ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/340670653

Spatial Phylogenetics of the North American Flora

Article *in* Journal of Systematics and Evolution · April 2020

DOI: 10.1111/jse.12590

CITATIONS
35

READS
667

8 authors, including:

Guralnick Robert
University of Florida
357 PUBLICATIONS **12,846** CITATIONS
SEE PROFILE

Pamela S Soltis
University of Florida
762 PUBLICATIONS **72,354** CITATIONS
SEE PROFILE

Stephen Andrew Smith
University of Michigan
180 PUBLICATIONS **19,984** CITATIONS
SEE PROFILE

Narayani Barve
University of Kansas
71 PUBLICATIONS **3,904** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

Pika Project View project

Project

Data mobilization and integrative approaches for boosting the taxonomic coverage in studies of insect biodiversity View project

All content following this page was uploaded by Julie M Allen on 20 April 2020.

The user has requested enhancement of the downloaded file.

Check for updates

Brent Mishler ORCID iD: 0000-0001-5727-4916

Shawn Laffan ORCID iD: 0000-0002-5996-0570

Research Article

Spatial Phylogenetics of the North American Flora

Brent D. Mishler[1*], Robert Guralnick[2,3], Pamela S. Soltis[2,3.4], Stephen A. Smith[5], Douglas E. Soltis[2,3.4,6], Narayani Barve[2], Julie M. Allen[7], and Shawn W. Laffan[8]

1. University and Jepson Herbaria, and Dept. of Integrative Biology, University of California, Berkeley, Berkeley CA 94720-2465, USA

2. Florida Museum of Natural History, University of Florida, Gainesville FL 32611, USA

3. Genetics Institute, University of Florida, Gainesville, FL 32608, USA

4. Biodiversity Institute, University of Florida, Gainesville, FL 32611, USA

5. Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48103, USA

6. Department of Biology, University of Florida, Gainesville, FL 32611, USA

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jse.12590.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

7. Department of Biology, University of Nevada Reno, Reno, NV 89557, USA

8. School of Biological, Earth and Environmental Sciences, The University of New South Wales, Australia

* Author for correspondence. E-mail: bmishler@berkeley.edu

Received 26 March 2019; Accepted 9 April 2020

Abstract

North America is a large continent with extensive climatic, geological, soil, and biological diversity. That biota is under threat from habitat destruction and climate change, making a quantitative assessment of biodiversity of critical importance. Rapid digitization of plant specimen records and accumulation of DNA sequence data enable a much-needed broad synthesis of species occurrences with phylogenetic data. Here we attempted the first such synthesis of a flora from such a large and diverse part of the world: all seed plants for the North American continent (here defined to include Canada, United States, and Mexico) with a focus on examining phylogenetic diversity and endemism. We collected digitized plant specimen records and chose a coarse grain for analysis, recognizing that this grain is currently necessary for reasonable completeness per sampling unit. We found that raw richness and endemism patterns largely support previous hypotheses of biodiversity hotspots. Application of phylogenetic metrics and a randomization test revealed novel results, including significant phylogenetic clustering across the continent, a striking east-west geographic difference in the distribution of branch lengths, and the discovery of centers of neo- and paleo-endemism in Mexico, the

This article is protected by copyright. All rights reserved.

Accepted Article

southwestern USA, and the southeastern USA. Finally, our examination of phylogenetic beta-diversity provides a new approach to comparing centers of endemism. We discuss the empirical challenges of working at the continental scale, and the need for more sampling across large parts of the continent, for both DNA data for terminal taxa and spatial data for poorly understood regions, to confirm and extend these results.

Keywords: biodiversity, North America, seed plants, phylogeny, endemism, phylogenetic diversity, phylogenetic endemism

1 Introduction

North America encompasses striking diversity in climate, geology, and soil, all of which have dynamic histories. The floristic history is equally dynamic and complex and is only now beginning to be understood from an evolutionary perspective. Spatial patterns of plant diversity in North America, as well as historical floristic affinities and endemism (e.g. Qian, 1998, 1999), have long been of interest to evolutionary biologists and botanists, especially global-scale disjunctions such as affinities of Eastern Asian and Eastern North American taxa (Halenius, 1750; Gray, 1846; Boufford & Spongberg, 1983; Wen, 1999). However, while quantitative approaches are lacking (but see Qian, 1998, for richness estimates at the generic level), particularly those that incorporate phylogenies, such approaches are critical for extending beyond simple pattern-based summaries and towards a more complete understanding of the historical forces that have shaped the continent’s vegetation. Furthermore, given the rapid pace of human-caused habitat destruction and climate change, advances in understanding spatial patterns of plant diversity are essential for developing sound conservation priorities.

This article is protected by copyright. All rights reserved.

Accepted Article

Recent phylogenetic approaches to biodiversity assessment have been developed to examine the spatial distribution of lineages across a region and to identify areas of particularly high lineage diversity and endemism. These approaches take advantage of recent large-scale digitization of natural history collections, rapid accumulation of DNA sequence data from many taxa, and expeditious scaling-up of methods for building large phylogenies. All of these major developments have made it possible to apply a large-scale evolutionary approach to the assessment of biodiversity and endemism that can be broadly termed *spatial phylogenetics* (Thornhill et al., 2016). The goals of spatial phylogenetics are to understand ecological, evolutionary, and biogeographic processes that have shaped the biota of an area; this information can also be used to prioritize areas of highest conservation need (Thornhill et al., 2016, 2017; Kling, 2018). A summary of spatial phylogenetic concepts and approaches is provided in Box 1.

The prospects for applying spatial phylogenetic methods across North America have been rapidly improving, given advances in digitization of museum specimens stimulated by the U.S. National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program and coordination of digitized specimen data by iDigBio (https://www.idigbio.org), as well as by development of resources such as the Open Tree of Life (Hinchliff et al., 2015), which provides a synthesis of deposited trees. The time is ripe for initiating a continent-wide analysis, given that: (*i*) herbarium digitization has been the largest target of ADBC, (*ii*) large-scale phylogenies are available for all named seed plants (Smith & Brown, 2018), and (*iii*) regional spatial phylogenetic studies of seed or vascular plants have now been completed for parts of North America (Alberta, Zhang et al., 2015; selected seed plants across North America,

This article is protected by copyright. All rights reserved.

Accepted Article

Zhang et al., 2017; California, Thornhill et al., 2017; Mexico, Sosa et al., 2018; Wisconsin, Spalink et al., 2018; Cyperaceae of North America, Spalink et al., 2018; and Florida, Allen et al., 2019). The analyses here were restricted to seed plants, although future studies could be expanded to more of the biota (e.g., Link-Pérez & Laffan, 2018). We aimed to: (1) conduct a preliminary spatial phylogenetic analysis for seed plants of North America using available data; (2) use the results to detect initial, broad patterns in the distribution of plant diversity; and (3) identify current gaps in both spatial and genetic data that need to be filled.

2 Methods

2.1 Assembly of spatial data set

Spatial data cleaning. We restricted our analysis to North America (defined here as Canada, the United States, and Mexico). Instead of taking the administrative boundary of Mexico, we used a biogeographic barrier in Yucatan (following The Nature Conservancy's terrestrial ecoregion map, http://maps.tnc.org/gis_data.html) as the extent of our study boundary for Mexico and downloaded data for kingdom *Plantae* from GBIF and iDigBio. We combined raw data from GBIF and iDigBio to further clean all records and generate occurrence points per species. All occurrence records with 'NA' or '0' values for latitude and longitude were removed, as were those that had coordinates outside North America (using a spatial overlay function in R). The total number of occurrence records before cleaning was 13,661,743. Plants in botanic gardens and greenhouses may artificially inflate diversity in those pixels, and we therefore buffered locations of 454 botanical gardens and herbaria

This article is protected by copyright. All rights reserved.

(https://github.com/ejedwards/reanalysis_zanne2014/tree/master/handling_climate_data) by 10 km and removed all points within those buffers.

The occurrence data were still likely to contain errors due to incorrectly georeferenced coordinates. Given the scope of this work, it was impossible to remove all problematic records. However, we filtered out those records with inconsistent reporting, for example those with coordinates in North America but that were listed as being from a country outside our delimited extent. We also used an outlier detection approach that we developed for this analysis and implemented in the R statistical language (R Core Team, 2013), where for each species, we determined the geographical centroid based on all species occurrences and empirically determined those occurrences away from the centroid based on various values of standard deviations; we used 3 different values of standard deviation: 2, 3 and 4. Tests of this method showed that it effectively filtered spatial outlier occurrences across multiple exemplar species and we utilized SD=4 for the following analyses. Following cleaning steps, the data set contained 11,067,080 records for 44,171 species, after performing the name reconciliation step (see name reconciliation, below). This full spatial dataset is archived online at DRYAD:xxx. As described in the following section, many of these species were not present in the phylogeny, so a reduced spatial data set with 6,940,643 records for just the 19,649 species present in the phylogeny was ultimately used in the analyses presented here. This reduced spatial dataset is archived online at DRYAD:xxx.

Final record set and name reconciliation. In order to filter the occurrence records to our study group of seed plants, and to assure standardized names, we utilized a multistep cleaning and name validation process codified into a function in R. This

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

function extracted ‘Spermatophyta’ (seed plants) names based on the accepted name table supplied by the OpenTree Taxonomy (Hinchliff et al., 2015). These accepted names were then matched against *The Plant List*, a comprehensive list of plants with synonyms and which is integrated into and helps feed the GBIF backbone taxonomy for plants (https://www.gbif.org/dataset/d9a4eedb-e985-4456-ad46-3df8472e00e8). After this first pass, we then used the gnr_resolve function from the Taxize (Chamberlain & Szöcs, 2013) package in R. This function gives a matching score based on passed names. We chose all the names having a matching score higher than 0.9, which was empirically determined to provide good results. From this list of accepted names, we then selected all of the associated synonyms from the OpenTree Taxonomy synonym table. We used this compiled list of names and extracted only binomial names, which produced 876,328 names of seed plants, covering a global scope and including valid names and synonyms.

Our next step was to match North American occurrence records from above with the list of global seed plant names. For each species name that matched occurrences, we created a separate species occurrence file, 51,747 species in total, representing each species names with at least one occurrence. This list of files could still represent synonyms rather than valid names, so we checked for synonyms one last time using the R package TaxonStand, which utilizes a version of *The Plant List*, in order to lump synonyms back to valid names. This led to a total of 44,171 accepted names and occurrence data files. These names were then finally intersected back to the phylogeny (inferred as described below), resulting in 19,649 matching tips on the tree. We note that the Open Tree phylogeny is not the same thing as comparing to the Open Tree taxonomy resources but that the purpose of using both was to ensure we could link species

This article is protected by copyright. All rights reserved.

Accepted Article

occurrence data to the tips of the phylogeny.

2.2 Phylogenetic tree assembly

The phylogeny used was constructed by conducting a hierarchical analysis of GenBank release 218. We provide a brief description here, but the methods and data analyses used for constructing this tree are described in Smith & Brown (2018). For each major clade (i.e., for angiosperms as recognized by APG IV 2016 and the Angiosperm Phylogeny Website, and for gymnosperms as recognized by the Angiosperm Phylogeny Website), we conducted clustering analyses starting at the tips and moving toward the root, conducting profile alignments with MAFFT v.7.305b (Katoh & Standley, 2013) between lineages. We then constructed 61 supermatrices based on the overall sampling within each gene, including genes that were well sampled across the entire clade as well as those genes that may be sampled well within a lineage within the clade. Initial fast likelihood trees were constructed for each alignment in order to check for outlying taxa (e.g., that may be the result of misidentification or problematic sequence) and overall quality (details are given in Smith & Brown, 2018). We then conducted full maximum likelihood analyses using RAxML v. 8.2.11 (Stamatakis, 2014) and the GTR+Γ molecular model, partitioning by gene region in the supermatrix.

Divergence times were calculated using penalized likelihood (Sanderson, 2002) as implemented in treePL (Smith & O'Meara, 2012). We obtained all calibrations by determining which clades in our trees were concordant with those in Magallón et al. (2015), extracting the dates from Magallón et al. (2015), and applying those as constraints in treePL. This approach resulted in dated trees for individual clades that we

This article is protected by copyright. All rights reserved.

Accepted Article

then combined using the backbone of the Open Tree of Life (vers. 9.1), yielding a dated phylogeny for seed plants consisting of 79,881 species (Smith & Brown, 2018). The unpruned tree is available from github.com/FePhyFoFum/big_seed_plant_trees (v. 0.1) with sequence alignments linked therein. For our analyses in the present study, this tree was pruned to the 19,649 North American species that had spatial data, as described above. The pruned tree is available from DRYAD:xxx.

2.3 Spatial analyses

The spatial data set and the phylogeny described above were analyzed using *Biodiverse version 3.0* (Laffan et al., 2010). The spatial coordinate information for each record was converted to presence within 50 × 50 km grid cells (Albers equal-area, EPSG:3310). Using the methods and metrics described by Mishler et al. (2014) and Thornhill et al. (2016), observed patterns of taxon richness (TR), weighted endemism of taxa (WE), corrected weighted endemism of taxa (CWE), phylogenetic diversity (PD), phylogenetic endemism (PE), and corrected phylogenetic endemism (CPE, calculated as PE ÷ PD; e.g., Millar et al., 2017) were mapped. We also calculated sampling redundancy for each cell (Garcillán et al., 2003), calculated as: [1 – (richness ÷ number of specimens)].

Spatial randomizations as described in Mishler et al. (2014) were run on the data set and compared against the observed results to estimate significance of PD, relative phylogenetic diversity (RPD), and relative phylogenetic endemism (RPE) as applied in CANAPE. Our goal is to detect areas that have markedly higher or lower values than expected given the number of species present. The randomization algorithm used

This article is protected by copyright. All rights reserved.

repeatedly shuffles the identities of the taxa found in each grid cell from an occurrence pool, holding constant the number of taxa per cell and the number of cells per taxon. This null model assumes that a taxon's occurrences display no spatial autocorrelation and no correlation with the occurrence of other taxa, and thus incorporates the expected correlation between species richness and endemism and the corresponding phylogeny-based measures.

Three randomizations were applied, each using different species pools. The first approach allocated species from the full North American species pool and is referred to below as the spatially unconstrained randomization. The second approach was spatially constrained by using three regional pools corresponding approximately to: (1) Canada plus Alaska, (2) the conterminous USA, and (3) Mexico. The third approach was spatially constrained by using four regional pools obtained by further splitting the conterminous USA pool into west-east portions based on the level 1 *Ecoregions of North America* boundary separating the Eastern Temperate Forest from the Great Plains (https://www.epa.gov/eco-research/ecoregions). In the two spatially constrained randomizations, a cell could only be assigned a species from within the region, thus capturing some of the expected differences in sampling effort for the regions and of the expected spatial autocorrelation of distributions across very different biomes.

2.4 Groups of cells showing significant PE

We identified groupings of significant CANAPE cells using an agglomerative UPGMA cluster analysis in *Biodiverse*, using the range-weighted phylogenetic turnover

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

metric (Laffan et al., 2016). By focusing on the shared range-restricted branches, this analysis highlights geographic regions within which the evolutionary makeup of the endemic flora is relatively homogeneous (see also Thornhill et al., 2017; Link-Pérez & Laffan, 2018).

3 Results

3.1 Observed patterns of diversity and endemism

The map of sample redundancy (fig. 1) shows considerable heterogeneity across North America. Mexico, the western USA, and some areas in the Midwest and eastern USA showed high redundancy, while Alaska and the northern and southeastern USA showed medium levels, and other parts of the USA and much of Canada have low redundancy or no samples at all.

Raw patterns of richness and endemism in species-based measures are shown in fig. 2. Richness is to some extent visually similar to redundancy, but not entirely, in that the Pacific Northwest, parts of the northeastern USA, and Florida showed high species richness, but only moderate redundancy. WE was highest in the west coast of North America, Florida, the mid-Atlantic USA, and southern Mexico. CWE was more restricted than WE and was highest in some parts of California, Florida, and Mexico.

PD showed highly similar patterns to species richness; likewise PE was quite similar to species endemism (fig. 3). Both associations are to be expected because if species occur at random with respect to each other, then as more taxa occur in a place, these taxa are likely to cover a broader span of the phylogeny. CPE showed a more

This article is protected by copyright. All rights reserved.

Accepted Article

distinct pattern, in that several areas had high CPE values that were not high in WE, CWE, or PE.

3.2 Results of statistical tests

Phylogenetic diversity. When the randomization was spatially unconstrained, i.e., species were randomly reassigned anywhere on the continent (fig. 4A), almost all grid cells were significantly low in PD (i.e., phylogenetic clustering, Webb et al., 2002). This result was also observed for both regional pool randomizations (see supplementary figures 1A and 2A).

Relative phylogenetic diversity. A striking pattern was seen in both the spatially unconstrained randomization and the randomization using 3 regional pools (fig. 4B, supplementary figure 1B), with the western half of the USA and Canada, plus central Mexico, tending to be significantly low and the eastern half of the USA and Canada, plus both coasts of Mexico, tending to be significantly high in RPD. The west coast of Canada and the Pacific Northwest showed a contrasting pattern to the rest of western North America, tending to be significantly high in RPD. The geographic areas that are significantly high or significantly low in RPD have longer or shorter branches present than expected, respectively.

In contrast, in the randomization using 4 regional pools (i.e., where the region corresponding to the conterminous USA was split; supplementary figure 2B) a different pattern in RPD significance was seen within that region. More cells with significantly longer branches than expected (blue) were seen at the western and eastern edges of the

This article is protected by copyright. All rights reserved.

Accepted Article

western region, and more cells with significantly shorter branches than expected (red) were seen in the eastern region.

CANAPE. Significance patterns detected with CANAPE differed considerably between the spatially unconstrained randomization and the two regional randomizations (compare figs. 5, 6, and supplementary figure 3). The spatially unconstrained randomization (fig. 5) showed only scattered centers of significant endemism except in the southwestern and southeastern USA and in Mexico; Mexico was almost entirely significant, with centers of neo-endemism in the interior and centers of mixed endemism everywhere else. On the other hand, for both regional randomizations (fig. 6 and supplementary figure 3), the far north of Canada (and much of Alaska) showed significant concentrations of neo- and mixed endemism, while California and the southeastern USA showed much more extensive areas of significant endemism. Additionally, Mexico did not show as many significant centers of endemism as in the spatially unconstrained randomization; the major areas remaining as significant centers of endemism were in Baja California and extreme southern Mexico.

Phylogenetic turnover among centers of endemism discovered in CANAPE. With the spatially unconstrained randomization (fig. 7), the greatest dissimilarity was between Canada and the rest of the continent. The Mexican centers of endemism formed a discrete cluster, with a north-south split within Mexico. With the randomization constrained within three latitudinal polygons (fig. 8), one cluster grouped centers of endemism in the USA and Mexico, while another cluster grouped centers of endemism in Canada and Alaska. Similarly, with the randomization constrained within four latitudinal polygons

This article is protected by copyright. All rights reserved.

Accepted Article

(supplementary figure 4), one cluster grouped centers of endemism in the USA and Mexico, while another cluster grouped centers of endemism in Canada and Alaska.

4 Discussion

4.1 Distribution of diversity and endemism of seed plants in North America

Four of the world's 35 (Myers et al., 2000), or 36 (Noss et al., 2015), biodiversity hotspots have been recognized in North America, and all of these were characterized by high species richness, PD, and PE (figs. 2 and 3). To qualify as a hotspot, a region must be home to 1500 endemic vascular plant species and retain 30% or less of its native vegetation. The California Floristic Province is one such hotspot and represents a region of high species diversification and endemism (Thornhill et al., 2017). The high PD and PE seen here (fig. 3) for western North America (the California Floristic Province as well as the Pacific Northwest) emphasize that the region is not only actively generating many new species, but also harboring many older phylogenetic branches. The Madrean Pine-Oak Woodlands of central and northern Mexico and parts of Arizona and New Mexico, USA, is another biodiversity hotspot that showed high PD near its southern limits (fig. 3) and considerable concentrations of neo-endemism (fig. 5), most likely due to recent diversification in lineages such as pines and oaks and the accumulation of closely related species (e.g. pines: Gernandt et al., 2003; Hernandez-Leon et al., 2013; oaks: Hipp et al., 2018). To the south of the Madrean Pine-Oak Woodlands, the Mesoamerica hotspot showed high mixed PE (fig. 5), likely reflecting both high recent diversification within some clades and the presence of many ancient clades, such as Lauraceae, with long branch lengths.

This article is protected by copyright. All rights reserved.

Accepted Article

Regions of southern Florida likewise appear to harbor complex assemblages of ancient and recent lineages, contributing to high phylogenetic diversity and endemism (also see Allen et al., 2019). The North American Coastal Plain (NACP) biodiversity hotspot was recently proposed (Noss et al., 2015) to call attention to this floristically diverse and threatened area. The region essentially corresponds to the Geological Coastal Plain and the Coastal Plain Floristic Province (CPFP; recognized by Takhtajan, 1986, as the most sharply defined floristic province in North America) but includes southern Florida, southern Texas, and northeastern Mexico, which the CPFP does not. The southeastern USA, especially Florida, has many centers of high PE, and Florida has 'mini-hotspots' of its own, i.e., southern Florida and the Apalachicola Bluffs region in the Florida Panhandle, both of which showed particularly high phylogenetic diversity on both this continental scale and from the perspective of Florida alone (Allen et al., 2019).

Phylogenetic turnover analyses of the areas identified as significantly high in in PE, using all three spatial randomizations (figs. 7, 8, and supplementary figure 4), showed similar floristic relationships. The hotspots of endemism in Mexico were more similar to each other than to those in the USA, with three exceptions: (1) northern Baja California, which was confirmed as part of the California Floristic Province, (2) north central Mexico, which was confirmed as part of the Madrean Pine-Oak Woodlands. (3) far southern Mexico and the southern tip of Florida showed affinities as parts of the Mesoamerica hotspot. The fact that regions of high PD and PE seen here effectively identified areas recognized as biodiversity hotspots on the basis of other criteria suggests that large-scale analyses such as this, despite issues of data availability, sampling bias,

This article is protected by copyright. All rights reserved.

and the like (discussed below), are still valuable for detecting hotspots worthy of protection.

Accepted Article

4.2 Phylogenetic clustering in the North American flora

Essentially all of North America exhibited significantly low PD (e.g., fig. 4A), indicating that at this broad geographic scale, co-occurring species are more closely related to each other than would be expected by chance. This finding has not been seen in previous spatial phylogenetic analyses, which were carried out on smaller and less heterogeneous regions. Significantly low PD might have been expected for the analyses presented here given the large, ecologically heterogeneous biogeographic area investigated. North America comprises a number of distinct biomes (Barbour et al., 2000), spanning tropical to arctic latitudes, and sea level to high-montane habitats, thus widespread phylogenetic clustering is not surprising, as scale affects the deviation from random phylogenetic assembly of species (Cavendar-Bares et al., 2009). Similar analyses, conducted at a smaller scale and over a more homogeneous landscape, recover more areas with significantly high measures for PD (e.g., Thornhill et al., 2016, 2017; Scherson et al., 2017; Allen et al., 2019).

4.3 Accumulation of old floras and diversification of new floras

RPD showed a striking separation (fig. 4B and supplementary fig. 1) of significantly high RPD regions, indicating a concentration of long phylogenetic branches (in eastern North America, some coastal regions of western North America, and southern Mexico), and significantly low RPD regions, indicating a concentration of short phylogenetic branches (in drier parts of western North America and across northern

This article is protected by copyright. All rights reserved.

Accepted Article

Canada and Alaska). This distinction between eastern and western North America is dramatic and parallels the long recognition of a mid-continental biogeographic break (Soltis et al., 2006). This pattern reflects the highly distinct floristic differences of these regions. Eastern North America, along with eastern Asia and portions of western North America and Europe, is home to remnants of mixed mesophytic forest that dominated much of the Northern Hemisphere until the Miocene (Boufford & Spongberg, 1983; Wen, 1999). This flora is rich in diverse woody, as well as herbaceous, lineages broadly spanning the seed plants. Global cooling and drying caused the extinction of many lineages of vascular plants from the mid-Miocene to the Quaternary, approximately a 10-my timeframe (Xiang et al., 2000), and resulted in separated pockets of this ancient floristic assemblage in western North America and southern Europe, as well as the broader expanses in eastern North America and eastern Asia (Li, 1952; Boufford & Spongberg, 1983; Wen, 1999; Ricklefs et al., 2004).

The Mississippi River has been recognized as a major source of discontinuity for both plants and animals (reviewed in Soltis et al., 2006). North America west of the Mississippi River comprises a set of more recently assembled floras than the eastern deciduous forest and its remnant relative in the Pacific Northwest. For example, prairie arose in response to aridification of central North America, and dry-adapted steppe and desert species in western and southwestern North America, including Mexico, represent adaptation to further aridification, much of which occurred during the late Middle Miocene (14.8-12.2 mya; Eronen et al., 2012). In addition, orogeny in western North America has generated recent communities, and ongoing volcanism continues to provide new habitats and opportunities for species diversification. As a result, these geologically

This article is protected by copyright. All rights reserved.

Accepted Article

young regions support younger floras that are composed largely of recent radiations. The consequence is closely related species with short branch lengths, resulting in the significantly low RPD observed. These regions, because of their geological youth, lack the accumulated diversity of disparate lineages present in eastern North America, for example.

The spatial randomization using 4 regional pools (i.e., where the region corresponding to the conterminous USA was split into two subregions; supplementary figure 2B) showed a different RPD pattern in the conterminous USA part of the map, which is to be expected given the new restrictions placed on the randomization. The simple west-east distinction seen in the unconstrained randomization is not observed because the randomization only places taxon occurrences within their original subregion. Instead, concentrations of long branches were observed along the West Coast and the eastern edge of the western subregion, while significant concentrations of short branches were seen in the western part of the eastern subregion; in both cases *local* patterns within subregions are uncovered, demonstrating the effect of scale discussed in the following section.

Despite recent orogeny, volcanism, and aridification, some coastal areas in the Pacific Northwest, the California Floristic Province, and Mexico have accumulated long branches, reflected in the significantly high RPD observed there (fig. 4B). These patterns are perhaps due to the modulating effects of the Mediterranean and subtropical to tropical climates, respectively, while recently opened habitats in interior western North America lack this accumulation. Continental-scale analyses such as this can identify patterns and generate hypotheses for more focused study, and we note the critical need for historical

This article is protected by copyright. All rights reserved.

Accepted Article

biogeographic analyses of specific clades as well as intensive analyses of climate versus geology and orogeny as drivers of the observed patterns.

4.4 Methodological issues with geographic sampling

This analysis illustrates the importance of considering scaling in geographic analysis. There are questions at both ends of the spatial scale. How large a portion of the world should be studied at once? And what should the geographic resolution (here the cell size) of the analysis be?

The extent of the study area has a substantial effect on expected patterns. In this very broad-scale analysis, incorporating most climate regions between the tropics and the arctic, the uniform observation of significantly low PD (i.e., phylogenetic clustering; fig. 4A) is to be expected. Habitat preference tends to be phylogenetically conservative, so relatives tend to co-occur in a given region or biome. At a smaller study scale, within biomes or communities, cases of phylogenetic over-dispersion are more likely, perhaps due to ecological processes such as competition. The boundary between ecology and biogeography lies in this transition zone where evolutionary and historical processes at larger geographic scales switch to ecological processes at finer scales (Webb et al., 2002). That line might best be defined by the cut-off below which it is reasonable to assume all terminal taxa could potentially occur anywhere in the study area if not for their traits prohibiting them from certain physical or biological environments (i.e., the "no dispersal limitation" assumption).

This consideration of scale of course also impacts the choice of spatial randomization model. The spatially unconstrained model applied here and in many other

This article is protected by copyright. All rights reserved.

Accepted Article

studies assumes that all terminal taxa can potentially occur anywhere in a study area. This can be unreasonable over broad areas such as in this study, and randomizations should ideally be spatially constrained in some way such that the pool of taxa sampled meets the "no dispersal limitation" assumption. However, a challenge then emerges to define regions that reflect natural dispersal constraints but that do not impose circularity on the results, i.e., without too many *a priori* assumptions about how taxa are distributed. We applied two spatially constrained randomizations here as examples, one using three polygons broadly corresponding to Canada and Alaska, the conterminous USA, and Mexico, and a second using four polygons (where the conterminous USA was broken into west and east polygons). CANAPE showed clear differences between the three randomizations (compare figs. 5, 6, and supplementary fig. 3). The patterns of endemism within the polygons could be more clearly seen in the constrained randomization, although there is the worry about edge effects as discussed below. Although divisions based on political boundaries are not ideal, we chose these polygons for convenience to test the effects of spatially constrained and unconstrained analyses; future analyses should seek biologically based areas as the basis for constrained analysis. Further work is needed into how to define "natural" subregions to serve as constraints for spatial randomizations. It would be worth exploring using phylogenetic turnover analyses, including range-weighted turnover (Laffan et al., 2016), to objectively determine relatively homogeneous regions that might be used as a spatial constraint for these randomizations.

A related consideration is the effect of edges of regions on patterns of endemism seen in the region. If a lineage is widespread outside a study region, yet rare within it, it is

This article is protected by copyright. All rights reserved.

Accepted Article

treated as highly range-restricted in the region, increasing its contribution to the endemism analyses. In this study, such an effect is likely present near the southern border of Mexico in all three randomizations, near the southern border of the middle polygon in the three region randomization, as well as the west-east boundary in the four region randomization. Many lineages near those edges are likely widespread outside the region, thus significant endemism seen near the edge may only represent local endemism. This edge effect is a factor to be considered in any spatial phylogenetic study that covers less than the whole world. Many internal branches will contain descendants that are not found in the study region, and thus their geographic ranges are underestimated even with full sampling of the study region. This is an effect that increases towards the root of the tree. However, for data sets such as this, many internal branch ranges are sufficiently large that the difference in range weighted lengths does not have a great effect on the PE results.

The edge effect on ranges can be examined in several ways. For example, when major centers of endemism appeared adjacent to the border of California, Thornhill et al. (2017) re-calculated CANAPE using only terminal taxa that are completely restricted to California, in their case finding the same results in most cases. As another example, to better represent range sizes outside their study area of Florida, Barve et al. (in prep.) included broader ranges of terminal taxa in the Western Hemisphere when calculating endemism. In any case, a regional study of endemism always needs to be regarded as potentially only documenting local endemism relative to that study area, although even then it is not an artifact if understood properly. Local endemism is often a target for conservation within political units, and land managers need to know where the rare plants are concentrated within their jurisdiction.

This article is protected by copyright. All rights reserved.

Accepted Article

Another issue of spatial scale is the resolution of the spatial units (often referred to as the grain), which for this study is the cell size. In principle, smaller grid cells are better, but the optimal grid cell size depends on knowledge of distribution of the study organisms and the spatial density of available sample data. As in the present study, distributions of terminal taxa can be estimated from point occurrence records. In most cases, these are taken from museum or herbarium databases, but may also come from unvouchered reports in the literature or crowdsourced media such as *iNaturalist* (https://www.inaturalist.org/*)*. Redundancy (fig. 1) is a good measure to estimate sampling density -- when it is high, there are many collections given the number of taxa, and it is at the lowest when there is only one collection per taxon, and the loss of one record implies a reduction of one unit of taxon richness. Baldwin et al. (2017) have an extensive discussion of choosing a grid cell size for analysis with point data.

It is clear that the number of species recognized in a region is related to both collecting effort and the extent of taxonomic endeavor, as was clearly observed in this study (i.e., comparing figs. 1 and 2). The causal relationship between collecting efforts and richness is complicated, however, by the typical behavior of plant collectors, who are attracted to areas of higher richness and endemism. Collection of herbarium specimens is about as far from random sampling as possible, being biased towards both richness and endemism, as discussed in detail by Baldwin et al. (2017). Because occurrence records from herbarium specimens are known to be biased by multiple factors (e.g. Baldwin et al., 2017; Daru et al., 2017), attempts are sometimes made to fill in ranges beyond the known occurrences by either niche modeling (i.e., through analysis of environmental correlates of known occurrences of a taxon and making predictions of where suitable

This article is protected by copyright. All rights reserved.

Accepted Article

habitat is elsewhere in the region) or by using continuous range maps authored by experts on a taxon based on their personal field experience. However, both of these approaches have potential problems of their own. The former assumes that the macro-environmental variables used in the model are the limiting factors for the range of a taxon, as opposed to biotic interactions or dispersal limitations (see further discussion in Thornhill et al., 2017). The latter assumes that expert knowledge is reliable and may still only be usable at relatively coarse grain sizes (Jetz et al. 2012). There are trade-offs involved in that grid cell sizes can often be smaller, especially for modeled distributions, but there is a danger that the apparent precision of the range may be inaccurate. No matter which methods are used to represent a taxon's range, improved point occurrence sampling is an important component in improving assemblage completeness and accuracy.

It is currently a good time to consider integrating checklist and inventory data (Mutke & Barthlott, 2005; see Ulloa Ulloa et al., 2017 for an example of integrated checklists across the Americas) often maintained by state and national parks, counties, and state and federal agencies, These resources allow us to gather the most complete assessment of species presences, and potentially absences, where possible. Such integration introduces new challenges including the verifiability of unvouchered observations. However, it is worth pursuing if the goal is to generate the most complete possible list of species over spatial units at a relatively broad scale. Next-step efforts are still needed to properly assemble those different types of data and use them in models that can best account for spatial and temporal grain and uncertainty of input data (Jetz et al., 2012).

This article is protected by copyright. All rights reserved.

Accepted Article

4.5 Methodological issues with taxon sampling

It is important to realize that there is no such thing as "complete" taxon sampling. Even if one could have every known extant species in North America in the phylogeny, that would still not include the unknown extant lineages, and most importantly would not include the extinct ones. Therefore, it is necessary to think of all these biodiversity measures as estimates of the minimum net balance of evolutionary events (i.e., lineage splitting, lineage extinction, and amount of change along branches; see Kling et al., 2018 for elaboration) that were present in the history of the terminal taxa that are currently present in a grid cell. The “minimum” clause is important since there could have been many more unobserved events, but there could not have been fewer as long as the phylogeny is accurate.

We need to have a more complete understanding of taxonomy in the North American flora, particularly of terminal taxa (TT), being as careful as possible that these are clades (Mishler & Wilkins, 2018), and we need DNA data from as many TT as possible to include in the phylogeny. The most complete phylogenetic tree that was possible to assemble at present contained only 19,649 (38%) of the 51,747 species for which we found North American spatial data. Thus our current data set includes only a subset of TT and whether this limited sampling induces a bias likely depends on the distribution of missing terminals on the phylogeny. If there is phylogenetic signal in the unsampled TT, that could introduce a bias for geographic regions that contain those TT, whereas if missing TT are randomly scattered on the tree, bias in PD-based measures is unlikely from that source.

This article is protected by copyright. All rights reserved.

Accepted Article

It is likely that additional field work may result in the discovery of new species, even in well-sampled regions (e.g., conspicuous mints, *Conradina etoniah*, Kral & McCartney, 1991; *C. cygniflora*, Edwards et al., 2009; and many new taxa from California, Ertter, 2000), but whether this would change the patterns observed here would depend on how those additional species are related to currently recognized species and how much they geographically overlap. For example, increased taxonomic study might in fact reduce RPD through discovery of cryptic species/complexes (e.g., Scheffers et al., 2012), which if they co-occur, would reduce the branch lengths within a grid cell. Thus, the interplay between collecting effort and taxonomic attention may or may not alter current patterns. More attention needs to be given to data gaps and biases, and to how the extent of scientific study may affect metrics of species vs. phylogenetic diversity.

4.6 Conclusion: needs for the future

Better sampling, both of DNA sequences representing taxa and of their geographic distributions, is a critical concern for improved understanding of spatial phylogenetic patterns and the processes that generate them. Establishing a complete species list for a given area, even a well-studied area, is plagued by many issues. The full data set produced here after cleaning had 51,747 species. But further checking of these names against *The Plant List* using the R package Taxonstand (Cayuela et al., 2012) suggests that only 44,171 of these names may be valid, with the rest being synonyms or unresolved names. Ulloa Ulloa et al. (2017) list approximately 23,000 species for Mexico and 15,000 for the United States and Canada, including the Caribbean, for a total of 38,000 species, based on a compilation of checklists. Thus, for the region corresponding to our area of study, their list would include approximately 30,000-35,000 species.

This article is protected by copyright. All rights reserved.

Govaerts (2001) lists 34,455 seed plant species for Northern America (defined as Greenland, Canada, the United States excluding Hawaii and Puerto Rico, and Mexico), although he admitted that this number seems low. Our higher number of species than these other two estimates may result from multiple factors. For example, our list includes both native and non-native species, and there clearly are still issues of synonymy and taxonomy as well. Our efforts to resolve cases of synonymy at multiple steps in our workflow reduced the estimated number of species somewhat. However, merging the distributions of the synonyms had no impact on the overall patterns of phylodiversity observed when comparing preliminary analyses to the final ones shown here

Strategic digitization and mobilization to further fill spatial gaps is key to gathering a more complete assessment of seed plant diversity (Meyer et al., 2016). Such efforts have been strongly catalyzed by the national efforts supported by the US National Science Foundation's ADBC program and its hub, iDigBio. However, further efforts geared towards developing new networks of people, data, and institutions around herbarium digitization are still needed, especially because some regions in North America are underrepresented despite high likelihood of available analog (but not digital) data. For example, we note limited availability of digitized seed plant data in the Canadian high arctic, which is an area containing often unique boreal and tundra species and that is also potentially under high threat from amplified effects of climatic change (Holland & Bitz, 2003).

The present study has yielded general and broad biodiversity patterns that are likely to be confirmed with better sampling. However, there remains much to be

Accepted Article

This article is protected by copyright. All rights reserved.

discovered about pattern of and processes generating, floristic diversity, especially for the more poorly sampled parts of North America

Accepted Article

Acknowledgements

Partially supported by the following NSF grants: DEB-1207915, DEB-1354552, DBI-1458640, DBI-1458466, and DBI-1547229.

Literature Cited

Allen JM, Germain-Aubrey CC, Barve N, Neubig KM, Majure LC, Laffan SW, Mishler BD, Owens HL, Smith SA, Whitten WM, Abbott JR, Soltis DE, Guralnick R, Soltis PS. 2019. Spatial phylogenetics of Florida vascular plants: the effects of calibration and uncertainty on diversity estimates. *iScience* 11: 57–70.

Baldwin BG, Thornhill AH, Freyman WA, Ackerly DD, Kling MM, Morueta-Holme N, Mishler BD. 2017. Species richness and endemism in the native flora of California. *American Journal of Botany* 104: 487–501.

Barbour MG, Billings WD. 2000. *North American terrestrial vegetation*. Cambridge, UK: Cambridge Univ. Press.

Boufford DE, Spongberg SA. 1983. Eastern Asian–eastern North American phytogeographical relationships—A history from the time of Linnaeus to the twentieth century. *Annals Missouri Botanical Garden* 70: 423 – 439.

Cavender-Bares J, Kozak KH, Fine PVA, Kembel SW. 2009. The merging of community ecology and phylogenetic biology. *Ecology Letters* 12: 693-715.

This article is protected by copyright. All rights reserved.

Accepted Article

Cayuela L, Granzow-de la Cerda Í, Albuquerque FS, Golicher DJ. 2012. Taxonstand: An r package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution* 3: 1078-1083.

Chamberlain SA, Szöcs E. 2013. Taxize: taxonomic search and retrieval in R. *F1000Research* 2:191. doi:10.12688/f1000research.2-191.v2

Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfeld TJ, Seidler TG, Sweeney PW, Foster DR, Ellison AM., Davis CC. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217: 939-955.

Edwards CE, Judd WS, Ionata GM, Herring B. 2009. Using population genetic data as a tool to identify new species: *Conradina cygniflora* (Lamiaceae), a new, endangered species from Florida. *Systematic Botany* 34:747-759.

Eronen JT, Fortelius M, Micheels A, Portmann FT, Puolamäki K, Janis CM. Neogene aridification of the Northern Hemisphere. *Geology* 40: 823-826

Ertter B. 2000. Floristic surprises in North America north of Mexico. *Annals of the Missouri Botanical Garden* 87: 81–109.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1–10.

This article is protected by copyright. All rights reserved.

Accepted Article

Garcillán PP, Ezcurra E, Riemann H. 2003. Distribution and species richness of woody dryland legumes in Baja California, Mexico. *Journal of Vegetation Science* 14: 475-486.

Gernandt DS, Liston A, Piñero D. 2003. Phylogenetics of *Pinus* subsections Cembroides and Nelsoniae inferred from cpDNA sequences. *Systematic Botany* 28: 657-673.

Govaerts R. 2001. How many species of seed plants are there? *Taxon* 50: 1085–1090.

Graham CH, Fine PVA. 2008. Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters* 11: 1265-1277.

Gray A. 1846. Analogy between the flora of Japan and that of the United States. *American Journal of Science and Arts* 2: 135-136.

Halenius J. 1750. Plantae Rariores Camschatcenses. Thesis, Univ. Uppsala, Uppsala.

Hernández-León S, Gernandt DS, Pérez de la Rosa JA, Jardón-Barbolla L. 2013. Phylogenetic relationships and species delimitation in *Pinus* Section Trifoliae inferrred from plastid DNA. *PLoS ONE* 8: e70501.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazisg R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Reed RH, Rees JA, Soltis DE, Williams T, Cranston KA. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings National Academy of Sciences* 112: 12764–12769.

This article is protected by copyright. All rights reserved.

Accepted Article

Hipp AL, Manos PS, Gonzalez-Rodriguez A, Hahn M, Kaproth M, McVay JD, Avalos SV, and Cavender-Bares J. 2018. Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. *New Phytologist* 217: 439–452.

Holland MM, Bitz CM. 2003. Polar amplification of climate change in coupled models. *Climate Dynamics* 21:221–232.

Jetz W, McPherson JM, Guralnick RP. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution* 27: 151–159.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kling MM, Mishler BD, Thornhill AH, Baldwin BG, Ackerly DD. 2018. Facets of phylodiversity: evolutionary diversification, divergence, and survival as conservation targets. *Philosophical Transactions Royal Society B* 374: 20170397.

Kral R, McCartney RB. 1991. A new species of *Conradina* (Lamiaceae) from northeastern peninsular Florida. *Sida* 14: 391–398.

Laffan SW, Lubarsky E, Rosauer DF. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography* 33, 643-647.

This article is protected by copyright. All rights reserved.

Accepted Article

Laffan SW, Rosauer DF, Di Virgilio G, Miller JT, Gonzales-Orozco C, Knerr N, Thornhill AH, Mishler BD. 2016. Range-weighted metrics of species and phylogenetic turnover can better resolve biogeographic breaks and boundaries. *Methods in Ecology and Evolution* 7: 580-588.

Li HL. 1952. Floristic relationships between eastern Asia and eastern North America. *Transactions of the American Philosophical Society* 42: 371-429.

Link-Pérez MA, Laffan SW. 2018. Fern and lycophyte diversity in the Pacific Northwest: Patterns and predictors. *Journal of Systematics and Evolution* 56: 498-522.

Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *The New Phytologist* 207: 437–453.

Millar TR, Heenan PB, Wilton AD, Smissen RD, Breitwieser I. Spatial distribution of species, genus and phylogenetic endemism in the vascular flora of New Zealand, and implications for conservation. *Australian Systematic Botany* 30(2):134-148.

Mishler BD, Wilkins JS. 2018. The hunting of the SNaRC: a snarky solution to the species problem. *Philosophy, Theory, and Practice in Biology* 10: 1-18.

Mishler BD, Knerr N, González-Orozco CE, Thornhill AH, Laffan SW, Miller JT. 2014. Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian *Acacia*. *Nature Communications* 5: 4473.

This article is protected by copyright. All rights reserved.

Accepted Article

Meyer C, Jetz W, Guralnick R, Fritz SA, Kreft H. 2016. Species-level biases in distribution records are driven by range geometry and socio-economics, not by species' detection or collection probabilities. *Global Ecology and Biogeography* 25: 1181-1193.

Mutke J, Barthlott W. 2005. Patterns of vascular plant diversity at continental to global scales. *Biologiske Skrifter* 55: 521-531.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853–8.

Noss, R. F., W. J. Platt, B. A. Sorrie, A. S. Weakley, D. B. Means, J. Costanza, and R. K. Peet. 2015. How global biodiversity hotspots may go unrecognized: lessons from the North American Coastal Plain. *Diversity and Distributions* 21: 236–244.

Qian, H. 1998. Large-scale biogeographic patterns of vascular plant richness in North America: An analysis at the generic level. *Journal of Biogeography* 25: 829-836.

Qian, H. 1999. Spatial pattern of vascular plant diversity in North America north of Mexico and its floristic relationship with Eurasia. *Annals of Botany* 83: 271-283.

R Core Team. 2013. R: A language environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ricklefs RE, Qian H, White PS. 2004. The region effect on mesoscale plant species richness between eastern Asia and eastern North America. *Ecography* 27:129-136.

This article is protected by copyright. All rights reserved.

Rosauer, DF, Laffan SW, Crisp MD, Donnellan SC, Cook LG. 2009. Phylogenetic endemism: a new approach to identifying geographical concentrations of evolutionary history. *Molecular Ecolog*y 18: 4061-4072.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19: 101–109.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF. 2012. What we know and don’t know about Earth’s missing biodiversity. *Trends in Ecology and Evolution* 27: 501–510.

Scherson RA, Thornhill AH, Urbina-Casanova R, Freyman WA, Pliscoff PA, Mishler BD. 2017. Spatial phylogenetics of the vascular flora of Chile. *Molecular Phylogenetics and Evolution* 112: 88-95.

Smith SA, Brown JW. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* 105: 302–314.

Smith SA, Walker JF. 2018. PyPHLAWD: A python tool for phylogenetic dataset construction. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13096

Smith SA, and O’Meara BC. 2012. TreePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28: 2689–2690.

Soltis DE, Morris A, McLachlan J, Manos P, Soltis PS. 2006. Comparative phylogeography of eastern North America. *Molecular Ecology* 15: 4261-4293.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Sosa V, De-Nova JA, Vásquez-Cruz M. 2018. Evolutionary history of the flora of Mexico: Dry forests cradles and museums of endemism. *Journal of Systematics and Evolution* 56: 523-536.

Spalink D, Pender J, Escudero M, Hipp AL, Roalson EH, Starr JR, Waterway MJ, Bohs L, Sytsma KJ. 2018. The spatial structure of phylogenetic and functional diversity in the United States and Canada: An example using the sedge family (Cyperaceae). *Journal of Systematics and Evolution* 56: 449-465.

Spalink D, Kriebel R, Li, P, Pace MC, Drew BT, Zaborsky JG, Rose J, Drummond CP, Feist MA, Alverson WS, Waller DM, Cameron KM, Givnish TJ, Sytsma KJ. 2018. Spatial phylogenetics reveals evolutionary constraints on the assembly of a large regional flora. *American Journal of Botany* 105: 1–13.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Takhtajan A. 1986. *Floristic regions of the world*. Berkeley: University of California Press.

The Angiosperm Phylogeny Group: Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.

The Plant List, 2013. Version 1.1. http://www.theplantlist.org/

This article is protected by copyright. All rights reserved.

Accepted Article

Thornhill AH, Mishler BD, Knerr N, Gonzalez-Orozco CE, Costion CM, Crayn DM, Laffan SW, Miller JT. 2016. Continental-scale spatial phylogenetics of Australian angiosperms provides insights into ecology, evolution and conservation. *Journal of Biogeography* 43: 2085–2098.

Thornhill AH, Baldwin BG, Freyman WA, Nosratinia S, Kling MM, Morueta-Holme N, Madsen TP, Ackerly DD, Mishler BD. 2017. Spatial phylogenetics of the native California flora. *BMC Biology* 15:96.

Ulloa Ulloa C, Acevedo-Rodríguez P, Beck S, Belgrano MJ, Bernal R, Berry PE, Brako L, Celis M, Davidse G, Forzza RC, Gradstein SR, Hokche O, León B, León-Yánez S, Magill RE, Neill DA, Nee M, Raven PH, Stimmel H, Strong MT, Villaseñor JL, Zarucchi JL, Zuloaga FO, Jørgensen PM. 2017. An integrated assessment of the vascular plant species of the Americas. *Science* 358: 1614–1617.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33: 475–505.

Wen J. 1999. Evolution of the eastern Asian and eastern North American disjunct distributions in flowering plants. *Annual Review of Ecology and Systematics* 30: 421-455.

Xiang QY, Soltis DE, Soltis PS, Manchester SR, Crawford DJ. 2000. Timing the eastern Asian–eastern North American floristic disjunction: molecular clock corroborates paleontological estimates. *Molecular Phylogenetics and Evolution* 15: 462–472.

This article is protected by copyright. All rights reserved.

Accepted Article

Zhang J, Nielsen SE, Stolar J, Chen Y, Thuiller W. 2015. Gains and losses of plant species and phylogenetic diversity for a northern high-latitude region. *Diversity and Distributions* 21: 1441-1454.

Zhang J, Nielsen SE, Chen Y, Georges D, Qin Y, Wang S-S, Svenning JC. Thuiller,W. 2017. Extinction risk of North American seed plants elevated by climate and land-use change. *J Appl Ecol*, 54: 303-312.

Box 1. Review of key phylodiversity measurements and approaches.

In spatial phylogenetics, alpha-diversity is primarily measured using the *phylogenetic diversity* metric (PD; Faith, 1992), the sum of branch lengths connecting the terminal taxa present in a given location (usually to the root of the tree). These phylogenetic methods have the advantage of being taxonomically rank-free. The taxonomic rank of terminals is unimportant as long as the entity the terminal represents is monophyletic and its geographic distribution can be characterized; the methods are thus relatively robust to lumping and splitting decisions by taxonomists. Beta-diversity, or turn-over on the landscape, is likewise measured by comparing the lengths of branches shared among pairs of locations, as discussed below.

Spatial phylogenetics also incorporates a measure of *phylogenetic endemism* (PE; Rosauer et al., 2009), which is like PD, but where the branches are weighted by the fraction of their geographic range found in that location. For a location that is a cell in a grid of equal-sized cells, this is equivalent to using a range-weighted tree (RWT). The RWT is produced by dividing each original branch length by its range size; thus,

This article is protected by copyright. All rights reserved.

Accepted Article

widespread branches shrink to insignificance and narrowly restricted branches are up-weighted.

Two derived metrics are also employed, *relative phylogenetic diversity* and *relative phylogenetic endemism* (RPD and RPE, respectively; Mishler et al., 2014). These measures are ratios of PD or PE measured on the original tree to PD or PE measured using a comparison tree that retains the same topology and sum of branch lengths, but which has each branch adjusted to be of equal length. The purpose is to see if unexpected concentrations of long branches or short branches are present in a location. RPE is employed in a method called *Categorical Analysis of Neo- And Paleo-Endemism* (CANAPE, Mishler et al., 2014). CANAPE searches for geographic centers of endemism using PE and classifies them using RPE by the branch lengths of the range-restricted taxa within them, allowing a clear, quantitative distinction between centers of neo- and paleo-endemism.

A spatial randomization process is used to test statistical significance of PD, RPD, and RPE (via CANAPE). The most commonly used algorithm re-assigns terminal taxon occurrences on the map, subject to two constraints: the range size of each taxon and the richness of each location are held constant. Significantly low or high PD means, respectively, that the terminal taxa that occur together at a location are either more closely related to each other than expected at random (phylogenetic clustering), or more distantly related to each other than expected at random (phylogenetic overdispersion). Significantly low or high RPD also indicates, respectively, that shorter or longer branches than expected at random occur at a location. The randomization as applied in CANAPE identifies locations that are significantly high in PE; then those locations (defined as

This article is protected by copyright. All rights reserved.

Accepted Article

centers of PE) are classified using RPE in a two-tailed test. Significantly low or high RPE means, respectively, that shorter rare branches or longer rare branches are more concentrated at a location than expected at random (i.e., centers of neo-endemism and paleo-endemism, respectively). Locations that have significantly high PE, but that are not significantly high or low in RPE, are interpreted as centers of mixed PE (see Mishler et al., 2014, for details).

In conjunction with these alpha-diversity measures, beta-diversity can also be studied phylogenetically. Regular phyloturnover compares the branch lengths found in two locations that are shared and not shared, using standard measures such as Jaccard's or Sørensen's coefficients (Graham & Fine, 2008). A more recent approach called *phylogenetic range-weighted turnover* (PhyloRWT, Laffan et al., 2016) emphasizes the branches that have smaller ranges. PhyloRWT serves as a particularly useful measure of phylobetadiversity for purposes of understanding changes in phylogenetic assemblages across the landscape, enabling applications such as bioregionalization, ecological studies of causes of beta-diversity, and analyses for applied conservation studies.

This article is protected by copyright. All rights reserved.

Accepted Article

Figures

Fig. 1. Redundancy (1 - [richness/#specimens]). Values close to 1 (brown) are well-sampled while zero (tan) means there is no redundancy in the sampling. Note that large parts of northern Canada are poorly sampled, along with a few areas in the USA.

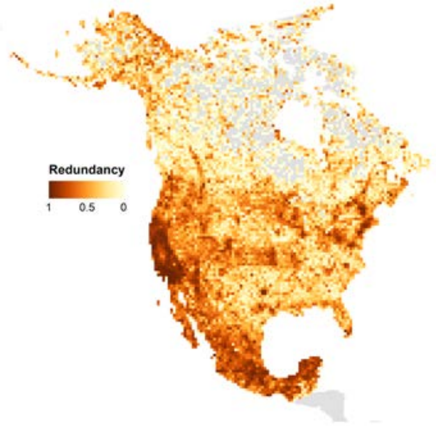
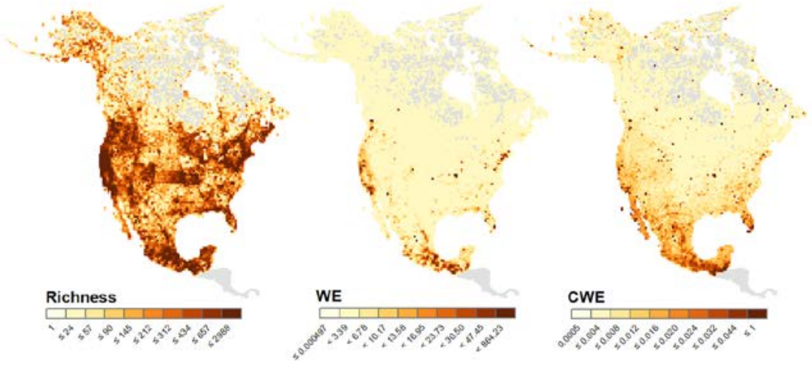


Fig. 2. Raw Richness, WE (∑ 1/range size of each OTU), and CWE (WE/richness).



This article is protected by copyright. All rights reserved.

Accepted Article

Fig. 3. Raw PD, PE, and CPE (PE/PD). These are very similar to the raw values for OTUs (fig. 2), which is to be expected if OTUs are being randomly sampled from the tree. Randomization tests are needed to detect locations where OTUs are not a random sample.

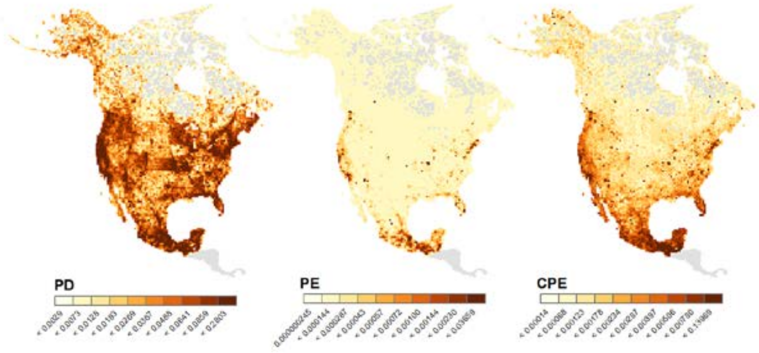
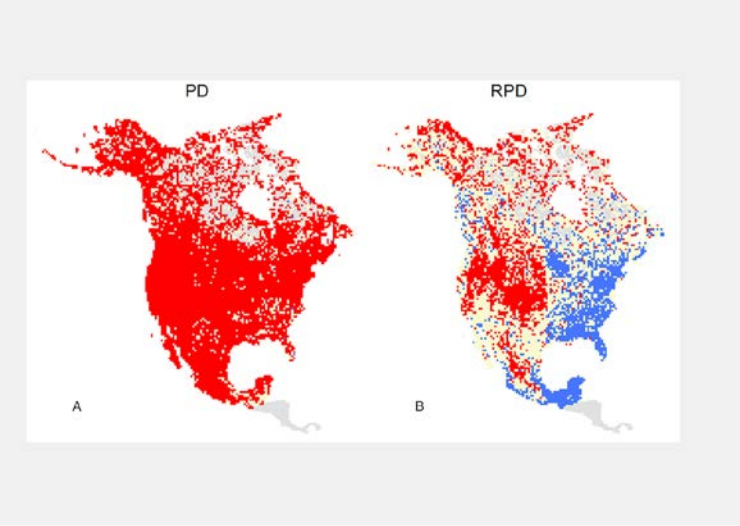


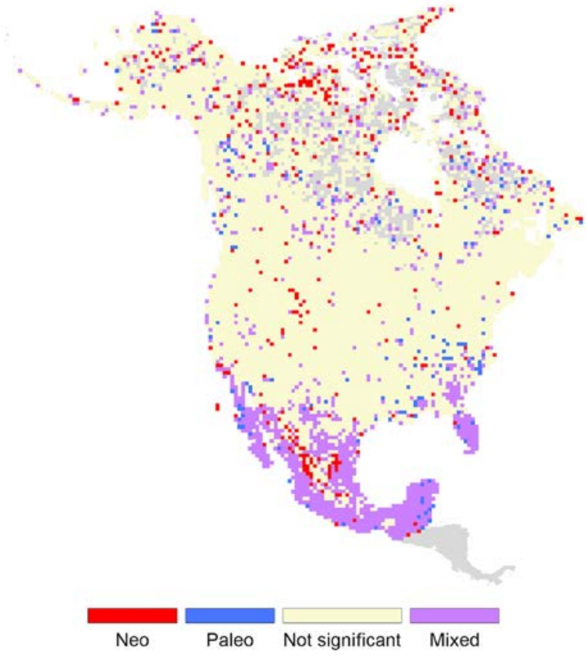
Fig. 4. A, PD significance with unconstrained randomization. Virtually the entire continent is significantly low in PD, meaning that taxa are more closely related than expected by chance. **B,** RPD significance with unconstrained randomization. Areas in blue have a concentration of significantly longer branches then expected; areas in red have a concentration of significantly shorter branches than expected. Note the striking difference between the western and eastern parts of the continent.



This article is protected by copyright. All rights reserved.

Accepted Article

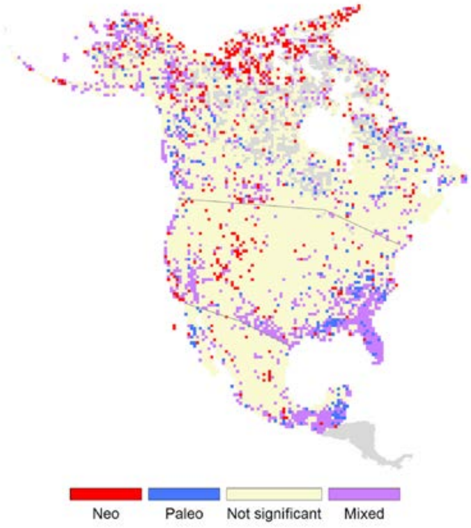
Fig. 5. CANAPE results, with unconstrained randomization. All cells that are colored are significantly high in PE (first step in CANAPE), then classified into three categories (second step in CANAPE): concentrations of rare short branches (neo-endemism), concentrations of rare long branches (paleo-endemism), and mixtures of the two.



This article is protected by copyright. All rights reserved.

Accepted Article

Fig. 6. CANAPE results, with randomization constrained within three latitudinal polygons indicated by the lines in the figure. All cells that are colored are significantly high in PE (first step in CANAPE); then classified into three categories (second step in CANAPE): concentrations of rare short branches (neo-endemism), concentrations of rare long branches (paleo-endemism), and mixtures of the two.



This article is protected by copyright. All rights reserved.

Accepted Article

Fig. 7. Range-weighted phylogenetic turnover among those cells found to be significant centers of endemism in CANAPE with the unconstrained randomization (shown in fig. 5).

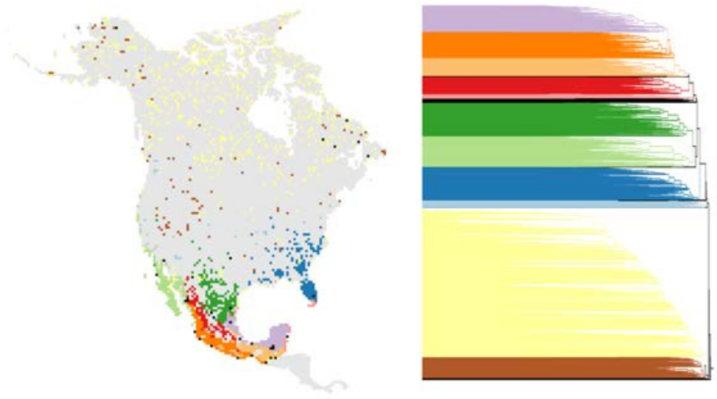
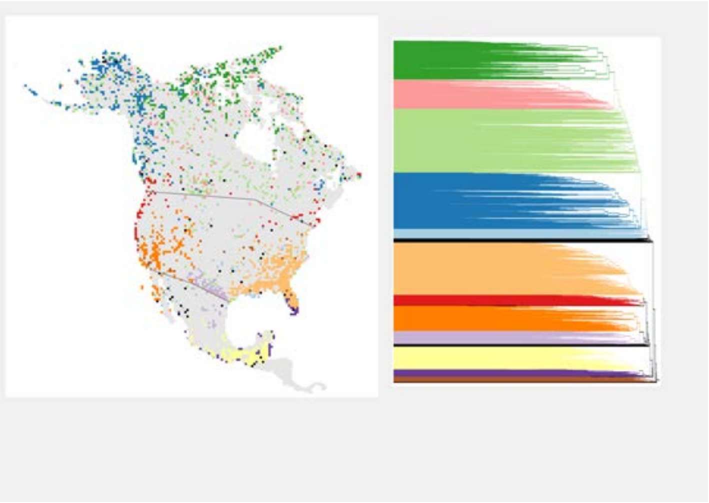


Fig. 8. Range-weighted phylogenetic turnover among those cells found to be significant centers of endemism in CANAPE with the randomization constrained within three latitudinal polygons indicated by the lines in the figure (shown in fig. 6).



This article is protected by copyright. All rights reserved.

View publication stats