. flavipes* using ABC analyses. Overall, we determined that the location of key refugia has changed over time (e.g., from one glacial period to the next), rather than a single refugium repeatedly serving as a reservoir of genetic diversity, whereby successive glacial-interglacial cycles reinforce the same genetic signatures of contraction and expansion.

### 2.4.1 Climate change as a driver of distributional shifts and genetic divergence

Determining whether distributional shifts have occurred in the history of a species can lead to a better understanding of processes that have shaped present-day genetic variation. Our SDMs suggested that in the period between the LIG and LGM, suitable habitat for *R. flavipes* shifted from the East Coast and the Gulf Coast northward toward the former southern edge of the Laurentide ice sheet (Figure 2.3). Consistent with this, our genetic analyses confirmed that the Northern cluster diverged between the LIG and LGM (ABC: 26.4–115.0 kya; BEAST: 83.6–195.0 kya). As suitable habitat expanded southward following the LGM (Figure 2.3), the Central cluster diverged during the LGM-Holocene transition (ABC: 2.8–22.5 kya; BEAST: 21.5–56.7 kya) and continued to expand in the Holocene, both in terms of geographic range (Figure 2.3) and population size (Figure 2.5). Our inferences about the long-term population history of *R. flavipes* are not dissimilar from reconstructions of glacial-interglacial colonization routes followed by many plant and animal species in the eastern U.S. For example, the pitcher-plant mosquito, *Wyeomyia smithii*, initially dispersed from the Gulf Coast northward along the East Coast, and subsequently moved southward into the southern Appalachians[169]. Similarly, the red salamander, *Pseudotriton ruber*, persisted in the Coastal Plain in the early Pliocene, and then expanded its range toward Appalachian upland habitat as cooling trends started in the early Pleistocene[170]. Thus, despite different life history traits, at least a few forest-dependent organisms may have responded similarly to climatic fluctuations in the past.

### 2.4.2 A northern refugium during the LGM and divergence of the Central cluster in the Holocene

Our analyses suggested that a northern refuge played a key role in subsequent colonization by *R. flavipes* of the central region of the southern Appalachians. Pollen records indicate that climatic conditions suitable for temperate forests existed over large areas of the southeastern U.S. during the LGM[37]. Furthermore, fossil and genetic evidence suggests that some tree species, including red oak, red maple and beech, were widespread in this region during that time[171,172]. Although somewhat unexpected, the existence of northern refugia close to the southern edge of the Laurentide ice sheet during the LGM is plausible owing to localized warm areas in close proximity to glaciers (e.g.,[37,171–175]). Despite the broad geographic range of the *R. flavipes* Central cluster (Figure 2.4a), this group contained the lowest genetic diversity (Table 2.1). We suggest that this is likely the result of founder effects associated with the relatively recent colonization of the central portion of the southern Appalachians from the north. Although subsequent population expansion seems to have occurred in the central region, more time may be needed to replace lost genetic variation. Assessment of changes in Ne over time showed that the Central cluster had increased in size over the last 10,000 years (Figure 2.5b), which is consistent with inferences based on non-genetic data that indicated the amount of suitable habitat in the central region increased since the LGM (Figure 2.3).

### 2.4.3 The potential role of environmental variables in promoting range expansions

Given the desiccation susceptibility of soft-bodied arthropods, range expansions and population growth in *R. flavipes* may be have been influenced by local-scale site-specific environmental variables such as precipitation. The southeastern U.S. was much warmer during the mid-Holocene (cf. LGM[176]), when tupelo and oak forest types dominated over pine, indicating wetter conditions[177].

The *R. flavipes* Central cluster likely diverged from the Northern cluster following a cooling trend in the Younger Dryas ( 12.9–11.7 kya). While this was a global cooling period, locally in the southeastern U.S., this period was characterized by a warmer and wetter climate, reflecting the trapping of heat in the western subtropical gyre due to reduced Atlantic meridional overturning circulation[178]. Accordingly, if high precipitation was important for facilitating range expansion, these conditions seem to have been in place at a time that coincides with colonization of the central region. Furthermore, seasonal differences in precipitation between the southern and northern portions of the study region[179] may have led to different flight phenologies and thus seasonal isolation and niche partitioning. Consistent with this, dbRDA revealed that in addition to spatial structuring of genetic variation, wet-season precipitation accounted for the remainder of genetic differentiation of the Southern cluster compared to the other two. We suggest that the influence of local-scale environmental variables upon the capacity for termite population growth and range expansion warrants further investigation.

### 2.4.4 The influence of spatial scale on genetic structure

Compared to previous work on *R. flavipes*, the spatial scale over which we detected genetic structure is notable. For example, based on mtDNA sequence and microsatellite genotypic data, Perdereau et al.[25] identified three distinct genetic clusters of *R. flavipes* in the eastern and southeastern U.S. across an area spanning at least twice the distance covered by sampling in the present study. However, with the exception of a few collection sites in West Virginia, those authors did not include *R. flavipes* sampled from the southern Appalachians. This contrast supports the view that fine-scale genetic structuring may be particularly prevalent in topographically complex montane areas (e.g.,[109,110,180]). Along a 1,000 km transect traversing the southern Appalachians, a wood-feeding cockroach (*Cryptocercus punctulatus*) that is syntopic with *R. flavipes* consists of five distinct genetic groups[44,181]. Interestingly, both of these saproxylic taxa have a zone of parapatry between genetic groups in the central region. Comparative phylogeographic analyses would be informative about the extent to which spatial-genetic patterns seen in dead-wood-associated insects correspond with shared microevolutionary processes that underpin them.

### 2.4.5 Caveats and future directions

An early understanding of genetic consequences of Pleistocene range expansions came from study systems that either repeatedly experienced severe glaciation (e.g.,[99]), or were relatively simplified linear systems (e.g.,[182]). In these cases, unidirectional expansion out of a single major refuge was commonly inferred, often based on signatures of repeated founder effects and serial reduction in genetic diversity at the leading edge. However, an expanded view of the geography of range expansion may be needed when considering unglaciated, topographically complex, montane landscape settings. In this study, we considered distributional shift (see Introduction) to be a plausible phylogeographic scenario for the southern Appalachian Mountains. However, further work is needed to understand the circumstances under which distributional shift scenarios are distinguishable from single-refuge contraction-expansion scenarios. Indeed, inferring Pleistocene distributional shifts using genetic data can be challenging, as multiple historical factors can contribute to current genetic variation.

Although our ABC analyses identified distributional shift as the best-fit scenario, it did not receive unambiguously superior support relative to the next-best scenario, and the estimated error in scenario choice was large (Table B.8). Accordingly, we must consider our ABC-based inference to be a preliminary working hypothesis, to be re-evaluated and re-tested with new data. Notwithstanding some limitations of our ABC inferences, it is notable that a common feature of the best-fit and second-best hypotheses is the expansion of the Central cluster. Specifically, both scenarios include the Northern cluster giving rise to the Central cluster. Additionally, both scenarios include a direct long-distance dispersal event. Buckley[183] advocated for an iterative approach to phylogeography, highlighting the value of working hypotheses for focusing subsequent analytical efforts on scenarios that have some empirical support. This study contributes to a growing body of literature that highlights an important role for multiple refugia—including those located further north than previously expected—in phylogeographic structuring of plants[172], vertebrates[184], and invertebrates[169]. Having characterized contemporary fine-scale spatial structure and historical climate-based distributions for *R. flavipes*, the present study has also revealed specific geographic locations that warrant dedicated sampling (e.g., the Southern genetic cluster has a relatively small range that requires better representation, and based on SDMs, sampling in the Gulf Coast and Coastal Plain areas would be particularly valuable).

Data Accessibility: The Supplementary Material and additional SDM, BAPS, BEAST, and ABC data are available for download from DRYAD via `http://datadryad.org` under repository entry DOI: `https://doi.org/10.5061/dryad.5hr7f31`. All appendices are included in Supplementary Material – File 1 (Supplementary Methods and Supplementary Results). All DNA sequence data are included in Supplementary Material – File 2, with Genbank accession numbers provided in the file. Posterior probabilities and error rates for all phylogeographic hypotheses tested in this study are included in Supplementary Material – File 3.

CHAPTER 3:

# CANOPY COVER AND TREE SPECIES RICHNESS MODULATE EPIGENETIC CHANGES IN EASTERN SUBTERRANEAN TERMITES IN APPALACHIAN FOREST ECOSYSTEMS

CITATION: Hyseni C, Garrick RC. Canopy cover and tree species richness modulate epigenetic changes in eastern subterranean termites in Appalachian forest ecosystems. **In Prep.** *2020*.

ABSTRACT: The eastern subterranean termite, *Reticulitermes flavipes*, native to the eastern United States, has been unintentionally introduced into other parts of the country, as well as in South America and Europe. Epigenetic mechanisms, such as DNA methylation, may play a role in facilitating biological invasions. Furthermore, expansion into human-altered habitats in the native range may precede establishment of species in similar human-altered habitats elsewhere. Thus, we hypothesized that disturbance of forest ecosystems in a portion of the native range of *R. flavipes* (i.e., the southern Appalachian Mountains) would have increased epigenetic variation in this termite species. Ultimately, if true, this may have played a role in the species becoming invasive elsewhere in the U.S. and the world. To characterize DNA methylation changes in *R. flavipes*, we screened 167 individuals from 45 sampling sites for variation in DNA methylation using the methylation-sensitive amplified fragment length polymorphism method. We assessed evidence of epigenetic divergence among individuals and used machine learning algorithms to classify individuals into distinct epigenetic groups (i.e., clusters). In addition to long-term influences leading to epigenetic divergence, we also assessed evidence of short-term environmental effects on epigenetic variation. Overall, we detected four epigenetic clusters. In addition, we found that wet-season precipitation and

summer temperature exerted a long-term influence on epigenetic variation. Importantly, disturbance of forest ecosystems, indirectly captured by tree canopy cover and tree species richness, had short-term effects on methylation at individual loci. This is the first study to show an effect of canopy cover on intraspecific epigenetic variation in termites.

## 3.1 Introduction

### 3.1.1 Phenotypic plasticity, epigenetics, and eusocial insects

Phenotypic plasticity is an important biological phenomenon that allows organisms to modulate their phenotypes in response to different biotic and abiotic environments. Eusocial insects display remarkable phenotypic plasticity (e.g.,[185]). To date, epigenetic mechanisms have been associated with the phenotypic differences observed among castes (e.g., workers, soldiers, reproductives) in ants[10], bees[11] and wasps[12], as well as termites[13–15]. Epigenetic mechanisms affect gene expression without changes in DNA sequence. Three main mechanisms of epigenetic control of gene expression have been characterized: methylation of nucleic acids (DNA and RNA), covalent modifications of histone tails, and non-coding RNAs[186].

### 3.1.2 DNA methylation and biological invasions

Invasive species may rely on phenotypic plasticity to deal with stressful environmental conditions in non-native environments. For instance, DNA methylation variance has been shown to increase in response to stressful conditions (e.g.,[187,188]). Initial genetic variation can be low due to genetic bottlenecks associated with the introduction of a small number of individuals into non-native ranges. Thus, changes in DNA methylation may be the primary means of dealing with habitat change, especially in novel environments during biological invasions[189–192].

Partly due to their extraordinary capacity for phenotypic plasticity, eusocial insects represent some of the most important invasive species in the world. As an example, two invasive termite species have been shown to shift their reproductive phenology (i.e., timing of spring migration and breeding) in a non-native environment[193]. This plasticity of phenology may be underpinned by epigenetic mechanisms. For instance, in barn swallows, methylation of the photoperiodic *Clock* gene plays a major role in regulating phenology[194].

The number of invasive termite species has increased from 17 in 1969 to 28 at present[195], and these species are likely to further expand their geographic ranges in the near future. Using species distribution modeling (SDM), Buczkowski et al.[21] predicted geographic expansion by 2050 for 12 of the 13 termite species they examined, including the eastern subterranean termite, *Reticulitermes flavipes* (Kollar). This species is native to the eastern United States, and has been unintentionally introduced into other parts of the U.S. (e.g., Oregon[22]), as well as other countries, in the Americas (e.g., Canada, Chile, and Uruguay), and in Europe (Austria, France, Germany, Italy, and even the Canary Islands)[23–26].

### 3.1.3 The southern Appalachian Mountains and subterranean termites

The southern Appalachian Mountains extend latitudinally from northeast Alabama to northwest Virginia. Steep altitudinal precipitation gradients, a complex heavily dissected topography, and a temperate climate, have shaped southern Appalachian forests into some of the most diverse environments in the eastern United States[31]. This diversity of environments supports high levels of species richness (e.g., darters[196]), including organisms that inhabit dead wood or use it for shelter (e.g., salamanders[39], millipedes[42]). Dead wood is a key factor in maintaining biodiversity and the functioning of forest ecosystems.

Dead-wood-associated arthropods are functionally important members of montane temperate forests[46–50]. Wood-feeding insects (together with wood-decaying fungi) are key ecosystem engineers that make major contributions to dead wood decomposition and nutrient cycling in forests[48]. Of these insect taxa, *R. flavipes* is an important early colonizer of standing moribund trees and snags, as well as fallen logs on the forest floor[46–50]. In areas of the southern Appalachians where commercial forestry operations occur, woody debris generated during logging is quickly colonized by *R. flavipes*. Thus, the species is capable of sustaining viable colonies in a variety of forest types, from unmanaged wilderness to intensively managed production forests.

### 3.1.4 Population expansion of *R. flavipes* and human-altered forest ecosystems

The distribution of *R. flavipes* covers a wide range of environments compared to two other co-occurring species in the eastern U.S., *R. malletei* and *R. virginicus*[179]. Based on outcomes from SDMs, *R. flavipes* is potentially able to exclude the other two species in the northern portion of the southern Appalachians, including western Kentucky, southern Ohio and Indiana, the majority of West Virginia and Pennsylvania, and parts of Virginia and North Carolina[179]. By modeling past changes in the geographic distribution of *R. flavipes* in the eastern U.S., Hyseni and Garrick[197] showed that the species has likely persisted in northern refugia during Pleistocene glaciation. Time-series SDMs, encompassing a period from 120,000 years ago to the present, suggested that the distribution of *R. flavipes* has cycled latitudinally, shifting northward toward the southern edge of the Laurentide ice sheet (e.g., Indiana, Ohio, Pennsylvania) during the Last Glacial Maximum ( 22,000 years ago), then shifting southward in the Holocene, with *R. flavipes* populations having undergone expansion in the last 9,000 years[197].

In the last five centuries–since the European settlement of North America–the expansion of *R. flavipes* has coincided with ever-increasing human-induced environmental change, including disturbance and degradation of forest ecosystems in the eastern U.S. These forests were historically dominated by fire-tolerant oak (*Quercus*) and pine (*Pinus*) species[198,199]. These open old-growth forests of less shade-tolerant oak and pine were common and succession to more shade-tolerant species, such as beech (*Fagus grandifolia*), was rare in the eastern U.S. before the 1600–1800s[200]. Extensive harvest and exclusion of fire has affected the composition of eastern U.S. forests. These forests are now on average only 40–80 years old[201].

With the climate of the last 9,000 years being conducive to population expansion of *R. flavipes*, and the new context of human-induced disturbance of forest ecosystems, the species has expanded its niche to include human-altered habitats. As a mechanism to deal with novel environments, phenotypic plasticity underpinned by DNA methylation may have played a part in the survival and establishment of *R. flavipes* in human-altered habitats in the species' native range in the eastern U.S. If so, this may have been the prelude to *R. flavipes* becoming invasive in other parts of the world. This would not be surprising, as there are numerous examples of species that become 'invasive' (i.e., dominant) in their native range[202–205].

Our goal here was to determine whether any increases in epigenetic variation of *R. flavipes* can be attributed to human-altered habitats within the native range of the species, focusing on the southern Appalachian Mountains. Given that human-induced changes to forest ecosystems in the eastern U.S. resulted in recent (40–80 years) re-structuring of these forests[201], we specifically investigated the potential effect of tree canopy cover and tree species richness on epigenetic variation in *R. flavipes*. Additionally, we assessed evidence for any effect of proximity to urban areas on epigenetic variation in *R. flavipes*.

## 3.2 Methods

### 3.2.1 Workflow

To address the goal of this study, we first assessed evidence for population stratification. Caste identity (workers and soldiers) constituted one layer of population stratification. Since termite colonies are composed of different castes, which interact with their environments differently (e.g., workers can digest cellulose, while other castes cannot, and have to be fed by workers[206]), our sampling included both workers and soldiers. If present, epigenetic divergence among individuals would constitute the second layer of population stratification. To characterize epigenetic divergence, we identified distinct epigenetic groups (i.e., clusters) and classified individuals into these epigenetic clusters. This portion of epigenetic variation is likely dependent on genetic variation.

After characterizing population stratification, we examined evidence for long-term and short-term influences on epigenetic variation. First, we determined the portion of epigenetic variation explained by the following long-term influences: 1) population stratification, 2) spatial structure or autocorrelation (here we refer to it as geography), and 3) environment. Then, after controlling for any long-term influences, we identified short-term influences on the remaining portion of epigenetic variation. To do so, we: 1) determined whether any colonies consisted of multiple epigenetic clusters (as a measure of within-colony epigenetic variation), and 2) whether increased within-colony epigenetic variation was associated with specific environments. Additionally, we identified any loci significantly correlated with environmental predictors (after controlling for long-term influences), and determined the effect of the environment on methylation state at these loci.

### 3.2.2 Data collection

#### 3.2.2.1 Geographic sampling

To identify *R. flavipes*, we performed molecular taxonomic identification (one termite per log) using Garrick et al.'s[54] PCR-RFLP assay. This method generates diagnostic species-specific banding patterns using sequential digestion with three restriction enzymes (RsaI, TaqI, and MspI) of a 376-bp region of the mitochondrial COII gene.

We aimed to capture environmentally-driven epigenetic variation (if it exists) in *R. flavipes* from the southern Appalachian Mountains. We used a sampling design which included a diverse set of environments found in this region. Specimens of *R. flavipes* were collected between June and October of 2016. We collected samples from rotting logs within forests at 45 sampling sites (one log per site). The sampling included the following ecoregions: the Appalachian Plateaus (21 sites), the Blue Ridge (11), the Valley and Ridge (10), and a few sites (3) in the Piedmont region (Table C.1; Figure 3.1). Spatial coordinates and elevation of each rotting log were recorded with a handheld GPS unit (Table C.1), and specimens were stored in 95% ethanol at 4°C. The mean elevation of *R. flavipes* sampling sites was 347 m in the Appalachian Plateaus, 680 m in the Blue Ridge, 496 m in the Valley and Ridge, and 312 m in the Piedmont.

#### 3.2.2.2 Epigenetic data

Termites sampled from each log were identified as soldiers or workers (no alates or secondary reproductives were sampled). We collected epigenetic data for 0-1 soldiers and 1-4 workers per log, totaling 167 individuals from 45 sampling sites (Table C.1). We screened these samples for variation in DNA methylation using the methylation sensitive amplified fragment length polymorphism (MS-AFLP) method[207] (see Supplementary Material), which modifies the standard AFLP protocol by substituting the MseI enzyme with the methylation-sensitive isoschizomeric enzymes MspI and HpaII (Promega, Wisconsin, USA). See Supplementary Material for details on digestion reactions, adapter construction and ligation, and PCR conditions (primer sequences are provided in Table C.2, and a schematic of the MS-AFLP protocol in Figure C.1).

Each termite DNA sample was digested twice, in separate reactions, using the restriction enzyme EcoRI either with MspI or HpaII. According to the restriction enzyme database, REBASE (`http://rebase.neb.com/rebase/rebase.html`), MspI can cleave non-methylated CCGG sequences and hemi- (one strand only) or fully methylated C*m*CGG sequences but not hemi- and fully methylated *m*CCGG and *m*C*m*CGG sequences, whereas HpaII digests only non-methylated CCGG sequences and hemi-methylated *m*CCGG sequences from all possible methylated CCGG variants.

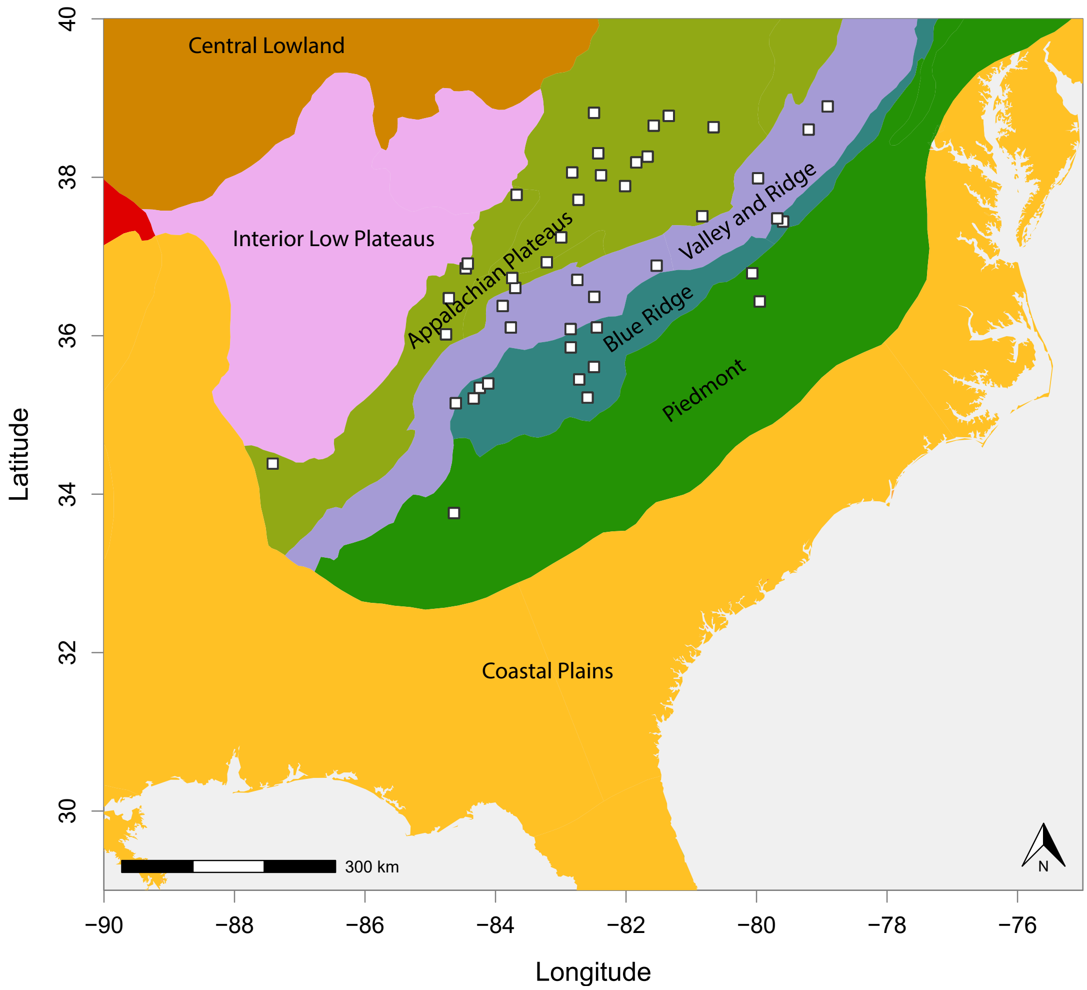



**Figure 3.1:** *Appalachian ecoregions and* R. flavipes *sampling sites.* Ecoregions are color coded and labeled. Sampling sites are shown as white squares with black outlines.

Using two reactions per sample, we can score four different methylation states: unmethylated (**CCGG**), when both enzymes cut at the restriction site; hemi- or fully methylated internal cytosine (**C*m*CGG**), when MspI cuts and HpaII does not cut; hemi-methylated outer cytosine (***m*CCGG**), when MspI does not cut and HpaII cuts; and the fourth state, when neither enzyme cuts, due to: a) the outer cytosine being fully methylated, or b) both cytosines being hemi- or fully methylated, or c) the restriction site having mutated. While Zhang et al.[208] showed that fragment absences actually represent methylation polymorphisms (i.e., possibilities a or b) rather than sequence variation (possibility c), we opted for the more conservative approach and consider this fourth state uninformative. We used "Mixed Scoring 2," as per Schulz et al.[209], which allowed us to distinguish between the three informative methylation states, by converting loci from binary presence/absence of MspI and HpaII fragments to three binary methylation states per locus (see Table 3.1).

To determine genotyping error rates, we ran 32 replicates from PCR to fragment analysis. Scoring of fragments and genotyping error rate analyses were carried out in the R environment[210] (code provided at `https://github.com/chazhyseni/msaflp`).

**Table 3.1:** *Scoring of MS-AFLP loci.* Loci are converted from binary presence/absence of MspI and HpaII fragments to three binary methylation states. With this scoring method, the fourth (uninformative) state cannot be directly discerned (e.g., individual 4 has a 0 for all three methylation states. The table shows a one-locus example. Individuals are abbreviated as "Ind.", enzymes as "Enz.", and loci as "Loc.".

| **Ind.** | **Enz.** | **Loc.1** |
|---|---|---|
| *Ind.1* | MspI | 1 |
| *Ind.1* | HpaII | 1 |
| *Ind.2* | MspI | 1 |
| *Ind.2* | HpaII | 0 |
| *Ind.3* | MspI | 0 |
| *Ind.3* | HpaII | 1 |
| *Ind.4* | MspI | 0 |
| *Ind.4* | HpaII | 0 |

--→

| **Ind.** | **Loc.1 (CCGG)** | **Loc.1 (C*m*CGG)** | **Loc.1 (*m*CCGG)** |
|---|---|---|---|
| *Ind.1* | 1 | 0 | 0 |
| *Ind.2* | 0 | 1 | 0 |
| *Ind.3* | 0 | 0 | 1 |
| *Ind.4* | 0 | 0 | 0 |

### 3.2.2.3 Environmental data

To construct a set of environmental predictors relevant to subterranean termite survival, we used climatic (precipitation and temperature seasonality) and soil moisture variables. Subterranean termite workers and soldiers are soft-bodied and thus prone to dessication; they require high humidity for survival[112]. At lower temperatures, they experience lower body water loss[211]. At high temperatures, high humidity increases survival[112]. Thus, precipitation and temperature seasonality are important factors. To capture this seasonality, we used a set of four weakly correlated precipitation and temperature "factors" obtained from `https://doi.org/10.5061/dryad.5hr7f31`[197,212]. These factors were calculated via factor analysis[179,197] from the 19 strongly correlated WorldClim (`http://www.worldclim.org`) bioclimatic variables, which represent long-term averages for a period from 1960-1990.

Since subterranean termite habitat includes soil in addition to above-ground rotting logs, we obtained soil property data from the International Soil Reference and Information Center database (`https://www.isric.org/explore/soilgrids`), a collection of maps of the spatial distribution of soil properties across the globe interpolated from soil profile observations (`https://www.isric.org/explore/wosis`). As an indicator of the soil's ability to retain water, we used available water capacity (AWC), both at 5 cm (AWC5cm) and 30 cm (AWC30cm) depths.

In order to capture disturbance of forest ecosystems, we used variables that may correlate with disturbance, such as tree canopy cover and tree species richness. We obtained remote-sensed satellite data for tree (canopy) cover (`https://lpdaac.usgs.gov/products/gfcc30tcv003/`), i.e., estimates of the percentage of horizontal ground in each 30-m pixel covered by woody vegetation > 5 m in height. We used tree cover data for a period from 2007-2013, as this preceded our sampling in 2016. To capture species richness of historically dominant pine and oak, we inferred pine and oak species richness using stacked species distribution modeling (SSDM)[213] based on species occurrence records for pine (13 species) and oak (6 species) obtained from the USGS Biodiversity Information Serving Our Nation database (`https://bison.usgs.gov`) and the four environmental factors[179,197] as predictors. Ensemble SSDM was performed by calculating weighted averages of probabilities of occurrence predicted separately by three machine learning algorithms: artificial neural networks[214], boosted regression trees[215], and the random forest algorithm[216].

The dataset of environmental predictors included nine variables, but after using a cutoff of $r = 0.7$ for Pearson's correlation coefficient, we excluded the temperature range[179,197] and AWC5cm variables. Thus, the final environmental dataset comprised seven environmental predictors (Figure 3.2). All environmental data were compiled from online databases and processed using R scripts, including resampling to bring all environmental layers to the same resolution, 1 km.

Aside from these environmental predictors, to address whether increases in epigenetic variation were associated with human-altered habitats, we calculated the distance from urban areas for each sampling site. We were operating under the assumption that greater human-induced environmental disturbance would be reflected by shorter distances from urban areas. To calculate these distances, we first obtained remote-sensed data on urban extent[217] from the Socioeconomic Data and Applications Center (`https://sedac.ciesin.columbia.edu/`). We then calculated distances from each sampling site to the nearest location in the urban extent layer. We used these distances to determine whether proximity of urban areas had an effect on within-colony epigenetic variation.

### 3.2.3 Data Analysis

#### 3.2.3.1 Epigenetic clustering

In order to infer the number of epigenetic clusters, we used two machine learning techniques, non-negative matrix factorization (NMF), as implemented in the R package 'LEA'[218], and discriminant analysis of principal components (DAPC), as implemented in the 'adegenet' package[219]. NMF is a case of unsupervised machine learning. To determine the number of epigenetic clusters that best fit the data, we performed clustering for several values of K (1 to 15). We performed clustering with 500,000 iterations. We used cross-entropy minimization to evaluate the validity of clusters and thus infer the optimal value of K. Cross-entropy values increased beyond K = 5, thus we did not have to consider values of K beyond the preliminarily chosen value of 15.

Discriminant analysis of principal components (DAPC) is an unsupervised-supervised machine learning technique. In DAPC, data is first transformed using a principal components analysis (PCA) and subsequently clusters are identified using discriminant analysis (DA). PCA is unsupervised, whereas DA is supervised, meaning that we require prior knowledge of data classification. Thus, in order to classify each individual into groups, we used K-means clustering. We performed clustering for several values of K (1 to 15) and evaluated cluster validity using the Bayesian information criterion (BIC)[220], as implemented in the 'find.clusters' function of the 'adegenet' package. Using 30 principal components (PCs) to represent the original binary MS-AFLP data, BIC was calculated for 100 replicates of K-

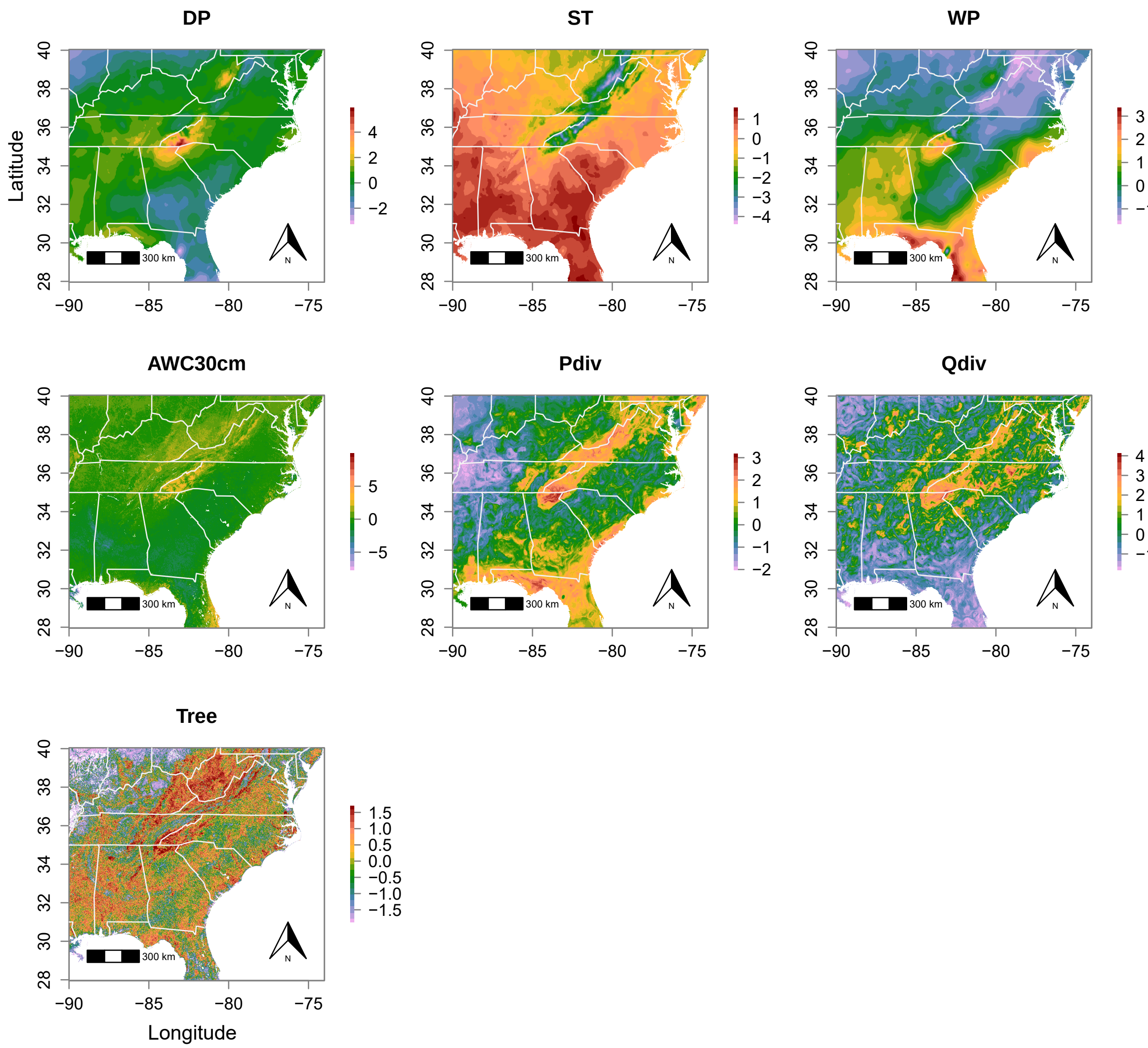



**Figure 3.2:** *Environmental predictors.* Maps of scaled (mean = 0, unit variance) environmental variables. Correlations among all seven variables shown here were below $r$ = 0.7. DP = dry-season precipitation; ST = summer temperature; WP = wet-season precipitation; AWC30cm = available water capacity at a soil depth of 30 cm; Pdiv = pine (*Pinus*) species richness; Qdiv = oak (*Quercus*) species richness; Tree = tree (canopy) cover. High positive values are shown in dark red, low negative values are shown in pink.

means clustering, for K = 1 to 15. As with NMF and cross-entropy, BIC increased continuously beyond K = 5, so the final upper limit was K = 15.

#### 3.2.3.2 Long-term influences on epigenetic variation

*Multivariate modeling*. To detect sources of variance in DNA methylation (i.e., epigenetic variation), we performed distance-based redundancy analysis (dbRDA)[87] using the 'capscale' function in the R package 'vegan'[86]. The response variable was a matrix of distances between individuals computed using the Sorensen–Dice index[221,222] implemented in the 'dist.binary' function of 'vegan'. The multivariate predictors were geography (i.e., spatial structure), environment (seven environmental predictors), and population stratification (epigenetic clustering and caste identity). To capture spatial structure, we transformed Euclidean geographic distances to a continuous rectangular vector by Principal Coordinates analysis of Neighbor Matrices (PCNM) using the 'pcnm' function in 'vegan'.

First, we performed principal coordinates analysis (i.e., multidimensional scaling, MDS), to visualize separation of epigenetic clusters in a two-dimensional MDS space. Then, we performed dbRDA (i.e., constrained/canonical analysis of principal coordinates, CAP), with multivariate predictors as constraints (fixed effects) or conditions (random effects). Then, to estimate the contributions of these multivariate predictors to epigenetic variation, we used the 'varpart' function in 'vegan'.

Finally, to determine significance of spatial, environmental, as well as population and caste predictors, we used multivariate F-statistics with 9999 permutations. To determine the significance of spatial predictors, PCNM axes were included as constraints in the model. To determine the significance of population and caste stratification, we used the interaction of these factors–the number of categories was number of clusters multiplied by the number of castes–as a constraint. To determine the significance of environmental predictors, we ran separate models with environmental variables as constraints: a) without conditions, b) conditioned on geography, and c) conditioned on geography as well as population stratification. Only significant PCNM axes were used when accounting for geography.

*Univariate modeling*. If, based on dbRDA modeling, the environment contributed significantly to epigenetic variation, we then used univariate modeling to determine whether any population strata occurred in significantly different environments. To test that the difference in means of the environmental predictors for any two strata was not zero, we employed two-tailed t-tests. We used the non-parametric Games-Howell posthoc test[223], which does not assume equal sample sizes or variances. To perform the Games-Howell posthoc test, we used the 'posthocTGH' function in the R package 'userfriendlyscience'[224].

#### 3.2.3.3 Short-term influences on epigenetic variation

To determine whether the environment played a role in within-colony epigenetic variation, we compared means for all seven environmental variables among sites (i.e., colonies, since we only sampled one log per site) grouped by the number of clusters to which individuals were assigned. The expected maximum number of clusters per colony was four, since a maximum of four individuals were screened for epigenetic variation within a colony. To test whether the difference in means between groups was not equal to zero, we performed two-tailed t-tests. Since we did not expect equal variances among groups, we performed the non-parametric Games-Howell posthoc test.

*Latent factor mixed modeling*. To determine co-association of environmental variables with methylation state at each locus, we used latent factor mixed modeling (LFMM[225]), a form of mixed-effects modeling where the random effects are latent (unobserved) variables. We used latent factor mixed modeling as implemented in the 'lfmm' function in the 'LEA' package. We used LFMM to model the fixed effects of environmental predictors while controlling for random effects of population stratification and geography.

*Univariate modeling*. After identifying outlier loci using LFMM, we used mixed-effects logistic regression to test the effect of single environmental predictors (fixed effect) on methylation state at each locus, while controlling for population stratification (random effect). Mixed-effects logistic regression was performed with the 'glmer' function in the R package 'lme4'[226].

## 3.3 Results

### 3.3.1 Epigenetic data

The MS-AFLP analysis resulted in 169 polymorphic loci, which were named based on the selective bases at the EcoRI restriction site (AT or AG; Figure C.1, Table C.2) and fragment size. For instance, a fragment of size 104 bp amplified with the E_AT primer (E = EcoRI restriction site; AT = selective bases) is named AT104. After creating 3 variables per locus (3 methylation states x 169 loci), the final dataset contained 470 polymorphic variables (CCGG = 139 loci; *m*CCGG = 165; and C*m*CGG = 166) across 167 individuals. The genotyping error rate based on 32 replicates was 3.6%.

### 3.3.2 Epigenetic clustering

Using NMF and DAPC, and the associated cluster validation techniques (cross-entropy and BIC, respectively), we identified four epigenetic clusters. The lowest cross-entropy value was 0.301 at K = 4 (Figure 3.3). Additionally, the lowest mean BIC was recorded for K = 4 (BIC = 265.17). Therefore, the final choice of K was 4. Herein, we refer to the four clusters as clusters 1 through 4 (Figure 3.4). It should be noted that the biggest decreases in BIC were from K = 1 to 3 (Figure 3.3). However, cluster 4 was a valid cluster. The small decrease from K = 3 to 4 was due to cluster 2 being relatively less differentiated from the other three clusters (Figure 3.5).

Out of 167 individuals, 159 were assigned to each of the four clusters with probability > 0.6. Of the 159 individuals, 14 were assigned to cluster 1, 43 to cluster 2, 58 to cluster 3, and 44 to cluster 4 (Table C.3). Of the eight unassigned individuals (Table C.3), five were assigned to cluster 2 with probabilities of 0.52-0.58, two were assigned with probability 0.56 to cluster 3 and 4, respectively, and the final individual was assigned with probability 0.50 to cluster 3 and 0.46 to cluster 4.

All four clusters were represented in each of the four major ecoregions where we sampled *R. flavipes* (Table 3.2a). Out of all individuals sampled in the Appalachian Plateaus, 45% were assigned to cluster 3 (Table 3.2b). The Valley and Ridge had 40% of individuals assigned to Cluster 4. A large proportion (64%) of cluster 1 was found in the Blue Ridge (Table 3.2c). While cluster 2 was spread across all four ecoregions, large proportions of clusters 3 and 4–57% and 48% respectively–were found in the Appalachian Plateaus (Table 3.2c).

### 3.3.3 Long-term influences on epigenetic variation

Distance-based redundancy analysis (Figure 3.5) showed that only tree cover did not contribute significantly (Table C.4) to epigenetic variation in *R. flavipes*. After accounting for geography by partialling out nine significant spatial components (PCNM axes 1 through 3, 7, 14, 17 through 19, and 22), summer temperature and wet-season precipitation, and pine and oak species richness remained

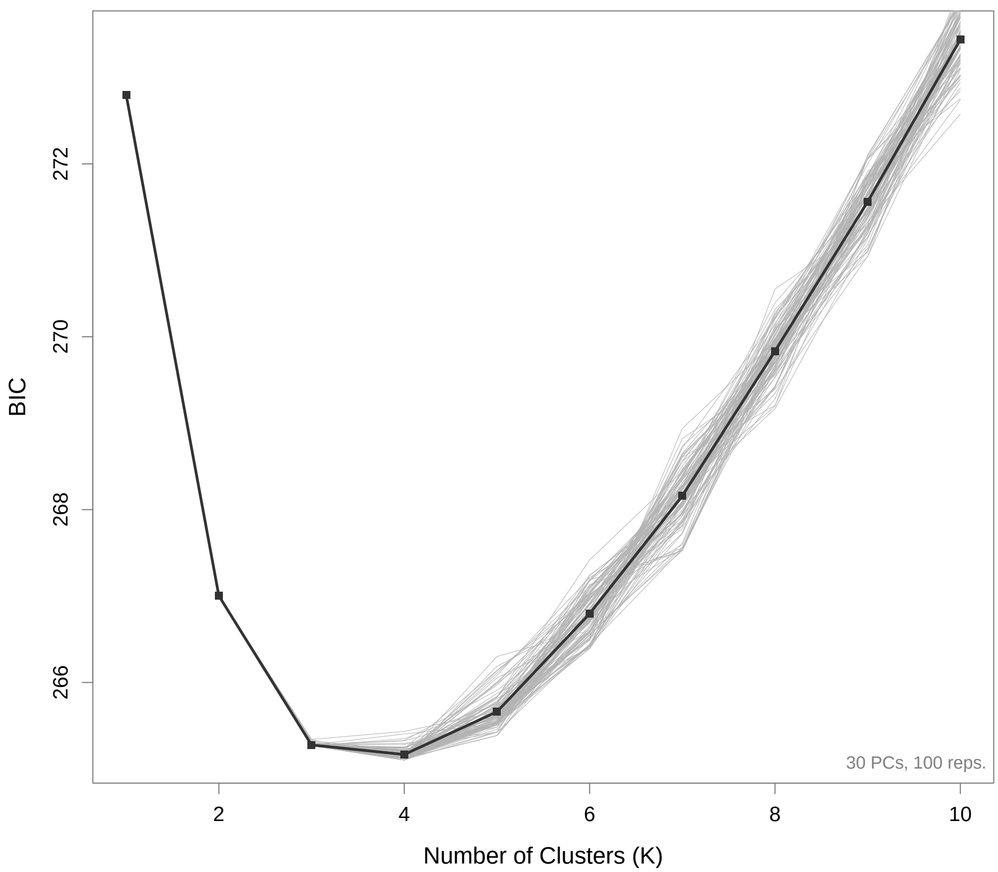



**Figure 3.3:** *Validation of* k*-means clustering.* Using 30 principal components (PCs) to represent the original binary MS-AFLP data, the Bayesian information criterion (BIC) was calculated for 100 replicates of *k*-means clustering, for K = 1 to 15 (11 through 15 cut off intentionally, for plotting purposes; BIC continues to increase). BIC was lowest at K = 4 (mean BIC = 265.17).

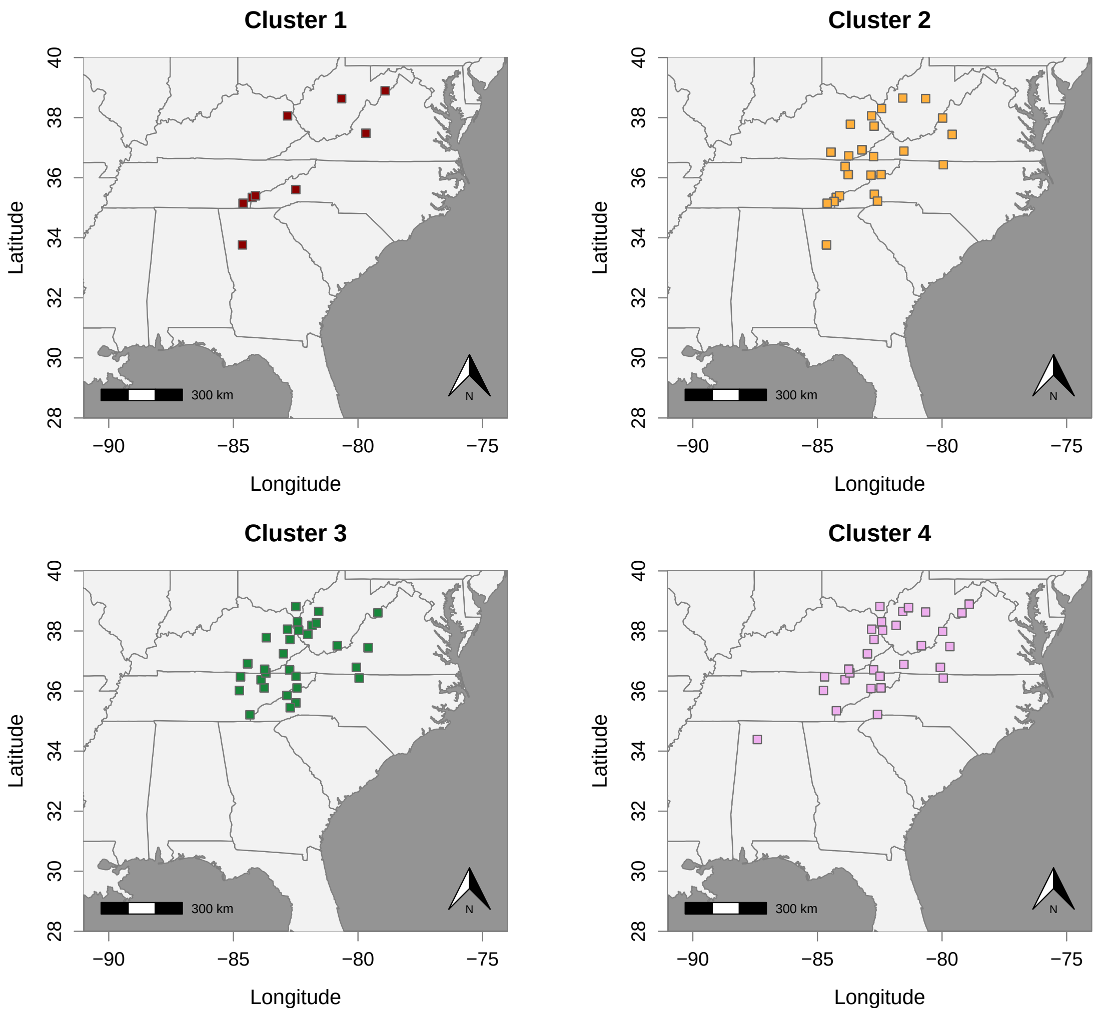



**Figure 3.4:** *Map of geographic sampling of* R. flavipes *with epigenetic cluster assignment of individuals.* The four panels show individuals, with the different colors representing the epigenetic cluster to which each individual was assigned. Only individuals (159 out of 167) with probability > 0.6 of belonging to a cluster are shown.

**Table 3.2:** *Distribution of epigenetic clusters of* R. flavipes *across southern Appalachian ecoregions.* **a**. Number of individuals with membership in each of the four clusters is shown for each ecoregion (see Figure 3.1). **b**. For each ecoregion, the proportion of individuals assigned to each cluster was calculated. **c**. Also shown is the proportion of individuals sampled in each ecoregion with membership in a given cluster. As a visual aid, low values are presented on a white background and high values on red.

| **a** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | |
|---|---|---|---|---|---|
| **Appalachian Plateaus** | 3 | 16 | 33 | 21 | 73 |
| **Blue Ridge** | 9 | 13 | 11 | 6 | 39 |
| **Piedmont** | 1 | 4 | 4 | 3 | 12 |
| **Valley and Ridge** | 1 | 10 | 10 | 14 | 35 |
| | 14 | 43 | 58 | 44 | Total Inds. |

| **b** | **Proportion of Ecoregion in Cluster** | | | |
|---|---|---|---|---|
| **Appalachian Plateaus** | 0.04 | 0.22 | 0.45 | 0.29 |
| **Blue Ridge** | 0.23 | 0.33 | 0.28 | 0.15 |
| **Piedmont** | 0.08 | 0.33 | 0.33 | 0.25 |
| **Valley and Ridge** | 0.03 | 0.29 | 0.29 | 0.40 |

| **c** | **Proportion of Cluster in Ecoregion** | | | |
|---|---|---|---|---|
| **Appalachian Plateaus** | 0.21 | 0.37 | 0.57 | 0.48 |
| **Blue Ridge** | 0.64 | 0.30 | 0.19 | 0.14 |
| **Piedmont** | 0.07 | 0.09 | 0.07 | 0.07 |
| **Valley and Ridge** | 0.07 | 0.23 | 0.17 | 0.32 |

significant (Figure 3.5; Table C.4). However, when population stratification was controlled for in addition to spatial structure, summer temperature and wet-season precipitation, available water capacity and pine species richness were significant, while the $p$-value for oak species richness was 0.060 (Figure C.2; Table C.4).

Based on these results, climatic conditions and tree species richness exerted a long-term influence on epigenetic variation, in addition to geography and population stratification. However, using the Games-Howell posthoc test, we found that the only near-significant differences occur between workers in cluster 2 compared to workers in clusters 3 ($p = 0.098$) and 4 ($p = 0.069$) with respect to wet-season precipitation. Specifically, cluster 2 occurred in areas with higher wet-season precipitation (Figure C.3). These results do not contradict dbRDA, which showed that in addition to wet-season precipitation, pine species richness also was correlated with axis 1 (i.e., "CAP1"; Figure 3.5). This is the axis that separated clusters 1 and 2 from clusters 3 and 4, with the former two being associated with higher wet-season precipitation and pine species richness. However, after accounting for geography, pine species richness was correlated with CAP2, while wet-season precipitation remained correlated with CAP1 (Figure C.2). In addition, after accounting for geography, clusters 3 and 4 were associated with higher summer temperatures (Figure C.2).

Geography, environment, and population stratification combined accounted for 17.2% of the observed variation in epigenetic data. Population stratification was responsible for 8.2% of the epigenetic variation, geography explained 7.1%, and environmental factors accounted for 3.0%. In terms of population stratification, caste differences alone explained 0.8%, while cluster differences explained 5.8%, and caste and cluster combined 8.2% of the variation. After partialling out the geographic component of the epigenetic variation, caste and cluster differences combined explained 7.5% of the variation, whereas environment alone (after partialling out geography, caste, and cluster) explained 2.6% of the variation. Some of the unexplained epigenetic variation (82.8%) should be attributable to short-term environmental influences.

### 3.3.4 Short-term influences on epigenetic variation

Eight out of forty-five sites (i.e., colonies) had all their individuals assigned to the same cluster. Twenty-four colonies consisted of individuals assigned to two clusters, while twelve colonies comprised individuals assigned to three clusters. In one colony–sampled at site 35 in northeastern Kentucky–all four individuals were assigned to different clusters. Eleven of the twelve colonies containing individuals with membership in three clusters are located in the Appalachian Plateaus, specifically eastern Tennessee, western Kentucky, and central West Virginia. The only other three-cluster colony was sampled at site 5 in the Piedmont region in South Carolina (Tables C.1 and C.3). The other two colonies in the Piedmont–sampled at sites 4 and 22–were two-cluster colonies (Tables C.1 and C.3).

Means were significantly different between one-cluster vs. three-or-more-cluster colonies only for canopy cover, which was significantly greater ($p = 0.032$) at one-cluster colonies (Figures 3.6 and 3.7). While the $p$-value was not significant ($p = 0.068$), one-cluster colonies were also associated with greater canopy cover than two-cluster colonies.

Mean distance from urban areas for one-cluster colonies was 23.25 km, while two-cluster and three-or-more-cluster colonies were 15.64 km and 16.08 km away. Although one-cluster colonies were farther from urban areas than the other two colony types, comparisons of means did not result in any significant differences (Figure C.4).

Oak and pine species richness were lowest at sampling sites in the Appalachian

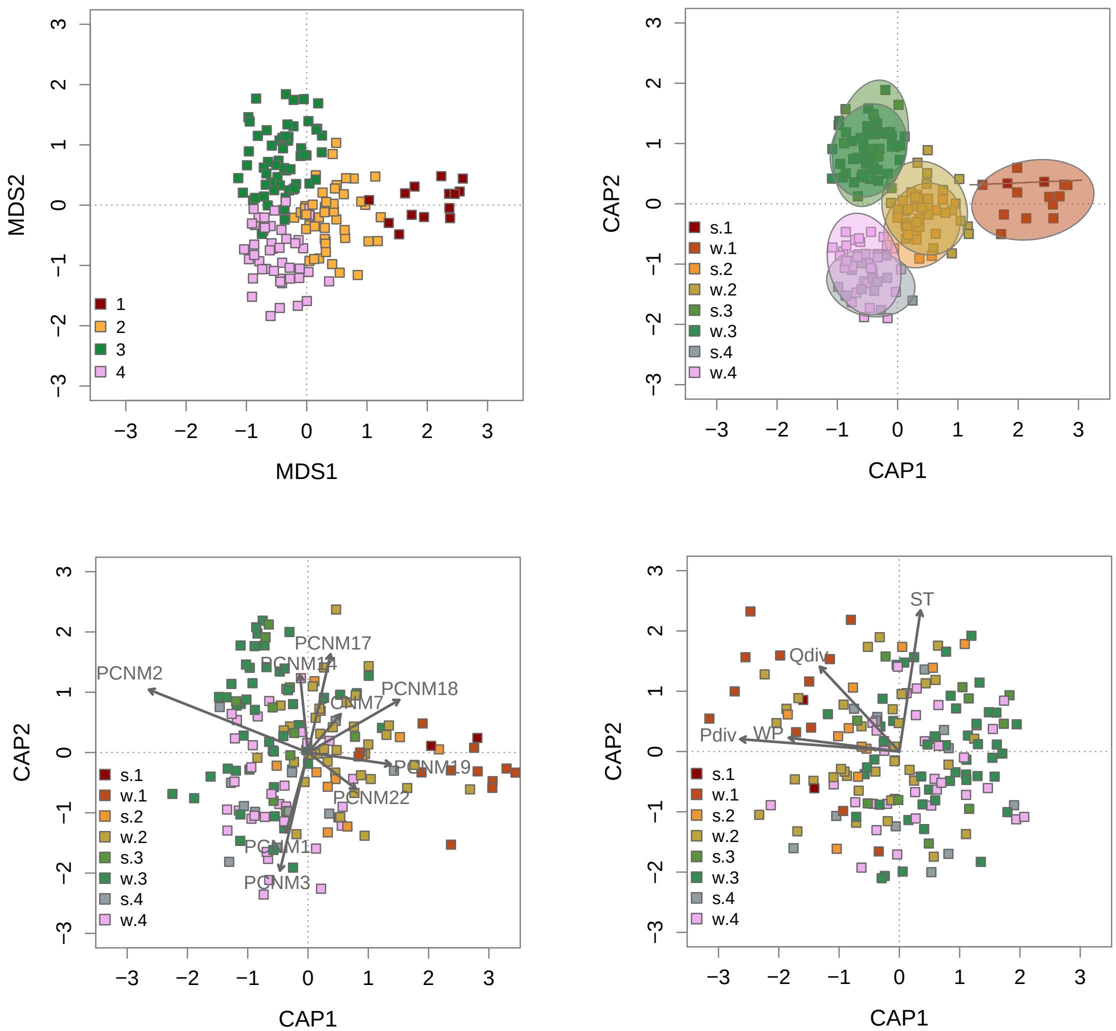



**Figure 3.5:** *Distance-based Redundancy Analysis (dbRDA).* Four *R. flavipes* epigenetic clusters are labeled 1 through 4 and two castes are labeled 's' (soldier) and 'w' (worker). The top left panel shows a plot of unconstrained dbRDA (i.e, multidimensional scaling, MDS), with each individual (square) color coded by cluster membership. The top right panel shows constrained dbRDA (i.e., constrained analysis of principal coordinates, CAP) with epigenetic clustering and caste identity (a factor with 8 categories: 2 castes x 4 clusters) as a predictor. The bottom left panel shows geography-constrained dbRDA, where variance in the epigenetic data is explained by geography (i.e., eigenvectors obtained via principal coordinates analysis of neighbor matrices, PCNM). Only significant PCNMs are shown. The bottom right panel shows environment-constrained dbRDA, where variance in epigenetic data is explained by environmental variables (only significant variables shown).

Plateaus (mean scaled value: oak = 0.93, pine = 0.33), the Piedmont (oak = 1.02, pine = 0.37), and the Valley and Ridge (oak = 0.88, pine = 1.20), and highest in the Blue Ridge (oak = 2.00, pine = 1.64).

Using LFMM, we detected twenty-one loci significantly correlated with environmental variables, after controlling for population stratification and spatial structure. To determine the effect on methylation state of each environmental predictor separately, we used mixed-effects logistic regression. Methylation state was significantly correlated with tree cover at nine loci (Figures C.5–C.7): four each positively and negatively correlated, with the ninth (locus AG113) being positively correlated for *m*CCGG and negatively correlated for C*m*CGG methylation (Table 3.3). Five loci (out of which four positively correlated) were significantly correlated with oak species richness, and five were significantly negatively correlated with pine species richness (Table 3.3). Methylation state was influenced by summer temperature and wet-season precipitation at three and two loci, respectively (Table 3.3).

Locus AG113 was significantly correlated with caste, pine and oak species richness, as well as canopy cover, while AG174 was correlated with dry-season precipitation, oak species richness and canopy cover (Table 3.3). These variables were significantly correlated with different AG113 methylation states: 1) probability of *m*CCGG was higher for soldiers than workers ($z$-score = worker - soldier), 2) pine species richness was negatively correlated with CCGG, 3) oak species richness was positively correlated with *m*CCGG, and 4) canopy cover was negatively correlated with C*m*CGG but positively correlated with *m*CCGG (Table 3.3). Dry-season precipitation, oak species richness, and canopy cover were all positively correlated with the same methylation state (C*m*CGG) at locus AG174 (Table 3.3).

**Table 3.3:** *Mixed-effects logistic regression results*. To account for structure in the data, when the fixed effect was one of the seven environmental variables, the random effects were caste (soldier/worker) and epigenetic clustering (four clusters). When the fixed effect was caste, the random effect was epigenetic clustering. Methylation state at each locus was the binary response variable. Only the loci/methylation states with significant fixed effects (evaluated separately) are shown. Positive associations ($z$-scores) are highlighted in green, whereas negative $z$-scores are shown in red. DP = dry-season precipitation; ST = summer temperature; WP = wet-season precipitation; AWC30cm = available water capacity at a soil depth of 30 cm; Pdiv = pine (*Pinus*) species richness; Qdiv = oak (*Quercus*) species richness; Tree = tree (canopy) cover.

| Fixed Effect | Locus | Methylation | Estimate | Std. Error | $z$-score | $p$-value |
|---|---|---|---|---|---|---|
| **Caste** | AG113 | *m*CCGG | -0.896 | 0.453 | -1.976 | 0.048 |

| Fixed Effect | Locus | Methylation | Estimate | Std. Error | *z*-score | *p*-value |
|---|---|---|---|---|---|---|
| **DP** | AT112 | CCGG | 0.506 | 0.212 | 2.389 | 0.017 |
| | AG174 | C*m*CGG | 0.715 | 0.264 | 2.705 | 0.007 |
| **WP** | AT188 | C*m*CGG | 0.908 | 0.422 | 2.155 | 0.031 |
| | AT206 | *m*CCGG | -1.418 | 0.598 | -2.372 | 0.018 |
| **ST** | AG134 | *m*CCGG | 0.758 | 0.346 | 2.194 | 0.028 |
| | AG148 | CCGG | 0.790 | 0.308 | 2.563 | 0.010 |
| | AG212 | *m*CCGG | 1.095 | 0.564 | 1.942 | 0.052 |
| **Pdiv** | AG113 | CCGG | -0.797 | 0.329 | -2.421 | 0.015 |
| | AG134 | *m*CCGG | -0.473 | 0.245 | -1.928 | 0.054 |
| | AT104 | CCGG | -0.619 | 0.255 | -2.425 | 0.015 |
| | AT166 | C*m*CGG | -0.853 | 0.341 | -2.503 | 0.012 |
| | AT216 | CCGG | -1.301 | 0.562 | -2.316 | 0.021 |
| **Qdiv** | AG113 | *m*CCGG | 0.530 | 0.246 | 2.154 | 0.031 |
| | AG174 | C*m*CGG | 0.755 | 0.388 | 1.947 | 0.052 |
| | AT166 | C*m*CGG | -0.775 | 0.311 | -2.491 | 0.013 |
| | AT188 | C*m*CGG | 0.753 | 0.348 | 2.163 | 0.031 |
| | AT240 | C*m*CGG | 0.672 | 0.307 | 2.192 | 0.028 |
| **Tree** | AG113 | C*m*CGG | -0.642 | 0.333 | -1.930 | 0.054 |
| | AG113 | *m*CCGG | 0.904 | 0.423 | 2.136 | 0.033 |
| | AG145 | C*m*CGG | -0.902 | 0.405 | -2.228 | 0.026 |
| | AG174 | C*m*CGG | 2.465 | 1.114 | 2.214 | 0.027 |
| | AG262 | CCGG | -1.457 | 0.543 | -2.684 | 0.007 |
| | AT104 | CCGG | 0.765 | 0.307 | 2.495 | 0.013 |
| | AT118 | CCGG | -0.781 | 0.383 | -2.038 | 0.042 |
| | AT126 | *m*CCGG | 1.101 | 0.441 | 2.496 | 0.013 |
| | AT206 | *m*CCGG | -1.222 | 0.409 | -2.987 | 0.003 |
| | AT240 | C*m*CGG | 1.409 | 0.622 | 2.265 | 0.023 |

## 3.4 Discussion

In this study, we gained insights into how disturbance of forest ecosystems in the southern Appalachian Mountains, resulting in changes in tree canopy cover and tree species richness, could have affected genome-wide DNA methylation in the eastern subterranean termite, *R. flavipes*. To our knowledge, this is the first study to show an effect of canopy cover on intraspecific epigenetic variation in termites.

To understand long-term influences on DNA methylation changes in *R. flavipes*, we examined evidence of epigenetic clustering. We also assessed differences in DNA

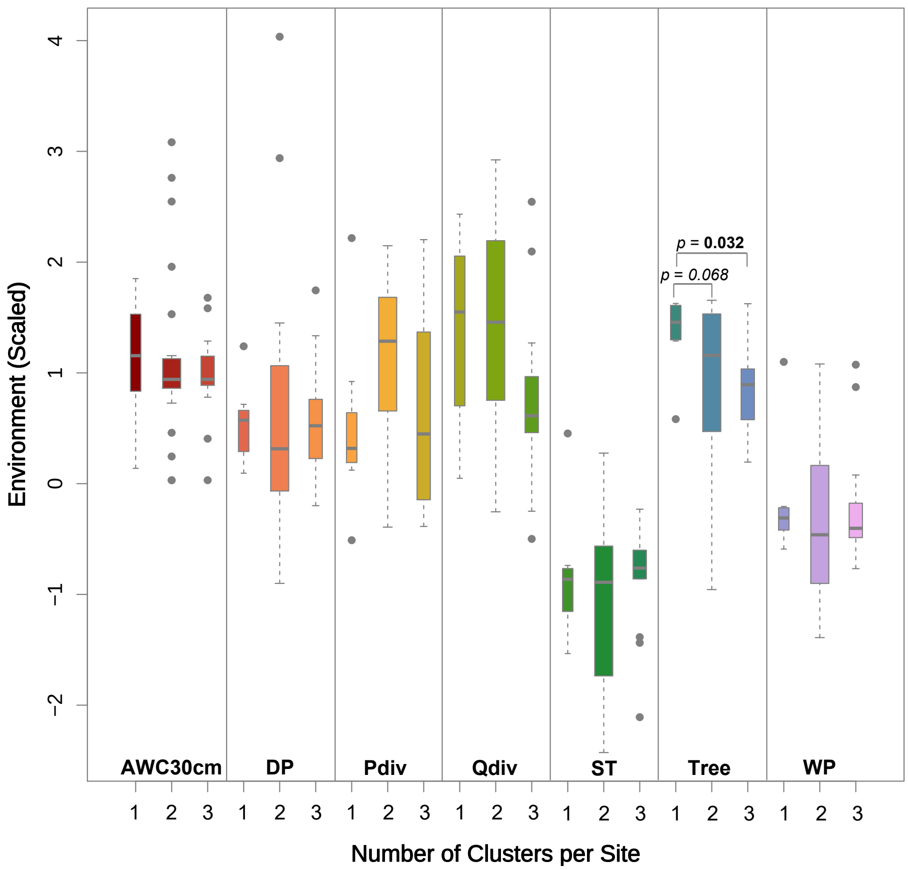



**Figure 3.6:** *Box plots of scaled values for seven environmental variables for different sampling site categories.* Sampling sites (i.e., rotting logs) were grouped based on the number of clusters that individuals were assigned to at each site: 1, 2, and 3. The last category, 3, includes, in addition to sites with three clusters, one site where all four individuals were assigned to a different cluster. The one significant *p*-value is shown, as well as the lowest non-significant *p*-value. AWC30cm = available water capacity at a soil depth of 30 cm; DP = dry-season precipitation; Pdiv = pine (*Pinus*) species richness; Qdiv = oak (*Quercus*) species richness; ST = summer temperature; Tree = tree (canopy) cover; WP = wet-season precipitation.

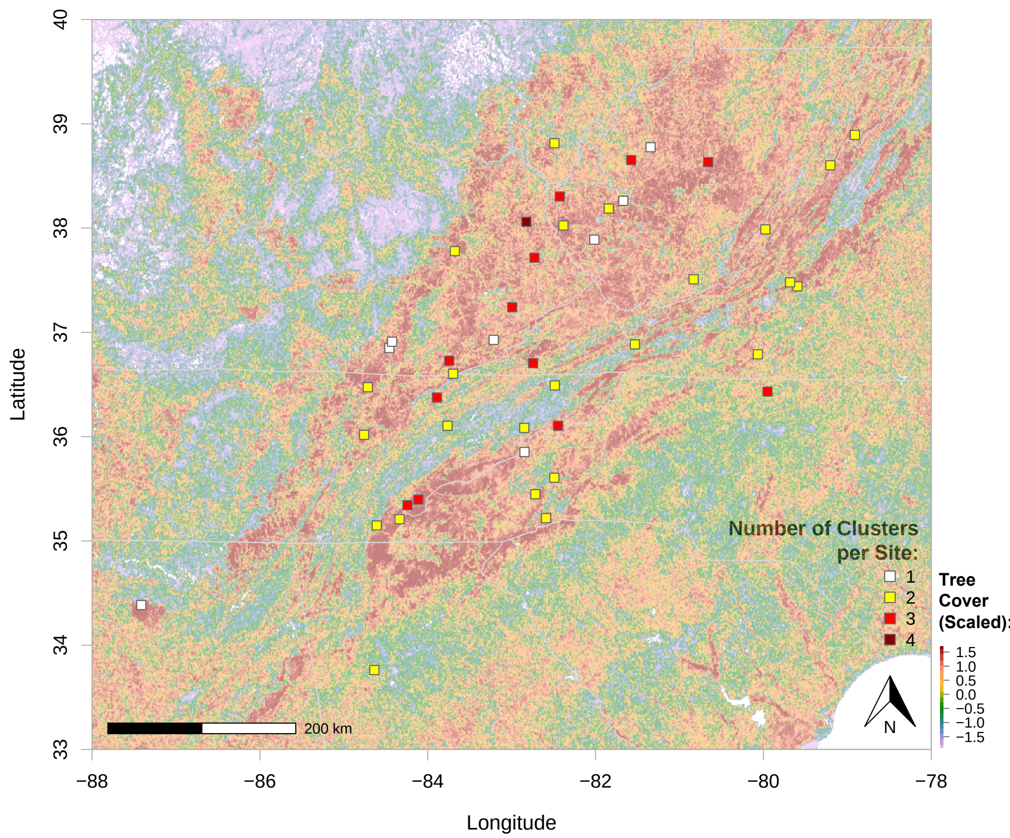



**Figure 3.7:** *Number of* R. flavipes *clusters detected at each sampling site superimposed on tree cover.* Numbers–1 through 4–represent the number of clusters that individuals at each sampling site were assigned to. Tree cover is shown as scaled values from high tree cover (dark red = 100% tree cover) to low tree cover (pink = 0% tree cover).

methylation between the soldier and worker castes, as these members of a colony may be differentially impacted by environmental conditions. In addition to long-term influences, we also assessed evidence of short-term environmental influences. Overall, we detected four epigenetic clusters, which overlapped geographically (but see below). Beyond the four epigenetic clusters, environmental factors were inferred to have exerted a long-term influence leading to stable methylation differences (see Figure C.3). In addition, short-term environmental effects were inferred, resulting in epigenetically mixed colonies (see Figures 3.6 and 3.7) and methylation differences at individual loci (see Table 3.3). Importantly, we found that tree canopy cover and tree species richness had significant impacts on DNA methylation.

### 3.4.1 The geography of *R. flavipes* epigenetic variation

While the four epigenetic clusters that were detected have overlapping ranges, there were some notable differences in the geography of these clusters. For instance, (64%) of cluster 1 was found in the Blue Ridge, whereas cluster 2 was distributed across all four ecoregions (from 9% in the Piedmont to 37% in the Plateaus), while large proportions of clusters 3 and 4 (57% and 48%, respectively) were found in the Appalachian Plateaus. The Appalachian Plateaus were also home to 11 of the 12 colonies that consisted of individuals assigned to three or more clusters. These Appalachian Plateaus sampling sites had low canopy cover, as well as low oak and pine species richness. Some evidence that low canopy cover and tree species richness in this region is a result of human-mediated disturbance, rather than a feature of the terrain (e.g., low canopy cover on rocky outcrops), comes from areas adjacent to our sampling sites. In southeastern Ohio, agricultural clearing, followed by abandonment and forest regeneration, has favored the fast-growing tulip poplar (*Liriodendron tulipifera*), or red maple (*Acer rubrum*)[227], thus affecting oak species richness. Similarly, in Pennsylvania, Fei and Steiner[228] found that red maple outcompetes oak following disturbance.

### 3.4.2 Long- and short-term influences on epigenetic variation

Climatic variables are expected to have exerted a long-term influence on epigenetic variation. In the present study, the focal climatic variables summarized a period from 1960-1990 and captured what has remained a relatively stable climate in the southern Appalachian Mountains, despite human-induced climate change[229]. Hyseni and Garrick[197] found that wet-season precipitation contributed significantly to genetic variation in *R. flavipes*. Similarly, in this study, we found that wet-season precipitation, as well as summer temperature, contributed to epigenetic variation. Indeed, clusters 1 and 2 were associated with higher wet-season precipitation, while clusters 3 and 4 were associated with higher summer temperature.

Within-cluster variation was predominantly driven by canopy cover and oak and pine species richness, operating at both long and short timescales. C*m*CGG methylation at locus AG174 was potentially trans-generationally inherited, reflecting long-term influences of dry-season precipitation, oak species richness, and canopy cover, which were all positively correlated with C*m*CGG methylation at locus AG174. Herrera and Bazaga[230] recovered isolation-by-distance patterns in lavender, *Lavandula latifolia*, that were similar between genetic data and C*m*CCGG methylation, suggesting similar trans-generational inheritance. Locus AG113, on the other hand, likely reflects somatic changes in methylation influenced by environmental factors. For instance, *m*CCGG at this locus was positively correlated with canopy cover and oak species richness, while C*m*CGG was negatively correlated with canopy cover (Table 3.3). Herrera and Bazaga[230] suggested that variation in *m*CCGG methylation in *L. latifolia*–which did not follow the same pattern as genetic data and C*m*CGG methylation–was likely due to somatic instability caused by factors such as water and light availability.

Based on our MS-AFLP data, only one locus showed significant differences in DNA methylation between soldiers and workers within epigenetic clusters. *mCCGG* methylation was significantly higher in soldiers than workers at the AG113 locus. DNA methylation at this locus potentially reflects somatic instability induced by contemporary environmental stressors. In *Reticulitermes* termites, soldiers are a terminal caste that can live at least five years after differentiating from workers[231]. A recent study of age polyethism (division of labor) in termite soldiers found that old soldiers were recruited to the front line of defense significantly more frequently than young soldiers[232]. Since we sampled termites immediately upon detection within a log, it is likely that the soldiers we sampled were older soldiers that attempted to protect the colony. Given the possibility that the sampled soldiers are older than the workers, they would have been exposed to environmental stressors (e.g., low canopy cover) for a longer period.

### 3.4.3 Trees and termites: Canopy cover influences on DNA methylation

Reductions in tree canopy cover and tree species richness–likely due to human-mediated disturbances of forest ecosystems in the eastern U.S.–appear to have influenced variation in DNA methylation in *R. flavipes*. Indeed, environmental stressors have been shown to influence epigenetic variation in other systems. For instance, DNA methylation changes have occurred in violets[233], marsh perennials[234], and lavender[230] as a consequence of herbivory, salinity, and artificial disturbance, respectively.

Here, we found that epigenetic variation was higher in areas with lower canopy cover compared to epigenetic variation in environments with high tree canopy cover and tree species richness. Specifically, we found epigenetically mixed colonies (i.e., those containing individuals with membership in two or more clusters) occurring under conditions of lower canopy cover compared to colonies in which all sampled individuals were assigned to the same epigenetic cluster. Additionally, even after accounting for population stratification, DNA methylation differences were significantly associated with differences in canopy cover at nine loci.

Previous studies have shown an effect of canopy cover on termite species richness. For instance, along a land-use intensification gradient in central Sumatra (Indonesia), termite species richness and relative abundance were highly correlated with reduction in canopy cover[235]. Also, based on data from two primary forest national parks in Ecuador, canopy cover was a significant driver of termite diversity, specifically wood- and wood-and-litter-feeding termites[236]. If canopy cover is low, and this negatively affects species richness, the persisting species may be released from competition. Such a release from competition may lead to niche expansion[237]. Epigenetic variation may also play a part in niche breadth evolution (reviewed in[238]).

### 3.4.4 Native invaders and plasticity

The occurrence, persistence, and spread of invasive species is facilitated by climate change, environmental disturbance and degradation, along with increasing connectedness mediated by global trade and travel. Human-induced disturbance of habitats may release certain species from previous ecological constraints (e.g., enemies and competitors[239]), leading to 'invasive' characteristics (e.g., high densities or reproductive rates). These characteristics can also be found within native ranges[202] when species expand into human-altered habitats.

Hufbauer et al.[240] proposed that expansion into and adaptation to human-altered habitats in a species' native range may facilitate the establishment and spread of that species in similar human-altered habitats elsewhere. While we commonly think of invasive species as species that become established and spread in new areas outside their native range, there are numerous examples of species that become 'invasive' (i.e., dominant) in their native range[202–205]. These species are aptly named "native invaders"[241]. R. flavipes may be an example of a native invader. Habitat disturbances (i.e., harvesting practices and fire exclusion leading to re-structuring of forests) in the native range of *R. flavipes* may have contributed to the ability of *R. flavipes* to invade environments in other parts of the world that are similarly altered by humans (e.g., in France[23,25,28,29])

Novel environments in non-native ranges, or stressful environments in native ranges are known to induce epigenetic changes, which could put a premium on phenotypic plasticity, and this may ultimately facilitate subsequent invasion success[189–192]. Interestingly, extensive methylation has been detected in *R. flavipes*. When comparing methylation levels across most insect orders, Bewick et al.[242] found that DNA methylation was highest in *Blattodea* (cockroaches and termites), while within *Reticulitermes*, they found that *R. flavipes* had much higher levels of DNA methylation than *R. virginicus*. Across the entire genome, methylation was at 5.7% in *R. flavipes* vs. 0.1% in *R. virginicus*. For coding regions, these percentages were 18.1% and 0.7% in *R. flavipes* and *R. virginicus*, respectively.

### 3.4.5 Conclusions and caveats

While wet-season precipitation and summer temperature exerted a long-term influence on epigenetic variation, canopy cover as well as oak and pine species richness may have induced both trans-generationally inherited and within-generation somatic methylation changes. However, given that the present study was based on observation in natural populations, rather than a multi-generational experiment, we cannot demonstrate trans-generational inheritance, which warrants follow-up experimental studies to verify that human-mediated disturbance and low canopy cover can induce heritable epigenetic changes. Nonetheless, this study provided important insights, including an increase of epigenetic variation in *R. flavipes* resulting from reduced tree canopy cover and tree species richness, which likely reflect contemporary human-mediated disturbance of forest ecosystems. However, the local terrain (e.g., rocky outcrops or ridgelines), and not necessarily human-induced disturbance, may have been the cause of the low canopy cover observed at some sampling sites.

Our finding that epigenetically mixed colonies were associated with lower canopy cover, leaves open the question of how this mixing takes place. It could be the result of the geographic overlap of the four detected epigenetic clusters, resulting in epigenetically mixed colonies when a king and queen from different epigenetic clusters start a new colony. However, we detected colonies with membership in more than two clusters. This could be a consequence of the reproductive plasticity found in several *Reticulitermes* species, which use asexual queen succession (AQS) to produce the next generation of queens, while other colony members are produced through sexual reproduction with the king. AQS was first reported by Matsuura et al.[243] in the Asian *R. speratus* and has since been identified in the North American *R. virginicus*[244] and the European *R. lucifugus*[245]. AQS in *Reticulitermes* occurs through automictic parthenogenesis, which involves meiosis, specifically terminal fusion (i.e., fusion of anaphase II products)[246]. Thus, these parthenogens are homozygous for a single maternal allele at almost all loci. Production of new queens that are homozygous for different loci may lead to sexually produced workers with membership in different epigenetic clusters. Thus, future studies should look at the impact AQS may have on the epigenetic composition of termite colonies.

Data Accessibility: The Supplementary Material and additional data, as well as R scripts, are available online at `https://github.com/chazhyseni/msaflp`.

CHAPTER 4:

# A NOVEL METRIC THAT CAPTURES FUNCTIONAL LANDSCAPE CONNECTIVITY AT MULTIPLE SCALES, FROM ALLELES TO COMMUNITIES

Citation: Hyseni C, Symula RE, Garrick RC, Caccone A. A novel metric that captures functional landscape connectivity at multiple scales, from alleles to communities. **In Prep.** *2020*.

Abstract: We introduce a new metric, ($MS_{Conn}$), and discuss its properties. The metric measures functional landscape connectivity at multiple scales (i.e., among individuals, local populations, or species), based on genetic data. We used simulations to evaluate the sensitivity of the metric to environmental heterogeneity, selection, and migration, To evaluate the metric at the individual and population scales, we simulated several landscapes, and then used these as the substrate on which to simulate individuals/genotypes. At these scales, $MS_{Conn}$ is applicable to the field of landscape genetics. Briefly, we simulated four different landscapes (grids of 10 x 10 cells), with the values for each cell ranging from 0 to 100, representing carrying capacity (i.e., the maximum number of individuals a cell can support). By design, the four landscapes captured different scales of environmental heterogeneity: 1) fine-scale heterogeneity (*Gaussian*), 2) medium-scale (*gradient*), and two coarse-grain landscapes: 3) clustered (*random*), and *uniform*. Using these landscapes, we ran individual-based forward-in-time spatially explicit simulations of diploid genotypes at neutral and selected (s = 0.1) loci. We also evaluated three parameter values for migration ('low', 'medium', and 'high'), thus simulating a total of 24 datasets (4 landscapes x 2 types of loci x 3 strengths of migration). $MS_{Conn}$ is sensitive to migration rate as well as selection. When migration rate was low, overall differences

in connectivity grids between neutral versus selected loci were higher for landscapes with fine-scale environmental heterogeneity. Differences between neutral and selected loci were not as pronounced when the scale of environmental heterogeneity was high. To evaluate the metric at the species scale, we used three of the above landscapes (excluding the *uniform* landscape), each representing a distinct species. We then assessed how the connectivity metric at this scale captures spatially heterogeneous community composition. Cells with high connectivity values were the ones that had close to equal numbers of all three species (i.e., high evenness). Thus, $MS_{Conn}$ is applicable to both the field of landscape genetics and community ecology.

## 4.1 Introduction

### 4.1.1 Functional connectivity

Taylor et al.[247] defined landscape connectivity as the effect that the landscape has on "movement along resource patches". The spatial arrangement of favorable versus unfavorable environments determines how organisms move across the landscape. This is referred to as functional connectivity (cf. structural connectivity, which represents spatial autocorrelation of environmental features, regardless of how individuals/populations/species may interact with these features)[248]. Functional connectivity has a spatial (i.e., arrangement of environmental features in space) and a temporal component (i.e., persistence of organisms through time)[249]. Both the spatial and temporal components of functional connectivity can be captured by assessing gene flow between organisms residing in different resource patches. Indeed, this is the purview of the field of landscape genetics[250].

### 4.1.2 Landscape genetics

Landscape genetics integrates methods from landscape ecology, spatial (multivariate) statistics and population genetics[250–252], with the goal of understanding how environmental features influence the movement of organisms across the landscape. Wright introduced the concept of isolation by distance to describe genetic differentiation as a function of geographic distance[253]. Similarly, landscape genetics involves quantifying the effect of landscape resistance (inverse of connectivity) on genetic differentiation, which has been termed 'isolation by resistance'[254,255]. However, this requires knowledge of the magnitude of resistance that each environmental feature represents to dispersal (and, indirectly, gene flow, assuming that dispersal is accompanied by successful reproduction). Such an approach has limitations[256] given that resistance values are rarely known, and therefore often assigned subjectively. However, methods have been developed to objectively optimize the assignment of resistance values[257]. An alternative approach is to model genetic differentiation as a function of dissimilarities between environments in which populations occur[258,259], by using methods such as multiple matrix regression[260], or distance-based redundancy analysis[87].

Genetic differentiation is usually represented by pairwise distance metrics such as Wright's $F_{ST}$[261], Nei's genetic distance[262], or Cavalli-Sforza and Edwards' chord distance[263]. However, in some circumstances, these metrics may be unable to capture the complex dynamics of gene flow. For instance, when there is no gene flow between two sink populations, but they both receive migrants from the same source population, a metric such as $F_{ST}$ will lead to a false inference of gene flow between the two sink populations[264]. Alternatively, network theory methods have been used to infer genetic differentiation among populations (i.e., population graphs)[265] or individuals[266]. Unlike $F_{ST}$ and other pairwise methods, network methodology considers genetic relationships between all components simultaneously, thus improving performance under complex scenarios of gene flow. However, despite these advantages, network methods yield between-group measures, such as the population-graph-derived conditional genetic distance, cGD[267]. Thus, despite fast advances since the field was first conceived[250], landscape genetics lacks a metric that is an attribute of a single entity (such as an individual or a population) while at the same time reflecting gene flow (i.e., functional connectivity) among entities. Such a metric would make it possible to directly model the influence of environmental features on connectivity rather than computing pairwise dissimilarities[260] or resistance distances[255] and assessing their effect on pairwise genetic distances.

### 4.1.3 Multi-scale connectivity

Here, we introduce a new connectivity metric, which is an attribute of a single entity. Furthermore, this metric can capture connectivity at multiple scales, from individuals to populations to species, and even communities. In this paper, we describe and evaluate its properties using simulations, and discuss potential applications. In addition to landscape genetics, this metric can be applied in community ecology. The field of community ecology is concerned with distribution and abundance of species as well as interaction among species occupying the same geographic area. It is now recognized that there is more than one spatial scale at which species interact[268], such as a network of local communities (i.e., a meta-community) linked by dispersal[269], thus leading to a multi-scale approach to community ecology[270]. The multi-scale connectivity metric introduced here can provide a measure of connectivity between species that form such meta-communities.

## 4.2 Methods

### 4.2.1 Description of the multi-scale connectivity metric

To help describe the scalability of the metric, we draw on self-similarity, a concept borrowed from mathematics, referring to the property of fractals of having parts that are similar or identical to the whole (Figure 4.1 is drawn as a fractal). Briefly, self-similarity means that the same statistical properties are displayed at different scales. Here, we define self-similarity at a few (out of many possible) scales. At the scale of individuals, it is defined as the probability that an individual has two identical alleles (i.e., probability of homozygosity). At the scale of populations (within a larger group, such as meta-population or species), it is defined as the probability that a population contains two individuals from the same (larger) group. This can be further extended to the scale of meta-populations or species within communities or ecosystems (Figure 4.2).

The $MS_{Conn}$ connectivity metric, as defined here:

$$MS_{Conn} = \frac{1}{n} \frac{1}{\sum_{g=1}^{n} {P_g}^2} \tag{4.1}$$

represents the scaled (i.e., divided by $n$, the number of $g$ groups) inverse of self-similarity, where self-similarity, ${P_g}^2$, is defined as the probability that any two entities (e.g., individuals or populations) belong to the same group, $g$. When $g$ represents alleles, probability ${P_g}^2$ is the proportion of homozygous individuals in a population. When $g$ represents a meta-population, probability ${P_g}^2$ is the proportion of a population's individuals assigned to the meta-population.

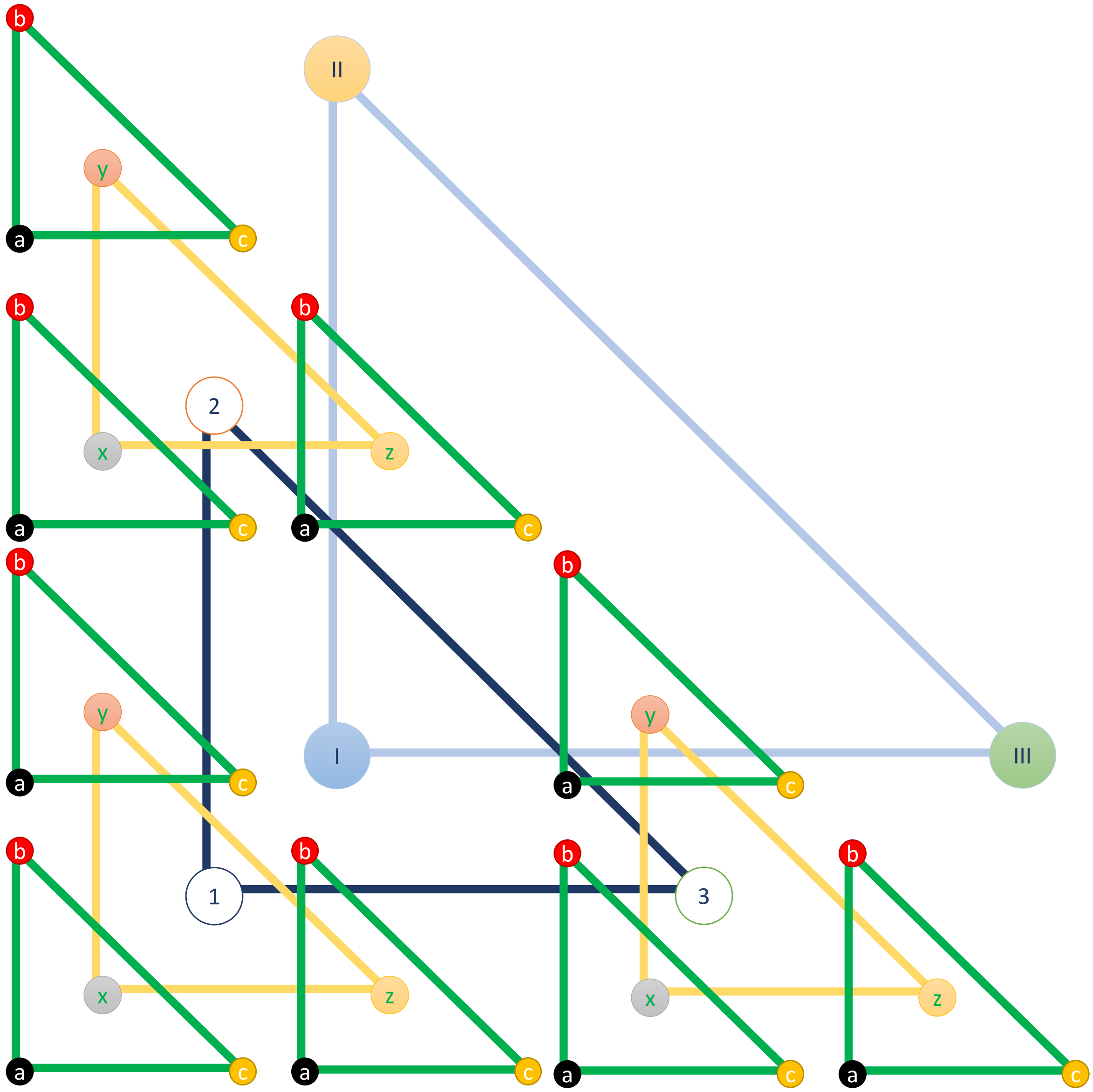



**Figure 4.1:** *Multi-scale connectivity*. Connectivity is represented at different levels: from connectivity of alleles (a, b, c) within individuals (x, y, z), to connectivity of individuals within populations (1, 2, 3), to connectivity of populations within larger groups (e.g., species or communities: I, II, III). The yellow lines represent connectivity between individuals (i.e., mating within a population), where the green lines represent two different alleles coming together in a heterozygous individual (cf. homozygotes at triangle vertices). The dark blue lines represent connectivity between populations (i.e., gene flow), and the light blue lines represent connectivity between species (i.e., species interactions within a community).

## Multi-scale Connectivity (Inverse Self-Similarity)

$$\frac{\boldsymbol{1}}{\boldsymbol{n}}\frac{\left(\sum_{g=1}^{n}\boldsymbol{P}_{g}\right)^{2}}{\sum_{g=1}^{n}\boldsymbol{P}_{g}{}^{2}}=\frac{\boldsymbol{1}}{\boldsymbol{n}\sum_{g=1}^{n}\boldsymbol{P}_{g}{}^{2}}$$

**Between-population Connectivity**

**populations categorized into groups (*g*):**

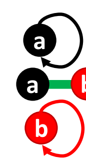
- **genetic clusters or meta-populations**
- **species**
- **communities**

| | Pop. 1 | Pop. 2 | Pop. 3 |
|---|---|---|---|
| Sp. I | 0.1 | - | - |
| Sp. II | 0.8 | - | - |
| Sp. III | 0.1 | - | - |

**Within-population Connectivity**

**(i.e., inter-individual connectivity)**

- **individuals represented as genotypes:**
  **$a^2$, ab, $b^2$**




| | Pop. 1 | Pop. 2 | Pop. 3 |
|---|---|---|---|
| | 0.25 | 0.10 | 0.79 |
| | 0.50 | 0.44 | 0.20 |
| | 0.25 | 0.46 | 0.01 |
| **Connectivity:** | **1** | **0.89** | **0.62** |

**Figure 4.2:** *Multi-scale connectivity equation.* The connectivity equation remains unchanged at all scales. The within-population (among individuals) connectivity example has three populations (Pop. 1, 2, and 3) composed of two alleles (a and b), which form three genotypes ($a^2$, ab, and $b^2$). Population connectivity within larger groups (i.e., between-population connectivity) is also shown. An example is given with three populations and three species (Sp. I, II, and III).

### 4.2.2 Simulations

We evaluated the performance of the metric at the scale of individuals, populations, as well as species. Here, we refer to the connectivity at these scales as within-population, between-population, and within-community connectivity. To evaluate the metric at the individual and population scales, we simulated several landscapes, and then used these as the substrate on which to simulate individuals (i.e., genotypes). We used the simulated individuals to test the connectivity metric at the within-population and between-population scales. In addition, to evaluate the metric at the within-community scale, we used each of the simulated landscapes to represent a distinct species, with the sum of the landscapes representing a meta-community and each cell a distinct community.

#### 4.2.2.1 Within- and between-population connectivity

We simulated four different landscapes (grids of 10 x 10 cells), with the values for each cell ranging from 0 to 100, representing carrying capacity (i.e., the maximum number of individuals a cell can support). The four landscapes captured different scales of environmental heterogeneity: 1) fine-scale heterogeneity (*Gaussian*), 2) medium-scale (*gradient*), and two coarse-grain landscapes: 3) clustered (*random*), and *uniform* (Figure 4.3). We used the R package 'NLMR'[271] to generate these landscapes. We used the R[210] package 'landsim'[272], an individual-based spatially-explicit forward-time simulation algorithm, to simulate genotypes on the four simulated landscapes (Figure 4.3). Since 'landsim' simulates genotypes at one locus with two alleles, we performed 1,000 (or 100; see below) replicates, thus resulting in a dataset of 1,000 bi-allelic markers (e.g., single nucleotide polymorphisms, SNPs).

We simulated both neutral-marker and selected-marker datasets. We simulated 1,000 neutral loci and 100 selected loci. The selected loci were simulated such that one allele has a selective advantage of $s = 0.1$. Additional simulations were performed at $s = 0.05$, but are not presented here, since $s = 0.1$ showcased differences to neutral loci better. We also evaluated three alternative parameter values for migration ('low', 'medium', and 'high'), thus simulating a total of 24 datasets (4 landscapes x 2 types of loci x 3 strengths of migration).

To simulate migration, we used a Gaussian kernel function (represented by $\sigma$ and the radius, $\rho$). We kept $\rho$ constant at 0.01 and varied $\sigma$ (0.2, 0.5, and 1). $\rho$ represents the maximum migration distance, whereas high values of $\sigma$ represent long-distance dispersal. We tested the effect of changing $\sigma$ from 0.2 to 2 (Figure 4.4). Since a value of 2 meant that all individuals could disperse to all cells in the grid, we decided to only use the first three values (Figure 4.4), $\sigma = 0.2, 0.5, or 1$, corresponding to 'low', 'medium', or 'high' migration. To compare the effect of different parameters on connectivity at different scales, we used root mean square error (RMSE) to quantify pairwise differences (across all cells on the grid) between connectivity outputs for different parameter values.

For within-population connectivity, ${P_a}^2$ (where a stands for allele) represents the expected proportion of homozygotes. Alternatively, since the output of 'landsim' simulations is number of each genotype per cell, we can divide this by the total number of individuals in each cell, and thus calculate observed proportions. Thus, the metric at this scale can be used to quantify deviation from the Hardy-Weinberg equilibrium in a spatially explicit manner. We did not explore this here, since we expected deviation from Hardy-Weinberg equilibrium, having included both migration and selection as simulation parameters. For between-population connectivity, we used (${P_K}^2$; where K stands for cluster). First, in order to infer the number of genetic clusters, we used non-negative matrix factorization (an unsupervised machine learning technique), as implemented in 'snmf' function of the R package 'LEA'[218]. We performed clustering, with 500,000 iterations, arbitrarily setting the value of K to 5, which was an overestimate in most cases. We did not attempt to determine the best value of K, since differences in K were not shown to affect between-population connectivity systematically (results not provided here). Additionally, a set value of K allowed us to compare between-population connectivity differences across selection and migration parameter values. We used averages for all individuals within a cell of probabilities of cluster membership.

#### 4.2.2.2 Within-community connectivity

To represent a meta-community, where each cell in the grid represents a community consisting of three species, we used the sum of the simulated landscapes (3, excluding the *uniform* landscape; Figure 4.3), with each landscape representing a species. To illustrate another way to compute the $MS_{Conn}$ metric, using raw numbers instead of proportions, we used the expanded form of the metric,

$$MS_{Conn} = \frac{1}{n}\frac{1}{\sum_{g=1}^{n} {P_g}^2} = \frac{1}{n}\frac{(\sum_{g=1}^{n} P_g)^2}{\sum_{g=1}^{n} {P_g}^2} = \frac{1}{n}\frac{(\sum_{g=1}^{n} N_g)^2}{\sum_{g=1}^{n} {N_g}^2} \tag{4.2}$$

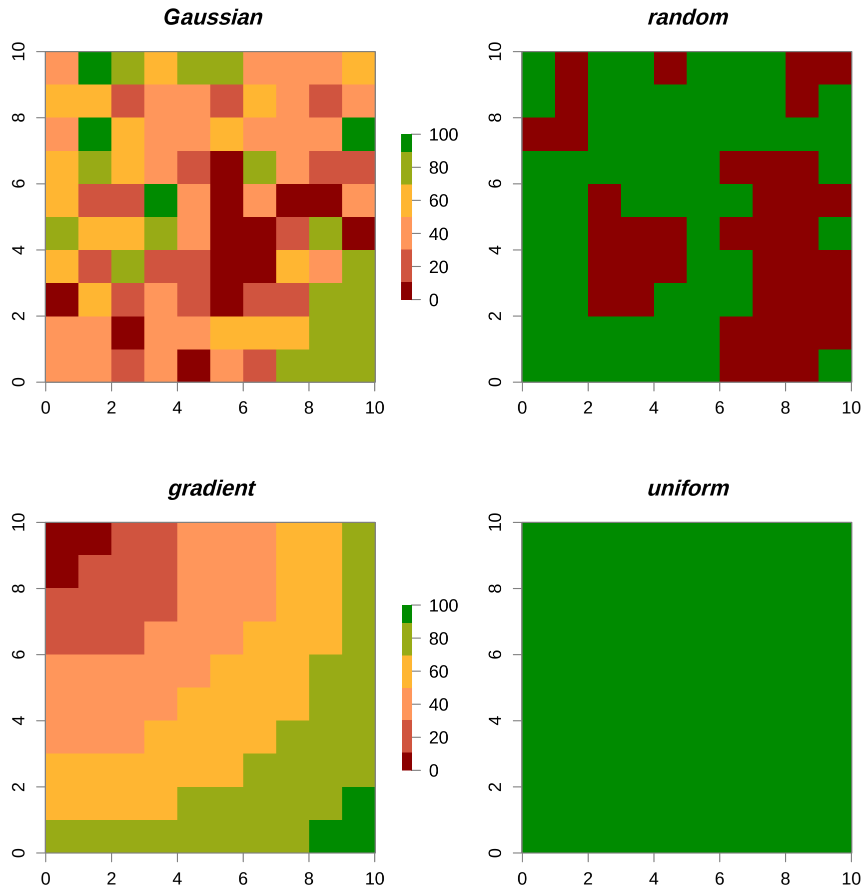



**Figure 4.3:** *Simulated landscapes used for genotype simulations.*The top left panel represents habitat with a Gaussian distribution carrying capacity (a maximum of 100 individuals). The top right panel represents random clusters of habitat (carrying capacity = 100 for all of them), plus unsuitable habitat outside the random clusters. The bottom left panel represents a distance gradient, where habitat unsuitability is increased with distance from suitable habitat. The bottom right panel represents a landscape of uniformly suitable (maximum carrying capacity) habitat.

where $N$ is the number of individuals belonging to species $g$ in a given cell (i.e., community).

## 4.3 Results

### 4.3.1 Properties of the connectivity metric

When scaled, the highest value of the $MS_{Conn}$ metric is 1, and the lowest is $1/n$. For instance, in the within-population scenario, maximum connectivity is achieved when alleles occur at equal frequencies, whereas minimum connectivity occurs when an allele becomes fixed (Figure D.1). At other scales, maximum con-

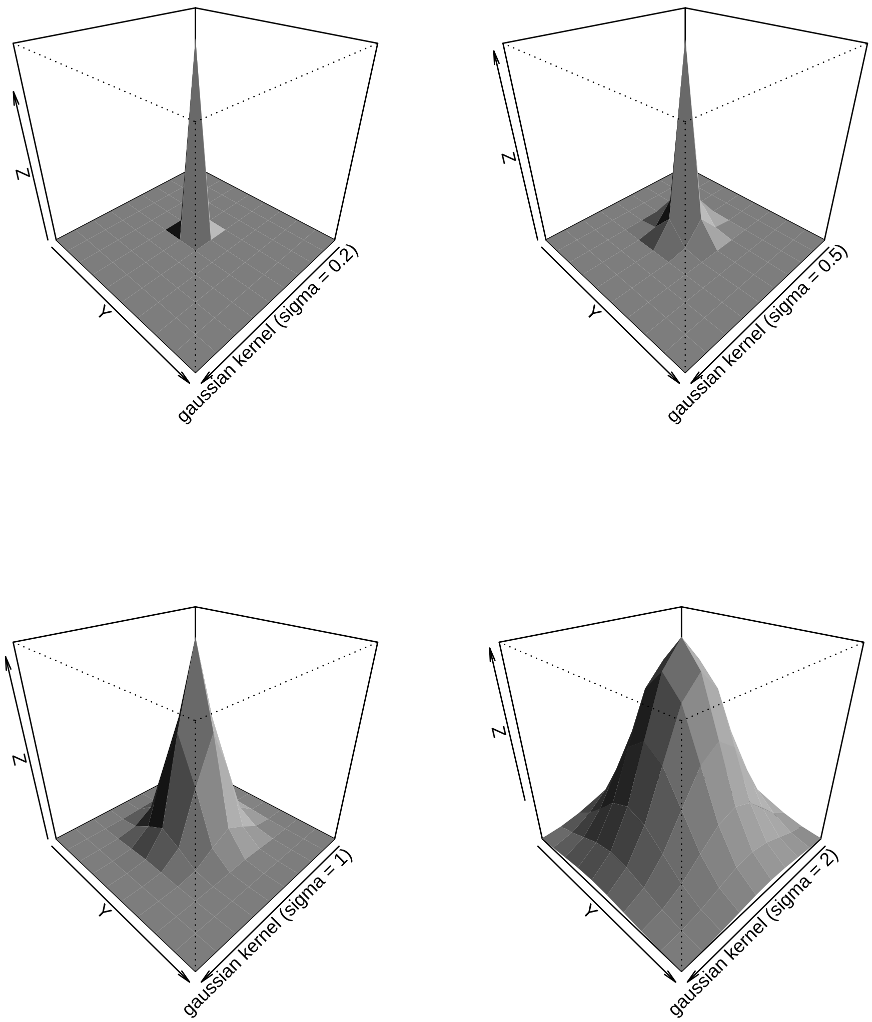



**Figure 4.4:** *Gaussian dispersal kernel*. Data were simulated using four different values of σ (0.2, 0.5, 1, and 2). Here, we show how those values affect the migration surface. Neighborhood and number of migrants (z-axis) increases when σ increases.

nectivity is achieved when individuals with membership in different groups (e.g., population, meta-population, or species) occur at equal frequencies. Minimum connectivity occurs when all individuals belong to the same group.

### 4.3.2 Within- and between-population connectivity

#### 4.3.2.1 Comparing results for different parameters

The $MS_{Conn}$ metric is sensitive to migration rate as well as selection. When migration rate was low, overall differences in connectivity grids between neutral loci versus loci under selection were higher Figures 4.5–4.7 in the *Gaussian* landscape with fine-scale environmental heterogeneity.

Differences between neutral loci and loci under selection were not as pronounced when the scale of environmental heterogeneity was higher, such as the medium-scale *gradient* landscape, (Figures D.2–D.4), or the coarse-grain *random* (Figures D.5–D.7) and *uniform* landscapes (Figures D.8–D.10).

For the *Gaussian* landscape, the largest differences in functional connectivity values were observed when comparing low-migration to medium- and high-migration simulations, both at neutral (RMSE = 0.211 and 0.207, respectively) and selected loci (RMSE = 0.357 and 0.381; Table 4.1). When migration rate was low, selection (cf. neutrality) played a big role in connectivity values (RMSE = 0.388), but less so when migration rate was medium (RMSE = 0.137) or high (RMSE = 0.110; Table 4.1). The largest difference in connectivity for the *gradient* landscape was also seen when comparing the low-migration simulation to the medium- and high-migration simulations (RMSE = 0.259 and 0.243; Table D.1). Compared to the fine- and medium-grain heterogeneity landscapes, RMSE values were lower for the coarse-grain heterogeneity *random* and *uniform* landscapes (Tables D.2 and D.3).

### 4.3.3 Within-community connectivity

Cells with high connectivity values (Figure 4.8) are the ones that have close to equal numbers of all three species, which, in community ecology, corresponds to high values of Pielou's evenness index, *J'*[273], calculated by dividing Shannon and Weaver's diversity index, *H'*[274], by the log of the number of species, *s*, in the community,

$$J' = \frac{H'}{\log s} = \frac{-{}_{i=1}^{s} \sum p_i \log p_i}{\log s} \tag{4.3}$$

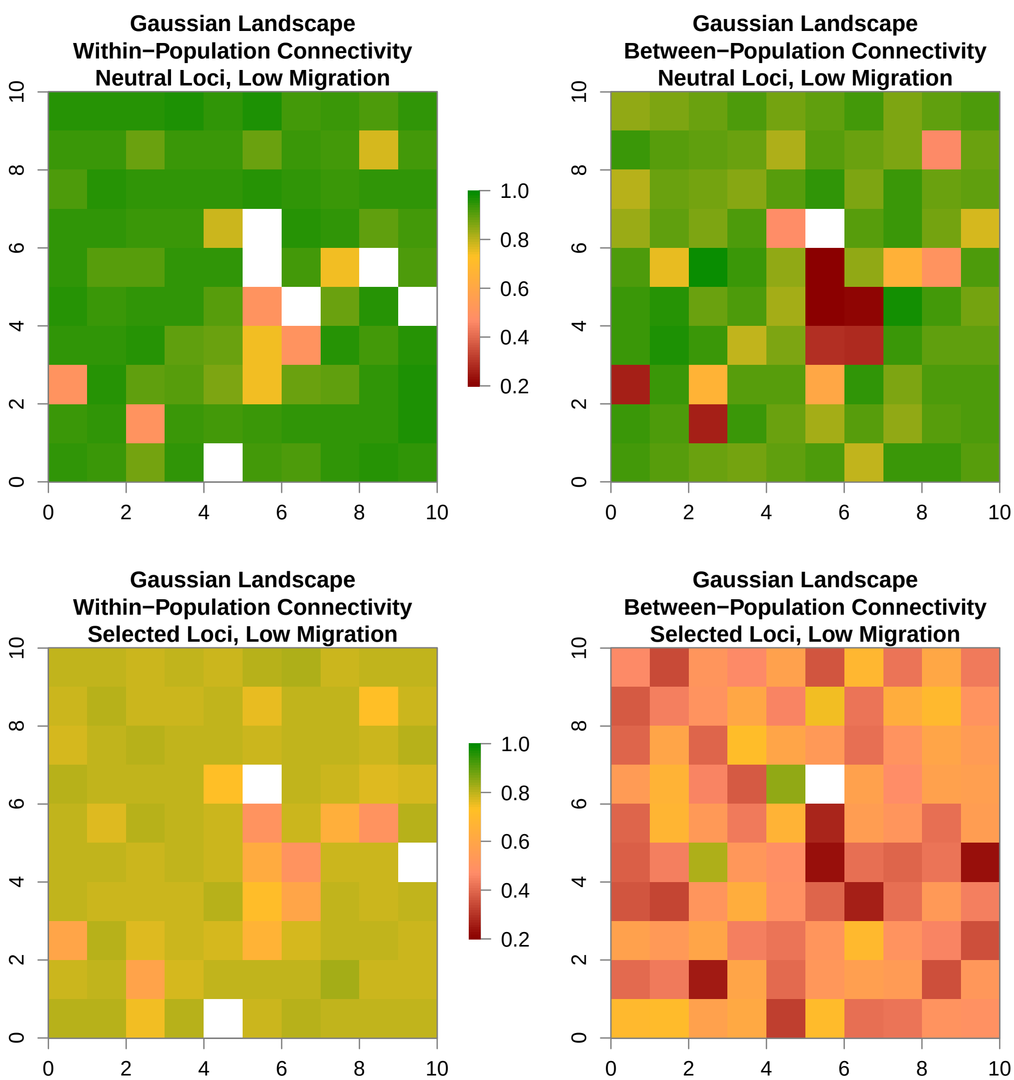



**Figure 4.5:** *Within- and between-population connectivity for genotypes simulated on the* Gaussian *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for non-neutral loci (s = 0.1). The simulations shown here were performed with σ = 0.2 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

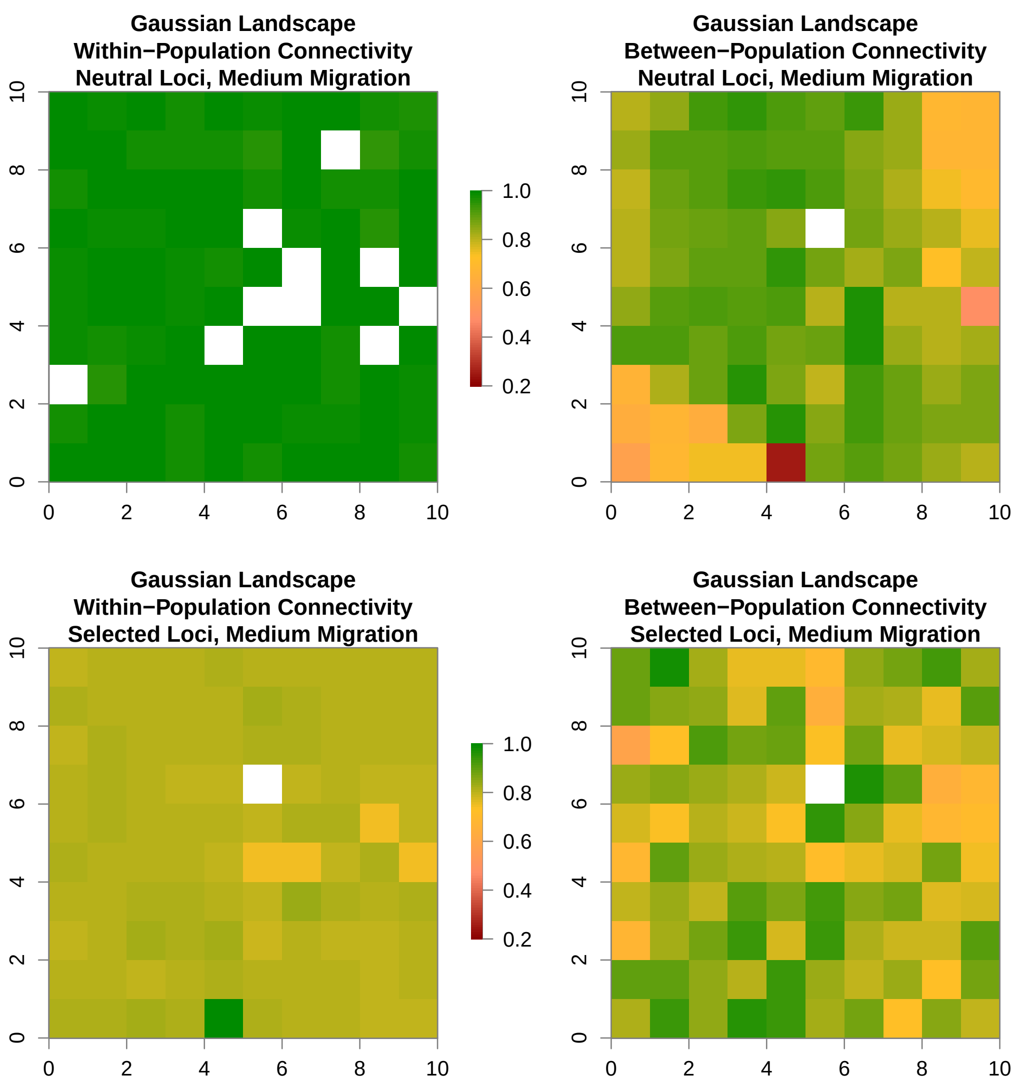



**Figure 4.6:** *Within- and between-population connectivity for genotypes simulated on the* Gaussian *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for non-neutral loci (s = 0.1). The simulations shown here were performed with σ = 0.5 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

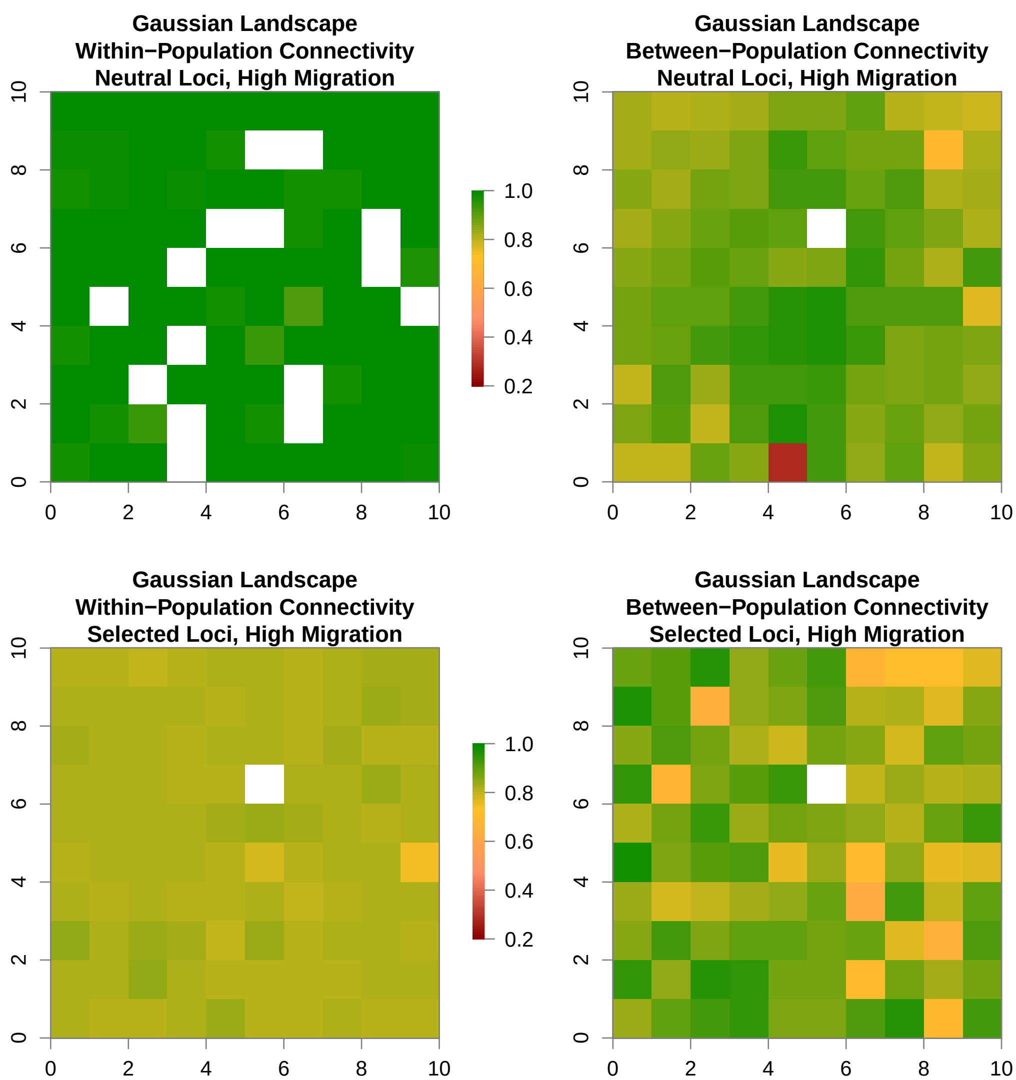



**Figure 4.7:** *Within- and between-population connectivity for genotypes simulated on the* Gaussian *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for non-neutral loci (s = 0.1). The simulations shown here were performed with σ = 1.0 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

**Table 4.1:** *Root mean square error (RMSE) of connectivity comparisons for the* Gaussian *landscape.* RMSE values shown here were used to quantify differences for within- and between-population connectivity based on neutral versus non-neutral loci (s = 0.1) for different degrees of long-distance dispersal (σ = 0.2, 0.5, and 1). RMSE values greater than 0.150 are italicized, whereas values greater than 0.250 are shown in bold.

**Root Mean Square Error of Functional Connectivity**

| **Gaussian Landscape** | | | **Within-Population** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Within** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | 0.140 | | | | | |
| | | $\sigma = 1.0$ | 0.132 | 0.027 | | | | |
| | *Selected* | $\sigma = 0.2$ | 0.141 | | | | | |
| | | $\sigma = 0.5$ | | *0.199* | | 0.071 | | |
| | | $\sigma = 1.0$ | | | *0.193* | 0.082 | 0.018 | |

| **Gaussian Landscape** | | | **Between-Population** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Between** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | *0.211* | | | | | |
| | | $\sigma = 1.0$ | *0.207* | 0.083 | | | | |
| | *Selected* | $\sigma = 0.2$ | **0.388** | | | | | |
| | | $\sigma = 0.5$ | | 0.137 | | **0.357** | | |
| | | $\sigma = 1.0$ | | | 0.110 | **0.381** | 0.116 | |

where $p_i$ is the proportion of species $i$ in the community.

Cells with the highest within-community connectivity correspond to cells where species II has 100 individuals (cf. to all other cells where this species is not present) and the other two species are present at similarly high numbers. Portions of the landscape were at least one species was absent (Figure 4.8) had low connectivity values.

## 4.4 Discussion

Landscape genetic studies focus predominantly on isolation by resistance, i.e., on the effect the intervening landscape has on genetic differentiation among individuals or populations[275], while effects of local environmental conditions are often neglected (but see[276]). Being an attribute of a sampling location–regardless whether this includes one or more individuals or an entire population–rather than capturing between-population dissimilarity, the $MS_{Conn}$ metric lends itself to being modeled in continuous space. Thus, the $MS_{Conn}$ metric would make it possible to model the effects on gene flow of both the intervening landscape and the local environmental conditions.

Modeling the effects of resistance alone represents an oversimplification of the dispersal process as just an escape from environmental unsuitability[277]. Individuals do not always escape unsuitable environments, and they can display a variety of dispersal strategies, even within a population or species[277]. A metric with properties such as $MS_{Conn}$ would enable the field of landscape genetics to move from pattern-oriented to process-oriented approaches.

The purview of landscape genetics is evolving to include interactions between the environment and adaptive genetic variation in natural populations. With the increasing availability of putative loci under selection for inference of local adaptation, new methods are appearing for detecting selection in landscape genomics studies (reviewed in[278]). As we have shown, the $MS_{Conn}$ metric is sensitive to gene flow as well as selection, and can thus be used in the new era of landscape genomics to quantify functional landscape connectivity with respect to both neutral and adaptive genetic variation.

$MS_{Conn}$ can be used in a hypothesis-testing framework, where neutral and non-neutral genotypes are simulated on an empirical landscape (i.e., study region), and then $MS_{Conn}$ can be calculated for all replicates of simulated neutral and selected genotypes, and compared against $MS_{Conn}$ calculated for the empirical genetic

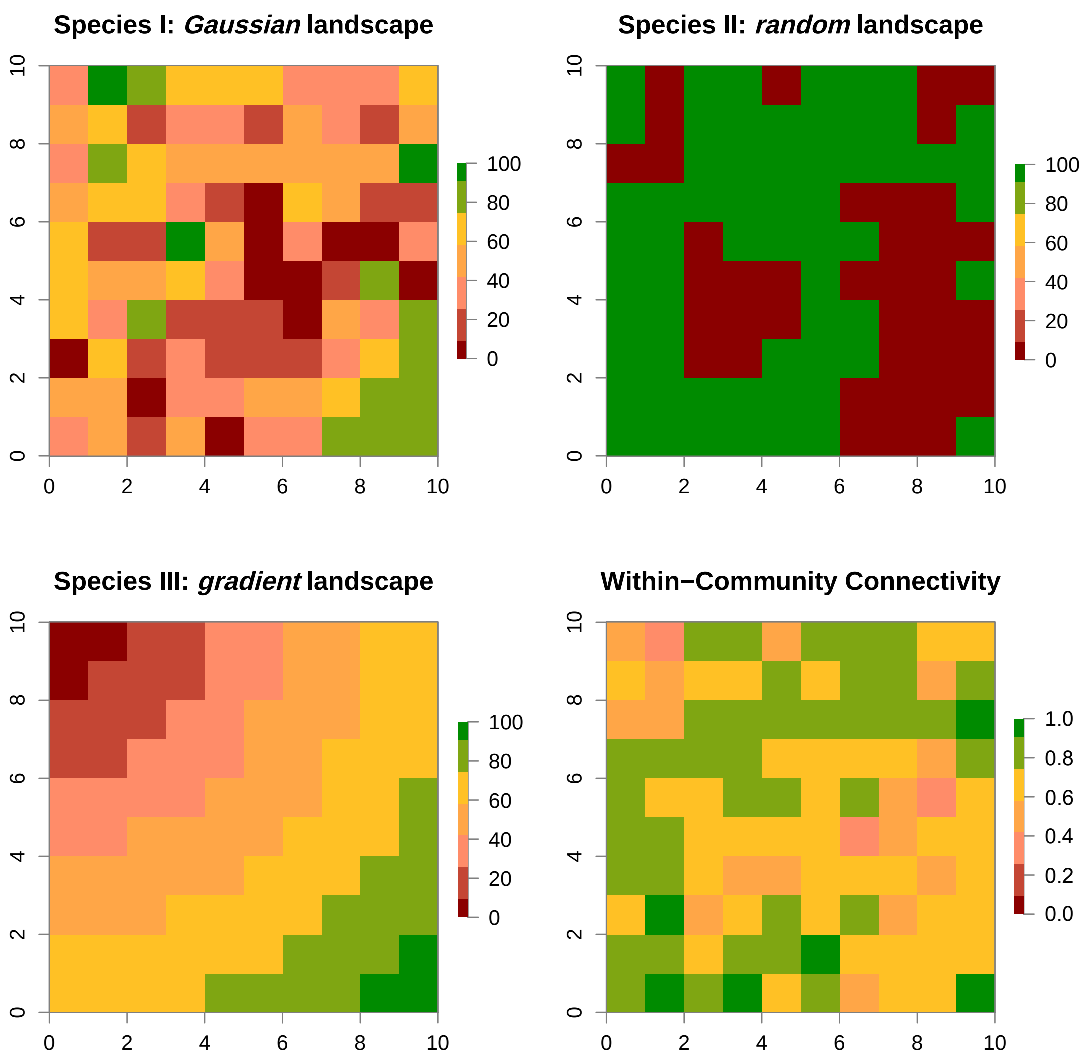



**Figure 4.8:** *Between-species (within-community) connectivity for three simulated species.* The top two panels show the *Gaussian* and *random* simulated landscapes, which represent species I and II, respectively, with a carrying capacity of 100 individuals per cell in the 10 x 10 grid. The bottom left panel represents species III (*uniform* landscape). The community in each cell is the sum of all three species' individuals. The bottom right panel shows within-community connectivity.

data, in order to infer whether the collected genetic data capture neutral or adaptive genetic variation. As we have shown that $MS_{Conn}$ is sensitive to the scale/grain of environmental heterogeneity, this hypothesis-testing framework should work for a range of empirical landscapes.

Environmental heterogeneity and its effect on connectivity can influence species interactions at the scale of meta-communities[270]. While effects of connectivity on biological diversity are largely consistent across levels of biological organization, scale represents a problem when capturing impacts of dispersal processes relevant to different levels of biological organization[279]. The scalability of $MS_{Conn}$ can be applied to addressing the role of connectivity in driving biodiversity patterns at all levels of biological organization, from alleles to communities.

Furthermore, $MS_{Conn}$ could be used in bringing together the fields of landscape genetics and community ecology. It has been recognized that interactions among species within a community can influence their genetic diversity (e.g.,[280]). In this new field of "landscape community genomics"[281], $MS_{Conn}$ could be used in constructing a framework to explicitly consider the effect on genetic variation of both biotic and abiotic factors. Given that there is overlap between landscape genetics and phylogeography[282], as part of such a framework, $MS_{Conn}$ could even be used to answer questions that are commonly asked in comparative phylogeography, such as whether species within a community have responded similarly to past and current geographic and ecological contexts.

### 4.4.1 Conclusions and future directions

$MS_{Conn}$ is applicable to the field of landscape genetics and community ecology, among others. At the within- and between-population scales, we observed sensitivity to environmental heterogeneity, as well as migration and selection. At the within-community scale, the metric shows similar properties to the Pielou evenness index. However, $MS_{Conn}$ could also be applied at a higher scale (e.g., communities that together form ecosystems), thus capturing connectivity between communities. To assess how well it captures between-community connectivity, the metric should be tested further on empirical data. Indeed, the next step is to apply the metric to empirical data at within- and between- scales, both at the population (landscape genetics) and community level (community ecology). In landscape genetics, $MS_{Conn}$ should be evaluated as the basis of a hypothesis-testing framework, involving comparisons of connectivity based on empirical genetic data versus simulations. Furthermore, more simulations are required to test the performance of the metric, especially as it pertains to sampling design, including the scale at which genetic data are collected, and the uniformity of sampling across the landscape, (e.g. clustered, randomly-, or uniformly-distributed sampling locations).

Data Accessibility: The Supplementary Material and additional data, as well as R scripts, are available online at `https://github.com/chazhyseni/MSconn`.

# CONCLUSION

Chapter 1 highlighted the roles that temperature and precipitation have played in driving niche divergence among a set of sympatric *Reticulitermes* species. The distribution of *R. flavipes* covers a wide range of environments compared to two other co-occurring species in the eastern U.S., *R. malletei* and *R. virginicus*. While the mid-latitudes of the southern Appalachians, characterized by complex topography and multiple ecoregions, provide suitable habitat to support at least three *Reticulitermes* species, competitive exclusion is a plausible explanation for apparent rare local-scale co-occurrence (i.e., micro-allopatry). Based on distribution modeling, *R. flavipes* is potentially able to exclude the other two species in the northern portion of the southern Appalachians, including western Kentucky, southern Ohio and Indiana, the majority of West Virginia and Pennsylvania, and parts of Virginia and North Carolina. Furthermore, there is separation in niche space among species, particularly *R. flavipes* and *R. virginicus*. Indeed, this study represents the first evidence of significant regional-scale niche divergence between *R. flavipes* and *R. virginicus*.

Chapter 2 showed that the distribution of *R. flavipes* has cycled latitudinally, first shifting northward toward the southern edge of the Laurentide ice sheet (e.g., Indiana, Ohio, Pennsylvania) during the Last Glacial Maximum (LGM; 22,000 years ago), then later shifting southward in the Holocene. Analyses of geo-referenced DNA sequence data identified three genetically distinct and geographically cohesive populations, corresponding with northern, central, and southern portions of the study region. Divergence between the Northern and Southern populations was the oldest, estimated to have occurred 65,000 years ago, while the Central and Northern populations diverged in the mid-Holocene, 9,000 years ago, after which the Central population continued to expand. This study contributes to a growing body of literature that highlights an important role for multiple refugia—

including those located further north than previously expected. Indeed, a northern refuge played a key role in subsequent colonization by *R. flavipes* of the central region of the southern Appalachians. Although somewhat unexpected, the existence of northern refugia close to the southern edge of the Laurentide ice sheet during the LGM is plausible owing to localized warm areas in close proximity to glaciers (e.g.,[37,171–175]). Thus, distributional shifts have resulted in populations contracting and becoming isolated, followed by expansions, with different populations developing associations with different environments, such as the Southern population being associated with higher wet-season precipitation than the other two populations.

Unlike glacial-interglacial oscillations, the dynamics of forest ecosystems may have had a more contemporary effect on intraspecific variation in *R. flavipes*, especially given that extensive harvest and exclusion of fire has affected the composition of eastern U.S. forests. Accordingly, Chapter 3 explored contributions of human disturbance of forest ecosystems to epigenetic variation in *R. flavipes*. I found that epigenetic variation was higher in more disturbed environments (i.e., lower canopy cover) compared to epigenetic variation in environments with high tree canopy cover and species richness. This study is the first to record an effect of canopy cover on intraspecific epigenetic variation in termites. In addition, DNA methylation differences at nine loci were significantly associated with differences in canopy cover. With the climate of the last 9,000 years being conducive to population expansion of *R. flavipes*, and the new context of human-induced disturbance of forest ecosystems, the species has expanded its niche to include human-altered habitats. As a mechanism to deal with novel environments, phenotypic plasticity underpinned by DNA methylation likely played a part in the survival and establishment of *R. flavipes* in human-altered habitats in the species' native range in the eastern U.S. Indeed, this may have been the prelude to *R. flavipes* becoming invasive in other parts of the world.

Together, these three chapters used the subterranean termite system to illustrate the influence of the geographic and ecological context on evolutionary processes, at historical and contemporary timescales. Indeed, environmental heterogeneity, as well as interactions among species, can influence gene flow. In Chapter 4, I developed a new landscape connectivity metric, $MS_{Conn}$. At the population level, this metric was sensitive to gene flow as well as selection, and could thus be used to quantify functional landscape connectivity with respect to both neutral and adaptive genetic variation. At the species level (within communities), the metric showed properties similar to the Pielou evenness index. However, $MS_{Conn}$ could also be applied at a higher scale (e.g., communities within ecosystems), thus capturing connectivity between communities. Ultimately, the $MS_{Conn}$ metric could be integrated into an eco-evolutionary framework, and thus bring together the fields of landscape genetics and community ecology, by making it possible to quantify the effect of biotic and abiotic environments on gene flow between populations, as well as the effect of gene flow on species interactions within and between communities.

# BIBLIOGRAPHY

[1] Hansen J, Sato M, Russell G, Kharecha P (2013) Climate sensitivity, sea level and atmospheric carbon dioxide. *Phil. Trans. R. Soc. A* 371(2001):20120294.

[2] Hudson R, Slatkin M, Maddison W (1992) Estimation of levels of gene flow from dna sequence data. *Genetics* 132:583–589.

[3] Laland KN, Odling-Smee FJ, Feldman MW (1999) Evolutionary consequences of niche construction and their implications for ecology. *Proc. Natl. Acad. Sci. U. S. A* 96(18):10242–10247.

[4] Hairston, Jr NG, Ellner SP, Geber MA, Yoshida T, Fox JA (2005) Rapid evolution and the convergence of ecological and evolutionary time. *Ecol. Lett* 8(10):1114–1127.

[5] Whitham TG, et al. (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nat. Rev. Genet* 7(7):510–523.

[6] Post DM, Palkovacs EP (2009) Eco-evolutionary feedbacks in community and ecosystem ecology: interactions between the ecological theatre and the evolutionary play. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 364(1523):1629–1640.

[7] Pelletier F, Garant D, Hendry AP (2009) Eco-evolutionary dynamics. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 364(1523):1483–1489.

[8] Fischer BB, et al. (2014) Phenotypic plasticity influences the eco-evolutionary dynamics of a predator—prey system. *Ecology* 95(11):3080–3092.

[9] Hendry AP (2016) Key questions on the role of phenotypic plasticity in eco-evolutionary dynamics. *J. Hered* 107(1):25–41.

[10] Bonasio R, et al. (2012) Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Curr. Biol* 22(19):1755–1764.

[11] Cardoso-Júnior CAM, et al. (2017) Epigenetic modifications and their relation to caste and sex determination and adult division of labor in the stingless bee *Melipona scutellaris*. *Genet. Mol. Biol* 40(1):61–68.

[12] Weiner SA, et al. (2013) A survey of DNA methylation across social insect species, life stages, and castes reveals abundant and caste-associated methylation in a primitively social wasp. *Naturwissenschaften* 100(8):795–799.

[13] Lo N, Li B, Ujvari B (2012) DNA methylation in the termite *Coptotermes lacteus*. *Insectes Soc* 59(2):257–261.

[14] Glastad KM, Hunt BG, Goodisman MAD (2013) Evidence of a conserved functional role for DNA methylation in termites. *Insect Mol. Biol* 22(2):143–154.

[15] Glastad KM, Gokhale K, Liebig J, Goodisman MAD (2016) The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Sci. Rep* 6:37110.

[16] Fittkau EJ, Klinge H (1973) On biomass and trophic structure of the central amazonian rain forest ecosystem. *Biotropica* 5(1):2–14.

[17] Moe SR, Mobæk R, Narmo AK (2009) Mound building termites contribute to savanna vegetation heterogeneity. *Plant Ecol* 202(1):31.

[18] Erpenbach A, Bernhardt-Römermann M, Wittig R, Thiombiano A, Hahn K (2013) The influence of termite-induced heterogeneity on savanna vegetation along a climatic gradient in west africa. *J. Trop. Ecol* 29(1):11–23.

[19] Vargo E, Husseneder C (2009) Biology of subterranean termites: insights from molecular studies of *Reticulitermes* and *Coptotermes*. *Annu. Rev. Entomol* 54:379–403.

[20] Jackson B, et al. (2009) Species diversity and composition in old growth and second growth rich coves of the southern appalachian mountains. *Castanea* 74:27–38.

[21] Buczkowski G, Bertelsmeier C (2017) Invasive termites in a changing climate: a global perspective. *Ecol. Evol* 7(3):974–985.

[22] McKern J, Szalanski A, Austin J (2006) First record of *Reticulitermes flavipes* and *Reticulitermes hageni* in oregon (isoptera: Rhinotermitidae). *Fla. Entomol* 89(4):541–542.

[23] Austin JW, et al. (2005) Genetic evidence for the synonymy of two *Reticulitermes* species: *Reticulitermes flavipes* and *Reticulitermes santonensis*. *Ann. Entomol. Soc. Am* 98(3):395–401.

[24] Ghesini S, Messenger MT, Pilon N, Marini M (2010) First report of *Reticulitermes flavipes* (isoptera: Rhinotermitidae) in italy. *Fla. Entomol* 93(2):327–328.

[25] Perdereau E, et al. (2013) Global genetic analysis reveals the putative native source of the invasive termite, *Reticulitermes flavipes*, in france. *Mol. Ecol* 22:1105–1119.

[26] Hernández-Teixidor D, Suárez D, García J, Mora D (2019) First report of the invasive *Reticulitermes flavipes* (kollar, 1837) (blattodea, rhinotermitidae) in the canary islands. *J. Appl. Entomol* 143(4):478–482.

[27] Bourguignon T, et al. (2016) Oceanic dispersal, vicariance and human introduction shaped the modern distribution of the termites *Reticulitermes*, *Heterotermes* and *Coptotermes*. *Proc. Biol. Sci* 283(1827):20160179.

[28] Suppo C, Robinet C, Perdereau E, Andrieu D, Bagnères AG (2018) Potential spread of the invasive north american termite, *Reticulitermes flavipes*, and the impact of climate warming. *Biol. Invasions* 20(4):905–922.

[29] Perdereau E, et al. (2019) Invasion dynamics of a termite, *Reticulitermes flavipes*, at different spatial scales in france. *Insects* 10(1).

[30] Clark S (2001) Birth of the mountains: The geologic story of the southern appalachian mountains, (Washington, DC, USA), USGS Reports; US Government Printing Office.

[31] Pittillo J, Hatcher R, Buol S (1998) Introduction to the environment and vegetation of the southern blue ridge province. *Castanea* 63:202–216.

[32] Hayes M, Moody A, White P, Costanza J (2007) The influence of logging and topography on the distribution of spruce-fir forests near their southern limits in great smoky mountains national park, usa. *Plant Ecol* 189:59–70.

[33] Whittaker R (1956) Vegetation of the great smoky mountains. *Ecol. Monogr* 26:1–80.

[34] Bennett K (1985) The spread of fagus grandifolia across eastern north america during the last 18000 years. *J. Biogeogr* 12:147–164.

[35] Bennett K (1986) The rate of spread and population increase of forest trees during the postglacial. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 314:523–531.

[36] Delcourt H, Delcourt P (1988) Quaternary landscape ecology: Relevant scales in space and time. *Landsc. Ecol* 2:23–44.

[37] Williams J, Post D, Cwynar L, Lotter A, Levesque A (2002) Rapid and widespread vegetation responses to past climate change in the north atlantic region. *Geology* 30:971–974.

[38] Williams J, Shuman B, Webb, III T, Bartlein P, Leduc P (2004) Late-quaternary vegetation dynamics in north america: scaling from taxa to biomes. *Ecol. Monogr* 74:309–334.

[39] Petranka J (1998) *Salamanders of the United States and Canada*. (Smithsonian Institution Press, Washington, DC, USA).

[40] Rissler L, Smith W (2010) Mapping amphibian contact zones and phylogeographical break hotspots across the united states. *Mol. Ecol* 19:5404–5416.

[41] Snyder B (2008) A preliminary checklist of the millipedes (diplopoda) of the great smoky mountains national park, usa. *Zootaxa* pp. 16–32.

[42] Marek P (2010) A revision of the appalachian millipede genus *Brachoria* chamberlin, 1939 (polydesmida: Xystodesmidae: Apheloriini). *Zool. J. Linn. Soc* 159:817–889.

[43] Nalepa C, Shimada K, Maekawa K, Luykx P (2017) Distribution of karyotypes of the *Cryptocercus punctulatus* species complex (blattodea: Cryptocercidae) in great smoky mountains national park. *J. Insect Sci* 17:69.

[44] Garrick R, Sabree Z, Jahnes B, Oliver J (2017) Strong spatial-genetic congruence between a wood-feeding cockroach and its bacterial endosymbiont, across a topographically complex landscape. *J. Biogeogr* 44:1500–1511.

[45] Garrick R, Newton K, Worthington R (2018) Cryptic diversity in the southern appalachian mountains: genetic data reveal that the red centipede, *Scolopocryptops sexspinosus*, is a species complex. *J. Insect Sci* 22:799–805.

[46] Ulyshen M (2013) Strengthening the case for saproxylic arthropod conservation: a call for ecosystem services research. *Insect Conserv. Divers* 6:393–395.

[47] Ulyshen M, Wagner T (2013) Quantifying arthropod contributions to wood decay. *Methods Ecol. Evol* 4:345–352.

[48] Ulyshen M (2014) Interacting effects of insects and flooding on wood decomposition. *PLoS ONE* 9:101867.

[49] Ulyshen M (2015) Insect-mediated nitrogen dynamics in decomposing wood. *Ecol. Entomol* 40:97–112.

[50] Ulyshen M (2016) Wood decomposition as influenced by invertebrates. *Biol. Rev* 91:70–85.

[51] Stokland J, Siitonen J, Jonsson B (2012) *Biodiversity in Dead Wood*. (Cambridge University Press, Cambridge, UK).

[52] Harmon M, et al. (1986) Ecology of coarse woody debris in temperate ecosystems in *Advances in Ecological Research*, eds. MacFadyen A, Ford E. (Academic Press, Cambridge, MA, USA) Vol. 15, pp. 133–302.

[53] Lim S, Forschler B (2012) *Reticulitermes nelsonae*, a new species of subterranean termite (rhinotermitidae) from the southeastern united states. *Insects* 3:62–90.

[54] Garrick R, Collins B, Yi R, Dyer R, Hyseni C (2015) Identification of eastern united states *Reticulitermes* termite species via pcr-rflp, assessed using training and test data. *Insects* 6:524–537.

[55] Elith J, Leathwick J (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst* 40:677–697.

[56] Kabir M, et al. (2017) Habitat suitability and movement corridors of grey wolf (*Canis lupus*) in northern pakistan. *PLoS ONE* 12:0187027.

[57] Milicic M, Vujic A, Jurca T, Cardoso P (2017) Designating conservation priorities for southeast european hoverflies (diptera: Syrphidae) based on species distribution models and species vulnerability. *Insect Conserv. Divers* 10:354–366.

[58] Bosso L, et al. (2018) Nature protection areas of europe are insufficient to preserve the threatened beetle *Rosalia alpina* (coleoptera: Cerambycidae): evidence from species distribution models and conservation gap analysis. *Ecol. Entomol* 43:192–203.

[59] Da Mata R, et al. (2017) Stacked species distribution and macroecological models provide incongruent predictions of species richness for drosophilidae in the brazilian savanna. *Insect Conserv. Divers* 10:415–424.

[60] Macfadyen S, McDonald G, Hill M (2018) From species distributions to climate change adaptation: knowledge gaps in managing invertebrate pests in broad-acre grain crops. *Agric. Ecosyst. Environ* 253:208–219.

[61] Kriticos D, et al. (2017) The potential global distribution ofthe brown marmorated stink bug, *Halyomorpha halys*, a critical threat to plant biosecurity. *J. Pest Sci* 90:1033–1043.

[62] Barbet-Massin M, Rome Q, Villemant C, Courchamp F (2018) Can species distribution models really predict the expansion of invasive species? *PLoS ONE* 13:e0193085.

[63] Hijmans R, Cameron S, Parra J, Jones P, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol* 25:1965–1978.

[64] Kriticos D, et al. (2012) Climond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods Ecol. Evol* 3:53–64.

[65] Title P, Bemmels J (2018) Envirem: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* 41:291–307.

[66] Araújo M, Peterson A (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology* 93:1527–1539.

[67] Kozak K, Wiens J (2006) Does niche conservatism promote speciation? a case study in north american salamanders. *Evolution* 60:2604–2621.

[68] Peterson A, Soberon J, Sanchez-Cordero V (1999) Conservatism of ecological niches in evolutionary time. *Science* 285:1265–1267.

[69] Rödder D, Lötters S (2009) Niche shift versus niche conservatism? climatic characteristics of the native and invasive ranges of the mediterranean house gecko (*Hemidactylus turcicus*). *Glob. Ecol. Biogeogr* 18:674–687.

[70] Maynard D, Crowther T, King J, Warren R, Bradford M (2015) Temperate forest termites: ecology, biogeography, and ecosystem impacts. *Ecol. Entomol* 40:199–210.

[71] Thuiller W, Lafourcade B, Engler R, Araújo M (2009) Biomod—a platform for ensemble forecasting of species distributions. *Ecography* 32:369–373.

[72] Thuiller W, Georges D, Engler R, Breiner F (2016) *biomod2: ensemble platform for species distribution modeling*.

[73] R Core Team (2018) *R: a language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).

[74] Alarcón D, Cavieres L (2018) Relationships between ecological niche and expected shifts in elevation and latitude due to climate change in south american temperate forest plants. *J. Biogeogr* 45:2272–2287.

[75] Smeraldo S, et al. (2018) Ignoring seasonal changes in the ecological niche of non-migratory species may lead to biases in potential distribution models: Lessons from bats. *Biodivers. Conserv* 27:2425–2441.

[76] Zacarias D, Loyola R (2018) Distribution modelling and multi-scale landscape connectivity highlight important areas for the conservation of savannah elephants. *Biol. Conserv* 224:1–8.

[77] Evans M, Smith S, Flynn R, Donoghue M (2009) Climate, niche evolution, and diversification of the "bird-cage" evening primroses (oenothera, sections anogra and kleinia. *Am. Nat* 173:225–240.

[78] Heibl C, Calenge C (2018) phyloclim: integrating phylogenetics and climatic niche modeling.

[79] Schoener T (1968) The anolis lizards of bimini: resource partitioning in a complex fauna. *Ecology* 49:704–726.

[80] Warren D, Glor R, Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62:2868–2883.

[81] Calabrese J, Certain G, Kraan C, Dormann C (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Glob. Ecol. Biogeogr* 23:99–112.

[82] Liu C, Berry P, Dawson T, Pearson R (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28:385–393.

[83] Liu C, White M, Newell G (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *J. Biogeogr* 40:778–789.

[84] Liu C, Newell G, White M (2016) On the selection of thresholds for predicting species occurrence with presence-only data. *Ecol. Evol* 6:337–348.

[85] Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol* 43:1223–1232.

[86] Oksanen J, et al. (2018) vegan: Community ecology package.

[87] Legendre P, Anderson M (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr* 69:1–24.

[88] Loehle C (2007) Predicting pleistocene climate from vegetation in north america. *Clim. Past* 3:109–118.

[89] Swenson N, Howard D (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in north america. *Am. Nat* 166:581–591.

[90] Soltis D, Morris A, McLachlan J, Manos P, Soltis P (2006) Comparative phylogeography of unglaciated eastern north america. *Mol. Ecol* 15:4261–4293.

[91] Walker M, Stockman A, Marek P, Bond J (2009) Pleistocene glacial refugia across the appalachian mountains and coastal plain in the millipede genus *Narceus*: evidence from population genetic, phylogeographic, and paleoclimatic data. *BMC Evol. Biol* 9:25.

[92] Leponce M, Roisin Y, Pasteels J (1997) Structure and dynamics of the arboreal termite community in new guinean coconut plantations. *Biotropica* 29:193–203.

[93] Hubbell S (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton University Press, Princeton, NJ, USA).

[94] Evans T, et al. (2009) Termites eavesdrop to avoid competitors. *Proc. Biol. Sci* 276:4035–4041.

[95] Oberst S, Bann G, Lai J, Evans T (2017) Cryptic termites avoid predatory ants by eavesdropping on vibrational cues from their footsteps. *Ecol. Lett* 20:212–221.

[96] Perdereau E, Bagnères AG, Dupont S, Dedeine F (2010) High occurrence of colony fusion in a european population of the american termite *Reticulitermes flavipes*. *Insectes Soc* 57:393–402.

[97] Perdereau E, Dedeine F, Christidès JP, Dupont S, Bagnères AG (2011) Competition between invasive and indigenous species: an insular case study of subterranean termites. *Biol. Invasions* 13:1457–1470.

[98] Hyodo F (2015) Use of stable carbon and nitrogen isotopes in insect trophic ecology. *Entomol. Sci* 18:295–312.

[99] Hewitt G (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc. Lond* 58:247–276.

[100] Avise J (2000) *Phylogeography: The History and Formation of Species*. (Harvard University Press, Cambridge, USA).

[101] Hewitt G (2004) Genetic consequences of climatic oscillations in the quaternary. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 359:183–195.

[102] Gomez A, Lunt D (2007) Refugia within refugia: Patterns of phylogeographic concordance in the iberian peninsula in *Phylogeography of Southern European Refugia: Evolutionary Perspectives on the Origins and Conservation of European Biodiversity*, eds. Weiss S, Ferrand N. (Springer, Dordrecht, Netherlands), pp. 155–188.

[103] Shafer A, Cullingham C, Cote S, Coltman D (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern north america. *Mol. Ecol* 19:4589–4621.

[104] Byrne M (2008) Evidence for multiple refugia at different time scales during pleistocene climatic oscillations in southern australia inferred from phylogeography. *Quat. Sci. Rev* 27:2576–2585.

[105] Carstens B, Knowles L (2007) Shifting distributions and speciation: species divergence during rapid climate change. *Mol. Ecol* 16:619–627.

[106] Hibbard K, Meehl G, Cox P, Friedlingstein P (2007) A strategy for climate change stabilization experiments. *Eos Trans. Amer. Geophys. Union* 88:217–221.

[107] Crandall K, Buhay J (2008) Global diversity of crayfish (astacidae, cambaridae, and parastacidae—decapoda) in freshwater in *Freshwater Animal Diversity Assessment*, eds. Balian E, Leveque C, Segers H, Martens K. (Springer, Dordrecht, Netherlands), pp. 295–301.

[108] Marek P, Bond J (2009) A mullerian mimicry ring in appalachian millipedes. *Proc. Natl. Acad. Sci. U. S. A* 106:9755–9760.

[109] Hedin M, Wood D (2002) Genealogical exclusivity in geographically proximate populations of *Hypochilus thorelli* marx (araneae, hypochilidae) on the cumberland plateau of north america. *Mol. Ecol* 11:1975–1988.

[110] Thomas S, Hedin M (2008) Multigenic phylogeographic divergence in the paleoendemic southern appalachian opilionid *Fumontana deprehendor* shear (opiliones, laniatores, triaenonychidae. *Mol. Phylogenet. Evol* 46:645–658.

[111] Myer A, Forschler BT (2019) Evidence for the role of subterranean termites (*Reticulitermes* spp.) in temperate forest soil nutrient cycling. *Ecosystems* 22.

[112] Wiltz B (2015) Effect of temperature and humidity on survival of *Coptotermes formosanus* and *Reticulitermes flavipes* (isoptera: Rhinotermitidae. *Sociobiology* 59:381–394.

[113] Vargo E, Carlson J (2006) Comparative study of breeding systems of sympatric subterranean termites (*Reticulitermes flavipes* and *R. hageni*) in central north carolina using two classes of molecular genetic markers. *Environ. Entomol* 35:173–187.

[114] Thorne B, Traniello J, Adams E, Bulmer M (1999) Reproductive dynamics and colony structure of subterranean termites of the genus *Reticulitermes* (isoptera: Rhinotermitidae): a review of the evidence from behavioral, ecological, and genetic studies. *Ethol. Ecol. Evol* 11:149–169.

[115] Cruzan M, Templeton A (2000) Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record. *Trends Ecol. Evol* 15:491–496.

[116] Knowles L (2009) Statistical phylogeography. *Annu. Rev. Ecol. Evol. Syst* 40:593–612.

[117] Hickerson M, et al. (2010) Phylogeography's past, present, and future: 10 years after avise. *Phylogenet. Evol* 54:291–301.

[118] Richards C, Carstens B, Knowles L (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *J. Biogeogr* 34:1833–1845.

[119] Alexandrino J, Teixeira J, Arntzen J, Ferrand N (2007) Historical biogeography and conservation of the golden-striped salamander (*Chioglossa lusitanica*) in northwestern iberia: integrating ecological, phenotypic and phylogeographic data in *Phylogeography of Southern European Refugia: Evolutionary Perspectives on the Origins and Conservation of European Biodiversity*, eds. Weiss S, Ferrand N. (Springer, Dordrecht, Netherlands), pp. 189–205.

[120] Waltari E, et al. (2007) Locating pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS One* 2:e563.

[121] Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecol. Lett* 8:993–1009.

[122] Pielou E (1991) *After the Ice Age: The Return of Life to Glaciated North America*. (University of Chicago Press, Chicago, USA).

[123] Knowles L, Alvarado-Serrano D (2010) Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. *Mol. Ecol* 19:3727–3745.

[124] Hugall A, Moritz C, Moussalli A, Stanisic J (2002) Reconciling paleodistribution models and comparative phylogeography in the wet tropics rainforest land snail *Gnarosophia bellendenkerensis* (brazier 1875. *Proc. Natl. Acad. Sci. U.S.A* 99:6112–6117.

[125] Garrick R, et al. (2004) Phylogeography recapitulates topography: very fine-scale local endemism of a saproxylic 'giant' springtail at tallaganda in the great dividing range of south-east australia. *Mol. Ecol* 13:3329–3344.

[126] Sunnucks P, et al. (2006) A tale of two flatties: different responses of two terrestrial flatworms to past environmental climatic fluctuations at tallaganda in montane southeastern australia. *Mol. Ecol* 15:4513–4531.

[127] Wang C, et al. (2009) Survey and identification of termites (isoptera: Rhinotermitidae) in indiana. *Ann. Entomol. Soc. Am* 102:1029–1036.

[128] Kearse M, et al. (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649.

[129] Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet* 68:978–989.

[130] Barbet-Massin M, Jiguet F, Albert C, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol* 3:327–338.

[131] Dormann C, et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36:27–46.

[132] Cheng L, Connor T, Siren J, Aanensen D, Corander J (2013) Hierarchical and spatially explicit clustering of dna sequences with baps software. *Mol. Biol. Evol* 30:1224–1228.

[133] Lanfear R, Calcott B, Ho S, Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol* 29:1695–1701.

[134] Darriba D, Taboada G, Doallo R, Posada D (2012) jmodeltest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9(772).

[135] Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol* 22:160–174.

[136] Bouckaert R, et al. (2014) Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol* 10:e1003537.

[137] Drummond A, Ho S, Phillips M, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.

[138] Brower A (1994) Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial dna evolution. *Proc. Natl. Acad. Sci. U. S. A* 91:6491–6495.

[139] Luchetti A, Marini M, Mantovani B (2005) Mitochondrial evolutionary rate and speciation in termites: data on european *Reticulitermes* taxa (isoptera, rhinotermitidae. *Insectes Soc* 52:218–221.

[140] Beaumont M, Zhang W, Balding D (2002) Approximate bayesian computation in population genetics. *Genetics* 162:2025–2035.

[141] Rambaut A, Drummond A, Xie D, Baele G, Suchard M (2018) Tracer v1.7. Available from.

[142] Librado P, Rozas J (2009) Dnasp v5: a software for comprehensive analysis of dna polymorphism data. *Bioinformatics* 25:1451–1452.

[143] Nei M (1987) *Molecular Evolutionary Genetics*. (Columbia University Press, New York, USA).

[144] Tajima F (1983) Evolutionary relationship of dna sequences in finite populations. *Genetics* 105:437–460.

[145] Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol* 7:256–276.

[146] Paradis E, Claude J, Strimmer K (2004) Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics* 20:289–290.

[147] Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Mol. Biol. Evol* 10:512–526.

[148] Cornuet JM, et al. (2014) Diyabc v2.0: a software to make approximate bayesian computation inferences about population history using single nucleotide polymorphism, dna sequence and microsatellite data. *Bioinformatics* 30:1187–1189.

[149] Pelletier T, Carstens B (2014) Model choice for phylogeographic inference using a large set of models. *Mol. Ecol* 23:3028–3043.

[150] Espindola A, et al. (2016) Identifying cryptic diversity with predictive phylogeography. *Proc. R. Soc. B* 283(20161529).

[151] Stone G, et al. (2017) Tournament abc analysis of the western palaearctic population history of an oak gall wasp, *Synergus umbraculus*. *Mol. Ecol* 26:6685–6703.

[152] Garrick R, et al. (2014) Lineage fusion in galapagos giant tortoises. *Mol. Ecol* 23:5276–5290.

[153] Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* 123:585–595.

[154] Fagundes N, et al. (2007) Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U.S.A* 104:17614–17619.

[155] Cornuet JM, et al. (2008) Inferring population history with diy abc: a user-friendly approach to approximate bayesian computation. *Bioinformatics* 24:2713–2719.

[156] Cornuet JM, Ravigne V, Estoup A (2010) Inference on population history and model checking using dna sequence and microsatellite data with the software diyabc (v1.0). *BMC Bioinformatics* 11(401).

[157] Fu Y, Li W (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

[158] Ramos-Onsins S, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol* 19:2092–2100.

[159] Rogers A, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol* 9:552–569.

[160] Zeng K, Shi S, Wu CI (2007) Compound tests for the detection of hitchhiking under positive selection. *Mol. Biol. Evol* 24:1898–1908.

[161] Heled J, Drummond A (2008) Bayesian inference of population size history from multiple loci. *BMC Evol. Biol* 8(289).

[162] Zamudio K, Savage W (2003) Historical isolation, range expansion, and secondary contact of two highly divergent mitochondrial lineages in spotted salamanders (*Ambystoma maculatum*). *Evolution* 57:1631–1652.

[163] Crespi E, Rissler L, Browne R (2003) Testing pleistocene refugia theory: phylogeographical analysis of *Desmognathus wrighti*, a high-elevation salamander in the southern appalachians. *Mol. Ecol* 12:969–984.

[164] Jones M, Voss S, Ptacek M, Weisrock D, Tonkyn D (2006) River drainages and phylogeography: an evolutionary significant lineage of shovel-nosed salamander (*Desmognathus marmoratus*) in the southern appalachians. *Mol. Phylogenet. Evol* 38:280–287.

[165] Kuchta S, Haughey M, Wynn A, Jacobs J, Highton R (2016) Ancient river systems and phylogeographical structure in the spring salamander, *Gyrinophilus porphyriticus*. *J. Biogeogr* 43:639–652.

[166] Jones K, Weisrock D (2018) Genomic data reject the hypothesis of sympatric ecological speciation in a clade of *Desmognathus* salamanders. *Evolution* 72:2378–2393.

[167] Nalepa C, Luykx P, Klass KD, Deitz L (2002) Distribution of karyotypes of the *Cryptocercus punctulatus* species complex (dictyoptera: Cryptocercidae) in the southern appalachians: Relation to habitat and history. *Ann. Entomol. Soc. Am* 95:276–287.

[168] Caterino M, Langton-Myers S (2018) Long-term population persistence of flightless weevils (*Eurhoptus pyriformis*) across old- and second-growth forests patches in southern appalachia. *BMC Evol. Biol* 18(165).

[169] Merz C, et al. (2013) Replicate phylogenies and post-glacial range expansion of the pitcher-plant mosquito, *Wyeomyia smithii*, in north america. *PLoS One* 8:e72262.

[170] Folt B, Garrison N, Guyer C, Rodriguez J, Bond J (2016) Phylogeography and evolution of the red salamander (*Pseudotriton ruber*). *Mol. Phylogenet. Evol* 98:97–110.

[171] Magni C, Ducousso A, Caron H, Petit R, Kremer A (2005) Chloroplast dna variation of *Quercus rubra* l. in north america and comparison with other fagaceae. *Mol. Ecol* 14:513–524.

[172] McLachlan J, Clark J, Manos P (2005) Molecular indicators of tree migration capacity under rapid climate change. *Ecology* 86:2088–2098.

[173] Jackson S, et al. (2000) Vegetation and environment in eastern north america during the last glacial maximum. *Quat. Sci. Rev* 19:489–508.

[174] Rowe K, Heske E, Brown P, Paige K (2004) Surviving the ice: northern refugia and postglacial colonization. *Proc. Natl. Acad. Sci. U. S. A* 101:10355–10359.

[175] Bennett K, Provan J (2008) What do we mean by "refugia"? *Quat. Sci. Rev* 27:2449–2455.

[176] Bartlein PJ, et al. (1998) Paleoclimate simulations for north america over the past 21,000 years: features of the simulated climate and comparisons with paleoenvironmental data. *Quat. Sci. Rev* 17:549–585.

[177] LaMoreaux H, Brook G, Knox J (2009) Late pleistocene and holocene environments of the southeastern united states from the stratigraphy and pollen content of a peat deposit on the georgia coastal plain. *Palaeogeogr. Palaeoclimatol. Palaeoecol* 280:300–312.

[178] Grimm E, et al. (2006) Evidence for warm wet heinrich events in florida. *Quat. Sci. Rev* 25:2197–2211.

[179] Hyseni C, Garrick R (2019) Ecological drivers of species distributions and niche overlap for three subterranean termite species in the southern appalachian mountains, usa. *Insects* 10(33).

[180] Garrick R (2011) Montane refuges and topographic complexity generate and maintain invertebrate biodiversity: Recurring themes across space and time. *J. Insect Conserv* 15:469–478.

[181] Everaerts C, et al. (2008) The *Cryptocercus punctulatus* species complex (dictyoptera: Cryptocercidae) in the eastern united states: comparison of cuticular hydrocarbons, chromosome number, and dna sequences. *Mol. Phylogenet. Evol* 47:950–959.

[182] Nason J, Hamrick J, Fleming T (2002) Historical vicariance and postglacial colonization effects on the evolution of genetic structure in *Lophocereus*, a sonoran desert columnar cactus. *Evolution* 56:2214–2226.

[183] Buckley D (2009) Toward an organismal, integrative, and iterative phylogeography. *Bioessays* 31:784–793.

[184] Fontanella F, Feldman C, Siddall M, Burbrink F (2008) Phylogeography of *Diadophis punctatus*: extensive lineage diversity and repeated patterns of historical demography in a trans-continental snake. *Mol. Phylogenet. Evol* 46:1049–1070.

[185] Kennedy P, et al. (2017) Deconstructing superorganisms and societies to address big questions in biology. *Trends Ecol. Evol* 32(11):861–872.

[186] Allis CD, Jenuwein T (2016) The molecular hallmarks of epigenetic control. *Nat. Rev. Genet* 17(8):487–500.

[187] Verhoeven KJF, Jansen JJ, van Dijk PJ, Biere A (2010) Stress-induced DNA methylation changes and their heritability in asexual dandelions. *New Phytol* 185(4):1108–1118.

[188] Weinhold A (2018) Transgenerational stress-adaption: an opportunity for ecological epigenetics. *Plant Cell Rep* 37(1):3–9.

[189] Huang X, et al. (2017) Rapid response to changing environments during biological invasions: DNA methylation perspectives. *Mol. Ecol* 26(23):6621–6633.

[190] Sheldon EL, Schrey A, Andrew SC, Ragsdale A, Griffith SC (2018) Epigenetic and genetic variation among three separate introductions of the house sparrow (*Passer domesticus*) into australia. *R. Soc. Open Sci* 5(4):172–185.

[191] Manfredini F, Arbetman M, Toth AL (2019) A potential role for phenotypic plasticity in invasions and declines of social insects. *Front. Ecol. Evol* 7:375.

[192] Marin P, et al. (2020) Biological invasion: the influence of the hidden side of the (epi) genome. *Funct. Ecol* 34(2):385–400.

[193] Chouvenc T, Helmick EE, Su NY (2015) Hybridization of two major termite invaders as a consequence of human activity. *PLoS One* 10(3):e0120745.

[194] Saino N, et al. (2017) Migration phenology and breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. *Sci. Rep* 7:45412.

[195] Evans TA, Forschler BT, Grace JK (2013) Biology of invasive termites: a worldwide review. *Annu. Rev. Entomol* 58:455–474.

[196] Near TJ, et al. (2011) Phylogeny and temporal diversification of darters (percidae: Etheostomatinae). *Syst. Biol* 60(5):565–595.

[197] Hyseni C, Garrick RC (2019) The role of glacial-interglacial climate change in shaping the genetic structure of eastern subterranean termites in the southern appalachian mountains, usa. *Ecol. Evol* 9:4621–4636.

[198] Nowacki GJ, Abrams MD (2008) The demise of fire and "mesophication" of forests in the eastern united states. *Bioscience* 58(2):123–138.

[199] Hanberry BB, Brzuszek RF, Foster HT, Schauwecker TJ (2019) Recalling open old growth forests in the southeastern mixed forest province of the united states. *Écoscience* 26(1):11–22.

[200] Hanberry BB, Bragg DC, Hutchinson TF (2018) A reconceptualization of open oak and pine ecosystems of eastern north america using a forest structure spectrum. *Ecosphere* 9(10):e02431.

[201] Pan Y, et al. (2011) Age structure and disturbance legacy of north american forests. *Biogeosciences* 8:715–732.

[202] Valéry L, Fritz H, Lefeuvre JC, Simberloff D (2009) Invasive species can also be native.... *Trends Ecol. Evol.* 24(11):585.

[203] Carey MP, Sanderson BL, Barnas KA, Olden JD (2012) Native invaders–challenges for science, management, policy, and society. *Front. Ecol. Environ.* 10(7):373–381.

[204] Simberloff D, Souza L, Nuñez MA, Barrios-Garcia MN, Bunn W (2012) The natives are restless, but not often and mostly when disturbed. *Ecology* 93(3):598–607.

[205] Valéry L, Fritz H, Lefeuvre JC (2013) Another call for the end of invasion biology. *Oikos* 122(8):1143–1146.

[206] Buczkowski G, Wang C, Bennett G (2007) Immunomarking reveals food flow and feeding relationships in the eastern subterranean termite, *Reticulitermes flavipes* (kollar). *Environ. Entomol.* 36(1):173–182.

[207] Reyna-López GE, Simpson J, Ruiz-Herrera J (1997) Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. *Mol. Gen. Genet* 253(6):703–710.

[208] Zhang MS, et al. (2007) Endosperm-specific hypomethylation, and meiotic inheritance and variation of DNA methylation level and pattern in sorghum (*Sorghum bicolor* l.) inter-strain hybrids. *Theor. Appl. Genet* 115(2):195–207.

[209] Schulz B, Eckstein L, Durka W (2013) Scoring and analysis of methylation-sensitive amplification polymorphisms for epigenetic population studies. *Mol. Ecol. Resour* 13:642–653.

[210] R Core Team (2019) *R: a language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).

[211] Sponsler RC, Appel AG (1990) Aspects of the water relations of the formosan and eastern subterranean termites (isoptera: Rhinotermitidae). *Environ. Entomol* 19(1):15–20.

[212] Hyseni C, Garrick RC (2020) Data from: The role of glacial-interglacial climate change in shaping the genetic structure of eastern subterranean termites in the southern appalachian mountains, usa. *Dryad*.

[213] Schmitt S, Pouteau R, Justeau D, de Boissieu F, Birnbaum P (2017) Ssdm: an r package to predict distribution of species richness and endemism based on stacked species distribution models. *Methods Ecol. Evol* 8(12):1795–1803.

[214] Ripley B (1996) *Pattern Recognition and Neural Networks*. (Cambridge University Press).

[215] Friedman J (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat* 29:1189–1232.

[216] Breiman L (2001) Random forests. *Mach. Learn* 45:5–32.

[217] Liu X, de Sherbinin A, Zhan Y (2019) Mapping urban extent at large spatial scales using machine learning methods with VIIRS nighttime light and MODIS daytime NDVI data. *Remote Sens* 11(10):1247.

[218] Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol* 6(8):925–929.

[219] Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics*.

[220] Schwarz G (1978) Estimating the dimension of a model. *Ann. Stat* 6(2):461–464.

[221] Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology* 26(3):297–302.

[222] Sorensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of vegetation on danish commons. *Biol. Skr* 5:1–34.

[223] Games PA, Howell JF (1976) Pairwise multiple comparison procedures with unequal n's and/or variances: a monte carlo study. *J. Educ. Behav. Stat* 1(2):113–125.

[224] Peters GJY (2018) *userfriendlyscience: Quantitative analysis made accessible*. R package version 0.7.2.

[225] Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol* 30(7):1687–1699.

[226] Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J. Stat. Softw* 67(1):1–48.

[227] Dyer JM (2001) Using witness trees to assess forest change in southeastern ohio. *Can. J. For. Res* 31(10):1708–1718.

[228] Fei S, Steiner KC (2009) Rapid capture of growing space by red maple. *Can. J. For. Res* 39(8):1444–1452.

[229] Eck MA, Perry LB, Soulé PT, Sugg JW, Miller DK (2019) Winter climate variability in the southern appalachian mountains, 1910-2017. *Int. J. Climatol* 39(1):206–217.

[230] Herrera CM, Bazaga P (2016) Genetic and epigenetic divergence between disturbed and undisturbed subpopulations of a mediterranean shrub: a 20-year field experiment. *Ecol. Evol* 6(11):3832–3847.

[231] Buchli H (1958) L'origine des castes et les potentialités ontogéniques des termites européens du genre *Reticulitermes* holmgren. *Ann. Sci. Nat. Zool* 20:263–429.

[232] Yanagihara S, Suehiro W, Mitaka Y, Matsuura K (2018) Age-based soldier polyethism: old termite soldiers take more risks than young soldiers. *Biol. Lett* 14(3).

[233] Herrera CM, Bazaga P (2010) Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet *Viola cazorlensis*. *New Phytol* 187(3):867–876.

[234] Foust CM, et al. (2016) Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. *Mol. Ecol* 25(8):1639–1652.

[235] Jones DT, et al. (2003) Termite assemblage collapse along a land-use intensification gradient in lowland central sumatra, indonesia. *J. Appl. Ecol* 40(2):380–391.

[236] Dahlsjö CAL, Valladares Romero CS, Espinosa Iñiguez CI (2020) Termite diversity in ecuador: a comparison of two primary forest national parks. *J. Insect Sci* 20(1).

[237] Emery NC, Ackerly DD (2014) Ecological release exposes genetically based niche variation. *Ecol. Lett* 17(9):1149–1157.

[238] Sexton JP, Montiel J, Shay JE, Stephens MR, Slatyer RA (2017) Evolution of ecological niche breadth. *Annu. Rev. Ecol. Evol. Syst* 48(1):183–206.

[239] Catford JA, Bode M, Tilman D (2018) Introduced species that overcome life history tradeoffs can cause native extinctions. *Nat. Commun* 9(1):2131.

[240] Hufbauer RA, et al. (2012) Anthropogenically induced adaptation to invade (AIAI): contemporary adaptation to human-altered habitats within the native range can promote invasions. *Evol. Appl* 5(1):89–101.

[241] Simberloff D (2011) Native invaders in *Encyclopedia of Biological Invasions*, eds. Simberloff D, Rejmánek M. (University of California Press, Berkeley and Los Angeles, CA, USA).

[242] Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ (2017) Evolution of DNA methylation across insects. *Mol. Biol. Evol* 34(3):654–665.

[243] Matsuura K, et al. (2009) Queen succession through asexual reproduction in termites. *Science* 323(5922):1687.

[244] Vargo EL, Labadie PE, Matsuura K (2012) Asexual queen succession in the subterranean termite *Reticulitermes virginicus*. *Proc. Biol. Sci* 279(1729):813–819.

[245] Luchetti A, Velonà A, Mueller M, Mantovani B (2013) Breeding systems and reproductive strategies in italian *Reticulitermes* colonies (isoptera: Rhinotermitidae). *Insectes Soc* 60(2):203–211.

[246] Matsuura K (2017) Evolution of the asexual queen succession system and its underlying mechanisms in termites. *J. Exp. Biol* 220:63–72.

[247] Taylor PD, Fahrig L, Henein K, Merriam G (1993) Connectivity is a vital element of landscape structure. *Oikos* 68(3):571–573.

[248] Tischendorf L, Fahrig L (2000) On the usage and measurement of landscape connectivity. *Oikos* 90(1):7–19.

[249] Auffret AG, Plue J, Cousins SAO (2015) The spatial and temporal components of functional connectivity in fragmented landscapes. *Ambio* 44:51–59.

[250] Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol* 18(4):189–197.

[251] Storfer A, et al. (2007) Putting the "landscape" in landscape genetics. *Heredity* 98(3):128–142.

[252] Manel S, Holderegger R (2013) Ten years of landscape genetics. *Trends Ecol. Evol* 28(10):614–621.

[253] Wright S (1943) Isolation by distance. *Genetics* 28(2):114–138.

[254] McRae BH (2006) Isolation by resistance. *Evolution* 60(8):1551–1561.

[255] McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci. U. S. A* 104(50):19885–19890.

[256] Spear SF, Balkenhol N, Fortin MJ, McRae BH, Scribner K (2010) Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Mol. Ecol* 19(17):3576–3591.

[257] Peterman WE (2018) ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods Ecol. Evol* 9(6):1638–1647.

[258] Wang IJ, Summers K (2010) Genetic structure is correlated with phenotypic divergence rather than geographic isolation in the highly polymorphic strawberry poison-dart frog. *Mol. Ecol.* 19(3):447–458.

[259] Wang IJ, Bradburd GS (2014) Isolation by environment. *Mol. Ecol* 23:5649–5662.

[260] Wang IJ (2013) Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution* 67(12):3403–3411.

[261] Wright S (1931) Evolution in mendelian populations. *Genetics* 16(2):97–159.

[262] Nei M (1972) Genetic distance between populations. *Am. Nat* 106(949):283–292.

[263] Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. models and estimation procedures. *Am. J. Hum. Genet* 19:233–257.

[264] Dyer RJ (2015) Population graphs and landscape genetics. *Annu. Rev. Ecol. Evol. Syst* 46(1):327–342.

[265] Dyer RJ, Nason JD (2004) Population graphs: the graph theoretic shape of genetic structure. *Mol. Ecol* 13(7):1713–1727.

[266] Greenbaum G, Templeton AR, Bar-David S (2016) Inference and analysis of population structure using genetic data and network theory. *Genetics* 202(4):1299–1312.

[267] Dyer RJ, Nason JD, Garrick RC (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Mol. Ecol* 19(17):3746–3759.

[268] Wiens JA (1989) Spatial scaling in ecology. *Funct. Ecol* 3(4):385–397.

[269] Wilson DS (1992) Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. *Ecology* 73(6):1984–2000.

[270] Leibold MA, et al. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett* 7:601–613.

[271] Sciaini M, Fritsch M, Scherer C, Simpkins CE (2018) Nlmr and landscapetools: an integrated environment for simulating and modifying neutral landscape models in r. *Methods Ecol. Evol* 9(11):2240–2248.

[272] Ralph P (2019) *landsim: simulate populations on landscapes*. R package version 0.0.0.1.

[273] Pielou EC (1966) The measurement of diversity in different types of biological collections. *J. Theor. Biol* 13:131–144.

[274] Shannon CE (1948) A mathematical theory of communication. *Bell Labs Tech. J* 27(3):379–423.

[275] Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? *Mol. Ecol* 19(17):3496–3514.

[276] Pflüger FJ, Balkenhol N (2014) A plea for simultaneously considering matrix quality and local environmental conditions when analysing landscape impacts on effective dispersal. *Mol. Ecol.* 23(9):2146–2156.

[277] Baguette M, Blanchet S, Legrand D, Stevens VM, Turlure C (2013) Individual dispersal, landscape connectivity and ecological networks. *Biol. Rev. Camb. Philos. Soc* 88(2):310–326.

[278] Storfer A, Patton A, Fraik AK (2018) Navigating the interface between landscape genetics and landscape genomics. *Front. Genet* 9:68.

[279] Fletcher RJ, Burrell NS, Reichert BE, Vasudev D, Austin JD (2016) Divergent perspectives on landscape connectivity reveal consistent effects from genes to communities. *Curr. Landsc. Ecol. Rep* 1(2):67–79.

[280] James PMA, Coltman DW, Murray BW, Hamelin RC, Sperling FAH (2011) Spatial genetic structure of a symbiotic beetle-fungal system: toward multi-taxa integrated landscape genetics. *PLoS One* 6(10):e25359.

[281] Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G (2015) Landscape community genomics: understanding eco-evolutionary processes in complex environments. *Trends Ecol. Evol* 30(3):161–168.

[282] Rissler LJ (2016) Union of phylogeography and landscape genetics. *Proc. Natl. Acad. Sci. U. S. A* 113(29):8079–8086.

[283] Phillips S, Anderson R, Schapire R (2006) Maximum entropy modeling of species geographic distributions. *Ecol. Modell* 190:231–259.

[284] Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G (2010) Uncertainty in ensemble forecasting of species distribution. *Glob. Chang. Biol* 16:1145–1157.

[285] Revelle W (2018) *psych: procedures for psychological, psychometric, and personality research*. R package version 1.8.10.

[286] Harman H, Jones W (1966) Factor analysis by minimizing residuals (minres). *Psychometrika* 31:351–368.

[287] Cattell R (1966) The scree test for the number of factors. *Multivariate Behav. Res* 1:245–276.

[288] Horn J (1965) A rationale and test for the number of factors in factor analysis. *Psychometrika* 30:179–185.

[289] Cronbach L (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334.

[290] Otto-Bliesner B, et al. (2006) Last glacial maximum and holocene climate in ccsm3. *J. Clim* 19:2526–2544.

[291] Feytaud J (1920) Sur les jeunes colonies du termite lucifuge. *C. R. Acad. Sci* 171:203–206.

[292] Beard R (1974) Termite biology and bait-block method of control. *Conn. Agric. Exp. Stn. Bull* 748:1–9.

[293] Truett GE, et al. (2000) Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). *Biotechniques* 29(1):52–54.

[294] Smith TW (2014) *binner: read fsa fragment files from an ABI Genetic Analyzer*. R package version 0.1.

[295] Whitlock R, Hipperson H, Mannarelli M, Butlin R, Burke T (2008) An objective, rapid and reproducible method for scoring aflp peak-height data that minimizes genotyping error. *Mol. Ecol. Resour* 8:725–735.

# LIST OF APPENDICES

## Appendix A:

# Chapter 1: Supplementary Material

Environmental variables and Ecological Niche Modeling methods Ecological Niche Models (ENMs) were constructed using the 'biomod2' package[71] in R[73]. To construct ENMs, in addition to presence records, we used pseudo-absence points, selected following Barbet-Massin et al.[130]. To do this, we first ran a rectilinear surface range envelope model[71], and then, from outside the area predicted as suitable habitat, we picked 100 random points. We created 20 independent sets of pseudo-absences, each of which were combined with the same 91 presence records. Four modeling algorithms were run: artificial neural networks[214], generalized boosted models or boosted regression trees[215], random forest[216], and maximum entropy[283]. We used 5 cross-validation runs per algorithm, for a total of 400 runs (4 algorithms x 5 cross-validations x 20 datasets), with 5,000 iterations per run. To assess model performance, 75% of the data were used for training, with 25% set aside as "out-of-bag" test data. To maximize the accuracy of presence/absence classification, we used the True Skill Statistic (TSS = sum of sensitivity and specificity – 1)[85], where ENMs with mean TSS above 0.2 were retained. We then used the ensemble framework[284] to obtain a weighted average of all ENMs, where ENMs were weighted according to TSS values. Nineteen bioclimatic variables[63] were obtained from the WorldClim database v.1.4 (`http://www.worldclim.org`). To reduce the number of predictors, and correlation among them, we performed factor analysis in successive stages using the 'psych' package[285], until two criteria were met: 1) each factor must be highly correlated (absolute value of $r > 0.5$) with at least two variables, and 2) each variable must be highly correlated with only one factor and show low correlation (absolute value of $r < 0.3$) with any other factor. We used ordinary least squares to find the minimum residual (MR) solution[286]. Oblique rotations were used, since strong correlations between factors were expected. Cattell's[287] scree test and Horn's[288] parallel analysis determined the number of factors to retain, and these were then inspected for reliability using Cronbach's[289] $\alpha$, with an acceptance criterion of $\alpha > 0.7$. The factors were named according to the bioclimatic variables they were most strongly correlated with. "Temperature Range" (TR; strongly correlated with bio4: "Temperature Seasonality" and bio7: "Temperature Annual Range"); "Dry-season Precipitation" (DP; strongly correlated with bio14: "Precipitation of Driest Month" and bio17: "Precipitation of Driest Quarter"); "Summer Temperature" (ST; strongly correlated with bio5: "Maximum Temperature of Warmest Month" and bio10: "Mean Temperature of Warmest Quarter"); "Wet-season Precipitation" (WP; strongly correlated with bio13: "Precipitation of Wettest Month" and bio17: "Precipitation of Wettest Quarter").

**Table A.1:** *Sampling sites with number of species occurrences at each site and number of logs per site.* Geographic coordinates and altitude (alt.) in meters for each site are reported. *R. flavipes*, *R. malletei*, and *R. virginicus* are abbreviated as Rf, Rm, and Rv, respectively. The number (#) of logs refers the number of logs sampled, from which termites were collected and identified to species (note that site 37 is the only site where two species were detected in the same log). Only non–redundant occurrence records were used for subsequent analyses.

| Site | Longitude | Latitude | Alt. (m) | Rf | Rm | Rv | # of Logs |
|---|---|---|---|---|---|---|---|
| 1 | −84.63805 | 34.77972 | 764 | 1 | 0 | 0 | 1 |
| 2 | −85.06536 | 34.57297 | 450 | 1 | 0 | 0 | 1 |
| 3 | −85.21630 | 34.64336 | 386 | 0 | 1 | 0 | 1 |
| 4 | −85.24268 | 34.56515 | 408 | 1 | 0 | 0 | 1 |
| 5 | −85.24043 | 34.56416 | 427 | 0 | 0 | 1 | 1 |
| 6 | −79.38618 | 38.82374 | 528 | 2 | 0 | 0 | 2 |
| 7 | −79.38506 | 38.82585 | 548 | 1 | 0 | 0 | 1 |
| 8 | −79.48494 | 38.72694 | 936 | 1 | 0 | 0 | 1 |
| 9 | −85.25067 | 34.54107 | 341 | 0 | 2 | 0 | 2 |
| 10 | −86.07185 | 33.20099 | 301 | 0 | 0 | 2 | 2 |
| 11 | −85.80658 | 33.47105 | 621 | 1 | 0 | 1 | 2 |
| 12 | −85.77732 | 33.49199 | 413 | 0 | 0 | 1 | 1 |
| 13 | −85.69289 | 33.57288 | 340 | 0 | 0 | 1 | 1 |
| 14 | −85.59404 | 33.70745 | 360 | 0 | 0 | 1 | 1 |
| 15 | −85.62832 | 33.67281 | 427 | 0 | 1 | 1 | 2 |
| 16 | −85.87318 | 33.40451 | 460 | 2 | 0 | 0 | 2 |
| 17 | −85.93159 | 33.36097 | 440 | 0 | 0 | 1 | 1 |
| 18 | −86.02572 | 33.33344 | 313 | 0 | 1 | 0 | 1 |
| 19 | −87.36352 | 34.23058 | 273 | 0 | 0 | 1 | 1 |
| 20 | −85.70074 | 33.56059 | 425 | 1 | 0 | 0 | 1 |
| 21 | −87.38140 | 34.29811 | 279 | 0 | 0 | 1 | 1 |
| 22 | −87.33273 | 34.41979 | 321 | 1 | 0 | 0 | 1 |
| 23 | −87.27680 | 34.17659 | 248 | 1 | 0 | 0 | 1 |

| Site | Longitude | Latitude | Alt. (m) | Rf | Rm | Rv | # of Logs |
|---|---|---|---|---|---|---|---|
| 24 | −85.58357 | 34.45540 | 395 | 1 | 0 | 0 | 1 |
| 25 | −85.59611 | 34.55167 | 526 | 0 | 1 | 0 | 1 |
| 26 | −85.67106 | 34.35716 | 392 | 0 | 1 | 0 | 1 |
| 27 | −85.45730 | 33.96340 | 300 | 1 | 0 | 0 | 1 |
| 28 | −85.84679 | 34.14676 | 188 | 1 | 0 | 0 | 1 |
| 29 | −85.26428 | 34.12260 | 232 | 1 | 0 | 0 | 1 |
| 30 | −85.81731 | 33.46215 | 485 | 1 | 0 | 0 | 1 |
| 31 | −84.71650 | 34.15014 | 272 | 0 | 0 | 1 | 1 |
| 32 | −83.10755 | 34.86200 | 536 | 2 | 0 | 0 | 2 |
| 33 | −83.05563 | 35.01376 | 887 | 1 | 0 | 0 | 1 |
| 34 | −83.08929 | 34.94523 | 744 | 1 | 0 | 0 | 1 |
| 35 | −83.12841 | 34.80557 | 481 | 0 | 0 | 1 | 1 |
| 36 | −83.22783 | 34.72782 | 394 | 2 | 1 | 0 | 3 |
| 37 | −83.31242 | 34.77755 | 469 | 1 | 0 | 1 | 1 |
| 38 | −83.29258 | 33.72088 | 132 | 2 | 0 | 2 | 4 |
| 39 | −86.07201 | 33.20150 | 291 | 1 | 0 | 0 | 1 |
| 40 | −84.71137 | 34.87866 | 354 | 1 | 0 | 0 | 1 |
| 41 | −84.65486 | 34.93135 | 485 | 0 | 1 | 0 | 1 |
| 42 | −84.33880 | 34.77507 | 730 | 1 | 0 | 0 | 1 |
| 43 | −84.25093 | 34.68311 | 810 | 1 | 0 | 0 | 1 |
| 44 | −83.73265 | 34.74192 | 766 | 0 | 0 | 1 | 1 |
| 45 | −83.51849 | 35.65682 | 780 | 1 | 0 | 0 | 1 |
| 46 | −83.35717 | 35.70232 | 653 | 1 | 0 | 0 | 1 |

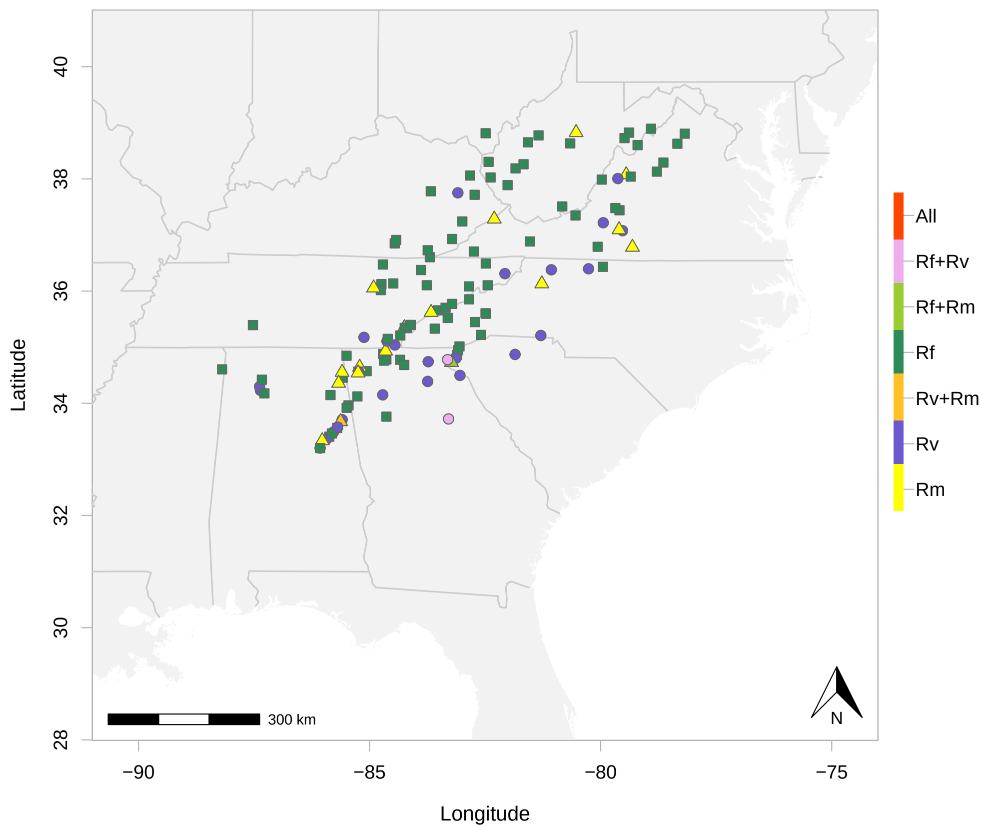



**Figure A.1:** *Map of* Reticulitermes *sampling depicting occurrences of one or more species at each site.* Abbreviations used for *R. flavipes*, *R. malletei*, and *R. virginicus* are Rf, Rm, and Rv, respectively. Sites are color coded based on the number of species detected. There were no sites with all three species (”All”). The sites with two species are shown in the legend as ”Rf + Rv,” ”Rf + Rm,” and ”Rv + Rm.”

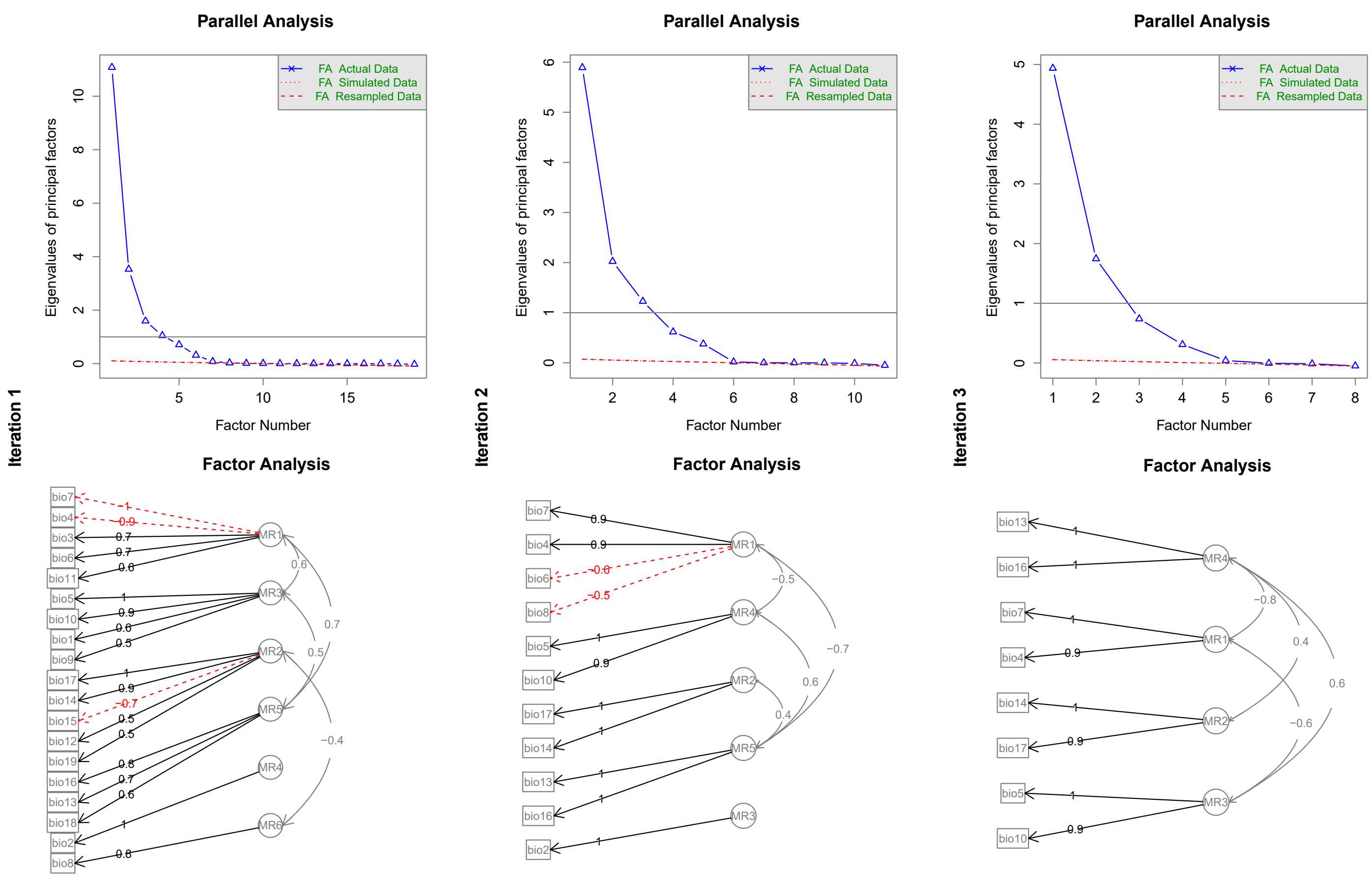



**Figure A.2:** *Factor analysis*. Each column of panels represents one of three iterations of factor analysis. The top row depicts scree plots showing eigenvalues in descending order, the traditional threshold where eigenvalue = 1, and the confidence interval (red dotted lines) obtained via parallel analysis. The bottom row shows the factors and strength of correlation with the original bioclimatic variables. In the third and final iteration, abbreviations are as follows: MR1: temperature range; MR2: dry-season precipitation; MR3: summer temperature; MR4: wet-season precipitation.

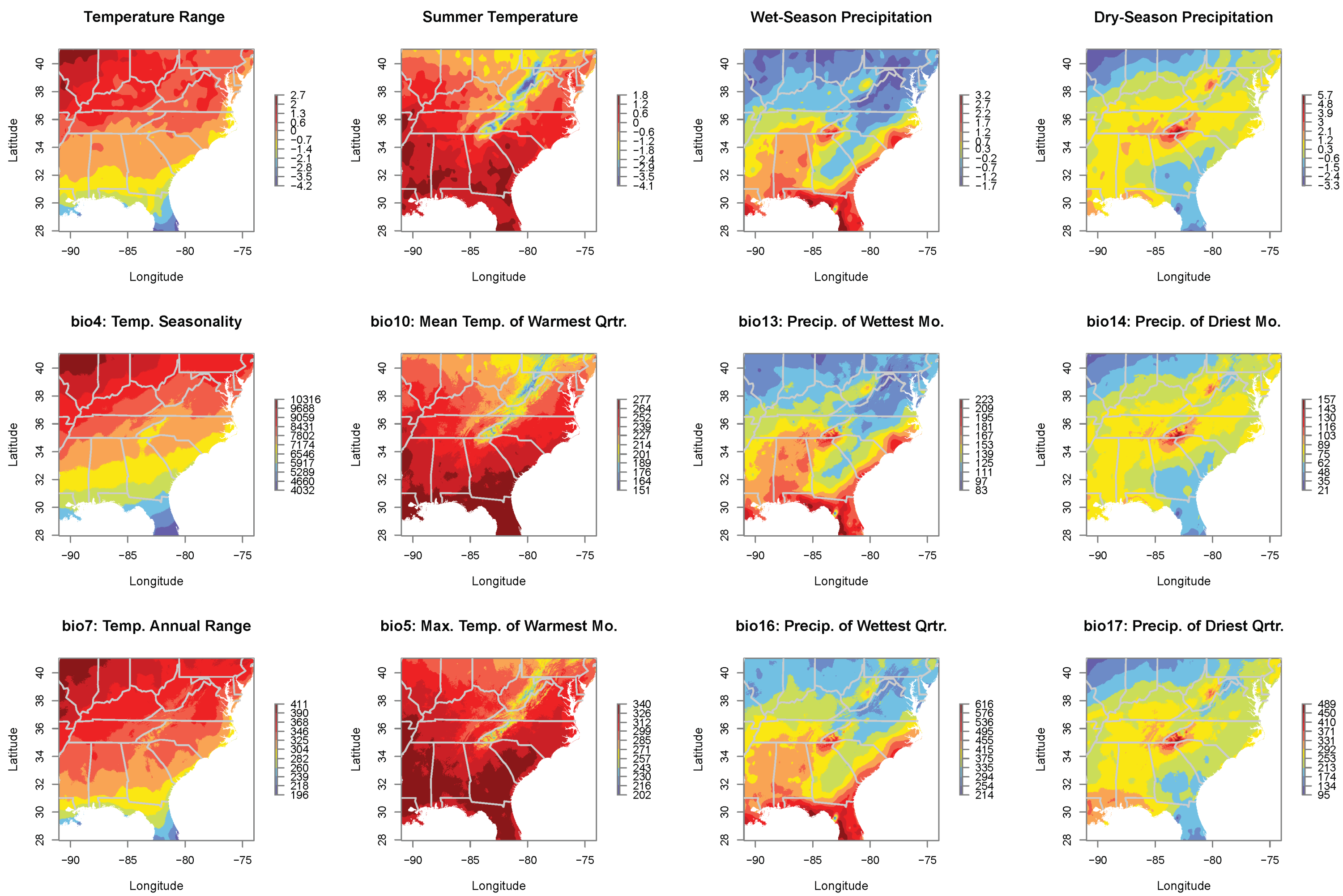



**Figure A.3:** *Environmental factors and bioclimatic variables*. The top row of panels shows the four environmental factors obtained via factor analysis (see Figure A.2). In each column of panels, the top panel shows the factor that explains the variation in the original bioclimatic variables, whereas the middle and bottom panels show the bioclimatic variables that correlate most strongly with the factor in the top panel. Note that the scales are different for each panel, but the colors go from dark blue (lowest value) to dark red (highest value). The environmental factors are unitless and go from negative to positive values. The unit for temperature variables is °C x 10. The unit for precipitation variables is mm.

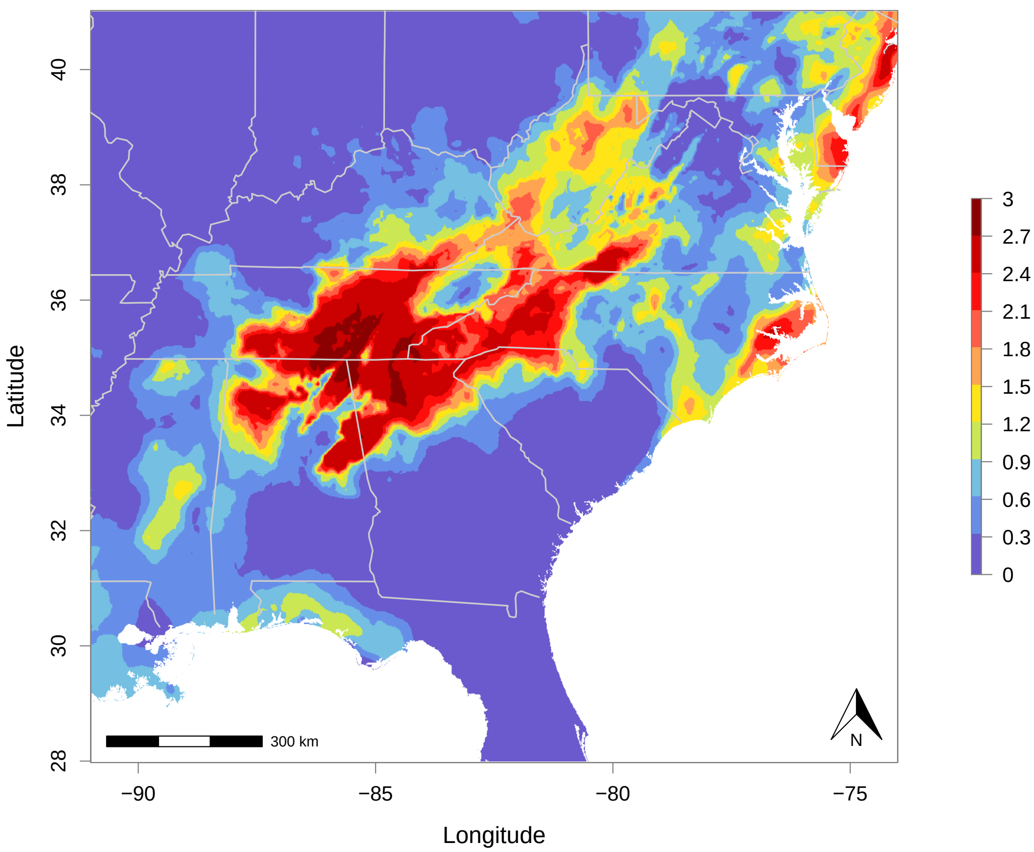



**Figure A.4:** *Distributional overlap of* Reticulitermes *species*. Overlap is depicted based on the sum of individual species' occurrence probabilities, with the highest value being 3 (dark red), where all three species co-occur at a probability of 1. Areas with occurrence probability above 1 (green to red) must have more than one species. Areas with probability below 1 (blues) could have more than one species with probabilities lower than 0.5. Absence of all three species is shown in dark blue.

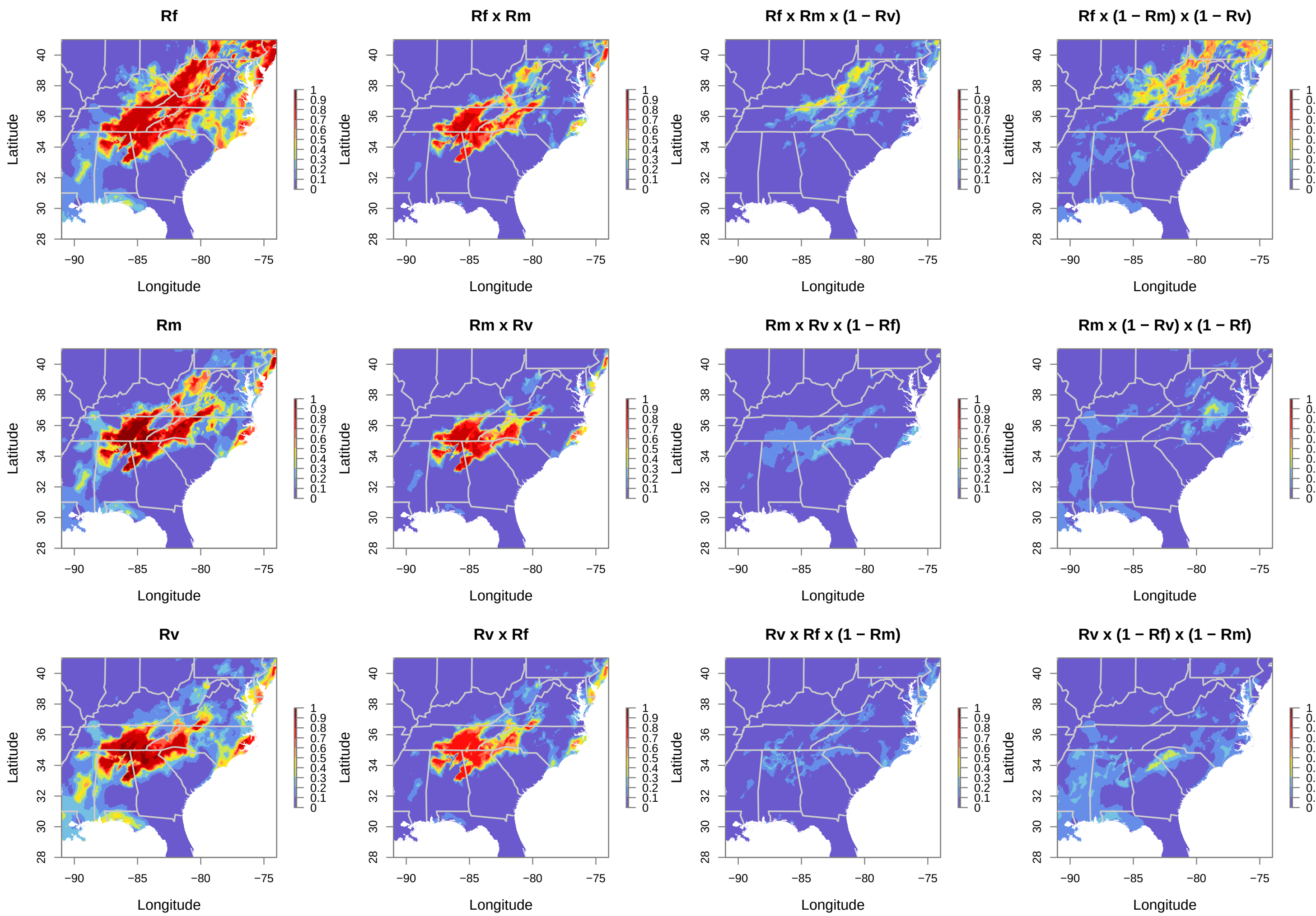



**Figure A.5:** *Probability of joint and exclusive occurrence of* Reticulitermes *species*. The leftmost column of panels shows probability of occurrence of *R. flavipes*, *R. malletei*, and *R. virginicus* (abbreviated as Rf, Rm, and Rv, respectively), whereas probability of absence is denoted as (1 – Rf), (1 – Rm), and (1 – Rv). Probability of occurrence is shown on a scale from 0 (dark blue) to 1 (dark red). The second column of panels shows probability of joint occurrence of two species (without excluding the third), expressed as products: "Rf x Rv," "Rf x Rm," and "Rv x Rm." The third column shows areas where two species co-occur, but the third species is absent (probability of absence: 1 – Rf, 1 – Rm, 1 – Rv). Probability of occurrence of a single species, while excluding the other two, is shown in the rightmost column.

Appendix B:

# Chapter 2: Supplementary Material

## 2.1 Supplementary Methods

### 2.1.1 Population sampling

Reticulitermes flavipes termites were collected for genetic analyses between 2012 and 2014 from 46 locations in the southern Appalachian Mountains. The presence of R. flavipes was confirmed at an additional 45 locations from 2015 to 2016. Since it is not possible to reliably distinguish among several co-distributed species on the basis of morphology when only members of the worker caste are collected[127], termites were identified using a molecular assay[54]. Briefly, a short (376-bp) region of the mitochondrial COII gene was amplified (using PCR primers RetCo2-F and RetCo2-R), and products were then separately digested with three restriction enzymes (RsaI, TaqI, and MspI), which in combination generate diagnostic species-specific banding patterns.

**Table B.1:** *Geographic locations from which* Reticulitermes flavipes *termites were sampled.* Each site has a unique ID, and associated state and county information is shown. Spatial coordinates are reported in decimal degrees, and elevation is in meters. Occurrence of *R. flavipes* was confirmed at 91 sites, and these were all used for Species Distribution Modeling. Genetic data were collected from individuals sampled from the first 46 sites.

| Site ID | State | County | Longitude | Latitude | Elevation | Genetic Data |
|---|---|---|---|---|---|---|
| A03 | Georgia | Gilmer | -84.6381 | 34.77972 | 764 | Yes |
| A04 | Georgia | Gordon | -85.0654 | 34.57297 | 450 | Yes |
| A09 | Georgia | Chattooga | -85.2427 | 34.56515 | 408 | Yes |
| A13 | Alabama | Cleburne | -85.7007 | 33.56059 | 425 | Yes |
| A14 | Alabama | Clay | -85.8173 | 33.46215 | 485 | Yes |
| A16 | Alabama | Clay | -86.072 | 33.2015 | 291 | Yes |
| A18 | Georgia | Murray | -84.7114 | 34.87866 | 354 | Yes |
| A21 | Georgia | Gilmer | -84.3388 | 34.77507 | 730 | Yes |
| A22 | Georgia | Fannin | -84.2509 | 34.68311 | 810 | Yes |

| Site ID | State | County | Longitude | Latitude | Elevation | Genetic Data |
|---|---|---|---|---|---|---|
| A30 | Tennessee | Sevier | -83.5185 | 35.65682 | 780 | Yes |
| A31 | Tennessee | Sevier | -83.3572 | 35.70232 | 653 | Yes |
| A32 | North Carolina | Swain | -83.3108 | 35.52117 | 666 | Yes |
| A37 | Tennessee | Cocke | -83.2134 | 35.7714 | 575 | Yes |
| A40 | Georgia | Murray | -84.6914 | 34.75931 | 804 | Yes |
| A41 | Alabama | Cleburne | -85.4976 | 33.91858 | 257 | Yes |
| A52 | Virginia | Giles | -80.5451 | 37.34757 | 1121 | Yes |
| A56 | Virginia | Albemarle | -78.7837 | 38.12902 | 814 | Yes |
| A60 | Virginia | Greene | -78.6431 | 38.29123 | 761 | Yes |
| A62 | Virginia | Rappahannock | -78.1815 | 38.80508 | 755 | Yes |
| A64 | Virginia | Madison | -78.3406 | 38.62592 | 1032 | Yes |
| A70 | Virginia | Augusta | -79.3498 | 38.04052 | 784 | Yes |
| A73 | Tennessee | Lawrence | -87.5268 | 35.39384 | 304 | Yes |
| A75 | Tennessee | Morgan | -84.7448 | 36.12452 | 378 | Yes |
| A76 | Tennessee | Morgan | -84.4883 | 36.13606 | 496 | Yes |
| A85 | Georgia | Dade | -85.4997 | 34.84695 | 315 | Yes |
| A86 | Mississippi | Tishomingo | -88.193 | 34.60502 | 177 | Yes |
| A87 | Tennessee | Monroe | -84.2476 | 35.34883 | 327 | Yes |
| A88 | Tennessee | Monroe | -84.1938 | 35.34534 | 425 | Yes |
| A92 | North Carolina | Swain | -83.5919 | 35.32969 | 593 | Yes |
| A97 | North Carolina | Buncombe | -82.4874 | 35.59535 | 722 | Yes |
| A106 | West Virginia | Pendleton | -79.3862 | 38.82374 | 528 | Yes |
| A107 | West Virginia | Pendleton | -79.3851 | 38.82585 | 548 | Yes |
| A108 | West Virginia | Pendleton | -79.4849 | 38.72694 | 936 | Yes |
| A117 | Alabama | Clay | -85.8066 | 33.47105 | 621 | Yes |
| A124 | Alabama | Clay | -85.8732 | 33.40451 | 460 | Yes |
| A131 | Alabama | Lawrence | -87.3327 | 34.41979 | 321 | Yes |
| A133 | Alabama | Winston | -87.2768 | 34.17659 | 248 | Yes |
| A134 | Alabama | DeKalb | -85.5836 | 34.4554 | 395 | Yes |
| A137 | Alabama | Cherokee | -85.4573 | 33.9634 | 300 | Yes |
| A138 | Alabama | Etowah | -85.8468 | 34.14676 | 188 | Yes |
| A139 | Georgia | Floyd | -85.2643 | 34.1226 | 232 | Yes |
| A141 | South Carolina | Oconee | -83.1076 | 34.862 | 536 | Yes |
| A142 | South Carolina | Jackson | -83.0556 | 35.01376 | 887 | Yes |
| A143 | South Carolina | Oconee | -83.0893 | 34.94523 | 744 | Yes |
| A145 | South Carolina | Oconee | -83.2278 | 34.72782 | 394 | Yes |
| A146 | South Carolina | Oconee | -83.3124 | 34.77755 | 469 | Yes |
| A150 | Georgia | Greene | -83.2926 | 33.72088 | 132 | No |
| T1 | Virginia | Scott | -82.7454 | 36.70494 | 460 | No |
| T2 | Virginia | Botetourt | -79.6821 | 37.47978 | 759 | No |

| Site ID | State | County | Longitude | Latitude | Elevation | Genetic Data |
|---|---|---|---|---|---|---|
| T3 | Virginia | Smyth | -81.5317 | 36.88458 | 731 | No |
| T4 | Virginia | Patrick | -80.0662 | 36.78941 | 386 | No |
| T6 | North Carolina | Rockingham | -79.9509 | 36.43191 | 256 | No |
| T10 | Virginia | Bedford | -79.5934 | 37.4409 | 692 | No |
| T12 | Virginia | Bath | -79.9771 | 37.98723 | 508 | No |
| T13 | West Virginia | Pendleton | -79.2019 | 38.60225 | 591 | No |
| T15 | West Virginia | Hardy | -78.9102 | 38.89373 | 589 | No |
| T16 | Ohio | Gallia | -82.4906 | 38.81387 | 277 | No |
| T17 | Tennessee | Morgan | -84.7587 | 36.01773 | 572 | No |
| T19 | Tennessee | Scott | -84.7143 | 36.47398 | 479 | No |
| T20 | Kentucky | McCreary | -84.4575 | 36.84983 | 412 | No |
| T21 | Kentucky | McCreary | -84.4248 | 36.91024 | 336 | No |
| T22 | Tennessee | Knox | -83.764 | 36.10415 | 399 | No |
| T23 | Tennessee | Union | -83.8904 | 36.37519 | 490 | No |
| T24 | Kentucky | Bell | -83.6973 | 36.60349 | 352 | No |
| T25 | Kentucky | Bell | -83.7441 | 36.72807 | 390 | No |
| T26 | Kentucky | Harlan | -83.2143 | 36.92808 | 767 | No |
| T27 | Tennessee | Sullivan | -82.4869 | 36.49101 | 427 | No |
| T29 | Kentucky | Knott | -82.9939 | 37.24096 | 318 | No |
| T31 | Georgia | Douglas | -84.6363 | 33.76154 | 295 | No |
| T32 | North Carolina | Buncombe | -82.4913 | 35.60575 | 770 | No |
| T33 | North Carolina | Henderson | -82.7176 | 35.44758 | 1205 | No |
| T34 | North Carolina | Henderson | -82.5896 | 35.21877 | 809 | No |
| T35 | Tennessee | Monroe | -84.2412 | 35.34314 | 413 | No |
| T36 | Tennessee | Monroe | -84.112 | 35.39665 | 553 | No |
| T37 | Tennessee | Polk | -84.3359 | 35.20793 | 513 | No |
| T39 | Tennessee | Polk | -84.6082 | 35.14822 | 588 | No |
| T46 | North Carolina | Madison | -82.8472 | 35.85284 | 656 | No |
| T47b | Tennessee | Greene | -82.8497 | 36.08371 | 408 | No |
| T48 | Tennessee | Unicoi | -82.4466 | 36.10384 | 522 | No |
| T55 | Kentucky | Powell | -83.6773 | 37.77913 | 256 | No |
| T57 | Kentucky | Floyd | -82.7283 | 37.71582 | 213 | No |
| T58 | Kentucky | Lawrence | -82.8253 | 38.05997 | 209 | No |
| T59 | West Virginia | Wayne | -82.4262 | 38.30313 | 186 | No |
| T60 | West Virginia | Wayne | -82.3832 | 38.02512 | 402 | No |
| T61 | West Virginia | Logan | -82.0147 | 37.88885 | 260 | No |
| T62 | West Virginia | Lincoln | -81.8428 | 38.18754 | 201 | No |
| T63 | West Virginia | Kanawha | -81.6695 | 38.26121 | 267 | No |
| T64 | West Virginia | Jackson | -81.5756 | 38.652 | 241 | No |
| T65 | West Virginia | Roane | -81.3447 | 38.77533 | 240 | No |

| Site ID | State | County | Longitude | Latitude | Elevation | Genetic Data |
|---|---|---|---|---|---|---|
| T66 | West Virginia | Braxton | -80.6589 | 38.63269 | 394 | No |
| T68 | West Virginia | Summers | -80.8312 | 37.50894 | 550 | No |

**Table B.2:** *Geographic locations from which* Reticulitermes *out-group taxa were sampled.* Site ID and associated state and county information is shown Spatial coordinates are reported in decimal degrees, and elevation is in meters.

| Site ID | State | County | Longitude | Latitude | Elevation | Species |
|---|---|---|---|---|---|---|
| A06 | Georgia | Walker | -85.2163 | 34.64336 | 386 | *R. malletei* |
| A12 | Alabama | Cleburne | -85.6939 | 33.57157 | 328 | *R. nelsonae* |
| A10 | Georgia | Chattooga | -85.2404 | 34.56416 | 427 | *R. virginicus* |
| A146 | South Carolina | Oconee | -83.3124 | 34.77755 | 469 | *R. virginicus* |
| A25 | Georgia | White | -83.7327 | 34.74192 | 766 | *R. virginicus* |

### 2.1.2 DNA isolation and genetic markers

Mitochondrial cytochrome c oxidase subunit I (COI) and II (COII) genes, and an intronic portion of the nuclear endo-beta-1,4-glucanase (EB14G) gene, were targeted. Each of these DNA regions were amplified separately via Polymerase Chain Reaction (PCR) in 15 μL volumes containing 5 to 50 ng of genomic DNA, 5 picomoles of each of two primers (Table B.3), and the following amounts of Promega (Madison, WI) reagents: 0.8 nanomoles of each dNTP, 32 nanomoles of $MgCl_2$, 0.5 units of GoTaq, and 5 μg of bovine serum albumin, in a 1x final concentration of PCR buffer. Reactions were performed in a Bio-Rad (Hercules, CA) T100 Thermal Cycler with the following conditions: initial denaturation at 95 °C for 3 min, 35 cycles of 95°C for 30 s, 52°C for 30 s, and 72°C for 1 min, followed by a final extension at 72°C for 5 mins. PCR products were viewed following agarose gel electrophoresis and cleaned with ExoSAP-IT (USB, Cleveland, OH).

**Table B.3:** *Primer sequences and locus information.* Primer sequences and their sources are reported here, including length of quality-filtered, trimmed mitochondrial (mtDNA) and nuclear (nDNA) sequence alignments measured in base pairs (bp).

| | Gene region | Primers | | | Alignment |
|---|---|---|---|---|---|
| | | Name | Sequence | Source | |
| mtDNA | COI | LCO-1490 | 5'- GGTCAACAAATCATAAAGATATTGG-3' | Folmer et al., 1994 | 563 bp |
| | | HCO-2198 | 5'- TAAACTTCAGGGTGACCAAAAAATCA -3' | Folmer et al., 1994 | |
| | COII | CO2-forward | 5'- AGAGCWTCACCTATTATAGAAC-3' | Park et al., 2004 | 554 bp |
| | | TK-N-3785 | 5'- GTTTAAGAGACCAGTACTTG -3' | Simon et al., 1994 | |
| nDNA | EB14G | Ret_EB14G_F | 5'-ATGGAGGTCGCAGCTACGTC-3' | This study | 251 bp |
| | | Ret_EB14G_R | 5'-GGCGCTGTTGTACGTGTTCCAG-3' | This study | |

### 2.1.3 Construction of species distribution models

#### 2.1.3.1 Model evaluation and calibration

We used the 'biomod2' package[71,72] in R for Species Distribution Model (SDM) construction. We used presence records, and pseudo-absence points selected following Barbet-Massin et al.[130], who showed that for machine learning methods it is better to use multiple replicates of pseudo-absence points, with the number of pseudo-absences in each replicate close to the number of occurrence points. Thus, we first ran a rectilinear surface range envelope model[71], and then, from outside the area predicted as suitable habitat, we picked 100 random points, creating 20 independent sets of pseudo-absences, each of which were combined with the same 91 presence records. Four modeling algorithms were run: artificial neural networks[214], generalized boosted models or boosted regression trees[215], random forest[216], and maximum entropy[283]. We used 5 cross-validation runs per algorithm, for a total of 400 runs (4 algorithms x 5 cross-validations x 20 datasets), with 5,000 iterations per run. To assess model performance, 75% of the data were used for training, with 25% set aside as "out-of-bag" test data. To maximize the accuracy of presence/absence classification, we used the True Skill Statistic (TSS = sum of sensitivity and specificity – 1)[85], where SDMs with mean TSS above 0.2 were retained. We then used the ensemble framework[284] to obtain a weighted average of all SDMs, where SDMs were weighted according to TSS values.

#### 2.1.3.2 Climate data

Present-day SDMs were based on mean climatological data spanning a period from 1960–1990, with all variables used at 1-km resolution. Historical distributions were modeled for the Mid-Holocene (MH; 6 thousand years ago, kya), the Last Glacial Maximum (LGM, 22 kya), and the Last Interglacial (LIG, 120–140 kya). For each period, 19 bioclimatic variables[63] were obtained from the WorldClim database v.1.4 (`http://www.worldclim.org`). Using the 1960–1990 climatological data as the baseline, MH and LGM paleoclimatic data were downscaled from simulations with Global Climate Models, from CMIP5 (`http://cmip-pcmdi.llnl.gov/cmip5`). LIG paleoclimatic data were downscaled from Otto-Bliesner et al.[290].

#### 2.1.3.3 Factor analysis

To reduce the number of predictors, and correlation among them, we performed factor analysis in successive stages using the 'psych' package[285], until two criteria were met: 1) each factor must be highly correlated (absolute value of r > 0.5) with at least two variables, and 2) each variable must be highly correlated with only one factor and show low correlation (absolute value of r < 0.3) with any other factor. We used ordinary least squares to find the minimum residual (MR) solution[286]. Oblique rotations were used, since strong correlations between factors were expected. Cattell's[287] scree test and Horn's[288] parallel analysis determined the number of factors to retain, and these were then inspected for reliability using Cronbach's[289] $\alpha$, with an acceptance criterion of $\alpha > 0.7$.

#### 2.1.3.4 Factor names

MR1: "Temperature Range" (TR; strongly correlated with bio4: "Temperature Seasonality" and bio7: "Temperature Annual Range"); MR2: "Dry-season Precipitation" (DP; strongly correlated with bio14: "Precipitation of Driest Month" and bio17: "Precipitation of Driest Quarter"); MR3: "Summer Temperature" (ST; strongly correlated with bio5: "Maximum Temperature of Warmest Month" and bio10: "Mean Temperature of Warmest Quarter"); MR4: "Wet-season Precipitation" (WP; strongly correlated with bio13: "Precipitation of Wettest Month" and bio17: "Precipitation of Wettest Quarter").

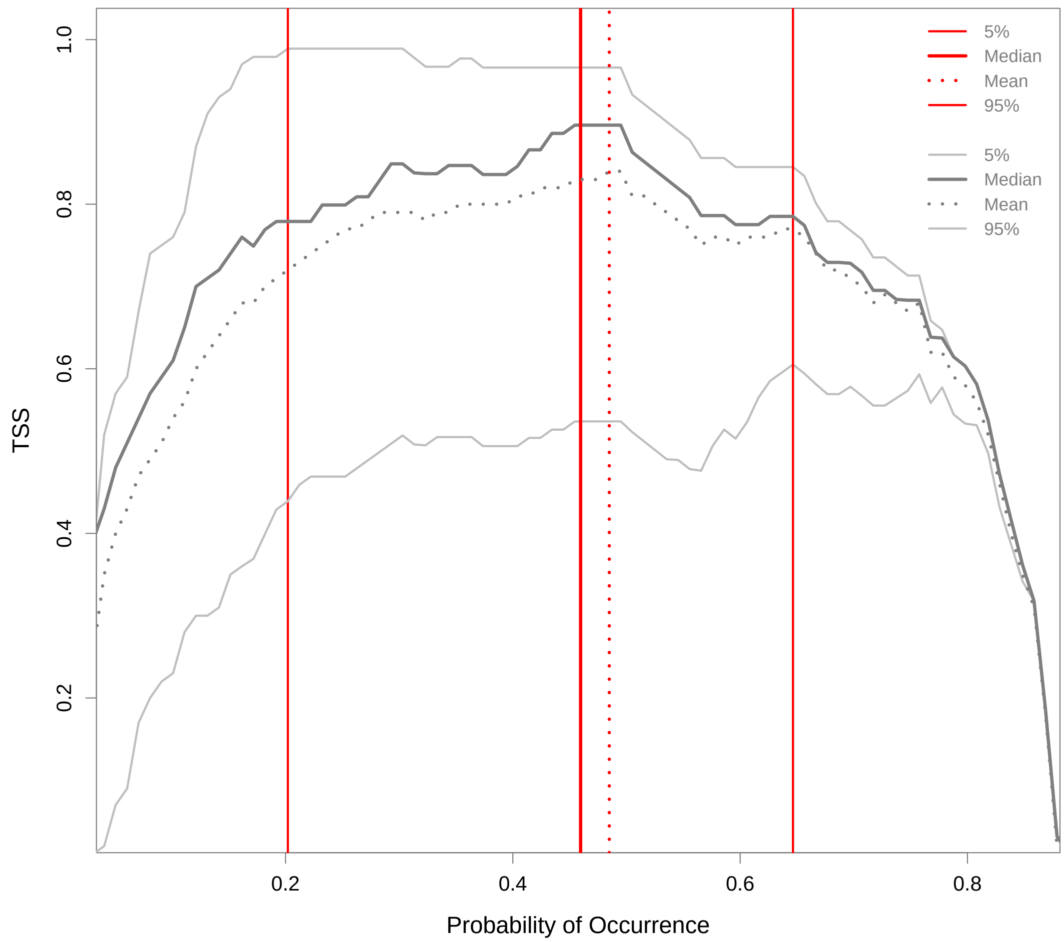



**Figure B.1:** *Optimal probability of occurrence threshold for conversion to binary presence/absence.* For each probability of occurrence value, True Skill Statistic (TSS; equal to the sum of sensitivity and specificity – 1) was calculated based on 91 occurrence records and 100 pseudo-absence points. We computed confidence intervals using 20 pseudo-absence replicates.

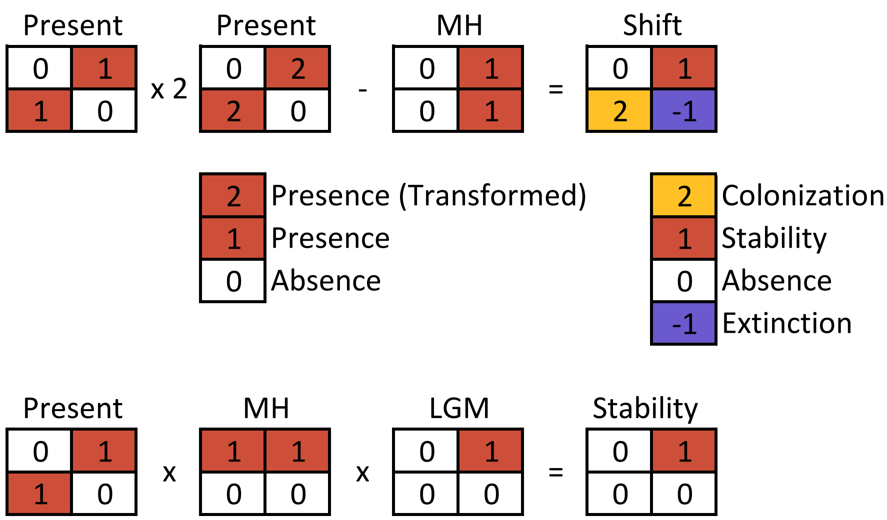



**Figure B.2:** *Schematic of distributional shift and stability calculations.* Occurrence probability was converted to binary occurrence (0 = absence; 1 = presence) based on a threshold of 0.2. To calculate the distributional shift from the Mid-Holocene (MH) to the present, we took the difference of the two, after multiplying the binary occurrence map for the present by 2. This multiplication ensures that we obtain four categories in the distributional shift calculation: colonization (difference = 2), stability (1), absence (0), and extinction (-1). To calculate stability across several time periods, we multiplied the binary occurrence maps. The Last Glacial Maximum is abbreviated as LGM.

**Table B.4:** *Environmental data by category (precipitation and temperature).* The bioclimatic variables shown here represent four different periods: present-day, Mid-Holocene, Last Glacial Maximum, and Last Interglacial. All data were obtained from WorldClim 1.4. All environmental variables have been scaled to 1-km resolution.

| Category | Environmental Variable | Code | Abbreviation |
|---|---|---|---|
| Precipitation | Annual Precipitation | BioClim 12 | bio12 |
| | Precipitation of Wettest Month | BioClim 13 | bio13 |
| | Precipitation of Driest Month | BioClim 14 | bio14 |
| | Precipitation Seasonality | BioClim 15 | bio15 |
| | Precipitation of Wettest Quarter | BioClim 16 | bio16 |
| | Precipitation of Driest Quarter | BioClim 17 | bio17 |
| | Precipitation of Warmest Quarter | BioClim 18 | bio18 |
| | Precipitation of Coldest Quarter | BioClim 19 | bio19 |
| Temperature | Annual Mean Temperature | BioClim 1 | bio1 |
| | Mean Diurnal Range | BioClim 2 | bio2 |
| | Isothermality | BioClim 3 | bio3 |
| | Temperature Seasonality | BioClim 4 | bio4 |
| | Max. Temperature of Warmest Month | BioClim 5 | bio5 |
| | Min Temperature of Coldest Month | BioClim 6 | bio6 |
| | Temperature Annual Range | BioClim 7 | bio7 |
| | Mean Temperature of Wettest Quarter | BioClim 8 | bio8 |
| | Mean Temperature of Driest Quarter | BioClim 9 | bio9 |
| | Mean Temperature of Warmest Quarter | BioClim 10 | bio10 |
| | Mean Temperature of Coldest Quarter | BioClim 11 | bio11 |

### 2.1.4 Comparison of scenarios using approximate Bayesian computation

To identify the best-fit model, we used approximate Bayesian computation (ABC; Beaumont et al. 2002), implemented in DIYABC v.2.1.0[148]. Within the ABC framework, two classes of model parameters were used to characterize the phylogeographic hypotheses described above: effective population sizes (Ne), and divergence times (T). We performed two rounds of modeling: 1) the preliminary round with broad priors, and 2) the final round with narrower priors. Based on posterior probabilities from the preliminary round, for the Northern and Southern clusters, we used uniform priors of Ne = 25,000–250,000. For the Central cluster, posterior probabilities from the preliminary round were not informative for narrowing Ne range, so we used a broad log-uniform prior of Ne = 500,000–5,000,000. All competing scenarios had two divergence events: any two of $T_N$, $T_C$ or $T_S$, (the subscript is the first letter abbreviation of the new cluster, i.e., Northern, Central, or Southern). The prior range for the more recent event encompassed the Mid-Holocene (MH) and the Last Glacial Maximum (LGM) (i.e., T = 2,000–25,000 years ago), while the priors of the older event ranged from the LGM to the Last Interglacial (LIG) (i.e., T = 20,000–120,000). Given the overlap between these divergence time priors, we enforced a condition such that the latter event was required to occur before the former. *Reticulitermes flavipes* colonies produce alates once a year, approximately two years after colony foundation[291], but colonies can grow to 70 individuals in their first year[292]. Thus, we assumed a 1-year generation time. We included brief bottlenecks (1–10 generations duration) at the beginning of each divergence event, in order to mimic founder events.

**Table B.5:** *ABC priors.* N, C, and S represent the effective population sizes of the Northern, Central, and Southern clusters. $T_N$, $T_C$, and $T_S$ represent the time of divergence of N, C, and S. The parameters $b_N$, $b_C$, and $b_S$ represent duration (number of generations) of bottleneck events, whereas $N_b$, $C_b$, and $S_b$ represent effective population sizes during bottleneck events. In the vicariance scenario (see Figure B.5), $N_{Anc}$ and $T_{SN}$ are the effective population size before divergence, and time of divergence of the ancestor of S and N. The parameters $\mu_{mt}$ and $\mu_{nuc}$ are mutation rates of the mtDNA and nDNA loci.

| *Parameter* | *Distribution* | *Minimum* | *Maximum* |
|---|---|---|---|
| N | Uniform | 25,000 | 250,000 |
| C | Log-Uniform | 500,000 | 5,000,000 |
| S | Uniform | 25,000 | 250,000 |
| $T_N$ or $T_S$ or $T_{SN}$ | Uniform | 20,000 | 120,000 |
| $T_C$ | Uniform | 2,000 | 25,000 |
| $b_N$ | Uniform | 1 | 10 |
| $b_C$ | Uniform | 1 | 10 |
| $b_S$ | Uniform | 1 | 10 |
| $N_b$ | Log-Uniform | 500 | 50,000 |
| $C_b$ | Log-Uniform | 100 | 10,000 |
| $S_b$ | Log-Uniform | 500 | 50,000 |
| $N_{Anc}$ | Log-Uniform | 5,000 | 500,000 |
| $\mu_{mt}$ | Uniform | $5 \times 10^{-9}$ | $5 \times 10^{-7}$ |
| $\mu_{nuc}$ | Uniform | $5 \times 10^{-10}$ | $2.5 \times 10^{-8}$ |

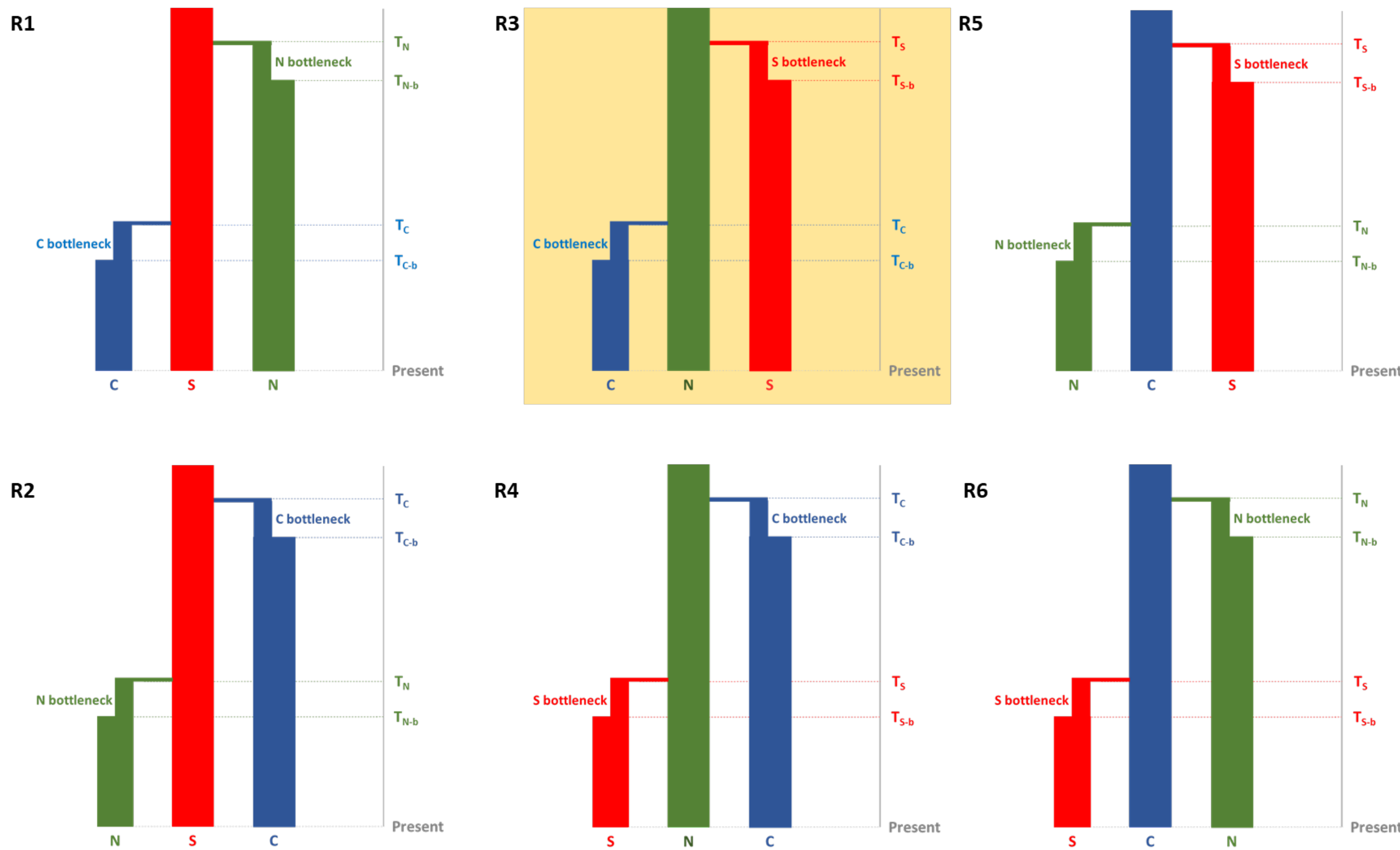



**Figure B.3:** *Refugial scenarios*. Scenarios compared in the first step of the first tier of ABC analyses. These "refugial scenarios" involved persistence in a single refugium, such that the other areas were colonized via successive expansions out of that refugium. We considered three refugial locations: Southern (S) = red, Northern (N) = green, and Central (C) = blue.

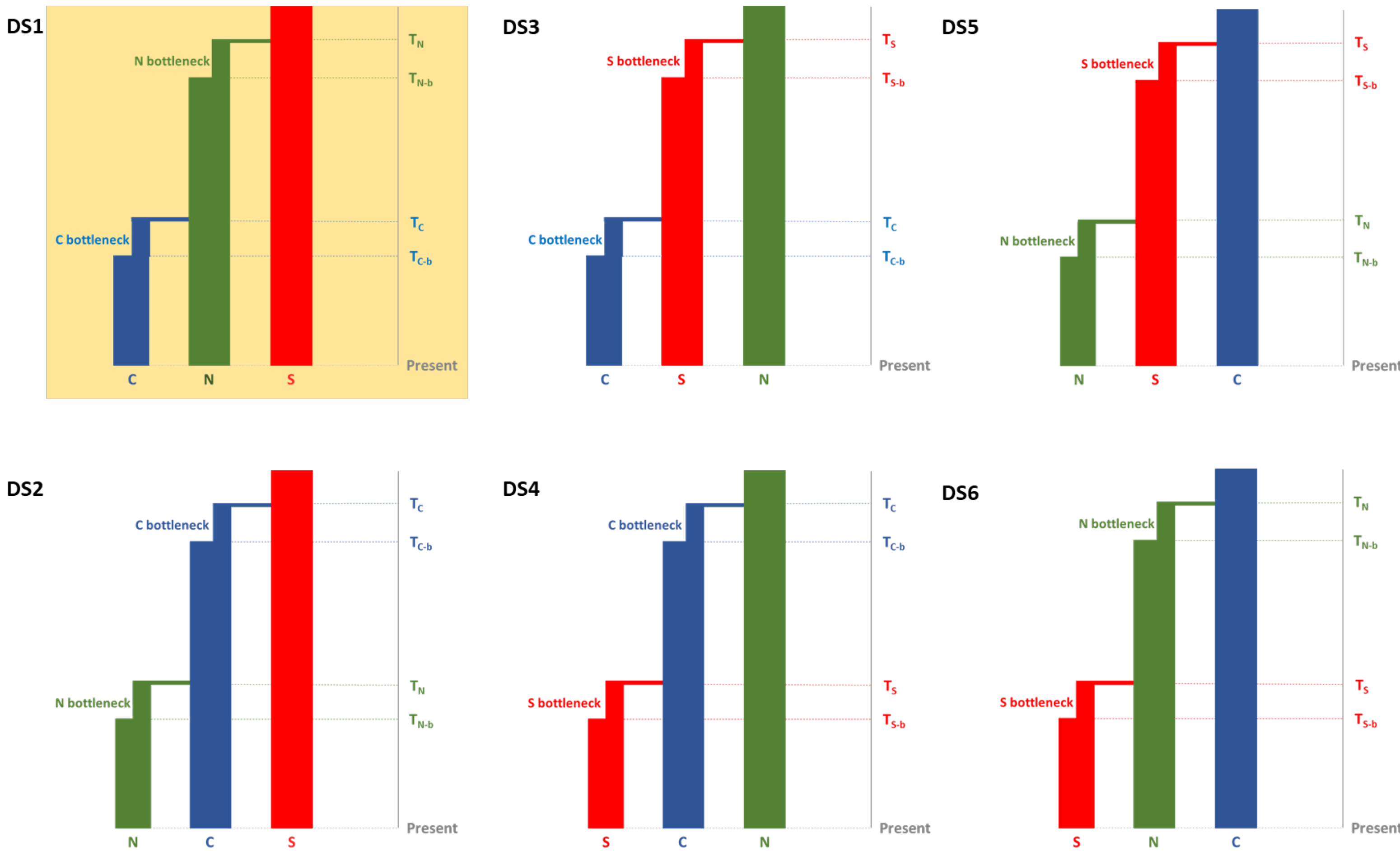



**Figure B.4:** *Distributional shift scenarios.* Scenarios compared in the second step of the first tier of ABC analyses. "Distributional shift" scenarios involved divergence in a stepping-stone fashion, where one population gave rise to a descendant population, which later became the progenitor of the third population. The Southern (S) cluster is shown in red, the Northern (N) in green, and the Central (C) in blue.

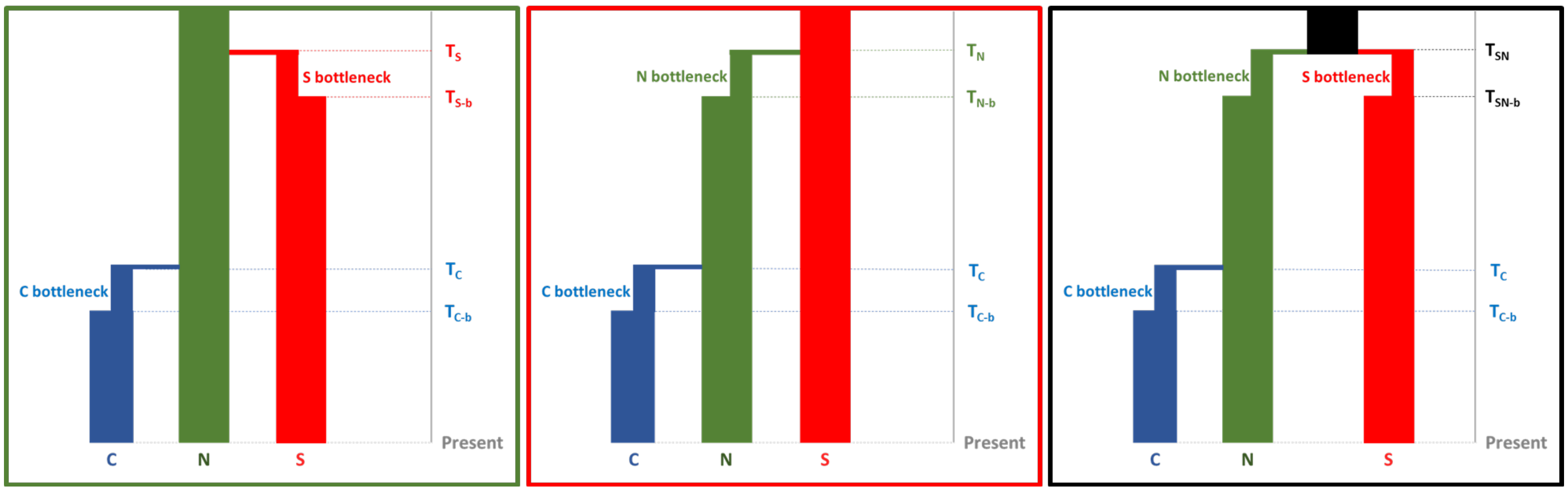



**Figure B.5:** *Alternative scenarios in the second tier of ABC hypothesis testing.* All three of these scenarios involve the Central (C) population diverging from the Northern (N) population. In the refugial scenario (R3; left panel), first the Southern (S) cluster, then the Central cluster, diverged from the Northern cluster (i.e., the primary refugium). In the distributional shift scenario (DS1; middle panel), N diverged from S, and then C diverged from N in a stepping-stone fashion. The vicariance scenario (V; right panel) involves the separation of an ancestral population into S and N, followed by C diverging from N.

## 2.2 Supplementary Results

### 2.2.1 Environmental factors used in species distribution models

**Table B.6:** *Correlations among environmental factors.* The table shows Pearson correlation coefficients among four environmental factors (MR) in each of four time periods: present-day, Mid-Holocene (MH), Last Glacial Maximum (LGM), and Last Interglacial (LIG).

**Correlation of Environmental Factors**

| Present | MR1 | MR2 | MR3 |
|---|---|---|---|
| MR2 | -0.29 | | |
| MR3 | -0.55 | 0.04 | |
| MR4 | -0.82 | 0.38 | 0.60 |

| MH | MR1 | MR2 | MR3 |
|---|---|---|---|
| MR2 | -0.28 | | |
| MR3 | -0.03 | 0.24 | |
| MR4 | 0.34 | -0.14 | -0.61 |

| LGM | MR1 | MR2 | MR3 |
|---|---|---|---|
| MR2 | 0.13 | | |
| MR3 | 0.30 | 0.39 | |
| MR4 | 0.70 | -0.28 | 0.04 |

| LIG | MR1 | MR2 | MR3 |
|---|---|---|---|
| MR2 | 0.36 | | |
| MR3 | -0.49 | -0.16 | |
| MR4 | 0.88 | 0.41 | -0.72 |

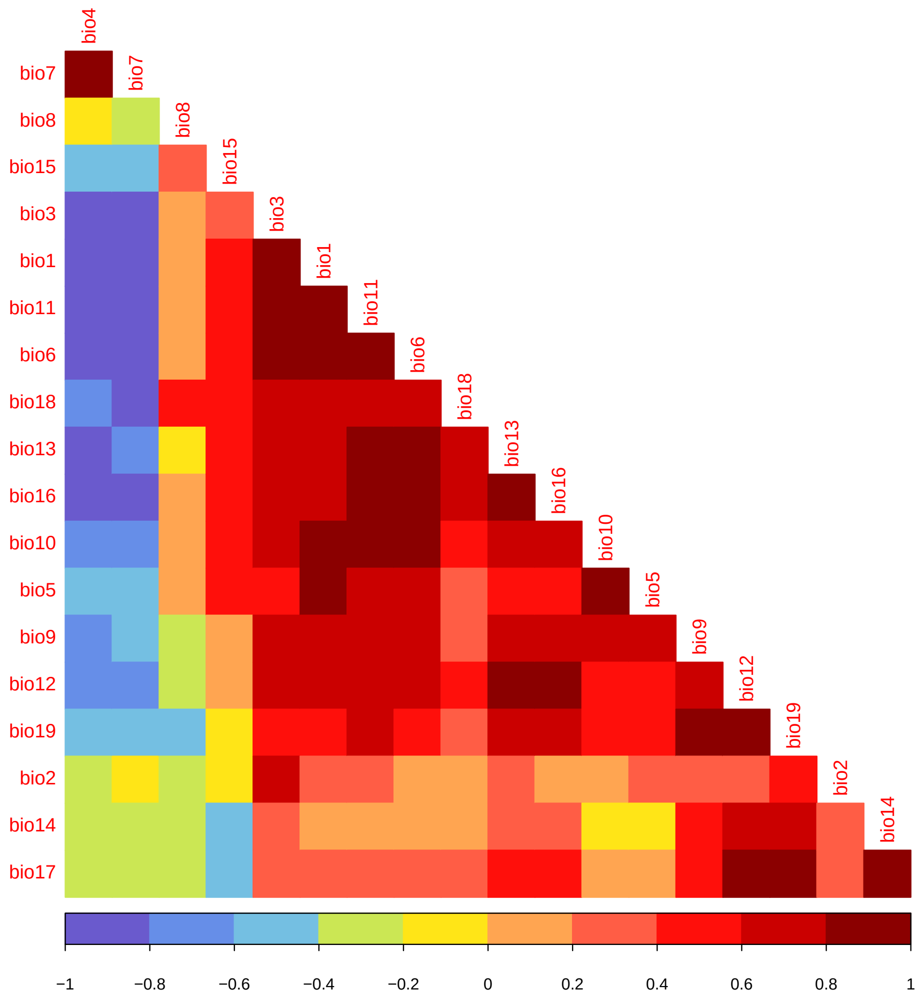



**Figure B.6:** *Pearson correlation among 19 bioclimatic variables*. The plot of correlation coefficients (color-coded as a heat map, with strong positive correlation shown in red vs. negative in blue) among 19 bioclimatic (bio) variables, representing the "present" (1960–1990).

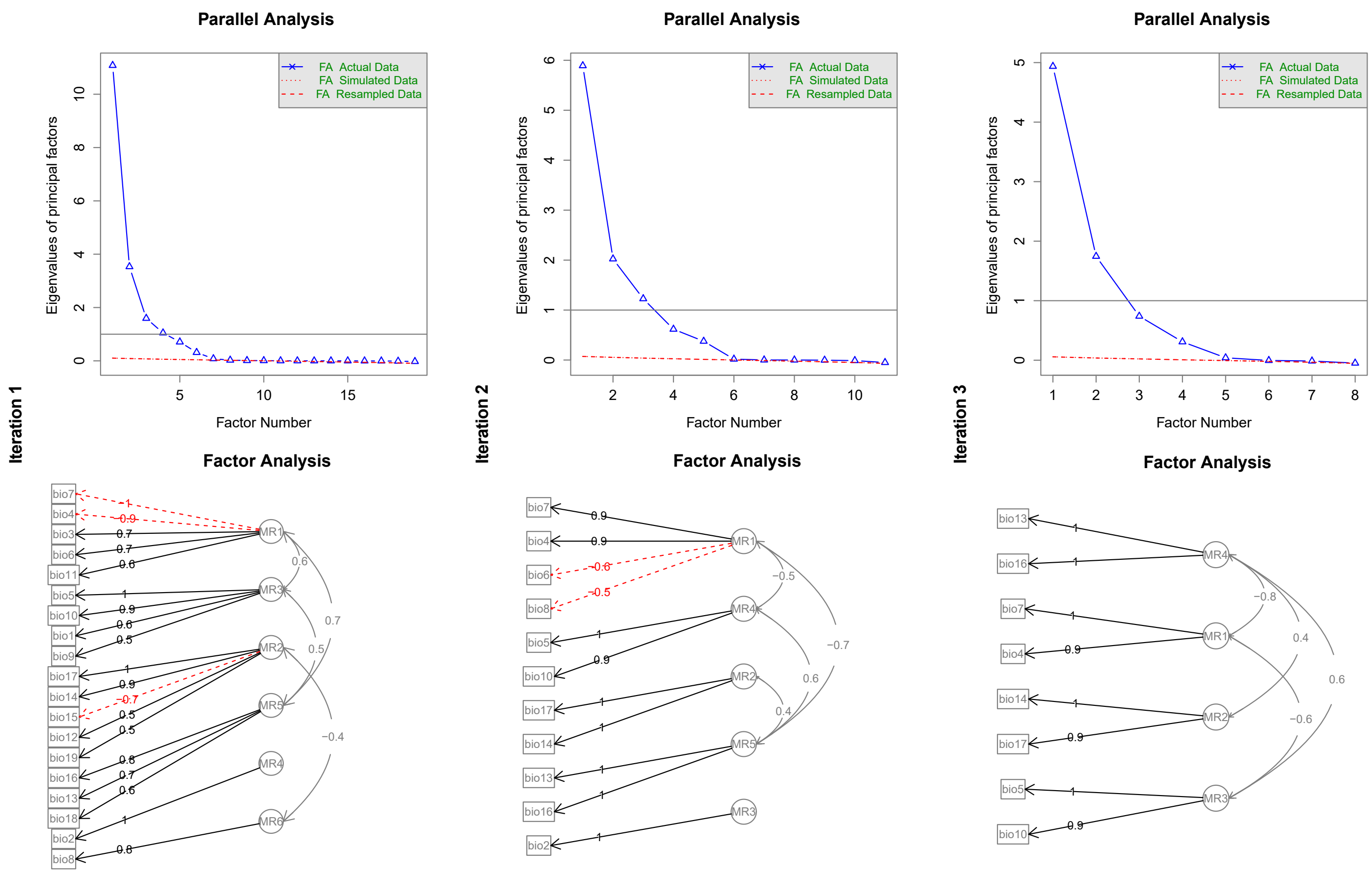



**Figure B.7:** *Factor analysis*. The results shown here are for the present. Each column of panels represents one of three iterations of factor analysis. The top row depicts scree plots showing eigenvalues in descending order, the traditional threshold where eigenvalue = 1, and the confidence interval (red dotted lines) obtained via parallel analysis. The bottom row shows the factors and strength of correlation with the original bioclimatic variables. In the third and final iteration, abbreviations are as follows: MR1: temperature range; MR2: dry-season precipitation; MR3: summer temperature; MR4: wet-season precipitation.

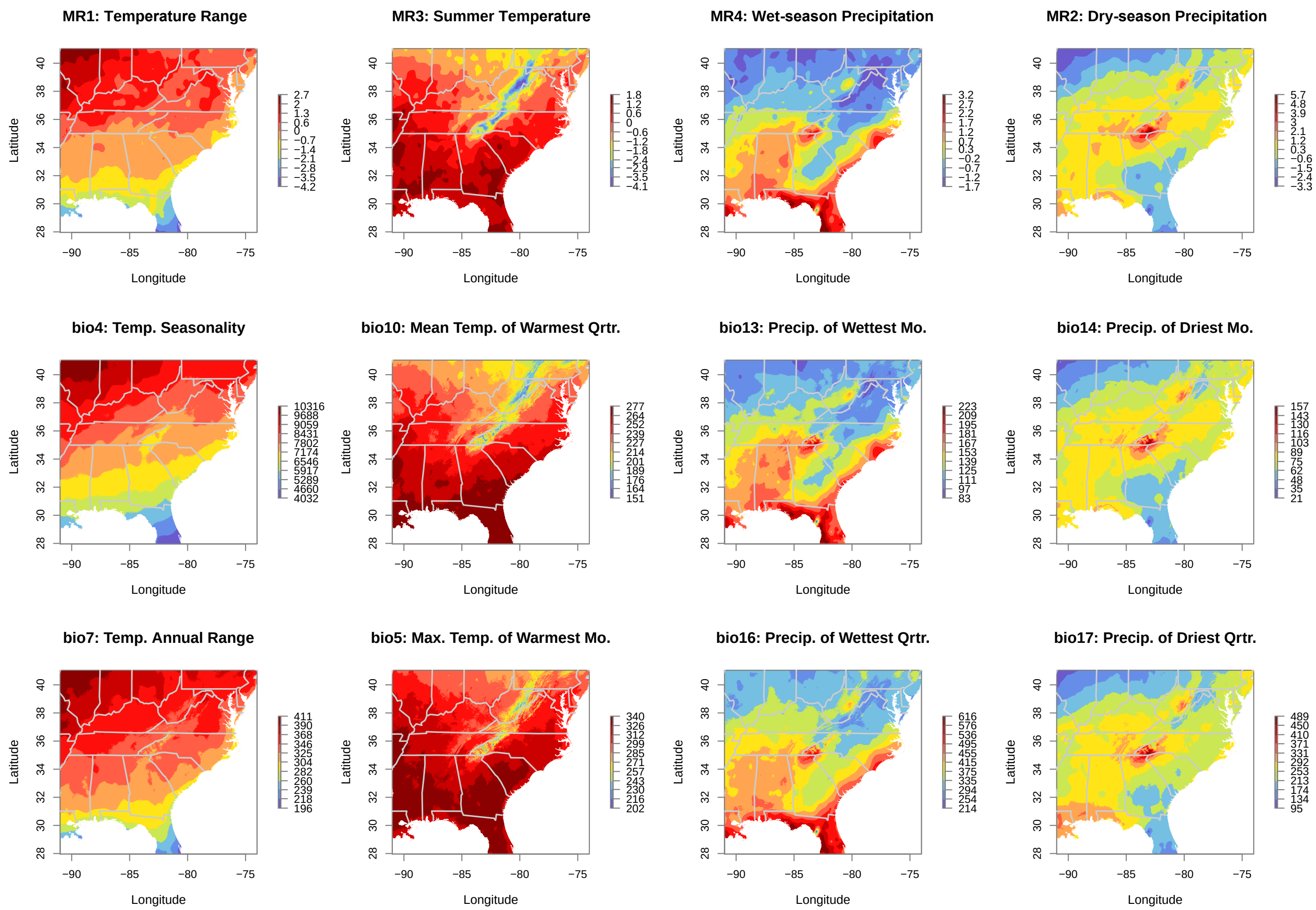



**Figure B.8:** *Environmental factors and bioclimatic variables.* The top row of panels shows the four environmental factors obtained via factor analysis. In each column of panels, the top panel shows the factor that explains the variation in the original bioclimatic variables, whereas the middle and bottom panels show the bioclimatic variables that correlate most strongly with the factor in the top panel. The scales are different for each panel, but the colors go from dark blue (lowest value) to dark red (highest value). The environmental factors are unitless and go from negative to positive values. The unit for temperature variables is °C x 10. The unit for precipitation variables is mm.

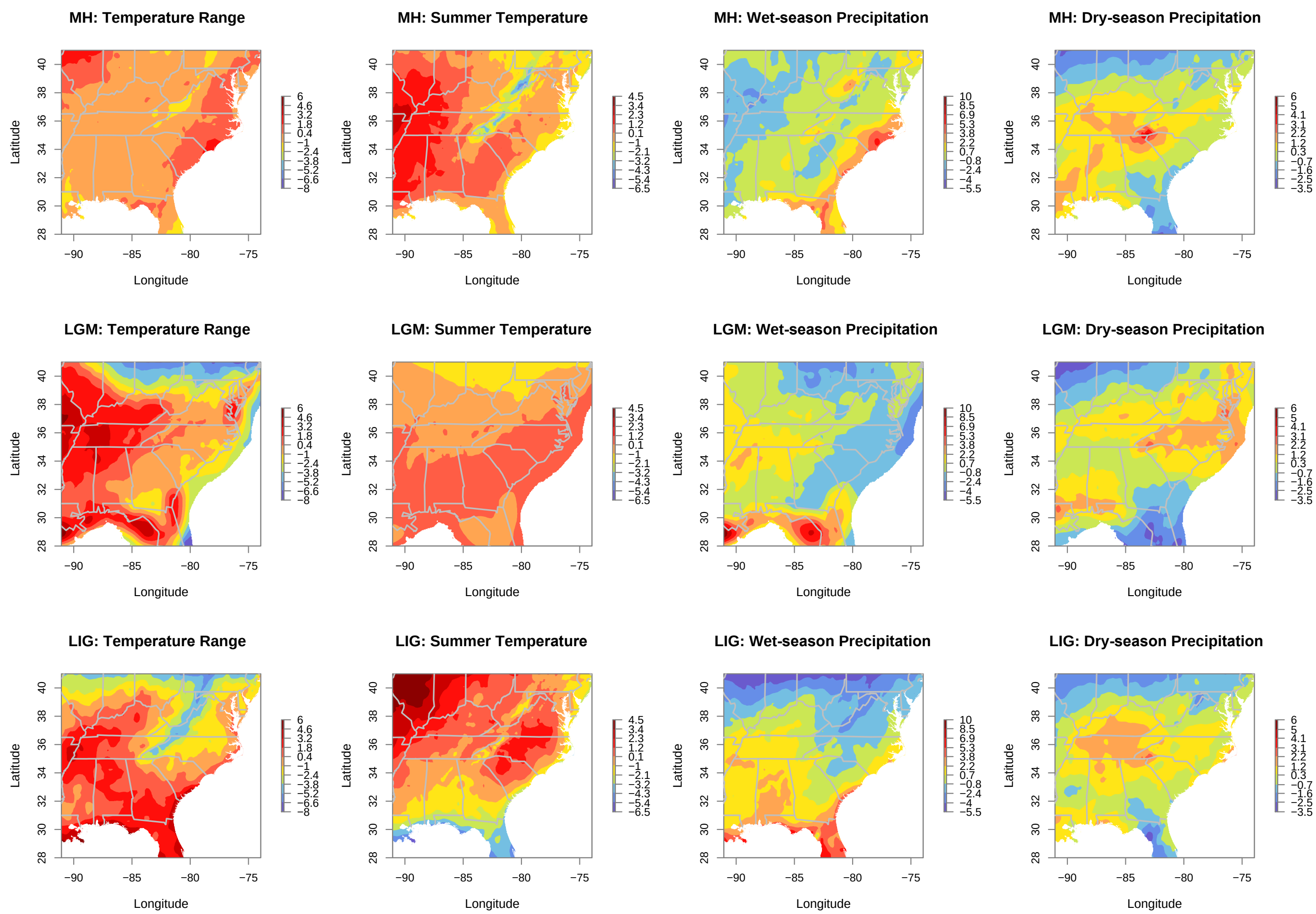



**Figure B.9:** *Paleoclimatic factors.* Each column of panels shows one of the four environmental factors. The top row of panels depicts the four factors for the Mid-Holocene (MH), the middle and bottom rows shows the factors for the Last Glacial Maximum (LGM) and the Last Interglacial (LIG), respectively. The environmental factors are unitless and go from negative (dark blue) to positive (dark red) values.

### 2.2.2 Genetic divergence, environment, and spatial structure

To measure divergence among genetic populations, the following statistics were calculated: average number of nucleotide substitutions per site (Dxy[143]), net number of nucleotide substitutions per site (Da[143]), average number of pairwise nucleotide differences (Kxy[144]), and $F_{ST}$[2].

**Table B.7:** *Genetic divergence.* Da = number of net nucleotide substitutions per site between populations; Dxy = average number of nucleotide substitutions per site between populations; Kxy = average number of pairwise nucleotide differences. Calculation of $F_{ST}$ is based on[2], treating each polymorphic site as a separate locus. Pairwise comparisons were performed among the Northern (N), Central (C), and Southern (S) genetic clusters.

| Locus | Comparison | Fixed Differences | Da | Dxy | Kxy | $F_{ST}$ |
|---|---|---|---|---|---|---|
| mtDNA | S-N | 9 | 0.013 | 0.026 | 28.611 | 0.494 |
| | S-C | 15 | 0.018 | 0.028 | 31.066 | 0.659 |
| | N-C | 3 | 0.005 | 0.012 | 13.690 | 0.447 |

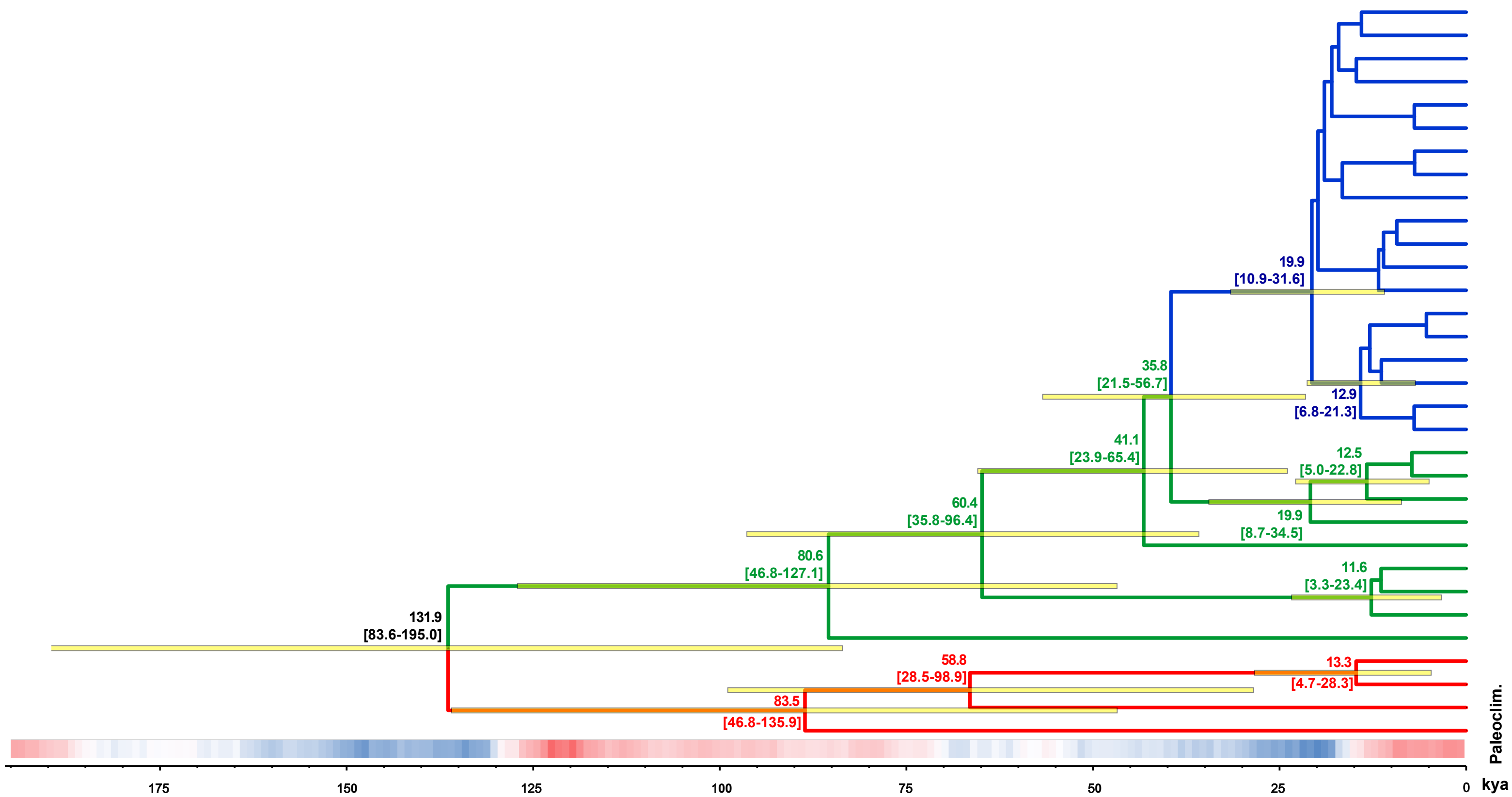



**Figure B.10:** *MtDNA phylogeny with divergence times*. Median values of divergence times and 95% confidence intervals are shown at nodes, color coded by genetic cluster (red: southern; green: northern; blue: central). Bars at nodes represent 95% confidence intervals. Median divergence times below 10 kya (kya = 1,000 years ago) are not shown. Paleoclimate (global surface air temperature data from[1]) is indicated by a bar at the bottom coded from blue (9.4°C) to red (15.6°C), representing cold to hot periods, respectively. The time scale is shown at the bottom.

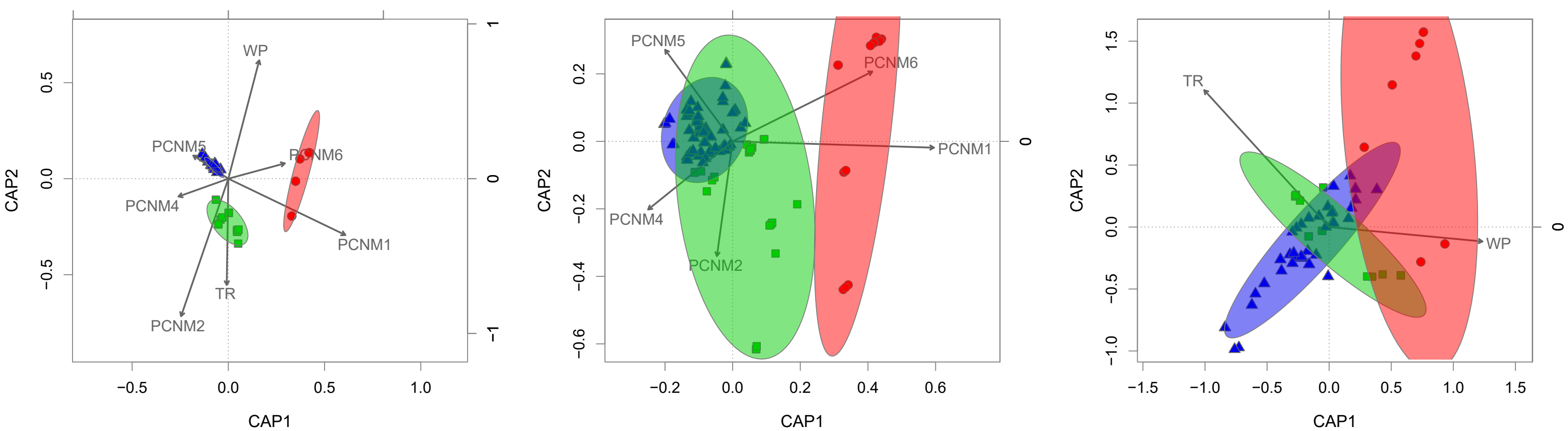



**Figure B.11:** *Distance-based Redundancy Analysis (dbRDA).* The three panels show multivariate dbRDA-partitioned variation in the mtDNA sequence data explained by geography (eigenvectors obtained via Principal Coordinates analysis of Neighbor Matrices, PCNM) and the contemporary environmental data (factors obtained via factor analysis). The left panel shows the full model, the middle panel shows geography (eigenvectors with significant contribution to genetic variation: PCNM1, 2, 4, 5, and 6) after removing contributions of the environment, and the right panel shows the environment (factors with significant contribution to genetic variation: TR = “temperature range” and WP = “wet-season precipitation”) after factoring out geography. CAP stands for Constrained Analysis of Principal coordinates. CAP1 and CAP2 denote axes 1 and 2. The Northern cluster is shown in green, the Central in blue, and the Southern in red. The ellipses represent 95% confidence intervals.

### 2.2.3 Phylogeographic scenarios: error rates and parameter estimates

**Table B.8:** *Type I and II error rates*. Type I (false positive) and type II (false negative) error rates for three alternative scenarios in the second tier of ABC hypothesis testing (see Figure B.5).

| Scenario | Type I error rate | Type II error rate |
|---|---|---|
| **DS1** | 0.509 | 0.328 |
| **R3** | 0.446 | 0.354 |
| **V** | 0.734 | 0.159 |

**Table B.9:** *Parameters of the best-fit scenario estimated using ABC*. N, C, and S represent the effective population sizes of the Northern, Central, and Southern clusters. $T_N$ and $T_C$ represent the time of divergence of the N and C clusters. $b_N$ and $b_C$ represent duration (number of generations) of bottleneck events. The parameters $N_b$ and $C_b$ represent effective population sizes during bottleneck events, and $\mu_{mt}$ and $\mu_{nuc}$ are mutation rates of the mtDNA and nDNA loci. Precision of parameter estimation is shown using the mean, median, and mode of the relative median of the absolute error (RMAE) for 500 data sets simulated using values drawn from posterior distributions.

| *DS1: S-N;N-C* | | | | | | |
|---|---|---|---|---|---|---|
| | | | | *RMAE* | | |
| *Parameter* | *Median* | *Quantile 2.5%* | *Quantile 97.5%* | *Mean* | *Median* | *Mode* |
| N | 82,700 | 32,300 | 213,000 | 0.326 | 0.324 | 0.374 |
| C | 1,170,000 | 516,000 | 4,530,000 | 0.581 | 0.546 | 0.624 |
| S | 174,000 | 63,100 | 245,000 | 0.226 | 0.227 | 0.309 |
| $T_N$ | 64,800 | 26,400 | 115,000 | 0.270 | 0.264 | 0.351 |
| $b_N$ | 5.68 | 1 | 10 | 0.398 | 0.426 | 0.75 |
| $N_b$ | 4,980 | 559 | 44,700 | 1.096 | 0.865 | 0.905 |
| $T_C$ | 8,630 | 2,750 | 22,500 | 0.390 | 0.362 | 0.438 |
| $b_C$ | 8.520 | 1 | 10 | 0.352 | 0.366 | 0.666 |
| $C_b$ | 168 | 101 | 4,570 | 7.410 | 2.409 | 0.453 |
| $\mu_{mt}$ | $1.21 \times 10^{-7}$ | $3.33 \times 10^{-8}$ | $4.13 \times 10^{-7}$ | 0.411 | 0.422 | 0.477 |
| $\mu_{nuc}$ | $5.76 \times 10^{-9}$ | $1.79 \times 10^{-9}$ | $1.79 \times 10^{-8}$ | 0.392 | 0.382 | 0.453 |

## 2.2.4 Population size changes: standard and compound neutrality tests

**Table B.10:** *Compound tests of neutrality in the Central cluster.* Both sampling site ID and genetic population membership of the out-group sequences used to perform the tests are shown. D = Tajima's D; H = Fay and Wu's H; EW = Ewens and Watterson statistic; DH = Combination of D and H; HEW = Combination of H and EW; DHEW = Combination of D, H, and EW. Significant values are shown in bold. The statistics and p-values are reported in separate rows, which have been labeled accordingly. Note that there are no compound statistics, only p-values associated with the compound tests.

| **Standard and Compound Neutrality Tests: Central Population** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Out-group: Site | Out-group: Cluster | D | H | EW | DH | HEW | DHEW | |
| A70 | N | **-1.886** | 0.254 | 0.058 | | | | *statistic* |
| | | **0.015** | 0.431 | 1 | 0.237 | 1 | 1 | *p-value* |
| A106 | N | **-1.886** | 0.254 | 0.058 | | | | *statistic* |
| | | **0.014** | 0.430 | 1 | 0.237 | 1 | 1 | *p-value* |
| A142 | N | **-1.886** | -0. 531 | 0.058 | | | | *statistic* |
| | | **0.014** | 0.184 | 1 | 0.083 | 1 | 1 | *p-value* |
| A60 | N | **-1.853** | -0.578 | 0.058 | | | | *statistic* |
| | | **0.017** | 0.176 | 1 | 0.079 | 1 | 1 | *p-value* |

# Appendix C:

# Chapter 3: Supplementary Material

## 3.1 Supplementary Methods

### 3.1.1 MS-AFLP protocol

Total genomic DNA was extracted using the Hot Sodium Hydroxide and Tris (HotSHOT) protocol[293], with a modified lysis step. We used 90 µL lysis solution (pH = 12.3) consisting of 25 mM NaOH and 0.2 mM Na2EDTA, as per[293], but we used a heating time of 30 min at 95°C, followed by slow cooling in the thermocycler from 95°C to 4°C (-0.2°C every 1 min). The lysis solution was then neutralized with 90 µL of 400 mM Tris-HCl (pH = 5.3) The HotSHOT protocol allowed for high-throughput isolation of DNA from 177 individuals.

We constructed 4.5 µM EcoRI and 45 µM HpaII/MspI adapters using adapter buffer consisting of 100 mM Tris, 10 mM EDTA, and 500 mM NaCl. For the EcoRI adapter, we used 60 µL of buffer and 300 µL each of 10 µM E_A1 and E_A2 primers (Table C.2; Figure C.1). For HpaII/MspI, we used 600 µL of 100 µM HM_A1 and HM_A2 primers. The ligation adapter mixtures were heated in a thermocycler for 3 min at 95°C, followed by slow cooling from 95°C to 12°C (-1°C every 1 min).

For restriction digests, 50-100 ng of DNA was digested with 8 units of EcoRI (Promega) and 8 or 6 units of HpaII or MspI (Promega), respectively. The reaction was incubated at 37°C for 3 hrs and inactivated at 65°C for 1 hr. Then, the product was combined with 1 unit of T4 DNA ligase (Promega), and final concentrations of 1x for T4 DNA ligase buffer, and 0.45 µM and 4.5 µM for EcoRI and HpaII/MspI adapters, respectively. The reaction was incubated at 37°C for 5 hrs.

The fragments resulting from these restriction digestions were ligated to two adapters compatible with EcoRI- and MspI/HpaII-generated ends (Figure C.1).

Ligated fragments were pre-amplified via PCR using pre-selective primers complementary to the adapters, followed by amplification with a pair of selective primers (Figure C.1).

For the pre-selective amplification, 5 μL of ligation product was combined with 10 μl of PCR mix containing 1 unit of Taq polymerase (Promega), 5 pmol each of preselective primers E_preA and HM_preT (Table C.2), 4 nmol of each dNTP (Promega), 25 nmol of $MgCl_2$, 6 μg of bovine serum albumin (Promega), and 1x buffer (Promega) and 4% dimethyl sulfoxide (final concentrations).

In order to remove excess primer and adapter dimers, the pre-selective PCR product was cleaned using Sera-Mag Speedbeads (GE Healthcare, Illinois, USA). We prepared the Speedbead solution (protocol available upon request) and used it at a 1:1 ratio with PCR product, in order to remove all fragments < 100 bp. For the selective amplification step, we used 3.5 μL of cleaned preselective PCR product.

For both pre-selective and selective PCR, we used the same touchdown thermocycler profile: 1) initial extension at 72°C (5 min) and denaturation at 95°C (3 min), followed by 2) 8 cycles of denaturation at 95°C (30 sec), annealing at 58-51°C (1 min; decreasing temperature by 1°C with every cycle), and extension at 72°C (1 min), with another 3) 37 cycles at an annealing temperature of 50°C (same denaturation and extension temperatures), and lastly, 4) a final extension at 72°C (10 min) and 60°C (15 min).

We performed preliminary testing of 18 (3 EcoRI x 6 HpaII/MspI) combinations: EcoRI + AC/AG/AT and HpaII/MspI + TAC/TAG/TCA/TCT/TGA/TGT. We selected the combination that maximized polymorphism while making it possible to use four primers in the same selective reaction: EcoRI + AT/AG and HpaII/MspI + TCA/TCT. EcoRI + AT was labeled with 6-FAM fluorescent dye, whereas EcoRI + AG was labeled with HEX. Thus, we were able to separate and visualize 6-FAM-labeled from HEX-labeled selective PCR products.

### 3.1.2 Genotyping

Fragment separation and detection was done on an ABI3730XL DNA capillary sequencer (Applied Biosystems, California, USA) at the DNA Analysis Facility on Science Hill at Yale University (`http://dna-analysis.research.yale.edu`). In order to improve fragment separation and detection, each reaction was run in duplicate at an injection voltage of 9 kV and again at 12 kV. We used a 50-500 bp ABI ROX ladder (Gel Company, California, USA) for sizing the MS-AFLP fragments. Since we only cleaned pre-selective PCR products, primer and adapter dimers were present in selective PCR products. Thus, fragments < 100 bp were not considered.

Binning of fragments was performed using a peak height threshold of 250 relative fluorescence units. We used the R package 'binner'[294] to score fragments in an automated fashion, using optimized parameters, followed by 'AFLPscore,'[295] a method for scoring AFLP peak-height data that minimizes genotyping error, then, to make sure all bins were more than 1 bp apart, re-binning and re-scoring, using R scripts (provided here: `https://github.com/chazhyseni/msaflp`). Fragment profiles for each individual were then visualized and checked manually with the program GeneMarker 2.4.0 (SoftGenetics, State College, PA).

**Table C.1:** *Sampling sites*. Termites were collected from one log per site. State, county, and ecoregion information is shown for each site. Geographic coordinates and altitude (in meters) data were collected using a handheld GPS device.

| Site | State | County | Ecoregion | Longitude | Latitude | Alt. (m) |
|---|---|---|---|---|---|---|
| 1 | Virginia | Scott | Valley and Ridge | -82.74536 | 36.70494 | 460 |
| 2 | Virginia | Botetourt | Blue Ridge | -79.68214 | 37.47978 | 759 |
| 3 | Virginia | Smyth | Valley and Ridge | -81.53171 | 36.88458 | 731 |
| 4 | Virginia | Patrick | Piedmont | -80.06615 | 36.78941 | 386 |
| 5 | North Carolina | Rockingham | Piedmont | -79.95090 | 36.43191 | 256 |
| 6 | Virginia | Bedford | Blue Ridge | -79.59341 | 37.44090 | 692 |
| 7 | Virginia | Bath | Valley and Ridge | -79.97707 | 37.98723 | 508 |
| 8 | West Virginia | Pendleton | Valley and Ridge | -79.20190 | 38.60225 | 591 |
| 9 | West Virginia | Hardy | Valley and Ridge | -78.91022 | 38.89373 | 589 |
| 10 | Ohio | Gallia | Appalachian Plateaus | -82.49056 | 38.81387 | 277 |
| 11 | Tennessee | Morgan | Appalachian Plateaus | -84.75872 | 36.01773 | 572 |
| 12 | Tennessee | Scott | Appalachian Plateaus | -84.71430 | 36.47398 | 479 |
| 13 | Kentucky | Mccreary | Appalachian Plateaus | -84.45749 | 36.84983 | 412 |
| 14 | Kentucky | Mccreary | Appalachian Plateaus | -84.42480 | 36.91024 | 336 |
| 15 | Tennessee | Knox | Valley and Ridge | -83.76402 | 36.10415 | 399 |
| 16 | Tennessee | Union | Valley and Ridge | -83.89043 | 36.37519 | 490 |
| 17 | Kentucky | Bell | Valley and Ridge | -83.69725 | 36.60349 | 352 |
| 18 | Kentucky | Bell | Appalachian Plateaus | -83.74413 | 36.72807 | 390 |
| 19 | Kentucky | Harlan | Appalachian Plateaus | -83.21425 | 36.92808 | 767 |
| 20 | Tennessee | Sullivan | Valley and Ridge | -82.48692 | 36.49101 | 427 |
| 21 | Kentucky | Knott | Appalachian Plateaus | -82.99386 | 37.24096 | 318 |
| 22 | Georgia | Douglas | Piedmont | -84.63633 | 33.76154 | 295 |

| Site | State | County | Ecoregion | Longitude | Latitude | Alt. (m) |
|---|---|---|---|---|---|---|
| 23 | North Carolina | Buncombe | Blue Ridge | -82.49127 | 35.60575 | 770 |
| 24 | North Carolina | Henderson | Blue Ridge | -82.71758 | 35.44758 | 1205 |
| 25 | North Carolina | Henderson | Blue Ridge | -82.58961 | 35.21877 | 809 |
| 26 | Tennessee | Monroe | Blue Ridge | -84.24123 | 35.34314 | 413 |
| 27 | Tennessee | Monroe | Blue Ridge | -84.11201 | 35.39665 | 553 |
| 28 | Tennessee | Polk | Blue Ridge | -84.33586 | 35.20793 | 513 |
| 29 | Tennessee | Polk | Blue Ridge | -84.60815 | 35.14822 | 588 |
| 30 | North Carolina | Madison | Blue Ridge | -82.84724 | 35.85284 | 656 |
| 31 | Tennessee | Greene | Valley and Ridge | -82.84973 | 36.08371 | 408 |
| 32 | Tennessee | Unicoi | Blue Ridge | -82.44664 | 36.10384 | 522 |
| 33 | Kentucky | Powell | Appalachian Plateaus | -83.67732 | 37.77913 | 256 |
| 34 | Kentucky | Floyd | Appalachian Plateaus | -82.72829 | 37.71582 | 213 |
| 35 | Kentucky | Lawrence | Appalachian Plateaus | -82.82529 | 38.05997 | 209 |
| 36 | West Virginia | Wayne | Appalachian Plateaus | -82.42619 | 38.30313 | 186 |
| 37 | West Virginia | Wayne | Appalachian Plateaus | -82.38316 | 38.02512 | 402 |
| 38 | West Virginia | Logan | Appalachian Plateaus | -82.01469 | 37.88885 | 260 |
| 39 | West Virginia | Lincoln | Appalachian Plateaus | -81.84275 | 38.18754 | 201 |
| 40 | West Virginia | Kanawha | Appalachian Plateaus | -81.66953 | 38.26121 | 267 |
| 41 | West Virginia | Jackson | Appalachian Plateaus | -81.57557 | 38.65200 | 241 |
| 42 | West Virginia | Roane | Appalachian Plateaus | -81.34470 | 38.77533 | 240 |
| 43 | West Virginia | Braxton | Appalachian Plateaus | -80.65887 | 38.63269 | 394 |
| 44 | West Virginia | Summers | Appalachian Plateaus | -80.83122 | 37.50894 | 550 |
| 45 | Alabama | Lawrence | Appalachian Plateaus | -87.41399 | 34.38517 | 314 |

**Table C.2:** *Adapter and primer sequences.*

**Adapters:**

| *Primer Name:* | *Primer Sequence:* |
|---|---|
| E_A1 | 5'-CTCGTAGACTGCGTACC-3' |
| E_A2 | 5'-AATTGGTACGCAGTCTAC-3' |
| HM_A1 | 5'-GACGATGAGTCTAGAA-3' |
| HM_A2 | 5'-CGTTCTAGACTCATC-3' |

**Pre-Selective Primers:**

| *Primer Name:* | *Primer Sequence:* |
|---|---|
| E_preA | 5'-GACTGCGTACCAATTCA-3' |
| HM_preT | 5'-GATGAGTCTAGAACGGT-3' |

**Selective Primers:**

| *Primer Name:* | *Primer Sequence:* |
|---|---|
| E_AT | 5'-/6-FAM/GACTGCGTACCAATTCAT3' |
| E_AG | 5'-/HEX/GACTGCGTACCAATTCAG-3' |
| HM_TCA | 5'-GATGAGTCTAGAACGGTCA-3' |
| HM_TCT | 5'-GATGAGTCTAGAACGGTCT-3' |

**Table C.3:** *Sampling sites with clustering and caste information.* The table shows numbers of individuals at each site assigned to the four clusters. There were 8 individuals that were assigned with probability less than 0.6. These individuals appear in the 'Unassigned' column. Additionally, all individuals at each site were identified as soldiers or workers. We collected epigenetic data for 0-1 soldiers and 1-4 workers per site.

| Site | Unassigned | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Soldiers | Workers |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 2 | 1 | 1 | 3 |
| 2 | 0 | 3 | 0 | 0 | 1 | 0 | 4 |
| 3 | 0 | 0 | 1 | 0 | 2 | 1 | 2 |
| 4 | 0 | 0 | 0 | 2 | 2 | 1 | 3 |
| 5 | 0 | 0 | 1 | 2 | 1 | 0 | 4 |
| 6 | 0 | 0 | 2 | 2 | 0 | 1 | 3 |
| 7 | 0 | 0 | 2 | 0 | 1 | 0 | 3 |
| 8 | 0 | 0 | 0 | 1 | 2 | 1 | 2 |
| 9 | 0 | 1 | 0 | 0 | 2 | 0 | 3 |
| 10 | 0 | 0 | 0 | 2 | 1 | 1 | 2 |
| 11 | 0 | 0 | 0 | 3 | 1 | 1 | 3 |
| 12 | 0 | 0 | 0 | 2 | 2 | 1 | 3 |
| 13 | 0 | 0 | 4 | 0 | 0 | 1 | 3 |
| 14 | 0 | 0 | 0 | 4 | 0 | 1 | 3 |
| 15 | 0 | 0 | 2 | 2 | 0 | 1 | 3 |
| 16 | 0 | 0 | 2 | 1 | 1 | 1 | 3 |
| 17 | 0 | 0 | 0 | 3 | 1 | 0 | 4 |
| 18 | 0 | 0 | 1 | 1 | 2 | 1 | 3 |
| 19 | 0 | 0 | 4 | 0 | 0 | 1 | 3 |
| 20 | 0 | 0 | 0 | 1 | 3 | 1 | 3 |
| 21 | 1 | 0 | 0 | 1 | 2 | 0 | 4 |
| 22 | 0 | 1 | 3 | 0 | 0 | 1 | 3 |
| 23 | 0 | 1 | 0 | 2 | 0 | 1 | 2 |
| 24 | 1 | 0 | 1 | 2 | 0 | 1 | 3 |
| 25 | 0 | 0 | 2 | 0 | 2 | 1 | 3 |
| 26 | 0 | 1 | 2 | 0 | 1 | 1 | 3 |
| 27 | 1 | 2 | 1 | 0 | 0 | 1 | 3 |
| 28 | 1 | 0 | 2 | 1 | 0 | 1 | 3 |
| 29 | 0 | 2 | 2 | 0 | 0 | 1 | 3 |
| 30 | 0 | 0 | 0 | 3 | 0 | 1 | 2 |
| 31 | 1 | 0 | 2 | 0 | 1 | 1 | 3 |
| 32 | 0 | 0 | 1 | 1 | 2 | 1 | 3 |
| 33 | 1 | 0 | 1 | 2 | 0 | 1 | 3 |
| 34 | 0 | 0 | 2 | 1 | 1 | 1 | 3 |
| 35 | 0 | 1 | 1 | 1 | 1 | 1 | 3 |
| 36 | 0 | 0 | 1 | 2 | 1 | 1 | 3 |
| 37 | 0 | 0 | 0 | 2 | 1 | 1 | 2 |
| 38 | 0 | 0 | 0 | 3 | 0 | 0 | 3 |
| 39 | 0 | 0 | 0 | 2 | 2 | 1 | 3 |
| 40 | 0 | 0 | 0 | 4 | 0 | 1 | 3 |
| 41 | 1 | 0 | 1 | 1 | 1 | 1 | 3 |
| 42 | 1 | 0 | 0 | 0 | 2 | 0 | 3 |
| 43 | 0 | 2 | 1 | 0 | 1 | 1 | 3 |
| 44 | 0 | 0 | 0 | 2 | 2 | 1 | 3 |
| 45 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

**Table C.4:** *dbRDA analysis of variance.* Degrees of freedom, sums of squares, *F*- and *p*-values are shown for constrained dbRDA: geography (i.e., spatial structure), environment, environment conditioned on geography (9 significant PCNMs), and environment conditioned on population stratification (8 categories = 2 castes * 4 clusters) and geography.

**Geography**

| | d.f. | Sum of Squares | F | *p* |
|---|---|---|---|---|
| PCNM1 | 1 | 0.497 | 1.267 | **0.009** |
| PCNM2 | 1 | 0.716 | 1.826 | **0.001** |
| PCNM3 | 1 | 0.550 | 1.402 | **0.002** |
| PCNM7 | 1 | 0.472 | 1.203 | **0.030** |
| PCNM14 | 1 | 0.479 | 1.221 | **0.029** |
| PCNM17 | 1 | 0.466 | 1.188 | **0.038** |
| PCNM18 | 1 | 0.523 | 1.334 | **0.004** |
| PCNM19 | 1 | 0.498 | 1.270 | **0.010** |
| PCNM22 | 1 | 0.477 | 1.217 | **0.029** |
| Residual | 157 | 61.582 | | |

**Environment**

| | d.f. | Sum of Squares | F | *p* |
|---|---|---|---|---|
| DP | 1 | 0.463 | 1.169 | **0.037** |
| ST | 1 | 0.502 | 1.269 | **0.008** |
| WP | 1 | 0.500 | 1.263 | **0.008** |
| AWC30cm | 1 | 0.500 | 1.263 | **0.007** |
| Pdiv | 1 | 0.569 | 1.438 | **0.001** |
| Qdiv | 1 | 0.486 | 1.228 | **0.016** |
| Tree | 1 | 0.410 | 1.037 | 0.341 |
| Residual | 159 | 62.920 | | |

**Environment | Geography**

| | d.f. | Sum of Squares | F | *p* |
|---|---|---|---|---|
| DP | 1 | 0.431 | 1.107 | 0.147 |
| ST | 1 | 0.463 | 1.188 | **0.034** |
| WP | 1 | 0.544 | 1.398 | **0.003** |
| AWC30cm | 1 | 0.458 | 1.175 | *0.064* |
| Pdiv | 1 | 0.483 | 1.241 | **0.017** |
| Qdiv | 1 | 0.459 | 1.179 | **0.041** |
| Tree | 1 | 0.423 | 1.086 | 0.179 |
| Residual | 150 | 58.416 | | |

**Environment | Caste*Cluster + Geography**

| | d.f. | Sum of Squares | F | *p* |
|---|---|---|---|---|
| DP | 1 | 0.406 | 1.084 | 0.178 |
| ST | 1 | 0.449 | 1.199 | **0.025** |
| WP | 1 | 0.496 | 1.324 | **0.001** |
| AWC30cm | 1 | 0.447 | 1.191 | **0.031** |
| Pdiv | 1 | 0.491 | 1.309 | **0.003** |
| Qdiv | 1 | 0.429 | 1.143 | *0.060* |
| Tree | 1 | 0.408 | 1.088 | 0.170 |
| Residual | 143 | 53.599 | | |

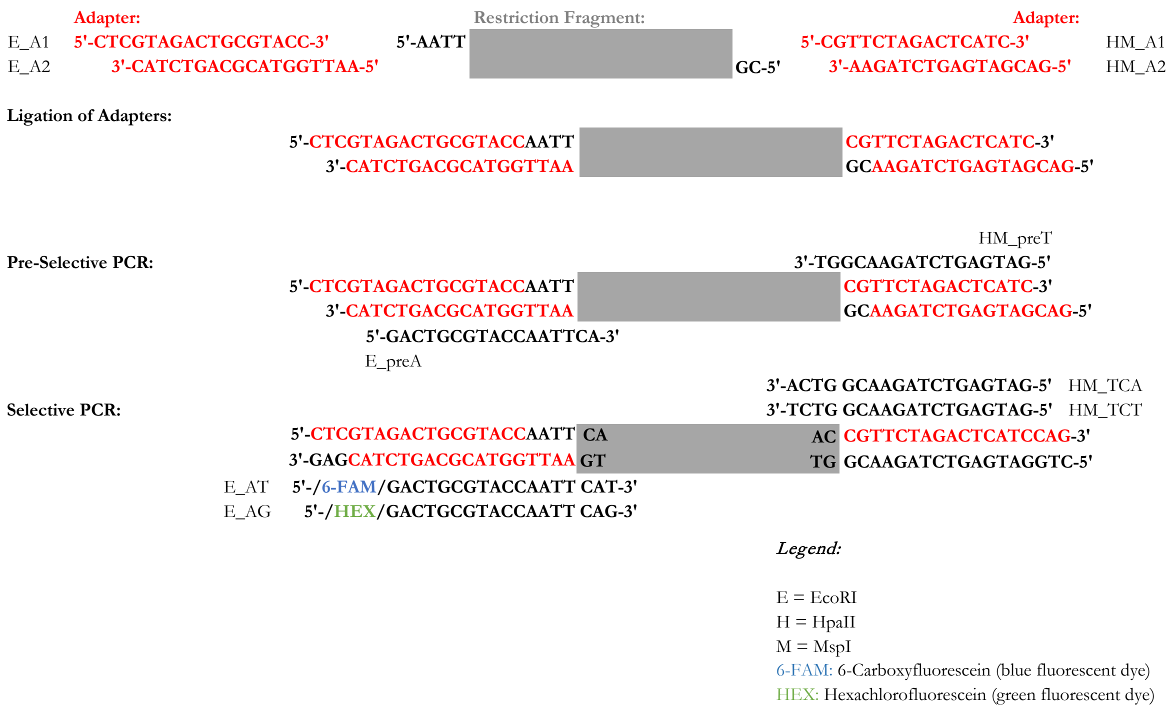



**Figure C.1:** *Methylation-sensitive Amplified Fragment Length Polymorphism schematic.*

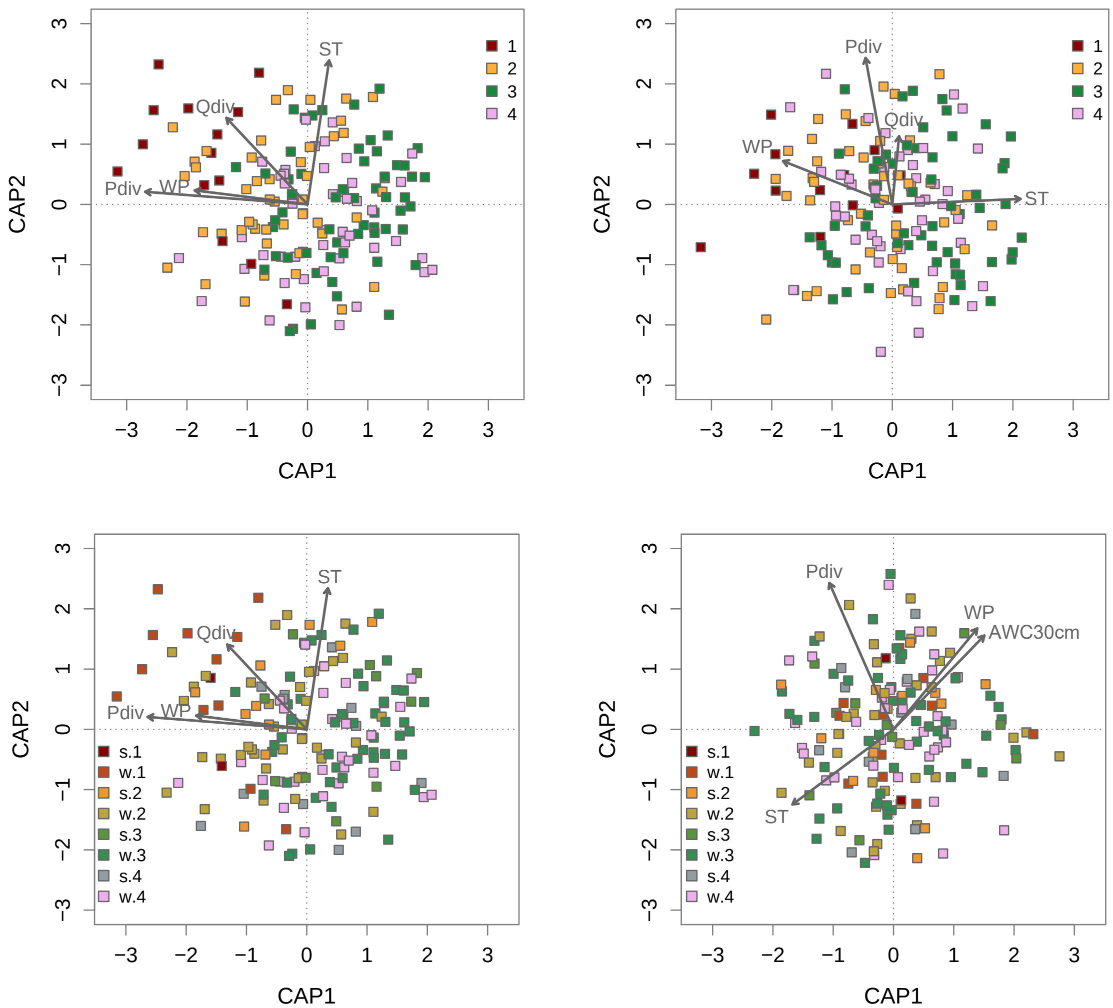



**Figure C.2:** *Distance-based Redundancy Analysis (dbRDA).* Four epigenetic clusters are labeled 1 through 4, and two castes are labeled 's' (soldier), and 'w' (worker). The two left panels show dbRDA constrained by environment alone, with the top panel showing cluster membership for each individual, while the bottom panel shows caste identity as well. The two right panels shows environment-constrained dbRDA, conditioned on geography alone in the top panel, while the bottom panel represents environment-constrained dbRDA after accounting for geography, caste identity, and epigenetic clustering. Only significant environmental variables are shown.

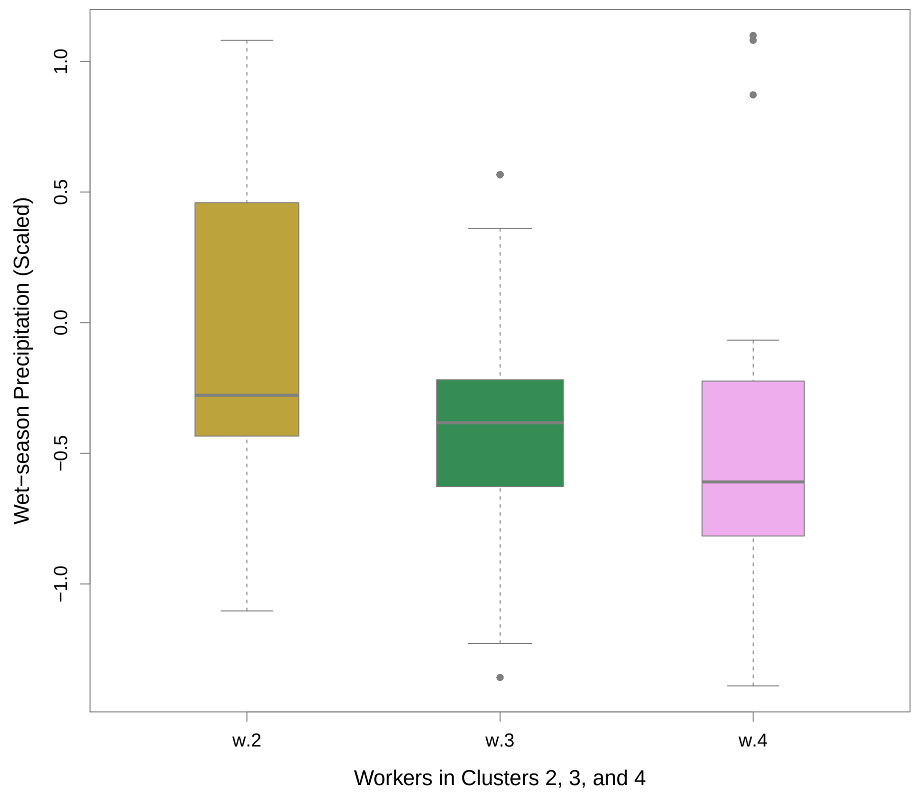



**Figure C.3:** *Box plots of wet-season precipitation for workers in clusters 2, 3, and 4.* Non-parametric Games-Howell posthoc test *p*-values: $p = 0.069$ for the w.2–w.4 comparison, $p = 0.098$ for the w.2–w.3 comparison, and $p = 0.987$ for the w.3–w.4 comparison.

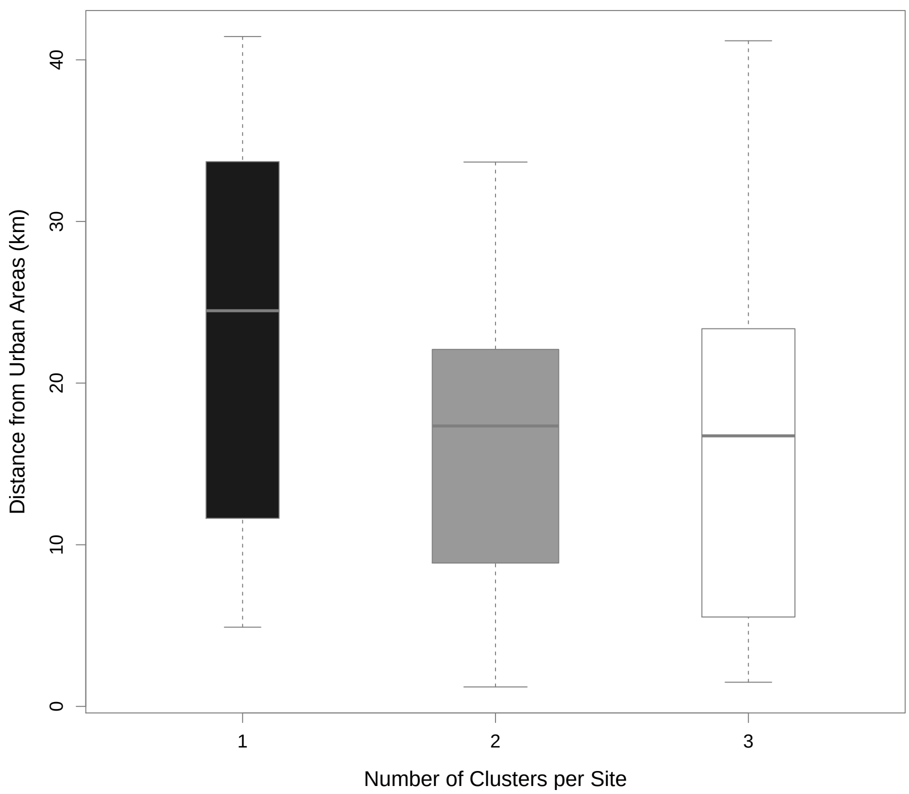



**Figure C.4:** *Box plots of distance from urban areas for different site categories.* Sites were grouped based on the number of clusters that individuals were assigned to at each site: 1, 2, and 3. The last category, 3, includes, in addition to sites with three clusters, one site where all four individuals were assigned to a different cluster. Non-parametric Games-Howell posthoc test $p$-values were $> 0.05$ for all comparisons.

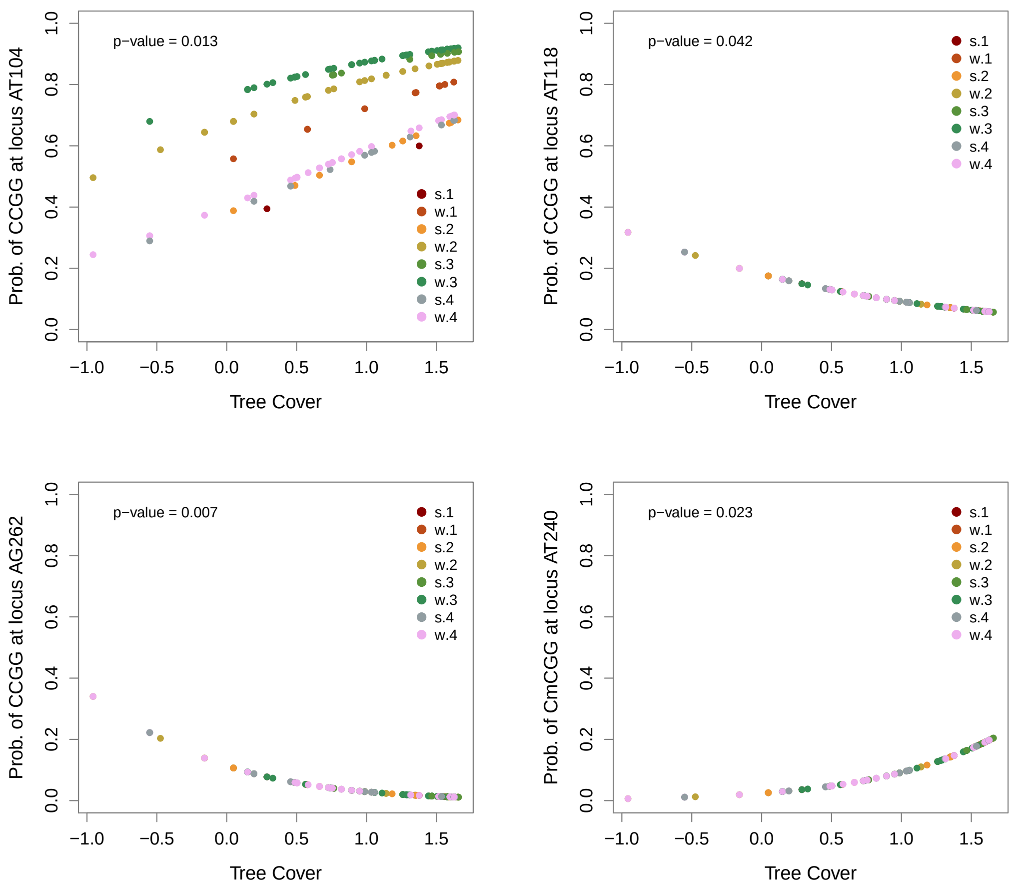



**Figure C.5:** *Effect of tree cover on methylation states at loci AT104, AT118, AG262, and AT240.* CCGG at locus AT104 (top left) and C*m*CGG at locus AT240 (bottom right) are positively correlated with tree cover. CCGG at locus AT118 (top right) and AG262 (bottom left) are negatively correlated with tree cover. Soldiers (s) and workers (w) in each of the four clusters are color coded.

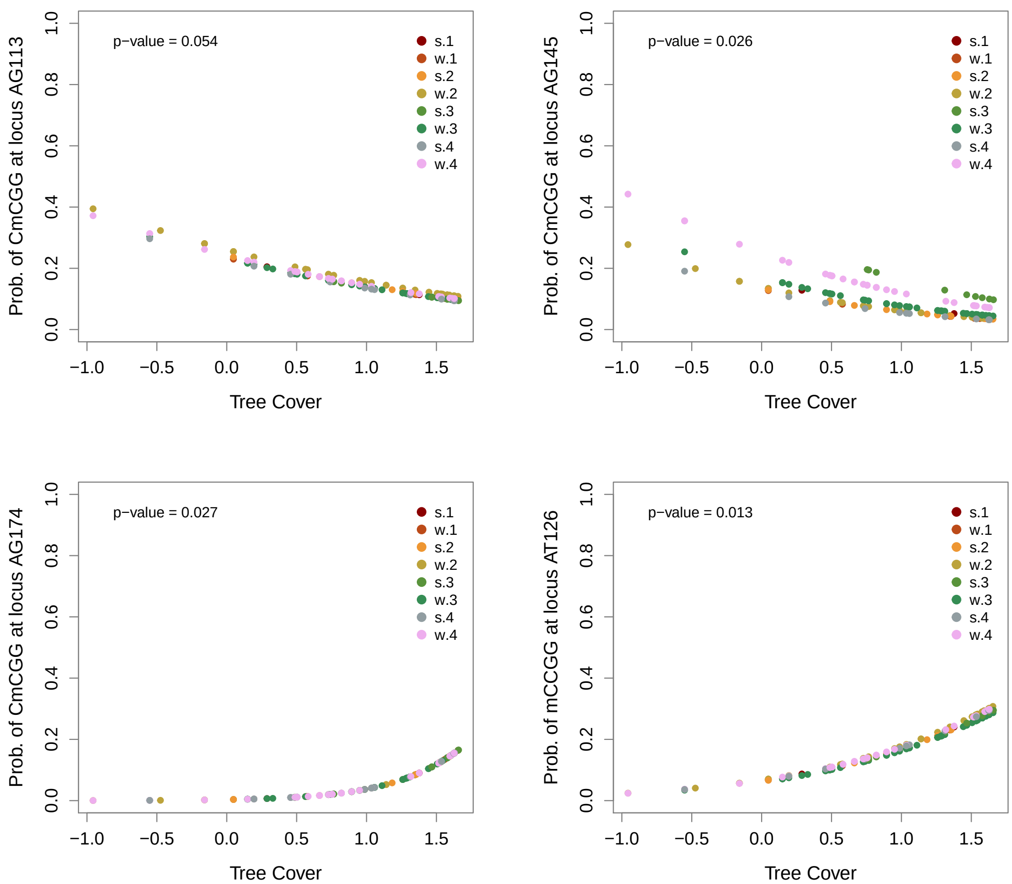



**Figure C.6:** *Effect of tree cover on methylation states at loci AG113, AG145, AG174, and AT126.* The top two panels show that probability of C*m*CGG methylation at loci AG113 and AG145 is negatively correlated with tree cover. C*m*CGG at locus AG174 (bottom left) and *m*CCGG at locus AT126 (bottom right) are positively correlated with tree cover. Soldiers (s) and workers (w) in each of the four clusters are color coded.

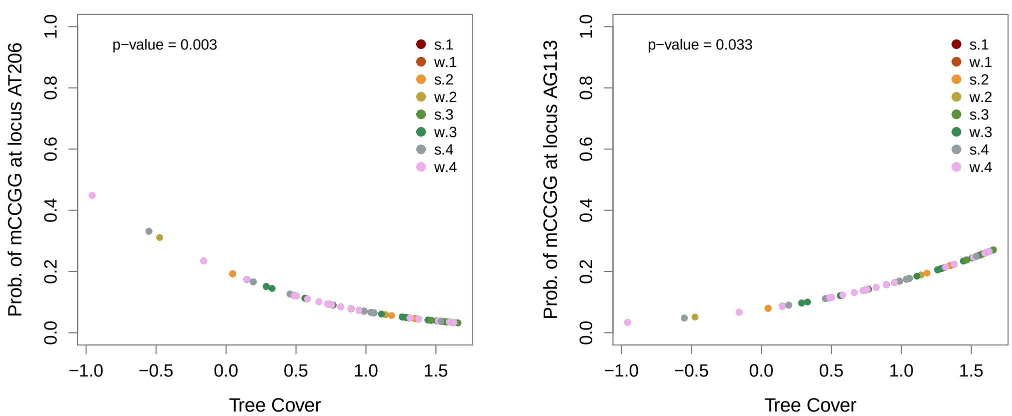



**Figure C.7:** *Effect of tree cover on methylation states at loci AT206 and AG113.* mCCGG at locus AT206 (left) is negatively correlated with tree cover, while *m*CCGG at locus AG113 (right) is positively correlated with tree cover. Soldiers (s) and workers (w) in each of the four clusters are color coded.

# Appendix D:

# Chapter 4: Supplementary Material

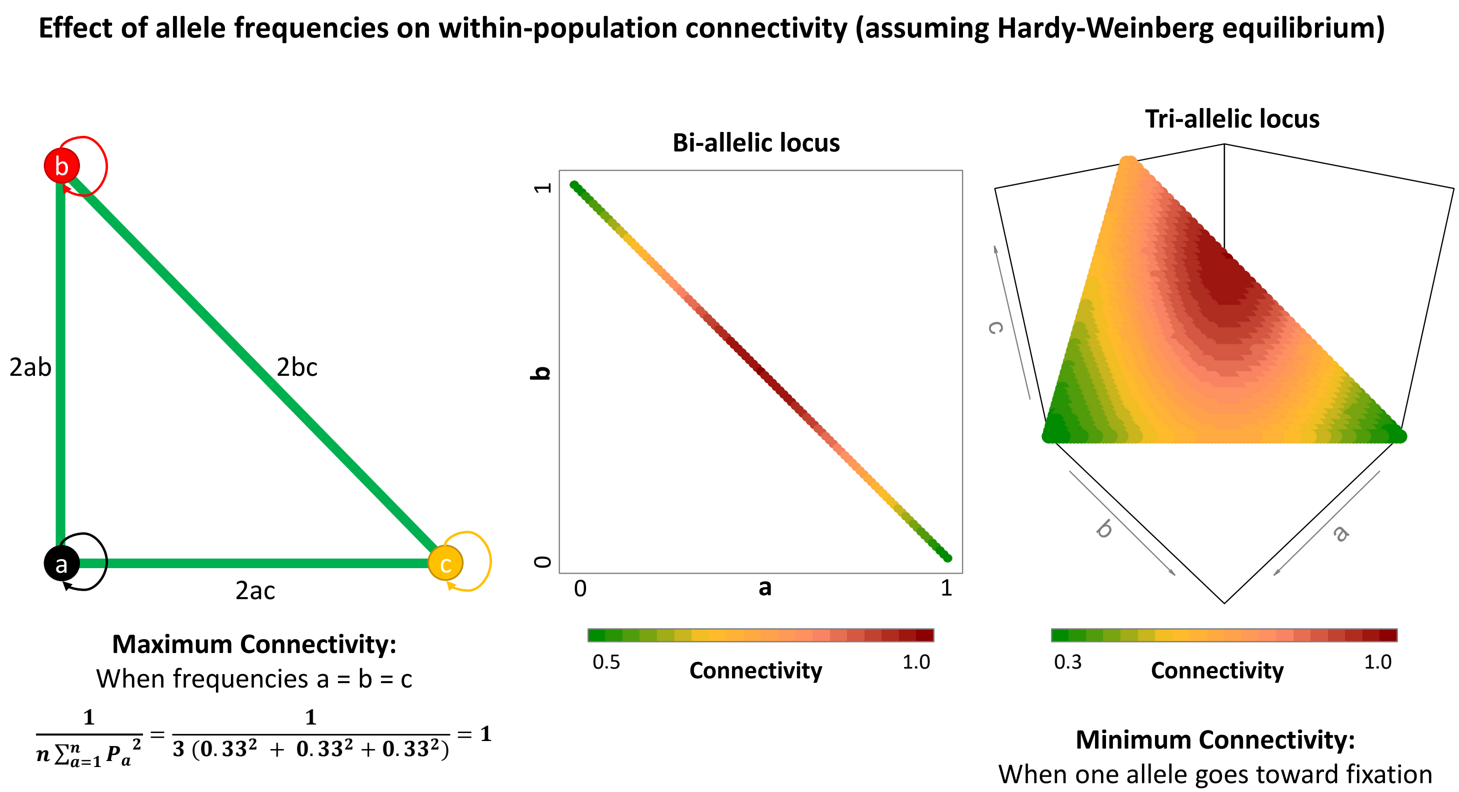



**Figure D.1:** *Properties of the connectivity metric.* Connectivity ranges from 1/n (fixation) to 1 (equal frequencies) based on allele frequencies.

**Table D.1:** *Root mean square error (RMSE) of connectivity comparisons for the* gradient *landscape*. RMSE values shown here were used to quantify differences for within- and between-population connectivity based on neutral versus non-neutral (s = 0.1) loci for different degrees of long-distance dispersal (σ = 0.2, 0.5, and 1). RMSE values greater than 0.150 are italicized, whereas values greater than 0.250 are shown in bold.

| Root Mean Square Error of Functional Connectivity | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Gradient Landscape** | | | **Within-Population** | | | | | |
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Within** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | 0.084 | | | | | |
| | | $\sigma = 1.0$ | 0.076 | 0.028 | | | | |
| | *Selected* | $\sigma = 0.2$ | 0.146 | | | | | |
| | | $\sigma = 0.5$ | | *0.188* | | 0.033 | | |
| | | $\sigma = 1.0$ | | | *0.188* | 0.033 | 0.009 | |
| **Gradient Landscape** | | | **Between-Population** | | | | | |
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Between** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | *0.183* | | | | | |
| | | $\sigma = 1.0$ | *0.217* | 0.072 | | | | |
| | *Selected* | $\sigma = 0.2$ | *0.161* | | | | | |
| | | $\sigma = 0.5$ | | 0.131 | | **0.259** | | |
| | | $\sigma = 1.0$ | | | *0.152* | *0.243* | 0.107 | |

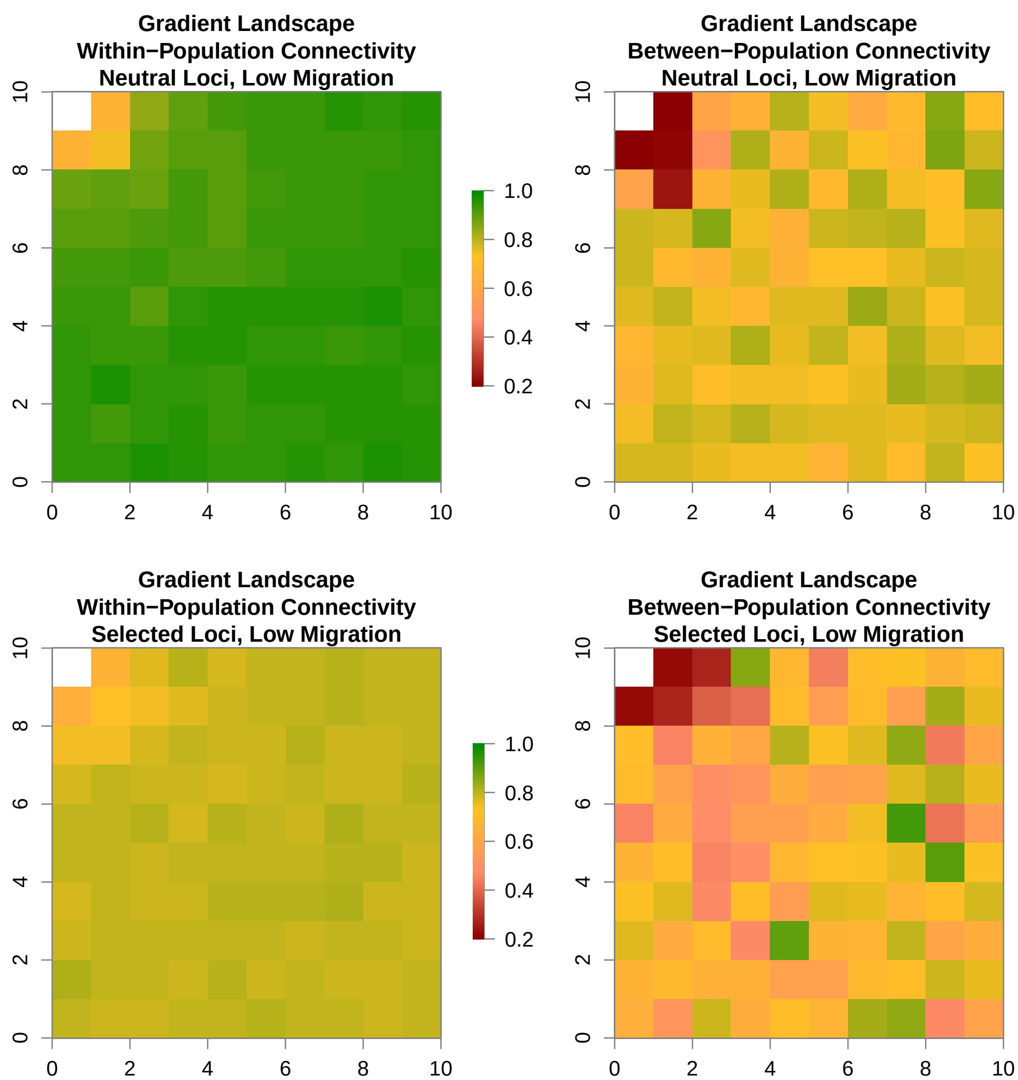



**Figure D.2:** *Within- and between-population connectivity for genotypes simulated on the* gradient *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.2 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

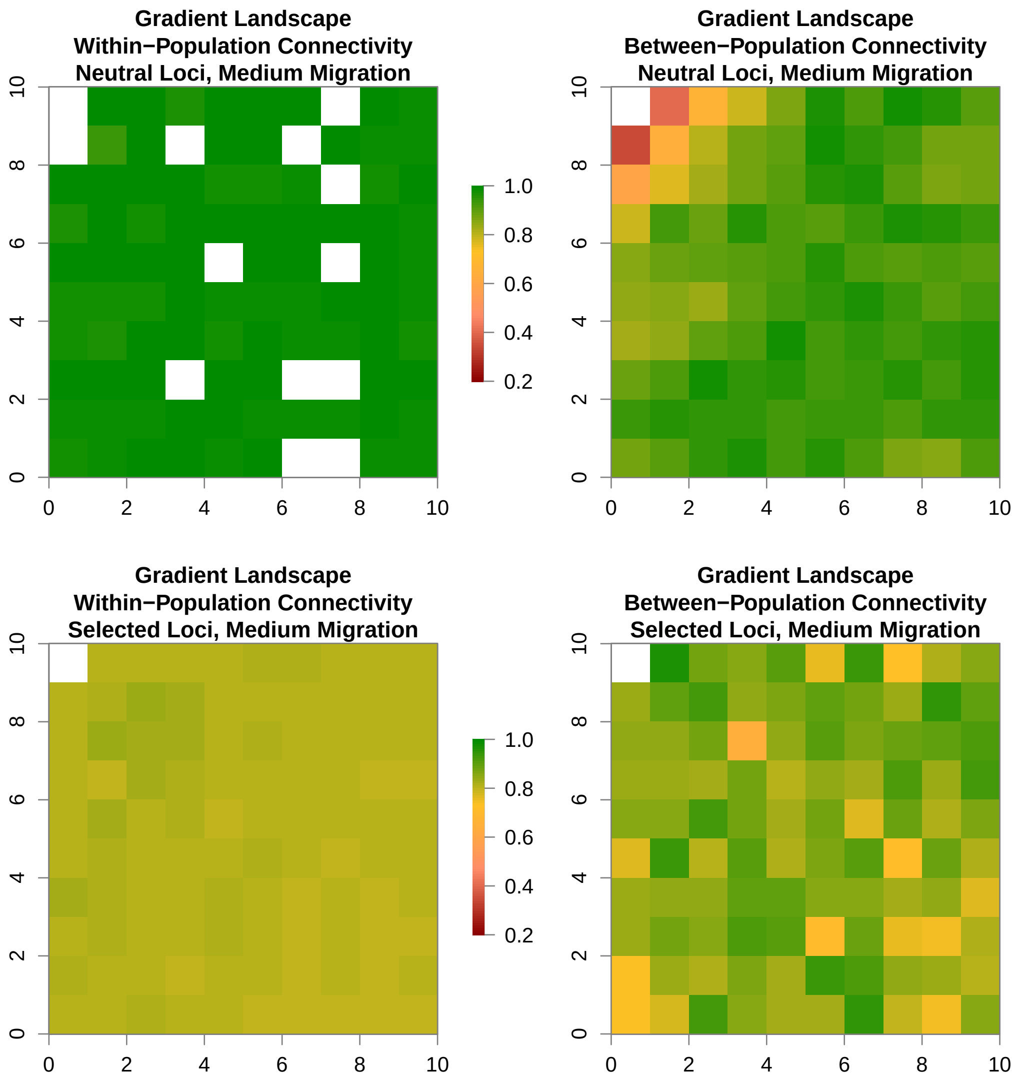



**Figure D.3:** *Within- and between-population connectivity for genotypes simulated on the* gradient *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.5 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

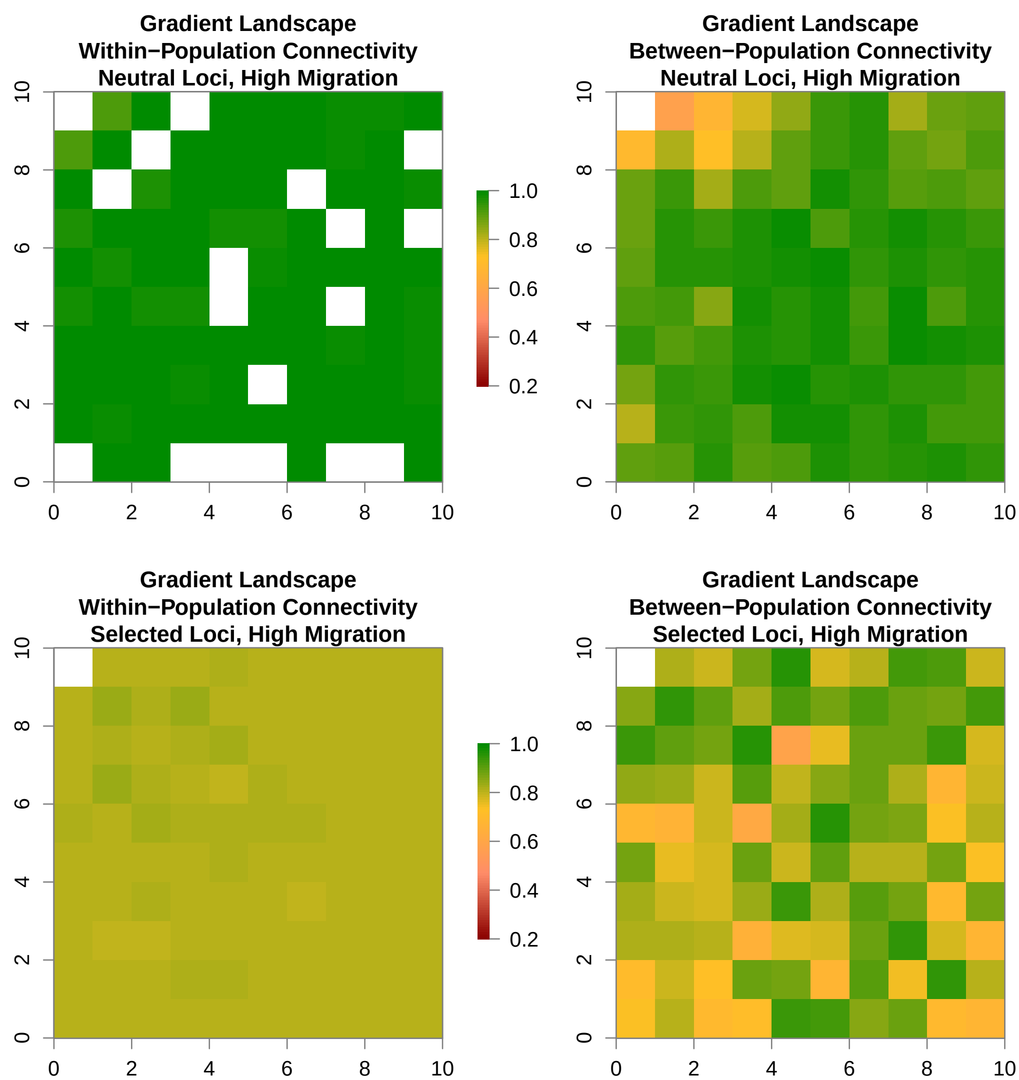



**Figure D.4:** *Within- and between-population connectivity for genotypes simulated on the* gradient *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 1.0 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

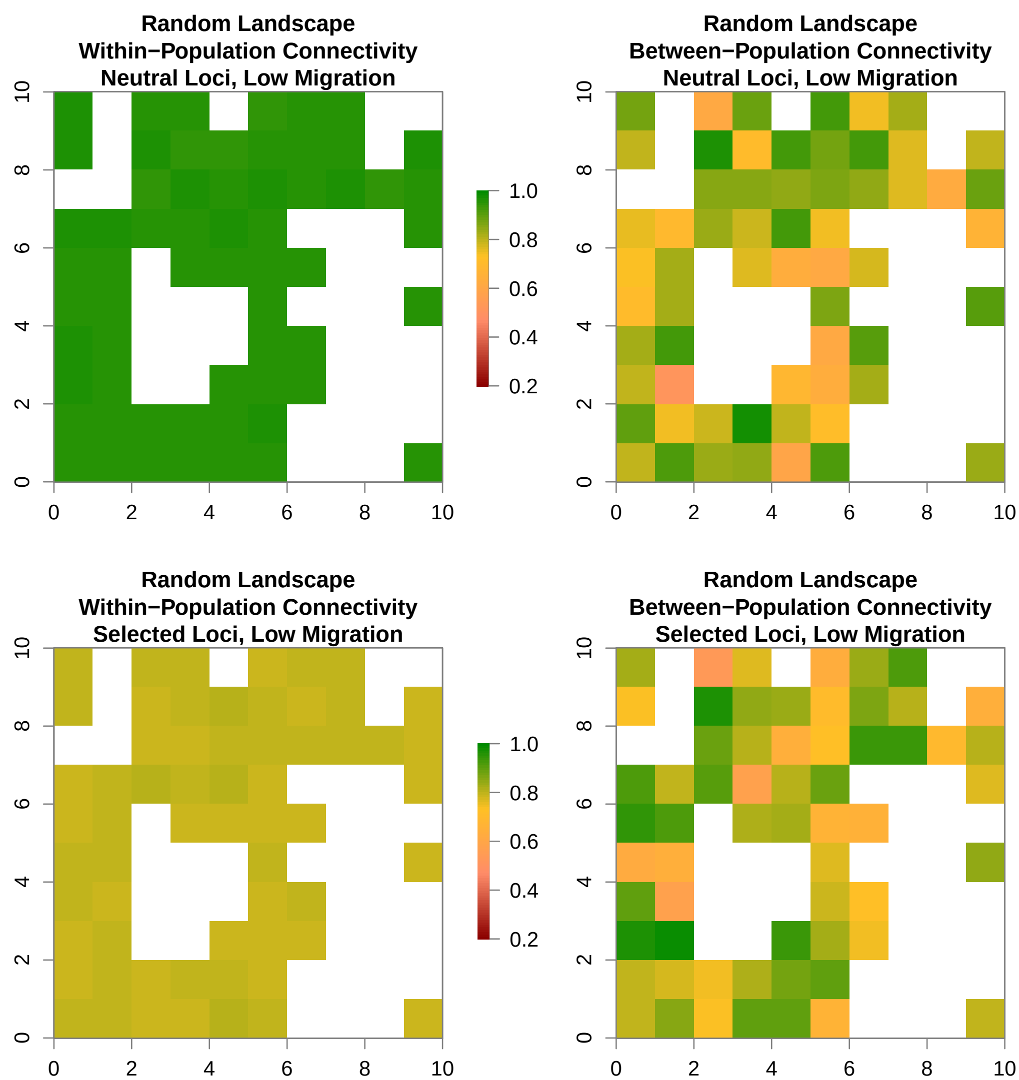



**Figure D.5:** *Within- and between-population connectivity for genotypes simulated on the* random *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.2 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

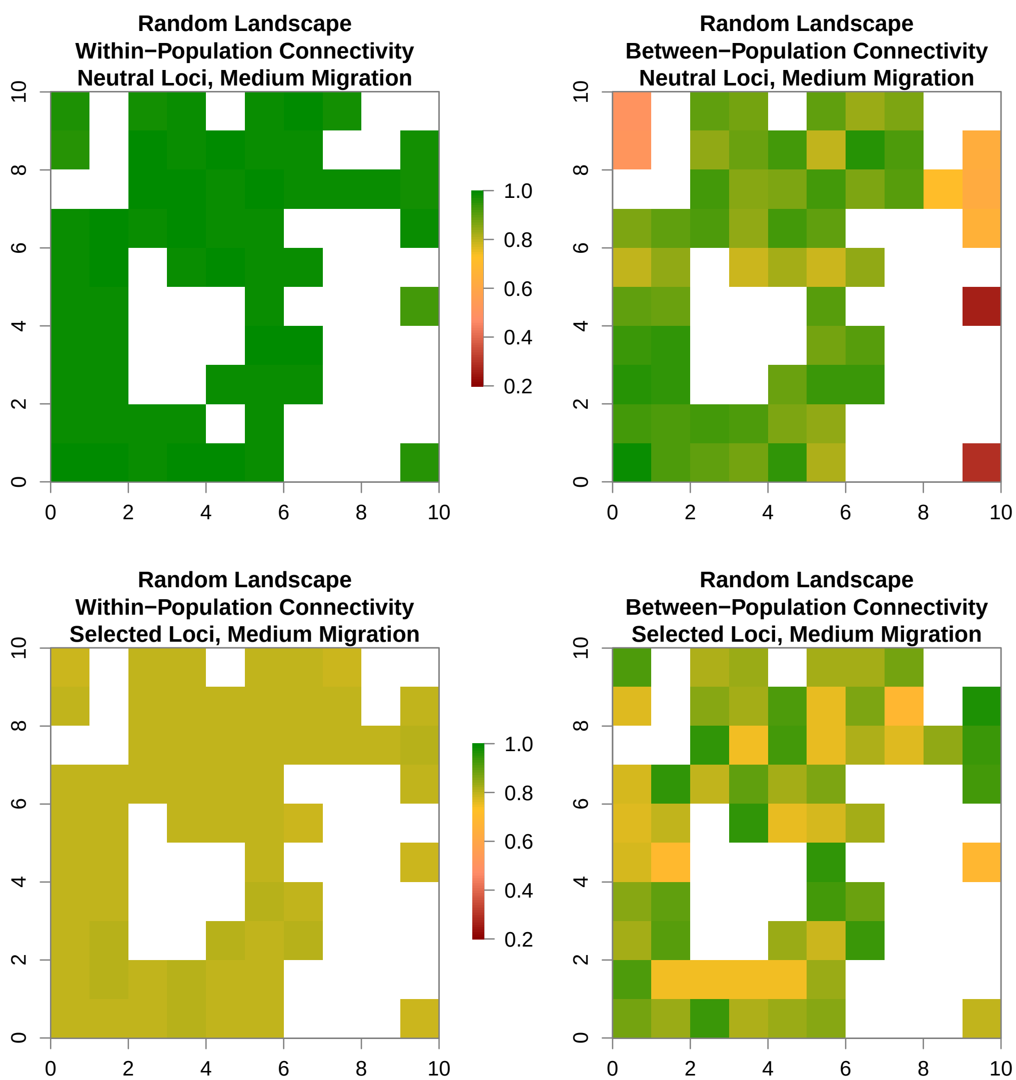



**Figure D.6:** *Within- and between-population connectivity for genotypes simulated on the* random *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.5 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

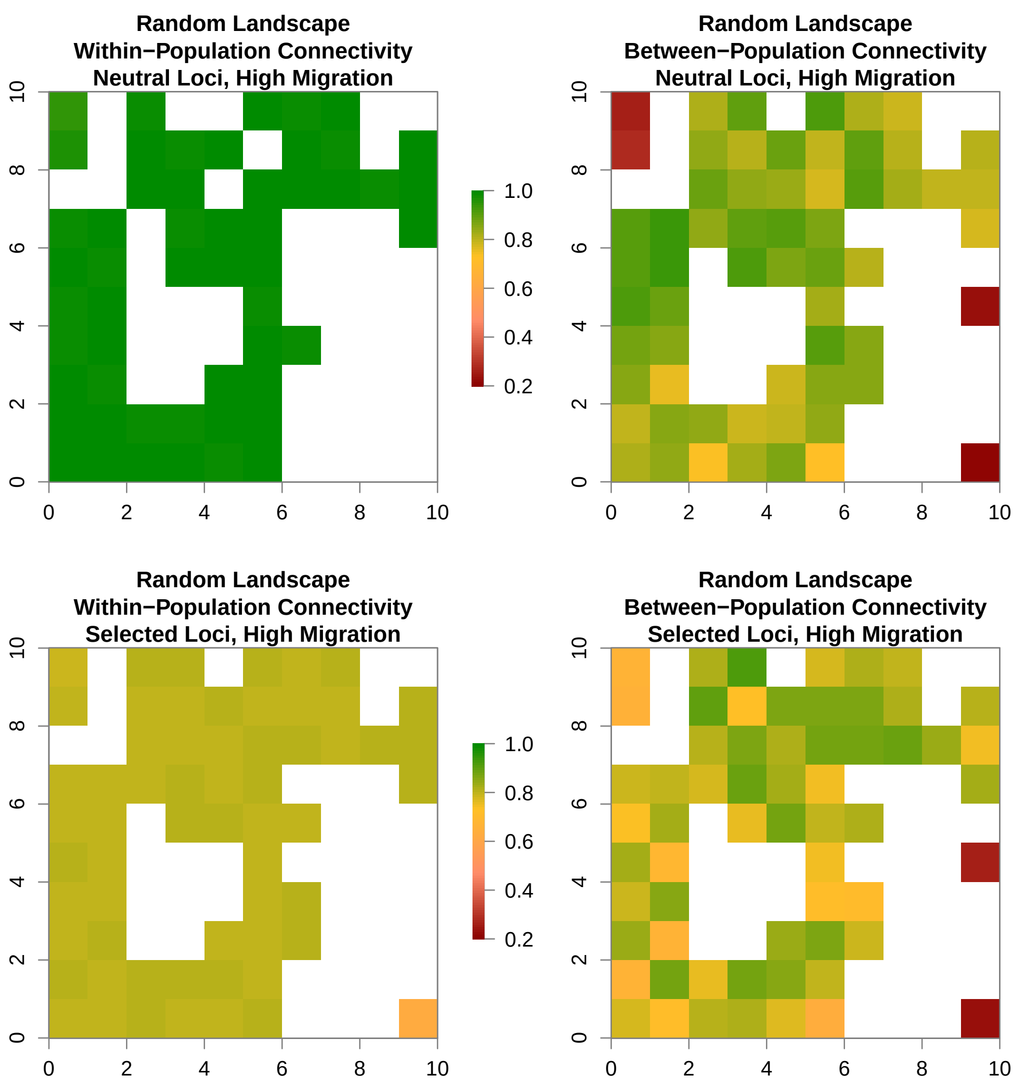



**Figure D.7:** *Within- and between-population connectivity for genotypes simulated on the* random *landscape*. The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 1.0 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

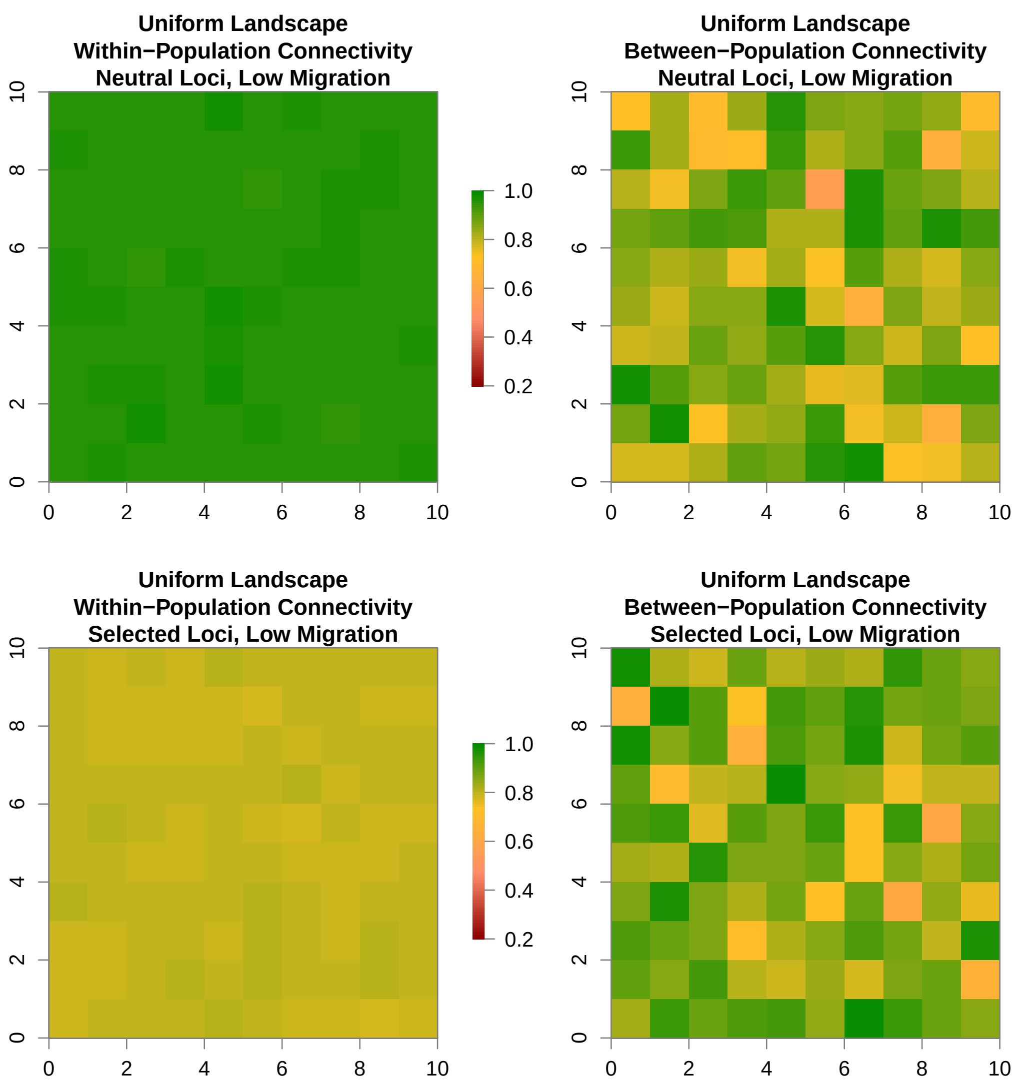



**Figure D.8:** *Within- and between-population connectivity for genotypes simulated on the* uniform *landscape.* The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.2 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

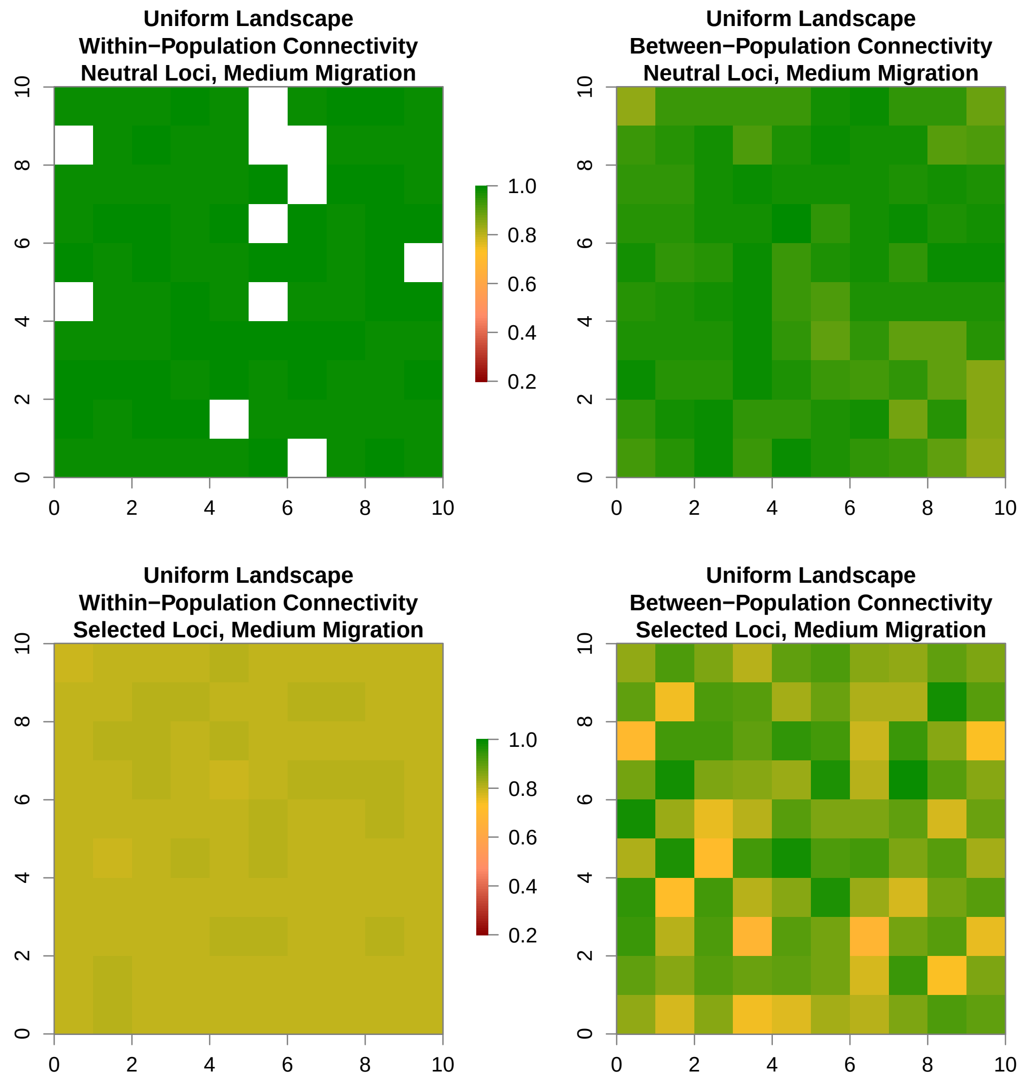



**Figure D.9:** *Within- and between-population connectivity for genotypes simulated on the* uniform *landscape.* The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 0.5 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

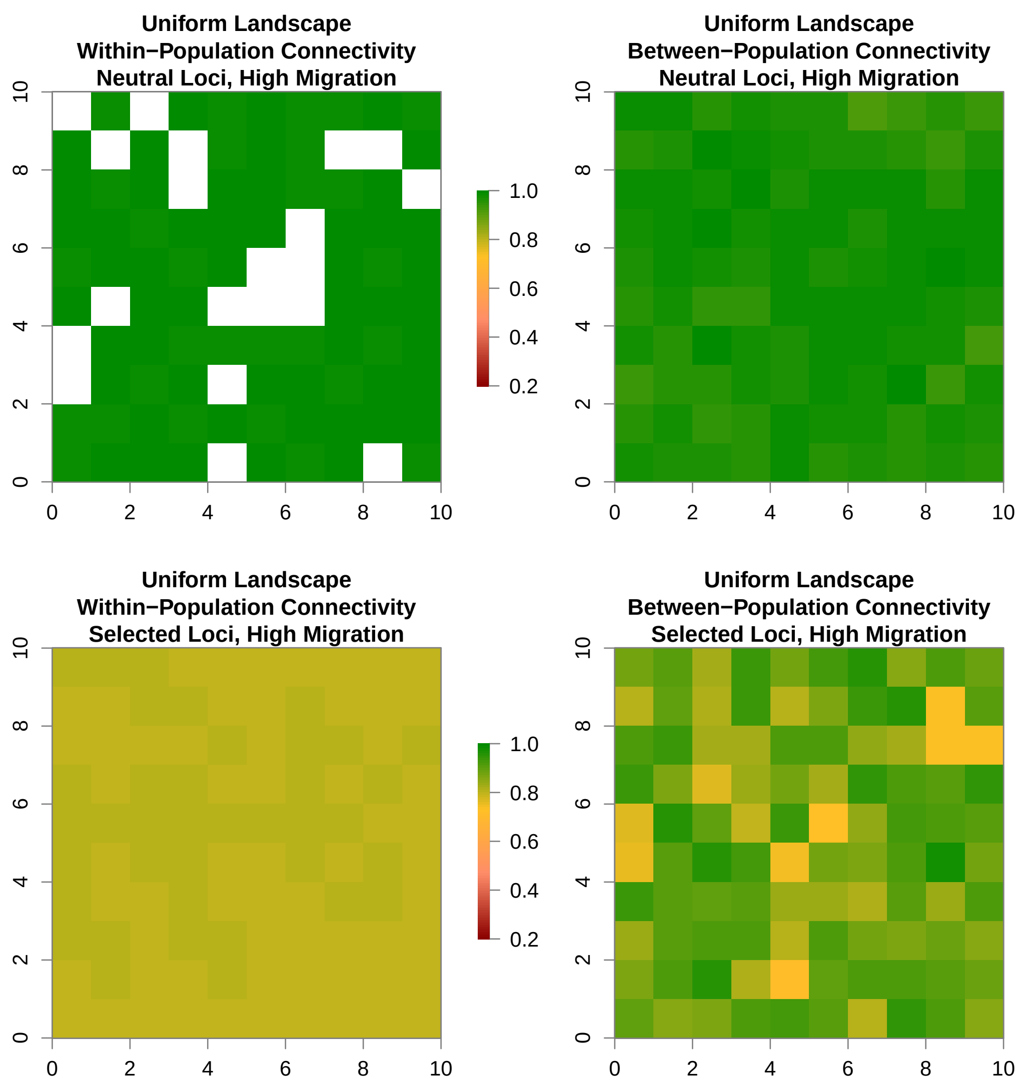



**Figure D.10:** *Within- and between-population connectivity for genotypes simulated on the* uniform *landscape.* The top two panels show within- and between-population connectivity for neutral loci, while the bottom two show connectivity for loci under selection (s = 0.1). The simulations shown here were performed with σ = 1.0 for the dispersal kernel. Connectivity is represented on a scale from 0.2 (dark red) to 1 (dark green).

**Table D.2:** *Root mean square error (RMSE) of connectivity comparisons for the* random *landscape*. RMSE values shown here were used to quantify differences for within- and between-population connectivity based on neutral versus non-neutral (s = 0.1) loci for different degrees of long-distance dispersal (σ = 0.2, 0.5, and 1). RMSE values greater than 0.150 are italicized, whereas values greater than 0.250 are shown in bold.

| **Root Mean Square Error of Functional Connectivity** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Random Landscape** | | | **Within-Population** | | | | | |
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Within** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | 0.033 | | | | | |
| | | $\sigma = 1.0$ | 0.037 | 0.010 | | | | |
| | *Selected* | $\sigma = 0.2$ | *0.168* | | | | | |
| | | $\sigma = 0.5$ | | *0.194* | | 0.008 | | |
| | | $\sigma = 1.0$ | | | *0.195* | 0.024 | 0.021 | |
| **Random Landscape** | | | **Between-Population** | | | | | |
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Between** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | *0.188* | | | | | |
| | | $\sigma = 1.0$ | *0.200* | 0.095 | | | | |
| | *Selected* | $\sigma = 0.2$ | *0.154* | | | | | |
| | | $\sigma = 0.5$ | | *0.155* | | 0.141 | | |
| | | $\sigma = 1.0$ | | | 0.109 | *0.162* | 0.146 | |

**Table D.3:** *Root mean square error (RMSE) of connectivity comparisons for the* uniform *landscape*. RMSE values shown here were used to quantify differences for within- and between-population connectivity based on neutral versus non-neutral (s = 0.1) loci for different degrees of long-distance dispersal (σ = 0.2, 0.5, and 1). RMSE values greater than 0.150 are italicized, whereas values greater than 0.250 are shown in bold.

**Root Mean Square Error of Functional Connectivity**

| **Uniform Landscape** | | | **Within-Population** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Within** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | 0.035 | | | | | |
| | | $\sigma = 1.0$ | 0.037 | 0.008 | | | | |
| | *Selected* | $\sigma = 0.2$ | *0.169* | | | | | |
| | | $\sigma = 0.5$ | | *0.198* | | 0.009 | | |
| | | $\sigma = 1.0$ | | | *0.198* | 0.010 | 0.005 | |

| **Uniform Landscape** | | | **Between-Population** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Neutral* | | | *Selected (s = 0.1)* | | |
| | | | *0.2* | *0.5* | *1.0* | *0.2* | *0.5* | *1.0* |
| **Between** | *Neutral* | $\sigma = 0.2$ | | | | | | |
| | | $\sigma = 0.5$ | 0.144 | | | | | |
| | | $\sigma = 1.0$ | *0.160* | 0.041 | | | | |
| | *Selected* | $\sigma = 0.2$ | 0.123 | | | | | |
| | | $\sigma = 0.5$ | | 0.121 | | 0.121 | | |
| | | $\sigma = 1.0$ | | | 0.115 | 0.104 | 0.101 | |

## VITA

Born in Geislingen, Germany in 1980, and having moved to Kosovo in 1985, Chaz Hyseni graduated from Frang Bardhi High School in Mitrovica, Kosovo in 1999. He then went on to work for three years as an interpreter for the United Nations Mission in Kosovo. In 2002, when he was admitted to Yale University, he moved to the United States to start college. He received a B.A. in Environmental Studies in 2007. As part of his degree, he traveled to the Galapagos Islands in the summer of 2006, and completed his thesis, "Galapagos giant tortoise conservation on the island of Santa Cruz: morphological and genetic distinctiveness of a newly discovered taxon" under the supervision of Dr. Adalgisa Caccone. Upon graduation, he worked as a research assistant in the lab of Dr. Caccone until 2012. After a one-year stint at Cornell University, he enrolled in the Department of Biology at the University of Mississippi in January 2014, under the supervision of Dr. Ryan Garrick.