DR. BRIAN ANDREW GILL (Orcid ID : 0000-0001-9567-8989)

DR. TYLER R. KARTZINEL (Orcid ID : 0000-0002-8488-0580)

Article type : Resource Article

Plant DNA-barcode library and community phylogeny for a semi-arid East African savanna

Brian A. Gill,[1] Paul M. Musili,[2] Samson Kurukura,[3] Abidikadir A. Hassan,[3] Jacob R. Goheen,[4] W. John Kress,[5] Maria Kuzmina,[6] Robert M. Pringle,[7] Tyler R. Kartzinel[1,8]

[1]Institute for Environment and Society, Brown University, Providence, RI, USA, [2]East African Herbarium, National Museums of Kenya, Nairobi, Kenya, [3]Mpala Research Centre, Nanyuki, Kenya, [4]Departments of Zoology & Physiology, University of Wyoming, Laramie, WY, USA, [5]National Museum of Natural History, Smithsonian Institution, Washington, DC, USA, [6]Center for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada, [7]Department of Ecology & Evolutionary Biology, Princeton University, Princeton, NJ, USA, [8]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

Keywords: biodiversity, barcode gap, comparative phylogenetics, East Africa, Forest Global Earth Observatory, Mpala Research Centre

Corresponding author: Tyler Kartzinel

85 Waterman Street, Providence, RI 02912

Fax: 401-863-3839; tyler_kartzinel@brown.edu

Running title: Plant DNA barcodes and phylogeny from Kenya

Abstract

Applications of DNA barcoding include identifying species, inferring ecological and evolutionary relationships between species, and DNA metabarcoding. These applications require reference libraries that are not yet available for many taxa and geographic regions. We collected, identified, and vouchered plant specimens from Mpala Research Center in Laikipia, Kenya, to develop an extensive DNA-barcode library for a savanna ecosystem in equatorial East Africa. We amassed up to five DNA barcode markers (*rbcL*, *matK*, *trnL-F*, *trnH–psbA*, and ITS) for 1,781 specimens representing up to 460 species (~92% of the known flora), increasing the number of plant DNA barcode records for Africa by ~9%. We evaluated the ability of these markers, singly and in combination, to delimit species by calculating intra- and inter-specific genetic distances. We further estimated a plant community phylogeny and demonstrated its utility by testing if evolutionary relatedness could predict the tendency of members of the Mpala plant community to have or lack "barcode gaps", defined as disparities between the maximum intra- and minimum inter-specific genetic distances. We found barcode gaps for 72–89% of taxa depending on the marker or markers used. With the exception of the markers *rbcL* and ITS, we found that evolutionary relatedness was an important predictor of barcode-gap presence or absence for all of the markers in combination and for matK, trnL-F, and trnH–psbA individually. This plant DNA barcode library and community phylogeny will be a valuable resource for future investigations.

Introduction

DNA barcoding (Hebert, Cywinska, Ball, & DeWaard, 2003) is a tool to identify species and infer the ecological and evolutionary relationships between them. Determining that a "barcode gap" is present at a locus enables researchers to reliably differentiate species from each other. Biologists use DNA barcodes to study food webs (Pompanon et al., 2012), biological invasions (Dejean et al., 2012), and for biomonitoring (Baird & Hajibabaei, 2012), wildlife forensics, (Dawnay, Ogden, McEwing, Carvalho, & Thorpe, 2007) and natural product validation (Filonzi, Chiesa, Vaghi, & Marzano, 2010). Genetic information provided by DNA barcodes can also be used to estimate phylogenies that yield insights into community assembly (Erickson et al., 2014; Swenson, 2012), trait evolution (Gill et al., 2016), and lineage diversification (Polato et al., 2018). To attain this wide range of uses, investigators rely upon the availability of DNA-barcode libraries, which do not yet exist for many taxa and for many of the world's most biodiverse regions.

As of January 2019, the Barcode of Life Data System (BOLD)(Ratnasingham & Hebert, 2007) included 20,867 records from vascular plant specimens in Africa. Although Africa comprises ~20% of the global landmass and harbors ~15% of the global plant diversity (Linder, 2014), African plants comprise only ~8% of the vascular plant records available in BOLD. Within Africa, most of the available DNA barcodes for plants are from Western and Southern Africa (13,998 specimens; 67% of African specimens)(Bezeng et al., 2017; Lahaye et al., 2008; Parmentier et al., 2013). These global and continental biases in the availability of genetic resources for plants limit both basic and applied research priorities in Africa (Daru, Berger, & van Wyk, 2016).

Existing African plant DNA barcodes have helped resolve the systematics of ecologically and economically important taxa, including rosewoods (Hassold et al., 2016), acacias (Boatwright, Maurin, & van der Bank, 2015; Kyalangalilwa, Boatwright, Daru, Maurin, & van der Bank, 2013), aloes (Daru et al., 2013; Manning, Boatwright, Daru, Maurin,

& van der Bank, 2014), and the Combretaceae (Gere et al., 2013; Jordaan, Van Wyk, & Maurin, 2011b, 2011a; Maurin, Chase, Jordaan, & Van der Bank, 2010). Researchers in the field of community phylogenetics (also called phylogenetic community ecology), have used African plant DNA barcodes to understand plant community responses to herbivory (Yessoufou et al., 2013), classify biogeographical regions of Southern Africa (Daru, van der Bank, et al., 2016), and to assess the evolutionary history of African cycads (Yessoufou, Bamigboye, Daru, & van der Bank, 2014), underground trees (geoxyles) (Maurin et al., 2014) and thorny savanna plant assemblages (Charles-Dominique et al., 2016). Increasing the taxonomic and geographic coverage of DNA barcodes for African plants will enrich our understanding of these species, communities, and ecosystems.

In equatorial East Africa, extensive research into conservation biology and the structure and function of savanna ecosystems has taken place at the Mpala Research Centre (MRC) in the Laikipia Highlands of central Kenya (0.293 N, 36.898 E; 1700–2000 m asl; Fig. 1). This ~200-km^2 unfenced conservancy and working ranch sustains diverse wildlife and domestic livestock species, and is the site of several long-term manipulative experiments that aim to elucidate the effects of herbivory and environmental change on semi-arid savanna ecosystems (Goheen et al., 2018). These studies include the Kenya Long-term Exclosure Experiment (KLEE) (Young, Okello, Kinyua, & Palmer, 1998), the Glade Legacies And Defaunation Experiment (GLADE) (Augustine & McNaughton, 2006), and the Ungulate Herbivory Under Rainfall Uncertainty experiment (UHURU) (Goheen et al., 2013; Kartzinel et al., 2014). Additionally, MRC hosts the only savanna site currently participating in the Forest Global Earth Observatory (ForestGEO) network; accordingly, botanical research at MRC contributes to global comparisons of long-term vegetation dynamics.

In 2012, we established a pipeline for the collection of plant vouchers and DNA barcoding at MRC (Kartzinel et al., 2015). We collected both woody and herbaceous plant specimens throughout the property and neighboring landscapes. For each morphospecies

identified in the field, we sequenced up to five plant DNA-barcode markers for one to four specimens per taxon. At the National Museums of Kenya's East Africa Herbarium (EA), expert botanists examined specimens and made taxonomic determinations, revising field-based morphospecific identifications as necessary. Here, we present our DNA barcode reference library for MRC, estimate a community phylogeny, and demonstrate the utility of these resources by testing if relatedness can predict the presence or absence of barcode gaps within this plant community.

Materials and methods

Site description, specimen collection, and taxonomic identification

The vegetation at MRC includes semi-arid savanna, acacia bushland, and wooded grasslands with interspersed riparian zones and rocky hills. On average, temperatures range from 11–24°C and precipitation totals ~600 mm yr^{-1}, accumulating mainly during three rainy periods (April-May, July-August, November) (Franz, Caylor, Nordbotten, Rodríguez-Iturbe, & Celia, 2010). Distinct habitat types occur on soil types characterized as red sandy loams (northern and southeastern areas), heavy-clay "black-cotton" vertisols (southwestern), and transitional soils between red-sand and black-cotton habitats (Pringle, Prior, Palmer, Young, & Goheen, 2016). This diversity of habitats types is represented across the ~200-km^2 MRC landscape, and similar soil types supporting similar plant communities occur throughout the 9,500-km^2 Laikipia region and more broadly across East Africa, including Nairobi National Park and parts of the Serengeti-Mara ecosystem.

From 2012 until 2018, we sampled plant species as extensively and thoroughly as possible from all vegetation zones in this ecosystem. Initial collections occurred within the UHURU (Goheen et al., 2013; Kartzinel et al., 2014) and KLEE (Young et al., 1998) experiments, where plant surveys are conducted at regular intervals. Subsequent collections spanned the extensive road network of MRC and the surrounding landscapes. Collection

efforts were led by parataxonomists with >10 years of botanical research experience at MRC, with specialized training directed by botanical experts at EA, and supplemented by input from ecologists at multiple institutions. We collected voucher specimens from three individuals per morphospecies and deposited them in the EA, the Smithsonian Institution herbarium (US), and MRC's research and teaching collection. A tissue sample for genetic analyses was collected from each of these specimens, along with up to one additional individual from the field (not vouchered).

At the time of collection, we provisionally identified the specimens, took photographs, and recorded GPS coordinates. Multiple researchers contributed to this collection effort and not all specimens could reliably be identified to species upon initial collection; consequently, some taxa were collected on multiple occasions and we obtained DNA barcodes from as many as 16 specimens per taxon. Expert botanists at the EA identified the specimens to the finest taxonomic level possible (~96% species-level identifications). We reserve use of the word "species" to refer to the 433 taxa assigned accepted Latin binomials in our dataset, and we more inclusively use the word "taxa" with reference to an additional 27 provisionally distinguishable taxonomic entities that are currently only resolved to family- or genus-level (N = 460 taxa in total). We present these data using the taxonomic nomenclature currently recognized by the Angiosperm Phylogeny Group (Chase et al., 2016) and The Plant List (The Plant List, 2013). Because this nomenclature follows the controversial splitting of African *Acacia* spp. into the genera *Senegalia* and *Vachellia* (Smith & Figueiredo, 2011), these latter names appear in our tables and figures, but we refer to "acacias" inclusively.

Generation of DNA barcodes

We used standard protocols to bidirectionally sequence five markers commonly used for plant DNA barcoding: *rbcL*, *matK*, *trnL-F*, *trnH–psbA*, and ITS (protocols in Text S1; primers in Table S1) (Fazekas, Kuzmina, Newmaster, & Hollingsworth, 2012; Ivanova,

Fazekas, & Hebert, 2008; Kartzinel et al., 2015; Kuzmina et al., 2017). Sequences (4,696) from a subset of specimens that we collected between 2012 and 2015 were initially included in a reference library used to support a DNA-metabarcoding study of herbivore diets (Kartzinel et al. 2015). The current dataset includes 1,762 new sequences and improvements in the accuracy of the taxonomic determinations that we completed between 2015 and 2018. To build consensus sequences, we trimmed and assembled forward and reverse reads in Geneious R11 (Kearse et al., 2012). The most current version of these DNA barcode sequences is provided as a publicly accessible dataset on BOLD.

Sequence Alignment

To align DNA sequences for genetic-distance, barcode-gap, and phylogenetic analyses, we used the R (R Core Team, 2017) package DECIPHER (Wright, 2016). These plant DNA barcode datasets included substantial sequence-length variation arising both from insertion-deletion polymorphisms and incomplete Sanger sequence reads. We considered including all sequences in these datasets meeting the minimum acceptable sequence-length thresholds set by Genbank—100 bp for coding genes and 200 bp for non-coding genes—but found that confidently establishing significant sequence homology in alignments required us to reduce the number of short sequences in further analyses. We therefore excluded the shortest ~25% of sequences obtained for each marker from further analyses. To evaluate the identities of specimens that are not identified to the species-level (N = 27 out of 460 taxa), representatives of these taxa were included in the alignment used to estimate the community phylogeny. However, to prevent taxonomic uncertainty from influencing genetic-distance and barcode gap analyses, we included only specimens that had been identified to species-level (N = 433 species).

For *rbcL*, we aligned all sequences simultaneously using AlignSeqs(). For *matK*, we used AlignTranslation() to align the amino acid translation of our DNA sequences and back-translated to DNA. Because distantly related species often have highly divergent sequences for *trnL-F*, *trnH–psbA*, and ITS, we split the sequence data to align by family using the R package PHYLOTOOLS (Zhang, 2017) and aligned using AlignSeqs(). For analyses considering data from all markers together, we concatenated all alignments to create a supermatrix of all markers, including markers that we had split into separate alignments by plant family (W. J. Kress et al., 2009).

Calculation of intra- and inter-specific genetic distances

To assess levels of genetic variation within and among plant species in our dataset, we calculated uncorrected intra- and inter-specific genetic distances for each marker separately (*rbcL*, *matK*, *trnL-F*, *trnH–psbA*, and ITS) and for all markers together (*rbcL* + *matK* + *trnL-F* + *trnH–psbA* + ITS) using DistanceMatrix() in DECIPHER (Wright, 2016). To limit the influence of missing data, terminal gaps, gap-to-letter matches, and gap-to-gap matches were not included in the calculation of distances. Pairwise comparisons of genetic distances among all species were conducted for the global alignments of *rbcL*, *matK*, and the supermatrix of all markers. By-family alignments for *trnL-F, trnH–psbA*, and ITS allowed for pairwise comparisons at the family level only. We evaluated the presence of "barcode gaps" based on these genetic distances, which can be identified based on the disparities between the maximum intra-specific and minimum inter-specific genetic distances (Hebert et al., 2003).

Estimation of community phylogeny

We built an alignment that included only a single specimen per taxon, choosing the specimen represented by the most available sequence data across at least three markers. We tested for the best model of nucleotide substitution and partitioning scheme to use in subsequent phylogenetic reconstruction steps using the program Partition Finder 2 (Lanfear, Frandsen, Wright, Senfeld, & Calcott, 2016). Based on AIC_c, the best model of nucleotide substitution was GTR + I + Γ which we applied to each of the following partitions: 1) *rbcL* codon position one, 2) *rbcL* codon position two, 3) *rbcL* codon position three, 4) *matK* codon positions one and two, 5) *matK* codon position three, 6) *trnL-F*, 7) *trnH–psbA*, and 8) ITS. To ensure accurate reconstruction of established phylogenetic relationships, we constrained family-level relationships known *a priori* using the "R20120829" tree, available from Phylomatic (Webb & Donoghue, 2005) in the R package BRRANCHING (Chamberlain, 2016). Using the supermatrix, constraint tree, GTR + I + Γ model of nucleotide substitution, and partitioning scheme determined using Partition Finder 2, we ran a maximum-likelihood analysis to estimate a phylogeny in the program RAxML (Stamatakis, 2014) through the CIPRES Science Gateway (Miller, Pfeiffer, & Schwartz, 2010). To maximize the RAxML phylogeny's utility for this and future studies, we needed to rescale its branch lengths to represent absolute time. We applied Phylocom's bladj function (Webb, Ackerly, & Kembel, 2008) to the RAxML phylogeny based on 36 fossil calibration dates (Bell, Soltis, & Soltis, 2010; Gastauer & Meira-Neto, 2016). We used the R package MONOPHY (Schwery & O'Meara, 2016) to test the monophyly of genera. For this analysis, we maintained supported nodes and collapsed unsupported nodes (bootstrap support < 63) to polytomies (Farris, Albert, Kallersjo, Lipscomb, & Kluge, 1996).

Testing for phylogenetic signal in barcode gaps

Phylogenetic signal is the tendency of closely related species to resemble each other because of shared evolutionary history (Felsenstein, 1985). While phylogenetic signal is ubiquitous, some evolutionary processes (Hansen & Martins, 1996) and properties of data sets such as incomplete taxonomic sampling (Blomberg, Garland, & Ives, 2003; Cavender-Bares, Keen, & Miles, 2006) can result in estimates of weak or no phylogenetic signal. Thus, the presence of phylogenetic signal in a particular local plant assemblage cannot be assumed. To demonstrate the utility of this phylogeny for studies of phylogenetic community ecology, we tested for phylogenetic signal in the presence or absence of barcode gaps for each marker and for all markers combined using continuous-time Markov models of discrete-trait evolution using the R package GEIGER (Harmon, Weir, Brock, Glor, & Challenger, 2008). Specifically, we compared model fits when the tree-transformation parameter λ (Pagel, 1999) was set to zero (no influence of phylogeny) against those obtained when λ was determined by maximum likelihood (potential influence of phylogeny). Significantly better model fits when the tree transformation parameter λ was determined by maximum likelihood indicate significant phylogenetic signal in barcode gap presence or absence. Significant phylogenetic signal in barcode gap presence or absence in the flora of MRC would allow us to identify clades in which DNA-based species identifications would be straightforward or problematic using these markers.

Results

Collection and sequencing success

Of the approximately 500 plant species known or thought to occur at the MRC, to date we have obtained 1,843 specimens from at least 433 species and an additional 27 provisionally distinguished taxa currently resolved to family- or genus-level (N = 460 taxa in total; ~92% of the known flora). These specimens belong to two phyla, three classes, 29

orders, 66 families, and 245 genera. Collectively, this barcode library includes data for 1,781 specimens, including 1,523 *rbcL*, 1,197 *matK*, 1,595 *trnL-F*, 1,260 *trnH–psbA*, and 883 ITS sequences. All sequence data have been published on BOLD as "DS-UHURUR2" (dx.doi.org/10.5883/DS-UHURUR2) and on Genbank (accessions in Supplementary Dataset S1).

Genetic distances and barcode gaps

The information provided by all markers combined in the supermatrix revealed barcode gaps for 311 of the 429 well-identified species that we assessed (72%). Considering each of the five markers separately, 73–89% of species exhibited a barcode gap (*rbcL* = 231 of 316 (73%), *matK* = 243 of 315 (77%), *trnL-F* = 257 of 324 (79%), *trnH–psbA* = 227 of 289 (79%) and ITS = 159 of 178 (89%); Fig. 2; Supplementary Dataset S2).

Barcode phylogeny and phylogenetic signal in barcode gaps

Our final phylogeny includes 324 taxa (70%) for which we obtained sufficiently long sequence data (Fig. 3). We visualized the results in detail using subtrees for the monocots (Fig. S1), superastrids (Fig. S2), and superrosids (Fig. S3). Of the 186 genera included in the tree, 40 were monophyletic, 20 were not monophyletic, and 126 were monotypic (precluding tests of monophyly). Many of the non-monophyletic genera were members of the orders Poales, Asterales, Lamiales, and Malvales (Figs. S4–S6).

We detected significant phylogenetic signal in the presence or absence of barcode gaps for all markers together ($\chi2 = 14.976$, df = 1, $P < 0.001$), and singly for *matK* ($\chi2 = 21.085$, df = 1, $P < 0.001$) and *trnH–psbA* ($\chi2 = 6.746$, df = 1, $P = 0.009$). Phylogenetic signal for *trnL-F* ($\chi2 = 3.122$, df = 1, $P = 0.077$) was marginally significant. We did not detect significant phylogenetic signal for *rbcL* ($\chi2 = 2.357$, df = 1, $P = 0.125$) or ITS ($\chi2 = 1.875$, df

= 1, P = 0.171). Thus, for four of the six markers or marker combinations tested, relatedness was an important predictor of whether a particular species exhibits a barcode gap in this community. Orders with many species lacking barcode gaps included Poales, Malvales, Lamiales, and Fabales (Table S2; Figs S7–S12). These classifications corresponded to many ecologically important savanna species including grasses (Poaceae), sedges (Cyperaceae), acacias (Fabaceae), and mallows (Malvaceae), and represent many of the same clades that included non-monophyletic genera in this community.

Discussion

Our DNA barcode dataset provides a publicly accessible record of an ongoing and long-term botanical inventory of MRC and the surrounding region. The concordance between morphological species identifications and DNA barcodes indicates that this reference library is generally reliable for species identification (72–89% marker resolution, depending on the marker used). Furthermore, the robust community phylogeny presented here will enable more detailed analyses of this plant community and local species interactions in an evolutionary context. Over time, we expect to increase our barcode coverage of the MRC flora and intend to publish updated versions of record for the dataset.

The development of a comprehensive plant DNA-barcode library for MRC has proven challenging and will require continued collection and taxonomic efforts. Some plant species rarely produce fertile specimens, are only distinguishable by their reproductive structures, or are dioecious, undescribed, or cryptic (Dick & Kress, 2009). Furthermore, even if specimens of all local species can ultimately be identified with complete accuracy, DNA barcoding rarely provides perfect discriminatory ability. Here, consistent with the limitations reported in other plant DNA-barcode studies (Braukmann, Kuzmina, Sills, Zakharov, & Hebert, 2017; CBOL Plant Working Group et al., 2009; Chase et al., 2005; Hollingsworth, Graham, & Little, 2011; W. J. Kress et al., 2009; W. John Kress & Erickson, 2007; W John Kress, Wurdack, Zimmer,

Weigt, & Janzen, 2005; Lahaye et al., 2008; Li et al., 2015; Newmaster, Fazekas, Steeves, & Janovec, 2008), this dataset achieves a high degree of taxonomic resolution that is nevertheless imperfect.

Analyses of intra- and inter-specific genetic distances and the presence or absence of barcode gaps reveal the relative utility of different DNA barcode markers and can help identify future research priorities. Phylogenetic signal in barcode gaps indicates that the efficacy of DNA barcoding for species identification in this plant community can be predicted based on knowledge of the evolutionary relationships among species. Species representing diverse and abundant savanna plants are particularly prone to the absence of a barcode gap, and thus sequencing efforts in the future can be scaled up or down depending on the expected marker resolution for the species groups of interest. Studies involving easily resolved clades might require only a subset of the markers utilized here to achieve perfect discrimination capabilities, while other clades might require more detailed genomic analyses (Li et al., 2015). Indeed, prime candidates for whole-chloroplast sequencing or "genome skimming" (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016) include many grasses (Poaceae), sedges (Cyperaceae), acacias (Fabaceae), and mallows (Malvaceae) for which these standard DNA barcoding protocols appear insufficient.

This DNA barcode release increases the number of plant DNA barcodes from Africa in the BOLD database by 9%, providing a valuable resource for research in the region and filling a recognized gap in the availability of genetic resources for East Africa, and for drylands and savannas and in general (Daru, Berger, et al., 2016). These ecosystems cover more than half of the African continent (Werner, 1991) and traverse a variety of temperature, rainfall, and soil conditions (House, Archer, Breshears, Scholes, & NCEAS Tree–Grass Interactions Participants, 2003). The plants that occur within them, are shaped by herbivory, fire, and land use (Charles-Dominique et al., 2016), and exist at an unstable equilibrium that is prone to phase shifts toward a steady state as either forest or grassland (Staver, Archibald, & Levin, 2011). Worldwide, the study of savanna floras is critical because

although savannas vary in vegetation composition and environmental conditions, their characteristic coexistence of trees and grasses provides important habitat for biodiversity and underpins human livelihoods (House et al., 2003; Scholes & Archer, 1997).

We expect that this plant DNA-barcode library will support on-going and future research by providing a well-curated taxonomy for the local flora and improving opportunities to compare and coordinate among the many research programs hosted by MRC. It will further support phytogeographical research across Africa and worldwide through MRC's participation in the global ForestGEO network. Ongoing investigations enabled by the construction of this library include the evaluation of putative new species, community and comparative phylogenetic analyses, and forensic ecological investigations through dietary DNA metabarcoding. This research will contribute to our collective understanding of savanna biodiversity and provide much-needed genetic resources for the region's flora.

Acknowledgements

Grace Charles, Elise DeFranco, Rhianna Hohbein, Mwadime, Truman Young contributed to specimen collection and identification. Patricia Chen, David Erickson, Christina Hansen, Johan Pansu, and Caroline Puente assisted with laboratory analyses. We thank the Government of Kenya for permission to conduct this research, which was supported a NatureNet Fellowship from The Nature Conservancy to TRK, National Science Foundation DEB-1457679 and DEB-1556728 to JRG, and DEB-1355122 and IOS-1656527 to RMP. The UHURU experiment was built with a Natural Sciences and Engineering Council Research Tools and Instruments grant.

References

Augustine, D. J., & McNaughton, S. J. (2006). Interactive effects of ungulate herbivores, soil fertility, and variable rainfall on ecosystem processes in a semi-arid savanna. *Ecosystems*, *9*, 1242–1256. doi:10.1007/s10021-005-0020-y

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, *21*(8), 2039–2044. doi:10.1111/j.1365-294X.2012.05519.x

Bell, C. D., Soltis, D. E., & Soltis, P. S. (2010). The age and diversification of the angiosperms re-revisited. *American Journal of Botany*, *97*(8), 1296–1303. doi:10.3732/ajb.0900346

Bezeng, B. S., Davies, T. J., Daru, B. H., Kabongo, R. M., Maurin, O., Yessoufou, K., … van der Bank, M. (2017). Ten years of barcoding at the African Centre for DNA Barcoding. *Genome*, *60*, 629–638. doi:10.1139/gen-2016-0198

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile. *Evolution*, *57*(4), 717–745. doi:doi.org/10.1111/j.0014-3820.2003.tb00285.x

Boatwright, J. S., Maurin, O., & van der Bank, M. (2015). Phylogenetic position of Madagascan species of Acacia s.l. and new combinations in Senegalia and Vachellia (Fabaceae, Mimosoideae, Acacieae). *Botanical Journal of the Linnean Society*, *179*, 288–294. doi:10.1111/boj.12320

Braukmann, T. W. A., Kuzmina, M. L., Sills, J., Zakharov, E. V., & Hebert, P. D. N. (2017). Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS ONE*, *12*(1), 1–19. doi:10.1371/journal.pone.0169515

Cavender-Bares, J., Keen, A., & Miles, B. (2006). Phylogenetic Structure of Floridian Plant Communities Depends on Taxonomic and Spatial Scale. *Ecology*, *87*(7).

CBOL Plant Working Group, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797. doi:10.1073/pnas.0905845106

Chamberlain, S. (2016). brranching: Fetch “Phylogenies” from Many Sources. Retrieved from https://cran.r-project.org/package=brranching

Charles-Dominique, T., Davies, T. J., Hempson, G. P., Bezeng, B. S., Daru, B. H., Kabongo, R. M., … Bond, W. J. (2016). Spiny plants, mammal browsers, and the origin of African savannas. *Proceedings of the National Academy of Sciences*, *113*(38), E5572–E5579. doi:10.1073/pnas.1607493113

Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., … Weber, A. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, *181*(1), 1–20. doi:10.1111/boj.12385

Chase, M. W., Salamin, N., Wilkinson, M., Dunwell, J. M., Kesanakurthi, R. P., Haidar, N., & Savolainen, V. (2005). Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1889–1895. doi:10.1098/rstb.2005.1720

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428. doi:10.1111/mec.13549

Daru, B. H., Berger, D. K., & van Wyk, A. E. (2016). Opportunities for unlocking the potential of genomics for African trees. *New Phytologist*, *210*(3), 772–778. doi:10.1111/nph.13826

Daru, B. H., Manning, J. C., Boatwright, J. S., Maurin, O., Maclean, N., Schaefer, H., … van der Bank, M. (2013). Molecular and morphological analysis of subfamily Alooideae (Asphodelaceae) and the inclusion of Chortolirion in Aloe. *International Association for Plant Taxonomy*, *62*(1), 62–76.

Daru, B. H., van der Bank, M., Maurin, O., Yessoufou, K., Schaefer, H., Slingsby, J. A., & Davies, T. J. (2016). A novel phylogenetic regionalization of phytogeographical zones of southern Africa reveals their hidden evolutionary affinities. *Journal of Biogeography*, *43*, 155–166. doi:10.1111/jbi.12619

Dawnay, N., Ogden, R., McEwing, R., Carvalho, G. R., & Thorpe, R. S. (2007). Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, *173*(1), 1–6. doi:10.1016/j.forsciint.2006.09.013

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: The example of the American bullfrog Lithobates catesbeianus. *Journal of Applied Ecology*, *49*, 953–959. doi:10.1111/j.1365-2664.2012.02171.x

Dick, C. W., & Kress, W. J. (2009). Dissecting tropical plant diversity with forest plots and a molecular toolkit. *BioScience*, *59*(9), 745–755. doi:10.1525/bio.2009.59.9.6

Erickson, D. L., Jones, F. A., Swenson, N. G., Pei, N., Bourg, N., Chen, W., … Kress, W. J. (2014). Comparative evolutionary diversity and phylogenetic structure across multiple forest dynamics plots: A mega-phylogeny approach. *Frontiers in Genetics*, *5*, 1–14. doi:10.3389/fgene.2014.00358

Farris, J. S., Albert, V. A., Kallersjo, M., Lipscomb, D., & Kluge, A. G. (1996). Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, *12*, 99–124. doi:10.1111/j.1096-0031.1996.tb00196.x

Fazekas, A. J., Kuzmina, M. L., Newmaster, S. G., & Hollingsworth, P. M. (2012). DNA barcoding methods for land plants. In W. J. Kress & D. Erickson (Eds.), *DNA Barcodes. Methods in Molecular Biology (Methods and Protocols)* (pp. 223–252). Totowa, NJ: Humana Press.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15.

Filonzi, L., Chiesa, S., Vaghi, M., & Marzano, F. N. (2010). Molecular barcoding reveals mislabelling of commercial fish products in Italy. *Food Research International*, *43*(5), 1383–1388. doi:10.1016/j.foodres.2010.04.016

Franz, T. E., Caylor, K. K., Nordbotten, J. M., Rodríguez-Iturbe, I., & Celia, M. A. (2010). An ecohydrological approach to predicting regional woody species distribution patterns in dryland ecosystems. *Advances in Water Resources*, *33*(2), 215–230. doi:10.1016/j.advwatres.2009.12.003

Gastauer, M., & Meira-Neto, J. A. A. (2016). An enhanced calibration of a recently released megatree for the analysis of phylogenetic diversity. *Brazilian Journal of Biology*, *76*(3), 619–628. doi:10.1590/1519-6984.20814

Gere, J., Yessoufou, K., Daru, B. H., Mankga, L. T., Maurin, O., & van der Bank, M. (2013). Incorporating trnH-psbA to the core DNA barcodes improves significantly species discrimination within southern African Combretaceae. *ZooKeys*, *365*, 127–147. doi:10.3897/zookeys.365.5728

Gill, B. A., Kondratieff, B. C., Casner, K. L., Encalada, A. C., Flecker, A. S., Gannon, D. G., … Funk, W. C. (2016). Cryptic species diversity reveals biogeographic support for the 'mountain passes are higher in the tropics' hypothesis. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20160553.

Goheen, J. R., Augustine, D. J., Veblen, K. E., Kimuyu, D. M., Palmer, T. M., Porensky, L. M., … Young, T. P. (2018). Conservation lessons from large-mammal manipulations in East African savannas: The KLEE, UHURU, and GLADE experiments. *Annals of the New York Academy of Sciences*, *1429*(1), 31–49. doi:10.1111/nyas.13848

Goheen, J. R., Palmer, T. M., Charles, G. K., Helgen, K. M., Kinyua, S. N., Maclean, J. E., … Pringle, R. M. (2013). Piecewise disassembly of a large-herbivore community across a rainfall gradient: The UHURU experiment. *PLoS ONE*, *8*(2), e55192. doi:10.1371/journal.pone.0055192

Hansen, T. F., & Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, *50*(4), 1404–1417. doi:10.2307/2410878

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*(1), 129–131. doi:10.1093/bioinformatics/btm538

Hassold, S., Lowry, P. P., Bauert, M. R., Razafintsalama, A., Ramamonjisoa, L., & Widmer, A. (2016). DNA barcoding of Malagasy Rosewoods: Towards a molecular identification of CITES-Listed dalbergia species. *PLoS ONE*, *11*, 1–17. doi:10.1371/journal.pone.0157881

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. doi:10.1098/rspb.2002.2218

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA

barcode. *PLoS ONE*, *6*(5), e19254. doi:10.1371/journal.pone.0019254

House, J. I., Archer, S., Breshears, D. D., Scholes, R. J., & NCEAS Tree–Grass Interactions Participants. (2003). Conundrums in mixed woody–herbaceous plant systems. *Journal of Biogeography*, *30*, 1763–1777. doi:10.1046/j.1365-2699.2003.00873.x

Ivanova, N. V., Fazekas, A. J., & Hebert, P. D. N. (2008). Semi-automated, membrane-based protocol for DNA isolation from plants. *Plant Molecular Biology Reporter*, *26*, 186–198. doi:https://doi.org/10.1007/s11105-008-0029-4

Jordaan, M., Van Wyk, A. E., & Maurin, O. (2011a). A conspectus of Combretum (Combretaceae) in southern Africa, with taxonomic and nomenclatural notes on species and sections. *Bothalia*, *41*(1), 135–160.

Jordaan, M., Van Wyk, A. E., & Maurin, O. (2011b). Generic status of Quisqualis (Combretaceae), with notes on the taxonomy and distribution of Q. parviflora. *Bothalia*, *41*(1), 161–169.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., … Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, *112*(26), 8019–8024. doi:10.1073/pnas.1509325112

Kartzinel, T. R., Goheen, J. R., Charles, G. K., DeFranco, E., MacLean, J. E., Otieno, T. O., … Pringle, R. M. (2014). Plant and small-mammal responses to large-herbivore exclusion in an African savanna: Five years of the UHURU experiment. *Ecology*, *95*(3), 787. doi:10.1890/13-1023R.1

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. doi:10.1093/bioinformatics/bts199

Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS ONE*, *2*(6), e508. doi:10.1371/journal.pone.0000508

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences*, *106*(44), 18621–18626. doi:10.1073/pnas.0909820106

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences*, *102*(23), 8369–8374. doi:10.1073/pnas.0503123102

Kuzmina, M. L., Braukmann, T. W. A., Fazekas, A. J., Graham, S. W., Dewaard, S. L., Rodrigues, A., … Hebert, P. D. N. (2017). Using Herbarium-Derived DNAs to Assemble a Large-Scale DNA Barcode Library for the Vascular Plants of Canada. *Applications in Plant Sciences*, *5*(12), 1700079. doi:10.3732/apps.1700079

Kyalangalilwa, B., Boatwright, J. S., Daru, B. H., Maurin, O., & van der Bank, M. (2013). Phylogenetic position and revised classification of Acacia s.l. (Fabaceae: Mimosoideae) in Africa, including new combinations in Vachellia and Senegalia. *Botanical Journal of the Linnean Society*, *172*(4), 500–523. doi:10.1111/boj.12047

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., … Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences*, *105*(8), 2923–2928. doi:10.1073/pnas.0709936105

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2016). Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772–773. doi:10.1093/molbev/msw260

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: From gene to genome. *Biological Reviews of the Cambridge Philosophical Society*, *90*, 157–166. doi:10.1111/brv.12104

Linder, H. P. (2014). The evolution of African plant diversity. *Frontiers in Ecology and Evolution*, *2*(July), 1–14. doi:10.3389/fevo.2014.00038

Manning, J., Boatwright, J. S., Daru, B. H., Maurin, O., & van der Bank, M. (2014). A molecular phylogeny and generic classification of Asphodelaceae subfamily Alooideae: A final resolution of the prickly issue of polyphyly in the Alooids? *Systematic Botany*, *39*(1), 55–74. doi:10.1600/036364414X678044

Maurin, O., Chase, M. W., Jordaan, M., & Van der Bank, M. (2010). Phylogenetic relationships of Combretaceae inferred from nuclear and plastid DNA sequence data: Implications for generic classification. *Botanical Journal of the Linnean Society*, *162*, 453–476. doi:10.1111/j.1095-8339.2010.01027.x

Maurin, O., Davies, T. J., Burrows, J. E., Daru, B. H., Yessoufou, K., Muasya, A. M., … Bond, W. J. (2014). Savanna fire and the origins of the “underground forests” of Africa. *New Phytologist*, *204*, 201–214. doi:10.1111/nph.12936

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Proceedings of the Gateway Computing Environments Workshop* (pp. 1–8).

Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D., & Janovec, J. (2008). Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, *8*, 480–490. doi:10.1111/j.1471-8286.2007.02002.x

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884. doi:10.1038/44766

Parmentier, I., Duminil, J., Kuzmina, M., Philippe, M., Thomas, D. W., Kenfack, D., … Hardy, O. J. (2013). How Effective Are DNA Barcodes in the Identification of African Rainforest Trees? *PLoS ONE*, *8*(4). doi:10.1371/journal.pone.0054921

Polato, N. R., Gill, B. A., Shah, A. A., Gray, M. M., Casner, K. L., Barthelet, A., … Zamudio,

K. R. (2018). Narrow thermal tolerance and low dispersal drive higher speciation in tropical mountains. *Proceedings of the National Academy of Sciences*, *115*(49), 12471–12476. doi:doi.org/10.1073/pnas.1809326115

Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: Diet assessment using next generation sequencing. *Molecular Ecology*, *21*(8), 1931–1950. doi:10.1111/j.1365-294X.2011.05403.x

Pringle, R. M., Prior, K. M., Palmer, T. M., Young, T. P., & Goheen, J. R. (2016). Large herbivores promote habitat specialization and beta diversity of African savanna trees. *Ecology*, *97*(10), 2640–2657. doi:10.1002/ecy.1522

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. doi:10.1111/j.1471-8286.2007.01678.x

Scholes, R. J., & Archer, S. R. (1997). Tree-Grass Interactions. *Annual Review of Ecology and Systematics*, *28*, 517–544.

Schwery, O., & O'Meara, B. C. (2016). MonoPhy: A simple R package to find and visualize monophyly issues. *PeerJ Computer Science*, *2*, e56. doi:10.7717/peerj-cs.56

Smith, G. F., & Figueiredo, E. (2011). Conserving Acacia Mill. with a conserved type: What happened in Vienna? *Taxon*, *60*, 1504–1506.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Staver, A. C., Archibald, S., & Levin, S. (2011). The global extent and determinants of savanna and forest as alternative biome states. *Science*, *334*(6053), 230–232. doi:10.1126/science.1210465

Swenson, N. G. (2012). Phylogenetic analyses of ecological communities using DNA barcode data. In W. J. Kress & D. L. Erickson (Eds.), *DNA Barcodes. Methods in Molecular Biology (Methods and Protocols)* (pp. 409–414). Totowa, NJ: Humana Press.

The Plant List (2013). Version 1.1. Published on the Internet; http://www.theplantlist.org/ (accessed 6th September).

Webb, C. O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, *24*(18), 2098–2100. doi:10.1093/bioinformatics/btn358

Webb, C. O., & Donoghue, M. J. (2005). Phylomatic: Tree assembly for applied phylogenetics. *Molecular Ecology Notes*, *5*(1), 181–183. doi:10.1111/j.1471-8286.2004.00829.x

Werner, P. (1991). *Savanna Ecology and Management: Australian Perspectives and Intercontinental Comparisons*. (Blackwell Science, Ed.). London.

Wright, E. S. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal*, *8*(1), 352–359. doi:V12242009

Yessoufou, K., Bamigboye, S. O., Daru, B. H., & van der Bank, M. (2014). Evidence of constant diversification punctuated by a mass extinction in the African cycads. *Ecology and Evolution*, *4*(1), 50–58. doi:10.1002/ece3.880

Yessoufou, K., Davies, T. J., Maurin, O., Kuzmina, M., Schaefer, H., van der Bank, M., & Savolainen, V. (2013). Large herbivores favour species diversity but have mixed impacts on phylogenetic community structure in an African savanna ecosystem. *Journal of Ecology*, *101*(3), 614–625. doi:10.1111/1365-2745.12059

Young, T. P., Okello, B. D., Kinyua, D., & Palmer, T. M. (1998). KLEE: A long-term multi-species herbivore exclusion experiment in Laikipia, Kenya. *African Journal of Range and Forage Science*, *14*(3), 94–102. doi:10.1080/10220119.1997.9647929

Zhang, J. (2017). phylotools: Phylogenetic tools for Eco-phylogenetics. Retrieved from https://github.com/helixcn/phylotools

Data accessibility:

-R code, alignments, and phylogeny: Dryad doi.org/10.5061/dryad.qk85bp8

-Specimen data and DNA barcodes: BOLD dx.doi.org/10.5883/DS-UHURUR2 and Genbank (accessions listed by record in Supplementary Dataset S1)

Author contributions

RMP and TRK conceived and designed the research; PMM, JRG, SK, and AAH conducted field collections and taxonomic identifications; TRK, MK, and WJK developed the barcoding strategy and performed laboratory analyses; TRK and BAG analyzed the data; BAG drafted the manuscript with subsequent contributions from all authors.

Figure 1. Localities where plant specimens were collected from Mpala Research Centre and surrounding areas in Laikipia, Kenya. Precise location data are available for 1,690 of the 1,781 specimens for which we successfully sequenced at least one marker. This dataset is the result of extensive botanical searches across Mpala's extensive road network, and these points represent the subset of locations where specimens representing a novel species for our dataset were first collected.

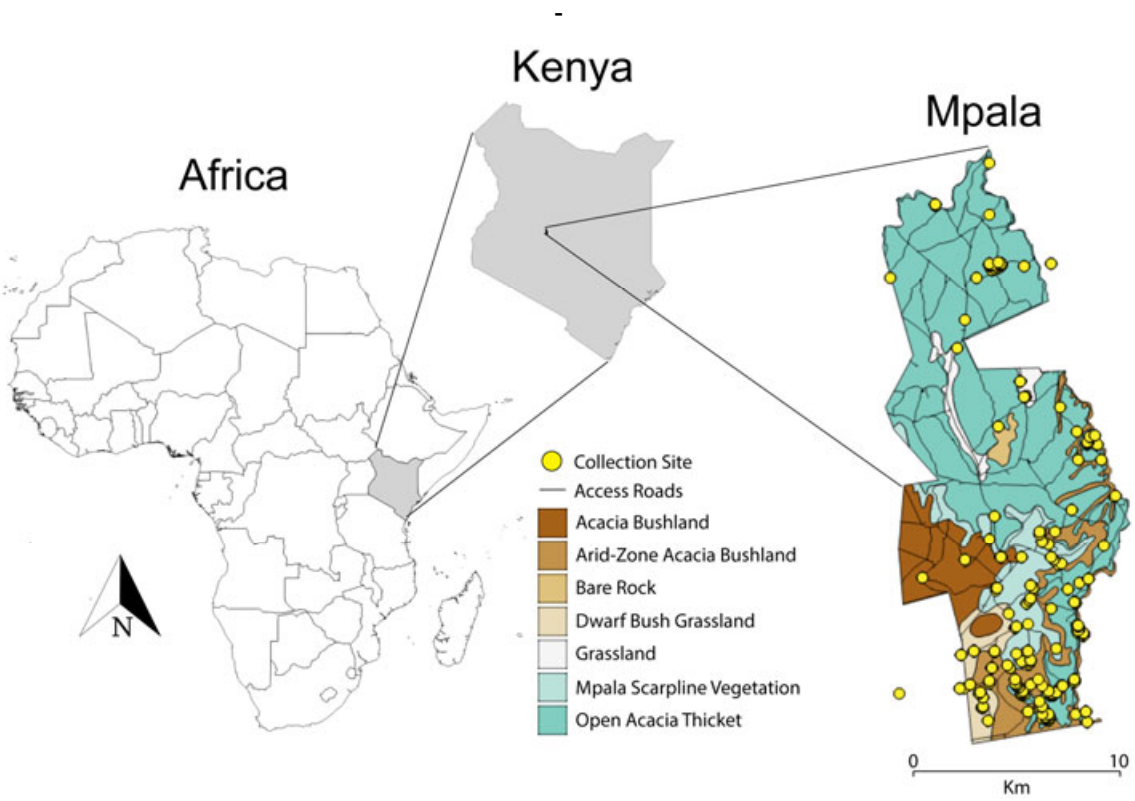

Figure 2. Barcode gap plots for species from Mpala Research Centre, showing the relationships between the maximum intra-specific genetic distance (horizontal axes) and minimum inter-specific distance (vertical axes) for each DNA-barcode marker and all markers combined. A barcode gap is indicated by points that fall above the solid line. Points have been made translucent such that high densities of points result in darker colors. Numbers on upper and lower halves of plots represent the number of species with (top) or without (bottom) barcode gaps.

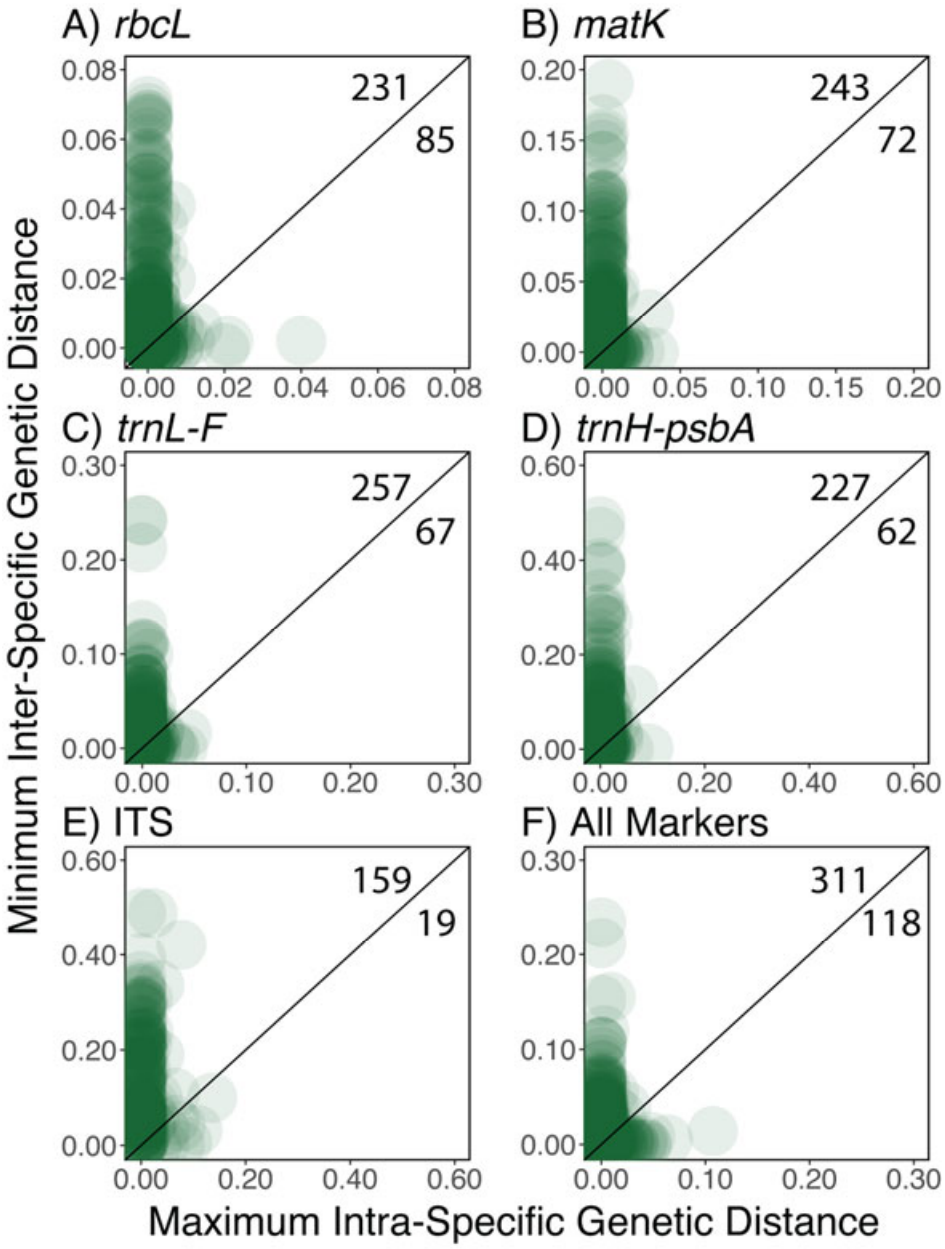

Figure 3. Community phylogeny for the plants of Mpala Research Center, Kenya. The size of wedges represents the number of species within orders. Four major lineages are shown, including Polypodiopsida (“P”), monocots (“M”), superrosids (“SR”), and superasterids (“SA”). Detailed subtrees for 89 monocots, 102 superrosids, and 131 superasterids are shown in Figs. S1–S3. Numbers after order names indicate the number of species represented.