A METANALYSIS OF SEQUENCES OF VASCULAR PLANTS IN THE WORLD'S BIODIVERSITY HOTSPOTS WITH A SPECIAL SECTION ON MADAGASCAR

by

John Michael Adrian Wojahn

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Biology

Boise State University

August 2020

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

John Michael Adrian Wojahn

Thesis Title: A Metanalysis of Sequences of Vascular Plants in the World's Biodiversity Hotspots with a Special Section on Madagascar

Date of Final Oral Examination: 03 June 2020

The following individuals read and discussed the thesis submitted by student John Michael Adrian Wojahn, and they evaluated their presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Sven Buerki, Ph.D.	Chair, Supervisory Committee
James Smith, Ph.D.	Member, Supervisory Committee
Stephen Novak, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Sven Buerki, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

DEDICATION

I dedicate this thesis to my mother, Patricia Mary Wojahn (née Houghton) who, despite having advanced cancer, has supported me through my graduate journey.

ACKNOWLEDGMENTS

I would like to acknowledge the superb mentorship of Sven Buerki (Boise State University). I would also like to acknowledge the help and insights provided by my network of collaborators: Félix Forest (Royal Botanic Gardens Kew), Porter P. Lowry II (Missouri Botanical Garden and Université des Antilles), Peter B. Phillipson (Missouri Botanical Garden and Muséum National d'Histoire Naturelle de Paris), Martin W. Callmander (Conservatoire et Jardin botaniques de la Ville de Genève), Steve Goodman (University of Chicago), and Laurent Gautier (Conservatoire et Jardin botaniques de la Ville de Genève) I also want to acknowledge the help and support of my thesis committee, James Smith (Boise State University) and Stephen Novak (Boise State University). I additionally want to acknowledge support of my family, friends, and fellow graduate students. Finally, I would like to acknowledge the gracious financial support provided to be by the National Geographic Society, Lush Fresh Handmade Cosmetics, and Boise State University.

ABSTRACT

Humans have become a major factor in reshaping the Earth's biosphere. One of the major effects of human changes to the environment is an increase in the rate of species extinction as compared to background rates. Biodiversity hotspots are areas whose species assemblages are very rich (50% of the world's plants and 42% of land vertebrates) yet very threatened with extinction (>70% habitat destruction), and which ought to be foci for conservation efforts. The intense peril in which the flora of these endangered regions are requires an equally intense response from the scientific community. This study investigated the benefits of adding genomic information to voucher specimens to alleviate the Linnaean (lack of species description), Wallacean (lack of data on species distribution) and Darwinian (lack of data on species evolution) shortfalls.

An open-source R bioinformatic pipeline was developed to determine the percentage of vascular plant species present in biodiversity hotspots with at least one reproducible DNA sequence deposited on GenBank. Reproducible DNA sequences were defined as being underpinned by traceable material and methods and accurate taxonomic identifications. A vascular plant species checklist for the 36 biodiversity hotspots was inferred using 32,914,892 GBIF occurrences, comprising 204,044 species. A total of 736,532 GenBank accessions (representing DNA barcodes) were downloaded for those species. Associated abstracts and metadata were mined from 3,127 publications deposited on PubMed to assess DNA sequences reproducibility. The reproducibility of each study was tested by a sentiments (natural language processing) analysis.

Overall, the analyses indicated that the reproducibility crisis also extended to the realm of biodiversity. There was a significant shortfall in genetic information available for biodiversity hotspots, where 80.3% of the sequences produced (591,431) were not reproducible. This meant that only 19.7% of sequences—representing only 37,637 species (18% of the total)— were reproducible. This phenomenon was named the Wu-Meyersian shortfall to recognize that we are critically lacking DNA sequence data for threatened biodiversity. This shortfall was named in honor of Ray Wu (the father of DNA sequencing; 1928-2008) and Norman Meyers (a pioneer in establishing biodiversity hotspots; 1934-2019). Working on this shortfall could contribute to alleviating the Linnean, Wallacean and Darwinian shortfalls and support conservation. Information was particularly lacking in tropical biodiversity hotspots, but no biodiversity hotspot other than Japan had > 50% of its flora reproducibly sequenced. Older biodiversity hotspots were less known than those established more recently. This is concerning since those are among the most diverse and threatened (e.g. Madagascar, Sundaland). From a DNA region perspective, ITS (23,422 species), *matK* (17,164 species), and *rbcL* (16,509 species) were the most commonly used barcodes. From a lineage perspective, gymnosperms (N=895) are exceptionally well-sequenced, with three quarters of their species having been reproducibly sequenced. Angiosperms are comparatively poorly sequenced (18%), but this may be explained by their extreme diversity (N=195,433). Finally, ferns and their allies (N=7,716) are poorly sequenced (22%). This is especially troubling because extinction of these species would represent the loss of hundreds of millions of years of unique evolutionary history. This study finally proposed best practices to ensure maximizing reproducibility of DNA sequences produced by the scientific community.

The bioinformatic pipeline can be applied to systems at multiple geographical scales and any taxonomic groups and is therefore appealing to a wide range of stakeholders. We recommended using it periodically to monitor progress towards alleviating the Wu-Meyersian shortfall.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BSU	Boise State University
GC	Graduate College
TDC	Thesis and Dissertation Coordinator
GBIF	Global Biodiversity Information Facility
DNA	Deoxyribonucleic Acid
ITS	Internal Transcribed Spacer
APG	Angiosperm Phylogeny Group
PPG	Pteridophyte Phylogeny Group
CBOL	Consortium for the Barcode of Life
NHM	Natural History Museum, London
MNdHNP	Muséum national d'Histoire naturelle Paris
RBGK	Royal Botanic Gardens Kew
JMAW	John Michael Adrian Wojahn
SB	Sven Buerki

CHAPTER ONE: A METANALYSIS OF REPRODUCIBLE SEQUENCES OF VASCULAR PLANTS IN THE WORLD'S BIODIVERSITY HOTSPOTS

Introduction

Humans have become a major factor in reshaping the Earth's biosphere (Waters et al., 2016). One of the major effects of human changes to the environment is an increase in the rate of species extinction as compared to background rates (Otto, 2018). Biodiversity hotspots are areas of the earth's biosphere whose species assemblages are very rich (a minimum of 0.5% of all vascular plants as endemics in each hotspot; 50% of the world's plants and 42% of land vertebrates in all hotspots total) yet very threatened with extinction (over 70% habitat destruction), and which ought to be foci for conservation efforts (Myers, 1988; Myers et al., 2000;). The intense perils in which the flora of these endangered regions are under requires an equally intense response from the scientific community. In this study, we investigate the benefits of adding genomic information to voucher specimens to support large-scale scientific and conservation endeavors.

Genomic information about the world's flora provides insights that traditional taxonomical and botanical survey methods cannot provide. For instance, genomic and DNA barcoding data support a number of possible analyses, such as i) promoting rapid new species discovery (i.e. Buerki et al., 2017), ii) assessing processes underpinning plant community assembly (i.e. Buerki et al., 2013), iii) monitoring the illegal trade of endangered organisms (i.e. Williamson et al., 2016), iv) promoting breeding programs of threatened species (i.e. Devey et al., 2013), v) supporting prioritizing species conservation (i.e. Forest et al., 2018), and vi) predicting plant chemistry/opening new fair trade ventures using local species (i.e. Grace et al., 2016). Impact of this latter research is only relevant if genomic data accompanying species identifications are based on voucher specimens (deposited in herbaria) and if those were identified by taxonomists (vouchering also allows possible re-identifications based on new evidence). Thus, genomic data only make sense if they were *reproducibly sequenced*. Assessing reproducibility of DNA sequences is even more important since the scientific community acknowledged the existence of a "reproducibility crisis" in science. Indeed, a survey published in *Nature* (Baker, 2016) revealed that more than 70% of researchers admitted trying and failing to reproduce other scientist experiments and more than half also admitted failing reproducing their own experiments. The research fields of evolution and ecology were sadly no exception to this rule and therefore are calling for the need to ensure that only reproducible DNA sequences are used in meta-analyses. For instance, annotations and linkages of DNA sequences in major data repositories are not consistent and several studies have questioned the quality or availability of data on GenBank (Bidartondo, 2008; Lindberg, 2000; Bilofsky et al., 1986). This study aims at alleviating this challenge by developing an open-source and free bioinformatic pipeline to rapidly assess the reproducibility of DNA sequences deposited on GenBank and their taxonomical identifications. We achieve this goal for 204,044 species by mining 736,532 sequences and retrieving unique PubMed accessions associated with them. We then downloaded abstracts for those 3,127 publications and performed a sentiments analysis (a natural language processing method developed for business and sociological studies). The algorithm assessed the polarity of each word in each abstract so as to determine the reproducibility of the abstract as a whole—positive for words indicating likely reproducible methodologies, neutral for words that could be associated with either both or neither reproducible or non-reproducible methodologies, and negative for words associated with likely non-reproducible/metagenomic methodologies. It then discarded studies it determined to be insufficiently reproducible. This automated approach was applied to a checklist of vascular plants occurring in the world biodiversity hotspots (CEPS, 2016). The species checklist was assembled by extracting GBIF (Global Biodiversity Information Facility, 2001) occurrence data occurring in the world's biodiversity hotspots and curating it based on taxonomy from the Plant List (The Plant List, 2013).

To estimate the fraction of vascular species sequenced across biodiversity hotspots, two specific questions were investigated: 1) how reproducible are vascular plant DNA sequences available on GenBank? 2) what are the most commonly utilized DNA regions? and 3) is there a correlation between the date of establishment of a biodiversity hotspot and its number of species sequenced? Indeed, we could predict that biodiversity hotspots that have been established early would have more plants sequenced than those recently established. A gap analysis was then conducted to identify regions and taxa that should be prioritized for large-scale DNA sequencing initiatives as well as potential DNA barcodes used to support this endeavor. Finally, guidelines to ensure best practices for DNA sequence production in biodiversity regions are presented here.

Materials and Methods

The bioinformatic approach applied in this study to establish a list of vascular plant species in biodiversity hotspots for which at least one reproducible and validly identified

DNA sequence is available is summarized below (see Appendix A for full details). The bioinformatic pipeline itself was implemented into the R package *ReproduciblePlants*, which is deposited on GitHub (https://github.com/wojahn/ReproduciblePlants). The approach consists of three main steps: i) inferring a taxonomically curated species checklist for the study area (here vascular plants occurring in world biodiversity hotspots), ii) querying GenBank to determine whether target DNA sequences are available for species in the checklist, iii) assessing reproducibility of produced DNA sequences and their species identifications by performing a sentiments analysis on abstracts from publications associated to DNA sequences (by querying the PubMed database; see below for more details) and by inspecting journal policies by manually searching their instructions for authors to (associated to data transparency and reproducibility) or in the case of DNA sequences without associated published studies available in PubMed by accounting for GenBank submission dates and authorships (see below for more details). Finally, the approach is integrating results from the above analyses to produce species lists of varying reproducibility and accounts for taxonomic lineages (major lineages within vascular plants) and geographic regions (each of the 36 biodiversity hotspots).

Inferring a Taxonomically Curated Species Checklist for the Study Area

A taxonomically curated species checklist for vascular plants occurring in biodiversity hotspots was inferred by downloading GBIF occurrence data underpinned by specimens and using the *ReproduciblePlants* package to overlap it with a shapefile of biodiversity hotspots (CEPS, 2016). The algorithm then adapted the taxonomy of the preliminary species checklist following accepted names published in the Plant List (The Plant List, 2013).

Querying GenBank to Determine Whether DNA Sequences Are Available for Each Species in the Checklist

The algorithm used the taxonomically curated species checklist to query GenBank and retrieve DNA accessions associated with each species for each of the 14 CBOL plant barcodes (Hollingsworth et al., 2011) and for nuclear and plastid genomes. The output is a species list with associated GenBank accessions corresponding to DNA sequences.

Assessing Reproducibility of DNA Sequences and their Species Identifications

GenBank accessions were used by the algorithm to mine the PubMed database (which is a literature repository linking DNA sequences deposited on GenBank to scientific publications) to download abstracts for the studies underpinning each accession and retrieved journal names and the list of authors. The reproducibility of the study was assessed by inferring the type of material that was used to generate DNA sequences. Here, we are specifically estimating whether the material at origin of DNA sequence is a physical plant voucher deposited in herbaria or part of living collections (highly reproducible) or if sequences originated from an environmental sample such as a feces (not reproducible, meaning that there are no opportunity to validate species identification and those were most likely obtained by applying a BLAST approach). To further validate species identifications, we succeeded at assessing if studies were conducted to advance taxonomic/systematics knowledge of plant biodiversity. In this context, we have assumed that species identifications were validated by taxonomic experts and were most likely underpinned by vouchers (this was later confirmed by looking at journal policies). To achieve this goal, a custom dictionary comprising three lists of key words was built reflecting confidence in study reproducibility and species identifications: i) systematics (incl. evolution, biogeography): high reproducibility and high confidence in species identifications (referred to as being of positive polarity); ii) applied sciences (e.g. agriculture, medicine, biotechnology, and pharmaceuticals): neutral reproducibility and unknown species identifications (referred to as being of neutral polarity); and iii) environmental research (eDNA, metagenomics): low reproducibility and low confidence in species identifications (referred to as being of negative polarity). A custom sentiments analysis was then performed by the algorithm on abstracts and associated custom dictionary (reflecting positive, neutral and negative polarities) by using the *sentimentr* package (Rinker, 2019). Sentiments analysis uses natural language processing and the list of keywords and their associated polarities to assess the polarity of whole sentences/documents and extract polarizing words for further analysis (Taboada and Brooke, 2011). Abstracts matching only neutral keywords, or which matched none of the keywords, were manually curated to assign polarity. For each of the abstracts mined from PubMed the overall polarity scores, matching positive words (if any), and matching negative scores (if any) were written into an output matrix. A Venn diagram illustrating the overlap and proportionality of the sentiments results' polarities was created by the algorithm. Finally, a list of authors, which published studies deemed highly reproducible was produced and used to estimate reproducibility and confidence in species identification of DNA sequences not underpinned by PubMed accessions (see below).

Evaluating Upload Date and Authorship for Unpublished Sequences

A large proportion of DNA sequences available on GenBank were not underpinned by PubMed accessions. Although we do not know the exact reasons for such a trend, we assumed that it was associated with a time lag between release of DNA sequences on GenBank and acceptance of publications. Finally, there might also be a time lag between entry of the publication in PubMed and the linkage of PubMed accessions with their associated DNA sequences. For those reasons, we have decided to estimate whether a DNA sequence without a PubMed number was reproducible and that its associated species identification is likely accurate by using two criteria: time since submission to GenBank and authors submitting the DNA sequence. The algorithm analysed the date of submission for sequences from species not represented by any published sequences. It discarded those older than 5 years from the date of analysis as being unlikely to be published in the future. Next the algorithm checked who submitted the DNA sequences representing the remaining species, only keeping species represented by at least one sequence submitted by authors who had published a reproducible sequence before (i.e. corresponding to authors assigned to the positive polarity dataset; see above). The output was a list indicating which of the species represented solely by unpublished sequences passed the date and authorship test.

Verifying that the Algorithm Actually Worked: Sentiments Analysis Efficacy Evaluation

To examine whether the sentiments analysis code in the algorithm was doing its job correctly, 50 studies were randomly sampled from the finished sentiments list and manually checked. 96% (48) of the abstracts were correctly sorted, with the remaining 4% (2) being incorrectly rejected because of their having contained words associated with metagenomics (i.e. studying DNA from environmental samples, Pace et al., 1986) in the background portions of their abstracts. The authors believe this error rate is tolerable because the negative words that disqualified the 2 good studies were heavily associated with the metagenomic studies analyzed for the initial compiling of the 344 keywords, and also because the errors resulted in an underestimation rather than overestimation (and overestimation could provide a false sense of completeness).

The sentiments analysis was performed rather than a simple word search because some authors may not directly state that they used vouchers or other reproducible methods in their abstracts. In fact, only 0.22% of the studies mentioned vouchers (or any of its semantic equivalents) at all in their abstracts. The sentiments analysis is also much quicker than a simple word search, taking less than half the time of the latter to complete, classifying the abstracts into categories that are easily interpretable by the algorithm (Figure 1.1).

Integrating Results to Produce Species Lists of Varying Reproducibility and Species Identification

The algorithm used the raw GenBank query output to create the first of the three species lists; A, which contained all species having at least one DNA region deposited in GenBank. It then used the result of the sentiments analysis to create the second species lists: B, which contained species with at least one reproducible study (studies scored as positive by the sentiments analysis). The algorithm then used the output from the date and authorship analysis of unpublished sequences to create the third species list: C, representing species currently without a PubMed number. The species list C was constructed to allow for forecasting the total number of species in biodiversity hotspots which have been sequenced at least once.

Lineage-Wise and Geographical Analysis of Reproducible Sequences

This analysis focused on species in list B because it is our most accurate estimate of knowledge for vascular plants in the world's biodiversity hotspots. The algorithm broke the composition of list B down by class lineage, sorting them into Angiosperms, Gymnosperms, and Ferns and Allies. Heatmaps showing the percentage of species sequenced per biodiversity hotspot were inferred. The maps were also inferred for each lineage. The rate of species sequencing through time was investigated through inferring a cumulative curve illustrating the number of newly sequenced species per year.

Data and Code Availability Statement

All data and code used in this work are available on GitHub at wojahn/ReproduciblePlants.

Results and Discussion

The Reproducibility Crisis Also Applies to Plant DNA Sequencing

Our analyses confirm that the vast majority of DNA sequences deposited on GenBank are not reproducible, therefore confirming large-scale studies on this topic (see Baker, 2016). Our sentiments analysis showed that 43.77% (89,314 species) of vascular plant species in biodiversity hotspots could have at least one DNA sequence in GenBank, but after testing for reproducibility the algorithm found that less than half of those—only

18.47% (37,687 species, Table 1.1)—are deemed reproducible (Table 1.1; Fig. 1.1). In fact, only 0.22% of the studies mentioned vouchers (or any of its semantic equivalents) in their abstracts, indicating that authors are not emphasizing their implementation of reproducibility (see guidelines at the end of the discussion section for best practices guidelines).

On a brighter note, several sequencing initiatives have been launched in the last decades and fostered our genomic knowledge of vascular plants. For instance, the Angiosperm Phylogeny Group (APG) established in 1998 (APG, 1998), the Consortium for the Barcode of Life (CBOL) established in 2003 (Hebert et al., 2003), and the Pteridophyte Phylogeny Group (PPG) established in 2015 (PPG, 2016). These initiatives may have helped drive the pace of research by providing the scientific community with tools and structure: the number of species with at least one reproducible sequence increased after the introduction of the APG (though it was also increasing before its advent, possibly because of the emergence of systematics two years prior), and the trend continued after the introduction of the CBOL (Fig. 1.3). A couple years after the introduction of the PPG, the number of reproducibly sequenced fern and ally species had a burst. However, it is very important to note that as no causative/correlative analyses were performed and therefore the analysis cannot show that these events actually influenced the number of reproducible sequences produced.

Four journals published over 50% of the studies underpinning the list of reproducibly sequenced species—*Molecular Phylogenetics and Evolution*, *American Journal of Botany*, *PLoS ONE*, and *Annals of Botany* (Table 1.2). However, only three of the four journals have requirements (not just recommendations) in their authors guidelines requesting authors to submit data/materials to public repositories, with the most prolifically publishing journal not having this requirement (Table 1.2). It is very important for a journal to require their authors to adhere to open data policies because the utility of sequences produced for future studies depends on public archiving and traceability of methods and material.

What are the Top Utilized DNA Regions?

The nuclear ribosomal internal transcribed spacer (ITS) region, *matK*, and *rbcL* barcodes were the top three most commonly utilized barcodes for reproducibly sequenced species (Table 1.3). It is not surprising that ITS, *matK*, and *rbcL* are the most common barcodes in the analysis, as combinations of them have been proposed as a universal plant barcode akin to the C oxidase 1 (*CO1*) barcode commonly used in animals (Hollingsworth et al., 2011).

Even though barcodes are very useful, having the genome of a plant is more informative because it allows researchers to study the inner workings of the plant to elucidate how it interacts with its environment and with other species. Our results indicate a massive shortfall in the number of species that have had their plastome reproducibly sequenced (98.5 % of plant species have not, Table 1.3) and an even larger shortfall in the number of species that have had their nuclear genome reproducibly sequenced (99.995 % of plant species have not, Table 1.3).

Nonetheless, genome sequencing is more difficult and time-consuming than sequencing DNA barcodes, suggesting that barcode sequencing should be prioritized so that the maximum number of plants can have one or more DNA barcodes sequenced. Plant material from which barcodes have been reproducibly sequenced may be used in the future for genomic sequencing, meaning that prioritizing barcoding now does not mean letting go the possibility of genomic sequencing later.

The Age of Biodiversity Hotspots is Inversely Proportional to How Many of its Species Have Been Reproducibly Sequenced

Unexpectedly, it appears that the longer ago a biodiversity hotspot was established, the lower the percentage of its flora that has been reproducibly sequenced (Figure 1.4, note that the curves account for sequences added even before the establishment of a hotspot). A Spearman's correlation analysis was performed to test this observation and a significantly moderate negative trend was found (r = -0.485638, p-value = 0.002678). This is a very concerning result, because many of the first biodiversity hotspots are among the most diverse and threatened by deforestation (e.g. Madagascar, Sundaland, the Tropical Andes). Indeed, this result is in line with Buerki et al. (2013) showing that only 59.3% of Malagasy endemic genera of angiosperms (184 of the 310 endemic genera) had at least one species sequenced. The lack of genetic knowledge on taxa unique to highly threatened regions such as Madagascar are a testament of the work remaining to be conducted to complete sequencing of vascular plants (see Figures 1.2, 1.4).

Tropical Biodiversity Hotspots are Receiving Less Attention than their Temperate Counterparts

All hotspots have not received the same amount of attention. Of the top 10 best-represented biodiversity hotspots, nine occur fully or primarily outside of the tropics (i.e.

more than 50% of their shapefile areas are in the temperate regions), and of the ten most poorly represented biodiversity hotspots, eight were fully or primarily within the tropics (Figure 1.2). For primarily temperate hotspots, on average 35.48% of species have been reproducibly sequenced, whereas for primarily tropical hotspots 24.28% of species have been reproducibly sequenced. The least reproducibly sequenced temperate hotspot was Southwest Australia (18.96% reproducibly sequenced), and the least reproducibly sequenced tropical hotspot was the Tropical Andes (14.68% reproducibly sequenced).

Japan is unique among the biodiversity hotspots in that it is the only one to have reproducibly sequenced more than 50% of its flora. This could be due to the proximity of the flora to research institutions, facilitating easy fieldwork. This proximity factor may also explain why the next two most-sequenced hotspots—the North American Coastal Plain and the California Floristic Provence—are so well represented (both being in the United States). Inequality of resources and scientific infrastructure between the Global South and the Global north may also help explain this disparity—for example, Madagascar has no genetic or genomic labs present on the island itself.

Overall, it appears that more attention needs to be focused on the flora of the tropics, with special focus given to the top five least-represented regions: the Tropical Andes (14.65% reproducibly sequenced) the Cerrado (17.35% reproducibly sequenced), Madagascar and the Indian Ocean Islands (17.97% reproducibly sequenced), the Atlantic Forest (18.33% reproducibly sequenced), and Sundaland (18.52% reproducibly sequenced) (Figure 1.2).

The presence of Sundaland in this bottom-five list is especially alarming, since this region has the highest deforestation rate in the world (Conservation International, 2011).

This may necessitate it being labeled the most important biodiversity hotspot in which to practice reproducible sequencing.

Ferns are Evolutionarily Unique, But They Are Not Receiving Enough Attention

Ferns have a meager diversity of morphological characters to use for species identification and differentiation, so researchers have turned to genetics to do that job (PPG, 2016). Oddly, the algorithm shows that despite this molecular-forward approach less than a quarter of ferns and allies have been reproducibly sequenced. This is especially troubling because ferns and their allies are the oldest lineages of vascular plants, meaning that if the species go extinct before they are reproducibly sequenced their rich evolutionary history will be lost (Arrigo et al., 2013; Stein et al., 2007).

A potential explanation for this paucity in reproducible fern sequences may be that the vast majority of ferns are held at a few institutions—the Natural History Museum London (U.K.), the Muséum National d'Histoire Naturelle (France), and the Royal Botanic Gardens Kew (U.K.)—and many of those vouchers are unmounted and thus unavailable (Carine et al., 2018; NHM, 2020; Morton, 1968). Another explanation could be that very few systematists have focused on ferns. We advocate that priority should be given to obtaining reproducible sequences of ferns and their allies.

DNA Sequencing of Gymnosperms Is Nearly Complete

Gymnosperms are exceptionally well represented, with three-quarters of their species having been reproducibly sequenced. In addition, four hotspots have had all of their known gymnosperm species reproducibly sequenced (New Caledonia, the Western Ghats and Sri Lanka, the Succulent Karoo, and the Coastal Forests of Eastern Africa) (Table 1.1, Figure 1.2). The boom-bust cycle of gymnosperm reproducible publication (and resultant staircase-like shape of their curve in Figure 1.3) may be due to seminal studies like Forest et al. (2018) contributing large numbers of sequences all at once. A possible explanation for their unusually high number of species reproducibly sequenced as compared to the other vascular plant lineages is that—since most gymnosperms are in biodiversity hotspots—their level of threat is much higher: 40% of gymnosperm species (more than double the rate for all species) are at high risk of extinction (Brummitt et al., 2015). This heightened risk may have driven an increased rate of research relative to the other lineages.

Angiosperms Are Poorly Sequenced, But Their Extreme Diversity Provides At Least Some Explanation for That

There are many more Angiosperms relative to the gymnosperms (Burger, 1981), which may help explain their low percentage of reproducibly sequenced species relative to the gymnosperms (18.02%, Table 1.1, Figure 1.3). However, since angiosperms are the dominant plant lineage on the Earth today unravelling their evolutionary history (and thus estimating their biodiversity) is especially important, and consortia such as the Angiosperm phylogeny group have been established to investigate it. However, the APG is primarily doing research at the family and generic level rather than at the species level, so their efforts on their own may not be enough to close the angiosperm sequencing gap (APG, 2016). More species-level focused studies may be warranted.

Proposed Guidelines to Ensure Best Reproducible Practices for DNA Sequencing

The authors propose a set of guidelines to be followed by the botanical community to help ensure the reproducibility of barcode sequences and genomes produced in the future:

i. DNA barcodes or genome sequences produced with the intention of serving as references to be used for species identification should be associated with a voucher that has been deposited in an herbarium.

ii. Vouchers which have been sequenced should have their taxonomical identity verified by a taxonomist/expert in that plant family

iii. The fact that they have an associated voucher (along with any information needed to locate that voucher, i.e. name of herbarium, collection, etc…) should be included in the abstract and the key words of any study using sequences derived from that voucher so as to facilitate sentiments analyses as well as to facilitate future duplication/confirmation of the study.

iv. Any sequence produced with the intention of serving as references to be used for species identification should have the string "reproducibly produced" noted somewhere in its GenBank definition line so that it is possible to narrow searches via the rentrez TITL term so that custom databases of reproducible reference sequences can be created more easily.

v. Researchers should focus on generating barcodes that can be used to infer deep phylogenetic relationships (e.g. *rbcL* and *matK*) and on generating barcodes that can be used to identify sequences to the species level (e.g. ITS).

Conclusion

In conclusion, there is a long way to go before the vascular plants of the world's biodiversity hotspots have been fully sequenced. The unprecedented rate of deforestation means that description and sequencing must be performed faster than ever. However, this effort must be reproducible to ensure that the time and resources spent are not wasted on producing nonreproducible sequences. The package and pipeline we produced will be run again every two years to assess the progress made by the world's scientists in this endeavor.

CHAPTER TWO: UPDATE ON THE TREE OF LIFE OF MALAGASY ANGIOSPERMS: A TOOL TO UNRAVEL THE ORIGIN, EVOLUTION AND BIOGEOGRAPHY OF THIS HIGHLY DIVERSE AND THREATENED FLORA

Introduction

The exceptional richness of Malagasy floristic diversity and its remarkable levels of endemism have been acknowledged since the first botanical collections were made on the island. The 20th century scientist Henri Perrier de la Bâthie and several 21st century researchers have confirmed Madagascar's botanical exceptionalism (Perrier de la Bâthie, 1936; Callmander et al., 2011; Buerki et al., 2012; Lowry et al., 2018). For example, Madagascar is home to an estimated 14,000 species of vascular plants of which over 87% are endemic (Lowry et al., 2018). The angiosperm component represents 95% of the whole vascular plant flora, with 10,650 species (84% endemic) currently described, distributed among 1621 genera (19% endemic) (Callmander et al., 2011). Madagascar has been designated one of the world's most important biodiversity hotspots, mainly because of the high level of diversity and endemism coupled with its unprecedented rate of deforestation, which threatens the survival of its biodiversity and the sustainability of its ecosystems (Myers et al., 2000). The island retains less than 10% of the surface of its natural habitats compared with their original extent before the arrival of the first humans, estimated to have been perhaps 10 millennia ago. Some ecosystems have been reduced to less than 1% of their original area, and the entire eastern rainforest could be entirely eliminated by 2070 (Moat and Smith, 2007; Morelli et al., 2020).

During the last two centuries, botanists have focused primarily on providing a taxonomic framework for the flora of Madagascar, but little is known about the evolutionary processes involved in shaping it (e.g. Buerki et al., 2012). One of the main barriers impeding a better understanding of these processes is the limited availability of well-supported molecular phylogenetic inferences that have been dated using robust calibrations from the fossil record. Buerki et al. (2012) reviewed current knowledge on this topic for endemic Malagasy genera of angiosperms and found phylogenetic information for only 184 of the 310 genera (59.3%), and divergence time estimates were available for only 67 of these genera (21.6%). The authors concluded that we were still in the infancy of our understanding of phylogenetic relationships of the island's unique flora, and they called for more studies. In this contribution, we evaluate the current state of knowledge on the phylogenetic position of Malagasy angiosperms with the ultimate objective of inferring a unified phylogenetic framework of this unparalleled region of the world. Such a framework would provide a unique tool in support of our effort to unravel the evolutionary processes and biogeographic processes that have shaped the Malagasy flora, and it would also be an asset for the formulation of conservation strategies. Indeed, combining phylogenetic data with IUCN Red List assessments (IUCN, 2012) could inform an expansion of the process of prioritizing species conservation, going beyond one based simply on threats by factoring in evolutionary uniqueness. Known as the Evolutionarily Distinct and Globally Endangered (EDGE) approach (https://www.edgeofexistence.org), it was recently applied to gymnosperms worldwide (Forest et al., 2018). Such an ambitious endeavor would require synergy between conservationists and evolutionary biologists, but the result would be well worth the effort as it would provide a unique roadmap to support the conservation of Madagascar's exceptional flora.

Objectives

In this contribution, we provide an updated assessment of our knowledge on the available phylogenetic information for Malagasy angiosperms by using taxonomic data from the Catalogue of the Vascular Plants of Madagascar (Madagascar Catalogue, 2020) via the Global Biodiversity Information Facility (GBIF, 2019) and DNA sequences deposited in GenBank (as of 18 December 2019; https://www.ncbi.nlm.nih.gov/genbank/; NCBI, 1988). We aim to answer the following questions:

i) What proportion of the Malagasy angiosperm flora has been sequenced?

ii) What are the most commonly used DNA barcodes/regions?

iii) Is there a geographical bias among the taxa sequenced?

In association with addressing each of these questions, we provide recommendations on how best to facilitate the completion of the phylogenetic framework of the Malagasy angiosperm flora. The methodology used is the same as that used in chapter one, but without assessing the reproducibility of sequences.

What Proportion of the Malagasy Angiosperm Flora has been Sequenced?

Knowledge provided by genomic data and DNA barcoding opens up a number of possible scientific analyses, such as inferring the phylogenetic position of a species or assessing its spatial and temporal origins (Hebert et al., 2003; Hollingsworth et al., 2012).

Using DNA sequences that are associated with both peer-reviewed articles and voucher specimens is an important and unfortunately often-overlooked aspect of DNA barcode analyses. Without the centuries of accumulated knowledge generated by taxonomists—as provided through the study of herbarium specimens and associated published scientific articles and monographs—the correct identification of the source material of DNA sequences cannot be assured. Our analyses demonstrate that only 4335 species (representing 31.0% of the estimated 14,000 angiosperm species on Madagascar) have at least one DNA sequence available on GenBank (28,386 Malagasy sequences total), representing 1,366 genera (84.3% of the island's angiosperm genera). Of these DNA sequences, only 9,973 (35.1%) are underpinned by publications registered with PubMed (the literature database associated to GenBank accessions; https://www.ncbi.nlm.nih.gov/pubmed). For the 64.9% that are not documented by a publication, the validity of the taxonomic identification of the sample used to produce the sequences has not been verified through needed scientific processes. On average, 14 DNA sequences were produced per publication, with seven publications supplying more than 200 DNA sequences each. A review of the publications containing the most DNA sequences showed that some studies produced data specifically for use in future DNA barcode libraries (e.g., Aubriot et al., 2013), which are underpinned by vouchers, while others generated environmental DNA results, which can never be taxonomically verified since there is no voucher (i.e. the authors have used a metabarcoding approach applied on feces or soil samples, for example; see for example Kartzinel et al., 2015). In any case, the results summarized above show that, despite our best efforts to study this highly diverse and threatened flora, significant effort still needs to be allocated to complete its phylogenetic framework.

The percentages of species with at least one DNA sequence barcode registered with GenBank in the 50 most species-rich families are presented in Table 4. This and subsequently cited percentages represent the proportion of just the currently described species sequenced, not the estimated total number of species in a given family. Over half of the species in 24 families have been sequenced (Table 2.1). The high proportion of Fabaceae is most likely due to the efforts of the Legume Phylogeny Working Group (Azani et al., 2017). Rubiaceae also have a coordinated group of researchers studying them, likely also resulting in high level of sequencing (e.g. Razafimandimbison et al., 2002). Cyperaceae, although less species-rich, are often used as an ecological indicator and thus might have received more attention than would be expected otherwise. Distressingly, despite being the most species-rich family on Madagascar, Orchidaceae (869 species) have DNA sequences available in GenBank representing only about a third of its species. This anomaly might be due to the protected status of orchids under the CITES regulations, limiting collections and the exportation of material. Moreover, for orchids, DNA must usually be extracted from flowers rather than leaves due to their mucilaginous and coriaceous nature. This latter feature makes obtaining tissue from orchid herbarium specimens extremely difficult since it implies destructive sampling of morphologically vital components of the specimen. Overall, since sequencing coverage differs from family to family, the amount of effort required to ensure that each family has some DNA sequences available for each of their species will vary between families. DNA sequences are totally lacking for 10 families representing 28 species on Madagascar. These 10 families are Achariaceae, Cardiopteridaceae, Cytinaceae, Dichapetalaceae, Hydnoraceae, Ixonanthaceae, Kirkiaceae, Peraceae, Picrodendraceae, and Trigoniaceae (Table 2.2).

Madagascar has five endemic families, each of which represents an evolutionary lineage unique to the island. The largest of these (Sarcolaenaceae, 80 species; Table 2.3 and Aubriot et al., 2013) has been comparatively well-sequenced, but more than 40% of its species still need to be surveyed. Only the two smallest families have been comprehensively sequenced, Physenaceae (two species) and Barbeuiaceae (one species).

Our analyses shows that, as you consider families of increasing diversity, the number of species that have at least one DNA sequence deposited in GenBank also increases (Figure 2.1). However, there remains a significant "sequencing gap" (shown by the distance between the number of species sequenced in a family and the total number of species, indicated by the optimum line) that will have to be overcome in order to complete the phylogenetic framework of Malagasy angiosperms (Figure 2.1). At this stage, very few families have some DNA sequences available for each of their species, and these are families with only one or two species in Madagascar.

What are the Most Used DNA Barcodes/Regions?

Most of the DNA sequences available on GenBank were produced using the Sanger sequencing approach. This approach, invented in the mid-seventies (Sanger and Coulson, 1975), involves sequencing DNA regions obtained from polymerase chain reactions (PCR). The first thirty DNA sequences were produced for Malagasy angiosperms 18 years after the invention of Sanger sequencing, all for the *rbcL* region (Figure 2.2). More than 28,000 DNA sequences are now available (see Figure 2.2 and Table 2.4). The rate of DNA sequencing was linear between 1993 and 2010 (during which 5,800 sequences were produced), at which point it began to increase exponentially, with more than 22,500 sequences generated in less than a decade (Figure 2.2). This dramatic change could in part reflect the momentum provided by the DNA barcoding initiative (stimulated by the CBOL Plant Working Group, 2009, which relied on data produced by the Angiosperm Phylogeny Working Group) and the emergence of next-generation sequencing techniques, which first became available in 2005.

The nuclear ribosomal ITS region and plastid coding *rbcL* and *matK* regions are the most frequently used for the study of Malagasy angiosperms, with 9,940, 6,622, and 6,338 DNA sequences, respectively, representing 2,467, 1,906, and 2,200 species (see Figure 2.2 and Table 2.4). Thus, although a total of 4313 species have been sequenced (see above), each of these three DNA regions provides a dataset with limited species coverage. An increase in the sampling of *rbcL* and *matK*—both of which are highly conserved regions in the chloroplast genome useful in elucidating deep relationships—would make it possible to infer a stronger phylogenetic framework of the Malagasy angiosperms, whereas adding sequences of the nuclear ITS region—which is more variable and thus can be used to elucidate more recent divergences—would provide insight into species relationships. One could also imagine taking advantage of target-enriched library techniques (see Johnson et al., 2019) to develop a set of RNA probes allowing high-throughput sequencing of the top 10 DNA regions shown in Table 2.4 for all Malagasy angiosperm taxa. Such libraries could then be pooled and sequenced on next-generation sequencing machines. This approach would be very cost effective and would make it possible to analyze a large set of taxa very rapidly using available bioinformatics pipelines.

Tissue and DNA banks housed in major institutions working on the Malagasy flora (e.g. the Missouri Botanical Garden and the Royal Botanic Gardens, Kew) could provide a second major impetus for the rapid production of DNA sequences for taxa that are currently unsampled. Next-generation sequencing techniques could also be used to tap into historical collections (as discussed in Buerki and Baker, 2016). As the cost of producing genomic data continues to decrease, we argue in favor of producing a suite of barcodes for all taxa that can be used to infer a unified phylogenetic framework of Madagascar's angiosperm flora. However, it will be important to make sure that high-quality genomic DNA extractions are stored for these taxa for subsequent analyses (such as efforts to improve our understanding of fine-scale evolutionary processes) and that all samples used are fully vouchered and reliably identified.

Is there a Geographical Bias among the Taxa Sequenced?

Here, we discuss progress made towards the completion of sequencing species in each of Madagascar's biomes and we assess the percentage of species occurring exclusively outside of protected areas that remain to be sequenced. Our results demonstrate that the effort applied to sequencing species has not been even across the island (Figure 2.3). In general, sequenced species show the same pattern as species richness, with geographical clusters of DNA sequencing effort occurring at the boundaries between biomes (Figure 2.3). Although some geographical regions were sequenced more than others, there is still a vast majority of species requiring to be sequenced, therefore reinforcing the "sequencing gap" as defined above (Figure 2.1). This gap will have to be closed to obtain enough phylogenetic information to study plant communities. Sadly, it is difficult for DNA to be extracted from some plant material, especially collections that were preserved in ethanol to protect them from rotting in the tropical climate, which significantly degrades plant DNA, and which were not accompanied by leaf samples dried in silica-gel, which is a preferred medium for preserving DNA for extraction (Chase and Hillis, 1991). Targeted fieldwork will be required to build a comprehensive tissue bank of the Malagasy flora.

Our analyses showed that species present in protected areas were over-represented in the overall sampling used for DNA sequencing. The rate of sequencing of species recorded inside protected areas (42.1% of 7,651 species) is 1.4 times higher than the rate of sequencing of all Malagasy angiosperm species (31.0%; see Goodman et al. 2018 for a map and description of the terrestrial protected area system of Madagascar). Our analyses indicate that there are ca. 2960 species not known to occur within any protected area, and that 42.6% of them have at least one DNA sequence in GenBank. Although we have not critically evaluated the taxonomical identity of species in this list, we hypothesize that this list would contain many narrowly distributed species, which are of high conservation value. Overall, this analysis suggests that botanists should focus more effort on securing DNA material of species that occur only outside of protected areas since they are, on average, in greater danger of extinction yet have received roughly the same amount of sequencing.

Perspectives

Although botanists have been studying the Malagasy flora for centuries, sequencing technology has only been applied to the island’s plants since 1993 (Figure 2.2). Given that more than 22,000 DNA sequences have been generated during just the last decade, we are very optimistic about the prospects for completing the phylogenetic framework for the angiosperms present on Madagascar. We advocate a coordinated international effort between in-country and international specialists, taxonomists, and phylogeneticists to bridge the remaining sampling gap and to produce sequences for an adequately informative set of DNA barcodes for the remaining ca. 9500 species. These sequences could then be used to infer a complete phylogenetic framework for the angiosperms of Madagascar, providing an unparalleled opportunity to unravel the evolutionary and biogeographic mechanisms that had shaped the origin of this remarkable flora.

REFERENCES

Chapter 1

Arrigo, N., Therrien, J., Anderson, C.L., Windham, M.D., Haufler, C.H. and Barker, M.S. 2013. A total evidence approach to understanding phylogenetic relationships and ecological diversity in *Selaginella* subg. *Tetragonostachys*. *American Journal of Botany*, 100: 1672-1682. doi:10.3732/ajb.1200426.

APG. 1998, An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85 (4): 531–553.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533: 7604.

Bidartondo, M.I. 2008. Preserving Accuracy in Genbank, *Science, New Series,* Vol. 319, No. 5870 p. 1616.

Bilofsky H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I, Rindone, W.P., Swindell, C.D., Tung, C.S. 1986. The GenBank genetic sequence databank, *Nucleic Acids Research*, Volume 14, Issue 1, Pages 1–4, https://doi.org/10.1093/nar/14.1.1.

Bivand R., Keitt T., Rowlingson B. 2019. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. *R package* version 1.4-8. https://CRAN.R-project.org/package=rgdal.

Bivand R., Lewin-Koh, N. 2019. maptools: Tools for Handling Spatial Objects. *R package* version 0.9-9.

Brummitt, N.A., Bachman, S.P., Griffiths-Lee, J., Lutz, M., Moat, J.F., et al. 2015. Green Plants in the Red: A Baseline Global Assessment for the IUCN Sampled Red List Index for Plants. *PLOS ONE* 10(8): e0135152.

Buerki, S., Lowry II, P.P., Munzinger, J., Tuiwawa, M., Naikatini, A., Callmander, M.W. 2017. *Alectryon vitiensis*: A new species of Sapindaceae endemic to Fiji. *Novon*, 25: 4210-429. doi:10.3417/D-16-00006.

Buerki S, Devey DS, Callmander MW, Phillipson PB, Forest F. 2013. Spatio-temporal history of the endemic genera of Madagascar. *Botanical Journal of the Linnean Society*, 171: 304-329. doi:10.1111/boj.12008.

Burger, W.C. 1981. Why Are There So Many Kinds of Flowering Plants?, *BioScience* 31: 572–581.

Carine, M.A. 2018. "Examining the spectra of herbarium uses and users". *Botany Letters* (2381-8107), 165 (3-4), p. 328.

Cayuela, L., Macarro, I., Stein, A., Oksanen, J. 2019. Taxonstand: Taxonomic Standardization of Plant Species Names. *R package version 2.2.* https://CRAN.R-project.org/package=Taxonstand.

Chamberlain S, Szoecs E, Foster Z, Arendsee Z, Boettiger C, Ram K, Bartomeus I, Baumgartner J, O'Donnell J, Oksanen J, Tzovaras BG, Marchand P, Tran V, Salmon M, Li G, Grenié M (2020). taxize: Taxonomic information from around the web. *R package version 0.9.95*, https://github.com/ropensci/taxize.

Conservation International 2011. [https://www.conservation.org/NewsRoom/pressreleases/Pages/The-Worlds-10-Most-Threatened-Forest-Hotspots.aspx#ranking].

CEPS. 2016. Centre for European Policy Studies. CEPS. https://www.ceps.eu/.

Devey, D.S., Forest, F., Rakotonasolo, F., Ma, P., Dentinger, B.T.M., Buerki, S. 2013. A snapshot of extinction in action: The decline and imminent demise of the endemic Eligmocarpus Capuron (Caesalpinioideae, Leguminosae) serves as an example of the fragility of Madagascan ecosystems. *South African Journal of Botany, Special Issue: Towards a New Classification System for Legumes,* 89: 273-280. doi:10.1016/j.sajb.2013.06.013.

Duncan Temple Lang 2020. XML: Tools for Parsing and Generating XML Within R and S-Plus. *R package* version 3.99-0.3. https://CRAN.R-project.org/package=XML .

Forest, F., Baloch, E., Brummitt, N.A., Bachman, S., Moat, J., Ickert-Bond, S., Hollingsworth, P.M., Liston, A., Little, D.P., Mathews, S., Rai, H., Rydin, C., Stevenson, D.W., Thomas, P., Buerki, S. 2018. Gymnosperms on the EDGE. *Scientific Reports,* 8: 6053. doi:10.1038/s41598-018-24365-4.

Fox B. 1989. Bash is in beta release! Newsgroup: gnu.announce.

GBIF.org. 2020. GBIF Occurrence Download https://doi.org/10.15468/dl.xnjza0.

Global Biodiversity Information Facility. 2001. *GBIF Memorandum of Understanding.*

Grace, J., Anderson, T., Seabloom, E. et al. Integrative modelling reveals mechanisms linking productivity and plant species richness. Nature 529, 390–393.

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B: Biological Sciences* 270: 313–321.

Hoffman M, Koenig K, Bunting G, Costanza J, & Williams KJ. 2016. Biodiversity Hotspots (version 2016.1) (Version 2016.1) [Data set]. *Zenodo*. http://doi.org/10.5281/zenodo.3261807.

Hollingsworth, P. M., Graham, S. W., and Little, D. P. 2011. Choosing and Using a Plant DNA Region. *PLoS ONE* 6: 19254.

Kalland A 1993. Management by totemization: whale symbolism and the anti-whaling campaign. *Arctic* 46: 124–133.

Knegtering E, Hendrickx L, van der Windt HJ, Uiterkamp A 2002. Effects of species' characteristics on nongovernmental organizations' attitudes toward species conservation policy. *Environment and Behavior* 34: 378–400.

Kovalchik, S. 2017. RISmed: Download Content from NCBI Databases. R package version 2.1.7. https://CRAN.R-project.org/package=RISmed.

Lang D. T. 2020. XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.99-0.3. https://CRAN.R-project.org/package=XML.

Larsson, J. 2020. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. *R package* version 6.1.0, <URL:https://cran.r-project.org/package=eulerr>.

Lindberg, D.A. 2000. "Internet access to the National Library of Medicine" (PDF). *Effective Clinical Practice*. **3** (5): 256–60. PMID 11185333. Archived from the original (PDF) on 2 November 2013.

Microsoft Corporation and Stephen Weston 2019. doSNOW: Foreach Parallel Adaptor for the 'snow' Package. *R package* version 1.0.18. https://CRAN.R-project.org/package=doSNOW.

Morton CV. 1968 The fern collections in some European herbaria. *American Fern Journal* 58: 158-168.

Myers, N. 1988. Threatened biotas: "Hot spots" in tropical forests. *Environmentalist* 8, 187–208. https://doi.org/10.1007/BF02240252.

Myers, N., Mittermeier, R., Mittermeier, C. *et al.* 2000. Biodiversity hotspots for conservation priorities. *Nature* 403:853–858.

National Center for Biotechnology Information (NCBI) 1988. Bethesda (MD): National Library of Medicine (US), *National Center for Biotechnology Information;* [1988] – [2020 Jan 29]. Available from: https://www.ncbi.nlm.nih.gov/.

Natural History Museum. 2020. Fern Collections. https://www.nhm.ac.uk/our-science/collections/botany-collections/fern-collections.html.

Otto Sarah P. 2018. Adaptation, speciation and extinction in the Anthropocene. *Proc. R. Soc. B* 285: 20182047.

Pace NR, Stahl DA, Lane DJ, Olsen GJ 1986. "The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences". In Marshall KC (ed.). *Advances in Microbial Ecology*. 9. *Springer* US. pp. 1–55. doi:10.1007/978-1-4757-0611-6_1. ISBN 978-1-4757-0611-6.

Pagès H., Aboyoun P., Gentleman R., DebRoy S. 2019. Biostrings: Efficient manipulation of biological strings. *R package* version 2.54.0.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. *R News* 5 (2), https://cran.r-project.org/doc/Rnews/.

PPG. 2016. A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution* 54 (6): 563–603.

R Core Team 2017. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rinker, T. W. 2019. sentimentsr: Calculate Text Polarity Sentiments version 2.7.1. http://github.com/trinker/sentimentsr.

Stein, W. E.; Mannolini, F.; Hernick, L. V.; Landling, E.; Berry, C. M. 2007. "Giant cladoxylopsid trees resolve the enigma of the Earth's earliest forest stumps at Gilboa". *Nature.* 446 (7138): 904–907. Bibcode:2007Natur.446..904S. doi:10.1038/nature05705. PMID 17443185.

The Plant List 2013. Version 1.1. *Published on the Internet*; http://www.theplantlist.org/ (accessed 1st January).

Tierney L., Rossini A.J., Li N., Sevcikova, H. 2018. snow: Simple Network of Workstations. *R package* version 0.4-3. https://CRAN.R-project.org/package=snow.

Taboada, Maite; Brooke, Julian 2011. "Lexicon-based methods for sentiments analysis". *Computational Linguistics*. 37 (2): 272–274. CiteSeerX 10.1.1.188.5517. doi:10.1162/coli_a_00049.

Waters C.N., Zalasiewicz, J., Summerhayes, C., Barnosky, A., Poirier, C., et al. 2016. The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* 351: aad2622.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag.*

Wickham, H., Hester, J., Francois, R. 2018. readr: Read Rectangular Text Data. *R package* version 1.3.1. New York.

Wickham, H., Hester, J., Chang, W. 2020. devtools: Tools to Make Developing R Packages Easier. *R package* version 2.2.2. https://CRAN.R-project.org/package=devtools.

Williamson, J., Maurin, O., Shiba, S.N.S., van der Bank, H., Pfab M., Pilusa, M., Kabongo, R.M., van der Bank, M. 2016. Exposing the illegal trade in cycad species (Cycadophyta: Encephalartos) at two traditional medicine markets in South Africa using DNA barcoding. *Genome* 59:771-781, https://doi.org/10.1139/gen-2016-0032.

Winter, D. J. 2017. rentrez: an R package for the NCBI eUtils API. *The R Journal* 9(2):520-526 https://CRAN.R-project.org/package=readr.

Wu, R., Padmanabhan, R., Bambara, R. 1974. Nucleotide sequence analysis of bacteriophage DNA. *Methods in Enzymology* 29: 231-253.

Zhorn Media. 2017. Caffeine. https://zhornsoftware.co.uk/caffeine/index.html.

Chapter 2

Aubriot, X., Lowry II, P. P., Cruaud, C., Couloux, A., and Haevermans, T. 2013. DNA barcoding in a biodiversity hot spot: Potential value for the identification of Malagasy *Euphorbia* L. listed in CITES Appendices I and II. *Molecular Ecology Resources* 13: 57-65.

Azani, N., Babineau, M., Bailey, C.D., Banks, H., Barbosa, A.R., Pinto, R.B., Boatwright, J.S., Borges, L.M., Brown, G.K., Bruneau, A., Candido, E., Cardoso, D., Chung, K., Clark, R.P., Conceição, A.d.S., Crisp, M., Cubas, P., Delgado-Salinas, A., Dexter, K.G., Doyle, J.J., Duminil, J., Egan, A.N., de la Estrella, M., Falcão, M.J., Filatov, D.A., Fortuna-Perez, A.P., Fortunato, R.H., Gagnon, E., Gasson, P., Rando, J.G., de Azevedo Tozzi, A.M.G., Gunn, B., Harris, D., Haston, E., Hawkins, J.A., Herendeen, P.S., Hughes, C.E., Iganci, J.R., Javadi, F., Kanu, S.A., Kazempour-Osaloo, S., Kite, G.C., Klitgaard, B.B., Kochanovski, F.J., Koenen, E.J., Kovar, L., Lavin, M., le Roux, M., Lewis, G.P., de Lima, H.C., López-Roberts, M.C., Mackinder, B., Maia, V.H., Malécot, V., Mansano, V.F., Marazzi, B., Mattapha, S., Miller, J.T., Mitsuyuki, C., Moura, T., Murphy, D.J., Nageswara-Rao, M., Nevado, B., Neves, D., Ojeda, D.I., Pennington, R.T., Prado, D.E., Prenner, G., de Queiroz, L.P., Ramos, G., Filardi, F.L.R., Ribeiro, P.G., de Lourdes Rico-Arce, M., Sanderson, M.J., Santos-Silva, J., São-Mateus, W.M., Silva, M.J., Simon, M.F., Sinou, C., Snak, C., de Souza, É.R., Sprent, J., Steele, K.P., Steier, J.E., Steeves, R., Stirton, C.H., Tagane, S., Torke, B.M., Toyama, H., da Cruz, D.T., Vatanparast, M., Wieringa, J.J., Wink, M., Wojciechowski, M.F., Yahara, T., Yi, T. and Zimmerman, E. 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). *Taxon* 66: 44-77.

Buerki, S., and Baker, W. J. 2016. Collections-based research in the genomic era. *Biological Journal of the Linnaean Society* 117: 5-10.

Buerki, S., Devey, D., Callmander, M., Phillipson, P., and Forest, F. 2012. Spatio-temporal history of the endemic genera of Madagascar. *Botanical Journal of the Linnean Society* 171: 304-329.

Callmander, M. W., Phillipson, P. B., Schatz, G. E., Andriambololonera, S., Rabarimanarivo, M., Rakotonirina, N., Raharimampinona, J., Chatelain, C., Gautier, L., Lowry II, P. P. 2011. The endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecology and Evolution* 144: 121-125.

CBOL Plant Working Group 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Science* 106: 12794-12797.

Chase, M. W., Hills, H. H. 1991. Silica gel: an ideal material for field preservation of leaf samples for DNA studies. *Taxon* 40: 215-220.

Forest, F., Moat, J., Baloch, E., Brummitt, N. A., Bachman, S. P., Ickert-Bond, S., Hollingsworth, P. M., Liston, A. Litle, D. P., Mathews, S., Rai, H. Rydin, C., Stevenson, D. W., Thomas, P., Buerki, S. 2018. Gymnosperms on the EDGE. *Scientific Reports* 8: 6053.

GBIF 2019. *GBIF Home Page* [https://www.gbif.org].

Goodman, S. M., Raherilalao, M. J., and Wohlhauser, S. (eds.) 2018. *Les aires protégées terrestres de Madagascar : Leur histoire, description et biote / The terrestrial protected areas of Madagascar: Their history, description, and biota*. Antananarivo: Association Vahatra.

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B: Biological Sciences* 270: 313-321.

Hollingsworth, P. M., Graham, S. W., and Little, D. P. 2011. Choosing and using a plant DNA region. *PLoS One* 6: 19254.

IUCN. 2012. *IUCN Red List Categories and Criteria: Version 3.1*. Ed.2. Gland and Cambridge: IUCN Species Survival Commission.

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G. K., Baker, W. J., & Wickett, N. J. 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Systematic biology, 68: 594–606.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., Rubenstein, D. I., Wang, W., and Pringle, R. M. 2015. DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Science* 112: 8019-8024.

Lowry, P. P., II, Phillipson, P. B., Andriamahefarivo, L., Schatz, G. E., Rajaonary, F., and Andriambololonera, S. 2018. *Flore/Flora*. In *Les aires protégées terrestres de Madagascar : Leur histoire, description et biote / The terrestrial protected areas of Madagascar: Their history, description, and biota,* eds. S. M. Goodman, M. J. Raherilalao, and S. Wohlhauser, pp. 243-255. Antananarivo: Association Vahatra.

Madagascar Catalogue Project. 2019. *Catalogue of the plants of Madagascar.* St. Louis & Antananarivo [http://www.tropicos.org/project/mada].

Moat, J., and Smith, P. 2007. *Atlas of the vegetation of Madagascar*. Kew: Kew Publishing.

Morelli, T. N., Smith, A. B. Mancini, A. N., Balko, E. A., Borgerson, C. et al. 2019. The fate of Madagascar's rainforest habitat. *Nature Climate Change* 10: 89.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., and Kent, J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853-858.

Perrier de la Bâthie, H. 1936. *Biogéographie des plantes de Madagascar*. Paris: Société d'éditions géographiques, maritimes et coloniales.

Razafimandimbison, S. G., & Bremer, B. 2002. Phylogeny and classification of Naucleeae s.l. (Rubiaceae) inferred from molecular (ITS, rBCL, and tRNT-F) and morphological data. *American Journal of Botany* 89(7): 1027–1041.

Sanger, F., and Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94: 441-448.

Schatz, G. E. 2002. Taxonomy and herbaria in service of plant conservation: Lessons from Madagascar's endemic families. *Annals of the Missouri Botanical Garden* 89: 145-152.

TABLES

Table 1.1 A matrix showing the number of species and percentage of global flora for the species lists at 4 different curation levels. Species in A have at least one sequence in GenBank. Species in B are reproducibly sequenced. Species in C are species represented solely by one or more potential future reproducibly sequences. B + C is a representation of what B could look like if the species in C are all reproducibly sequenced. The percentages in the parentheses represent the percent of all taxa or the percent of that lineage those species represent.

Curation Level	All Vascular Plants	Angiosperms	Gymnosperms	Ferns and Allies
A	89,314 (43.77%)	85,086 (43.54%)	816 (91.17%)	3,412 (44.22%)
B	37,687 (18.47%)	35,217 (18.02%)	671 (74.97%)	1,749 (22.67%)
C	16,045 (7.86%)	15,301(7.83%)	36 (4.02%)	708 (9.18%)
B + C	53,732 (26.33%)	50,518 (25.85%)	707 (78.99%)	2,457 (31.85%)

Table 1.2 A matrix showing the top 25 journals (names in ISO abbreviation) by number of reproducible studies published in them, showing the number and percent total of reproducible studies and data policy (as manually collected from the journals' instructions for authors) for each journal.

Journal ISO abbreviation	Number of reproducible studies underpinning species in list B	Percent total reproducible studies underpinning species in list B	Requires (NOT just recommends) data to be publicly archived
Mol. Phylogenet. Evol.	532	20.76%	No
Am. J. Bot.	412	16.07%	Yes
PLoS ONE	252	9.83%	Yes
Ann. Bot.	93	3.63%	Yes
Mol. Ecol.	87	3.39%	Yes
BMC Evol. Biol.	72	2.81%	Yes
New Phytol.	59	2.30%	Yes
Proc. Natl. Acad. Sci. U.S.A.	57	2.22%	Yes
Mitochondrial DNA A DNA Mapp Seq Anal	56	2.18%	Yes
Sci Rep	54	2.11%	Yes
J. Plant Res.	51	1.99%	No
Mol. Biol. Evol.	48	1.87%	Yes
Mol Ecol Resour	40	1.56%	Yes
Syst. Biol.	37	1.44%	Yes
Evolution	35	1.37%	Yes
Genome Biol Evol	35	1.37%	Yes
BMC Plant Biol.	34	1.33%	No

J. Mol. Evol.	25	0.98%	No
BMC Genomics	19	0.74%	No
Curr. Genet.	19	0.74%	No
Front Plant Sci	19	0.74%	Yes
Gene	19	0.74%	Yes
Genome	17	0.66%	Yes
Plant Biol (Stuttg)	17	0.66%	Yes
Biol. Pharm. Bull.	15	0.59%	Yes

Table 1.3 A matrix showing the most commonly utilized barcodes for studies contributing species to list B ordered by the number of species represented by at least on barcode of that type. The number of sequences and percent of global hotspot flora represented by those species are also indicated.

Barcode	Number of sequences	Number of species	% total global hotspot flora
ITS	58,791	23,422	11.48%
matK	36,269	17,164	8.41%
rbcL	38,265	16,509	8.09%
trnL	15,646	9,091	4.46%
psbA	13,322	6,330	3.10%
rpoB	6,686	3,173	1.56%
rpoC1	6,857	3,147	1.54%
atpF	5,404	2,781	1.36%
atpH	4,887	2,598	1.20%
psbK	4,433	2,408	1.18%
psbI	4,132	2,270	1.11%
trnH	4,141	1,981	0.97%
Plastid genome	3,123	300	0.15%
Nuclear genome	193	10	0.005%

Table 2.1 Number of Malagasy species represented in GenBank per family, the percent of total species in Madagascar sequenced, and the species richness rank (based on data from Madagascar Catalogue 2020) for the top 50 most sequenced families of the Malagasy flora.

Family	Number of species in GenBank	Percent of family sequenced	Species richness rank as per the Madagascar Catalogue
Fabaceae	415	59.5	2
Poaceae	377	75.4	4
Rubiaceae	354	54.2	3
Orchidaceae	289	36.0	1
Euphorbiaceae	189	48.2	6
Compositae	155	31.4	5
Apocynaceae	153	47.7	8
Cyperaceae	139	54.3	10
Malvaceae	126	34.5	7
Acanthaceae	65	21.7	9
Melastomataceae	56	22.2	11
Convolvulaceae	55	62.5	22
Oleaceae	49	79.0	36
Solanaceae	48	75.0	31
Araliaceae	46	70.8	30
Annonaceae	45	62.5	26
Arecaceae	44	22.3	13
Sarcolaenaceae	44	65.7	28
Phyllanthaceae	42	42.9	19
Lamiaceae	40	17.1	12

Sapindaceae	39	38.2	18
Xanthorrhoeaceae	37	34.9	17
Sapotaceae	36	50.0	27
Cucurbitaceae	35	54.7	32
Crassulaceae	33	55.9	37
Anacardiaceae	32	48.5	29
Burseraceae	32	94.1	54
Ebenaceae	31	40.3	24
Gentianaceae	30	47.6	33
Bignoniaceae	28	45.2	35
Pandanaceae	27	30.3	21
Balsaminaceae	27	24.6	16
Passifloraceae	26	70.3	49
Asparagaceae	24	49.0	41
Moraceae	24	64.9	50
Primulaceae	23	20.7	15
Vitaceae	23	67.7	55
Meliaceae	22	26.5	23
Amaranthaceae	22	51.2	43
Gesneriaceae	22	51.2	44
Dioscoreaceae	21	51.2	46
Lauraceae	21	18.4	14
Celastraceae	21	58.3	53
Urticaceae	21	37.5	40

Hypericaceae	20	66.7	59
Boraginaceae	17	47.2	52
Violaceae	17	60.7	64
Piperaceae	17	37.8	42
Araceae	16	64.0	67
Myrtaceae	16	21.3	25

Table 2.2 Number of genera and species in each of the ten families in Madagascar with no sequences in GenBank. Number of species and genera in each family are based on data from Madagascar Catalogue (2020).

Family	Number of genera	Number of species
Achariaceae	1	5
Cardiopteridaceae	1	2
Cytinaceae	1	2
Dichapetalaceae	1	8
Hydnoraceae	1	1
Ixonanthaceae	1	1
Kirkiaceae	1	1
Peraceae	1	1
Picrodendraceae	3	8
Trigoniaceae	1	1

Table 2.3 Number of species represented in GenBank for each of Madagascar's endemic families, the percent of total species sequenced, and the species richness rank (based on data from Madagascar Catalogue 2020).

Family	Number of species in GenBank	Percent of family sequenced	Number of currently recognized species
Sarcolaenaceae	44	56.4	78
Sphaerosepalaceae	3	15.0	20
Physenaceae	2	100.0	2
Asteropeiceae	2	25.0	8
Barbeuiaceae	1	100.0	1

Table 2.4 Number of sequences from Malagasy material for each of the 12 phylogenetically useful DNA regions as defined by Hollingsworth et al. (2011), number of species represented, and percent of species of total Malagasy species sequenced.

Region	Type	Number of sequences	Number of species	Percent of species
ITS region	Nuclear	9940	2467	25.2%
matK	Plastid	6338	2200	22.5%
rbcL	Plastid	6622	1906	19.5%
trnL	Plastid	3038	1448	14.8%
trnF	Plastid	1314	767	7.8%
psbA	Plastid	1025	533	5.5%
trnH	Plastid	579	328	3.4%
atpH	Plastid	94	59	0.6%
psbK	Plastid	67	37	0.4%
psbI	Plastid	38	36	0.4%
rpoC1	Plastid	0	0	0.0%
rpoB	Plastid	0	0	0.0%

FIGURES

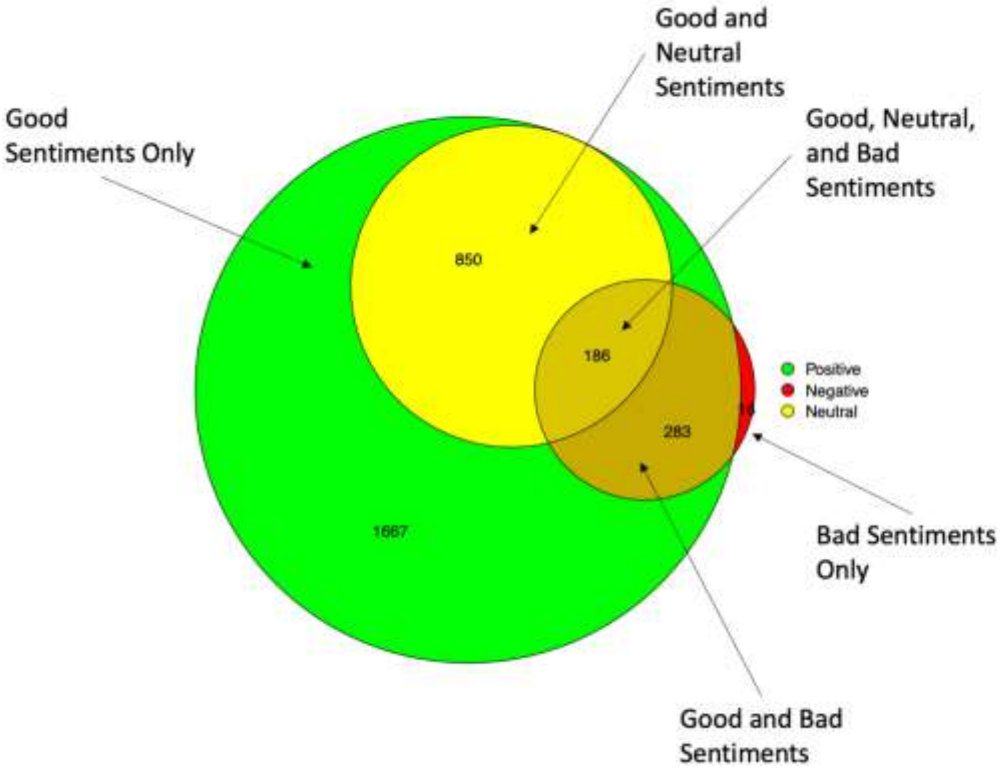
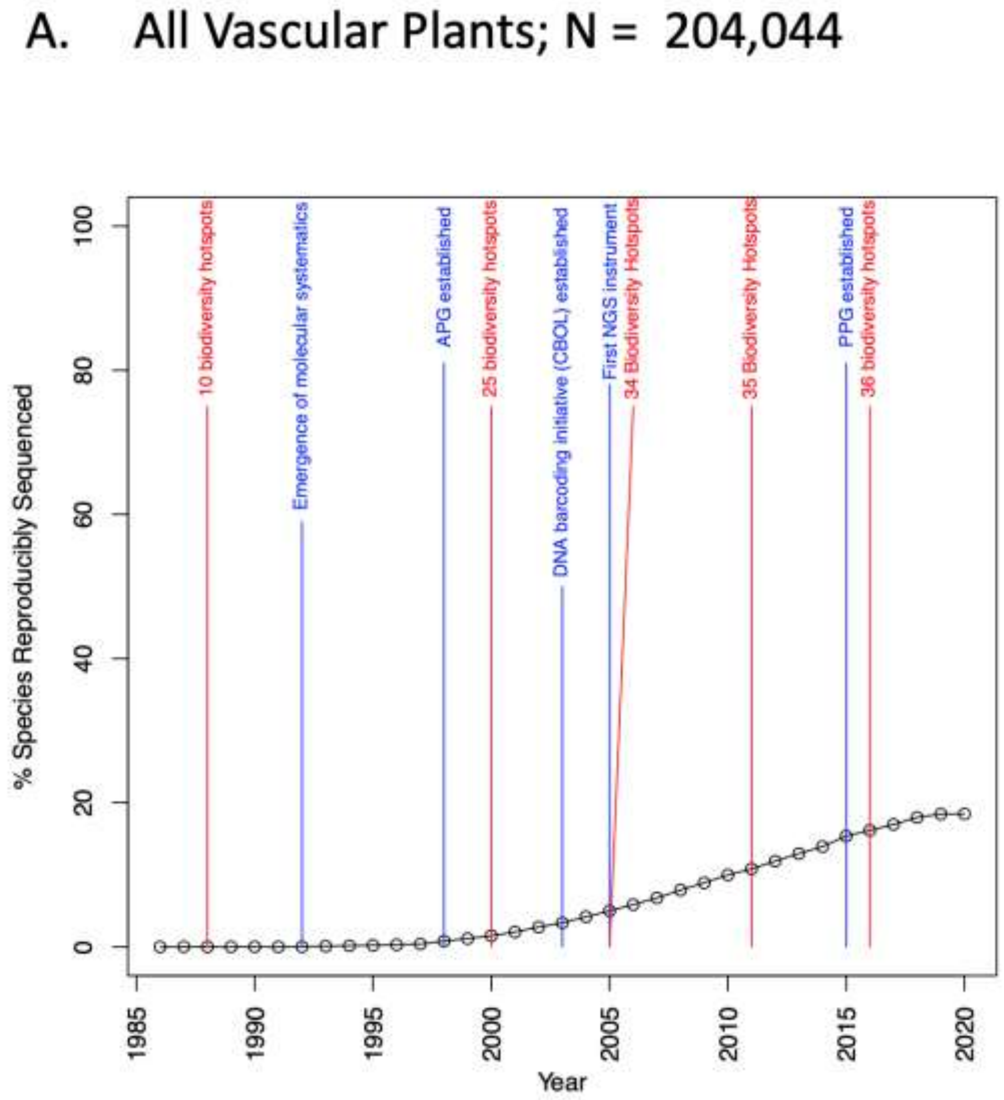


Figure 1.1 An elliptical Venn diagram illustrating the overlap and proportionality of the sentiments classifications of the abstracts mined from PubMed after manual curation.



A. All Vascular Plants; N = 204,044

Figure 1.2A Cumulative curve representing the percent of species reproducibly sequenced for each year 1986 to 2020 for all vascular plants. APG stand for the Angiosperm Phylogeny Group and PPG stands for the Pteridophyte Phylogeny Group.

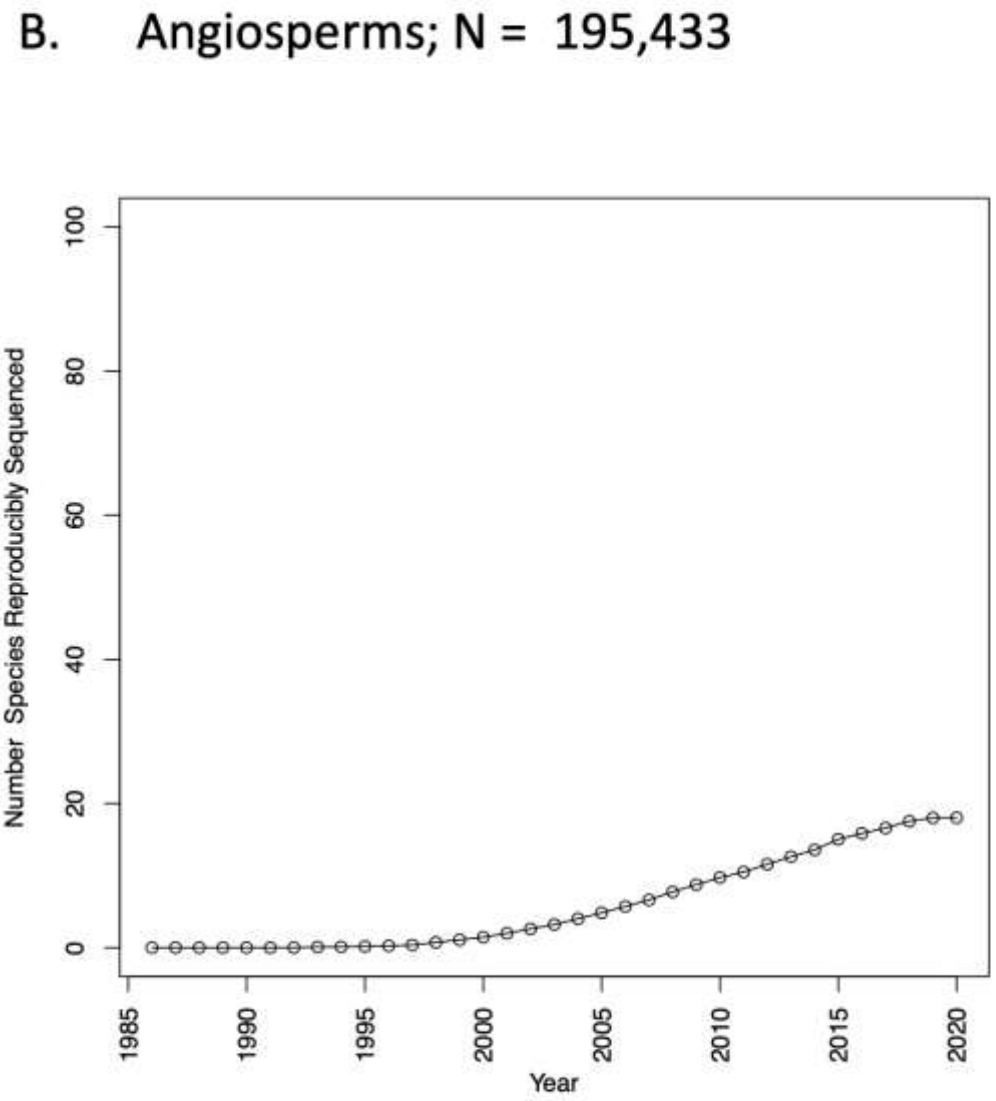
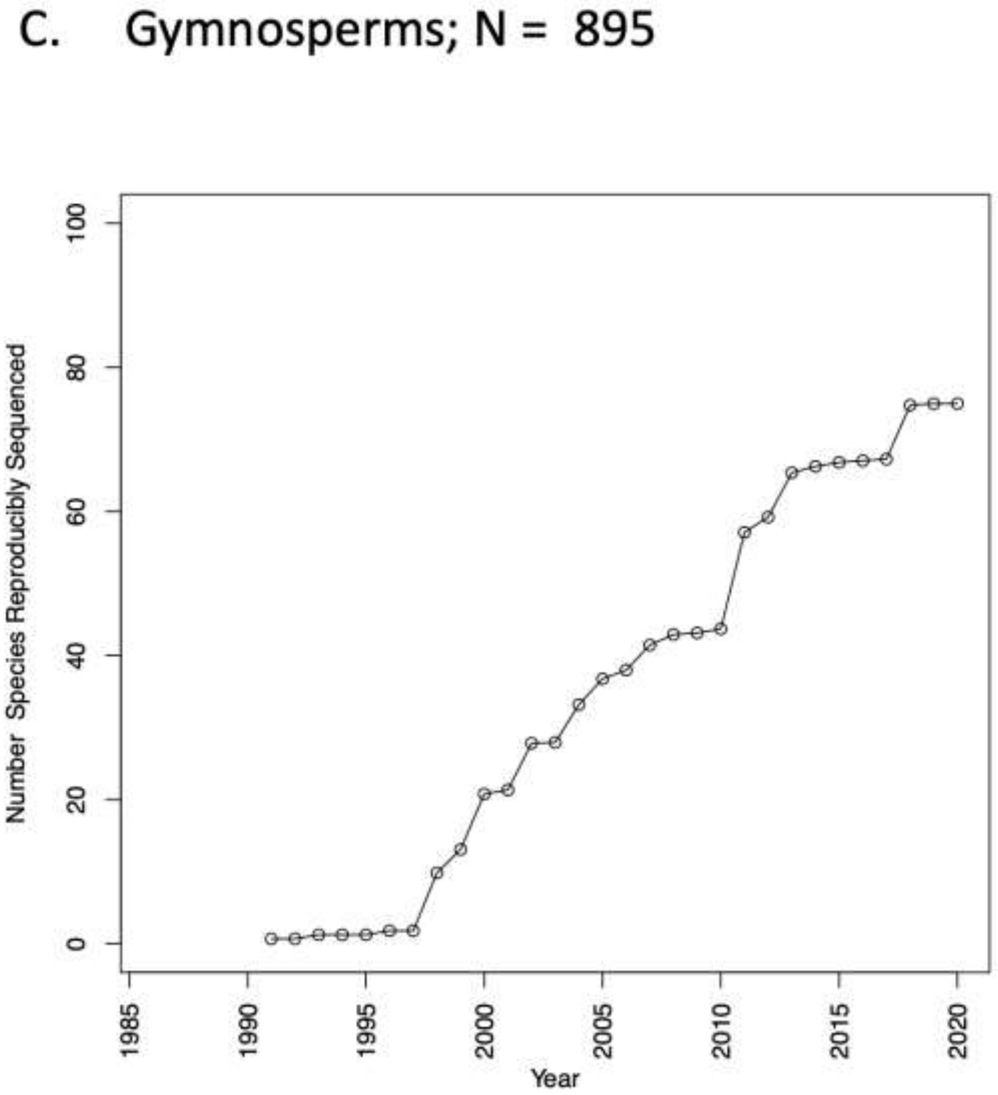
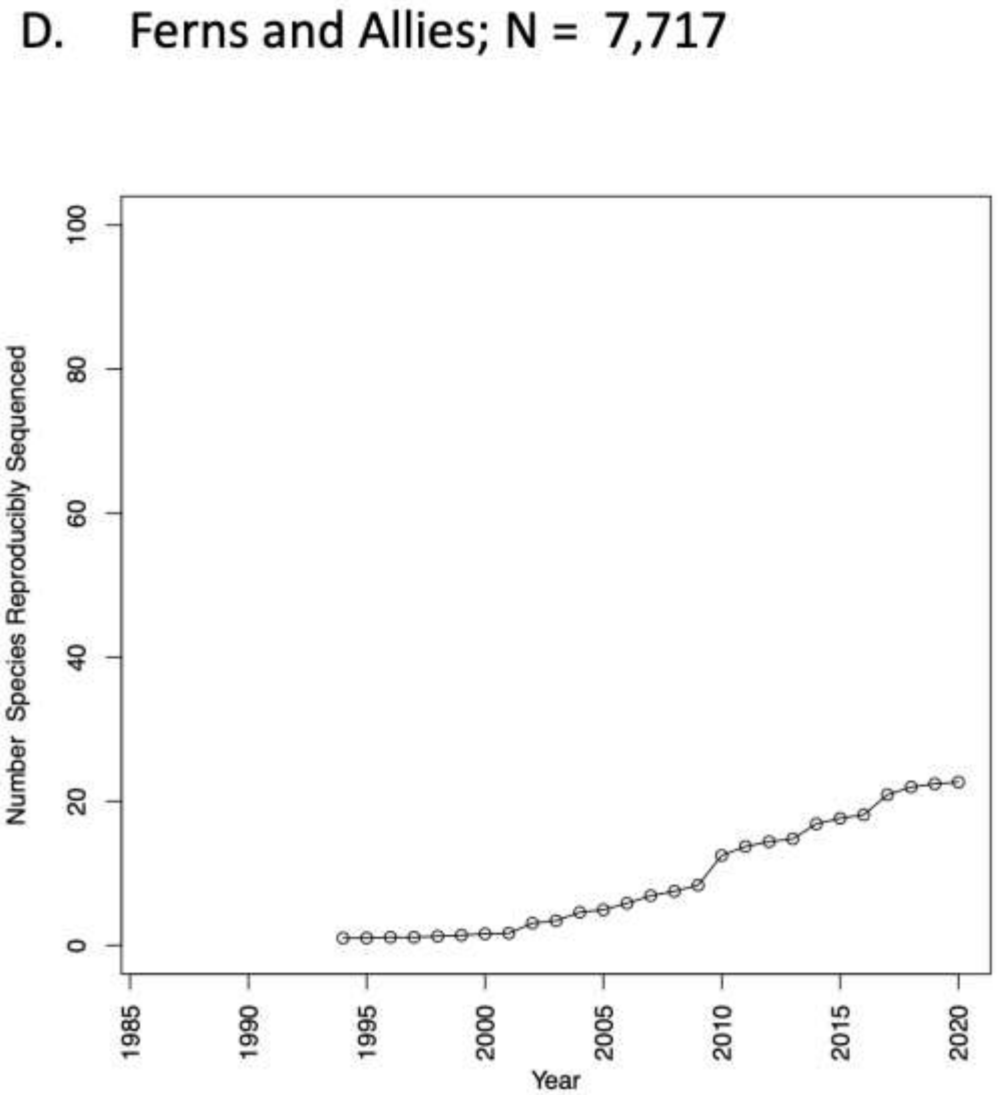


Figure 1.2B Cumulative curve representing the percent of species reproducibly sequenced for each year 1986 to 2020 for angiosperms. APG stand for the Angiosperm Phylogeny Group and PPG stands for the Pteridophyte Phylogeny Group.



C. Gymnosperms; N = 895

Figure 1.2C Cumulative curve representing the percent of species reproducibly sequenced for each year 1986 to 2020 for gymnosperms. APG stand for the Angiosperm Phylogeny Group and PPG stands for the Pteridophyte Phylogeny Group.



D. Ferns and Allies; N = 7,717

Figure 1.2D Cumulative curve representing the percent of species reproducibly sequenced for each year 1986 to 2020 for ferns and allies. APG stand for the Angiosperm Phylogeny Group and PPG stands for the Pteridophyte Phylogeny Group.

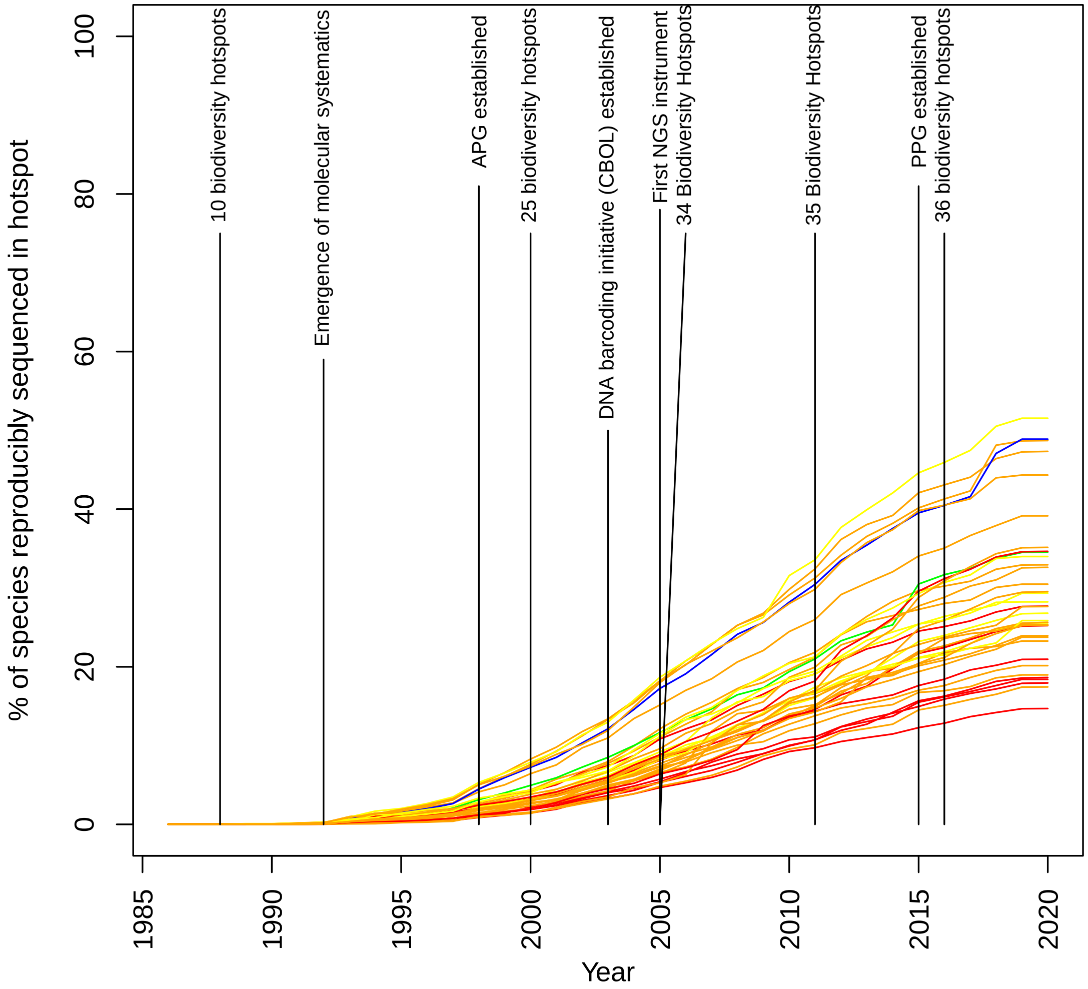


Figure 1.3 A plot of cumulative curves representing the percent of species reproducibly sequenced for each year 1986 to 2020 for the 36 biodiversity hotspots colored according to their year of establishment (see key). Notice that the older biodiversity hotspots are less thoroughly reproducibly sequenced than are the newer hotspots. Note that the curves account for sequences added even before the establishment of a hotspot.

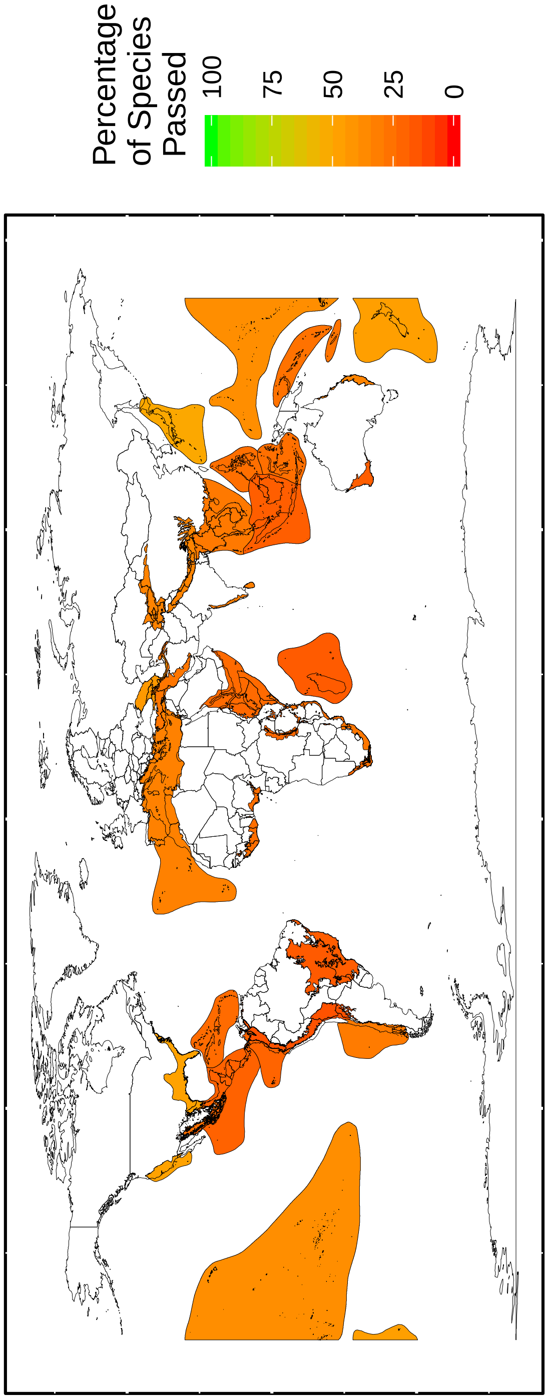


Figure 1.4A Map for all vascular plants showing all 36 biodiversity hotspots colored according to a scale where redder shades mean fewer species are represented by at least one reproducible sequence (i.e. fewer species passed) and greener shades mean more species are represented by at least one reproducible sequence.

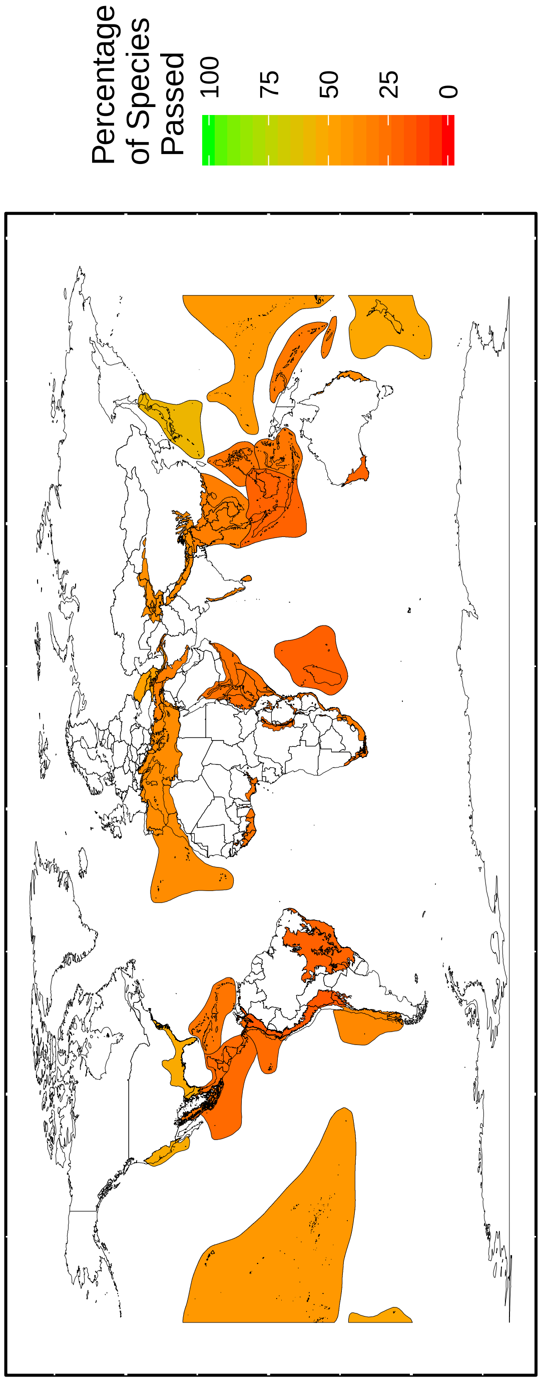


Figure 1.4B Map for angiosperms showing all 36 biodiversity hotspots colored according to a scale where redder shades mean fewer species are represented by at least one reproducible sequence (i.e. fewer species passed) and greener shades mean more species are represented by at least one reproducible sequence.

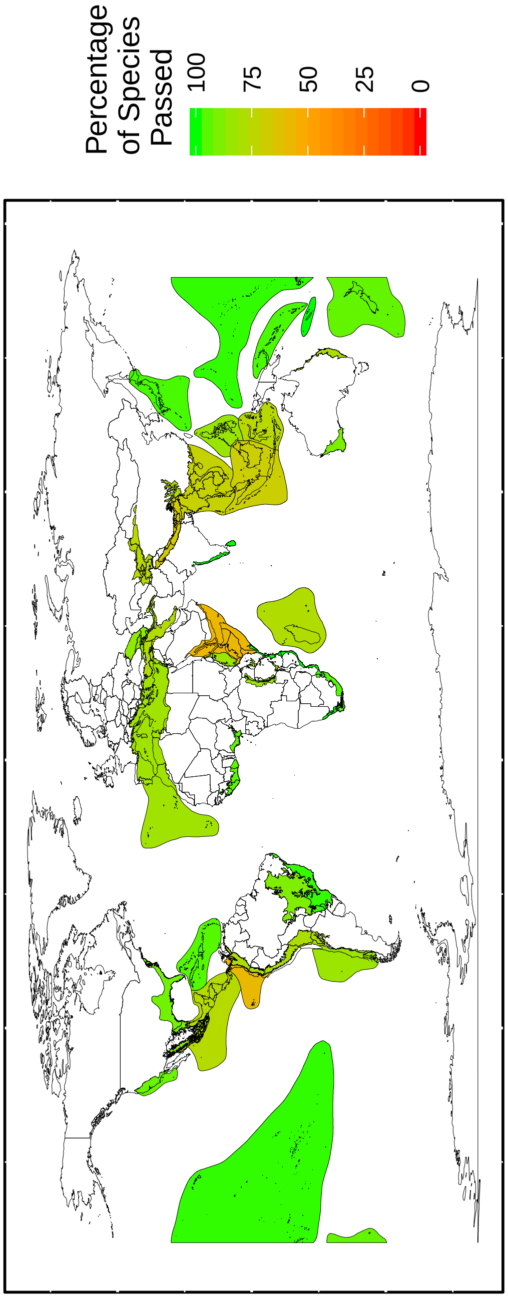


Figure 1.4C Map for gymnosperms showing all 36 biodiversity hotspots colored according to a scale where redder shades mean fewer species are represented by at least one reproducible sequence (i.e. fewer species passed) and greener shades mean more species are represented by at least one reproducible sequence.

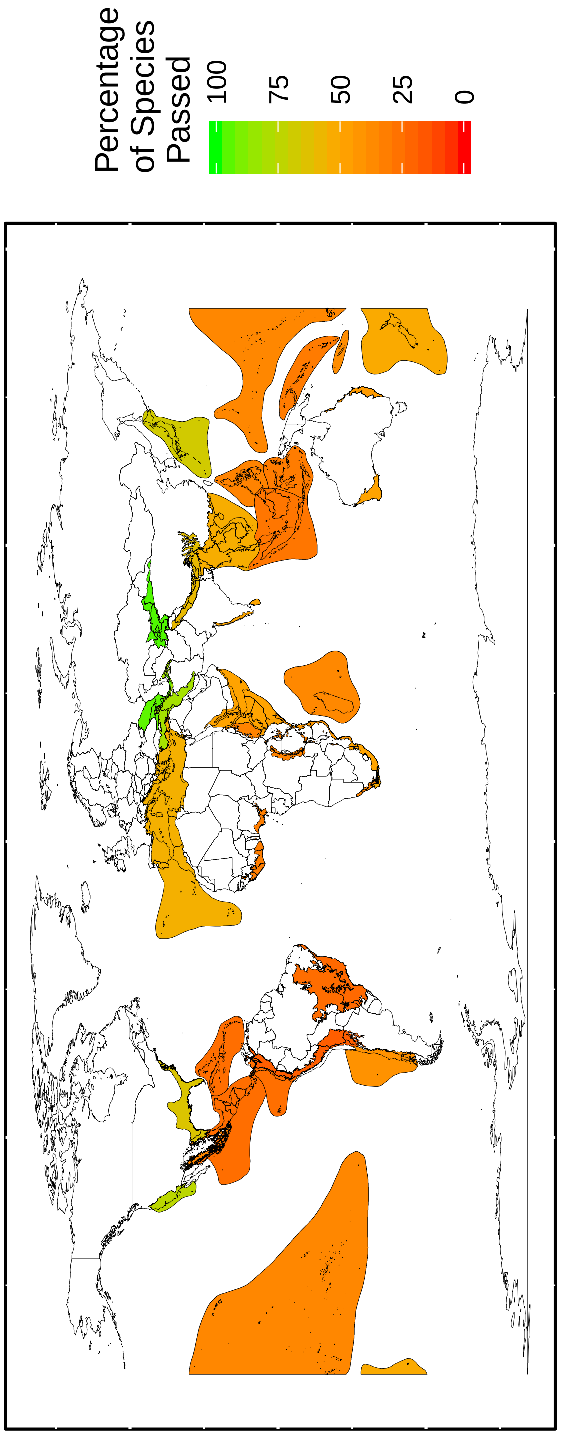


Figure 1.4D Map for Ferns and Allies showing all 36 biodiversity hotspots colored according to a scale where redder shades mean fewer species are represented by at least one reproducible sequence (i.e. fewer species passed) and greener shades mean more species are represented by at least one reproducible sequence.

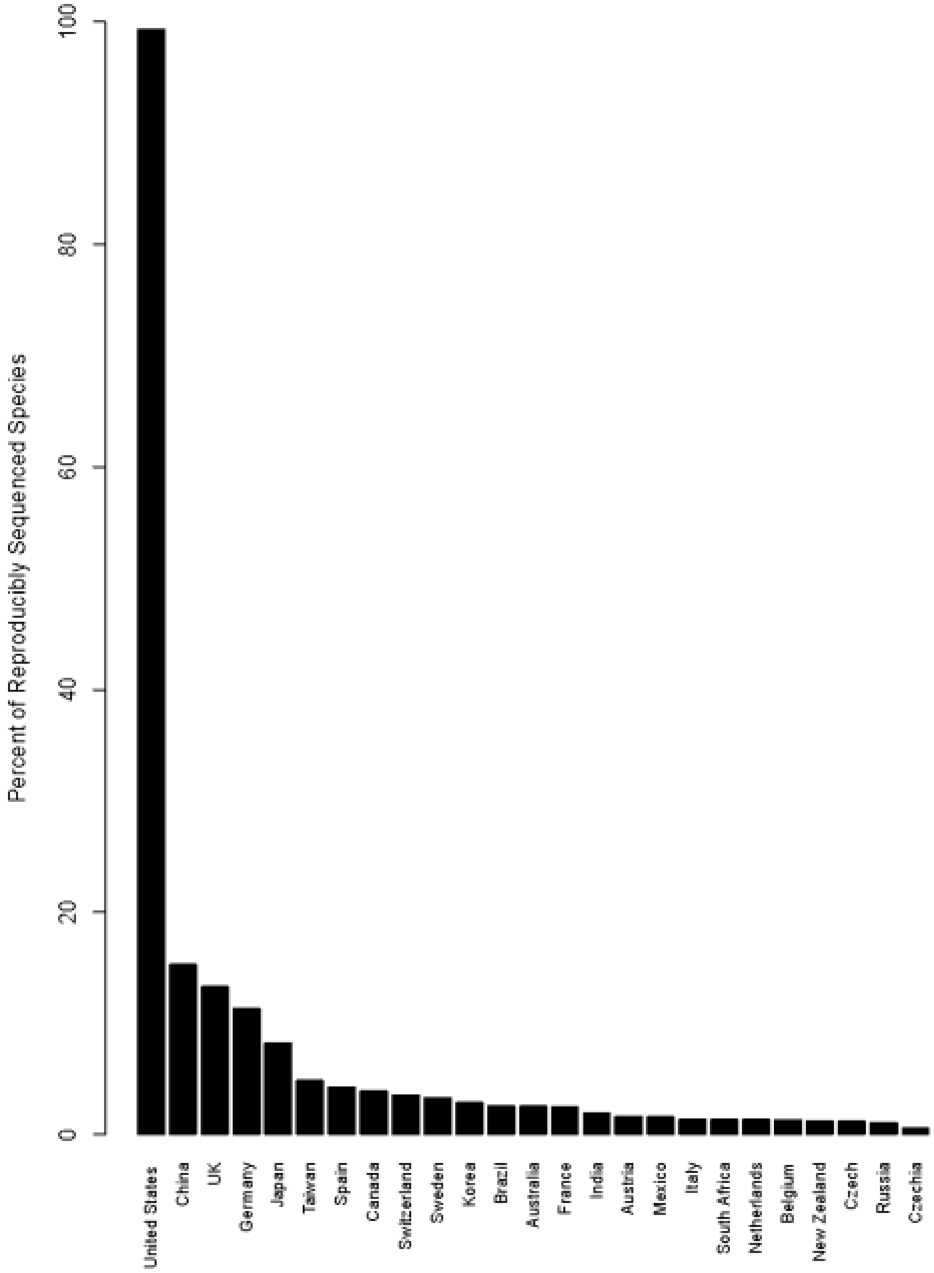


Figure 1.5 A bar plot showing the percent of reproducibly sequenced species (100% here means 18.45% of all vascular hotspot species) by country of first and/or last author.

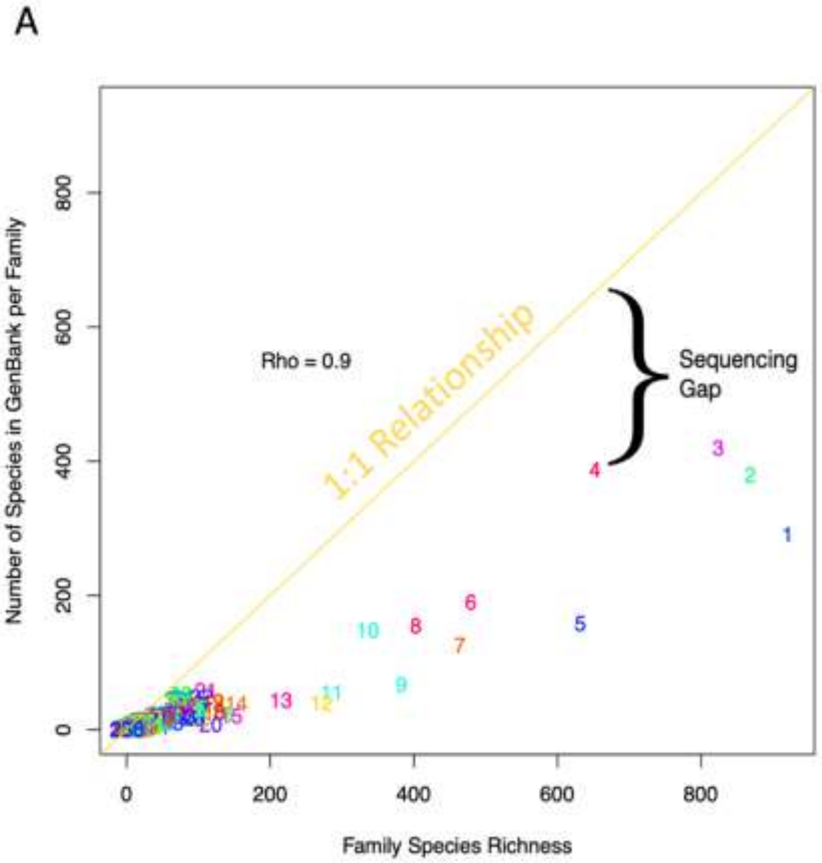


Figure 2.1A Graph of the relationship between species richness per family and the number of species in GenBank. A positive correlation was inferred for this relationship (Rho = 0.9). The numbers represent the species richness rank as inferred from GBIF in Table 1 (i.e. the row numbers); the colors of the numbers are provided simply to facilitate their visual discrimination, especially where values are clumped. The golden line represents a 1:1 relationship, i.e., all the species within a family have been sequenced

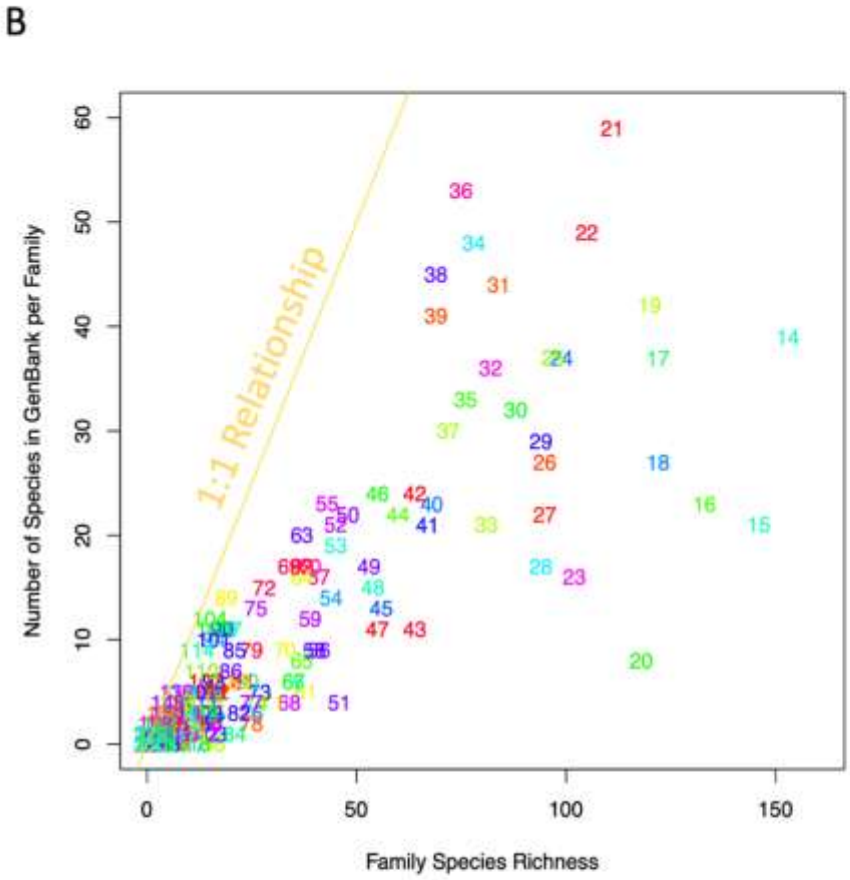


Figure 2.1B A zoomed-in version of the bottom left section of part A so that the number labels can be more easily visualized. Note that the axes are not proportioned the same in part B as they are in part A.

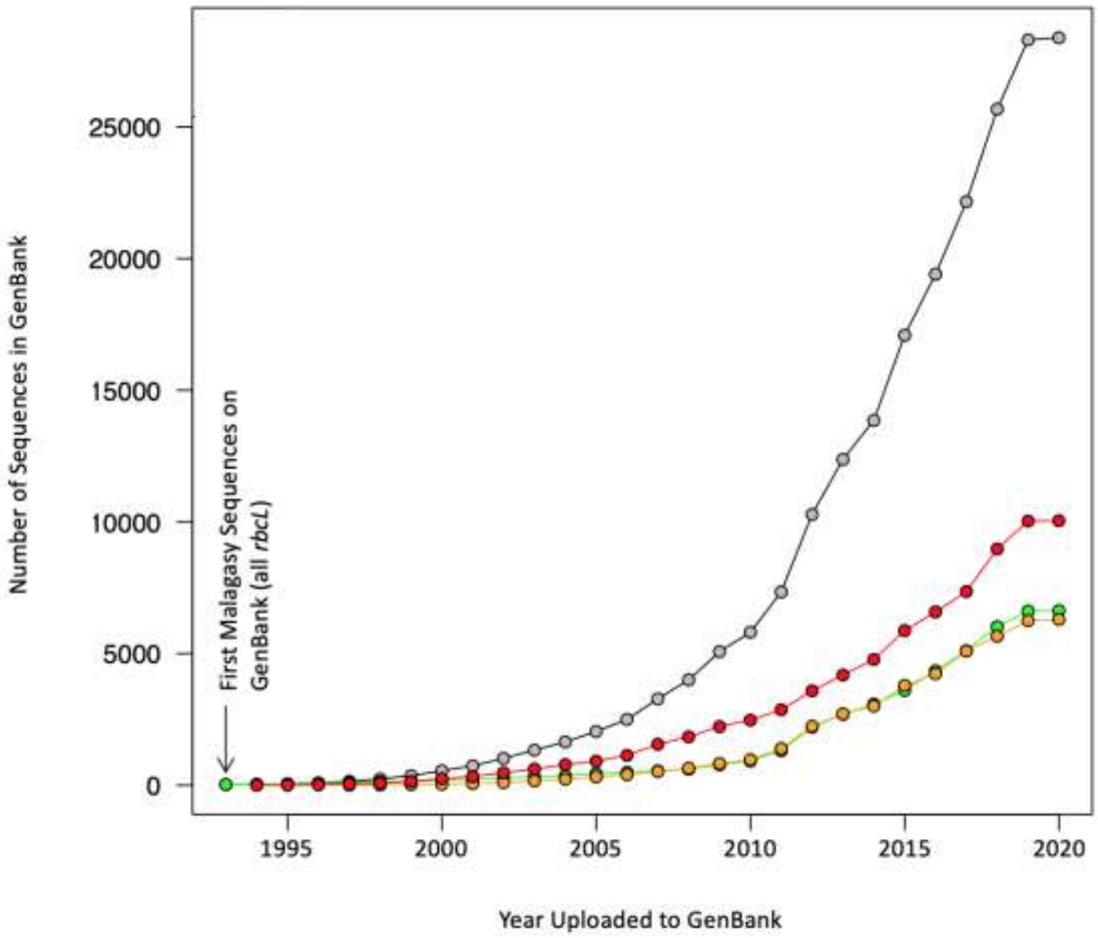


Figure 2.2. Graph of the number of Malagasy plant barcode sequences uploaded to GenBank by year. The total number of sequences is represented by the black line, *rbcL* by the green line, *matK* by the orange line, and ITS by the red line.

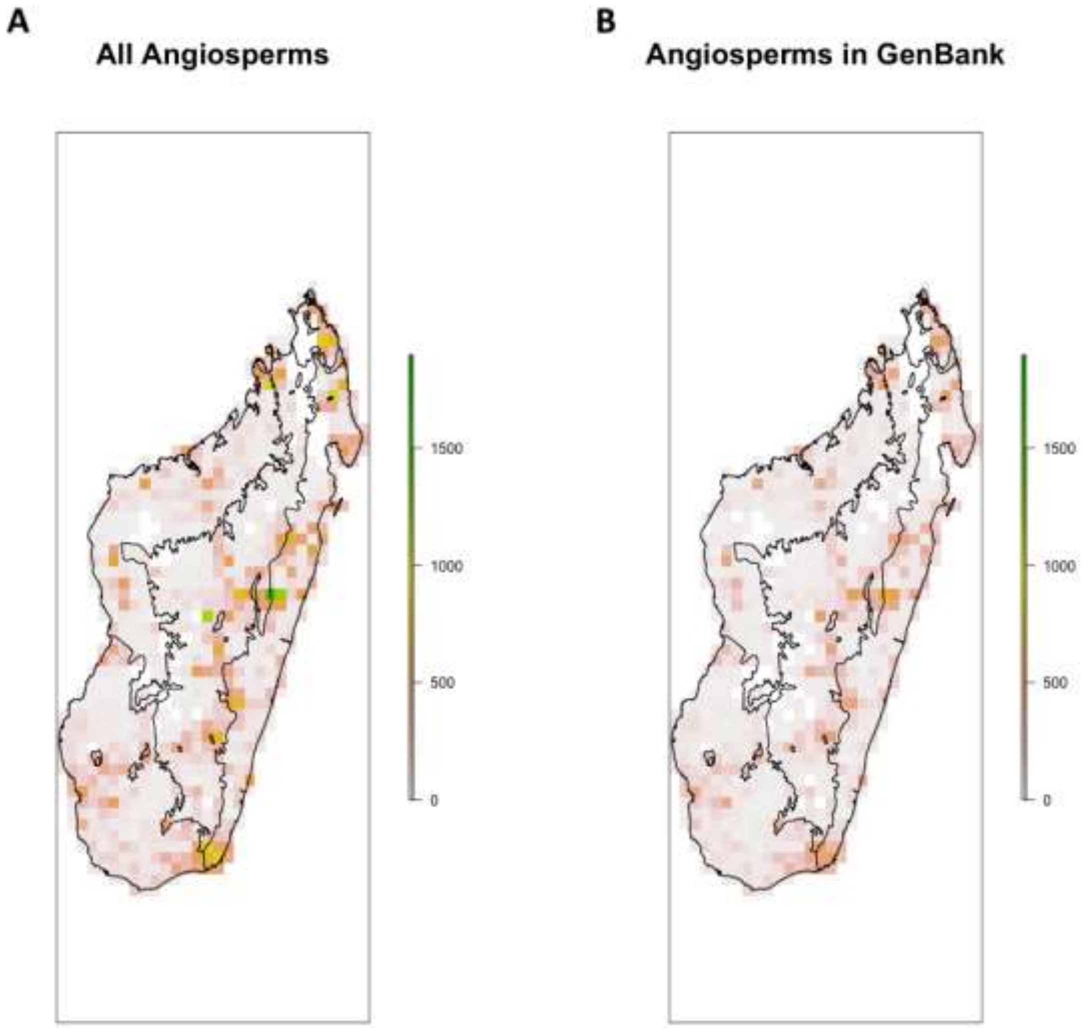


Figure 2.3. Two species richness maps showing (a) species richness for angiosperms in Madagascar and for (b) species richness of angiosperms with at least one DNA sequence in GenBank.

APPENDIX

Detailed Materials and Methods

Part 1

The analysis pipeline is divided into 3 parts, each associated with its own script. It can be run on any size data set. In part 1, the raw occurrence data were downloaded from the web, formatted, and uploaded to R. Next a list of species in each hotspot was generated by overlapping occurrence geographical coordinates with a shapefile of all biodiversity hotspots. Then the resultant species list was taxonomically curated. In part 2 the curated species list was used to mine GenBank to get accessions associated with each species. In part 3 the reproducibility of the studies that produced the accessions was assessed and three lists were inferred: the first (A) contained species with at least one accession in GenBank; the second (B) contained species with at least one accession in GenBank that was associated with a study our algorithm has determined to be reproducible; the third list (C) contained species represented only by accessions that had been submitted to GenBank within the last 5 years but which had not yet been published but had been submitted to GenBank by an author that had previously published accessions associated with at least one reproducible study contributing at least one species to list B. If lists B and C are combined, they provide an approximation of what B may be composed of in the future. World hotspot heatmaps representing the percent of species reproducibly sequenced (list B) were produced for all vascular plants and for each of the three plant lineages. A barcodewise analysis identified the most commonly used barcodes for list B. A cumulative curve illustrating the date of acceptance for the studies used to create list B was inferred.

Pre-Analysis Preparation

Analysis started with installing the ReproduciblePlants package from ‘wojahn/ReproduciblePlants’ on GitHub using the devtools (Wickham et al., 2020) package. Caffeine (Zhorn, 2017) was used to prevent the computer from going to sleep during the analyses because several functions took days to over a week to run. The desired output directory was set as a string in the object mainDirect so that it could be passed to the functions that require it as an argument. Next GBIF occurrence data were downloaded from the GBIF web portal for all tracheophytes with GPS coordinate metadata as a secondary requirement (GBIF, 2020). Bash (Fox, 1989) was used to index out the species, latitudinal coordinates, and longitudinal coordinates from the main file (the file is too large to be handled in R) and placed them into a new file. That file was then read into R (R Core Team, 2019) using the readr (Wickham et al., 2018) package. Next a shapefile containing all of the biodiversity hotspots’ geographical coordinates was downloaded from the Critical Ecosystem Partnership Fund (CEPS, 2016) and read into R using the rgdal (Bivand, 2019) package.

Overlapping Occurrence Data with a Shapefile of Known Biodiversity Hotspots

The purpose of this step was to determine which species have been recorded in each biodiversity hotspot. This was done by the function ReproduciblePlants::HotspotOverlappeR, which outputted a longform species and associated hotspot matrix. It used the packages sp (Pebesma et al., 2005) and maptools (Bivand et al., 2019) internally.

Taxonomical Curation

The purpose of taxonomical curation was to ensure that the species continuing through the pipeline were not synonyms and were validly published. This was done by the function ReproduciblePlants::CurateTaxomony which took the list of species derived from the occurrence data and curated them using the taxize (Chamberlain et al., 2013) and Taxonstand (Cayuela et al., 2019) packages. It returned a matrix whose first column contained the curated species names and whose second column contained the curated family names.

Part 2

Pre-existing GenBank Barcode Analysis

The purpose of performing pre-existing GenBank barcode analysis was to determine the quantity and identity of barcodes and genomes in GenBank for the curated list of species produced above. To do this, ReproduciblePlants::GenBankMineR queried GenBank for all of the curated species. This function used the rentrez (Winter, 2017) package internally. It searched GenBank for the CBOL (Hollingsworth et al., 2011) barcodes as well as plastid, mitochondrial, and nuclear genomes for each species, taking up to 100 accessions for each category. The files this function produced were not human-readable, so ReproduciblePlants::CleanGenBankOutput was used to create a more human-friendly version. It used the Biostrings (Pagès, 2019) package internally.

Part 3

PubMed Mining

The purpose of mining PubMed was to determine if the sequences mined from GenBank were from studies that were successfully published. A list of accessions was created from the output of ReproduciblePlants::MakeAccessionsVector by ReproduciblePlants::GenBankMineR. The list of accessions was run through ReproduciblePlants::PubMedQuerieR. This function checked whether each accession had an associated publication registered with PubMed and, if it did, it downloaded the author names, year of publication, year of acceptance, name of the journal, country of publication, and full abstract. It used the rentrez, XML (Lang, 2020) and RISmed (Kovalchik, 2017) packages internally.

Abstract Sentiments Analysis

The purpose of performing sentiments analysis on the abstracts of the studies mined from PubMed representing the sequences from GenBank was to sift out studies (and the species of which they were the sole representatives) that did not follow reproducible methodologies. The sentiments analysis was performed rather than a simple word search because some authors may not have directly stated that they used vouchers or other reproducible study methods in their abstracts. A list of keywords was compiled by a brainstorming session between JMAW and SB, as well as through a visual search of numerous metagenomics-oriented studies sieved out during the initial rounds of coding the sentiments analysis. The presence/absence of 344 keywords was inferred from each abstract for all of the unique papers representing species from studies registered in

PubMed using a customized sentiment analysis. This was done using ReproduciblePlants::KeywordsSentimentAnalyzeR. It used the sentimentr (Rinker, 2017) package internally. Each keyword was classified as either positive, neutral, or negative based on its likely impact on the reproducibility of the study in which abstract it occurred. Abstracts which matched nothing or which only matched neutral keywords were manually curated in Excel (Microsoft Corporation, 2020) and reuploaded to R. A Venn diagram visualizing the quantities and overlap of the positive, negative, and neutral lists was created using ReproduciblePlants::SentimentVenneR. This function used the eulerr (Larsson, 2020) package internally.

Restricting Species Not in PubMed by Date of Submission to GenBank

The purpose of restricting species not in PubMed by date of submission was to exclude sequences that are not likely to be published (i.e. are more than 5 years old) and the species that are represented by solely by them. This and the following step were done to try and ensure that species that may be in the ‘publication backlog” have representation in our analysis. GenBank publication dates for each of the accessions not represented by a publication in PubMed were ascertained using ReproduciblePlants::ProcessGenBankDates. The viability of each accession was determined by flagging any sequence older than 5 years as suspicious.

Restricting Species Not in PubMed by Shared Authorship with Passed PubMed Species

The purpose of restricting species not in PubMed by their authors was to exclude species that are only represented by accessions submitted by authors who have not submitted a reproducible publication as determined by our pipeline. Authorship for each of the accessions not represented by a publication in PubMed were ascertained using ReproduciblePlants::AuthorGetteR. This function used rentrez and XML internally. ReproduciblePlants::AuthorRestrictoR used the output of the above function to assess the viability of each accession, having flagged any sequence not sharing at least one author (last name and first initial or intitials) with a passed PubMed species as suspicious.

Making A, B, and C Species Lists

The purpose of creating three separate lists was to show the potential diversity and depth of sequencing efforts at different levels of reproducibility. the first (A) is a list of species with at least one accession in GenBank; the second (B) is a list of species with at least one accession in GenBank that is associated with a study our algorithm has determined to be reproducible (i.e. had positive sentiments only or positive and neutral sentiments only, or which contains the word voucher or any of its semantic equivalents regardless of its sentimentality); the third list (C) is a list containing species represented only by accessions that have been submitted to GenBank within the last 5 years but which have not yet been published but have been submitted to GenBank by an author that has published accessions associated with reproducible studies. If lists B and C are combined they provide an approximation of what B may be composed of in the future. ReproduciblePlants::FinalListsMakeR was used to compile the A, B, and C lists from the outputs matrices of AuthorRestrictoR, ProcessGenBankDates, the automatic and manually-curated matrices of KeywordsSentimentAnalyzeR, and GenBankMineR.

Lineage-Wise Analysis

The purpose of performing lineage-wise analyses of the A, B, and C lists was to determine whether or not all lineages were proportionately represented in each list. The number and percentage of passing species for the world, for angiosperms, for gymnosperms, and for ferns and their allies were calculated by ReproduciblePlants::LineagePercentsPassed for each A, B, and C list.

World Biodiversity Hotspot Maps

The purpose of creating world maps of the percentage of passing species for each of the three lists for all tracheophytes and the three major lineages was to allow for geographical patterns of sequencing effort to be easily visualized. World maps of the percentage of species passing were compiled for all tracheophytes, angiosperms, gymnosperms, and ferns and their allies for each alpha, beta, and gamma list by ReproduciblePlants::PercentPassedMapsMakeR. It used the ggplot2 (Wickham, 2016) package internally.

Barcode Analyses

The purpose of performing barcode analyses was to determine what barcodes are the most commonly used. The frequency of barcodes for species in the B list was calculated by ReproduciblePlants::MakeBarcodeTable.

Cumulative and Rate Curves for Date of Acceptance

Cumulative and rate curves illustrating the date of acceptance for the studies used to create list B was inferred by ReproduciblePlants::MakeCumulativeCurve.

Verifying that the Algorithm Actually Worked: Sentiments Analysis Efficacy Evaluation

To examine whether the sentiments analysis code in the algorithm was doing its job correctly, 50 studies were randomly sampled from the finished sentiments list and manually checked. 96% (48) of the abstracts were correctly sorted, with the remaining 4% (2) being incorrectly rejected because of their having contained words associated with metagenomics (i.e. studying DNA from environmental samples, Pace et al., 1986) in the background portions of their abstracts. The authors believe this error rate is tolerable because the negative words that disqualified the 2 good studies were heavily associated with the metagenomic studies analyzed for the initial compiling of the 344 keywords, and also because the errors resulted in an underestimation rather than overestimation (and overestimation could provide a false sense of completeness).

The sentiments analysis was performed rather than a simple word search because some authors may not directly state that they used vouchers or other reproducible methods in their abstracts. In fact, only 0.22% of the studies mentioned vouchers (or any of its semantic equivalents) at all in their abstracts. The sentiments analysis is also much quicker than a simple word search, taking less than half the time of the latter to complete, classifying the abstracts into categories that are easily interpretable by the algorithm (Figure 1).

Analysis of Author Countries

The addresses of the first and last authors of each publication underpinning species in list B at the time of publication were mined and the country indicated in the address was noted. The country(ies) of the authors of each study were then associated with their respective accessions and a pivot table was constructed. A bar plot was then constructed illustrating the pivot table results. Whether or not each country contained a biodiversity hotspot was established through web searching of current maps of them and their associated overseas territories/departments/states/kingdoms/associates/colonies.

Note on Parallelization

The ReproduciblePlants functions CurateTaxonomy, GenBankMineR, PubMedQuerieR, UnpublishedByAge, and AuthorGetteR were all run in parallel using the snow (Tierney et al., 2018) and doSNOW (Microsoft Corporation et al., 2020) packages using one less (7) than the total number of logical cores (8). Outputs were written either every 1,000 or 100 iterations and were then bound together into a finalized output file after the parallelized functions had completed.

Note on Machine Used to Perform Analyses and Location of Pipeline and ReproduciblePlants package

The analyses were run on a mid-2015 15-inch MacBook Pro retina with a 2.8 GHz Quad-Core Intel Core i7 and 6 GB 1600 MHz DDR3 with an Intel Iris Pro 1536 MB graphics card running macOS Catalina 10.15.3 (19D76). Overall the analyses took about 3 weeks (~504 hours) of analysis time to run (in reality the analyses were run piecemeal and took longer total than that to run, but if they had been run end-to-end it would have taken that long). The ReproduciblePlants package, the pipeline used, all of the input data and all of the output data can be found on wojahn/ReproduciblePlants on GitHub.