Introducing an \textit{rbcL} and \textit{trnL} reference library to aid in the metabarcoding analysis of foraged plants from semi-arid eastern South African savannas

Danielle Botha\textsuperscript{1}, Mornè du Plessis\textsuperscript{2}, Frances Siebert\textsuperscript{1}, Sandra Barnard\textsuperscript{1}

\textsuperscript{1}Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa.

\textsuperscript{2}Core Sequencing Facility, National Institute for Communicable Diseases of the National Health Laboratory Service, Sandringham, Johannesburg, South Africa

*Corresponding author

E:mail: sandra.barnard@nwu.ac.za

\& SB and FS conceived and designed the research and conducted field collections; FS conducted taxonomic identifications; DB and MduP performed analyses. DB drafted the manuscript with subsequent contributions from all authors.
Abstract

The success of a metabarcoding study is determined by the extent of taxonomic coverage and the quality of records available in the DNA barcode reference database used. This study aimed to create an *rbc*L and *trn* L (UAA) DNA barcode sequence reference database of plant species that are potential herbivore foraging targets and commonly found in semi-arid savannas of eastern South Africa. A study-area-specific species list of 755 species was compiled. Thereafter, reference libraries for *rbc*L and *trn* L (UAA) sequences were created mined from sequence databases according to specific quality criteria to ensure accurate taxonomic coverage and resolution. The taxonomic reliability of these reference libraries was evaluated by testing for the presence of a barcode gap, identifying a data-appropriate identification threshold, and determining the identification accuracy of reference sequences via primary distance-based criteria. The final *rbc*L reference dataset consisted of 1238 sequences representing 318 genera and 562 species. The final *trn* L dataset consisted of 921 sequences representing 270 genera and 461 species. Barcode gaps were found for 76% of the taxa in the *rbc*L barcode reference dataset and 68% of the taxa in the *trn* L barcode reference dataset. The identification success rate, calculated with the *k*-nn criterion was 85.86% for the *rbc*L dataset and 73.72% for the *trn* L dataset. The datasets for *rbc*L and *trn* L combined during this study are not presented as a complete DNA reference library, but rather as two datasets that should be used in unison to identify plants present in the semi-arid eastern savannas of South Africa.

1. Introduction
Dietary analysis is a fundamental part of constructing habitat selection and utilization models as well as assessing the influence of land use type on the plant community and how this, in turn, can influence foraging strategies [1, 2]. Determining and analysing food items in ecosystems will also aid in identifying key environmental resources for the design of reliable conservation and management strategies [3, 4]. To determine the composition of animal diets and how this composition reflects the plant community in question, animal faeces can be examined to identify food items [2,5], as well as the relative abundance thereof ingested per individual sampled [6].

The metabarcoding approach allows for species identification from heterogeneous and degraded environmental samples based on the amplification and subsequent analysis of DNA barcodes [7,8], which are short gene sequences from a standardized region of the genome that displays low intraspecies and high interspecies variability [9,10]. DNA metabarcoding provides an attractive alternative to the traditional methods of large-scale species identifications and it can be applied to microscopic [11], cryptic [12], and digested materials [2,13].

The rbcLa and trnL (UAA) DNA barcodes have been proposed as two single-locus gene regions that can be used for the analysis of herbivorous- and omnivorous animal diets via DNA metabarcoding [2,13-16]. These two barcodes were chosen as the core barcoding regions for this study due to their coverage of different savanna plant species and because of their sequence availability on global databases.

Except for the choice of the barcoding marker to be used, the efficiency and usefulness of metabarcoding as a tool are underpinned by the extent of taxonomic coverage and the quality of records available in the DNA barcode reference database to which the query sequences will be compared and identified [17,18]. However, there is a lack of...
plant taxonomic sequence information for some ecological regions on global DNA sequence databases [19]. According to [19], as of 2019 DNA sequences available on the Barcode of Life Data (BOLD) database represent only ~8% of the vascular plants in Africa, while Africa compromises ~20% of the earth’s landmass and owns approximately 15% of the global plant diversity [20]. Furthermore, forbs (i.e., non-graminoid herbaceous vascular plants: [21,22], which contribute more than 70% to herbaceous species richness in grasslands, temperate deciduous forests, and savanna ecosystems, have been neglected in vegetation studies and knowledge of their ecological role in savanna ecosystems are limited [23]. As a result, forbs are also weakly represented on global DNA sequence databases.

Furthermore, it is well known that public DNA reference databases may contain sequencing errors and unverified taxonomies by a lack of voucher specimen information and links to metadata. This is especially true for GenBank where the curation of the database is left largely to scientists submitting sequences [24,25]. Since the DNA reference database will determine the accuracy of taxonomic assignations i.e., the success of a barcode in distinguishing between species [26], it is necessary to create a subset of the data available on public DNA databases that are manually curated for sequence quality, taxonomic validity, geographic validity as well as to test the efficiency of this subset reference database to discriminate between and within species with similarity, distance- and tree-based methods.

With this study we therefore aim to: (i) create a comprehensive rbcLa and trnL (UAA) DNA barcode reference database of plant species commonly found in a semi-arid South African savanna, and (ii) evaluate the efficiency of this reference database in discriminating between and within species by testing for the presence of a barcode gap, identifying a data appropriate identification threshold as well as determining the
identification accuracy of reference sequences via primary distance-based criteria. The fulfilment of these aims will result in a DNA reference database that can be used confidently to delineate between species found in the faeces of herbivores foraging in a semi-arid eastern South African savanna as well as to use the successful taxonomic identifications to make certain deductions about the plant community to aid in management and conservations strategies.

2. Materials and methods

2.1. Barcode reference list assembly

The compilation of this DNA reference database of species found in a semi-arid eastern South African savanna is challenging since scientific publications listing all taxa likely to be encountered in the diets of herbivores foraging in these ecosystems were largely unavailable. The semi-arid eastern savannas of South African, located in the Limpopo and Mpumalanga provinces is defined as savannas receiving a mean annual precipitation of below ~650 mm rendering them ‘unstable’, therefore needing disturbances such as fire and herbivory to maintain the co-existence of trees and grasses [27]. The species lists used as reference for sequence mining was built around two sites at which floristic surveys have been undertaken for the purpose of community assemblage changes with grazing [28]. Dominant plant species recorded at these sites were sequenced by the North-West University (NWU) if they were poorly represented in DNA sequence libraries. Syferkuil is an experimental farm in the Limpopo province, and Welverdiend (Mpumalanga province) a rural settlement with open spaces for communal grazing. To expand the reference database accuracy this list was also enriched with applicable floristic and DNA sequencing data from other
studies based in African savannas from which DNA sequences were available (e.g. [2,19]. A subset of rbcLa and trnL (UAA) records available on public databases was created by using the compiled species list (Supplementary Table 1) to mine sequences according to specific quality criteria to ensure accurate taxonomic coverage and resolution. These criteria included sourcing a maximum of three individuals per species per marker, not including sequences from the same herbarium, and giving preference to sequences with voucher specimens and sequences with the least ambiguous bases. Accession numbers of barcodes that match the quality criteria and have been sequenced with the rbcL primers, rbcLa-F: ATGTCACCACAAACAGACTAAAGC [29] and rbcLa-R: GTAAAATCAAGTCCACCRCG [30] as well as with the trnL primers [31], trnL-c: CGAAATCGGTAGACGCTACG and trnL-d: GGGGATAGAGGGACTTGAAC, can be found in Supplementary Table 2. The public databases used to source sequences in this study were the GenBank nucleotide database (www.ncbi.nlm.nih.gov) and the BOLD System public database (www.boldsystems.org).

This study also enriched the sourced dataset by supplying twenty-four sequences from plants found in semi-arid eastern South African Savanna ecosystems. Most of these sequences are known to not occur at all, or not frequently on public DNA sequence databases. A pipeline for the collection, DNA isolation, and sequencing of rbcLa and trnL barcodes has been established in 2015. Leaf material and voucher specimens of woody and herbaceous plants were collected from Welveridend, Syferkuil and neighbouring landscapes. Botanists at the NWU examined the voucher specimens to confirm their taxonomy and they were entered into the local Goossens Herbarium of the NWU. The barcode sequences were added to the GenBank nucleotide database and included in the reference database compiled and evaluated in this study. GenBank
accessions numbers for the rbcLa barcode range from MZ461574.1 to MZ461594.1 and the trnL barcode from MZ461547.1 to MZ461570.1. (Supplementary Table 3)

2.2. Phylogeny

The candidate barcode reference sequences for both markers to be used for phylogenetic analysis were respectively aligned using Geneious (v. 2021.2.2) via MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [32], with a maximum number of eight iterations. The sequences for the trnL marker were not split into families, as they will be for downstream analysis, due to the requirement of robust representations of the alignment and clusters formed for the subsequent “quality” analysis.

The aligned candidate barcode reference sequences were used to reconstruct neighbour-joining (NJ) phylogenetic trees for the rbcLa and trnL dataset, as implemented by Geneious with bootstrap testing of 1000 replicates using the Jukes-Cantor model. Within angiosperms, Amborella is reported as the sister to all remaining flowering plants [33,34], and accordingly, the rbcLa and trnL (UAA) genes of Amborella trichopoda were included as an outgroup for the rbcLa and trnL sequences and aligned with the rest of the datasets. The individuals in both the rbcLa and trnL phylogenetic tree were also reviewed manually to validate clusters and sub-clusters that were formed. This evaluation was put in place to improve further analyses recommending the removal of sequences showing an evident deviation from their cluster, be it a cluster of life forms (such as trees, shrubs, grasses, forbs) or plant families. The sequences that were identified to have problematic placements were then replaced with another individual sequence of the species with equal quality if possible, or it was removed from further analysis.
2.3. DNA barcode assembly for reference database

DNA sequences for both marker datasets were exported to RStudio and aligned to prepare sequences for downstream analysis. Alignment was again facilitated with MUSCLE, but without the outgroup sequences, and subsequent determination of genetic distance and barcode-gap analysis took place. All sequences for \textit{rbcLa} were aligned simultaneously using the \textit{AlignSeqs()} function in the DECIPHER package [35], and \textit{trnL} sequences were split into families [19] and accordingly aligned with the same function due to the often highly divergent sequences displayed for distantly related species. Distance matrices for both datasets were calculated using the Kimura-2 parameter (K2P) model [36] which is the build-in model of evolution in the APE package’s \textit{dist.dna()} function [37]. Paradis & Schliep

2.4. Barcode gap analysis

The barcode analysis for each marker dataset used the K2P model to calculate genetic distances for each sequence per marker and was done using the R package SPIDER (v.1.5) [38]. A barcode gap analysis is performed by two approaches, as recommended by [39], which will allow for barcode gap analyses of the complete datasets, as well as barcode gaps for individual sequences represented in the respective datasets. The first approach compares the median of inter – and intraspecific distances for each marker dataset and assesses the significance of the differences between genetic distances with a Wilcoxon rank-sum test with the R function \textit{wilcox.test()} in the R STATS package. The second approach follows the determination of the presence of a barcode gap following [40], where the difference between maximum intraspecific distance and the minimum interspecific distance is
calculated for each sequence per marker dataset. Both approaches were applied during this study.

2.5. Analysis with distance-based criteria

The optimisation of the species identification threshold for each marker dataset is performed to increase identification success [26]. During this study, we used the `threshOpt()` function in the SPIDER package of R (v.1.5) [38]: to systematically subject the marker dataset to a range of threshold values (between 0.1 to 0.2% K2P distance). The optimal identification threshold was then calculated based on the number of true positive, false positive, true negative, and false negative identifications for each threshold in the range. The optimal threshold was identified as the value predicting the lowest cumulative error, i.e., the lowest amount of false-positive and negative identifications according to [38].

The predicted accuracy of taxonomic assignation of the reference sequences was determined using three distance-based analyses via the SPIDER package of R (v.1.5) [38]. These primary distance-based criteria included the nearest neighbour (k-nn) criterion [41], Meier’s best close match (BCM) criterion [42] as well as BOLD identification criterion [43]. The functions by which these measures are executed (`nearNeighbor()`, `bestCloseMatch()`, and `threshID()`), respectively performed an internal taxonomic assignation, where individual sequences were treated as unknown queries and the rest of the marker dataset was treated as the DNA reference database to match corresponding taxonomies within the identified optimal threshold.

3. Results

3.1. DNA reference dataset: phylogenetic analysis and summary statistics
The final species list (Supplementary Table 1) consists of 755 species, 356 genera, 99 families as well as 3 classes (Magnoliopsida, Liliopsida, and Polypodiopsida) in the phylum Tracheophyta. Of these 755 species, sequences for 108 could not be found for either the \textit{rbcL} or \textit{trnL} barcode. Upon inspection of the NJ trees of \textit{rbcL} and \textit{trnL}, it was advised to replace 265 individuals with other sequences from the same \textit{species} of equal quality since they did not group with other members of their family or life form. Sequences for 59 individuals could be replaced, but 206 sequences had to be removed. This led to a loss of taxonomic coverage in some species, now only represented by < 3 individuals and the complete loss of 57 unique species from the reference dataset. The remaining species and accession numbers, or process identities (process IDs) in the case of BOLD sequences, can be found in the supplementary table which now constitutes the complete DNA reference database. The composition of the complete reference database is illustrated in Figure 1, where the family Poaceae is the largest representative with 47 genera and 109 species. The second-largest representative family is the Fabaceae with 35 genera and 87 species. The \textit{rbcLa} reference dataset consists of 1238 sequences representing 318 genera and 562 species. The \textit{trnL} dataset consists of 921 sequences which represent 270 genera and 461 species. Naturally, there is an overlap between the datasets as sequences were mined for the same species for both markers: sequences for 345 species of 224 genera are found in both marker datasets.

Figure 1. The composition of the final DNA sequence reference database in terms of number of genera (histogram) and species (line plot) per family.

### 3.2. Reference dataset analysis
According to the first approach of the barcode gap analysis, by which the presence of a barcode was assessed for both marker datasets by the comparisons of medians for inter-and intraspecific genetic distances, it was found that for the \textit{rbcL} dataset, the interspecific K2P distances ranged from 0,00 – 0,94 with a median of 0.01. The interspecific distances were statistically larger than the intraspecific K2P distances (Wilcoxon test: $P < 2.2e^{-16}$) with ranges of 0 – 1.45 and a median of 0.00. The \textit{trnL} dataset followed the same trend with generally larger interspecific distances (ranges from 0,00 – 1,01 and a median of 0,01) than intraspecific distances (ranges: 0 – 1,07; median: 0; Wilcoxon test: $p < 2.2e^{-16}$). The second approach [42], revealed a positive difference between maximum intra- and minimum interspecific K2P distances for 76% of the \textit{rbcL} sequences and 68% for the \textit{trnL} sequences. These results are reported as scatterplots (Fig. 2) where the sequences that showed a gap appear above the 1:1 slope.

Figure 2: Scatterplots of the Barcode gap s for species present A) in the \textit{rbcL} dataset and (B) \textit{trnL} dataset. Scatterplots show the relationships between minimum interspecific distance (y-axis) and maximum intraspecific distance (x-axis) where the barcode gap is indicated by points falling above the 1:1 slope. High densities of plots are indicated by the darker colours.

To improve the accuracy of the DNA reference database, optimal identification thresholds were determined with the threshOpt function in the R package SPIDER. The default threshold of species optimisation used by identification algorithms such as BLAST and BOLD is 1%, which may not always be appropriate for all reference datasets. Accordingly, the determined optimal threshold for the \textit{rbcL} dataset was 0.4%
and that of the \textit{trnL} dataset was 0.6\%, which clearly indicates that the identification accuracy of the datasets would not have been accurately portrayed by the default threshold value of 1\%. These optimal threshold values will be used during the simulated taxonomic identification via the three distance-based analyses which will aid in predicting the accuracy of the reference datasets.
Table 1: Discriminatory power of the *rbcL* and *trnL* reference datasets predicted by the application of three distance-based measures: nearest neighbour, Meier’s best close match, as well as BOLD identification criterion, for both default and optimized thresholds for the inclusion and exclusion of singletons for the respective datasets. All three analyses were performed to identify queries within the respective identification thresholds to genus level.

| Dataset | Singleton | Total number of taxa | Threshold value (%) | Nearest Neighbour | Best Close Match | BOLD Identification |
|---------|-----------|----------------------|--------------------|-------------------|------------------|---------------------|
|         |           |                      |                    | False | True | Ambiguous | Correct | Incorrect | No ID | Ambiguous | Correct | Incorrect | No ID |
| *rbcL*  | Included  | 1238                 | 1                  | 175   | 1063 | 104       | 968     | 72        | 94    | 498       | 604     | 42        | 94    |
|         |           |                       | 0,4                | NA    | NA   | 102       | 877     | 38        | 221   | 194       | 794     | 29        | 221   |
|         | Excluded  | 1171                 | 1                  | 108   | 1063 | 104       | 968     | 48        | 51    | 498       | 604     | 18        | 51    |
|         |           |                       | 0,4                | NA    | NA   | 102       | 877     | 29        | 163   | 194       | 7943    | 20        | 1631  |
| *trnL*  | Included  | 921                  | 1                  | 242   | 679  | 49        | 550     | 93        | 229   | 183       | 457     | 52        | 229   |
|         |           |                       | 0,6                | NA    | NA   | 49        | 508     | 77        | 287   | 129       | 459     | 46        | 287   |
|         | Excluded  | 859                  | 1                  | 180   | 679  | 49        | 550     | 76        | 184   | 183       | 457     | 35        | 184   |
|         |           |                       | 0,6                | NA    | NA   | 49        | 508     | 63        | 239   | 129       | 459     | 32        | 239   |

NA – The identification simulation using the nearest-neighbour criterion does not allow for supplying an alternative identification threshold and only uses the build-in default of 1%.

*Nearest-Neighbour (k-NN) criterion result definitions:
False – the nearest species index is not the same as the tested individual; True – the nearest species index is the same as the tested individual. Note that the optimal threshold could not be changed from its default in the nearestNeighbour() function.

**Best Close Match (BCM) criterion result definitions:**

Ambiguous – Correct and incorrect species are identified as closest matches within the identification threshold; Correct – all close matches within the identification threshold of the query are the same species; Incorrect – all matches within the identification threshold are different species than the query; No ID – no matches could be made within the identification threshold for any individual.

***BOLD identification criterion result definitions:**

Ambiguous – Correct and incorrect species are identified as closest matches within the identification threshold; Correct – all close matches within the identification threshold of the query are the same species; Incorrect – all matches within the identification threshold are different species than the query; No ID – no matches could be made within the identification threshold for any individual.
The nearest-neighbour (k-NN) criterion performed the best among the distance-based measures during the identification simulations (Table 1). The identification success rate, calculated as the number of matches that share the same species index as the query was 85.86% for the complete *rbcL* dataset and 73.72% for the *trnL* dataset. These identification success rates can be increased by excluding singletons, resulting in an identification success rate of 90.78% and therefore an increase of 4.92% for the *rbcL* dataset. Excluding singletons led to an increase of 5.33% resulting in a success rate of 79.05% for the *trnL* dataset. Meier’s [41] best close match (BCM) criterion performed similarly to the k-NN identification criterion with an identification success rate of 82.66% correct matches with the optimized threshold value for the exclusion of singletons for the *rbcL* dataset and 64.03% for the *trnL* dataset with matches performed in the optimized threshold with the exclusion of singletons. The BOLD identification criterion failed to show the same level of identification success rates with the inclusion of an optimized identification threshold. The BOLD identification criterion performed the worst among the other measures with the highest discriminating power shown for the *trnL* dataset with the exclusion of singletons, reaching 53.20% with the default threshold value of 1%.

4. Discussion

Reference library development

Despite several studies in which DNA barcoding was applied in the analyses of herbivore diets [3,44,45], a focus on herbivore diets in African savannas remains limited [2, 19]. Until present, none focused on South African savannas, and there is not currently a robust DNA reference sequence library available for the semi-arid eastern South African Savanna. During this study we have curated a DNA reference
library consisting of both and the *rbcL* and a *trnL* dataset. This is the first step towards facilitating metabarcoding studies aimed at South African savanna ecosystems, specifically for the identification of plant materials obtained in herbivorous faeces.

The ideal construction of the reference library would be to sample, identify, and sequence the relevant barcodes of all the species on the species list. However, this would be a costly and laborious research campaign. Mining sequences from global databases is a much more cost-effective option and the method illustrated here can be reproduced for any nature of metabarcoding study aimed at representing the composition of an ecosystem without manually sequencing all specimens. The use of GenBank as a taxonomic resource has been questioned since there exists an absence of preserved voucher specimens, non-justified species identifications, and low-quality data [42,46]. The same is true for BOLD, although this database is better curated due to higher quality submission standards but may still contain incorrectly identified taxa and low-quality sequence data [42,43]. Being aware of these limitations enables the user to create certain download criteria which result in a database of sequences with known localities, known voucher specimens as well as the inclusion of barcodes with the least ambiguous bases. This database, as well as any database mined from global reference databases, is not impervious to imperfection, but this method of using a species list and applying download quality criteria will certainly lift its standard above a taxonomic assignation against whole databases such as GenBank and BOLD. The lowest level of taxonomic resolution achieved in metabarcoding studies is typically genus-level [44,47]; Turunen *et al.*, 2021) and the assignation of taxonomy at lower taxonomic levels is often infeasible and inaccurate. Therefore, to ensure that this database of barcode sequences is
robust enough to be used as an identification tool the barcode gap was analysed together with the barcode phylogeny.

Disparities between inter-and intraspecific distances between and among species are defined as a barcoding gap that, if present at a locus, enables the reliable differentiation of species [26, 38]. A barcode gap was evident in both approaches used and for both datasets of which the reference library consists of. A barcode gap for 76% of the species was obtained for the \textit{rbcl} barcode sequences and for 68% of the species for the \textit{trnL} sequences. This implies that reliable species differentiation would not be possible for 24% of the species in the \textit{rbcl} dataset and 32% in the \textit{trnL} dataset. However, limitations have also been reported in other DNA-barcode studies such as [19, 48 and 30]. Gill et al., (2019) [19] reported a barcode gap for only 73% of the species investigated for the \textit{rbcl} primer and 79% for the \textit{trnL-F} primer in their barcode library for the semi-arid eastern South African Savanna plants. Mishra et al., (2017) [48] have shown that even within one genus (\textit{Terminalia}) the three different barcodes of \textit{matK}, ITS and \textit{rbcl} contained a barcode gap for <70% of the species. As stated by Gill et al., (2019) [19] savannas are comprised of a diverse range of species that are prone to the absence of a barcoding gap.

The distance-based measures of \textit{k}-nn, BCM, and BOLD identification criteria infer identification success rates by considering the K2P distance matrix of each dataset to simulate taxonomic identifications of one sequence against the rest of the databases to match taxonomies within the identification threshold [37]. Usually, the distance-based matrices are applied at a species level, but for this metabarcoding study, it was decided to predict the identification success of the datasets based on genus-level identifications (Table 1). The best success rates across all three methods were seen with the exclusion of singletons, in this case, genera.
represented by only one individual. The analysis of the identification of success
rates, with the exclusion of singletons, is also a prerequisite for the alignment that
precedes the barcode gap analysis. Singletons are typically a problem in barcoding
studies since the identification simulations will treat the singleton as a query, and it
will not have a match available in the reference dataset and will result in either
“incorrect” or “no identification” [37].

Analyses revealed that the rbcL dataset has a predicted identification success rate of
90.78% and trnL 79.05% with the k-NN method. This would imply that a barcode gap
analysis is not always an accurate predictor of the species discrimination success of
sequences in a reference database, as concluded in certain studies [40,49, 50]. This
is also reflected by the findings in this study as illustrated by the inconsistencies
between identification success rate simulations and the evaluation of the barcode
gap. However, in this case, the barcode gap analysis can be accepted as a more
efficient tool to predict database accuracy since it displays the identification success
rate achieved by species-level identifications, whereas the distance-based analysis
implemented genus-level taxonomic assignments. Furthermore, the discrimination
success rates demonstrated by both datasets are above (rbcL = 76%) or very close
(trnL = 68%) to the discrimination success rate of 72% as proposed by the CBOL
Plant Working Group (2009) [51]. This implies that an adequate discrimination
success rate would be possible to the lowest taxonomic level of genus when taking
the predicted identification success rates into account.

The phylogenetic analysis served as a re-identification strategy of sequences
included in the rbcL and trnL reference database. The exclusion of singletons, as
recommended by the identification rates of the various distance-based measures
(Table 1) was not considered in this part of the study, as this would have led to a
great loss of taxonomic coverage and identifications as seen by a drop of 5.41% and
7.22% in the total taxonomy included in the rbcL and trnL databases, respectively.
Excluding singletons from the reference datasets would lead to a loss of the families
of flowering plants, namely: Aristolochiaceae, Dioscoreaceae, Gentianaceae,
Kirkiaceae, Passifloraceae, Piperaceae, Plantaginaceae, Urticaceae in the rbcL
datasets as well as the loss of Hypoxidaceae, Menispermaceae, Nyctaginaceae,
Passifloraceae, Piperaceae, Rutaceae, Scrophulariaceae and Stilbaceae in the trnL
dataset. We, therefore, sacrificed some identification success for the inclusion of
singletons, and therefore an accurate representation, in terms of species present, in
a South African savanna. The NJ trees revealed that sequences formed cohesive
clusters of orders for the rbcL dataset. However, an obvious low taxonomic
resolution for some orders in the trnL dataset was again obvious, such as for the
orders Poales and Commelinales and Fabales. Similar results were also shown by
[19] for the families Poaceae, Fabaceae, and Malvaceae. Low taxonomic resolutions
were expected since many genera which are known to occur in South African
savannas are under-sampled in global reference databases or are yet to be
barcoded with either rbcL or trnL, leading to a lack of representatives for certain
genera or the inclusion of singletons in the reference database. Species represented
in the South African savannas are very diverse as can be seen from the sum of the
branch lengths of the NJ trees, namely 108.23 for the 922 trnL-barcoding taxa and
382.34 for the 1239 rbcL-barcoding taxa.

Conclusion

During this study, we developed a DNA barcode reference library that is robust to
identify taxons on the list of species that we curated for some foraged plants from the
semi-arid eastern South African savanna. All the different tests used to validate the
use and accuracy of this library indicate that it can be used with confidence to assign
taxonomies for plants found in the eastern savannas of South Africa. We envisage
that it will add to other similar research done not only on the local flora but also to the
work done on savannas elsewhere on the African continent. The datasets for rbcL
and trnL are not presented as a complete DNA reference library, but rather as two
databases that should be used in unison to identify species foraged on in the semi-
arid eastern South African Savanna.

Acknowledgments

We thank Miss Siena Bieler for assistance with field collections, DNA isolation, and
sequencing of barcodes. We thank the North-West University for its support of this
research.

References

1. Valentini, A, Pompanon, F, Taberlet, P. DNA barcoding for ecologists. Trends in
Ecology and Evolution, 2008; 4(2):110-117.
https://doi.org/10.1016/j.tree.2008.09.011

2. Kartzinel, TR, Chen, PA, Coverdale, TC, Erickson, DL, Kress, WJ, Kuzmina, ML,
et al. DNA metabarcoding illuminates dietary niche partitioning by African large
herbivores. Proceedings of the National Academy of Sciences, 2015;
112(26):8019-8024. https://doi.org/10.1073/pnas.1503283112

3. Kowalczyk, R, Wójcik, JM, Taberlet, P, Kamiński, T, Miquel, C, Valentini, A, et al.
Foraging plasticity allows a large herbivore to persist in a sheltering forest
habitat: DNA metabarcoding diet analysis of the European bison. Forest
Ecology and Management, 2019; 449:117474.
https://doi.org/10.1016/j.foreco.2019.
4. Gebremedhin, B, Flagstad, Ø, Bekele, A, Chala, D, Bakkestuen, V, Boessenkool, S, et al. DNA metabarcoding reveals diet overlap between the endangered *Walia* *ibex* and domestic goats—implications for conservation. *PLoS One*, 2016; 11(7), p.e0159133. https://doi.org/10.1371/journal.pone.0159133

5. Lee, T, Alemseged, Y, Mitchell, A. Dropping Hints: Estimating the diets of livestock in rangelands using DNA metabarcoding of faeces. *Metabarcoding and Metagenomics*, 2018; 2:1-17. https://doi.org/10.3897/mbmg.2.22467

6. Deagle, BE, Thomas, AC, McInnes, JC, Clarke, LJ, Vesterinen, EJ, Clare, EL, et al. Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, 2019; 28(2):391-406. https://doi.org/10.1111/mec.14734

7. Pegard, A, Miquel, C, Valentini, A, Coissac, E, Bouvier, F, Francois, D, et al. Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, 2009; 57(13):5700-5706. https://doi.org/10.1021/jf803680c

8. Bhargava, M, Sharma, A. DNA barcoding in plants: evolution and applications of in silico approaches and resources. *Molecular Phylogenetics and Evolution*, 2013; 67(3):631-641. https://doi.org/10.1016/j.ympev.2013.03.002

9. Hebert, PD, Ratnasingham, S, De Waard, JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2003; 270: S96–S99. https://doi.org/10.1098/rsbl.2003.0025

10. Galan, M, Pons, JB, Tournayre, O, Pierre, E, Leuchtmann, M, Pontier, D, et al. Metabarcoding for the parallel identification of several hundred predators and their...
prey: Application to bat species diet analysis. Molecular Ecology Resources, 2018; 18(3):474-489. https://doi.org/10.1111/1755-0998.12749

11. Kang, W, Anslan, S, Börner, N, Schwarz, A, Schmidt, R, Künzel, S, et al. Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: The effect of sample-size and bioinformatics on the identified communities. Ecological Indicators. 2021; 121. https://doi.org/10.1016/j.ecolind.2020.107070.

12. Collins, RA, Cruickshank, RH. The seven deadly sins of DNA barcoding. Molecular Ecology Resources, 2013; 13(6):969-975. https://doi.org/10.1111/1755-0998.12046

13. Mallott, EK, Garber, PA, Malhi, RS. trnL outperforms rbcL as a DNA metabarcoding marker when compared with the observed plant component of the diet of wild white-faced capuchins (Cebus capucinus, Primates). PloS one, 2018; 13(6). https://doi.org/10.1371/journal.pone.0199556

14. Taberlet, P, Coissac, E, Pompanon, F, Gielly, L, Miquel, C, Valentini, A, et al. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Research, 2007; 35(3). https://doi.org/10.1093/nar/gkl938

15. Valentini, A, Miquel, C, Nawaz, MA, Bellemain, EVA, Coissac, E, Pompanon, F, et al. New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Molecular Ecology Resources, 2009; 9(1):51-60. https://doi.org/10.1111/j.1755-0998.2008.02352.x
16. Kress, WJ, García-Robledo, C, Uriarte, M, Erickson, DL. DNA barcodes for ecology, evolution, and conservation. Trends in Ecology & Evolution, 2015; 30(1):25-35. https://doi.org/10.1016/j.tree.2014.10.008

17. Matesanz, S, Pescador, DS, Pías, B, Sánchez, AM, Chacón-Labella, J, Illuminati, A, et al. Estimating belowground plant abundance with DNA metabarcoding. Molecular Ecology Resources, 2019; 19(5): 1265-1277. https://doi.org/10.1111/1755-0998.13049

18. Ando, H, Fujii, C, Kawanabe, M, Ao, Y, Inoue, T, Takenaka, A. Evaluation of plant contamination in metabarcoding diet analysis of a herbivore. Scientific Reports 2018; 8(1):1-10. https://doi.org/10.1038/s41598-018-32845-w

19. Gill, BA, Musili, PM, Kurukura, S, Hassan, AA, Goheen, JR, Kress, WJ, et al. Plant DNA-barcode library and community phylogeny for a semi-arid East African savanna. Molecular Ecology Resources, 2019; 19(4): 838-846. https://doi.org/10.1111/1755-0998.13001

20. Linder, HP. The evolution of African plant diversity. Frontiers in Ecology and Evolution, 2014; 2:38. https://doi.org/10.3389/fevo.2014.00038

21. Scott-Shaw, R, Morris, CD. Grazing depletes forb species diversity in the mesic grasslands of KwaZulu-Natal, South Africa. African Journal of Range & Forage Science, 2015; 32(1): 21-31. https://doi.org/10.2989/10220119.2014.901418

22. Zaloumis, NP, Bond, WJ. Reforestation or conservation? The attributes of old growth grasslands in South Africa. Philosophical Transactions of the Royal Society B: Biological Sciences, 2016; 371(1703). https://doi.org/10.1098/rstb.2015.0310.
23. Siebert, F, Dreber, N. Forb ecology research in dry African savannas: Knowledge, gaps, and future perspectives. Ecology and Evolution, 2019; 9(13): 7875-7891. https://doi.org/10.1002/ece3.5307

24. Harris, DJ. Can you bank on GenBank?. Trends in Ecology & Evolution, 2003; 18(7): 317-319. https://doi.org/10.1016/S0169-5347(03)00150-2

25. Hollingsworth, PM, Graham, SW, Little, DP. Choosing and using a plant DNA barcode. PloS one, 2011; 6(5). https://doi.org/10.1371/journal.pone.0019254.

26. Meyer, CP, Paulay, G. DNA barcoding: error rates based on comprehensive sampling. PLoS Biology, 2005; 3(12). https://doi.org/10.1371/journal.pbio.0030422.

27. Sankaran, M., Hanan, NP, Scholes, RJ, Ratnam, J, Augustine, DJ, Cade, BS, et al. Determinants of woody cover in African savannas. Nature, 2005; 438(7069): 846-849. https://doi.org/10.1038/nature04070

28. Klem, J. Drought responses of forb and grass communities in communal and protected rangelands [dissertation]. Potchefstroom (SA): North-West University; 2019.

29. Levin, RA, Wagner, WL, Hoch, PC, Nepokroeff, M, Pires, JC, Zimmer, EA, et al. Family-level relationships of Onagraceae based on chloroplast rbcL and ndhF data. American Journal of Botany, 2003; 90:107-115. https://doi.org/10.3732/ajb.90.1.107

30. Kress, WJ, Erickson, DL. A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS one, 2007; 2(6). https://doi.org/10.1371/journal.pone.0000508
31. Taberlet, P, Gielly, L, Pautou, G, Bouvet, J. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant molecular biology, 1991; 17(5):1105-1109. https://doi.org/10.1007/BF00037152

32. Edgar, RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004; 32(5):1792-1797. https://doi.org/10.1093/nar/gkh340

33. Drew, BT, Ruhfel, BR, Smith, SA, Moore, MJ, Briggs, BG, Gitzendanner, M., et al. Another look at the root of the angiosperms reveals a familiar tale. Systematic Biology, 2014; 63(3):368-382. https://doi.org/10.1093/sysbio/syt108

34. Gitzendanner, MA, Soltis, PS, Wong, GKS, Ruhfel, BR, Soltis, DE. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. American Journal of Botany, 2018; 105(3):291-301. https://doi.org/10.1002/ajb2.1048

35. Wright, ES. Using DECIPHER v2.0 to analyze big biological sequence data in R. R Journal, 2016; 8(1):352-359. https://doi.org/10.32614/RJ-2016-025

36. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 1980; 16(2):111-120. https://doi.org/10.1007/BF01731581

37. Paradis, E, Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 2019; 35(3):526-528. https://doi.org/10.1093/bioinformatics/bty633

38. Brown, SD, Collins, RA, Boyer, S, Lefort, MC, Malumbres-Olarte, JAGOBA, Vink, CJ, et al. Spider: an R package for the analysis of species identity and evolution,
with particular reference to DNA barcoding. Molecular Ecology Resources, 2012; 12(3):562-565. https://doi.org/10.1111/j.1755-0998.2011.03108.x

39. Niemann, HJ, Bezeng, BS, Orton, RD, Kabongo, RM, Pilusa, M, Van der Bank, M. Using a DNA barcoding approach to facilitate biosecurity: Identifying invasive alien macrophytes traded within the South African aquarium and pond plant industry. South African Journal of Botany, 2022; 144:364-376. https://doi.org/10.1016/j.sajb.2021.08.041

40. Meier, R, Zhang, G, Ali, F. The use of mean instead of smallest interspecific distances exaggerates the size of the “barcoding gap” and leads to misidentification. Systematic Biology, 2008; 57(5):809-813. https://doi.org/10.1080/10635150802406343

41. Duda, R., Hart, PE, Stork, DG. Pattern classification and scene analysis. 2nd ed. New York: Wiley. 1973; Vol. 3, pp. 731-739.

42. Meier, R, Shiyang, K, Vaidya, G, Ng, PK. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology, 2006; 55(5):715-728. https://doi.org/10.1080/10635150600969864

43. Ratnasingham, S, Hebert, PD. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes, 2007; 7(3):355-364. https://doi.org/10.1111%2Fj.1471-8286.2007.01678.x

44. Goldberg, AR, Conway, CJ, Tank, DC, Andrews, KR, Gour, DS, Waits, LP. Diet of a rare herbivore based on DNA metabarcoding of feces: Selection, seasonality, and survival. Ecology and Evolution, 2020; 10(14):7627-7643. https://doi.org/10.1002/ece3.6488
45. Iwanowicz, DD, Vandergast, AG, Cornman, RS, Adams, CR, Kohn, JR, Fisher, RN, et al. Metabarcoding of fecal samples to determine herbivore diets: A case study of the endangered Pacific pocket mouse. PloS one, 2016; 11(11): p.e0165366. https://doi.org/10.1371/journal.pone.0165366

46. Meiklejohn, KA, Damaso, N, Robertson, JM. Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PloS one, 2019; 14(6):p.e0217084. https://doi.org/10.1371/journal.pone.0217084

47. Turunen, J, Mykrä, H, Elbrecht, V, Steinke, D, Braukmann, T, Aroviita, J. The power of metabarcoding: Can we improve bioassessment and biodiversity surveys of stream macroinvertebrate communities? Metabarcoding and Metagenomics, 2021; 5:99-110. https://doi.org/10.3897/mbmg.5.68938

48. Mishra, P, Kumar, A, Sivaraman, G, Shukla, AK, Kaliamooorthy, R, Slater, A, et al. Character-based DNA barcoding for authentication and conservation of IUCN Red listed threatened species of genus Decalepis (Apocynaceae). Scientific Reports, 2017; 7(1):1-12. https://doi.org/10.1038/s41598-017-14887-8.

49. Cognato, AI. Standard percent DNA sequence difference for insects does not predict species boundaries. Journal of Economic Entomology, 2006; 99(4), pp.1037-1045. https://doi.org/10.1093/jee/99.4.1037

50. Little, DP, Knopf, P, Schulz, C. DNA barcode identification of Podocarpaceae— the second largest conifer family. PLoS One, 2013; 8(11): p.e81008. https://doi.org/10.1371/journal.pone.0081008

51. CBOL Plant Working Group. A DNA barcode for land plants. Proceedings of the National Academy of Sciences, 2009; 106(31):12794-12797. https://doi.org/10.1073/pnas.0905845106
Genera and species per family contained in the DNA reference database

Figure
