Species Range Shifts in Dynamic Geological and Climatic Landscapes: Studies in Temperate and Tropical Trees

by

Jordan Brian Bemmels

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2018

Doctoral committee:

Professor Christopher W. Dick, Chair
Associate Professor Inés Ibáñez
Professor L. Lacey Knowles
Assistant Professor Stephen A. Smith

Jordan B. Bemmels

jbemmels@umich.edu

ORCID ID: 0000-0001-9996-6996

ACKNOWLEDGEMENTS

Chris Dick, my advisor, for welcoming me into his lab and providing his support in any research avenue I wished to pursue. | **Lacey Knowles**, if not an official co-advisor then an honourary one. | **Inés Ibáñez** and **Stephen Smith**, for their guidance and feedback. | **Na Wei**, for the mentorship at critical times. | **Nancy Garwood, Lúcia Lohmann, Alison Nazareno, Joaquín Ortego, Álvaro Pérez, Simon Queenborough, Renato Valencia**, and **Joe Wright**, for the collaborations. | **Pascal Title**, for taking collaboration the extra mile. | **The National Science Foundation (Doctoral Dissertation Improvement Grant; Graduate Research Fellowships Program), the Natural Sciences and Engineering Research Council of Canada (Postgraduate Scholarship-M)**, **Rackham Graduate School**, and **the University of Michigan Department of Ecology and Evolutionary Biology**, for financial support. | **Berea College Forest, Connecticut State Forests (SFs), Daniel Boone National Forest (NF), Davey Crockett NF, French Creek State Park (SP), George Washington NF, Global Biodiversity Information Facility, Holly Springs NF, Hoosier NF, Jefferson NF, Kisatchie NF, Land Between the Lakes National Recreation Area, Lower Wisconsin State Riverway, Mark Twain NF, Matthaei Botanical Gardens, Monongahela NF, Murray State University Hancock Biological Station, Nantahala NF, New York SFs, Oconee NF, Ozark NF, Pontificia**

Universidad Católica del Ecuador, Prentice Cooper SF, Shimek SF, Stephen F. Austin State University and Experimental Forest, Smithsonian Tropical Research Institute, Society of Ontario Nut Growers, Sumter NF, Talladega NF, Tombigbee NF, Torreya SP, UMich E.S. George Reserve, UMich Logistics Transportation and Parking, and **Uwharrie NF**, for making fieldwork and use of field datasets possible. | **The Centre for Applied Genomics, Raquel Marchán-Rivadeneira** at the **UMich Genomic Diversity Lab**, and the **UMich DNA Sequencing Core**, for making labwork possible. | Members of the **Dick and Knowles labs**, for both technical advice and friendships. | Faculty, staff, fellow grad students, and postdocs in **EEB**, whom I interacted with over the years and who made this a fantastic place to go to graduate school. | **Pame Martínez, Marian Schmidt, Déa Thomaz**, and **Oscar Vargas**, whose friendships made Ann Arbor an exciting place to live. | **Breaden Belcher** and **Rafael D'Andrea**, co-adventurers. | **My family**, for the love and support.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

ABSTRACT

Spatial patterns in the geographic distributions of ecosystems, species, and genetic variation are the result of both ecological conditions and evolutionary dynamics that have unfolded in a long, historical process. Attempts to account for spatial patterns in biodiversity provided some of the earliest inspiration for the development of the theory of evolution, but more than a century and a half later, biologists are still discovering precisely how and why these patterns arise. As a field, historical biogeography has emphasized the importance of Earth's geological and climatic history in understanding dynamic spatial biodiversity patterns. It is well known that species and populations respond to geological and climatic change, but the details about how these responses unfolded are often vaguely specified, and the precise ecological mechanisms mediating them are often unknown. My dissertation tests specific hypotheses about historical range shifts and their genetic consequences, using recent advances such as next-generation sequencing, demographic and coalescent modelling, Approximate Bayesian Computation, and large datasets of species occurrence records and species traits. This work spans several arboreal study systems, and aims to provide new insight into unanswered questions and controversial historical biogeographic hypotheses.

I begin by considering the effects of glaciation on range shifts in two hickory species from eastern North America. It is not precisely known where most temperate

deciduous tree species from this region survived the Last Glacial Maximum, and whether northern populations existed and contributed to postglacial recolonization. I show that both species were likely fairly geographically widespread and genetically connected, and that postglacial recolonization occurred from a northern source in one species and a southern source in another. Next, I develop and test competing hypotheses about the ecological factors mediating postglacial range shifts in canyon live oak from California. I aim to attain mechanistic insight into the drivers of historical range shifts in this geographically complex Mediterranean-climate region, and find that summer drought tolerance was likely a key ecological factor mediating these shifts. Lastly, I zoom out in scale and ask how long-term migrational responses to geological and climatic change affect the assembly of entire communities. I examine the biogeographic distributions of >1,000 Neotropical rainforest woody plant species, and show that traits such as drought tolerance and elevational range have impacted the ability of species to disperse across or around the biogeographic filter created by Pliocene uplift of the northern Andes and formation of dry habitats in northeastern South America.

Overall, my dissertation work has revealed the importance of species-specific responses to geological and climatic change, and how these responses affect the geographic distribution of biodiversity (from genetic to species to community levels). I test hypotheses concerning the geographic locations from which range shifts occurred (in hickories), the ecological factors mediating range shifts in complex environments (in canyon live oak), and the community-wide impact of species traits on biogeographic dispersal (in Neotropical trees). This work contributes to a growing body of knowledge

helping transform historical biogeography from a realm of broad patterns, into a field where new insight can be gained by accounting for species-specific histories and the ecological processes that mediate range shifts.

CHAPTER I

Introduction

A historical perspective on biogeography

Explaining how biodiversity is spatially distributed across the globe is a fundamental question in the study of ecology and evolution, and has interested many of the foremost biologists and geographers of the past centuries. It was once widely held that the flora and fauna characteristic of different regions were entirely the product of physical and climatic conditions, and that different regions with similar climates could be expected to have similar biotas (Nelson, 1978). However, as the biotas of different continents became better known to biologists throughout the 18th and 19th centuries, it became increasingly clear that regions that were geographically distant but with nearly identical climates typically shared few species in common (Nelson, 1978).

Once it became apparent that different regions had different floras and faunas, a central question became: why? The observation of differences among regions was a key piece of evidence (Costa, 2009; Gunnell, 2013) in the development of the theory of evolution by Charles Darwin and Alfred Russel Wallace (Darwin & Wallace, 1858; Darwin, 1859). Importantly, the theory of evolution – which suggested that species were derived from formerly existing species that presumably had their own geographic

distributions – provided a mechanistic explanation for the development of different biotas in different regions, and suggested that fully accounting for the earth's biodiversity would require a historical perspective. Although Darwin and other early evolutionary biologists were aware that not only life, but also the geological features of the earth itself had slowly changed over time (Lyell, 1830), they did not realize the extent to which geological change continually influenced evolution, and instead primarily focused on the role of dispersal limitation among regions assumed to be relatively geologically static over evolutionary time scales (Nelson, 1978). It was not until development of the theory of plate tectonics in the mid-20th century that it became clear that the earth's geological history was a key driver of biogeographic patterns, or as Léon Croizat put it, that "Earth and life evolved together" (Croizat, 1962, p. 46; cited in Nelson, 1978). The development of historical climatology, including the discovery that the earth's climate had undergone radical changes such as those of the most recent Ice Age (Agassiz, 1840), additionally suggested that species' geographic distributions were dynamic, and had responded to not only geological but also climatic change.

At the same time that the foundational principles of historical biogeography were being worked out to explain the geographic distributions of different species, parallel developments were leading to the realization that intraspecific phenotypic variation exists and tends to be geographically distributed. The knowledge that different natural populations of the same species differ subtly in their inherent characteristics is surprisingly old – far predating any modern conceptualizations of genetics or inheritance. In 1745, Henri-Louis Duhamel du Monceau began the first known experimental research into geographic variation in growth-related traits among tree populations, using

a common garden of *Pinus sylvestris* populations collected from across Europe (Langlet, 1971). Just as early biologists suspected that climates alone determined species' geographic distributions, so too did early botanists believe that local climates and physiography were the direct cause of intraspecific variation among plant populations (Langlet, 1971).

If we consider only observable phenotypic traits related to growth and survival, these early botanists were not far off from the modern view – most temperate tree populations do exhibit geographic differences in fitness-related traits and are genetically adapted to local climatic conditions (Savolainen et al., 2007; Alberto et al., 2013; Aitken & Bemmels, 2016). However, advances in genetics, including the discovery of the principles of Mendelian inheritance (Mendel, 1866), the modern synthesis of evolutionary biology (Huxley, 1942), and the development of the neutral theory of molecular evolution (Kimura, 1983), eventually led to the view that most species contain substantial molecular genetic variation, and most of this variation is adaptively neutral. Due to inheritance of genetic material from one generation to the next, it was soon realized that the geographic distribution of neutral genetic variation reflects historical, microevolutionary relationships among populations. The field of phylogeography did not develop until the end of the 20th century, but it greatly expanded the scope of biogeography by demonstrating that neutral genetic variation could be examined in order to uncover signatures of how climate change and geological context have influenced the demographic history of individual species (Avise et al., 1987; Avise, 2009; Hickerson et al., 2010).

The importance of geoclimatic history

One overarching theme revealed from the brief perspective on historical biogeography presented above is that history matters. Evolution is inherently a historical discipline, but biogeography makes the additional claim that in order to explain evolutionary and ecological patterns, it is necessary to also consider earth's *geological* and *climatic* history. Linking evolutionary history to geology and climate, or understanding how "Earth and life evolved together" (Croizat 1962, p. 46), is a central goal of biogeography and a key motivation behind my dissertation. I aim to explore how species have responded to geological and climatic (hereafter, "geoclimatic") change, and the impacts of these responses on patterns of genetic variation, species distributions, and regional community composition.

The role of geoclimatic history in generating macroevolutionary and global patterns is well recognized, although the limitations of its power to explain broad-scale phenomena continue to be debated. It is well known that the distribution of lineages across continents reflects their geographic origins and historical opportunities for dispersal. For example, many plant families (e.g., Araucariaceae, Nothofagaceae, Podocarpaceae) are co-distributed across southern South America, New Zealand, and New Guinea, which seems like an odd pattern until one considers the likely origin of these families in the humid southern Gondwanan forest that once connected these distant regions via Antarctica (Cox et al., 2001). Although terrestrial connections may explain the cross-continental distributions of many lineages, rare events like oceanic dispersal may also explain such distributions in other taxa (Dick et al., 2007; Dick &

Heuertz, 2008).

Geoclimatic history has also had a large influence on species richness patterns. For example, mammal species richness is higher in tectonically active regions than inactive regions, likely due in part to increased population fragmentation and speciation rates in tectonically active regions (Badgley, 2010). While geoclimatic change can be a driver of speciation, regions that experience extreme geoclimatic change may be at risk for increased extinction rates. For example, variation in extinction rates among continents due to differences in the severity of Pleistocene glaciation could partly explain why East Asian temperate forests have higher tree species diversity than those from eastern North America or Europe (Qian & Ricklefs, 2000). Nonetheless, present-day climates may also partly explain differences in temperate tree diversity among these continents (Adams & Woodward, 1989), highlighting how it can often be difficult to distinguish between biogeographic and ecological explanations for broad-scale patterns. The latitudinal diversity gradient – paradoxically one of the most well studied yet enigmatic global patterns in biodiversity – is another case where the relative importance of history vs. ecology in generating and maintaining biogeographic patterns continues to be debated. The time-area hypothesis proposes that tropical areas have the highest biodiversity because for most of Earth's history the majority of its land surface had a tropical climate, allowing for a greater accumulation of species adapted to tropical conditions (Fine & Ree, 2006). However, numerous ecological hypotheses to explain the latitudinal diversity gradient that are unrelated to geoclimatic history have also been proposed (Willig et al., 2003).

While the importance of considering history has long been recognized in

macroevolutionary studies, prior to the emergence of the field of phylogeography (Avise et al., 1987) – which I consider a subfield within historical biogeography (Arbogast & Kenagy, 2001) – very few population biologists adopted a historical perspective in their work (Hickerson et al., 2010). However, the explosion of phylogeographic research over the past decades has seen biologists explicitly address questions about the histories of individual species and populations. For example, the effects of Pleistocene glaciation on temperate taxa continues to be a major topic of phylogeographic interest (Shafer et al., 2010; Qiu et al., 2011; Tzedakis et al., 2013; Lumibao et al., 2017). Clear expectations have been developed and tested regarding the genetic consequences of population fragmentation into distinct glacial refugia and recolonization of formerly glaciated or otherwise uninhabitable regions from these refugia. Some of these expectations include elevated genetic diversity in refugial regions, high genetic differentiation among refugia, declining genetic diversity with increasing latitude in the Northern Hemisphere due to founder effects as population migrated northward, and elevated genetic diversity in areas of secondary contact and admixture (Hewitt, 2000, 2004; Petit et al., 2003).

Understanding genetic expectations of different scenarios is critical to distinguishing among competing hypotheses about species' biogeographic histories. Descriptive interpretation of genetic patterns, particularly from mitochondrial DNA in animals and chloroplast DNA in plants, was a mainstay of early phylogeography (Knowles & Maddison, 2002; Avise, 2009; Hickerson et al., 2010). However, genetic expectations of specific scenarios are not always known. Conceptual and methodological advances have increased our ability to statistically test complex hypotheses (Knowles, 2009), and have made phylogeography an increasingly

integrative discipline (Hickerson et al., 2010). Many phylogeographic studies now rely on multiple lines of evidence, including genetic data, the fossil record, and species distribution models (Gavin et al., 2014). In addition, the next-generation sequencing revolution (Koboldt et al., 2013) has moved phylogeography beyond inference from only a few loci to characterization of genome-wide patterns of genetic variation (McCormack et al., 2013). Finally, application of coalescent theory has allowed researchers to statistically account for stochasticity in lineage sorting and mutation (Knowles, 2009; Hickerson et al., 2010), providing an improved understanding of the range of genetic patterns that might be expected to result from a given historical scenario.

Perhaps one of the most exciting technical advances in recent years has been the development of coalescent simulation-based methods that integrate information from multiple sources to generate genetic predictions of different scenarios that can be tested with Bayesian statistical techniques (e.g., Neuenschwander et al., 2008; Carnaval et al., 2009; He et al., 2013, 2017). Because these approaches rely on comparing simulated data to empirical data without the need to calculate likelihoods from population genetic theory (Beaumont et al., 2002; Csilléry et al., 2010), they can be used to test a potentially unlimited number of scenarios of arbitrary complexity, assuming the existence of the technical ability to perform simulations and of sufficiently informative genetic summary statistics for distinguishing among scenarios (Peter & Slatkin, 2013; Alvarado-Serrano & Hickerson, 2016). It is now becoming increasingly possible to statistically compare support for complex, competing hypotheses that only a few years ago might have been essentially indistinguishable (Knowles, 2009). In addition, integrative methodological advances are allowing researchers to test species-

specific hypotheses that highlight the importance of ecological traits in mediating phylo- and biogeographic histories (Massatti & Knowles, 2016; Papadopoulou & Knowles, 2016).

Overview of chapters

Biogeography has thus developed from a field primarily documenting broad patterns, into one testing detailed, ecologically motivated hypotheses about the impacts of geoclimatic history at both the macro- and microevolutionary scale. Within this context, the thrust of my dissertation is to explore how geoclimatic history has shaped the geographic distribution of biodiversity, from the level of genetic diversity to species ranges to regional community composition. I make use of a variety of methods, including population genomics, ecological niche modelling, statistical phylogeography, coalescent modelling, and macroecology, and consider multiple study systems and spatiotemporal contexts. Each chapter has its own specific question of interest, but my overall goal is to make use of novel datasets, methodologies, and statistical techniques to test an unsolved or controversial historical biogeographic hypothesis in forest tree species. Because forest trees are often dominant species and may be ecological indicators of the types of communities they inhabit, they serve as ideal model systems to understand historical patterns and processes potentially relevant to forest ecosystems as a whole. Where possible, I attempt to bridge gaps between historical and ecological explanations for biodiversity patterns, by testing species-specific, ecologically motivated biogeographic hypotheses. My work aims to move away from

vague, general statements about how geoclimatic change has impacted each study system, and instead to provide detailed insight into testable historical scenarios.

In Chapters 2 and 3, I explore where two eastern North American (ENA) hickories (bitternut hickory, *Carya cordiformis*; shagbark hickory, *Carya ovata*) survived the Last Glacial Maximum (LGM) ca. 21,500 years ago, and how they recolonized their current ranges after glacial retreat. In Chapter 2, I examine the population genetic structure of both species using rangewide datasets of >1,000 nuclear single nucleotide polymorphisms (SNPs). Although in this initial chapter I interpret patterns *ad hoc* without testing any specific model, a general presentation of population genomic structure of a widespread deciduous tree from ENA has to my knowledge never previously been published. A large number of different locations have been proposed as glacial refugia for temperate deciduous trees from ENA (Jackson et al., 2000; Soltis et al., 2006; Jaramillo-Correa et al., 2009), showing that the LGM distribution of species from this region is still controversial and more detailed exploration of the neutral population genomics of widespread, common tree species that could serve as model taxa for this region is warranted. One of the largest unresolved controversies in temperate phylogeography is whether temperate deciduous trees survived the LGM in climatically and topographically constrained microrefugia located much farther north than southern locations that have traditionally been identified as major refugia (Stewart & Lister, 2001; McLachlan et al., 2005; Provan & Bennett, 2008; Tzedakis et al., 2013; Rull, 2014). In Chapter 3, I pair the previous chapter's genomic dataset with spatially explicit demographic and coalescent simulations and Approximate Bayesian Computation (ABC) in the X-ORIGIN pipeline (He et al., 2017), in order to infer the geographic origins

of postglacial expansion. I explicitly use models designed to account for the possibility of expansion from a northern source.

While detecting the geographic origins of population expansion is a question with its own merits, another perhaps more fundamental question is not *wherefrom*, but rather, *why* range shifts occur. Certainly species distributions respond to climate change, but are there specific ecological factors or aspects of the changing environment that are most directly responsible for mediating climate-driven range shifts, and can these factors be distinguished? In Chapter 4, I address these questions in canyon live oak (*Quercus chrysolepis*) from the Mediterranean climates of the California Floristic Province. Given the high topographic and climatic complexity of this region, it is not clear *a priori* how tree species might be expected to have shifted in distribution in response to Pleistocene glacial cycles. I develop six hypotheses about which ecological factors may have mediated range shifts since the LGM and driven genetic structure in this species. I then test these hypotheses using integrative distributional, demographic and coalescent (iDDC) modelling (He et al., 2013) and ABC.

Finally, I turn my attention to tropical trees, deeper timescales, and whole communities in Chapter 5. Geoclimatic change has had a profound influence on the biogeographic history of Neotropical taxa. For example, the Miocene to Pliocene formation of the Isthmus of Panama (Coates & Stallard, 2013; Bacon et al., 2015; Montes et al., 2015) sparked the iconic Great American Biotic Interchange, allowing numerous taxa to disperse between North and South America (Gentry, 1982; Cody et al., 2010; Bacon et al., 2015). However, it is unclear how such a large proportion of lowland rainforest tree species have come to be distributed in both continents given that

the northern Andes and adjacent dry habitats are a biogeographic barrier to terrestrial dispersal between the Amazon and Central America for rainforest-adapted species. In this chapter, I continue my investigation of the ecological factors mediating range shifts and dispersal. In particular, I examine biogeographic distributions and species traits of >1,000 woody plant species from two lowland rainforest communities in the Amazon and Central America. I test whether species with certain ecological and life-history traits are more likely to be found on both sides of the Andes, which could suggest that the northern Andes region has acted as a biogeographic filter to dispersal between continents and impacted the assembly of regional species pools.

Co-authorship status

Specific contributions of co-authors are described in each chapter individually. Authorship on each chapter is as follows:

Chapter II. Jordan B. Bemmels (University of Michigan), Christopher W. Dick (University of Michigan).

Chapter III. Jordan B. Bemmels, L. Lacey Knowles (University of Michigan), Christopher W. Dick.

Chapter IV. Jordan B. Bemmels, Pascal O. Title (University of Michigan), Joaquín Ortego (Estación Biológica de Doñana), L. Lacey Knowles.

Chapter V. Jordan B. Bemmels, S. Joseph Wright (Smithsonian Tropical Research Institute), Nancy C. Garwood (Southern Illinois University), Simon A. Queenborough (Yale University), Renato Valencia (Pontificia Universidad Católica del

Ecuador; Smithsonian Tropical Research Institute), Christopher W. Dick.

Literature cited

Adams, J.M. & Woodward, F.I. (1989) Patterns in tree species richness as a test of the glacial extinction hypothesis. *Nature*, **339**, 699–701.

Agassiz, L. (1840) *Études sur les glaciers.* Jent et Gassmann, Neuchâtel.

Aitken, S.N. & Bemmels, J.B. (2016) Time to get moving: assisted gene flow of forest trees. *Evolutionary Applications*, **9**, 271–290.

Alberto, F.J., Aitken, S.N., Alía, R., González-Martínez, S.C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., & Savolainen, O. (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Alvarado-Serrano, D.F. & Hickerson, M.J. (2016) Spatially explicit summary statistics for historical population genetic inference. *Methods in Ecology and Evolution*, **7**, 418–427.

Arbogast, B.S. & Kenagy, G.J. (2001) Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography*, **28**, 819–825.

Avise, J.C. (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., & Saunders, N.C. (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Bacon, C.D., Silvestro, D., Jaramillo, C., Smith, B.T., Chakrabarty, P., & Antonelli, A. (2015) Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proceedings of the National Academy of Sciences*, **112**, 6110–6115.

Badgley, C. (2010) Tectonics, topography, and mammalian diversity. *Ecography*, **33**, 220–231.

Beaumont, M.A., Zhang, W., & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Carnaval, A.C., Hickerson, M.J., Haddad, C.F.B., Rodrigues, M.T., & Moritz, C. (2009)

Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

Coates, A.G. & Stallard, R.F. (2013) How old is the Isthmus of Panama? *Bulletin of Marine Science*, **89**, 801–813.

Cody, S., Richardson, J.E., Rull, V., Ellis, C., & Pennington, R.T. (2010) The Great American Biotic Interchange revisited. *Ecography*, **33**, 326–332.

Costa, J.T. (2009) The Darwinian revelation: tracing the origin and evolution of an idea. *BioScience*, **59**, 886–894.

Cox, C.B., Cottage, F., Close, B., & Barry, C. (2001) The biogeographic regions reconsidered. *Journal of Biogegraphy*, **28**, 511–523.

Croizat, L. (1962) *Space, Time, Form: the Biological Synthesis.* Caracas.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., & François, O. (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Darwin, C. (1859) *On the Origin of Species.* John Murray, London.

Darwin, C. & Wallace, A.R. (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London: Zoology*, **3**, 45–62.

Dick, C.W., Bermingham, E., Lemes, M.R., & Gribel, R. (2007) Extreme long-distance dispersal of the lowland tropical rainforest tree *Ceiba pentandra* L. (Malvaceae) in Africa and the Neotropics. *Molecular Ecology*, **16**, 3039–3049.

Dick, C.W. & Heuertz, M. (2008) The complex biogeographic history of a widespread tropical tree species. *Evolution*, **62**, 2760–2774.

Fine, P.V.A. & Ree, R.H. (2006) Evidence for a time‐integrated species‐area effect on the latitudinal gradient in tree diversity. *The American Naturalist*, **168**, 796–804.

Gavin, D.G., Fitzpatrick, M.C., Gugger, P.F., et al. (2014) Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, **204**, 37–54.

Gentry, A.H. (1982) Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleisotcene fluctuations, or an accident of the Andean orogeny? *Annals of The Missouri Botanical Garden*, **69**, 557–593.

Gunnell, G.F. (2013) Biogeography and the legacy of Alfred Russel Wallace. *Geologica Belgica*, **16**, 211–216.

He, Q., Edwards, D.L., & Knowles, L.L. (2013) Integrative testing of how environments

from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

He, Q., Prado, J.R., & Knowles, L.L. (2017) Inferring the geographic origin of a range expansion: latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program X-ORIGIN. *Molecular Ecology*, **26**, 6908–6920.

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

Hewitt, G.M. (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**, 183–195.

Hickerson, M.J., Carstens, B.C., Cavender-Bares, J., Crandall, K.A., Graham, C.H., Johnson, J.B., Rissler, L., Victoriano, P.F., & Yoder, A.D. (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Huxley, J. (1942) *Evolution: the Modern Synthesis.* Allen and Unwin, London.

Jackson, S.T., Webb, R.S., Anderson, K.H., Overpeck, J.T., Webb III, T., Williams, J.W., & Hansen, B.C.S. (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews*, **19**, 489–508.

Jaramillo-Correa, J.P., Beaulieu, J., Khasa, D.P., & Bosquet, J. (2009) Inferring the past from the present phylogeography structure of North American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, **39**, 286–307.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Knowles, L.L. (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Knowles, L.L. & Maddison, W.P. (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., & Mardis, E. (2013) The next-generation sequencing revolution and its impact on genomics. **155**, 27–38.

Langlet, O. (1971) Two hundred years genecology. *Taxon*, **20**, 653–721.

Lumibao, C.Y., Hoban, S.M., & McLachlan, J. (2017) Ice ages leave genetic diversity "hotspots" in Europe but not in Eastern North America. *Ecology Letters*, .

Lyell, C. (1830) *Principles of Geology.* John Murray, London.

Massatti, R. & Knowles, L.L. (2016) Contrasting support for alternative models of

genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. *Molecular Ecology*, **25**, 3974–3986.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., & Brumfield, R.T. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

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

Mendel, G.J. (1866) Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, **1866**, 3–47.

Montes, C., Cardona, A., Jaramillo, C., Pardo, A., Silva, J.C., Valencia, V., Ayala, C., Pérez-Angel, L.C., Rodriguez-Parra, L.A., Ramirez, V., & Niño, H. (2015) Middle Miocene closure of the Central American Seaway. *Science*, **348**, 226–229.

Nelson, G. (1978) From Candolle to Croizat: comments on the history of biogeography. *Journal of the History of Biology*, **11**, 269–305.

Neuenschwander, S., Largiadèr, C.R., Ray, N., Currat, M., Vonlanthen, P., & Excoffier, L. (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.

Papadopoulou, A. & Knowles, L.L. (2016) A paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences*, **113**, 8018–8024.

Peter, B.M. & Slatkin, M. (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Müller-Starck, G., Demesure-Musch, B., Palmé, A., Martín, J.P., Rendell, S., & Vendramin, G.G. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

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

Qian, H. & Ricklefs, R.E. (2000) Large-scale processes and the Asian bias in temperate plant species diversity. *Nature*, **407**, 180–182.

Qiu, Y.X., Fu, C.X., & Comes, H.P. (2011) Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics and Evolution*, **59**, 225–244.

Rull, V. (2014) Macrorefugia and microrefugia: a response to Tzedakis *et al. Trends in Ecology and Evolution*, **29**, 243–244.

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.

Shafer, A.B.A., Cullingham, C.I., Côté, S.D., & Coltman, D.W. (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.

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

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

Tzedakis, P.C., Emerson, B.C., & Hewitt, G.M. (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology and Evolution*, **28**, 696–704.

Willig, M.R., Kaufman, D.M., & Stevens, R.D. (2003) Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 273–309.

CHAPTER II

Genomic evidence of a widespread southern distribution during the Last Glacial Maximum for two eastern North American hickory species

Abstract

Aim

Phylogeographic studies of temperate forest taxa often infer complex histories involving population subdivision into distinct refugia during the Last Glacial Maximum (LGM). However, some temperate deciduous trees may have been broadly distributed in southeastern North America during the LGM. We investigate genome-wide genetic structure in two widespread eastern North American tree species to determine if range expansion from genetically isolated refugia or from a broader, less genetically subdivided region better explains their postglacial history.

Location

Eastern North America (ENA).

Taxa

Bitternut hickory (*Carya cordiformis* (Wangenh.) K.Koch) and shagbark hickory (*Carya ovata* (Mill.) K.Koch).

Methods

Genetic diversity and differentiation indices were calculated from >1,000 nuclear SNP loci genotyped in ca. 180 individuals per species sampled across ENA. Genetic structure was investigated using principle component analysis and genetic clustering algorithms. As an additional tool for inference, areas of suitable habitat during the LGM were predicted using species distribution models (SDMs).

Results

Populations across all latitudes showed similar levels of genetic diversity. Most genetic variation was weakly differentiated across ENA, with the exception of an outlier population of *Carya ovata* in Texas. Genetic structure in each species exhibited an isolation-by-distance pattern. SDMs predicted high LGM habitat suitability over much of the southeastern United States.

Main conclusions

Both hickory species likely survived the LGM in low-density populations that were broadly distributed across southeastern North America and not highly genetically differentiated, except that the range-edge Texas population of *Carya ovata* may represent a separate glacial refugium. Over most of ENA, genetic structure in both species is best explained by simple latitudinal range shifts and high gene flow among

populations, rather than expansions from multiple, genetically isolated refugia as is characteristic of taxa from other Northern Hemisphere temperate regions of the world.

Introduction

Temperate forest ecosystems have long served as models for understanding how historical forces give rise to population genetic structure in terrestrial organisms (Hewitt, 1999, 2000; Petit et al., 2003; Soltis et al., 2006; Shafer et al., 2010; Qiu et al., 2011). Migrational responses to Pleistocene glaciation often left large impacts on genetic structure, but the severity and effects of glaciation varied in different temperate regions of the world (Hewitt, 2000; Shafer et al., 2010; Qiu et al., 2011; Lumibao et al., 2017). In Europe, where many classic phylogeographic paradigms were first established (Lumibao et al., 2017), widespread temperate taxa typically retreated to glacial refugia in Mediterranean regions (Hewitt, 1999, 2000; Petit et al., 2003). Glacial refugia, as defined here, are relatively small, geographically distinct regions, among which genetic connectivity is low (Bennett and Provan 2008). After expanding out of refugia following glacial retreat, many European species experienced a progressive loss of genetic diversity due to founder effects during northward migration (Hewitt, 1999, 2000). However, mid-latitude areas often exhibit elevated genetic diversity due to admixture of lineages from different refugia (Petit et al., 2003).

In other temperate regions of the world, phylogeographic patterns were structured by very different geographies and glacial histories. In eastern North America (ENA), early studies tended to emphasize genetic breaks between populations

separated by rivers and mountain ranges (Soltis et al., 2006; Jaramillo-Correa et al., 2009). In western North America, major refugia existed in the Pacific Northwest and Beringia, with smaller refugia on offshore islands and between continental ice sheets (Shafer et al., 2010). In East Asia, responses to glaciation included not only latitudinal migration, but also elevational and longitudinal migration and *in situ* persistence (Qiu et al., 2011). The complexity of these classic paradigms has recently been expanded in all four Northern Hemisphere regions to include small, low-density cryptic refugia in areas previously thought unsuitable for habitation by temperate species (Stewart & Lister, 2001; Willis & Van Andel, 2004; McLachlan et al., 2005; Provan & Bennett, 2008; Qiu et al., 2011). However, the molecular and fossil evidence supporting the existence of cryptic refugia is not universally accepted (Tzedakis et al., 2013).

Compared to the other three Northern Hemisphere temperate forest regions, ENA is phylogeographically unique for at least three reasons. Firstly, its geography is relatively simple, characterized by a large contiguous landmass with only a single north-south mountain range of modest height (i.e., the Appalachians), and generally gradual transitions between ecosystem types. Secondly, latitudinal temperature gradients during the Last Glacial Maximum (LGM, ca. 21.5 ka; Jackson et al., 2000) were particularly steep, with warm areas located in close proximity to glaciers (Tzedakis et al., 2013). Thirdly, despite numerous phylogeographic studies, well-delineated glacial refugia generally shared by most species have not conclusively been identified. Proposed refugial locations include the Gulf Coast, the Atlantic Coast, Florida, Texas, the Ozark Plateau, the Lower Mississippi River Valley, the Appalachians, and interior areas near ice sheets (Griffin & Barrett, 2004; Magni et al., 2005; Soltis et al., 2006; Jaramillo-

Correa et al., 2009; Morris et al., 2010; Barnard-Kubow et al., 2015; McCarthy & Mason-Gamer, 2016; Peterson & Graves, 2016), which together sum to nearly the entire unglaciated region of eastern North America. While some species may have survived in one or more of these distinct refugia, other temperate taxa were likely not restricted to distinct LGM refugia, but were widespread over vast areas of the southeastern United States (Bennett, 1985; Magni et al., 2005; McLachlan et al., 2005; Peterson & Graves, 2016; Lumibao et al., 2017).

In addition to genetic evidence, the fossil record has provided valuable insight into vegetation dynamics in ENA since the LGM, but has not yet resulted in a definitive account of the phylogeographic history of temperate deciduous tree species for several reasons. Macrofossils of temperate deciduous trees are known from the Lower Mississippi River Valley (e.g., Delcourt et al. 1980), but there are few other LGM macrofossil sites available from areas with climates likely to have supported these taxa (Jackson et al. 2000). Fossil pollen of temperate deciduous tree species is broadly distributed during the LGM, but typically represents only a minor portion of the total pollen from most assemblages (Jackson et al. 2000). Instead, coniferous species (e.g., *Picea, Pinus*) dominate most pollen assemblages (Davis 1983, Jackson et al. 2000), and plant communities over large areas of ENA may have had no modern analog (Jackson et al. 2000, Jackson and Williams 2004). The lack of clearly identifiable temperate deciduous forest communities in the LGM fossil record suggests that localized glacial refugia may not have existed in ENA for these communities as a whole, or if they did exist, they may not be represented in the fossil record. Nonetheless, temperate deciduous species were evidently present in some conifer-dominated LGM

communities (Jackson et al. 2000), raising questions about their geographic ranges, population sizes, genetic connectivity among populations, and which populations contributed most to postglacial recolonization.

Given the diversity of phylogeographic hypotheses that have been evoked to explain genetic patterns in ENA taxa, studies assessing genome-wide patterns of genetic variation in widely distributed model species would provide valuable insight into the history of temperate deciduous trees from this region. Surprisingly, we are aware of no such studies, although Eckert et al. (2010) and Nadeau et al. (2015) have conducted genome-wide studies of more narrowly distributed conifers, and many non–genome-wide studies of temperate deciduous trees exist (e.g., McLachlan et al. 2005). Here, we use genome-wide genetic variation to examine the phylogeographic history of two widespread, ENA tree species: bitternut hickory (*Carya cordiformis* (Wagenh.) K.Koch) and shagbark hickory (*Carya ovata* (Mill.) K.Koch). We construct and analyze single-nucleotide polymorphism (SNP) datasets from nearly rangewide collections of each species in order to characterize geographic patterns of genetic diversity and differentiation across ENA, and build paleodistribution models in order to infer areas that may have had suitable climates for survival during the LGM. In particular, we aim to determine if genetic structure is best explained by recolonization from genetically isolated and geographically distinct refugia (and if so, where these refugia were located), or by expansion from a much larger region that was not strongly genetically subdivided.

Methods

Study species

Carya cordiformis and *Carya ovata* are wind-pollinated, animal-dispersed trees co-distributed from southern Quebec to eastern Texas (Fig. 1). Their ranges roughly correspond to the overall geographic distribution of temperate deciduous forests in ENA. *Carya ovata* additionally occurs in several small, disjunct populations in the Sierra Madre Oriental of northern Mexico (Little, 1971). *C. cordiformis* occupies many habitats but occurs most frequently on mesic soils and bottomlands (Smith, 1990), whereas *C. ovata* is common on a wider variety of sites but is most frequent on drier uplands (Graney, 1990).

Phylogeographic knowledge is completely lacking in *C. cordiformis*. In *C. ovata*, analysis of cpDNA haplotypes has revealed no clear pattern, as some haplotypes are widespread throughout the entire range and others are more spatially restricted, including in formerly glaciated areas (Lumibao et al., 2017). *Carya* pollen is not typically distinguished to the species level, but LGM-age pollen of *Carya* has been found at low density over large areas of the southeastern USA (Prentice et al., 1991; Jackson et al., 2000). *Carya* macrofossils dating to the LGM have been found as far north as western Tennessee (35ºN; Jackson et al., 2000), along with trace amounts of pollen in the central portion of the state (36ºN; Liu et al., 2013).

Many of the ca. 13 North American *Carya* species readily hybridize with one another (Fralish & Franklin, 2002). Geographically structured hybridization may impact phylogeographic inferences in tree species (Saeki et al., 2011; Thomson et al., 2015),

and one limitation of our analyses is that we are unable to assess patterns of hybridization with other *Carya*. However, no stable hybrid zones exist in our species and we consider it unlikely that occasional hybridization would systematically bias genetic structure in a similar way across thousands of loci.

DNA sampling and SNP genotyping

Silica-dried leaf tissue was collected from 182 individuals of each species from populations across ENA (Fig. 1; Tables S1.1-S1.2). Sampled individuals within each population were separated by a minimum of 50 m (but sometimes up to dozens of km) in order to minimize the chance of sampling siblings and other close relatives. Sample size varied greatly among populations depending on the number of individuals meeting these requirements that could be located (mean N = 7.8; Tables S1.1-S1.2). A representative voucher specimen from each population was deposited in the University of Michigan Herbarium (MICH; DOI: JBB 79-164).

DNA samples were extracted using Nucleospin Plant II extraction kits (*Macherey-Nagel*; Düren, Germany), and libraries were prepared using a modified double digest Restriction Associated DNA (ddRAD) sequencing protocol following Peterson et al., (2012), with restriction enzymes *EcoRI* and *MseI.* Full details of extraction methods and library preparation are provided in Appendix A. Seven libraries of 72 samples each were sequenced at The Hospital for Sick Children (Toronto, ON) on an *Illumina HiSeq* (Illumina; San Diego, CA) using single-end 50-bp sequencing. In order to ensure adequate depth of coverage, at least one million raw reads per sample were required to

process a sample, and individuals not meeting this target were resequenced in subsequent libraries.

Loci were identified and single nucleotide polymorphisms (SNPs) were genotyped using STACKS v.1.44-v.1.46 (Catchen et al., 2011, 2013). Full details of SNP discovery are provided in Appendix A. After SNPs were successfully identified, one SNP genotype per locus was exported from STACKS using the *populations* tool, retaining only SNPs with a minimum genotyping rate of 75% (-r 0.75) and a minimum minor allele frequency (MAF) of 3.3% (--min_maf 0.033), the lowest detectable MAF in at least one population of each species, following Massatti & Knowles (2014). Minimum MAF is an important parameter to consider because it can impact inference of genetic structure (De la Cruz & Raska, 2014). We therefore explored preliminary principal component analyses (see below) with minimum MAF = 1% and 5%, but found that using the higher minimum MAF (5%) made little qualitative difference in preliminary results. With the lower minimum MAF (1%), tiny clusters of a few individuals appearing very distinct from all other genotypes tended to appear in *C. cordiformis*. This pattern may arise when closely related individuals are genetically similar to one another in terms of rare alleles due to shared, local demographic history (De la Cruz & Raska, 2014). In contrast, as we are interested in broader-scale differences among populations, our choice of minimum MAF = 3.3% was likely appropriate.

In order to retain only putatively nuclear SNPs, we removed any SNPs from loci that aligned with a maximum of two mismatches (-v 2) to the *Juglans regia* (Juglandaceae) chloroplast genome (Genbank accession NC_028617.1) or the *Cucurbita pepo* (Cucurbitaceae) mitochondrion genome (NC_014050.1) using BOWTIE

v.1.2 (Langmead et al., 2009). Extremely variable loci were also excluded as these may represent locus assembly errors; we defined these loci as those with values of θ (Watterson, 1975) above the 95th percentile, with θ calculated for each locus individually using the *R* package 'pegas' v.0.10 (Paradis, 2010). Individual samples with unusually high levels of missing data across all loci (based on visual inspection) were also excluded.

Genetic diversity and divergence

Three genetic diversity parameters were calculated overall and for each population: observed and expected heterozygosity (H_o and H_e, respectively), and nucleotide diversity (π). Genetic differentiation (F_{ST}) (Nei, 1987) was calculated overall and pairwise between each pair of populations. H_o, H_e, and F_{ST} were calculated in the *R* package 'hierfstat' v.0.04-22 (Goudet, 2005), while π was calculated using *populations* in STACKS v.1.46 (Catchen et al., 2011, 2013). Genetic diversity and differentiation measures are not reported for populations represented by a single individual.

Population genetic structure

In order to test for isolation by distance (IBD), Mantel tests (Mantel, 1967) were performed to assess the relationship between population pairwise F_{ST} values and geographic distances. Principal component analysis (PCA) was used to investigate genetic relationships among individuals and populations using the *dudi.pca* function in the *R* package 'adegenet' v.2.0.1 (Jombart, 2008; Jombart & Ahmed, 2011). The NC_e population (Table S1.1) was excluded for *C. cordiformis* because some, but not all

individuals from this population formed a distinct genetic cluster, which suggests that several closely related individuals were unintentionally sampled and the high genetic similarity between these individuals could have biased initial PCA results.

Genetic clusters were characterized using FASTSTRUCTURE v.1.0 (Raj et al., 2014), with all populations and individuals included, using the recommended procedure for detecting subtle genetic structure. Initially, the simple prior model was used and the number of clusters (*K*) was varied from 1 to 6 for each species, and *K* was selected using the *chooseK* tool in FASTSTRUCTURE. Then, FASTSTRUCTURE was rerun 100 times using the logistic prior model for the optimal value(s) of *K*, and final estimates of genetic membership of individuals in each genetic cluster were obtained as the average membership from the five runs with the highest likelihood, following Raj et al. (2014). After investigating the broadest level of structure within each dataset, we reran FASTSTRUCTURE on individual genetic clusters in order to test for substructure within clusters.

Paleodistribution modelling

Species distribution models (SDMs) were constructed in order to predict the potential distribution of each species during the Last Glacial Maximum (LGM; 21.5 ka). Complete details of SDM construction and data sources are given in Appendix A. Briefly, occurrence records were obtained from the US Forest Service Forest Inventory Analysis Database (O'Connell et al., 2012), while environmental variables were obtained at 2.5-arcminute resolution from the WORLDCLIM v.1.4. (Hijmans et al., 2005) and ENVIREM (Title & Bemmels, 2018) databases. SDMs were constructed using MAXENT v.3.4.1

(Phillips et al., 2004, 2006, 2017) in the *R* package 'dismo' (Hijmans et al., 2015), with models optimized according to best practices, following Title & Bemmels (2018). Habitat suitability was projected for the LGM according to each of the CCSM4, MIROC-ESM, and MPI-ESM-P general circulation models (GCMs), but since projections were similar for all three GCMs, results were averaged into a single map.

Results

Genetic diversity and differentiation

The final genetic datasets for *C. cordiformis* and *C. ovata* contained 177 individuals genotyped at 1,046 SNPs, and 180 individuals genotyped at 1,018 SNPs, respectively. The overall genotyping rate for both species was 89%.

While some populations were represented by very few individuals (Tables S1.1-S1.2), very small sample sizes are typically sufficient to obtain accurate population genomic measures of genetic diversity and differentiation if calculated across thousands of SNPs (Willing et al., 2012; Nazareno et al., 2017). For this reason, we did not perform rarefaction in our analyses to match the lowest population sample size. As further empirical justification for this decision, we found that genetic diversity estimates were uncorrelated with population sample size (Fig. S1.1), except for a negative relationship between H_o and sample size in *C. ovata* ($R^2 = 0.21$, $p = 0.035$). However, as H_o is computed on a per-individual basis, there is no theoretical reason to explain how sample size could affect estimates of H_o and we suspect that this correlation is spurious.

Genetic diversity showed little variation among populations for both species (Fig. 2). A significant decline in genetic diversity with increasing latitude was not observed for any genetic diversity measure (H_o, H_e, π) for either species. Instead, a significant increase in H_o with increasing latitude was observed in *C. cordiformis* (R^2 = 0.23, p = 0.021), as was a marginally non-significant increase in H_o with increasing latitude in *C. ovata* (R^2 = 0.19, p = 0.051). Although genetic variation was fairly uniform across latitudes, far northern and far southern populations sometimes showed slightly lower values of H_e and π than typical of mid-latitude populations (Fig. 2), as expected for range-edge populations (Jaramillo-Correa et al., 2009). Among-population genetic differentiation (F_{ST}) is low in both species overall (*C. cordiformis*: 0.047; *C. ovata*: 0.038), and among most pairs of populations (Tables S1.3-S1.4).

Spatial genetic structure

Spatial genetic structure in both species is weak and dominated by a pattern of isolation by distance (IBD). Mantel tests of IBD were statistically significant in both species (*C. cordiformis*: r = 0.36, p = 0.0017; *C. ovata*: r = 0.47, p = 8.1 x 10^{-5}; Fig. 3).

Principal component analysis (PCA) of genetic variation also revealed an IBD-like pattern, with gradual transitions and substantial overlap among geographic regions, and without clearly defined, distinct genetic clusters over most of the species range (Fig. 4). One clear exception to this pattern is that in *C. ovata*, the Texas population (TX; Fig. 4b) forms a separate cluster that does not overlap with any other populations. In *C. cordiformis,* PC1 represents a north-south geographic transition, while PC2 does not

appear strongly related to geography. In *C. ovata*, PC1 distinguishes TX from all other populations, while PC2 represent a north-south transition.

Lack of strong genetic structure was also suggested by FASTSTRUCTURE results. Under the model with simple priors, the optimal number of genetic clusters was $K = 1$ for both species. Under the model with logistic priors, which is more useful for detecting subtle structure (Raj et al., 2014), optimal K ranged from 1-6. However, the logistic priors model is prone to overfitting (Raj et al., 2014) and $K > 2$ did not produce results that were biologically interpretable. We therefore note that $K = 1$ or 2 is likely the optimal model complexity to explain genetic structure. In both species with $K = 2$, genetic variation was geographically structured but with transition zones between clusters (Figs. 5, S1.2). In *C. cordiformis*, the transition was from north to south, whereas in *C. ovata*, the transition was primarily from east to west. No substructure was evident within any genetic cluster for any species, except that within the western cluster for *C. ovata*, optimal $K = 2$ and the Texas population (TX) forms a distinct subcluster relative to the other four populations (AR, IA, ON, WI; data not shown). However, the grouping of these four populations into a distinct subcluster may be only a statistical artifact reflecting the substantial additional membership of each these four western populations in the main eastern cluster.

Paleodistribution modelling

The same four climatic variables were coincidentally retained in the SDMs for both species: maximum temperature of the coldest month, potential evapotranspiration of the warmest quarter, mean annual precipitation, and climatic moisture index (Table

S1.5). For both species, models were able to predict the current species distribution (Fig. 1) very well along the northern and western range edges, but performed more poorly at delineating the southern range edge (Fig. 5). This poorer performance may reflect the fact that both species are rare in the southern portion of their ranges, where presence and absence may be determined more by soil type and topography (Graney, 1990; Smith, 1990) than by broad-scale climatic differences among sites. For both species, a large, continuous area of high LGM habitat suitability is predicted to have extended over much of the southeastern US, from central Texas to the coast of North Carolina (Fig. 5).

Discussion

Carya cordiformis and *Carya ovata* likely survived the LGM over a broad geographic area covering much of the southeastern United States. This scenario is supported by our genetic results showing weak genetic structure and an isolation-by-distance (IBD) pattern, and is compatible with predictions of our paleodistribution models and the fossil record. Genetic differentiation is weakly geographically structured, without sharp phylogeographic breaks over most of ENA. However, a Texas population (TX) of *C. ovata* is genetically distinct compared to other populations and may be derived from a separate, genetically isolated glacial refugium. We find no evidence of any further subdivision into separate refugia in either species. Although the geographic distribution of both species was likely quite broad during the LGM, the fossil record indicates that most temperate deciduous trees occurred at low density in communities with no modern analogs (Prentice et al., 1991; Jackson, et al. 2000; Jackson and

Williams, 2004). Further research into the phylogeographic history of other widespread ENA temperate deciduous trees is needed, but our results suggest that both hickory species may not have experienced the complex refugial dynamics that structured genetic variation in other Northern Hemisphere temperate forest regions.

A widespread distribution during the LGM

Weak genetic structure, lack of strong phylogeographic breaks across most of ENA, and lack of distinct areas of elevated genetic diversity suggest that both species likely occurred over a broad area of southern ENA during the LGM. Nonetheless, these patterns might also be observed if high gene flow has homogenized populations and eroded historical demographic signatures (e.g., He et al., 2013). Low population genetic differentiation (*C. cordiformis*, F_{ST} = 0.047; *C. ovata*, F_{ST} = 0.038) suggests that gene flow among populations is indeed quite high, which is typical of widespread, wind-pollinated forest trees, due to their large population sizes and capacity for long-distance pollen-mediated gene flow (Hamrick et al., 1992; Savolainen et al., 2007; Alberto et al., 2013). However, the amount of time that has passed since the LGM may have been insufficient for gene flow to completely erode genetic signatures of expansion from a geographically restricted or fragmented LGM distribution. Both species are slow growing and long lived, with peak reproduction occurring in *C. cordiformis* from ages 50 to 125 (Smith, 1990), and in *C. ovata* from ages 60 to 200 (Graney, 1990). Generation times are more than an order of magnitude shorter in species such as small vertebrates and herbaceous plants, yet these species frequently retain phylogeographic structure interpreted to reflect the effects of glaciation (Soltis et al. 2006).

On the other hand, long generation times and large effective population sizes of nuclear DNA (relative to cpDNA or mtDNA) might mean that if periods of range fragmentation and refugial isolation were brief, strongly genetically differentiated populations might not have had sufficient time to develop. Thus, while we hypothesize that a widespread, genetically connected distribution was maintained throughout the LGM in the southeastern United States, the existence of separate, briefly isolated refugia that have since merged into a more genetically homogenous distribution is also potentially compatible with our results.

While genetic diversity of European taxa was often lost during northward expansion from southern refugia, this pattern is not observed in most temperate tree species from ENA (Lumibao et al., 2017). The relatively uniform levels of population genetic diversity (H_e, π) across the species range (Fig. 2) suggest that historical recolonization occurred over a large, slowly expanding region with little loss of diversity during migration (Jaramillo-Correa et al., 2009). However, high gene flow could also have reduced differences in genetic diversity among populations. Whereas H_e and π are relatively uniform across populations, observed heterozygosity (H_o) increases with increasing latitude in both species (Fig. 2). Both species are less common in southern areas than in the north (Fig. 5; Graney 1990; Smith 1990), and it is possible that southern populations could generally be smaller, more isolated, and more prone to reductions in H_o due to inbreeding.

Predictions of paleodistribution models also suggest that climatic conditions were favourable for survival of both species over a broad geographic area (Fig. 5). However, SDMs should be interpreted with caution, especially when projecting models to non-

analog climates in novel time periods or geographic areas (Owens et al., 2013). Model projections for the LGM also depend on simulations from general circulation models, which may not accurately reflect true paleoclimatic conditions (Varela et al. 2015). Despite these concerns, the fossil pollen record is largely compatible with our paleodistribution models, as low amounts of *Carya* pollen have been found in LGM assemblages across much of the southeastern United States (Prentice et al., 1991; Jackson et al. 2000). As climates warmed, major increases in *Carya* pollen occurred ca. 16-13 ka in sites as geographically distant as the Missouri Ozarks (Jones et al. 2017), central Tennessee (Liu et al. 2013), and South Carolina (Watts 1980). Although long-distance migration to these sites is possible, expansion from nearby sources is also plausible and would hint that *Carya* may have been broadly longitudinally distributed prior to postglacial expansion.

Glacial refugia and sources of postglacial recolonization

We find no evidence of genetically isolated refugia in *C. cordiformis*, but a separate glacial refugium in Texas is strongly suggested by our PCA results for *C. ovata* (Fig. 4b). Glacial refugia have previously been inferred in Texas and northern Mexico for several southern *Pinus* and *Prunus* species (Schmidtling & Hipkins, 1998; Schmidtling, 2003; Shaw & Small, 2005; Eckert et al., 2010). However, a separate refugium is not the only possible explanation for the genetic distinctiveness TX, as we speculate that the ancestors of this population might have historically experienced gene flow with *C. ovata* populations in the mountains of northern Mexico (Little, 1971). Although we have not included any high-elevation Mexican populations in our genetic analyses or

paleodistribution models, we cannot exclude the possibility that these populations may have migrated to lower elevations and come into contact with TX during the LGM. Alternatively, because TX is a range-edge population (Fig. 1b), it is possible that genetic drift due to small population size and limited gene flow with other populations could have caused substantial changes in allele frequency in this population. Denser population sampling across the southwestern portion of the species range (i.e., Louisiana, southern Arkansas) and of Mexican populations could help distinguish among these scenarios.

The east-west population structuring in *C. ovata* (Fig. 1b) detected in FASTSTRUCTURE also supports our inference that TX likely represents a separate glacial refugium. Some northern populations show admixture between the western and eastern clusters, but most populations across the species range show substantial membership in the eastern genetic cluster, suggesting that the contribution of the western (TX) lineage to postglacial recolonization was relatively minor. In contrast, the northern and southern genetic clusters in *C. cordiformis* are broadly longitudinally distributed (Fig. 1a) and likely do not represent distinct refugia. This pattern suggests that northern LGM populations may have made the greatest contribution to postglacial recolonization in *C. cordiformis*, as has previously been inferred for *Acer rubrum* and *Fagus grandifolia* (McLachlan et al., 2005). Alternatively, recolonization may have occurred from southern areas, but northern areas may have experienced subtle shifts in allele frequency due to genetic drift during northward migration.

While our FASTSTRUCTURE results present intriguing hypotheses, we note that these results must be interpreted with caution because IBD is known to bias tests of

hierarchical genetic structure (Frantz et al., 2009; Meirmans, 2012). In particular, such tests are susceptible to incorrect inference of multiple genetic clusters when populations are geographically subsampled from within a single larger cluster subject to IBD (Frantz et al., 2009; Meirmans, 2012). A rangewide IBD pattern is present in our datasets (Fig. 3), and genetic structure is weak in both species (optimal K = 1 to 2), with geographic transition zones between inferred genetic clusters (Fig. 1) and substantial overlap among populations along the first and second principal component axes of genetic variation (Fig. 4). Rather than indicating the true presence of biologically meaningful genetic clusters, our FASTSTRUCTURE results might be a statistical artifact of underlying IBD, especially in *C. cordiformis* for which no sharp genetic breaks were detected (Fig. 4a).

Implications and future directions

Despite decades of phylogeographic study (Soltis et al. 2006) and synthesis of the fossil record (Davis 1983, Jackson et al. 2000), a general phylogeographic history of ENA temperate forest taxa remains elusive. Given that *C. cordiformis* and *C. ovata* are common, widespread tree species with a geographic distribution roughly matching that of modern temperate deciduous forests in ENA, insights from these species may be relevant to predicting the phylogeographic histories of other geographically widespread temperate deciduous trees from this region. While different species likely responded in their own individual ways to Pleistocene glaciation, our results suggest that at least some temperate tree species were fairly widespread throughout southern ENA during the LGM. One of the most striking findings of our analyses is that both species appear

to lack the strong phylogeographic breaks characteristic of taxa from other temperate regions of the world (Hewitt, 2000; Shafer et al., 2010; Qiu et al., 2011; Lumibao et al., 2017), except for the Texas population of *C. ovata*. That both species show similar phylogeographic patterns over most of ENA is somewhat surprising, given substantial differences in their ecology. *Carya ovata* is a habitat generalist (Graney, 1990), but low genetic structure in a primarily mesic, bottomland species like *C. cordiformis* (Smith, 1990) provides strong evidence that even temperate tree species with more specific habitat requirements were not necessarily highly genetically fragmented during the LGM.

Although both species were likely fairly broadly distributed during the LGM, the fossil pollen record suggests that most LGM forest communities across southern ENA were conifer-dominated with no modern analogs (Jackson 2000, Jackson and Williams 2004). The absence of temperate deciduous forest communities is also noted from the pollen record from other Northern Hemisphere temperate regions (Prentice et al., 2000; Binney et al., 2017), and the LGM extent of the temperate deciduous forest biome, or whether it existed at all, remains unknown (Prentice et al., 2000). Despite lack of evidence for expansive temperate deciduous forests during the LGM, several other widespread temperate deciduous tree and plant species are also believed to have expanded from a large area covering much of southern ENA, including *Acer rubrum* (Bennett 1985; Magni et al. 2005; McLachlan et al., 2005; Peterson & Graves, 2016). In general, these results support the emerging understanding (Lumibao et al., 2017) that the genetic consequences of Pleistocene glaciation on widespread ENA temperate tree species were very different from those produced by expansion from distinct refugia in Europe (Hewitt, 1999, 2000; Petit et al., 2003).

However, many other temperate plant species experienced more complex phylogeographic histories (e.g., Griffin & Barrett, 2004; Gonzales et al., 2008; Eckert et al., 2010; Barnard-Kubow et al., 2015; Nadeau et al., 2015; Zinck & Rajora, 2016). Species adapted to a narrower range of climatic conditions than *C. cordifomris* and *C. ovata* could have been more likely to experience geographic fragmentation during the LGM. We hypothesize that taxa with a strictly southern, warm-temperate distribution may have become fragmented into distinct far-southern refugia in Florida, Texas, or along the Gulf or Atlantic Coasts (e.g., Gonzales et al., 2008; Eckert et al., 2010). In contrast, cool-temperate and more climatically widespread species such as *C. cordiformis* and *C. ovata* could have survived in more expansive inland areas of cool-temperate conditions that extended farther north into continental areas (e.g., Fig. 5; see also McLachlan et al., 2005).

Genetically distinct southern populations of temperate species, such as the Texas population of *C. ovata*, have often been identified as high conservation priority (Petit et al., 2003; Hampe & Petit, 2005; Médail & Diadema, 2009). We suggest that any conservation efforts in *C. ovata* should ensure inclusion of populations from Texas and surrounding regions. However, across most of the range of either species, we see little reason to prioritize conservation of southern populations, because these populations are neither genetically distinct nor do they exhibit elevated genetic diversity. On the other hand, most temperate tree species exhibit geographically structured climatically adaptive genetic variation (Savolainen et al., 2007; Aitken & Bemmels, 2016) unlikely to be captured by the putatively neutral SNP markers we employed. Conserving

populations from a variety of climates across the species range would therefore likely maximize conservation of adaptively relevant genetic diversity.

To our knowledge, this study is the first to investigate genome-wide genetic variation in a geographically widespread temperate deciduous tree from ENA. Much future work remains to be done to test whether the patterns we observe are generally applicable to other ENA trees. In addition to further phylogeographic studies of genome-wide variation from taxa with diverse traits and geographic distributions, application of demographic and coalescent modelling techniques would greatly enhance our understanding of the demographic history of ENA taxa. Such techniques would allow statistical tests of hypotheses regarding levels of genetic connectivity among populations, changes in population size over time (Barthe et al., 2017), locations of source populations from which postglacial recolonization occurred (He et al., 2017), and which ecological factors have most strongly impacted historical migration patterns (Bemmels et al., 2016).

Acknowledgements

The authors thank B.J. Belcher, P. Cousineau, R. D'Andrea, J. Fiske, J. Ronson, D. Saenz and P. Tichenor for assistance with fieldwork, and A.T. Thomaz for assistance with genetic analyses. Graduate student support and research funding was provided to J.B.B. by the National Science Foundation (GRFP fellowship; DDIG 1501159), and the University of Michigan EEB Department and Rackham Graduate School.

Author contributions

JBB and CWD conceived the project; JBB performed fieldwork and labwork, and analyzed the data. JBB wrote the chapter with input from CWD.

Literature cited

Aitken, S.N. & Bemmels, J.B. (2016) Time to get moving: assisted gene flow of forest trees. *Evolutionary Applications*, **9**, 271–290.

Alberto, F.J., Aitken, S.N., Alía, R., González-Martínez, S.C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., & Savolainen, O. (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Barnard-Kubow, K.B., Debban, C.L., & Galloway, L.F. (2015) Multiple glacial refugia lead to genetic structuring and the potential for reproductive isolation in a herbaceous plant. *American Journal of Botany*, **102**, 1842–1853.

Barthe, S., Binelli, G., Hérault, B., Scotti-Saintagne, C., Sabatier, D., & Scotti, I. (2017) Tropical rainforests that persisted: inferences from the Quaternary demographic history of eight tree species in the Guiana shield. *Molecular Ecology*, **26**, 1161–1174.

Bemmels, J.B., Title, P.O., Ortego, J., & Knowles, L.L. (2016) Tests of species-specific models reveal the importance of drought in postglacial range shifts of a Mediterranean-climate tree: insights from integrative distributional, demographic and coalescent modelling and ABC model selection. *Molecular Ecology*, **25**, 4889–4906.

Bennett, K.D., & Provan, J. (2008) What do we mean by 'refugia'? *Quaternary Science Reviews*, **27**, 2449-2455.

Bennett, K.D. (1985) The spread of *Fagus grandifolia* across eastern North America during the last 18 000 years. *Journal of Biogeography*, **12**, 147–164.

Binney, H., Edwards, M., Macias-Fauria, M., Lozhkin, A., Anderson, P., Kaplan, J.O., Andreev, A., Bezrukova, E., Blyakharchuk, T., Jankovska, V., Khazina, I., Krivonogov, S., Kremenetski, K., Nield, J., Novenko, E., Ryabogina, N., Solovieva, N., Willis, K., & Zernitskaya, V. (2017) Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. *Quaternary Science Reviews*,

157, 80–97.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., & Cresko, W.A. (2013) STACKS: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J.H. (2011) STACKS: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

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

De la Cruz, O. & Raska, P. (2014) Population structure at different minor allele frequency levels. *BMC proceedings*, **8**, S55.

Delcourt, Paul A., Delcourt, Hazel R., Brister, Ronald C., & Lackey, Laurence, E. (1980) Quaternary vegetation history of the Mississippi Embayment. *Quaternary Research*, **13**, 111-132.

Eckert, A.J., Van Heerwaarden, J., Wegrzyn, J.L., Nelson, C.D., Ross-Ibarra, J., González-Martínez, S.C., & Neale, D.B. (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Fralish, J.S. & Franklin, S.B. (2002) *Taxonomy and Ecology of Woody Plants in North American Forests.* John Wiley & Sons, New York.

Frantz, A.C., Cellina, S., Krier, A., Schley, L., & Burke, T. (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

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

Goudet, J. (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Graney, D.L. (1990) *Carya ovata* (Mill.) K. Koch. *Silvics of North America: 2. Hardwoods. Agricultural Handbook 654* (ed. by R.M. Burns and B.H. Honkala), US Department of Agriculture, Forest Service, Washington, DC.

Griffin, S.R. & Barrett, S.C.H. (2004) Post-glacial history of *Trillium grandiflorum* (Melianthiaceae) in eastern North America: inferences from phylogeography. *American Journal of Botany*, **91**, 465–473.

Hampe, A. & Petit, R.J. (2005) Conserving biodiversity under climate change: the rear

edge matters. *Ecology Letters*, **8**, 461–467.

Hamrick, J.L., Godt, J.W., & Sherman-Broyles, S.L. (1992) Factors influencing levels of genetic diversity in woody plant species. *New Forests*, **6**, 95–124.

He, Q., Prado, J.R., & Knowles, L.L. (2017) Inferring the geographic origin of a range expansion: latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program X-ORIGIN. *Molecular Ecology*, in press.

He, Q., Edwards, D.L., & Knowles, L.L. (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

Hewitt, G. (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

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

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans, R.J., Phillips, S., Leathwick, J., & Elith, J. (2015) Package "dismo". R package. http://cran.r-project.org/web/packages/dismo/index.html.

Jackson, S.T., & Williams, J.W. (2004) Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences*, **32**, 495-537.

Jackson, S.T., Webb, R.S., Anderson, K.H., Overpeck, J.T., Webb III, T., Williams, J.W., & Hansen, B.C.S. (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews*, **19**, 489–508.

Jaramillo-Correa, J.P., Beaulieu, J., Khasa, D.P., & Bosquet, J. (2009) Inferring the past from the present phylogeography structure of North American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, **39**, 286–307.

Jombart, T. (2008) ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jombart, T. & Ahmed, I. (2011) ADEGENET 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Jones, R.A., Williams, J.W., & Jackson, S.T. (2017) Vegetation history since the last glacial maximum in the Ozark highlands (USA): a new record from Cupola Pond, Missouri. *Quaternary Science Reviews*, **170**, 174-187.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S.L. (2009) Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Little, E.L. (1971) *Atlas of United States trees, volume 1. Conifers and important hardwoods.* USDA Miscellaneous Publication 1146, Washington, DC.

Liu, Y., Andersen, J.L., Williams, J.W., & Jackson, S.T. (2013) Vegetation history in central Kentucky and Tennessee (USA) during the last glacial and deglacial periods. *Quaternary Research*, **79**, 189-198.

Lumibao, C.Y., Hoban, S.M., & McLachlan, J. (2017) Ice ages leave genetic diversity "hotspots" in Europe but not in Eastern North America. *Ecology Letters*, in press.

Magni, C.R., Ducousso, A., Caron, H., Petit, R.J., & Kremer, A. (2005) Chloroplast DNA variation of *Quercus rubra* L. in North America and comparison with other Fagaceae. *Molecular Ecology*, **14**, 513–524.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Massatti, R. & Knowles, L.L. (2014) Microhabitat differences impact phylogeographic concordance of codistributed species: genomic evidence in montane sedges (*Carex* L.) from the Rocky Mountains. *Evolution*, **68**, 2833–2846.

McCarthy, D.M. & Mason-Gamer, R.J. (2016) Chloroplast DNA-based phylogeography of *Tilia americana* (Malvaceae). *Systematic Botany*, **41**, 865–880.

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

Médail, F. & Diadema, K. (2009) Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, **36**, 1333–1345.

Meirmans, P.G. (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.

Morris, A.B., Graham, C.H., Soltis, D.E., & Soltis, P.S. (2010) Reassessment of phylogeographical structure in an eastern North American tree using Monmonier's algorithm and ecological niche modelling. *Journal of Biogeography*, **37**, 1657–1667.

Nadeau, S., Godbout, J., Lamothe, M., Gros-Louis, M.-C., Isabel, N., & Ritland, K. (2015) Contrasting patterns of genetic diversity across the ranges of *Pinus monticola* and *P. strobus*: a comparison between eastern and western North American postglacial colonization histories. *American Journal of Botany*, **102**, 1342–1355.

Nazareno, A.G., Bemmels, J.B., Dick, C.W., & Lohmann, L.G. (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species.

Molecular Ecology Resources, in press.

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

O'Connell, B.M., LaPoint, E.B., Turner, J.A., Ridley, T., Boyer, D., Wilson, A.M., Waddell, K.L., & Conkling, B.L. (2012) *The Forest Inventory and Analysis Database: database description and users manual version 5.1.4 for Phase 2. Gen. Tech. Rep.,* US Department of Agriculture, Forest Service, Washington, DC.

Little, E.L. (1971) Atlas of United States trees, volume 1. Conifers and important hardwoods. USDA Miscellaneous Publication 1146. .

Raj, A., Stephens, M., & Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Paradis, E. (2010) Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.

Peterson, B.J. & Graves, W.R. (2016) Chloroplast phylogeography of *Dirca palustris* L. indicates populations near the glacial boundary at the Last Glacial Maximum in eastern North America. *Journal of Biogeography*, **43**, 314–327.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., & Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Müller-Starck, G., Demesure-Musch, B., Palmé, A., Martín, J.P., Rendell, S., & Vendramin, G.G. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., & Blair, M.E. (2017) Opening the black box: an open-source release of MAXENT. *Ecography*, **40**, 887–893.

Phillips, S.J., Anderson, R.P., & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, **2004**, 655–662.

Prentice, C., Bartlein, P.J., & Webb III, T. (1991) Vegetation and climate change in eastern North America since the Last Glacial Maximum. *Ecology*, **72**, 2038–2056.

Prentice, I.C., Jolly, D., & BIOME 6000 participants. (2000) Mid-Holocene and glacial-maximum vegetation geography of the northern continents and Africa. *Journal of Biogeography*, **27**, 507–519.

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

Qiu, Y.X., Fu, C.X., & Comes, H.P. (2011) Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics and Evolution*, **59**, 225–244.

Raj, A., Stephens, M., & Pritchard, J.K. (2014) FASTSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Saeki, I., Dick, C.W., Barnes, B. V., & Murakami, N. (2011) Comparative phylogeography of red maple (*Acer rubrum* L.) and silver maple (*Acer saccharinum* L.): impacts of habitat specialization, hybridization and glacial history. *Journal of Biogeography*, **38**, 992–1005.

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.

Schmidtling, R.C. (2003) The southern pines during the Pleistocene. *Acta Horticulturae*, **615**, 203–209.

Schmidtling, R.C. & Hipkins, V. (1998) Genetic diversity in longleaf pine (*Pinus palustris*): influence of historical and prehistorical events. *Canadian Journal of Forest Research*, **28**, 1135–1145.

Shafer, A.B.A., Cullingham, C.I., Côté, S.D., & Coltman, D.W. (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.

Shaw, J. & Small, R.L. (2005) Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *American Journal of Botany*, **92**, 2011–2030.

Smith, H.C. (1990) *Carya cordiformis* (Wangenh.) K. Koch. *Silvics of North America: 2. Hardwoods. Agricultural Handbook 654* (ed. by R.M. Burns and B.H. Honkala), US Department of Agriculture, Forest Service, Washington, DC.

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

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

Thomson, A.M., Dick, C.W., & Dayanandan, S. (2015) A similar phylogeographical structure among sympatric North American birches (*Betula*) is better explained by introgression than by shared biogeographical history. *Journal of Biogeography*, **42**,

339–350.

Title, P.O. & Bemmels, J.B. (2018) ENVIREM: An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291-307.

Tzedakis, P.C., Emerson, B.C., & Hewitt, G.M. (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology and Evolution*, **28**, 696–704.

Varela, S., Lima-Ribeiro, M.S., & Terribile, L.C. (2015) A short guide to the climatic variables of the Last Glacial Maximum for biogeographers. *PLoS ONE*, **10**, e0129037.

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Watts, W.A. (1980) Late-Quaternary vegetation history at White Pond on the inner coastal plain of South Carolina. *Quaternary Research*, **13**, 187-199.

Williams, J.W. (2002) Variations in tree cover in North America since the last glacial maximum. *Global and Planetary Change*, **35**, 1–23.

Willing, E.M., Dreyer, C., & van Oosterhout, C. (2012) Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE*, **7**, e42649.

Willis, K.J. & Van Andel, T.H. (2004) Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. *Quaternary Science Reviews*, **23**, 2369–2387.

Zinck, J.W.R. & Rajora, O.P. (2016) Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single refugium, multiple routes. *BMC Evolutionary Biology*, **16**, 56.

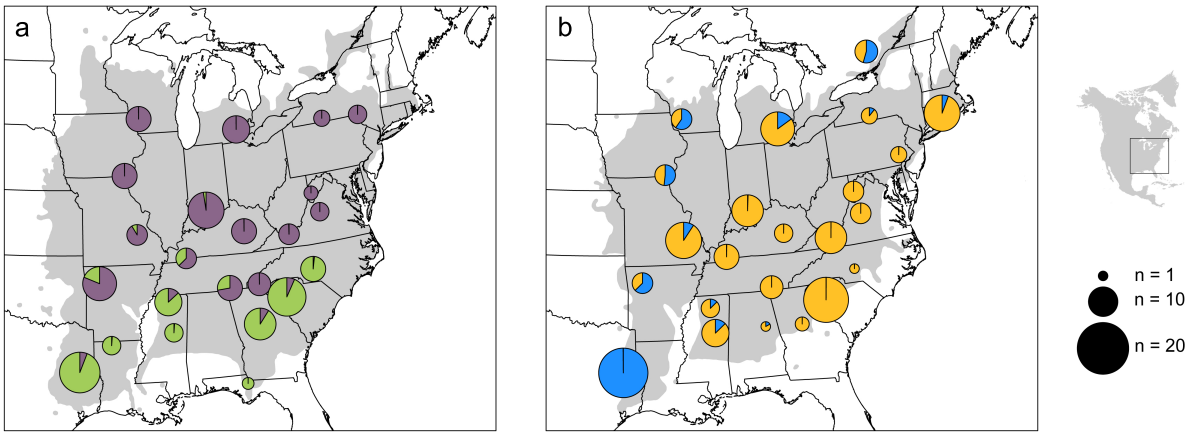

Figure 2.1. Membership of hickory populations in genetic clusters from FASTSTRUCTURE. Membership of (a) *Carya cordiformis* and (b) *Carya ovata* populations in genetic clusters (different colours on pie charts; *K* = 2) identified using FASTSTRUCTURE (Raj et al., 2014). The geographic distribution of each species is shown in grey (Little, 1971), and the sample size (n) of each population is proportional to the size of the pie chart. Note that optimal *K* = 1-2, but *K* = 1 is not shown as all individuals of each species would belong to the same genetic cluster.

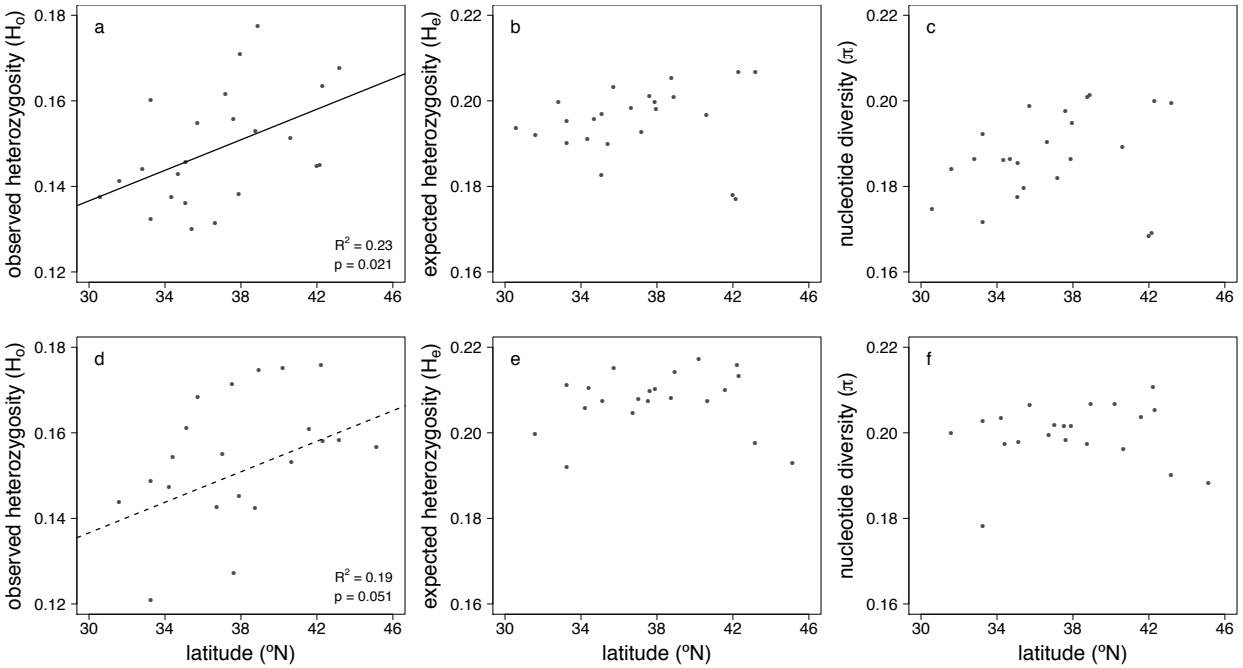

Figure 2.2. Genetic diversity vs. population latitude in hickories. Genetic diversity vs. latitude for populations of *Carya cordiformis* (a-c) and *Carya ovata* (d-f). Statistically significant relationships are portrayed as solid lines, and marginally significant relationships as dashed lines.

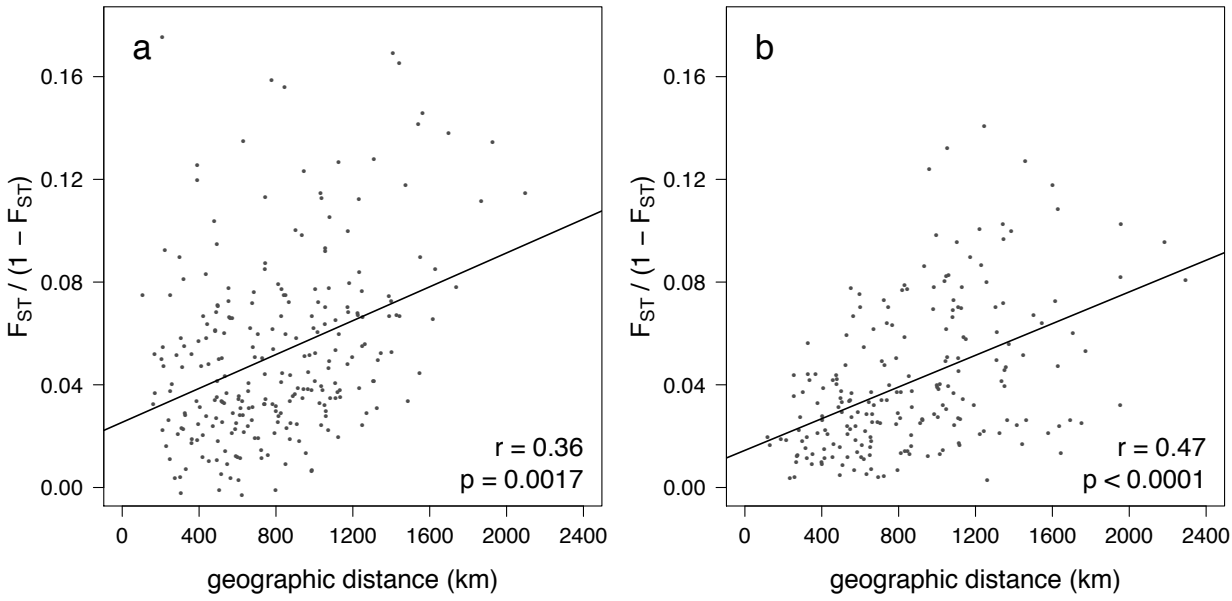

Figure 2.3. Mantel test of isolation by distance in hickories. Mantel test showing isolation by distance in (a) *Carya cordiformis* and (b) *Carya ovata*. Each dot represents a pair of populations and the y-axis is a measure of genetic differentiation between populations.

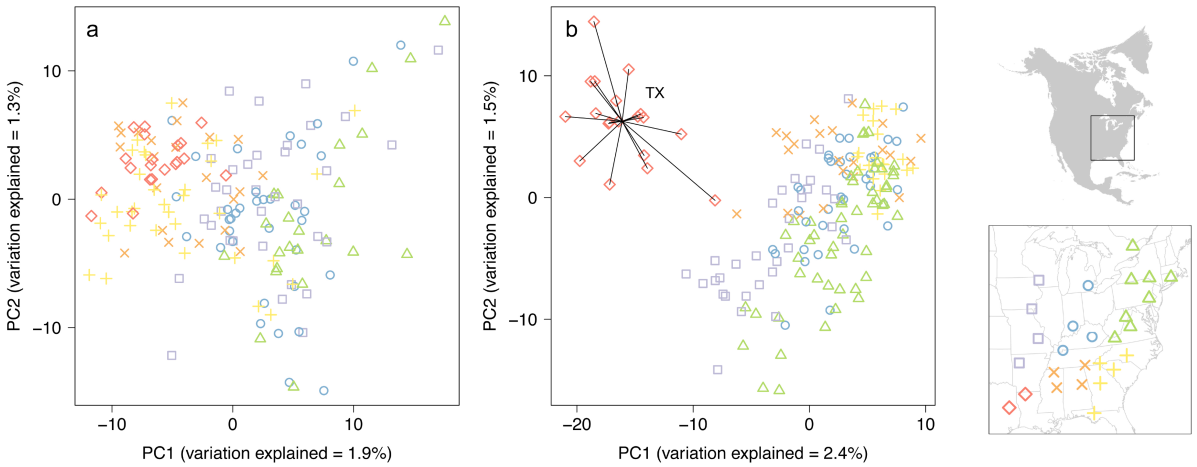

Figure 2.4. Clustering of hickory individuals along PC axes of genetic variation. Clustering of individuals along the first and second principal component (PC) axes of genetic variation in (a) *Carya cordiformis* and (b) *Carya ovata.* Each symbol represents a single individual, and symbols and colours correspond to the geographic location of the individual (shown in the inset map of eastern North America; geographic regions are delimited here for visualization purposes only). Individuals from the Texas population (TX) of *C. ovata* discussed in the manuscript are connected by black lines in (b). Note that sample size varies greatly for each population, and some populations were not sampled for both species (see Fig. 2.1).

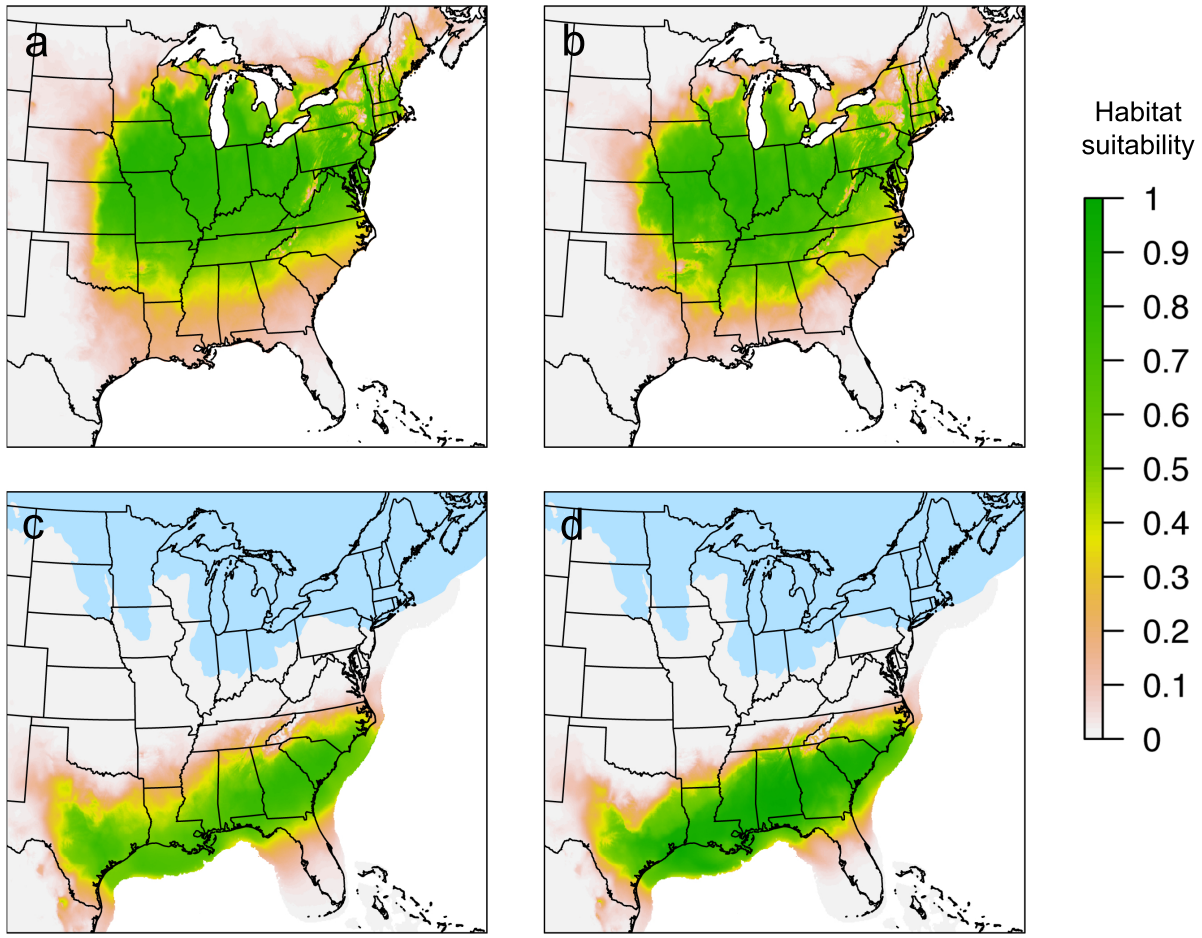

Figure 2.5. Predicted habitat suitability for the present and the LGM in hickories. Species distribution models showing predicted habitat suitability (colour scale: grey, habitat not suitable; green, maximum habitat suitability) for (a, c) *Carya cordiformis* and (b, d) *Carya ovata*, for climates of both (a-b) the current time period and (c-d) the Last Glacial Maximum (ca. 21.5 ka). Glaciated areas are shown in blue.

APPENDIX A

Supplementary information from Chapter II

Supplementary methods

DNA extraction and library preparation

DNA was extracted by twice rinsing homogenized leaf tissue in diluted β-mercaptoethanol solution to remove secondary compounds, then using *Nucleospin Plant II* extraction kits (*Macherey-Nagel*; Düren, Germany) with SDS reagents following the manufacturer's protocol. DNA libraries were prepared using double digest Restriction Associated DNA (ddRAD) sequencing following Peterson et al. (2012), with several modifications. Specifically, (1) adaptors and sequencing primers were designed following Parchman et al. (2012) for restriction enzymes *EcoRI* and *MseI*; (2) the 96 individual barcodes were designed following Meyer & Kircher (2010); (3) PCR amplification was performed individually prior to pooling and size selection, based on initial results that suggested that performing the library-preparation steps in this order resulted in a more even distribution of the number of reads obtained among samples; (4) two individual PCRs were performed on each sample using *iProof* high-fidelity DNA polymerase (*Bio-Rad*; Hercules, CA) then pooled for size selection on a *Pippin Prep* machine (*Sage Science*; Beverly, MA) with a target fragment length of 375-475 bp; and

(5) DNA-cleaning steps were performed using a 1.5X ratio of *Agencourt AMPure XP* beads (*Beckman Coulter*; Brea, CA) in all steps, except that a 0.8X ratio was used prior to PCR amplification and pooling, in order to pre-select larger fragments to improve equivalence of standardized concentrations among samples. All other steps in library preparation were similar to those described by Peterson et al. (2012).

SNP discovery

DNA sequences with suspected adaptor contamination were identified and removed using CUTADAPT v.1.11 (Martin, 2011). Loci and single nucleotide polymorphisms (SNPs) were then identified and genotyped using STACKS v.1.44-1.46 (Catchen et al., 2011, 2013). The individuals components of the STACKS pipeline were each run individually using default settings with the following exceptions: (1) when demultiplexing reads in *process_radtags*, up to two mismatches were allowed between barcodes (--barcode_dist_1 2), barcodes were rescued (-r), and reads with low quality scores were discarded (-q); (2) when building initial stacks in *ustacks*, a minimum of five reads were required to create a stack (-m 5), and lumberjack stacks (-r) and over-merged tags were removed (-d); and (3) when building a catalog of loci in *cstacks*, a maximum of two mismatches were allowed per locus when merging stacks (-n 2), and all individuals were used to build separate catalogs for each species. The optional component *rxstacks* was also run to remove problematic loci from the entire population, with non-biological haplotype pruning enabled (--prune_haplo), the maximum proportion of confounded loci in the population set to 25% (--conf_lim 0.25), and a minimum log likelihood of -5.0 required to retain a locus (--lnl_lim -5.0). After running *rxstacks*, the

pipeline components *cstacks* and *sstacks* were rerun using the same settings previously described. The invariant 5-bp *EcoRI* cut site was removed from the beginning of each read using *fastx_timmer* in FASTX-TOOLKIT v.0.0.13 (hannonlab.cshl.edu), resulting in a final locus length of 36 bp.

Paleodistribution modelling

Presence and absence data were obtained from the US Forest Service Forest Inventory Analysis Database for Phase 2 (O'Connell et al., 2012), using plot data from all states in the eastern and central US. All plots where the species was present during the most recent complete inventory were used to define occurrence records (*Carya cordiformis*: n = 3,897; *Carya ovata*: n = 5,547), and 10,000 plots where the species was not present were randomly sampled to define background points for each species. Although the density of FIA plots varies widely among and within states, the spatial sampling bias of background points should approximate that of occurrence records, such that spatial thinning of occurrence records as a means of reducing sampling bias was not necessary.

Environmental data for temperate ENA were obtained at 2.5-arcminute resolution from the WORLDCLIM v.1.4. (Hijmans et al., 2005) and ENVIREM (Title & Bemmels, 2018) databases, and converted to the North America Albers Equal Area Conic projection to ensure equal area of raster grid cells, as recommended by Budic et al. (2016). Environmental variables were obtained for both current conditions as well as the LGM according to three separate general circulation models (GCMs): CCSM4, MIROC-ESM, and MPI-ESM-P. Several environmental variables that were unlikely to be reliable

predictors over the study area were removed, including topographic variables; Emberger's pluviometric quotient, which is most relevant to Mediterranean climates (Daget, 1977); and climatic variables of the wettest and driest quarters, which may reflect winter conditions in some regions and summer conditions in other regions. A total of 26 climatic variables were retained for potential inclusion in SDMs.

SDMs were constructed using MAXENT v.3.4.1 (Phillips et al., 2004, 2006, 2017) with the *R* package 'dismo' (Hijmans et al., 2015). An iterative process was used to perform variable selection: first, a model was built with the complete set of climate variables; then, variables that were highly correlated with the variable of greatest permutation importance ($r > 0.8$) were removed. This process was repeated until only uncorrelated variables remained, and then all remaining variables with a permutation importance <5% were removed to define the final set of variables. After variable selection, model parameters were optimized using the *R* package 'ENMeval' (Muscarella et al., 2014). Models were constructed across a range of regularization parameter (β) values, using either all feature classes or only the hinge feature class, following (Elith et al., 2010). The optimal model was chosen by comparing AICc scores, and visually examining model output to ensure that the selected model did not show signs of overfitting. After building a final model for each species for current climatic conditions, models were projected to the LGM for all three GCMs. Predicted LGM habitat suitability for all three GCMs was similar, so predictions were combined to produce a single habitat-suitability map.

Table A.1. *Carya cordiformis* collecting localities and genetic diversity statistics. *Carya cordiformis* population collecting localities, sample sizes (n), and genetic diversity statistics (H_o, observed heterozygosity; H_e, expected heterozygosity; π, nucleotide diversity). The first two letters of population IDs correspond to official abbreviations of US states. Reported latitudes and longitudes are the average values for all individuals in the population.

ID	Collecting location	Latitude	Longitude	n	H_o	H_e	π
AR	Ozark National Forest	35.70	-93.59	12	0.155	0.203	0.199
FL	Torreya State Park	30.58	-84.95	2	0.138	0.194	0.175
GA	Oconee National Forest	33.26	-83.79	11	0.160	0.195	0.192
IA	Shimek State Forest	40.61	-91.70	8	0.151	0.197	0.189
IN	Hoosier National Forest	38.77	-86.39	13	0.153	0.205	0.201
KY_e	Berea College Forest; Daniel Boone National Forest	37.61	-84.06	8	0.156	0.201	0.198
KY_w	Land Between the Lakes National Recreation Area; Murray State University Hancock Biological Station	36.64	-87.98	6	0.131	0.198	0.190
LA	Kisatchie National Forest	32.81	-92.96	5	0.144	0.200	0.186
MI	University of Michigan E.S. George Reserve, Horner Woods, and Matthaei Botanical Gardens	42.29	-83.72	9	0.163	0.207	0.200
MO	Mark Twain National Forest	37.88	-91.05	6	0.138	0.200	0.187
MS_n	Holly Springs National Forest	34.70	-89.32	9	0.143	0.196	0.187
MS_s	Tombigbee National Forest	33.26	-89.11	5	0.132	0.190	0.172
NC_e	Uwharrie National Forest	35.42	-80.02	8	0.130	0.190	0.180
NC_w	Nantahala National Forest	35.09	-83.52	7	0.146	0.197	0.186
NY_e	East Branch Rest Area	41.99	-75.16	5	0.145	0.178	0.169
NY_w	New York State Forests	42.16	-77.67	4	0.145	0.177	0.169
SC	Sumter National Forest	34.33	-81.93	14	0.138	0.191	0.186
TN	Prentice Cooper State Forest	35.08	-85.39	8	0.136	0.183	0.177
TX	Davy Crockett National Forest; Stephen F. Austin Experimental Forest and State University	31.59	-94.92	15	0.141	0.192	0.184
VA_n	George Washington National Forest	37.96	-78.97	5	0.171	0.198	0.195
VA_s	Jefferson National Forest	37.18	-81.19	6	0.162	0.193	0.182
WI	Lower Wisconsin State Riverway	43.18	-90.52	8	0.168	0.207	0.199
WV	Monongahela National Forest	38.88	-79.32	3	0.177	0.201	0.201
			Overall	**177**	**0.149**	**0.196**	**0.204**

Table A.2. *Carya ovata* collecting localities and genetic diversity statistics. *Carya ovata* population collecting localities, sample sizes (n), and genetic diversity statistics (H_o, observed heterozygosity; H_e, expected heterozygosity; π, nucleotide diversity). The first two letters of population IDs correspond to official abbreviations of US states or Canadian provinces. Reported latitudes and longitudes are the average values for all individuals in the population. Genetic diversity statistics are not reported for populations represented by a single individual.

ID	Collecting location	Latitude	Longitude	n	H_o	H_e	π
AL	Talladega National Forest	33.34	-86.02	1	-	-	-
AR	Ozark National Forest	35.72	-93.53	6	0.168	0.215	0.207
CT	Connecticut State Forests	41.59	-72.25	13	0.161	0.210	0.204
GA	Oconee National Forest	33.26	-83.78	3	0.149	0.192	0.178
IA	Shimek State Forest	40.65	-91.78	6	0.153	0.207	0.196
IN	Hoosier National Forest	38.74	-86.42	11	0.142	0.208	0.197
KY_e	Berea College Forest; Daniel Boone National Forest	37.53	-84.22	5	0.171	0.207	0.202
KY_w	Land Between the Lakes National Recreation Area; Murray State University Hancock Biological Station	36.72	-88.06	8	0.143	0.205	0.200
MI	University of Michigan E.S. George Reserve, Horner Woods, and Matthaei Botanical Gardens	42.31	-83.74	12	0.158	0.213	0.205
MO	Mark Twain National Forest	37.62	-90.77	13	0.127	0.210	0.198
MS_n	Holly Springs National Forest	34.41	-89.34	5	0.154	0.210	0.197
MS_s	Tombigbee National Forest	33.25	-89.13	9	0.121	0.211	0.203
NC_e	Uwharrie National Forest	35.43	-80.05	1	-	-	-
NY_w	New York State Forests	42.22	-77.24	4	0.176	0.216	0.211
ON	Lavant (Lanark Highlands)	45.13	-76.53	7	0.157	0.193	0.188
PA	French Creek State Park	40.20	-75.80	4	0.175	0.217	0.207
SC	Sumter National Forest	34.22	-82.11	17	0.147	0.206	0.204
TN	Prentice Cooper State Forest	35.12	-85.39	7	0.161	0.207	0.198
TX	Davy Crockett National Forest; Stephen F. Austin Experimental Forest and State University	31.58	-94.81	19	0.144	0.200	0.200
VA_n	George Washington National Forest	37.90	-79.04	6	0.145	0.210	0.202
VA_s	Jefferson National Forest	37.02	-81.22	11	0.155	0.208	0.202
WI	Lower Wisconsin State Riverway	43.17	-90.44	6	0.158	0.198	0.190
WV	Monongahela National Forest	38.95	-79.26	6	0.175	0.214	0.207
			Overall	**180**	**0.154**	**0.208**	**0.215**

Table A.3. *Carya cordiformis* pairwise population differentiation (F_{ST}). Pairwise population genetic differentiation (F_{ST}) in *Carya cordiformis.* Population IDs are given in Table A.1. Cells are shaded according to F_{ST} value for ease of visual interpretation (white: 0.00-0.04; light grey: 0.05-0.08; dark grey: 0.09-0.12; black: ≥0.13).

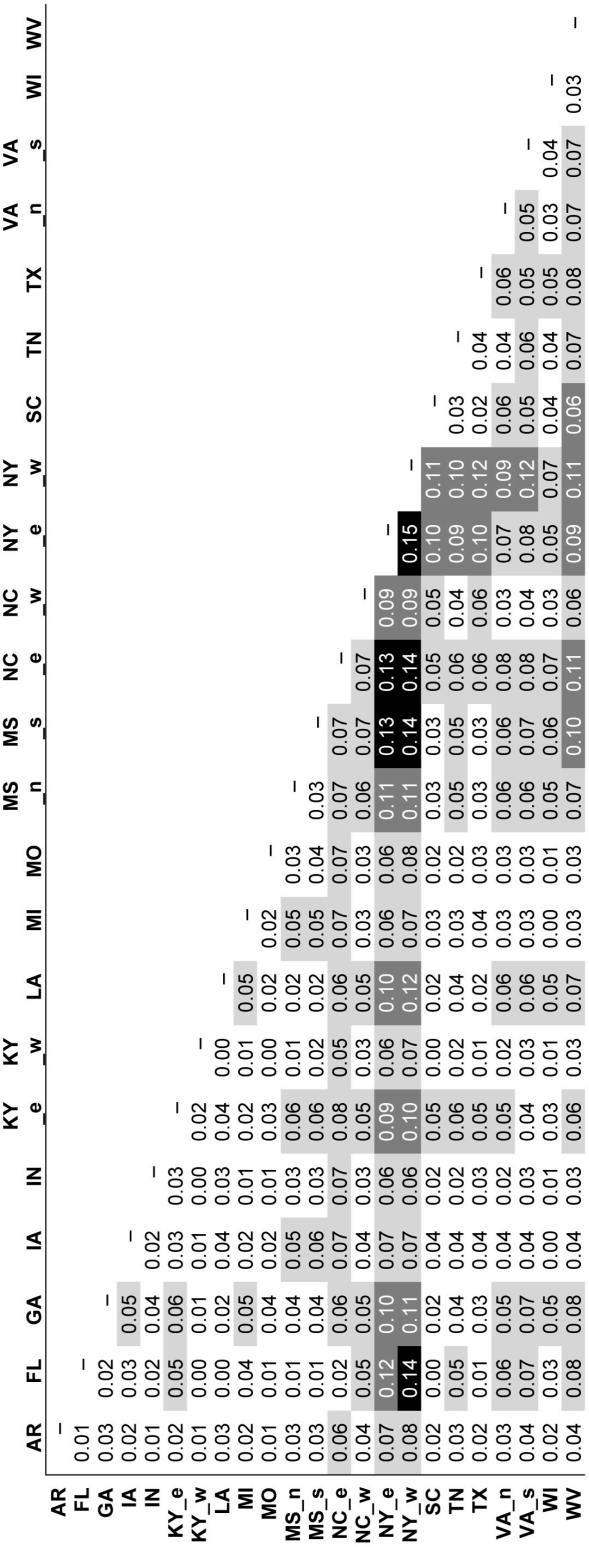

Table A.4. *Carya ovata* pairwise population differentiation (F_{ST}). Pairwise population genetic differentiation (F_{ST}) in *Carya ovata.* Population IDs are given in Table A.2. Cells are shaded according to F_{ST} value for ease of visual interpretation (white: 0.00-0.04; light grey: 0.05-0.08; dark grey: ≥0.09).

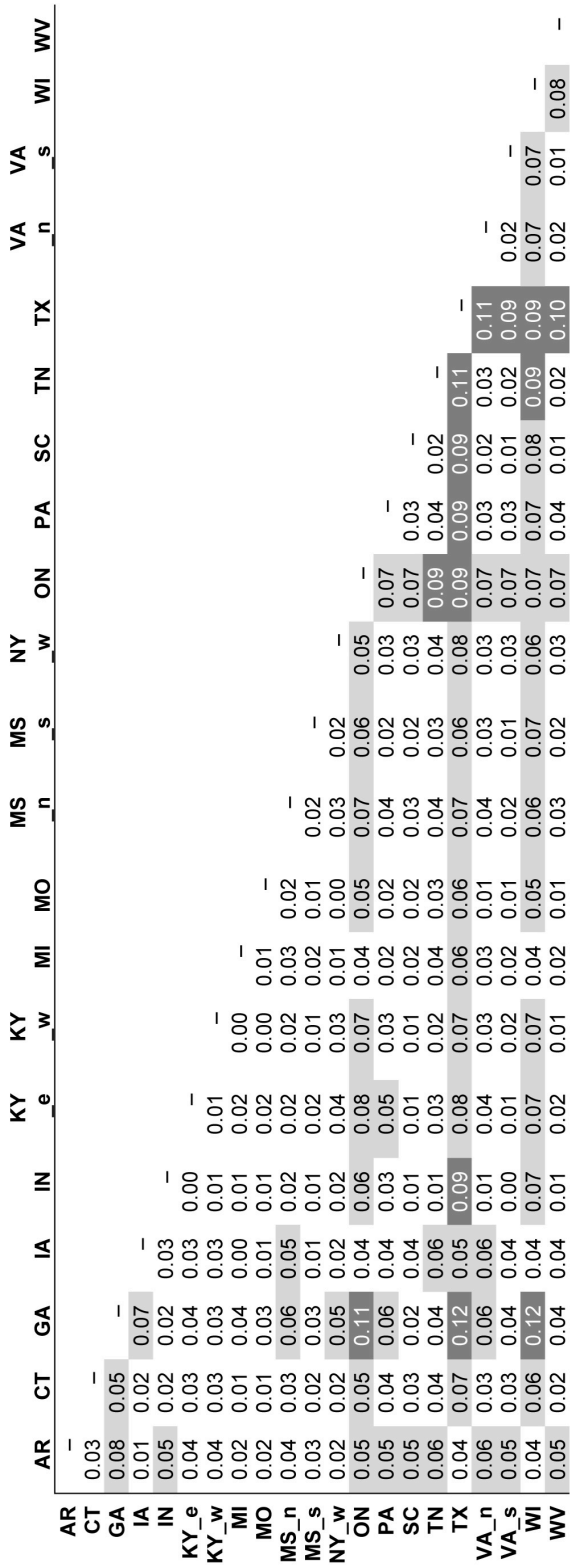

Table A.5. Species distribution model parameters and performance in hickories. Species distribution model (SDM) parameters and performance for (a) *Carya cordiformis* and (b) *Carya ovata.* Climatic variables are the predictor variables retained in the final model, and the percent contribution of each variable to the SDM and the permutation importance were assessed by MAXENT (Phillips et al., 2004, 2006, 2017). For (a) *Carya cordiformis*, model performance (area under the curve, AUC) = 0.718, model complexity β = 2.0, and all feature classes were used to construct the model. For (b) *Carya ovata*, AUC = 0.729, β = 0.4, and only hinge feature classes were used.

(a) *Carya cordiformis*

Climatic variable	Percent contribution	Permutation importance	Citation
maximum temperature of the coldest month	66.4	68.2	Title & Bemmels (2018)
mean annual precipitation (BIO12)	11.9	15.1	Hijmans et al. (2005)
potential evapotranspiration of the warmest quarter	18.4	9.0	Title & Bemmels (2018)
climatic moisture index	3.3	7.7	Title & Bemmels (2018)

(b) *Carya ovata*

Climatic variable	Percent contribution	Permutation importance	Citation
maximum temperature of the coldest month	68.3	53.6	Title & Bemmels (2018)
mean annual precipitation (BIO12)	6.6	22.9	Hijmans et al. (2005)
climatic moisture index	5.1	18.7	Title & Bemmels (2018)
potential evapotranspiration of the warmest quarter	20.0	4.9	Title & Bemmels (2018)

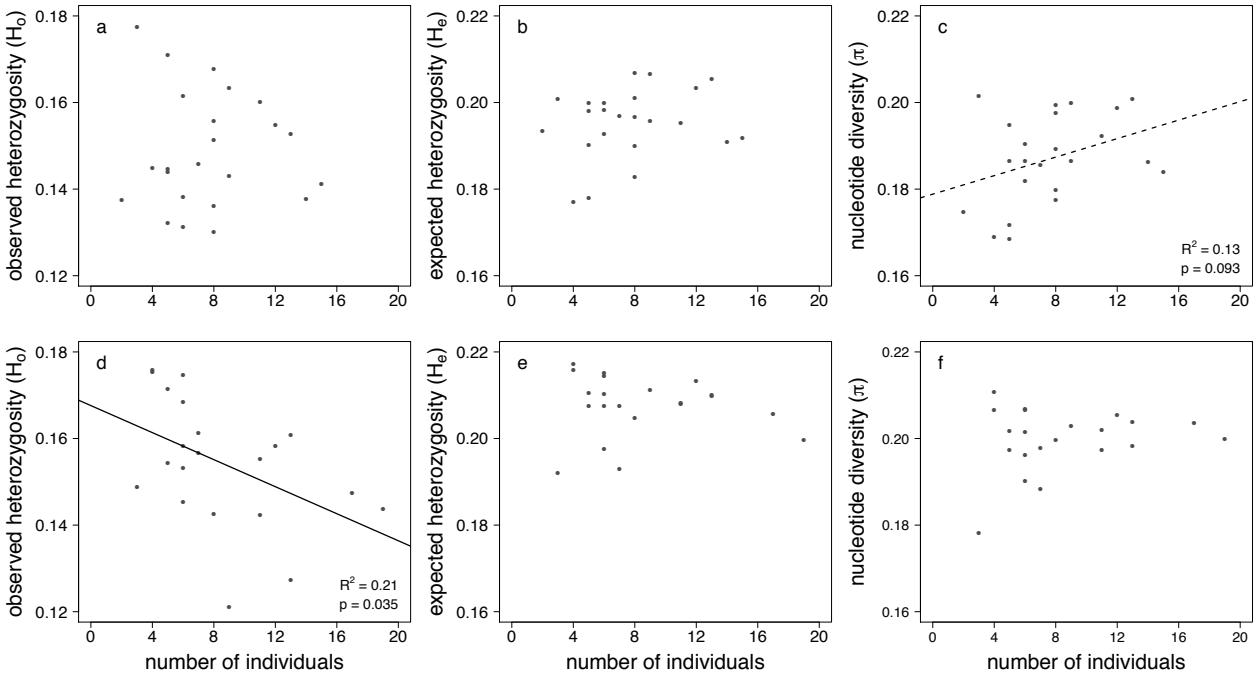

Figure A.1. Genetic diversity vs. number of individuals per population in hickories. Genetic diversity vs. number of individuals (n) for populations of *Carya cordiformis* (a-c) and *Carya ovata* (d-f). Statistically significant relationships are portrayed as solid lines, and marginally significant relationships as dashed lines.

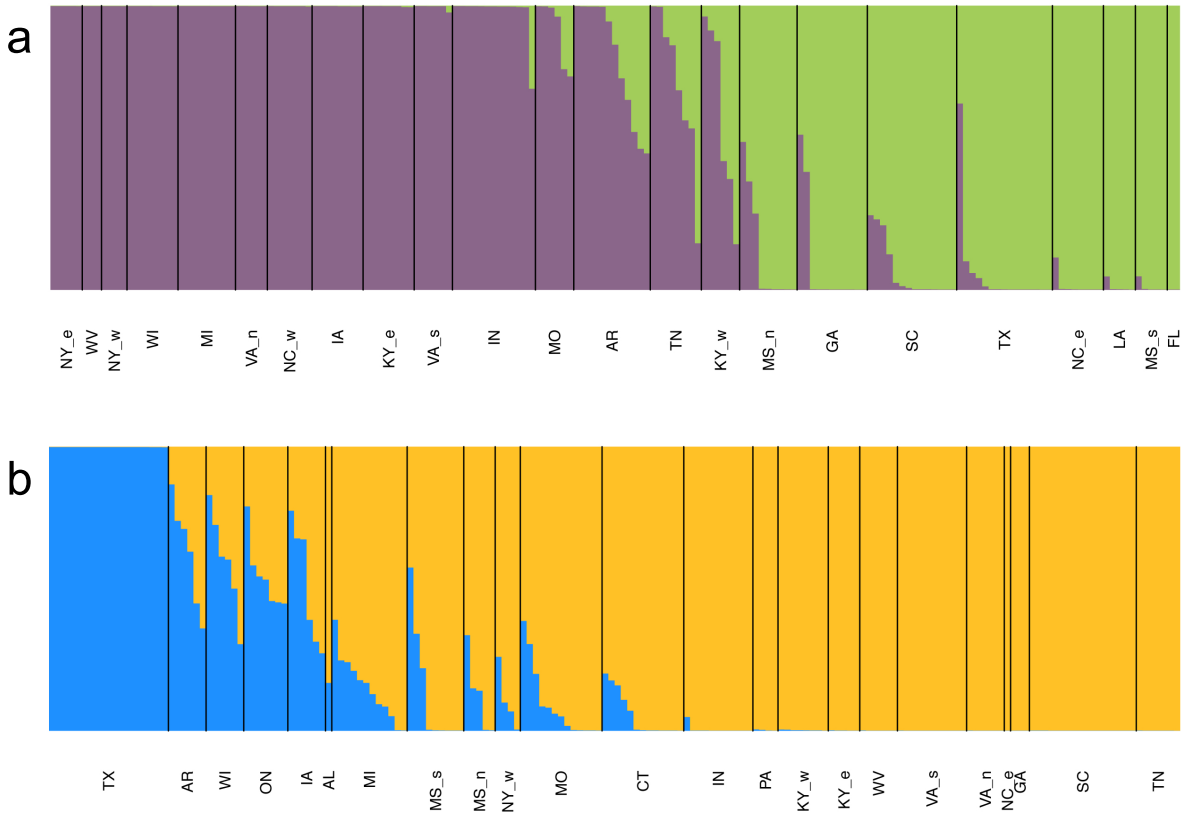

Figure A.2. Membership of hickory individuals in genetic clusters from FASTSTRUCTURE. Membership of (a) *Carya cordiformis* and (b) *Carya ovata* individuals in genetic clusters identified using FASTSTRUCTURE (Raj et al., 2014). Each vertical bar represents one individual, and colours represent the proportion of membership for that individual in each of two genetic clusters (K = 2). Individuals are grouped by population, with populations arranged in order of decreasing membership in Cluster 1 for each species. Population IDs are given in Tables A.1-A.2.

Supplementary literature cited

Budic, L., Didenko, G., & Dormann, C.F. (2016) Squares of different sizes: effect of geographical projection on model parameter estimates in species distribution modeling. *Ecology and Evolution*, **6**, 202–211.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., & Cresko, W.A. (2013) STACKS: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J.H. (2011) STACKS: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Daget, P. (1977) Le bioclimat méditerranéen: analyse des formes climatiques par le système d'Emberger. *Vegetatio*, **34**, 87–103.

Elith, J., Kearney, M., & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans, R.J., Phillips, S., Leathwick, J., & Elith, J. (2015) Package "dismo". R package, http://cran.r-project.org/web/packages/dismo/index.html.

Martin, M. (2011) CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10-12.

Meyer, M. & Kircher, M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **5**, 1–11.

Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M., & Anderson, R.P. (2014) ENMEVAL: an R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, **5**, 1198–1205.

O'Connell, B.M., LaPoint, E.B., Turner, J.A., Ridley, T., Boyer, D., Wilson, A.M., Waddell, K.L., & Conkling, B.L. (2012) *The Forest Inventory and Analysis Database: database description and users manual version 5.1.4 for Phase 2. Gen. Tech. Rep.,* US Department of Agriculture, Forest Service, Washington, DC.

Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., & Buerkle, A. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., & Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., & Blair, M.E. (2017) Opening the black box: an open-source release of MAXENT. *Ecography*, **40**, 887–893.

Phillips, S.J., Anderson, R.P., & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, **2004**, 655–662.

Raj, A., Stephens, M., & Pritchard, J.K. (2014) FASTSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Title, P.O. & Bemmels, J.B. (2018) ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291-307.

CHAPTER III

Contrasting northern and southern origins of postglacial expansion inferred for two eastern North American hickories (*Carya* spp.) from demographic and coalescent simulations and Approximate Bayesian Computation

Abstract

The locations where temperate taxa survived the Last Glacial Maximum (LGM) often remain vaguely defined, especially in Eastern North America (ENA). In particular, debate continues for all northern-hemisphere temperate deciduous forest regions about the existence of northern refugia and their potential contribution to postglacial recolonization relative to that of southern areas. We infer the geographic origins of postglacial expansion in two widespread, co-distributed hickories (*Carya*, Juglandaceae) from ENA, using the X-ORIGIN pipeline. Demographic and coalescent simulations are initiated from different origins based on a broad geographic prior, and spatial summary statistics are calculated for empirical and simulated datasets of >1,000 single nucleotide polymorphisms from 17-20 rangewide populations of each species. Approximate Bayesian Computation is used to compare empirical and simulated data, in order to infer the geographic origin of expansion. For bitternut hickory (*Carya cordiformis*), the estimated expansion origin is located in the northern Lower Mississippi

River Valley, whereas for shagbark hickory (*Carya ovata*) it is in the eastern Gulf Coastal Plain between the Mississippi and Apalachicola-Chattahoochee Rivers. In *C. cordiformis*, postglacial expansion likely occurred from relatively far north, whereas the lineages that contributed most to postglacial expansion in *C. ovata* had a more southern origin, highlighting the species-specific nature of migrational responses to Pleistocene glacial cycles.

Introduction

Temperate taxa have repeatedly experienced major range shifts in response to Pleistocene glacial cycles (Hewitt 1999; Shafer *et al.* 2010; Qiu *et al.* 2011). Understanding these range shifts has been a key focus of many phylogeographic studies, in part because of their implications for explaining contemporary population genetic structure (Hewitt 2000), identifying regions of long-term conservation priority (Hampe & Petit 2005), predicting migration capacity given future climate change (Feurdean *et al.* 2013), and inferring how dispersal limitation has impacted extinction rates and species richness patterns (Qian & Ricklefs 2000; Svenning & Skov 2007). The concept of glacial refugia has been central to explaining survival of temperate species during the Last Glacial Maximum (LGM). Although different definitions of glacial refugia exist (Bennett & Provan 2008), here we use the term broadly to refer to locations where temperate species survived the LGM (ca. 21,500 ± 1,500 years bp; Jackson *et al.* 2000), due to the persistence of favourable climates at either a regional or local scale.

Glacial refugia for temperate species were once thought to have occurred primarily in southern areas of all three temperate deciduous forest regions of the northern hemisphere. Major, well-known refugia include the Iberian, Italian, and Balkan peninsulas of Europe (Hewitt 1999), subtropical and tropical regions of East Asia (Harrison *et al.* 2001; Cao *et al.* 2015), and several different areas of southeastern North America (Soltis *et al.* 2006; Jaramillo-Correa *et al.* 2009). However, recent evidence has suggested that many temperate plants may have also survived in previously unknown northern refugia (Stewart & Lister 2001; Willis & Van Andel 2004; Magri *et al.* 2006; Hu *et al.* 2008; Magri 2008; Provan & Bennett 2008; Tian *et al.* 2009; Bai *et al.* 2010; Kimura *et al.* 2014; Zeng *et al.* 2015). In some cases, specific locations of putative northern refugia have been identified, such as the Carpathians in Europe (Provan & Bennett 2008; Daneck *et al.* 2016), and the Qin and Changbai Mountains in China (Bai *et al.* 2010; Zeng *et al.* 2015). In other cases, the locations of northern refugia have remained cryptic or poorly defined. Northern range limits during the LGM may be particularly difficult to delineate for species that survived in small areas known as microrefugia where unusual local microclimates allowed isolated populations to survive within a broader region of generally inhospitable climate (Stewart & Lister 2001; Willis & Van Andel 2004; Provan & Bennett 2008; Rull 2009, 2014). Whether northern micorefugia for temperate deciduous trees in fact existed and contributed to postglacial expansion remains a subject of controversy (Tzedakis *et al.* 2013; Rull 2014).

In Eastern North America (ENA), it has long been suspected that some temperate deciduous trees were fairly widespread during the LGM (Bennett 1985), but it is unclear whether distinct northern refugia existed for these species (Sewell *et al.* 1996;

Morris *et al.* 2010), or if they survived in a broader, less fragmented distribution covering much of unglaciated ENA (Bennett 1985; Magni *et al.* 2005; McLachlan *et al.* 2005). Unlike in Europe where the locations of major refugia have been well established, in ENA the general geographic extent of temperate deciduous trees during the LGM is not well known. LGM-age macrofossils of temperate deciduous trees have been found adjacent to the Lower Mississippi River Valley (Delcourt *et al.* 1980; Givens & Givens 1987), providing strong evidence that some temperate species survived at least locally within this region. In contrast, over much of southern ENA, the fossil pollen record suggests that LGM communities were conifer-dominated with no modern analogues, but frequently mixed with trace to low levels of pollen from temperate deciduous trees or of species characteristic of more open vegetation (Davis 1983; Bennett 1985; Jackson *et al.* 2000; Jackson & Williams 2004).

Even if complete LGM species distributions could be identified, not all LGM populations necessarily contributed to postglacial recolonization (Daneck *et al.* 2016). Phylogeographic studies are able to provide unique insight into patterns of postglacial recolonization because they allow inferences of genetic relationships among populations from different parts of the species range (Bennett & Provan 2008). Phylogeographic studies have not reached a general consensus about the distribution of temperate deciduous trees during the LGM, but have instead identified numerous regions that may have served as refugia, including the Gulf and Atlantic Coastal Plains, the Lower Mississippi River Valley, the Southern Appalachians, peninsular Florida, central Texas, and interior areas not far from ice sheets (McLachlan *et al.* 2005; Shaw & Small 2005; Soltis *et al.* 2006; Potter *et al.* 2008; Jaramillo-Correa *et al.* 2009; Morris *et*

al. 2010). However, traditional descriptive analyses of phylogeographic data, even in combination with interpretations of the fossil record and predictions of species distribution models, are subject to interpretation (Tzedakis *et al.* 2013). Recent advances in statistical phylogeography that incorporate coalescent theory (e.g., Carnaval *et al.* 2009; Knowles 2009; Barthe *et al.* 2017), including simulation-based frameworks that combine spatially explicit demographic and coalescent models paired with Approximate Bayesian Computation (ABC) (He *et al.* 2013, 2017), have allowed testing of a variety of highly specific phylogeographic hypotheses (Bemmels *et al.* 2016; Massatti & Knowles 2016) and may provide insight into questions unanswered by traditional approaches.

Given the uncertainty regarding the locations of glacial refugia and sources of postglacial expansion for temperate deciduous trees from ENA, tests of model-based scenarios using a statistical phylogeographic framework would help clarify the demographic expansion history of taxa from this region. Here, we use one such framework – the X-ORIGIN pipeline (He *et al.* 2017) – to infer the origin of postglacial expansion for two widespread, co-distributed hickory species (bitternut hickory, *Carya cordiformis* (Wangenh.) K.Koch, Juglandaceae; shagbark hickory, *C. ovata*, (Mill.) K.Koch). We use X-ORIGIN to conduct spatially explicit demographic and coalescent simulations initiated from different geographic expansion origins in order to generate genetic predictions for >1,000 single nucleotide polymorphism (SNP) loci. We then compute spatial genetic summary statistics and compare simulated datasets to empirical SNP data using ABC, in order to infer the most likely geographic origin of expansion. In particular, we aim to determine how far north the most likely origins of

expansion were located, and whether these origins were the same or differed between species.

Methods

Study species and genetic datasets

Carya cordiformis and *Carya ovata* are both widespread, common tree species native to the temperate deciduous forest biome of ENA, with ranges extending from eastern Texas to southern Quebec (Fig. B.1). Several disjunct populations of *C. ovata* also occur in the mountains of northern Mexico, but Mexican populations were not included in our analysis. Both species are able to inhabit a range of different sites (Graney 1990; Smith 1990), but *C. cordiformis* is generally a more mesic species common on fertile soils and bottomlands, whereas *C. ovata* is more frequently found in drier, upland habitats (Barnes & Wagner 2004). Both species are wind-pollinated and dispersed by animals. Previous descriptive phylogeographic analysis has revealed weak genetic structure and an isolation-by-distance pattern in both species, interpreted to suggest that both species were fairly geographically widespread over southern ENA during the LGM (Bemmels & Dick in review). There are no strong phylogeographic breaks in either species, except between a Texas population and all other populations of *C. ovata* (Bemmels & Dick in review).

Single nucleotide polymorphism (SNP) datasets of 1,046 and 1,018 unlinked, nuclear SNPs (Bemmels & Dick in review) were reutilized for *C. cordiformis* and *C. ovata*, respectively. However, populations represented by fewer than five individuals

were excluded due to low sample size, as well as the Texas population of *C. ovata* because it may correspond to a separate glacial refugium (Bemmels & Dick in review). This resulted in a total of 168 individuals from 20 populations in *C. cordiformis*, and 148 individuals from 17 populations in *C. ovata*, with an overall genotyping rate of 88% in both species.

Paleodistribution models

Species distribution models (SDMs) were also previously generated by Bemmels & Dick (in review). SDMs were constructed in *Maxent* v.3.4.1 (Phillips *et al.* 2004, 2006, 2017), using occurrence from the US Forest Service Forest Inventory Analysis Database (O'Connell *et al.* 2012), and environmental variables from the *WorldClim* v.1.4 (Hijmans *et al.* 2005) and ENVIREM (Title & Bemmels 2018) databases. Model optimization and variable selection was performed, and habitat suitability was projected for both the current time period, and the LGM for each of the CCSM4, MIROC-ESM, and MPI-ESM-P general circulation models (GCMs). Results were averaged into a single projection for the LGM because projections of all three GCMs were similar. For further details, see Bemmels & Dick (in review).

Demographic and coalescent simulations

Demographic and coalescent model simulations were performed for both species using the X-ORIGIN pipeline (He *et al.* 2017). The key components of X-ORIGIN are to (1) simulate a spatially explicit demographic model of population expansion, with relative carrying capacities of demes and migration rates scaled relative to habitat suitability

from SDMs; (2) simulate genetic data for each demographic simulation according to a coalescent topology determined by the simulated demographic history; and, (3) use summary statistics to compare simulated data to empirical genetic data and estimate the geographic origin of population expansion (Ω). Priors are specified on the latitude and longitude of Ω and other parameters in the demographic simulations, and posterior estimates of these parameters are obtained using Approximate Bayesian Computation (ABC). We do not describe full details of X-ORIGIN here (for details, see He *et al.* 2017), but we outline key information specific to our study below.

In order to generate the landscapes required to conduct demographic simulations, SDMs for three time periods were converted into downscaled raster landscapes at 25-arcmin resolution (approximately 46.3 km), with relative carrying capacities scaled according to habitat-suitability values from the SDMs and converted into 10 carrying-capacity bins of equal magnitude. The landscape for the intermediate time period was generated by averaging carrying capacities during the LGM and current time period, following He *et al.* (2013). Because one of our goals was to test for the possibility of expansion from northern refugia, which might have occurred in areas not predicted by SDMs to contain suitable habitat, all unglaciated areas with zero habitat suitability in the LGM landscape were converted to the lowest non-zero habitat-suitability bin so that it was possible to initiate simulations from these regions.

Demographic simulations were conducted in SPLATCHE2 (Ray *et al.* 2010) and were initiated from a source population with latitude and longitude of the central deme chosen from terrestrial areas of a uniform rectangular prior covering 25 to 39ºN and -103 to -74ºW (Table 3.1, Fig. 3.1). The southern, eastern, and western borders of this

area were selected so as to include areas predicted by SDMs to contain suitable habitat for each species during the LGM, while the northern border was selected to extend to the southern edge of glaciation. Source populations were initiated in a diamond-shaped area represented by a central deme, plus an extension of three additional demes in all cardinal directions. This created an area seven demes wide (for a total of 175 arcmin, or approximately 324 km; Fig. 3.1).

By using this procedure, we were able to initiate demographic simulations from a central area ca. 324 km wide instead of from a point location represented by a single deme. Preliminary tests revealed that simulations initiated from a single deme generated low levels of genetic diversity relative to the empirical data, which could indicate that collapsing all lineages into a single ancestral deme created an unrealistically severe genetic bottleneck. The choice to model a starting area seven demes wide was arbitrary, but we chose an area this size in order to provide sufficient initial demes to avoid enforcing a severe bottleneck, without initiating simulations from such a large area that the geographic source of expansion would become overly broad and ill defined. Although the size of the starting area was fixed (and not a parameter of interest in our simulations), the actual number of individuals contained in the initial demes varied in different simulations, depending on the prior for maximum carrying capacity.

Demographic simulations were initiated at 21.5 ka (i.e., the LGM) and individuals were allowed to colonize the landscape according to the carrying capacity (K; scaled relative to K_{max} based on habitat suitability) of each deme, and the proportional migration rate (m). Priors on K_{max} and m followed a log-uniform distribution (Table 3.1).

Habitat suitability maps representing the LGM, intermediate time period, and current conditions were each successively utilized for one third of the total generations (He et al. 2013). A generation time of 50 years was selected, roughly corresponding to the minimum time to reach peak reproduction (50-125 years in *C. cordiformis*, and 60-200 years in *C. ovata*; Graney 1990; Smith 1990), for a total of 430 generations over the 21,500 years of each simulation.

After the demographic portion of each simulation, genetic data were generated for the same number of individuals per population and number of SNPs as in the empirical data. Coalescent genetic simulations were performed for sampled individuals, with the ancestry of alleles traced from the present backwards to the initial demes at 21.5 ka. At this time, all initial demes were collapsed into a single ancestral source population of size N_{anc} (following a log-uniform prior; Table 3.1) in which all alleles were allowed to coalesce. The SNP mutation model in SPLATCHE2 was used to generate genotypes for each individual according to the simulated pattern of coalescence, with a minimum minor allele frequency (MAF) fixed at 0.033 to match the MAF specified in the empirical data (Bemmels & Dick in review). A total of 500,000 demographic and genetic simulations were performed for each species.

Estimating the origin of expansion

Summary statistics that contain spatial information (pairwise Ψ and F_{ST}; He *et al.*, 2017) were generated for both the empirical data (Table B.1-B.2) and all simulated datasets, using ARLEQUIN v.3.5 (Excoffier & Lischer 2010) and scripts provided with X-ORIGIN. For the purpose of calculating the directionality index Ψ (Peter & Slatkin 2013) in

the empirical data, ancestral alleles for each SNP were defined as the allele at highest frequency in populations from areas predicted by the SDMs to have contained suitable habitat during the LGM.

Approximate Bayesian Computation (ABC; Beaumont *et al.* 2002) was used to compare empirical to simulated summary statistics, and infer the geographic expansion origin (Ω) and other demographic parameters. Summary statistics were transformed into principal components (PCs), and the 2,500 simulations (0.5%) with transformed summary statistics most closely matching the empirical data were selected using ABCTOOLBOX (Wegmann *et al.* 2010). A point location for Ω was estimated from a two-dimensional kernel density of expansion origins in the 2,500 retained simulations, using scripts provided with X-ORIGIN. The default method of computing the kernel density in X-ORIGIN assigns different weights to retained simulations based on the inverse of the computed distance between the summary statistics for each retained simulation and the empirical data. However, we applied a $\log_{10}$ transformation to the default weighting values prior to calculating the kernel density and Ω, in order to reduce the effect of a few extreme outliers caused by incredibly small distances between the retained and empirical summary statistics, which tended to overwhelm the signal of other retained simulations. Posterior distributions of other demographic parameters (K_{max}, N_{anc}, m) were estimated from the retained simulations using ABC-GLM adjustment (Leuenberger & Wegmann 2010) implemented in ABCTOOLBOX (Wegmann *et al.* 2010).

In order to determine how well the models are able to generate the observed data, we calculated Wegmann's *et al*. (2010) *p*-value, which is the proportion of simulations with a smaller likelihood than that of the empirical data. Small *p*-values

indicate that the model is unable to generate the observed data (Wegmann *et al.* 2010). We also examined *p*-values across different numbers of retained PC axes for spatial summary statistics, in order to determine the optimal number of PC axes to retain for estimating *Ω*. We aimed to retain the largest number of PC axes possible (thus explaining the maximum amount of variation in the summary statistics of the simulated datasets) for which the *p*-value was still high (assessed *post hoc*).

In addition, we calculated the average error in estimating *Ω*, by estimating *Ω* from pseudo-observed datasets (PODs) with known expansion origins using the same estimation procedure as for the empirical data. Ten PODs were systematically generated from each possible deme contained within the geographic prior, and we calculated the geographic distance between the inferred *Ω* and the true simulated expansion origin for each POD using the *R* package *geosphere* (Hijmans *et al.* 2017).

Results

The optimal number of PC axes to retain for *C. cordiformis* was four (Table 3.2a), together explaining 67.7% of the overall variation and resulting in a *p*-value of 0.87. The estimated expansion origin (*Ω*) was located in the northern Lower Mississippi River Valley (LMRV) in northwestern Arkansas (Fig. 3.2a). Additional area with a high likelihood of serving as the expansion origin extended over the northern LMRV and northwest of this region. For *C. ovata*, 3 PC axes were retained (Table 3.2b), together explaining 63.4% of the overall variation with a *p*-value of 1.00, while *Ω* was located in the Florida Panhandle (Fig. 3.2b). However, areas with high likelihood of serving as the

expansion origin extended over much of the eastern Gulf Coastal Plain between the Mississippi and Apalachicola-Chattahoochee Rivers, and the southern LMRV.

We note that Wegmann's *p*-value was very high when only the first few PC axes of variation in summary statistics were retained, but rapidly declined as more PC axes were retained (Table 3.2). This suggests that it was easy for models to generate simulated datasets that closely matched the empirical data in the first several PC axes, but it was difficult to generate simulated datasets that also closely matched the empirical data in subsequent PC axes. However, the first few PC axes explained the majority of variation among simulations (Table 3.2). We also note that the estimated expansion origin was somewhat sensitive to the number of axes retained (Fig. B.2-B.3). As the number of retained PC axes increased, the inferred Ω in *C. cordiformis* shifted slightly southwestward into the Ozark and Ouachita Mountains, and in *C. ovata*, it shifted westward to exposed continental shelf off the coast of Louisiana. However, Wegmann's *p*-value was low in these cases, suggesting that our models are a poorer fit to the empirical data for higher PC dimensions and that the inferred Ω in these cases should be interpreted with caution.

High values were obtained for posterior estimates (Fig. 3.3) of migration rate (m) and maximum carrying capacity (K_{max}), likely because substantial population connectivity is required to generate the low overall genetic differentiation among populations present in the empirical data (Table B.1-B.2). High m is reasonable given that *Carya* likely has great potential for pollen-mediated gene flow by wind and long-distance seed dispersal by animal vectors. High K_{max} is also not unreasonable given that both species are common and form major components of forest cover in some

areas. Ancestral population size (N_{anc}) was poorly estimated, suggesting the simulated summary statistics did not contain information relevant to estimating this parameter. However, estimating K_{max}, m, and N_{anc} was not our primary interest.

The average error in estimating Ω from pseudo-observed datasets (PODs) was low in *C. cordiformis* (median = 250 km; mean = 338 km) and moderately low in *C. ovata* (median = 325 km; mean = 421 km). These errors are comparable to the width of the area from which geographic simulations were initiated (7 demes wide, ca. 324 km). However, the average error varied widely depending on the true geographic origin of expansion (Fig. 3.4). When expansion occurred from corners and edges of the geographic prior, error was much higher than when expansion occurred from near the central portion of the geographic prior. In both species, the inferred Ω from the empirical dataset was located in a region of low error (Fig. 3.4).

Discussion

By using demographic and coalescent simulations paired with ABC in the X-ORIGIN pipeline (He *et al.* 2017), we have estimated the geographic origins of postglacial expansion for two widespread eastern North American deciduous trees. In *Carya cordiformis*, the area with a high likelihood of serving as the expansion origin was mostly restricted to the northern Lower Mississippi River Valley (LMRV), whereas in *Carya ovata* this area covered parts of the eastern Gulf Coastal Plain between the Mississippi and Apalachicola-Chattahoochee Rivers and the southern LMRV. The inferred expansion origin for *C. cordiformis* was located relatively far north, whereas a

more southern origin was inferred for *C. ovata,* suggesting that the species responded differently to Pleistocene glacial cycles. Our models reveal the putative source of genetic lineages involved in population expansion but do not reveal complete LGM geographic distributions and likely do not capture all aspects of a species' postglacial demographic history. Nonetheless, they provide unique insight that complements other methods of phylogeographic inference. Similar studies in other species, as well as tests of competing alternative hypotheses, would help increase the understanding of the phylogeographic history of ENA as a whole.

The origins of population expansion

Our results highlight the central role of the northern LMRV and the eastern Gulf Coastal Plain as sources of postglacial expansion in *C. cordiformis* and *C. ovata*, respectively. The LMRV has previously been identified as an important refugium for temperate deciduous trees from the macrofossil record (Delcourt *et al.* 1980; Givens & Givens 1987; Jackson *et al.* 2000) and it is noteworthy that the inferred expansion origin for *C. cordiformis* occurs near one of the few areas where the LGM presence of *Carya* has been firmly established (Delcourt *et al.* 1980). Due to a general paucity of LGM-age fossil sites from ENA (Jackson *et al.* 2000), the complete geographic distribution of *Carya* during the LGM is unknown. However, interpolated isopollen maps suggest the potential for trace to low amounts of *Carya* pollen over a fairly broad geographic area, including much of the eastern Gulf Coastal Plain (Prentice *et al.* 1991; Jackson *et al.* 2000) in areas with a high likelihood of expansion origin for *C. ovata*.

The geographic area with a high likelihood of expansion origin is relatively small in both species, but we note that X-ORIGIN predicts the origin of lineages that contributed to postglacial recolonization (He *et al.* 2017), not the complete LGM geographic distribution of the entire species. It is possible that additional populations existed in areas identified as having a low likelihood, but that these populations made little or no contribution to postglacial expansion. There is some evidence to suggest that this is the case in *C. ovata*. A Texas population of this species (excluded in our analyses, see *Methods*) was previously identified as genetically distinct from other populations and interpreted as possibly representing a separate glacial refugium (Bemmels & Dick in review), suggesting that *C. ovata* was also present in the western Gulf Coastal Plain during the LGM. Similarly, the inferred northern expansion origin in *C. cordiformis* does not indicate that the species was absent from more southern areas. Instead, genetic clustering analysis previously revealed weakly differentiated northern and southern clusters in *C. cordiformis*, suggesting that populations from the species' northern range edge during the LGM likely made the greatest contribution to postglacial recolonization, with little contribution of southern lineages (Bemmels & Dick in review).

Our results suggest that the Atlantic Coastal Plain was extremely unlikely to have served as the expansion origin for either species (Fig. 3.2). Although it is possible that both species were absent from this area during the LGM, we also speculate that if LGM populations did exist in this region, the southern Appalachians might have impeded the northward migration of these populations relative to that of populations from west of the Appalachians. The southern Appalachians have previously been inferred to represent a phylogeographic barrier in some tree species, but for other species they had little effect

on migration, and in some cases may even have served as a refugium (Jaramillo-Correa *et al.* 2009). Thus, we cannot conclusively infer the cause of the lack of expansion signal from the Atlantic Coastal Plain.

Species-specific range shifts

Although *C. cordiformis* and *C. ovata* are roughly codistributed today throughout ENA (Fig. B.1), the inferred origin for *C. cordiformis* was located much farther north than for *C. ovata* (Fig. 3.2), suggesting that the two species experienced different demographic histories. Identifying concordant demographic signatures among different taxa has been an important tool of comparative phylogeography (Soltis *et al.* 2006; Avise 2009), yet species differing in traits such as habitat affinity and dispersal ability often exhibit different phylogeographic patterns and population genetic structure (Stewart *et al.* 2010; Correa Ribeiro *et al.* 2016; Massatti & Knowles 2016; Papadopoulou & Knowles 2016). The prevalence of non-analogue pollen assemblages over much of ENA during and following the LGM (Davis 1983; Overpeck *et al.* 1992; Jackson *et al.* 2000; Jackson & Williams 2004) also strongly suggests that responses to Pleistocene glacial cycles were generally species-specific.

Our models are unable to determine why the inferred expansion origins differ between *C. cordiformis* and *C. ovata*, but we speculate that differences in habitat affinity could have conceivably generated the patterns we observe. In particular, LGM microclimates near the northern LMRV may have been more suitable for long-term persistence of a mesic species such as *C. cordiformis* than of more upland species such as *C. ovata*. The regionally most abundant trees in the LMRV during the LGM

were *Picea, Larix*, and other boreal species, but local presence of *Carya, Fagus, Juglans, Populus*, and *Quercus* has been established from the blufflands immediately adjacent to the LMRV proper (Delcourt *et al.* 1980). During the LGM, glacial outwash from ice sheets created a massive valley train dozens of kilometers wide covering low-elevation areas of the LMRV (Smith 1996), which Delcourt *et al.* (1980) suggest likely impacted the microclimate of surrounding regions. In particular, they suggest that frequent advective fogs may have created a cool, humid microclimate in the LMRV blufflands, allowing local persistence of mesic temperate deciduous species (Delcourt *et al.* 1980). This type of microclimate might have supported populations of *C. cordiformis*, but is less likely to have been favourable for long-term persistence of *C. ovata.* Instead, the inferred expansion origin for *C. ovata* is located farther south in the eastern Gulf Coastal Plain, an area characterized by low hills and rolling plains. The warmer climates and more varied topography of this region might have provided more expansive areas of suitable habitat for a less mesic, predominantly upland species like *C. ovata.*

Implications of a northern expansion origin

In *C. cordiformis*, the estimated expansion origin is located relatively far north (36.0ºN). Regional dominance of *Picea* and *Pinus* in LGM pollen assemblages over much of ENA at similar latitudes suggests that temperate deciduous trees occurred only locally or were minor components of communities where they occurred (Delcourt *et al.* 1980; Davis 1983; Jackson *et al.* 2000). As such, the inferred expansion origin likely corresponds to a region containing microrefugia where *C. cordiformis* would have been able to survive in climatically and topographically ideal microsites. Our findings for *C.*

cordiformis tentatively support the emerging understanding that microrefugia, in addition to or instead of better known southern macrorefugia, played an important role in LGM survival and postglacial recolonization of some temperate tree species (Stewart & Lister 2001; Willis & Van Andel 2004; Provan & Bennett 2008; Rull 2009; Hampe & Jump 2011).

Increasing evidence for the existence of northern microrefugia suggests that estimates of postglacial migration rates of some temperate tree species may need to be revised (Feurdean *et al.* 2013). The inability of typically observed dispersal distances to explain the rapid migration rates required in order for temperate tree species to have colonized their current northern range limits out of southern refugia is known as Reid's paradox (Reid 1899; Clark *et al.* 1998). Characterization of dispersal kernels and has generally highlighted the importance of long-distance dispersal in efforts to account for Reid's paradox (Clark *et al.* 1998). However, expansion out of northern microrefugia would mean that some species have migrated much shorter distances than previously thought (Feurdean *et al.* 2013). Interestingly, fossil pollen records suggest that following glacial retreat, *Carya* became abundant in the Midwest by about 12 to 9 ka, but did not reach the northern Appalachians until 9 to 6 ka (Davis 1983; Prentice *et al.* 1991), which Prentice *et al.* (1991) attribute to changes in regional climatic patterns. Alternatively, the Midwest may have been colonized long before the northern Appalachians simply due to the geographic proximity of the Midwest to the inferred expansion origin of *C. cordiformis*. However, the precise species driving this pattern have not been identified, and the expansion origins of most other *Carya* species are unknown. In contrast, the

inferred expansion origin for *C. ovata* is located far to the south (30.5ºN), suggesting that it was able to rapidly migrate great distances since glacial retreat.

It has frequently been noted that populations from glacial refugia should be a conservation priority, due to the high genetic diversity and genetic uniqueness of these regions (Petit *et al.* 2003; Hampe & Petit 2005; Médail & Diadema 2009). Although little is known about what happens to non-refugial populations when climates deteriorate (Stewart *et al.* 2010), they are often believed to go extinct, rather than progressively tracking climates back into refugia (Bennett *et al.* 1991; Dalén *et al.* 2007). Conserving populations from areas that contain suitable habitats during both glacial and interglacial periods may therefore be critical not only for conservation of genetic diversity, but also for long-term species survival (Hampe & Petit 2005; Stewart *et al.* 2010; Hampe & Jump 2011). However, in both *C. cordiformis* and *C. ovata*, the estimated expansion origin is located at the periphery or outside of the species' current distribution (Fig. B.1), rather than in areas of long-term habitat stability. These results may explain why no areas of elevated or unique genetic diversity were identified in either species (Bemmels & Dick in review). Better characterizing rear-edge populations close to the inferred source of expansion could help determine if they contain unique genetic features of conservation interest (Hampe & Petit 2005).

Model limitations and future directions

Validation of the X-ORIGIN procedure using pseudo-observed datasets (PODs) revealed that average error in estimating the expansion origin was low (Fig. 3.4). These results suggest that – if the assumptions of the model are realistic – the estimated

origins are likely to be fairly accurate to within a few hundred kilometers. However, in *C. ovata* the area of high likelihood of the expansion origin is spread over several parts of the eastern Gulf Coastal Plain and southern LMRV (Fig. 3.2b), reducing our confidence in the precise geographic coordinates of *Ω*. In addition, given that results were somewhat sensitivity to the number of PC axes retained (Fig. B.2-B.3), we suggest that the point estimates of the expansion origin should be treated as approximate. However, the general region of ENA in which *Ω* is located is well supported.

Like any modelling exercise, our models contain a number of assumptions and simplifications, and likely do not capture the entire, true postglacial history of either species. This may be one reason why our models are unable to generate the observed data for higher numbers of PC axes (as suggested by low *p*-values). The higher-order PC axes may increasingly reflect the effects of locally distributed alleles in the empirical data that are only found at high frequency in a few populations, or the effects of unusual patterns of gene flow between particular populations, which our rangewide models are not designed to accommodate.

In particular, our models make at least two major assumptions that may strongly affect the range of genetic patterns able to be simulated. Firstly, we modelled only a single origin of population expansion. This decision was based on previous results (Bemmels & Dick in review) that revealed no signatures of multiple genetically isolated refugia in either species, except for in a Texas population of *C. ovata* that we excluded here. Secondly, we modelled range expansion from a specific geographic region (Fig. 3.1), rather than from a diffuse or continuous distribution. We expected that if expansion occurred from a multiple refugia or a very diffuse area, we might obtain results with high

error or with signal of expansion from multiple regions. Instead, we obtained strong signal for a high likelihood of expansion from a specific region in both species (Fig. 3.2). The procedure we utilized could easily be modified into a model-testing framework (e.g., He *et al.* 2013; Bemmels *et al.* 2016; Massatti & Knowles 2016), in order to assess support for alternative scenarios such expansion from a single vs. multiple refugia, or from a diffuse latitudinal band rather than a localized refugium. The use of X-ORIGIN and other ABC frameworks to test these and other scenarios would be worthwhile for expanding our knowledge of historical range shifts for other taxa from ENA, especially those that might exhibit very different phylogeographic patterns from our study species.

Acknowledgements

The authors thank the following organizations for assistance coordinating fieldwork: Berea College Forest, Connecticut State Forests (SFs), Daniel Boone National Forest (NF), Davey Crockett NF, George Washington NF, Holly Springs NF, Hoosier NF, Jefferson NF, Kisatchie NF, Land Between the Lakes National Recreation Area, Lower Wisconsin State Riverway, Mark Twain NF, Matthaei Botanical Gardens, Monongahela NF, Murray State University Hancock Biological Station, Nantahala NF, Oconee NF, Ozark NF, Prentice Cooper SF, Shimek SF, Stephen F. Austin State University and Experimental Forest, Society of Ontario Nut Growers, Sumter NF, Tombigbee NF, the University of Michigan E.S. George Reserve, and Uwharrie NF. The authors also thank B.J. Belcher, P. Cousineau, R. D'Andrea, J. Ronson, D. Saenz, and P. Tichenor for assistance with fieldwork; Q. He for extensive guidance with

implementing the X-ORIGIN pipeline; and the National Science Foundation (GRFP fellowship; DDIG 1501159), the University of Michigan Department of Ecology and Evolutionary Biology, and Rackham Graduate School for graduate student support and research funding.

Author contributions

JBB and CWD conceived the project; LLK suggested analysis methods; JBB analyzed the data, with input from LLK; JBB wrote the chapter with input from CWD and LLK.

Literature cited

Alberto FJ, Aitken SN, Alía R *et al.* (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Bai WN, Liao WJ, Zhang DY (2010) Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytologist*, **188**, 892–901.

Barnes B V., Wagner WH (2004) *Michigan Trees*. University of Michigan Press, Ann Arbor.

Barthe S, Binelli G, Hérault B *et al.* (2017) Tropical rainforests that persisted: inferences from the Quaternary demographic history of eight tree species in the Guiana shield. *Molecular Ecology*, **26**, 1161–1174.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Bemmels JB, Dick CW (in review) Genomic evidence of a widespread southern

distribution during the Last Glacial Maximum for two eastern North American hickory species. *Journal of Biogeography,* manuscript ID JBI-17-0564.R1.

Bemmels JB, Title PO, Ortego J, Knowles LL (2016) Tests of species-specific models reveal the importance of drought in postglacial range shifts of a Mediterranean-climate tree: insights from integrative distributional, demographic and coalescent modelling and ABC model selection. *Molecular Ecology*, **25**, 4889–4906.

Bennett KD (1985) The spread of *Fagus grandifolia* across eastern North America during the last 18 000 years. *Journal of Biogeography*, **12**, 147–164.

Bennett KD, Provan J (2008) What do we mean by "refugia"? *Quaternary Science Reviews*, **27**, 2449–2455.

Bennett AKD, Tzedakis PC, Willis KJ (1991) Quaternary refugia of north European trees. *Journal of Biogeography*, **18**, 103–115.

Cao X, Herzschuh U, Ni J, Zhao Y, Böhmer T (2015) Spatial and temporal distributions of major tree taxa in eastern continental Asia during the last 22,000 years. *Holocene*, **25**, 79–91.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

Clark JS, Fastie C, Hurtt G *et al.* (1998) Reid's paradox of rapid plant migration. *BioScience*, **48**, 13–24.

Correa Ribeiro P, Lemos-Filho JP, de Oliveira Buzatti RS, Lovato MB, Heuertz M (2016) Species-specific phylogeographical patterns and Pleistocene east-west divergence in Annona (Annonaceae) in the Brazilian Cerrado. *Botanical Journal of the Linnean Society*, **181**, 21–36.

Dalén L, Nyström V, Valdiosera C *et al.* (2007) Ancient DNA reveals lack of postglacial habitat tracking in the arctic fox. *Proceedings of the National Academy of Sciences*, **104**, 6726–6729.

Daneck H, Fér T, Marhold K (2016) Glacial survival in northern refugia? Phylogeography of the temperate shrub *Rosa pendulina* L. (Rosaceae): AFLP vs. chloroplast DNA variation. *Biological Journal of the Linnean Society*, **119**, 704–718.

Davis MB (1983) Quaternary history of deciduous forests of eastern North America and Europe. *Annals of the Missouri Botanical Garden*, **70**, 550–563.

Delcourt PA, Delcourt HR, Brister RC, Lackey LE (1980) Quaternary vegetation history of the Mississippi Embayment. *Quaternary Research*, **13**, 111–132.

Excoffier L, Lischer HEL (2010) ARLEQUIN suite ver 3.5: a new series of programs to

perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Feurdean A, Bhagwat SA, Willis KJ *et al.* (2013) Tree migration-rates: narrowing the gap between inferred post-glacial rates and projected rates. *PLoS ONE*, **8**, e71797.

Givens CR, Givens FM (1987) Age and significance of fossil white spruce (*Picea glauca*), Tunica Hills, Louisiana-Mississippi. *Quaternary Research*, **27**, 283–296.

Graney DL (1990) *Carya ovata* (Mill.) K. Koch. In: *Silvics of North America: 2. Hardwoods. Agricultural Handbook 654* (eds Burns RM, Honkala BH). US Department of Agriculture, Forest Service, Washington, DC.

Hampe A, Jump AS (2011) Climate relicts: past, present, future. *Annual Review of Ecology, Evolution, and Systematics*, **42**, 313–333.

Hampe A, Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters*, **8**, 461–467.

Hamrick JL, Godt JW, Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woody plant species. *New Forests*, **6**, 95–124.

Harrison SP, Yu G, Takahara H, Prentic IC (2001) Diversity of temperate plants in east Asia. *Nature*, **413**, 129–130.

He Q, Edwards DL, Knowles LL (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

He Q, Prado JR, Knowles LL (2017) Inferring the geographic origin of a range expansion: latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program X-ORIGIN. *Molecular Ecology*, **26**, 6908–6920.

Hewitt G (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

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

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans RJ, Williams E, Vennes C (2017) Package “geosphere”. R package version 1.5-7. https://cran.r-project.org/web/packages/geosphere/index.html.

Hu LJ, Uchiyama K, Shen HL *et al.* (2008) Nuclear DNA microsatellites reveal genetic variation but a lack of phylogeographical structure in an endangered species, Fraxinus mandshurica, across north-east China. *Annals of Botany*, **102**, 195–205.

Jackson ST, Webb RS, Anderson KH *et al.* (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews*, **19**, 489–508.

Jackson ST, Williams JW (2004) Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences*, **32**, 495–537.

Jaramillo-Correa JP, Beaulieu J, Khasa DP, Bosquet J (2009) Inferring the past from the present phylogeography structure of North American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, **39**, 286–307.

Kimura MK, Uchiyama K, Nakao K *et al.* (2014) Evidence for cryptic northern refugia in the last glacial period in Cryptomeria japonica. *Annals of Botany*, **114**, 1687–1700.

Knowles LL (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.

Magni CR, Ducousso A, Caron H, Petit RJ, Kremer A (2005) Chloroplast DNA variation of *Quercus rubra* L. in North America and comparison with other Fagaceae. *Molecular Ecology*, **14**, 513–524.

Magri D (2008) Patterns of post-glacial spread and the extent of glacial refugia of European beech (*Fagus sylvatica*). *Journal of Biogeography*, **35**, 450–463.

Magri D, Vendramin GG, Comps B *et al.* (2006) A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist*, **171**, 199–221.

Massatti R, Knowles LL (2016) Contrasting support for alternative models of genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. *Molecular Ecology*, **25**, 3974–3986.

McLachlan JS, Clark JS, Manos PS (2005) Molecular indicators of tree migration capacity under rapid climate change. *Ecology*, **86**, 2088–2098.

Médail F, Diadema K (2009) Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, **36**, 1333–1345.

Morris AB, Graham CH, Soltis DE, Soltis PS (2010) Reassessment of phylogeographical structure in an eastern North American tree using Monmonier's algorithm and ecological niche modelling. *Journal of Biogeography*, **37**, 1657–1667.

O'Connell BM, LaPoint EB, Turner JA *et al.* (2012) *The Forest Inventory and Analysis Database: database description and users manual version 5.1.4 for Phase 2. Gen.*

Tech. Rep. Washington, DC.

Overpeck JT, Webb RS, Webb T (1992) Mapping eastern North American vegetation change over the past 18,000 years: no analogs and the future. *Geology*, **20**, 1071–1074.

Papadopoulou A, Knowles LL (2016) A paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences*, **113**, 8018–8024.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Petit RJ, Aguinagalde I, de Beaulieu J-L *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME (2017) Opening the black box: an open-source release of MAXENT. *Ecography*, **40**, 887–893.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips SJ, Dudík M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, **2004**, 655–662.

Potter K, Dvorak W, Crane B *et al.* (2008) Allozyme variation and recent evolutionary history of eastern hemlock (*Tsuga canadensis*) in the southeastern United States. *New Forests*, **35**, 131–145.

Prentice C, Bartlein PJ, Webb III T (1991) Vegetation and climate change in eastern North America since the Last Glacial Maximum. *Ecology*, **72**, 2038–2056.

Provan J, Bennett KD (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology and Evolution*, **23**, 564–571.

Qian H, Ricklefs RE (2000) Large-scale processes and the Asian bias in temperate plant species diversity. *Nature*, **407**, 180–182.

Qiu YX, Fu CX, Comes HP (2011) Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics and Evolution*, **59**, 225–244.

Ray N, Currat M, Foll M, Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics*, **26**, 2993–2994.

Reid C (1899) *The Origin of the British Flora*. Dulau, London.

Rull V (2009) Microrefugia. *Journal of Biogeography*, **36**, 481–484.

Rull V (2014) Macrorefugia and microrefugia: a response to Tzedakis *et al*. *Trends in Ecology and Evolution*, **29**, 243–244.

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.

Sewell M, Parks C, Chase M (1996) Intraspecific chloroplast DNA variation and biogeography of North American *Liriodendron* L. (Magnoliaceae). *Evolution*, **50**, 1147–1154.

Shafer ABA, Cullingham CI, Côté SD, Coltman DW (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.

Shaw J, Small RL (2005) Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *American Journal of Botany*, **92**, 2011–2030.

Smith HC (1990) *Carya cordiformis* (Wangenh.) K. Koch. In: *Silvics of North America: 2. Hardwoods. Agricultural Handbook 654* (eds Burns RM, Honkala BH). US Department of Agriculture, Forest Service, Washington, DC.

Smith LM (1996) Fluvial geomorphic features of the Lower Mississippi alluvial valley. *Engineering Geology*, **45**, 139–165.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Stewart JR, Lister AM (2001) Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology and Evolution*, **16**, 608–613.

Stewart JR, Lister AM, Barnes I, Dalen L (2010) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 661–671.

Svenning JC, Skov F (2007) Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology Letters*, **10**, 453–460.

Tian B, Liu R, Wang L *et al.* (2009) Phylogeographic analyses suggest that a deciduous species (*Ostryopsis davidiana* Decne., Betulaceae) survived in northern China during the Last Glacial Maximum. *Journal of Biogeography*, **36**, 2148–2155.

Title PO, Bemmels JB (2018) ENVIREM: an expanded set of bioclimatic and topographic

variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291–307.

Tzedakis PC, Emerson BC, Hewitt GM (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology and Evolution*, **28**, 696–704.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCTOOLBOX: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

Willis KJ, Van Andel TH (2004) Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. *Quaternary Science Reviews*, **23**, 2369–2387.

Zeng YF, Wang WT, Liao WJ, Wang HF, Zhang DY (2015) Multiple glacial refugia for cool-temperate deciduous trees in northern East Asia: the Mongolian oak as a case study. *Molecular Ecology*, **24**, 5676–5691.

Table 3.1. Priors on the expansion origin and demographic parameters in hickories. Priors on the geographic origin of expansion and demographic parameters were the same for both species.

Parameter	Description	Distribution	Minimum	Maximum
latitude	latitude of Ω (º)	uniform	25	39
longitude	longitude of Ω (º)	uniform	-103	-74
N_{anc}	ancestral population size prior to initiating expansion (number of individuals)	log-uniform	$10^{3.3}$	$10^{5.0}$
K_{max}	maximum carrying capacity of a deme (number of individuals)	log-uniform	$10^{3.3}$	$10^{5.0}$
m	migration rate between neighbouring demes (proportion of individuals migrating per generation)	log-uniform	$10^{-2.0}$	$10^{-0.3}$

Table 3.2. Summary of model results from X-ORIGIN in hickories. Summary of results for (a) *Carya cordiformis* and (b) *Carya ovata* when different numbers of principal component (PC) axes are retained in estimating *Ω*. The cumulative % variation refers to the cumulative amount of variation in spatial summary statistics explained by the retained PC axes. Also reported are the marginal density of the model, Wegmann's *p*-value (Wegmann *et al.* 2010), and the estimated expansion origin (*Ω*). The optimal number of PC axes to retain based on changes in Wegmann's *p*-value (assessed *post hoc*) is indicated with an arrow.

(a)

Retained PC axes	Cumulative % variation explained	Marginal density	*p*-value	*Ω*	
2	56.5	1.27×10^{-2}	1.00	28.0ºN, 97.1ºW	
3	62.8	3.96×10^{-3}	0.99	34.3ºN, 90.4ºW	
4	67.7	1.65×10^{-3}	0.87	36.0ºN, 90.8ºW	←
5	72.0	1.29×10^{-4}	0.12	33.9ºN, 92.9ºW	

(b)

Retained PC axes	Cumulative % variation explained	Marginal density	*p*-value	*Ω*	
2	56.3	3.14×10^{-2}	1.00	30.5ºN, 87.0ºW	
3	63.4	4.09×10^{-2}	1.00	30.5ºN, 87.0ºW	←
4	67.9	3.60×10^{-3}	0.18	29.2ºN, 92.9ºW	
5	72.2	6.56×10^{-3}	0.34	28.4ºN, 91.2ºW	

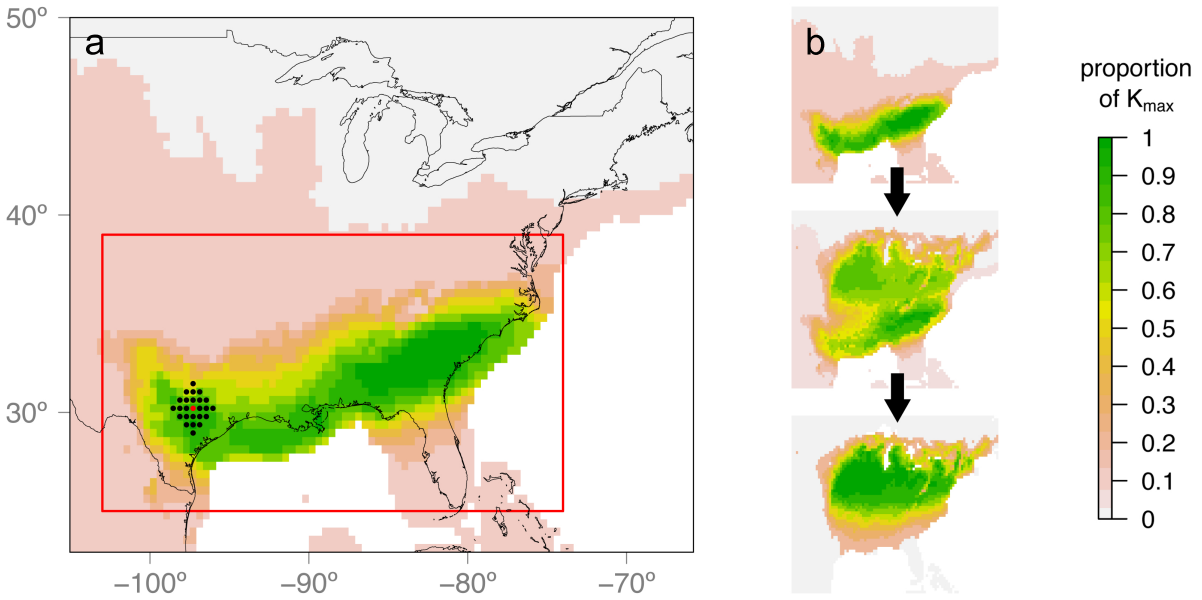

Figure 3.1. Schematic overview of X-ORIGIN simulations in hickories. Schematic overview of demographic simulations, with an example provided for *Carya cordiformis*. (a) Simulations were initiated in the Last Glacial Maximum (LGM) landscape from a central deme (red dot) plus an area extending three additional demes (black dots) in all directions. The location of the central deme was selected from a uniform geographic prior (red rectangle). Carrying capacity of all demes was scaled relative to maximum carrying capacity (K_{max}) according to habitat suitability from a species distribution model (SDM). Note that all unglaciated areas with zero habitat suitability were converted to the lowest LGM habitat suitability bin (K_{max}*0.1) so that simulations could be initiated from these areas to test for the possibility of expansion from microrefugial areas not predicted by the SDM to have high LGM habitat suitability. Individuals were allowed to colonize the landscape, with (b) relative carrying capacities shifting from the LGM to present. After demographic simulations were completed, genetic coalescent simulations were performed (not shown) tracing ancestry of alleles backward from sampled populations into the initial demes from (a), which were then collapsed into a single ancestral source population of size N_{anc} and allowed to coalesce.

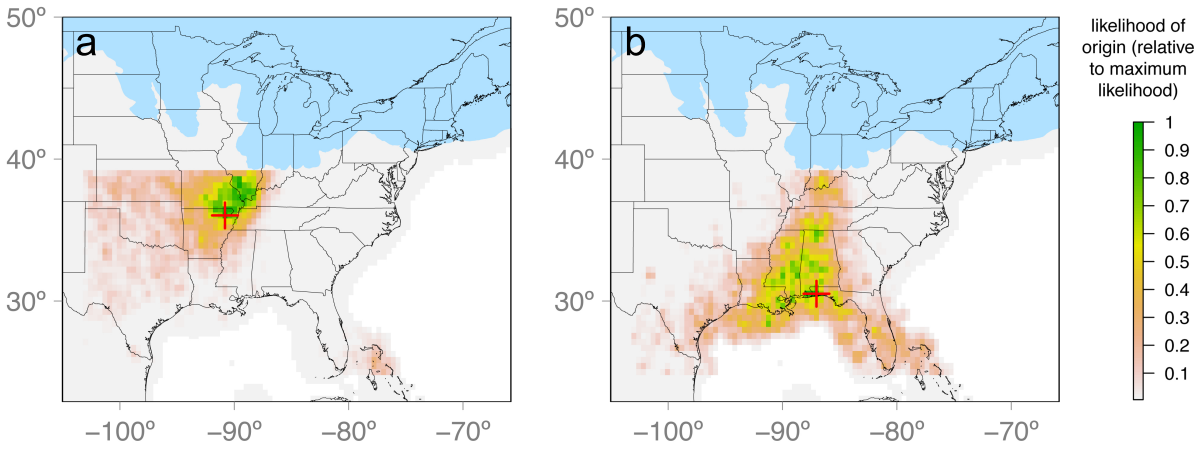

Figure 3.2. Estimated expansion origins of hickories. Estimated expansion origin (*Ω*; red cross) in (a) *Carya cordiformis* and (b) *Carya ovata*. The shading of pixels depicts the kernel density showing the likelihood that each pixel served as the expansion origin, relative to the pixel with the maximum likelihood (i.e., *Ω*). Glaciated regions are shown in blue. The results presented in (a) and (b) are based on retention of 4 and 3 PC axes of variation in summary statistics, respectively. For results with a different number of PC axes retained, see Fig. B.2-B.3.

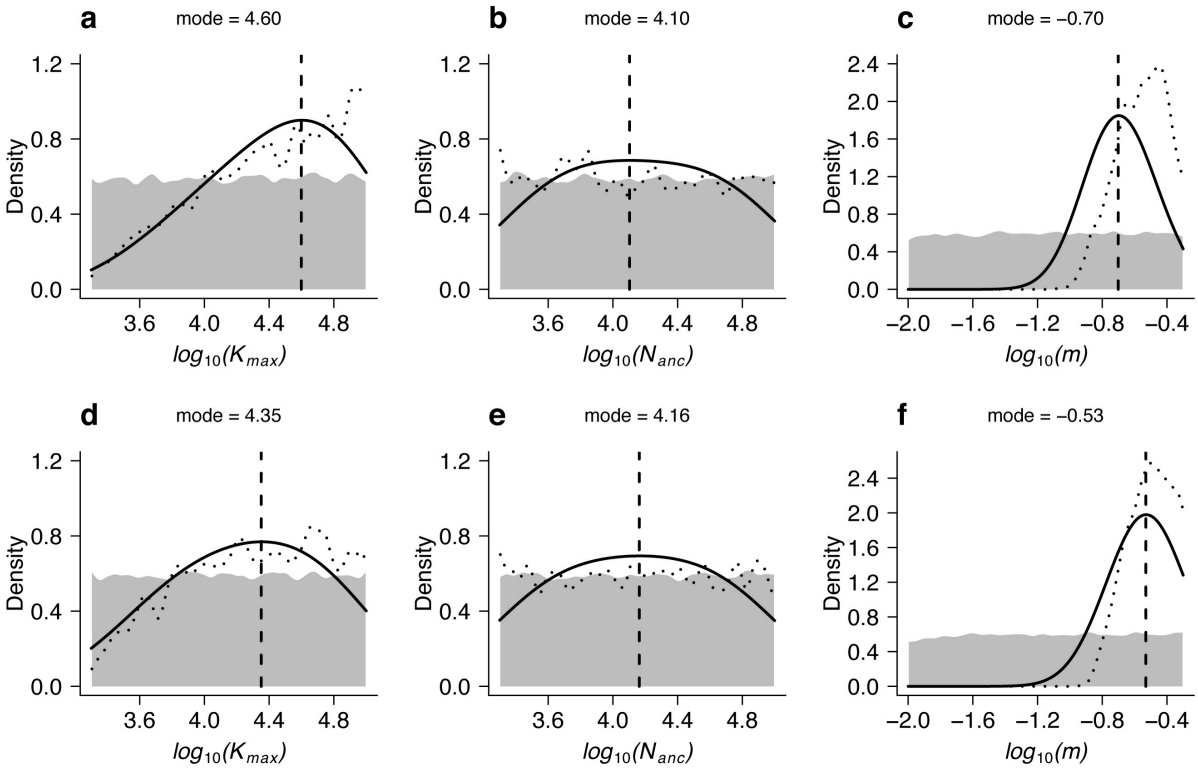

Figure 3.3. Posterior estimates of demographic parameters for hickories. Parameter estimates (dashed vertical lines) and prior and posterior distributions for (a-c) *Carya cordiformis* and (d-f) *Carya ovata*. Grey shading: prior distribution; dotted line: posterior distribution of raw retained simulations; solid black line: posterior distribution after ABC-GLM regression adjustment (Leuenberger & Wegmann 2010); K_{max}: maximum carrying capacity; N_{anc}: ancestral population size, m: migration rate.

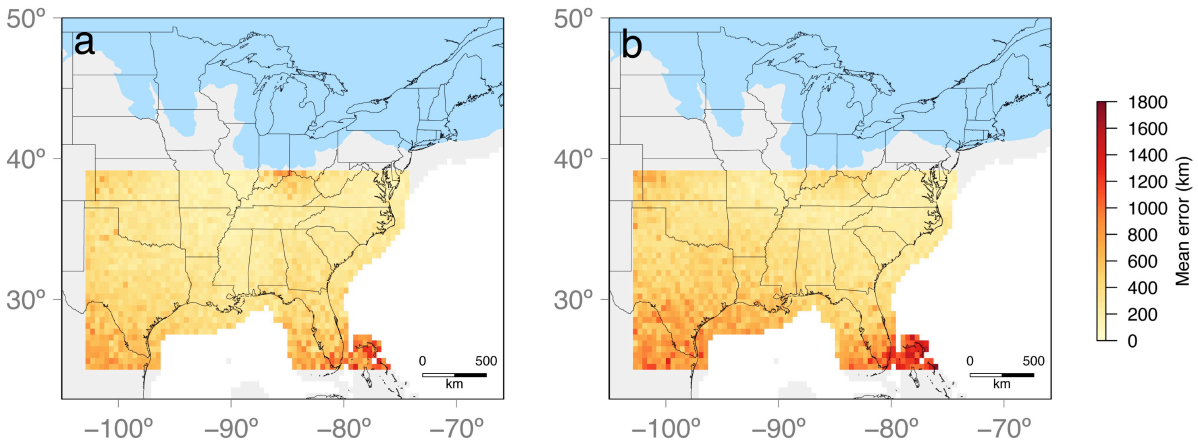

Figure 3.4. Mean error in estimating the expansion origin of hickories. Mean error in estimating Ω from pseudo-observed datasets (PODs) for (a) *Carya cordiformis* and (b) *Carya ovata.* Colours of each pixel show the mean geographic distance between the inferred Ω and the true expansion origin, when the true origin was located in that pixel. Mean errors of each pixel are calculated from a total of 10 PODs, except for several pixels in the southwest corner of the map where not all of the 10 simulations resulted in successful colonization of all populations.

APPENDIX B

Supplementary information from Chapter III

Table B.1. *Carya cordiformis* empirical summary statistics (pairwise F_{ST} and Ψ). Empirical summary statistics for *Carya cordiformis*. Below the diagonal: population pairwise F_{ST}; above the diagonal: population pairwise Ψ. For geographic locations of populations, see Fig. B.1a.

	AR	GA	IA	IN	KYe	KYw	LA	MI	MO	MSn	MSs	NCe	NCw	NYe	SC	TN	TX	VAn	VAs	WI
AR	–	0.00	-0.01	0.00	-0.03	-0.02	-0.03	-0.02	-0.04	-0.01	-0.04	-0.04	-0.03	-0.10	0.02	-0.03	0.01	-0.06	-0.04	-0.02
GA	0.04	–	-0.02	0.00	-0.03	-0.03	-0.03	-0.02	-0.05	-0.02	-0.05	-0.04	-0.03	-0.11	0.02	-0.03	0.01	-0.05	-0.05	-0.02
IA	0.03	0.05	–	0.01	-0.02	-0.01	-0.02	0.00	-0.03	0.00	-0.03	-0.03	-0.02	-0.09	0.03	-0.02	0.03	-0.05	-0.03	-0.01
IN	0.02	0.04	0.03	–	-0.03	-0.02	-0.04	-0.02	-0.04	-0.02	-0.05	-0.04	-0.03	-0.10	0.02	-0.03	0.02	-0.06	-0.04	-0.02
KYe	0.03	0.06	0.04	0.03	–	0.01	0.00	0.01	-0.02	0.02	-0.02	-0.01	0.00	-0.07	0.05	0.00	0.04	-0.03	-0.02	0.00
KYw	0.02	0.02	0.02	0.02	0.03	–	-0.01	0.01	-0.02	0.00	-0.02	0.00	-0.01	-0.08	0.04	-0.01	0.04	-0.04	-0.02	0.01
LA	0.05	0.03	0.07	0.06	0.06	0.03	–	0.01	-0.01	0.01	-0.01	0.00	0.00	-0.08	0.05	0.00	0.05	-0.02	-0.01	0.02
MI	0.02	0.04	0.02	0.01	0.03	0.02	0.07	–	-0.03	0.00	-0.03	-0.01	-0.01	-0.08	0.04	-0.01	0.04	-0.04	-0.03	0.00
MO	0.02	0.04	0.04	0.02	0.04	0.02	0.05	0.03	–	0.03	0.01	0.01	0.01	-0.05	0.06	0.01	0.06	-0.02	0.00	0.03
MSn	0.03	0.04	0.05	0.03	0.07	0.03	0.04	0.05	0.05	–	-0.03	-0.02	-0.02	-0.09	0.03	-0.02	0.03	-0.05	-0.03	0.00
MSs	0.04	0.04	0.06	0.03	0.07	0.04	0.05	0.05	0.06	0.05	–	0.00	0.02	-0.06	0.06	0.00	0.06	-0.02	0.00	0.03
NCe	0.05	0.04	0.06	0.06	0.08	0.05	0.08	0.07	0.08	0.08	0.08	–	0.01	-0.06	0.06	0.00	0.06	-0.02	-0.01	0.01
NCw	0.04	0.05	0.05	0.04	0.06	0.04	0.08	0.03	0.05	0.07	0.09	0.08	–	-0.06	0.05	0.00	0.05	-0.03	-0.02	0.01
NYe	0.08	0.09	0.07	0.06	0.09	0.08	0.12	0.06	0.08	0.12	0.14	0.15	0.09	–	0.12	0.06	0.12	0.04	0.05	0.08
SC	0.02	0.02	0.04	0.02	0.05	0.01	0.04	0.03	0.03	0.03	0.03	0.05	0.06	0.11	–	-0.05	0.00	-0.08	-0.06	-0.04
TN	-0.04	-0.06	-0.03	-0.05	0.00	-0.03	0.00	-0.04	-0.02	0.01	0.01	0.05	-0.01	0.06	0.01	–	0.05	-0.03	-0.02	0.01
TX	0.02	0.01	0.04	0.03	0.05	0.02	0.03	0.04	0.04	0.04	0.03	0.07	0.06	0.11	0.03	0.01	–	-0.07	-0.06	-0.04
VAn	0.04	0.06	0.05	0.04	0.05	0.03	0.08	0.04	0.04	0.07	0.06	0.08	0.05	0.08	0.06	-0.03	0.07	–	0.01	0.04
VAs	0.05	0.07	0.05	0.04	0.04	0.05	0.09	0.04	0.05	0.07	0.08	0.08	0.06	0.09	0.06	-0.01	0.05	0.06	–	0.02
WI	0.02	0.05	0.01	0.02	0.04	0.02	0.07	0.01	0.03	0.05	0.07	0.08	0.05	0.06	0.04	0.00	0.06	0.04	0.04	–

Table B.2. *Carya ovata* empirical summary statistics (pairwise F_{ST} and Ψ).
Empirical summary statistics for *Carya ovata*. Below the diagonal: population pairwise F_{ST}; above the diagonal: population pairwise Ψ. For geographic locations of populations, see Fig. B.1b.

	AR	CT	IA	IN	KYe	KYw	MI	MO	MSn	MSs	ON	SC	TN	VAn	VAs	WI	WV
AR	–	0.03	-0.01	0.02	-0.01	0.02	0.04	0.03	-0.02	0.03	-0.03	0.05	0.01	-0.03	0.02	-0.03	-0.01
CT	0.04	–	-0.03	0.00	-0.04	-0.01	0.00	0.00	-0.05	0.00	-0.06	0.01	-0.02	-0.05	-0.01	-0.05	-0.04
IA	0.02	0.03	–	0.04	0.00	0.02	0.04	0.03	-0.02	0.04	-0.02	0.05	0.02	-0.02	0.03	-0.02	-0.01
IN	0.06	0.02	0.05	–	-0.03	-0.01	0.01	0.00	-0.04	0.00	-0.05	0.01	-0.02	-0.04	0.00	-0.06	-0.04
KYe	0.05	0.03	0.04	0.02	–	0.03	0.04	0.04	-0.02	0.04	-0.02	0.05	0.02	-0.02	0.03	-0.01	-0.01
KYw	0.04	0.03	0.04	0.02	0.02	–	0.02	0.01	-0.03	0.02	-0.04	0.02	-0.01	-0.04	0.00	-0.04	-0.03
MI	0.02	0.01	0.01	0.02	0.03	0.01	–	0.00	-0.06	-0.01	-0.06	0.01	-0.03	-0.06	-0.01	-0.06	-0.05
MO	0.03	0.02	0.03	0.02	0.03	0.02	0.01	–	-0.05	0.00	-0.05	0.02	-0.02	-0.05	-0.01	-0.06	-0.04
MSn	0.05	0.04	0.07	0.04	0.03	0.04	0.04	0.05	–	0.06	-0.01	0.07	0.03	-0.01	0.04	-0.01	0.00
MSs	0.04	0.03	0.04	0.02	0.03	0.02	0.02	0.03	0.05	–	-0.06	0.01	-0.03	-0.06	-0.01	-0.06	-0.05
ON	0.00	-0.02	-0.01	0.02	0.03	0.01	0.00	0.01	0.04	0.03	–	0.07	0.04	0.00	0.05	0.00	0.01
SC	0.05	0.02	0.05	0.01	0.02	0.01	0.03	0.02	0.04	0.02	0.04	–	-0.03	-0.07	-0.02	-0.07	-0.06
TN	0.07	0.04	0.07	0.02	0.04	0.02	0.04	0.04	0.06	0.04	0.04	0.02	–	-0.04	0.01	-0.03	-0.03
VAn	0.07	0.03	0.07	0.03	0.05	0.03	0.04	0.03	0.06	0.05	0.04	0.02	0.04	–	0.04	0.00	0.01
VAs	0.05	0.02	0.04	0.01	0.02	0.02	0.03	0.02	0.04	0.01	0.04	0.02	0.02	0.03	–	-0.04	-0.03
WI	0.04	0.07	0.05	0.08	0.08	0.08	0.05	0.07	0.08	0.09	0.02	0.08	0.10	0.08	0.08	–	0.01
WV	0.05	0.02	0.04	0.02	0.02	0.02	0.03	0.02	0.04	0.03	0.06	0.03	0.03	0.03	0.02	0.08	–

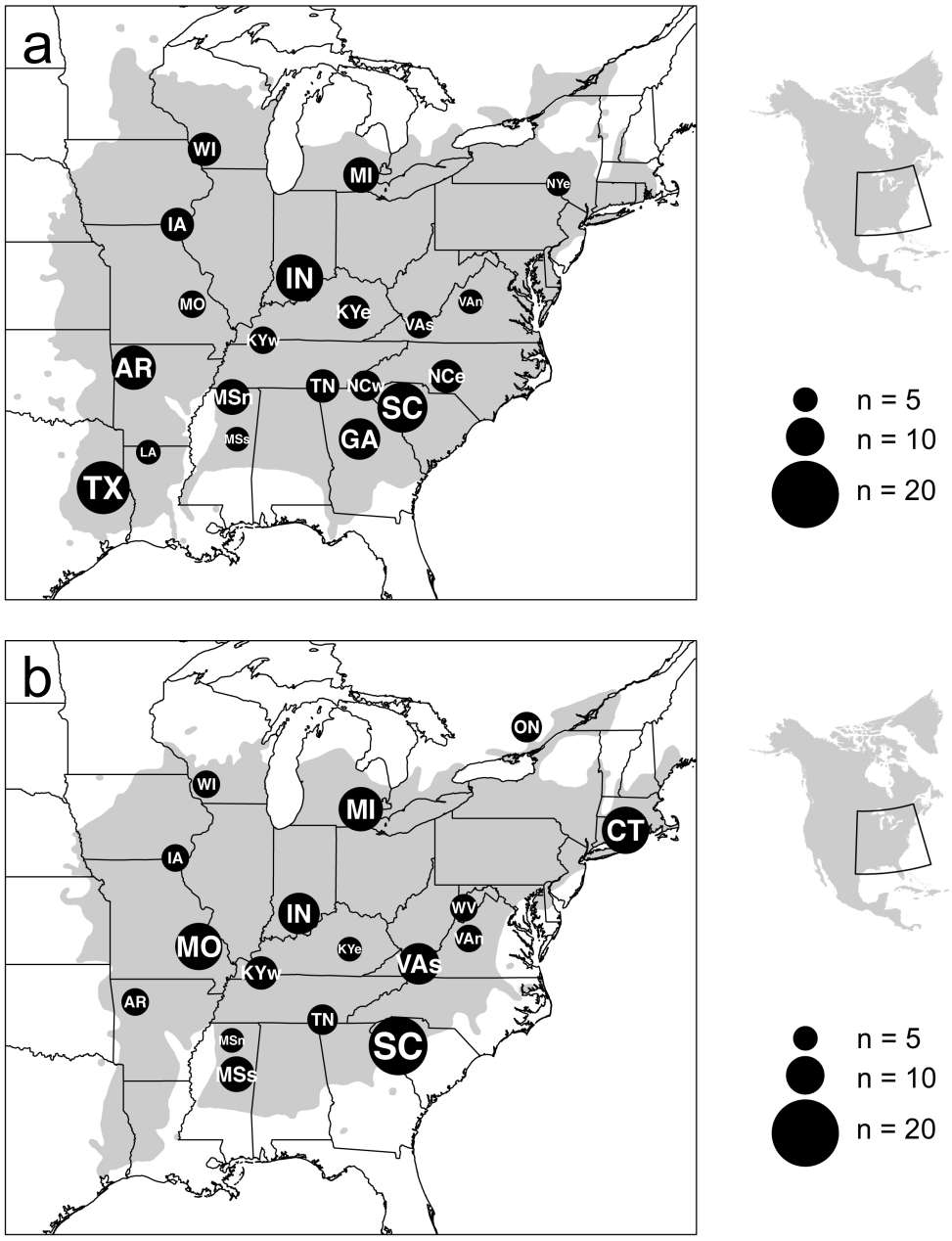

Figure B.1. Geographic locations of sampled populations of hickories. Geographic locations of sampled populations for (a) *Carya cordiformis* and (b) *Carya ovata*. The size of the black circle is proportional to the number of individuals sampled in each population, with population IDs written in white text. The geographic distribution of each species is shown in grey shading (Little 1971).

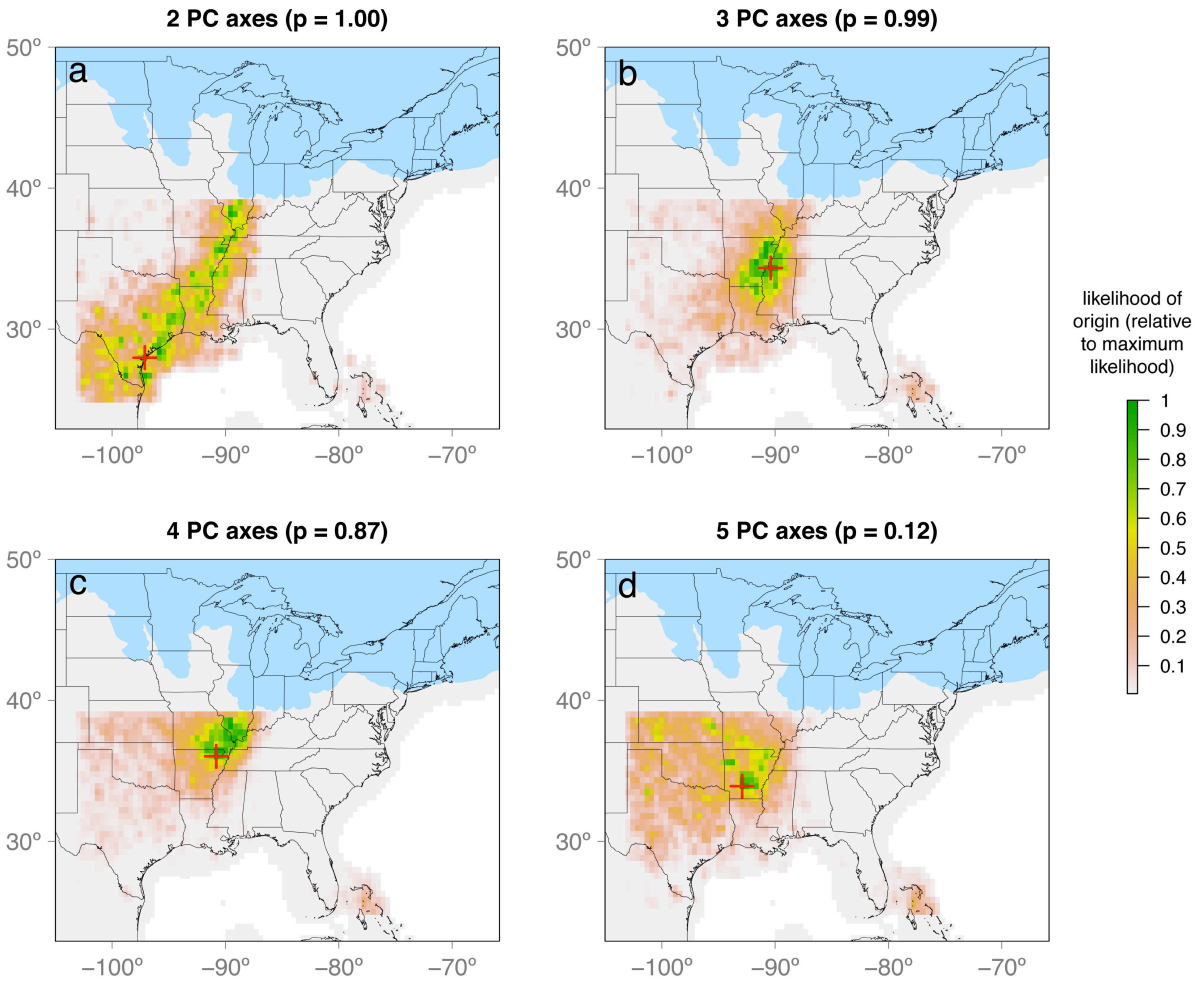

Figure B.2. *Carya cordiformis* expansion origin inferred from different numbers of PC axes. Estimated expansion origin (*Ω*; red cross) in *Carya cordiformis* when different numbers of PC axes are retained. The shading of pixels depicts the kernel density showing the likelihood that each pixel served as the expansion origin, relative to the pixel with the maximum likelihood (i.e., *Ω*). Glaciated regions are shown in blue. Wegmann's *p*-value of the model (Wegmann *et al.* 2010) is reported for each number of PC axes retained. Note that low *p*-values indicate the model is unable to easily generate the empirical data for this number of PC axes, and results for low *p*-values should be interpreted with caution.

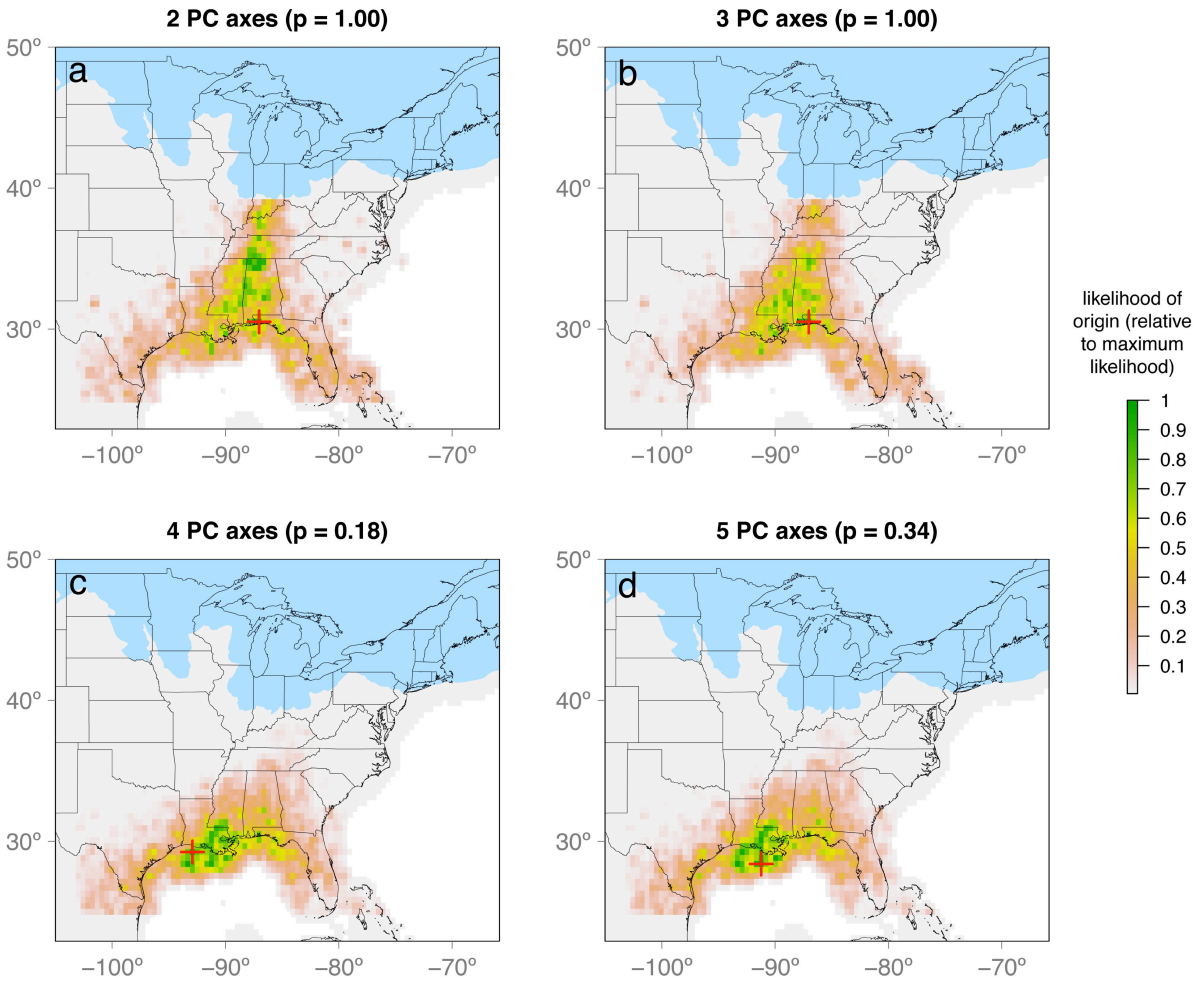

Figure B.3. *Carya ovata* expansion origin inferred from different numbers of PC axes. Estimated expansion origin (*Ω*; red cross) in *Carya ovata* when different numbers of PC axes are retained. The shading of pixels depicts the kernel density showing the likelihood that each pixel served as the expansion origin, relative to the pixel with the maximum likelihood (i.e., *Ω*). Glaciated regions are shown in blue. Wegmann's *p*-value of the model (Wegmann *et al.* 2010) is reported for each number of PC axes retained. Note that low *p*-values indicate the model is unable to easily generate the empirical data for this number of PC axes, and results for low *p*-values should be interpreted with caution.

Supplementary literature cited

Little EL (1971) Atlas of United States trees, volume 1. Conifers and important hardwoods. USDA Miscellaneous Publication 1146, Washington, DC.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier, L (2010) ABCTOOLBOX: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

CHAPTER IV

Tests of species-specific models reveal the importance of drought in postglacial range shifts of a Mediterranean-climate tree: insights from integrative distributional, demographic, and coalescent modelling and ABC model selection

Abstract

Past climate change has caused shifts in species distributions and undoubtedly impacted patterns of genetic variation, but the biological processes mediating responses to climate change, and their genetic signatures, are often poorly understood. We test six species-specific biologically informed hypotheses about such processes in canyon live oak (*Quercus chrysolepis*) from the California Floristic Province. These hypotheses encompass the potential roles of climatic niche, niche multidimensionality, physiological trade-offs in functional traits, and local-scale factors (microsites and local adaptation within ecoregions) in structuring genetic variation. Specifically, we use ecological niche models (ENMs) to construct temporally dynamic landscapes where the processes invoked by each hypothesis are reflected by differences in local habitat suitabilities. These landscapes are used to simulate expected patterns of genetic variation under each model and evaluate the fit of empirical data from 13 microsatellite loci genotyped in 226 individuals from across the species range. Using Approximate Bayesian

Computation (ABC), we obtain very strong support for two statistically indistinguishable models: a trade-off model in which growth rate and drought tolerance drive habitat suitability and genetic structure, and a model based on the climatic niche estimated from a generic ENM, in which the variables found to make the most important contribution to the ENM have strong conceptual links to drought stress. The two most probable models for explaining patterns of genetic variation thus share a common component, highlighting the potential importance of seasonal drought in driving historical range shifts in a temperate tree from a Mediterranean climate where summer drought is common.

Introduction

Shifts in species distributions in response to climate change are a key factor structuring population genetic variation in both temperate and tropical species (Taberlet *et al.* 1998; Soltis *et al.* 2006; Carnaval *et al.* 2009; Morgan *et al.* 2011; Massatti & Knowles 2016). However, the biological mechanisms governing these shifts and their potential impact on patterns of neutral genetic variation are often poorly understood. For example, some plant species may be associated with ecological microsites partly or wholly defined by non-climatic factors (e.g., John *et al.* 2007; Frei *et al.* 2012; Allié *et al.* 2015) that could constrain responses to regional-scale climate change (Kroiss & HilleRisLambers 2015). Likewise, geographic distributions may be limited by different abiotic stresses (e.g. cold temperatures, drought) among species (Normand *et al.* 2009), or by different factors in different geographic regions of a single species' range (Morin *et al.* 2007). Consequently, more detailed species-specific hypotheses about the causes of

range shifts and their impacts on population genetic structure are needed (Papadopoulou & Knowles 2016). To this end, we develop and test a suite of competing biologically-informed models (Table 4.1) to explain the genetic structure of canyon live oak (*Quercus chrysolepis* Liebm., Fagaceae). These models make different predictions about patterns of genetic variation, depending upon the relative importance of climatic niche, niche multidimensionality, physiological trade-offs in functional traits, and local-scale factors (e.g., microsites and local adaptation within ecoregions) in governing the species' distribution and demographic history since the Last Glacial Maximum (LGM, 21.5 ka).

Considering that canyon live oak is a member of the climatically and ecologically heterogeneous California Floristic Province (CFP) of western North America and is distributed across a wide range of elevations (90 to 2,740 m; Thornburgh 1990), the response of this species to shifts in climate might be associated with different aspects of its ecology. For example, canyon live oak grows on many soil types and in many forest and chaparral communities (Thornburgh 1990), but is found exclusively in regions of high topographic complexity (Little 1971). Likewise, it is common throughout California, Oregon, and Baja California (Fig. 4.1), but is most abundant in sheltered canyons and on steep, rocky slopes, where it may be the dominant tree species (Thornburgh 1990). Consequently, while regions with climates similar to those of its present distribution likely existed in California's flat Central Valley during the LGM (Ortego *et al.* 2015), the climatic niche by itself may not accurately represent past distributional shifts in regions where topographic complexity is very low. Alternatively, it is possible that shifts in distributions due to past climate change might reflect constraints due to trade-offs in

functional and physiological traits. For example, a trade-off between drought tolerance and growth rate may exist in species from climates with hot, dry summers (Howe *et al.* 2003; Alberto *et al.* 2013; Aitken & Bemmels 2016), and drought determines range limits of some plant species, including trees (Morin *et al.* 2007; Normand *et al.* 2009; Linares & Tíscar 2011; Rasztovits *et al.* 2014; Urli *et al.* 2014). Moreover, in many temperate trees, a trade-off between growth rate and cold tolerance drives population-level local adaptation (Howe *et al.* 2003; Savolainen *et al.* 2007; Alberto *et al.* 2013; Aitken & Bemmels 2016) and may determine species range limits (Loehle 1998; but, see Morin *et al.* 2007 for a counterperspective). Given geographic variation in functional traits in many tree species, it is also possible that geographic range shifts in response to climate change will depend strongly on individual responses of specific populations to unique environmental factors (e.g., Davis *et al.* 2001; Pearman *et al.* 2010; Benito Garzón *et al.* 2011; Valladares *et al.* 2014; Gotelli & Stanton-Geddes 2015; Hällfors *et al.* in press). Lastly, the response to past climate change might simply reflect shifts in habitat suitability as it relates to basic climate variables, without the need to invoke complex, species-specific nuances of niche or mechanistic trade-offs in functional traits. Basic climate variables (e.g., temperature, precipitation) are frequently used in correlative ecological niche models (ENMs) to model species distributions and to predict how distributions have changed over time (Alvarado-Serrano & Knowles 2014). In canyon live oak specifically, previous work has shown that patterns of genetic connectivity and admixture among populations are correlated with areas of high habitat suitability since the LGM, as predicted by a climatic ENM (Ortego *et al.* 2015).

It is these types of biologically informed hypotheses that motivate this study (as opposed to generic statistical phylogeographic tests; reviewed in Papadopoulou & Knowles 2016). Specifically, through tests of six models (Table 4.1) we explore the relative support for alternative hypotheses about the niche of canyon live oak and factors that may have driven its response to climate change, including basic climate variables, microsites, niche multidimensionality, trade-offs in functional traits, and local adaptation within ecoregions. We use integrative distributional, demographic and coalescent (iDDC) modelling (Knowles & Alvarado-Serrano 2010; Brown & Knowles 2012; He *et al.* 2013) to generate genetic expectations under each model, and Approximate Bayesian Computation (ABC; Beaumont *et al.* 2002; Csilléry *et al.* 2010) to evaluate the fit of empirical data characterized from 13 microsatellite loci in 226 indiviuals sampled across the species range to the genetic predictions of each model. We highlight how careful extraction of spatially explicit information from ENMs reflecting the different processes that may influence range shifts in response to past climate change is a key step in translating biologically-informed species-specific hypotheses into testable genetic predictions about a species' response to climate change.

Materials and methods

Sampling and genotyping

We collected leaf tissue from a total of 257 adult individuals from 46 localities across California (Fig. 4.1; Table C.2); 160 individuals were sampled by Ortego *et al.* (2015), and 97 additional individuals were collected to provide complete geographic

sampling for this study. Samples were genotyped at 13 polymorphic nuclear microsatellite markers developed for use in *Quercus* (Steinkellner *et al.* 1997; Kampfer *et al.* 1998; Durand *et al.* 2010). Full characterization of microsatellite loci and DNA extraction and microsatellite genotyping followed the procedures described by Ortego *et al.* (2014, 2015). Only individuals that were successfully genotyped at 10 or more of the 13 loci were retained for subsequent analyses (see Table C.2), resulting in a dataset with a total of 226 individuals from 44 localities.

Assignment of individuals into populations

Populations were initially classified geographically based on major mountain ranges. Individuals were also assigned to different genetic clusters on the basis of their microsatellite genotypes using the Bayesian analysis implemented in STRUCTURE v.2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007; Hubisz *et al.* 2009). The likelihood of different genetic clusters (K = 1 to 10) was estimated from 10 independent runs with one million MCMC cycles, following a burn-in step of 100,000 iterations. STRUCTURE was run both with and without a prior conditioned on either individual sampling localities or the mountain ranges of sampled localities (Hubisz *et al.* 2009). Genetic clusters generally corresponded well to mountain ranges, except for localities from the Sierra Nevada. Sierra Nevada localities were often assigned to two different genetic clusters – a group of northern and of southern localities (Fig. C.1). As a result of these analyses, we divided the 226 individuals from 44 localities into six populations, which included the Peninsular Ranges, Transverse Ranges, Southern Sierra Nevada, Northern Sierra Nevada, Southern Coast Ranges, and Northern Coast Ranges and Klamath Mountains

(Fig 1, Table C.2). A Mojave Desert population was excluded from all further analyses due to small sample size (*n* = 6).

Translating hypotheses into ecological niche models

Ecological niche models (ENMs) were used to generate habitat suitability maps for canyon live oak in the present and during the Last Glacial Maximum (LGM, 21.5 ka), using maximum entropy modelling with MAXENT v.3.3.3k (Phillips *et al.* 2004, 2006). Details of the general niche modelling procedure and data sources are given in Appendix C. To construct ENMs, specific environmental variables were selected as proxies for the biological mechanisms hypothesized to determine habitat suitability, as summarized below (see Table C.1 for complete details of all variables included in each model, and Appendix C for more detailed justification of variable selection):

(1) *GeneralENM*: This model does not invoke a specific mechanism determining geographic range, but focuses on the assumption that basic climatic variables (Table C.1; Hijmans *et al.* 2005) characterize habitat suitability according to a generic climatic ENM.

(2) *Microsite*: This model focuses on the assumption that habitat suitability may be limited by the availability of specific microsites such as canyons, steep slopes, and mountain ridges where canyon live oak could have a competitive advantage over other tree species (Thornburgh 1990). We assume that the four topographic variables that are included in this model (elevation, slope, aspect, and terrain roughness index; Amante & Eakins 2009; Hijmans *et al.* 2015a; Title & Bemmels 2018) have not substantially changed within the CFP since the LGM, except for exposed continental shelf due to

lower sea levels and increased extent of glaciation during the LGM (see Appendix C).

(3) *Multidimension*: This model assumes that a combination of basic climate variables, microsite, and additional climate variables putatively more closely related to ecological processes (Table C.1; Wang *et al.* 2006, 2012; Golicher 2012; Metzger *et al.* 2013; Title & Bemmels 2018) determines habitat suitability. These variables include all variables from the *GeneralENM* and *Microsite* models (but excluding elevation), as well as additional ecologically-relevant variables summarizing evapotranspiration, thermicity, aridity, growing degree days, and length of the growing season (Table C.1). Note that elevation was excluded because the relationship between elevation and climate under current conditions is very different from the relationship that existed during the LGM (Ritter & Hatoff 1975).

(4) *GrowCold:* This model focuses on a possible trade-off between growth rate and cold tolerance that may constrain suitable habitat of canyon live oak. The model is constructed from variables hypothesized to reflect the level of abiotic stress and selective pressure experienced by the species and its fitness relative to competitors in relation to this trade-off (Table C.1). We include variables related to cold-induced stress (e.g., mean temperature of the coldest quarter) as well as ameliorating variables indicating opportunity for growth during non-stressful conditions (e.g., growing degree days ≥ 5ºC).

(5) *GrowDrought*: This model focuses on a possible trade-off between growth and drought tolerance that may constrain suitable habitat of canyon live oak. As in the *GrowCold* model, chosen variables are hypothesized to reflect the level of abiotic stress experienced by the species and potential impacts on its fitness relative to competitors in

relation to this trade-off (see Table C.1); both stressor and ameliorating variables were included (as discussed above).

(6) *LocalAdaptation*: As in the *Multidimension* model, all available climatic and topographic variables (except elevation) are used to construct the ENM for this model, but with the difference that populations within each region are hypothesized to be strongly locally adapted. As such, habitat suitability in this model is predicted by unique climatic and topographic variables for each region separately, rather than the species as a whole (see also Gray & Hamann 2013). Given that genetic expectations are generated for the entire species range (as detailed below), regional habitat-suitability maps were standardized and combined into a single map (i.e., the habitat-suitability value of each grid cell in the combined map was set equal to the highest habitat suitability for the corresponding grid cell in any of the individual regional maps). Regions of local adaptation were delimited using Commission for Environmental Cooperation North American Level III Ecoregions (CEC 1997), retaining only ecoregions with at least 25 occurrence records. A total of six ecoregions met this criterion: California Coastal Sage, Chaparral, and Oak Woodlands; Coast Range; Klamath Mountains; Mojave Basin and Range; Sierra Nevada; and Southern and Baja California Pine-Oak Mountains. Each ecoregion comprised an average of 231 occurrence records (range: 47 to 401). The ecoregion-based population definitions described here were used only for the purpose of constructing ENMs in the *LocalAdaptation* model. Note also that such localized effects of ecoregion-specific habitat suitabilities were only investigated with respect to the same bioclimatic variables as in the *Multidimension* model (and not with respect to the subsets of bioclimatic variables featured in each of the other four models) because of

computational limitations.

Genetic predictions of each model

The integrative distributional, demographic and coalescent (iDDC) approach (He *et al.* 2013) was used to generate genetic predictions under each model (Fig. 4.2). For each separate model, (i) relative habitat suitabilities were extracted from the spatially explicit distributional model provided by the ENM, and were then rescaled to inform carrying capacities and migration rates of (ii) a demographic expansion across the landscape. For each of the six models tested, demographic simulations were conducted on landscapes representing three consecutive time periods, with corresponding shifts in habitat-suitability in response to changes in climate since the Last Glacial Maximum (LGM) for each time period. Specifically, maps for the present time period and for the LGM were generated directly from projections of the ENMs; a map representing intermediate conditions was also generated, in which the value of each grid cell corresponds to the mean value of that grid cell in the present and LGM maps. Parameters from the spatially explicit demographic model were then used to (iii) generate genetic predictions under a spatially explicit coalescent simulation. Finally, datasets simulated under the iDDC models were compared with the empirical data using an Approximate Bayesian Computation (ABC) framework for model selection and parameter estimation (Beaumont *et al.* 2002).

Demographic simulations were conducted in SPLATCHE2 (Ray *et al.* 2010) and were initiated at 21.5 ka from hypothesized ancestral source populations for each model. Ancestral source populations were defined as all grid cells of the LGM map with habitat

suitability greater than the median habitat suitability of all grid cells of the current climate map containing an occurrence record (Brown & Knowles 2012). This threshold averaged 0.57 among models (range: 0.52 to 0.59). Note that relative LGM habitat suitability was obtained from each model directly as output of the ENM produced in MAXENT (on a scale from 0 to 1). Next, habitat-suitability values for all maps across all time periods were categorized into 20 bins of equal magnitude, and maps were then used to perform the spatially explicit demographic simulations. In the demographic simulations, population carrying capacities and migration rates of each grid cell were rescaled proportionally according to habitat-suitability bins (with carrying capacity and migration rate ranging from zero to the maximum value of these parameters in a given simulation, as sampled from the prior distribution; see below). Note that because a single map is required by SPLATCHE2, custom *Python* scripts were used to convert the three maps of 20 bins each (39 bins for the intermediate map to account for intermediate values averaged between two bins; see above) into a single map with a theoretical maximum of 20^2x39 categories, with each category representing a unique combination of habitat-suitability bins across the three time periods. This makes it possible to model a dynamic landscape where habitat suitabilities change over time. Habitat-suitability bins representing each of the three temporal periods (LGM, intermediate, current) were consecutively applied for one third of the total number of generations each. Given that reproductive maturity in canyon live oak occurs after 15-20 years but individuals may live up to 300 years (Thornburgh 1990), average generation time was assumed to be 50 years, resulting in 430 generations from the LGM to present.

Following each time-forward demographic simulation, a time-backward coalescent genetic simulation was performed, in which the ancestry of an allele was traced back from the present into ancestral source populations. Before the the onset of population expansion from suitable areas at 21.5 ka modelled by the ENMs (see Fig. 4.3), alleles coalesced in a single large ancestral population (a maximum of 10^7 generations used in the simulations provided ample time for coalescence).

Individuals in simulated datasets were sampled from the same grid cells corresponding to the geographic locations from which the empirical data were sampled, and genetic data for these individuals were simulated along the coalescent genealogies at each locus using a strict stepwise microsatellite mutational model assuming no indels of more than one repeat unit, no recombination, and a maximum number of alleles equal to the number of repeat units separating the largest and smallest allele for each locus in the empirical data.

Model selection and parameter estimation using ABC

For the empirical data (Table C.3) and each simulated genetic dataset, 24 summary statistics were calculated (mean, total, and population heterozygosity, *H*; total and population pairwise population differentiation, F_{ST}) using ARLEQUIN v.3.5 (Excoffier & Lischer 2010). Although the number of alleles, *K*, has previously been used as a summary statistic (He *et al.* 2013), it was not used here because *K* was difficult to fit to empirical data in simulations across all models (i.e., all models had a consistent tendency to generate values of *K* substantially lower than in the empirical data; see Table C.4). We were thus concerned that the distance threshold between empirical and

simulated datasets would need to be very large in order to retain a sufficient number of simulations for parameter estimation, which may have reduced the precision of parameter estimates (Beaumont *et al.* 2002). To check whether excluding *K* would have a major impact on model selection, we conducted simulations to validate our model selection procedure (validation methods described below) with and without *K*, and found that including *K* had very little impact on our ability to distinguish among models (results not shown). We also note that our models are highly capable of producing datasets with properties that match the empirical data with respect to the 24 summary statistics used here (Table 4.1; Table C.4).

Rather than estimating parameter posterior distributions directly from summary statistics, partial least squares (PLS) components were calculated from summary statistics in order to reduce the number of summary statistics and account for correlations among them (Boulesteix & Strimmer 2006) using the *transformer* tool in ABCTOOLBOX with boxcox transformation for the pooled first 10,000 runs of each model (following He *et al.* 2013). In order to determine the optimal number of PLS components to retain, root mean squared error (RMSE) plots were examined and five PLS components were retained for calculating the distance between simulations and the empirical observations, because RMSE of the four parameters in our models does not decrease substantially with additional PLS components (results not shown).

Approximate Bayesian Computation (ABC) was used to estimate parameters and select among our six models using the wrapper program ABCTOOLBOX (Wegmann *et al.* 2010) on a high-performance computing cluster (Advanced Research Computing at the University of Michigan). One million datasets were simulated for each model across a

broad range of parameter values (i.e., maximum carrying capacity, K_{max}; migration rate, *m*; ancestral effective population size before population expansion, N_{anc}; and microsatellite mutation rate, μ) under a uniform prior on the base 10 logarithm of each parameter. The priors for parameter values were the same among models (i.e., $log(K_{max})$, 2.7 to 4.0; $log(N_{anc})$, 3.0 to 6.0; $log(\mu)$, -6.0 to -2.0; and *log(m)*, -3.0 to -1.7; Fig. 4.4), with the exception of the *GrowCold* model, for which higher values of *log*(*m)* were used (*log(m)*, -2.6 to -1.3) to ensure colonization of interior areas (see below). Note that the *GrowCold* model was the only model for which exclusively coastal ancestral source populations were inferred (Fig. C.2). Because the same range of parameter values was used in all models, this different prior in the *GrowCold* model is unlikely to have biased model selection given that the density of simulations for the given range of parameter space was the same in all models.

In all models, priors on migration rate were carefully considered in order to reflect (a) biologically realistic values of migration rate, and (b) values that would result in colonization of the landscape within the time spanning the LGM to the present. For example, true migration rates of our species are not known, but the prior $-3.0 \leq log(m) \leq -1.7$ covers potentially high values of migration rates at the spatial and temporal scale of our simulations (5-arcminute or ~9km x 9km grid cells; 50 years per generation) and we tested a variety of migration rates (and carrying capacities) in initial simulations to identify a range of migration rates that would result in colonization of the landscape within the time spanning the LGM to the present. Specifically, we identified a minimum value of *log(m)* for which complete landscape colonization was achieved (i.e., lower

values were not included in the prior for *log(m)* because the landscape would not be completely colonized, which could bias model selection). Likewise, we did not apply exceptionally high *log(m)* values because such values resulted in such rapid colonization that the differences among models in terms of their colonization patterns would be lost.

For each model, 5,000 simulations (0.5% of the total number of simulations per model) that most closely matched those of the empirical data were retained (He *et al.* 2013) and used to generate posterior distributions of parameters, using ABC-GLM (general linear model) adjustment (Leuenberger & Wegmann 2010). Bayes factors were approximated in order to assess relative support for the most strongly supported model compared to each other model; the approximate Bayes factor in favour of model X over model Y is calculated as the marginal density of model X divided by the marginal density of model Y (Leuenberger & Wegmann 2010).

Validation of model choice and parameter estimates

To determine whether the alternative models can be accurately distinguished with ABC given the data, we simulated 100 pseudo-observed datasets (PODs) under each model and analyzed them using our ABC procedure for model choice, using a subset of total simulations (100,000 per model) for computational efficiency. For each model, we calculated the proportion of the PODs for which the true model was either correctly or incorrectly identified. For PODs for which the true model was correctly chosen, the strength of support for the true model was calculated as the mean logarithm of the Bayes factor comparing the true model to the model with the second-highest marginal density. This represents how strongly the true model is identified to the exclusion of all

other models. When an incorrect model was chosen, the strength of support for the incorrect model was calculated as the mean logarithm of the Bayes factor comparing the incorrectly chosen model to the true model used to generate the POD. This value determines how strongly the incorrect model is favoured over the true model. Lastly, to assess the ability of each model to generate the empirical data, Wegmann's *et al.* (2010) *p*-value was calculated from 5,000 retained simulations. This *p*-value is the proportion of simulated datasets with a smaller or equal likelihood than the empirical data under the ABC-GLM (Wegmann *et al.* 2010).

To assess the accuracy of parameters estimated with ABC, we calculated the posterior quantiles of true parameter values from 1,000 PODs for the models with highest support. A Kolmogorov-Smirnov test was used to test these quantiles against a uniform distribution. Deviation from a uniform distribution indicates bias in parameter estimation (Cook *et al.* 2006; Wegmann *et al.* 2010).

To determine whether there are specific summary statistics that are easier or more difficult to fit to the empirical data in specific models, we generated a distribution of the simulated values of each summary statistic from 100,000 simulations per model (with simulation parameters drawn from the prior). We then calculated the percentile corresponding to the empirical value of each summary statistic within its simulated distribution, and calculated the distance between this percentile and the median (i.e., 50^{th} percentile) of the simulated distribution.

Results

Multiple disjunct putative ancestral source populations based on habitat suitability during the LGM were estimated under each of the six models (Fig. 4.3; Fig. C.2). These sources included locations in both coastal and inland mountain ranges, with the exception of exclusively coastal ancestral source populations estimated for the *GrowCold* model. Predicted habitat suitability during the LGM and intermediate time periods differed substantially among the six models, with the exception of the *GeneralENM* and *GrowDrought* models, which had very similar predictions for these time periods. In contrast, the current distribution of predicted suitable habitat was very similar for all models, except that the *Microsite* model also predicted large areas outside of the species' current range to contain suitable habitat (Fig. C.3).

With respect to the relative probabilities of the six models, two models – the *GeneralENM* model and the *GrowDrought* model – had the highest support (highest marginal density; Table 4.1). However, the Bayes factor comparing these two models was less than three, suggesting that there is not a statistically significant difference in the support for one model over the other (Kass & Raftery 1995). In other words, the *GeneralENM* and *GrowDrought* models are approximately equally well supported, in contrast to the much lower support for all the other models (Table 4.1). These two most probable models are also highly capable of generating simulated data comparable with the empirical data (see *p*-values, Table 4.1), despite uncertainty in parameter estimates (Fig. 4.4). Even with fairly broad posterior distributions for some parameter estimates (Fig. 4.4), the data contain information relevant to estimating the parameters (i.e., the

posterior distribution differs from the prior), and there is evidence of increased accuracy of parameter estimates following GLM (general linear model) adjustment (Fig. 4.4). There is little evidence of bias in most parameter estimates (Fig. 4.5), except for slight deviations from uniformity detected from the quantiles of the mutation rate (μ) parameter for the *GeneralENM* and possibly the *GrowDrought* models (p = 0.0243 and 0.0503, respectively), and of the ancestral population size (N_{anc}) parameter for the *GrowDrought* model (p = 0.0082). A slight tendency to potentially overestimate each of these parameter values was detected (Fig. 4.5).

Validation of model selection using pseudo-observed datasets (PODs) showed that for most models, the true model is correctly identified the majority of the time (Table 4.2a) and average relative support for the true model is strong to very strong (Table 4.2b; Kass & Raftery 1995). Selection of an incorrect model with strong relative support is extremely uncommon. In the rare cases when an incorrect model is inferred, average relative support for the incorrectly chosen model compared to the true model is typically very low (Table 4.2c), indicating that even if an incorrect model is identified as most likely, support is not strong enough to decisively exclude the true model from consideration. In contrast, for the *GeneralENM* and *GrowDrought* models there is limited ability to discern under which of these two models the PODs were simulated (Table 4.2). This is not surprising, given the similar relative support for these models in the empirical data (Table 4.1). Nonetheless, the *GeneralENM* and *GrowDrought* models are extremely unlikely to be confused with any of the other four models (Table 4.2).

Most models generated values of mean and total heterozygosity in agreement with empirical data, but simulated values of overall F_{ST} were typically higher than those of the empirical data in the *Multidimension*, *GrowCold*, and *LocalAdaptation* models (Table C.4). These models also tended to produce certain population-specific simulated heterozygosity values that were lower than in the empirical data, and simulated pairwise F_{ST} values that were higher than in the empirical data. In contrast, the *Microsite* model tended to produce simulated pairwise F_{ST} values that were substantially lower than in the empirical data for comparisons involving the Northern Sierra Nevada population (and to a lesser extent, the Northern Coast Ranges and Klamath Mountains population). Simulated pairwise F_{ST} values involving the Northern Sierra Nevada population also tended to be lower than empirical values in the two most-supported models (*GeneralENM* and *GrowDrought*), although most other summary statistics in these models were similar to the empirical data.

Discussion

Considering the biologically informed hypotheses we focus upon in our study, our goal was to consider whether we could distinguish among possible processes that might determine habitat suitability for canyon live oak and consequently, how the species distribution has shifted in response to changing climatic conditions. Differences in relative support among the models (Table 4.1) not only demonstrate differences in how influential these processes have likely been, but also how drought in particular may mediate the response to climate change in canyon live oak. Specifically, strong relative

support based on ABC model selection for two statistically indistinguishable models (Table 4.1) suggests that either climatic variables predictive of the species distribution that are related to drought stress (*GeneralENM* model), or a physiological trade-off between growth rate and summer drought tolerance (*GrowDrought* model), or both (see Table C.1), are primary determinants of habitat suitability. More generally, this shared component of the two most highly supported models highlights the potential importance of drought in driving historical range shifts in a temperate tree from the predominately Mediterranean climate of the California Floristic Province (CFP), a region characterized by summer drought. Below, we discuss how our work contributes to an expanding literature about the factors that limit species distributions based on work from other disciplines, and compare and contrast our results with knowledge of factors important to other tree species from less seasonally dry regions of the temperate zone. We also discuss the implications of our work for evaluating support for alternative hypotheses (e.g., cold tolerance, microsite variation, and local adaptation) using explicit predictions for patterns of genetic variation, and the general challenges of our approach and the limitations of such inferences (see also Papadopoulou & Knowles 2016; Massatti & Knowles 2016).

Drought tolerance as a determinant of distributional shifts and genetic structure

In the Mediterranean climate of the CFP, summer is the driest season (Hijmans *et al.* 2005), and plants must tolerate or avoid summer drought stress. As such, summer drought is likely an important environmental condition determining relative habitat suitability for plants, either directly through abiotic stress or indirectly through effects on

relative fitness in relation to competitors. The high support for the *GeneralENM* and *GrowDrought* models demonstrates that summer drought may not only be a key determinant of habitat suitability, but it may also drive demographic responses to climate change that ultimately impact population genetic structure of canyon live oak. In both of these models, the climatic variables making the largest contribution to the ENMs are strongly related to summer drought stress, and to the ability of a plant to tolerate or avoid this stress (see Table C.1). The *GeneralENM* model uses a generic ENM in which drought was not explicitly modelled and in which other climatic variables unrelated to drought were considered, but the four climatic variables making the greatest contribution to the ENM reflect precipitation during the summer and winter, and precipitation and temperature seasonality. As such, they represent the degree to which summers are hot and dry, and winters are cool and wet. Summer conditions likely directly reflect drought stress, whereas these winter conditions are hypothesized to reflect soil moisture availability during early spring, which may be the period of maximum growth for trees from Mediterranean environments prior to the onset of summer drought (Montserrat-Martí *et al.* 2009; Pinto *et al.* 2011). In comparison, the *GrowDrought* model features an ENM using climatic variables explicitly selected to reflect a possible trade-off between growth rate and summer drought tolerance. The climatic variables contributing most strongly to this ENM (Table C.1) are preciptiation of the driest quarter and Emberger's pluviothermic quotient, which captures annual climatic dryness as experienced by plants with particular relevance to Mediterranean climates (Daget 1977).

The shared component of the two most supported models (i.e., drought stress) complements knowledge from other fields suggesting that drought limits geographic

distributions and drives adaptation of some temperate tree species, especially those from Mediterranean climates. For example, across 1,577 European plant species, summer drought determines latitudinal range limits in 22% of species (Normand *et al.* 2009). Although drought stress does not generally limit the ranges of most of these plant taxa, its role in structuring plant distributions is especially common in the Mediterranean biomes of southern Europe, and in central Europe at the transition between Mediterranean and less seasonally dry biomes (Normand *et al.* 2009). Plant taxa with distributions limited by drought include trees specifically; for example, among European trees drought stress has been implicated in determining dry-edge range limits of *Fagus sylvatica* (Rasztovits *et al.* 2014)*, Pinus nigra* (Linares & Tíscar 2011), and *Quercus robur* (Urli *et al.* 2014). Drought mortality was also found to be regionally important (e.g., in the Great Plains and at high-elevation sites) in limiting the ranges of at least 12 North America tree species (out of 17 studied; Morin *et al.* 2007).

In addition to setting range limits, drought tolerance is a trait of adaptive significance among populations of some tree species. For example, a trade-off between growth rate and drought tolerance has been documented among populations of Douglas-fir (*Pseudotsuga menziesii*; White 1987), and is hypothesized to underlie several adaptive differences in functional traits such as growth rate, growth phenology, growth pattern (i.e., determinate versus indeterminate), and root to shoot ratio (White 1987; Joly *et al.* 1989; Kaya *et al.* 1994). Putatively adaptive clines in phenotypic traits along precipitation gradients have also been observed in height growth and timing of bud flush in several western North American tree species (Aitken & Bemmels 2016). Although weak or non-adaptive clines along precipitation gradients may emerge when

strong adaptive clines along temperature gradients exist (see below) if precipitation and temperature are geographically correlated, it is noteworthy that clines associated with precipitation are substantially stronger than those associated with temperature gradients in several species (e.g., *Picea pungens*, *Pinus attenuata, Pinus monticola, Populus trichocrapha*, and possibly *Pseudotsuga menziesii* and *Quercus garryana;* Aitken & Bemmels 2016).

While our procedure identified seasonal drought tolerance as an ecological factor that has likely shaped the response of canyon live oak to climate change and left signatures in patterns of genetic variation, our approach considers only the historically most important factors structuring genetic variation since the LGM. We tested only dynamic models (i.e., models where habitat suitability changes over time) because we have strong reason to believe that accounting for demographic history will be required to fully explain genetic structure in this study system. In particular, canyon live oak has a long generation time (we assumed only 430 generations since the LGM) and limited seed-dispersal ability by acorns (Thornburgh 1990), such that genetic signatures of past range shifts in response to climate change are unlikely to have been completely erased by contemporary patterns of gene flow (see Ortego *et al.* 2015). It is possible that ecological factors other than drought tolerance may be more important in driving contemporary processes affecting gene flow among populations, but testing these processes under contemporary climatic conditions was beyond the scope of our models.

Lack of support for competing explanations for genetic structure

Patterns of genetic variation in canyon live oak did not identify several commonly invoked competing factors (including cold tolerance, microsite variation, and local adaptation) as primary determinants of shifting geographic distributions in the face of climate change (Table 4.1). It is possible that this finding reflects differences in which environmental factors (e.g., temperature versus precipitation) are important for determining distributions and driving adaptation among different temperate tree species (see Howe *et al.* 2003; Normand *et al.* 2009; Aitken & Bemmels 2016). Yet, the lack of support for some of the models is nonetheless somewhat surprising, especially given that these models consider alternative ecological processes that are generally recognized to be broadly relevant across many taxa. For example, temperature is widely believed to limit cold-edge distributions in temperate trees through various physiological mechanisms (Sakai & Weiser 1973; Pigott & Huntley 1981; Morin *et al.* 2007; Normand *et al.* 2009; Mellert *et al.* 2011; Kollas *et al.* 2014; Lenz *et al.* 2014; Siefert *et al.* 2015). Furthermore, numerous tree species exhibit a trade-off between growth rate and cold tolerance at the population level, with more cold-tolerant populations exhibiting slower growth rate, earlier bud set, and (less frequently) shifts in phenology of bud flush (Howe *et al.* 2003; Savolainen *et al.* 2007; Alberto *et al.* 2013; Aitken & Bemmels 2016). This trade-off may also determine range limits at the species level, with warm-edge distributions limited by competition from faster-growing species and cold-edge distributions limited by low temperatures (Loehle 1998; but see also Morin *et al.* 2007). However, it is possible that the adaptive and ecological significance of drought in temperate trees has been understudied relative to that of cold temperatures because of

biases in the choice of taxa studied. For example, most of the taxa studied are from temperate deciduous and conifer forests (Howe *et al.* 2003; Morin *et al.* 2007; Savolainen *et al.* 2007; Normand *et al.* 2009; Aitken & Bemmels 2016), whereas less attention has been paid to taxa from more seasonally dry regions of the temperate zone such as Mediterranean climates (e.g., Morin *et al.* 2007; Aitken & Bemmels 2016). In temperate broadleaf forests in particular, seasonal summer drought is uncommon and is unlikely to be a major source of abiotic stress (Morin *et al.* 2007). The response to seasonal drought may also differ across biomes (Allen *et al.* 2010; Vicente-Serrano *et al.* 2013). In other words, temperate trees from Mediterranean climates may simply be subject to fundamentally different primary ecological and adaptive constraints than those from wetter, colder, and less seasonally dry climates within the temperate zone.

Lack of support for models reflecting alternative processes that could possibly affect habitat suitability (Table 4.1), especially those associated with local conditions, does not necessarily mean these processes do not play a role in response to climate change, but perhaps that their effects are minor at the regional scale studied here. In particular, lack of support for models incorporating local-scale factors (i.e., *Microsite* and *LocalAdaptation* models) suggests that responses to Pleistocene glacial cycles were primarily driven by climatic factors affecting habitat suitability over broad spatial scales. Consequently, although under current climatic conditions canyon live oak is distributed primarily in mountainous areas (Little 1971; Thornburgh 1990) and terrain roughness index (TRI) is one of the variables most highly predictive of current habitat suitability (*Multidimension* model; Table C.1), TRI covaries with other predictor variables and may not itself be the driver of the species distribution. This interpretation also seems likely

considering that both the *GeneralENM* and *GrowDrought* models receive high support, even though under these models the species is predicted to have been distributed in areas of low topographic complexity in the past (e.g., in California's northern Central Valley; Fig. 4.2). Our results are therefore consistent with the hypothesis that canyon live oak, despite its abundance in sheltered canyons and on steep, rocky slopes, was capable of making shifts to topographically novel habitats such as the northern Central Valley during the LGM (Fig. 4.2), which may reflect the ability of this species to grow on a wide variety of soil types and in multiple community assemblages (Thornburgh 1990).

Likewise, lack of support for the *LocalAdaptation* model (Table C.1) suggests that the response of canyon live oak to climate change is not localized. Given that populations of many temperate and boreal tree species are locally adapted to climate (Savolainen *et al.* 2007; Alberto *et al.* 2013; Aitken & Bemmels 2016), local adaptation has been hypothesized to have been an important factor affecting Pleistocene range shifts in trees (Davis *et al.* 2001), and is often considered to be a key factor that will determine the effects of future climate change on the potential geographic distributions of tree populations (e.g., Pearman *et al.* 2010; Benito Garzón *et al.* 2011; Gray & Hamann 2013; Valladares *et al.* 2014; Gotelli & Stanton-Geddes 2015; Hällfors *et al.* in press) and of adaptive genomic variation (Fitzpatrick & Keller 2015). In some cases, local adaptation may also leave a signature in patterns of neutral genetic variation (through its mediating effects on patterns of gene flow; e.g., Lee & Mitchell-Olds 2011). While the *LocalAdaptation* model was not the most probable model identified in our study, we note that it did receive very strong relative support compared to the *Multidimension* model (Bayes factors = 234; Table 4.1) in which exactly the same

environmental variables were used to generate species-wide predictions of habitat suitability (Table C.1). This suggests that further investigation into localized effects of other predictors of habitat suitability may indeed be worthwhile, especially with regards to the highly supported models identified here (Table 4.1).

In addition to identifying the most probable models and determining that these models are indeed capable of generating the data (Table 4.1), we also compared the simulated summary statistics under each model with the empirical data (Table C.4) to examine what made a model a poor fit. This revealed that the empirical data did not match the low heterozygosity and high pairwise F_{ST} values for certain populations predicted by the *Multidimension*, *GrowCold*, and *LocalAdaptation* models. This lack of fit suggests the generally small, disjunct ancestral source populations and spatially restricted LGM habitat suitability predicted by these models (Fig. 3; Fig. C.2) is not well supported by the data. In contrast, in the *Microsite* model, relatively low pairwise F_{ST} values in the simulated data compared with the empirical data, especially for comparisons involving the two northernmost populations, suggest that large areas of high habitat suitability predicted since the LGM in the northern portion of this species' range in this model (Fig. 3; Fig. C.2) are not well supported. A qualitatively similar pattern (but with a smaller observed differences between simulated and empirical data) was observed in both of the most well supported models (*GeneralENM, GrowDrought*), suggesting even the most probable models do not capture the complex history of the northern Sierra Nevada populations (Table C.4). Exploring whether the northern Sierra Nevada historically contained smaller, more demographically isolated populations than

suggested by our current models (Fig. 3; Fig. C.2) could be a hypothesis to test in future studies.

The California Floristic Province during the late Pleistocene

The California Floristic Province (CFP) is a plant biodiversity hotspot (Myers *et al.* 2000; Lancaster & Kay 2013) characterized by high topographic, climatic, and ecological heterogeneity. The maintenance of high biodiversity within the CFP has been hypothesized in part to reflect long-term regional-scale climatic stability that kept extinction rates low even through periods of intense global climatic change (Lancaster & Kay 2013). LGM habitat-suitability predictions for canyon live oak from the two most supported models (in fact, from all models except the *GrowCold* model; Fig. 3 and Fig. C.2) are in agreement with this hypothesis. Both the *GeneralENM* and *GrowDrought* models predict high habitat suitability in some portion of every major mountain range in the CFP currently inhabited by the species, with the exception of the Mojave Desert and the northernmost portion of the range in the Klamath Mountains. The possible existence of these areas of high habitat suitability since the LGM throughout geographically disparate regions of the CFP suggests that canyon live oak is unlikely to have gone locally extinct in most regions of its current geographic distribution, and that only modest range shifts were needed in most regions in order for the species to track changes in suitable habitat.

This scenario contrasts with the major continental-scale changes in climate in response to glacial cycles that characterized other temperate regions such as eastern North America and Europe (Taberlet *et al.* 1998; Soltis *et al.* 2006; Gavin *et al.* 2014). At

smaller spatial scales, pronounced effects of climate change did occur within the CFP. For example, alpine glaciers in the Sierra Nevada expanded in size (Gillespie *et al.* 2004), and pollen records indicate local changes in species abundance and shifts in the distribution of vegetation types to lower elevations (Roosma 1958; Cole 1983; Litwin *et al.* 1999; Heusser *et al.* 2015; McGann 2015), by as much as 600 to 750 m in the western Sierra Nevada (Ritter & Hatoff 1975). Nevertheless, at a regional scale, steep elevational gradients and the moderating effects of orographic precipitation may have provided a 'climatic buffering' effect preventing extreme regional-scale fluctuations in climate (Lancaster & Kay 2013). As a result, species from the CFP were likely able to track geographic shifts in suitable climate by migrating over relatively short distances (Davis *et al.* 2008; Lancaster & Kay 2013). For canyon live oak in particular, large regions of moderately stable habitat during both glacial and interglacial periods may have served as reservoirs of genetic diversity and driven patterns of genetic connectivity and admixture among populations (Ortego *et al.* 2015).

Utility of species-specific genetic predictions for testing hypotheses

Because different processes can produce similar patterns of genetic variation, phylogeographic studies rely upon model-based inferences in which expectations for patterns of genetic variation under particular processes are specified. However, the approach applied here differs from other model-based inferences (see Knowles 2009; Hickerson et al. 2010). Specifically, biologically informed hypotheses about factors that may determine how taxa respond to climate change are explicitly modelled here by considering their predicted effects on the movement of species across a landscape. As

such, our work adds to the growing number of studies that use spatially explicit models to capture how population dynamics (e.g., changes in population size and dispersal probabilities) impact patterns of genetic variation (e.g., Neuenschwander *et al.* 2008; He *et al.* 2013; Massatti & Knowles 2014).

A key aspect of our approach – the generation of species-specific predictions for patterns of genetic variation given different factors that might determine the habitat suitability of a species – is a novel application that differs fundamentally from other approaches for using patterns of genetic variation to study the effects of climate change on geographic distributions of taxa. In particular, our approach considers that the best characterization of habitat suitability for taxa may not be one based on a typical ENM analysis of bioclimatic variables, as generally assumed in studies that rely on measures of habitat suitability to test hypotheses about the effects of climate change using genetic data (e.g., Knowles 2009; Lanier *et al.* 2015) There are nonetheless caveats with our approach that should be considered, especially regarding the use of different environmental variables as proxies for competing biological processes hypothesized to determine habitat suitability. Specifically, we do not have an explicit means of determining if these environmental variables truly capture the processes they are intended to represent. This limitation is not unique to our approach. Instead, it is a broader conceptual concern with any approach in which predictions from correlative ENMs are used because it is not possible to ascertain whether environmental variables determine distributions directly, or are correlated with some other variable that is actually the source of causation but was not incorporated into the ENM (Austin 2002). While mechanistic ENMs that directly model functional traits of species could provide

information to avoid misleading inferences about causal variables (Kearney & Porter 2009), the detailed information required for such functional modelling is frequently not available, which contrasts with the broad applicability of the approach applied here.

There are additional aspects of our study that should be kept in mind, some of which are not specific to our study, but are general issues with model-based inference. Our study provides a robust evaluation of competing models for observed patterns of genetic variation, as we evaluate not only the relative probabilities of models, but also conduct validations of our approach (i.e., we determine that the models are capable of generating the data, and that there is sufficient power to accurately distinguish among models given the quantity of genetic data collected in our study). As such, we can make strong statements about which of the different models best fit the data. However, we acknowledge there may of course be additional factors not considered here that might contribute to patterns of genetic variation, and therefore, our approach does not identify the optimal model (nor does any model-based approach). Recognizing the limits of the inference space is important for avoiding possible misinterpretations of model-based approaches, but it does not discount the insights gained with respect to the study goals. Instead, our work demonstrates that with thoughtful consideration of the factors that might determine habitat suitability (including not only climatic variables, but also potential trade-offs in functional traits that may impact a taxon's ability to tolerate physiological stresses or compete, as well as localized effects related to microsite variation and adaptive differences), such hypotheses can be translated into models for studying which factors mediate the effects of climate change on species distributions. Likewise, even though many assumptions are made in the procedures applied here (e.g., converting

measures of habitat suitability into population demographic parameters; for details see Brown and Knowles 2012), these assumptions are arguably not more problematic than many assumptions implicitly made in other model-based approaches (e.g., not considering the spatial mosaic of habitat suitabilities that impacts both local population sizes and migration probabilities, despite the clear effects of such heterogeneity on patterns of genetic variation; see Knowles & Alvarado-Serrano 2010). Lastly, spatially explicit models, despite some of their limitations discussed above (see also Massatti & Knowles 2016), provide a window into a diversity of questions that would continue to go unexplored without their application.

Conclusions

We compare the relative statistical support for six different models concerning distributional shifts in canyon live oak in response to climate change, each of which is motivated by a different hypothesis about the mechanistic factors that may determine habitat suitability. We obtain very strong relative statistical support for two models that share a common conceptual link to summer drought, and show through validation of the model-selection procedure that we can be highly confident in the fit of data under these models, as well as in our ability to accurately discriminate among the different models. We suggest that drought tolerance may not only be a critical factor determining habitat suitability and mediating distributional shifts in response to climate change since the LGM in canyon live oak, but its importance may be generalized to other plants. Specifically, by comparison with studies of other temperate trees that have emphasized other processes but where focal taxa have typically been from less seasonally dry

regions of the temperate zone, our work suggests that summer drought may play a key adaptive and ecologically important role in other trees from Mediterranean climates. Moreover, our approach demonstrates how different factors hypothesized to determine habitat suitability may be tested by using spatially explicit information from ENMs to generate specific patterns of genetic variation for testing biologically informed hypotheses about the effects of climate change on species distributions. As such, the models supported in our study are a general example of the type of biologically informed, species-specific hypotheses that contribute to our broader understanding of the importance of biotic factors in structuring genetic variation (reviewed in Papadopoulou & Knowles 2016).

Acknowledgements

The authors thank P.F. Gugger for assistance with sampling; Q. He, R. Massatti, and A. Papadopoulou for their help and guidance in performing the ABC analyses; Q. He for providing custom scripts for iDDC modelling; and three anonymous reviewers for their suggestions, which greatly improved the chapter. Funding was provided for fieldwork and microsatellite genotyping by an internal EBD 'Microproyectos' grant to J.O., financed by the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Program for Centres of Excellence in R+D+I (SEV-2012-0262); for a high-performance computing cluster allocation for simulations by a Collegiate Professor Honorarium at the University of Michigan (L.L.K.); for support to J.B.B. by an NSF GRFP fellowship (DEB: 1256260) and a University of Michigan Department of Ecology and

Evolutionary Biology Edwin H. Edwards Scholarship in Biology; and for support to J.O. by a Ramón y Cajal Fellowship (RYC-2013-12501). This research was also supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

Author contributions

All authors designed the research project. JO collected and genotyped samples and performed STRUCTURE analyses. JBB and POT implemented the iDDC modelling and ABC procedures, under guidance of LLK. JBB and LLK wrote the chapter, with suggestions from POT and JO.

Literature cited

Aitken SN, Bemmels JB (2016) Time to get moving: assisted gene flow of forest trees. *Evolutionary Applications*, **9**, 271–290.

Alberto FJ, Aitken SN, Alía R *et al.* (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Allen CD, Macalady AK, Chenchouni H *et al.* (2010) A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, **259**, 660–684.

Allié E, Pélissier R, Engel J *et al.* (2015) Pervasive local-scale tree-soil habitat association in a tropical forest community. *PLOS ONE*, **10**, 1–17.

Alvarado-Serrano DF, Knowles LL (2014) Ecological niche models in phylogeographic studies: applications, advances and precautions. *Molecular Ecology Resources*, **14**, 233–248.

Amante C, Eakins BW (2009) ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24. *National Geophysical Data Center, NOAA*.

Austin, MP (2002) Spatial predictions of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101-118.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Benito Garzón M, Alía R, Robson TM, Zavala MA (2011) Intra-specific variability and plasticity influence potential tree species distributions under climate change. *Global Ecology and Biogeography*, **20**, 766–778.

Boulesteix A-L, Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44.

Brown JL, Knowles LL (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of Pleistocene glaciation and recent climate change on the American Pika. *Molecular Ecology*, **21**, 3757–3775.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

CEC (1997) *Ecological regions of North America: toward a common perspective*. Commision for Environmental Cooperation, Montréal.

Cole K (1983) Late pleistocene vegetation of Kings Canyon, Sierra Nevada, California. *Quarternary Reserach*, **19**, 117–129.

Cook SR, Gelman A, Rubin DB (2006) Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, **15**, 675–692.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Daget P (1977) Le bioclimat méditerranéen: analyse des formes climatiques par le système d'Emberger. *Vegetatio*, **34**, 87–103.

Davis EB, Koo MS, Conroy C, Patton JL, Moritz C (2008) The California Hotspots Project: identifying regions of rapid diversification of mammals. *Molecular Ecology*, **17**, 120–138.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. *Science*, **292**, 673–679.

Durand J, Bodénès C, Chancerel E *et al.* (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics*, **11**, 570.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Fitzpatrick MC, Keller SR (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, **18**, 1–16.

Frei ES, Scheepens JF, Stöklin J (2012) Dispersal and microsite limitation of a rare alpine plant. *Plant Ecology*, **213**, 395–406.

Gavin DG, Fitzpatrick MC, Gugger PF *et al.* (2014) Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, **204**, 37–54.

Gent PR, Danabasoglu G, Donner LJ *et al.* (2011) The Community Climate System Model version 4. *Journal of Climate*, **24**, 4973–4991.

Gillespie A, Porter S, Atwater BF (2004) *Quaternary glaciations - extent and chronology. Part II: North America. Additional CD-ROM*. Elsevier, San Diego, CA.

Golicher D (2012) Implementing a bucket model using WorldClim layers. https://rpubs.com/dgolicher/2964. *RPubs*.

Gotelli NJ, Stanton-Geddes J (2015) Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, **42**, 1577–1585.

Gray LK, Hamann A (2013) Tracking suitable habitat for tree populations under climate change in western North America. *Climatic Change*, **117**, 289–303.

Hällfors MH, Liao J, Dzurisin JDK *et al.* (in press) Addressing potential local adaptation in species distribution models: implications for conservation under climate change. *Ecological Applications*.

Harrison SP, Bartlein PJ, Brewer S *et al.* (2014) Climate model benchmarking with

glacial and mid-Holocene climates. *Climate Dynamics*, **43**, 671–688.

He Q, Edwards DL, Knowles LL (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

Heusser LE, Kirby ME, Nichols JE (2015) Pollen-based evidence of extreme drought during the last Glacial (32.6-9.0 ka) in coastal southern California. *Quaternary Science Reviews*, **126**, 242–253.

Hickerson, MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, **54**, 291-301.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans RJ, van Etten J, Cheng J *et al.* (2015a) Package "raster". https://cran.r-project.org/web/packages/raster/index.html.

Hijmans RJ, Phillips S, Leathwick J, Elith J (2015b) Package "dismo". http://cran.r-project.org/web/packages/dismo/index.html.

Howe GT, Aitken SN, Neale DB *et al.* (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, **81**, 1247–1266.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

John R, Dalling JW, Harms KE *et al.* (2007) Soil nutrients influence spatial distributions of tropical tree species. *PNAS*, **104**, 864–869.

Joly RJ, Adams WT, Stafford SG (1989) Phenological and morphological responses of mesic and dry site sources of coastal Douglas-fir to water deficit. *Forest Science*, **35**, 987–1005.

Kampfer S, Lexer C, Glössl J, Steinkellner H (1998) Characterization of $(GA)_n$ microsatellite loci from *Quercus robur*. *Hereditas*, **129**, 183–186.

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kaya Z, Adams WT, Campbell RK (1994) Adaptive significance of intermittent shoot

growth in Douglas-fir seedlings. *Tree Physiology*, **14**, 1277–89.

Kearney, M, Porter W (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334-350.

Knowles LL (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Knowles LL, Alvarado-Serrano DF (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. *Molecular Ecology*, **19**, 3727–3745.

Kollas C, Körner C, Randin CF (2014) Spring frost and growing season length co-control the cold range limits of broad-leaved trees. *Journal of Biogeography*, **41**, 773–783.

Kroiss SJ, HilleRisLambers J (2015) Recruitment limitation of long-lived conifers: Implications for climate change responses. *Ecology*, **96**, 1286–1297.

Lancaster LT, Kay KM (2013) Origin and diversification of the California flora: re-examining classic hypotheses with molecular phylogenies. *Evolution*, **67**, 1041–1054.

Lanier HC, Massatti R, He Q, Olson LE, Knowles LL (2015) Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genomic variation in Collared Pikas (*Ochotona collaris*). *Molecular Ecology*, 3688–3705.

Lee C-R, Mitchell-Olds T (2011) Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Molecular Ecology*, **20**, 4631–4642.

Lenz A, Vitasse Y, Hoch G, Körner C (2014) Growth and carbon relations of temperate deciduous tree species at their upper elevation range limit. *Journal of Ecology*, **102**, 1537–1548.

Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.

Linares JC, Tíscar PA (2011) Buffered climate chang eeffects in a Mediterranean pine species: range limit implications from a tree-ring study. *Oecologia*, **167**, 847–859.

Little EL (1971) *Atlas of United States trees, volume 1. Conifers and important hardwoods. USDA Miscellaneous Publication 1146.* Washington, DC.

Litwin RJ, Smoot JP, Durika NJ, Smith GI (1999) Calibrating Late Quaternary terrestrial climate signals: radiometrically dated pollen evidence from the southern Sierra Nevada, USA. *Quaternary Science Reviews*, **18**, 1151–1171.

Loehle C (1998) Height growth rate tradeoffs northern determine and southern range limits for trees. *Journal of Biogeography*, **25**, 735–742.

Massatti R, Knowles LL (2016) Tests of species-specific predictions for genetic variation in a wet versus dry-adapted species show that microhabitat preference of alpine plants may mediate the effects of climate change. *Molecular Ecology, in press*.

Massatti R, Knowles LL (2014) Microhabitat differences impact phylogeographic concordance of codistributed species: genomic evidence in montane sedges (*Carex* L.) from the Rocky Mountains. *Evolution*, **68**, 2833–2846.

McGann M (2015) Late Quaternary pollen record from the central California continental margin. *Quaternary International*, **387**, 46–57.

Mellert KH, Fensterer V, Küchenhoff H *et al.* (2011) Hypothesis-driven species distribution models for tree species in the Bavarian Alps. *Journal of Vegetation Science*, **22**, 635–646.

Metzger MJ, Bunce RGH, Jongman RHG *et al.* (2013) A high-resolution bioclimate map of the world: A unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, **22**, 630–638.

Montserrat-Martí G, Camarero JJ, Palacio S *et al.* (2009) Summer-drought constrains the phenology and growth of two coexisting Mediterranean oaks with contrasting leaf habit: Implications for their persistence and reproduction. *Trees*, **23**, 787–799.

Morgan K, O'Loughlin SM, Chen B *et al.* (2011) Comparative phylogeography reveals a shared impact of pleistocene environmental change in shaping genetic diversity within nine *Anopheles* mosquito species across the Indo-Burma biodiversity hotspot. *Molecular Ecology*, **20**, 4533–4549.

Morin X, Augspurger C, Chuine I (2007) Process-based modeling of species' distributions: what limits temperate tree species' range boundaries? *Ecology*, **88**, 2280–2291.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Neuenschwander S, Largiadèr CR, Ray N *et al.* (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.

Normand S, Treier UA, Randin C *et al.* (2009) Importance of abiotic stress as a range-limit determinant for European plants: insights from species responses to climatic gradients. *Global Ecology and Biogeography*, **18**, 437–449.

Ortego J, Bonal R, Muñoz A, Aparicio JM (2014) Extensive pollen immigration and no evidence of disrupted mating patterns or reproduction in a highly fragmented holm oak stand. *Journal of Plant Ecology*, **7**, 384–395.

Ortego J, Gugger PF, Sork VL (2015) Climatically stable landscapes predict patterns of genetic structure and admixture in the Californian canyon live oak. *Journal of Biogeography*, **42**, 328–338.

Papadopoulou A, Knowles LL (2016) A paradigm shift in comparative phylogeography driven by trait-based hypotheses. *PNAS,* **113**, 8018-8024.

Pearman PB, D'Amen M, Graham CH, Thuiller W, Zimmermann NE (2010) Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography*, **33**, 990–1003.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips SJ, Dudík M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, 655–662.

Pigott C, Huntley JP (1981) Factors controlling the distribution of *Tilia cordata* at the northern limits of its geographical range. III. Nature and causes of seed sterility. *New Phytologist*, **87**, 817–839.

Pinto CA, Henriques MO, Figueiredo JP *et al.* (2011) Phenology and growth dynamics in Mediterranean evergreen oaks: effects of environmental conditions and water relations. *Forest Ecology and Management*, **262**, 500–508.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rasztovits E, Berki I, Mátyás C *et al.* (2014) The incorporation of extreme drought events improves models for beech persistence at its distribution limit. *Annals of Forest Science*, **71**, 201–210.

Ray N, Currat M, Foll M, Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics*, **26**, 2993–2994.

Ritter EW, Hatoff BW (1975) Late Pleistocene pollen and sediments: an analysis of a central California locality. *Texas Journal of Science*, **29**, 195–207.

Roosma A (1958) A climatic record from Searles Lake, California. *Science*, **128**, 716.

Sakai AA, Weiser CJ (1973) Freezing resistance of trees in North America with reference to tree regions. *Ecology*, **54**, 118–126.

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.

Sayre R, Comer P, Warner H, Cress J (2009) *A new map of standardized terrestrial ecosystems of the conterminous United States: US Geological Survey Professional Paper 1768*. Reston, VA.

Siefert A, Lesser MR, Fridley JD (2015) How do climate and dispersal traits limit ranges of tree species along latitudinal and elevational gradients? *Global Ecology and Biogeography*, **24**, 581–593.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Steinkellner H, Fluch S, Turetschek E *et al.* (1997) Identification and characterization of $(GA/CT)_n$-microsatellite loci from *Quercus petraea*. *Plant Molecular Biology*, **33**, 1093–1096.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Thornburgh DA (1990) Canyon live oak. In: *Silvics of North America: 1. Conifers; 2. Hardwoods. US Department of Agriculture Handbook 654.* (eds Burns RM, Honkala BH). Washington, DC.

Thornthwaite CW (1948) An approach toward a rational classification of climate. *Geographical Review*, **38**, 55–94.

Title, PO, Bemmels JB (2018) ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291-307.

Urli M, Lamy J-B, Sin F *et al.* (2014) The high vulnerability of *Quercus robur* to drought at its southern margin paves the way for *Quercus ilex*. *Plant Ecology*, **216**, 177–187.

Valladares F, Matesanz S, Guilhaumon F *et al.* (2014) The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. *Ecology Letters*, **17**, 1351–1364.

Vicente-Serrano SM, Gouveia C, Camarero JJ *et al.* (2013) Response of vegetation to

drought time-scales across global land biomes. *PNAS*, **110**, 52–57.

Wang T, Hamann A, Spittlehouse DL, Aitken SN (2006) Development of scale-free climate data for western Canada for use in resource management. *International Journal of Climatology*, **26**, 383–397.

Wang T, Hamann A, Spittlehouse DL, Murdock TQ (2012) ClimateWNA—high-resolution spatial climate data for western North America. *Journal of Applied Meteorology and Climatology*, **51**, 16–29.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

White TL (1987) Drought tolerance of southwestern Oregon Douglas-fir. *Forest Science*, **33**, 283–293.

Wilson MFJ, O'Connell B, Brown C, Guinan JC, Grehan AJ (2007) Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy*, **30**, 3–35.

Zhao L, Xia J, Xu C-Y *et al.* (2013) Evapotranspiration estimation methods in hydrological models. *Journal of Geographical Sciences*, **23**, 359–369.

Zomer RJ, Trabucco A, Bossio DA, Verchot L V. (2008) Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems & Environment*, **126**, 67–80.

Zomer RJ, Trabucco A, Van Straaten O, Bossio DA (2006) *Carbon, Land and Water: A Global Analysis of the Hydrologic Dimensions of Climate Change Mitigation through Afforestation/Reforestation. International Water Management Institute Research Report 101.* Colombo, Sri Lanka.

Table 4.1. Summary of model results from the ABC procedure for canyon live oak. Summary of models and relative support from the ABC procedure for each model. A higher marginal density corresponds to higher support for the model, while *p*-values close to 1.0 indicate that the model is able to reproduce data in agreement with the empirical data (Wegmann *et al.* 2010). Bayes factors represent the degree of relative support for the most highly supported model (*GeneralENM*) over the other models. Bayes factors greater than 20 indicate strong support, while those greater than 150 indicate very strong support (Kass and Rafterty 1995).

Model	Hypothesized factors mediating species response to climate change	Marginal density	Wegmann's *p*-value	Bayes factor
GeneralENM	basic climatic variables of a generic ecological niche model	2.35×10^{-2}	0.9900	-
Microsite	availability of topographic microsites	1.27×10^{-7}	0.0024	1.86×10^{5}
Multidimension	basic and ecologically-informed climate variables; microsites	8.20×10^{-9}	0.0038	2.87×10^{6}
GrowCold	trade-off between growth rate and cold tolerance	3.21×10^{-7}	0.0046	7.34×10^{4}
GrowDrought	trade-off between growth rate and drought tolerance	8.43×10^{-3}	0.9272	2.79
LocalAdaptation	unique factors in each locally adapted ecoregion	3.51×10^{-7}	0.0044	6.70×10^{4}

Table 4.2. Validation of the ABC procedure for canyon live oak. Validation of the ABC procedure for model selection using pseudo-observed datasets (PODs; see text for explanation). (a) Confusion matrix showing the ability of the ABC procedure to correctly identify the model used to generate the POD. Numbers in the table represent the percent of PODs (n = 100 for each model) determined by the ABC procedure to be most highly supported by each of the models. Bold numbers on the diagonal indicate that the true model was identified, while numbers off the diagonal indicate incorrect model identification. (b-c) Average level of support, measured as the mean logarithm of Bayes factors, $log_{10}(BF)$, for (b) the true model compared to the second-most-supported model, when the true model is chosen, and (c) the incorrectly chosen model compared to the true model, when an incorrect model is chosen. Values in (b) represent the strength with which the ABC procedure unambiguously supports the true model to the exclusion of all other models, when the true model is chosen. Values in (c) represent the average strength with which the ABC procedure incorrectly favours the chosen model over the true model, when an incorrect model is chosen. Asterisk (*): mean $log_{10}(BF) \geq 1.30$, indicating strong relative support for the chosen model; dagger (†): mean $log_{10}(BF) \geq 2.18$, indicating very strong support (Kass and Raferty 1995).

A)

True model	**Model selected by ABC procedure**					
	GeneralENM	*Microsite*	*Multidimension*	*GrowCold*	*GrowDrought*	*LocalAdaptation*
GeneralENM	**52**	7	7	6	19	9
Microsite	6	**80**	4	6	1	3
Multidimension	0	1	**74**	23	1	1
GrowCold	0	1	25	**74**	0	0
GrowDrought	29	11	4	4	**47**	5
LocalAdaptation	3	3	0	2	2	**90**

B)

True model	**Model selected by ABC procedure**					
	GeneralENM	*Microsite*	*Multidimension*	*GrowCold*	*GrowDrought*	*LocalAdaptation*
GeneralENM	**0.26**					
Microsite		**2.12***				
Multidimension			**1.90***			
GrowCold				**1.38***		
GrowDrought					**0.69**	
LocalAdaptation						**5.00†**

C)

True model	**Model selected by ABC procedure**					
	GeneralENM	*Microsite*	*Multidimension*	*GrowCold*	*GrowDrought*	*LocalAdaptation*
GeneralENM		0.48	0.44	0.68	0.32	0.59
Microsite	0.26		0.70	0.54	1.41*	0.29
Multidimension	-	0.38		0.43	1.00	0.48
GrowCold	-	0.62	0.37		-	-
GrowDrought	0.26	0.72	0.73	0.92		0.79
LocalAdaptation	0.33	0.27	-	0.72	0.80	

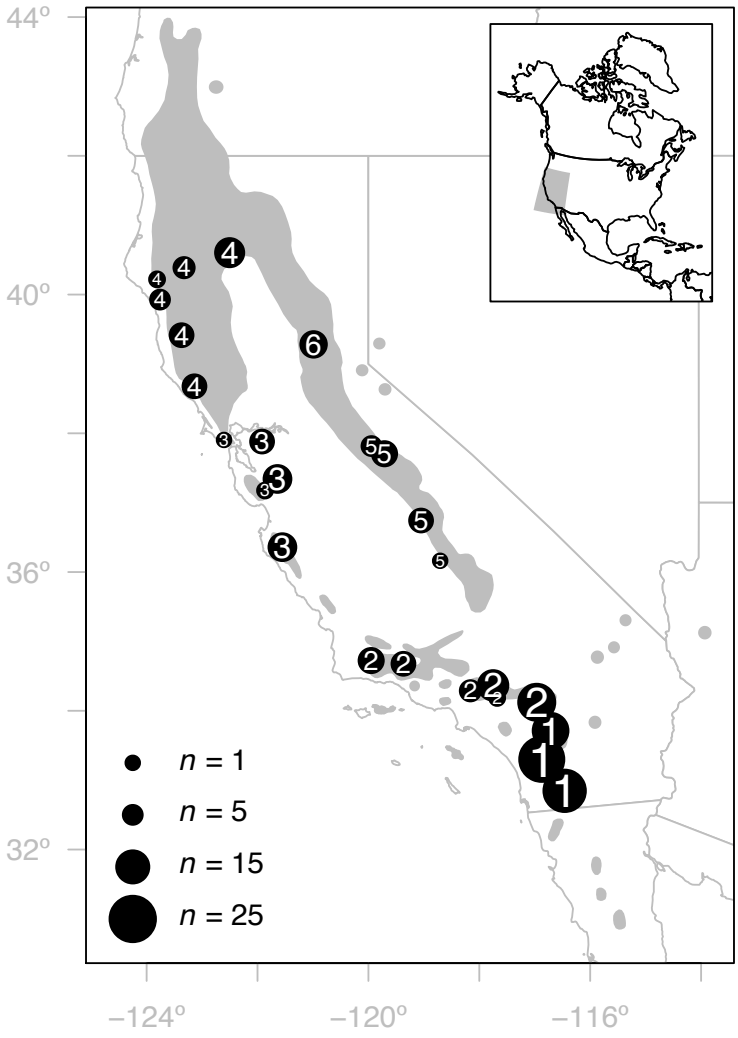

Figure 4.1. Geographic distribution and sampling localities of canyon live oak. Geographic distribution of canyon live oak (grey shading; according to Little 1971) and sampling localities, where the size of the black circle corresponds to the number of individuals collected (sampling localities that are very close together were combined). Numbers on the black circles indicate populations as follows: (1) Peninsular Ranges, (2) Transverse Ranges, (3) Southern Coast Ranges, (4) Northern Coast Ranges and Klamath Mountains, (5) Southern Sierra Nevada, and (6) Northern Sierra Nevada. Several small, disjunct portions of the species distribution located east of the depicted range are not shown.

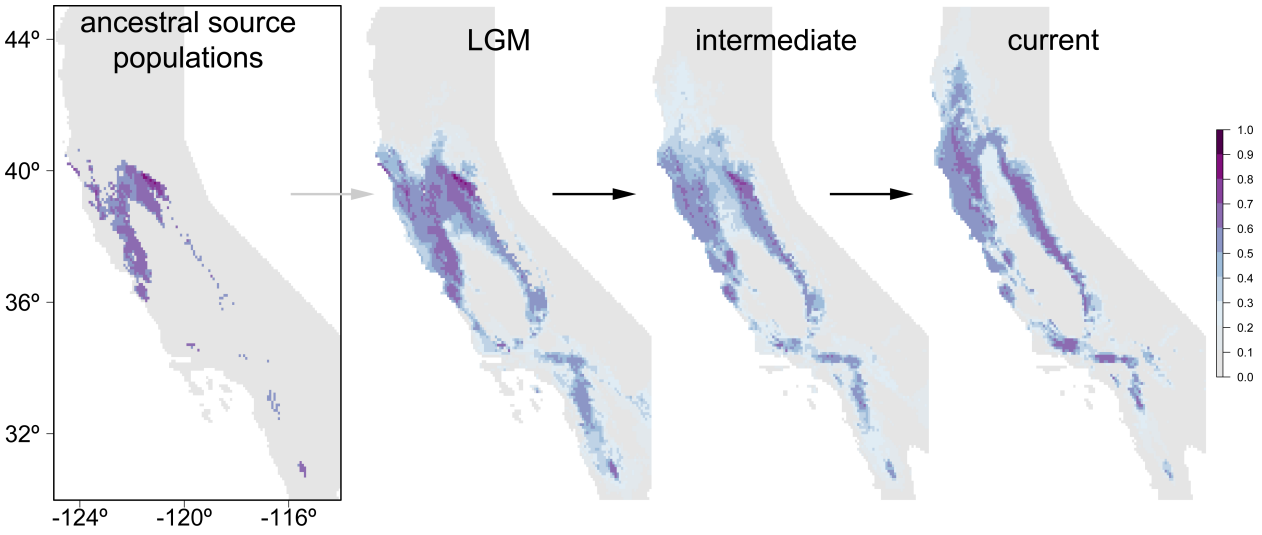

Figure 4.2. Example dynamic ecological niche model for canyon live oak. Dynamic ecological niche model used for demographic simulations, with an example illustrated for the *GeneralENM* model. Demes representing ancestral source populations (extracted from the areas of highest habitat suitability during the Last Glacial Maximum, LGM; see *Materials and Methods* for details) are initiated (grey arrow) within the LGM landscape at 21.5 ka. Demes are allowed to colonize the landscape, with carrying capacity and migration rate of each deme scaled relative to habitat suitability (coloured grid cells). Habitat suitability then shifts (black arrows) to that of intermediate and current time periods as the simulation progresses. One third of the total number of generations is simulated under each of the LGM, intermediate, and current landscapes.

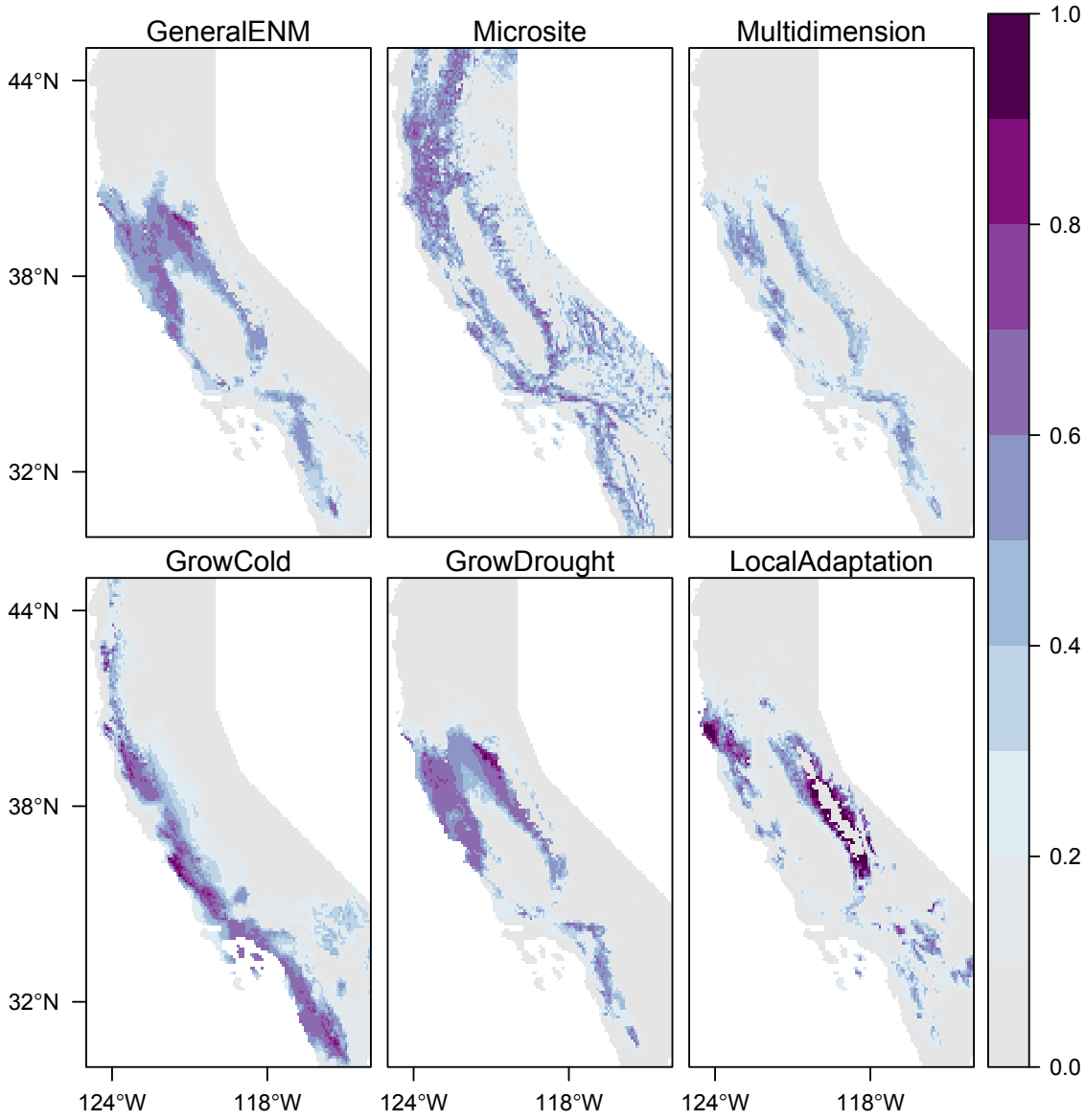

Figure 4.3. LGM habitat suitability of canyon live oak according to different models. Habitat suitability for canyon live oak during the Last Glacial Maximum (21.5 ka) from ecological niche models constructed for each of the iDDC models.

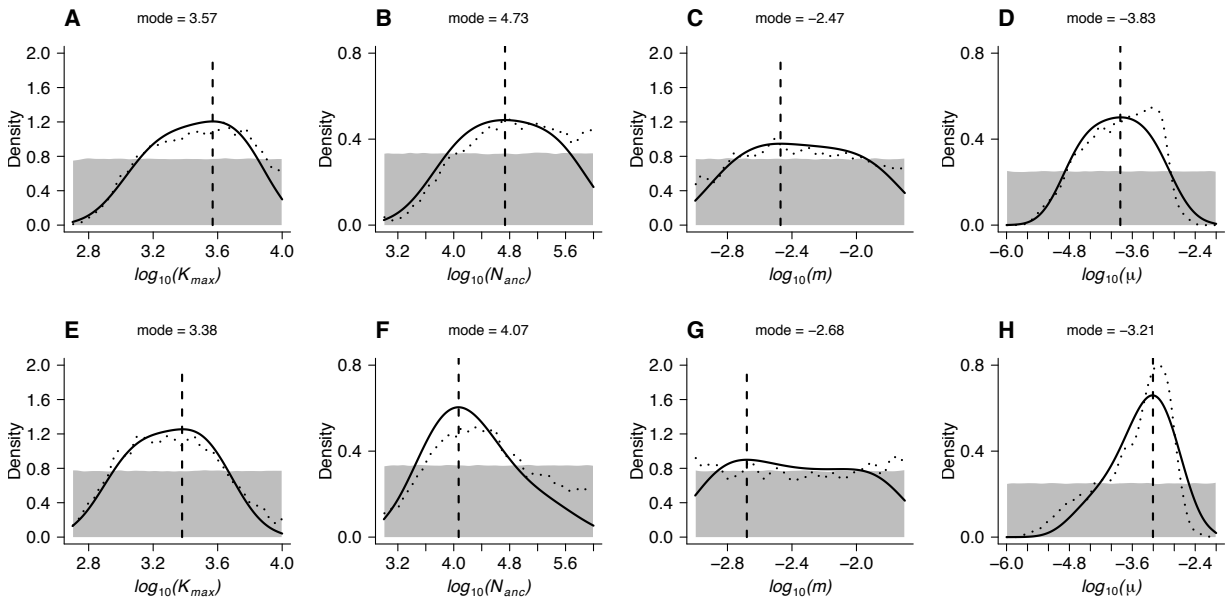

Figure 4.4. Posterior estimates of demographic parameters for canyon live oak. Prior and posterior distributions of model parameters for the two most supported models, *GeneralENM* (a-d) and *GrowDrought* (e-h). Grey shading: prior distribution; dotted black line: posterior distribution before the ABC-GLM procedure; solid black line: final posterior distribution following ABC-GLM. K_{max}, carrying capacity; N_{anc}, ancestral population size; m, migration rate; μ, microsatellite mutation rate.

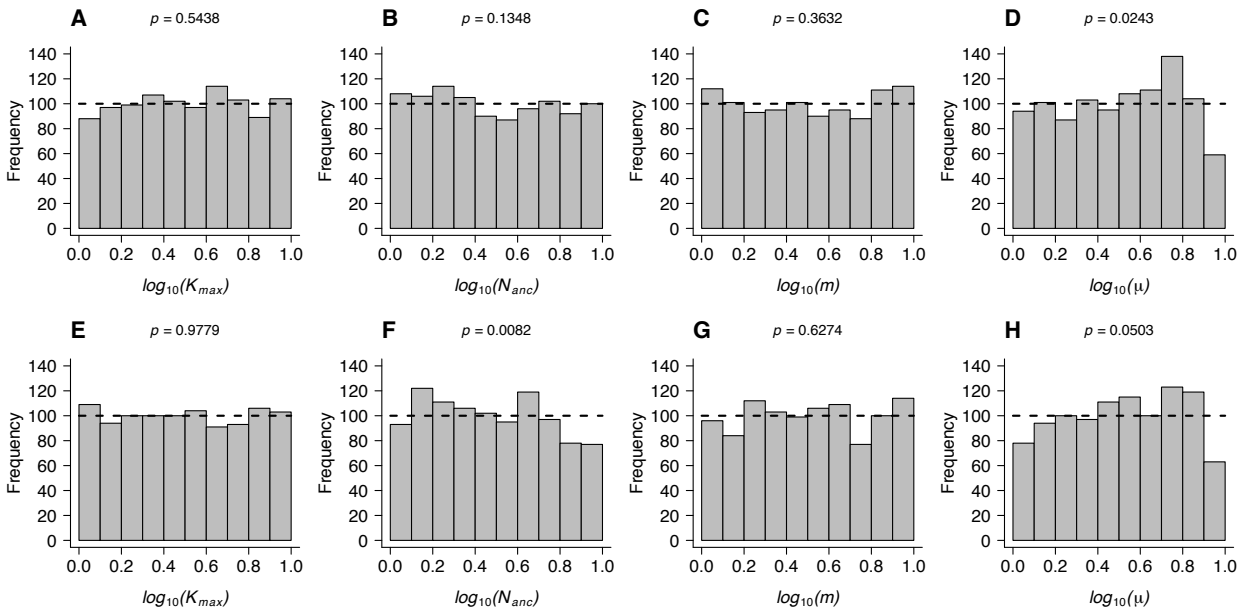

Figure 4.5. Bias in parameter estimation from the ABC procedure in canyon live oak. Distribution of posterior quantiles of true parameter values from 1,000 pseudo-observed datasets, used to assess bias in parameter estimation for the two most supported models, the *GeneralENM* (a-d) and *GrowDrought* (e-h) models. Posterior quantiles (grey bars) are compared to a uniform distribution (dashed black line). The *p*-values test for deviation from a uniform distribution using a Kolmogorov-Smirnov test, with *p*-values less than 0.05 indicating bias in parameter estimation. K_{max}, carrying capacity; N_{anc}, ancestral population size; *m*, migration rate; μ, microsatellite mutation rate.

APPENDIX C

Supplementary information from Chapter IV

Supplementary methods

Generation of ecological niche models

Ecological niche models (ENMs) were used to generate habitat suitability maps for canyon live oak in the present and during the Last Glacial Maximum (LGM, 21.5 ka), using maximum entropy modelling with MAXENT v.3.3.3k (Phillips *et al.* 2004, 2006) in the *R* package *dismo 1.0-12* (Hijmans *et al.* 2015b).

Species occurrence records were assembled using methods described by Ortego *et al.* (2015). Records were obtained from herbarium databases (Consortium of California Herbaria, *http://ucjeps.berkeley.edu/consortium*; Consortium of Pacific Northwest Herbaria, *http://www.pnwherbaria.org*; University of Arizona Herbarium, *http://ag.arizona.edu/herbarium*; Global Biodiversity Information Facility, *http://www.gbif.org*), as well as our own sampling localities. Only occurrence records from Oregon, California, and Baja California were used to construct ENMs, as our focus was on this portion of the species range. Relict, disjunct populations from Arizona, New Mexico, and Chihuahua might not have been connected with California populations since the LGM and could be adapted to different climatic conditions. Including these

populations would potentially bias ENMs intended to predict habitat suitability for California populations. Occurrence records were filtered to remove duplicate records, records with coordinate precision greater than 1 km, records that were within 1 km of another record, and duplicate records that fell within the same grid cell of our climatic data. A total of 1,406 unique observations were retained and used to construct ENMs.

Climatic and topographic data were obtained for both current and LGM conditions. All climate data for LGM conditions were derived from the Community Climate System Model v.4 (CCSM4; Gent *et al.* 2011), which has been shown to perform well for predicting reconstructed terrestrial climate conditions during the LGM (Harrison *et al.* 2014). Thirty-seven climate and topography variables were obtained from which to construct ENMs according to different hypotheses about the determinants of canyon live oak's geographic range (Table C.1). Nineteen bioclimatic variables representing temperature and precipitation regimes were obtained directly from the WorldClim Global Climate Dataset (www.worldclim.org; Hijmans *et al.* 2005). We calculated 17 additional variables of interest ourselves that could be derived from the WorldClim variables (Hijmans *et al.* 2005) combined with elevation (also used as a variable by itself; Amante & Eakins 2009) and solar radiation (www.cgiar-csi.org; Zomer *et al.* 2006, 2008). The calculation of these variables is summarized by Title & Bemmels (2018) from formulae originally described elsewhere (Thornthwaite 1948; Daget 1977; Wang *et al.* 2006, 2012; Zomer *et al.* 2006, 2008; Wilson *et al.* 2007; Sayre *et al.* 2009; Metzger *et al.* 2013; Hijmans *et al.* 2015a). We also estimated annual actual evapotranspiration (AET) using a bucket model (D. Golicher, ECOSUR, San Cristóbal, Mexico; Golicher 2012). AET represents the combination of plant transpiration and

evaporation but is difficult to directly measure or estimate because it is impacted by numerous factors (e.g., plant physiology, soil moisture, energy balance, watershed hydrology) operating at different spatial and temporal scales (Zhao *et al.* 2013). The bucket-model method of calculating AET is applicable only for averaging over long time periods at regional scales where local-scale factors such as watershed runoff caused by daily variation in rainfall are less relevant (Golicher 2012).

ENMs were constructed using MAXENT 3.3.3k (Phillips *et al.* 2004, 2006) with the *R* package *dismo 1.0-12* (Hijmans *et al.* 2015b) at 2.5-arcminute resolution, and were afterwards downscaled to 5-arcminute resolution (approximately 9 by 9 km) to decrease the number of grid cells and improve computational speed of demographic simulations. Habitat-suitability maps were generated for each model for current and LGM conditions. Glaciated regions of the Sierra Nevada and other mountain ranges (Gillespie *et al.* 2004) were masked in LGM habitat-suitability maps by converting habitat-suitability values of these regions to zero.

Selection of environmental variables included in each ENM

The specific environmental variables used to construct the ecological niche models (Table C.1) were selected in order to best reflect the ecological factors invoked in each hypothesis, summarized as follows:

(1) *GeneralENM:* as this model tested the impact of response to basic climate variables in a generic ENM, we used all 19 WorldClim bioclimatic variables (Hijmans *et al.* 2005). These variables are frequently used in ecological niche modelling, and represent basic characterizations of overall climate in terms of temperature and

precipitation.

(2) *Microsite*: because this model tested the effect of topographic microsite availability, we deliberately did not include any climatic variables. In particular, we hypothesized that areas of high topographic complexity may be most likely to contain ideal microsites for canyon live oak. We therefore included terrain roughness index and slope as measures of topographic complexity. We included elevation and aspect as additional descriptors of topography.

(3) *Multidimension*: because this model tested whether considering multiple aspects of ecological niche is necessary to understand response to climate change, we included all available variables (except elevation; see *Materials and Methods*). The 19 WorldClim bioclimatic variables (Hijmans *et al.* 2005) we included reflect basic climate; the topographic variables consider the impact of microsite availability; and we hypothesized that some of the additional climatic variables we included (Title & Bemmels 2018) may characterize climate in a manner more directly physiologically meaningful to determining habitat suitability for a plant. These additional variables included measures of actual and potential evapotranspiration, heat and moisture indices, length of the growing season, and annual heat accumulation (i.e., growing degree-days).

(4) *GrowCold:* for our two trade-off hypotheses, we selected variables that we hypothesized were most relevant to the physiological factors governing the trade-off (in terms of abiotic stress levels, and potential for avoiding or tolerating stress by growing during less stressful periods of the year). Rather than choosing every environmental variable that could in some way potentially be linked to each trade-off, we only selected a few that we assumed were most directly related to the trade-off in particular. We also

chose quarterly rather than monthly estimates (when both estimates were available), because quarterly estimates may be better descriptors of overall seasonal climate patterns exerting a general influence on plant physiology and phenology than extremes restricted to a single month.

To construct a model characterizing the potential trade-off between growth rate and cold tolerance, we first chose several temperature variables: mean annual temperature and Sayre's *et al.* thermicity index as descriptors of overall temperatures; mean temperature of the coldest quarter, which is likely related to the timing and severity of potentially damaging spring and fall frosts when risk of cold injury is highest (Howe *et al.* 2003); and mean temperature of the warmest quarter, which could reflect whether ideal temperatures for growth exist during the summer instead, when cold injury is not a concern. We also chose length of the frost-free period and growing degree-days ≥0ºC and ≥5ºC because these variables describe length of the growing season and available heat accumulation for growth. Conceptually, these variables are likely to reflect respectively the amount of cold stress and overall growth potential experienced by the plant. Finally, we also chose potential evapotranspiration as an additional descriptor of growth potential given unlimited water supply (since precipitation was not hypothesized to be a limiting factor for determining habitat suitability in this model).

(5) *GrowDrought*: To reflect the potential trade-off between growth rate and drought tolerance (note that drought stress primarily occurs during the summer in the California Floristic Province) we chose several variables related to overall and seasonal precipitation: mean annual precipitation and precipitation of the driest and warmest

quarters. We further used Thornthwaite's aridity, humidity, and moisture indices, as well as the UNEP aridity index, to characterize overall climatic dryness and moisture availability, and thus the amount of drought stress likely experienced by a plant. Actual evapotranspiration was chosen as a measure of overall growth potential of the plant, given constraints to water supply. Finally, Emberger's pluviothermic quotient, *Q*, was also included, which was originally developed as an index for differentiating among Mediterranean vegetation zones based on a physiologically inspired characterization of annual climatic dryness (Daget 1977). *Q* may thus reflect a continuum between areas of high growth potential and areas of high drought stress.

(6) *LocalAdaptation*: This model tested whether different factors have different relative importance for determining habitat suitability in different ecoregions, within which populations may be locally adapted (see *Materials and Methods*). Because we did not hypothesize a priori which factors would be most important in each specific ecoregion, we included all available variables (except elevation), as in the *Multidimension* hypothesis.

Table C.1. Environmental variables included in each model for canyon live oak. Climatic and topographic variables included when constructing ecological niche models (ENMs) for each iDDC model. An X in the table indicates that a given variable was included in an ENM, and variables with ≥ 5% permutation importance in the MAXENT ENM algorithm are indicated with an asterisk (*), or with a numerical superscript ($^{1\text{-}6}$) indicating the number of population models for which permutation importance was ≥ 5% for the *LocalAdaptation* model. Citations for variables sources and calculations are as follows: A: Hijmans *et al.* (2005); B: Golicher (2012); C: Title & Bemmels (2018); D: Zomer *et al.* (2006, 2008); E: Metzger *et al.* (2013); F: Thornthwaite (1948); G: Sayre *et al.* (2009); H: Daget (1977); I: Wang *et al.* (2006, 2012); J: Amante & Eakins (2009); K: Hijmans *et al.* (2015a); L: Wilson *et al.* (2007).

		Model					
Variable	**Citation**	*General-ENM*	*Micro-site*	*Multi-dimension*	*Grow Cold*	*Grow Drought*	*Local Adaptation*
Climate variables							
Bio1 - mean annual temp	A	X		X	X		X
Bio2 - mean diurnal temp range	A	X		X			X
Bio3 - isothermality	A	X		X			X[1]
Bio4 - temp seasonality	A	X*		X*			X[2]
Bio5 - max temp of warmest month	A	X		X			X
Bio6 - min temp of coldest month	A	X		X			X
Bio7 - temp annual range	A	X		X			X
Bio8 - mean temp of wettest quarter	A	X		X			X
Bio9 - mean temp of driest quarter	A	X		X			X
Bio10 - mean temp of warmest quarter	A	X		X	X		X
Bio11 - mean temp of coldest quarter	A	X		X	X*		X
Bio12 - annual precip	A	X		X		X	X
Bio13 - precip of wettest month	A	X		X			X
Bio14 - precip of driest month	A	X*		X*			X[4]
Bio15 - precip seasonality	A	X*		X*			X[4]
Bio16 - precip of wettest quarter	A	X		X			X
Bio17 - precip of driest quarter	A	X		X		X*	X[1]
Bio18 - precip of warmest quarter	A	X		X		X	X
Bio19 - pecip of coldest quarter	A	X*		X			X
actual evapotranspiration (AET)	B			X		X	X[2]
potential evapotranspiration (PET)	C, D			X	X*		X
PET seasonality	C, E			X			X[1]
Thornthwaite's aridity index	C, E, F			X		X	X
Thornthwaite's humidity index	C, E, F			X		X	X[1]
Thornthwaite's moisture index	C, F			X		X	X
UNEP aridity index	C, E			X		X	X[1]
Sayre's et al. thermicity index	C, E, G			X	X		X[1]
Emberger's pluviothermic quotient	C, E, H			X		X*	X[1]
length of frost-free period	I			X	X		X
annual number of frost-free days	I			X			X
growing degree days ≥ 0ºC	C, E			X	X*		X
growing degree days ≥ 5ºC	C, E			X	X*		X
monthCount10deg	C, E			X			X
Topographic variables							
elevation	J		X*				
aspect	C, K		X	X			X
slope	C, K		X	X			X
terrain roughness index (TRI)	C, K, L		X*	X*			X[2]

Table C.2. Population geographic coordinates and sample sizes for canyon live oak. Population and locality geographic coordinates and sample sizes for microsatellite genotyping. CR: Coast Ranges; NF: National Forest; Mtns: Mountains.

Pop	Population name	Locality name	Latitude	Longitude	*n*
1	Peninsular Ranges	Laguna Mountain	32.84524	-116.43885	12
1	Peninsular Ranges	Laguna Mountain	32.84954	-116.48535	10
1	Peninsular Ranges	Palomar Mtns.	33.28688	-116.80194	2
1	Peninsular Ranges	Palomar Mtns.	33.293433	-116.890233	7
1	Peninsular Ranges	Palomar Mtns.	33.30272	-116.87217	2
1	Peninsular Ranges	Palomar Mtns.	33.30433	-116.87156	4
1	Peninsular Ranges	Palomar Mtns.	33.30513	-116.87831	5
1	Peninsular Ranges	Palomar Mtns.	33.30585	-116.87193	2
1	Peninsular Ranges	Palomar Mtns.	33.31366	-116.87095	2
1	Peninsular Ranges	San Jacinto Mtns.	33.68201	-116.68956	3
1	Peninsular Ranges	San Jacinto Mtns.	33.6855	-116.69991	1
1	Peninsular Ranges	San Jacinto Mtns.	33.68962	-116.70644	2
1	Peninsular Ranges	San Jacinto Mtns.	33.69373	-116.71179	4
1	Peninsular Ranges	San Jacinto Mtns.	33.7283	-116.72005	4
1	Peninsular Ranges	San Jacinto Mtns.	33.74875	-116.73753	1
1	Peninsular Ranges	San Jacinto Mtns.	33.79186	-116.74465	2
2	Transverse Ranges	San Bernardino Mtns.	34.10532	-116.97227	3
2	Transverse Ranges	San Bernardino Mtns.	34.10846	-116.97408	3
2	Transverse Ranges	San Bernardino Mtns.	34.11334	-116.97994	4
2	Transverse Ranges	San Bernardino Mtns.	34.11794	-116.97784	3
2	Transverse Ranges	San Bernardino Mtns.	34.13028	-116.9825	2
2	Transverse Ranges	San Bernardino Mtns.	34.16885	-116.89307	3
2	Transverse Ranges	San Gabriel Mtns.	34.17832	-117.67668	1
2	Transverse Ranges	San Gabriel Mtns.	34.19299	-117.67851	1
2	Transverse Ranges	San Gabriel Mtns.	34.35678	-117.74315	4
2	Transverse Ranges	San Gabriel Mtns.	34.37137	-117.75443	5
2	Transverse Ranges	San Gabriel Mtns.	34.37299	-117.75391	4
2	Transverse Ranges	San Gabriel Mtns.	34.2985	-118.14864	2
2	Transverse Ranges	San Gabriel Mtns.	34.31516	-118.1368	2
2	Transverse Ranges	San Gabriel Mtns.	34.25204	-118.19614	2
2	Transverse Ranges	Los Padres NF	34.67807	-119.36795	8
2	Transverse Ranges	Figueroa Mountain	34.72447	-119.95008	9
3	Southern CR	Hastings	36.35896	-121.551	11
3	Southern CR	Almaden Quicksilver County Park	37.175667	-121.864365	2
3	Southern CR	Lick Observatory	37.342607	-121.639584	11
3	Southern CR	Mt Diablo State Park	37.88094	-121.92024	4
3	Southern CR	Mt Diablo State Park	37.881764	-121.915026	4
3	Southern CR	Talmapais State Park	37.904302	-122.604107	1
4	Northern CR & Klamath Mtns.	Sonoma	38.67822	-123.13693	8
4	Northern CR & Klamath Mtns.	Willits	39.38811	-123.42426	4
4	Northern CR & Klamath Mtns.	Willits	39.44133	-123.31432	4
4	Northern CR & Klamath Mtns.	Redwood Highway	39.92774	-123.75834	5
4	Northern CR & Klamath Mtns.	Avenue of the Giants	40.21926	-123.81198	2
4	Northern CR & Klamath Mtns.	Shasta-Trinity NF-36 Road	40.38203	-123.29874	5
4	Northern CR & Klamath Mtns.	Shasta-Trinity NF-36 Road	40.41838	-123.45644	1
4	Northern CR & Klamath Mtns.	Redding	40.60477	-122.5023	12
5	Southern Sierra Nevada	Sequoia NF	36.16261	-118.70589	1
5	Southern Sierra Nevada	Sequoia National Park	36.74134	-119.0313	4
5	Southern Sierra Nevada	Sequoia National Park	36.74482	-119.06886	4
5	Southern Sierra Nevada	Stanislaus NF - Yosemite National Park	37.66518	-119.80762	2
5	Southern Sierra Nevada	Stanislaus NF - Yosemite National Park	37.71377	-119.72743	2
5	Southern Sierra Nevada	Stanislaus NF - Yosemite National Park	37.715669	-119.677205	4
5	Southern Sierra Nevada	Stanislaus NF - Yosemite National Park	37.731082	-119.604807	1
5	Southern Sierra Nevada	Stanislaus NF - Yosemite National Park	37.8161	-119.94467	5
6	Northern Sierra Nevada	Tahoe NF	39.28197	-120.98866	10

Table C.3. Empirical genetic summary statistics in canyon live oak. Genetic summary statistics from empirical microsatellite data. Numerical subscripts on summary statistic names refer to population numbers (Fig. 4.1; Table C.2). Mean values are the mean of the individual population values, while total values are calculated over all populations combined. Asterisk (*): number of alleles (*K*) was not included as a summary statistic for model selection (see *Materials and Methods* for explanation).

Summary statistic	Observed value
Heterozygosity	
H_1	0.7506
H_2	0.7541
H_3	0.7761
H_4	0.7546
H_5	0.7662
H_6	0.6893
H_{mean}	0.7485
H_{total}	0.7779
Population differentiation	
F_{ST}	0.0330
$F_{ST(2,1)}$	0.0077
$F_{ST(3,1)}$	0.0356
$F_{ST(3,2)}$	0.0326
$F_{ST(4,1)}$	0.0434
$F_{ST(4,2)}$	0.0420
$F_{ST(4,3)}$	0.0094
$F_{ST(5,1)}$	0.0312
$F_{ST(5,2)}$	0.0264
$F_{ST(5,3)}$	0.0204
$F_{ST(5,4)}$	0.0310
$F_{ST(6,1)}$	0.0713
$F_{ST(6,2)}$	0.0814
$F_{ST(6,3)}$	0.0817
$F_{ST(6,4)}$	0.0635
$F_{ST(6,5)}$	0.0595
*Number of alleles**	
K_1	13.23
K_2	13.31
K_3	12.15
K_4	10.92
K_5	10.15
K_6	5.08
K_{mean}	10.81
K_{total}	17.85

Table C.4. Fit between empirical and simulated summary statistics in canyon live oak. Ability of each model to generate simulated summary statistics in agreement with summary statistics of the empirical data. A simulated distribution of summary-statistic values was generated for each model (10^5 simulations per model; parameter values drawn from prior distribution), and the percentile corresponding to the empirical summary-statistic value within this distribution was calculated. Reported in the table is the distance between the empirical value and the median (50th percentile) of the simulated distribution (possible values: -50 to +50); values approaching ± 50 indicate summary statistics that are difficult to fit, whereas values close to zero indicate that simulated values were similar to those observed in the empirical data. For example, a value of (+) 9.7 indicates that the empirical value is 9.7 percentiles above the simulated median value (i.e., the 59.7th percentile overall). Shading has been added to cells for visual clarity: no shading, empirical value within ± 20 percentiles of the simulated median value; light grey, within ± 20-30 percentiles; dark grey, ± 30-40 percentiles; black, ± 40-50 percentiles. Asterisk (*): number of alleles (*K*) was not included as a summary statistic for model selection (see *Materials and Methods* for explanation).

	Model					
Summary statistic	***General-ENM***	***Microsite***	***Multi-dimension***	***GrowCold***	***Grow-Drought***	***Local-Adaptation***
Heterozygosity						
H_1	(+) 9.7	(+) 6.3	(+) 11.4	(+) 10.1	(+) 10.5	(+) 9.2
H_2	(+) 12.9	(+) 8.5	(+) 10.7	(+) 14.9	(+) 10.7	(+) 22.8
H_3	(+) 19.6	(+) 27.3	(+) 35.3	(+) 41.5	(+) 17.1	(+) 25.9
H_4	(+) 9.4	(+) 7.3	(+) 30.7	(+) 27.6	(+) 8.4	(+) 7.7
H_5	(+) 9.4	(+) 9.0	(+) 8.7	(+) 8.7	(+) 8.8	(+) 23.8
H_6	(+) 3.5	(+) 0.7	(+) 3.6	(+) 5.3	(+) 2.0	(+) 4.4
H_{mean}	(+) 10.7	(+) 9.7	(+) 19.9	(+) 21.8	(+) 9.4	(+) 16.2
H_{total}	(+) 9.0	(+) 7.8	(+) 8.1	(+) 10.1	(+) 8.4	(+) 9.3
Population differentiation						
$F_{ST(2,1)}$	(-) 43.1	(-) 28.1	(-) 45.0	(-) 43.0	(-) 43.9	(-) 45.9
$F_{ST(3,1)}$	(-) 14.9	(-) 19.4	(-) 39.6	(-) 43.0	(-) 21.2	(-) 26.7
$F_{ST(3,2)}$	(-) 21.2	(-) 23.8	(-) 39.2	(-) 43.6	(-) 14.9	(-) 39.7
$F_{ST(4,1)}$	(+) 7.4	(+) 29.9	(-) 38.8	(-) 33.7	(-) 1.0	(+) 8.6
$F_{ST(4,2)}$	(-) 2.0	(+) 19.1	(-) 38.5	(-) 35.7	(+) 5.3	(-) 28.8
$F_{ST(4,3)}$	(-) 34.2	(-) 39.6	(-) 45.1	(-) 45.3	(-) 33.0	(-) 39.0
$F_{ST(5,1)}$	(-) 0.3	(+) 18.2	(-) 18.6	(+) 5.1	(-) 12.6	(-) 35.7
$F_{ST(5,2)}$	(-) 15.3	(+) 2.6	(-) 14.8	(-) 14.0	(-) 7.9	(-) 42.9
$F_{ST(5,3)}$	(-) 19.3	(-) 32.0	(-) 42.1	(-) 44.7	(-) 16.7	(-) 42.7
$F_{ST(5,4)}$	(+) 8.9	(+) 12.8	(-) 41.6	(-) 37.2	(+) 11.4	(-) 34.0
$F_{ST(6,1)}$	(+) 25.8	(+) 44.7	(+) 6.9	(+) 17.8	(+) 27.4	(+) 12.4
$F_{ST(6,2)}$	(+) 23.3	(+) 41.4	(+) 18.6	(+) 19.4	(+) 33.2	(-) 13.5
$F_{ST(6,3)}$	(+) 17.9	(+) 5.5	(-) 22.7	(-) 33.5	(+) 25.1	(-) 5.8
$F_{ST(6,4)}$	(+) 25.1	(+) 39.3	(-) 30.3	(-) 25.5	(+) 33.5	(+) 17.3
$F_{ST(6,5)}$	(+) 27.8	(+) 37.3	(+) 17.9	(+) 25.3	(+) 36.2	(-) 23.4
F_{ST}	(-) 12.6	(+) 5.6	(-) 36.9	(-) 35.5	(-) 9.4	(-) 38.4
*Number of alleles**						
K_1	(+) 28.4	(+) 26.8	(+) 29.8	(+) 26.8	(+) 27.8	(+) 27.3
K_2	(+) 30.3	(+) 27.7	(+) 28.8	(+) 30.5	(+) 28.3	(+) 34.5
K_3	(+) 45.3	(+) 47.5	(+) 49.1	(+) 49.9	(+) 44.4	(+) 47.1
K_4	(+) 26.4	(+) 25.6	(+) 43.4	(+) 40.5	(+) 24.8	(+) 24.8
K_5	(+) 21.5	(+) 21.6	(+) 20.9	(+) 20.6	(+) 20.2	(+) 30.0
K_6	(+) 1.5	(-) 0.4	(+) 1.0	(+) 1.6	(+) 0.6	(+) 0.8
K_{mean}	(+) 25.6	(+) 24.9	(+) 28.8	(+) 28.0	(+) 24.0	(+) 27.4
K_{total}	(+) 28.5	(+) 28.2	(+) 28.0	(+) 28.7	(+) 27.5	(+) 28.2

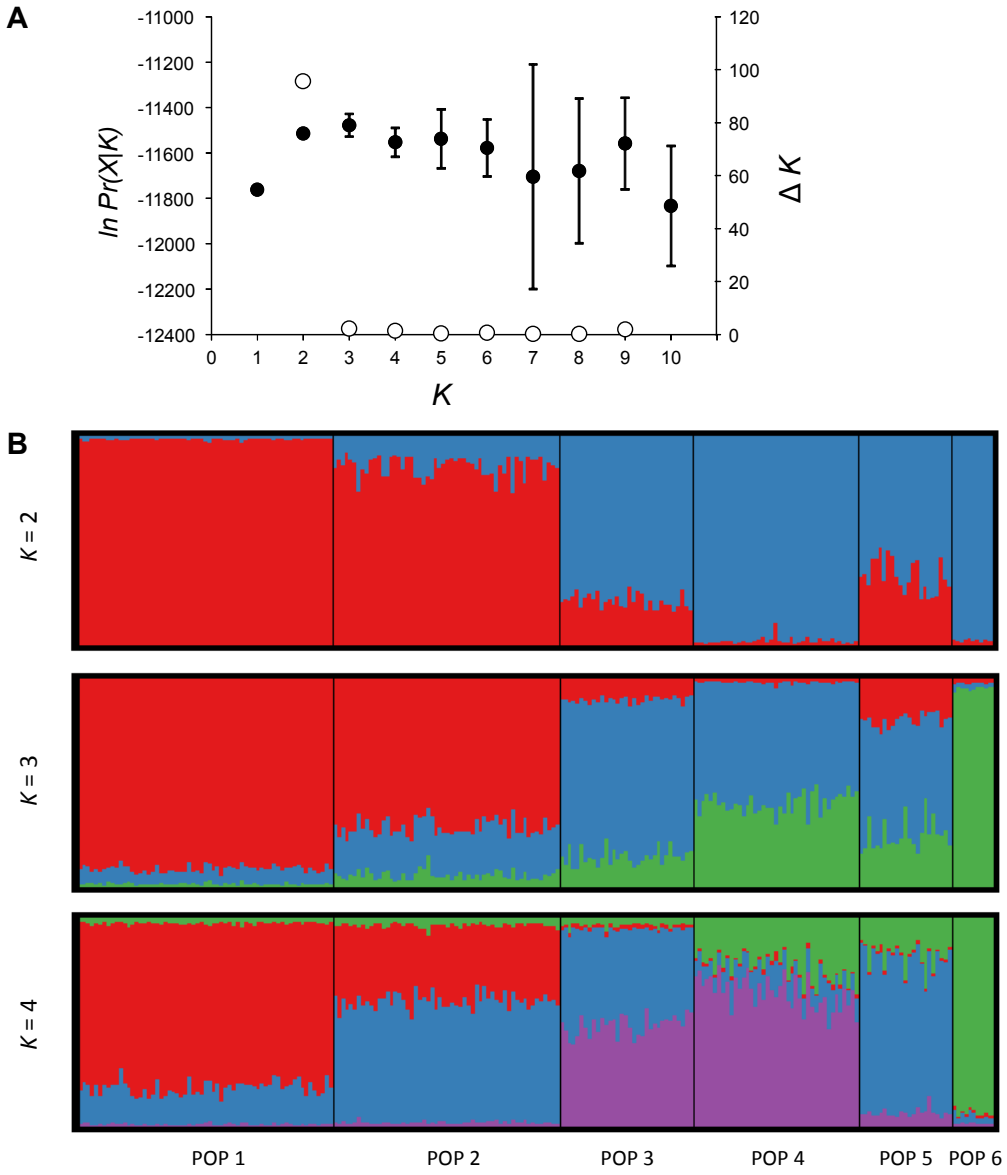

Figure C.1. Membership of canyon live oak individuals in genetic clusters from STRUCTURE. Results of Bayesian clustering analyses in STRUCTURE. (a) Plots show the mean (± SD) log probability of the data (*ln Pr(X|K)*) over 10 runs (left axis, black dots and error bars) and the magnitude of Δ*K* (right axis, open dots) for each value of *K* (number of clusters). (b) Genetic assignments of individuals based on the Bayesian analyses implemented in the program STRUCTURE. Each individual is represented by a vertical bar, which is partitioned into *K* coloured segments showing the individual's probability of belonging to the cluster with that colour. Thin vertical black lines separate individuals from the main geographical regions considered in this study according to population (see Fig. 4.1). Populations are defined as follows: (1) Peninsular Ranges, (2) Transverse Ranges, (3) Southern Coast Ranges, (4) Northern Coast Ranges and Klamath Mountains, (5) Southern Sierra Nevada, and (6) Northern Sierra Nevada.

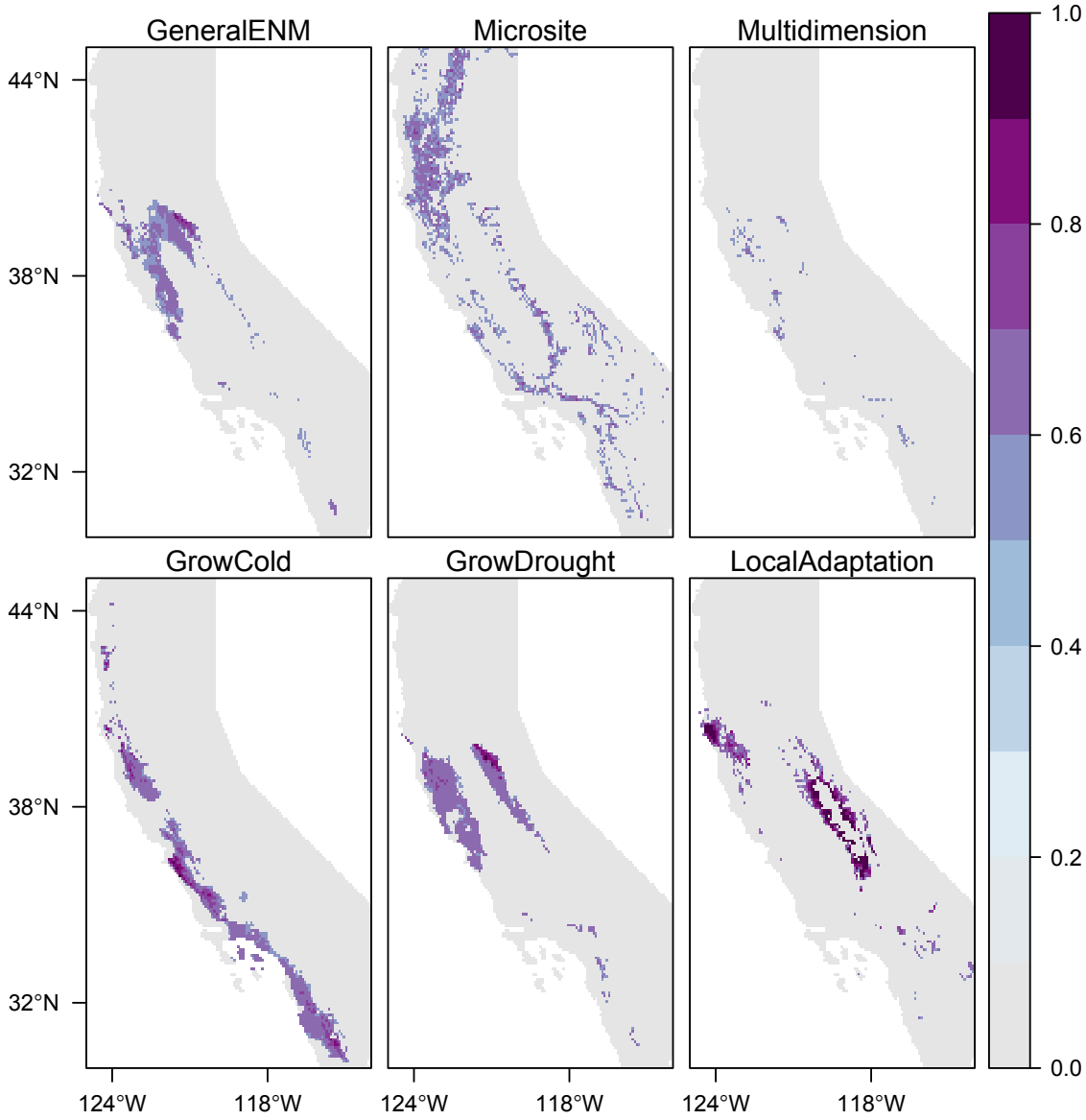

Figure C.2. Ancestral source populations of canyon live oak according to each model. Habitat suitability in ancestral source populations of canyon live oak from which demographic simulations were initiated for each model.

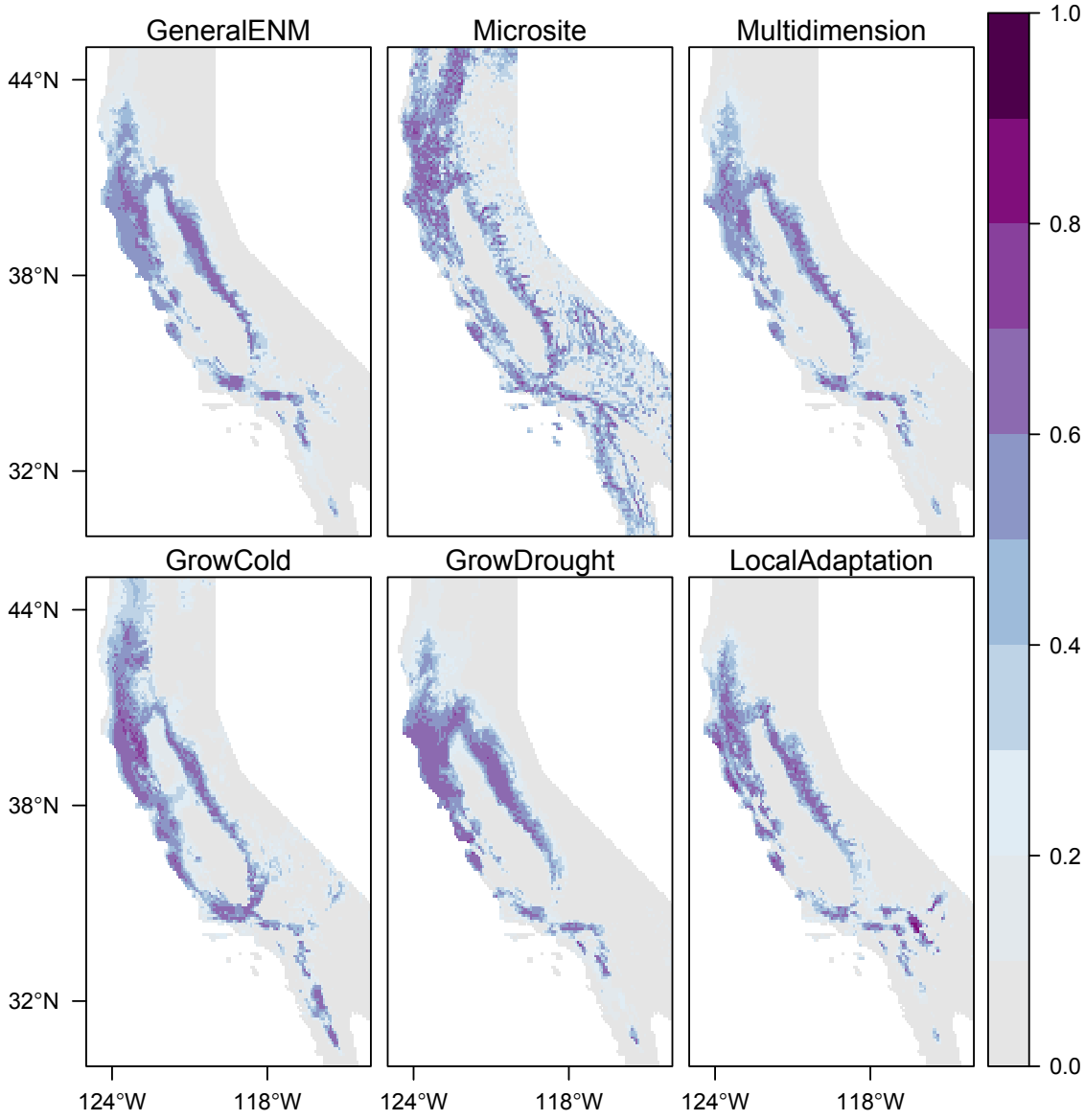

Figure C.3. Current habitat suitability of canyon live oak according to each model. Current habitat suitability for canyon live oak from ecological niche models constructed for each iDDC model.

Supplementary literature cited

Amante C, Eakins BW (2009) ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24. *National Geophysical Data Center, NOAA*, Boulder, CO.

Daget P (1977) Le bioclimat méditerranéen: analyse des formes climatiques par le système d'Emberger. *Vegetatio*, **34**, 87–103.

Gent PR, Danabasoglu G, Donner LJ *et al.* (2011) The Community Climate System Model version 4. *Journal of Climate*, **24**, 4973–4991.

Gillespie A, Porter S, Atwater BF (2004) *Quaternary glaciations - extent and chronology. Part II: North America. Additional CD-ROM*. Elsevier, San Diego.

Golicher D (2012) Implementing a bucket model using WorldClim layers. https://rpubs.com/dgolicher/2964. *RPubs*.

Harrison SP, Bartlein PJ, Brewer S *et al.* (2014) Climate model benchmarking with glacial and mid-Holocene climates. *Climate Dynamics*, **43**, 671–688.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans RJ, van Etten J, Cheng J *et al.* (2015a) Package "raster". https://cran.r-project.org/web/packages/raster/index.html.

Hijmans RJ, Phillips S, Leathwick J, Elith J (2015b) Package "dismo". http://cran.r-project.org/web/packages/dismo/index.html.

Howe GT, Aitken SN, Neale DB *et al.* (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, **81**, 1247–1266.

Metzger MJ, Bunce RGH, Jongman RHG *et al.* (2013) A high-resolution bioclimate map of the world: A unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, **22**, 630–638.

Ortego J, Gugger PF, Sork VL (2015) Climatically stable landscapes predict patterns of genetic structure and admixture in the Californian canyon live oak. *Journal of Biogeography*, **42**, 328–338.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips SJ, Dudík M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, 655–662.

Sayre R, Comer P, Warner H, Cress J (2009) *A new map of standardized terrestrial ecosystems of the conterminous United States: US Geological Survey Professional Paper 1768*. Reston, VA.

Thornthwaite CW (1948) An approach toward a rational classification of climate. *Geographical Review*, **38**, 55–94.

Title, PO, Bemmels JB (2018) ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291-307.

Wang T, Hamann A, Spittlehouse DL, Aitken SN (2006) Development of scale-free climate data for western Canada for use in resource management. *International Journal of Climatology*, **26**, 383–397.

Wang T, Hamann A, Spittlehouse DL, Murdock TQ (2012) ClimateWNA—high-resolution spatial climate data for western North America. *Journal of Applied Meteorology and Climatology*, **51**, 16–29.

Wilson MFJ, O'Connell B, Brown C, Guinan JC, Grehan AJ (2007) Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy*, **30**, 3–35.

Zhao L, Xia J, Xu C-Y *et al.* (2013) Evapotranspiration estimation methods in hydrological models. *Journal of Geographical Sciences*, **23**, 359–369.

Zomer RJ, Trabucco A, Bossio DA, Verchot L V. (2008) Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems & Environment*, **126**, 67–80.

Zomer RJ, Trabucco A, Van Straaten O, Bossio DA (2006) *Carbon, Land and Water: A Global Analysis of the Hydrologic Dimensions of Climate Change Mitigation through Afforestation/Reforestation. International Water Management Institute Research Report 101.* Colombo, Sri Lanka.

CHAPTER V

Filter-dispersal assembly of lowland Neotropical rainforests across the Andes

Abstract

Numerous Neotropical rainforest species are distributed in both Amazonia and Central America, reflecting a rich history of biotic interchange between regions. However, some plant lineages are endemic to one region, due in part to the dispersal barrier posed by the northern Andean cordilleras and adjacent savannas. To investigate the role of biogeographic filtering across the northern Andes in regional community assembly, we examined environmental tolerances, functional traits, and biogeographic distributions of >1000 woody plant species (trees, shrubs, lianas) locally co-occurring in forest plots in lowland Panama (542 species) and Amazonian Ecuador (667 species). High regional abundance was strongly predictive of the probability of being geographically widespread (i.e., present on both sides of the Andes). However, we also found that species with broad environmental tolerances (those that are able to inhabit high elevations and areas of low mean annual precipitation) were more likely to have a cross-Andean distribution even after accounting for regional abundance, suggesting that biogeographic filtering for these traits has mediated cross-Andean dispersal. Regional abundance and environmental tolerances were additionally associated with a suite of

life-history traits related to high dispersal-colonization ability, but most traits reflecting dispersal-colonization ability were not predictive of biogeographic distribution. Our results highlight how the process of biogeographic filtering, based primarily on environmental tolerances, has mediated regional-scale floristic assembly of Neotropical rainforests. The impacts of this process, which we term filter-dispersal assembly, are likely to be especially important to forests in Central America, where biotic interchange with Amazonia has heavily influenced regional community composition.

Introduction

Biotic interchange between North and South America has profoundly impacted regional community assembly in the hyperdiverse rainforests of the Neotropics (Croat and Busey 1975, Pennington and Dick 2004, Cody et al. 2010, Bacon et al. 2015a, b). This impact has been especially important in greater Central America (Gentry 1982, Dick et al. 2005), which we define broadly to include Central America proper, the adjacent Chocó region of South America, and southern Mexico. Gentry observed that Central American lowland rainforest trees are largely derived from a subset of Amazonian lineages, a pattern he attributed to dispersal from South America during the Great American Biotic Interchange following emergence of the Isthmus of Panama (Gentry 1982). Closure of the Isthmus has traditionally been dated to ca. 3 Ma (e.g., Coates and Stallard 2013). Recent geological and biological evidence has suggested a much earlier closure in the Middle to Late Miocene (Montes et al. 2015, Bacon et al.

2015a, b), but the precise timing remains contentious (O'Dea et al. 2016, Jaramillo et al. 2017, Molnar 2017).

Although closure of the Isthmus of Panama created a terrestrial connection between North and South America, other simultaneously emerging landscape features constrained opportunities for dispersal between lowland rainforests in the Amazon and Central America (Fig. 5.1). Continued Pliocene tectonic activity in northwestern South America resulted in rapid uplift of the Eastern Cordillera of the Colombian Andes, which had attained only around 40% of its current height prior to 4 Ma (Gregory-Wodzicki 2000). Given the relatively narrow range of temperatures to which many tropical species are adapted, the steep temperature gradients of the Andes likely present a formidable barrier to dispersal of lowland species (Janzen 1967). Some portions of the northernmost Eastern Cordillera of Colombia and Venezuela are lower in elevation and represent less of an elevational barrier, but the Llanos (savanna and grassland) and other dry habitats to the east and north of this region may prevent dispersal around the Andes for rainforest-adapted species. These habitats originated when the pan-Amazonian rainforest and wetland formerly covering most of lowland northern South America (Hoorn et al. 2010) was gradually replaced by dry habitats in the northwestern part of the continent from the late Miocene to early Pliocene (Jaramillo et al. 2010). On the opposite (western) side of the Andes, dry habitats are unlikely to have posed a major dispersal barrier, as conditions moist enough to support lowland rainforest taxa persisted in lower Central America and the Chocó throughout the Pliocene (Graham and Dilcher 1998) and Pleistocene (Bush and Colinvaux 1990, González et al. 2006).

Despite the dispersal barriers presented by the Andes and associated dry habitats (Fig. 5.1), numerous lowland rainforest plant species are distributed on both sides of the Andes. Ancient vicariance could explain the cross-Andean distributions of rainforest species with Miocene origins (Dick et al. 2013). Some of these species may have experienced oceanic dispersal between continents (Pennington and Dick 2004, Dick et al. 2007, Cody et al. 2010), or else migrated into Central America from the South American western side of the Andes following closure of the Isthmus of Panama. In contrast, vicariance is unable to explain cross-Andean distributions of species of post-Miocene origin (e.g., *Inga*; Richardson et al. 2001). Most of these species likely dispersed either over or around the Andes. Recent oceanic dispersal is also possible but may have been relatively rare, given the great distances separating Central American and Amazonian rainforests by the shortest marine and fluvial routes.

Considering these dispersal barriers, why are some lowland rainforest species widespread throughout the Neotropics, whereas others are restricted to only one side of the Andes? Broad environmental tolerances and life-history traits associated with high dispersal-colonization ability may have equipped species to navigate the challenges posed by the mountains and savannas of the northern Andes region. As such, this region may have acted as a biogeographic filter to intercontinental dispersal. We consider four mutually compatible hypotheses concerning the potential for biogeographic filtering across this region: firstly, species with broad elevational ranges may have been able to disperse across the incipient Eastern Cordillera of the Colombian Andes prior to rapid Pliocene uplift (Gregory-Wodzicki 2000) or along piedmont in lower montane forest. Secondly, drought tolerant species may have been

able to disperse around the northernmost Andes through dry environments, especially along habitat corridors such as the gallery forests of the Llanos that are similar in species composition to continuous rainforest (Behling and Hooghiemstra 1998). Thirdly, life-history traits associated with high dispersal-colonization ability could have facilitated dispersal across and colonization of new landscapes in general, including those of the northern Andes region. In tropical woody plants, these traits include wind dispersal, small seed size, large maximum stem diameter (a proxy for height, which is positively correlated with dispersal distance; Thomson et al. 2011), and traits associated with rapid growth (including low wood density and short-lived leaves characterized by low leaf mass per area) (Croat and Busey 1975, Swaine and Whitmore 1988, Nathan et al. 2002, Muller-Landau et al. 2008, Wright et al. 2010, Thomson et al. 2011). Alternatively, there might be few functional traits impacting the ability of lowland rainforest species to cross the Andes. If all species are ecologically equivalent (e.g., Hubbell 2001) in their ability to disperse across the Andes, then dispersal probability could depend solely on population size, with more abundant species being most likely to disperse between regions.

To test these four hypotheses, we examine associations between species traits and biogeographic distributions of locally co-occurring species from communities in the Amazon (Yasuní, Ecuador) and Central America (Barro Colorado Island, Panama) (Fig. 5.1; Table D.1). We identify a suite of traits distinguishing geographically widespread species (those found on both sides of the Andes) from regionally distributed species (found exclusively east or west of the Andes; Fig. 5.2), and consider how these patterns may have arisen due to biogeographic filtering across the northern Andes region.

Methods

Study sites and taxa

Lists of 810 and 621 woody plant taxa identified to the species level were assembled from Pontificia Universidad Católica del Ecuador (PUCE) and Smithsonian Tropical Research Institute (STRI) Center for Tropical Forest Science (CTFS) 50-ha plots at Yasuní, Ecuador and Barro Colorado Island (BCI), Panama, respectively (Fig. 5.1; Table D.1). The species list for Yasuní is incomplete for climbing species, as climbers were only censused in 30 subplots with a combined area of 1.2 ha (Romero-Saltos 2011). The list from BCI includes many species that do not occur in the 50-ha plot itself, but occur in similar habitat nearby.

Occurrence records

Georeferenced occurrence records were obtained for each species from the Global Biodiversity Information Facility (www.gbif.org) using the *R* package *rgbif* version 0.8.0 (Chamberlain et al. 2015). Quality controls were applied to ensure that occurrence records were correctly taxonomically identified and accurately georeferenced. In particular, records with the following issues were excluded using *rgbif*: mismatch between the country (or continent) implied by the geographic coordinates and the stated country (or continent) of origin; invalid coordinates or geodetic datum; failed or suspicious coordinate reprojection; coordinates with a value of zero; species name with no match, a fuzzy match, or a match to a higher taxonomic level; basis of observation

corresponding to a fossil specimen, human observation, living specimen, or machine observation; and records from areas outside the Neotropics. Following initial quality controls, occurrence records were spatially thinned to remove duplicate samples and reduce spatial clustering, and records with low precision were removed prior to use in subsequent analyses. A maximum of one occurrence record for each species was retained per cell of a 30-arcsecond grid (~0.93 km), and only occurrence records with geographic coordinates recorded to at least two decimal places of precision (~1.1 km) were retained.

Characterizing biogeographic distributions

Occurrence records were used to classify species by biogeographic distribution, depending on whether they inhabit only one side or both sides of the Andes. Yasuní species were classified as either $\text{eastern}_{\text{Yas}}$ (i.e., occurring only east of the Andes) or $\text{widespread}_{\text{Yas}}$, whereas BCI species were classified as either $\text{western}_{\text{BCI}}$ (i.e., occurring only west of the Andes) or $\text{widespread}_{\text{BCI}}$ (Fig. 5.2). Note that we define the area 'west of the Andes' to include not only the Chocó region of South America and lower Central America, which are directly west of the northern Andes, but also the extension of this region farther northward into Mesoamerica (Fig. 5.2c). To reduce, but likely not entirely eliminate, the effects of undetected errors in taxonomic identification (Baker et al. 2017) and georeferencing that escaped our quality-control procedure, we required a minimum of three unique occurrence records on the opposite side of the Andes (relative to Yasuní or BCI) in order to classify a species as geographically widespread. Species with zero occurrence records on the opposite side of the Andes were classified as occurring

exclusively east or west of the Andes, and those with 1-2 occurrence records on the opposite side were not assigned a biogeographic distribution and were excluded from further analysis. Although this approach excluded a substantial number of species from analysis, caution is needed when assessing biogeographic distributions based on occurrence records because vouchers of tropical trees can be incorrectly identified (Baker et al. 2017) and it is not clear if voucher misidentification could systematically bias our distribution estimates.

The distribution of species within specific ecoregions was characterized using global terrestrial ecoregions defined by The Nature Conservancy (maps.tcn.org) based on World Wildlife Fund major habitat types (Olson et al. 2001). A more conservative threshold was used requiring a minimum of five occurrence records from a given ecoregion to classify a species as present in an ecoregion, to reduce the impact of atypical occurrences in transitional, ecotonal areas.

Functional traits

Regional abundance of each species was defined as the number of occurrence records east of the Andes for species from Yasuní, and the number of occurrence records west of the Andes for species from BCI. Functional traits associated with life-history strategy (seed mass, leaf mass per area (LMA), wood density, maximum stem diameter, and dispersal mode; hereafter referred to as “life-history traits”) were assembled from datasets associated with the 50-ha plots and curated by N.C.G., S.A.Q., R.V. and S.J.W., as well as from published data (Kraft and Ackerly 2010) (see also Appendix D, Supplementary methods). Note that all analyses involving stem diameter

were performed on non-climbing species only, because the interpretation of stem diameter may differ for climbers compared to non-climbers.

Two additional traits reflecting environmental tolerances (elevational range and drought tolerance) were calculated for each species by characterizing distributions along elevation and precipitation gradients. Elevation (elev) and mean annual precipitation (MAP) were obtained at 30-arcsecond resolution from *WorldClim* (Hijmans et al. 2005) and extracted for each occurrence record using the *R* package *raster* version 2.3-40 (Hijmans et al. 2015). Elevational range was defined as the 90th percentile of the elevation ($Q90_{elev}$) of occurrence records for a given species, and reflects the ability of lowland species to survive at high elevations. The 90th percentile was used instead of the absolute maximum elevation in order to capture the general trend for the species and to reduce the effect of outliers that might be due to errors undetected by our quality-control procedure. Drought tolerance ($-Q10_{MAP}$) was defined as the 10th percentile of MAP of occurrence records for a given species, multiplied by negative one so that high values (close to zero) indicate high drought tolerance. Because $-Q10_{MAP}$ is based on average annual climatic conditions, it represents the typical environmental tolerance of a species but is not informative of ability to respond to atypical, acute drought episodes.

In order to compare environmental traits (i.e., elevational range and drought tolerance) among species, values must be calculated using occurrence records spanning geographic space that is potentially available for all species to inhabit. However, widespread species by definition have access to potential habitat on both sides of the Andes that exclusively eastern or western species do not have access to.

To avoid this potential source of bias, environmental traits were characterized using occurrence records only from east of the Andes for all Yasuní species, and using occurrence records only from west of the Andes for all BCI species (Fig. 5.2), as opposed to using occurrence records throughout the entire Neotropics. To ensure adequate sample size, indices were only calculated for species with at least 20 unique occurrence records from east or west of the Andes for Yasuní and BCI species, respectively.

Geographic sampling bias undoubtedly exists in our occurrence records, as many areas of the Neotropics – especially within the Amazon basin (Hopkins 2007) – are remote and poorly sampled. To test the effects of geographic sampling bias and whether the entire regions east and west of the Andes were appropriate spatial scales over which to calculate environmental traits, we also recalculated environmental traits using only occurrence records from more intensely sampled, smaller geographic areas with fewer potential biogeographic dispersal barriers: the western Amazon Basin (i.e., Colombia, Ecuador, and Peru east of the Andes) for Yasuní species, representing 50% of all occurrence records east of the Andes; and Costa Rica and Panama for BCI species, representing 43% of all occurrence records west of the Andes. Correlations among indices calculated over these narrow and more intensely sampled vs. broader and more sporadically sampled geographic spaces were moderate to high (r = 0.69-0.93), so we present analyses for elevational range and drought tolerance calculated only from complete occurrence records for each of the entire regions east and west of the Andes, as described above.

Finally, in addition to using mean annual precipitation (MAP) to calculate -$Q10_{MAP}$, we also generated two alternative indices of drought tolerance using precipitation of the driest quarter and precipitation seasonality instead of MAP. All three indices were highly correlated for both communities ($r \geq 0.74$), so we present analyses only for -$Q10_{MAP}$.

Several functional traits (regional abundance, elevational range, seed mass, LMA, and stem diameter) were $\log_{10}$ transformed in all analyses in order to approximate a normal distribution of trait values.

Principal component analysis

Principal component analysis (PCA) was used to define major axes of trait variation and examine whether species with different biogeographic distributions form distinct clusters along PC trait axes. All traits except dispersal mode (because it is a categorical trait) and all non-climbing species (including those without an assigned biogeographic distribution) were included in PCA, except that species were not included for which observations were missing for more than two of the seven traits. The remaining missing observations (except for regional abundance, for which complete data were available) were imputed using the joint modelling (JointM) method following Dray and Josse (2015), taking the mean of five imputations computed in the *R* package *Amelia* (Honaker et al. 2011). This resulted in 14.8% imputed data for Yasuní, and 12.5% imputed data for BCI. PCA was performed using the *prcomp* function in *R*, using a correlation matrix with values centred to a mean of 0 and units rescaled to standard deviation of 1. Separate analyses were performed for Yasuní and BCI because our indices of elevational range and drought tolerance were calculated over different

geographic areas (the regions east and west of the Andes, respectively), and it is not appropriate to treat these variables as directly equivalent for the purposes of defining PC trait axes.

Associations between biogeographic distributions and species traits

Regressions were performed using a phylogenetic generalized linear mixed model (PGLMM) for binary dependent traits (Ives and Helmus 2011, Ives and Garland Jr. 2014) to test for associations between species traits and biogeographic distributions. Binary PGLMM jointly estimates regression parameters and phylogenetic signal (σ^2) of regression residuals in the phylogenetic variance-covariance matrix (Ives and Helmus 2011, Ives and Garland Jr. 2014). Simultaneously estimating phylogenetic signal in regression residuals along with regression parameters has been shown to perform well when the need for phylogenetic correction is not known *a priori* (Revell 2010).

A single phylogeny containing all species in both communities (n = 1,292) was constructed for PGLMM regressions. First, the *R* program *brranching* (Chamberlain 2016) was used to query the *phylomatic* reference phylogeny *R20120829* for plants (Webb and Donoghue 2005). Relationships among taxa were extracted from the reference phylogeny, with missing tips inserted and unresolved relationships within clades treated as polytomies. All tips corresponding to species not present in our dataset were removed. Branch lengths were then adjusted using the *bladj* function in *Phylocom* (Webb et al. 2008), with internal node ages adjusted according to the *ages_exp* estimates of Gastauer and Meira-Neto (2016), based on fossil calibrations (Hedges and Kumar 2009, Bell et al. 2010).

PGLMM regressions were performed using the function *binaryPGLMM()* in the *R* program *ape* (Paradis et al. 2004). These regressions test whether individual traits are significant predictors of biogeographic distribution when accounting for phylogenetic relationships among taxa. Both univariate regressions (testing the effect of each trait individually) and multivariate regressions (testing the effects of all traits considered simultaneously, and using datasets including imputed trait values) were performed. For univariate regressions, Holm-Bonferonni correction for multiple comparisons (Holm 1979) was used to calculate adjusted *p*-values for trait effects, across each set of traits at Yasuní and BCI separately. To assess the goodness of fit of regression models and to approximate the amount of variation explained by individual parameters, R^2_{ce} of the models overall and partial R^2_{ce} of each trait were calculated using the *R* package *rr2* Ives (2017).

Results

Species distributions and traits

Geographic distributions (Fig. 5.2) were unambiguously assigned to 667 (out of 810 total) woody plant species from Yasuní and 542 (out of 621) species from BCI. 120 of these species were present in both communities. Approximately 70% of species had widespread distributions spanning both sides of the Andes, and ≥25% occurred at least occasionally in dry ecoregions including savannas and shrubland (Table D.2). Approximately 10-20% of widespread species, but almost no exclusively eastern or western species, were present in the Caribbean (Table D.2). The spatial distribution of

occurrence records varied widely across the Neotropics, likely reflecting the combined effects of both spatial sampling bias as well as the true geographic distributions of species (Fig. 5.2). For example, the high number of occurrence records from lower Central America and the western Amazon was expected, given that our study species are by definition known to occur in these regions. In contrast, the complete lack of occurrence records over much of the Brazilian Amazon likely reflects severe under-sampling of this region.

Sufficient high-quality occurrence records were available to generate indices of elevational range ($Q90_{elev}$) and drought tolerance ($-Q10_{MAP}$) for 564 species from Yasuní and 478 species from BCI. As life-history traits were gathered from existing databases (see Appendix D, Supplementary methods), the number of species for which these traits were available varied widely (Table D.3). Correlations among traits were generally modest (mean absolute value of r = 0.16; Table 5.1).

Associations between traits and biogeographic distributions

Several species traits were individually associated with a widespread biogeographic distribution (Fig. 5.3-5.4). Results of univariate regressions controlling for phylogenetic relatedness among species (Table D.4-D.5) showed that at both communities, the probability that a species had a widespread distribution increased with increasing regional abundance, increasing elevational range and increasing drought tolerance. In addition, the probability of being widespread increased at Yasuní with decreasing seed mass and decreasing wood density, and at BCI with increasing maximum stem diameter and with wind dispersal. However, the effects of wood density

at Yasuní and dispersal mode at BCI were no longer significant after correcting for multiple comparisons (Holm 1979). Although not one of our main hypotheses, we also tested whether biogeographic distribution was predicted by local abundance (the number of stems in the 50-ha plots) but found no effect at either site (Yasuní, $p = 0.13$; BCI, $p = 0.31$).

In multivariate regressions that simultaneously account for the effects of multiple traits, fewer traits were significant predictors of biogeographic distribution at each community (Table 5.2). The probability of having a widespread distribution increased at Yasuní with increasing regional abundance and increasing elevational range, and at BCI with increasing regional abundance and increasing drought tolerance. In both communities, regional abundance explained the greatest proportion of overall variance in biogeographic distribution, when controlling for the effects of other traits (partial R^2_{ce} = 0.14 and 0.12 at Yasuní and BCI, respectively; Table 5.2), but elevational range at Yasuní and drought tolerance at BCI also made substantial contributions (partial R^2_{ce} = 0.07 and 0.05, respectively).

Results of principal component (PC) analyses revealed that a similar suite of traits characterized the first and second axes of trait variation at each community, and that PC scores along these axes were significant predictors of biogeographic distribution at both communities (Fig. 5.5; Table D.4-D.5). In particular, widespread species at both sites tended to have low values of PC1 and high values of PC2. These values were associated with high regional abundance, broad elevational range, and high drought tolerance at both sites, as well as low seed mass and low LMA at Yasuní, and low wood density and large maximum stem diameter at BCI. Dispersal mode was

not included in PCA because it is a categorical trait. Species with different dispersal modes tended to cluster in distinct areas of PC trait space, but there was little correspondence between these clusters and the clusters occupied by species with different biogeographic distributions (compare Fig. 5.5 and Fig. D.1).

The proportion of widespread and spatially restricted species did not differ among families at Yasuní ($p = 0.12$), and differed at BCI ($p = 0.015$) for only two moderately species-poor families (Annonaceae and Myrtaceae; Table D.6) representing 5.7% of all species. These results suggest that biases in the biogeographic distributions of specific species-rich families are unlikely to have driven the overall patterns we observed.

Discussion

The northern Andes region as a biogeographic filter

Biotic interchange of lowland species between the Amazon and Central America has been mediated, in part, by biogeographic filtering across the Northern Andes region. Although high regional abundance was the strongest predictor of a widespread (cross-Andean) biogeographic distribution, several other species traits, including broad elevational range and high drought tolerance, were also strong predictors of a widespread distribution in at least one community and added additional explanatory power beyond the effects of regional abundance alone (Fig. 5.3; Table 5.2). These results suggest that species with broad ecological amplitude along temperature (i.e.,

elevation) and precipitation gradients were more likely to have dispersed across or around the Andes than species with narrower ecological amplitudes.

Several possible mechanisms may underlie the strong correlation between high regional abundance and a widespread biogeographic distribution (Fig. 5.3, Table 5.2). If all species have ecologically equivalent abilities to disperse across the Andes, then dispersal may be a chance event with a probability roughly proportional to population size. Such a model could be considered an ecologically neutral (*sensu* Hubbell 2001) null hypothesis of species' distributions in relation to the Andes. Alternatively, well-known tropical tree species with large geographic ranges may be more frequently sampled and correctly identified than poorly known species with restricted geographic ranges (Ruokolainen et al. 2002, Baker et al. 2017). Sampling bias could strengthen the relationship between regional abundance (i.e., number of occurrence records) and biogeographic distribution, but is unlikely to be the sole cause of this pattern given that regional abundance varies over several order of magnitude (Fig. 5.3), likely reflecting real differences in the relative commonness of different species. We also note that very poorly known taxa (i.e., unidentified morphospecies and species with ≤ 20 occurrence records east or west of the Andes) were excluded from our multivariate analyses, likely reducing the potential effects of this bias.

While the aforementioned mechanisms do not involve biogeographic filtering, biogeographic filtering could also lead to a positive correlation between regional abundance and a widespread distribution if, for example, the same traits that affect regional abundance also affect cross-Andean dispersal ability. Positive relationships between landscape-scale abundance and geographic range size have previously been

documented in tropical trees (Kristiansen et al. 2009, ter Steege et al. 2013, Dexter and Chave 2016). The causes of variation in abundance and range size are still contentious (Kristiansen et al. 2009), but substantial phylogenetic signal in these traits at the generic level suggests that they are both determined by intrinsic characteristics of taxa (Dexter and Chave 2016). While we cannot distinguish the relative contributions of neutral dispersal dynamics, sampling bias, and trait-based biogeographic filtering processes on the observed correlation between regional abundance and biogeographic distribution, it is clear that regional abundance is not the sole driver of cross-Andean dispersal ability because several other traits were also strongly predictive of biogeographic distribution (Fig. 5.3), and added additional explanatory power to our analyses not captured by regional abundance alone (Table 5.2).

In particular, broad elevational range and high drought tolerance were also strongly correlated with a widespread biogeographic distribution at Yasuní and BCI, respectively (Table 5.2). Given that many lowland Neotropical species were able to inhabit higher elevations during periods of warmer climate (Weng et al. 2007), we suggest that some species may have dispersed over the Eastern Cordillera of the Colombian and Venezuelan Andes during warm periods. The Huancabamba Depression in northern Peru might have been an alternative route for dispersal, but we consider this route less likely because it has a more ancient uplift history (Gregory-Wodzicki 2000, Hoorn et al. 2010) and higher elevation than the lowest mountain passes of the Eastern Cordillera near the Colombian-Venezuelan border (2,145 m vs. ca. 1,300 m, respectively; Weigend 2002). In addition to elevation, precipitation is another key environmental determinant of the distribution of lowland rainforest plants at

broad spatial scales (Engelbrecht et al. 2007, Condit et al. 2013, Esquivel-Muelbert et al. 2017) and only a subset of rainforest trees is also able to inhabit dry areas (Esquivel-Muelbert et al. 2017). Dispersal through the Llanos and other lowland dry habitats around the northernmost Andes, especially along habitat corridors such as gallery forests, represents an alternative to high-elevation dispersal, but may have been available to only the most drought tolerant species.

Although Yasuní and BCI are both lowland rainforest communities, we note that BCI has lower precipitation and a more pronounced dry season than Yasuní, and contains a greater proportion of species that also inhabit dry ecoregions (Table D.1, Table D.2c). These habitat differences are one possible explanation for the differences among sites in terms of which environmental tolerances are most predictive of a widespread biogeographic distribution (Table 5.2). The effect of drought tolerance was significant (after accounting for the effects of other traits) only at BCI, whereas the effect of elevational range was significant only at Yasuní. Biogeographic filtering across the Andes evidently had different effects on the assembly of the different communities, but in both communities, widespread species appear to be more likely than exclusively eastern or western species to be broadly distributed across landscape-scale ecotonal transitions between habitat types in the surrounding area. For example, the widespread component of the BCI community exhibits higher drought tolerance than the exclusively western component and may be more broadly distributed along the transition from rainforest to seasonally dry forest across central Panama. In contrast, the widespread component of the Yasuní community exhibits greater elevational range than the exclusively eastern component, and may be more broadly distributed along the

transition from lowland to premontane forest in the Ecuadorean Amazon and lower slopes of the Andes.

Widespread and weedy rainforest plants

Beyond regional abundance and environmental tolerances, we found little evidence that additional functional traits were predictive of biogeographic distribution (Fig. 5.3-5.4). The life-history traits we examined were individually either non-significantly or only weakly predictive (Fig. 5.3-5.4), and did not add explanatory power to multivariate analyses not already captured by other traits (Table 5.2). Individual life-history traits therefore appear to be relatively unimportant direct determinants of ability to disperse across the Andes. However, principal component analyses revealed that several of these traits contributed to the same axes of trait variation as high regional abundance and broad environmental tolerances (Fig. 5.5). These traits included low seed mass and low LMA at Yasuní, and low wood density and large maximum stem diameter at BCI. While not individually predictive of biogeographic distribution, these traits jointly contribute to a general life-history strategy we consider to characterize 'weedy' lowland Neotropical woody plants (see also Dick et al. 2005). The general characteristics of these weedy species (reflected by trait loadings on PC axes, Fig. 5.5) include high regional abundance, broad ecological amplitude along elevation and precipitation gradients, and high dispersal-colonization ability.

In contrast to regional abundance and ecological amplitude, traits reflecting dispersal-colonization ability were largely not predictive of biogeographic distribution in relation to the Andes. These traits are nonetheless likely related to dispersal-

colonization ability and overall weediness at smaller (i.e., local) scales. Increasing maximum height (which we assessed using maximum stem diameter as a proxy) and decreasing seed size are known to be correlated with increased dispersal distance in both wind- and animal-dispersed tropical trees (Muller-Landau et al. 2008, Thomson et al. 2011). Once dispersal occurs, species that grow and reproduce rapidly may be able to quickly establish populations and colonize new areas (Swaine and Whitmore 1988). A functional trade-off exists between growth rate and mortality in tropical forest trees, with rapidly growing species characterized by low wood density, and to a lesser extent, by small seeds, low LMA characteristic of short-lived leaves, and low maximum height (Wright et al. 2010). These traits each contributed to the PC axes characterizing weediness in at least one community (Fig. 5.5), except that weedy species tended to have high (not low) maximum height (i.e., maximum stem diameter; BCI community only, Fig. 5.3, Fig. 5.5). However, we note that this relationship was predicted given the positive correlation between maximum height and dispersal ability described above (Muller-Landau et al. 2008, Thomson et al. 2011). Previous studies have also documented a positive correlation between maximum height and geographic range size in Amazonian trees (e.g., palms; Ruokolainen et al. 2002, Kristiansen et al. 2009).

Filter-dispersal assembly of Central American rainforests

Dispersal has played a major role in the floristic assembly of Neotropical rainforests, not only within contiguous regions such as the Amazon basin (Dexter et al. 2017), but also between North and South America (Gentry 1982, Cody et al. 2010, Bacon et al. 2015a) and from other areas of the world into the Neotropics (Pennington

and Dick 2004, Antonelli et al. 2015). While Dexter et al. (2017) recently proposed that 'dispersal assembly' drives the composition of species pools across geographic sub-regions of the Amazon basin, our results suggest that constraints on dispersal imposed by biogeographic filtering have played a much larger role in the floristic assembly of larger regions of the Neotropics (such as Central America or the entire Amazon basin), according to a process we term 'filter-dispersal assembly'. Filter-dispersal assembly occurs when biogeographic filtering across a dispersal barrier has a large impact on regional community composition on one side of the barrier, relative to the total species pool of a broader geographical area.

Filter-dispersal assembly across the northern Andes region was likely especially important in the assembly of Central American rainforests because Central American lowland forests have been strongly influenced by asymmetric biotic interchange with South America (Simpson 1980, Gentry 1982, Dick et al. 2005). Plant species that dispersed from Amazonia or are derived from dispersing lineages comprise a large proportion of the regional species pool in Central America, but not vice versa (Gentry 1982). Gentry (1982) observed that dispersing plant lineages tend to represent only a subset of Amazonian lineages with exceptionally large geographic ranges; we have additionally shown that species with a widespread (cross-Andean) distribution are more likely to have broad elevational ranges and high drought tolerance, highlighting the importance of biogeographic filtering in determining which species are able to disperse between continents.

Filter-dispersal assembly of Central American rainforests contrasts with dispersal assembly (Dexter et al. 2017) of Amazonian rainforests and suggests that different

dynamics may have driven the regional floristic assembly of these two regions. Within Amazonia, floristic inventories have revealed that most tree (Pitman et al. 1999, Wittmann et al. 2013) and liana (Romero-Saltos et al. 2001) species have broad geographic ranges, indicating that there are few major barriers to dispersal of woody plants. Pitman et al. (2002) compared tree species from plots at Yasuní to those at Manu National Park, Peru, which is similar to BCI in terms of species richness, precipitation regime and geographic distance from Yasuní, but is not separated from Yasuní by a biogeographic barrier. In that study, 30% of taxonomically identified species were shared between both Yasuní and Manu (Pitman et al. 2002), compared to only 9% between Yasuní and BCI (121 of 1,307 species; this study). However, the number of unidentified morphospecies at Manu is likely much higher than at BCI. Accounting for unidentified morphospecies, which tend to have small geographic ranges (Ruokolainen et al. 2002), could reduce the percentage of species shared between Yasuní and Manu.

In addition, Dexter et al. (2017) found that regional species pools of four major tree clades show complete lack of geographic phylogenetic structure across the Amazon basin, and concluded that dispersal likely homogenizes species pools across Amazonia over evolutionary timescales. Interestingly, however, Central America shows significant phylogenetic clustering in three of the same four clades, reflecting the relative isolation of Central American rainforests compared to those across the Amazon basin (Dexter et al. 2017). Thus, the northern Andes biogeographic filter has likely posed constraints on dispersal and prevented full homogenization of the Central American flora relative to that of Amazonia, even though substantial biotic interchange has

evidently still occurred (evidenced by the large number of species present in both regions; Table D.2).

Our analyses were restricted to single communities in Central America and in the Amazon, and it would not be meaningful to directly compare traits across communities because local-scale ecological filtering in different habitats (Table D.1) may result in different trait distributions at BCI compared to Yasuní. However, one implication of asymmetric filter-dispersal assembly is that relative to Amazonian rainforests, Central American rainforests in general may be enriched in species with the set of weedy traits we have identified; this prediction remains to be tested. We also note that filter-dispersal assembly is likely to have large importance to understanding species-level biogeographic distributions, but is likely less important in understanding regional community assembly at the genus and family levels, as the origins of most Neotropical plant genera and families and the cross-continental distributions of many of these lineages substantially predate Pliocene uplift of the northern Andes and formation of dry habitats (Cody et al. 2010, Bacon et al. 2015a).

Impacts of multiples processes

We have emphasized the impact of biogeographic filtering on species distributions, but acknowledge that contemporary distributions have been shaped by multiple dispersal and diversification processes over time periods spanning millions of years. The impacts of these processes are likely to obfuscate, rather than reinforce, any general patterns we might detect. In contrast, the strong signal for an effect of species traits on biogeographic distribution, in agreement with predictions of our hypotheses,

suggests that biogeographic filtering has had an enduring and general impact on species-level biotic interchange across the Andes.

Nonetheless, not all widespread species attained a cross-Andean distribution by terrestrial dispersal over or around the Andes. Some species likely attained a cross-Andean distribution through ancient vicariance (Dick et al. 2013). We could not identify these putatively vicariant species because we do not know the timing of origin of individual taxa, as the best phylogeny we have available is taxonomically incomplete, not well resolved at available nodes, and lacks robust estimates of divergence times of most extant species. However, we predict that vicariant widespread species (with no history of dispersal) may be more similar to exclusively eastern or western species than to recently dispersing widespread species, which would likely weaken, rather than strengthen, the patterns we observe. In addition to ancient vicariance, some species may have experienced long-distance dispersal over water barriers. The finding that 10-20% of widespread species also occurred on Caribbean islands (Table D.2) provides strong evidence that some species are capable of oceanic dispersal, but the relative contribution of oceanic dispersal to biotic interchange between Central and South America remains unknown.

On-going speciation and extinction processes may also confound our analyses, especially if some exclusively eastern or western species are derived from lineages with a recent cross-Andean dispersal history. This situation could arise if, for example, a single formerly widespread ancestral species went locally extinct on one side of the Andes, or rapidly evolved into separate eastern and western species. In either scenario, the resulting eastern and western species would be expected to possess traits more

similar to widespread species, further reducing detectable differences between species with different biogeographic distributions.

While we have documented broad patterns and hypothesized several dispersal routes that are highly plausible given the geological and climatic history of northern South America, future work will aid in uncovering the finer details of how biotic interchange proceeded for particular species. Genetic and phylogeographic studies will be needed in order to determine the geographic origins of widespread taxa; to infer specific routes and timing of dispersal; to determine which population splits between the Amazon and Central America predate Andean uplift (e.g., Dick et al. 2013) and represent vicariance events; and to quantify the relative contribution of dispersal and *in situ* diversification to regional species pools in both regions (e.g., Xing and Ree 2017). These studies will help reinforce our understanding of how biogeographic filtering across the northern Andes region, in combination with other processes, has mediated biotic interchange and driven floristic assembly of Neotropical communities, and Central American rainforests in particular.

Acknowledgements

The authors thank J.B. Pease for discussions about the project and H.G. Romero-Saltos for the list of liana species at Yasuní. Graduate student support was provided to J.B.B. by an NSF GRFP fellowship (DEB 1256260) and the University of Michigan. Additional NSF support was provided to C.W.D. (FESD 1338694 and DEB 1240869).

Author contributions

JBB and CWD conceived the project; SJW, NCG, SAQ, and RV provided data; JBB performed data analyses, with input from SJW and CWD; JBB wrote the chapter, with input from CWD and other authors.

Literature cited

Antonelli, A. et al. 2015. An engine for global plant diversity: highest evolutionary turnover and emigration in the American tropics. - Front. Genet. 6: 1–14.

Bacon, C. D. et al. 2015a. Biological evidence supports an early and complex emergence of the Isthmus of Panama. - Proc. Natl. Acad. Sci. 112: 6110–6115.

Bacon, C.D. et al. 2015b. Reply to Lessios and Marko et al.: early and progressive migration across the Isthmus of Panama is robust to missing data and biases. - Proc. Natl. Acad. Sci. Lett. 112: E5767–5768.

Baker, T. R. et al. 2017. Maximising synergy among tropical plant systematists, ecologists, and evolutionary biologists. - Trends Ecol. Evol. 32: 258-267.

Behling, H. and Hooghiemstra, H. 1998. Late Quaternary palaeoecology and palaeoclimatology from pollen records of the savannas of the Llanos Orientales in Colombia. - Palaeogeogr. Palaeoclimatol. Palaeoecol. 139: 251–267.

Bell, C. D. et al. 2010. The age and diversification of the angiosperms re-revisited. - Am. J. Bot. 97: 1296–1303.

Bush, M. B. and Colinvaux, P. A. 1990. A pollen record of a complete glacial cycle from lowland Panama. - J. Veg. Sci. 1: 105–118.

Chamberlain, S. 2016. brranching: fetch “phylogenies” from many sources. R package version 0.2.0. https://github.com/ropensci/brranching.

Chamberlain, S. et al. 2015. rgbif: Interface to the Global Biodiversity Information Facility. R package version 0.8.0. http://CRAN.R-project.org/package=rgbif.

Coates, A.G. and Stallard, R. F. 2013. How old is the Isthmus of Panama? - Bull. Mar.

Sci. 89: 801–813.

Cody, S. et al. 2010. The Great American Biotic Interchange revisited. - Ecography 33: 326–332.

Condit, R. et al. 2013. Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. - PNAS 110: 5064–5068.

Croat, T. B. and Busey, P. 1975. Geographical affinities of the Barro Colorado Island Flora. - Brittonia 27: 127–135.

Dexter, K. and Chave, J. 2016. Evolutionary patterns of range size, abundance and species richness in Amazonian angiosperm trees. - PeerJ 4: e2402.

Dexter, K. G. et al. 2017. Dispersal assembly of rain forest tree communities across the Amazon basin. - PNAS 114: 2645–2650.

Dick, C. W. et al. 2005. Biogeographic history and the high beta-diversity of Panamanian rainforest trees in Panama. - In: Harmon, R. S. (ed), The Río Chagres, Panama: a multidisciplinary profile of a tropical watershed. Springer, pp. 259–269.

Dick, C. W. et al. 2007. Extreme long-distance dispersal of the lowland tropical rainforest tree *Ceiba pentandra* L. (Malvaceae) in Africa and the Neotropics. - Mol. Ecol. 16: 3039–3049.

Dick, C. W. et al. 2013. Neogene origins and implied warmth tolerance of Amazon tree species. - Ecol. Evol. 3: 162–169.

Dray, S. and Josse, J. 2015. Principal component analysis with missing values: a comparative survey of methods. - Plant Ecol. 216: 657–667.

Engelbrecht, B. M. J. et al. 2007. Drought sensitivity shapes species distribution patterns in tropical forests. - Nature 447: 80–82.

Esquivel-Muelbert, A. et al. 2017. Seasonal drought limits tree species across the Neotropics. - Ecography 40: 618–629.

Gastauer, M. and Meira-Neto, J. A. A. 2016. An enhanced calibration of a recently released megatree for the analysis of phylogenetic diversity. - Brazilian J. Biol. 76: 619–628.

Gentry, A. H. 1982. Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleisotcene fluctuations, or an accident of the Andean orogeny? - Ann. Missouri Bot. Gard. 69: 557–593.

González, C. et al. 2006. Late Quaternary vegetation and climate change in the Panama Basin: Palynological evidence from marine cores ODP 677B and TR 163-38. - Palaeogeogr. Palaeoclimatol. Palaeoecol. 234: 62–80.

Graham, A. and Dilcher, D. L. 1998. Studies in Neotropical paleobotany. XII. A palynoflora from the Pliocene Rio Banano formation of Costa Rica and the Neogene vegetation of Mesoamerica. 85: 1426–1438.

Gregory-Wodzicki, K. M. 2000. Uplift history of the Central and Northern Andes: a review. - Geol. Soc. Am. Bull. 112: 1091–1105.

Hedges, S. B. and Kumar, S. 2009. The Timetree of Life. - Oxford University Press.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. - Int. J. Climatol. 25: 1965–1978.

Hijmans, R. J. et al. 2015. Package “raster”. R package version 2.3-40. https://cran.r-project.org/web/packages/raster/index.html.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. - Scand. J. Stat. 6: 65–70.

Honaker, J. et al. 2011. *Amelia II*: a program for missing data. - J. Stat. Softw. 45: 1–47.

Hopkins, M. J. G. 2007. Modelling the known and unknown plant biodiversity of the Amazon Basin. J. Biogeogr. 34: 1400-1411.

Hoorn, C. et al. 2010. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. - Science 330: 927–931.

Hubbell, Stephen P. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton, NJ.

Ives, 2017. 2017. R^2s for correlated data: phylogenetic models, LMMs, and GLMMs. - bioRxiv doi: 10.1101/144170.

Ives, A. R. and Helmus, M. R. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. - Ecol. Monogr. 81: 511–525.

Ives, A. R. and Garland Jr., T. 2014. Phylogenetic regression for binary dependent variables. - In: Garamszegi, L. Z. (ed), Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology. Springer-Verlag, pp. 231–261.

Janzen, D. H. 1967. Why mountain passes are higher in the tropics. - Am. Nat. 101: 233–249.

Jaramillo, C. et al. 2010. The origin of the modern Amazon rainforest: implications of the palynological and paleobotanical record. - In: Hoorn, C. and Wesselingh, F. P. (eds), Amazonia, landscape and species evolution: a look into the past. Blackwell, pp. 317–334.

Jaramillo, C. et al. 2017. Comment (1) on “Formation of the Isthmus of Panama” by

O'Dea et al. - Sci. Adv. 3: e1602321.

Kraft, N. J. B. and Ackerly, D. D. 2010. Functional trait and phylogenetic tests of community assembly across spatial scales in an Amazonian forest. - Ecol. Monogr. 80: 401–422.

Kristiansen, T. et al. 2009. Commonness of Amazonian palm (Arecaceae) species: Cross-scale links and potential determinants. - Acta Oecologica 35: 554–562.

Molnar, P. 2017. Comment (2) on "Formation of the Isthmus of Panama" by O'Dea et al. – Sci. Adv. 3: e1602320.

Montes, C. et al. 2015. Middle Miocene closure of the Central American Seaway. - Science 348: 226–229.

Muller-Landau, H. C. et al. 2008. Interspecific variation in primary seed dispersal in a tropical forest. - J. Ecol. 96: 653–667.

Nathan, R. et al. 2002. Mechanisms of long-distance dispersal of seeds by wind. - Nature 418: 409–413.

O'Dea, A. et al. 2016. Formation of the Isthmus of Panama. - Sci. Adv. 2: e1600883.

Olson, D. M. et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth. - Bioscience 51: 933–938.

Paradis, E. et al. 2004. APE: Analyses of Phylogenetics and Evolution in *R* language. - Bioinformatics 20: 289–290.

Pennington, R. T. and Dick, C. W. 2004. The role of immigrants in the assembly of the South American rainforest tree flora. - Philos. Trans. R. Soc. B 359: 1611–1622.

Pitman, N. C. A. et al. 1999. Tree species distributions in an upper Amazonian forest. - Ecology 80: 2651–2661.

Pitman, N. C. A. et al. 2002. A comparison of tree species diversity in two upper Amazonian forests. - Ecol. Soc. Am. 83: 3210–3224.

Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. - Methods Ecol. Evol. 1: 319–329.

Richardson, J. E. et al. 2001. Rapid diversification of a species-rich genus of Neotropical rain forest trees. - Science 293: 2242–2245.

Romero-Saltos, H. G. 2011. Community and functional ecology of lianas in the Yasuní Forest Dynamics Plot, Amazonian Ecuador. PhD thesis, University of Miami.

Romero-Saltos, H. et al. 2001. Patrones de diversidad, distribución y rareza de plantas leñosas en el Parque Nacional Yasuní y la Reserva Étnica Huaorani, Amazonía

ecuatoriana. - In: Duivenvoorden, J. F. et al. (eds), Evaluación de recursos vegetales no maderables en la Amazonía noroccidental. IBED, Universiteit van Amsterdam, pp. 131–162.

Ruokolainen, K. et al. 2002. Two biases in estimating range sizes of Amazonian plant species. - J. Trop. Ecol. 18: 935–942.

Simpson, G. G. 1980. *Splendid Isolation: The Curious History of South American Mammals.* - Yale University Press.

Swaine, M. D. and Whitmore, T. C. 1988. On the definition of ecological species groups in tropical rain forests. - Vegetatio 75: 81–86.

ter Steege, H. et al. 2013. Hyperdominance in the Amazonian tree flora. - Science 342: 1243092.

Thomson, F. J. et al. 2011. Seed dispersal distance is more strongly correlated with plant height than with seed mass. - J. Ecol. 99: 1299–1307.

Webb, C. O. and Donoghue, M. J. 2005. *Phylomatic*: tree assembly for applied phylogenetics. - Mol. Ecol. Notes 5: 181–183.

Webb, C. O. et al. 2008. *Phylocom*: software for the analysis of phylogenetic community structure and trait evolution. - Bioinformatics 24: 2098–2100.

Weigend, M. 2002. Observations on the biogeography of the Amotape-Huancabamba Zone in northern Peru. - Bot. Rev. 68: 38–54.

Weng, C. et al. 2007. Response of pollen diversity to the climate-driven altitudinal shift of vegetation in the Colombian Andes. - Philos. Trans. R. Soc. B 362: 253–262.

Wittmann, F. et al. 2013. Habitat specifity, endemism and the neotropical distribution of Amazonian white-water floodplain trees. - Ecography 36: 690–707.

Wright, S. J. et al. 2010. Functional traits and the growth-mortality trade-off in tropical trees. - Ecology 91: 3664–3674.

Xing, Y. and Ree, R. H. 2017. Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. - Proc. Natl. Acad. Sci. 114: e3444–e3451.

Table 5.1. Correlations among species traits in Neotropical trees. Correlations (Pearson's *r*) between environmental and life-history traits at Yasuní (above diagonal) and Barro Colorado Island (below diagonal).

	regional abundance	elevational range	drought tolerance	seed mass	LMA	wood density	stem diameter
regional abundance	-	0.19	0.46	-0.16	-0.08	-0.04	-0.08
elevational range	0.38	-	0.41	-0.14	-0.15	-0.22	-0.07
drought tolerance	0.46	0.17	-	-0.15	-0.10	-0.06	0.17
seed mass	-0.17	-0.09	-0.09	-	0.21	0.27	0.13
LMA	-0.15	-0.07	-0.14	0.32	-	0.25	0.01
wood density	-0.10	-0.22	0.02	0.17	0.19	-	-0.30
stem diameter	0.00	0.05	0.24	0.00	0.08	-0.12	-

Table 5.2. Multivariate models predicting biogeographic distributions of Neotropical trees. Multivariate regressions performed using a phylogenetic generalized linear mixed model for binary dependent traits (Ives and Helmus 2011, Ives and Garland Jr. 2014) for species from (a) Yasuní and (b) Barro Colorado Island. Estimates show the effect of species traits on predicting a widespread biogeographic distribution. The parameter σ^2 is a measure of phylogenetic signal in the phylogenetic variance-covariance matrix. Statistical significance of adjusted p-values is indicated as follows: . p < 0.10, * p < 0.05; ** p < 0.01, *** p < 0.001. R^2_{ce} for the overall models and partial R^2_{ce} for each trait are measures of explained variance, and were calculated following Ives (2017).

(a)	Yasuní (n = 447, R^2_{ce} = 0.31)					
parameter	**estimate**	**standard error**	**z-score**	**p**		**partial R^2_{ce}**
σ^2	6.29×10^{-2}			0.35		
intercept	-18.32	4.09	-4.48	7.51×10^{-6}	***	
regional abundance	3.56	0.48	7.39	1.42×10^{-13}	***	0.14
elevational range	3.75	0.70	5.38	7.54×10^{-8}	***	0.07
drought tolerance	1.50×10^{-4}	3.85×10^{-4}	0.39	0.70		< 0.01
seed mass	-0.09	0.16	-0.58	0.56		< 0.01
leaf mass per area	-0.04	1.25	-0.03	0.98		< 0.01
wood density	1.24	1.03	1.20	0.23		< 0.01
stem diameter	0.10	0.47	0.21	0.84		< 0.01

(b)	Barro Colorado Island (n = 253, R^2_{ce} = 0.27)					
parameter	**estimate**	**standard error**	**z-score**	**p**		**partial R^2_{ce}**
σ^2	1.85×10^{-10}			0.50		
intercept	-11.48	4.09	-2.80	5.05×10^{-3}	**	
regional abundance	2.53	0.59	4.32	1.56×10^{-5}	***	0.12
elevational range	1.47	0.97	1.51	0.13		0.02
drought tolerance	1.26×10^{-3}	4.39×10^{-4}	2.87	4.07×10^{-3}	**	0.05
seed mass	-0.22	0.21	-1.04	0.30		< 0.01
leaf mass per area	1.61	1.50	1.07	0.28		0.01
wood density	1.65	1.45	1.14	0.26		0.01
stem diameter	0.71	0.38	1.86	0.06	.	0.01

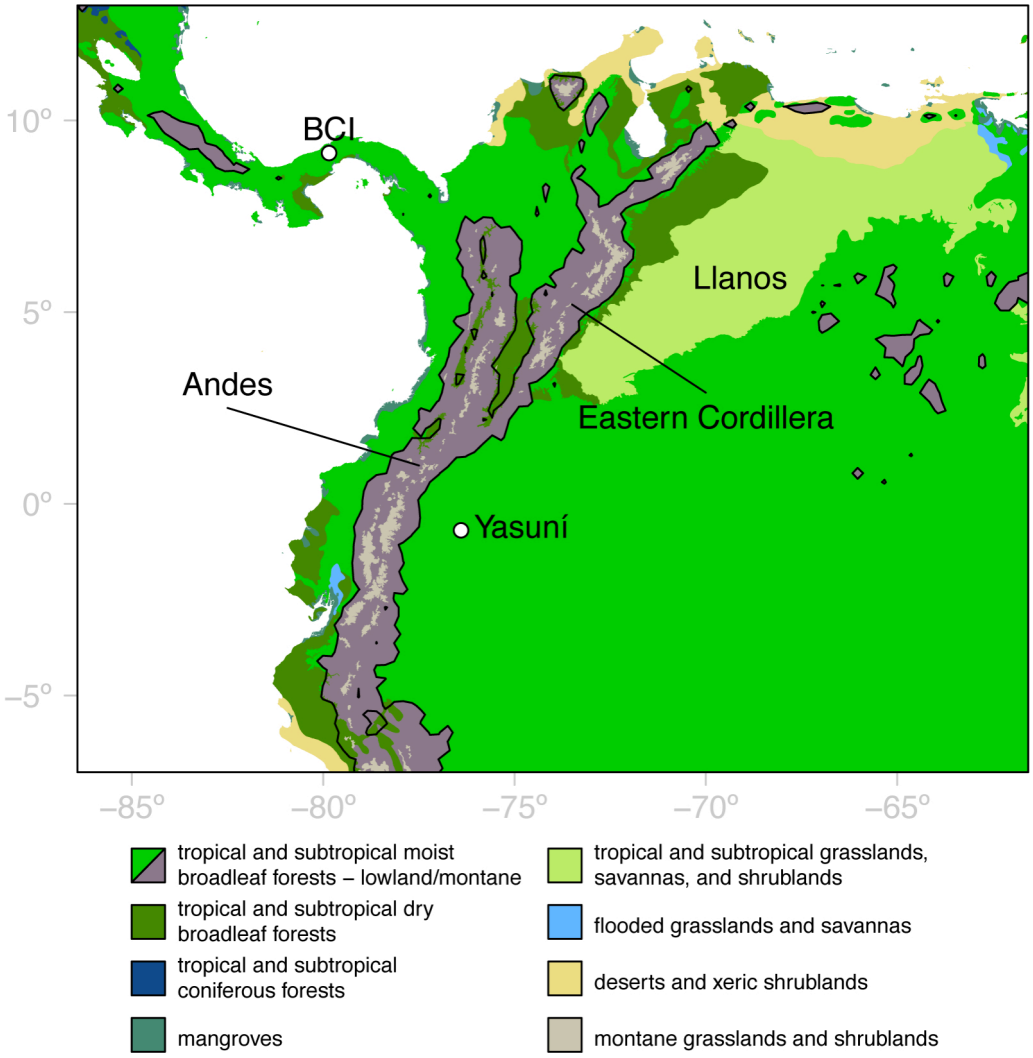

Figure 5.1. Map of the landscapes of northwestern South America. Landscapes of northwestern South America showing the location of lowland rainforest (bright green) and other ecoregions in relation to major biogeographic barriers, including the Eastern Cordillera of the Colombian Andes and the Llanos. The Andes are outlined by black contour lines 1000 m in elevation, with moist broadleaf forests above this elevation (purple) primarily consisting of montane forest. Study sites are shown as white dots. Ecoregion distributions are derived from Olson et al. (2001).

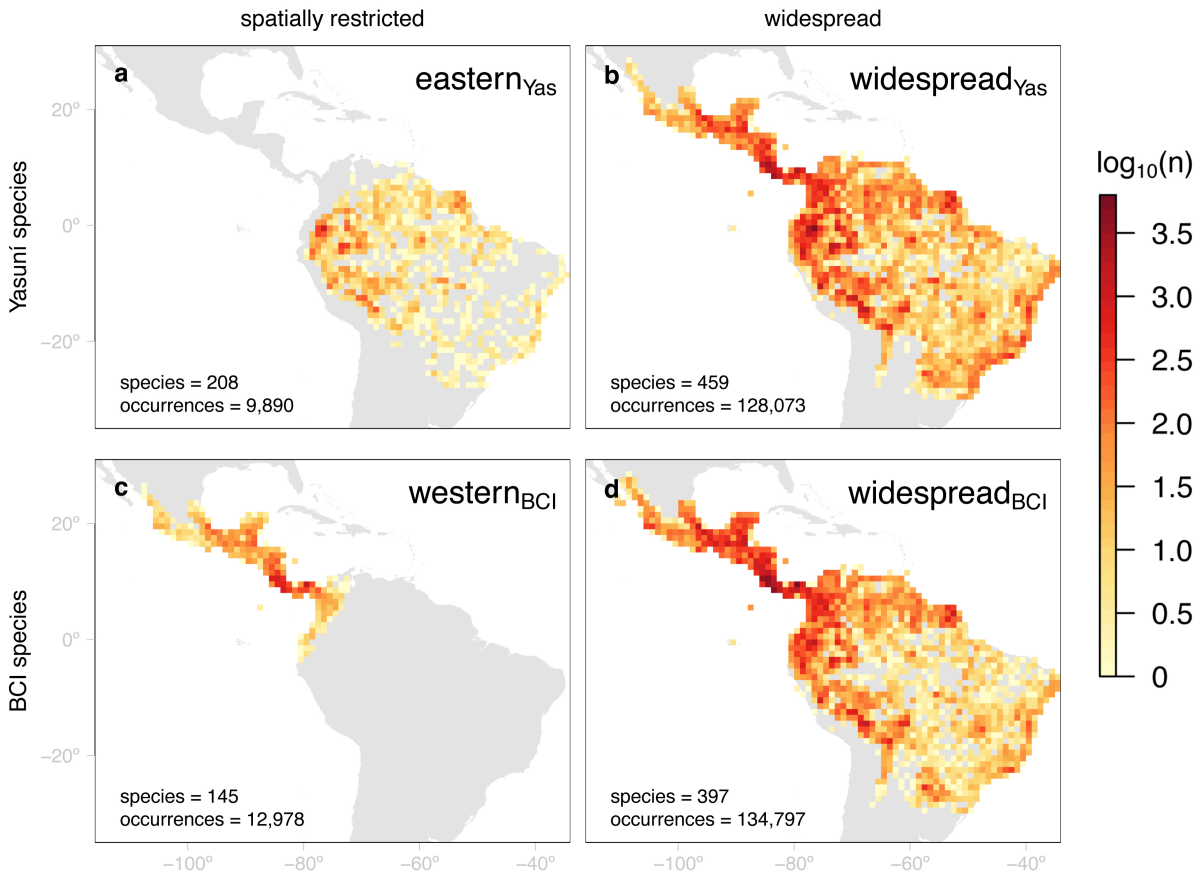

Figure 5.2. Biases in spatial distributions of occurrence records of Neotropical trees. Spatial distribution of occurrence records for species with each of the four types of biogeographic distributions considered in this study (see *Methods* for definitions). Colours depict the number of occurrence records (n) per pixel of 1º latitude by 1º longitude, and vary according to a log10 scale. Number of species and total number of occurrence records across all species are given for each biogeographic distribution. 120 species are represented in both panels (b) and (d) because they were present at both Yasuní and BCI. 143 species and 75 species were excluded for Yasuní and BCI, respectively, because there were only 1-2 occurrence records on the opposite side of the Andes such that biogeographic distributions could not confidently be assigned to these species. Some species are also present in the Caribbean (see Table D.2), but occurrences in the Caribbean were not included in any of our analyses. Note that the distribution of occurrence records reflects the combined effects of both spatial sampling bias (e.g., many pixels in the Brazilian Amazon have no occurrence records, likely due to insufficient sampling in this region), and the actual geographic distributions of the species in this study (e.g., large numbers of occurrence records are from the western Amazon and lower Central America, as would be expected given that these areas are geographically proximate to and have habitat similar to that of study sites Yasuní and Barro Colorado Island, respectively).

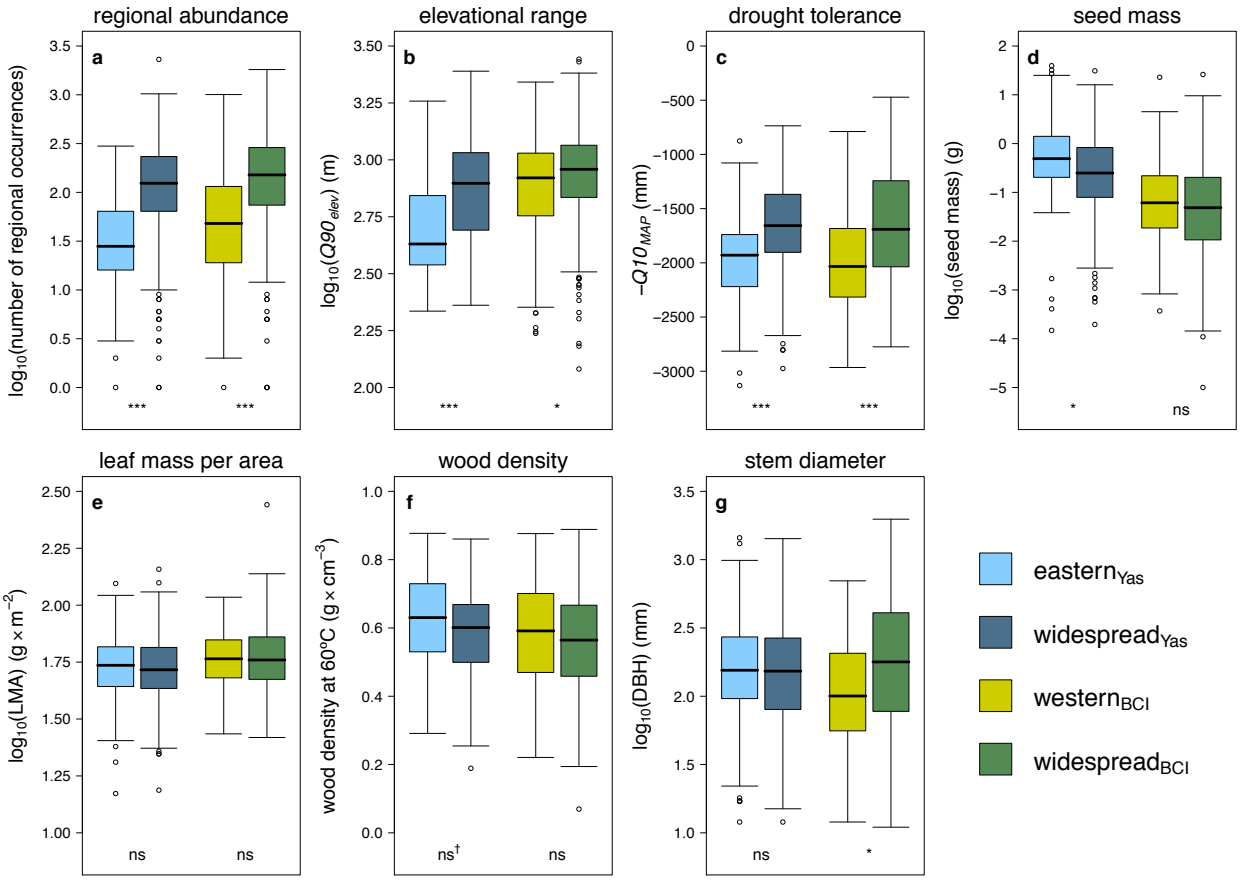

Figure 5.3. Species traits of Neotropical trees, by biogeographic distribution. Boxplots showing the distribution of traits among species with different biogeographic distributions for Yasuní (blue) and BCI (green). Adjusted *p*-values (ns, not significant, * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$) indicate whether each species trait at each community is a significant predictor of biogeographic distribution, from results of binary regressions that account for phylogenetic relatedness among taxa (for complete results, see Table D.4-D.5), after correction for multiple comparisons (Holm 1979; † indicates $p < 0.05$ prior to correction). For number of species included in each comparison see Table D.3.

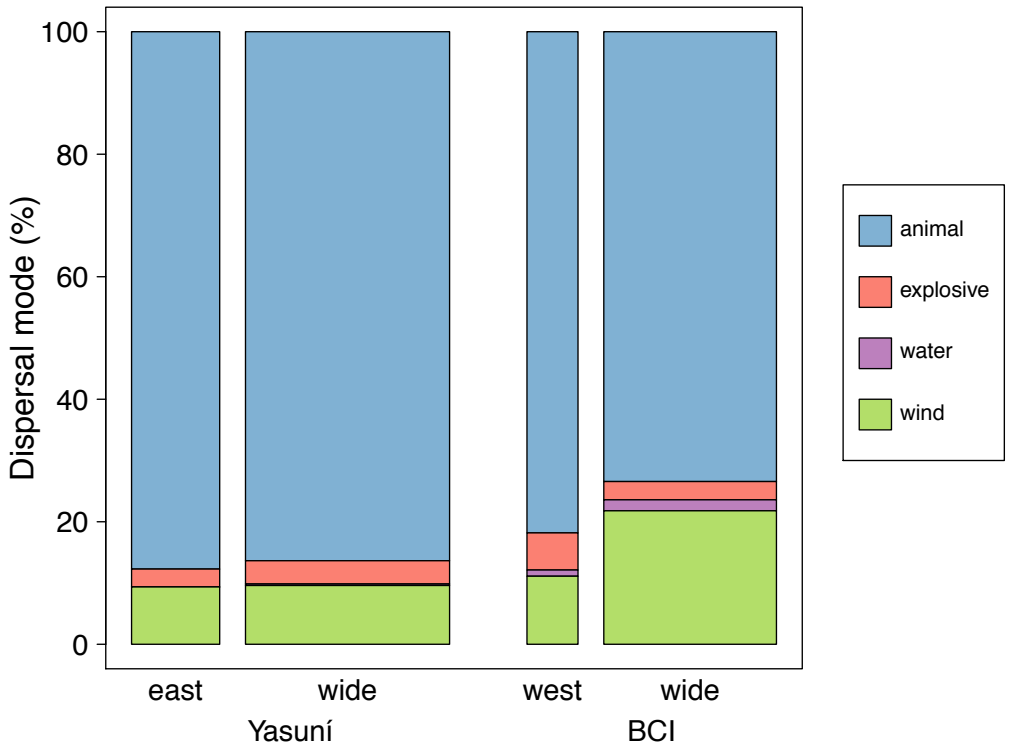

Figure 5.4. Dispersal modes of Neotropical trees, by biogeographic distribution. Dispersal modes of species with different biogeographic distributions (wide = widespread; east = east of the Andes; west = west of the Andes) at Yasuní and BCI. The width of each bar is proportional to the number of species in each biogeographic category. Relative to animal dispersal, wind dispersal is a significant predictor (p = 0.027) of widespread biogeographic distribution at BCI in binary regressions that account for phylogenetic relatedness among taxa (no longer significant after correction for multiple comparisons (Holm 1979), adjusted p = 0.11; for complete results, see Table D.4-D.5).

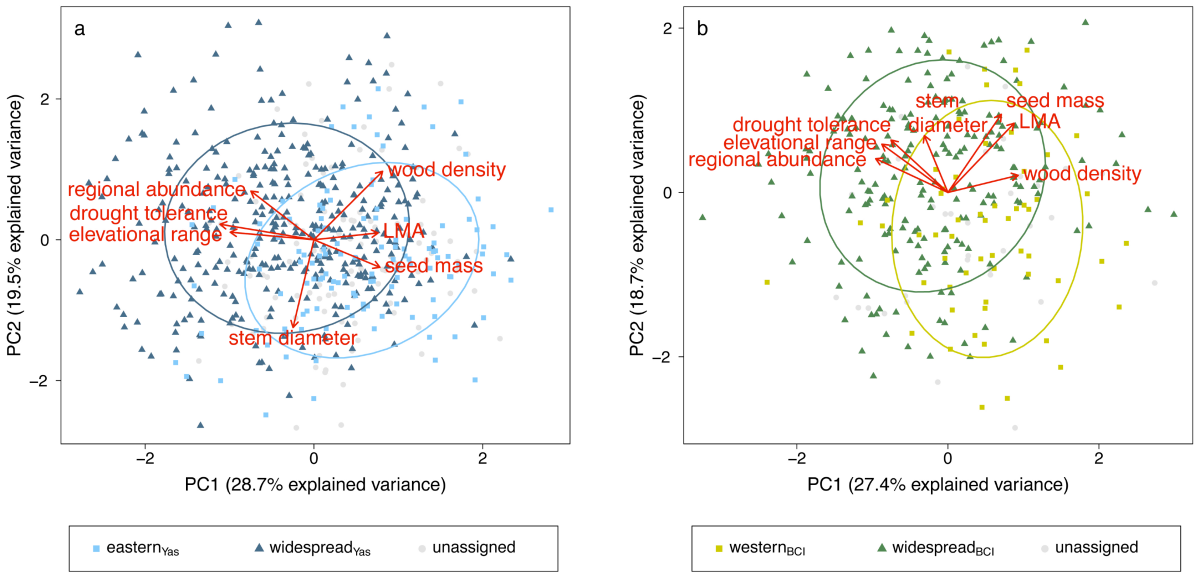

Figure 5.5. Clustering of Neotropical trees along PC trait axes, by geographic distribution. Biplot showing clustering of species with different biogeographic distributions along the first two principal component (PC) axes of variation in environmental and life-history traits at (a) Yasuní (number of species, n = 462) and (b) Barro Colorado Island (n = 260). Dots represent individual species, and normal ellipses contain ~68% of all species for each dispersal mode. Trait loadings on each PC axis are shown as red arrows.

APPENDIX D

Supplementary information from Chapter V

Supplementary methods

Assembling life-history traits

Leaf mass per area (LMA) for Yasuní species was obtained from Kraft and Ackerly (2010). Other life-history traits were assembled from unpublished datasets curated by N.C. Garwood and S.A. Queenborough (Yasuní: seed mass and dispersal mode), R. Valencia (Yasuní: maximum stem diameter), and S.J. Wright (Yasuní: wood density, see also Hietz *et al.* (2013); BCI: all life-history traits).

Seed mass and wood density were measured using dry material (for wood density specifically, wood was oven dried at a standardized 60ºC). Leaf mass per area was measured for entire leaves (or entire leaflets for very large compound leaves) growing in indirect light under a closed canopy. Maximum stem diameter at each site was defined as the 95th percentile of DBH of all stems (≥1 cm DBH) censused in each of the 50-ha plots. This method may substantially underestimate the maximum potential stem diameter of rare, shade-tolerant species represented only by juveniles in either of the plots. However, when comparing maximum stem diameter at BCI to maximum size categories known across Panama, we identified very few species likely to represent

severe underestimates relative to other species in the same size category (data not shown), and no significant correlation between maximum stem diameter and the number of stems in the plot (both traits log-transformed, Pearson's $r = 0.11$, $p = 0.061$). Thus, although maximum stem diameter may be underestimated for a few rare species, this is unlikely to be a major source of error. All analyses involving stem diameter (including principal component analyses) were conducted on non-climbing species only, because the interpretation of stem diameter may differ for climbers compared to non-climbers.

Dispersal mode was assessed at Yasuní at the species level when possible (Vásquez Martínez 1997, Pennington et al. 2004), or at the genus level (Cornejo and Janovec 2010) if species-level data were unavailable and variation was not known to exist within the genus. At BCI, dispersal mode was assessed based on fruit morphology (Croat 1978) and unpublished observations by O. Calderon, B. DeLeon, C.O. Handley Jr., and S.J. Wright. At both sites, finer details about putative animal dispersal vectors were available, but because many animal-dispersed tropical trees are dispersed by multiple vectors (Muller-Landau et al. 2008), it is often not possible to assign a single dispersal agent to a given species. Due to this difficulty as well as different methodologies used at Yasuní and BCI to determine animal vectors, we combined all animal-dispersed species into a single category. Preliminary analyses using more finely-delineated dispersal categories based on presumed primary animal vectors (Yasuní: terrestrial animals, as well as canopy animals dispersing small, medium, and large seeds; BCI: bats, birds, big birds, and mammals) did not reveal any significant effect of presumed primary animal dispersal vectors on biogeographic distribution (data not

shown), suggesting that the decision to collapse all animal dispersers into a single category was justified.

Table D.1. Characteristics of the 50-ha plots at study sites Yasuní and BCI. Characteristics of the 50-ha plots at Yasuní and Barro Colorado Island (BCI). Note that number of tree species refers to the total number of putative taxa within the 50-ha plots, whereas the number of woody plant species in this study includes all woody plants (including shrubs and climbers) for which valid species-level taxonomic identifications could be obtained. The list of woody plant species for BCI is much larger than the number of tree species because it also includes many species not found within the 50-ha plot, but found in similar nearby habitats. Plot data and tree species numbers are from http://www.ctfs.si.edu/plots/summary and climate data are from the WorldClim database (Hijmans et al. 2005).

	Yasuní	**BCI**
latitude (º)	-0.6859	9.1543
longitude (º)	-76.3953	-79.8461
elevation (m)	ca. 230	ca. 140
country	Ecuador	Panama
biogeographic realm	east of the Andes	west of the Andes
tree species (total, 50-ha plots only)	1,114	299
woody plant species (this study)	810	617
woody plant genera (this study)	370	344
woody plant families (this study)	79	77
mean annual precipitation (mm)	3,115	2,635
precipitation of the driest quarter (mm)	621	125

Table D.2. Geographic distributions of Neotropical trees from each study site. Geographic distributions of study species from Yasuní and Barro Colorado Island (BCI). (a) Biogeographic distribution of species in relation to the Andes (see Fig. 5.2). Note that unassigned species have at least one occurrence record on the opposite side of the Andes relative to Yasuní or BCI, but fewer occurrence records than the minimum threshold (3 unique records) required to classify a species as widespread (see *Methods*). (b) Presence in the Caribbean (≥3 occurrence records), according to biogeographic distribution in relation to the Andes. (c) Number and percent of species from each community inhabiting selected Neotropical ecoregions (Olson et al. 2001). Species are considered present in an ecoregion if they are represented by at least five unique occurrence records within the ecoregion.

a)	**Yasuní**		**BCI**	
Biogeographic distribution	**n**	**%**	**n**	**%**
unassigned	143	-	75	-
assigned				
eastern	208	31.2	-	-
western	-	-	145	26.8
widespread	459	68.8	397	73.2
total (assigned species only)	**667**	**100.0**	**542**	**100.0**

b)	**Yasuní**		**BCI**	
Presence in Caribbean	**n**	**%**	**n**	**%**
unassigned	143	-	75	-
assigned				
not present in Caribbean	605	90.7	428	81.7
present in Caribbean				
eastern	0	0.0	-	-
western	-	-	5	1.0
widespread	62	9.3	109	20.8
total (assigned species only)	**667**	**100.0**	**542**	**100.0**

c)	**Yasuní**		**BCI**	
Presence in ecoregions	**n**	**%**	**n**	**%**
tropical and subtropical moist broadleaf forests	800	98.8	600	97.2
Brazilian Atlantic Forest	134	16.5	108	17.5
tropical and subtropical dry broadleaf forests	236	29.1	330	53.5
tropical and subtropical grasslands, savannas and shrublands	190	23.5	162	26.3
Llanos	79	9.8	91	14.7
flooded grasslands and savannas	23	2.8	17	2.8
deserts and xeric shrublands	72	8.9	94	15.2
total (all species)	**810**	**100.0**	**617**	**100.0**

Table D.3. Number of species for which trait data were available for Neotropical trees. Number of species (n) of each biogeographic distribution included in phylogenetic regressions (Table 5.2; Tables D.4-D.5), for individual environmental and life-history traits (univariate models), and for all traits combined including imputed trait values (multivariate model). Note that this table reports number of species used in phylogenetic regressions, but occasionally there were a few additional species for which trait data were available, but phylogenetic information was lacking. These additional species were included in graphical depictions in Fig. 5.3-5.5 in order to visually represent all of the available data for each trait, but were not included in phylogenetic regressions or in the counts reported in this table.

	Yasuní			Barro Colorado Island		
Trait	**eastern$_{Yas}$**	**widespread$_{Yas}$**	**total**	**western$_{BCI}$**	**widespread$_{BCI}$**	**total**
univariate models						
regional abundance	207	456	663	143	391	534
elevational range	136	424	560	107	364	471
drought tolerance	136	424	560	107	364	471
seed mass	99	227	326	55	215	270
LMA	163	361	524	62	207	269
wood density	91	210	301	111	306	417
diameter	163	370	533	60	196	256
dispersal mode	170	396	566	97	329	426
PC1	123	339	462	58	202	260
PC2	123	339	462	58	202	260
multivariate model						
all traits (including imputed values)	109	338	447	53	200	253

Table D.4. Univariate models predicting biogeographic distributions of Yasuní trees. Regressions performed using a phylogenetic generalized linear mixed model for binary dependent traits (Ives and Helmus 2011, Ives and Garland Jr. 2014) for species from Yasuní. Trait estimates show the relationship between species traits and a widespread biogeographic distribution. The parameter σ^2 is a measure of phylogenetic signal in the phylogenetic variance-covariance matrix. Adjusted p-values for species traits are given after correction for multiple comparisons (Holm 1979). Statistical significance of adjusted p-values is indicated as follows: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$. R^2_{ce} for the overall models and partial R^2_{ce} for each trait are measures of explained variance, and were calculated following Ives (2017).

parameter	estimate	standard error	z-score	p-value	adjusted p		partial R^2_{ce}
regional abundance (n = 663, R^2_{ce} = 0.25)							
σ^2	2.36×10^{-11}			0.50			
intercept	-3.55	0.39	-9.03	$< 2.2 \times 10^{-16}$			
$\log_{10}$(number of regional occurrences)	2.45	0.22	11.04	$< 2.2 \times 10^{-16}$	2.20×10^{-15}	***	0.25
elevational range (n = 560, R^2_{ce} = 0.14)							
σ^2	0.05			0.35			
intercept	-9.47	1.40	-6.74	1.61×10^{-11}			
$\log_{10}(Q90_{elev})$ (m)	3.81	0.51	7.45	9.34×10^{-14}	7.47×10^{-13}	***	0.12
drought tolerance (n = 560, R^2_{ce} = 0.09)							
σ^2	1.90×10^{-13}			0.50			
intercept	4.40	0.50	8.83	$< 2.20 \times 10^{-16}$			
$Q10_{MAP}$ (mm)	1.79×10^{-3}	2.59×10^{-4}	6.92	4.43×10^{-12}	3.10×10^{-11}	***	0.07
seed mass (n = 326, R^2_{ce} = 0.04)							
σ^2	0.01			0.48			
intercept	0.60	0.14	4.22	2.49×10^{-5}			
$\log_{10}$(seed mass) (g)	-0.46	0.15	-3.04	2.36×10^{-3}	0.01	*	< 0.01
LMA (n = 524, R^2_{ce} = 0.03)							
σ^2	0.08			0.23			
intercept	1.78	1.19	1.50	0.13			
$\log_{10}$(LMA) (g m^{-2})	-0.57	0.68	-0.84	0.40	1.00		< 0.01

Table D.4 continued.

parameter	estimate	standard error	z-score	p-value	adjusted p		partial R^2_{ce}
wood density							
(n = 301, R^2_{ce} = 0.02)							
σ^2	6.22×10^{-15}			0.50			
intercept	2.12	0.63	3.38	7.32×10^{-4}			
wood density at 60ºC (g cm^{-3})	-2.13	1.01	-2.11	0.03	0.14		0.02
diameter							
(n = 533, R^2_{ce} = 0.01)							
σ^2	0.03			0.37			
intercept	1.08	0.55	1.96	0.05			
$\log_{10}$(DBH) (mm)	-0.12	0.25	-0.48	0.63	1.00		< 0.01
dispersal mode							
(n = 566, R^2_{ce} = 0.02)							
σ^2	0.04			0.36			
intercept	0.82	0.12	6.63	3.37×10^{-11}			
explosive	0.49	0.58	0.84	0.40	1.00		< 0.01
water	-	-	-	-	-		(all disp.
wind	0.02	0.32	0.07	0.94	-		modes)[1]
PC1							
(n = 462, R^2_{ce} = 0.20)							
σ^2	0.17			0.13			
intercept	1.21	0.20	6.21	5.32×10^{-10}			
PC1	-0.76	0.10	-7.56	4.18×10^{-14}	3.76×10^{-13}	***	0.16
PC2							
(n = 462, R^2_{ce} = 0.11)							
σ^2	0.24			0.06			
intercept	0.98	0.21	4.78	1.78×10^{-6}			
PC2	0.45	0.11	4.30	1.72×10^{-5}	1.03×10^{-4}	***	0.06

[1] Partial R^2_{ce} is calculated for the effect of all dispersal modes combined.

Table D.5. Univariate models predicting biogeographic distributions of BCI trees. Regressions performed using a phylogenetic generalized linear mixed model for binary dependent traits (Ives and Helmus 2011, Ives and Garland Jr. 2014) for species from Barro Colorado Island. Trait estimates show the relationship between species traits and a widespread biogeographic distribution. The parameter σ^2 is a measure of phylogenetic signal in the phylogenetic variance-covariance matrix. Adjusted p-values for species traits are given after correction for multiple comparisons (Holm 1979). Statistical significance of adjusted p-values is indicated as follows: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$. R^2_{ce} for the overall models and partial R^2_{ce} for each trait are measures of explained variance, and were calculated following Ives (2017).

parameter	estimate	standard error	z-score	p-value	adjusted p		partial R^2_{ce}
regional abundance (n = 534, R^2_{ce} = 0.19)							
σ^2	0.63			6.16×10^{-3}			
intercept	-1.79	0.73	-2.46	0.01			
$\log_{10}$(number of regional occurrences)	1.56	0.21	7.32	2.45×10^{-13}	2.45×10^{-12}	***	0.10
elevational range (n = 471, R^2_{ce} = 0.02)							
σ^2	0.06			0.43			
intercept	-2.55	1.45	-1.76	0.08			
$\log_{10}(Q90_{elev})$ (m)	1.29	0.49	2.62	8.84×10^{-3}	0.04	*	0.02
drought tolerance (n = 471, R^2_{ce} = 0.07)							
σ^2	5.65×10^{-13}			0.50			
intercept	3.46	0.48	7.27	3.68×10^{-13}			
$Q10_{MAP}$ (mm)	1.24×10^{-3}	2.45×10^{-4}	5.05	4.46×10^{-7}	4.01×10^{-6}	***	0.06
seed mass (n = 270, R^2_{ce} = 0.20)							
σ^2	2.19			2.92×10^{-3}			
intercept	1.31	1.15	1.14	0.26			
$\log_{10}$(seed mass) (g)	-0.01	0.18	-0.04	0.97	1		< 0.01
LMA (n = 269, R^2_{ce} = 0.08)							
σ^2	0.30			0.12			
intercept	0.95	1.91	0.50	0.62			
$\log_{10}$(LMA) (g m^{-2})	0.08	1.07	0.08	0.94	1		< 0.01

Table D.5 continued.

parameter	estimate	standard error	z-score	p-value	adjusted p		partial R^2_{ce}
wood density							
(n = 417, R^2_{ce} = 0.11)							
σ^2	1.00			0.01			
intercept	1.57	0.92	1.72	0.09			
wood density at 60ºC (g cm^{-3})	-1.05	0.83	-1.26	0.21	0.63		0.01
diameter							
(n = 256, R^2_{ce} = 0.10)							
σ^2	0.27			0.15			
intercept	-0.87	0.74	-1.18	0.24			
$\log_{10}$(DBH) (mm)	0.92	0.33	2.75	6.04×10^{-3}	0.04	*	0.03
dispersal mode							
(n = 426, R^2_{ce} = 0.15)							
σ^2	1.38			0.01			
intercept	1.11	0.92	1.21	0.23			
explosive	-0.52	0.58	-0.89	0.38	-		0.02
water	0.87	1.13	0.77	0.44	-		(all disp.
wind	0.89	0.40	2.22	0.03	0.11		modes)[1]
PC1							
(n = 260, R^2_{ce} = 0.14)							
σ^2	0.22			0.16			
intercept	1.30	0.24	5.34	9.13×10^{-8}			
PC1	-0.59	0.13	-4.59	4.39×10^{-6}	3.51×10^{-5}	***	0.09
PC2							
(n = 260, R^2_{ce} = 0.12)							
σ^2	0.13			0.28			
intercept	1.26	0.21	5.97	2.41×10^{-9}			
PC2	0.61	0.15	4.20	2.70×10^{-5}	1.89×10^{-4}	***	0.07

[1] Partial R^2_{ce} is calculated for the effect of all dispersal modes combined.

Table D.6. Geographic distributions by taxonomic family in Neotropical trees. Test of whether the proportion of widespread vs. exclusively eastern or western species differs among families at (a) Yasuní and (b) Barro Colorado Island. Results are from Pearson's chi-squared tests with p-values simulated by Monte Carlo simulation with 10^6 replicates (Yasuní: p = 0.12; BCI: p = 0.01) to account for low expected frequencies for many families. Complete residuals are not shown, but families with absolute values of residuals >1.96 are highlighted in bold and underlined, indicating statistically significant differences between expected and observed frequencies. Only families represented by at least five species were included for each community.

a)	$eastern_{Yas}$		$widespread_{Yas}$		all_{Yas} species	
	number of species	%	number of species	%	number of species	%
Annonaceae	**12**	**6.4**	**6**	**1.1**	**18**	**2.5**
Apocynaceae	2	1.1	9	1.7	11	1.5
Arecaceae	4	2.1	12	2.3	16	2.2
Bignoniaceae	3	1.6	14	2.6	17	2.4
Boraginaceae	2	1.1	6	1.1	8	1.1
Burseraceae	3	1.6	9	1.7	12	1.7
Celastraceae	0	0.0	5	0.9	5	0.7
Chrysobalanaceae	**6**	**3.2**	**4**	**0.8**	**10**	**1.4**
Clusiaceae	2	1.1	7	1.3	9	1.3
Combretaceae	4	2.1	3	0.6	7	1.0
Convolvulaceae	3	1.6	2	0.4	5	0.7
Dilleniaceae	1	0.5	5	0.9	6	0.8
Euphorbiaceae	4	2.1	17	3.2	21	2.9
Fabaceae	27	14.4	48	9.0	75	10.4
Lauraceae	14	7.4	25	4.7	39	5.4
Lecythidaceae	3	1.6	6	1.1	9	1.3
Malpighiaceae	6	3.2	7	1.3	13	1.8
Malvaceae	5	2.7	12	2.3	17	2.4
Melastomataceae	7	3.7	25	4.7	32	4.4
Meliaceae	6	3.2	18	3.4	24	3.3
Monimiaceae	2	1.1	3	0.6	5	0.7
Moraceae	11	5.9	23	4.3	34	4.7
Myristicaceae	2	1.1	9	1.7	11	1.5
Myrtaceae	4	2.1	10	1.9	14	1.9
Olacaceae	1	0.5	4	0.8	5	0.7
Piperaceae	1	0.5	8	1.5	9	1.3
Rhamnaceae	2	1.1	3	0.6	5	0.7
Rubiaceae	14	7.4	29	5.5	43	6.0
Salicaceae	**0**	**0.0**	**11**	**2.1**	**11**	**1.5**
Sapindaceae	6	3.2	12	2.3	18	2.5
Sapotaceae	9	4.8	20	3.8	29	4.0
Simaroubaceae	3	1.6	4	0.8	7	1.0
Solanaceae	4	2.1	7	1.3	11	1.5
Ulmaceae	0	0.0	5	0.9	5	0.7
Urticaceae	3	1.6	6	1.1	9	1.3
Verbenaceae	3	1.6	4	0.8	7	1.0
Violaceae	0	0.0	7	1.3	7	1.0
Total	**179**	**100.0**	**405**	**100.0**	**584**	**100.0**

Table D.6 continued.

b)	$western_{BCI}$		$widespread_{BCI}$		all_{BCI} species	
	number of species	%	number of species	%	number of species	%
Anacardiaceae	1	0.8	5	1.5	6	1.3
Annonaceae	**12**	**10.1**	**6**	**1.8**	**18**	**3.9**
Apocynaceae	3	2.5	6	1.8	9	2.0
Arecaceae	3	2.5	8	2.4	11	2.4
Bignoniaceae	2	1.7	20	5.9	22	4.8
Boraginaceae	1	0.8	5	1.5	6	1.3
Burseraceae	2	1.7	4	1.2	6	1.3
Celastraceae	1	0.8	4	1.2	5	1.1
Chrysobalanaceae	2	1.7	5	1.5	7	1.5
Clusiaceae	2	1.7	8	2.4	10	2.2
Combretaceae	1	0.8	4	1.2	5	1.1
Dilleniaceae	1	0.8	5	1.5	6	1.3
Euphorbiaceae	2	1.7	14	4.1	16	3.5
Fabaceae	15	12.6	53	15.7	68	14.9
Lauraceae	1	0.8	12	3.6	13	2.8
Malpighiaceae	5	4.2	7	2.1	12	2.6
Malvaceae	3	2.5	15	4.4	18	3.9
Melastomataceae	6	5.0	14	4.1	20	4.4
Meliaceae	1	0.8	12	3.6	13	2.8
Moraceae	4	3.4	19	5.6	23	5.0
Myristicaceae	7	5.9	8	2.4	15	3.3
Myrtaceae	**5**	**4.2**	**3**	**0.9**	**8**	**1.8**
Piperaceae	2	1.7	5	1.5	7	1.5
Polygonaceae	2	1.7	4	1.2	6	1.3
Rubiaceae	14	11.8	41	12.1	55	12.0
Salicaceae	4	3.4	9	2.7	13	2.8
Sapindaceae	3	2.5	23	6.8	26	5.7
Sapotaceae	4	3.4	8	2.4	12	2.6
Solanaceae	3	2.5	5	1.5	8	1.8
Urticaceae	4	3.4	3	0.9	7	1.5
Verbenaceae	3	2.5	3	0.9	6	1.3
Total	**119**	**100.0**	**338**	**100.0**	**457**	**100.0**

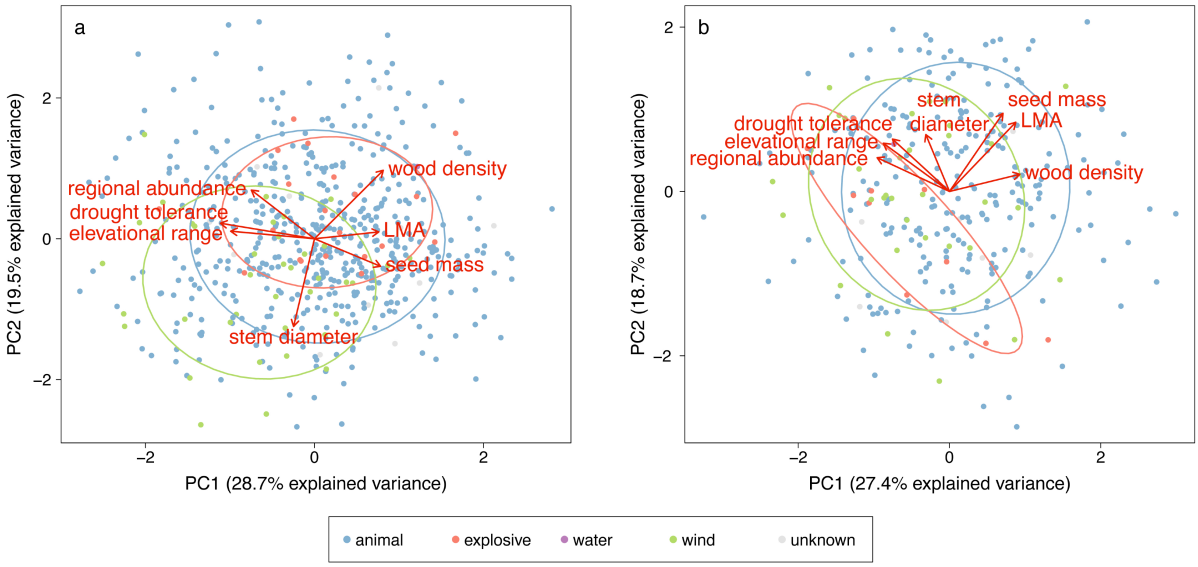

Figure D.1. Clustering of Neotropical trees along PC trait axes, by dispersal mode. Biplot showing clustering of species with different dispersal modes along the first two principal component (PC) axes of variation in environmental and life-history traits at (a) Yasuní (number of species, n = 462) and (b) Barro Colorado Island (n = 260). Dots represent individual species, and normal ellipses contain ~68% of all species for each dispersal mode. Trait loadings on each PC axis are shown as red arrows. Note that dispersal mode was not used as a trait in PC analyses, but is shown here for comparison with Fig. 5.5. The same underlying ordinations were used here as in Fig. 5.5; only the colours of the dots have been changed in order to represent dispersal mode instead of biogeographic distribution.

Supplementary literature cited

Cornejo, F. and Janovec, J. 2010. Seeds of Amazonian Plants. - Princeton University Press.

Croat, T. 1978. Flora of Barro Colorado Island. - Stanford University Press.

Hietz, P. et al. 2013. Strong radial variation in wood density follows a uniform pattern in two neotropical rain forests. - Funct. Ecol. 27: 684–692.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. - Int. J. Climatol. 25: 1965–1978.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. - Scand. J. Stat. 6: 65–70.

Ives, 2017. R^2s for correlated data: phylogenetic models, LMMs, and GLMMs. - bioRxiv doi: 10.1101/144170.

Ives, A. R. and Helmus, M. R. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. - Ecol. Monogr. 81: 511–525.

Ives, A. R. and Garland Jr., T. 2014. Phylogenetic regression for binary dependent variables. - In: Garamszegi, L. Z. (ed), Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology. Springer-Verlag, pp. 231–261.

Kraft, N. J. B. and Ackerly, D. D. 2010. Functional trait and phylogenetic tests of community assembly across spatial scales in an Amazonian forest. - Ecol. Monogr. 80: 401–422.

Muller-Landau, H. C. et al. 2008. Interspecific variation in primary seed dispersal in a tropical forest. - J. Ecol. 96: 653–667.

Olson, D. M. et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth. - Bioscience 51: 933–938.

Pennington, T. D. et al. 2004. Illustrated guide to the trees of Peru. - D. Hunt.

Vásquez Martínez, R. 1997. Flórula de las reservas biológicas de Iquitos, Perú (AR Lleras and CM Taylor, Eds.). - Missouri Botanical Garden Press.

CHAPTER VI

Conclusion

Synthesis and reflections

My dissertation has addressed the nature of species range shifts by pulling together diverse tools from population genomics, phylogeography, species distribution modelling, demographic and coalescent simulations, Bayesian statistics, and macroecology. As such, it has highlighted the interdisciplinary nature of historical biogeography, as well as the insight to be gained from novel methodological techniques and analysis of large datasets. While I initially set out to explore how historical climate change and geological activity have caused range shifts and driven spatial patterns in biodiversity, it became increasingly clear as my research progressed that accounting for Earth's history is necessary, but not sufficient, to explain biogeographic patterns. Instead, the emerging picture from the chapters of this dissertation is that biogeographic processes cannot be fully understood without due attention to species ecology. Instead of treating geoclimatic change as a proximal explanation for biodiversity patterns, deeper insight can be gained by considering the ways in which species individually experience biogeographic change in response to their unique ecological characteristics.

My work with hickories from eastern North America revealed that even species that are geographically co-distributed today may have experienced different phylogeographic histories. I found that *Carya cordiformis* and *Carya ovata* were likely both fairly widespread during the Last Glacial Maximum, but that the populations that made the largest genetic contribution to postglacial expansion differed between species. In *C. cordiformis*, a northern expansion origin was inferred from an area previously suspected to have contained cool, moist microrefugia for mesic temperate deciduous tree species (Delcourt et al., 1980), whereas a southern expansion origin was inferred for the more upland species *C. ovata.* These results add to the growing body of knowledge suggesting that northern microrefugia may have been important in the postglacial recolonization of some species (Provan & Bennett, 2008), and suggest that species-specific phylogeographic histories may partly explain the unresolved controversy over whether or not northern microrefugia generally existed (Tzedakis et al., 2013; Rull, 2014).

Rather than searching for ecological explanations for patterns *post hoc*, as was the case in my hickory work, I explicitly tested the role of ecological factors in mediating postglacial range shifts in canyon live oak from the California Floristic Province. This work revealed that changing summer drought conditions were likely a critical factor in mediating these shifts. In general, temperature is thought to be an important constraint on geographic ranges (Loehle, 1998) and is known to drive patterns of population-level local adaptation (Howe et al., 2003; Savolainen et al., 2007) in temperate trees. However, drought conditions may be more important than temperature for determining distributional limits of trees from Mediterranean environments that experience drought

stress during the growing season (Howe et al., 2003; Normand et al., 2009; Alberto et al., 2013). The results for canyon live oak support this hypothesis, and additionally suggest that long-term response to changing summer drought conditions has been an important driver of genetic structure.

Lastly, my work with Neotropical lowland rainforest woody plants expanded my exploration of ecological factors mediating biogeographic processes from the scale of individual species to that of entire communities. This work showed that species traits like elevational range and drought tolerance have impacted the probability of dispersal between the Amazon and Central America. This finding highlights the role of the northern Andes region as a biogeographic filter, and suggests that biogeographic filtering could have mediated regional community assembly on opposite sides of the Andes. The effects of biogeographic filtering were likely particularly important to assembly of forest communities in Central America, where numerous lowland tree species have an Amazonian origin (Gentry, 1982). However, much work remains to be done in order to test whether biogeographic filtering has caused Central American rainforests be enriched in species with weedy ecological traits associated with broad ranges and high capacity for biogeographic dispersal, relative to their Amazonian counterparts (Gentry, 1982; Dick et al., 2005).

Limitations and future directions

My dissertation has shown that changing geoclimatic landscapes may affect co-distributed species differently; that species-specific responses to geoclimatic change

may be mediated by ecological traits; and that individual responses among species scale up to create patterns across entire communities. Although I have focused primarily on attempting to understand the causes of range shifts and their genetic and biogeographic consequences, range shifts are far from the only way in which tree species respond to geoclimatic change. Instead of or in addition to migrating through space to track changing climates, tree populations may also genetically adapt to novel conditions, and those that fail to migrate or adapt risk going locally extinct (Aitken et al., 2008). Over evolutionary time scales, responses to geoclimatic change also drive speciation and extinction processes (Qian & Ricklefs, 2000; Badgley, 2010), which I have not considered in this dissertation. Thus, my dissertation is far from a complete account of the ways in which species respond to geoclimatic change and how this affects biodiversity patterns, but instead is narrowly focused on range shifts, which are only one aspect of a bigger picture.

Furthermore, despite my attempts to integrate ecological and historical perspectives on species range shifts, the ecological perspective of my dissertation has been extremely limited in scope. In particular, my work on hickories and canyon live oak failed to consider biotic interactions or life-history traits that might mediate range shifts, and instead focused on how species interact with the abiotic environment. My work on Neotropical trees did consider other ecological factors like functional and life-history traits, and biotic interactions (i.e., dispersal mode), but did not find that these factors were important to biogeographic filtering relative to the importance of interactions with the abiotic environment. Despite these findings, in other contexts, biotic interactions such as pollinator shifts or changes in interactions with herbivores or mutualists may

strongly influence how species respond to climate change (Van der Putten et al., 2010; HilleRisLambers et al., 2013; but see Katz & Ibáñez, 2016 for a counterexample), or may mediate climate-driven diversification processes (Kay et al., 2005; Becklin et al., 2016). However, the focus on abiotic factors in understanding range shifts is not a unique shortcoming of my own work, but rather reflects the emphasis currently employed in the field (Van der Putten et al., 2010). While historical responses to geoclimatic change may involve interactions with both the biotic and abiotic environment, our understanding of how the biotic environment affects biogeographic processes is still in its infancy.

The use of demographic and coalescent modelling techniques and Approximate Bayesian Computation to study species responses to geoclimatic change is also barely beyond its infancy (e.g., Beaumont et al., 2002; Carnaval et al., 2009; Knowles, 2009; He et al., 2013; Barthe et al., 2017). These methods allow statistical testing of a variety of very specific hypotheses, but only insofar as methodologies exist for modelling these hypotheses and simulating genetic expectations that can accurately distinguish among scenarios. Bio- and phylogeographers will need to continue developing creative methodologies to expand the scope of these models, in order to better test how species with differing life-history traits responded to historical geoclimatic change, and how responses might be mediated by biotic interactions. In addition to statistical phylogeography, macroecology is another promising approach to historical biogeography, although opinions differ as to whether macroecology is a distinct field or a subset of biogeography (Blackburn & Gaston, 2002; Fisher, 2002; Nee, 2002). Nonetheless, the advent of large datasets of species occurrence records (e.g., Global

Biodiversity Information Facility, www.gbif.org), fossil records (e.g., Neotoma, www.neotomadb.org), present-day and paleoclimates (Hijmans et al., 2005; Title & Bemmels, 2018), functional traits (Kattge et al., 2011), and phylogenetic relationships among species (Hinchliff et al., 2015), will also make macroecological approaches an increasingly feasible and exciting perspective in biogeography.

Range shifts past and future

Given that historical climate-driven range shifts were a central focus of my dissertation, it would be remiss not to at least briefly consider the implications of what was learned for conserving biodiversity and predicting range shifts in response to contemporary, anthropogenic climate change. Phylogeographic studies that characterize the geographic distribution of genetic diversity and identify refugial regions can inform conservation strategies (Petit et al., 2003; Hampe & Petit, 2005) and establish historical baselines for tree migration rates (Feurdean et al., 2013). However, I remain sceptical as to whether the lessons learned from the study of historical millennial-scale range shifts are likely to be directly relevant to understanding future range shifts over human time scales.

There may be isolated conclusions to be drawn. The work on hickories suggests that species-specific responses to climate change may result in non-analogue future communities. Such a situation would not be unusual, as non-analogue communities were common throughout eastern North America over much of the late Pleistocene and early Holocene (Jackson et al., 2000; Jackson & Williams, 2004). The work on canyon

live oak suggests that changing summer drought conditions might be more important than changing temperatures in influencing how ranges of Mediterranean-climate tree species will shift in response to climate change. However, these types of predictions may be better assessed with provenance trials that are already widely employed in forest genetics research (Langlet, 1971; Alberto et al., 2013). Finally, the work on Neotropical trees suggests that biogeographic barriers may become more or less pronounced for particular species as climates change. For example, the Andes may become a less pronounced barrier if warming climates allow lowland species to migrate substantially upslope (Weng et al., 2007) and to cross mountain passes. Such a scenario might potentially facilitate greater homogenization of the floras of Central America and the Amazon in the coming centuries and millennia.

However, in terms of more immediate responses to climate change, it is unclear to what extent the processes driving historical range shifts are relevant to understanding future range shifts. The present rate at which climates are warming far exceeds historical rates (IPCC, 2014), and the ecological context within which climate change is occurring dramatically differs from that of any historical time period. Human activities have greatly modified opportunities for dispersal of tree species by causing unprecedented habitat fragmentation of Earth's forests (Haddad et al., 2015); contributing to extinctions of Pleistocene megafauna (Koch & Barnosky, 2006), some of which were likely important dispersal agents of large-seeded trees (Pires et al., 2018); and creating new opportunities for global-scale dispersal and biological invasions (Hulme, 2009). It would be unwise to attempt to predict future range shifts, or mitigate their consequences through conservation actions, without considering the human-

mediated context within which they will occur. Far into the Anthropocene epoch, historical biogeographers of future millennia will have a much more complex task at hand, attempting to explain not only how “Earth and life evolved together” (Croizat, 1962, p. 46), but rather, the coevolution of Earth and life and human.

Literature cited

Aitken, S.N., Yeaman, S., Holliday, J.A., Wang, T., & Curtis-McLane, S. (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, **1**, 95–111.

Alberto, F.J., Aitken, S.N., Alía, R., González-Martínez, S.C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., & Savolainen, O. (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Badgley, C. (2010) Tectonics, topography, and mammalian diversity. *Ecography*, **33**, 220–231.

Barthe, S., Binelli, G., Hérault, B., Scotti-Saintagne, C., Sabatier, D., & Scotti, I. (2017) Tropical rainforests that persisted: inferences from the Quaternary demographic history of eight tree species in the Guiana shield. *Molecular Ecology*, **26**, 1161–1174.

Beaumont, M.A., Zhang, W., & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Becklin, K.M., Anderson, J.T., Gerhart, L.M., Wadgymar, S.M., Wessinger, C.A., & Ward, J.K. (2016) Examining plant physiological responses to climate change through an evolutionary lens. *Plant Physiology*, **172**, 635-649.

Blackburn, T.M. & Gaston, K.J. (2002) Macroecology is distinct from biogeography. *Nature*, **418**, 723.

Carnaval, A.C., Hickerson, M.J., Haddad, C.F.B., Rodrigues, M.T., & Moritz, C. (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

Croizat, L. (1962) *Space, Time, Form: the Biological Synthesis.* Caracas.

Delcourt, P.A., Delcourt, H.R., Brister, R.C., & Lackey, L.E. (1980) Quaternary

vegetation history of the Mississippi Embayment. *Quaternary Research*, **13**, 111–132.

Dick, C.W., Condit, R., & Bermingham, E. (2005) Biogeographic history and the high beta-diversity of Panamanian rainforest trees in Panama. *The Río Chagres, Panama: a multidisciplinary profile of a tropical watershed* (ed. by R.S. Harmon), pp. 259–269. Springer, Dordrecht.

Feurdean, A., Bhagwat, S.A., Willis, K.J., Birks, H.J.B., Lischke, H., & Hickler, T. (2013) Tree migration-rates: narrowing the gap between inferred post-glacial rates and projected rates. *PLoS ONE*, **8**, e71797.

Fisher, H.J. (2002) Macroecology: new, or biogeography revisited? *Nature*, **417**, 787.

Gentry, A.H. (1982) Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleisotcene fluctuations, or an accident of the Andean orogeny? *Annals of The Missouri Botanical Garden*, **69**, 557–593.

Haddad, N.M., Brudvig, L.A., Clobert, J., et al. (2015) Habitat fragmentation and its lasting impact on Earth's ecosystems. *Science Advances*, **1**, e1500052.

Hampe, A. & Petit, R.J. (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters*, **8**, 461–467.

He, Q., Edwards, D.L., & Knowles, L.L. (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

HilleRisLambers, J., Harsch, M.A., Ettinger, A.K., Ford, K.R., & Theobald, E.J. (2013) How will biotic interactions influence climate change-induced range shifts? *Annals of the New York Academy of Sciences*, **1297**, 112–125.

Hinchliff, C.E., Smith, S.A., Allman, J.F., et al. (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, **112**, 12764–12769.

Howe, G.T., Aitken, S.N., Neale, D.B., Jermstad, K.D., Wheeler, N.C., & Chen, T.H. (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, **81**, 1247–1266.

Hulme, P.E. (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, **46**, 10–18.

IPCC (2014) *Climate Change 2014: Synthesis Report. Contribution of Working Groups I,*

II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. (ed. by Core WorkingTeam, Pachauri, R.K., & Meyer, L.A.). IPCC, Geneva.

Jackson, S.T., Webb, R.S., Anderson, K.H., Overpeck, J.T., Webb III, T., Williams, J.W., & Hansen, B.C.S. (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews*, **19**, 489–508.

Jackson, S.T. & Williams, J.W. (2004) Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences*, **32**, 495–537.

Kattge, J., Díaz, S., Lavorel, S., et al. (2011) TRY - a global database of plant traits. *Global Change Biology*, **17**, 2905–2935.

Katz, D.S.W. & Ibáñez, I. (2016) Biotic interactions with natural enemies do not affect potential range expansion of three invasive plants in response to climate change. *Biological Invasions*, **18**, 3351–3363.

Kay, K.M., Reeves, P.A., Olmstead, R.G., & Schemske, D.W. (2005) Rapid speciation and the evolution of hummingbird pollination in neotropical *Costus* subgenus *Costus* (Costaceae): evidence from nrDNA its and ETS sequences. *American Journal of Botany*, **92**, 1899–1910.

Knowles, L.L. (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Koch, P.L. & Barnosky, A.D. (2006) Late Quaternary extinctions: state of the debate. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 215–250.

Langlet, O. (1971) Two hundred years genecology. *Taxon*, **20**, 653–721.

Loehle, C. (1998) Height growth rate tradeoffs northern determine and southern range limits for trees. *Journal of Biogeography*, **25**, 735–742.

Nee, S. (2002) Thinking big in ecology. *Nature*, **417**, 229–230.

Normand, S., Treier, U.A., Randin, C., Vittoz, P., Guisan, A., & Svenning, J.-C. (2009) Importance of abiotic stress as a range-limit determinant for European plants: insights from species responses to climatic gradients. *Global Ecology and Biogeography*, **18**, 437–449.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Müller-Starck, G., Demesure-Musch, B., Palmé, A., Martín, J.P., Rendell, S., & Vendramin, G.G. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Pires, M.M., Guimarães, P.R., Galetti, M., & Jordano, P. (2018) Pleistocene megafaunal extinctions and the functional loss of long-distance seed-dispersal services. *Ecography*, **41**, 153–163.

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

Qian, H. & Ricklefs, R.E. (2000) Large-scale processes and the Asian bias in temperate plant species diversity. *Nature*, **407**, 180–182.

Rull, V. (2014) Macrorefugia and microrefugia: a response to Tzedakis et al. *Trends in Ecology and Evolution*, **29**, 243–244.

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.

Title, P.O. & Bemmels, J.B. (2018) ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, **41**, 291–307.

Tzedakis, P.C., Emerson, B.C., & Hewitt, G.M. (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology and Evolution*, **28**, 696–704.

Van der Putten, W.H., Macel, M., & Visser, M.E. (2010) Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 2025–2034.

Weng, C., Hooghiemstra, H., & Duivenvoorden, J.F. (2007) Response of pollen diversity to the climate-driven altitudinal shift of vegetation in the Colombian Andes. *Philosophical Transactions of the Royal Society B*, **362**, 253–262.