



Identifying biodiversity knowledge gaps for conserving South Africa's endemic flora

Lerato N. Hoveka¹ · Michelle van der Bank¹ · Bezeng S. Bezeng² · T. Jonathan Davies^{1,3}

Received: 8 September 2019 / Revised: 30 April 2020 / Accepted: 25 May 2020 /

Published online: 13 June 2020

© The Author(s) 2020

Abstract

As a megadiverse country with a rapidly growing population, South Africa is experiencing a biodiversity crisis: natural habitats are being degraded and species are becoming threatened with extinction. In an era of big biodiversity data and limited conservation resources, conservation biologists are challenged to use such data for cost-effective conservation planning. However, while extensive, key genomic and distributional databases remain incomplete and contain biases. Here, we compiled data on the distribution of South Africa's > 10,000 endemic plant species, and used species distribution modelling to identify regions with climate suitable for supporting high diversity, but which have been poorly sampled. By comparing the match between projected species richness from climate to observed sampling effort, we identify priority areas and taxa for future biodiversity sampling. We reveal evidence for strong geographical and taxonomic sampling biases, indicating that we have still not fully captured the extraordinary diversity of South Africa's endemic flora. We suggest that these knowledge gaps contribute to the insufficient protection of plant biodiversity within the country—which reflect part of a broader *Leopoldean shortfall* in conservation data.

Keywords Endemism · Biodiversity knowledge · Sampling gaps · Species richness · Wallacean and Darwinian shortfalls

Communicated by Daniel Sanchez Mata.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10531-020-01998-4>) contains supplementary material, which is available to authorized users.

✉ Lerato N. Hoveka
leratohoveka@gmail.com

¹ African Centre for DNA Barcoding, University of Johannesburg, APK Campus, PO Box 524, Johannesburg 2006, South Africa

² BirdLife South Africa, Private Bag X16, Pinegowrie, Johannesburg 2123, South Africa

³ Biodiversity Research Centre, University of British Columbia, 2212 Main Mall, Vancouver, BC V6T 1Z4, Canada

Introduction

Life on earth is unevenly distributed and so too is our knowledge about it. In the past two decades, efforts to identify biodiversity knowledge gaps have attracted considerable interest among scientist and conservation practitioners (e.g. Scarascia-Mugnozza et al. 2000; Jarnevich et al. 2006; Costion et al. 2015; Oliveira et al. 2019). This interest has been sparked by the increasing evidence suggesting losses of species due to habitat destruction, overexploitation of natural resources, pollution, the spread of invasive species and climate change (Ceballos et al. 2015). Despite concerted attempts to curb biodiversity loss, global biodiversity is thought to be declining at unprecedented rates (Pereira et al. 2012; Ceballos et al. 2015). Due to limited time, financial resources and human capacity, it is essential that conservation efforts are prioritized effectively to maximize conservation returns. However, efficient conservation planning is hampered by key knowledge gaps (Scott et al. 1993; Jennings 2000; Meyer et al. 2015): the Linnaean shortfall describes the mismatch between the number of described species and the true number of species, the Wallacean shortfall reflects gaps in our knowledge of species distributions, and the Darwinian shortfall refers to our lack of knowledge on the evolutionary relationship among species (Hortal et al. 2015). We here refer to the lack of data to make informed conservation decisions as the *Leopoldean shortfall*, after Aldo Leopold, in recognition of his contributions to the modern conservation movement.

In many regions, we lack knowledge of species taxonomy (Paton et al. 2008), distributions (Meyer et al. 2016; Sporbert et al. 2019) and genetic diversity (Hoban et al. 2013) due to biases in species collection, lack of funding and research infrastructure, the declining number of taxonomists and complexity in identifying and describing species, and the inadequate training of conservation practitioners to collect and incorporate new types of bioinformatics data into biodiversity databases. As a result, efficient biodiversity conservation is frequently hampered by incomplete and unrepresentative data. Identifying and addressing such biodiversity knowledge gaps is a particular challenge for megadiverse, developing countries, where the economic commitment of governments to biodiversity conservation is frequently far less than what is required (He 2009; Yu 2010; Silveira et al. 2018). South Africa, the third most biologically diverse country globally, and with more than 50 percent of the population living below the poverty line (Statistics South Africa 2019), provides an object lesson.

South Africa's biological diversity is estimated at between 250,000 to 1,000,000 plant and animal species combined (Wynberg 2002). This astounding richness is thought to be due, in part, to the country's climatic diversity, ranging from desert to humid forest climates, and from cool temperate to warm subtropical climates, and varied topography (Cowling and Hilton-Taylor 1994). Over 13,000 of the plant species occurring within South Africa are found nowhere else (Raimondo et al. 2009). Much of this endemic diversity is restricted to the country's three global biodiversity hotspots: the Cape Floristic Region (CFR), the Succulent Karoo, and Maputaland-Pondoland-Albany (Myers et al. 2000). The CFR is the only plant kingdom confined within the borders of a single country, it is also a centre of diversity for endemic mammals, reptile and amphibian (Cowling et al. 2003). This hotspot contains approximately 8200 plant species, 6210 of which are endemic to the region (Raimondo et al. 2009). The Succulent Karoo includes a third of the world's succulent plant species (Brownlie and Wynberg 2001) and is also home to over 2400 endemic plants. The Maputaland-Pondoland-Albany biodiversity hotspot, with approximately 600 tree species, has the highest tree richness of any temperate forest on the planet

(Tarrant 2012), and is home to approximately 1900 endemic plants (Raimondo et al. 2009). To safeguard this rich biodiversity, taxonomic, genomic and spatial knowledge gaps need to be identified, and filled.

Various studies have shown that sampling bias and the decline in taxonomic studies present fundamental challenges to biodiversity conservation in South Africa (e.g. Robertson and Barker 2006; Thuiller et al. 2006; Von Staden et al. 2013; Tolley et al. 2019). The perilous state of plant taxonomic research in South Africa has been well-recognised for over a decade (Victor and Smith 2011). Only 62% of the South African flora has been revised since 1970, 13% of the flora has no revision, and taxonomic data for 25% of the flora is outdated (Von Staden et al. 2013)—a Linnaean shortfall. Locality data from herbarium specimens is curated within the Botanical Database of South Africa (BOD-ATSA; Ranwashe 2015)—the single largest database of species georeferenced occurrences and herbarium voucher data in South Africa, with more than a million specimens (Williams and Crouch 2017). Nonetheless, even basic data on the geographic distributions of many species is lacking. More than a 1000 species do not have any recorded occurrence data, while > 2600 species have less than five occurrence records—a Wallacean shortfall. Despite rapid advances in sequencing technology, and major sequencing initiatives, such as the DNA barcoding effort (Hebert et al. 2003) led by the International Barcode of Life (iBOL) consortium (<https://ibol.org/>), supported by the African Centre for DNA Barcoding in South Africa (Bezeng et al. 2017), most species still lack sequence data that would allow them to be placed on the plant tree-of-life—a Darwinian shortfall.

Here, we examine biodiversity data gaps for the endemic flora of South Africa. First, we use species distribution models (SDMs) to identify areas with broad climatic conditions suitable for (encompass the climate niche envelope of) many species. Importantly, we do not expect our projected occurrences to necessarily reflect species realised distributions. Indeed, we would expect species not to occur within much of their area of projected occurrence, even within highly suitable environments, as fine scale edaphic landscape features and other environmental requirements likely occur in only a small fraction of suitable climate space, and our grid cells additionally encompass urban areas, farmland, and other highly transformed habitat. Our models should thus be viewed as defining the coarse grained area of extent with suitable climate for species, and not their actual area of occupancy (for a conceptually similar approach see Engelbrecht et al. 2016). Second, we contrast our SDM projections with the geographic distribution of biodiversity sampling effort, and identify locations and taxa that are poorly represented in existing biodiversity database. Third, we merge taxonomic and spatial data with information from GenBank/NCBI (<https://www.ncbi.nlm.nih.gov/genbank>) to explore unevenness in the representation of genetic data across space and species. We hope our study will be useful in helping guide future sampling efforts, and contribute to addressing the shortfall in critical biodiversity data necessary for informed conservation decision making.

Methods

Endemic flora and occurrence data

We compiled a database of South Africa's endemic flora representing 175 families, 1061 genera, 10,965 species, and 762,655 distribution records (413,491 unique species × location occurrence records) of bryophytes, pteridophytes, gymnosperms, and angiosperms, using the Checklist of South African plants (Germishuizen et al. 2006) to crosscheck

species endemism. Species taxonomy was standardized using The Plant List (www.plantlist.org) and the Angiosperm Phylogeny Group (APG IV; Chase et al. 2016) as taxonomic authorities for plant names and families respectively, and then matched to geographic occurrence records from the Botanical Database of South Africa (BODATSA). Records for subspecies and varieties were merged into single species.

Herbarium specimens in BODATSA (specimens for endemic species, $n = 292,393$) are georeferenced to quarter degrees squares (QDS), approximately $25 \text{ km} \times 25 \text{ km}$, representing the approximate spatial precision of historical records. Recent and more precise plant occurrence records were sourced from the Protea Atlas Protect ($n = 245,407$), the Custodians of Endangered Wildflowers (CREW) programme, ($n = 15,244$) and the National Vegetation Map project (VEGMAP; $n = 135,165$), ACOCKS database ($n = 66,068$) and smaller projects ($n = 138,957$). Data were obtained from the South African National Biodiversity Institute (SANBI), with the assistance of LW Powrie. Several species not included in these databases ($n = 1054$) were manually added by georeferencing location data from protologues, Floras, and revisions, and extracting latitude–longitude coordinates from Google Earth (<https://www.earth.google.com>).

Genetic data

To assess taxonomic bias in DNA sequence data, we queried GenBank (<https://www.ncbi.nlm.nih.gov>)—a comprehensive public database containing nucleotides for approximately 260,000 formally described species (Benson et al. 2012)—for records for each of South Africa’s endemic species using the package SeqinR (Charif and Lobry 2007), in R version 3.5.2 (R Development Core Team 2006), returning the number of DNA sequence available for each species.

Species Distribution Modelling

Nineteen raster-based bioclimatic variables were sourced from the WorldClim database (<https://worldclim.org>; Hijmans et al. 2005) at a spatial resolution of 10 arc minutes, approximating the resolution of the species occurrence data. These bioclimatic variables were then used as environmental predictors to generate species distribution models (SDMs) using an ensemble forecast (Hijmans and Elith 2013) of three models: generalized linear models (GLMs; Guisan et al. 2002), random forests (RFs; Breiman 2001), and the gradient boosting machines (GBM; Friedman et al. 2000) fitted in R version 3.5.2 (R Development Core Team 2006), using the *gbm*, *lmtree*, *boot*, *dismo*, and *randomForest* packages (Hothorn et al. 2019; Ridgeway 2006; Chang and Hanna 2005). These standard modelling approaches use presence-absence data for predicting habitat suitability for species. Since the dataset lacks true absences, pseudo-absences were generated from background data, bounded to the country borders of South Africa. Twenty-five percent of occurrence records were used for testing the model and 75% of occurrence records were used for training the model. Duplicate records were removed to prevent model over-fitting. Ensemble predictions were generated by combining the individual model outputs weighted by the average of their AUC (Area Under the Curve), which gives more weight to the algorithm that gives a better estimation (Ranjitkar et al. 2014; Breiner et al. 2015). Predictions with AUC values < 0.5 were given a weight of zero. Average predictions were transformed into binary presence-absence maps by applying the threshold that maximizes the sum of actual-positive rate and actual-negative rate (Manel et al. 2001).

Spatial predictions from SDMs are influenced by the number of occurrence points, with accuracy decreasing as they are informed by less data; we, therefore, fit SDMs only to species with five or more occurrences ($n = 8295$). For species with three or four occurrence points ($n = 691$), the species range was defined using a convex hull in ArcGIS 10.5 (Esri, CA, USA), following Goldsmith et al. (2016). For species with one or two occurrences ($n = 1961$), the QDS in which they occurred was considered as their range.

Areas of high climate suitability

To identify areas with broad climate suitability (within the climate envelope) for many species, rasters from each of the output SDMs were stacked in R, using the Mass library (Ripley et al. 2012). Raster values for each summary output were then extracted onto an equal-area $25 \text{ km} \times 25 \text{ km}$ lattice in ArcGIS 10.6 (Esri, CA, USA) to match the resolution of the underlying species occurrence records (QDS) in the BODATSA database. High values cells represent areas with mean climate matching to that within a high number of projected species' distributions, although we would not expect species to fill their climate niche as additional niche factors varying over finer spatial scales, species interactions, and historical contingency all likely influence species' realised distributions. Coastal grid cells with $< 50\%$ of land were excluded because the projected richness for such cells would likely overestimate their true richness.

To generate maps of observed species richness and sampling density, we conducted a spatial join in ArcMap v.10.6 (Esri, CA, USA) to summarise the number of species from occurrence records and total number of occurrence records within each cell of the polygon lattice, used above.

Spatial and taxonomic gaps

We identified potential geographical gaps in taxon sampling effort by quantifying the fraction of observed species from occurrence records relative to projected richness from SDMs—which we refer to as the sampling fraction. As a proxy for sampling intensity, we also generated a density map of documented species coordinates and calculated the fraction of occurrence records relative to projected richness from SDMs—which we refer to as the sampling density. To test for the “road effect”—the tendency for collections to be greater in sites that are easily accessible—we overlaid a map of national roads on the sampling density map (<https://mapcruzin.com/free-south-africa-arcgis-maps-shapefiles.htm>).

We further quantified the fraction of the richness of species with genetic data relative to projected richness from SDMs to identify geographical areas which represent potential targets for genetic sampling. We then classified those species with sequence data by their IUCN Red List threat status (<https://redlist.sanbi.org>; Raimondo et al. 2009) to explore whether threatened species were more likely to lack sequence information.

Last, we examined whether biodiversity data gaps along one axis aligned with data gaps among other axes across space and taxa. First, we evaluated spatial correlation strengths using Pearson's correlation coefficients, adjusting degrees of freedom to account for spatial non-independence among grids cells using Moran's I, as implemented in SAM V.40 (Rangel et al. 2010). Second, we evaluated congruence across taxa using linear regression models constructed in R. For the spatial analyses we evaluated correlations between: (1) sampling fraction—ratio of observed species richness to projected richness—and sampling density—ratio of documented plant records to projected richness; (2) sampling fraction and

ratio of species with genetic data; and (3) the residuals from (1) and (2) above. For the taxonomic analysis we evaluated correlations between: (1) the number of endemic species per family and the number of georeferenced occurrence records for endemic species per family; (2) the number of endemic species per family and the number of sequences for endemic species per family; and (3) the number of georeferenced occurrence records for endemic species per family and the number of sequences for endemic species per family.

Results

Species richness, sampling effort and the spatial distribution of biodiversity data gaps

Spatial patterns of observed and projected endemic richness were mapped onto 1790 grid cells (25 km × 25 km) by (1) stacking projections from SDMs, and (2) recording observed species richness from occurrence records (Fig. 1a, b). We show that 4% of the country has not been sampled for endemics—70 grid cells with no endemic species recorded—while our SDMs indicate that all cells enclose climate space that falls within the climatic niche space of at least 69 species (although we would not necessarily expect this to translate into realised richness). The overall spatial structure in relative richness is similar for SDMs and observed records (Fig. S1 Supplementary Information: $r^2 = 0.7$, $p < 0.05$), but total grid cell richness is much lower in the latter (the highest number of observed species per grid cell is 1838, while the equivalent projected richness from SDMs is 5303 species per grid cell).

The projected endemic richness from SDMs matches well to current understanding of species diversity across South African biomes (Fig. 1). The Maputo-Pondoland-Albany biodiversity hotspot, the Soutpansberg, and the Wolkberg centres of endemism coincide with areas of high projected endemic-richness. The Savanna Biome is the largest and one of the most species-rich biomes in South Africa; here we show projected endemic richness is greatest in the northeast and lower in the northwest (Fig. 1a). The Succulent Karoo Biome, is synonymous with the Succulent Karoo biodiversity hotspot, and has projected grid cell endemic richness peaking at 4908 species (mean projected endemic richness: 1896, range = 136–4908; Fig. 1a). The Albany-Thicket and the Indian Ocean Coastal Belt biomes encompass the grid cells with the highest projected endemic richness within the Maputo-Pondoland-Albany biodiversity hotspot (mean projected endemic richness: 952, range 296–4661; Fig. 1a). The biome with the highest overall mean projected richness is the Fynbos. This biome falls within the Cape Floristic Region biodiversity hotspot, where projected grid cell endemic richness peaks at 5,303 species (mean projected endemic richness: 2765 range 526–5303; Fig. 1a). The Grassland and Nama Karoo biomes have the lowest mean projected endemic richness, however, even within these biomes some cells have high projected richness, for example, those that coincide with the with the Maputo-Pondoland-Albany hotspots, the Drakensberg escarpment, and the Sekhukhuneland and Barberton regional centres of endemism.

Despite the overall strong correlation between observed (Fig. 1b) and projected endemic richness (Fig. 1a), there is spatial structure in the residuals of the relationship (Fig. 1c)—the sampling fraction. For example, much of the Nama-Karoo and Savanna biomes have low sampling fraction (see also a conceptually similar analysis by Robertson and Barker 2006), whereas the sampling fraction is much greater in the generally species-rich Fynbos and Succulent-Karoo, and species-poor Grassland biomes.

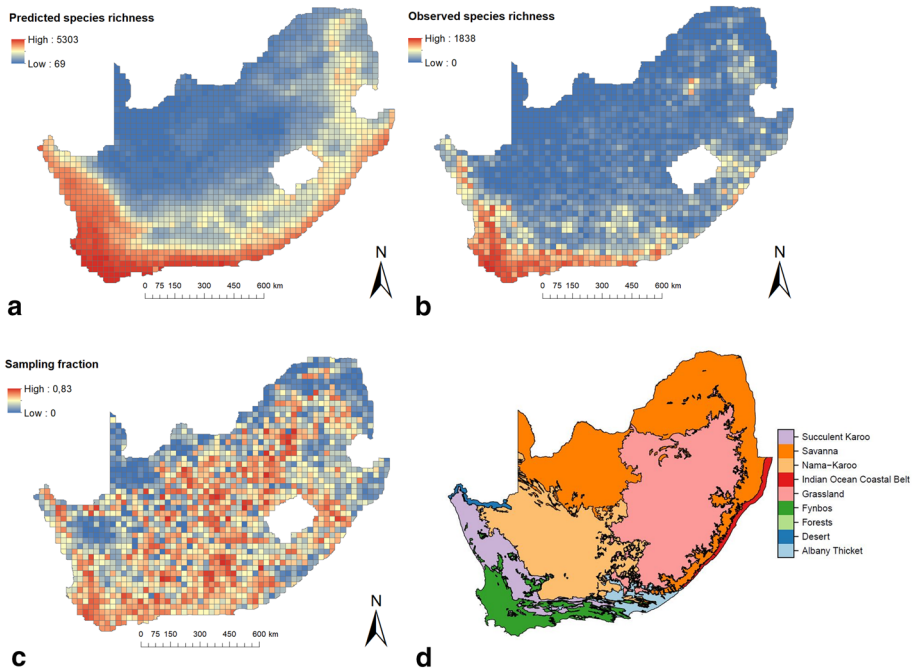


Fig. 1 Shortfalls in our knowledge of the distribution of endemic plants in South Africa. **a** Projected endemic species richness estimated from Species Distribution Models (SDMs). We do not necessarily expect true richness to match to predicted richness as it is likely that many fine scale process not captured in our SDMs limit species realised distributions; nonetheless, there is an obviously high correlation between observed and predicted richness (compare maps **a** and **b**), and we suggest model predictions are informative for identifying potential sampling gaps. **b** Observed endemic species richness from occurrence records (sum of the unique species records in each cell). Cells are shaded using a graduated colour scheme: red = high species richness, blue = low species richness. **c** Sampling fraction: ratio of observed endemic species richness to predicted endemic species richness – cells with low sampling fraction indicate areas with climate suitable for supporting high endemic richness, but for which there are relatively few occurrence records. Red = high species sampling, blue = low species sampling. Grid cell resolution **a–c** 25 km × 25 km. **d** Map of the biomes of South Africa, after Mucina and Rutherford (2006); data from <https://bgis.sanbi.org/SpatialDataset>

The Fynbos Biome and Gauteng Province are the most intensively sampled regions (Fig. 2a)—estimated from the total number of occurrence records. However, we find that there is a generally higher sampling density—ratio of documented plant occurrence records to predicted endemic species richness—in areas near roads (mean sampling density = 0.16 and 0.04 for grid cells with road and grid cells without roads, respectively; $t = 20.27$, $p < 0.01$; Fig. 2b). There is also a strong correlation between sampling density and sampling fraction—the ratio of observed endemic species richness from occurrence records to projected endemic richness from SDMs (Pearson’s $r = 0.862$, d.f. = 269, $p < 0.001$, after adjusting degrees of freedom to correct for spatial non-independence). For example, we again identify much of the Nama-Karoo and part of the Savanna Biome as under-sampled (low sampling density), whereas the Fynbos and much of the Grassland Biome have higher sampling density, with a greater number of documented occurrence records per species. As an important biodiversity hotspot, the relatively low sampling density in the Maputo-Pondoland-Albany region, which peaks at around only 0.2, is notable.

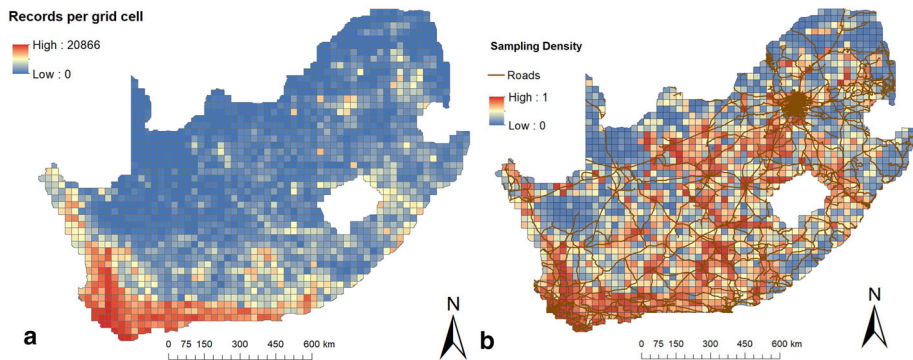


Fig. 2 Distribution of endemic plant collection records in South Africa. **a** Total number of georeferenced occurrence records for endemic plants per grid cell. Cells are shaded using a graduated colour scheme: red = high number of records, blue = low number of records. **b** Sampling density (ratio of documented plant records to projected endemic species richness [see Fig. 1a]) with the main road network overlaid. The Fynbos biodiversity hotspot has been relatively well sampled, while proportional sampling density in the species-rich Maputo-Pondoland-Albany hotspot peaks at around 0.2. Red = high sampling density richness, blue = low sampling density

In general, the sampling of genetic data for endemic species is poor relative to the projected richness; only 5% (80 of 1790) of grid cells have more than 50% of projected species with sequence data (Fig. 3a), with the interior of the country a notable ‘coldspot’ of genetic sampling (Fig. 3b). Several areas along the South African border have been relatively well sampled for genetic data, and these might represent lower genetic sampling priorities. Notably, there is no correlation between areas in need of better taxonomic sampling (low sampling fraction) and areas in need of genetic sampling (Pearson’s $r = -0.004$, d.f. = 111, $p = 0.963$, adjusted degrees of freedom).

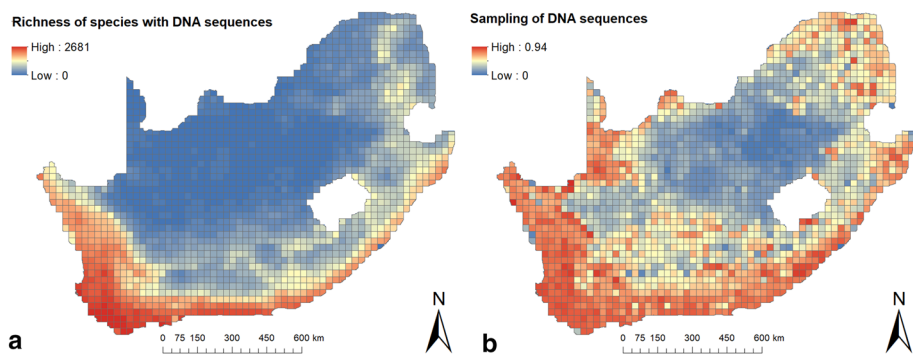
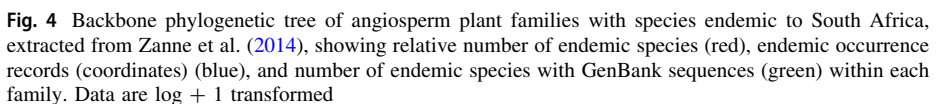


Fig. 3 Shortfalls in our knowledge of the sampling of genetic data for endemic plants: **a** projected richness of endemic species with genetic data (species with at least one sequence in GenBank) and **b** Sampling of DNA sequences (proportion of species with at least one sequence in GenBank relative to total endemic species richness per cell). Cells are shaded using a graduated colour scheme: blue cells indicate poorer genetic sampling of taxa, while red cells indicate higher genetic sampling. There is high sampling effort needed in the interior and northern regions of the country, while species-poor, these regions have been largely overlooked by past genetic sampling efforts. The Fynbos and parts of the Karoo and Albany thicket appear to be better sampled

Our database includes plants from 175 families and 1061 genera, with large variation in the taxonomic and phylogenetic distribution of biodiversity data (Fig. 4). The ten families with the highest number of endemic species within South Africa are listed in Table S1 (Supplementary Information). These ten families comprise 61% of all endemic species in the database. Forty families are represented by just one endemic species. Families with the highest number of unique species \times location occurrence records include Proteaceae, Asteraceae, and Fabaceae (Table S2: Supplementary Information). The top ten families by sampling (Table S2) comprise 69% of all the occurrence records in the database. Three families are represented by a single record in our analyses, all are monotypic (Ditrichaceae, Potamogetonaceae and Thelypteridaceae). In general, more species rich families have been better sampled than less species rich families (Fig. S2 Supplementary Information: $r^2 = 0.78$; slope = 1.11; p-value < 0.05), as would be expected if all species had an equal probability of being sampled. However, there is some notable variation in sampling intensity across families. For example, Anemiaceae has only one endemic species but is represented by 508 records; perhaps of more conservation concern are the several families that are relatively under-sampled.



There is large taxonomic variance in the availability of DNA sequence data, indicating a bias in the species targeted for sequencing (Fig. 4). Only 36% of endemic species have DNA sequences available on GenBank, and less species-rich families tend to be sampled less, as might be expected, although the strength of the correlation is not particularly high (Fig. S3 Supplementary Information: $r^2 = 0.60$; slope = 1.11; p-value < 0.05). Families without a GenBank record for endemic species include Fissidentaceae (3 endemic species), Lythraceae (4 endemic species), and Pylaisiadelphaceae (4 endemic species), as well as some families with higher endemic richness (e.g. Ricciaceae [26 endemic species] and Celestraceae [33 endemic species]). Nonetheless, some moderately species poor families have also been intensively sampled for genetic data. For example, Zamiaceae, with an endemic richness of 29 species has, at the time of writing, 995 sequences on GenBank. For the set of species with DNA sequences, 77% are of low conservation concern, 20% are threatened and 3% are data deficient. In contrast, for the set of species lacking DNA sequences, 68% are of low conservation concern, 20% are threatened and 12% are data deficient. There is no significant relationship between families that lack sequence data and families that lack georeferenced occurrence records (Fig S4 Supplementary Information: $r^2 = 0.004$; slope = 0; p-value = 0.84, from the linear regression of the number of sequence per family against and number of coordinates per family), hence these two gaps also need to be targeted separately. A list of top families ranked by number of sequences is provided in Table S3.

Discussion

South Africa is characterized by an interior with wide-ranging plains and plateaux, bounded by remarkable mountain landscapes and undulating coastal plains. Endemic-rich areas are found in a virtually continuous arc around the Great Escarpment, mainly within the three global biodiversity hotspots—the Succulent Karoo, Cape Floristic Region (CFR) and Maputo-Pondoland-Albany—and the Barberton, Sekhukhune, Soutpansberg and Wolkberg, centres of endemism. The southwestern part of the Fynbos is the most species-rich area in the country. Past climatic conditions are thought to be one of the main factors contributing to the high species richness in this region. During the Pleistocene, an epoch of high glacial-interglacial climate variability, the southwestern Cape remained relatively stable (Sniderman et al. 2013), while the eastern part of the country experienced greater climatic fluctuations. This resulted in higher speciation and lower extinction rates in the southwest relative to the southeast, leading to a greater accumulation of species over time in the former (Cowling and Lombard 2002; Cowling et al. 2004). As a consequence of these evolutionary dynamics, a high proportion of the native flora is composed of range restricted endemics (Goldblatt 1997), many of which are vulnerable to extinction (<https://redlist.sanbi.org/>) yet remain under-researched and poorly represented in biodiversity databases. To adequately protect this rich diversity, we must address these biodiversity data gaps. Here, focussing on endemic plants, we have explored the distribution of biodiversity data across space and phylogeny to identify regions and taxa that have been under-sampled as a guide to help future data gathering efforts.

The Wallacean and Linnean shortfalls

The observed richness of species obtained from occurrence records differs importantly to the richness estimated from species distribution models. Species distribution models

trained on observed occurrence data and broad-scale climate variables will likely overestimate species realised distributions, which are shaped by various additional processes and more fine scale niche partitioning (e.g. see Dubois et al. 2013). Our estimates of projected richness should thus be viewed as defining the coarse grained area of extent with suitable climate for species, and not their actual area of occupancy (see Elith and Leathwick 2009, for related discussion). Nonetheless, variation in the ratio of observed and projected richness highlights potential geographical biases in sampling effort. For example, the number of recorded species from occurrence records in the Nama-Karoo and Savanna biomes appears to be lower than that expected from projected species distribution models relative to observations across the Fynbos and Grassland biomes. These discrepancies are informative as they allow us to identify potential sampling gaps—areas where increased sampling effort is needed to fully characterise species geographic distributions—and thus help address the Wallacean shortfall.

We suggest important areas for future sampling include much of the Nama-Karoo, and some of the Savanna Biome, as highlighted above, and also the Maputo-Pondoland-Albany biodiversity hotspots. One reason for apparent under-sampling in these regions may be that there are fewer roads and centres of research nearby (Reddy and Dávalos 2003). Our results show that areas near roads are better sampled, likely because they are more accessible (Daru et al. 2018; Meyer et al. 2016). For example, the province of Gauteng has been relatively well-sampled, perhaps reflecting its status as the economic hub of South Africa, with a high density of roads and research institutes.

In comparison with the Nama-Karoo and the Maputo-Pondoland-Albany hotspots, the CFR has been relatively well sampled, and it is one of the regions with the greatest density of species records in the country. The CFR is recognised as a distinct floristic kingdom within the Mediterranean biome—the most threatened biome in the world (Cox and Underwood 2011)—and has thus attracted national and international research attention. Several non-government conservation agencies, including the World Wildlife Fund (WWF), Wildlife Protection Society of South Africa (WESSA), Earth Life, and CAPE, have offices located in the region, and support research on and conservation of the Fynbos flora. In addition, government programs, such as the Millennium Seed Bank (MSB) and the Custodian of Rare and Endangered Wildflowers (CREW), make use of volunteers and citizen scientists to sample remnants of natural vegetation in the region. While the considerable research effort focussed on the CFR is, of course, very welcome, other species-rich regions require equal attention.

In the past decade, plant collection efforts have decreased substantially in the country, reflected by the 14,000 plant collection records between 2006 and 2010 in comparison to the 94,000 records between 1976 and 1980 (Williams and Crouch 2017). We show that there is positive spatial correlation between areas of low sampling fraction—ratio of observed species richness to predicted species richness—and areas of low sampling density—ratio of documented plant records to predicted species richness—indicating that we are missing records for much of the diversity in areas that have been poorly sampled taxonomically, and raising the possibility that we may also be missing undescribed species in these areas (there is no evidence that the rate of new species description is declining over time; Victor et al. 2015)—the Linnaean shortfall.

Bias in plant collection has not only been spatial, but also taxonomic. Large families have been better sampled than smaller ones. Societal interest also plays a role in the sampling of taxa: more charismatic species are more likely to attract funds and research attention (Wilson et al. 2007; Troudet et al. 2017). For example, Proteaceae—a large family of significant agricultural and horticultural value—is the most intensively sampled

family in our database, and has been the target of large-scale ecological research through the Protea Atlas Project (Rebello 1993). It is also possible that smaller families occur in regions that have been less well sampled, or are more likely to be comprised of narrow ranged endemics, thus making them less likely to be included in general biodiversity surveys (Eberhard et al. 2009; Hemp 2006). However, there is large variation in sampling intensity across families independent from species richness, and other idiosyncratic or historical explanations likely contribute to taxonomic differences in sampling representation.

The Darwinian shortfall

Species richness has been used as an index for classifying important areas of biodiversity for decades (Pimm et al. 2014; Veach et al. 2017). However, a narrow focus on species may fail to capture genetic and functional diversity. There have been numerous calls to incorporate phylogenetic diversity, as a surrogate for functional or feature diversity, more directly into conservation planning (e.g. Cadotte and Davies 2010; Rolland et al. 2011; Winter et al. 2013; Faith 2015). Phylogenies are important for understanding structural and functional aspects of biodiversity in an evolutionary context, and allow us to assess how the tree of life will be affected by global change (Rolland et al. 2011). However, the use of phylogenetic data in conservation decision making remains a challenge, particularly in developing countries, where genetic data is often scarce or incomplete, and DNA sequencing is costly (Rodrigues and Gaston 2002)—the Darwinian shortfall.

While there is a strong need to gather more genomic data, it must be done efficiently to avoid escalating costs. Optimization strategies for data collection include the targeting of regions for which there is a high probability that data-poor species occur, and the selection of localities where many target species can be found (Parra-Quijano et al. 2012). In this study, we find that only a third of South Africa's endemic species have DNA sequences available, and that IUCN data deficient species are disproportionately under-represented, which makes the incorporation of genetic data into systematic conservation planning in South Africa even more of a challenge. We identify locations with climates suited to supporting high diversity but for which only a small fraction of projected species have sequence data, and suggest these as priority areas for tissue sampling. Species distribution models have been previously used for guiding the collecting genetic data to good effect (Ramírez-Villegas et al. 2010; van Zonneveld et al. 2014; Khoury et al. 2015). Here we show that many locations in the interior of the country have not been well-sampled for genetic data, whereas the exterior of the country has been better sampled, partly reflecting the success of DNA barcoding initiatives across the three biodiversity hotspots (e.g. see Lahaye et al. 2008; Bezeng et al. 2017; Powell et al. 2018).

On average, species-rich families have been better sampled for genetic data than species-poor families. *Zamiaceae* (a relatively small family) is an exception, with a high number of sequences per species. This family has been the subject of intense research, and its deep evolutionary history has made it a model taxon for studies on plant evolution and biogeography (e.g. Gregory and Chemnick 2004; Calonje et al. 2019). In addition, several species within the family are valuable medicinal, ornamental and commercial plants, attracting increased research effort (e.g. Ndawonde et al. 2007; Ravele and Makhado 2010; Cousins et al. 2011).

Genetic data is not only important for ecological and evolution studies, but is increasingly a fundamental component of taxonomy. Currently 611 endemic species are

listed as data deficient by the IUCN as a consequence of taxonomic uncertainty. DNA sequencing and phylogenetic studies could assist in addressing this issue, and thus facilitate appropriate IUCN Red Listing, which might provide increased conservation protection. A further 291 species are data deficient due to lack of ecological information, and DNA sequence data could help here also. Genetic data can be used to predict the conservation status of a species, for example, via phylogenetic imputation of traits or extinction risk (Bland et al. 2015; González-del-Pliego et al. 2019). Targeted sequencing efforts could thus help address both the Linnean and Darwinian shortfalls. However, there is a no significant correlation between areas that need sampling for occurrence data (Wallacean shortfall) and areas that need sampling for genetic data.

The Leopoldean shortfall

In this study, we have identified important biodiversity knowledge gaps. Strong geographical and taxonomic sampling biases indicate that we have not fully captured the extraordinary diversity of South Africa's endemic Flora in biodiversity databases. We suggest that these conservation data gaps represent a *Leopoldean shortfall*—contributing to the insufficient protection of plant biodiversity within the country. We identify areas and taxa that are in need of increased research attention. However, we show that the Wallacean and the Darwinian shortfalls need to be targeted separately, as gaps in our knowledge of species' distributions do not overlap with gaps in our knowledge of species' genomes. One way to help address these shortfalls is for scientist to reach out to non-professional to assist in data collection, as exemplified by the Protea Atlas Project (Rebelo 1993). Most importantly, there is a renewed call for scientists across the globe to make use of emerging and new technologies such as artificial intelligence, image-recognition algorithms, remote sensing, metagenomics etc. to collect data, identify, locate, and track species (see Pimm et al. 2015). By making use of these innovative and non-invasive approaches, the research community will be able to better address the data shortfalls we highlight here, and contribute to protecting and conserving biodiversity.

Acknowledgements This work was supported by the National Research Foundation, South Africa. We thank LW Powrie for providing us with access to the BODATSA and associated plant distribution databases, and Ross Stewart for assistance with figures.

Funding This work was supported by the National Research Foundation, South Africa.

Data Availability All data are available from the sources cited in the Methods or from the authors upon request.

Compliance with ethical standards

Conflict of interest We have no conflicts of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. *Nucleic Acids Res* 41:36–42
- Bezeng BS, Davies TJ, Daru BH, Kabongo RM, Maurin O, Yessoufou K, van der Bank H, Van der Bank M (2017) Ten years of barcoding at the African Centre for DNA Barcoding. *Genome* 60:629–638
- Bland LM, Collen BEN, Orme CDL, Bielby JON (2015) Predicting the conservation status of data-deficient species. *Conserv Biol* 29:250–259
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Breiner FT, Guisan A, Bergamini A, Nobis MP (2015) Overcoming limitations of modelling rare species by using ensembles of small models. *Methods Ecol Evol* 6:1210–1218
- Brownlie, S, Wynberg, R (2001) Integration of biodiversity into National Environmental Assessment procedures <https://www.cdbint/impact/casse-studies/csimapact-ibneap-za-en.pdf>. Accessed 20 June 2019
- Cadotte MW, Davies TJ (2010) Rarest of the rare: advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. *Divers Distrib* 16:376–385
- Calonje M, Meerow AW, Griffith MP, Salas-Leiva D, Vovides AP, Coiro M, Francisco-Ortega J (2019) A time-calibrated species tree phylogeny of the New World cycad genus *Zamia* L. (Zamiaceae, Cycadales). *Int J Plant Sci* 180:286–314
- Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM (2015) Accelerated modern human—induced species losses: Entering the sixth mass extinction. *Sci Adv* 1:1400253
- Chang, JC, Hanna, SR (2005) Technical descriptions and user's guide for the BOOT statistical model evaluation software package, version 20.
- Charif D, Lobry JR (2007) SeqinR 10-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman E, Vendruscolo M (eds) *Structural approaches to sequence evolution*. Springer, Berlin, pp 207–232
- Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc* 181:1–20
- Costion CM, Simpson L, Pert PL, Carlsen MM, Kress WJ, Crayn D (2015) Will tropical mountaintop plant species survive climate change? Identifying key knowledge gaps using species distribution modelling in Australia. *Biol Conserv* 191:322–330
- Cousins SR, Williams VL, Witkowski ET (2011) Quantifying the trade in cycads (Encephalartos species) in the traditional medicine markets of Johannesburg and Durban, South Africa. *Econ Bot* 65:356–370
- Cowling RM, Lombard AT (2002) Heterogeneity, speciation/extinction history and climate: explaining regional plant diversity patterns in the Cape Floristic Region. *Divers Distrib* 8:163–179
- Cowling RM, Hilton-Taylor C (1994) Patterns of plant diversity and endemism in southern Africa: an overview. *Strelitzia* 1:31–52
- Cowling RM, Pressey RL, Rouget M, Lombard AT (2003) A conservation plan for a global biodiversity hotspot—the Cape Floristic Region, South Africa. *Biol Conserv* 112:191–216
- Cowling RM et al (2004) Climate stability in Mediterranean type ecosystems: implications for the evolution and conservation of biodiversity. In: Arianoutsou M (ed), *Proc10th MEDECOS international conference on ecology, conservation and management of mediterranean climate ecosystems* Millpress, pp 1–11
- Cox RL, Underwood EC (2011) The importance of conserving biodiversity outside of protected areas in Mediterranean ecosystems. *PLoS ONE* 6:0014508
- Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfield TJ, Seidler TG, Sweeney PW, Foster DR, Ellison AM, Davis CC (2018) Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol* 217:939–955
- Dubuis A, Giovanettina S, Pellissier L, Pottier J, Vittoz P, Guisan A (2013) Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *J Veg Sci* 24:593–606
- Eberhard SM, Halse SA, Williams MR, Scanlon MD, Cocking J, Barron HJ (2009) Exploring the relationship between sampling efficiency and short-range endemism for groundwater fauna in the Pilbara region, Western Australia. *Freshw Biol* 54:885–901
- Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* 40:677–697
- Engelbrecht I, Robertson M, Stoltz M, Joubert JW (2016) Reconsidering environmental diversity (ED) as a biodiversity surrogacy strategy. *Biol Conserv* 197:171–179
- Faith DP (2015) Phylogenetic diversity, functional trait diversity and extinction: avoiding tipping points and worst-case losses. *Philos Trans R Soc Lond B* 370:1–10

- Friedman J, Hastie T, Tibshirani R (2000) Special invited paper additive logistic regression: a statistical view of boosting. *Ann Stat* 28:337–374
- Germishuizen G, Meyer NL, Steenkamp Y, Keith M (2006) A checklist of South African Plants Southern African. Botanical Diversity Network Report, No 41. SABONET, Pretoria
- Goldblatt P (1997) Floristic diversity in the Cape flora of South Africa. *Biodiv Cons* 6:359–377
- Goldsmith GR, Morueta-Holme N, Sandel B, Fitz ED, Boyle B, Casler N, Engemann K, Jørgensen PM, Kraft NJ, McGill B (2016) Plant-O-Matic: a dynamic and mobile guide to all plants of the Americas. *Methods Ecol Evol* 7:960–965
- González-del-Pliego P, Freckleton RP, Edwards DP, Koo MS, Scheffers BR, Pyron RA, Jetz W (2019) Phylogenetic and trait-based prediction of extinction risk for data-deficient amphibians. *Curr Biol* 29:1557–1563
- Gregory TJ, Chemnick J (2004) Hypotheses on the relationship between biogeography and speciation in Dioon (Zamiaceae) Cycad classification: concepts and recommendations Wallingford. CABI Publishing, Oxon, pp 137–148
- Guisan A, Edwards TC Jr, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol Model* 157:89–100
- He F (2009) Price of prosperity: economic development and biological conservation in China. *J Appl Ecol* 46:511–515
- Hebert PD, Cywinka A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. *Proc R Soc B* 270:313–321
- Hemp A (2006) Vegetation of Kilimanjaro: hidden endemics and missing bamboo. *Afr J Ecol* 44:305–328
- Hijmans RJ, Elith J (2013) Species distribution modelling with R R CRAN Project
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high-resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978
- Hoban SM, Haufler HC, Pérez-Espona S, Arntzen JW, Bertorelle G, Bryja J, Frith K, Gaggiotti OE, Galbusera P, Godoy JA, Hoelzel AR (2013) Bringing genetic diversity to the forefront of conservation policy and management. *Conserv Gen Res* 5:593–598
- Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu Rev Ecol Evol Syst* 46:523–549
- Hothorn T, Zeileis A, Farebrother RW, Cummins C, Millo G, Mitchell D, Zeileis MA (2019) Package ‘lmtest’: diagnostic checking in regression relationships R version 3.5.2.
- Jarnevich CS, Stohlgren TJ, Barnett D, Kartesz J (2006) Filling in the gaps: modelling native species richness and invasions using spatially incomplete data. *Divers Distrib* 12:511–520
- Jennings MD (2000) Gap analysis: concepts, methods, and recent results. *Landsc Ecol* 15:5–20
- Khoury CK, Castañeda-Alvarez NP, Achicanoy HA, Sosa CC, Bernau V, Kassa MT, Norton SL, van der Maesen LJG, Upadhyaya HD, Ramírez-Villegas J, Jarvis A (2015) Crop wild relatives of pigeonpea [*Cajanus cajan* (L) Millsp]: distributions, ex situ conservation status, and potential genetic resources for abiotic stress tolerance. *Biol Conserv* 184:259–270
- Lahaye R, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the Floras of biodiversity hotspots. *Proc Natl Acad Sci USA* 105:2923–2928
- Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *J App Ecol* 38:921–931
- Meyer C, Kreft H, Guralnick R, Jetz W (2015) Global priorities for an effective information basis of biodiversity distributions. *Nat Commun* 6:1–8
- Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol Lett* 19:992–1006
- Mucina L, Rutherford MC (2006) The vegetation of South Africa, Lesotho and Swaziland. *Strelitzia* 19. South African National Biodiversity Institute, Pretoria
- Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature* 403:853–858
- Ndawonde BG, Zobolo AM, Dlamini ET, Siebert SJ (2007) A survey of plants sold by traders at Zululand muthi markets, with a view to selecting popular plant species for propagation in communal gardens. *AFR J Range For Sci* 24(2):103–107
- Oliveira U, Soares-Filho BS, Santos AJ, Paglia AP, Brescovit AD, de Carvalho CJ, Silva DP, Rezende DT, Leite FSF, Batista JAN, Barbosa JPPP (2019) Modelling highly biodiverse areas in Brazil. *Sci Rep*. <https://doi.org/10.1038/s41598-019-42881-9>
- Parra-Quijano M, Iriondo JM, Torres E (2012) Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodivers Conserv* 21:79–96

- Paton AJ, Brummitt N, Govaerts R, Harman K, Hinchcliffe S, Allkin B, Lughadha EN (2008) Towards target 1 of the global strategy for plant conservation: a working list of all known plant species—progress and prospects. *Taxon* 57:602–611
- Pereira HM, Navarro LM, Martins IS (2012) Global biodiversity change: the bad, the good, and the unknown. *Annu Rev Environ Resour* 37:25–50
- Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344:987–998
- Pimm SL, Alibhai S, Bergl R, Dehgan A, Giri C, Jewell Z, Joppa L, Kays R, Loarie S (2015) Emerging technologies to conserve biodiversity. *Trends Ecol Evol* 30:685–696
- Powell RF, Magee AR, Boatwright JS (2018) Decoding ice plants: challenges associated with barcoding and phylogenetics in the diverse succulent family Aizoaceae. *Genome* 61:815–821
- R Development Core Team (2006) R: A Language and Environment for Statistical Computing Vienna: R Foundation for Statistical Computing.
- Raimondo D, Staden LV, Foden W, Victor JE, Helme NA, Turner RC, Kamundi DA, Manyama PA (2009) Red list of South African plants. South African National Biodiversity Institute, Pretoria
- Ramírez-Villegas J, Khoury C, Jarvis A, Debouck DG, Guarino L (2010) A gap analysis methodology for collecting crop gene pools: a case study with Phaseolus beans. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0013497>
- Rangel TF, Diniz-Filho JAF, Bini LM (2010) SAM: a comprehensive application for spatial analysis in macroecology. *Ecography* 33:46–50
- Ranjitkar S, Xu J, Shrestha KK, Kindt R (2014) Ensemble forecast of climate suitability for the Trans-Himalayan Nyctaginaceae species. *Ecol Model* 282:18–24
- Ranwashe F (2015) BODATSA: Botanical Collections v11 South African National Biodiversity Institute. https://ipt.sanbi.org.za/iptsanbi/resource?r=brahms_online&v=1.1. Accessed 12 Jan 2017.
- Ravele AM, Makhado RA (2010) Exploitation of *Encephalartos transvenosus* outside and inside Mphaphuli cycads nature reserve, Limpopo Province, South Africa. *Afr J Ecol* 48:105–110
- Reboul T (1993) Protea Atlas Project-A spectacular year of atlasing. *Veld & Flora* 79:26–27
- Reddy S, Dávalos LM (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *J Biogeogr* 30:1719–1727
- Ridgeway G (2006) Generalized Boosted Regression Models Documentation on the R Package ‘gbm’, version 1.5–7, <https://www.i-pensiericom/gregr/gbmshtml>. Accessed 18 May 2019.
- Ripley, B, Hornik, K, Gebhardt, A, Firth, D (2012) Package ‘MASS’: support functions and datasets for venables and Ripley’s MASS R package version 7.3–17
- Robertson MP, Barker NP (2006) A technique for evaluating species richness maps generated from collections data. *South Af J Sci* 102:77–84
- Rodrigues AS, Gaston KJ (2002) Maximising phylogenetic diversity in the selection of networks of conservation areas. *Biol Conserv* 105:103–111
- Rolland J, Cadotte MW, Davies J, Devictor V, Lavergne S, Mouquet N, Pavoine S, Rodrigues A, Thuiller W, Turcati L, Winter M (2011) Using phylogenies in conservation: new perspectives. *Biol Lett*. <https://doi.org/10.1098/rsbl.2011.1024>
- Scarascia-Mugnozza G, Oswald H, Piussi P, Radoglou K (2000) Forests of the Mediterranean region: gaps in knowledge and research needs. *For Ecol Manage* 132:97–109
- Scott JM, Davis F, Csuti B, Noss R, Butterfield B, Groves C, Anderson H, Caicco S, D’Erchia F, Edwards TC Jr, Ulliman J (1993) Gap analysis: a geographic approach to protection of biological diversity. *Wildl Monogr* 123:3–41
- Silveira FA, Teixeira AL, Zanetti M, Pádua JG, Andrade ACSD, Costa MLND (2018) Ex situ conservation of threatened plants in Brazil: a strategic plan to achieve Target 8 of the Global Strategy for Plant Conservation. *Rodriguésia* 69:1547–1555
- Sniderman JMK, Jordan GJ, Cowling RM (2013) Fossil evidence for a hyper sclerophyll Flora under a non-Mediterranean-type climate. *PNAS* 110:3423–3428
- Sporbert M, Bruehlheide H, Seidler G, Keil P, Jandt U, Austrheim G, Biurrun I, Campos JA, Čarni A, Chytrý M, Csiky J (2019) Assessing sampling coverage of species distribution in biodiversity databases. *J Veg Sci* 30:620–632
- Statistics South Africa (2019) Five facts about poverty in South Africa. <https://www.statssagovza/?p=12075>. Accessed 13 June 2019
- Tarrant, J (2012) Conservation assessment of threatened frogs in KwaZulu-Natal and a national assessment of chytrid infection in threatened South African species. Doctoral dissertation. University of North West.

- Thuiller W, Midgley GF, Rouget M, Cowling RM (2006) Predicting patterns of plant species richness in megadiverse South Africa. *Ecography* 28:733–744
- Tolley KA, Weeber J, Maritz B, Verburgt L, Bates MF, Conradie W, Hofmeyr MD, Turner AA, da Silva JM, Alexander GJ (2019) No safe haven: protection levels show imperilled South African reptiles not sufficiently safe-guarded despite low average extinction risk. *Biol Conserv* 233:61–72
- Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F (2017) Taxonomic bias in biodiversity data and societal preferences. *Sci Rep*. <https://doi.org/10.1038/s41598-017-09084-6>
- van Zonneveld M, Dawson I, Thomas E, Scheldeman X, van Etten J, Loo J, Hormaza JI (2014) Application of molecular markers in spatial analysis to optimize in situ conservation of plant genetic resources In *Genomics of plant genetic resources*. Springer, Dordrecht, pp 67–91
- Veatch V, Di Minin E, Pouzols FM, Moilanen A (2017) Species richness as criterion for global conservation area placement leads to large losses in coverage of biodiversity. *Divers Distrib* 23:715–726
- Victor JE, Smith GF (2011) The conservation imperative and setting plant taxonomic research priorities in South Africa. *Biodiv Conserv* 20:1501
- Victor J, Smith G, Van Wyk A, Ribeiro S (2015) Plant taxonomic capacity in South Africa. *Phytotaxa* 238:149–162
- Von Staden L, Raimondo D, Dayaram A (2013) Taxonomic research priorities for the conservation of the South African Flora. *S Afr J Sci* 109:1–10
- Williams VL, Crouch NR (2017) Locating sufficient plant distribution data for accurate estimation of geographic range: the relative value of herbaria and other sources. *S Afr J Bot* 109:116–127
- Wilson JR, Procheş Ş, Braschler B, Dixon ES, Richardson DM (2007) The (bio) diversity of science reflects the interests of society. *Front Ecol Environ* 5:409–414
- Winter M, Devictor V, Schweiger O (2013) Phylogenetic diversity and nature conservation: where are we? *Trends Ecol Evol* 28:199–204
- Wynberg R (2002) A decade of biodiversity conservation and use in South Africa: tracking progress from the Rio Earth Summit to the Johannesburg World Summit on Sustainable Development. *S Afr J Sci* 98:233–243
- Yu X (2010) Biodiversity conservation in China: barriers and future actions. *Int J Environ Sci* 67:117–126
- Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL (2014) Three keys to the radiation of angiosperms into freezing environments. *Nature* 506:89–92