

1 **Widespread sampling biases in herbaria revealed from large-scale digitization**

2

3

4

5 Barnabas H. Daru^{1,*}, Daniel S. Park^{1,*}, Richard B. Primack², Charles G. Willis¹, David S.
6 Barrington³, Timothy J. S. Whitfeld⁴, Tristram G. Seidler⁵, Patrick W. Sweeney⁶, David
7 R. Foster⁷, Aaron M. Ellison⁷ and Charles C. Davis¹

8

9 ¹Department of Organismic and Evolutionary Biology and Harvard University Herbaria,
10 Harvard University, Cambridge, MA 02138, USA; ²Biology Department, Boston
11 University, Boston, MA 02215, USA; ³Pringle Herbarium, Plant Biology Department,
12 University of Vermont, Torrey Hall, 27 Colchester Ave, Burlington, VT 05405, USA;
13 ⁴Brown University Herbarium, Department of Ecology and Evolutionary Biology, Brown
14 University, 34 Olive Street, Box G-B225, Providence, Rhode Island 02912 USA;
15 ⁵Biology Department, University of Massachusetts, 611 North Pleasant Street, Amherst,
16 MA 01003, USA; ⁶Division of Botany, Peabody Museum of Natural History, Yale
17 University, New Haven, CT, USA; ⁷Harvard University, Harvard Forest, 324 North Main
18 Street, Petersham, Massachusetts, 01366 USA.

19

20 *These authors contributed equally to the study

21

22 ¹To whom correspondence should be addressed. Email: barnabas_daru@fas.harvard.edu

23

24 **Short Title:** Sampling bias in herbarium specimens

25 **Manuscript information:** 5268 words (Introduction = 723 words, Materials and
26 Methods = 1560 words, Results = 957 words, Discussion = 2028 words | 8 figures (8
27 color figures) | 3 Tables | 1 supporting information

28

29

30 **SUMMARY**

- 31 1. Non-random collecting practices may bias conclusions drawn from analyses of
32 herbarium records. Recent efforts to fully digitize and mobilize regional floras offer a
33 timely opportunity to assess commonalities and differences in herbarium sampling
34 biases.
- 35 2. We determined spatial, temporal, trait, phylogenetic, and collector biases in ~5
36 million herbarium records, representing three of the most complete digitized floras of
37 the world: Australia (AU), South Africa (SA), and New England (NE).
- 38 3. We identified numerous shared and unique biases among these regions. Shared biases
39 included specimens i) collected close to roads and herbaria; ii) collected more
40 frequently during spring; iii) of threatened species collected less frequently; and iv) of
41 close relatives collected in similar numbers. Regional differences included i) over-
42 representation of graminoids in SA and AU and of annuals in AU; and ii) peak
43 collection during the 1910s in NE, 1980s in SA, and 1990s in AU. Finally, in all
44 regions, a disproportionately large percentage of specimens were collected by a few
45 individuals. These mega-collectors, and their associated preferences and
46 idiosyncrasies, may have shaped patterns of collection bias via ‘founder effects’.
- 47 4. Studies using herbarium collections should account for sampling biases and future
48 collecting efforts should avoid compounding these biases.

49 **Keywords:** Herbarium, collector bias, geographic bias, regional flora, sampling bias,
50 temporal bias, trait bias

51

52 INTRODUCTION

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

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

69 Sampling biases fall into several broad categories. Taxonomic or phylogenetic
70 bias is the unbalanced sampling of certain taxa or clades over others, typically resulting
71 from the interests of a collector or the attractiveness of plants (Hortal *et al.*, 2007).
72 Geographic bias occurs when specimens are collected more frequently in one place than
73 another often because of differential accessibility (Hijmans *et al.*, 2000). Temporal bias
74 occurs when collection activity is favored in certain years or parts of the year (Cotterill *et al.*
75 *et al.*, 1994; Funk & Morin, 2000; Norris *et al.*, 2001). Meyer *et al.* (2016) evaluated
76 worldwide terrestrial plant occurrence data using 120 million records from the Global
77 Biodiversity Information Facility (GBIF; Edwards *et al.*, 2000). Their analyses revealed
78 large taxonomic gaps in global plant occurrence data (< 25% of species of land plants
79 were sampled); extensive spatial gaps across regions that harbor high concentrations of
80 plant diversity, especially in Asia, Central Africa, and Amazonia; and strong temporal
81 discontinuities in occurrence records across decades, which can hamper inferences about
82 the effects on plants of recent and future environmental change.

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

94 Expanding upon Meyer *et al.*'s work, we explored spatial, temporal, and
95 taxonomic/phylogenetic sampling biases in collections from three of the most extensively
96 collected, digitized, and mobilized regional floras in the world: South Africa (SA),
97 Australia (AU), and the New England (NE) region of the United States. The SA flora is a
98 compilation of digitized herbarium specimens from all major herbaria across the country
99 available in a single online portal (South African National Biodiversity Institute [SANBI],
100 2016; le Roux *et al.*, 2017). The Australian Virtual Herbarium (AVH, 2016) is the main
101 database for AU. It contains digitized herbarium specimens from all the major herbaria in
102 Australia. The Consortium of the Northeast Herbaria database contains digitized
103 specimens from 15 participating herbaria in the NE region of the United States (Schorn *et*
104 *al.*, 2016). We also examined trait bias – sampling bias due to intrinsic life-history
105 characteristics, including life cycle (annual *vs.* perennial), plant height, and growth form
106 (woody *vs.* herbaceous), and species conservation status. Finally, we examined the
107 contributions of individual collectors to each flora. We identified biases in all five of
108 these categories within each of these regional floras. Our results revealed both
109 commonalities and differences in regional collection biases and identified new sampling
110 foci as collections grow in the future.

111

112 **MATERIAL AND METHODS**

113 **Sources and description of data**

114 We obtained 12,488,200 herbarium specimen records of vascular plants from AU
115 (Australia Virtual Herbarium [AVH], 2016); 2,049,905 herbarium specimen records from
116 SA including Lesotho and Swaziland (South African National Biodiversity Institute
117 [SANBI], 2016); and 879,388 herbarium specimen records from the NE (USA) flora
118 (Consortium of Northeastern Herbaria [CNH], 2016). The records were cleaned in two
119 steps (Fig. S1). First, we standardized the taxonomy of all species using the Taxonomic
120 Name Resolution Service v.4.0 (Boyle *et al.*, 2013). Second, we removed specimens that
121 were duplicates from the same collection locality and date; specimens with clearly
122 erroneous locations (*i.e.*, in oceans); specimens missing exact collection date and/or
123 georeferenced location data; and field observation records not tied to a physical specimen.
124 Following this data cleaning, we retained 32% of the initial specimens for further analysis,
125 including: 24% of the AU records (31,966 taxa; 661,370 records); 49% (20,824 taxa,
126 4,579,320 records) from SA; and 75% (11,447 taxa, 1,008,206 records) from NE.

127

128 **Analyses**

129 ***Spatial biases***

130 First, we evaluated the density of sampling localities across the focal regions using
131 Delaunay triangulation polygons, which measure the land area covered by each sampling
132 locale (Fortune, 1992). Larger triangles indicate sparser collecting effort, whereas smaller
133 triangles indicate more concentrated effort. Second, we examined infrastructure bias by
134 calculating the minimum distance of each collection locality to the nearest major road
135 (GADM, 2015) and herbarium (following Thiers, 2016). Our dataset of roads assumes
136 that the network of major roads in these regions has remained largely unchanged over the
137 past century (Forman *et al.*, 1995). Thus, we focused only on major roads in these three
138 regions and excluded all street roads and dirt tracks in our analysis. We then compared
139 these distances to those generated by a null model (1000 iterations) in which the same
140 number of sample points was randomly (Poisson) distributed across each geographic
141 region. Third, we mapped geographic biases in sampling density, defined as areas of
142 excessive (hotspots) or insufficient (coldspots) collection (Hijmans *et al.*, 2000). Hotspots
143 and coldspots were determined at a spatial grain of $0.25^\circ \times 0.25^\circ$ based on the number of
144 specimens per grid cell, and identified using the 2.5% threshold (Ceballos & Ehrlich,

145 2006; Orme *et al.*, 2005; Daru *et al.*, 2015), based, respectively, on the 97.5th and 2.5th
146 percentile values in the number of specimens collected per grid cell. Spatial distance
147 calculations were done with the functions *dist2Line* and *spDists* in the R packages *sp*
148 (Bivand *et al.*, 2013) and *geosphere* (Hijmans, 2015), respectively. In our final predictive
149 model of sampling density, we also included human population density (CIESIN, 2016),
150 sampling localities, infrastructure (distance to herbaria and roads), number of specimens
151 collected, and elevation or topographic relief.

152

153 ***Temporal bias***

154 For each regional flora, we explored bias at several temporal scales. Collection dates
155 ranged from 20 May 1664 to 9 January 2016 (AU), 15 November 1656 to 6 June 2016
156 (SA), and 28 July 1687 to 4 May 2016 (NE). We hypothesized that collectors tended to
157 avoid fieldwork during unfavorable conditions (*e.g.*, weekdays, winter, war time). To test
158 for temporal bias, we first re-coded collection dates as days of the week (Sunday = 1,
159 Monday = 2, *etc.*), day of year (DOY; where January 1 = 1 DOY and December 31 = 365
160 DOY, *etc.*), and decade (*e.g.*, 1901-1910, 1911-1920, *etc.*). We then used a Rayleigh test
161 of directional statistics in the R package *circular* (Agostinelli & Lund, 2013) to test
162 whether each of these collection dates were randomly distributed against all dates
163 spanning the entire duration of plant collection. If $P < \alpha = 0.05$, we rejected the null
164 hypothesis of temporal uniformity at scales of weeks, days of the year, or decades.

165

166 ***Trait bias***

167 We used customized R scripts to harvest information on growth duration (annual *vs.*
168 perennial), growth form (woody *vs.* herbaceous), and height for each species from online
169 regional databases (all accessed in June 2016), including: New South Wales Flora Online
170 (<http://plantnet.rbgsyd.nsw.gov.au>); JSTOR Global Plants (<https://plants.jstor.org>); Atlas
171 of Living Australia (<http://bie.ala.org.au>); Plants of Southwestern Australia
172 (<http://keys.lucidcentral.org>); the African Plant Database (<http://www.ville-ge.ch>); Plants
173 of Southern Africa (<http://www.plantzafrica.com>); Plant Resources of Tropical Africa
174 (<http://www.prota4u.org>); Flora of North America (<http://www.efloras.org>); and the
175 USDA Plants Database (<http://plants.usda.gov>). We then manually checked these trait

176 data for inconsistencies among sources, and adjusted the data accordingly. Extinction risk
177 assessments for each species were retrieved from the IUCN Red List database
178 (www.iucnredlist.org, accessed August 2016), which uses the following categories: Data
179 Deficient (DD), Least Concern (LC), Lower Risk/Conservation Dependent (LR/CD),
180 Near Threatened (NT), Vulnerable (VU), Endangered (EN), Critically Endangered (CR),
181 and Extinct (EX). We grouped these narrow categories into two broader threat categories,
182 threatened (EX+CR+EN+VU) or not threatened (LR/CD+NT+LC), following Yessoufou
183 *et al.* (2012).

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

192

193 ***Phylogenetic bias***

194 We assessed phylogenetic signal in collection frequency as a measure of phylogenetic
195 bias using two different tests (Wolkovich *et al.*, 2013). A strong phylogenetic signal –
196 closely related species sharing similar collection frequency – would suggest phylogenetic
197 bias in collections. We first assembled a phylogeny using Phylomatic (Webb &
198 Donoghue, 2005), enforcing a topology that assumed the APG III backbone (tree
199 R20120829). This phylogeny included all species in our analysis, but provided only an
200 approximate degree of relatedness based on taxonomic hierarchy at family level, thus
201 many relationships, especially within genera, were unresolved. This is problematic
202 because recent theoretical and empirical studies have shown that a lack of resolution in a
203 community phylogeny may mask significant patterns by reducing statistical power
204 (Schaefer *et al.*, 2011) or suggest significant phylogenetic patterns that are not supported
205 by more completely resolved phylogenies (Davies *et al.*, 2012).

206 To alleviate these concerns, we also tested for phylogenetic bias by including only
207 those species sampled in the dated molecular phylogeny inferred from seven genes for
208 32,223 plant species (Zanne *et al.*, 2014). Although this phylogeny has been criticized
209 (Edwards *et al.*, 2015), it nonetheless represents the single largest phylogeny to date for
210 flowering plants. The taxon sampling for testing phylogenetic bias included 5814 species
211 from AU, 3568 from SA, and 4269 from NE.

212 We estimated phylogenetic signal using three common metrics: Abouheif's C_{mean}
213 (Abouheif, 1999), Blomberg's K (Blomberg *et al.*, 2003), and Pagel's lambda (λ) (Pagel,
214 1999). Significance was assessed by comparing observed values to a null distribution
215 created by shuffling the trait values across the tips of the phylogeny 1000 times. Pagel's λ
216 uses a maximum-likelihood method with branch-length transformation to estimate the
217 best-fit of a trait against a Brownian model. Values of Pagel's λ range from 0 (no
218 phylogenetic signal) to 1 (strong phylogenetic signal). Both Blomberg's K (a significant
219 phylogenetic signal is indicated by a K value > 1) and Pagel's λ were calculated using the
220 R package phytools (Revell, 2012). Abouheif's C_{mean} was calculated using adephylo
221 (Jombart & Dray, 2008). We tested the sensitivity of our analysis by exploring
222 phylogenetic signal in collecting effort across nine well-sampled NE clades: Asteraceae,
223 Brassicaceae, Cyperaceae, Ericaceae, Fabaceae, Lamiaceae, Poaceae, Ranunculaceae,
224 and Rosaceae.

225 In addition to phylogenetic signal, we also used phylogenetic generalized least
226 squares regressions (PGLS) in the R package caper (Orme *et al.*, 2012) to model
227 collecting effort per species in each region as a function of species evolutionary ages,
228 evolutionary distinctiveness (ED), and “evolutionary distinctiveness and global
229 endangerment” (EDGE; Isaac *et al.*, 2007). Species ages were measured as the length of
230 terminal branches (BL) linking species on a phylogenetic tree. ED measures the degree of
231 phylogenetic isolation of a species, whereas the EDGE metric was determined by
232 calculating the ED score of each species (Isaac *et al.*, 2007) and combining it with global
233 endangerment (GE) from IUCN conservation categories: $\text{EDGE} = \ln(1 + \text{ED}) + \text{GE} \times$
234 $\ln(2)$, where GE represents expected probability of species extinction over a 100-year
235 period (Redding & Mooers, 2006) categorized as follows: least concern = 0.001, near

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

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

245

246 *Collector bias*

247 We determined collector bias by tabulating the number of specimens amassed by each
248 collector in the three floras. We then examined Pearson's product-moment correlation
249 between the numbers of specimens collected per collector with the number of species
250 collected per collector.

251

252 **Computation and availability of data and code**

253 All statistical analyses were conducted using the Research Computing Clusters of
254 Harvard University (<https://rc.fas.harvard.edu/>). Data files and custom R scripts will be
255 available from the Harvard Forest Data Archive
256 (<http://harvardforest.fas.harvard.edu/data-archive>).

257

258 **RESULTS**

259 *Spatial bias*

260 High sampling density was observed in southeast and southwest AU, the Cape region of
261 SA, and two of the six NE states (Connecticut and Massachusetts) relative to other parts
262 of those regions (Fig. 1a-c). When we weighted each sampling locale by the number of
263 specimens, we found a mismatch between hotspots (top 2.5% quantiles) and coldspots
264 (lowest 2.5% quantiles) of sampling intensity (Fig. 1d-f), suggesting hotspots and
265 coldspots are not randomly distributed. Hotspots of collecting tend to cluster around

266 coasts in AU and SA, whereas coldspots were abundant in interior areas. In NE, hotspots
267 were concentrated in the south and coldspots in the north.

268 Herbarium specimens tended to be collected closer than expected to roads and
269 herbaria ($p < 0.01$; Fig. 2a, b). More than 50% of herbarium specimens were collected
270 within 2 km of roadsides in all three floras ($p < 0.01$; Fig. 2a). Moreover, distance to
271 herbaria explained 45% of the variance in collecting effort in AU, 29% in SA and 12.3%
272 in NE (Table 1). Despite substantial gradients in altitudes in each region (-15 – 2022 m
273 a.s.l. in AU; 1 – 3254 m a.s.l. in SA; and -3 – 1485 m a.s.l. in NE), most specimens were
274 collected below 500 m a.s.l. in AU and NE (81%, 44%, and 93% of specimens in AU, SA,
275 and NE, respectively; Fig. 2c).

276

277 *Temporal bias*

278 There were historical biases in collection efforts in the three floras: low sampling until
279 1880 in AU and SA, and a burst of collections in NE in the early 20th century (Fig. 3).
280 Conversely, there was a dramatic increase in botanical collection in SA and AU after
281 World War II, peaking in the 1980s and 1990s, respectively (Fig. 3), 100 years after peak
282 collection activity in NE. Seasonally, specimen collections were biased toward spring and
283 summer for the three floras, with peak collection ranging from September to December in
284 AU and SA (Rayleigh $Z = 0.189$ and $Z = 0.251$ respectively, both $p < 0.001$), and May to
285 September in NE (Rayleigh $Z = 0.718$, $p < 0.001$; Fig. 4a). There was a non-significant
286 trend towards collection on weekends (Saturdays and Sundays) in NE (Rayleigh test $Z =$
287 1.0 , $p < 0.001$) and midweek in SA and AU (Rayleigh test $Z = 0.105$ and $Z = 1.0$,
288 respectively; both $p < 0.001$; Fig. 4a).

289

290 *Trait bias*

291 Perennials were more frequently collected than annuals in terms of specimens per species
292 in SA and NE, whereas the opposite was true for AU (Fig. 5a). Similarly, graminoid
293 specimens per species were over-represented relative to other habits in AU and SA,
294 whereas herbs and trees were over-represented in NE (Fig. 5b). Relatively short plants
295 were more frequently represented than taller plants in all three floras: 79.3%, 89.3% and

296 84.9% of the plants collected in AU, SA and NE, respectively were less than 5 m in
297 height (Fig. 5c).

298 Threatened species were collected significantly less often than non-threatened
299 plants across all three floras (all $p < 0.001$; Fig. 5d).

300

301 ***Phylogenetic bias***

302 There was a significant, but weak phylogenetic signal in the abundance of specimens per
303 species across all three floras (Table 2). Specifically, closely related species tended to
304 have a more similar number of specimens in a collection than expected (Table 2; Fig. 6).
305 Such phylogenetic bias was strongest in SA (Abuoheif's $C_{\text{mean}} = 0.15$ and $\lambda = 0.32$; both
306 $p < 0.01$, but $K = 0.0013$ [NS]). For instance, in SA, collections of the genus *Protea*
307 averaged 115 specimens per species whereas only two specimens were collected per
308 species of *Rytigynia* on average. Most *Amsonia* in NE were represented by < 10
309 specimens per species, whereas many fern species were represented by high specimen
310 numbers (e.g., *Onoclea* with 845 specimens/species). Australian collections showed the
311 weakest phylogenetic bias (Abuoheif's $C_{\text{mean}} = 0.12$ and $\lambda = 0.18$, both $p < 0.01$, but $K =$
312 0.00085 [NS]; Fig. 6). Phylogenetic signal varied at the family level as well in NE, with
313 Asteraceae showing the strongest collection bias (Fig. 7), followed by Cyperaceae,
314 Poaceae, and Rosaceae (Table S1).

315 'Evolutionary distinctiveness and global endangerment' (EDGE) was significant
316 predictor of collecting efforts in all three floras ($p < 0.001$), with variance ranging from
317 1.89% (NE) and 3.75% (AU), to 8.89% in SA. In general, EDGE species were generally
318 under-collected in terms of specimens per species (Table 3).

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

325

326 ***Collector bias***

327 The number of specimens per collector was highly skewed (Fig. 8). In AU, more than 50%
328 of the examined specimens were amassed by only 2% of the collectors, including A.C.
329 Beauglehole (46,728 specimens), B. Hyland (32,019 specimens), and P.I. Forster (30,280
330 specimens; Fig. 8a). In SA, more than 50% of the specimens were amassed by 9.5% of
331 collectors, including J.P.H Acocks (19,344 specimens), E.E. Esterhuysen (15,566
332 specimens), and E.E. Galpin (14,146 specimens; Fig. 8b). In NE, 50% of the specimens
333 were contributed by 3.2% of the collectors, including L.J. Mehrhoff (19,149 specimens),
334 M.L. Fernald (14,368 specimens), and A.S. Pease (12,238 specimens; Fig. 8c). The
335 number of specimens amassed by these collectors was positively correlated with the
336 number of species they collected, suggesting that these collectors were doing general
337 collecting rather than focusing on a particular group of plants ($r = 0.85$ in AU, 0.95 in SA
338 and 0.84 in NE; all $p < 0.01$; Fig. S2).

339

340 **DISCUSSION**

341 Historically, the primary function of herbaria has been to serve as an institution of
342 taxonomy, allowing users to construct classifications of plants, verify identifications,
343 determine the ranges and morphological characteristics of species, and develop local and
344 regional floras (Greve *et al.*, 2016). Over time, new uses for specimens have arisen, and
345 now more than ever, they are being used in ways that collectors rarely imagined (Pyke &
346 Ehrlich, 2010; Lavoie, 2013; Willis *et al.*, 2017a,b; Nualart *et al.*, 2017; Rudin *et al.*,
347 2017). Accordingly, attempts to assess and categorize biases inherent in these collections
348 have been made (Rich & Woodruff, 1992; Geri *et al.*, 2013; Schmidt-Lebuhn *et al.*, 2013;
349 Meyer *et al.*, 2016; Stropp *et al.*, 2016). Among these, the most comprehensive
350 investigation is by Meyer *et al.* (2016), who proposed an important conceptual
351 framework for analyzing gaps and biases along taxonomic, geographical, and temporal
352 dimensions. Although Meyer *et al.* (2016) focused more on observational records than
353 herbarium collections, they uncovered numerous biases in ‘digitally accessible
354 information’ regarding plants and provided an important baseline for evaluating and
355 improving global floristic coverage in collection data. However, vast geographic areas
356 remain where collections data are sparsely available, in part because numerous herbaria
357 have not yet been fully digitized and mobilized. Collection biases in these areas are

358 difficult to categorize, and may skew global patterns of bias when considered alongside
359 areas whose collection data now are readily available. By focusing on three of the most
360 well-collected and digitized floras in the world, we reduced effects of missing or
361 unavailable data, and most importantly, could evaluate commonalities and differences in
362 patterns of bias among regional collections.

363

364 *Spatial bias*

365 Our data confirmed the tendencies for botanists to collect along roadsides (*e.g.*, Funk &
366 Richardson, 2002), near herbaria (*e.g.*, Hijmans *et al.*, 2000), in more accessible areas
367 (Rich & Woodruff, 1992), and at lower elevations. Before automobiles became common
368 in the 1920s, people walked or rode domesticated animals (Botkin, 1968; Belasco, 1979).
369 As routes became established, they formed our modern infrastructure, including roads,
370 railroads, and cities that contain herbaria, and spatial biases associated with infrastructure
371 likely increased (Everill *et al.*, 2014). Because roads are known to fragment populations
372 and landscapes (*e.g.*, Forman & Alexander, 1998; Hui *et al.*, 2003; Griffith *et al.*, 2010;
373 Li *et al.*, 2014) and botanists and herbaria predominate in cities, specimens collected in
374 proximity to either are unlikely to represent a random sample across species distributions.
375 Species collected along roadsides are likely to be over-represented by species that thrive
376 with disturbance, and under-represented by forest interior and wetland species.

377 Collection bias towards lower elevations (< 500m) was most striking in SA,
378 despite extensive collection efforts along the hyper-diverse mountains of the Cape Fold
379 Belt. This is likely because of the presence of the arid and relatively depauperate Great
380 Karoo Plateau, which spreads across over a third of the country, but accounts for only a
381 small proportion of the region's biodiversity. As a result, the low-elevation collection
382 bias in SA may reflect actual species abundance.

383

384 *Temporal bias*

385 Collections in AU and SA have increased through time until only very recently, but those
386 in NE peaked in the early 1900s. These differences between regional collection activities

387 may parallel broader societal factors influencing plant collection. In NE, the
388 establishment of the New England Botanical Club during the 1890s (NEBC, 1899)
389 preceded a surge and peak in collecting activity associated with prolific botanical
390 expeditions of the region coinciding with the ‘Golden Age’ of plant collecting in Europe
391 and North America (Whittle, 1970; Musgrave *et al.*, 1999). In SA, collection efforts
392 began much later, peaked during the Apartheid Era from 1948 to 1994, and declined
393 thereafter under the New Democratic Rule, concomitant with the general economic
394 decline of the country and concern for public safety (Ferreira & Harmse, 2000; Lemanski,
395 2004). In AU, the mass immigration of Europeans in 1948 shortly after World War II that
396 included numerous highly skilled professionals (Price, 1998; Leuner, 2007) coincided
397 with a huge increase in botanical collecting. Botanical collecting may have declined more
398 recently because of legislation in AU and SA to regulate collections activities, especially
399 those designed to protect rare and endangered species.

400 Collecting efforts within a season revealed common patterns of bias: specimens in
401 the three regions were collected overwhelmingly in spring and summer. Sampling during
402 these time periods likely reflects efforts to represent the onset and peak of flowering in
403 these temperate regions. However, this seasonal bias likely overlooks key developmental
404 transitions (*e.g.*, Poethig, 2013), including bud formation, bud break, fruit maturation,
405 and leaf senescence (van der Schoot *et al.*, 2014). These temporal patterns also likely
406 reflect favorable conditions for fieldwork. Supporting this argument, these temporal
407 patterns were most pronounced in NE, which experiences the harshest winter climates
408 among the three regions. Collecting was also more likely during holidays and school
409 vacations in NE and AU.

410

411 ***Trait bias***

412 In all three regions, small to medium-height species were over-collected whereas tall
413 species (>5 m) were under-collected. This pattern presumably is related to the relative
414 ease of collecting specimens from shorter species, because reproductive materials are
415 more accessible for shorter specimens, and the dense distribution of numerous short,

416 frequently herbaceous species. Specimens of trees with woody twigs also are bulkier and
417 more difficult to prepare, which may reduce their collection frequency.

418 Threatened species were also greatly under-represented in all floras. This is
419 perhaps not surprising given their limited abundance (Palmer *et al.*, 2002) and imposed
420 collecting restrictions (Klemens & Thorbjarnarson, 1995; Pritchard, 1996; Gibbons *et al.*,
421 2000; Robinson, 2001). Regardless of formal restrictions, botanists now often avoid over-
422 collection of such species by following informal guidelines and collecting plants only in
423 areas with numerous individuals of the species (Iwanycki, 2009). Although careful
424 measures for collecting rare plants is important, under-collection of rare species may lead
425 to incorrect extinction risk assessments (that the species is rarer than it actually is) and
426 greatly limit opportunities to glean historic population and biogeographic data to guide
427 species conservation and restoration.

428 Annuals were over-represented relative to perennials in AU, while the opposite
429 was observed in SA and NE. Graminoids were over-represented in AU and SA, but they
430 were significantly under-represented in NE. This result may stem from the higher
431 likelihood of common species being collected multiple times by different individuals or
432 expeditions. Along these lines, much of AU is dominated by annual grasses, and the
433 savannas of SA are populated by a variety of native and non-native perennial grasses
434 interspersed with forbs and woody plants (Bond & Parr, 2010). New England, on the
435 other hand, is generally forested, with an abundance of shrubs and perennial herbs below.
436 Lianas and vines simultaneously represent the smallest proportion of growth forms and
437 comprise the least number of specimens per species in all three floras. Such trait-based
438 biases in botanical collections not only influence our perception of species abundance and
439 range, but can also lead to erroneous estimations of functional diversity and ecosystem
440 services, especially for studies relying on specimen databases (Schmidt-Lebuhn *et al.*,
441 2013).

442

443 ***Phylogenetic bias***

444 Taxonomic biases in collection data have been reported previously (Hijmans *et al.*, 2000;
445 Tobler *et al.*, 2007; Meyer *et al.*, 2016). However, our study is the first, to our knowledge,

446 to demonstrate explicit evidence for phylogenetic bias in herbarium collections.

447 Collection efforts in all three floras were concentrated in particular clades.

448 Previous examination of taxonomic bias has been hampered because they did not
449 evaluate clades lacking formal taxonomic ranks or overlooked other metrics of
450 evolutionary diversity. For instance, our phylogenetic approach not only captured
451 taxonomic bias, but revealed that evolutionarily distinct and globally endangered species
452 are underrepresented in herbarium records relative to other large clades (*e.g.*, Asteraceae,
453 Cyperaceae, Poaceae and Rosaceae in NE). Phylogenetically isolated species that are
454 threatened with extinction represent important targets for future collecting, and for
455 conservation prioritization.

456 Our findings also have strong links for understanding collector behavior. For
457 example, regional collectors with a particular interest in SA *Protea* species or NE *Rubus*
458 species would contribute to phylogenetic collecting bias. Similarly, phylogenetic bias
459 could result from collectors focusing on plants with certain phylogenetically conserved
460 traits, such as showy flowers or the presence of particular secondary metabolites (*e.g.*,
461 medicinal plants). Although our analysis of phylogenetic bias in herbarium collections
462 was far from conclusive and limited by taxa for which phylogenetic data were available,
463 our conclusions are supported by other studies demonstrating widespread taxonomic bias
464 on a global scale (Meyer *et al.*, 2016).

465

466 ***Collector bias***

467 In all regions we identified that a large percentage of specimens were gathered by only a
468 few collectors (Fig. 8). This implies that the habits and preferences of a few individuals
469 likely shaped the establishment and formation of herbarium collections. These “founder
470 effects” propagate across all the dimensions of collection bias examined above. For
471 example, certain collectors may focus on floristic regions and sample all species found
472 therein, whereas others may focus on collecting species of a particular clade across
473 various regions. Professional botanists may tend to collect specimens on weekdays during
474 any time of the year, whereas amateurs and faculty with teaching responsibilities may
475 focus their efforts on weekends and vacation months. Those interested in function and

476 physiology may only collect plants of certain habits or life-histories (*e.g.*, carnivorous or
477 succulent plants). These effects would be compounded when associated with mega-
478 collectors. For instance, the Harvard University Herbaria's collection of Asian plants
479 dates to the early establishment of the institution, and continues to attract scholars of the
480 flora of Asia and their collections. Investigating the historical significance and potential
481 biases created and propagated by these early pioneers is a ripe area for future research.

482

483 **Future collecting**

484 To ensure that herbaria continue to be vital centers for research beyond their importance
485 to taxonomy and systematics, herbarium directors and collectors should consider working
486 to accommodate and possibly reduce biases in plant collections. Biases can be accounted
487 for to a degree using statistical approaches that mitigate their effects (Droissart *et al.*,
488 2012; Feeley, 2012; Grass *et al.*, 2014; Engemann *et al.*, 2015). For instance,
489 comparisons of herbarium species collected near to, and far away from, urban areas and
490 other infrastructure (McCarthy *et al.*, 2012) or using rarefaction methods to predict
491 abundances (Schmidt-Lebuhn *et al.*, 2013) could be useful strategies to improve species
492 distribution models and predict future changes across a flora. Future collecting
493 expeditions should focus on "coldspots" of diversity (Hijmans *et al.*, 2000). Although
494 some of the coldspots we identified may represent inhospitable environments, they often
495 correspond to unique and irreplaceable ecosystems, including the Succulent Karoo of SA
496 and the North Maine Woods in northern NE. Some of these coldspots also may indicate
497 areas where herbarium specimens have yet to be digitized and mobilized, providing
498 additional focus for efforts to make collection data widely available.

499 Phylogenetic and trait biases can be alleviated by targeting collection efforts at
500 specimen gaps along these axes. Temporal bias is more difficult to address, as we cannot
501 add to historic collections. However, we can make efforts to maintain consistent regional
502 botanical records by conducting field surveys at regular intervals.

503 Some of the biases may also be attributed in part to longstanding curation
504 practices. As herbarium collections were amassed for qualitative floristic, taxonomic, and
505 systematic research, duplicate specimens of common species and similar specimens from

506 the same geographic area have been discarded or sent elsewhere. Indeed, many herbaria
507 refuse to accession new specimens belonging to regions or species that are already
508 represented in their current collections, despite increasing use of herbarium collections
509 for novel applications. This trend is becoming even more pronounced, as herbaria around
510 the world are increasingly constrained by funding, labor, and space. As new uses for
511 biological collections continue to proliferate, curation practices may have to change to
512 accommodate different avenues of research, such as climate change biology and rare
513 plant conservation. Finally, future collecting should strive to overcome “stamp collecting”
514 (*e.g.*, Vul *et al.*, 2014), the tendency to not collect additional specimens of a given species
515 from a specific location and time once another specimen has been collected there and
516 then. Although analyzing the impacts of all of these solutions is beyond the scope of this
517 study, future studies could statistically test these solutions using appropriate null models.
518

519

520 **ACKNOWLEDGMENTS**

521 We thank the Harvard University Herbaria for logistic and financial support, and the
522 virtual herbaria in the three regional floras for granting us access to their data: Australian
523 Virtual Herbarium (<http://avh.chah.org.au>), South African National Biodiversity Institute
524 (<http://newposa.sanbi.org/>) and Consortium for Northeast Herbaria
525 (<http://portal.neherbaria.org/portal/>). Digitization of most New England specimens was
526 funded by the ADBC program of the U.S. National Science Foundation (Awards
527 1208829, 1208835, 1208972, 1208973, 1208975, 1208989, 1209149). Special thanks to
528 T.J. Davies, E.K. Meineke, K.M. Peterson, and K.G. Dexter for valuable discussion
529 during the formation of this manuscript.

530

531

532 **Author Contributions:** Conceived the project: CCD. Designed the experiment: BHD,
533 DSP. Performed the experiments: BHD. Analyzed the data: BHD with help from
534 DSP. Contributed reagents/materials/analysis tools: BHD, DSP, CGW, AME,
535 CCD. Wrote the paper: BHD with significant comments and editing from all co-
536 authors.

538 **References**

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

541 **Agostinelli C, Lund U. 2013.** *R package 'circular': Circular Statistics (version 0.4-7).*

542 URL <https://r-forge.r-project.org/projects/circular/>

543 **AVH. 2016.** *Australia's Virtual Herbarium, Council of Heads of Australasian Herbaria,*

544 <http://avh.chah.org.au>, accessed on 09 June 2016.

545 **Belasco WJ. 1979.** *Americans on the road, from autocamp to motel 1910-1945.*

546 Baltimore: Johns Hopkins University Press.

547 **Bivand RS, Pebesma E, Gomez-Rubio V. 2013.** *Applied spatial data analysis with R,*

548 *Second edition.* Springer, NY. <http://www.asdar-book.org/>

549 **Blomberg SP, Garland T, Ives AR. 2003.** Testing for phylogenetic signal in

550 comparative data: behavioural traits are more labile. *Evolution* **57**: 717–745.

551 **Bond WJ, Parr CL. 2010.** Beyond the forest edge: ecology, diversity and conservation

552 of the grassy biomes. *Biological Conservation* **143**: 2395–2404.

553 **Botkin BA. 1968.** Automobile humor: from the horseless carriage to the compact car.

554 *The Journal of Popular Culture* **I**: 395–402.

555 **Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzherin D, Rees T, Matasci N, Narro ML,**

556 **Piel WH, Mckay SJ, et al. 2013.** The taxonomic name resolution service: an

557 online tool for automated standardization of plant names. *BMC Bioinformatics* **14**:

558 16.

559 **Ceballos G, Ehrlich PR. 2006.** Global mammal distributions, biodiversity hotspots, and

560 conservation. *Proceedings of the National Academy of Sciences USA* **103**: 19374–

561 19379.

562 **CIESIN. 2016.** *Center for International Earth Science Information Network, Columbia*

563 *University.* Gridded Population of the World, Version 4 (GPWv4): Population

564 Density. Palisades, NY: NASA Socioeconomic Data and Applications Center

565 (SEDAC). <http://dx.doi.org/10.7927/H4NP22DQ>. Accessed 29 August 2016.

566 **CNH. 2016.** Consortium of Northeastern Herbaria. <http://portal.neherbaria.org/portal/>

567 **Cotterill FPD, Hustler CW, Broadley DG. 1994.** Systematics and biodiversity. *Trends*

568 *in Ecology and Evolution* **9**: 228.

- 569 **Cotterill FPD. 1995.** Systematics, biological knowledge and environmental conservation.
570 *Biodiversity and Conservation* **4**: 183–205.
- 571 **Dalton R. 2003.** Natural history collections in crisis as funding is slashed. *Nature* **423**:
572 6940.
- 573 **Daru BH, Van der Bank M, Davies TJ. 2015.** Spatial incongruence among hotspots
574 and complementary areas of tree diversity in southern Africa. *Diversity and*
575 *Distributions* **21**: 769–780.
- 576 **Davies TJ, Kraft NJB, Salamin N, Wolkovich EM. 2012.** Incompletely resolved
577 phylogenetic trees inflate estimates of phylogenetic conservatism. *Ecology* **93**:
578 242–247.
- 579 **Davis CC, Willis CG, Connolly B, Kelly C, Ellison AM. 2015.** Herbarium records are
580 reliable sources of phenological change driven by climate and provide novel
581 insights into species’ phenological cueing mechanisms. *American Journal of*
582 *Botany* **102**: 1599–1609.
- 583 **Droissart V, Hardy OJ, Sonké B, Dahdouh-Guebas F, Stévant T. 2012.** Subsampling
584 herbarium collections to assess geographic diversity gradients: A case study with
585 endemic Orchidaceae and Rubiaceae in Cameroon. *Biotropica* **44**: 44–52.
- 586 **Edwards EJ, de Vos JM, Donoghue MJ. 2015.** Doubtful pathways to cold tolerance in
587 plants. *Nature* **521**: E5–E6.
- 588 **Edwards JL, Lane MA, Nielsen ES. 2000.** Interoperability of biodiversity databases:
589 biodiversity information on every desktop. *Science* **289**: 2312–2314.
- 590 **Engemann K, Enquist BJ, Sandel B, Boyle B, Jørgensen PM, Morueta-Holme N,**
591 **Peet RK, Violle C, Svenning J-C. 2015.** Limited sampling hampers “big data”
592 estimation of species richness in a tropical biodiversity hotspot. *Ecology and*
593 *Evolution* **5**: 807–820.
- 594 **Everill PH, Primack RB, Ellwood EE, Melaas EK. 2014.** Determining past leaf-out
595 times of New England’s deciduous forests from herbarium specimens. *American*
596 *Journal of Botany* **101**: 1–8.
- 597 **Feeley KJ. 2012.** Distributional migrations, expansions, and contractions of tropical plant
598 species as revealed in dated herbarium records. *Global Change Biology* **18**: 1335–
599 1341.

- 600 **Ferreira SLA, Harmse AC. 2000.** Crime and tourism in South Africa: international
601 tourists perception and risk. *South African Geographical Journal* **82**: 80–85.
- 602 **Forman RTT, Alexander LE. 1998.** Roads and their major ecological effects. *Annual*
603 *Review of Ecology and Systematics* **29**: 207–31
- 604 **Forman RTT, Friedman DS, Fitzhenry D, Martin JD, Chen AS, Alexander LE. 1995.**
605 Ecological effects of roads: Toward three summary indices and an overview for
606 North America. In: Canters K, ed. *Habitat fragmentation and infrastructure*.
607 Ministry of Transport, Public Works and Water Management: Maastricht and The
608 Hague, Netherlands, 40–54.
- 609 **Fortune S. 1992.** Voronoi diagrams and Delaunay triangulations. *Computing in*
610 *Euclidean Geometry* **1**: 193–233.
- 611 **Funk V. 2003.** The importance of herbaria. *Plant Science Bulletin* **49**: 94–95.
- 612 **Funk VA, Morin N. 2000.** A survey of the herbaria of the southeast United States. *Sida,*
613 *Botanical Miscellany* **18**: 35–52.
- 614 **Funk VA, Richardson K. 2002.** Biological specimen data in biodiversity studies: use it
615 or lose it. *Systematic Biology* **51**: 303–316.
- 616 **GADM. 2015.** *Global Administrative Areas*, version 2.8 (www.gadm.org).
- 617 **Geri F, Lastrucci L, Viciani D, Foggi B, Ferretti G, Maccherini S, Bonini I, Amici V,**
618 **Chiarucci A. 2013.** Mapping patterns of ferns species richness through the use of
619 herbarium data. *Biodiversity and Conservation* **22**: 1679–1690.
- 620 **Gibbons JW, Scott DE, Ryan T, Buhlmann K, Tuberville T, Greene J, Mills T,**
621 **Leiden Y, Poppy S, Winne C et al. 2000.** The global decline of reptiles, déj·vu
622 amphibians. *BioScience* **50**: 653–666.
- 623 **Grass A, Tremetsberger K, Hössinger R, Bernhardt K. 2014.** Change of species and
624 habitat diversity in the Pannonian region of eastern Lower Austria over 170 years:
625 Using herbarium records as a witness. *Natural Resources* **5**: 583–596.
- 626 **Greve M, Lykke AM, Fagg CW, Gereau RE, Lewis GP, Marchant R, Marshall AR,**
627 **Ndayishimiye J, Bogaert J, Svenning JC. 2016.** Realising the potential of
628 herbarium records for conservation biology. *South African Journal of Botany* **105**:
629 317–323.

- 630 **Griffith EH, Sauer JR, Royle JA. 2010.** Traffic effects on bird counts on North
631 American breeding bird survey routes. *Auk* **127**: 387–393.
- 632 **Hart R, Salick J, Ranjitkar S, Xu J. 2014.** Herbarium specimens show contrasting
633 phenological responses to Himalayan climate. *Proceedings of the National*
634 *Academy of Sciences USA* **111**: 10615–10619.
- 635 **Hijmans RJ, Garrett KA, Huaman Z, Zhang DP, Schreuder M, Bonierbale M. 2000.**
636 Assessing the geographic representation of genebank collections: the case of the
637 Bolivian wild potatoes. *Conservation Biology* **14**: 1755–1765.
- 638 **Hijmans RJ. 2015.** *geosphere: Spherical Trigonometry*. R package version 1.4-3.
639 <http://CRAN.R-project.org/package=geosphere>
- 640 **Hortal J, Lobo JM, Jiménez-Valverde A. 2007.** Limitations of biodiversity databases:
641 case study on seed-plant diversity in tenerife, canary islands. *Conservation*
642 *Biology* **21**: 853–863.
- 643 **Hui C, Shuang-cheng L, Yi-li Z. 2003.** Impact of road construction on vegetation
644 alongside Qinghai-Xizang highway and railway. *Chinese Geographical Science*
645 **13**: 340–346.
- 646 **Isaac NJ, Turvey ST, Collen B, Waterman C, Baillie JE. 2007.** Mammals on the
647 EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE* **2**: e296.
- 648 **Iwanycki N. 2009.** *Guidelines for collecting herbarium specimens of vascular plants*.
649 Royal Botanical Gardens Canada.
- 650 **Jombart T, Dray S. 2008.** adephylo: exploratory analyses for the phylogenetic
651 comparative method. *Bioinformatics* **26**: 1907–1909.
- 652 **Klemens MW, Thorbjarnarson JB. 1995.** Reptiles as a food resource. *Biodiversity and*
653 *Conservation* **4**: 281–298.
- 654 **Lavoie C. 2013.** Biological collections in an ever changing world: Herbaria as tools for
655 biogeographical and environmental studies. *Perspectives in Plant Ecology,*
656 *Evolution and Systematics* **15**: 68–76.
- 657 **le Roux MM, Wilkin P, Balkwill K, Boatwright JS, Bytebier B, Filer D, Klak C,**
658 **Klopper RR, Koekemoer M, Livermore L et al. 2017.** Producing a plant
659 diversity portal for South Africa. *Taxon* **66**: 421–431.

- 660 **Lees DC, Lack HW, Rougerie R, Hernandez-Lopez A, Raus T, Avtzis ND, Augustin**
661 **S, Lopez-Vaamonde C. 2011.** Tracking origins of invasive herbivores through
662 herbaria and archival DNA: the case of the horse-chestnut leaf miner. *Frontiers in*
663 *Ecology and the Environment* **9**: 322–328.
- 664 **Lemanski C. 2004.** A new apartheid? The spatial implications of fear of crime in Cape
665 Town, South Africa. *Environment & Urbanization* **16**: 101–111.
- 666 **Leuner B. 2007.** *Migration, multiculturalism and language maintenance in Australia.*
667 Peter Lang, Oxford.
- 668 **Li Y, Yu J, Ning K, Du S, Han G, Qu F, Wang G, Fu Y, Zhan C. 2014.** Ecological
669 effects of roads on the plant diversity of coastal wetland in the Yellow River Delta.
670 *The Scientific World Journal* **2014**: 952051.
- 671 **McCarthy KP, Fletcher JR RJ, Rota CT, Hutto RL. 2012.** Predicting species
672 distributions from samples collected along roadsides. *Conservation Biology* **26**:
673 68–77.
- 674 **Meyer C, Weigelt P, Kreft H. 2016.** Multidimensional biases, gaps and uncertainties in
675 global plant occurrence information. *Ecology Letters* **19**: 992–1006.
- 676 **Miller-Rushing A, Primack R, Mukunda S. 2006.** Photographs and herbarium
677 specimens as tools to document phenological changes in response to global
678 warming. *American Journal of Botany* **93**: 1667–1674.
- 679 **Musgrave T, Gardner C, Musgrave W. 1999.** *The plant hunters. Two hundred years of*
680 *adventure and discovery.* Seven Dials.
- 681 **NEBC. 1899.** Editorial announcement. *Rhodora* **1**: 1–2
- 682 **Newbold T. 2010.** Applications and limitations of museum data for conservation and
683 ecology, with particular attention to species distribution models. *Progress in*
684 *Physical Geography* **34**: 3–22.
- 685 **Norris WR, Lewis DQ, Widrlechner MP, Thompson JD, Pope RO. 2001.** Lessons
686 from an inventory of the Ames, Iowa, flora (1859–2000). *Journal of the Iowa*
687 *Academy of Science* **108**: 34–63.
- 688 **Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017.** Assessing the relevance of
689 herbarium collections as tools for conservation biology. *Botanical Review*
690 doi:10.1007/s12229-017-9188-z

- 691 **Orme CD, Davies RG, Burgess M, Eigenbrod F, Pickup N, Olson VA, Webster AJ,**
692 **Ding TS, Rasmussen PC, Ridgely RS, et al. 2005.** Global hotspots of species
693 richness are not congruent with endemism or threat. *Nature* **436**: 1016–1019.
- 694 **Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012.**
695 *caper: Comparative Analyses of Phylogenetics and Evolution in R.* R package
696 version 0.5.[http://CRAN.R-project.org/package = caper](http://CRAN.R-project.org/package=caper).
- 697 **Pagel M. 1999.** Inferring the historical patterns of biological evolution. *Nature* **401**: 877–
698 884.
- 699 **Palmer MW, Earls PG, Hoagland BW, White PS, Wohlgemuth T 2002.** Quantitative
700 tools for perfecting species list. *Environmetrics* **13**: 121–137.
- 701 **Poethig, RS. 2013.** Vegetative phase change and shoot maturation in plants. *Current*
702 *Topics in Developmental Biology* **105**: 125–152.
- 703 **Prather LA, Alvarez-Fuentes O, Mayfield MH, Ferguson CJ. 2004a.** The decline of
704 plant collecting in the United States: a threat to the infrastructure of biodiversity
705 studies. *Systematic Botany* **29**: 15–28.
- 706 **Prather LA, Alvarez-Fuentes O, Mayfield MH, Ferguson CJ. 2004b.** Implications of
707 the decline in plant collecting for systematic and floristic research. *Systematic*
708 *Botany* **29**: 216–220.
- 709 **Price CA. 1998.** Post-war immigration: 1945-1998. *Journal of the Australian Population*
710 *Association* **15**: 17.
- 711 **Pritchard PCH. 1996.** The Galápagos tortoises: nomenclatural and survival status.
712 Lunenburg (MA): Chelonian Research Foundation. *Chelonian Research*
713 *Monographs* **1**.
- 714 **Pyke GH, Ehrlich PR. 2010.** Biological collections and ecological/environmental
715 research: a review, some observations and a look to the future. *Biological Reviews*
716 **5**: 247–266.
- 717 **Redding DW, Mooers AØ. 2006.** Incorporating evolutionary measures into conservation
718 prioritization. *Conservation Biology* **20**: 1670–1678.
- 719 **Revell LJ. 2012.** phytools: An R package for phylogenetic comparative biology (and
720 other things). *Methods in Ecology and Evolution* **3**: 217–223.

- 721 **Rich TCG, Woodruff ER. 1992.** Recording bias in botanical surveys. *Watsonia* **19**: 73–
722 95.
- 723 **Robinson JG. 2001.** Using ‘sustainable use’ approaches to conserve exploited
724 populations. In: Reynolds JD, Mace GM, Redford KH, Robinson JG, eds.
725 *Conservation of exploited species*. Cambridge: Cambridge University Press, 485–
726 498.
- 727 **Rudin SM, Murray DW, Whitfeld TJS. 2017.** Retrospective analysis of heavy metal
728 contamination in Rhode Island based on old and new herbarium specimens.
729 *Applications in Plant Sciences* **5**: 1–13.
- 730 **SANBI. 2016.** *South African National Biodiversity Institute*. Botanical Database of
731 Southern Africa (BODATSA), <http://newposa.sanbi.org/>, accessed on 22 July
732 2016.
- 733 **Schaefer H, Hardy OJ, Silva L, Barraclough TG, Savolainen V. 2011.** Testing
734 Darwin’s naturalization hypothesis in the Azores. *Ecology Letters* **14**: 389–396.
- 735 **Schmidt-Lebuhn AN, Knerr NJ, Kessler M. 2013.** Non-geographic collecting biases in
736 herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and*
737 *Conservation* **22**: 905–919.
- 738 **Schorn C, Weber E, Bernardos R, Hopkins C, Davis CC. 2016.** The New England
739 Vascular Plants Project: 295,000 specimens and counting. *Rhodora* **118**: 324–325.
- 740 **Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Stielow B,**
741 **Geml J, Richardson JE, Bakker FT. 2013.** Genomic treasure troves: complete
742 genome sequencing of herbarium and insect museum specimens. *PLoS ONE* **8**:
743 e69189.
- 744 **Stropp J, Ladle RJ, Malhado ACM, Hortal J, Gaffuri J, Temperley, WH, Skøien JO,**
745 **Mayaux, P. 2016.** Mapping ignorance: 300 years of collecting flowering plants in
746 Africa. *Global Ecology and Biogeography* **25**: 1085–1096.
- 747 **Syfert MM, Smith MJ, Coomes DA. 2013.** The effects of sampling bias and model
748 complexity on the predictive performance of MaxEnt species distribution models.
749 *PLoS ONE* **8**: e55158.

- 750 **Thiers B. 2016.** *Index Herbariorum: A global directory of public herbaria and*
751 *associated staff.* New York Botanical Garden's Virtual Herbarium.
752 <http://sweetgum.nybg.org/science/ih/>.
- 753 **Tobler M, Honorio E, Janovec J, Reynel C. 2007.** Implications of collection patterns of
754 botanical specimens on their usefulness for conservation planning: an example of
755 two neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodiversity*
756 *and Conservation* **16**: 659–677
- 757 **van der Schoot C, Paul LK, Rinne PLH. 2014.** The embryonic shoot: a lifeline through
758 winter. *Journal of Experimental Botany* **65**: 1699–1712.
- 759 **Vul E, Goodman N, Griffiths TL, Tenenbaum JB. 2014.** One and done? Optimal
760 decisions from very few samples. *Cognitive Science* **38**: 599–637.
- 761 **Webb CO, Ackerly DD, Kembel SW. 2008.** PHYLOCOM: software for the analysis of
762 phylogenetic community structure and trait evolution. *Bioinformatics* **24**: 2098–
763 2100.
- 764 **Webb CO, Ackerly DD, McPeck MA, Donoghue MJ. 2002.** Phylogenies and
765 community ecology. *Annual Review of Ecology and Systematics* **33**: 475–505.
- 766 **Webb CO, Donoghue MJ. 2005.** Phylomatic: tree assembly for applied phylogenetics.
767 *Molecular Ecology Notes* **5**: 181–183.
- 768 **Whittle T. 1970.** *The Plant Hunters*. Heinemann, London.
- 769 **Willis CG, Ellwood ER, Primack RB, Davis CC, Pearson KD, Gallinato AS, Yost**
770 **JM, Nelson G, Mazer SJ, Rossington NL et al. 2017a.** Old plants, new tricks:
771 phenological research using herbarium specimens. *Trends in Ecology & Evolution*
772 **32**: 531–546.
- 773 **Willis CG, Law E, Williams AC, Franzone BF, Bernardos R, Brun L, Hopkins C,**
774 **Schorn C, Weber E, Parks DS et al. 2017b.** CrowdCurio: an online
775 crowdsourcing platform to facilitate climate change studies using herbarium
776 specimens. *New Phytologist* **215**: 479–488.
- 777 **Wolf A, Anderegg WRL, Ryan SJ, Christensen J. 2011.** Robust detection of plant
778 species distribution shifts under biased sampling regimes. *Ecosphere* **2**: 115.
- 779 **Wolkovich EM, Davies TJ, Schaefer H, Cleland EE, Cook BI, Travers SE, Willis**
780 **CG, Davis CC. 2013.** Temperature-dependent shifts in phenology contribute to

781 the success of exotic species with climate change. *American Journal of Botany*
782 **100**: 1407–1421.

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

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

788

789

790 LEGENDS TO FIGURES

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

798

799 **Fig. 2** Comparison of geographic sampling bias of herbarium records in relation to (a) the
800 minimum distance to roads, (b) minimum distance to herbaria (b), and (c) regional
801 altitudes at sampling locales. Black lines in (a) and (b) correspond to sampling locales
802 and red indicates an equal number of random points generated 1000 times. Dark grey
803 shading in (c) corresponds to sampling locales in relation to the regional altitudes *i.e.*, all
804 other altitudes (in red) for all three floras, Australia (left), South Africa (middle) and New
805 England (right). Dotted line in (c) indicates altitude at 500 m above sea level.

806

807 **Fig. 3** Timeline of herbarium specimen collection density in relation to major historical
808 events in time (indicated in red text) for the three floras: Australia, South Africa and New
809 England. Analysis of phylogenetic structure through time by binning sequences of
810 collection dates into decades and testing for overdispersion *vs.* clustering, are indicated in
811 black font. The red trend line indicates the gross domestic product of each region.

812

813 **Fig. 4** Temporal biases in herbarium collections. (a) Comparison of density plots of
814 collection dates by seasons of the year of herbarium records (blue line) with the dates
815 spanning the entire duration of collection (red line); blue lines outside the red lines
816 indicate over-collecting at a particular time of year, and (b) Distribution of collection
817 dates by days of the week for the three floras. Australia (n = 4,579,321 collection dates),
818 South Africa (n = 771,991 collection dates), and New England (n = 562,587 collection
819 dates).

820

821 **Fig. 5** Assessment of bias in plant traits. (a) growth duration, (b) growth form, (c) height,
822 and (d) extinction risk, for the floras of Australia (left pane), South Africa (middle pane)
823 and New England (right pane).

824

825 **Fig. 6** Distribution of phylogenetic bias, the tendency of closely related species to be
826 similarly collected in herbarium records for three floras: (a) Australia, (b) South Africa,
827 and (c) New England. Collecting effort is not phylogenetically random, but tends to be
828 clustered in few selected lineages. The color scales correspond to the log numbers of
829 specimens per species and ranges from red (low number of specimens per species) to blue
830 (high number of specimens per species).

831

832 **Fig. 7** Phylogenetic bias in collection frequency for exemplar families in New England
833 flora. Phylogenetic bias is indicated by significant phylogenetic signal in at least one of
834 three metrics (Abouheif's C_{mean} , Blomberg's K and Pagel's λ . The color bar illustrates
835 values within families: log numbers of specimens per species and ranges from red (low
836 number of specimens per species) to blue (high number of specimens per species).

837

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

843

844

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

AUSTRALIA	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept
	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	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept
	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	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept
	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	

846

847

848

849 **Table 2.** Results of the tests of phylogenetic signal in the number of specimens collected per species using three methods (Abouheif's
 850 C_{mean} , Blomberg's K and Pagel's λ). Phylogenetic data is derived from Zanne *et al.* (2014). All tests are based on 1000 randomizations.
 851 **P < 0.001; *P < 0.01; NS, P > 0.05

	Australia (n = 5814 species)	South Africa (n = 3568 species)	New England (n = 4269 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 lambda	0.18**	0.32**	0.29**

852

853

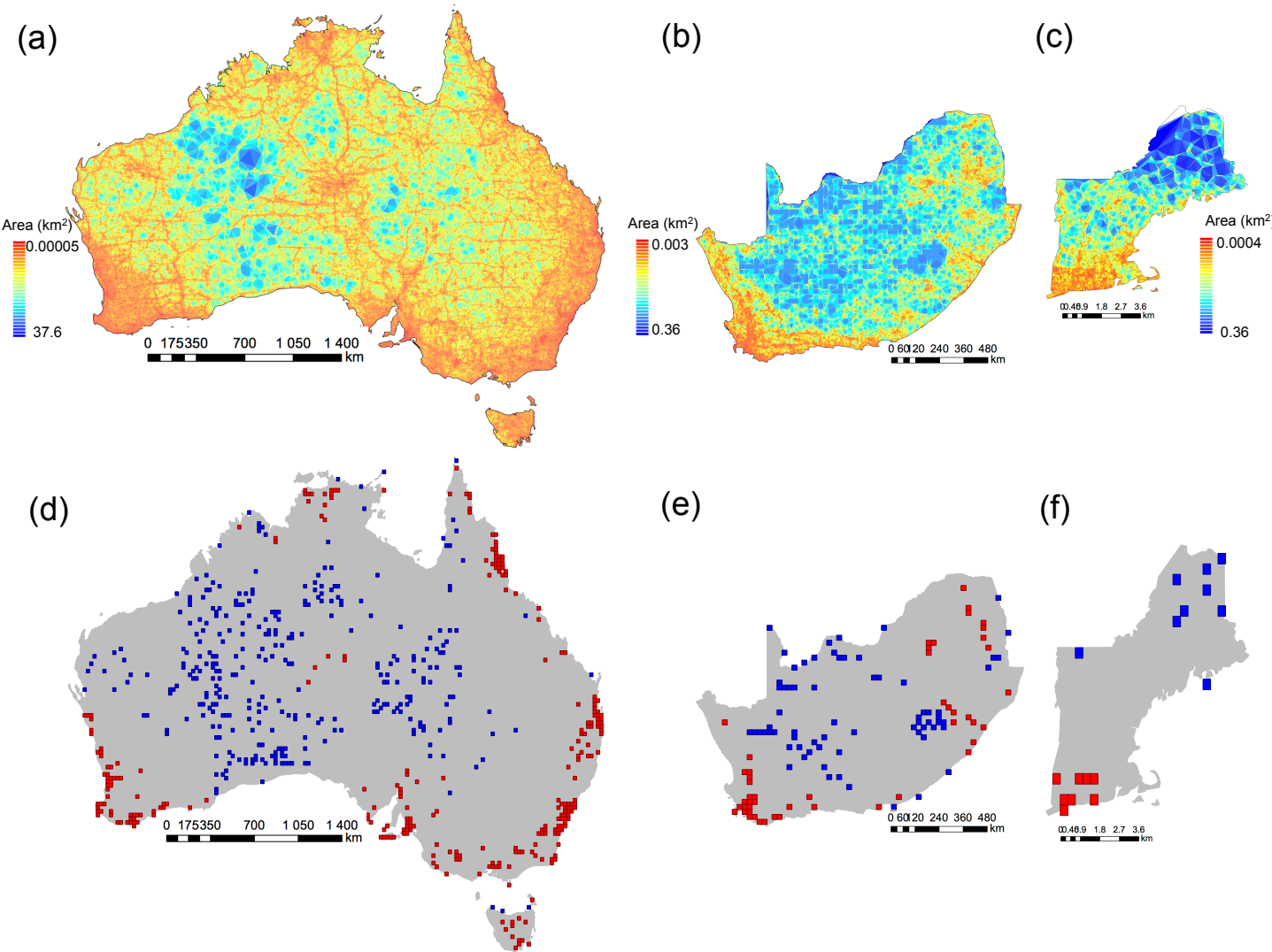
854

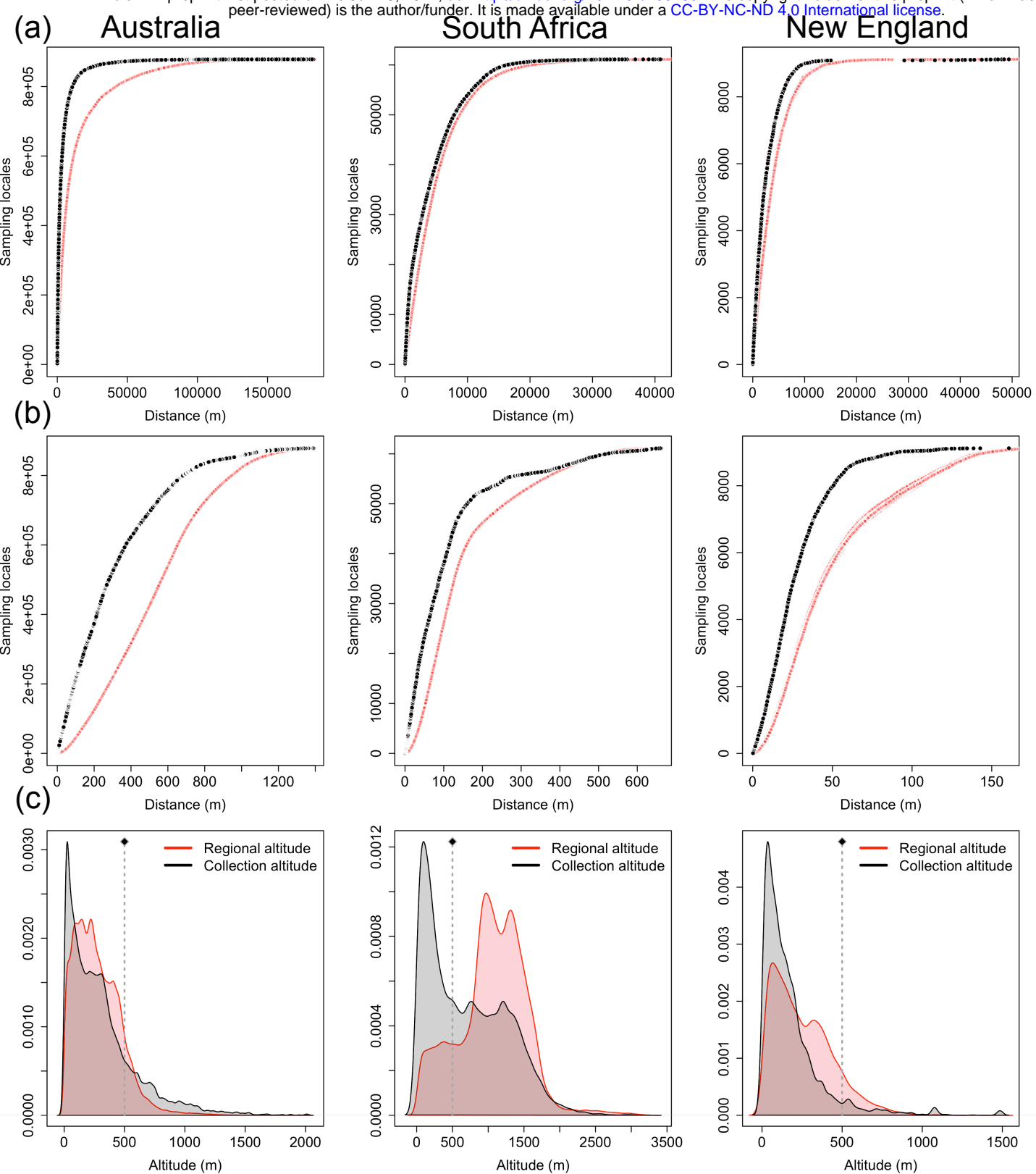
855 **Table 3** Multiple regressions of phylogenetic generalized least squares of collecting effort (frequency) of herbarium specimens with
 856 phylogenetic metrics of species uniqueness.

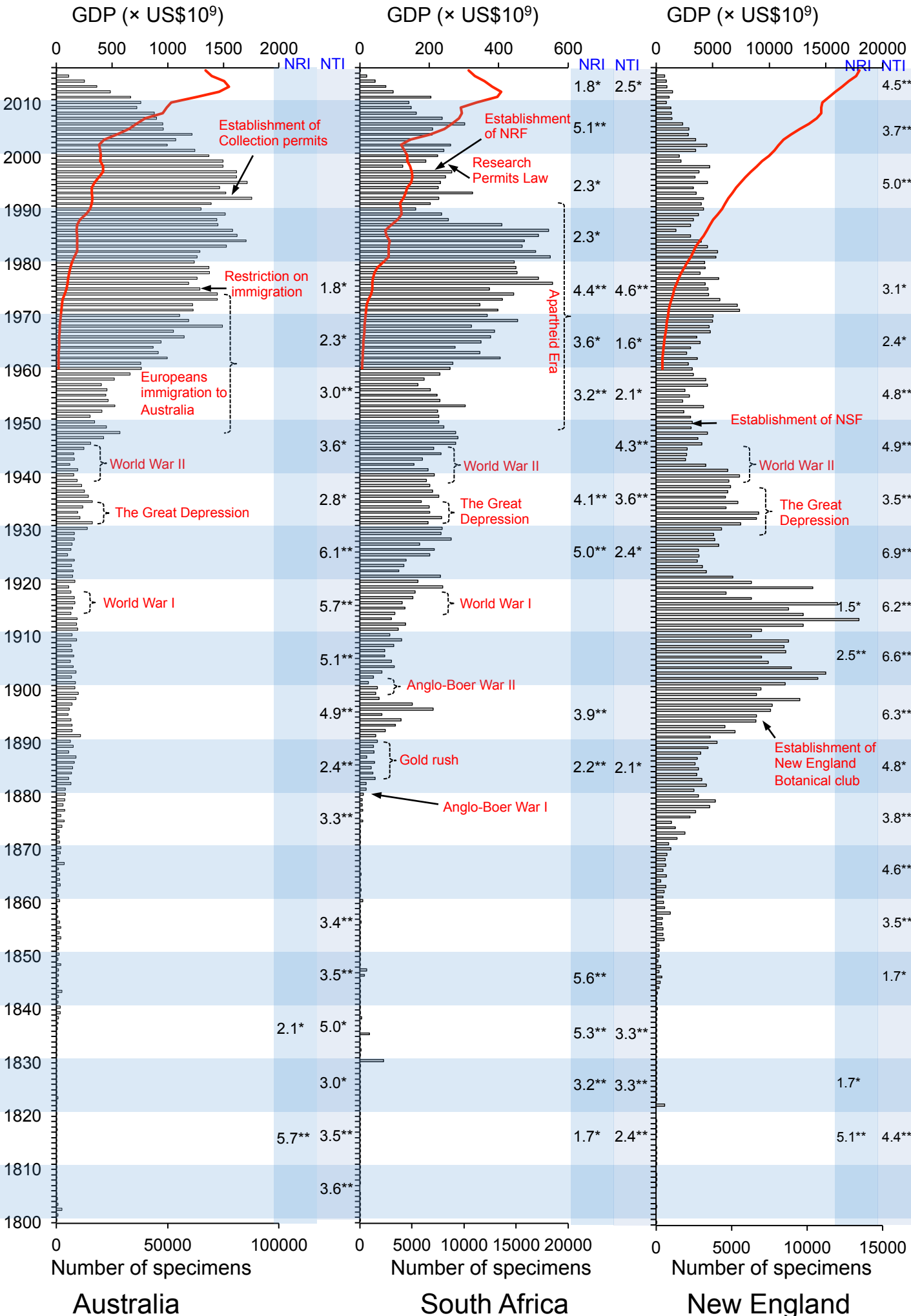
Australia	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept
	BL	1.36	0.7	0.049	0.035	4.37
	ED	0.2	0.008		0.44	
	EDGE	3.75	<0.001		-1.23	
South Africa	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept
	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	Predictors (log ₁₀ -transformed)	Percentage of variance explained (%)	P values	Model adjusted R ²	Model slope	Model intercept

	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	

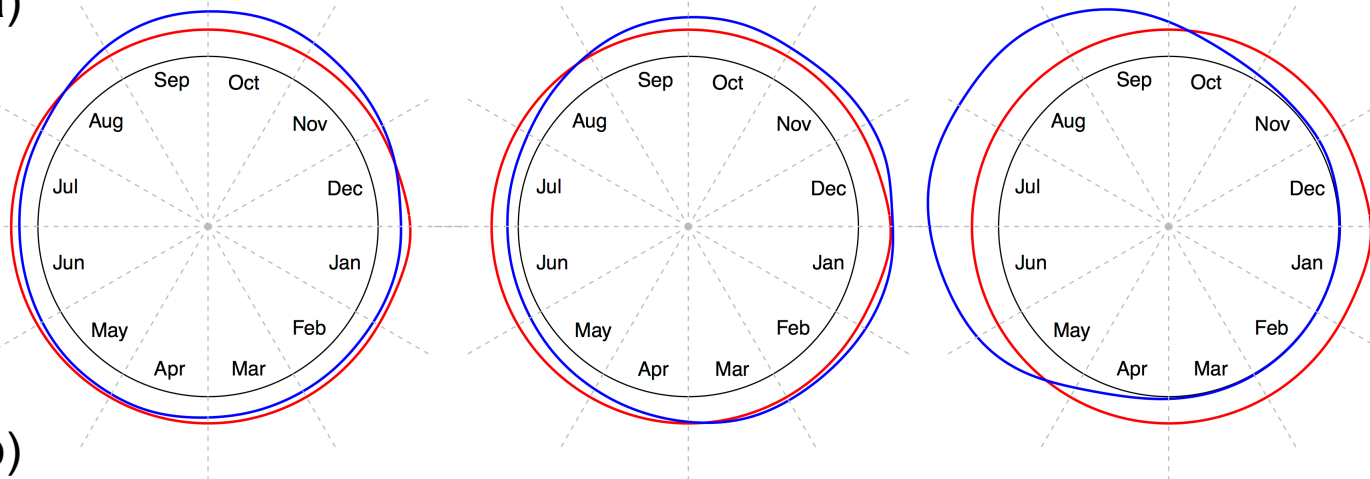
857



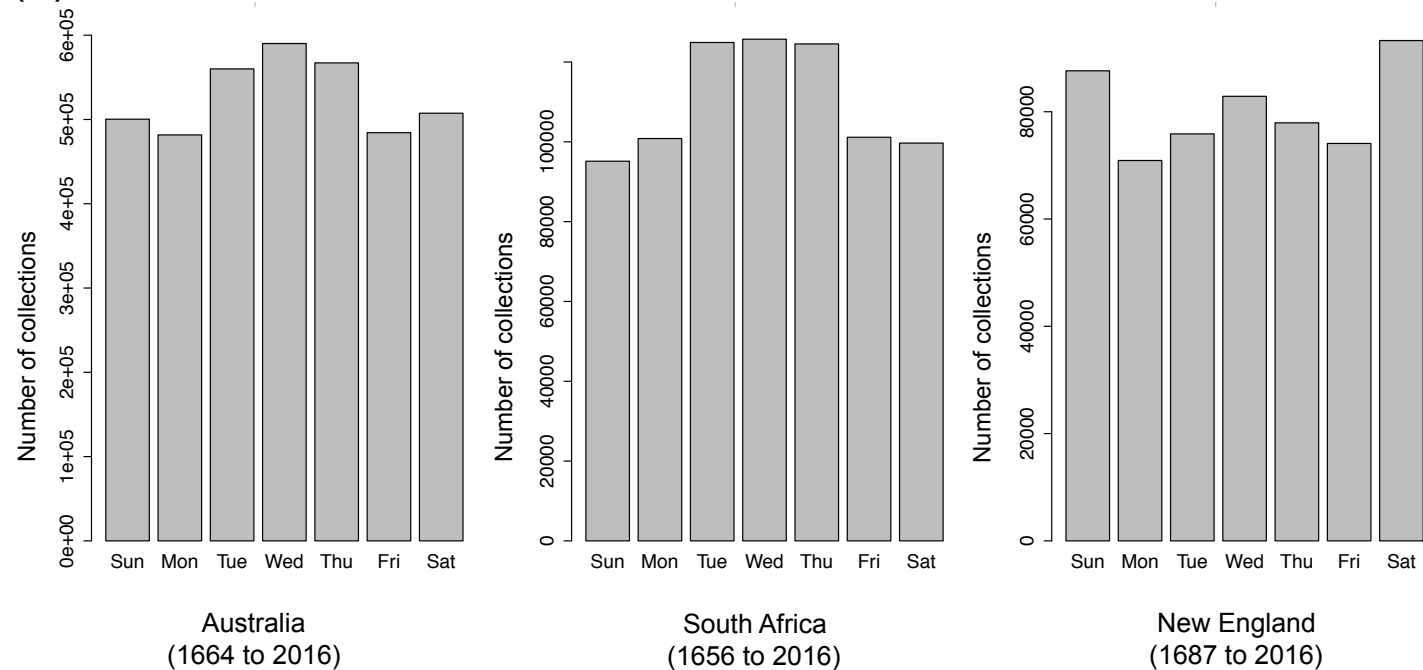


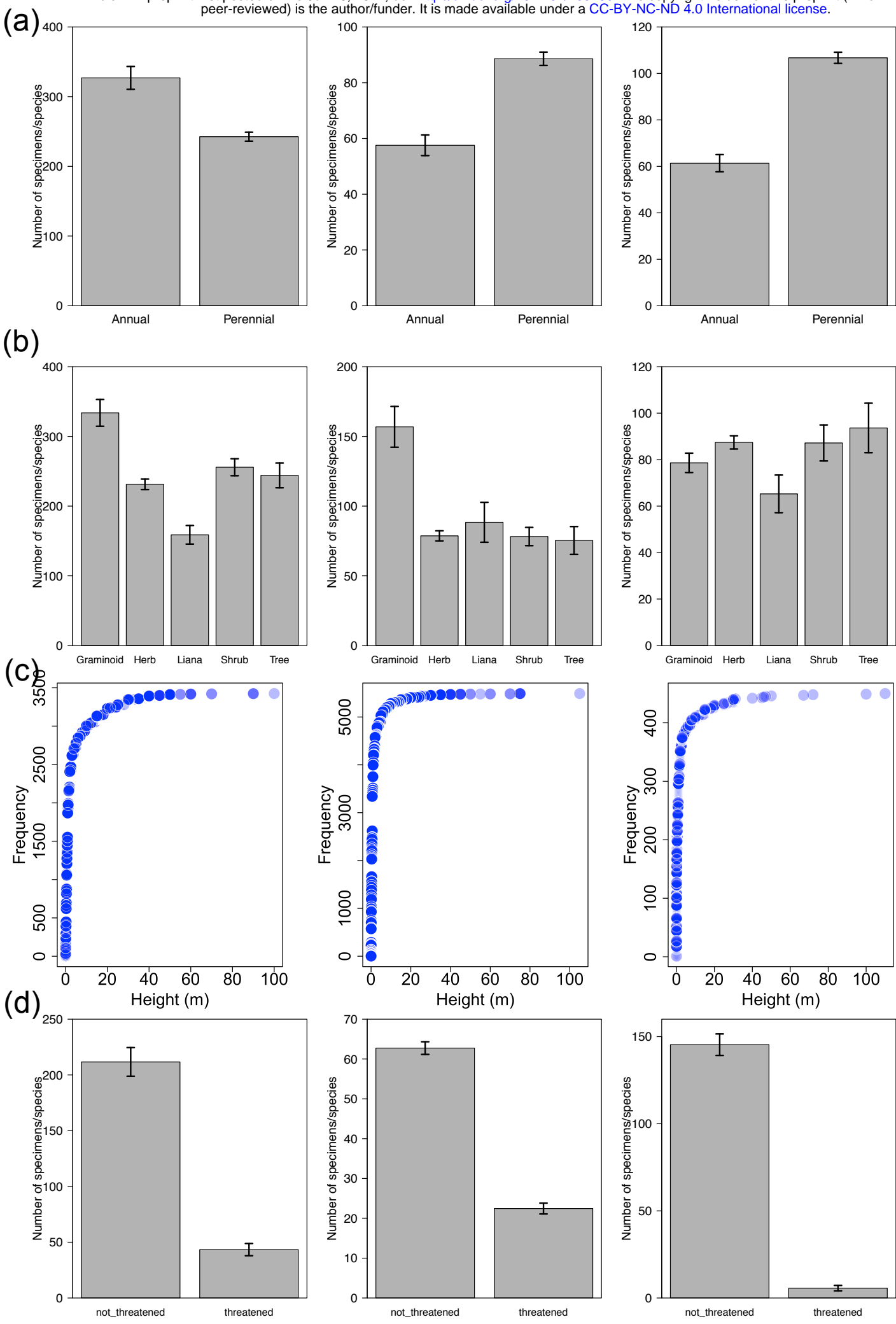


(a)

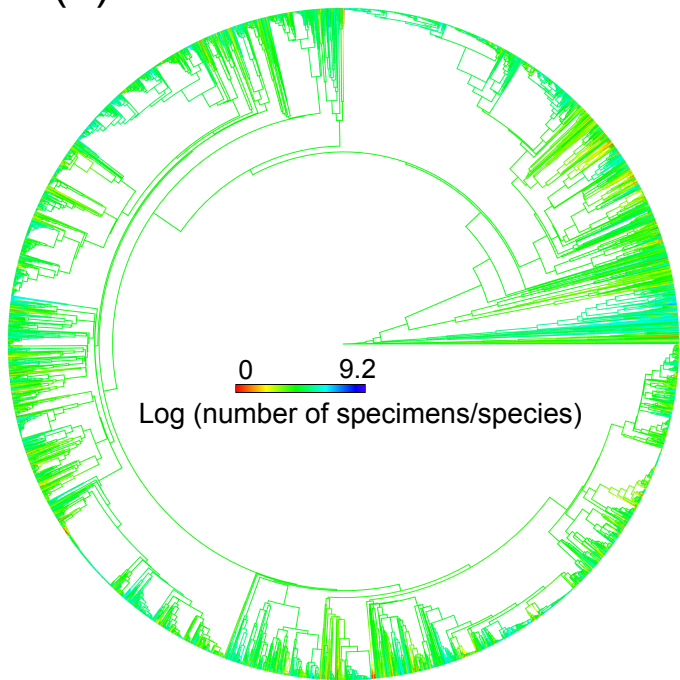


(b)

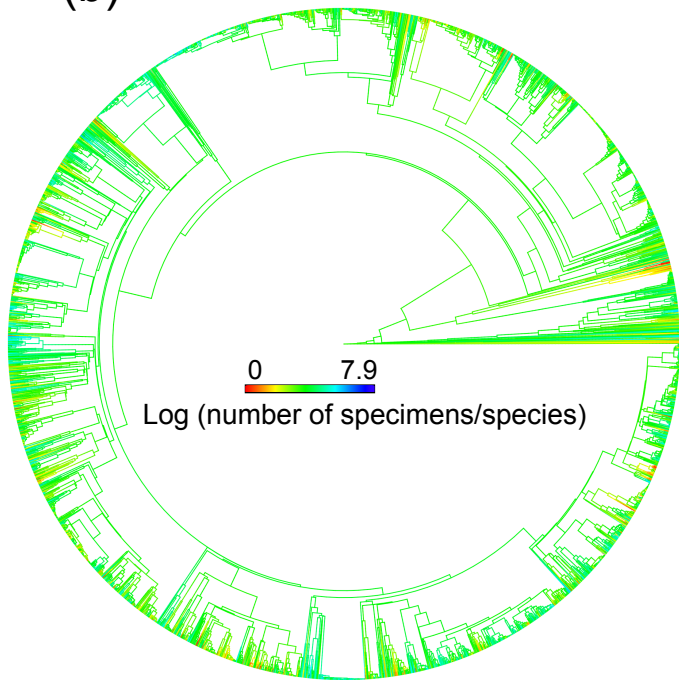




(a)



(b)



(c)

