Molecular Ecology

WILEY

MOLECULAR ECOLOGY

MINI REVIEW

Global Availability of Plant DNA Barcodes as Genomic Resources to Support Basic and Policy-Relevant Biodiversity Research

Tyler R. Kartzinel[1,2] | Hannah K. Hoff[1,2] | Timothy J. Divoll[3] | Bethan L. Littleford-Colquhoun[1,2] | Heidi Anderson[4] | Mary K. Burak[1,2] | Maria L. Kuzmina[5] | Paul M. Musili[6] | Haldre Rogers[7] | Alejandra J. Troncoso[8,9] | Rebecca Y. Kartzinel[1,10]

[1]Department of Ecology, Evolution, & Organismal Biology, Brown University, Providence, Rhode Island, USA | [2]Institute at Brown for Environment and Society, Brown University, Providence, Rhode Island, USA | [3]Center for Computation and Visualization, Brown University, Providence, Rhode Island, USA | [4]Yellowstone Center for Resources, Yellowstone National Park, Mammoth Hot Springs, Wyoming, USA | [5]Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada | [6]Botany Department, National Museums of Kenya, Nairobi, Kenya | [7]Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, Virginia, USA | [8]Instituto de Ecología y Biodiversidad (IEB), Santiago, Chile | [9]Departamento de Biología, Universidad de La Serena, La Serena, Chile | [10]Brown University Herbarium, Brown University, Providence, Rhode Island, USA

Correspondence: Tyler R. Kartzinel (tyler_kartzinel@brown.edu)

Received: 21 November 2024 | **Revised:** 15 February 2025 | **Accepted:** 19 February 2025

Handling Editor: Joanna Freeland

Funding: This study was supported by Division of Environmental Biology (1930820, 2026294, 2046797), Office of Integrative Activities (2033823), National Park Service (P22AC00332, P23AC00378).

Keywords: benefits sharing | Convention on Biological Diversity | digital sequence information | DNA metabarcoding | museums | Nagoya Protocol

ABSTRACT

Genetic technologies such as DNA barcoding make it easier and less expensive to monitor biodiversity and its associated ecosystem services, particularly in biodiversity hotspots where traditional assessments are challenging. Successful use of these data-driven technologies, however, requires access to appropriate reference data. We reviewed the >373,584 reference plant DNA barcodes in public repositories and found that they cumulatively cover a remarkable quarter of the ~435,000 extant land plant species (Embryophyta). Nevertheless, coverage gaps in tropical biodiversity hotspots reflect well-documented biases in biodiversity science – most reference specimens originated in the Global North. Currently, at least 17% of plant families lack any reference barcode data whatsoever, affecting tropical and temperate regions alike. Investigators often emphasise the importance of marker choice and the need to ensure protocols are technically capable of detecting and identifying a broad range of taxa. Yet persistent geographic and taxonomic gaps in the reference datasets show that these protocols rely upon risk undermining all downstream applications of the strategy, ranging from basic biodiversity monitoring to policy-relevant objectives – such as the forensic authentication of materials in illegal trade. Future networks of investigators could work strategically to improve data coverage, which will be essential in global efforts to conserve biodiversity while advancing more fair and equitable access to benefits arising from genetic resources.

1 | Introduction

DNA barcoding is a strategy to identify species using short, standardised genomic sequences (Hebert et al. 2003). If all species have unique barcodes, then a reference library covering known taxa can enable accurate specimen identifications (Ratnasingham and Hebert 2007). Available DNA barcode data can help verify and refine species' identities (Hebert et al. 2003,

2004) or be leveraged to characterise complex mixtures of genetic material via 'DNA metabarcoding' (Pompanon et al. 2012). Decades of 'upstream' research have yielded a versatile technology that has rapidly transformed an array of 'downstream' applications in taxonomy and systematics (Hebert and Gregory 2005), ecology and evolution (Kress et al. 2015), biogeography (Bezeng et al. 2017), paleoecology (Williams et al. 2023), anthropology (Maixner et al. 2018), forensic science (Willette et al. 2017), parasitology and biomedical research (Ondrejicka et al. 2014; Reese et al. 2019) and evaluations of the diets, microbiomes and nutrition of diverse species (Kartzinel et al. 2015) – among others. Because DNA barcodes can be used to support global biodiversity assessments, the technology has also gained traction with policymakers aiming to benchmark progress toward global biodiversity targets under the Convention on Biological Diversity, the United Nations' Sustainable Development Goals and Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (REDD+) (Bush et al. 2017).

Biological collections underpin our knowledge of life on earth (Heberling and Isaac 2017; Meineke, Davis, et al. 2018). Physical specimens, painstakingly identified by experts, are irreplaceable for all DNA barcoding applications that require identifying unknown genetic material (Hebert et al. 2003). The completeness and reliability of biodiversity collections are thus inevitably reflected in the types of gaps, errors and biases associated with DNA barcode libraries (Pinheiro et al. 2019; Lendemer et al. 2020; Davis 2023). Quantifying shortcomings in these reference libraries – including well-documented geographic and taxonomic biases that reflect the historical priorities of commercial interests and well-resourced investigators – is needed to ensure the data are interpreted properly (Meyer et al. 2016; Daru et al. 2018; Cooper et al. 2019; Meineke and Daru 2021). Investigators should beware of the constraints imposed by limited data coverage because a lack of access to relevant reference data could generate imprecise and potentially misleading results.

The most effective DNA reference libraries provide data from accurately identified and geographically relevant specimens (Goldstein and DeSalle 2011; Kuzmina et al. 2017; Kolter and Gemeinholzer 2021). Unfortunately, progress developing extensive repositories of plant DNA barcodes has lagged despite the utility of these data for advancing basic and applied research priorities (Kress et al. 2005; Hollingsworth et al. 2009; Braukmann et al. 2017). Substantial efforts were made to create globally relevant barcode libraries for the chloroplast genes *rbc*L (552 base pairs) and *mat*K (~800 base pairs), since these markers were proposed as the 'standard' plant DNA barcodes by the Consortium for the Barcode of Life Working Group (Hollingsworth et al. 2009). However, the relatively long size of these barcodes in base pairs has precluded their use in some applications that require shorter and less-supported barcodes, such as in studies of degraded environmental DNA that frequently use the *trn*L(UAA) intron (often 500–600 base pairs), its shorter *trn*L-P6 fragment (often < 100 base pairs) or the nuclear ribosomal DNA fragments called internal transcribed spacers (ITS1 and ITS2) (Taberlet et al. 2007; Riaz et al. 2011; Ivanova et al. 2016; Deagle et al. 2019; Bansch et al. 2020). Consequently, the most taxonomically complete reference databases may not correspond to those that are in highest demand for many downstream applications in the real world.

If a researcher relies on a reference database that lacks a matching record for the taxon under investigation, they will have to settle for the next closest match (Pompanon et al. 2012). There is thus a need to jointly evaluate both the taxonomic and geographic coverage of public DNA barcodes. The geographic coverage of DNA barcodes can be defined as the extent to which the expert-verified specimens used to build a reference database reflect the distribution of biodiversity across the earth. Geographic coverage will be limited whenever sampled specimens and the species they represent include only a narrow subset of the relevant species diversity. To begin characterising and addressing contemporary geographic coverage of publicly available barcode data, we quantified spatial, taxonomic and marker-based coverage of the plant DNA barcodes that are in widest use today.

2 | Methods

2.1 | Building the Datasets

We began our review of publicly available plant DNA barcodes by downloading data from 'BOLD' (i.e., Barcode of Life Data Systems), the largest curated repository of DNA barcodes for plants and animals (Ratnasingham and Hebert 2007; Meiklejohn et al. 2019). From the 11 phyla of land plants (Embryophyta) available in BOLD v. 4, as of June 2024, we obtained all public sequence records that were (1) assigned to one of 660 plant families and (2) associated with any of four common DNA barcode markers: *rbc*L and *mat*K, the 'standard' plant DNA barcodes (Kress et al. 2005; Hollingsworth et al. 2009); *trn*L, the most widely used reference sequence for plant DNA metabarcoding (Taberlet et al. 2007; Pansu et al. 2022); and ITS, the nuclear marker that is most widely used for plant barcoding and phylogenetic studies (Alvarez and Wendel 2003; Kolter and Gemeinholzer 2021). This dataset included 373,584 sequences from 281,669 specimen records representing at least 102,887 species from 651 families (Dataset S1). About half of these sequences (199,591/373,584 = 53%) and a larger fraction of so-called 'specimens' (198,551/281,669 = 70%) originated from GenBank, which BOLD mines for data in addition to accepting direct submissions (Ratnasingham and Hebert 2007). Most sequences, regardless of origin, included a GenBank accession number (262,922/373,584 = 70%), indicating that these repositories overlap, albeit incompletely.

Whereas BOLD was developed to support and curate the growth of specimen-based barcode libraries, complementary data-mining methods are often used to build reference databases by extracting sequences and associated metadata from public repositories (Riaz et al. 2011; Meiklejohn et al. 2019). This strategy can produce much larger databases but may leave end-users lacking some relevant metadata and prone to taxonomic errors and/or imprecision compared to those using platforms such as BOLD (Vilgalys 2003; Meiklejohn et al. 2019). Data-mining methods often involve searching large public repositories for pre-defined primer sequences, which can be an impediment since standard quality controls require data producers to remove sequences corresponding to the primers that they used at the bench prior to archiving their data (Riaz et al. 2011). We therefore quantified how much a sequence-mining strategy might improve (or fail to improve) taxonomic coverage compared to

a more traditional DNA barcoding strategy. Our comparison focused on the *trn*L(UAA) intron (often 500–600 base pairs) and a shorter, internal fragment of this marker called *trn*L-P6 (often < 100 base pairs) because the former is the least-barcoded marker in our review of the BOLD dataset – and thus there is the greatest possible room for improvement – while the latter has become one of the most widely used in plant DNA metabarcoding applications (Pompanon et al. 2012; Deagle et al. 2019). We did this by using '*in silico* PCR' to search for and extract the *trn*L-P6 barcode fragment from public plant records in the European Nucleotide Archive (ENA) release 143 ($N = 4{,}145{,}939$ accessions). We searched these records for the *trn*L-P6 primers g/h, allowing a maximum of three mismatches to each primer and retaining all sequences that were found between these primer-binding sites at an appropriate size range of 8–300 bp (Taberlet et al. 2007; Boyer et al. 2016). The resulting *in silico* database comprised reference sequences that are internal to the full-length *trn*L barcode and thus inevitably, on average, provide less taxonomic resolution. However, if the data-mining strategy yields substantially more reference data from the geographic localities or taxonomic groups that are most relevant to a study's objectives, then it could still provide significant advantages.

2.2 | Geographic Coverage

To evaluate the geographic coverage of plant barcodes, we analysed both country-level distributions and coordinate positions in the BOLD metadata. Country was listed for 150,347/281,669 (53%) of records, and coordinates were available for 111,111 (39%), though coordinate precision varied and was occasionally listed as country centroids. We visualised the global distribution of source specimen localities with respect to mean annual temperature and mean annual precipitation using WorldClim data with 1° resolution (Fick and Hijmans 2017). To identify geographic patterns in the global environments that have been under sampled, we plotted locations characterised by climates that are represented by <5% of specimen source localities in the database.

2.3 | Taxonomic Coverage

Evaluating taxonomic coverage in plant biodiversity data is complicated by the need to address data quality, differences among taxonomic authorities in how species are treated and the interoperability of databases (Thomas 2009). We approached these challenges by obtaining data on the relative size of plant families according to the Integrated Taxonomic Information System (ITIS) database, accessed via *taxize* (Chamberlain et al. 2018), which has 98% overlap with the Global Biodiversity Information Facility (GBIF) (National Museum of Natural History 2023). To correlate barcode availability with family size, we matched family names between ITIS and BOLD, returning 101,674 plant species in 730 ITIS-accepted land plant family names (Dataset S2). Plant family data from BOLD and ITIS overlapped broadly, with 609/730 (83%) of ITIS family names matching a public BOLD record (Dataset S2). A subset of ITIS families matched no BOLD records, partly due to differences in the extent to which each database addresses recent taxonomic revisions. Similarly, plant family names in the sequence-mining dataset largely overlapped with ITIS (611/730, 84%). To evaluate global trends in taxonomic coverage, we used a log–log correlation between the count of accepted species names in a family according to ITIS (predictor) and the number of specimens, sequences and named species in the barcode data (response). We tested hypotheses concerning potential sampling biases based on family size, considering whether the barcode coverage of each family reflected (i) an approximately 1:1 correlation with the relative size of each family, (ii) bias toward large families or (iii) bias towards small families (or none of the above).

2.4 | Case Study in Geographic Coverage of Site-Based Reference Data

As with all historical biodiversity collections, the growth of DNA barcode libraries is often driven by independent teams with motivations connected to their specific research foci, taxonomic groups or geographic locations. It is not uncommon for species to be omitted because collecting fertile voucher specimens can be phenologically challenging, labour-intensive or not related to the interests of the investigator. Yet each new barcode has the potential to provide coverage of taxa beyond specific sampling localities, and thus site-based barcode projects have the potential to meaningfully contribute to the growth of global barcode coverage. To illustrate the challenges and opportunities associated with the ongoing growth of a representative site-based barcode project, we present an example from Yellowstone National Park (Littleford-Colquhoun et al. 2024). The first public release of data from the Yellowstone Barcode Project included 319 of the 1386 plant species known to occur in Yellowstone National Park (Whipple 2001). We evaluated the contribution of this collection to global geographic coverage using data from GBIF (GBIF 2024). We obtained coordinate data from a total of 18,588,886 GBIF records corresponding to species in the barcode dataset and mapped them into globally distributed hexagons with a 69-km average edge length using Uber's Hexagonal Hierarchical Spatial Index (Kuethe 2022).

2.5 | Code Book

To empower investigators to conduct more comprehensive evaluations of plant barcode taxonomic and geographic coverage as relevant to their study objectives, we published a Code Book organised in chapters that correspond to each section of our review (see Data Accessibility Statement). This Code Book can be customised to support additional case studies or comparative evaluations of barcode coverage that may be of interest to data producers and end users alike.

3 | Results

3.1 | Globally Available Plant DNA Barcode Data

Our review of global plant barcode diversity in BOLD revealed high taxonomic precision, with 97% of specimens identified to species. Most specimens were associated with *rbc*L ($N = 47\%$), *mat*K (45%) and/or ITS (38%), while *trn*L was only sparsely represented (< 2%; Figure 1). The total of 102,887 plant species

covered by at least one marker in BOLD thus remains far from approaching the full diversity of plant species, but nevertheless provides coverage for a remarkable ~24% of extant plants.

3.2 | Geographic Coverage

The DNA barcode records in BOLD originated in 189 countries and Antarctica, though regional sampling was highly skewed (Figure 1A and Figure S1). Most specimens with coordinate data originated north of the Tropic of Cancer ($N = 61{,}865$; 56%), fewer were from the tropics ($N = 37{,}229$; 34%) and very few were from south of the Tropic of Capricorn ($N = 12{,}018$; 11%); these percentages better match the latitudinal distribution of landmass than plant species richness, as the tropics harbour 80% of biodiversity but comprise only 36% of global landmass. Specimen collection localities were concentrated in relatively warm and dry regions of the temperate zone (Figure 1A,D). Exacerbating disparities in collection localities, marker coverage was uneven: latitudinal coverage for *rbc*L and *mat*K was broad, but *trn*L and ITS sampling were far more localised (Figure 1B,C). Indeed, 75% of *trn*L barcodes in BOLD originated from five site-based project containers, including three developed to provide local reference libraries for dietary DNA metabarcoding at Mpala Research Centre, Kenya (project code 'UHURU', 30%) (Kartzinel et al. 2015), Gorongosa National Park, Mozambique (PNG, 9%) (Pansu et al. 2019) and Yellowstone National Park, United States (YNPBP; 9%) (Littleford-Colquhoun et al. 2024); the other two focused on high-latitude bryophytes (BRYCA, 21%; TMBRY, 6%).

3.3 | Taxonomic Coverage

We found strong, positive correlations between the relative size of plant families in the ITIS database and the corresponding amount of data available in BOLD (Figure S2A,B). The slope of each correlation was less than 1:1 on a log–log scale, indicating that some small families have been barcoded extensively, while some larger families may be undersampled relative to their diversity.

Posing a challenge, many ITIS-accepted family names did not match an entry in BOLD for any of the markers we surveyed ($N = 121/730$ families: 17%). These 'no-barcode families' included 823 species in the ITIS database (range = 1–203 species per family; Dataset S2). The majority of no-barcode families are bryophytes and ferns (~70%), which are taxonomically challenging

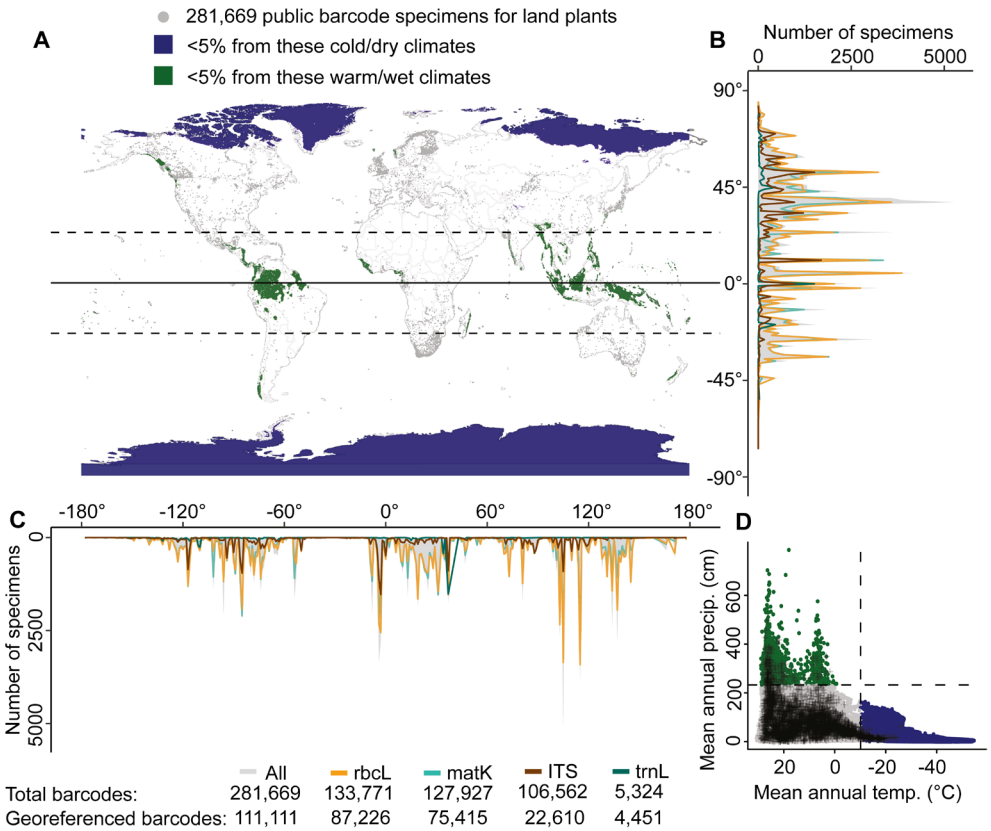

FIGURE 1 | Global geography of plant DNA barcodes. (A) Map shows georeferenced BOLD specimens (grey points) and the climate extremes under which <5% of specimens originated (blue = cold/dry; green = warm/wet). The (B) latitudinal and (C) longitudinal distributions of all barcodes are broadly matched for the three most widely used barcodes, *rbc*L, *mat*K, ITS; there is a strongly trimodal distribution of *trn*L barcodes. For each marker, the legend notes the total number of barcodes available in BOLD and the subset for which coordinate data are available. (D) Bioclimatic coverage of specimens (dark + signs) reveals a climatic envelope in which most specimens originated from relatively warm/dry environments (grey; dashed lines = 95 percentiles) and few originated from warm/wet (green) or cold/dry (blue). The climate envelope was established using 100,000 random localities across the terrestrial earth surface.

groups and the subjects of recent revisions (Christenhusz and Chase 2014; Schuettpelz et al. 2016; Li et al. 2024). Many of these groups were also monotypic ($N = 1$ species per family; 56%) or nearly so ($N < 5$ species per family; 77%). Of the 68 bryophyte families included on the no-barcode list, 23 were represented by only a single genus. Most of these gaps thus do not appear to involve many species, but some may. The most species-rich no-barcode family in this dataset was Hydrophyllaceae ($N = 203$ accepted names), which has also been affected by significant taxonomic reclassifications based on progress resolving the phylogeny of this and related groups (e.g., Boraginaceae, Cordiaceae, Ehretiaceae, Heliotropiaceae) (Luebert et al. 2016). As such, some no-barcode families could be artefacts arising from taxonomic reorganisations, highlighting challenges associated with the interoperability of databases.

While the no-barcode families that were completely missing from public BOLD data tended to present narrow targets for biodiversity research – comprising few known species, difficult taxonomy and/or recent taxonomic rearrangements – we found the coverage of plant families varied extensively across barcode markers. On a marker-by-marker basis, the fraction of no-barcode families increased from *rbc*L (129/370 families: 18%) to *mat*K (251/730: 34%) and to ITS (261/730: 36%) before jumping dramatically to include most plant families for *trn*L (518/730: 71%; Dataset S2). Said another way, only 29% of plant families can be positively identified using public *trn*L sequences in BOLD.

3.4 | Comparing Taxonomy-Driven Versus Sequence-Mining Coverage

Mining the ENA database for *trn*L-P6 sequences yielded 157,020 references from at least 666 family names reported for this one locus in this one database (Dataset S3) versus 651 family names reported inclusively across all four loci we considered in the BOLD dataset (Dataset S1). When matched against the ITIS-accepted plant family names, far fewer no-barcode families remained in this library of 'short' *trn*L-P6 sequences (119/730 = 16%) versus the 518 no-barcode families represented in the 'full-length' *trn*L dataset from BOLD (71%). This strategy thus provided ~30-fold more sequence records and coverage for ~3-fold more plant families, albeit with less consistently complete and curated metadata compared to the full-length *trn*L dataset from BOLD. There was again a strong, positive correlation between the size of a plant family and the numbers of sequences and species represented in the barcode data (Figure S2C,D). Plant family size is thus a generally good predictor of the data volume that may be obtained through either strategy, although both approaches may suffer from the undersampling of larger families.

3.5 | Case Study in Geographic Coverage of Site-Based Reference Data

Our example of a site-based DNA barcode project focused on the flora of Yellowstone National Park and comprised 319 species. It included a specimen of lodgepole pine – the most widespread conifer in western North America – which is an example of a specimen record that provided geographic coverage across North America in addition to Eurasia and Australasia (Figure 2A). Yet this dataset did not include big sagebrush, a common plant in Yellowstone's sagebrush-steppe, so local coverage of this taxon would require access to public data generated elsewhere in the region (Figure 2B). This contrast provides a clear example of how species' overlapping geographic ranges provide complementary coverage that can cumulatively enhance the completeness of taxonomic coverage within and beyond a study site. The potential for other researchers to benefit from accessing

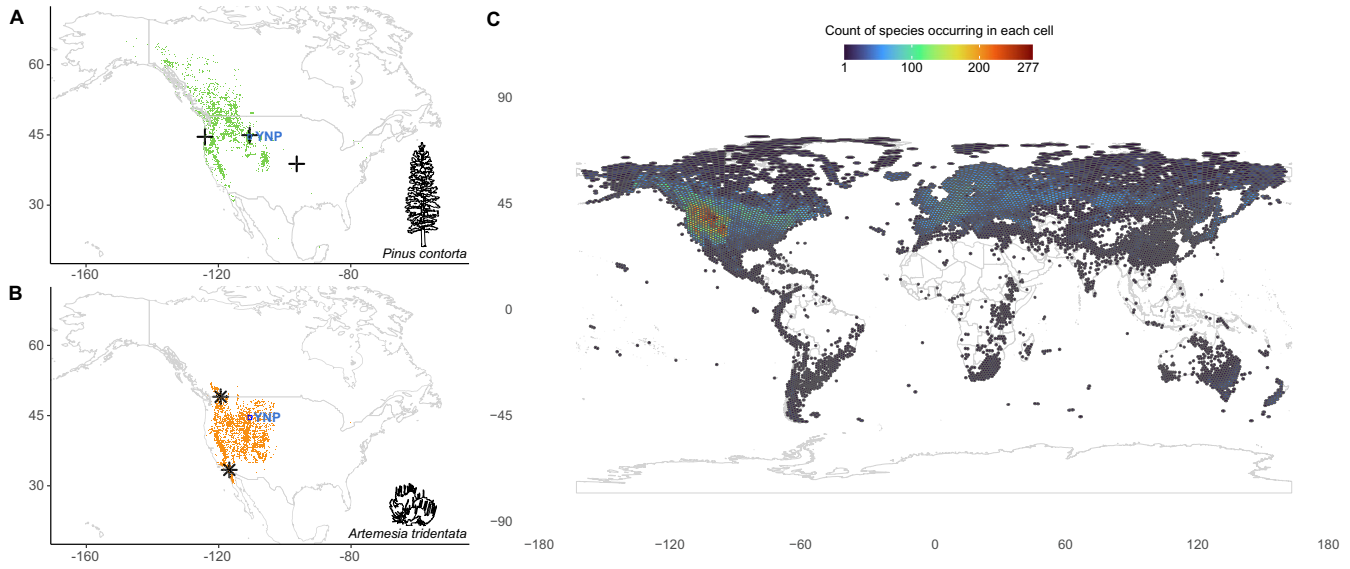

FIGURE 2 | Geographic coverage of site-based barcode data from Yellowstone National Park. Continent-scale coverage of (A) lodgepole pine (+ sign) and (B) big sagebrush DNA barcodes (* sign), with green or orange colouring indicating the presence of the taxon within a 10-min grid cell using GBIF. Lodgepole pine barcodes were available from specimens both inside and outside the park ('YNP'); big sagebrush was not included in the initial release of data from Yellowstone but available from other public data. (C) Case study of 319 barcoded species from the Yellowstone dataset matched 311 species' distribution records in GBIF, providing geographic coverage concentrated in the Rocky Mountains and extending across the globe.

the public data provided by the Yellowstone Barcode Project is clearly concentrated in the Rocky Mountains, but the overall geographic footprint of the dataset spans continents because it includes some species with very broad distributions (Figure 2C). The geographic reach of these data spanned a mean distance from Yellowstone of 8,493 km per species (range = 1–19,489 km; Figure 2).

4 | Discussion

We quantified progress toward a comprehensive plant barcode of life, focusing on the completeness of geographic and taxonomic coverage. We found the reference data in BOLD cover a remarkable quarter of the estimated 435,000 extant land plant species, about 40% of which are exceedingly rare (Enquist et al. 2019). Not unexpectedly, sampling biases leave striking gaps in the global coverage of available data – especially in, but not limited to, tropical biodiversity hotspots. Today, at least 17% of plant families appear to lack any public barcode data, and the largest plant families tend to be under-sampled relative to the numbers of species they contain. Marker coverage was also unequal, so marker choice can exacerbate other shortcomings in the data – nearly three-quarters of plant families were not included in the full-length *trn*L dataset (71%). We have thus clearly reached a major milestone in our progress with the plant barcode of life agenda, but persistent coverage gaps warrant attention. Failing to resolve them has the potential to undermine many common end-uses of the data.

4.1 | Equitable Access to Genomic Resources

Considerable attention has been given to the importance of ensuring that DNA barcoding assays are able to amplify and thus detect the full array of taxa under investigation (Hollingsworth et al. 2009; Riaz et al. 2011), but comparatively little attention has been given to geographic coverage. Geographic biases in the availability of reference DNA resources reflect persistent inequalities that pose barriers toward landmark international agreements such as the Convention on Biological Diversity and the Nagoya Protocol, which calls for 'access to genetic resources and the fair and equitable sharing of benefits arising from their utilization' (United Nations Environment Programme 1992; Paton and Lughadha 2011).

Our review highlights how tropical countries harbour the greatest concentrations of biodiversity, yet this diversity is badly underrepresented in DNA barcode databases: the places that most need to monitor environmental change because they harbour the most biodiversity and bear the greatest impacts of global change are those that cannot readily access the reference data that facilitate biodiversity monitoring. Gaps in barcode coverage reflect long-term disparities in the resources available to collect specimens, identify taxa and sequence DNA. Governmental policies developed since the Nagoya Protocol emerged as a legal framework under the Convention on Biological Diversity have the potential to both help and hinder efforts to improve global coverage of DNA barcodes (Watanabe 2017; Ambler et al. 2021; Colella et al. 2023). Unless proactive measures are taken and prove effective at enhancing the use of biodiversity data for conservation, the regions that are now underrepresented in public databases are liable to remain at a disadvantage when it comes to scientific research on biodiversity, the monitoring of endangered species, the detection of invasive species, the development of effective biodiversity policies and the forensic capabilities available to law enforcement (Moritz and Cicero 2004; Valentini et al. 2009).

Whether and how the international community might catalyse better coverage of DNA barcodes available for basic research and conservation is a timely question. The 2024 Conference of the Parties (COP16) to the Convention on Biological Diversity in Cali, Colombia, launched the 'Cali Fund' as a mechanism to share benefits from the use of Digital Sequence Information (DSI) (Conference of the Parties to the Convention on Biological Diversity 2024). Under the agreement, large companies and other major entities that benefit from accessing genomic resources, such as pharmaceutical or biotech companies, "should contribute" a percentage of their profits or revenues to the Cali Fund and thus ensure the resulting benefits can be shared with developing countries, Indigenous Peoples and local communities in the service of nature. The agreement includes exemptions for academic and public research, recognising important distinctions between commercial and non-commercial uses of sequence data, and it was established on the basis that the fund must be operated in ways that are "consistent with open access to data" (Conference of the Parties to the Convention on Biological Diversity 2024). As DNA barcode data are often shared and accessed in the open science tradition via public databases (Centre for Biodiversity Genomics University of Guelph 2021; von Wettberg and Khoury 2022), strategic coordination to enhance equitable access to DNA barcodes would be consistent with the requirement of the Cali Fund to "support the realization of the objectives of the Convention in developing countries" including with "scientific research on biodiversity" (Conference of the Parties to the Convention on Biological Diversity 2024). In this spirit, we now consider strategies that might improve the usefulness of plant DNA barcode libraries for basic research and conservation.

4.2 | No-Barcode Families

About 17% of land plant families lack public barcodes altogether, and the most common DNA barcode markers vary markedly in their coverage of taxa, which amplifies the challenge of addressing taxonomic uncertainties, omissions, and the interoperability of databases (Thomas 2009; Kolter and Gemeinholzer 2021). The prevalence of no-barcode families in both BOLD and our sequence-mining database is a problem because it is impossible to identify a barcode sequence as part of a family when that family is not included in the reference set. The best we can do in these situations is recognise a non-identification scenario and place an unidentified sequence within a higher taxonomic category (e.g., Order or higher); though more often, the algorithms used by search engines will attempt identifications based on the most similar sequences in a reference set, and the results will inevitably be wrong (Brower 2006). Therefore, obtaining the first barcodes for each family – across all relevant markers – is critical for improving accuracy. Examples include a small family of flowering plants endemic to Chile known as 'bridal wreaths'

(Francoaceae) and a species-rich family of pantropical liverworts (Balantiopsidaceae). Herbaria that host specimens from these families – especially type specimens (Renner et al. 2023) – could make important contributions to global barcode coverage. Enhancing sequence coverage of no-barcode taxa could become a thematic focus for consortia of herbaria, sequencing facilities and funders aiming to remedy rather than reinforce inequities in biodiversity science, noting that many smaller institutions host historically and taxonomically important collections but may require access to a broader network of funders and trusted collaborators if they are going to generate data for the greater good (Daru et al. 2018; Meineke, Davies, et al. 2018; Meineke, Davis, et al. 2018; Pinheiro et al. 2019).

4.3 | Site-Based Data

Investigators that lack access to a local DNA barcode library often rely on data mined from public repositories and therefore risk reporting errant species identifications. Knowledge of the extent to which site-based collections provide complementary coverage can thus help guide the growth and appropriate uses of reference data (Figure 2). As illustrated by the site-based project example at Yellowstone, some plant species have broad geographic distributions and thus field botanists should not have to survey every habitat to attain far better coverage than exists today. Accounting for these kinds of overlapping footprints between localities of substantial research and conservation value could enable more strategic coordination, minimise costs and maximise the reach of benefits beyond each study area. For example, the publication of a plant DNA barcode library in Kenya included data for only 460 species, yet increased barcode coverage for the plants of Africa by ~10%, helping improve precision in continent-scale analyses (Gill et al. 2019; Pansu et al. 2022). Especially in the wet tropics that harbour the greatest concentration of plant diversity, site-based research is likely to yield both improvements in data coverage and better recognition of hitherto unidentified species.

5 | Conclusions

All DNA-based identifications are hypotheses. The strength of evidence available to evaluate them is inextricably linked to the quality of reference data, so shortcomings in the coverage of biodiversity collections needed to support these data-driven technologies can limit their application. Failing to overcome historic disparities in the means to generate reference DNA barcodes risks exacerbating future inequities in the benefits that nations can derive from applying these resources in the service of biodiversity research and conservation. A major step toward improving coverage might be to guide future barcode data growth by leveraging geographic information associated with ongoing herbarium digitisation projects (Meyer et al. 2016) or site-based research networks, such as ForestGEO (Anderson-Teixeira et al. 2015), NutNet (Borer and Stevens 2022) and NEON (Keller et al. 2008). When it comes to prioritisation, our results show that species from big families with broad geographic ranges could generally accelerate progress because they tend to be under-sampled relative to their diversity (Figure 2). But this would not obviate the concurrent need to improve coverage of the rare, endemic and otherwise overlooked taxa that have historically represented narrow targets for biodiversity research (Figure S2). Complementary means of spurring progress could increasingly leverage emerging long-read or high-throughput sequencing technologies to overcome historical limitations arising from the early reliance on Sanger sequencing platforms for DNA barcoding (Hebert et al. 2025). It would further diminish constraints on end-users if whole-chloroplast genomes or multiplexed barcode markers were more routinely sequenced and archived in public repositories (Deagle et al. 2014; Littleford-Colquhoun and Kartzinel 2024). A concern is that the erosion of institutional support for biodiversity training and infrastructure could lead to a collapse in essential resources required for emerging data-driven technologies (Davis 2023).

Author Contributions

T.R.K. and R.Y.K. conceived the study. T.R.K., H.K.H., T.J.D. and B.L.L.-C. analysed data. T.R.K. wrote the first draft of the manuscript, and all authors contributed to the interpretation of patterns and writing of the final manuscript.

Acknowledgements

We thank all who contribute to the growth of publicly available, expert-verified plant DNA barcodes. We acknowledge NSF DEB-2046797, DEB-2026294, DEB-1930820 and OIA-2033823, as well as the National Park Service Cooperative Research and Training Program P22AC00332 and P23AC00378.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data tables required to complete the analyses are included in Datasets S1–S4 and the code for analyses is detailed in a published Quarto Code Book (https://trklab-metabarcoding.github.io/MolEco-MEC-24-1288/). Source code is available in a corresponding GitHub repository (https://github.com/trklab-metabarcoding/MolEco-MEC-24-1288).

Benefits Sharing Statement

This research presents a summary of public plant DNA barcodes as a prime example of genomic resources available for non-commercial benefit sharing. It highlights opportunities to enhance benefitsharing from access to these resources.

References

Alvarez, I., and J. F. Wendel. 2003. "Ribosomal ITS Sequences and Plant Phylogenetic Inference." *Molecular Phylogenetics and Evolution* 29: 417–434.

Ambler, J., A. A. Diallo, P. K. Dearden, P. Wilcox, M. Hudson, and N. Tiffin. 2021. "Including Digital Sequence Data in the Nagoya Protocol Can Promote Data Sharing." *Trends in Biotechnology* 39: 116–125.

Anderson-Teixeira, K. J., S. J. Davies, A. C. Bennett, et al. 2015. "CTFS-ForestGEO: A Worldwide Network Monitoring Forests in an Era of Global Change." *Global Change Biology* 21: 528–549.

Bansch, S., T. Tscharntke, R. Wunschiers, et al. 2020. "Using ITS2 Metabarcoding and Microscopy to Analyse Shifts in Pollen Diets of Honey Bees and Bumble Bees Along a Mass-Flowering Crop Gradient." *Molecular Ecology* 29: 5003–5018.

Bezeng, B. S., T. J. Davies, B. H. Daru, et al. 2017. "Ten Years of Barcoding at the African Centre for DNA Barcoding." *Genome* 60: 629–638.

Borer, E. T., and C. J. Stevens. 2022. "Nitrogen Deposition and Climate: An Integrated Synthesis." *Trends in Ecology & Evolution* 37: 541–552.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac. 2016. "Obitools: A Unix-Inspired Software Package for DNA Metabarcoding." *Molecular Ecology Resources* 16: 176–182.

Braukmann, T. W., M. L. Kuzmina, J. Sills, E. V. Zakharov, and P. D. N. Hebert. 2017. "Testing the Efficacy of DNA Barcodes for Identifying the Vascular Plants of Canada." *PLoS One* 12: e0169515.

Brower, A. V. Z. 2006. "Problems With DNA Barcodes for Species Delimitation: 'Ten Species' of Reassessed (Lepidoptera: Hesperiidae)." *Systematics and Biodiversity* 4: 127–132.

Bush, A., R. Sollmann, A. Wilting, et al. 2017. "Connecting Earth Observation to High-Throughput Biodiversity Data." *Nature Ecology & Evolution* 1: 0176.

Centre for Biodiversity Genomics University of Guelph. 2021. "The Global Taxonomy Initiative 2020: A Step-By-Step Guide for DNA Barcoding." Secretariat of the Convention on Biological Diversity, Montreal, Canada.

Chamberlain, S., E. Szoecs, Z. Foster, et al. 2018. "taxize: Taxonomic Information From Around the Web."

Christenhusz, M. J. M., and M. Chase. 2014. "Trends and Concepts in Fern Classification." *Annals of Botany* 113: 571–594.

Colella, J. P., L. Silvestri, G. Súzan, M. Weksler, J. A. Cook, and E. P. Lessa. 2023. "Engaging With the Nagoya Protocol on Access and Benefit-Sharing: Recommendations for Noncommercial Biodiversity Researchers." *Journal of Mammalogy* 104: 430–443.

Conference of the Parties to the Convention on Biological Diversity. 2024. "Digital Sequence Information on Genetic Resources: Draft Decision Submitted by the President." Sixteenth Meeting: CBD/COP/16/L.32/Rev.11.

Cooper, N., A. L. Bond, J. L. Davis, R. Portela Miguez, L. Tomsett, and K. M. Helgen. 2019. "Sex Biases in Bird and Mammal Natural History Collections." *Proceedings of the Biological Sciences* 286: 20192025.

Daru, B. H., D. S. Park, R. B. Primack, et al. 2018. "Widespread Sampling Biases in Herbaria Revealed From Large-Scale Digitization." *New Phytologist* 217: 939–955.

Davis, C. C. 2023. "The Herbarium of the Future." *Trends in Ecology & Evolution* 38: 412–423.

Deagle, B. E., S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet. 2014. "DNA Metabarcoding and the Cytochrome c Oxidase Subunit I Marker: Not a Perfect Match." *Biology Letters* 10: 20140562.

Deagle, B. E., A. C. Thomas, J. C. McInnes, et al. 2019. "Counting With DNA in Metabarcoding Studies: How Should We Convert Sequence Reads to Dietary Data?" *Molecular Ecology* 28: 391–406.

Enquist, B. J., X. Feng, B. Boyle, et al. 2019. "The Commonness of Rarity: Global and Future Distribution of Rarity Across Land Plants." *Science Advances* 5: eaaz0414.

Fick, S. E., and R. J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37: 4302–4315.

GBIF. 2024. "What is GBIF?" *in* T. G. B. I. Facility, editor.

Gill, B. A., P. M. Musili, S. Kurukura, et al. 2019. "Plant DNA-Barcode Library and Community Phylogeny for a Semi-Arid East African Savanna." *Molecular Ecology Resources* 19: 838–846.

Goldstein, P. Z., and R. DeSalle. 2011. "Integrating DNA Barcode Data and Taxonomic Practice: Determination, Discovery, and Description." *BioEssays* 33: 135–147.

Heberling, J. M., and B. L. Isaac. 2017. "Herbarium Specimens as Exaptations: New Uses for Old Collections." *American Journal of Botany* 104: 963–965.

Hebert, P. D., and T. R. Gregory. 2005. "The Promise of DNA Barcoding for Taxonomy." *Systematic Biology* 54: 852–859.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. DeWaard. 2003. "Biological Identifications Through DNA Barcodes." *Proceedings of the Royal Society B: Biological Sciences* 270: 313–321.

Hebert, P. D. N., R. Floyd, S. Jafarpour, and S. W. J. Prosser. 2025. "Barcode 100K Specimens: In a Single Nanopore Run." *Molecular Ecology Resources* 25: e14028.

Hebert, P. D. N., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs. 2004. "Ten Species in One: DNA Barcoding Reveals Cryptic Species in the Neotropical Skipper Butterfly *Astraptes fulgerator*." *Proceedings of the National Academy of Sciences of the United States of America* 101: 14812–14817.

Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, et al. 2009. "A DNA Barcode for Land Plants." *Proceedings of the National Academy of Sciences of the United States of America* 106: 12794–12797.

Ivanova, N. V., M. L. Kuzmina, T. W. Braukmann, A. V. Borisenko, and E. V. Zakharov. 2016. "Authentication of Herbal Supplements Using Next-Generation Sequencing." *PLoS One* 11: e0156426.

Kartzinel, T. R., P. A. Chen, T. C. Coverdale, et al. 2015. "DNA Metabarcoding Illuminates Dietary Niche Partitioning by African Large Herbivores." *Proceedings of the National Academy of Sciences* 112: 8019–8024.

Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman. 2008. "A Continental Strategy for the National Ecological Observatory Network." *Frontiers in Ecology and the Environment* 6: 282–284.

Kolter, A., and B. Gemeinholzer. 2021. "Plant DNA Barcoding Necessitates Marker-Specific Efforts to Establish More Comprehensive Reference Databases." *Genome* 64: 265–298.

Kress, W. J., C. García-Robledo, M. Uriarte, and D. L. Erickson. 2015. "DNA Barcodes for Ecology, Evolution, and Conservation." *Trends in Ecology & Evolution* 30: 25–35.

Kress, W. J., K. J. Wurdack, E. A. Zimmer, L. A. Weigt, and D. H. Janzen. 2005. "Use of DNA Barcodes to Identify Flowering Plants." *Proceedings of the National Academy of Sciences of the United States of America* 102: 8369–8374.

Kuethe, S. 2022. "H3: R Bindings for H3."

Kuzmina, M. L., T. W. A. Braukmann, A. J. Fazekas, et al. 2017. "Using Herbarium-Derived DNAs to Assemble a Large-Scale DNA Barcode Library for the Vascular Plants of Canada." *Applications in Plant Sciences* 5, no. 12: apps.1700079. https://doi.org/10.3732/apps.1700079.

Lendemer, J., B. Thiers, A. K. Monfils, et al. 2020. "The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education." *Bioscience* 70: 23–30.

Li, Y. F., L. Luo, Y. Liu, et al. 2024. "The Bryophyte Phylogeny Group: A Revised Familial Classification System Based on Plastid Phylogenomic Data." *Journal of Systematics and Evolution* 62: 577–588.

Littleford-Colquhoun, B. L., C. Geremia, L. M. McGarvey, et al. 2024. "Body Size Modulates the Extent of Seasonal Diet Switching by Large Mammalian Herbivores in Yellowstone National Park." *Royal Society Open Science* 11: 240136.

Littleford-Colquhoun, B. L., and T. R. Kartzinel. 2024. "A CRISPR-Based Strategy for Targeted Sequencing in Biodiversity Science." *Molecular Ecology Resources* 24: e13920.

Luebert, F., L. Cecchi, M. W. Frohlich, et al. 2016. "Familial Classification of the Boraginales." *Taxon* 65: 502–522.

Maixner, F., D. Turaev, A. Cazenave-Gassiot, et al. 2018. "The Iceman's Last Meal Consisted of Fat, Wild Meat, and Cereals." *Current Biology* 28: 2348.

Meiklejohn, K. A., N. Damaso, and J. M. Robertson. 2019. "Assessment of BOLD and GenBank – Their Accuracy and Reliability for the Identification of Biological Materials." *PLoS One* 14: e0217084.

Meineke, E. K., and B. H. Daru. 2021. "Bias Assessments to Expand Research Harnessing Biological Collections." *Trends in Ecology & Evolution* 36: 1071–1082.

Meineke, E. K., T. J. Davies, B. H. Daru, and C. C. Davis. 2018. "Biological Collections for Understanding Biodiversity in the Anthropocene." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374: 20170386.

Meineke, E. K., C. C. Davis, and T. J. Davies. 2018. "The Unrealized Potential of Herbaria for Global Change Biology." *Ecological Monographs* 88: 505–525.

Meyer, C., P. Weigelt, and H. Kreft. 2016. "Multidimensional Biases, Gaps and Uncertainties in Global Plant Occurrence Information." *Ecology Letters* 19: 992–1006.

Moritz, C., and C. Cicero. 2004. "DNA Barcoding: Promise and Pitfalls." *PLoS Biology* 2: 1529–1531.

National Museum of Natural History, S. I. 2023. "Integrated Taxonomic Information System (ITIS)."

Ondrejicka, D. A., S. A. Locke, K. Morey, A. V. Borisenko, and R. H. Hanner. 2014. "Status and Prospects of DNA Barcoding in Medically Important Parasites and Vectors." *Trends in Parasitology* 30: 582–591.

Pansu, J., J. A. Guyton, A. B. Potter, et al. 2019. "Trophic Ecology of Large Herbivores in a Reassembling African Ecosystem." *Journal of Ecology* 107: 1355–1376.

Pansu, J., M. C. Hutchinson, T. M. Anderson, et al. 2022. "The Generality of Cryptic Dietary Niche Differences in Diverse Large-Herbivore Assemblages." *Proceedings of the National Academy of Sciences* 119: e2204400119.

Paton, A., and E. N. Lughadha. 2011. "The Irresistible Target Meets the Unachievable Objective: What Have 8 Years of GSPC Implementation Taught Us About Target Setting and Achievable Objectives?" *Botanical Journal of the Linnean Society* 166: 250–260.

Pinheiro, H. T., C. S. Moreau, M. Daly, and L. A. Rocha. 2019. "Will DNA Barcoding Meet Taxonomic Needs?" *Science* 365: 873–874.

Pompanon, F., B. E. Deagle, W. O. C. Symondson, D. S. Brown, S. N. Jarman, and P. Taberlet. 2012. "Who Is Eating What: Diet Assessment Using Next Generation Sequencing." *Molecular Ecology* 21: 1931–1950.

Ratnasingham, S., and P. D. N. Hebert. 2007. "bold: The Barcode of Life Data System." *Molecular Ecology Resources* 7: 355–364. http://www.barcodinglife.org.

Reese, A. T., T. R. Kartzinel, B. L. Petrone, P. J. Turnbaugh, R. M. Pringle, and L. A. David. 2019. "Using DNA Metabarcoding to Evaluate the Plant Component of Human Diets: A Proof of Concept." *mSystems* 4, no. 5: e00458-19. https://doi.org/10.1128/mSystems.00458-19.

Renner, S. S., M. D. Scherz, C. L. Schoch, M. Gottschling, and M. Vences. 2023. "DNA Sequences From Type Specimens and Type Strains – How to Increase Their Number and Improve Their Annotation in NCBI GenBank and Related Databases." *Systematic Biology* 73: 486–494.

Riaz, T., W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, and E. Coissac. 2011. "ecoPrimers: Inference of New DNA Barcode Markers From Whole Genome Sequence Analysis." *Nucleic Acids Research* 39: e145.

Schuettpelz, E., H. Schneider, A. R. Smith, et al. 2016. "A Community-Derived Classification for Extant Lycophytes and Ferns." *Journal of Systematics and Evolution* 54: 563–603.

Taberlet, P., E. Coissac, F. Pompanon, et al. 2007. "Power and Limitations of the Chloroplast *trn*L (UAA) Intron for Plant DNA Barcoding." *Nucleic Acids Research* 35: e14.

Thomas, C. 2009. "Biodiversity. Biodiversity Databases Spread, Prompting Unification Call." *Science* 324: 1632–1633.

United Nations Environment Programme. 1992. "Convention on Biological Diversity, June 1992."

Valentini, A., F. Pompanon, and P. Taberlet. 2009. "DNA Barcoding for Ecologists." *Trends in Ecology & Evolution* 24: 110–117.

Vilgalys, R. 2003. "Taxonomic Misidentification in Public DNA Databases." *New Phytologist* 160: 4–5.

von Wettberg, E., and C. K. Khoury. 2022. "Biodiversity Data: The Importance of Access and the Challenges Regarding Benefit Sharing." *Plants, People, Planet* 4: 2–4.

Watanabe, M. E. 2017. "The Nagoya Protocol: Big Steps, New Problems." *Bioscience* 67: 400.

Whipple, J. J. 2001. "Annotated Checklist of the Vascular Plants of Yellowstone National Park." Yellowstone National Park, Wyoming, USA.

Willette, D. A., S. E. Simmonds, S. H. Cheng, et al. 2017. "Using DNA Barcoding to Track Seafood Mislabeling in Los Angeles Restaurants." *Conservation Biology* 31: 1076–1085.

Williams, J. W., T. L. Spanbauer, P. D. Heintzman, et al. 2023. "Strengthening Global-Change Science by Integrating aeDNA With Paleoecoinformatics." *Trends in Ecology & Evolution* 38: 946–960.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.