

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

A dataset for benchmarking molecular identification tools based on genome skimming

Renata C. Asprino^{1,2,3}, Liming Cai^{3,4,5}, Yujing Yan³, Peter J. Flynn³, Lucas C. Marinho^{1,3,6}, Xiaoshan Duan^{3,7}, Christiane Anderson⁸, Charles C. Davis³, and Bruno A. S. de Medeiros^{9,10,11}

Affiliations

- 1. Programa de Pós-Graduação em Botânica, Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil
- 2. Botany, School of Natural Sciences, Trinity College Dublin, Dublin, Ireland
- 3. Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge, Massachusetts 02138, USA
- 4. Department of Integrative Biology, The University of Texas at Austin, Austin, Texas 78712, USA
- 5. University of Florida, Gainesville, USA
- 6. Departamento de Biologia, Universidade Federal do Maranhão, São Luís, Maranhão, Brazil
- 7. College of Forestry, Northwest Agriculture & Forestry University, Yangling 712100, Shaanxi, China
- 8. University of Michigan Herbarium, Ann Arbor, Michigan 48108, USA
- 9. Field Museum of Natural History, Chicago, Illinois 60605, USA
- 10. Smithsonian Tropical Research Institute, Panama City, Panama
- 11. Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA

Corresponding co-senior author(s):
Bruno A. S. de Medeiros, bdemedeiros@fieldmuseum.org;
Charles C. Davis, cdavis@oeb.harvard.edu

46

47

48

49 Abstract

50

51 Genome skimming is an emerging tool allowing for scalable DNA barcoding efforts for
52 numerous biodiversity science applications. Despite its growing importance, there are
53 few standardized datasets for benchmarking genome skimming tools, making it
54 challenging to evaluate new methods (e.g., using machine learning), and comparing to
55 existing ones (e.g., conventional barcoding loci derived from Sanger-based sequencing).
56 To address this gap, we present four curated datasets designed for benchmarking
57 molecular identification tools using low-coverage genomes. These datasets comprise
58 vast phylogenetic and taxonomic diversity from closely related species to all taxa
59 currently represented on NCBI SRA. One of them consists of novel sequences from
60 taxonomically verified samples in the plant clade Malpighiales, while the other four
61 datasets compile publicly available data. All include raw genome skim sequences and
62 two-dimensional graphical representations of genomic data (chaos game
63 representations and varKodes), enabling comprehensive testing and validation of
64 molecular species identification methods. These datasets represent a reliable resource
65 for researchers to assess the accuracy, efficiency, and robustness of their tools in a
66 consistent and reproducible manner.

67

68 Background & Summary

69

70 Genome skimming has become a versatile tool for biodiversity science, with broad-
71 reaching applications spanning phylogenetics to species identification^{1,2,3,4,5}. Low-
72 coverage genomic sequencing facilitates the assembly of both traditional DNA-marker
73 barcodes⁶ as well as barcodes that include entire organellar genomes and many nuclear
74 ribosomal genes^{3,7}. Another advantage of genome skimming protocols is that they are
75 robust to DNA quality, being ideal for specimens from Natural History collections which
76 may present degraded DNA⁸. More recently, genome skimming data are being applied
77 for innovative assembly- and alignment-free species identification^{1,9,10}. A large number
78 of methods^{1,10,11,12,13,14,15,16,17,18} have been developed to apply molecular identification
79 and, typically, their accuracy and efficiency are evaluated with a custom dataset. The
80 customized nature of such datasets is potentially problematic because the success of a
81 given method may be dataset-dependent.

82

83 Here, we assert that this problem can be solved with a readily accessible and well-
84 annotated benchmark dataset. Specifically, the use of benchmarking datasets plays an
85 essential role in both testing novel methods and guiding the improvement of existing
86 methods by allowing unbiased method comparison and reduced errors due to data
87 variation^{19,20}. Benchmarking datasets also help to identify and address potentially
88 confounding variables affecting the performance of different methods. These datasets
89 are of widespread interest to computer scientists across different disciplines, each
90 addressing unique challenges within their respective fields. Fields as diverse as text
91 transcription^{21,22}, medical diagnostics^{23,24}, and bioinformatics^{25,26} have invested in

92 developing standardized datasets to facilitate the validation and comparison of
93 analytical tools.

94
95 A few such datasets also exist in the field of genomics, notably targeted to the tasks of
96 orthology, variant and function prediction. For the former case, OrthoBench^{27,28} has
97 emerged as the standard benchmarking dataset against which orthogroup inference
98 algorithms have been tested for over a decade. The major benchmark dataset for variant
99 prediction is VariBench¹⁹, which supports the development and evaluation of
100 computational methods for interpreting genetic variants, crucial for improving disease
101 diagnosis and understanding genetic differences across various applications. Finally,
102 there is a newly curated collection of benchmark datasets for genomic functional
103 sequence classification in humans, mice, and roundworms²⁰, facilitating the development
104 and evaluation of machine learning models predicting function from DNA sequence data.
105 These models play a crucial role in interpreting vast amounts of genomic data,
106 particularly in human genome investigations, and facilitate discoveries in genetics that
107 have significant implications for medicine and other biological fields.

108
109 Another critical challenge in biodiversity and genomic science is the development of
110 DNA-based taxonomic identification methods. In this case, however, we lack a publicly
111 available benchmark dataset similar to those described above. As part of developing
112 **varKoder**, a new method of DNA-based taxonomic identification based on low-coverage
113 genomic reads¹ (i.e., genome skimming), we have created a number of curated datasets
114 for organisms spanning different taxonomic ranks and phylogenetic depths, from closely
115 related populations, species, to all taxa represented on the NCBI Sequence Read Archive
116 (SRA, <https://www.ncbi.nlm.nih.gov/sra/>).

117
118 To facilitate future comparisons of emerging DNA barcoding methods, here we provide
119 these datasets with metadata and instructions for data access. These datasets are useful
120 for both conventional DNA barcodes^{29,30,31,32,33} and alternative methods that rely on low-
121 coverage genomic sequencing (i.e., DNA signatures^{1,34}). These data will enable future
122 comparisons to our newly developed approach using the same data that we applied for
123 testing. The datasets made available in this data descriptor include the following: (1)
124 newly sequenced and expert-curated low-coverage whole genome sequencing for
125 species in the flowering plant clade Malpighiales, spanning divergences from closely
126 related species to families, and with samples labeled at species, genus and family levels
127 (2) species-level datasets for plants, animals, fungi and bacteria obtained from the
128 literature, and samples labeled at the species level or below (3) a dataset including all
129 eukaryotic families from the NCBI SRA, labeled at the family level and (4) a dataset with
130 all taxa available from the NCBI SRA, labeled with their complete taxonomic
131 classification.

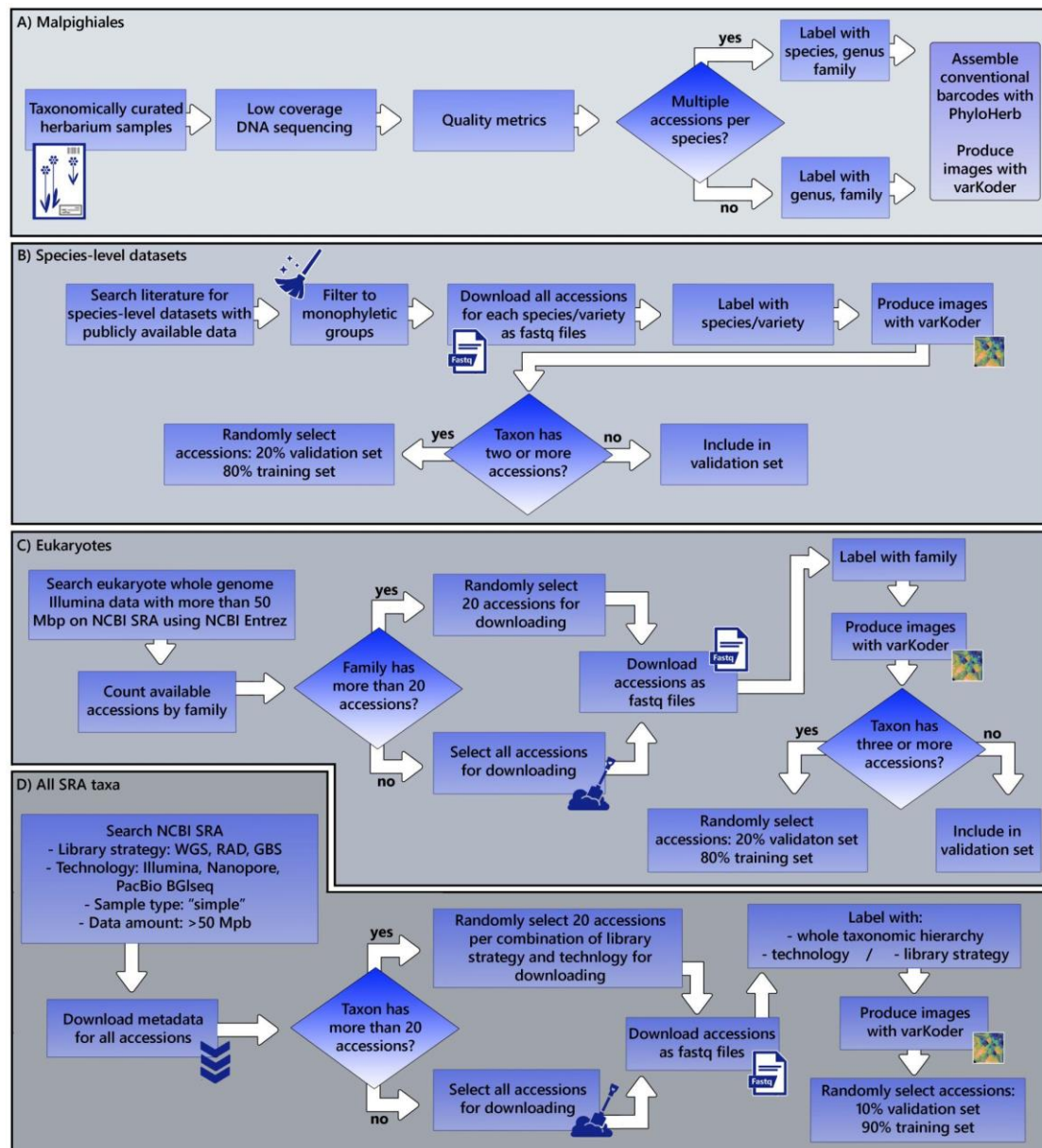
132

133 **Methods**

134

135 Each of the four datasets includes sequencing data and image representations derived
136 from them (i.e., varKodes and ranked frequency chaos game representations¹). **Figure 1**
137 provides an overview of the sampling strategy for each dataset and the workflow used to
138 assemble them.

139
140
141
142



143
144
145
146
147

Figure 1. An overview of data collection and the workflow used to create and curate each dataset. The datasets were compiled from newly generated sequences or from publicly available data, following filtering and processing steps shown here.

148
149

Taxon sampling with varying phylogenetic depths

150
151
152
153
154

Malpighiales dataset. This newly generated dataset tests hierarchical classification from species to family level in plants. Plants exhibit notoriously complex genomic architectures³⁵ that challenge the performance of conventional DNA barcoding³⁶, rendering them a good test case for molecular identification tools. This dataset includes three flowering plant families, all members of the large and morphologically diverse

155 order Malpighiales^{37,38,39}: Malpighiaceae, Elatinaceae, and Chrysobalanaceae. See below
156 for laboratory methods applied for collecting these newly generated sequences.

157

158 The Malpighiaceae data are the most taxonomically sampled and include 287 accessions
159 representing 195 species, which were sampled from 277 herbarium specimens and ten
160 silica-dried field collections. Among these data, the genus *Stigmaphyllon* were
161 comprehensively sampled to build, validate, and test identification methods at shallower
162 phylogenetic depths. A total of 100 *Stigmaphyllon* samples were collected, including 10
163 accessions per species across 10 species. One main advantage of sampling *Stigmaphyllon*
164 is that its taxonomy has been extensively revised, resulting in a diverse and clearly
165 classified set of samples^{40,41}. Moreover, the *Stigmaphyllon* clade represents a wide array
166 of divergence times that span distantly- (34.1 Myr) to very closely-related (0.6 Myr)
167 species^{1,42}.

168

169 The focus for the remainder of the sampling in Malpighiales (Malpighiaceae,
170 Chrysobalanaceae, and Elatinaceae) is to identify a given sample to genus or family. In
171 this case, among the non-*Stigmaphyllon* samples we included 3–9 species per genus
172 representing 30 genera of Malpighiaceae, eight of Chrysobalanaceae, and one of
173 Elatinaceae. Each sample representative was labeled with its corresponding genus and
174 family identification.

175

176 *Species- and subspecies-level datasets.* To test shallow-level classification at species or
177 lower taxonomic ranks, we compiled four datasets from publicly available genome
178 skimming data from the NCBI SRA using NCBI Entrez. These datasets include one
179 bacterial species and one genus each from plants, animals, and fungi.

180

181 First, we included a dataset from *Mycobacterium tuberculosis*, the species of pathogenic
182 bacteria that causes tuberculosis. The bacterial set consisted of clinical isolates from five
183 distinct, monophyletic lineages of *M. tuberculosis* (1.2.2.1, 2.2.1.1.1, 3.1.2, L4.1.i1.2.1, and
184 L4.3.i2) with seven clinical isolates per lineage, totaling 35 samples. This dataset enables
185 testing identification tools on an extremely recently diverged, clinically relevant
186 bacterial lineage⁴³. This dataset of clinical isolates from human-adapted lineages
187 exhibited 99.9% sequence similarity despite key differences in phenotypes, including
188 drug resistance, virulence, and transmissibility⁴³. *Mycobacterium tuberculosis* has
189 diversified quite rapidly in humans, with nine monophyletic lineages. Divergence time
190 estimates for the most recent common ancestor of *M. tuberculosis* are <6,000 years ago⁴⁴.
191 The validation set included 3–6 different samples from the five training lineages as well
192 as 1–4 samples from lineages not included in the training set (2.1, 4.10.i1, and
193 4.6.2.1.1.1.1), totaling 25 validation samples.

194

195 For plants, we included a dataset from a well-delineated clade of mycoheterotrophic
196 orchids⁴⁵ (genus *Corallorhiza*), that allows for assessing the infraspecific taxa variation.
197 *Corallorhiza striata* includes several well-known and easily identifiable varieties. For the
198 *Corallorhiza* training set, we included five species (or varieties) with at least five samples
199 per species (for *C. bentleyi*, *C. striata* var. *involuta*, *C. striata*), except for *C. striata* var.
200 *vreelandii* and *C. striata* var. *striata*, for which we included six and seven samples each,
201 respectively, totaling 28 samples. The validation set included 2–11 different samples

202 from three of the five training species/varieties (*C. striata*, *C. striata* var. *striata*, and *C.*
203 *striata* var. *vreelandii*) as well as one sample from *C. trifida* which was not included in the
204 training set, totaling 18 validation samples.

205

206 For animals, we assembled a *Bembidion* beetle dataset, which includes well-known
207 closely-related cryptic species that were the target of extensive low-coverage whole-
208 genome sequencing^{46,47}. The training set included five samples for each of five species
209 including *B. breve*, *B. ampliatum*, *B. lividulum*, *B. saturatum*, and *B. testatum*, totaling 25
210 samples. The validation set included 1–4 different samples from the five training species
211 as well as from species not included in the training set including *B. aeruginosum*, *B.*
212 *curtulum*, *B. geoparlis*, *B. neocoerulescens*, and *B. oromaia*, totaling 18 samples.

213

214 For fungi, we used *Xanthoparmelia*, a lichen-forming fungal genus whose species are
215 poorly understood and which often form paraphyletic species groupings⁴⁸. Samples for
216 *Bembidion*, *Corallorhiza*, and *Mycobacterium tuberculosis* isolates all formed
217 monophyletic groups, whereas *Xanthoparmelia* species did not. Since the
218 *Xanthoparmelia* species were paraphyletic, we subsampled only monophyletic groups
219 for model training. In this case, four species included three samples per species (*X.*
220 *camtschadalis*, *X. mexicana*, *X. neocumberlandia*, and *X. coloradoensis*) and one species
221 included five samples (*X. chlorochroa*) for the training set, totaling 17 samples. One
222 potential confounding factor is that *Xanthoparmelia* is a lichen-forming fungus and thus
223 genome-skim data represents a chimera of fungal and algal genomes representing both
224 partners in this unique symbiosis. Species of the algal symbiont *Trebouxia* are flexible
225 generalists across fungal *Xanthoparmelia* species. Since these genome skims are a mix of
226 both algal photobiont and fungus, we expect this to be a challenging identification
227 problem because of the more generalist nature of *Trebouxia*⁴⁹. The validation set
228 included 1–3 different samples from the five training species as well as one sample from
229 species not included in the training set including *X. maricopensis*, *X. plittii*, *X. psoromifera*,
230 *X. stenophylla*, *X. sublaevis*, totaling 15 validation samples.

231

232 *Eukaryote family-level dataset.* We retrieved DNA sequencing data from the NCBI SRA on
233 March 7, 2023 using NCBI Entrez, filtering for whole genome sequencing data with
234 random library selection from Eukaryotes (taxid:2759), requiring fastq file availability
235 and DNA as biomolecular type. For each record, we collected taxonomic information
236 using NCBI's Taxonomy database to retrieve family and kingdom classification. Records
237 were filtered to include only those sequenced on the Illumina platform with more than
238 50 million sequenced bases. To ensure balanced representation across taxa, we
239 randomly selected one sequencing run per taxon, and then randomly selected up to 20
240 taxa per family. For each sample, we used fastq-dump
241 (<https://hpc.nih.gov/apps/sratoolkit.html>) to download between 10,000 and 510,000
242 reads per sample. The resulting dataset comprises 8,222 accessions, including families of
243 animals (5,642 accessions, 1,426 families), plants (2,705 accessions, 401 families) and
244 fungi (1,572 accessions, 363 families).

245

246 *All-taxa dataset.* We retrieved DNA sequencing data from the NCBI SRA using NCBI
247 Entrez on January 9, 2024 and the following criteria: (1) fastq file availability, (2) DNA as
248 biomolecular type, (3) library strategies limited to Genotyping by Sequencing (GBS),

249 Restriction site Associated DNA sequencing (RAD-Seq), or Whole Genome Sequencing
250 (WGS), (4) sample type “simple”, (5) sequencing platform including Illumina, Oxford
251 Nanopore, PacBio SMRT, or BGISEQ, (6) more than 50 million sequenced bases. For each
252 record, we collected taxonomic information of the full taxonomic hierarchy using NCBI's
253 Taxonomy database. To ensure balanced representation across taxa and methodologies,
254 we randomly selected up to 20 records for each unique combination of taxonomic ID,
255 library strategy, and sequencing platform to avoid overrepresentation of model species
256 such as humans, mice, and *Escherichia coli*. For each sample, we calculated a target
257 number of reads estimated to yield 60 million bases from the SRA record metadata,
258 approximately three times the amount needed for 20 million bases of quality-filtered
259 sequence. We then used fastq-dump to download that amount of spots per sample (or at
260 least 10,000 spots, if the estimated number was smaller than that). The resulting dataset
261 includes 253,820 accessions including 28,636 taxonomic labels.

262

263 **Laboratory methods for newly generated data**

264

265 For our newly sequenced Malpighiales data we used total genomic DNA extractions. We
266 isolated total genomic DNA from 0.01–0.02 g of silica-dried leaf material or, more
267 commonly, herbarium collections using the Maxwell 16 DNA Purification Kit (Promega
268 Corporation, Inc., WI, USA) and quantified it using the Qubit 4.0 Fluorometer (Invitrogen,
269 CA, USA), with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Inc., MA, USA).
270 Our sampling of herbaria followed the guidelines for effective and ethical sampling of
271 these resources outlined by Davis et al.⁵⁰. Genomic libraries were prepared using ca. 70
272 ng of genomic DNA where possible, using 1/8 reactions of the Kapa HyperPlus Library
273 Preparation Kit (Roche, Basel, Switzerland). Libraries were indexed by using the IDT for
274 Illumina TruSeq DNA unique dual 8 bp barcodes (Illumina Inc., San Diego, CA, USA) or
275 the Nextflex-Ht barcodes (Bioo Scientific Corporation, TX, USA) for multiplexing up to
276 384 samples per sequencing lane. For library preparation, the genomic DNA was
277 sheared by enzymatic fragmentation to 350–400 base pairs (bp), depending on the
278 quality of the input DNA. Libraries' concentrations were verified with the Qubit 4.0
279 Fluorometer, using the Qubit dsDNA HS Assay Kit (Invitrogen, CA, USA), and average
280 sizes of DNA fragments were verified with the High Sensitivity HSD1000 ScreenTape
281 Assay in the 2200 TapeStation (Agilent Technologies, Waldbronn, Germany). Libraries
282 were diluted into 0.7 nM or 1.0 nM and pooled together. We used Real-Time PCR
283 (BioRad CFX96 Touch, BioRad Laboratories, Hercules, USA) with the NEBNext Library
284 Quant Kit (New England Biolabs, Ipswich, USA) for verifying the final concentration of
285 the libraries' pools. Sequencing of libraries was conducted using the Illumina Hi-Seq
286 2500 or the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA) for 125 bp or 150
287 bp pair-ended reads, at The Bauer Core Facility at Harvard University, MA, USA.

288

289 **Extracting conventional barcodes from genome skimming data**

290

291 For the Malpighiales dataset, we assembled conventional barcodes. To recover the
292 traditional plant barcodes *rbcl*, *matK*, *trnL-F*, *ndhF*, and ITS from our Malpighiales
293 genome skim data, we applied GetOrganelle v1.7.7.0⁵¹ and PhyloHerb v1.1.1⁵² to
294 automatically assemble and extract these DNA markers, respectively. Briefly, the
295 complete or subsampled genome skim data were first assembled into plastid genomes or

296 nuclear ribosomal regions using GetOrganelle⁵¹ with its default settings. Next, PhyloHerb
 297 was applied to extract the relevant barcode genes using its built-in BLAST database.

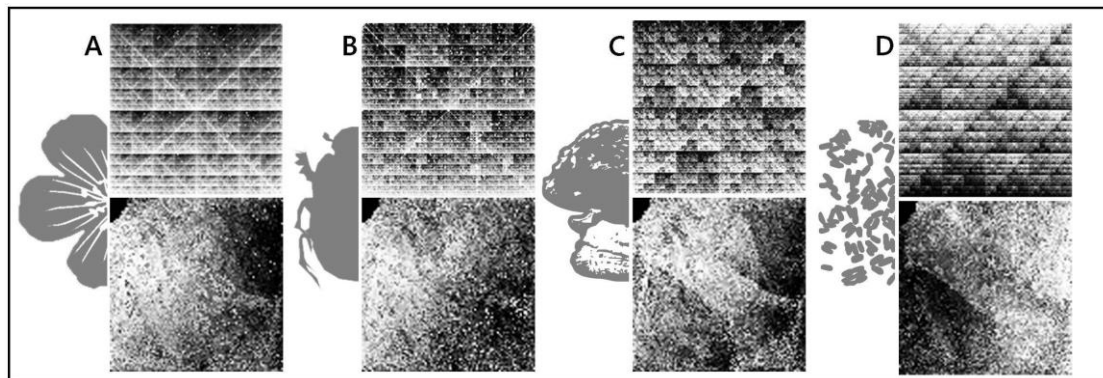
298

299 **Creation of varKode and CGR images from genome skimming data**

300

301 In addition to raw sequence data, we provide image representations of the genome
 302 signature (**Figure 2**) implied by these data for all samples included here. See our
 303 companion paper¹ for details on how these images are generated. In all cases, pixels in
 304 these images represent individual k-mer sequences. Brightness represents the frequency
 305 of a k-mer, transformed to ranks and digitized to 8 bits. The two kinds of representation
 306 provided differ in how k-mers are mapped to pixels. VarKodes are a compact
 307 representation in which kmer counts and their reverse complements are combined. The
 308 mapping of kmers to pixels in an image attempts to place more similar kmers closer
 309 together in the image space. Ranked frequency chaos game representation (rfCGR)
 310 images are similarly produced, but the mapping of k-mers to pixels follows the chaos
 311 game representation⁵³. rfCGRs present a fractal pattern, while varKodes generally
 312 present gradients spanning the whole image. In both cases, we used the “varKoder
 313 image” command to generate varKodes, and then used “varKoder convert” to generate
 314 rfCGRs from these varKodes. In all cases, we used k-mers of size seven, which were
 315 determined to yield optimal balance between classification accuracy and computing
 316 effort¹. These k-mer counts were used to generate images and we normalized counts by
 317 ranking and then rescaling and quantizing ranks to integer numbers ranging from 0 to
 318 255, which are the brightness levels supported by a png image. All images are saved in
 319 png format, including built-in exif metadata with the labels assigned to each sample.
 320 After producing images, we split datasets into training and validation sets. The following
 321 specific settings have been used for each dataset described below.

322



323

324 **Figure 2.** Demonstration of the two types of image representations of the genome
 325 signature included in our datasets. Examples of rfCGRs (top) and varKodes (bottom) are
 326 shown for four different clades: plants (a), animals (b), fungi (c), and bacteria (d). rfCGRs
 327 are larger images, and their relative sizes are shown to scale. In each case, both images
 328 were produced from the same sequence data. a) Local ID 1089 (plant, *Triaspis*
 329 *hypericoides*) b) SRA Accession SRR15249224 (beetle, *Mesosa* sp.). c) SRA Accession
 330 SRR15292413 (fungus, *Amanita* sp.). d) SRA Accession SRR2101396 (Bacteria,
 331 *Mycobacterium tuberculosis*).

332

333 *Malpighiales*. varKodes have been produced from data amounts varying from 500Kbp to
334 200 Mbp and k-mer size of 7. We applied leave-one-out cross-validation in all tests
335 following de Medeiros et al.¹, so the dataset has not been split into training and
336 validation sets. All accessions have been labelled with their genus and family
337 identification. For species in the genus *Stigmaphyllon*, we additionally labeled accessions
338 with their species identity.

339

340 *Species- and subspecies-level datasets*. varKodes have been produced from data amounts
341 varying from 500 Kbp to the maximum amount of data available for each accession and
342 k-mer size of 7. All accessions have received a single label: their species or variety name.
343 For species or varieties represented by at least four accessions, we randomly chose 20%
344 of the accessions for the validation set (with a minimum of 1) and 80% for the training
345 set. For species or varieties with three or less accessions, they were only included in the
346 validation set, to test whether a multi-label model correctly predicted no labels for that
347 accession.

348

349 *NCBI SRA Eukaryotes*. varKodes have been produced from data amounts varying from
350 500Kbp to 10Mbp and k-mer size of 7. All accessions have received a single label: their
351 family name. For families represented by at least three accessions, we randomly chose
352 20% of the accessions for the validation set (with a minimum of 1) and 80% for the
353 training set. Families with less than two accessions were only included in the validation
354 set, to test whether a multi-label model correctly predicted no labels for that accession.

355

356 *NCBI SRA all-taxa*. varKodes have been produced from data amounts varying from
357 500Kbp to 20Mbp and k-mer size of 7. All accessions received multiple labels, including:
358 (1) all NCBI taxonomy IDs related to that accession (i.e., the full taxonomic hierarchy, as
359 separate labels), (2) the library strategy, and (3) the sequencing platform. We randomly
360 selected 10% of the accessions for the validation set, regardless of their labels. Next, we
361 removed from the validation set any labels not present in at least one accession in the
362 training set.

363

364 **Metadata organization**

365

366 To maximize the utility of our datasets for benchmarking molecular identification tools,
367 we provide comprehensive metadata for each sample. The metadata is organized in a
368 consistent format across all datasets to enable easy comparison and reuse in future
369 investigations. Each dataset—*Malpighiales*, *Species and subspecies-level (Bembidion*
370 *beetles, Corallorhiza orchids, Xanthoparmelia fungi, Mycobacterium tuberculosis)*,
371 *Eukaryote families and All SRA taxa*—includes a metadata table detailing the raw
372 sequencing data for each sample, with taxonomic-, sequencing-, and sample-related
373 information. All datasets share 17 common metadata fields (**Table 1**). The *Malpighiales*
374 dataset, the only one containing new sequence data, includes five additional fields that
375 provide more specific details on voucher information (**Table 2**). The metadata is
376 provided in the same Harvard Dataverse repository as the data
377 (<https://dataverse.harvard.edu/privateurl.xhtml?token=e82d1443-ffe9-493c-8f81-9fd7d5ec9c8f>).
378

379

380 **Table 1.** Description of common metadata fields for all datasets.

FIELD	DESCRIPTION
SRA_Run_ID	The unique identifier for the run in the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra).
Local_ID	A unique identifier assigned to each sample as used in Medeiros et al. ¹ . This identifier serves as a local reference for linking metadata, sequence data and images.
Tax_ID	The taxonomic identifier associated with the organism, as per the NCBI taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy).
Taxon	The scientific name of the organism from which the sample was derived.
Taxonomy_Superkingdom	Broader taxonomic classification at the Superkingdom level (i.e., Eukaryota, Bacteria, Viruses or Archaea).
Taxonomy_Kingdom	Taxonomic classification at the Kingdom level. Helps contextualize the sample.
Taxonomy_Family	Family to which the sample belongs. Provides additional context for understanding the evolutionary relationships between samples.
BioSample_ID	The unique identifier for the sample in NCBI's BioSample database (https://www.ncbi.nlm.nih.gov/biosample), linking to additional metadata.
Download_Path	URL from which the sequence data in Lite Format (with simplified quality scores) can be downloaded from the NCBI SRA.
Library_Strategy	Describes the sequencing strategy (e.g., WGS, RAD-Seq), indicating how the data was generated.
Library_Source	Indicates the source from which the DNA was extracted (i.e., genomic DNA or metagenomic).
Library_Layout	Specifies the configuration of sequencing reads: either SINGLE (single-end) or PAIRED (paired-end).
Seq_Platform	The sequencing platform used, such as Illumina, PacBio, Oxford Nanopore, etc.

Seq_Model	The specific sequencing instrument model (e.g., Illumina NovaSeq 6000), for reproducibility.
Size_MB	The file size in megabytes (MB) of the sequence data from the NCBI SRA in Normalized Format (with full per-base quality scores).
Labels	All the labels assigned to a given accession. All labels were combined as a string separated by semicolon, allowing for more compact storage of information.
Set	Set in de Medeiros et al. ¹ . For the Malpighiales dataset, this column has empty values since samples were evaluated with cross-validation. For other datasets, there are three possible values: “train” for training set, “valid” for validation set and “valid_notrain” for accessions used in validation but with taxonomic labels not included in the training set, to test for false positives.

381

382 **Table 2.** Description of additional metadata fields exclusive in the Malpighiales dataset.

FIELD	DESCRIPTION
Taxonomy_Genus	Labels the genus to which the sample belongs, to support identification to genus level.
Voucher	Information on the collector and the collection number, which links the sample to its voucher specimen.
Collector	The name of the individual(s) responsible for collecting the specimen.
CollectorID	The specific number associated with the collector’s collection for this sample.
Collection	The acronym of the collection where the herbarium voucher of the sample is deposited.

383

384 **Data Records**

385

386 The dataset is available at Harvard Dataverse and the NCBI Sequence Read Archive. The
387 Harvard Dataverse repository
388 (<https://dataverse.harvard.edu/privateurl.xhtml?token=e82d1443-ffe9-493c-8f81-9fd7d5ec9c8f>) includes metadata tables, processed conventional DNA barcodes, and
389 DNA signature images (varKodes and rfCGRs). New sequences (i.e., Malpighiales) have
390 been uploaded to NCBI SRA under PRJNA1052627. All remaining sequence data were
391 already publicly available on NCBI SRA and can be retrieved from the accession numbers
392 in the metadata tables. The complete dataset comprises four major components,
393 summarized below.

394

395

396 **Malpighiales**

397 This dataset contains 287 newly sequenced accessions from three families in the order
398 Malpighiales. This includes families Malpighiaceae (251 accessions representing 31
399 genera), Elatinaceae (6 accessions for 1 genus), and Chrysobalanaceae (30 accessions for
400 8 genera). Malpighiaceae includes *Stigmaphyllon* with the most comprehensive species
401 sampling: 10 species and 10 accessions sampled per species. *Stigmaphyllon* accessions
402 are labeled with species, genus and family. All other accessions are labeled with genus
403 and family. This dataset is used for benchmarking molecular identification tools from
404 species to family levels under a realistic scenario of uneven diversity and sequencing
405 effort. The data provided includes raw sequencing data, processed conventional
406 barcodes (*rbcL*, *matK*, *trnL-F*, *ndhF*, and ITS), and image representations (varKodes and
407 rfCGRs).

408

409 **Species- and subspecies-level datasets**

410 This is composed of four datasets from published data of four clades – *Bembidion* beetles
411 (43 accessions from 10 species), *Corallorhiza* orchids (46 accessions from 6
412 species/varieties), *Xanthoparmelia* fungi (32 accessions from 10 species), and
413 *Mycobacterium* bacteria (60 accessions from 8 lineages). In each case, we include raw
414 sequencing data and image representations. These datasets are suitable for
415 benchmarking species-level identification, as well as variety, strain, or subspecies.

416

417 **Eukaryote families**

418 We compiled a dataset for identifying eukaryote families from the NCBI Sequence Read
419 Archive. This includes 9,910 accessions from 2,182 families of animals, plants and fungi.
420 Of these, 861 families (517 Metazoa, 197 plants, 147 fungi), represented by 8,222
421 accessions, had at least three accessions available and were included in the training set.
422 We include sequence data and image representations. This dataset serves to benchmark
423 family-level identification tools at a large scale.

424

425 **All SRA taxa**

426 This is the largest dataset compiled from the NCBI Sequence Read Archive, containing
427 data including all the taxonomic hierarchy and multiple sequencing methods (253,820
428 accessions including 28,636 taxonomic labels, three labels for library strategy, and four
429 labels for sequencing platform). We include sequence data and image representations.
430 This is the largest and most heterogeneous dataset provided here, benchmarking
431 identification at all taxonomic levels across different sequencing methodologies.

432

433 For raw sequence data, we provide accession numbers to NCBI SRA runs. These can be
434 downloaded in conventional formats (such as fastq) using the SRA toolkit
435 (<https://github.com/ncbi/sra-tools>).

436

437 Processed conventional barcodes are provided as fasta files. Each fasta file is named
438 after the gene region represented and includes individual sequences named after the
439 SRA accession number.

440

441 Image representations are provided as png images. These images follow a file name
442 convention that is interpreted by **varKoder** and include information about accession

443 number, k-mer size, type of representation and amount of DNA sequence data used to
 444 produce the image: “[local_ID]@[sequence base pairs]+[representation]+k[k-mer
 445 size].png”. For example, the file “SRR9036258@00010000K+varKode+k7.png”
 446 represents accession with local ID SRR9036258, 10 Mbp (i.e., 10,000 Kbp) of sequence
 447 data, varKode representation and k-mer size of 7. Labels associated with accession can
 448 be found in the metadata tables and also as image metadata contained in the png file.
 449 **varKoder** is able to read this image metadata, and it is also visible through general
 450 purpose programs that handle image metadata, such as exiftool (<https://exiftool.org>).
 451

452 **Technical Validation**

453

454 Quality metrics for new sequence data: We measured sequencing success using various
 455 quality metrics, including total input DNA for library preparation, sequencing yield (in
 456 megabases), percentage of bases with a QScore ≥ 30 , mean quality score, average GC
 457 content, and sequencing depth. These metrics were calculated for the newly sequenced
 458 data of Malpighiales’ representatives to ensure robustness and reliability of the
 459 sequencing results. A summary of these metrics are provided in Supplementary Table 1.
 460

461 Metrics from GetOrganelle: We used GetOrganelle to assess the quality of the assembled
 462 Malpighiales’ plastid genomes, examining factors like assembly success and
 463 completeness. These metrics are also provided in Supplementary Table 1.
 464

465 We have not further validated sequences that were already publicly available. In that
 466 case, we used data as downloaded from NCBI following the filters specified in materials
 467 and methods.
 468

469 **Usage Notes**

470

471 See de Medeiros et al¹ for a complete account of how these datasets have been used to
 472 develop and test varKoder. NCBI accession numbers can be used to download associated
 473 sequence data with the SRA toolkit (<https://github.com/ncbi/sra-tools>). Conventional
 474 barcode sequences in the fasta format can be used for sequence alignment and search.
 475 varKode and rfCGR images can be used as input to varKoder or other programs
 476 processing images in the PNG format. Conventional barcode sequences and PNG images
 477 can be found in the Harvard Dataverse repository accompanying this article
 478 (<https://dataverse.harvard.edu/privateurl.xhtml?token=e82d1443-ffe9-493c-8f81-9fd7d5ec9c8f>)
 479
 480

481 **Code Availability**

482

483 The code used to retrieve and process sequence data used here is available in a github
 484 repository (https://github.com/brunoasm/varKoder_development), archived in
 485 FigShare (<https://figshare.com/s/693d5194204f73fee6e2>). The source code for
 486 varKoder, which can process sequence data into varKodes and rfGRS, as well as train and
 487 use neural networks, is available at <https://github.com/brunoasm/varKoder>.
 488

489 Acknowledgments

490

491 BdM received postdoctoral fellowships from the Harvard University Museum of
492 Comparative Zoology and the Smithsonian Tropical Research Institute during the early
493 stages of this study. RCA and LCM were supported by the Coordenação de
494 Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. LC
495 was supported by Harvard University and by a Stengl Wyer scholarship from the
496 University of Texas at Austin. PF was supported by LVMH Research, Dior Science and
497 NSF PRFB. YY was supported by a postdoctoral fellowship from Harvard University
498 Herbaria. CCD was supported by Harvard University, LVMH Research, Dior Science, and
499 National Science Foundation grants DEB-1355064 and DEB-0544039. Computations
500 were performed at the Harvard Cannon Cluster and the Field Museum Grainger
501 Bioinformatics Center. We thank the Bauer Core Facility, and especially Claire Reardon,
502 at Harvard University for providing technical support during the laboratory process. We
503 thank Kylee Peterson for assistance in obtaining the newly sequenced data. Goia Lyra
504 helped train RCA on molecular lab techniques. Newly generated sequence data were
505 collected under Harvard University's binding Participation Agreement.

506

507 Author contributions

508

509 Renata C. Asprino compiled the herbarium samples, collected and curated the new DNA
510 sequence data, prepared the data repositories and wrote the manuscript.

511 Liming Cai curated the new DNA sequence data, processed conventional barcodes and
512 wrote the manuscript.

513 Yujing Yan collected and curated the new DNA sequence data and wrote the manuscript.

514 Peter J. Flynn collected, curated and processed the species-level datasets and wrote the
515 manuscript.

516 Lucas C. Marinho collected and curated the new DNA sequence data, and prepared
517 figures.

518 Xiaoshan Duan contributed to conceive the workflow, collected and curated the new
519 DNA sequence data.

520 Christiane Anderson helped to conceive the sampling and compiled the herbarium
521 samples.

522 Charles C. Davis designed the research, funded new DNA sequencing, compiled the
523 herbarium samples, collected and curated the new DNA sequence data, and wrote the
524 manuscript.

525 Bruno A. S. de Medeiros designed the research, designed varKodes, wrote the program
526 *varKoder*, curated the large SRA datasets, prepared the data repositories and wrote the
527 manuscript.

528 All authors revised and approved the manuscript.

529

530 **Competing interests**

531

532 CCD declares that he is supported by LVMH Research and Dior Science, a company
533 involved in the research and development of cosmetic products based on floral extracts.
534 He also serves as a member of Dior's Age Reverse Board.

535

536 **References**

537

- 538 1. de Medeiros, B. *et al.* A universal DNA barcode for the Tree of Life. Preprint
539 at <https://doi.org/10.32942/X24891> (2024).
- 540 2. Dodsworth, S. Genome skimming for next-generation biodiversity analysis.
541 *Trends Plant Sci.* **20**, 525–527 (2015).
- 542 3. Coissac, E., Hollingsworth, P. M., Laverigne, S. & Taberlet, P. From barcodes to
543 genomes: extending the concept of DNA barcoding. *Mol. Ecol.* **25**, 1423–1428
544 (2016).
- 545 4. Zeng, C.-X. *et al.* Genome skimming herbarium specimens for DNA barcoding
546 and phylogenomics. *Plant Methods* **14**, 43 (2018).
- 547 5. Quattrini, A. M. *et al.* Skimming genomes for systematics and DNA barcodes
548 of corals. *Ecol. Evol.* **14**, e11254 (2024).
- 549 6. Liu, S. *et al.* SOAPBarcode: revealing arthropod biodiversity through
550 assembly of Illumina shotgun sequences of PCR amplicons. *Methods Ecol.*
551 *Evol.* **4**, 1142–1150 (2013).
- 552 7. Gillett, C. P. D. T., Crampton-Platt, A., Timmermans, M. J. T. N., Jordal, B. H.,
553 Emerson, B. C., & Vogler, A. P. Bulk de novo mitogenome assembly from
554 pooled total DNA elucidates the phylogeny of weevils (Coleoptera:
555 Curculionoidea). *Mol. Biol. Evol.* **31**, 2223–2237 (2014).
- 556 8. Bakker, F. T. *et al.* Herbarium genomics: plastome sequence assembly from a
557 range of herbarium specimens using an Iterative Organelle Genome
558 Assembly pipeline. *Biol. J. Linn. Soc.* **117**, 33–43 (2016).
- 559 9. Bohmann, K., Mirarab, S., Bafna, V. & Gilbert, M. T. P. Beyond DNA barcoding:
560 The unrealized potential of genome skim data in sample identification. *Mol.*
561 *Ecol.* **29**, 2521–2534 (2020).
- 562 10. Sarmashghi, S., Bohmann, K., P. Gilbert, M. T., Bafna, V. & Mirarab, S. Skmer:
563 assembly-free and alignment-free sample identification using genome skims.
564 *Genome Biol.* **20**, 34 (2019).
- 565 11. Fiannaca, A. *et al.* Deep learning models for bacteria taxonomic classification
566 of metagenomic data. *BMC Bioinform.* **19**, 198 (2018).
- 567 12. Linard, B., Swenson, K. & Pardi, F. Rapid alignment-free phylogenetic
568 identification of metagenomic sequences. *Bioinform.* **35**, 3303–3312 (2019).
- 569 13. Desai, H. P., Parameshwaran, A. P., Sunderraman, R. & Weeks, M.
570 Comparative study using neural networks for 16S ribosomal gene
571 classification. *J. Comput. Biol.* **27**, 248–258 (2020).
- 572 14. Shang, J. & Sun, Y. CHEER: HierarCHical taxonomic classification for viral
573 mEtagEnomic data via deep leaRning. *Methods* **189**, 95–103 (2021).

- 574 15. Millán Arias, P., Alipour, F., Hill, K. A. & Kari, L. DeLUCS: Deep learning for
575 unsupervised clustering of DNA sequences. *PLoS ONE* **17**, e0261531 (2022).
- 576 16. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible
577 microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857
578 (2019).
- 579 17. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2:
580 memory friendly classification with the genome taxonomy database.
581 *Bioinform.* **38**, 5315–5316 (2022).
- 582 18. Weitschek, E., Fison, G. & Felici, G. Supervised DNA barcodes species
583 classification: analysis, comparisons and results. *BioData Mining* **7**, 4 (2014).
- 584 19. Shirvanizadeh, N. & Vihinen, M. VariBench, new variation benchmark
585 categories and data sets. *Front. Bioinform.* **3**, 1248732 (2023).
- 586 20. Grešová, K., Martinek, V., Čechák, D., Šimeček, P. & Alexiou, P. Genomic
587 benchmarks: a collection of datasets for genomic sequence classification.
588 *BMC Genom. Data.* **24**, 25 (2023).
- 589 21. Joshi, C., Sorenson, L., Wolfert, A., Clement, M., Price, J. & Buckles, K. CENSUS-
590 HWR: a large training dataset for offline handwriting recognition. Preprint at
591 <https://doi.org/10.48550/arXiv.2305.16275> (2023).
- 592 22. Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M. & Vidal, E. A set of
593 benchmarks for handwritten text recognition on historical documents.
594 *Pattern Recogn.* **94**, 122–134 (2019).
- 595 23. Kulyabin, M. *et al.* OCTDL: Optical coherence tomography dataset for image-
596 based deep learning methods. *Sci. Data* **11**, 365 (2024).
- 597 24. Pawłowska, A. *et al.* Curated benchmark dataset for ultrasound based breast
598 lesion analysis. *Sci. Data* **11**, 148 (2024).
- 599 25. Beery, S. *et al.* The Auto Arborist Dataset: a large-scale benchmark for
600 multiview urban forest monitoring under domain shift. Presented at the
601 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,
602 New Orleans, LA, USA. Available at:
603 <https://doi.org/10.1109/CVPR52688.2022.02061>. (2022)
- 604 26. Cañas, J. S. *et al.* A dataset for benchmarking Neotropical anuran calls
605 identification in passive acoustic monitoring. *Sci Data* **10**, 771 (2023).
- 606 27. Trachana, K. *et al.* Orthology prediction methods: A quality assessment using
607 curated protein families. *BioEssays* **33**, 769–780 (2011).
- 608 28. Emms, D. M. & Kelly, S. Benchmarking Orthogroup Inference Accuracy:
609 Revisiting Orthobench. *Genome Biol. Evol.* **12**, 2258–2266 (2020).
- 610 29. Hebert, P. D. N., Ratnasingham, S. & de Waard, J. R. Barcoding animal life:
611 cytochrome c oxidase subunit 1 divergences among closely related species.
612 *Proc. R. Soc. Lond. B* **270**, S96–S99 (2003).
- 613 30. Kress, W. J. Plant DNA barcodes: Applications today and in the future. *J. Syst.*
614 *Evol.* **55**, 291–307 (2017).
- 615 31. Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System.
616 *Mol. Ecol. Notes* **7**, 355–364 (2007).
- 617 32. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E.
618 Towards next-generation biodiversity assessment using DNA
619 metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).

- 620 33. Seifert, K. A. Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour.* **9**,
621 83–89 (2009).
- 622 34. de la Fuente, R., Díaz-Villanueva, W., Arnau, V. & Moya, A. Genomic signature
623 in evolutionary biology: A review. *Biology* **12**, 322 (2023).
- 624 35. Lynch, M. *The Origins of Genome Architecture*. (Sinauer Associates, 2007).
- 625 36. Gonzalez, M. A. *et al.* Identification of amazonian trees with DNA barcodes.
626 *PLoS ONE* **4**, e7483 (2009).
- 627 37. Cai, L. *et al.* The perfect storm: gene tree estimation error, incomplete
628 lineage sorting, and ancient gene flow explain the most recalcitrant ancient
629 angiosperm clade, Malpighiales. *Syst. Biol.* **70**, 491–507 (2021).
- 630 38. Xi, Z. *et al.* Phylogenomics and a posteriori data partitioning resolve the
631 Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U.S.A.*
632 **109**, 17519–17524 (2012).
- 633 39. Wurdack, K. J. & Davis, C. C. Malpighiales phylogenetics: gaining ground on
634 one of the most recalcitrant clades in the angiosperm tree of life. *Amer. J. Bot.*
635 **96**, 1551–1570 (2009).
- 636 40. Anderson, C. Monograph of *Stigmaphyllon* (Malpighiaceae). *Syst. Bot.*
637 *Monogr.* **51**, 1–313 (1997).
- 638 41. Anderson, C. Revision of *Rysopterys* and transfer to *Stigmaphyllon*
639 (Malpighiaceae). *Blumea* **56**, 73–104 (2011).
- 640 42. Cai, L. *et al.* Phylogeny of Elatinaceae and the tropical Gondwanan origin of
641 the Centroplacaceae (Malpighiaceae, Elatinaceae) clade. *PLoS ONE* **11**,
642 e0161881 (2016).
- 643 43. Freschi, L. *et al.* Population structure, biogeography and transmissibility of
644 *Mycobacterium tuberculosis*. *Nat. Commun.* **12**, 6099 (2021).
- 645 44. Sabin, S. *et al.* A seventeenth-century *Mycobacterium tuberculosis* genome
646 supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex.
647 *Genome Biol.* **21**, 201 (2020).
- 648 45. Barrett, C. F., Wicke, S. & Sass, C. Dense infraspecific sampling reveals rapid
649 and independent trajectories of plastome degradation in a heterotrophic
650 orchid complex. *New. Phytol.* **218**, 1192–1204 (2018).
- 651 46. Sproul, J. S., Barton, L. M. & Maddison, D. R. Repetitive DNA profiles reveal
652 evidence of rapid genome evolution and reflect species boundaries in
653 ground beetles. *Syst. Biol.* **69**, 1137–1148 (2020).
- 654 47. Sproul, J. S. & Maddison, D. R. Cryptic species in the mountaintops: species
655 delimitation and taxonomy of the *Bembidion breve* species group
656 (Coleoptera: Carabidae) aided by genomic architecture of a century-old type
657 specimen. *Zool. J. Linn. Soc.* **183**, 556–583 (2018).
- 658 48. Keuler, R. *et al.* Interpreting phylogenetic conflict: hybridization in the most
659 speciose genus of lichen-forming fungi. *Mol. Phylog. Evol.* **174**, 107543
660 (2022).
- 661 49. Leavitt, S. D. *et al.* Fungal specificity and selectivity for algae play a major
662 role in determining lichen partnerships across diverse ecogeographic
663 regions in the lichen-forming family Parmeliaceae (Ascomycota). *Mol. Ecol.*
664 **24**, 3779–3797 (2015).
- 665 50. Davis, C. C., Sessa, E., Paton, A., Antonelli, A. & Teisher, J. K. Guidelines for the
666 effective and ethical sampling of herbaria. *Nat. Ecol. Evol.* (2024).

- 667 51. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo
668 assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
- 669 52. Cai, L., Zhang, H. & Davis, C. C. PhyloHerb: A high-throughput phylogenomic
670 pipeline for processing genome skimming data. *Appl. Plant Sci.* **10**, e11475
671 (2022).
- 672 53. Jeffrey, H. J. Chaos game representation of gene structure. *Nucl. Acids Res.* **18**,
673 2163–2170 (1990).