Cambridge Large Two (2020), 1–4

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE TYPE

Filling in gaps: Herbaria as genomic resources

Pranav Gundrala

Abstract

Keywords: keyword entry 1, keyword entry 2, keyword entry 3

1. Introduction

Herbaria are collections of dried plant specimens stored and catalogued systematically for use in scientific research. These collections are often highly diverse and well preserved. Recent estimates suggest that over 396 million specimens are held in 3567 active herbaria worldwide (Thiers 2024). These specimens represent most of the world's known plant diversity from the last 400 years, including many rare, extinct, or highly endemic species.

Despite this immense volume of specimens, only a small proportion are being used for DNA sequencing and genomics research. A particular area of interest is DNA barcoding, which is a method to identify species using short, standardized genomic sequences called barcodes. This method requires a reference library of barcodes which have already been identified to species. By constructing such a library, barcode data obtained from biological samples can be used to identify a single species or characterize taxa within a mixture of genetic material, called "metabarcoding" (Kartzinel et al. 2025).

The two standard DNA barcodes for plants, found in chloroplast genes matK and rbcL, are generally larger (500-800 base pairs) than the newer trnL(UAA) region (350 base pairs) which is preferred for studies requiring shorter barcodes, such as those using mixed environmental DNA samples. This study focuses on two plant DNA barcode regions found within the chloroplast genome: rbcL (a segment of the gene encoding for RuBisCO) and trnL(UAA) (a generally shorter, intronic region of the leucine transfer RNA gene).

Of the over 300 thousand known plant species, only about one-tenth have vouchered sequences of the rbcL and matK barcodes stored in either GenBank or BOLD (Xu et. al. 2015). The lack of available reference sequences represents an ongoing challenge in DNA barcoding research, which aims to identify species within a sample using short "barcode" sequences, in a variety of fields. Herbarium specimens hold promise for barcoding because they are taxonomically verified to the species level. This can help to overcome the challenge of identifying inconspicuous plants in the field, especially when floral characteristics are not available (Xu et. al. 2015).

It has long been proposed that herbaria carry immense potential as sources of genetic material, but, in practice, the process of recovering useful sequences of high quality is difficult. Poor extraction and PCR amplification are among some of the challenges faced by herbarium barcoding, and current research efforts are focused on developing optimized protocols for handling the sensitive and fragmented DNA produced by these specimens (Särkinen et al. 2020).

1.1 *Characterizing the Brown University Herbarium (BRU)*

The Herbarium at Brown University (BRU) is a regional herbarium specializing in New England native species, representing the largest collection of specimens collected from Rhode Island. While the herbarium is particularly rich in specimens from those localities, the collection represents specimens collected from all 50 US states and over 100 countries. The distribution of specimen collection years centers around the late 19th to early 20th centuries, with a spike in more recent specimens due to active collecting by students, researchers, and associated botanists (Figure 1).

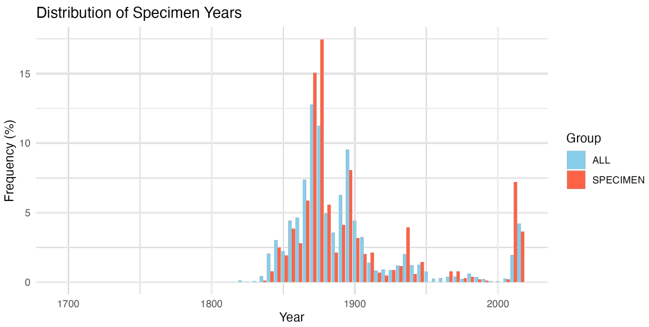


Figure 1. Frequency of specimens in each year from 1700 to 2023. Entire collections (blue) vs. candidate specimens for sampling (red) belonging to no-barcode families

Despite the regional nature of the herbarium, it still contains an impressive diversity of accessions, including incredibly old specimens (Sesamum indicum, 1782 from St. Lucia) and even now-extinct species (Castilleja guadalupensis, 1875 from Guadalupe Island). There exists a potential to leverage the collections at BRU to contribute to the ongoing DNA barcoding efforts for vascular plants. To test this viability in our collections, we selected candidate specimens in families with few to no existing barcodes in the BOLD (Barcode of Life Datasystem) database, with the goal of extracting and sequencing DNA at the trnL and rbcL barcode sites.

2. Materials and Methods

2.1 Specimen selection and sampling

Specimens in BOLD are placed in families designated by ITIS, which the BRU records do not perfectly align with. In order to identify specimens of interest, the ITIS families were first narrowed to those with less than 100 barcodes across all markers, prioritizing those families with no existing *trnL* barcodes (271 families). A list of genera was produced to search the BRU database, identifying candidate 1444 specimens from 72 families. These families were then ranked by least total barcodes and then least trnL barcodes. Of these, the youngest specimen each from the 30 least barcoded families were selected. After review, 18 of these specimens were good candidates for sampling: intact and clean of debris, sufficient herbaceous tissue, type-specimens excluded.

Destructive sampling was performed, recovering about 0.5 cm^2 of tissue from each specimen, ensuring the appearance and integrity of the specimens were not harmed. These samples were then either wholly or partially consumed for DNA extraction.

2.2 Generation of DNA barcodes

Tissue samples were cut into smaller fragments and placed into 2 mL conical tubes with MP Biomedicals Lysing Matrix A (Garnet Matrix, 1/4 in. Ceramic Sphere). DNA was extracted from samples using the FastDNA^{TM} SPIN Kit (MP Biomedicals) following the manufacturer's instructions. Bi-directional PCR amplification of the rbcL and trnL sites was performed using previously published primers (Gill et al. 2019, Kartzinel et al. 2024) and protocols described by Taberlet et al. 1991 (Appendix 1). To confirm PCR success, 2 uL of each product was run through a 1.5% agarose gel along with a PCR positive control (from *Spinacia oleracea*) and negative control of molecular water (Figure 3). Amplicons were cleaned using ExoSap-IT^{TM} (Applied Biosystems) and sequenced on an ABI 3730 DNA Analyzer.

To build consensus sequences, forward and reverse reads were trimmed and assembled in *Geneious* and aligned using the MUSCLE algorithm. For the resulting FASTA sequences, BLAST$^{®}$ searches were performed using the blastn suite and default parameters on the core nucleotide (core_nt) database.

3. Results

3.1 DNA extraction and amplification

DNA was extracted for PCR amplification of the *trnL* and *rbcL* sites and ran through agarose gels, showing varying degrees of success (Figure 2). There was no apparent trend in PCR amplification success in terms of specimen age, but the strongest signals did come from the youngest specimens (collected 2010 or later). Still, specimens like (9) *Anacampseros retusa* (2014) or (5) *Onoclea sensibilis* (2021) showed weak signals, low amplified DNA yield, and subsequent failure to produce viable barcodes.

Those specimens with weak signals after amplification were measured for their resulting DNA concentrations (Figure 3, 4). These concentrations are particularly low, and likely represent highly fragmented DNA and poor amplification success. There does not appear to be a significant trend in these data in terms of the age of the specimens.

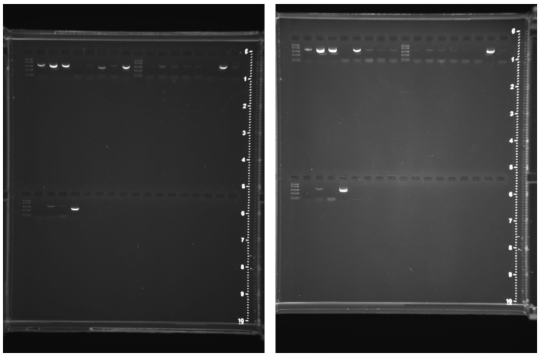


Figure 2. Gel runs of PCR products for trnL (right) and rbcL (left) amplification. Specimens are in order from left to right by sample number (Figure 4). First row: Ladder, samples 1-8, second ladder, samples 9-16. Second row: Ladder, samples 17-18, Negative control, Positive control

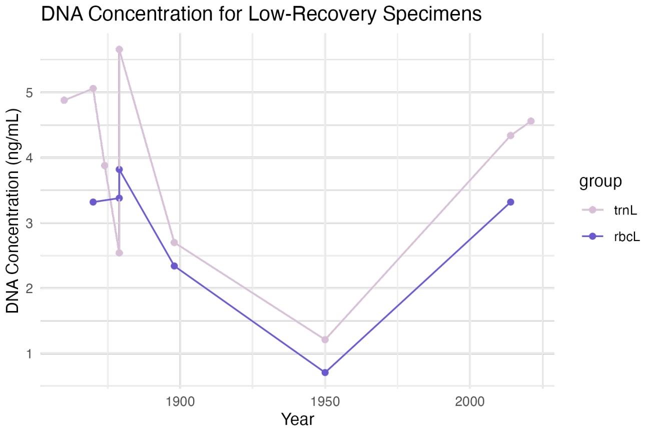


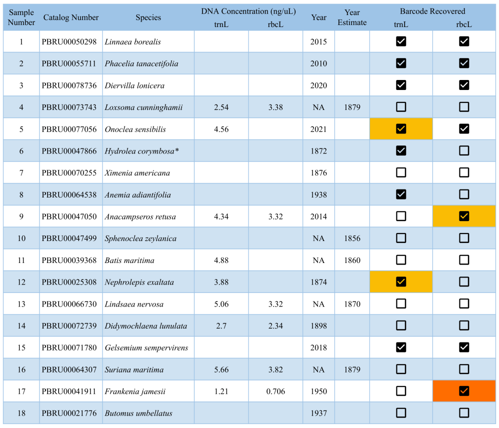
Figure 3. DNA concentrations (ng/uL) after PCR amplification of trnL (lilac) and rbcL (purple) for specimens showing weak banding, plotted against year-of-collection.

3.2 Sequence alignment and analysis

PCR products were sent for sequencing, and raw forward and reverse sequences were compared to produce consensus sequences for each specimen and then aligned using *Geneious*.

Many of the sequencing results were unclear, fragmented, or unusable with many ambiguous base calls, resulting in an inability to recover a consensus sequence. Those sequences that were recovered (Checked in Figure 5) were compared to references using a BLAST search to confirm their identity (Figure 5). However, because these specimens are from no-barcode families, and *none* have an existing trnL barcode in the BOLD database, these matches were largely to unattested accessions or whole chloroplast genome sequences.

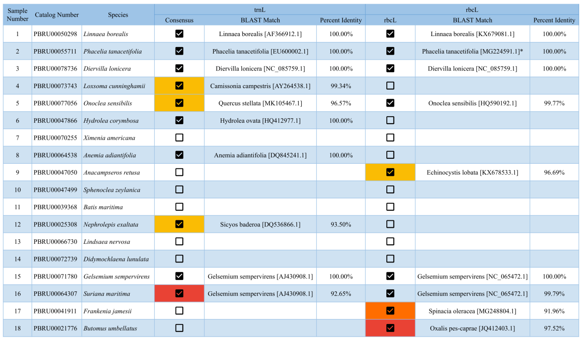
Ranked by highest percent identity, some barcodes matched to the correct taxon (Samples 1, 2, 3) at both sites. Others did not, but these taxa do not have any existing barcode or whole chloroplast genome data in GenBank (highlighted in yellow), so it is possible that these are truly novel barcodes. A few barcodes did not match to the correct taxon, and there *is* existing GenBank whole chloroplast genome data (highlighted in red), so it is unclear if these barcodes can correctly identify the species. The barcode for *Frankenia jamesii* matched to the species used as a positive control (highlighted in orange), but not with a percent identity of 100, meaning it is unclear if this outcome is due to contamination or a lack of data.



Sample Number	Catalog Number	Species	DNA Concentration (ng/uL)		Year	Year Estimate	Barcode Recovered	
			trnL	rbcL			trnL	rbcL
1	PBRU00050298	*Linnaea borealis*			2015		☑	☑
2	PBRU00055711	*Phacelia tanacetifolia*			2010		☑	☑
3	PBRU00078736	*Diervilla lonicera*			2020		☑	☑
4	PBRU00073743	*Loxsoma cunninghamii*	2.54	3.38	NA	1879	☐	☐
5	PBRU00077056	*Onoclea sensibilis*	4.56		2021		☑	☑
6	PBRU00047866	*Hydrolea corymbosa**			1872		☑	☐
7	PBRU00070255	*Ximenia americana*			1876		☐	☐
8	PBRU00064538	*Anemia adiantifolia*			1938		☑	☐
9	PBRU00047050	*Anacampseros retusa*	4.34	3.32	2014		☐	☑
10	PBRU00047499	*Sphenoclea zeylanica*			NA	1856	☐	☐
11	PBRU00039368	*Batis maritima*	4.88		NA	1860	☐	☐
12	PBRU00025308	*Nephrolepis exaltata*	3.88		1874		☑	☐
13	PBRU00066730	*Lindsaea nervosa*	5.06	3.32	NA	1870	☐	☐
14	PBRU00072739	*Didymochlaena lunulata*	2.7	2.34	1898		☐	☐
15	PBRU00071780	*Gelsemium sempervirens*			2018		☑	☑
16	PBRU00064307	*Suriana maritima*	5.66	3.82	NA	1879	☐	☐
17	PBRU00041911	*Frankenia jamesii*	1.21	0.706	1950		☐	☑
18	PBRU00021776	*Butomus umbellatus*			1937		☐	☐

**Hydrolea corymbosa* BLASTed with highest % identity to *Hydrolea ovata*, though it is possible the specimen itself was misidentified.

Figure 4. Table of all sampled specimens showing catalog number, taxa, DNA concentration after PCR, year-of-collection, year estimate (if no date existed on the label), and check boxes for usable/valid barcode recovery. **Some barcodes** did not BLAST to a matching taxa, but no such data exists (yellow). *Frankenia jamesii* (red) matched closest to *Spinacia oleracea* (used as a positive control), so it is unclear if this is due to contamination.



Sample Number	Catalog Number	Species	trnL			rbcL		
			Consensus	BLAST Match	Percent Identity	rbcL	BLAST Match	Percent Identity
1	PBRU00050298	*Linnaea borealis*	☑	Linnaea borealis [AF366912.1]	100.00%	☑	Linnaea borealis [KX679081.1]	100.00%
2	PBRU00055711	*Phacelia tanacetifolia*	☑	Phacelia tanacetifolia [EU600002.1]	100.00%	☑	Phacelia tanacetifolia [MG224591.1]*	100.00%
3	PBRU00078736	*Diervilla lonicera*	☑	Diervilla lonicera [NC_085759.1]	100.00%	☑	Diervilla lonicera [NC_085759.1]	100.00%
4	PBRU00073743	*Loxsoma cunninghamii*	☑	Camissonia campestris [AY264538.1]	99.34%	☐		
5	PBRU00077056	*Onoclea sensibilis*	☑	Quercus stellata [MK105467.1]	96.57%	☑	Onoclea sensibilis [HQ590192.1]	99.77%
6	PBRU00047866	*Hydrolea corymbosa*	☑	Hydrolea ovata [HQ412977.1]	100.00%	☐		
7	PBRU00070255	*Ximenia americana*	☐			☐		
8	PBRU00064538	*Anemia adiantifolia*	☑	Anemia adiantifolia [DQ845241.1]	100.00%	☐		
9	PBRU00047050	*Anacampseros retusa*	☐			☑	Echinocystis lobata [KX678533.1]	96.69%
10	PBRU00047499	*Sphenoclea zeylanica*	☐			☐		
11	PBRU00039368	*Batis maritima*	☐			☐		
12	PBRU00025308	*Nephrolepis exaltata*	☑	Sicyos baderoa [DQ536866.1]	93.50%	☐		
13	PBRU00066730	*Lindsaea nervosa*	☐			☐		
14	PBRU00072739	*Didymochlaena lunulata*	☐			☐		
15	PBRU00071780	*Gelsemium sempervirens*	☑	Gelsemium sempervirens [AJ430908.1]	100.00%	☑	Gelsemium sempervirens [NC_065472.1]	100.00%
16	PBRU00064307	*Suriana maritima*	☑	Gelsemium sempervirens [AJ430908.1]	92.65%	☑	Gelsemium sempervirens [NC_065472.1]	99.79%
17	PBRU00041911	*Frankenia jamesii*	☐			☑	Spinacia oleracea [MG248804.1]	91.96%
18	PBRU00021776	*Butomus umbellatus*	☐			☑	Oxalis pes-caprae [JQ412403.1]	97.52%

Figure 5. Table of highest % identity BLAST matches for recovered barcodes. Some barcodes did not match to the correct taxa, but no data exists (yellow). Other barcodes did not match to the correct taxa, but whole chloroplast genomes *do* exist in GenBank (red), even though BOLD barcodes may be absent or limited. *Frankenia jamesii* does not have existing barcode or genome data, but the barcode BLASTed to *Spinicia oleracea*, which was used as a positive control, indicating possible contamination.

4. Discussion

Although herbarium specimens are dessicated and preserved, DNA can still be degraded by a number of processes based on the methods of drying and storage (Bhoyar 2024). Immediately after collection, during the drying process, cellular and enzymatic processes can significantly damage DNA. Once desiccated and stored, non-enzymatic degradation, mediated by exposure to heat, corrosive chemicals, UV radiation, and moisture all continue to impact DNA, albeit to a lesser extent. The former has to do with the drying or preservation method, which is not easily inferred from specimens (Särkinen *et. al.* 2012). The latter has to do with the age of specimens, which has systematically been observed to have little to no impact on the success of DNA extraction and amplification (Erkens *et. al.* 2008).

This has been corroborated by some studies (Särkinen *et. al.* 2012, Kuzmina *et. al.* 2012, Staats *et. al.* 2011), but refuted by others (Xu *et. al.* 2015 and Brewer *et. al.* 2019) who did observe varying degrees of success due to age. However, the latter studies do note that these effects could be due to other factors. In our study, specimen age *did* seem to impact the ability to recover sequence, with the short *trnL* barcode recovering more sequences than *rbcL* (Figure 6). However, once again, there may have been other factors at play, given the small sample size represented here.

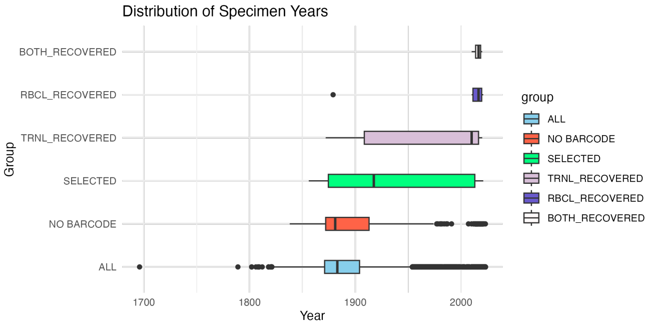


Figure 6. Boxplot distribution of specimen year-of-collection for nested groups: all BRU specimens (blue); those in no-barcode families (red); those that were selected for sampling (green); and specimens where barcodes for either trnL (lilac), rbcL (purple), or both sites (white) were recovered.

The particularities of this kind of sampling can impact the sequencing pipeline at any stage from DNA extraction and marker amplification (PCR) to eventual sequencing. The authors of Särken *et. al.* mention that specimen chemistry might impact PCR success, citing both artifacts of specimen preservation (like poisoning with mercury or arsenic) and taxon specific traits (phytochemistry like the presence of sap, resin, or high polyphenol concentration).

Other taxon specific traits may have impacted the drying itself, as certain leaf morphologies (thick or fleshy leaves) lend themselves to slower drying and therefore a higher chance for DNA degradation (Särkinen *et. al.* 2012). Amongst our specimens, *Batis maritima*, *Surina maritima*, and *Anacampseros retusa* are all succulent-leaved plants, and each sample showed low amplified DNA yield and low sequence quality, resulting in no usable barcodes recovered. While *Surina maritima* did produce readable sequence for the *rbcL* site, the consensus matched most closely to *Gelsemium sempervirens* (the sample directly neighboring it) in a BLAST search, indicating contamination.

Another taxon specific trait that might be explored is the preservation of DNA and barcoding success in fern (pteridophyte) samples. The aforementioned papers only characterize DNA preservation and barcoding in herbarium specimens of seed plants (Särkinen *et. al.* 2012) or, even more narrowly, flowering plants (Brewer *et. al.* 2019). From our samples, the following were of fern species: *Loxsoma cunninghamii* (4), *Onoclea sensibilis* (5), *Anemia adiantifolia* (8), *Nephrolepis exaltata* (12), *Lindsaea nervosa* (13), *Didymochlaena lunulata* (14). While the sample size here is small (n = 6), these specimens represent the majority of those that did not produce usable barcodes, including *all* of the specimens which produced barcode sequences that did not BLAST search to the correct taxon. Furthermore, these specimens widely range in age, with collection years from 1874 up to 2021. Current literature does not classify the preservation of DNA or barriers to marker amplification in these taxa, and more work might be needed to understand either the underlying morphological, phytochemical, and (possibly) genetic reasons for poor barcode recovery in this group.

5. Conclusion

This initial attempt at barcoding within the BRU collections has proved the utility of these collections for further barcoding experiments. While there are challenges to recovering DNA barcodes from herbarium specimens, those that are recovered are of high quality. The barcode sequences recovered here represent significant additions to existing the reference database, and each one is vouchered by a specimen held at the BRU. Those specimens that were not successful in producing barcodes appeared to share certain characteristics in terms of age, leaf morphology, and taxa. Older specimens seemed to produce weak PCR signals, low-quality DNA, and no usable sequence. Younger specimens that produced similarly low quality sequence were of two kinds: (1) succulent-leaved like Anacampseros retusa (2014) or (2) belonging to the pteridophyte clade like Onoclea sensibilis (2021). While the sample size here is too small to draw definitive conclusions about factors improving or limiting barcode recovery, the results lined up with existing research into herbarium specimen barcoding and DNA degradation in plants. Further barcoding work in these collections might focus on alternative extraction methods or classification of barcoding success in particular taxa like ferns.

6. Acknowledgments

This research was funded and supported by the Brown University Herbarium and the Tyler Kartzinel Lab and advised by Rebecca Kartzinel, Tyler Kartzinel, and Anna Jackson.

7. Data Availability Statement

Herbarium specimen data was accessed from the Consortium of Northeastern Herbaria website: https://neherbaria.org/portal/

Code and metadata can be found at: https://github.com/pgundral/herbarium-metabarcoding

Appendix 1. PCR Protocol

The PCR protocol provided by the Tyler Kartzinel Lab can be found at: https://docs.google.com/document/d/1_PYMVABPbbSCR_xliUy0EWCL8sn02TTS/edit?usp=sharing&ouid=1055246927578982636
rtpof=true&sd=true