A universal DNA barcode for the Tree of Life

Bruno de Medeiros[1,2,3], Liming Cai[4, 5], Peter J. Flynn[4], Yujing Yan[4], Xiaoshan Duan[4,6], Lucas C. Marinho[4, 7], Christiane Anderson[8], and Charles C. Davis[4]

[1]Field Museum of Natural History, Chicago, Illinois, 60605, USA

[2]Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts, 02138 USA

[3]Smithsonian Tropical Research Institute, Panama City, Panama

[4]Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge, Massachusetts, 02138 USA

[5]Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, 78712 USA

[6]College of Forestry, Northwest Agriculture & Forestry University, Yangling 712100, Shaanxi, China

[7]Departamento de Biologia, Universidade Federal do Maranhão, Av. dos Portugueses 1966, Bacanga 65080-805, São Luís, Maranhão, Brazil

[8]University of Michigan Herbarium, 3600 Varsity Drive, Ann Arbor, Michigan 48108, USA

Corresponding authors:

Bruno de Medeiros, Field Museum of Natural History, Chicago, IL, 60605; E-mail: bdemedeiros@fieldmuseum.org

Charles C. Davis, Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Cambridge, MA 02138, USA; E-mail: cdavis@oeb.harvard.edu

Abstract

Species identification using DNA barcodes has revolutionized biodiversity sciences and society at large. However, conventional barcoding methods do not reflect genomic complexity, may lack sufficient variation, and rely on limited genomic loci that are not universal across the Tree of Life. Here, we develop a novel barcoding method that uses exceptionally low-coverage genome skim data to create a “varKode”, a two-dimensional image representing the genomic landscape of a species. Using these varKodes, we then train neural networks for precise taxonomic identification. Applying an expertly annotated genomic dataset including hundreds of newly sequenced genomic samples from the plant clade Malpighiales, we demonstrate >91% precision when identifying species or genera. Remarkably, high accuracy remains despite minimal data amounts that lead to failure when applying alternative methods. We further illustrate the broad utility of varKodes across several focal clades of eukaryotes and prokaryotes. As a final test, we classify the entire NCBI eukaryote sequence-read archive to identify its 861 constituent families with >95% precision despite utilizing less than 10 Mbp of data per sample. Enhanced computational efficiency and scalability, minimal data inputs robust to degraded DNA, and modularity for further development make varKoding an ideal approach for biodiversity science.

Keywords: biodiversity science, computer vision, DNA barcoding, Malpighiaceae, natural history collections, neural networks, species identification, taxonomy

Introduction

For two decades, conventional DNA barcoding, which relies on standardized short sequences (400–800 bp) for species identification[1, 2, 3, 4, 5], has enabled novel and massively scalable science spanning evolution[4, 6, 7, 8, 9]; ecology[10, 11, 12, 13, 14] and paleontology[15, 16, 17, 18, 19]. Practical applications of barcoding have also made major contributions to environmental health, including the ability to authenticate medicinal plants[20], detect agricultural pests[21], and monitor poaching and the trade of endangered species[22, 23, 24, 25, 26, 27]. Despite these remarkable achievements, however, conventional DNA barcoding suffers from at least four limitations. First, barcodes are customized specifically for particular clades of organisms (e.g., plants, animals, and fungi), and therefore are not universal—in many cases even within focal clades. For example, commonly used plant barcodes from chloroplast genes such as *mat*K and *rbc*L cannot be applied as barcodes for all plants[28, 29], or for animals and fungi. Second, conventional barcode loci may fail to distinguish closely related taxa, a pervasive shortcoming in plants[2, 30]. Third, reliance on a single locus may lead to spurious results in the case of complex evolutionary scenarios such as hybridization in deep and shallow time[31, 32, 33, 34]. And fourth, the necessary comparison of homologous genes may fail when PCR primers are not universal[35], the source DNA is fragmented[27], or paralogy and the presence of pseudogenes confounds accurate orthology assessments[36, 37].

Newer alternatives to conventional barcoding have begun to address these challenges by leveraging two technological advancements: high-throughput sequencing and machine-learning applications powered by neural networks. High-throughput sequencing facilitates more comprehensive assessments of total genomic space[38, 39]. For example, presence/absence patterns among short DNA sequences (k-mers) from low-coverage reads (i.e., genome skims) can estimate overall sequence distances, bypassing genome alignments entirely as implemented in *Skmer*[40]. Machine learning enables more complex sequence comparisons than do more conventional methods that rely on homology and simple metrics[41]. Machine-learning models can cluster DNA sequences correctly without supervision[42, 43] and can classify sequences based on reference datasets[44, 45, 46, 47]. In particular, neural networks are exceptionally powerful for sophisticated computer-vision tasks, such as image classification[48]. Thus, the combination of low-coverage genome skimming data and neural networks holds enormous promise for accurate and scalable DNA barcoding, but its potential has yet to be fully realized.

Genomes differ substantially in many features beyond the simple nucleotide differences commonly used in conventional barcoding (e.g., repeat content), but these differences have been overlooked for species identification[49, 50, 51, 52]. We propose that i.) relevant genomic features can be captured by nucleotide composition with short k-mer counts and very small sequence coverage; and ii.) these counts can be used to distinguish species and higher taxa efficiently and accurately using machine learning. Inspired by prior work[42, 44, 53], we developed a novel barcoding method (**varKoding**) that integrates genome skim data with machine-learning models trained using two-dimensional images representing genome composition (a **varKode**) (**Figure 1A**). To assess the utility of varKoding for accurate species identification, we first generated a *de novo* genome skim dataset including hundreds of samples derived primarily from historical herbarium specimens for the diverse plant genus *Stigmaphyllon* (Malpighiaceae), which has received extensive phylogenetic and taxonomic treatment[54, 55, 56, 57, 58]. Upon establishing the power and robustness of our tool for identifying species of *Stigmaphyllon*, we explored the utility of varKodes at greater phylogenetic depths among flowering plant families and genera of species spanning three diverse clades within the order Malpighiales (Malpighiaceae, Chrysobalanaceae, and Elatinaceae). Finally, we demonstrate the generality and scalability of varKoding across the Tree of Life by testing it on several published species-level datasets from fungi, plants, animals, bacteria, and finally from a massive dataset including all families of eukaryotes from publicly available sequence data.

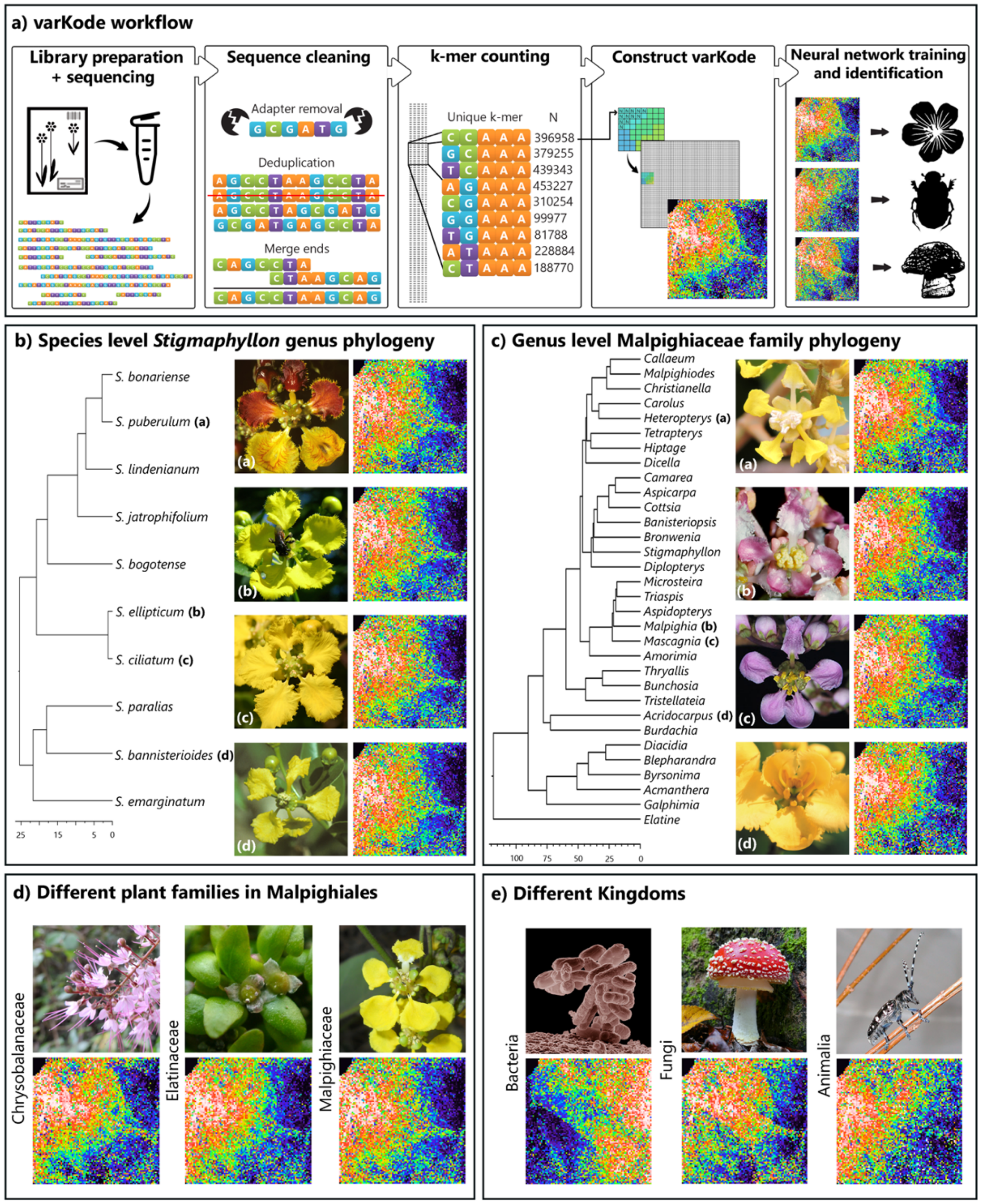

Figure 1. varKoding and training data overview. (**A**) varKode generation workflow. varKode images are natively grayscale, but here they are mapped to a rainbow color scale for increased contrast. (**B**) Phylogeny and example varKodes of *Stigmaphyllon* species. (**C**) Phylogeny and example varKodes of Malpighiaceae genera including their closest outgroup (*Elatine*, Elatinaceae). (**D**) Examples of varKodes from across plant families of Malpighiales, and (**E**) across kingdoms. Chronograms depicted for each representative set with timelines in millions of years (Myr) at the bottom of **B** and **C**.

Results and Discussion

varKodes can be classified with neural networks

An accurate and scalable DNA-barcoding method using neural networks has not previously been developed owing to two widely held misconceptions: i.) accurate barcoding by neural networks requires sufficiently large training data sets that they would be impractical for typical applications[59]; and ii.) existing neural network architectures for image classification are inadequate for species barcoding[42]. In contrast, our analysis demonstrates that carefully designed varKodes analyzed with existing neural network architectures optimized for image classification can identify taxa with very high accuracy even from modest amounts of data. varKodes use short k-mer counts from raw sequencing reads to create a snapshot of the total genomic landscape for a given sample. Variation in varKodes can be small but remain visually perceptible among species (e.,g., of *Stigmaphyllon*, **Figure 1B**) and genera (e.g., of Malpighiaceae, **Figure 1C**). Variation is more striking among higher levels of phylogenetic divergence, such as between families in the order Malpighiales (**Figure 1D**) or different kingdoms of eukaryotes and prokaryotes (**Figure 1E**). We expected, therefore, that neural network architectures developed for image classification, (e.g., resnets[60] or vision transformers[61]) would be able to differentiate varKodes.

We first optimized hyperparameters and training conditions to maximize accuracy for species-level identification of *Stigmaphyllon*. We identified that varkodes depicting k-mer length = 7 struck a good balance between accuracy and the amount of input sequence data (**Figure 2A**). Furthermore, models trained with augmented data from several subsampled images drawn from each individual exhibited substantially better performance and greater robustness (**Figure 2A**). A linear model demonstrated that neural network architectures and training methods designed for image classification of photographs[60, 62, 63, 64, 65] are extremely useful for varKode-based identification, contrary to suggestions that classification of similar images requires specialized architectures[42]. Specifically, we observed increased accuracy with more parameter-rich neural network architectures (*ResNeXt101*[66], among those tested), augmentation with lighting transformations, *CutMix*[65] and *MixUp*[64]. Label smoothing[67] and pretraining models on photographs decreased accuracy (**Figure 3**). We also identified that these approaches enabled training with very modest datasets: four samples per taxon was sufficient for 100% median accuracy (**Figure 2B**). Errors in species identification were concentrated among sequences derived from herbarium samples that demonstrated evidence of DNA damage as is sometimes reported for ancient DNA[68] (**Figure 2B**). However, we identified that the inclusion of low-quality training samples decreased validation accuracy only among other low-quality samples but not among high-quality ones (**Figure 4**).

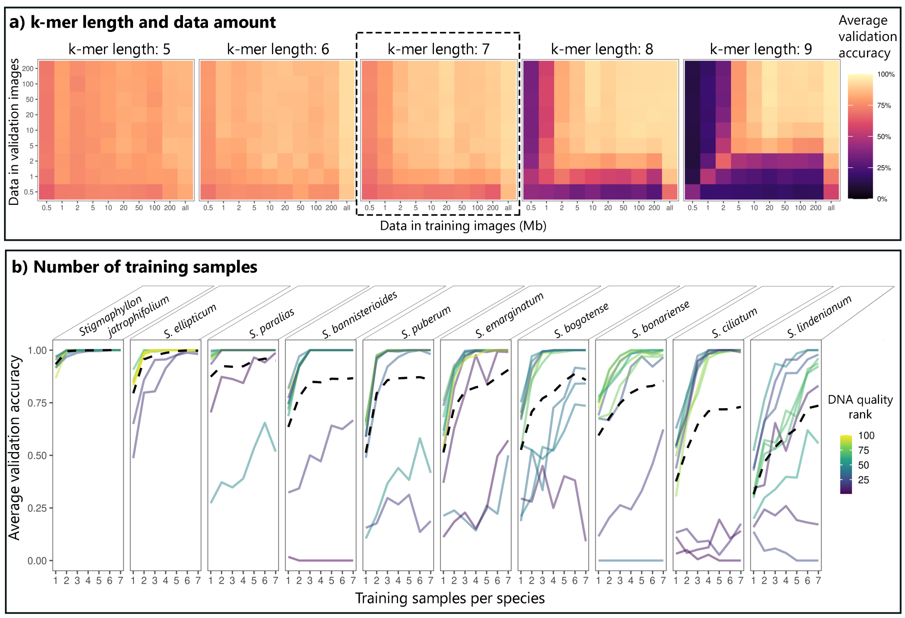

Figure 2. Neural network training of varKodes for species identification. (**A**) Effect of k-mer length and data amount used to produce varKodes on validation accuracy. Longer k-mers increase accuracy when more data are used. Mixing varKodes subsampled from different amounts of data improves accuracy. Box with dashed line (k-mer length = 7) strikes a good balance between model accuracy and amount of required data. (**B**) Validation accuracy improves with increased number of training samples per species, but even 3–4 samples are sufficient in most cases for achieving high accuracy. Each solid line represents one sample, colored by DNA quality (i.e., variation in base pair frequencies). Higher rank indicates better quality. Dashed lines represent averages across all samples.

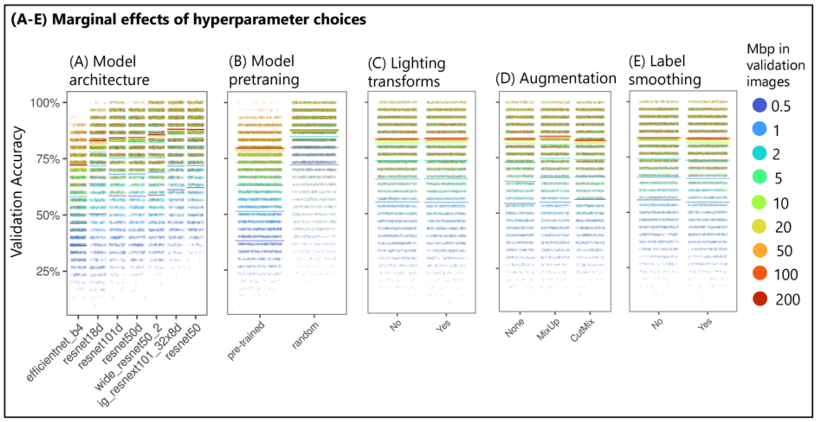

Figure 3. Marginal effects of neural network model and training options. Dots represent individual replicates, and bars depict averages. All parameters were identified to be significant in a linear model: more complex model architectures, lighting transformations, and augmentation methods *MixUp* and *CutMix* improved accuracy. However, pretraining with large image datasets and label smoothing decreased accuracy.

We hypothesized that lower-quality samples shared similar sequences resulting from common patterns of DNA damage and greater levels of microbial or human contaminants, resulting in spurious similarities in varKodes (**Figure 5**). Contaminants are thought to increase errors in genome skim methods[69]. To mitigate this problem, we applied multi-label classification[70] to our neural network models. While single-label classification models always return a single prediction (that is, an inferred label), multi-label models can return zero or more predictions, resulting in higher robustness to spurious patterns of similarity. For a set of samples with known labels used for validation, a prediction is a true positive if the predicted label matches the actual label, and a false positive if not. Failure to predict an actual label is deemed a false negative. For each validation sample, we summarized predictions as i.) correct (true positives only), ii.) incorrect (false positives only), iii.) ambiguous (multiple predictions, including true and false positives), or iv.) inconclusive (no prediction). For each test, we summarized results across all validation samples using two metrics: precision (the sum of all true positives divided by the sum of all true and false positives) and recall (the sum of all true positives divided by the sum of all true positives and negatives).

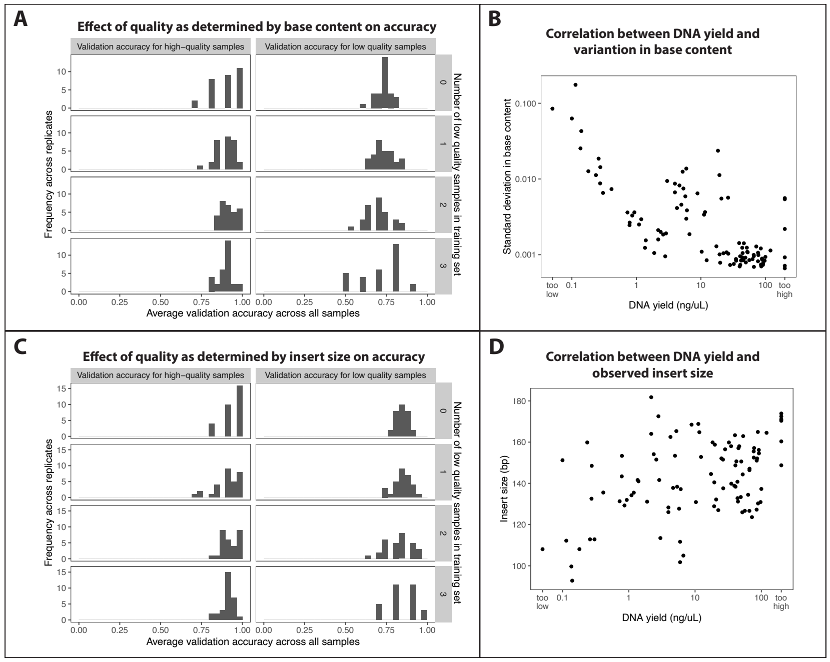

Figure 4. Effect of the inclusion of low-quality training samples, inferred from variation in base pair content (A, C) or insert size (B, D). Increasing the fraction of samples in the training set that were low-quality did not strongly affect the average validation accuracy, but it increased dispersion. Low-quality samples are the four samples with highest variation in base-pair content or shortest insert size in raw reads for each species. Panels **B** and **D** show the correlation of each quality metric with DNA extraction yield.

In summary, we developed and tested a robust and scalable method of DNA barcoding capable of training with small amounts of data, and implemented it in the ***varKoder*** software, which can process sequence data required to generate varKodes, train an image-classification neural network using varKodes, and query new data with a trained neural network. These tasks are accomplished with widely used tools for sequence processing[71, 72, 73, 74, 75] and for neural network training[76, 77, 78].

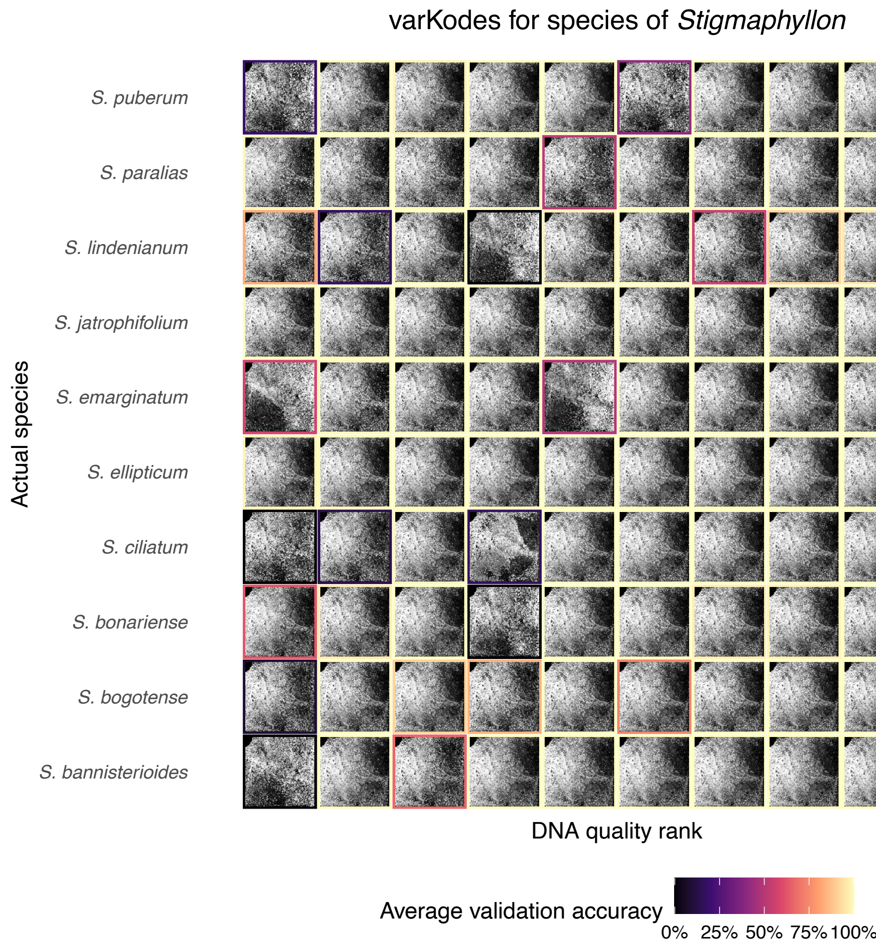

Figure 5. Low-quality DNA may lead to spurious patterns of similarity in varKodes. Samples with lower quality show varKode patterns divergent from their species more often than high-quality ones. These divergent patterns may be similar between low-quality samples across species. These samples also show reduced validation accuracy in a single-label model. For each sample, we show the varKodes produced from all DNA data available. Within each species, samples are organized from lowest (left) to highest (right) DNA quality. Bounding boxes around each sample indicate the average validation accuracy across 30 random replicates with 7 training samples per species.

varKodes are highly accurate for identification of species, genera, and families.

To test varKodes under a real-world scenario with heterogeneous data (e.g., large numbers of taxa, multiple replicates per taxon, varying sequence depth and sample quality), our *de novo* assembled genomic data set included 287 accessions: 100 samples of *Stigmaphyllon* from our initial development outlined above, plus additional genera in the families Malpighiaceae (30 genera; 151 samples), Chrysobalanaceae (8 genera; 30 samples), and Elatinaceae (1 genus; 6 samples) in the order Malpighiales. Using these data, we first demonstrated high cross-validation accuracies for species identity of *Stigmaphyllon* (83.0–93.4% correct, 91.5%-95.7% precision, 87%-96.7% recall depending on data input amount; **Figure 6A**). Most errors were inconclusive or ambiguous predictions, and not incorrect assignments.

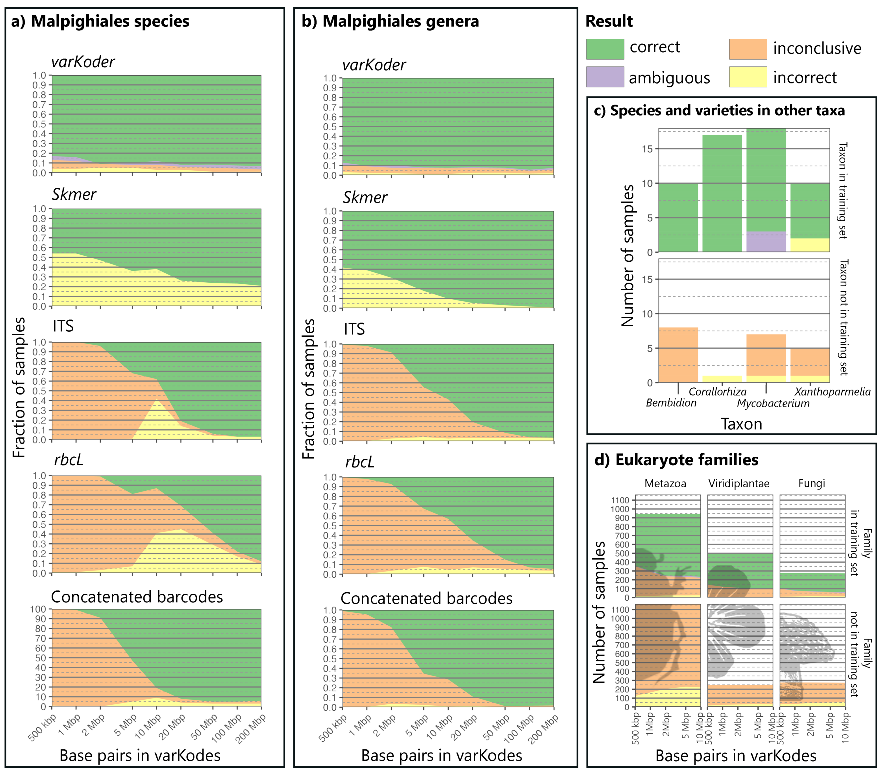

Figure 6. Performance of *varKoder* and alternative barcoding methodologies across different data sets. (**A**) Leave-one-out cross-validation to identify species of Malpighiales using different approaches and amounts of data to assemble query samples. (**B**) Same as (**A**), but for genera. (**C**) Performance for species-level identification across different publicly-available datasets: *Bembidion* beetles, *Corallorhiza* orchids, *Mycobacterium tuberculosis* bacteria, and *Xanthoparmelia* fungi. All query samples used as much data as were available. (**D**) Performance for Eukaryote family-level identification for different amounts of input data.

varKoder is also robust to the amount of input sequence data necessary for model training, performing well even at the lower range of input data (**Figure 6A**). Assuming an average genome size of about 2 Gbp for Malpighiaceae[79], the very small amount of genome skim data used to generate varKodes represented coverages of less than ~0.0002×–0.107×. Moreover, when compared to cross-validation accuracies of existing alternatives, *varKoder* accuracy is higher than *Skmer*, which showed 46% correct predictions (57.5% precision, 46% recall) with minimal data amounts and peaked at 79.1% for the larger data amounts (80% precision, 79.1% recall, **Figure 6A**). On the other hand, traditional barcodes including individual plastid genes and nuclear ribosomal ITS regions performed well for both BLAST-based (25–97% correct, 66.6–97.3% precision, 25–97% recall depending on the gene) and phylogenetic-based (94–95% correct, >99% precision, 97.2–98.4% recall for concatenated matrices) approaches when at least 50 Mbp of data was provided (**Figure 6A, Figure 7**). However, these results were much worse when <50 Mbp of data were available (down to zero correct for BLAST), with unsuccessful locus assembly leading to inconclusive predictions as the primary reason for the failure (**Figure 6A, Figure 7**). In summary, *varKoder* reaches much higher accuracy for species determination than existing methods for unprecedentedly small amounts of data and demonstrates similar accuracies for datasets when greater amounts of sequence data are available.

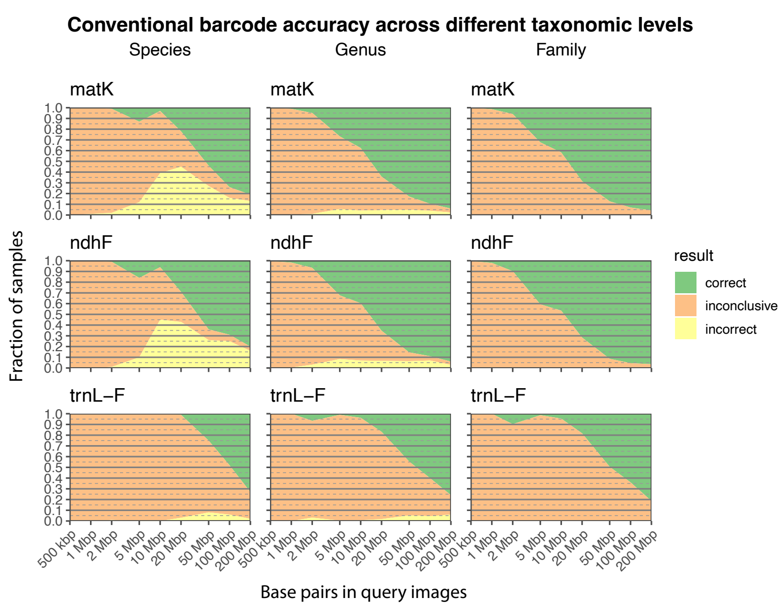

Figure 7. Accuracy of conventional barcode loci for species, genera and families within the Malpighiales.

Genus-level identification yielded similar high accuracies with *varKoder* (87.1–94.3% correct, 94.1%–97.4% precision, 89.1%–95.4% recall depending on input amount, **Figure 6B**), but with a higher rate of inconclusive predictions (2.8–7.6%). A linear model demonstrated that this higher uncertainty can be attributed to two factors: i.) samples exhibiting higher levels of DNA damage in genera other than *Stigmaphyllon* and ii.) genera trained with fewer replicates (e.g., down to 3 samples for some genera; **Figure 8**). Additionally, samples within genera share fewer genetic similarities than samples within species, which likely poses a more challenging classification problem. However, the incorrect rate is very small in all cases (1.4–3.1%) with most errors being inconclusive or ambiguous predictions. In contrast, *Skmer* exhibited better performance when larger amounts of data were used (99.2% correct, 99.2% precision, 99.2% recall for 200 Mbp), but performed poorly for lower amounts of data like those commonly generated from genome skim experiments (58.2% correct, 58.2% precision, 58.2% recall for 500 Kbp) (**Figure 6B**). Genus-level identifications using conventional barcodes in a concatenated phylogeny were up to 98.1% correct (99.2% precision, 97.2% recall) when a large amount of data (200 Mbp) was available (**Figure 6B**). But like its application at species-level identification, most predictions were inconclusive when less than 20 Mbp reads were used (**Figure 6B**). Although genome skimming can be used to sequence conventional barcodes, they are more often obtained with amplicon sequencing, which has failure rates ranging from 15–75% even with highly optimized protocols[80]. Therefore, conventional barcodes have a high number of inconclusive predictions also with amplicon sequencing. At the family level, *Skmer* and *varKoder* had near-perfect accuracy across all data amounts (>97% correct), while conventional varKodes performed well when there was sufficiently large amounts of data (**Figures 7, 9**). We note that 135 of our 287 *de novo* assembled genome skim samples had at least 200Mbp of available data (**Figure 8)**, and these are enriched for specimens that performed well in DNA library preparation and sequencing. As a result, the good performance across methods for the highest data amounts may result partly from higher-quality DNA yielding more reads with more even genome coverage.

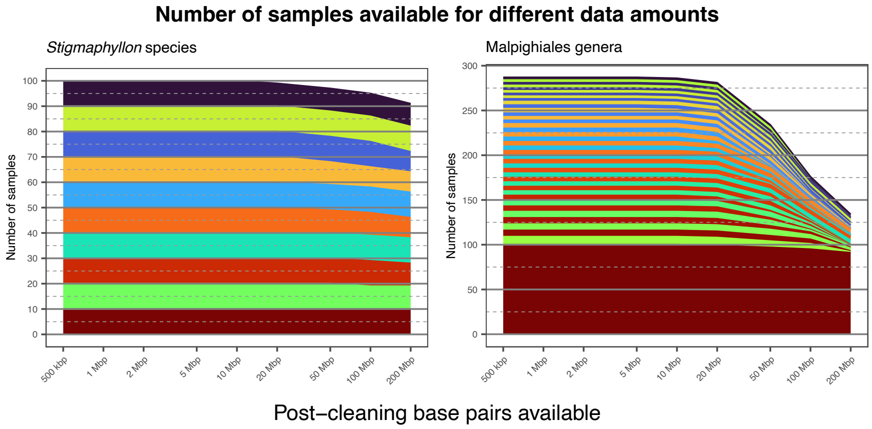

Figure 8. Number of samples available for different data amounts in each dataset. Arbitrary colors are assigned to individual taxa.

varKodes are universal and scalable across the Tree of Life.

To further test the universality of varKodes, we expanded the testing of our tool using published data from diverse clades of plants, fungi, animals, and bacteria (**Figure 6C**). These tests included species-level identification in insects (*Bembidion* beetles[81, 82]) and lichen-forming fungi (*Xanthoparmelia*[83]), species and infra–specific taxon identification in coralroot orchids (*Corallorhiza*[84]), and clinical isolate identification of evolved strains of human pathogenic bacteria (*Mycobacterium tuberculosis*[85]). In all cases, we tested the performance of *varKoder* on taxa included in the training set and on taxa not included in the training set. We identified perfect species identification (100% correct, 100% precision, 100% recall) for beetles and coralroot orchids included in the training set. For bacteria, 16% of the validation set returned ambiguous assignments; the remaining samples were correctly identified (85.7% precision, 100% recall). In lichen-forming fungi, which include DNA from both the fungal and algal partners, and thus are more challenging, 20% of the test samples returned incorrect assignments; the remainder were correct (80% precision, 80% recall). For all cases, species or varieties not included in the training set generally resulted in inconclusive results, with a minority yielding incorrect predictions **(Figure 6C)**.

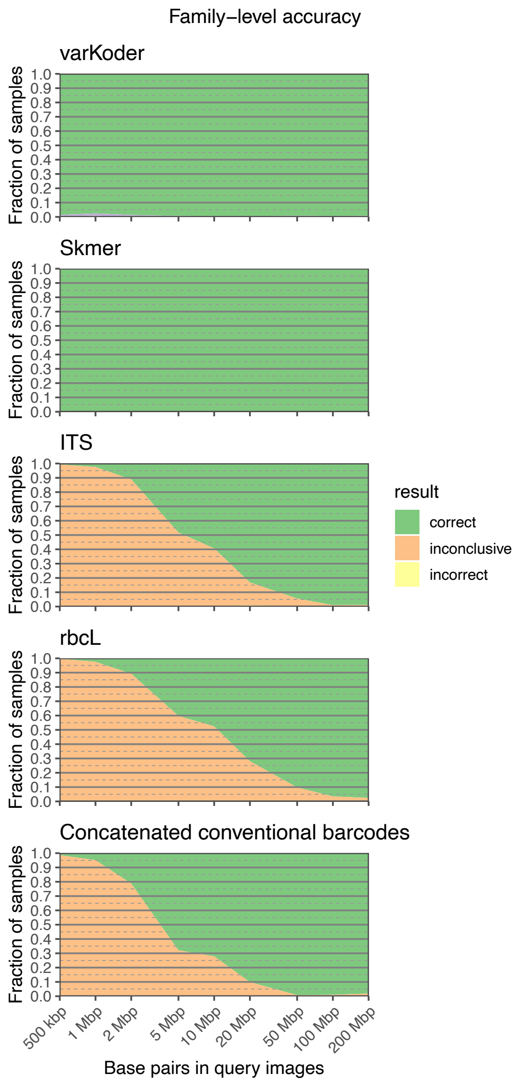

Figure 9. Comparison of *varKoder*, *Skmer*, and conventional barcode accuracy for identifying families of Malpighiales.

Finally, we tested the universality and scalability of varKodes by training a single model to identify all 861 eukaryotic families from at least three accessions per family compiled from the NCBI Sequence Read Archive. Owing to NCBI download bottlenecks, we restricted varKode construction to a more restricted amount of data per sample, downloading up to only 10 Mbp of data. This exercise achieved a rate of correct predictions of 62.1–79.6% across all kingdoms when families were included in the training set (**Figure 6D**), with most errors being inconclusive predictions (14.2–33.3%). Precision varied from 95% to 97% and recall from 65% to 78%. Similarly to the species- and variety-level exercise, families not included in the training set often yielded inconclusive predictions **(Figure 6D)**, suggesting a potential for varKoding to be used as a discovery tool when reasonably well-sampled training data sets are available.

As we note above, a single model classifying all eukaryotic families is not possible with conventional barcodes, since they are not universal. This is a central limitation of conventional barcodes. *Skmer*, the state-of-the-art genome skimming alternative, cannot be scaled to a dataset of this size: our attempt to apply it could not be finished after more than 40 days using 32 high-performance computing cores. In general, conventional barcodes, when derived from genome skimming data, require memory- and processor-intensive sequence assembly, and *Skmer* relies on pairwise all-by-all sample comparisons; its computing time and required storage both increase quadratically with the number of samples. Neural network models, on the other hand, have a fixed size, independent of the number of samples used in training, and training time scales linearly with the number of input samples. Our most complex model, trained with all eukaryote families, has about 1.3GB of disk size. varKodes images also are tiny (8.2 KB on average for k-mer length of 7) replacements to much larger genomic data sets (on average, 144 MB per sample here). A varKode model potentially trained on millions of species can therefore easily be ported to devices without continuous internet access, thus allowing for more widely distributed applications of varKoding, such as field-laboratory environments or proposed distributed genetic databases[86]. Hence, varKodes are not only comparable across the entire Tree of Life but also can leverage existing and widely available computer hardware to provide accurate and fast identifications commensurate to the scale of Earth's biodiversity.

Conclusions

varKoding represents a major advance in inventorying Earth's biodiversity. They are universal, accurate, efficient, and hold tremendous promise for scalability and adaptability. varKodes are applicable to organisms with simple or complex genomes. Although our focal test clade from Malpighiaceae specifically is known to exhibit high variation in ploidy across the family[87, 88], it did not interfere with our efforts. Indeed, further exploration may reveal that these sorts of macrostructural genomic properties form the basis of key varKode differences between some clades. In particular, varKodes i.) provide accurate identification with far less data than existing methods that use next-generation sequence data; ii.) are universal across the Tree of Life; iii.) demonstrate enhanced computational efficiency and scalability; and iv.) are modular and can improve with time alongside innovations in sequencing technologies, bioinformatics, and machine learning. Reference data for varKoding will be increasingly available from ambitious efforts including the Earth Biogenome Project[89], the African Biogenome Project[90], the 10,000 Plants Genome Project[91], and the Vertebrates Genome Project[92]. We also note that varKoding is much easier and cost-effective to obtain from low-coverage genome skims than high-quality contiguous genomes. For example, our cost for a 3× skim of herbarium samples is about $34 per sample, versus a high-quality genome which may cost tens-of-thousands of dollars each. Although varKodes inevitably will benefit from the aforementioned large-scale sequencing initiatives, a concerted effort to obtain genome skims from museum type specimens and other representative specimens could have a larger impact in a far shorter amount of time than sequencing high-quality genomes. For example, the majority of our Malpighiales samples were derived from herbarium specimens, some more than 110 years old and presently less suitable for chromosomal-level genome assembly. Thus, varKodes show tremendous promise for further automating species identification from herbaria and other natural history collections[93]. Such multiomic efforts represent a new frontier of biodiversity discovery but should be advanced effectively and ethically to preserve and protect the biodiversity heritage represented in global natural history collections for future use[94, 95].

We expect that varKoding will be invaluable to the biodiversity science community in numerous ways. One avenue to be explored is its utility for the identification of samples with poor-quality and degraded DNA, such as unidentified fragmentary fossil and subfossil remains in natural history collections[96, 97]. Because our method relies on counts of very short k-mers (7 bp), they are well-suited for varKodes while other barcoding methods are not possible. Moreover, we explicitly labeled and classified samples based on their taxonomic identities, but varKodes could in principle be used to classify a set of sequences based on any kind of metadata, as long as sufficient training data are available. For example, varKodes likely will be useful for environmental sampling initiatives in which the entire genomic composition of a sample spanning multiple species can be characterized (varKoded), even if *varKoder* is not optimized to recognize individual species or genes within a mixed sample. For example, we envisage that varKodes could be useful to correlate aquatic eDNA samples to location and water quality, to ascertain the origin of a sample for forensic study, or to or help trace the geographic origin of organisms seized during transit suspected of illegal harvesting.

Methods

Data

Taxon sampling, DNA sequencing, assembly, and annotation for newly acquired genetic data—Our newly generated plant data set included three flowering plant families, all members of the large and diverse order Malpighiales[34]: Malpighiaceae, Elatinaceae, and Chrysobalanaceae. The Malpighiaceae data are the most taxonomically comprehensive and include 251 accessions representing 161 species, which were sampled from 248 herbarium specimens and three silica-dried field collections. These represent 30 genera. Among these data, *Stigmaphyllon* has the most comprehensive species sampling, including 10 species and 10 accessions sampled per species. Elatinaceae includes 6 samples from 6 different species in the genus Elatine, and Chrysobalanaceae includes 30 accessions representing 30 species in 8 genera. All 100 *Stigmaphyllon* samples were sequenced specifically to build, validate, and test our identification models at shallower phylogenetic depths and were consequently labeled with species, genus, and family names. A key advantage of sampling *Stigmaphyllon* is that its taxonomy has been extensively revised by coauthor C. Anderson[57, 58]. Plants exhibit notoriously complex genomic architectures[98], rendering them a good test case for our investigation. Moreover, the *Stigmaphyllon* clade represents a wide array of divergence times that span distantly- (30.8 millions of years, Myr) to very closely-related (0.6 Myr) species **(Figure 1)**. The focus for the remainder of the Malpighiaceae, Chrysobalanaceae, and Elatinaceae sampling was to identify a given sample to genus. In this case, among the non-*Stigmaphyllon* samples we included 3–9 species per genus representing 29 genera of Malpighiaceae, eight of Chrysobalanaceae, and one of Elatinaceae. Each generic representative was labeled with its corresponding genus and family identification. Unlike *Stigmaphyllon,* where we included multiple accessions per species, there were no additional replicates per species for our genus-level sampling.

We used total genomic DNA extractions detailed previously for our newly included Malpighiales data[54, 99]. Where applicable, we isolated total genomic DNA from 0.01–0.02 g of silica-dried leaf material or, more commonly, herbarium collections using the Maxwell® 16 Tissue DNA Purification Kit (Promega Corporation, Inc., Madison, WI, USA). Genomic libraries were prepared using ca. 70 ng of genomic DNA per sample where possible. For DNA library preparation we used the Kapa HyperPlus library prep (Kapa Biosystems, Inc., MA, USA) with Nextflex-Ht barcodes (Bioo Scientific Corporation, TX, USA) and IDT TrueSeq barcodes (Integrated DNA Technologies, Inc., IO, USA), fragmenting DNA to 350–400 base pairs (bp), and indexing for Illumina multiplex sequencing. We verified the DNA concentration of these libraries, and fragment sizes using the Qubit dsDNA HS Assay Kit on a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA), and the Agilent TapeStation 2200 (Agilent Technologies, Inc., Waldbronn, Germany). All total genomic DNA libraries were diluted to 0.7 nM, pooled, and sequenced with the Illumina Hi-Seq 2x125 on the Genome Analyzer II (Illumina, Inc., San Diego, CA, USA) at the Bauer Core Genomics Sequencing Core Facility at Harvard University, MA, USA. The genome skimming pipeline we applied is described by Weitemier *et al.*[100] and has been extensively applied in studies by members of our coauthor group[101, 102, 103].

Public genomic data compilation—To further understand the versatility of *varKodes* more broadly across the Tree of Life, we tested species identification using genome skim data sets from four genera of plants, animals, fungi, and a bacterial species. This involved a plant data set from coralroot orchids (genus *Corallorhiza*), a well-delineated clade of mycoheterotrophic orchids[84]. This data set allowed us to assess the utility of varKodes for identifying infraspecific taxa: *Corallorhiza striata* includes several well-known and easily identifiable varieties. For animals, we assembled a *Bembidion* beetle data set, which includes well-known closely related cryptic species[81, 82]. For fungi, we used *Xanthoparmelia*, a lichen-forming genus with fungal symbionts whose species are poorly understood and which often form paraphyletic species groupings[83]. Finally, our bacterial data set was from *Mycobacterium tuberculosis*, the species of pathogenic bacteria that causes tuberculosis. This genomic data set consisted of clinical isolates from five distinct, monophyletic lineages of *M. tuberculosis* and enabled us to understand how varKodes function on an extremely recently diverged, clinically relevant bacterial lineage[85]. This data set of clinical isolates from human-adapted lineages exhibited 99.9% sequence similarity despite key differences in phenotypes, including drug resistance, virulence, and transmissibility[85]. *Mycobacterium tuberculosis* has diversified quite rapidly in humans, with nine monophyletic lineages. Divergence time estimates for the most recent common ancestor of *M. tuberculosis* are <6,000 years ago[104].

In all the above cases, we included taxa with at least two samples in the training set when using publicly available data. Our validation set consisted of randomly selected samples from these taxa. We additionally validated the model on samples from taxa with only one sample available, and, therefore, not included in the training set. Each of these four data sets were downloaded using the NCBI Sequence Read Archive.

In addition to these species-level datasets, we used NCBI Entrez to query all of the data available on SRA for Eukaryotes. We then filtered this list to accessions generated with Illumina technology and containing at least 50 million base pairs. From this filtered list, we selected all families with at least three subtaxa containing sequences. We then randomly selected one accession per subfamilial taxon, and up to 20 subtaxa per family. We used fastq-dump (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) to download up to 500,000 spots for each accession and used these to generate varKodes from 500kbp to 10Mbp of data. In each family, 80% of the accessions were used in training and the remaining 20% were used for validation. To validate model behavior for taxa not included in the training set, we downloaded all accessions from SRA in families of plants, animals, and fungi excluded from the training set but containing at least one sample with at least 50 million base pairs of data.

Initial varKode design and testing

varKode *sequence data preprocessing*—We designed images—**varKodes**—that portray relative frequencies of k-mers from low-coverage raw Illumina reads. These are similar to a 'chaos game representation' *sensu* Jeffrey[53], but optimized for raw reads in which sequence orientation is unknown (and therefore k-mers and their reverse complement are indistinguishable). We call these varKodes because they en*CODE* the *VAR*iation in k-mer frequencies in a sample.

To avoid sequencing artifacts, raw Illumina reads were lightly cleaned prior to k-mer counting and involved the following steps: identical reads were de-duplicated using *clumpify.sh* as implemented in BBtools[72, 105], adapters were removed, low-quality tails trimmed, and overlapping read pairs merged using *fastp*[74]. Next, we randomly selected subsets of cleaned reads with predefined data amounts, ranging from 500 kbp to 200 Mbp. These data subsets were used to generate a variety of input varKodes for a single sample and all such images were used for training (see main text and Figure 2A). Finally, we applied *dsk*[73] to count k-mers of a given length based on clean raw reads. *dsk* exhibits good performance with low memory requirements, which is ideal for potential applications using varKodes on low-memory devices. We note that analyses for species-level public datasets have low compute requirements and were performed on an Apple MacBook with ARM processor architecture. Bioinformatics and image classification application of this nature are typically thought to be possible only in more resourced computer servers[41], but our method demonstrates that this is not the case.

k-mer to image mapping—We developed a two-dimensional mapping of k-mers to pixels to create the varKode image. Each unique k-mer has a unique pixel location on the varKode. A desirable property of this mapping is that more similar k-mers exhibit greater spatial adjacency. We first began by listing all possible canonical k-mers to generate the mappings for k-mer lengths between 5 to 9. To identify which k-mers were more similar to each other, we counted, for each k-mer, the occurrence of smaller sub k-mers and then grouped them based on greater or lesser overall similarity. For example, each possible 5-base-pair sequence can be represented uniquely by the counts of subsequences of lengths 2 and 3 contained within it and compared similarly across other k-mers. Likewise, each possible 9-base-pair sequence can be represented uniquely by the counts of subsequences of lengths 2, 3, and 5. Moreover, since our method works with raw reads, the orientation of each sequence is unknown and therefore each k-mer represents itself and its reverse complement. For this reason, we averaged counts for each canonical k-mer and its reverse complement.

Next, we applied *t-SNE*[106], a non-linear dimensionality reduction method, to group k-mers based on their relative similarity. This allowed us to reduce canonical k-mer representation into a two-dimensional space. We noticed from this output that *t-SNE* separated k-mers mainly by AT richness, so we rotated coordinates to make this the main left-to-right axis.

Next, we transformed these data mapped in continuous space to pixels in a square grid forming the initial varKode. Our square grid was constructed with the minimum size required to fit each individual canonical k-mer to a unique grid cell (pixel). After rescaling continuous *t-SNE* coordinates, we assigned each k-mer to the closest available pixel, using randomization in the cases in which more than one k-mer overlapped in a single pixel. This procedure resulted in a mapping that uniquely assigns each k-mer to a pixel in the varKode.

Once we established the two-dimensional mapping of each k-mer to the varKode, we developed a method for transcribing k-mer counts to be represented as pixel brightness. To make varKodes as compact as possible, we used 8-bit grayscale images. As a result, for a typical 8-bit grayscale image format, we have 256 possible brightness levels per pixel. Therefore, raw k-mer counts had to be mapped to 256 values while maintaining relevant information on their variation. Because k-mer counts vary across many orders of magnitude, we first rank k-mers based on their absolute counts. We attempted alternative data transformations with the same goal in our early iterations, including log and square-root, but these were less successful in terms of final model accuracy. The ranks were subsequently sorted into 256 bins, and these represent the values used to translate ranks to pixel brightness to finalize each varKode. The varKode image is saved as a compressed png file. These operations use python libraries *numpy*[76] and *pillow*[107].

Testing the effect of k-mer length and data amount—We chose neural network models to compare varKodes because of their enhanced ability to handle images and identify complex patterns within them. We employed *fastai*[78] for this purpose, a high-level implementation of neural networks based on *pytorch*[77]. All of the model architectures we applied are image classification models available from the *timm* library[108], which have been widely tested using a variety of image types. To identify the optimal training hyperparameters for our neural network, we conducted a series of tests using our species-level data set for the genus *Stigmaphyllon*. We generated varKodes for each of the *Stigmaphyllon* samples using the workflow described above. We first tested the joint effect of k-mer length and input data amount for neural network classification accuracy by selecting three samples per species as a validation set; the remaining samples were used to train neural networks using different amounts of input data across 10 randomly generated training sets. As input data for both the validation and training sets, we randomly subsampled the original sequences into fastq files containing from 500 Kb to 200 Mb (equivalent to about 1,700 to 670,000 2x150bp Illumina reads). In this test, we only included samples that yielded at least 200 million base pairs after cleaning. We also tested the effect of combining images for all data amounts in training. For each replicate, we applied the widely used image classification neural network *resnet50* architecture[109] to classify varKodes and trained models for 30 epochs. We visualized the distribution of validation accuracy for each combination of input data amount and k-mer lengths to find a good balance between both.

Neural network optimization—After identifying an appropriate k-mer length and input data used to produce varKodes, we next tested a series of neural network training conditions. We varied the neural network model complexity, choosing from seven commonly used architectures: *resnet50*[109], *resnet-D*[60] with different depths (18, 50, 101), a wide *resnet50*[60], *efficientnet-B4*[110], and *ResNeXt101*[66]. We also tested the effect of the following: random initial weights vs pretrained weights from the *timm* library[108], presence or absence of lighting transforms, presence or absence of label smoothing, and presence or absence of augmentation strategies (i.e., *CutMix*[65] or *MixUp*[64]). Because these parameters may have complex interactions, we tested all combinations of architecture, pretraining, transforms, label smoothing, and augmentation, with 20 replicates for each combination of conditions. In each replicate, we randomly chose 20% of the samples for each species of *Stigmaphyllon* as validation and trained the model using the remainder for 30 epochs. Training was performed using all varKodes available for each sample (from 500kbp to 200Mbp). For validation, we separately evaluated whether each varKode with a different amount of data was correctly identified. For each replicate and amount of data used to validate varKodes, we recorded the average validation accuracy across the validation set. We then applied a linear model to predict the effect of all training parameters and amount of data in validation varKodes on the arc-sin transformed validation accuracy. We started from the full model containing all parameters and their interactions and reduced the model step-wise based on AIC scores, as implemented in the R function step.

Testing sample number requirements—A legitimate concern with complex neural networks is that they require vast amounts of training data and that typical skimming data sets might be insufficient for them to be useful. We tested the robustness of our models to the effect of the number of samples per species included in training by using from one to seven samples per species as training set and the remaining as validation, with 50 replicates per number of training samples. The batch size used in training was adjusted for the cases with very few samples included, so that each training epoch included about 10 batches. We included varKodes from 1Mbp to 200Mbp in both training and validation sets. In this case, we applied the training parameters informed by our previous analyses: a *resnext101* architecture, random initial weights, *CutMix* augmentation, and label smoothing for 30 epochs. We visualized the effect of the number of samples by plotting the average validation accuracy of each sample against the number of training samples used in each case.

Testing the effect of data quality—Most of the cases with low accuracy corresponded to samples with low DNA yield (**Figure 2B**). We identified that DNA extraction yield was significantly correlated with two metrics of DNA quality: average insert size and variation in nucleotide composition along reads[68] (**Figure 4**). varKodes produced from these samples may be visually distinct from other samples of the same species (**Figure 5**). For this reason, we further tested whether sample quality in training or validation impacted accuracy. Using both quality metrics, we identified the five lowest quality samples for each species. We next produced training sets using six randomly chosen samples per species, varying the number of low-quality samples included in training from zero to four. We included varKodes from 1Mbp to 200Mbp in both training and validation sets. We repeated this for 30 replicates for each number of low-quality samples. Like our tests with varying sample numbers, we applied the following training parameters: a *resnext101* architecture, random initial weights, *CutMix* augmentation, label smoothing for 30 epochs. For the validation set, we separately recorded the accuracy for high- and low-quality samples. We then visualized the effect of inclusion of low-quality samples in the training set by observing the distribution of validation accuracies for high-quality and low-quality samples across the range of number of low-quality samples included in the training set.

Implementation of varKoder—Following all of the tests described above, we implemented the optimal neural network training strategies in a python program named ***varKoder***. *varKoder* can process, train and query varKodes and is freely available on our GitHub: https://github.com/brunoasm/varKoder. Because it employs standard neural network frameworks (namely, *pytorch*[77], *fastai*[78], and *timm*[108]), any of the image classification models and training hyperparameters available now or in the future via these libraries can be easily adapted and applied to varKode classification. For example, since our initial tests, we have identified that a vision-transformer architecture[61] outperforms convolutional neural networks in varKode classification. This was also observed in other computer-vision tasks[111]. Moreover, we have implemented a multi-label model as the default to increase robustness to low-quality varKodes with little diagnostic information in the training set. This was done by using an asymmetric multi-label loss function[70] instead of the standard cross-entropy loss function used in single-label classification. A vision-transformer architecture and multi-label classification are now default in *varKoder* v.0.8.0, which was used in all subsequent analyses.

varKoder evaluation and comparison to alternatives using a de novo Malpighiales genomic dataset

varKoder—To test *varKoder* performance in a complex dataset spanning multiple taxonomic levels and varying phylogenetic depths, we used the Malpighiales dataset including genera in Elatinaceae, Chrysobalanaceae and Malpighiaceae. Species of *Stigmaphyllon* (Malpighiaceae) were labeled with species, genus, and family names; all other samples were labeled with genus and family names. We tested the performance of *varKoder* in each sample with leave-one-out cross-validation. For each sample, we retained it as validation and trained a neural network using all of the other samples. In preliminary assessments, we found that a vision transformer architecture combined with a multi-label model sometimes led to instability in training for some datasets. For that reason, we used a two-step approach. Models were pre-trained for 20 epochs as single-label, using the least inclusive taxonomic assignment available for each sample and a base learning rate of 0.05. Next, we trained for an additional 10 epochs using the pre-trained weights but with a much smaller learning rate (0.005) and a multi-label output. Training samples included varKodes from 500 Kbp to 200 Mbp, and we recorded validation accuracy separately for varKodes produced from each amount of data. We used an arbitrary confidence threshold of 0.7 to make predictions in the multilabel models. For validation samples, we deemed a prediction correct if only the correct taxon was predicted for each taxonomic rank (i.e., species, genus, family). We deemed a prediction incorrect if one or more predictions passed the threshold for a taxonomic rank, but none match the actual label. When predicted labels included both the correct and incorrect taxa, we deemed it ambiguous. If the output prediction included no taxon with confidence above the threshold, we considered it as inconclusive. As metrics across all samples, we used prediction and recall, averaged across all predictions. We visualized the fraction of correct, incorrect, ambiguous, and inconclusive samples for each taxonomic rank and each amount of data used to produce varKodes.

Skmer—To compare *varKoder* with alternative methods, we used fastq files cleaned and subsampled by *varKoder* as input files to *Skmer*. In this case, we also used leave-one-out cross-validation to evaluate performance. For each amount of input data (500Kbp to

200Mbp), we cycled through all samples, constructing a *Skmer* database with the "*skmer reference*" command and including all samples but one. We then used the "*skmer query*" command on the sample left out and deemed the identification as correct if the sample in the reference database with closest estimated genetic distance had the correct taxon label. Because *Skmer* could always query a sample and there is no objective criterion to consider matches beyond the best match, the output predictions can only be correct or incorrect, but not inconclusive or ambiguous. We visualized the results similarly as we did with *varKoder*.

Conventional plant barcodes—To infer phylogenies from our genome skim data (Figure 1), we applied the *PhyloHerb* bioinformatic pipeline[112], which has been recently applied by a variety of projects from algae to flowering plants[99, 101, 102]. Briefly, this pipeline works as follows: for plastid loci, *PhyloHerb* maps raw short reads to a database of land plant plastid genomes. Mapped reads are then assembled into scaffolds using *SPAdes*[113] and plastid loci are identified using nucleotide BLAST searches with a default e-value threshold of 1e-40. *PhyloHerb* then outputs orthologous plastid genes into individual FASTA files, which are fed directly into MAFFT v7.407[114] for alignment. Alignments are then concatenated into a super matrix using the 'conc' function within the *PhyloHerb* package. Phylogenies for both individual locus and the concatenated alignment were inferred with IQTREE v2.0.6 using the GTR+GAMMA model with 1000 ultrafast bootstrap replicates[115].

To recover the traditional plant barcodes, *rbc*L, *mat*K, *trn*L-F, *ndh*F, and ITS, from our Malpighiales genome skim data, we applied GetOrganelle v1.7.7.0[116] and *PhyloHerb* v1.1.1[112] to automatically assemble and extract these DNA markers, respectively. Briefly, the complete or subsampled genome skim data were first assembled into plastid genomes or nuclear ribosomal regions using *GetOrganelle*[116] with its default settings. Next, *PhyloHerb* was applied to extract the relevant barcode genes using its built-in BLAST database. To test whether these traditional barcodes provided accurate identification to species, genus, and family, we ran an all-by-all BLASTn analysis for each individual gene across the same data subsampling schemes as *Skmer* and *varKoder*. BLAST targets were always drawn from assemblies using all the data available for each specimen, whereas queries included assemblies from input data amounts varying from 500 Kbp to 200 Mbp.

Within each BLAST analysis for each one of the Malpighiales accessions, we deemed an identification to be correct if the best non-self BLAST hit came from the same taxon, and incorrect otherwise. We deemed it inconclusive if the locus could not be assembled for that amount of data or BLAST returned no results. For concatenated barcodes, we produced a phylogenetic tree for each amount of data, and deemed an identification to be correct if the sample with lowest patristic distance came from the same taxon. We deemed it to be inconclusive when none of the genes in the concatenated dataset could be assembled for a sample. We visualized results similarly to *varKoder*, separately for each conventional barcoding gene and for the concatenated dataset.

varKoder application in diverse published datasets

Species-level identification in plants, animals, fungi, and bacteria—For each of the four organismal clades, we trained a multi-label model that included five species with at least three samples per species. For *Bembidion*, we included five species with five samples per species. For *Corallorhiza*, we included five species (or varieties) with at least five samples per species, except for *C. striata* var. *vreelandii* and *C. striata* var. *striata,* for which we included six and seven samples each, respectively. For *Mycobacterium tuberculosis*, we included representatives of five monophyletic *M. tuberculosis* lineages (L1, L2, L3, L4.1.i1.2.1, and L4.3.i2) with seven clinical isolates per lineage. Samples for *Bembidion, Corallorhiza*, and *M. tuberculosis* isolates all formed monophyletic groups, whereas *Xanthoparmelia* species did not. Since the *Xanthoparmelia* species were paraphyletic, we subsampled only monophyletic groups for model training. In this case, four species included three samples per species (*X. camtschadalis*, *X. mexicana*, *X. neocumberlandia*, and *X. coloradoensis*) and one species included five samples per species (*X. chlorochroa*). One potential confounding factor for the *Xanthoparmelia* model is that *Xanthoparmelia* is a lichen-forming fungus and thus genome skim data represents a chimera of fungal and algal genomes representing both partners in this unique symbiosis. Species of the algal symbiont *Trebouxia* are flexible generalists across fungal species *Xanthoparmelia*. Since these genome skims are a mix of both algal photobiont and fungus, we hypothesize the accuracy of our model decreased because of the more generalist nature of *Trebouxia*[117].

For all four test cases, we applied default *varKoder* v.0.8.0 parameters for generating varKode images, training each model, and testing the accuracy of the trained model using the 'query' function. In all cases, we included all the available data for each training or validation sample. To test if trained models accurately predicted species identity, we queried them using genome skim samples not used for training but from the same species included in the model. We also tested genome skim samples of species within the same genus but not used in model training. As in the case of Malpighiales, we set the threshold to make a prediction to 0.7 and used the same criteria to consider a prediction correct, incorrect, inconclusive, or ambiguous. We separately evaluated results for taxa with representatives included in the training set and taxa used only as queries, without conspecific samples in the training set.

All eukaryotic families data set from SRA—Each accession obtained from SRA was labeled with its family identification obtained from NCBI. Because of the larger size of this dataset, a leave-one-out cross-validation approach would have been intractable. Therefore, we randomly selected 80% of the samples in each family as the training set and used the remainder for validation. Similarly to Malpighiales, we used a two-step training method by pre-training as a single-label model and finalizing with a multi-label model. However, because of the larger size of this dataset, we adjusted the base learning rate and batch size to accelerate training. Namely, pre-training was done with a learning rate of 0.1 and a batch size of 300 for 30 epochs. Final training was done with the same batch size but a smaller base learning rate of 0.01 in 5 epochs with frozen body weights and three epochs with unfrozen weights.

Acknowledgements

BdM received postdoctoral fellowships from the Harvard University Museum of Comparative Zoology and the Smithsonian Tropical Research Institute during the early stages of this study. LC was supported by Harvard University and by a Stengl Wyer scholarship from the University of Texas at Austin. PF was supported by LVMH Research, and Dior Science. YY was supported by a postdoctoral fellowship from Harvard University Herbaria. CCD was supported by Harvard University, LVMH Research, Dior Science, and National Science Foundation grants DEB-1355064 and DEB-0544039. Computations were performed at the Harvard Cannon Cluster and the Field Museum Grainger Bioinformatics Center. We thank the Bauer Core Facility, and especially Claire Reardon, at Harvard University for providing technical support during the laboratory process. We thank R. Asprino and K. Peterson for their assistance and the team at Sound Solutions for Sustainable Science for carefully editing early versions of our manuscript.

Author contributions

BdM conceived varKodes and wrote the program *varKoder*. BdM and CCD designed the research. CCD, XD, YY, LCM, and CA collected the new sequence data. BdM, CCD, LC, YY, PJF analyzed and interpreted the data. LCM prepared the figures. BdM and CCD wrote the manuscript with key contributions from LC, YY and PJF. All authors approved the manuscript.

Code Availability

The current version of varKoder is available at https://github.com/brunoasm/varKoder. A fastai model pre-trained on SRA data is available at https://huggingface.co/brunoasm/vit_large_patch32_224.NCBI_SRA

References

1. Hebert PDN, Ratnasingham S, de Waard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Royal Soc B* **270**, S96-S99 (2003).

2. Kress WJ. Plant DNA barcodes: applications today and in the future. *J Syst Evol* **55**, 291-307 (2017).

3. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes* **7**, 355-364 (2007).

4. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol* **21**, 2045-2050 (2012).

5. Seifert KA. Progress towards DNA barcoding of fungi. *Mol Ecol Res* **9 Suppl s1**, 83-89 (2009).

6. Sharkey MJ*, et al.* Minimalist revision and description of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. *Zookeys* **1013**, 1-665 (2021).

7. Lahaye R*, et al.* DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA*, 0709936105 (2008).

8. Kuzmina ML*, et al.* Using herbarium-derived DNAs to assemble a large-scale DNA barcode library for the vascular plants of Canada. *Appl Plant Sci* **5**, apps.1700079 (2017).

9. Muñoz-Rodríguez P*, et al.* A taxonomic monograph of *Ipomoea* integrated across phylogenetic scales. *Nat Plants* **5**, 1136-1144 (2019).

10. Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Natl Acad Sci USA*, (2004).

11. Zeale MR, Butlin RK, Barker GL, Lees DC, Jones G. Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Mol Ecol Res* **11**, 236-244 (2011).

12. Nitta JH, Meyer JY, Taputuarai R, Davis CC. Life cycle matters: DNA barcoding reveals contrasting community structure between fern sporophytes and gametophytes. *Ecol Monograph* **87**, 278-296 (2016).

13. Kress WJ, *et al.* Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc Natl Acad Sci USA* **106**, 18621-18626 (2009).

14. Willis CG, Franzone BF, Xi Z, Davis CC. The establishment of Central American migratory corridors and the biogeographic origins of seasonally dry tropical forests in Mexico. *Front Genet* **5**, 433 (2014).

15. Willerslev E, *et al.* Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science* **317**, 111-114 (2007).

16. Crump SE, *et al.* Ancient plant DNA reveals High Arctic greening during the Last Interglacial. *Proc Natl Acad Sci USA* **118**, e2019069118 (2021).

17. Kjær KH, *et al.* A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283-291 (2022).

18. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* **107**, 6477-6481 (2010).

19. Rollo F, Ubaldi M, Ermini L, Marota I. Ötzi's last meals: DNA analysis of the intestinal content of the Neolithic glacier mummy from the Alps. *Proc Natl Acad Sci USA* **99**, 12594-12599 (2002).

20. Yu J, Wu X, Liu C, Newmaster S, Ragupathy S, Kress WJ. Progress in the use of DNA barcodes in the identification and classification of medicinal plants. *Ecotoxicol Environ Saf* **208**, 111691 (2021).

21. Ashfaq M, Hebert PDN. DNA barcodes for bio-surveillance: regulated and economically important arthropod plant pests. *Genome* **59**, 933-945 (2016).

22. Eaton MJ, Meyers GL, Kolokotronis S-O, Leslie MS, Martin AP, Amato G. Barcoding bushmeat: molecular identification of Central African and South American harvested vertebrates. *Conserv Genet* **11**, 1389-1404 (2010).

23. Liu J, *et al.* Integrating a comprehensive DNA barcode reference library with a global map of yews (*Taxus* L.) for forensic identification. *Mol Ecol Res* **18**, 1115-1131 (2018).

24. Ogden R, Dawnay N, McEwing R. Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement. *Endanger Species Res* **9**, 179-195 (2009).

25. Williamson J*, et al.* Exposing the illegal trade in cycad species (Cycadophyta: *Encephalartos*) at two traditional medicine markets in South Africa using DNA barcoding. *Genome* **59**, 771-781 (2016).

26. Costa FO, Carvalho GR. The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of Fish. *Genomics, Society and Policy* **3**, 29 (2007).

27. Gao Z, Liu Y, Wang X, Wei X, Han J. DNA mini-barcoding: a derived barcoding method for herbal molecular identification. *Front Plant Sci* **10**, (2019).

28. Molina J*, et al.* Possible loss of the chloroplast genome in the parasitic flowering plant *Rafflesia lagascae* (Rafflesiaceae). *Mol Biol Evol* **31**, 793-803 (2014).

29. Cai L*, et al.* Deeply altered genome architecture in the endoparasitic flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). *Curr Biol* **31**, 1002-1011.e1009 (2021).

30. Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* **293**, 2242-2245 (2001).

31. Wang J, Luo J, Ma Y-Z, Mao X-X, Liu J-Q. Nuclear simple sequence repeat markers are superior to DNA barcodes for identification of closely related *Rhododendron* species on the same mountain. *J Syst Evol* **57**, 278-286 (2019).

32. Su X, Wu G, Li L, Liu J. Species delimitation in plants using the Qinghai–Tibet Plateau endemic Orinus (Poaceae: Tridentinae) as an example. *Ann Bot* **116**, 35-48 (2015).

33. Lu Z, Sun Y, Li Y, Yang Y, Wang G, Liu J. Species delimitation and hybridization history of a hazel species complex. *Ann Bot* **127**, 875-886 (2021).

34. Cai L*, et al.* The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. *Syst Biol* **70**, 491-507 (2021).

35. Clarke LJ, Soubrier J, Weyrich LS, Cooper A. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Mol Ecol Res* **14**, 1160-1170 (2014).

36. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc Natl Acad Sci USA* **105**, 13486-13491 (2008).

37. Xiong H*, et al.* Species tree estimation and the impact of gene loss following whole-genome duplication. *Syst Biol* **71**, 1348-1361 (2022).

38. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Amer J Bot* **99**, 349-364 (2012).

39. Bohmann K, Mirarab S, Bafna V, Gilbert MTP. Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Mol Ecol* **29**, 2521-2534 (2020).

40. Sarmashghi S, Bohmann K, P. Gilbert MT, Bafna V, Mirarab S. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol* **20**, 34 (2019).

41. Borowiec ML, Dikow RB, Frandsen PB, McKeeken A, Valentini G, White AE. Deep learning as a tool for ecology and evolution. *Methods Ecol Evol* **13**, 1640-1660 (2022).

42. Millán Arias P, Alipour F, Hill KA, Kari L. DeLUCS: Deep learning for unsupervised clustering of DNA sequences. *PLOS ONE* **17**, e0261531 (2022).

43. Kari L*, et al.* Mapping the space of genomic signatures. *PLOS ONE* **10**, e0119815 (2015).

44. Fiannaca A*, et al.* Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform* **19**, 198 (2018).

45. Linard B, Swenson K, Pardi F. Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinform* **35**, 3303-3312 (2019).

46. Desai HP, Parameshwaran AP, Sunderraman R, Weeks M. Comparative study using neural networks for 16S ribosomal gene classification. *J Comput Biol* **27**, 248-258 (2020).

47. Shang J, Sun Y. CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. *Methods* **189**, 95-103 (2021).

48. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* **521**, 436-444 (2015).

49. Cong Y, Ye X, Mei Y, He K, Li F. Transposons and non-coding regions drive the intrafamily differences of genome size in insects. *iScience* **25**, 104873 (2022).

50. Heckenhauer J*, et al.* Genome size evolution in the diverse insect order Trichoptera. *GigaScience* **11**, giac011 (2022).

51. Schley RJ*, et al.* The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. *New Phytol* **236**, 433-446 (2022).

52. Sproul JS, Barton LM, Maddison DR. Repetitive DNA profiles reveal evidence of rapid genome evolution and reflect species boundaries in ground beetles. *Syst Biol*, (2020).

53. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res* **18**, 2163-2170 (1990).

54. Davis CC, Anderson WR. A complete generic phylogeny of Malpighiaceae inferred from nucleotide sequence data and morphology. *Amer J Bot* **97**, 2031-2048 (2010).

55. Cai L, Xi Z, Peterson K, Rushworth C, Beaulieu J, Davis CC. Phylogeny of Elatinaceae and the tropical Gondwanan origin of the Centroplacaceae (Malpighiaceae, Elatinaceae) clade. *PLOS ONE* **11**, e0161881 (2016).

56. Davis CC, Anderson WR, Donoghue MJ. Phylogeny of Malpighiaceae: evidence from chloroplast *ndhF* and *trnL-F* nucleotide sequences. *Amer J Bot* **88**, 1830-1846 (2001).

57. Anderson C. Revision of *Ryssopterys* and transfer to *Stigmaphyllon* (Malpighiaceae). *Blumea* **56**, 73–104 (2011).

58. Anderson C. Monograph of *Stigmaphyllon* (Malpighiaceae). *Syst Bot Monogr* **51**, 1-313 (1997).

59. Christin S, Hervet É, Lecomte N. Applications for deep learning in ecology. *Methods Ecol Evol* **10**, 1632-1644 (2019).

60. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M. Bag of tricks for image classification with convolutional neural networks. *arXiv*, (2018).

61. Vaswani A*, et al.* Attention is all you need. (2017).

62. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*) (2016).

63. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. *arXiv*, (2018).

64. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: beyond empirical risk minimization. *arXiv*, (2018).

65. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: regularization strategy to train strong classifiers with localizable features. *arXiv*, (2019).

66. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. *arXiv* **1611.05431**, (2017).

67. Goodfellow I. *Deep learning*. The MIT Press (2016).

68. Weiß CL*, et al.* Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R Soc Open Sci* **3**, 160239 (2016).

69. Rachtman E, Balaban M, Bafna V, Mirarab S. The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. *Mol Ecol Res* **20**, 649-661 (2020).

70. Ben-Baruch E*, et al.* Asymmetric loss for multi-label classification. *arXiv*, (2021).

71. Bushnell B. BBMap.). https://sourceforge.net/projects/bbmap/ (2022).

72. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE* **12**, e0185056 (2017).

73. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. *Bioinform* **29**, 652-653 (2013).

74. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinform* **34**, i884-i890 (2018).

75. Tange O. *GNU Parallel 2018*. Ole Tange (2018).

76. Harris CR*, et al.* Array programming with NumPy. *Nature* **585**, 357-362 (2020).

77. Paszke A*, et al.* PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*). Curran Associates, Inc. (2019).

78. Howard J, Gugger S. Fastai: a layered API for deep learning. *Information* **11**, 108 (2020).

79. Pellicer J, Leitch IJ. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol* **226**, 301-305 (2020).

80. D'Ercole J, Prosser SWJ, Hebert PDN. A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)—patterns of nucleotide misincorporation. *PeerJ* **9**, e10420 (2021).

81. Sproul JS, Barton LM, Maddison DR. Repetitive DNA profiles reveal evidence of rapid genome evolution and reflect species boundaries in ground beetles. *Syst Biol* **69**, 1137-1148 (2020).

82. Sproul JS, Maddison DR. Cryptic species in the mountaintops: species delimitation and taxonomy of the *Bembidion* breve species group (Coleoptera: Carabidae) aided by genomic architecture of a century-old type specimen. *Zool J Linn Soc* **183**, 556-583 (2018).

83. Keuler R*, et al.* Interpreting phylogenetic conflict: hybridization in the most speciose genus of lichen-forming fungi. *Mol Phylog Evol* **174**, 107543 (2022).

84. Barrett CF, Wicke S, Sass C. Dense infraspecific sampling reveals rapid and independent trajectories of plastome degradation in a heterotrophic orchid complex. *New Phytol* **218**, 1192-1204 (2018).

85. Freschi L*, et al.* Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. *Nat Commun* **12**, 6099 (2021).

86. Kimura LT*, et al.* Amazon Biobank: a collaborative genetic database for bioeconomy development. *Funct Integr Genomics* **23**, 101 (2023).

87. Cameron KM, Chase MW, Anderson WR, Hills HG. Molecular systematics of Malpighiaceae: evidence from plastid *rbcL* and *matK* sequences. *Amer J Bot* **88**, 1847-1862 (2001).

88. Anderson WR. Chromosome numbers of neotropical Malpighiaceae. *Contr Univ Michigan Herb* **19**, 341–354 (1993).

89. Ebenezer TE*, et al.* Africa: sequence 100,000 species to safeguard biodiversity. *Nature* **603**, 388-392 (2022).

90. Lewin HA*, et al.* The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci USA* **119**, e2115635118 (2022).

91. Cheng S*, et al.* 10KP: A phylodiverse genome sequencing plan. *GigaScience* **7**, giy013 (2018).

92. Staff E. A reference standard for genome biology. *Nat Biotechnol* **36**, 1121-1121 (2018).

93. Davis CC. The herbarium of the future. *Trends Ecol Evol* **38**, 412-423 (2023).

94. Davis CC. Collections are truly priceless. *Science* **383**, 1035-1035 (2024).

95. Davis CC, Sessa EB, Paton A, Antonelli A, Teisher J. The destructive sampling conundrum and guidelines for effective and ethical sampling of herbaria. *EcoEvoRxiv*, (2024).

96. Card DC, Shapiro B, Giribet G, Moritz C, Edwards SV. Museum genomics. *Annu Rev Genet* **55**, 633-659 (2021).

97. Davis CC. The herbarium of the future. *Trends Ecol Evol*, (2022).

98. Lynch M. *The origins of genome architecture*. Sinauer Associates (2007).

99. Marinho LC*, et al.* Plastomes resolve generic limits within tribe Clusieae (Clusiaceae) and reveal the new genus Arawakia. *Mol Phylog Evol* **134**, 142-151 (2019).

100. Weitemier K*, et al.* Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Appl Plant Sci* **2**, 1400042 (2014).

101. Lucas MC*, et al.* Phylogenetic relationships of *Tovomita* (Clusiaceae): carpel number and geographic distribution speak louder than venation pattern. *Syst Bot* **46**, 102-108 (2021).

102. Lyra GdM*, et al.* Phylogenomics, divergence time estimation and trait evolution provide a new look into the Gracilariales (Rhodophyta). *Mol Phylog Evol* **165**, 107294 (2021).

103. Yan Y, Davis CC, Dimitrov D, Wang Z, Rahbek C, Borregaard MK. Phytogeographic history of the tea family inferred through high-resolution phylogeny and fossils. *Syst Biol* **70**, 1256-1271 (2021).

104. Sabin S*, et al.* A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biol* **21**, 201 (2020).

105. Bushnell B. BBtools v.37.61.) (2017).

106. van der Maaten L, Hinton G. Viualizing data using t-SNE. *JMLR* **9**, 2579-2605 (2008).

107. Clark A. Pillow, Version 9.4.0. Software. https://pypi.org/project/Pillow/. (2023).

108. Wightman R. PyTorch Image Models. (2019).

109. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv* **1512.03385**, 1512.03385-01512.03385 (2015).

110. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. *arXiv*, abs/1905.11946 (2019).

111. Dosovitskiy A, *et al.* An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*, (2020).

112. Cai L, Zhang H, Davis CC. PhyloHerb: A high-throughput phylogenomic pipeline for processing genome skimming data. *Appl Plant Sci* **10**, e11475 (2022).

113. Bankevich A, *et al.* SPAdes: a new genome assembly algorithm and Its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477 (2012).

114. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).

115. Minh BQ, *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).

116. Jin J-J, *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **21**, 241 (2020).

117. Leavitt SD, *et al.* Fungal specificity and selectivity for algae play a major role in determining lichen partnerships across diverse ecogeographic regions in the lichen-forming family Parmeliaceae (Ascomycota). *Mol Ecol* **24**, 3779-3797 (2015).