Title:

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

Running title:

Phylogeography of North American *Carya*

Authors and affiliations:

Jordan B. Bemmels[1,*] and Christopher W. Dick[1,2]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA
[2] Smithsonian Tropical Research Institute, Apartado 0843–03092, Balboa, Ancon, Panama; orcid.org/0000-0001-8745-9137

[*] Corresponding author: jbemmels@umich.edu
orcid.org/0000-0001-9996-6996

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.

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

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 a large region of continuous habitat and recolonized northern areas in a single expanding front that encountered few migration barriers. More complex scenarios, such as forest refugia, need not be invoked to explain genetic structure. The genetically distinct Texas population of *Carya ovata* could represent a separate glacial refugium, but other explanations are possible. Relative to that of other temperate forest regions, the phylogeographic history of ENA may have been exceptionally simple, involving a northward range shift but without well defined refugia.

Keywords
eastern North America, forest refugia, isolation by distance, Last Glacial Maximum, phylogeography, range expansion, temperate trees

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; Gavin et al., 2014; Lumibao et al., 2017). Migrational responses to Pleistocene glaciation often left large impacts on the genetic structure of temperate species, but the severity and effects of glaciation varied in different 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) and where glaciation during the Last Glacial Maximum (LGM, ca. 21.5 ka; Jackson et al., 2000) was particularly severe, widespread temperate taxa typically retreated to distinct glacial refugia in Mediterranean regions (Hewitt, 1999, 2000; Petit et al., 2003). Genetic diversity was progressively lost due to founder effects as populations recolonized northern areas following glacial retreat (Hewitt, 1999, 2000), except in mid-latitude areas where admixture of lineages from distinct refugia has often led to elevated genetic diversity (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; Soltis et al., 2006; 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 LGM 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, fossil and genetic evidence

120 suggests that some temperate taxa were widespread over vast areas of the
121 southeastern United States during the LGM (Bennett, 1985; Magni et al., 2005;
122 McLachlan et al., 2005; Peterson & Graves, 2016; Lumibao et al., 2017).
123 Given the diversity of hypotheses that have been evoked to explain
124 phylogeographic patterns in ENA taxa, studies assessing genome-wide patterns of
125 genetic variation in widely distributed model species would provide valuable insight into
126 the history of temperate forests from the region as a whole. Surprisingly, we are aware
127 of no such studies (but Eckert et al. (2010) and Nadeau et al. (2015) have conducted
128 studies of more narrowly distributed tree species). Here, we use genome-wide genetic
129 variation to examine the phylogeographic history of two widespread, ENA tree species:
130 bitternut hickory (*Carya cordiformis* (Wagenh.) K.Koch) and shagbark hickory (*Carya*
131 *ovata* (Mill.) K.Koch). We construct and analyze single-nucleotide polymorphism (SNP)
132 datasets from nearly rangewide collections of each species and build paleodistribution
133 models in order to characterize geographic patterns of genetic diversity and
134 differentiation across ENA. In particular, we aim to determine if genetic structure is best
135 explained by recolonization from distinct forest refugia (and if so, where these refugia
136 where located), or by a single expansion from a large, continuous forest region.
137
138 **Methods**
139
140 *Study species*
141 *Carya cordiformis* and *Carya ovata* are wind-pollinated, animal-dispersed trees
142 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, making them excellent model taxa to study the phylogeography of eastern deciduous forests as a whole. *Carya ovata* additionally occurs in several small, disjunct populations in the Sierra Madre Oriental of northern Mexico (Little, 1971). In the northern half of temperate ENA, *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 (Graney, 1990). In southern areas both species are less common and typically restricted to wet, fertile soils (Graney, 1990; Smith, 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 in southern Tennessee (35ºN; Jackson et al., 2000), and trace amounts of pollen as far north as southern Illinois (39ºN; Grüger, 1972). However, the presence of small amounts of pollen in northern areas is not necessarily indicative of local presence, as pollen can be reworked from layers representing different time periods, or transported by wind over long distances (Tzedakis 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),

166 and one limitation of our analyses is that we are unable to assess patterns of
167 hybridization with other *Carya*. However, no stable hybrid zones exist in our species and
168 we consider it unlikely that occasional hybridization would systematically bias genetic
169 structure in a similar way across thousands of loci.
170
171 *DNA sampling and SNP genotyping*
172 Silica-dried leaf tissue was collected from 182 individuals of each species from
173 populations across ENA (Fig. 1; Tables S1.1-S1.2). Sampled individuals within each
174 population were separated by a minimum of 50 m (but sometimes up to dozens of km)
175 in order to minimize the chance of sampling siblings and other close relatives. Sample
176 size varied greatly among populations depending on the number of individuals meeting
177 these requirements that could be located (mean N = 7.8; Tables S1.1-S1.2). A
178 representative voucher specimen from each population was deposited in the University
179 of Michigan Herbarium (MICH; DOI: XXXXX).
180 DNA samples were extracted using Nucleospin Plant II extraction kits (*Macherey-*
181 *Nagel*; Düren, Germany), and libraries were prepared using a modified double digest
182 Restriction Associated DNA (ddRAD) sequencing protocol following Peterson et al.,
183 (2012), with restriction enzymes *EcoRI* and *MseI.* Full details of extraction methods and
184 library preparation are provided in Appendix S1 in Supporting Information. Seven
185 libraries of 72 samples each were sequenced at The Hospital for Sick Children (Toronto,
186 ON) on an *Illumina HiSeq* (Illumina; San Diego, CA) using single-end 50-bp sequencing.
187 In order to ensure adequate depth of coverage, at least one million raw reads per

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

190 Loci were identified and single nucleotide polymorphisms (SNPs) were
191 genotyped using STACKS v.1.44-v.1.46 (Catchen et al., 2011, 2013). Full details of SNP
192 discovery are provided in Appendix S1. After SNPs were successfully identified, one
193 SNP genotype per locus was exported from STACKS using the *populations* tool, retaining
194 only SNPs with a minimum genotyping rate of 75% (-r 0.75) and a minimum minor allele
195 frequency (MAF) of 3.3% (--min_maf 0.033), the lowest detectable MAF in at least one
196 population of each species, following Massatti & Knowles (2014). Minimum MAF is an
197 important parameter to consider because it can impact inference of genetic structure
198 (De la Cruz & Raska, 2014). We therefore explored preliminary analyses with minimum
199 MAF = 1% and 5%, but found that using the higher minimum MAF (5%) made little
200 qualitative difference in preliminary results. With the lower minimum MAF (1%), broad-
201 scale patterns were overwhelmed by local-scale signatures, likely due to rare variants
202 shared among closely related individuals (De la Cruz & Raska, 2014). These results
203 suggested that our choice of minimum MAF = 3.3% was appropriate.

204 In order to retain only putatively nuclear SNPs, we removed any SNPs from loci
205 that aligned with a maximum of two mismatches (-v 2) to the *Juglans regia*
206 (Juglandaceae) chloroplast genome (Genbank accession NC_028617.1) or the
207 *Cucurbita pepo* (Cucurbitaceae) mitochondrion genome (NC_014050.1) using BOWTIE
208 v.1.2 (Langmead et al., 2009). Extremely variable loci were also excluded as these may
209 represent locus assembly errors; we defined these loci as those with values of θ
210 (Watterson, 1975) above the 95th percentile, with θ calculated for each locus individually

211 using the *R* package 'pegas' v.0.10 (Paradis, 2010). Individual samples with unusually
212 high levels of missing data across all loci (based on visual inspection) were also
213 excluded.
214
215 *Genetic diversity and divergence*
216 Three genetic diversity parameters were calculated overall and for each
217 population: observed and expected heterozygosity (H_o and H_e, respectively), and
218 nucleotide diversity (π). Genetic differentiation (F_{ST}) (Nei, 1987) was calculated overall
219 and pairwise between each pair of populations. H_o, H_e, and F_{ST} were calculated in the *R*
220 package 'hierfstat' v.0.04-22 (Goudet, 2005), while π was calculated using *populations*
221 in STACKS v.1.46 (Catchen et al., 2011, 2013). Genetic diversity and differentiation
222 measures are not reported for populations represented by a single individual.
223
224 *Population genetic structure*
225 In order to test for isolation by distance (IBD), Mantel tests (Mantel, 1967) were
226 performed to assess the relationship between population pairwise F_{ST} values and
227 geographic distances. Principal component analysis (PCA) was used to investigate
228 genetic relationships among individuals and populations using the *dudi.pca* function in
229 the *R* package 'adegenet' v.2.0.1 (Jombart, 2008; Jombart & Ahmed, 2011). The NC_e
230 population (Table S1.1) was excluded for *C. cordiformis* because some, but not all
231 individuals from this population formed a distinct genetic cluster, which suggests that
232 several closely related individuals were unintentionally sampled and the high genetic
233 similarity between these individuals could have biased initial PCA results.

234 Genetic clusters were characterized using FASTSTRUCTURE v.1.0 (Raj et al.,
235 2014), with all populations and individuals included, using the recommended procedure
236 for detecting subtle genetic structure. Initially, the simple prior model was used and the
237 number of clusters (*K*) was varied from 1 to 6 for each species, and *K* was selected
238 using the *chooseK* tool in FASTSTRUCTURE. Then, FASTSTRUCTURE was rerun 100
239 times using the logistic prior model for the optimal value(s) of *K*, and final estimates of
240 genetic membership of individuals in each genetic cluster were obtained as the average
241 membership from the five runs with the highest likelihood, following Raj et al. (2014).
242 After investigating the broadest level of structure within each dataset, we reran
243 FASTSTRUCTURE on individual genetic clusters in order to test for substructure within
244 clusters.
245
246 *Paleodistribution modelling*
247 Species distribution models (SDMs) were constructed in order to predict the
248 potential distribution of each species during the Last Glacial Maximum (LGM; 21.5 ka).
249 Complete details of SDM construction and data sources are given in Appendix S1.
250 Briefly, occurrence records were obtained from the US Forest Service Forest Inventory
251 Analysis Database (O'Connell et al., 2012), while environmental variables were
252 obtained at 2.5-arcminute resolution from the WORLDCLIM v.1.4. (Hijmans et al., 2005)
253 and ENVIREM (Title & Bemmels, 2017) databases, for both current and LGM conditions.
254 SDMs were constructed using MAXENT v.3.4.1 (Phillips et al., 2004, 2006, 2017) in the *R*
255 package 'dismo' (Hijmans et al., 2015), with models optimized according to best
256 practices, following Title & Bemmels (2017).

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). Furthermore, genetic diversity estimates were generally uncorrelated with population sample size (Fig. S1.1), except that a negative relationship was found 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 expect that sample size might bias 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

280 values of H_e and π than typical of mid-latitude populations (Fig. 2), as expected for
281 range-edge populations (Jaramillo-Correa et al., 2009). Among-population genetic
282 differentiation (F_{ST}) is low in both species overall (*C. cordiformis*: 0.047; *C. ovata*: 0.038),
283 and among most pairs of populations (Tables S1.3-S1.4).
284
285 *Spatial genetic structure*
286 Spatial genetic structure in both species is very weak and dominated by a pattern
287 of isolation by distance (IBD). Mantel tests of IBD were statistically significant in both
288 species (*C. cordiformis*: r = 0.36, p = 0.0017; *C. ovata*: r = 0.47, p = 8.1 x 10^{-5}; Fig. 3).
289 Principal component analysis (PCA) also revealed an IBD-like pattern, without
290 clearly defined, distinct genetic clusters (Fig. 4). One exception to this pattern is that in
291 *C. ovata*, the Texas population (TX; bright pink dots, Fig. 4b) forms a separate cluster
292 that does not overlap with any other populations. However, the relative position of TX
293 along PC axes is still closest to that of geographically proximate western and southern
294 populations, consistent with an IBD pattern.
295 Lack of strong genetic structure was also suggested by FASTSTRUCTURE
296 results. Under the model with simple priors, the optimal number of genetic clusters was
297 *K* = 1 for both species. Under the model with logistic priors, which is more useful for
298 detecting subtle structure (Raj et al., 2014), optimal *K* ranged from 1-6. However, the
299 logistic priors model is prone to overfitting (Raj et al., 2014) and *K* > 2 did not produce
300 results that were biologically interpretable. We therefore note that *K* = 1 or 2 is likely the
301 optimal model complexity to explain genetic structure. In both species with *K* = 2,
302 genetic structure was weak and similar to an IBD-like pattern (Figs. 5, S1.2). In *C.*

303 *cordiformis*, a gradual north-south transition between clusters was evident. In *C. ovata*,
304 the transition between clusters was primarily from east to west. No substructure was
305 evident within any genetic cluster for any species, except that within the western cluster
306 for *C. ovata*, optimal K = 2 and the Texas population (TX) forms a distinct subcluster
307 relative to the other four populations (AR, IA, ON, WI; data not shown). However, the
308 grouping of these four populations into a distinct subcluster may be only a statistical
309 artifact reflecting the substantial additional membership of each these four western
310 populations in the main eastern cluster.
311
312 *Paleodistribution modelling*
313 The same four climatic variables were coincidentally retained in the SDMs for
314 both species: maximum temperature of the coldest month, potential evapotranspiration
315 of the warmest quarter, mean annual precipitation, and climatic moisture index (Table
316 S1.5). For both species, models were able to predict the current species distribution (Fig.
317 1) very well along the northern and western range edges, but performed more poorly at
318 delineating the southern range edge (Fig. 5). This poorer performance may reflect the
319 fact that both species are rare in the southern portion of their ranges, where presence
320 and absence may be determined more by soil type and topography (Graney, 1990;
321 Smith, 1990) than by broad-scale climatic differences among sites. For both species, a
322 large, continuous area of high LGM habitat suitability is predicted to have extended over
323 much of the southeastern US, from central Texas to exposed continental shelf off the
324 coast of North Carolina, and south to northern Florida (Fig. 5).
325

Discussion

Carya cordiformis and *Carya ovata* likely survived the LGM in a large, contiguous forest region covering much of the southeastern United States, and recolonized northern areas in a single expanding front. This scenario is supported by our genetic results showing weak genetic structure and an isolation-by-distance (IBD) pattern, and by our paleodistribution models and the fossil record. Genetic differentiation is continuous across ENA, except that a Texas population of *C. ovata* is more genetically distinct than other populations. The cause of this distinctiveness remains unclear. Insights from both species are likely applicable to understanding the phylogeographic history of temperate forests from ENA as a whole, and suggest that ENA may lack the complex refugial dynamics characteristic of other temperate forest regions of the world.

Weak genetic structure

In both species, genetic structure is weak and geographic patterns of genetic variation are primarily characterized by IBD, rather than sharp genetic breaks among regions. Although some geographic structure was detected with FASTSTRUCTURE, 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). We note that 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 only gradual geographic transitions between inferred

349 genetic clusters (Fig. 1). Similarly, substantial overlap among populations exists along
350 the first and second principal component axes (Fig. 4), with the exception of the Texas
351 population of *C. ovata* (see below). Rather than indicating the true presence of
352 biologically meaningful genetic clusters, our FASTSTRUCTURE results are likely a
353 statistical artifact of underlying IBD and suggest that over the majority of the species
354 range, genetic differentiation is continuous.

355 In addition to lack of phylogeographic breaks, there are no identifiable regions of
356 elevated genetic diversity (Fig. 2). Similar to many other temperate tree species from
357 ENA (Lumibao et al., 2017), *C. cordiformis* and *C. ovata* do not exhibit declines in
358 genetic diversity with increasing latitude, and population genetic diversity (H_e, π) is
359 relatively uniform across the species range. The absence of regions of elevated
360 diversity suggests either historical recolonization in a single, slowly expanding migration
361 front experiencing little loss of diversity during migration, or else very high gene flow
362 among populations (Jaramillo-Correa et al., 2009). Both of these situations are likely
363 applicable to our study species. In particular, low population genetic differentiation (*C.*
364 *cordiformis*, F_{ST} = 0.047; *C. ovata*, F_{ST} = 0.038) provides further evidence that gene flow
365 among populations is likely high. Low population differentiation is commonly observed in
366 widespread, wind-pollinated forest trees, due to their large population sizes and
367 capacity for long-distance pollen-mediated gene flow (Hamrick et al., 1992; Savolainen
368 et al., 2007; Alberto et al., 2013).

369 Despite generally uniform genetic diversity across populations in terms of H_e and
370 π, observed heterozygosity (H_o) increases with increasing latitude in both species (Fig.
371 2). Higher genetic diversity in northern regions might reflect the effects of larger

372 populations in the north that are less subject to genetic drift (Griffin & Barrett, 2004), but
373 genetic drift would be expected to simultaneously affect H_e and π (not only H_o). A more
374 likely explanation is that decreased H_o in the south reflects an increase in homozygotes
375 due to inbreeding in generally smaller, more isolated southern populations, but that this
376 effect has not yet led to a loss of overall genetic diversity at the population level (H_e, π).
377 We also note that far northern and far southern populations of both species
378 sometimes exhibit slightly lower population genetic diversity (H_e, π) than mid-latitude
379 populations (Fig. 2). In this case, we do suspect that this pattern reflects reduced gene
380 flow between core and peripheral populations and loss of diversity due to genetic drift in
381 generally smaller, more isolated peripheral populations (Hampe & Petit, 2005; Jaramillo-
382 Correa et al., 2009). However, in some European taxa, elevated mid-latitude genetic
383 diversity is believed to instead reflect historical secondary contact and admixture of
384 lineages from distinct southern refugia (Petit et al., 2003). Nonetheless, we find no
385 evidence of distinct refugia or mid-latitude admixture among lineages, making
386 secondary contact a less likely explanation for the patterns we observe.
387
388 *Temperate forests in ENA during the LGM*
389 The phylogeographic history of both *C. cordiformis* and *C. ovata* is best
390 characterized by simple latitudinal range shifts. Our genetic data and species
391 distribution models suggest that both species inhabited large areas of the southeastern
392 United States during the Last Glacial Maximum and gradually expanded northward to
393 occupy their current distribution as glaciers retreated. We find no evidence of highly
394 genetically distinct geographic regions or of strong phylogeographic breaks that would

395 suggest population fragmentation into multiple refugia (except possibly range-edge *C.*
396 *ovata* from Texas; see below). We also find no evidence that postglacial recolonization
397 of northern areas has occurred via multiple routes or been impeded by major landscape
398 barriers. LGM survival over large areas of the southeastern United States and gradual
399 northward recolonization is also compatible with the fossil record for the genus *Carya*
400 (Prentice et al., 1991; Jackson et al., 2000).

401 While we do not detect any signatures of complex refugial dynamics or multiple
402 postglacial recolonization routes, an alternative scenario we must consider is that one or
403 both species might have experienced a more complex phylogeographic history, the
404 signatures of which have been subsequently obscured by extensive contemporary gene
405 flow (e.g., He et al., 2013). Low population genetic differentiation and an IBD pattern
406 suggest that gene flow among populations is indeed fairly high. However, both species
407 are slow growing and long lived, with peak reproduction occurring in *C. cordiformis* from
408 ages 50 to 125 (Smith, 1990), and in *C. ovata* from ages 60 to 200 (Graney, 1990).
409 Relatively few generations have therefore passed since the LGM, meaning that there
410 has likely been insufficient time for gene flow to erode genetic signatures of postglacial
411 expansion from multiple refugia.

412 Given that *C. cordiformis* and *C. ovata* are common, widespread tree species
413 with a geographic distribution roughly matching that of temperate forests in ENA, we
414 consider them to be model taxa for understanding the phylogeographic history of
415 temperate forests from this region as a whole. Both species occupy a variety of habitat
416 sites, but do show relevant differences in their ecology (Graney, 1990; Smith, 1990). In
417 particular, *C. ovata* is a habitat generalist, and phylogeographic patterns for this species

418 are likely to be broadly representative of those for ENA forest taxa in general. In
419 contrast, *C. cordiformis* is a more mesic, bottomland species in many parts of its range,
420 and thus would be expected to be particularly susceptible to fragmentation into refugia
421 and therefore exhibit strong phylogeographic structure. However, low genetic structure
422 and lack of any evidence of distinct refugia in *C. cordiformis* provide particularly strong
423 evidence that temperate forests in ENA during the LGM were not highly fragmented.
424 Several other widespread woody plant species from ENA are also believed to
425 have expanded from a large area covering much of the southeastern United States,
426 including *Acer rubrum* (McLachlan et al., 2005), *Dirca palustris* (Peterson & Graves,
427 2016), *Fagus grandifolia* (Bennett, 1985; McLachlan et al., 2005), and *Quercus rubra*
428 (Magni et al., 2005). The presence of climatic conditions able to support temperate trees
429 over large areas of the southeastern United States during the LGM is also well
430 supported by pollen records (Prentice et al., 1991; Jackson et al., 2000; Williams, 2002).
431 Despite these sources of phylogeographic concordance, more complex scenarios
432 involving divisions into distinct refugia or multiple recolonization routes have been
433 proposed for many other plant species, especially those that are less geographically
434 widespread (e.g., Griffin & Barrett, 2004; Gonzales et al., 2008; Eckert et al., 2010;
435 Barnard-Kubow et al., 2015; Nadeau et al., 2015; Zinck & Rajora, 2016). In addition,
436 many forest communities with no modern analogue were present in ENA during the
437 LGM (Jackson et al., 2000), suggesting that not all taxa responded to glaciation in the
438 same way. It is therefore unclear whether any general migrational responses to
439 glaciation are applicable to ENA forests, or whether most taxa exhibited species-specific,
440 idiosyncratic responses.

441 One possible explanation for different responses among species is that suitable
442 habitat for species adapted to a narrower range of climatic or edaphic conditions could
443 have been more geographically fragmented during the LGM than for more widespread
444 species. In particular, ENA taxa with a strictly southern, warm-temperate distribution
445 may have become fragmented into distinct far-southern refugia in Florida, Texas, or
446 along the Gulf or Atlantic Coasts (e.g., Gonzales et al., 2008; Eckert et al., 2010). In
447 contrast, more widespread species such as *C. cordiformis* and *C. ovata* could have
448 survived in more expansive inland areas of cool-temperate conditions that extended
449 farther north (e.g., Fig. 5). As model taxa, *C. cordiformis* and *C. ovata* may be most
450 appropriate for understanding the phylogeographic history of temperate forests as a
451 whole, and may be less representative of taxa with widely differing habitat requirements
452 and narrower ecological amplitudes.
453 Nonetheless, our results suggest that temperate forests from ENA may have
454 experienced a relatively simple phylogeographic history. Previous phylogeographic
455 studies have established that major rivers and the Appalachian Mountains are important
456 phylogeographic barriers for some ENA taxa, but these findings are not universal and
457 are typically found in small animals with limited dispersal ability (Soltis et al., 2006). In
458 contrast, rivers are unlikely to present substantial dispersal barriers to large trees, and
459 the phylogeographic impact (or lack thereof) of the Appalachians on tree populations
460 remains unclear (Jaramillo-Correa et al., 2009). Overall, there are few topographic or
461 climatic barriers likely to lead to complex phylogeographic histories for ENA forest trees,
462 especially in comparison to other temperate regions of the world (Hewitt, 2000; Shafer
463 et al., 2010; Qiu et al., 2011).

464

465 *Genetic distinctiveness of Texas shagbark hickory*

466 Although an IBD pattern characterizes genetic structure throughout most of ENA
467 in both species, the Texas population (TX) of *C. ovata* presents an exception to this
468 general pattern. TX forms its own PCA genetic cluster that does not overlap with any
469 other population (pink dots; Fig. 4b), and it is the only population with membership
470 almost exclusively in the western cluster identified by FASTSTRUCTURE.

471 There are several possible interpretations of these results. One possibility is that
472 TX may be derived from a separate glacial refugium. Glacial refugia have previously
473 been inferred in Texas and northern Mexico for several southern *Pinus* and *Prunus*
474 species (Schmidtling & Hipkins, 1998; Schmidtling, 2003; Shaw & Small, 2005;
475 Jaramillo-Correa et al., 2009; Eckert et al., 2010). Alternatively, ancestors of the TX
476 population may have experienced gene flow with *C. ovata* populations in the mountains
477 of northern Mexico (Little, 1971). Although we have not included any high-elevation
478 Mexican populations in our genetic analyses or paleodistribution models, we cannot
479 exclude the possibility that these populations may have migrated to lower elevations
480 and come into contact with other populations during the LGM or at another time during
481 the Pleistocene.

482 However, we do not necessarily need to invoke a separate refugium or contact
483 with Mexican populations to explain the genetic distinctiveness of the range-edge TX
484 population, because our results could also merely reflect low gene flow between TX and
485 other populations. There is likely very little gene flow between TX and populations we
486 sampled further to the east due to the absence of *C. ovata* in the Lower Mississippi

487 River Valley (Fig. 1). In addition, the Ozark and Ouachita Mountains could conceivably
488 reduce gene flow with populations sampled to the north. Denser population sampling
489 across the southwestern portion of the species range could help fill in geographic gaps
490 in our PC analyses, in order to determine whether TX is truly a distinct gene pool, or
491 represents the extreme end of a rangewide IBD pattern, but is subject to greater genetic
492 isolation than other populations.

493 Given the uncertain origin of the Texas population of *C. ovata*, we suggest that
494 any conservation efforts in this species should ensure inclusion of populations from
495 Texas and surrounding regions. Genetically distinct southern populations of temperate
496 species have often been identified as high conservation priority (Petit et al., 2003;
497 Hampe & Petit, 2005; Médail & Diadema, 2009). However, across most of the range of
498 *C. cordiformis* and *C. ovata*, we see little reason to prioritize conservation of southern
499 populations, due to their lack of genetic distinctiveness and lack of elevated genetic
500 diversity. On the other hand, most temperate tree species exhibit geographically
501 structured climatically adaptive genetic variation (Savolainen et al., 2007; Aitken &
502 Bemmels, 2016) unlikely to be captured by the putatively neutral SNP markers we
503 employed. Conserving populations from a variety of climates across the species range
504 would therefore likely maximize conservation of adaptively relevant genetic diversity.
505

506 *Conclusions*

507 Genome-wide patterns of genetic variation, predictions of paleodistibution
508 models, and the fossil record all suggest that *C. cordiformis* and *C. ovata* recolonized
509 postglacial ENM from a single, continuous region of temperate forest that likely covered

510 much of the southeastern United States during the LGM. We generally find no evidence
511 of distinct glacial refugia or strong phylogeographic breaks. However, as an exception to
512 this overall pattern, it is unclear whether the higher genetic distinctiveness of a Texas
513 population of *C. ovata* relative to other populations reflects the effects of a separate
514 Texas refugium, historical contact with Mexican populations, or contemporary patterns
515 of gene flow. The two *Carya* species we have studied represent excellent model taxa for
516 understanding the phylogeographic history of eastern deciduous forests as a whole.
517 Whereas complex refugial dynamics must be invoked to explain genetic structure in
518 other temperate regions of the world, our results suggest that the phylogeographic
519 history of temperate forests from ENA can largely be explained by simple latitudinal
520 range shifts, without division into distinct refugia.
521

522 **References**

523

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

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

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

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

535 Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., & Cresko, W.A. (2013)
536 STACKS: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–
537 3140.

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

541 Eckert, A.J., Van Heerwaarden, J., Wegrzyn, J.L., Nelson, C.D., Ross-Ibarra, J.,
542 González-Martínez, S.C., & Neale, D.B. (2010) Patterns of population structure and
543 environmental associations to aridity across the range of loblolly pine (*Pinus taeda*
544 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.

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.

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.

Grüger, E. (1972) Late quaternary vegetation development in south-central Illinois. *Quaternary Research*, **2**, 217–220.

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., 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., 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.

591 De la Cruz, O. & Raska, P. (2014) Population structure at different minor allele
592 frequency levels. *BMC proceedings*, **8**, S55.
593 Langmead, B., Trapnell, C., Pop, M., & Salzberg, S.L. (2009) Ultrafast and memory-
594 efficient alignment of short DNA sequences to the human genome. *Genome*
595 *Biology*, **10**, R25.
596 Little, E.L. (1971) *Atlas of United States trees, volume 1. Conifers and important*
597 *hardwoods.* USDA Miscellaneous Publication 1146, Washington, DC.
598 Lumibao, C.Y., Hoban, S.M., & McLachlan, J. (2017) Ice ages leave genetic diversity
599 "hotspots" in Europe but not in Eastern North America. *Ecology Letters*, in press.
600 Magni, C.R., Ducousso, A., Caron, H., Petit, R.J., & Kremer, A. (2005) Chloroplast DNA
601 variation of *Quercus rubra* L. in North America and comparison with other
602 Fagaceae. *Molecular Ecology*, **14**, 513–524.
603 Mantel, N. (1967) The detection of disease clustering and a generalized regression
604 approach. *Cancer Research*, **27**, 209–220.
605 Massatti, R. & Knowles, L.L. (2014) Microhabitat differences impact phylogeographic
606 concordance of codistributed species: genomic evidence in montane sedges
607 (*Carex* L.) from the Rocky Mountains. *Evolution*, **68**, 2833–2846.
608 McCarthy, D.M. & Mason-Gamer, R.J. (2016) Chloroplast DNA-based phylogeography
609 of *Tilia americana* (Malvaceae). *Systematic Botany*, **41**, 865–880.
610 McLachlan, J.S., Clark, J.S., & Manos, P.S. (2005) Molecular indicators of tree
611 migration capacity under rapid climate change. *Ecology*, **86**, 2088–2098.
612 Médail, F. & Diadema, K. (2009) Glacial refugia influence plant diversity patterns in the
613 Mediterranean Basin. *Journal of Biogeography*, **36**, 1333–1345.

614 Meirmans, P.G. (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**,
615 2839–2846.
616 Morris, A.B., Graham, C.H., Soltis, D.E., & Soltis, P.S. (2010) Reassessment of
617 phylogeographical structure in an eastern North American tree using Monmonier's
618 algorithm and ecological niche modelling. *Journal of Biogeography*, **37**, 1657–1667.
619 Nadeau, S., Godbout, J., Lamothe, M., Gros-Louis, M.-C., Isabel, N., & Ritland, K.
620 (2015) Contrasting patterns of genetic diversity across the ranges of *Pinus*
621 *monticola* and *P. strobus*: a comparison between eastern and western North
622 American postglacial colonization histories. *American Journal of Botany*, **102**,
623 1342–1355.
624 Nazareno, A.G., Bemmels, J.B., Dick, C.W., & Lohmann, L.G. (2017) Minimum sample
625 sizes for population genomics: an empirical study from an Amazonian plant species.
626 *Molecular Ecology Resources*, in press.
627 Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.
628 O'Connell, B.M., LaPoint, E.B., Turner, J.A., Ridley, T., Boyer, D., Wilson, A.M.,
629 Waddell, K.L., & Conkling, B.L. (2012) *The Forest Inventory and Analysis*
630 *Database: database description and users manual version 5.1.4 for Phase 2. Gen.*
631 *Tech. Rep.,* US Department of Agriculture, Forest Service, Washington, DC.
632 Paradis, E. (2010) Pegas: an R package for population genetics with an integrated-
633 modular approach. *Bioinformatics*, **26**, 419–420.
634 Peterson, B.J. & Graves, W.R. (2016) Chloroplast phylogeography of *Dirca palustris* L.
635 indicates populations near the glacial boundary at the Last Glacial Maximum in
636 eastern North America. *Journal of Biogeography*, **43**, 314–327.

637 Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., & Hoekstra, H.E. (2012) Double
638 digest RADseq: an inexpensive method for de novo SNP discovery and genotyping
639 in model and non-model species. *PLoS ONE*, **7**, .
640 Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R.,
641 Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Müller-Starck, G.,
642 Demesure-Musch, B., Palmé, A., Martín, J.P., Rendell, S., & Vendramin, G.G.
643 (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*,
644 **300**, 1563–1565.
645 Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., & Blair, M.E. (2017) Opening
646 the black box: an open-source release of MAXENT. *Ecography*, **40**, 887–893.
647 Phillips, S.J., Anderson, R.P., & Schapire, R.E. (2006) Maximum entropy modeling of
648 species geographic distributions. *Ecological Modelling*, **190**, 231–259.
649 Phillips, S.J., Dudík, M., & Schapire, R.E. (2004) A maximum entropy approach to
650 species distribution modeling. *Proceedings of the Twenty-First International*
651 *Conference on Machine Learning*, **2004**, 655–662.
652 Prentice, C., Bartlein, P.J., & Webb III, T. (1991) Vegetation and climate change in
653 eastern North America since the Last Glacial Maximum. *Ecology*, **72**, 2038–2056.
654 Provan, J. & Bennett, K.D. (2008) Phylogeographic insights into cryptic glacial refugia.
655 *Trends in Ecology and Evolution*, **23**, 564–571.
656 Qiu, Y.X., Fu, C.X., & Comes, H.P. (2011) Plant molecular phylogeography in China
657 and adjacent regions: tracing the genetic imprints of Quaternary climate and
658 environmental change in the world's most diverse temperate flora. *Molecular*
659 *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). **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)

683 Comparative phylogeography of unglaciated eastern North America. *Molecular*
684 *Ecology*, **15**, 4261–4293.
685 Stewart, J.R. & Lister, A.M. (2001) Cryptic northern refugia and the origins of the
686 modern biota. *Trends in Ecology and Evolution*, **16**, 608–613.
687 Thomson, A.M., Dick, C.W., & Dayanandan, S. (2015) A similar phylogeographical
688 structure among sympatric North American birches (*Betula*) is better explained by
689 introgression than by shared biogeographical history. *Journal of Biogeography*, **42**,
690 339–350.
691 Title, P.O. & Bemmels, J.B. (2017) ENVIREM: An expanded set of bioclimatic and
692 topographic variables increases flexibility and improves performance of ecological
693 niche modeling. *Ecography*, in press.
694 Tzedakis, P.C., Emerson, B.C., & Hewitt, G.M. (2013) Cryptic or mystic? Glacial tree
695 refugia in northern Europe. *Trends in Ecology and Evolution*, **28**, 696–704.
696 Watterson, G.A. (1975) On the number of segregating sites in genetical models without
697 recombination. *Theoretical Population Biology*, **7**, 256–276.
698 Williams, J.W. (2002) Variations in tree cover in North America since the last glacial
699 maximum. *Global and Planetary Change*, **35**, 1–23.
700 Willing, E.M., Dreyer, C., & van Oosterhout, C. (2012) Estimates of genetic
701 differentiation measured by F_{ST} do not necessarily require large sample sizes when
702 using many SNP markers. *PLoS ONE*, **7**, 1–7.
703 Willis, K.J. & Van Andel, T.H. (2004) Trees or no trees? The environments of central
704 and eastern Europe during the Last Glaciation. *Quaternary Science Reviews*, **23**,
705 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.

Data accessibility

SNP datasets and species occurrence records are available for download from the Dryad Digital Repository (DOI: XXXXX).

Biosketches

Jordan Bemmels is interested in biogeography of temperate and tropical trees and this work is part of his PhD thesis. Christopher Dick is his thesis advisor, and is broadly interested in tropical tree biogeography and evolution.

Author contributions: J.B.B. and C.W.D. conceived the project; J.B.B. performed fieldwork and labwork, and analyzed the data; J.B.B. wrote the manuscript with input from C.W.D.

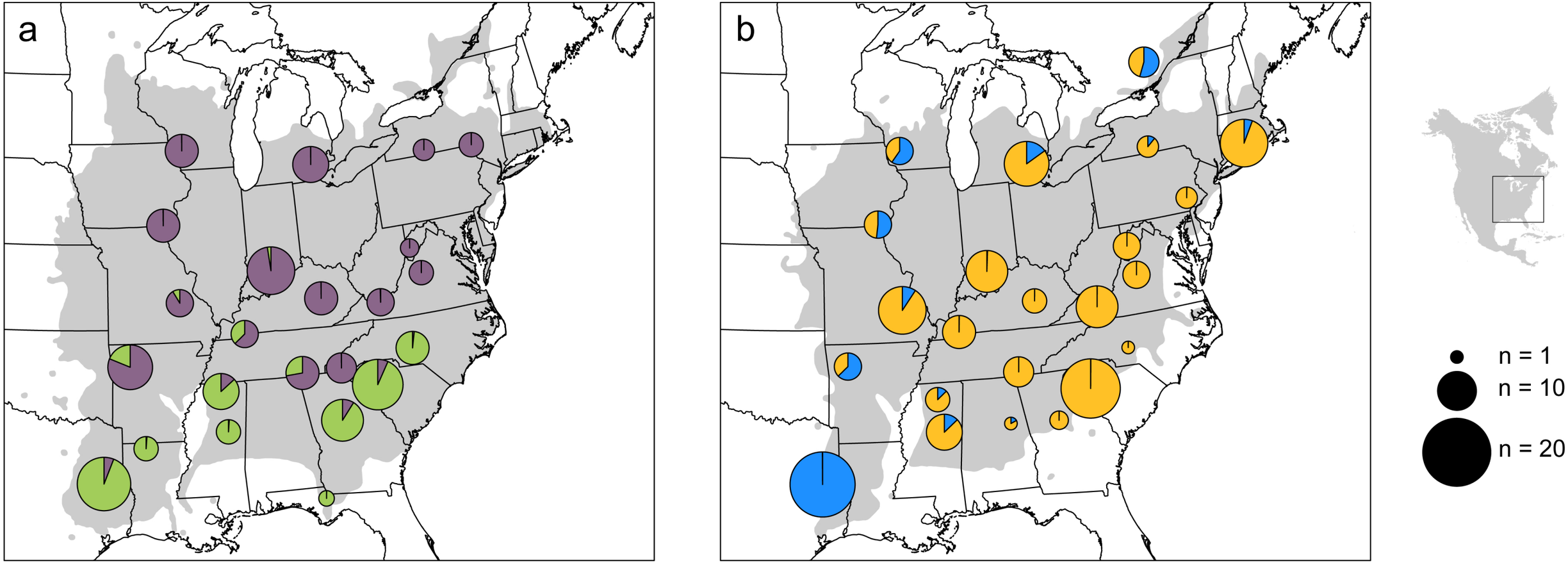
Figure 1. Membership of (a) *Carya cordiformis* and (b) *Carya ovata* populations in genetic clusters (different colours on pie charts; *K* = 2) identified using FASTSTRUCTURE (Raj et al., 2014). The geographic distribution of each species is shown in grey (Little, 1971), and the sample size (n) of each 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.

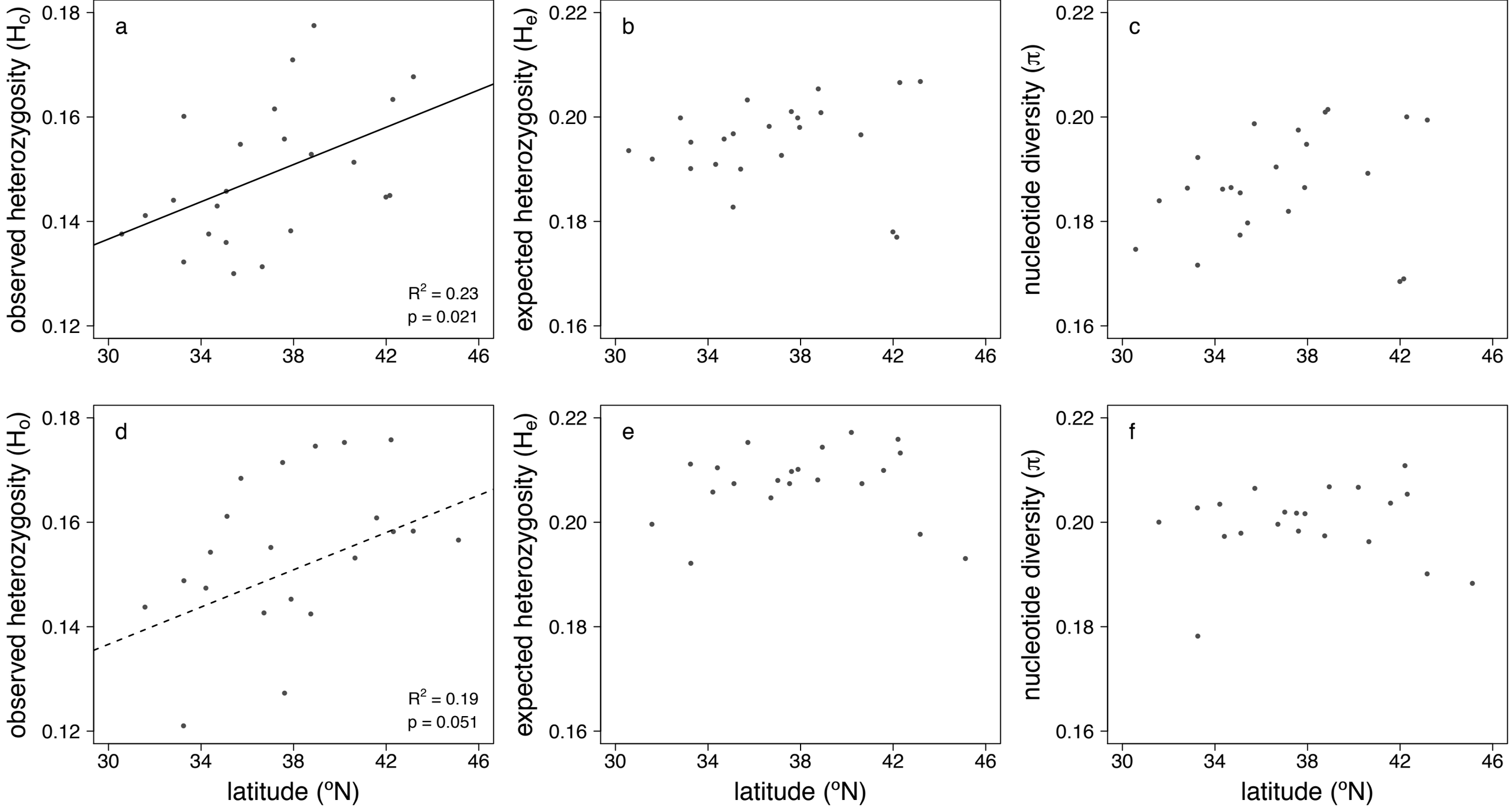
Figure 2. Genetic diversity vs. latitude for populations of *Carya cordiformis* (a-c) and *Carya ovata* (d-f). Statistically significant relationships are portrayed as solid lines, and marginally significant relationships as dashed lines.

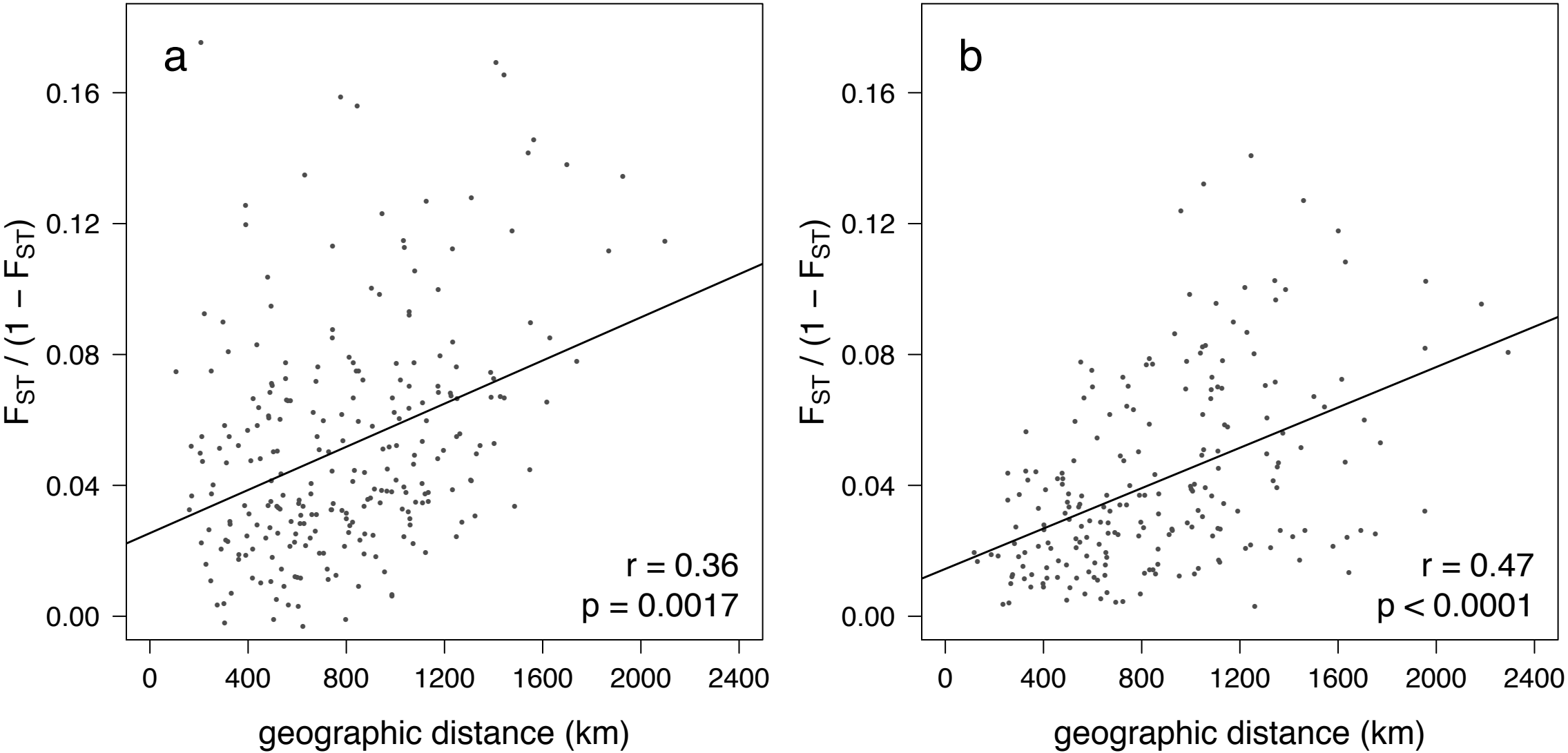
Figure 3. Mantel test showing isolation by distance in (a) *Carya cordiformis* and (b) *Carya ovata*. Each dot represents a pair of populations and the y-axis is a measure of genetic differentiation between populations.

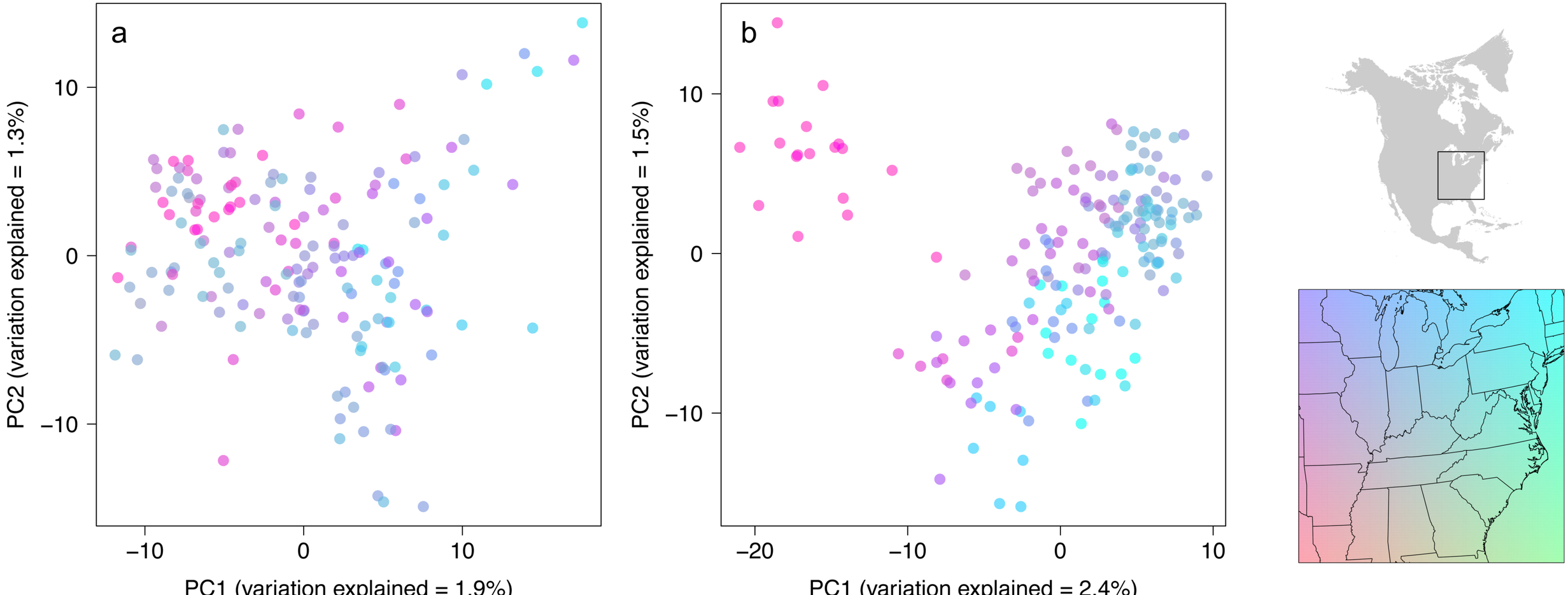
Figure 4. Clustering of individuals along the first and second principal component (PC) axes of genetic variation in (a) *Carya cordiformis* and (b) *Carya ovata.* Each dot represents a single individual, and colours correspond to the geographic location of the individual (shown in the inset map of eastern North America). Note that the Texas population (TX) of *C. ovata* discussed in the manuscript is represented by the pink dots in the upper left corner of (b).

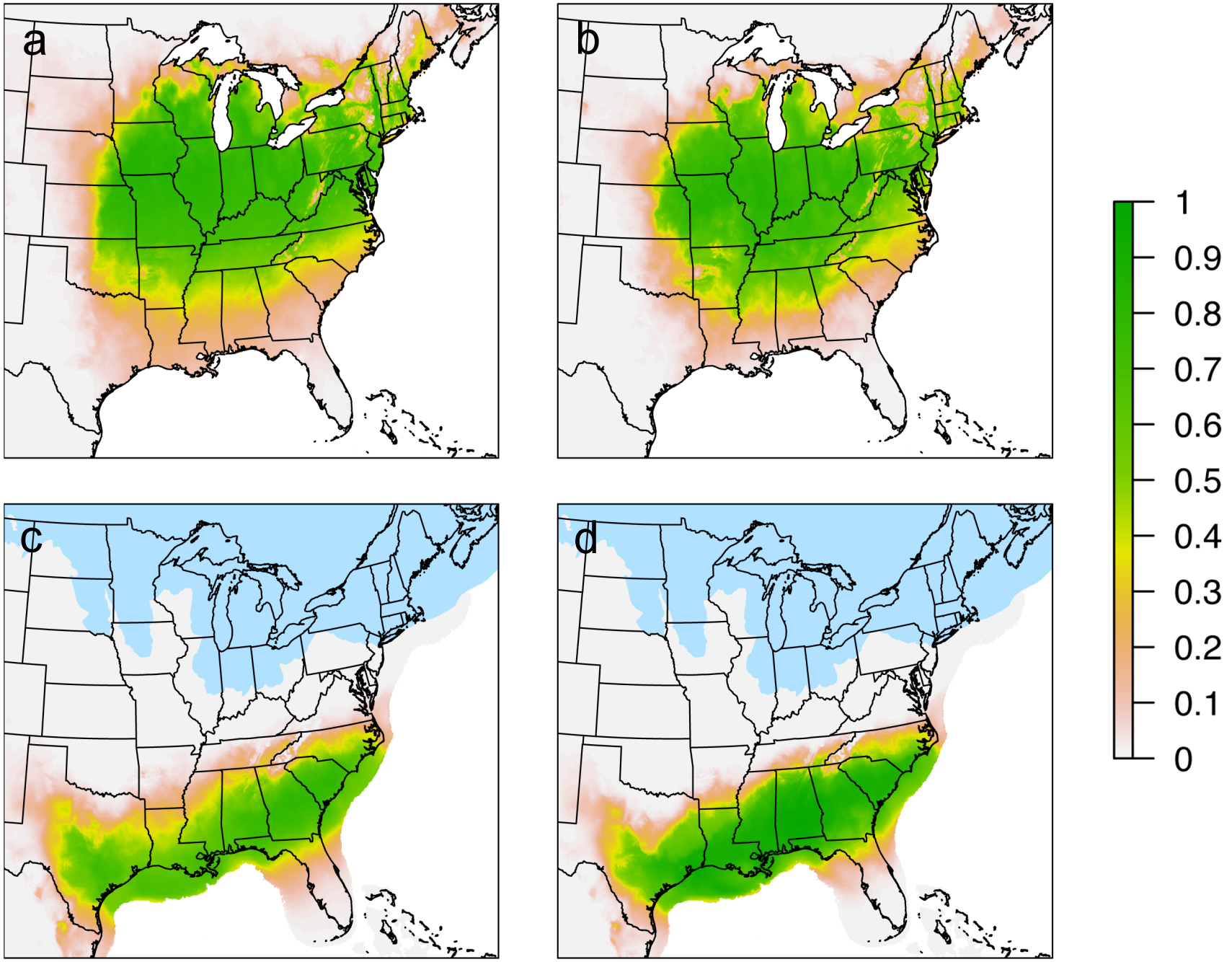
Figure 5. Species distribution models showing predicted habitat suitability (grey to green colour scale) for (a, c) *Carya cordiformis* and (b, d) *Carya ovata*, in both (a-b) the current time period and (c-d) the Last Glacial Maximum (ca. 21.5 ka). Glaciated areas are shown in blue.


a
b
n = 1
n = 10
n = 20


a
b
c
d
e
f
observed heterozygosity (H_o)
expected heterozygosity (H_e)
nucleotide diversity (π)
latitude (ºN)
R^2 = 0.23
p = 0.021
R^2 = 0.19
p = 0.051
0.12
0.14
0.16
0.18
0.20
0.22
30
34
38
42
46


a
b
0.16
0.12
0.08
0.04
0.00
$F_{ST} / (1 - F_{ST})$
0
400
800
1200
1600
2000
2400
geographic distance (km)
r = 0.36
p = 0.0017
r = 0.47
p < 0.0001


a
b
PC2 (variation explained = 1.3%)
PC1 (variation explained = 1.9%)
PC2 (variation explained = 1.5%)
PC1 (variation explained = 2.4%)
10
0
−10
−20


a
b
c
d
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0