Widespread sampling biases in herbaria revealed from large-scale digitization

Barnabas H. Daru[1]*, Daniel S. Park[1]*, Richard B. Primack[2], Charles G. Willis[1], David S. Barrington[3], Timothy J. S. Whitfeld[4], Tristram G. Seidler[5], Patrick W. Sweeney[6], David R. Foster[7], Aaron M. Ellison[7,8] and Charles C. Davis[1]

[1]Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge, MA 02138, USA; [2]Biology Department, Boston University, Boston, MA 02215, USA; [3]Pringle Herbarium, Plant Biology Department, University of Vermont, Torrey Hall, 27 Colchester Ave, Burlington, VT 05405, USA; [4]Brown University Herbarium, Department of Ecology and Evolutionary Biology, Brown University, 34 Olive Street, Box G-B225, Providence, RI 02912, USA; [5]Biology Department, University of Massachusetts, 611 North Pleasant Street, Amherst, MA 01003, USA; [6]Division of Botany, Peabody Museum of Natural History, Yale University, New Haven, CT 06511, USA; [7]Harvard Forest, Harvard University, 324 North Main Street, Petersham, MA 01366, USA; [8]Tropical Forests & People Research Centre, University of the Sunshine Coast, Maroochydore, Qld 4558, Australia

Author for correspondence:
Barnabas H. Daru
Tel: +1 857 218 0117
Email: barnabas_daru@fas.harvard.edu

Summary

- Nonrandom collecting practices may bias conclusions drawn from analyses of herbarium records. Recent efforts to fully digitize and mobilize regional floras online offer a timely opportunity to assess commonalities and differences in herbarium sampling biases.
- We determined spatial, temporal, trait, phylogenetic, and collector biases in *c.* 5 million herbarium records, representing three of the most complete digitized floras of the world: Australia (AU), South Africa (SA), and New England, USA (NE).
- We identified numerous shared and unique biases among these regions. Shared biases included specimens collected close to roads and herbaria; specimens collected more frequently during biological spring and summer; specimens of threatened species collected less frequently; and specimens of close relatives collected in similar numbers. Regional differences included overrepresentation of graminoids in SA and AU and of annuals in AU; and peak collection during the 1910s in NE, 1980s in SA, and 1990s in AU. Finally, in all regions, a disproportionately large percentage of specimens were collected by very few individuals. We hypothesize that these mega-collectors, with their associated preferences and idiosyncrasies, shaped patterns of collection bias via 'founder effects'.
- Studies using herbarium collections should account for sampling biases, and future collecting efforts should avoid compounding these biases to the extent possible.

Introduction

Herbaria contain a wealth of information about the ecological and evolutionary history of living and extinct species (Funk, 2003). Despite the continuous decline in plant collecting and declining support for herbaria (Dalton, 2003; Prather *et al.*, 2004a,b), there has been a recent surge of studies leveraging herbarium collections for diverse research projects not focused on systematics (Pyke & Ehrlich, 2010; Lees *et al.*, 2011; Feeley, 2012; Lavoie, 2013; Hart *et al.*, 2014). These studies include studies on plant demography, current and future species distributions, and temporal changes in phenology and morphology (e.g. Miller-Rushing *et al.*, 2006; Newbold, 2010; Pyke & Ehrlich, 2010; Lavoie, 2013; Staats *et al.*, 2013; Davis *et al.*, 2015; Willis *et al.*, 2017a,b).

Ideally, herbarium collections used for these studies would include statistically unbiased samples of plant diversity across space and time. However, as the majority of specimens were collected for qualitative taxonomic and/or systematic inquiries, they were usually collected nonrandomly and sampling designs were rarely quantified (Wolf *et al.*, 2011; Schmidt-Lebuhn *et al.*, 2013). Because nonrandom samples may be statistically biased, analyzing them without accounting for biases might lead to spurious results (Syfert *et al.*, 2013).

Sampling biases fall into several broad categories. Taxonomic or phylogenetic bias is the unbalanced sampling of certain taxa or clades over others, typically resulting from the scientific interests of a collector or the attractiveness of plants (Hortal *et al.*, 2007). Geographic bias occurs when specimens are collected more frequently in one place than another, often because of differential accessibility (Hijmans *et al.*, 2000). Temporal bias occurs when collection activity is favored in certain years or parts of the year (Cotterill *et al.*, 1994; Funk & Morin, 2000; Norris *et al.*, 2001). Meyer *et al.* (2016) evaluated worldwide terrestrial plant occurrence data using 120 million records from the Global

Biodiversity Information Facility (GBIF; Edwards *et al.*, 2000). Their analyses revealed large taxonomic gaps in global plant occurrence data (< 25% of species of land plants were sampled); extensive spatial gaps across regions that harbor high concentrations of plant diversity, especially in Asia, Central Africa, and Amazonia; and strong temporal discontinuities in occurrence records across decades, all of which can hamper inferences about the effects on plants of recent and future environmental change.

Although Meyer *et al.*'s (2016) study represents the most comprehensive effort to assess biases in plant collections at a global scale to date, the vast majority of herbarium collections have not been digitized, and of those that have been, many are unavailable, in whole or in part, on GBIF. Thus, Meyer *et al.*'s (2016) assessment of biases may itself be biased, or may inaccurately reflect biases in more complete, regional botanical collections that have been more fully mobilized. Furthermore, over two-thirds of the plant records in GBIF are not tied to physical specimens, and thus cannot be easily validated by others (Cotterill, 1995). For these reasons, we suspect that an analysis of finer grained collection data, focused on specific regions that have been predominantly digitized and validated, may reveal clearer patterns of sampling biases between regions than the global trends identified by Meyer *et al.* (2016) (cf. Hijmans *et al.*, 2000 for Bolivian potatoes; *Solanum*).

Expanding upon Meyer *et al.*'s work, we explored spatial, temporal, and taxonomic/phylogenetic sampling biases in collections from three of the most extensively collected, digitized, and mobilized regional floras in the world: South Africa (SA), Australia (AU), and the New England (NE) region of the USA. The SA flora is a compilation of digitized herbarium specimens from all major herbaria across the country available in a single online portal (South African National Biodiversity Institute (SANBI), 2016; le Roux *et al.*, 2017). The Australian Virtual Herbarium (AVH, 2016) is the main database for AU. It contains digitized herbarium specimens from all the major herbaria in AU. The Consortium of the Northeast Herbaria database contains digitized specimens from 15 participating herbaria in the NE region of the USA (Schorn *et al.*, 2016). We also examined trait bias – sampling bias attributable to intrinsic life-history characteristics, including life cycle (annual vs perennial), plant height, growth form (woody vs herbaceous), and species conservation status. Finally, we examined the contributions of individual collectors to each flora. We identified biases in all five of these categories within each of these regional floras. Our results revealed both commonalities and differences in regional collection biases and identified new sampling foci as collections grow in the future.

Materials and Methods

Sources and description of data

We obtained 12 488 200 herbarium specimen records of vascular plants from AU (Australia Virtual Herbarium [AVH], 2016); 2 049 905 herbarium specimen records from SA including Lesotho and Swaziland (South African National Biodiversity Institute (SANBI), 2016; and 879 388 herbarium specimen records from the NE (USA) flora (Consortium of Northeastern Herbaria (CNH), 2016). The records were cleaned in two steps (Supporting Information Fig. S1). First, we standardized the taxonomy of all species using the TAXONOMIC NAME RESOLUTION SERVICE v.4.0 (Boyle *et al.*, 2013). This online tool corrects and standardizes plant names against reference taxonomies, such as Missouri Botanical Garden's Tropicos (http://tropicos.org/) database or the PlantList (http://theplantlist.org/). Second, we removed specimens that were duplicates from the same collection locality and date; specimens with clearly erroneous locations (i.e. in oceans); specimens with zero coordinates and occurrences that fell outside the boundaries of our study; specimens missing exact collection date or georeferenced location data; and field observation records not tied to a physical specimen. Following this data cleaning, we retained 32% of the initial specimens for further analysis: 24% of the AU records (31 966 taxa; 2 958 195 records); 49% (20 824 taxa; 1 008 206 records) of the SA records; and 75% (3719 taxa; 661 370 records) of the NE records.

Analyses

Spatial biases First, we evaluated the density of sampling localities across the focal regions using Delaunay triangulation polygons, which measure the land area covered by each sampling locale (Fortune, 1992). Larger triangles indicate sparser collecting effort, whereas smaller triangles indicate more concentrated effort. Second, we examined infrastructure bias by calculating the minimum distance of each collection locality to the nearest major road (GADM, 2015) and herbarium (following Thiers, 2016). Our data set of roads derives from the publicly available Digital Chart of the World (http://maproom.psu.edu/dcw/), which was compiled by the US Defense Mapping Agency from 1 : 1 000 000 scale paper maps (ESRI, 1992). All roads appearing at this scale were included in our analyses. Although this data set includes only larger roads and has not been updated since 1992, it probably represents the most comprehensive digital record of roads around the world. We then compared these distances to those generated by a null model (1000 iterations) in which the same number of sample points was randomly (Poisson) distributed across each geographic region. Third, we mapped geographic biases in sampling density, defined as areas of excessive (hotspots) or insufficient (coldspots) collection (Hijmans *et al.*, 2000). Hotspots and coldspots were determined at a spatial grain of $0.25° \times 0.25°$ based on the number of specimens per grid cell, and identified using the 2.5% threshold (Orme *et al.*, 2005; Ceballos & Ehrlich, 2006; Daru *et al.*, 2015), based, respectively, on the 97.5[th] and 2.5[th] percentile values for the number of specimens collected per grid cell. Spatial distance calculations were computed with the functions DIST2LINE and SPDISTS in the R packages SP (Bivand *et al.*, 2013) and GEOSPHERE (Hijmans, 2015), respectively. In our final predictive model of sampling density, we also included human population density (CIESIN, 2016), sampling localities, infrastructure (distance to herbaria and roads), number of specimens collected, and elevation.

Temporal bias For each regional flora, we explored bias at several temporal scales. Collection dates ranged from 20 May 1664 to 9 January 2016 (AU), 15 November 1656 to 6 June 2016 (SA), and 28 July 1687 to 4 May 2016 (NE). We hypothesized that collectors tended to avoid fieldwork in unfavorable conditions (e.g. winter and wartime) or on certain days of the week (e.g. weekdays for nonprofessional botanists). To test for temporal bias, we first recoded collection dates as days of the week (Sunday = 1, Monday = 2, etc.), and day of the year (DOY; where 1 January = 1 DOY and 31 December = 365 DOY, etc.). We then used a Rayleigh test of directional statistics in the R package CIRCULAR (Agostinelli & Lund, 2013) to test whether each of these collection dates was randomly distributed against all dates spanning the entire duration of plant collection. If $P < \alpha = 0.05$, we rejected the null hypothesis of temporal uniformity at scales of weeks, days of the year, or decades.

Trait bias We used customized R scripts to harvest information on growth duration (annual vs perennial), growth form (woody vs herbaceous), and height for each species from online regional databases (all accessed in June 2016), including: New South Wales Flora Online (http://plantnet.rbgsyd.nsw.gov.au); JSTOR Global Plants (https://plants.jstor.org); Atlas of Living Australia (http://bie.ala.org.au); Plants of Southwestern Australia (http://keys.lucidcentral.org); the African Plant Database (http://ville-ge.ch); Plants of Southern Africa (http://plantzafrica.com); Plant Resources of Tropical Africa (http://www.prota4u.org); Flora of North America (http://efloras.org); and the USDA Plants Database (http://plants.usda.gov). We then manually checked these data for inconsistencies in terminologies for defining certain traits. For example, 'vines' vs 'lianas' for climbers, 'forbs' vs 'herbs' for herbaceous life forms, and 'biennial' for perennial growth duration. Extinction risk assessments for each species were obtained from the International Union for the Conservation of Nature (IUCN) Red List database (www.iucnredlist.org, accessed August 2016), which uses the following categories: Data Deficient (DD), Least Concern (LC), Lower Risk/Conservation Dependent (LR/CD), Near Threatened (NT), Vulnerable (VU), Endangered (EN), Critically Endangered (CR), and Extinct (EX). We grouped these narrow categories into two broader threat categories, threatened (EX + CR + EN + VU) and not threatened (LR/CD + NT + LC), following Yessoufou et al. (2012).

Trait bias was evaluated using a chi-squared test to contrast the number of observed specimens collected per species with the abundance of a species if specimen collection was equal across all species for each trait category. Because of dramatically unequal sampling effort in some species – for example *Senna artemisioides* with 10 167 specimens vs *Eucalyptus cordieri* with only one – and the low coverage of taxa with available trait data, we randomly sampled 50 specimens from each available species with trait data using 1000 randomizations. Species with < 50 specimens were excluded from this analysis.

Phylogenetic bias We assessed phylogenetic signal in collection frequency as a measure of phylogenetic bias using two

different tests following Wolkovich et al. (2013). A strong phylogenetic signal – closely related species sharing similar collection frequency – would suggest phylogenetic bias in collections. We first assembled a phylogeny using PHYLOMATIC (Webb & Donoghue, 2005), enforcing a topology that assumed the Angiosperm Phylogeny Group (APG) III (APG III, 2009) backbone (tree R20120829). This phylogeny included all species in our analysis, but provided only an approximate degree of relatedness based on taxonomic hierarchy at family level; many relationships, especially within genera, were unresolved. This is problematic because recent theoretical and empirical studies have shown that a lack of resolution in a community phylogeny may mask significant patterns by reducing statistical power (Schaefer et al., 2011; Daru et al., 2017) or suggest significant phylogenetic patterns that are not supported by more completely resolved phylogenies (Davies et al., 2012).

To alleviate these concerns, we also tested for phylogenetic bias by including only those species sampled in the dated molecular phylogeny inferred from seven genes for 32 223 plant species (Zanne et al., 2014). Although this phylogeny has been criticized (Edwards et al., 2015), it nonetheless represents the single largest phylogeny to date for flowering plants. The taxon sampling for testing phylogenetic bias included 5814 species from AU, 3568 from SA, and 3700 from NE.

We estimated phylogenetic signal using three common metrics: Abouheif's C_{mean} statistic (Abouheif, 1999), Blomberg's K (Blomberg et al., 2003), and Pagel's lambda (λ) (Pagel, 1999). Significance was assessed by comparing observed values to a null distribution created by shuffling the trait values across the tips of the phylogeny 1000 times. Pagel's λ uses a maximum-likelihood method with branch-length transformation to estimate the best fit of a trait against a Brownian model. Values of Pagel's λ range from 0 (no phylogenetic signal) to 1 (strong phylogenetic signal). Both Blomberg's K (a significant phylogenetic signal is indicated by a K value > 1) and Pagel's λ were calculated using the R package PHYTOOLS (Revell, 2012). Abouheif's C_{mean} was calculated using ADEPHYLO (Jombart & Dray, 2008). We tested the sensitivity of our analysis by exploring phylogenetic signal in collecting effort across nine well-sampled clades as represented in NE: Asteraceae, Brassicaceae, Cyperaceae, Ericaceae, Fabaceae, Lamiaceae, Poaceae, Ranunculaceae, and Rosaceae.

In addition to phylogenetic signal, we also used phylogenetic generalized least squares (PGLS) regressions in the R package CAPER (Orme et al., 2012) to model collecting effort per species in each region as a function of species evolutionary ages, evolutionary distinctiveness (ED), and 'evolutionary distinctiveness and global endangerment' (EDGE; Isaac et al., 2007). Species ages were measured as the length of terminal branches (BL) linking species on a phylogenetic tree. ED measures the degree of phylogenetic isolation of a species, whereas the EDGE metric was determined by calculating the ED score of each species (Isaac et al., 2007) and combining it with global endangerment (GE) from IUCN conservation categories: $EDGE = \ln(1 + ED) + GE \times \ln(2)$, where GE represents expected probability of species extinction over a 100-yr period (Redding & Mooers, 2006) categorized as follows: Least Concern = 0.001, Near Threatened and

Conservation Dependent = 0.01, Vulnerable = 0.1, Endangered = 0.67, and Critically Endangered = 0.999.

Last, we examined the phylogenetic structure of collecting efforts across decades to test for patterns of phylogenetic overdispersion and clustering through time. Temporal phylogenetic structure by decade (i.e. 1901–1910, 1911–1920, etc.) was evaluated using the net relatedness index (NRI) and nearest taxon index (NTI; Webb *et al.*, 2002, 2008). NRI describes a tree-wide pattern of phylogenetic dispersion, whereas NTI evaluates phylogenetic structure towards the tips of the phylogeny. Negative values of NRI or NTI indicate phylogenetic overdispersion whereas positive values indicate phylogenetic clustering.

Collector bias We determined collector bias by tabulating the number of specimens amassed by each collector in all three floras. We then examined Pearson's product-moment correlation between the number of specimens collected per collector and the number of species collected per collector.

Computation and availability of data and code

All statistical analyses were conducted using the RESEARCH COMPUTING CLUSTERS of Harvard University (https://rc.fas.harvard.edu/). Data files and custom R scripts are available from the Harvard Forest Data Archive, dataset HF296 (http://harvardforest.fas.harvard.edu/data-archive).

Results

Spatial bias

High sampling density was observed in southeastern and southwestern AU, the Cape region of SA, and two of the six NE states (Connecticut and Massachusetts) relative to other parts of those regions (Fig. 1a–c). When we weighted each sampling locale by the number of specimens, we found a mismatch between hotspots (top 2.5% quantiles) and coldspots (lowest 2.5%
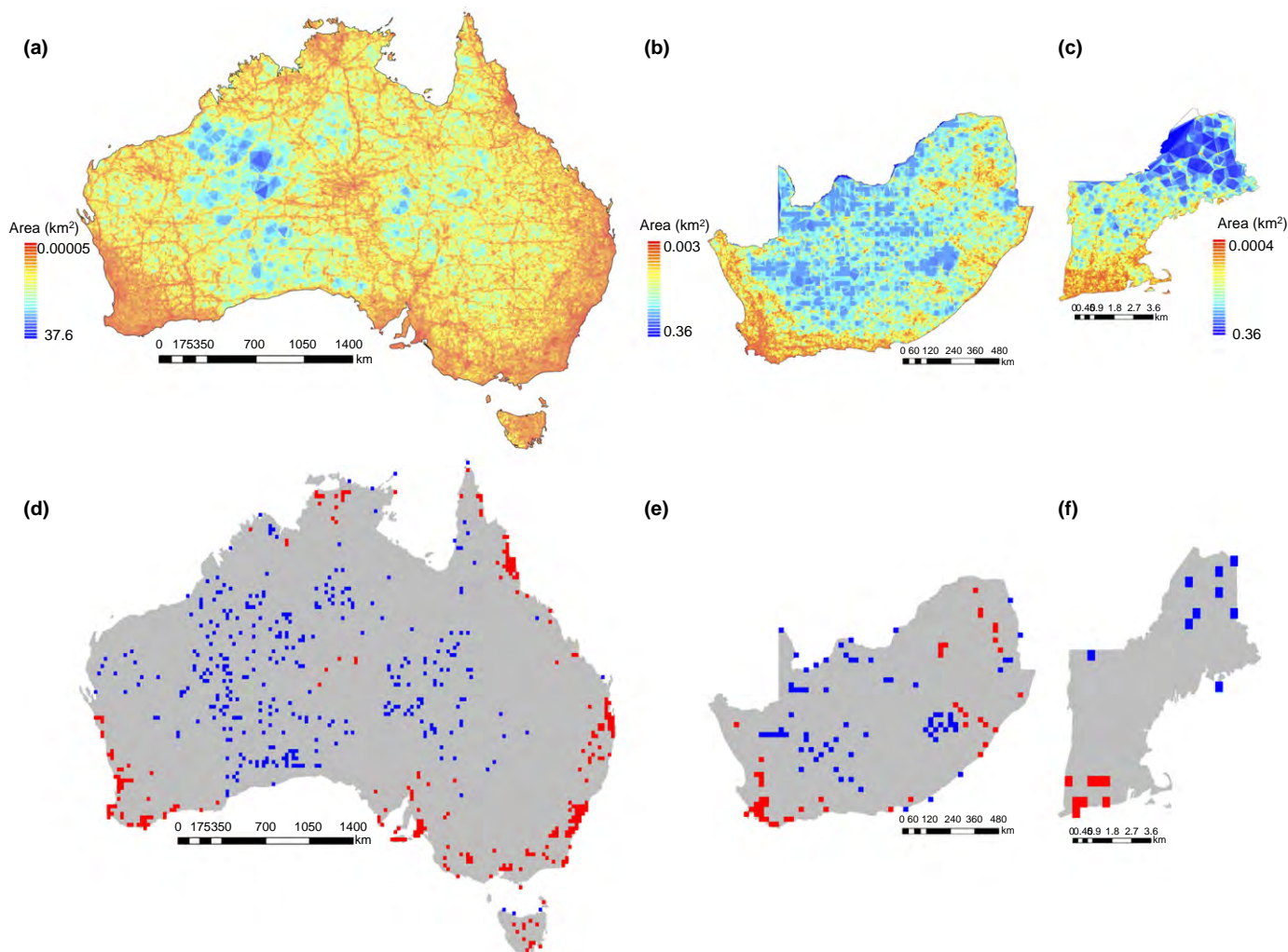

Fig. 1 Spatial bias in herbarium collections. (a−c) The geographic distribution of herbarium collecting activity depicting the spatial variation in sampling effort using Delaunay polygon tiles for (a) Australia (*n* = 857 245 locales), (b) South Africa (*n* = 61 130 locales), and (c) New England (*n* = 130 374 locales). (d−e) Hotspots (red) and coldspots (blue) of herbarium sampling within quarter degree grids for (d) Australia, (e) South Africa and (f) New England. The hotspots and coldspots are the top and lowest 2.5% quantiles, respectively, of the number of specimens per locale.

quantiles) of sampling intensity (Fig. 1d–f), suggesting that hotspots and coldspots were not randomly distributed. Hotspots of collecting tended to cluster around coasts in AU and SA, whereas coldspots were abundant in interior areas. In NE, hotspots were concentrated in the south and coldspots occurred in the north.

Herbarium specimens tended to be collected closer than expected to roads and herbaria ($P < 0.01$; Fig. 2a,b). More than 50% of herbarium specimens were collected within 2 km of roadsides in all three floras ($P < 0.01$; Fig. 2a). Moreover, distance to herbaria explained 45% of the variance in collecting effort in AU, 29% in SA and 12.3% in NE, with a higher density of specimens closer to herbaria (Table 1). Despite substantial gradients in altitudes in each region (−15 to 2022 m above sea level (asl) in AU; 1 to 3254 m asl in SA; and −3 to 1485 m asl in NE), most specimens were collected below 500 m asl in AU and NE (81, 44, and 93% of specimens in AU, SA, and NE, respectively; Fig. 2c). We also found a negative correlation between collecting effort and altitude in AU and SA, suggesting a tendency for specimens to be collected at lower elevations. However, the opposite was true for NE, where more specimens tended to be collected at higher elevations than expected by chance (Table 1).

Temporal bias

There were historical biases in collection efforts in the three floras: low sampling until 1880 in AU and SA, and a burst of collections in NE in the early 20th Century (Fig. 3). Conversely, there was a dramatic increase in botanical collection in SA and AU after World War II, peaking in the 1980s and 1990s, respectively (Fig. 3). This peaking occurred $c.$ 100 yr after peak collection activity in NE. Seasonally, specimen collections were biased toward spring and summer for all three floras, with peak collections ranging from September to December in AU and SA (Rayleigh $Z = 0.189$ and $Z = 0.251$, respectively; both $P < 0.001$), and May to September in NE (Rayleigh $Z = 0.718$; $P < 0.001$; Fig. 4a). There was a significant trend toward collection on weekends (Saturdays and Sundays) in NE (Rayleigh test $Z = 1.0$; $P < 0.001$) and midweek in SA and AU (Rayleigh test $Z = 0.105$ and $Z = 1.0$, respectively; both $P < 0.001$; Fig. 4a).

Trait bias

Perennials were more frequently collected than annuals in terms of specimens per species in SA and NE; the opposite was true for AU where there was a greater tendency for annuals to be collected (Fig. 5a). Similarly, graminoid specimens per species were overrepresented relative to other habits in AU and SA, whereas herbs and trees were overrepresented in NE (Fig. 5b). Relatively short plants were more frequently represented than taller plants in all three floras: 79.3, 89.3 and 84.9% of the plants collected in AU, SA and NE, respectively, were < 5 m in height (Fig. 5c).

Threatened species were collected significantly less often than nonthreatened plants across all three floras (all $P < 0.001$; Fig. 5d).

Phylogenetic bias

It is possible that closely related species were collected more similarly (either undercollected or overcollected) than expected by chance. We tested this assumption phylogenetically and found a significant, but weak phylogenetic signal in the abundance of specimens per species across all three floras (Table 2). Specifically, closely related species tended to have a more similar number of specimens than expected (Table 2; Fig. 6). This phylogenetic bias was strongest in SA (Abuoheif's $C_{mean} = 0.15$ and $\lambda = 0.32$; both $P < 0.01$, but $K = 0.0013$ (ns)). For instance, in SA, collections from the genus *Protea* averaged 115 specimens per species whereas only two specimens were collected for species in the genus *Rytigynia* on average. Most *Agoseris* in NE were represented by < 10 specimens per species, whereas many fern genera were represented by high specimen numbers (e.g. *Onoclea* with 845 specimens/species). Australian collections showed the weakest phylogenetic bias (Abuoheif's $C_{mean} = 0.12$ and $\lambda = 0.18$; both $P < 0.01$, but $K = 0.00085$ (ns); Fig. 6). Phylogenetic signal varied at the family level as well in NE, with Asteraceae showing the strongest collection bias (Fig. 7), followed by Cyperaceae, Poaceae, and Rosaceae (Table S1). These families were represented by much higher collection numbers than other families.

EDGE was a significant predictor of collecting efforts in all three floras ($P < 0.001$), with variance ranging from 1.89% (NE) and 3.75% (AU) to 8.89% (SA). In general, EDGE species (distinctive, endangered species) were generally undercollected in terms of specimens per species (Table 3).

Lastly, floristic collecting showed a general trend of phylogenetic clustering within decades for all three floras. The collection of different clades of plants was not evenly distributed across time. NTI was significantly positive in each flora, indicating that clustering occurred near the tips of the phylogeny (Fig. 3). We only observed significant phylogenetic clustering at the deeper nodes of the phylogeny, as indicated by NRI, in SA (Fig. 3); deeper phylogenetic clustering was weak in NE and AU (Fig. 3).

Collector bias

The number of specimens per collector was highly skewed (Fig. 8). In AU, > 50% of the examined specimens were amassed by only 2% of the collectors, including A. C. Beauglehole (46 728 specimens), B. Hyland (32 019 specimens), and P. I. Forster (30 280 specimens; Fig. 8a). In SA, > 50% of the specimens were amassed by 9.5% of collectors, including J. P. H. Acocks (19 344 specimens), E. E. Esterhuysen (15 566 specimens), and E. E. Galpin (14 146 specimens; Fig. 8b). In NE, 50% of the specimens were contributed by 3.2% of the collectors, including L. J. Mehrhoff (19 149 specimens), M. L. Fernald (14 368 specimens), and A. S. Pease (12 238 specimens; Fig. 8c). The number of specimens amassed by these collectors was very strongly positively correlated with the number of species they collected, suggesting that these collectors were mainly doing general collecting rather than focusing on a particular group of plants ($r = 0.85$ in AU, 0.95 in SA and 0.84 in NE; all $P < 0.01$; Fig. S2).
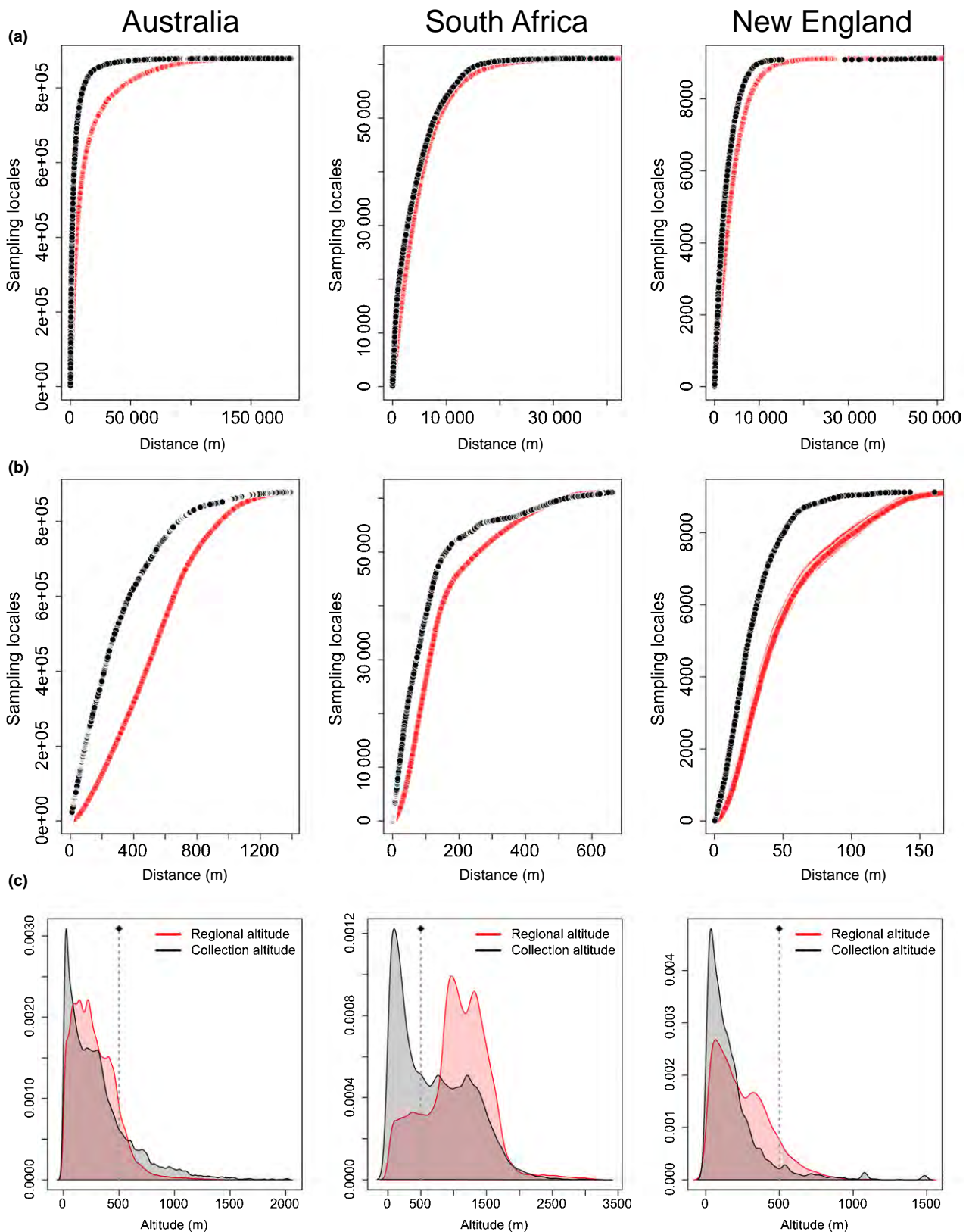
Fig. 2 Comparison of geographic sampling bias of herbarium records in relation to (a) the minimum distance to roads; (b) the minimum distance to herbaria; and (c) regional altitudes at sampling locales. Black lines in (a) and (b) correspond to sampling locales and red indicates an equal number of random points generated 1000 times. Dark gray shading in (c) corresponds to sampling locales and red shading corresponds to regional altitudes, that is, all other altitudes, for all three floras: Australia (left), South Africa (middle) and New England (right). The dotted line in (c) indicates an altitude of 500 m above sea level.

Table 1 Model coefficients for multiple regressions of collecting effort in terms of the number of specimens collected per locality

	Predictors (log$_{10}$-transformed)	Percentage of variance explained (%)	P-values	Model adjusted R^2	Model slope	Model intercept
Australia	Distance to roads	0.14	0.001	0.4571	−0.02	11.45
	Distance to herbaria	45.03	0.001		−0.89	
	Human population density	0.50	0.001		0.11	
	Altitude	0.041	0.001		−0.046	
South Africa	Distance to roads	0.00001	0.0003	0.3075	−0.011	11.33
	Distance to herbaria	29.13	0.001		−0.73	
	Human population density	0.0009	0.001		−0.03	
	Altitude	1.62	0.001		−0.15	
New England	Distance to roads	0.07	0.0009	0.17	0.13	7.03
	Distance to herbaria	12.3	0.001		−0.87	
	Human population density	4.68	0.001		0.30	
	Altitude	0.04	0.001		0.046	

Discussion

Historically, the primary function of herbaria has been to serve as a resource for botanists carrying out taxonomic and systematic research, allowing users to construct classifications of plants, verify identifications, determine the ranges and morphological characteristics of species, and develop local and regional floras (Greve *et al.*, 2016). Over time, new uses for specimens have arisen, and now more than ever, they are being used in ways that collectors rarely imagined (Pyke & Ehrlich, 2010; Lavoie, 2013; Nualart *et al.*, 2017; Rudin *et al.*, 2017; Willis *et al.*, 2017a,b). Accordingly, attempts to assess and categorize biases inherent in these collections have been made (Rich & Woodruff, 1992; Geri *et al.*, 2013; Schmidt-Lebuhn *et al.*, 2013; Meyer *et al.*, 2016; Stropp *et al.*, 2016). Among these, the most comprehensive investigation is by Meyer *et al.* (2016), who proposed an important conceptual framework for analyzing gaps and biases along taxonomic, geographic, and temporal dimensions. Although Meyer *et al.* (2016) focused more on observational records than herbarium collections, they uncovered numerous biases in 'digitally accessible information' regarding plants and provided an important baseline for evaluating and improving global floristic coverage in collection data. However, collection biases in large geographic areas are difficult to categorize because the collections' data are not yet digitized, and this may skew the global patterns of bias noted by Meyer *et al.* (2016). By focusing on three of the most well-collected and digitized floras in the world, we reduced effects of missing or unavailable data and, most importantly, could evaluate commonalities and differences in patterns of bias among regional collections.

Spatial bias

Our data confirmed the tendencies for botanists to collect along roadsides (e.g. Funk & Richardson, 2002), near herbaria (e.g. Hijmans *et al.*, 2000; Moerman & Estabrook, 2006; Pautasso & McKinney, 2007), in more accessible areas (Rich & Woodruff, 1992), and at lower elevations. Before automobiles became common in the 1920s, botanists often walked or rode domesticated animals to collection sites (Botkin, 1968; Belasco, 1979). As our modern infrastructure developed (e.g. roads, railroads, and cities that contain herbaria) and access to automobiles increased, spatial biases associated with infrastructure probably increased as botanists tended to travel and collect using automobiles (Everill *et al.*, 2014). Because roads are known to alter local environmental conditions and facilitate biological invasions (e.g. Forman & Alexander, 1998; Hui *et al.*, 2003; Griffith *et al.*, 2010; Li *et al.*, 2014) and botanists and herbaria predominate in cities, specimens collected in proximity to either are unlikely to represent a random sample across species distributions. Specifically, species collected along roadsides are likely to be overrepresented by species that thrive with disturbance, and underrepresented by forest interior and wetland species that are harmed by disturbance (Gutzwiller & Flather, 2011; Rivers-Moore & Cowden, 2012). As the road network continues to expand and as people become ever more concentrated in cities, this bias toward collecting near roads might become stronger in coming decades.

Collection bias toward lower elevations (< 500 m) was striking in SA and AU, despite extensive collection efforts in adjacent hyper-diverse hotspots such as the mountains in the Cape Fold Belt (SA), and Mount Lesueur-Eneabba (Western AU). This is probably attributable to the presence of the arid and relatively species-poor Great Karoo Plateau (SA) and Great Sandy Desert (AU), which each encompass over a third of the respective study sites, but account for only a small proportion of the biodiversity of each region. As a result, the low-elevation collection bias in the floras may reflect actual species abundance. In NE, the trend toward collecting at higher elevation might be a consequence of the strong tendency for botanists to visit the White Mountains and Mount Katadhin to collect alpine species.

Although we realize that patterns of species richness may not be randomly distributed across the landscape, accounting for underlying patterns of richness or abundance is difficult because our knowledge of such patterns often derives from (and is thus not independent of) these same (biased) collections. By comparing locations of samples (collections) against a Poisson set of points and specimens per species, and not total collection numbers, we tested only for the nonrandom distribution of collection
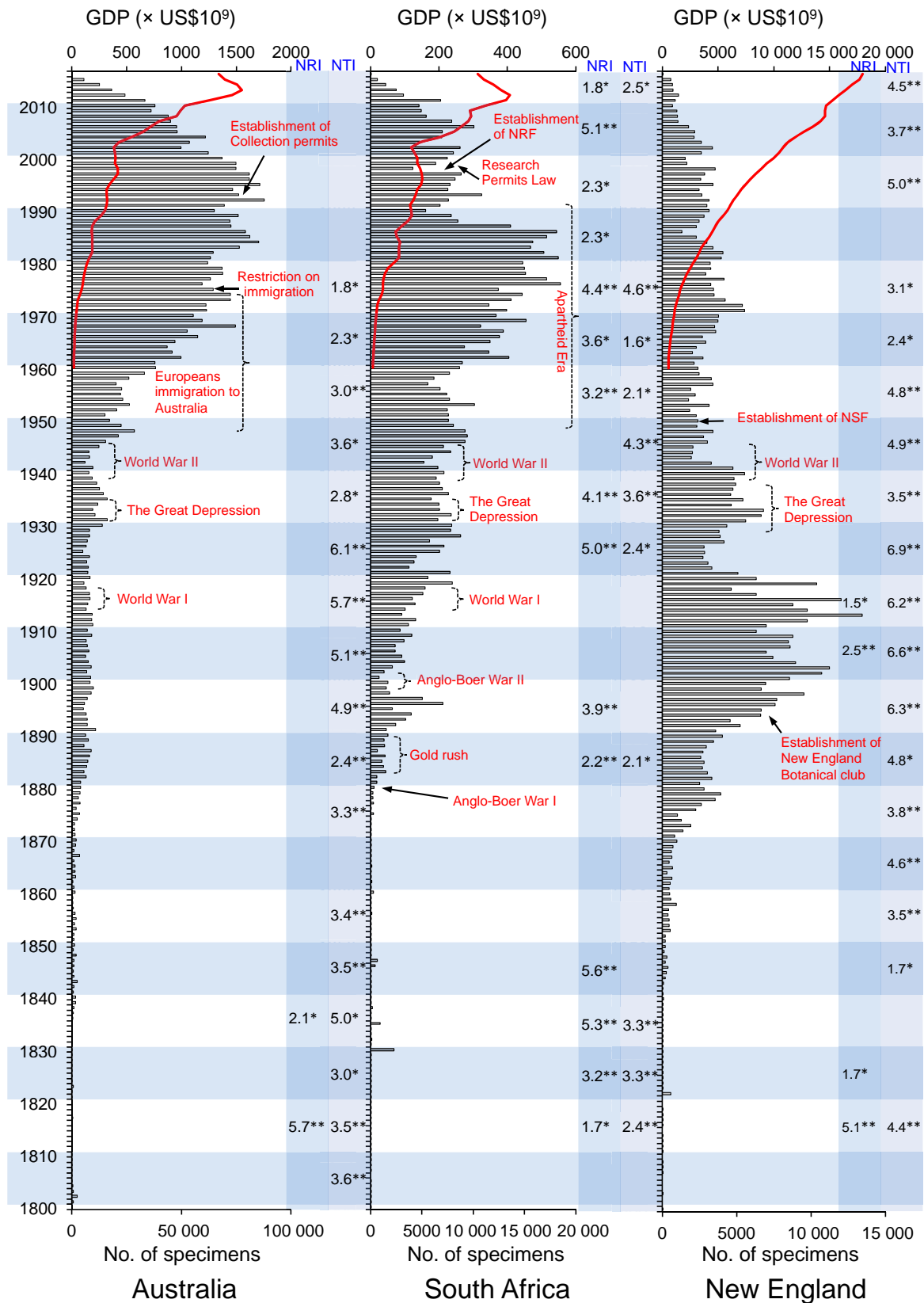
Fig. 3 Timeline of herbarium specimen collection density in relation to major historical events in time (indicated in red text) for the three floras: Australia, South Africa and New England. Analysis of phylogenetic structure through time by binning sequences of collection dates into decades and testing for overdispersion vs clustering is indicated in black font. The red trend line indicates the gross domestic product of each region. NRI, net relatedness index; NTI, nearest taxon index. **, $P < 0.001$; *, $P < 0.01$.

Fig. 4 Temporal biases in herbarium collections. (a) Comparison of density plots of collection dates by seasons of the year of herbarium records (blue line) with the dates spanning the entire duration of collection (red line); blue lines outside the red lines indicate overcollecting at a particular time of year. (b) Distribution of collection dates by days of the week for the three floras: Australia ($n = 4579\,321$ collection dates), South Africa ($n = 771\,991$ collection dates), and New England ($n = 562\,587$ collection dates).

locations on a landscape. And, indeed, we found that the collection locations were not spatially distributed randomly (Poisson) on the landscape. It is also possible that georeferencing might have introduced additional bias in some specimens. While ascertaining the degree of accuracy of georeferenced records might be challenging because such information is often unavailable, our cumulative curves are probably less affected.

Temporal and seasonal bias

Collections in AU and SA increased through time until a few decades ago, but those in NE peaked much earlier in the early 1900s. These differences between regional collection activities may parallel broader societal factors influencing plant collection. In NE, for example, the establishment of the New England Botanical Club during the 1890s (NEBC, 1899) preceded

a surge and peak in collecting activity associated with prolific botanical expeditions in the region coinciding with the 'Golden Age' of plant collecting in Europe and North America (Whittle, 1970; Musgrave *et al.*, 1999). In SA, collection efforts began much later, peaking during the Apartheid Era (1948–1994), and declined thereafter under the New Democratic Rule, concomitant with the general economic decline of the country and concern for public safety (Ferreira & Harmse, 2000; Lemanski, 2004). In AU, the mass immigration of Europeans in 1948 after World War II included numerous highly skilled professionals (Price, 1998; Leuner, 2007) and coincided with an enormous increase in botanical collecting. Botanical collecting may have declined more recently as a result of legislation in AU and SA to regulate collection activities, especially those designed to protect rare and endangered species.
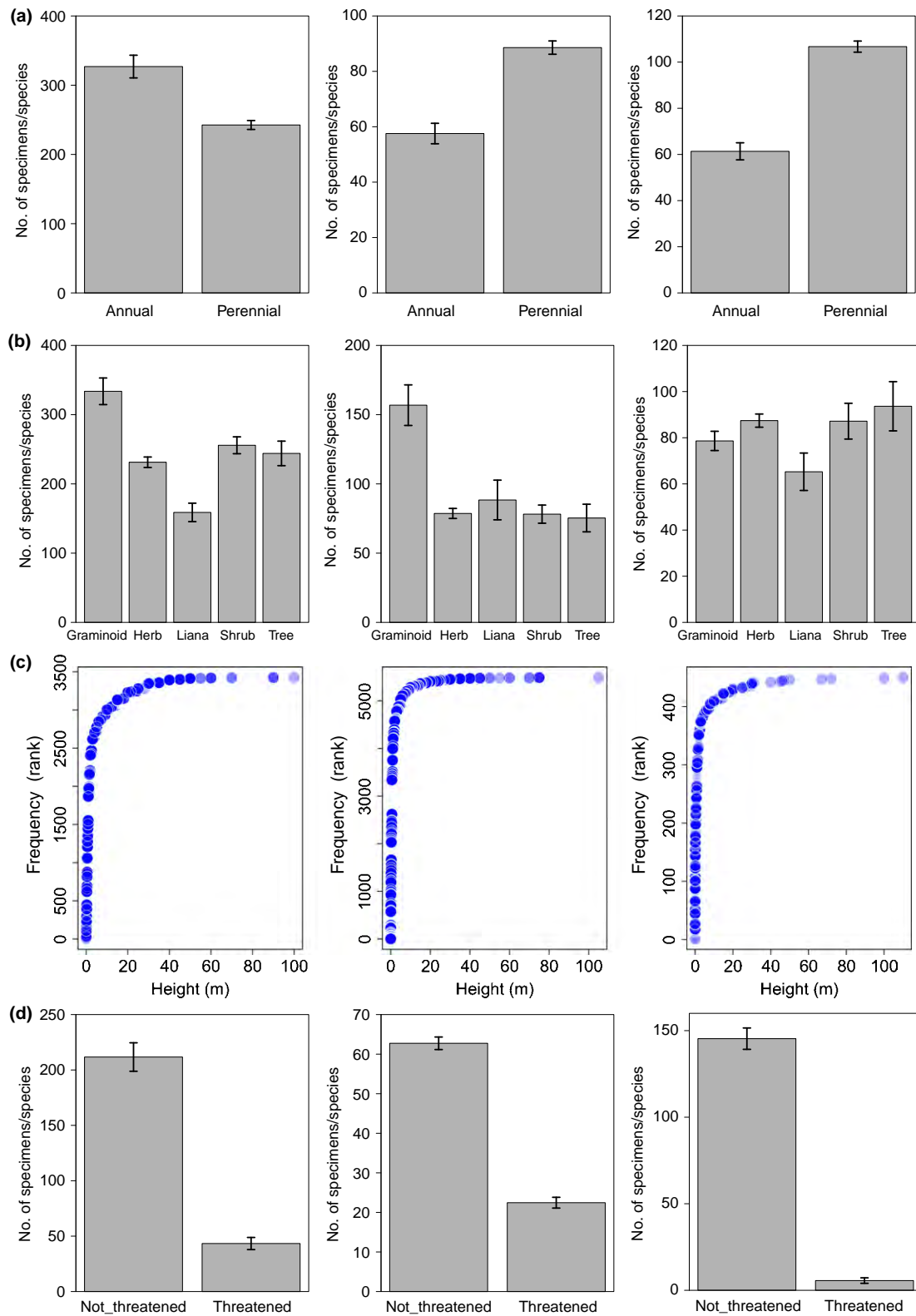
Fig. 5 Assessment of bias in plant traits: (a) growth duration; (b) growth form; (c) height; and (d) extinction risk for the floras of Australia (left pane), South Africa (middle pane) and New England (right pane). Error bars in (a, b, d) represent ± SE.

Collecting efforts within a season revealed common patterns of bias: specimens in all three regions were collected overwhelmingly in biological spring and summer. Sampling during these

time periods probably reflects efforts to collect plants in good flowering and fruiting condition. However, this seasonal bias probably overlooks key developmental transitions (e.g. Poethig,

Table 2 Results of the tests of phylogenetic signal in the number of specimens collected per species using three methods (Abouheif's C_{mean}, Blomberg's K and Pagel's λ)

	Australia ($n = 5814$ species)	South Africa ($n = 3568$ species)	New England ($n = 3700$ species)
Abouheif's C_{mean}	0.12**	0.15**	0.12**
Blomberg's K	0.00085[ns]	0.0013[ns]	0.0030[ns]
Pagel's λ	0.18**	0.32**	0.29**

Phylogenetic data are derived from Zanne et al. (2014). All tests are based on 1000 randomizations. **, $P < 0.001$; ns, $P > 0.05$.

2013), including bud formation, bud break, leaf out, fruit development, and leaf senescence (van der Schoot et al., 2014). Supporting this argument, these temporal patterns were most pronounced in NE, which experiences the harshest winter climates of the three regions. Plants collected during the winter season are almost always in dormant condition, and often lack the leaves and reproductive structures needed for taxonomic research. Collecting was also more likely during holidays and school vacations in NE and AU.

Trait bias

In all three regions, short to medium-height species were collected more frequently than tall species (> 5 m). This pattern is presumably related to the relative ease of collecting specimens from shorter, often herbaceous, species, and because reproductive materials are more accessible and potentially more abundant. Specimens of trees with woody twigs also are typically bulkier and more difficult to prepare, which may reduce their collection frequency.

Threatened species were also greatly underrepresented in all floras. This is perhaps not surprising given their limited abundance (Palmer et al., 2002) and imposed collecting restrictions

(Klemens & Thorbjarnarson, 1995; Pritchard, 1996; Gibbons et al., 2000; Robinson, 2001). However, it is also true that collectors sometimes oversample rare or threatened species because of their higher scientific value and avoid the more common ones (Garcillán et al., 2008; Garcillán & Ezcurra, 2011; Minteer et al., 2014). Regardless of past practices and contemporary formal restrictions, botanists now often avoid overcollection of such species by following informal guidelines and collecting plants only in areas with numerous individuals of the species (Iwanycki, 2009). Although great care in collecting rare plants is important, undercollection of rare species may lead to incorrect extinction risk assessments (i.e. that the species is rarer than it actually is) and greatly limit opportunities to glean historic population and biogeographic data to guide species conservation and restoration.

Annuals were overrepresented relative to perennials in herbarium collections in AU; the opposite was observed in SA and NE. There was also a high representation of graminoids in herbarium collections in AU and SA. This result may stem from the higher likelihood of common species being collected multiple times by different individuals or expeditions. Along these lines, much of AU is dominated by annual grasses, and the savannas of SA are populated by a variety of native and nonnative perennial grasses interspersed with forbs and woody plants (Bond & Parr, 2010). New England, in contrast, is generally forested and has an abundance of shade-tolerant shrubs and perennial herbs. Graminoids are also considered harder to identify and may be avoided by nonspecialists. Lianas and vines simultaneously represent the smallest proportion of growth forms and comprise the least number of specimens per species in all three floras. Such trait-based biases in botanical collections not only influence our perception of species abundance and range, but can also lead to erroneous estimations of functional diversity and ecosystem services, especially for studies relying on specimen databases (Schmidt-Lebuhn et al., 2013). Whether herbarium records represent true patterns of abundance and diversity remains difficult to untangle from human-mediated collecting biases. However, the large differences
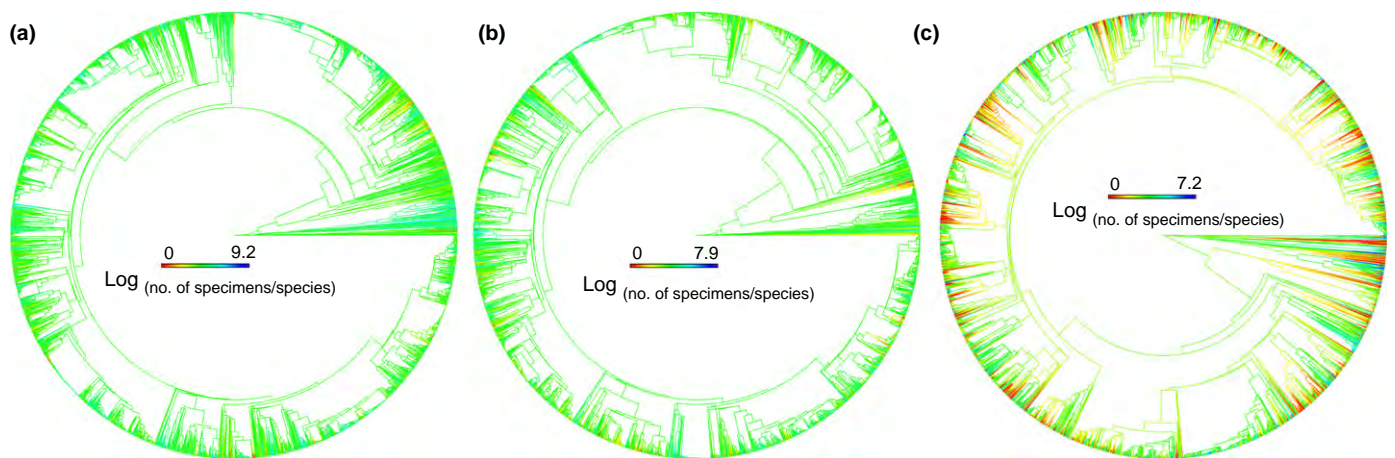

Fig. 6 Distribution of phylogenetic bias, the tendency of closely related species to be similarly collected in herbarium records for three floras: (a) Australia; (b) South Africa; and (c) New England. Collecting effort is not phylogenetically random, but tends to be clustered in a few selected lineages. The color scales correspond to $\log_{10}$ number of specimens per species and range from red (low number of specimens per species) to blue (high number of specimens per species).
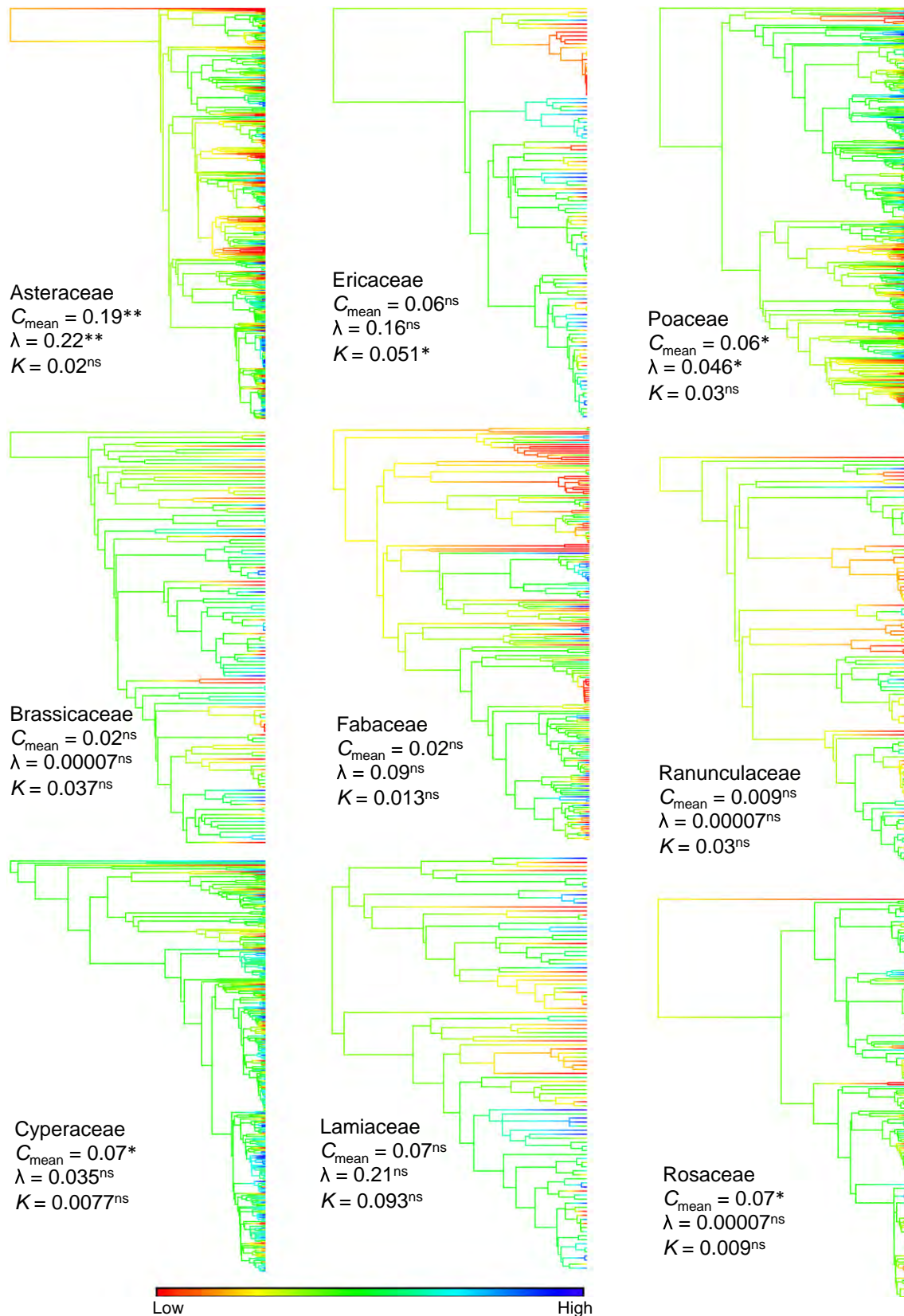
Fig. 7 Phylogenetic bias in collection frequency for exemplar families in New England flora. Phylogenetic bias is indicated by significant phylogenetic signal in at least one of three metrics (Abouheif's C_{mean}, Blomberg's K and Pagel's λ). The color bar illustrates values within families: $\log_{10}$ numbers of specimens per species and ranges from red (low number of specimens per species) to blue (high number of specimens per species). **, $P < 0.001$; *, $P < 0.01$; ns, $P > 0.05$.

Table 3 Multiple regressions of phylogenetic generalized least squares of collecting effort (frequency) of herbarium specimens with phylogenetic metrics of species uniqueness

	Predictors (log$_{10}$-transformed)	Percentage of variance explained (%)	P-values	Model adjusted R^2	Model slope	Model intercept
Australia	BL	1.36	0.7	0.049	0.035	4.37
	ED	0.2	0.008		0.44	
	EDGE	3.75	< 0.001		−1.23	
South Africa	BL	0.47	0.3	0.09	−0.063	3.63
	ED	0.000015	0.001		0.63	
	EDGE	8.89	< 0.001		−1.3	
New England	BL	0.09	0.94	1.70E-02	−0.0052	3.89
	ED	0.054	0.0045		0.79	
	EDGE	1.87	< 0.001		−2.28	

BL, terminal branch length; ED, evolutionary distinctiveness; EDGE, evolutionary distinctiveness and global endangerment.
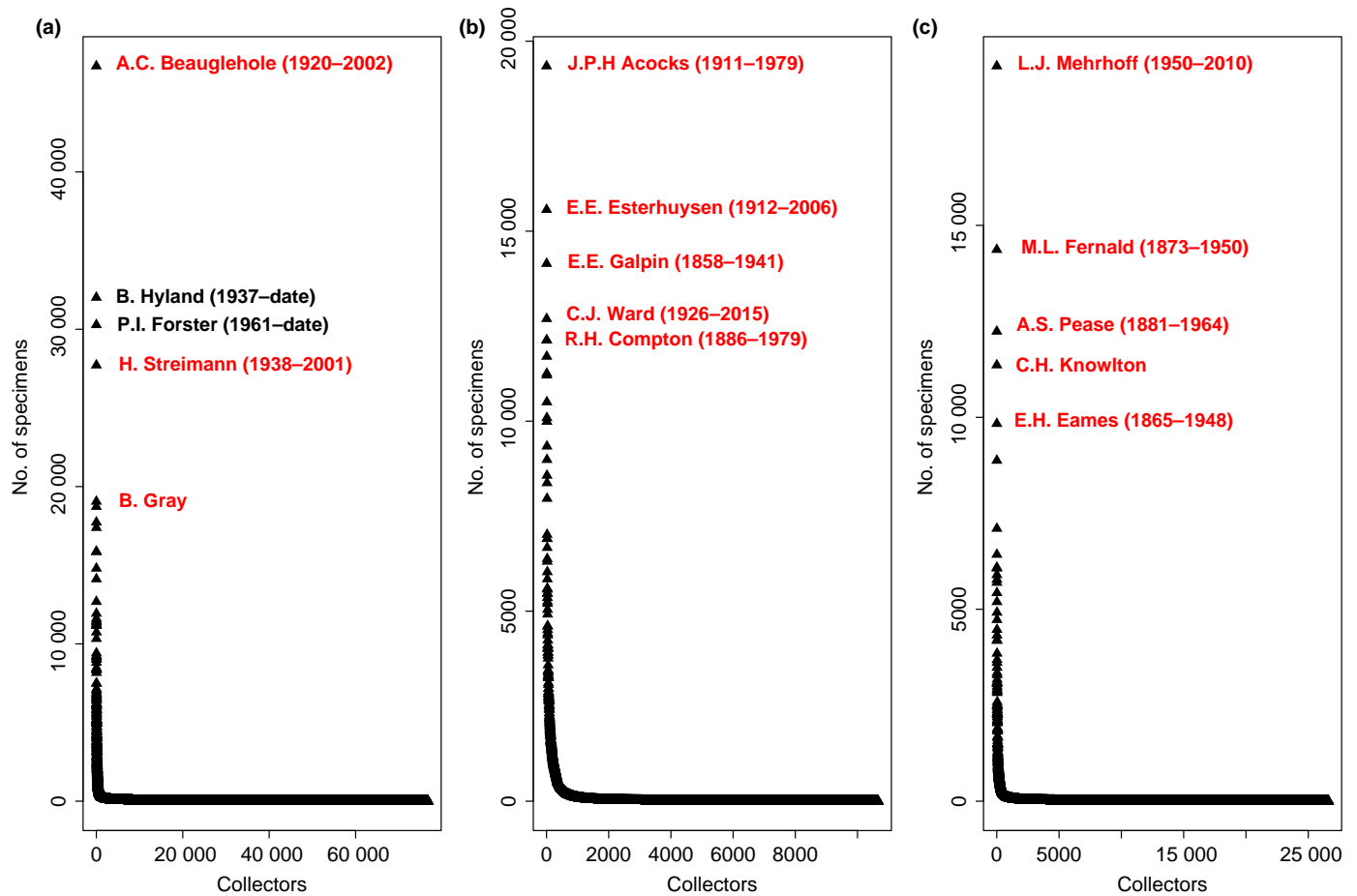

Fig. 8 Collector bias in herbarium collections. The number of herbarium specimens amassed per collector for three regional floras in (a) Australia, (b) South Africa, and (c) New England is shown. The top five collectors in each flora are highlighted in red. Numbers within parentheses correspond to lifespans of the collectors, with collectors that have died highlighted in red and currently living ones in black.

among the three floras in the traits of botanical collections almost certainly are reflective of genuine differences in species abundance and diversity.

Phylogenetic bias

Taxonomic biases in collection data have been reported previously (Hijmans *et al.*, 2000; Tobler *et al.*, 2007; Meyer *et al.*, 2016). However, our study is the first, to our knowledge, to demonstrate explicit evidence for phylogenetic bias in herbarium collections. Collection efforts in all three floras were concentrated in particular clades.

Previous examinations of taxonomic bias (e.g. Hijmans *et al.*, 2000; Tobler *et al.*, 2007; Meyer *et al.*, 2016) did not use the full complement of modern phylogenetic methods that included patterns of evolutionary relatedness, and so were limited in their

ability to detect details of taxonomic bias. By contrast, our phylogenetic approach not only captured taxonomic bias in favor of certain entire families (e.g. Asteraceae, Cyperaceae, Poaceae, and Rosaceae in NE), but revealed that evolutionarily distinct and globally endangered species are underrepresented in herbarium records relative to more common species. Such evolutionarily distinct species, which are threatened with extinction, represent important targets for future documentation or prioritization for conservation (Isaac *et al.*, 2007). However, collecting threatened taxa requires specialized training, compliance with regulations, and awareness of actual collection needs (Minteer *et al.*, 2014). Increasingly, DNA barcoding approaches, using small samples from living tissues, combined with GPS-referenced digital photography might be an avenue to document such species.

Collector bias

In all three regions, a large percentage of specimens was gathered by only a few collectors (Fig. 8). Thus, the habits and preferences of a few individuals probably shaped the establishment and formation of these herbarium collections. These 'founder effects' propagate across all the dimensions of collection bias examined above, and help us to understand past collection behavior (Bebber *et al.*, 2012). For example, certain collectors may focus on geographically circumscribed floristic zones, often near their place of residence, workplace, or vacation home, and sample all species found therein, whereas others may focus on collecting species of a particular clade across various regions. Professional botanists may tend to collect specimens on weekdays at any time of the year, whereas amateurs and faculty with teaching responsibilities may focus their efforts on weekends and vacation months. Those interested in function and physiology may only collect plants of certain habits or life histories (e.g. carnivorous, aquatic or succulent plants). These effects would probably be compounded when associated with mega-collectors. For instance, the Harvard University Herbaria's collection of Asian, especially woody plants was largely built by a few collectors and dates to the early establishment of the institution, and continues to attract scholars of the flora of Asia and their collections. Investigating the historical significance and potential biases created and propagated by these early pioneers is a ripe area for future research.

Future collecting

To ensure that herbaria continue to be vital centers for research beyond their importance to taxonomy and systematics, herbarium directors and collectors should account for and, whenever possible, reduce biases in plant collections. Biases can be accounted for to a degree using statistical approaches (Droissart *et al.*, 2012; Feeley, 2012; Grass *et al.*, 2014; Engemann *et al.*, 2015). For instance, inclusion of covariates for distances of collections from herbaria, roads, or other infrastructure (McCarthy *et al.*, 2012), using rarefaction methods to predict abundances (Schmidt-Lebuhn *et al.*, 2013), or including the collector as a variable would improve species distribution models and associated predictions of future changes across a flora. To remedy such

biases, future collecting expeditions should focus on 'coldspots' of collection intensity (Hijmans *et al.*, 2000), that is, places that are underrepresented in collections. Although some of the coldspots we identified probably represent more inaccessible environments, they often correspond to unique ecosystems, including the Succulent Karoo of SA and the Great North Woods in northern NE that contain many species of interest. Some of these coldspots also may indicate areas where herbarium specimens have yet to be mobilized, providing additional focus for efforts to make collection data widely available. Equally important is the need to continue modern collecting in well-established 'hotspots' so that there are multiple temporal benchmarks against which change can be measured. This is particularly true for nonnative invasive species that have rapidly expanding distributions and vulnerable native species that have ranges that are collapsing.

Phylogenetic and trait biases can be alleviated by targeting collection efforts where we know species have been undercollected. Temporal bias is more difficult to address, as we cannot add to historic collections. However, we can make efforts to maintain consistent regional botanical records by conducting field surveys at regular intervals. Also, by linking multiple herbaria into larger digital databases, the temporal biases of individual herbaria can be smoothed out to some extent.

We acknowledge that some of the biases also may be attributed to longstanding curation practices at the herbariums themselves. As herbarium collections were amassed for qualitative floristic, taxonomic, and systematic research, duplicate specimens of common species and nonreproductive material have sometimes been discarded, sent elsewhere, or not accepted in the first place. This trend is becoming even more pronounced as many herbaria around the world are increasingly constrained by funding, labor, and space. As new uses for biological collections continue to proliferate, curation practices should also change to accommodate different avenues of research, such as climate-change biology and rare plant conservation. This will often be most effective through continued collecting of specimens to overcome past biases. And, most importantly, researchers analyzing herbarium specimens in a widening array of studies needed to be aware of the biases in these collections, and apply appropriate statistical techniques.

comments of the Associate Editor and three anonymous reviewers on the submitted manuscript.

Author contributions

References

Abouheif E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research* 1: 895–909.

Agostinelli C, Lund U. 2013. *R package 'circular': circular statistics (version 0.4-7).* [WWW document] URL https://r-forge.r-project.org/projects/circular/ [accessed 26 August 2016].

APG, III (Angiosperm Phylogeny Group). 2009. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.

AVH. 2016. *Australia's virtual herbarium, Council of Heads of Australasian Herbaria.* [WWW document] URL http://avh.chah.org.au [accessed 9 June 2016].

Bebber DP, Carine MA, Davidse G, Harris DJ, Haston EM, Penn MG, Cafferty S, Wood JRI, Scotland RW. 2012. Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. *Proceedings of the Royal Society B* 279: 2269–2274.

Belasco WJ. 1979. *Americans on the road, from autocamp to motel 1910–1945.* Baltimore, MD, USA: Johns Hopkins University Press.

Bivand RS, Pebesma E, Gomez-Rubio V. 2013. *Applied spatial data analysis with R, 2nd edn.* New York, NY, USA: Springer. [WWW document] URL http://www.asdar-book.org/ [accessed 8 July 2016].

Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioural traits are more labile. *Evolution* 57: 717–745.

Bond WJ, Parr CL. 2010. Beyond the forest edge: ecology, diversity and conservation of the grassy biomes. *Biological Conservation* 143: 2395–2404.

Botkin BA. 1968. Automobile humor: from the horseless carriage to the compact car. *Journal of Popular Culture* I: 395–402.

Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ et al. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* 14: 16.

Ceballos G, Ehrlich PR. 2006. Global mammal distributions, biodiversity hotspots, and conservation. *Proceedings of the National Academy of Sciences, USA* 103: 19374–19379.

CIESIN. 2016. *Center for International Earth Science Information Network, Columbia University. Gridded population of the World, version 4 (GPWv4): population density.* Palisades, NY, USA: NASA Socioeconomic Data and Applications Center (SEDAC). [WWW document] URL https://doi.org/10.7927/h4np22dq [accessed 29 August 2016].

CNH. 2016. *Consortium of Northeastern Herbaria.* [WWW document] URL http://portal.neherbaria.org/portal/ [accessed 13 June 2016].

Cotterill FPD. 1995. Systematics, biological knowledge and environmental conservation. *Biodiversity and Conservation* 4: 183–205.

Cotterill FPD, Hustler CW, Broadley DG. 1994. Systematics and biodiversity. *Trends in Ecology & Evolution* 9: 228.

Dalton R. 2003. Natural history collections in crisis as funding is slashed. *Nature* 423: 6940.

Daru BH, Elliott TL, Park DS, Davies TJ. 2017. Understanding the processes underpinning patterns of phylogenetic regionalization. *Trends in Ecology & Evolution* 32: 845–860.

Daru BH, Van der Bank M, Davies TJ. 2015. Spatial incongruence among hotspots and complementary areas of tree diversity in southern Africa. *Diversity and Distributions* 21: 769–780.

Davies TJ, Kraft NJB, Salamin N, Wolkovich EM. 2012. Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. *Ecology* 93: 242–247.

Davis CC, Willis CG, Connolly B, Kelly C, Ellison AM. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany* 102: 1599–1609.

Droissart V, Hardy OJ, Sonké B, Dahdouh-Guebas F, Stévart T. 2012. Subsampling herbarium collections to assess geographic diversity gradients: a case study with endemic Orchidaceae and Rubiaceae in Cameroon. *Biotropica* 44: 44–52.

Edwards EJ, de Vos JM, Donoghue MJ. 2015. Doubtful pathways to cold tolerance in plants. *Nature* 521: E5–E6.

Edwards JL, Lane MA, Nielsen ES. 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science* 289: 2312–2314.

Engemann K, Enquist BJ, Sandel B, Boyle B, Jørgensen PM, Morueta-Holme N, Peet RK, Violle C, Svenning J-C. 2015. Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution* 5: 807–820.

ESRI. 1992. *Digital chart of the world, 1:1M.* Redlands, CA, USA: Environmental Systems Research Institute, Inc.

Everill PH, Primack RB, Ellwood EE, Melaas EK. 2014. Determining past leaf-out times of New England's deciduous forests from herbarium specimens. *American Journal of Botany* 101: 1–8.

Feeley KJ. 2012. Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology* 18: 1335–1341.

Ferreira SLA, Harmse AC. 2000. Crime and tourism in South Africa: international tourists perception and risk. *South African Geographical Journal* 82: 80–85.

Forman RTT, Alexander LE. 1998. Roads and their major ecological effects. *Annual Review of Ecology and Systematics* 29: 207–231.

Fortune S. 1992. Voronoi diagrams and Delaunay triangulations. *Computing in Euclidean Geometry* 1: 193–233.

Funk V. 2003. The importance of herbaria. *Plant Science Bulletin* 49: 94–95.

Funk VA, Morin N. 2000. A survey of the herbaria of the southeast United States. *Sida, Botanical Miscellany* 18: 35–52.

Funk VA, Richardson K. 2002. Biological specimen data in biodiversity studies: use it or lose it. *Systematic Biology* 51: 303–316.

GADM. 2015. *Global administrative areas, version 2.8.* [WWW document] URL www.gadm.org [accessed 20 September 2016].

Garcillán PP, Ezcurra E. 2011. Sampling procedures and species estimation: testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. *Journal of Vegetation Science* 22: 273–280.

Garcillán PP, Ezcurra E, Vega E. 2008. Guadalupe Island: lost paradise recovered? Overgrazing impact on extinction in a remote oceanic island as estimated through accumulation functions. *Biodiversity and Conservation* 17: 1613–1625.

Geri F, Lastrucci L, Viciani D, Foggi B, Ferretti G, Maccherini S, Bonini I, Amici V, Chiarucci A. 2013. Mapping patterns of ferns species richness through the use of herbarium data. *Biodiversity and Conservation* 22: 1679–1690.

Gibbons JW, Scott DE, Ryan T, Buhlmann K, Tuberville T, Greene J, Mills T, Leiden Y, Poppy S, Winne C et al. 2000. The global decline of reptiles, déjà·vu amphibians. *BioScience* 50: 653–666.

Grass A, Tremetsberger K, Hössinger R, Bernhardt K. 2014. Change of species and habitat diversity in the Pannonian region of eastern Lower Austria over 170 years: using herbarium records as a witness. *Natural Resources* 5: 583–596.

Greve M, Lykke AM, Fagg CW, Gereau RE, Lewis GP, Marchant R, Marshall AR, Ndayishimiye J, Bogaert J, Svenning JC. 2016. Realising the potential of herbarium records for conservation biology. *South African Journal of Botany* 105: 317–323.

Griffith EH, Sauer JR, Royle JA. 2010. Traffic effects on bird counts on North American breeding bird survey routes. *Auk* 127: 387–393.

Gutzwiller KJ, Flather CH. 2011. Wetland features and landscape context predict the risk of wetland habitat loss. *Ecological Applications* 21: 968–982.

Hart R, Salick J, Ranjitkar S, Xu J. 2014. Herbarium specimens show contrasting phenological responses to Himalayan climate. *Proceedings of the National Academy of Sciences, USA* 111: 10615–10619.

Hijmans RJ. 2015. *geosphere: spherical trigonometry. R package version 1.4-3.* [WWW document] URL http://CRAN.R-project.org/package=geosphere [accessed 2 August 2016].

Hijmans RJ, Garrett KA, Huaman Z, Zhang DP, Schreuder M, Bonierbale M. 2000. Assessing the geographic representation of genebank collections: the case of the Bolivian wild potatoes. *Conservation Biology* 14: 1755–1765.

Hortal J, Lobo JM, Jiménez-Valverde A. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in tenerife, canary islands. *Conservation Biology* 21: 853–863.

Hui C, Shuang-cheng L, Yi-li Z. 2003. Impact of road construction on vegetation alongside Qinghai-Xizang highway and railway. *Chinese Geographical Science* 13: 340–346.

Isaac NJ, Turvey ST, Collen B, Waterman C, Baillie JE. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE* 2: e296.

Iwanycki N. 2009. *Guidelines for collecting herbarium specimens of vascular plants.* Hamilton, ON, Canada: Royal Botanical Gardens Canada.

Jombart T, Dray S. 2008. adephylo: exploratory analyses for the phylogenetic comparative method. *Bioinformatics* 26: 1907–1909.

Klemens MW, Thorbjarnarson JB. 1995. Reptiles as a food resource. *Biodiversity and Conservation* 4: 281–298.

Lavoie C. 2013. Biological collections in an ever changing world: herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics* 15: 68–76.

Lees DC, Lack HW, Rougerie R, Hernandez-Lopez A, Raus T, Avtzis ND, Augustin S, Lopez-Vaamonde C. 2011. Tracking origins of invasive herbivores through herbaria and archival DNA: the case of the horse-chestnut leaf miner. *Frontiers in Ecology and the Environment* 9: 322–328.

Lemanski C. 2004. A new apartheid? The spatial implications of fear of crime in Cape Town, South Africa. *Environment & Urbanization* 16: 101–111.

Leuner B. 2007. *Migration, multiculturalism and language maintenance in Australia.* Oxford, UK: Peter Lang.

Li Y, Yu J, Ning K, Du S, Han G, Qu F, Wang G, Fu Y, Zhan C. 2014. Ecological effects of roads on the plant diversity of coastal wetland in the Yellow River Delta. *Scientific World Journal* 2014: 952051.

McCarthy KP, Fletcher JRRJ, Rota CT, Hutto RL. 2012. Predicting species distributions from samples collected along roadsides. *Conservation Biology* 26: 68–77.

Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19: 992–1006.

Miller-Rushing A, Primack R, Mukunda S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. *American Journal of Botany* 93: 1667–1674.

Minteer BA, Collins JP, Love KE, Puschendorf R. 2014. Avoiding (re) extinction. *Science* 344: 260–261.

Moerman DE, Estabrook GF. 2006. The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography* 33: 1969–1974.

Musgrave T, Gardner C, Musgrave W. 1999. *The plant hunters. Two hundred years of adventure and discovery.* London, UK: Seven Dials.

NEBC. 1899. Editorial announcement. *Rhodora* 1: 1–2.

Newbold T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* 34: 3–22.

Norris WR, Lewis DQ, Widrlechner MP, Thompson JD, Pope RO. 2001. Lessons from an inventory of the Ames, Iowa, flora (1859–2000). *Journal of the Iowa Academy of Science* 108: 34–63.

Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017. Assessing the relevance of herbarium collections as tools for conservation biology. *Botanical Review* 83: 303–325.

Orme CD, Davies RG, Burgess M, Eigenbrod F, Pickup N, Olson VA, Webster AJ, Ding TS, Rasmussen PC, Ridgely RS *et al.* 2005. Global hotspots of

species richness are not congruent with endemism or threat. *Nature* 436: 1016–1019.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012. *caper: comparative analyses of phylogenetics and evolution in R. R package version 0.5.* [WWW document] URL http://CRAN.R-project.org/package = caper [accessed 22 July 2016].

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877–884.

Palmer MW, Earls PG, Hoagland BW, White PS, Wohlgemuth T. 2002. Quantitative tools for perfecting species list. *Environmetrics* 13: 121–137.

Pautasso M, McKinney ML. 2007. The botanist effect revisited: plant species richness, county area, and human population size in the United States. *Conservation Biology* 21: 1333–1340.

Poethig RS. 2013. Vegetative phase change and shoot maturation in plants. *Current Topics in Developmental Biology* 105: 125–152.

Prather LA, Alvarez-Fuentes O, Mayfield MH, Ferguson CJ. 2004a. The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. *Systematic Botany* 29: 15–28.

Prather LA, Alvarez-Fuentes O, Mayfield MH, Ferguson CJ. 2004b. Implications of the decline in plant collecting for systematic and floristic research. *Systematic Botany* 29: 216–220.

Price CA. 1998. Post-war immigration: 1945–1998. *Journal of the Australian Population Association* 15: 17.

Pritchard PCH. 1996. *The Galápagos tortoises: nomenclatural and survival status.* Lunenburg, MA, USA: Chelonian Research Foundation in association with Conservation International and Chelonia Institute.

Pyke GH, Ehrlich PR. 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological Reviews* 5: 247–266.

Redding DW, Mooers AØ. 2006. Incorporating evolutionary measures into conservation prioritization. *Conservation Biology* 20: 1670–1678.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223.

Rich TCG, Woodruff ER. 1992. Recording bias in botanical surveys. *Watsonia* 19: 73–95.

Rivers-Moore NA, Cowden C. 2012. Regional prediction of wetland degradation in South Africa. *Wetlands Ecology and Management* 20: 491–502.

Robinson JG. 2001. Using 'sustainable use' approaches to conserve exploited populations. In: Reynolds JD, Mace GM, Redford KH, Robinson JG, eds. *Conservation of exploited species.* Cambridge, UK: Cambridge University Press, 485–498.

le Roux MM, Wilkin P, Balkwill K, Boatwright JS, Bytebier B, Filer D, Klak C, Klopper RR, Koekemoer M, Livermore L *et al.* 2017. Producing a plant diversity portal for South Africa. *Taxon* 66: 421–431.

Rudin SM, Murray DW, Whitfeld TJS. 2017. Retrospective analysis of heavy metal contamination in Rhode Island based on old and new herbarium specimens. *Applications in Plant Sciences* 5: 1–13.

SANBI. 2016. *South African national biodiversity institute. Botanical Database of Southern Africa (BODATSA).* [WWW document] URL http://newposa.sanbi. org/ [accessed 22 July 2016].

Schaefer H, Hardy OJ, Silva L, Barraclough TG, Savolainen V. 2011. Testing Darwin's naturalization hypothesis in the Azores. *Ecology Letters* 14: 389–396.

Schmidt-Lebuhn AN, Knerr NJ, Kessler M. 2013. Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation* 22: 905–919.

van der Schoot C, Paul LK, Rinne PLH. 2014. The embryonic shoot: a lifeline through winter. *Journal of Experimental Botany* 65: 1699–1712.

Schorn C, Weber E, Bernardos R, Hopkins C, Davis CC. 2016. The New England vascular plants project: 295,000 specimens and counting. *Rhodora* 118: 324–325.

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT. 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* 8: e69189.

Stropp J, Ladle RJ, Malhado ACM, Hortal J, Gaffuri J, Temperley WH, Skøien JO, Mayaux P. 2016. Mapping ignorance: 300 years of collecting

flowering plants in Africa. *Global Ecology and Biogeography* 25: 1085–1096.

Syfert MM, Smith MJ, Coomes DA. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE* 8: e55158.

Thiers B. 2016. *Index Herbariorum: a global directory of public herbaria and associated staff.* New York Botanical Garden's Virtual Herbarium. [WWW document] URL http://sweetgum.nybg.org/science/ih/ [accessed 29 September 2016].

Tobler M, Honorio E, Janovec J, Reynel C. 2007. Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodiversity and Conservation* 16: 659–677.

Webb CO, Ackerly DD, Kembel SW. 2008. PHYLOCOM: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* 24: 2098–2100.

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

Webb CO, Donoghue MJ. 2005. Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes* 5: 181–183.

Whittle T. 1970. *The plant hunters.* London, UK: Heinemann.

Willis CG, Ellwood ER, Primack RB, Davis CC, Pearson KD, Gallinato AS, Yost JM, Nelson G, Mazer SJ, Rossington NL *et al.* 2017a. Old plants, new tricks: phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32: 531–546.

Willis CG, Law E, Williams AC, Franzone BF, Bernardos R, Brun L, Hopkins C, Schorn C, Weber E, Parks DS *et al.* 2017b. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytologist* 215: 479–488.

Wolf A, Anderegg WRL, Ryan SJ, Christensen J. 2011. Robust detection of plant species distribution shifts under biased sampling regimes. *Ecosphere* 2: 115.

Wolkovich EM, Davies TJ, Schaefer H, Cleland EE, Cook BI, Travers SE, Willis CG, Davis CC. 2013. Temperature-dependent shifts in phenology

contribute to the success of exotic species with climate change. *American Journal of Botany* 100: 1407–1421.

Yessoufou K, Daru BH, Davies TJ. 2012. Phylogenetic patterns of extinction risk in the Eastern Arc ecosystems, an African biodiversity hotspot. *PLoS ONE* 7: e47082.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Analytical workflow representing the different steps in the development of this study from data compilation and collation to statistical analysis.

Fig. S2 Relationships of the number of specimens collected per species with the number of species collected in each flora for Australia, South Africa, and New England in the USA.

Table S1 Results of the tests of phylogenetic signal in the number of specimens collected per species

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

About *New Phytologist*

- *New Phytologist* is an electronic (online-only) journal owned by the New Phytologist Trust, a **not-for-profit organization** dedicated to the promotion of plant science, facilitating projects from symposia to free access for our Tansley reviews.

- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as ready' via *Early View* – our average time to decision is <26 days. There are **no page or colour charges** and a PDF version will be provided for each article.

- The journal is available online at Wiley Online Library. Visit **www.newphytologist.com** to search the articles and register for table of contents email alerts.

- If you have any questions, do get in touch with Central Office (np-centraloffice@lancaster.ac.uk) or, if it is more convenient, our USA Office (np-usaoffice@lancaster.ac.uk)

- For submission instructions, subscription and all the latest information visit **www.newphytologist.com**