varKodes extend DNA barcoding to ultrafine-grained identity, place, and state

Authors: Peter J. Flynn[1], Jackson Kehoe[1,2], T. Jonathan Davies[3], Lorène Gourguillon[4], Virginie Couturaud[5], Bruno A. S. de Medeiros[6], Charles C. Davis[1*]

[1] Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge, Massachusetts, 02138 USA;
[2] Chicago Botanic Garden, Glencoe, Illinois, 60022, USA;
[3] Department of Botany, V, Vancouver, BC V6T 1Z4, Canada;
[4]LVMH Research, 185 Avenue de Verdun, 45804 Saint Jean de Braye CEDEX, France;
[5]Department of Scientific Communication, Parfum Christian Dior, Neuilly sur Seine, France;
and
[6] Field Museum, Chicago, Illinois, 60605, USA

*Correspondence: cdavis@oeb.harvard.edu

Abstract

DNA barcoding has transformed biodiversity science, but locus-based markers often fail across taxa, at shallow phylogenetic depths, and in mixed or spatially structured samples. Here, we test whether varKoder, an alignment-free framework that converts low-coverage genomic reads into image-based genomic signatures (varKodes), can extend barcoding from taxonomic identification to broader inference of identity, place, and state. Across five public genomic and metagenomic datasets–domestic dogs, grape cultivars, soils, human skin microbiomes, and human stool microbiomes–varKodes resolved signals that span single-organism genomes and mixed microbial communities. They distinguished dog breeds and grape cultivars despite exceptionally recent divergence, hybrid ancestry, and complex pedigree histories. They also assigned skin microbiome samples to source individuals and body-site categories, demonstrating that host-associated microbial mosaics can function as individual-level genomic fingerprints. Beyond identity, varKodes recovered geographic provenance from soil metagenomes and distinguished skin microbiomes by country of origin, supporting metagenomic mosaics as location-specific signatures. Finally, varKodes classified inflammatory bowel disease status from stool metagenomes, providing a proof of concept for identifying biological state using the same kind of genomic signature. Performance was strongest for well-sampled or phylogenetically well-differentiated groups, indicating that denser reference data and retraining should improve accuracy. Together, these results demonstrate that DNA barcodes can be reconceived as retrainable, genome-wide signatures for supervised inference across a wide variety of biological scales, sample types, and applications.

Keywords: DNA barcoding | genome skimming | metagenomics | genomic signatures | geographic provenance | forensics | medical diagnostics

Introduction

High-throughput sequencing has transformed biodiversity research, and DNA barcoding is widely used for assigning unknown samples to named taxa[1–4]. However, traditional barcoding remains constrained by its reliance on one or a few loci and therefore barcodes are not universal across the tree of life[5]. Even when appropriate loci exist, short markers often provide insufficient resolution to distinguish closely related taxa or sub-specific taxa, including population level variation, recent domesticates, cultivars, or individuals[6]. These limitations are especially problematic for the fastest-growing applications of genomic sequencing, which involve complex or spatially structured samples, including eDNA, soil metagenomes, and host-associated microbiomes. In such cases, the scientific question is not only “What taxon is present?”, but may also include where a sample originated from geographically or what biological condition or state it reflects. Existing workflows for taxonomy, ecology, microbiology, biosecurity, food safety, forensics, medical diagnostics remain fragmented, because methods designed around particular marker loci, reference databases, or other features (e.g., spatial location, disease state) often do not transfer easily across the variety of organisms, sample types, or applications.

A major methodological shift in biodiversity science is therefore moving identification from locus-based markers to genome-wide signatures: standardized representations inferred from sparse genomic reads and linked to user-defined labels, from species and genera to ecological, spatial, or biological states. As part of this shift, we developed varKoder, an alignment-free method that converts very low coverage genome-skim data into flattened image-based genomic landscapes, or varKodes, for machine-learning classification[5]. varKodes do not require assembly, alignment, or predefined markers, and rely only on a small random fraction of a genome. They therefore expand the operational meaning of a barcode from a short, locus-specific sequence to a retrainable genome-wide signature that can be applied across taxa, mixed-community samples, and diverse classification objectives. Moreover, this assembly-free use of low-coverage short-read data aligns with a broader expansion of genome skimming and short-read sequencing as scalable tools for biodiversity monitoring, species identification, and comparative genomics[7]. In short, varKodes shift barcoding from a tool used primarily for taxonomic identification to a broader framework for genomic inference spanning multiple sample types, objectives, and disciplines (**Fig. 1**).

Despite the demonstrated potential of varKodes, two key aspects of our proposed expanded barcoding framework remain largely underexplored. First, the resolution limit of varKodes applied to low-coverage data is unclear. Many applied problems require

taxonomic discrimination among lineages separated by extremely shallow divergence, on the order of tens to hundreds of generations, including dog breeds and crop cultivars, where recent human selection, admixture, and complex naming histories complicate inference. Second, it remains uncertain whether varKodes can reliably leverage metagenomic mosaics, or whole-sample community DNA profiles, that may encode geo-nomic signatures: genomic or metagenomic patterns that carry information about geographic origin, ecological context, or biological condition. If such subtle differences could be detected, this would extend varKodes beyond taxonomic assignment to a broader three-part framework: identity, place, and state (**Fig. 1**). Moreover, if their resolution is sufficient, varKodes could be used to predict geographic provenance (e.g., from a confiscated sample to determine point of origin of illegally trafficked goods/exports) or other sample attributes (e.g., medical diagnosis [disease versus not diseased] or habitat condition [e.g., an agricultural field versus a forest]).

Together, these possibilities motivate a more direct exploration of whether varKodes can recover signals of identity, place, and state spanning single-organism genomes and mixed-community metagenomic mosaics, without requiring a custom method for each kind of sample or question. Here, we evaluate varKodes at these boundaries using five publicly available datasets to test their utility, and genomic signatures more generally, to finely discriminate identity, place, and biological state. Specifically, we test whether varKodes can accurately discriminate i.) domestic dog (*Canis lupus familiaris*) samples by breed; ii.) grapevine (*Vitis vinifera*) samples by cultivar in a dataset shaped by shallow divergence and hybridization; iii.) human-skin microbiomes by individual host identity, body site, and geographic source; iv.) soil metagenomes by geographic origin across structured spatial scales and land-use categories; and v.) stool metagenomes by country of origin and inflammatory bowel disease status as a proof-of-concept clinical application for varKode-based classification. Together, our analyses provide a robust empirical assessment of how little data are required for reliable classification; how accuracy and performance of varKodes varies with time since divergence, admixture; and whether metagenomic mosaics can support geo-nomic signatures for fine-scale spatial and medical diagnostic inference within a single workflow. By focusing on these challenging benchmark datasets, our study illustrates new opportunities for identification based on genome signatures and supports a much broader, more harmonized vision of barcoding as a general-purpose, rapid inferential tool for biodiversity and beyond.

Results and Discussion

A broader vision of DNA barcoding requires genomic signatures that can operate across broad organismal diversity, mixed samples, and different kinds of biological questions.

We organize results from our five assembled datasets around three forms of inference: identity, place, and state, while emphasizing both classification performance and the biological or sampling features that shaped accuracy. Sample composition, metadata labels, and coverage summaries are provided in **Supplementary Data 1**. To account for class imbalance among breeds, cultivars, geographic labels, and disease states, we report correct classification alongside macro-averaged precision and recall in the main text, with support-weighted summaries and full per-label performance summaries provided in **Supplementary Data 2**.

One varKode, many identities: breeds, cultivars, and individual microbial fingerprints

varKodes resolved identity across three increasingly fine and biologically distinct scales: recently formed lineages, cultivated varieties, and individual humans identified through host-associated microbiomes. In domestic dogs, varKodes predicted breed identity with 79.6% correct classification, 81.3% precision, and 75.3% recall (**Fig. 2Ai**), indicating that sparse, genome-wide signatures contain sufficient signal to distinguish breeds shaped by exceptionally recent selection and complex breeding histories. Breed assignment is a stringent test because many modern breeds are recent (<200 years old), closely related, and shaped by intensive artificial selection. Marker-based genomic approaches provide the most relevant comparison: Parker et al.[8], for example, assigned 99% of dogs to breed using microsatellite genotypes (**Table 1**). By contrast, varKoder recovered breed-level signal with approximately 80% correct classification using sparse genomic signatures, without predefined breed loci, reference-panel genotyping, alignment, or imputation. Performance varied among breeds, from high classification success in Yorkshire Terriers (94.5% correct; 97.4% precision; 95.6% recall) to lower performance in English Cocker Spaniels (20.0% correct; 68.0% precision; 37.8% recall). This variation was not explained simply by breed age or genome coverage (**Extended Data Table 1**), suggesting that classification difficulty more likely reflects biological similarity among breeds, complex ancestry, label identity, or reference-set composition. For poorly classified samples such as English Cocker Spaniel, the low recall may therefore reflect limited within-breed representation, close genomic similarity to other recently derived breeds, including possible admixtures, and varying quality in how breed labels were assigned or curated across public datasets, rather than insufficient sequencing depth alone. Misclassifications were also structured across breed labels, with some true breeds disproportionately incorrectly assigned to particular breeds. Such structured misclassifications may be informative because they can flag lineages whose current labels obscure biological complexity that may be worth further pursuing.

Grapevine cultivars provided a second stringent test of ultra-fine identity, because many cultivars have shallow divergence, complex pedigrees, and histories of synonymy and hybrid origin. Here, varKodes recover clear cultivar-level signal commensurate with work by others (**Table 1**). In particular, targeted SSR and SNP panels can uniquely identify cultivars under ideal conditions[9]. These earlier marker-based studies provide an important benchmark because they represent optimized diagnostic systems for cultivar identification. In this context, varKode's contribution is not to outperform targeted SNP or SSR panels, but to test whether cultivar-level signal can be recovered from sparse genome-skim data without predefined loci, alignment, or marker selection, effectively rendering our proposed approach as a lower barrier to entry. Across grape cultivars, varKodes achieved 82% correct classification, 90% precision, and 80% recall (**Fig. 2Bi–ii**). Performance varied among cultivars, with high classification success for Zinfandel (98.1% correct; 88.5% precision; 100% recall) and Thompson Seedless (89.5% correct; 85.6% precision; 92.0% recall), versus lower performance for Cabernet Sauvignon (50.8% correct; 83.3% precision; 71.4% recall) and Muscat Hamburg (36.1% correct; 87.5% precision; 38.9% recall). This heterogeneity is biologically expected in a crop shaped by recent hybridization, clonal propagation, and complex breeding structure[10]. For example, reduced performance for Muscat of Hamburg (87.5% precision, 38.9% recall), may be explained by its hybrid ancestry as a cross between Schiava Grossa and Muscat of Alexandria, as well as its placement within the broader Muscat group, which has long been recognized for complex naming and pedigree histories[11,12]. Cabernet Sauvignon likewise has a known hybrid origin as the offspring of Cabernet Franc × Sauvignon Blanc[13]. Thus, lower recall in some cultivars likely reflects biological complexity and sampling limitations, rather than failure of the varKode representation.

Skin microbiomes extended identity inference to a fine and conceptually distinct scale: individual-level discrimination from host-associated metagenomes. Unlike our dog and grape analyses, which classify samples using genomic variation in the focal organism, our skin analysis asked whether a composite microbial community could serve as a person-specific genomic signature. Using U.S. skin metagenomes collected from multiple body sites in Oh et al.[14], varKodes assigned skin samples to their correct host individual with approximately 75% correct classification, 77% precision, and 73% recall (**Fig. 3A–B**). Because each individual was represented by samples from distinct anatomical sites, successful classification indicates that varKodes recovered host-specific microbial signal despite strong body-site effects.

Existing benchmarks place this result in a useful context. A target capture skin-microbiome marker panel developed for forensic identification, hidSkinPlex, reported 78.0–83.7% individual classification accuracy[15], whereas Franzosa et al.[16] matched approximately 30% of individuals from skin microbiomes using person-specific shotgun

metagenomic codes (**Table 1**). Thus, varKodes recovered individual-level signal from untargeted, alignment-free shotgun data at a level approaching targeted forensic microbiome assays. Broader sampling across individuals, time points, and body sites will be required to further evaluate forensic-scale performance, but our results demonstrate that individual identity can be encoded in host-associated microbial mosaics[17].

This resolution also extended to even finer within-host scale discrimination: varKodes classified body-site category with 93% correct classification, 92% precision, and 91% recall (**Fig. 3C**). Accuracy varied by body site, with highest performance for foot (100% correct; 98% precision, 100% recall) and head samples (93% correct; 92% precision, 94% recall), versus lower performance for torso samples (70% correct; 79% precision, 72% recall; **Fig. 3D**), likely reflecting less distinct microbial signatures across torso sites than across foot or head sites. Together, these results demonstrate that varKodes can resolve identity across a continuum from breeds and cultivars to individuals and body sites. They also illustrate that categories of identity, place, and state are not mutually exclusive: host-associated microbial signatures can encode individual identity while also carrying information about anatomical location, geographic provenance, environmental exposure, and potentially health status.

One varKode, many places: metagenomic mosaics encode spatial provenance

Beyond identity, varKodes recovered geographic provenance from metagenomic mosaics, supporting the idea that mixed-community DNA can function as a location-specific genomic signature. Here, provenance refers to assignment to predefined geographic categories (i.e., continent, region, province, or state) rather than prediction of exact GPS coordinates. Thus, the global analysis evaluated continent-level assignment, whereas the regional analyses evaluated finer labels, including province-level labels in Asia, regional labels in North America (i.e., Southwest, North Central, etc), and U.S. state-level labels within North America. Under this framework, varKodes classified soil provenance with 81.9% correct classification globally (76.6% precision, 86.0% recall), 85.4% for province-level predictions within Asia (94.3% precision, 77.3% recall), 89.8% for regional predictions within North America (95.0% precision, 83.4% recall), and 94.4% for U.S. state-level origin (95.2% precision, 91.9% recall) (**Fig. 4A–D**). Higher accuracy at finer-scales should not be interpreted as an intrinsic effect of spatial scale; it more likely reflects differences in sampling density and label quality, especially the denser representation of North American and Asian sampling in the source dataset[18]. varKode's 82–94% correct classification falls within the range reported by other machine-learning geolocation frameworks (**Table 1**). The key contribution here is that comparable provenance signal was recovered directly from varKode images,

without first identifying the organisms present, estimating their relative abundances, or building custom reference-based features.

We next tested whether provenance predictions were driven primarily by broad habitat differences rather than geography. Habitat-only classifiers assigned Asian soil samples to land-use type with high overall performance (94.4% correct, 94.4% precision, 94.4% recall), and North American habitat-only classifiers also performed well (91.0% correct, 91.8% precision, 91.7% recall; **Extended Data Fig. 2**). Strikingly, habitat remained highly predictable when Asian and North American samples were analyzed together (93.1% correct, 94.3% precision, 93.1% recall; **Extended Data Fig. 2**), suggesting that land-use categories leave repeatable metagenomic signatures that can be detected across broad geographic regions. Geographic provenance also remained strong within habitat-stratified subsets. Across North American agricultural, grassland, and forest soils, varKodes achieved 100.0% correct classification, precision, and recall (**Extended Data Fig. 3**); across Asian agricultural, artificial-surface, and forest soils, varKodes also achieved 100.0% correct classification, precision, and recall (**Extended Data Fig. 3**). These results indicate that varKodes are not merely detecting coarse environmental contrasts such as forest versus grassland. Instead, these metagenomic mosaics retain discrete information about both habitat type *and* geographic origin.

Prediction accuracy depended on reference sampling density. Correctly classified North American soil samples were concentrated in densely sampled areas, whereas errors or lower accuracies were more common in sparsely sampled regions (**Extended Data Fig. 4**). Kernel density estimation (KDE) based sampling density differed significantly between correct and incorrect predictions (Wilcoxon $p < 0.001$), and logistic regression showed a positive relationship between local sampling density and the probability of correct classification ($\beta = 0.005$, $p < 0.001$). Thus, geo-nomic performance is not only a property of the classifier, but also of the reference samples used to train it (**Fig. 4E**). Dense, geographically representative sampling will be essential for extending varKode-based provenance inference to finer spatial scales and more heterogeneous regions.

Lastly, human-associated microbiomes showed parallel provenance patterns. Facial skin metagenomes from the U.S. and China were classified by country of origin with near-perfect accuracy (>99% correct, >99% precision, >99% recall; **Extended Data Fig. 5A**). Because these data were drawn from distinct sampling and sequencing efforts, country-level differences may reflect both biological geography and study-specific technical effects. Nevertheless, the convergence between soil and skin analyses suggests that metagenomic mosaics encode reproducible place-based signals across both environmental and host-associated contexts.

One varKode, many states: disease classification from stool metagenomes

Finally, we tested whether varKodes could classify biological state (here, disease versus non-diseased samples), using inflammatory bowel disease (IBD) as a proof-of-concept clinical application[19–21]. IBD encompasses chronic inflammatory disorders of the gastrointestinal tract, including Crohn's disease (CD) and ulcerative colitis (UC), which are often treated as related but distinct disease categories. IBD is associated with shifts in gut microbial community structure, and stool metagenomic profiles have previously been explored as a noninvasive biomarker of disease status and proxy for diagnosis[21–23].

Publicly-available stool metagenomes from the U.S. and Netherlands were first assigned to their country of origin with near-perfect performance (>99% correct, >99% precision, >99% recall; **Extended Data Fig. 5B**), indicating strong cohort- and spatially structured signals in gut microbial communities. Even when samples were pooled across countries and independently generated studies, varKodes distinguished healthy controls from IBD with approximately 82% correct classification (78% precision, 77% recall; **Fig. 5A–B**). It further differentiated Crohn's disease (CD) from ulcerative colitis (UC) with approximately 76% correct classification (76% precision, 76% recall; **Fig. 5C**), a clinically challenging differentiation to perform using standard diagnostic workflows[24]. Country-specific models produced similar results, indicating that pooling across did not materially reduce disease-classification performance (**Supplementary Data 2**). For healthy versus IBD classification, varKodes achieved an AUC of 0.90 (**Supplementary Data 2**), placing our approach within the range of recent microbiome-based IBD classifiers while requiring no targeted panel design or species-abundance profiling (**Table 1**).

These results should be interpreted as a proof of concept only, not as a substitute for proper clinical diagnosis, which depends on clinical features, laboratory markers, endoscopy, histology, and imaging. Because the U.S. and Netherlands cohorts were generated separately, country-level signals may also include unaccounted for technical variation (see **Methods**). Nevertheless, the ability to classify disease state from the same image-based genomic representation used for identity and provenance shows that the same varKode framework can be used to identify different kinds of biological information.

Conclusions

Our results demonstrate that sparse genomic data support inference well beyond conventional taxonomic assignment (**Fig. 1**). Across recently diverged breeds and

cultivars, spatially structured skin and soil metagenomes, and disease-associated stool microbiomes, varKodes recovered biologically meaningful signals using a single alignment-free framework inferred especially from low-coverage composite genomic signatures. The central implication is that DNA barcoding need not be defined by a small set of standardized loci[5]. Instead, barcodes can be broadly reconceived as retrainable genome-wide signatures that support supervised inference of identity, place, and state.

This shift is both conceptual and practical. Conceptually, varKodes expand barcoding from species identification to a more general framework for genomic discrimination. Practically, they point toward a transition from low-information marker workflows to shallow genomic sequencing coupled with computationally lightweight image generation, flexible reference libraries, and machine-learning classification. Across these expanded applications, varKode's practical advantage is that they can use shallow sequencing data without assembly, alignment, taxonomic profiling, or custom diagnostic panels. This reduces laboratory and computational overhead relative to workflows that depend on high-coverage genomes, predefined loci, or application-specific reference data. Such a transition could unify applications across biodiversity science, agriculture, conservation, food and medicine authentication, biosecurity, ecological monitoring, forensic geolocation, and noninvasive clinical screening. Because host-associated metagenomic profiles can encode individual-level information, future forensic or clinical applications should seek appropriate consent, privacy protection, and governance. Although individual-level data in this study were de-identifed, future forensic or clinical applications will require clear ethical guidance, appropriate data safeguards, and transparent consent processes.

At present, varKodes are best understood as reference-trained genomic signatures. Their predictions are strongest when labels are represented by dense, well-curated reference data; novel, undersampled, or biologically complex categories may yield ambiguous, inconclusive, or discordant predictions. This limitation also suggests a clear path forward: as reference datasets become larger, better curated, and more geographically representative, the accuracy and scope of varKode-based inference will improve. In this sense, varKodes are not simply an extension of DNA barcoding, but a blueprint for what barcoding can become in the era of shallow genomics and scalable machine learning.

Materials and Methods

Dataset selection and labeling

Samples from our five publicly available datasets were retained only when raw reads and well-defined metadata labels were available. Labels included breed, cultivar, host individual, body site, country, geographic region, habitat type, and diagnosis. Complete details for all datasets, including genomic coverage for each sample, are provided in **Supplementary Data 1**. Representative comparator studies, reported accuracies, and practical trade-offs for each dataset were summarized in **Table 1** to better place varKode performance in context.

The dog dataset comprised 245 genome samples from Parker et al.[25], here representing 17 breeds: Tibetan Mastiff, Portuguese Water Dog, Yorkshire Terrier, Norwich Terrier, Soft-Coated Wheaten Terrier, Border Terrier, German Shepherd Dog, Golden Retriever, English Cocker Spaniel, Belgian Tervuren, Belgian Sheepdog, Border Collie, Bearded Collie, Rottweiler, Bernese Mountain Dog, and Bull Terrier, and Grey Wolf as the outgroup. These breeds were selected to span the dog-breed phylogeny inferred by Parker et al.[25] and to include both recently established breeds and older lineages. The grape dataset comprised 84 genome-skim samples representing nine *Vitis vinifera* cultivars: Zinfandel, Thompson Seedless, Riesling, Red Globe, Chardonnay, Pinot Noir, Italia, Cabernet Sauvignon, and Muscat of Hamburg. Cultivar sampling was guided by the phylogenetic framework of Zhou et al.[26] and included both historically established cultivars, such as Riesling and Pinot Noir, and more recent derivatives or breeding lines, including Cabernet Sauvignon, Muscat of Hamburg, Italia, and Red Globe. For both the dog and grape datasets, we retained only labels represented by four or more samples, allowing breed- and cultivar-level models to be evaluated with minimal within-label replication.

The soil metagenome dataset included a subset of 2,111 samples from Ma et al.[18]. We initially considered the global dataset and then focused subsequent analyses on North America and Asia because these regions had the highest sampling density. Soil samples were labeled at multiple spatial scales. Geographic provenance was treated as a categorical classification task rather than as prediction of exact latitude and longitude. Accuracy therefore reflects assignment to the correct geographic label used in a given analysis, (e.g., continent, province, North American region, or U.S. state). North American samples were assigned to U.S. states and broader geographic regions, whereas Asian samples were assigned to regional or province-level labels, depending on the analysis. Samples were also labeled by habitat type using eight land-use categories: agricultural land, artificial surfaces, bare land, forest, grassland, shrubland, tundra, and wetland. For habitat-stratified geographic analyses, we retained the best-represented habitat classes within each continent and trained geographic classifiers separately within those subsets. Broad geographic-region labels were assigned using

the World Geographical Scheme for Recording Plant Distributions[27], whereas U.S. state and province labels were taken directly from sample metadata.

The human skin microbiome dataset comprised shotgun metagenomic samples from Oh et al.[14] and Li et al.[28]. Oh et al.[14] sampled multiple anatomical skin sites from individuals in the U.S., enabling tests of whether a held-out skin sample could be assigned to the correct host individual despite being collected from a distinct body location. Our host-identity analysis used 352 Oh et al. samples from 12 individuals. For the body-site analysis, anatomical sites from Oh et al.[14] were collapsed into three broader categories based on their sample metadata; these assignments are provided in **Supplementary Data 1**. This body-site dataset included 226 samples. Li et al.[28] was used only for the country-of-origin comparison because that dataset consisted of facial skin microbiomes from. To make the country-level comparison as comparable as possible, we compared facial skin metagenomes from Li et al. (China; n = 30) with facial skin metagenomes from Oh et al. (U.S.; n = 30). Thus, our host-identity and body-site analyses used the Oh et al. dataset, whereas the country-level skin analysis used facial samples from both Oh et al. and Li et al.

The stool microbiome dataset comprised 218 shotgun metagenomes from Franzosa et al.[19], including healthy controls and inflammatory bowel disease samples from cohorts in the U.S. and the Netherlands. These samples were used to evaluate both country-of-origin classification and disease-state classification, including healthy versus inflammatory bowel disease and Crohn's disease versus ulcerative colitis.

Sample sizes per label were not standardized; instead, all publicly available samples with usable reads and relevant metadata were retained. This preserves the natural imbalance typical of public genomic repositories but may contribute to uneven prediction performance across labels. For this reason, we report both macro-averaged metrics, which emphasize label-balanced performance, and support-weighted metrics, which reflect performance under the observed sample distribution. Because some datasets were assembled from separate studies, certain comparisons, particularly country-of-origin analyses in skin and stool microbiomes, may also reflect study-specific technical variation. This study reanalyzed previously published, publicly available human-associated metagenomic datasets and generated no new human-participant data **(Supplementary Data 1)**.

varKode generation and model training

Raw sequencing reads were downloaded for all samples and processed with modest quality filtering prior to varKode generation. To accommodate variation in sequencing depth across samples, reads were randomly subsampled to fixed total base-pair allotments before image generation. Image representations were produced using the

ranked-frequency chaos game representation (rfCGR) implemented in varKoder v1.5.0[5]. For each sample, varKodes were generated from randomly subsampled read sets spanning 500 kbp to the full available sequence length.

Models were trained using the default varKoder pipeline in two stages. First, single-label classifiers were trained for 20 epochs using the most specific available label for each sample, with a base learning rate of 0.05. Second, pretrained weights were used to initialize multi-label models trained for an additional 10 epochs with an asymmetric loss function and a reduced learning rate of 0.005. All analyses used the default vision transformer architecture pretrained on the all-SRA dataset of Asprino et al.[29], together with lighting transforms and MixUp augmentation.

Validation schemes and classification tasks

For the dog, grape, skin, and stool datasets, we evaluated performance using leave-one-out cross-validation, in which each sample was held out in turn and predicted from a model trained on all remaining samples. This design was chosen because these datasets were relatively small and label-rich. For the soil metagenome dataset, which contained more than 2,000 samples, we used a random 80/20 train-test split to reduce computational burden while preserving geographic representation.

Separate classification models were trained for each prediction task using the relevant labels and sample subsets. For the genome-derived datasets, models predicted dog breed identity or grape cultivar identity. For skin microbiomes, models predicted host individual identity, body-site category, and country of origin. The host-identity analysis should be interpreted as assigning a held-out body-site sample to an individual already represented by other skin samples in the reference set, rather than identifying a completely unsampled individual. For stool microbiomes, models predicted country of origin and disease status, including healthy versus inflammatory bowel disease and Crohn’s disease versus ulcerative colitis. For stool disease classification, we evaluated both pooled models across the U.S. and Netherlands cohorts and country-specific models trained within each cohort. For soil metagenomes, separate models predicted geographic provenance at multiple spatial scales, habitat type alone, and geographic labels within habitat-stratified subsets.

Prediction scoring and performance summaries

Predictions from multi-label models were evaluated using a fixed confidence threshold of 0.70, following de Medeiros et al.[5]. A prediction was scored as correct when only the true label exceeded the threshold, incorrect when one or more labels exceeded the threshold but none matched the true label, ambiguous when both true and false labels exceeded the threshold, and inconclusive when no label exceeded the threshold. This

scoring framework enabled consistent comparison across label types, including taxonomic and individual identity, geographic provenance, habitat type, and disease state. Unless otherwise stated, accuracies reported in the Results and Discussion reflect performance at the largest number of base pairs available per sample.

To address class imbalance and provide standard classifier-performance summaries, we calculated correct classification, precision, recall, F1, and area under the curve (AUC) for each classification task. Correct classification was defined as the proportion of samples scored as correct under the threshold-based prediction framework described above. Precision measures the proportion of predicted positives that were correct, recall measures the proportion of true positives that were recovered, F1 is the harmonic mean of precision and recall, and AUC summarizes how well the model separates samples that belong to a given label from samples that do not, based on raw prediction scores. For multiclass analyses, these metrics were first calculated separately for each label in a one-versus-rest framework. For a given label, true positives were samples for which that label was both present in the true label set and predicted label set; false positives were samples for which that label was predicted but not present in the true label set; and false negatives were samples for which that label was present in the true label set but not predicted. Per-label precision was calculated as TP/(TP + FP), per-label recall as TP/(TP + FN), and per-label F1 as 2 × precision × recall / (precision + recall). Label support was defined as the number of true samples for that label, equivalent to TP + FN.

We report label-averaged, or macro-averaged, precision and recall in the main text because these summaries were calculated as the unweighted mean of the per-label values, giving each label equal influence regardless of sample size. We also calculated support-weighted label averages, in which each per-label metric was weighted by that label's support before averaging across labels. These support-weighted summaries therefore reflect performance under the observed sample distribution, with more heavily sampled labels contributing more to the final value. AUC was calculated separately for each label using one-versus-rest binary truth values and the corresponding raw prediction score, then summarized using the same macro-averaged and support-weighted approaches. Full per-label metrics, macro-averaged summaries, support-weighted summaries, F1 values, AUC values, and exact-match accuracies are reported in **Supplementary Data 2**. Exact-match accuracy was recorded as the proportion of samples for which the normalized predicted label set exactly matched the normalized true label set. All downstream analyses and visualizations were performed in R.

Dog breed-age and genome-coverage analyses

Estimated dog-breed origination dates were compiled from published genetic studies and breed-history references[8,30–34]. For each breed, we recorded the approximate date of breed formation, historical recognition, or earliest well-supported breed description reported in these sources. When sources differed or provided approximate historical ranges, we used a conservative consensus estimate based on agreement across references. These dates were used both to visualize breed origins on the dated dog-breed cladogram in **Fig. 2Aii** and to test whether breed age explained variation in varKode prediction accuracy. Breed age was expressed in centuries and modeled as a continuous predictor using binomial logistic regression. Wolves were excluded from this analysis because they represent an outgroup rather than a domesticated dog breed. We analyzed both decisive predictions only, scored as correct versus incorrect, and a conservative formulation in which ambiguous and inconclusive predictions were treated as failures. Rank-based associations between breed age and prediction accuracy were additionally assessed using Spearman correlation.

To test whether prediction accuracy was explained by sequencing depth, we joined dog prediction results to estimated genome coverage values reported in **Supplementary Data 1**. Estimated coverage for dog samples was calculated relative to the CanFam3.1 reference genome (RefSeq accession GCF_000002285.3). For comparison across genome-derived datasets, estimated coverage for grape samples was calculated relative to the *Vitis vinifera* reference genome (GenBank accession GCA_030704535.1). Correct breed assignment at the largest query size was modeled as a binary response using binomial logistic regression, with estimated genome coverage as a continuous predictor. Predictions categorized as ambiguous, inconclusive, or incorrect were treated as not correct. We also summarized estimated genome coverage across prediction-outcome categories to evaluate whether failed predictions were concentrated among lower-coverage samples.

Sampling density and spatial accuracy in soil metagenomes

To test whether geographic prediction accuracy was associated with local sampling density, we estimated KDE from sample coordinates using *kde2d* in MASS on a 300 × 300 grid. KDE values were interpolated at each sample location and used as estimates of local reference density.

We compared KDE values between correctly and incorrectly classified samples using a Wilcoxon rank-sum test and modeled prediction correctness as a function of local density using binomial logistic regression. For visualization, KDE values were binned into 50 equal-width intervals and plotted against the proportion of correct classifications.

A North American map was generated with KDE density shown as a raster background and prediction outcomes overlaid as points.

Author contributions: P.J.F. and C.C.D. conceived and designed the study. P.J.F. performed the analyses, with analytical and data-processing contributions from J.K. V.C. and L.G. provided early input, planning guidance, and editorial feedback on the manuscript. Results were interpreted by P.J.F., J.K., B.A.S.d.M., and C.C.D. P.J.F. and C.C.D. wrote the initial manuscript draft. All authors contributed to manuscript revision and approved the final version.

Acknowledgements: P.J.F. was supported by an NSF Postdoctoral Research Fellowship in Biology and by LVMH Research and Dior Science. C.C.D. was supported by Harvard University, LVMH Research, and Dior Science. J.K. was supported by Harvard University with funds to C.C.D. B.A.S.d.M. was supported by the Walder Foundation. Computational analyses were performed on the Harvard FASRC Cannon cluster. We thank members of the Davis Lab and Sound Solutions for Sustainable Science for constructive feedback and careful editing of earlier versions of the manuscript.

References

1. Gostel, M. R. & Kress, W. J. The expanding role of DNA barcodes: indispensable tools for ecology, evolution, and conservation. Diversity 14, 213 (2022).
2. Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B Biol. Sci. 270, 313-321 (2003).
3. Hollingsworth, P. M., Graham, S. W. & Little, D. P. Choosing and using a plant DNA barcode. PLoS ONE 6, e19254 (2011).
4. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. Proc. Natl Acad. Sci. USA 102, 8369-8374 (2005).
5. de Medeiros, B. A. S. et al. A composite universal DNA signature for the tree of life. Nat. Ecol. Evol. 9, 1426-1440 (2025).
6. Huang, W. et al. DNA-based identification of plants and the genomic nature of plant species differences. Commun. Biol. 9, 673 (2026).
7. Bleidorn, C., Sandberg, F., Martin, S., Vogler, A. P. & Podsiadlowski, L. The untapped potential of short-read sequencing in biodiversity research. Trends Genet. 42, 137-149 (2026).
8. Parker, H. G. et al. Genetic structure of the purebred domestic dog. Science 304, 1160-1164 (2004).

9. Laucou, V. et al. Extended diversity analysis of cultivated grapevine Vitis vinifera with 10K genome-wide SNPs. PLoS ONE 13, e0192540 (2018).
10. Myles, S. et al. Genetic structure and domestication history of the grape. Proc. Natl Acad. Sci. USA 108, 3530-3535 (2011).
11. Crespan, M. The parentage of Muscat of Hamburg. Vitis 42, 193-197 (2003).
12. Crespan, M. & Milani, N. The Muscats: a molecular analysis of synonyms, homonyms and genetic relationships within a large family of grapevine cultivars. Vitis 40, 23-30 (2001).
13. Bowers, J. E. & Meredith, C. P. The parentage of a classic wine grape, Cabernet Sauvignon. Nat. Genet. 16, 84-87 (1997).
14. Oh, J. et al. Biogeography and individuality shape function in the human skin metagenome. Nature 514, 59-64 (2014).
15. Woerner, A. E. et al. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci. Int. Genet. 38, 130-139 (2019).
16. Franzosa, E. A. et al. Identifying personal microbiomes using metagenomic codes. Proc. Natl Acad. Sci. USA 112, E2930-E2938 (2015).
17. Fierer, N. et al. Forensic identification using skin bacterial communities. Proc. Natl Acad. Sci. USA 107, 6477-6481 (2010).
18. Ma, B. et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nat. Commun. 14, 7318 (2023).
19. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. 4, 293-305 (2019).
20. Vich Vila, A. et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci. Transl. Med. 10, eaap8914 (2018).
21. Zheng, J. et al. Noninvasive, microbiome-based diagnosis of inflammatory bowel disease. Nat. Med. 30, 3555-3567 (2024).
22. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 13, R79 (2012).
23. Pittayanon, R. et al. Differences in gut microbiota in patients with vs without inflammatory bowel diseases: a systematic review. Gastroenterology 158, 930-946.e1 (2020).
24. Kucharzik, T. et al. ECCO-ESGAR-ESP-IBUS guideline on diagnostics and monitoring of patients with inflammatory bowel disease: Part 1: initial diagnosis, monitoring of known inflammatory bowel disease, detection of complications. J. Crohns Colitis 19, jjaf106 (2025).
25. Parker, H. G. et al. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. Cell Rep. 19, 697-708 (2017).

26. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc. Natl Acad. Sci. USA 114, 11715-11720 (2017).
27. Brummitt, R. K. & Pando, F. World Geographical Scheme for Recording Plant Distributions (Hunt Institute for Botanical Documentation, Carnegie Mellon Univ., 2001).
28. Li, Z. et al. Characterization of the human skin resistome and identification of two microbiota cutotypes. Microbiome 9, 47 (2021).
29. Asprino, R. C. et al. A curated benchmark dataset for molecular identification based on genome skimming. Sci. Data 12, 906 (2025).
30. vonHoldt, B. M. et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464, 898-902 (2010).
31. Morris, D. Dogs: The Ultimate Dictionary of Over 1,000 Dog Breeds (Trafalgar Square, 2002).
32. American Kennel Club. The Complete Dog Book 20th edn (Ballantine Books, 2006).
33. Wilcox, B. & Walkowicz, C. Atlas of Dog Breeds of the World 4th edn (TFH Publications, 1993).
34. Clark, A. R. & Brace, A. H. (eds) The International Encyclopedia of Dogs (Howell Book House, 1995).
35. Cabezas, J. A. et al. A 48 SNP set for grapevine cultivar identification. BMC Plant Biol. 11, 153 (2011).
36. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. FEMS Microbiol. Rev. 35, 343-359 (2011).
37. Zhang, Y., McCarthy, L., Ruff, E. & Elhaik, E. Microbiome geographic population structure (mGPS) detects fine-scale geography. Genome Biol. Evol. 16, evae209 (2024).
38. Bhattacharya, C. et al. Supervised machine learning enables geospatial microbial provenance. Genes 13, 1914 (2022).
39. Su, Q. et al. Faecal microbiome-based machine learning for multi-class disease diagnosis. Nat. Commun. 13, 6818 (2022).
40. Larson, G. et al. Rethinking dog domestication by integrating genetics, archaeology, and biogeography. Proc. Natl Acad. Sci. USA 109, 8878-8883 (2012).
41. Noraz, R. et al. Ancient DNA reveals 4000 years of grapevine diversity, viticulture and clonal propagation in France. Nat. Commun. 17, 2494 (2026).
42. vonHoldt, B. M. & Driscoll, C. A. Origins of the dog: genetic insights into dog domestication. In The Domestic Dog: Its Evolution, Behavior and Interactions with People 2nd edn (ed. Serpell, J.) 22-41 (Cambridge Univ. Press, 2016).

697 **Tables**

Dataset	Reference	No. samples	Benchmark task	Method	Reported benchmark	How varKoder compares
Dog	Parker et al.[8]	414 dogs; 85 breeds	Breed identification	Microsatellite genotyping	99% correct breed assignment	Lower: 79.6%, but marker-free and alignment-free
Grape	Cabezas et al.[35]	1,200 accessions	Cultivar identification	48-SNP marker panel	100% accession discrimination	Lower: 82%, but no diagnostic SNP panel
Grape	Laucou et al.[9]	783 unique genotypes	Cultivar identification	Vitis18K SNP array	100% identification using 14 SNPs	Lower: 82%, but no selected SNP markers
Skin-individual	Woerner et al.[15]	51 individuals	Individual identification	Targeted skin microbiome markers	78.0–83.7% individual identification	Similar: ~75%, without targeted capture
Skin-individual	Franzosa et al.[16]	120 HMP subjects	Same-person skin matching	Shotgun metagenomic codes	~30% skin matching over 30–300 days	Higher: ~75%; validation differs
Skin-body site	Knights et al.[36]	268 train; 133 test	Skin-site classification	Random Forest on 16S OTUs	66% accuracy across 12 skin sites	Higher: 93%; broader site categories
Soil	Zhang et al.[37]	237 soil samples	Soil geolocation	XGBoost on microbial abundances	86% soil-site accuracy	Similar: 81.9–94.4%, without abundance tables
Soil	Bhattacharya et al.[38]	4,305 metagenomes	Microbial geolocation	ML on taxonomic profiles	85–89% city; 90–94% continental	Similar: 81.9–94.4%, without

						taxonomic profiling
Stool	Zheng et al.[21]	5,979 fecal metagenomes	IBD classification	18-species m-ddPCR panel	AUC 0.91 discovery; 0.81 validation	Similar: AUC 0.90, without targeted panel design
Stool	Su et al.[39]	2,320 train/test; 1,597 validation	Multi-disease stool classification	Random Forest on species profiles	AUROC 0.90–0.99 test; 0.69–0.91 validation	Similar: AUC 0.90, without species profiling

Table 1. Benchmark studies used to contextualize varKoder performance across datasets.

Representative studies that report broadly comparable identification, skin-site classification, geolocation, or biological-state classification results. Because datasets, methods, and performance metrics vary across studies, these comparisons are intended as context rather than direct one-to-one benchmarks. The final column summarizes whether varKode performance was lower than, similar to, or higher than the reported benchmark range, while highlighting key methodological trade-offs where relevant.

Figures

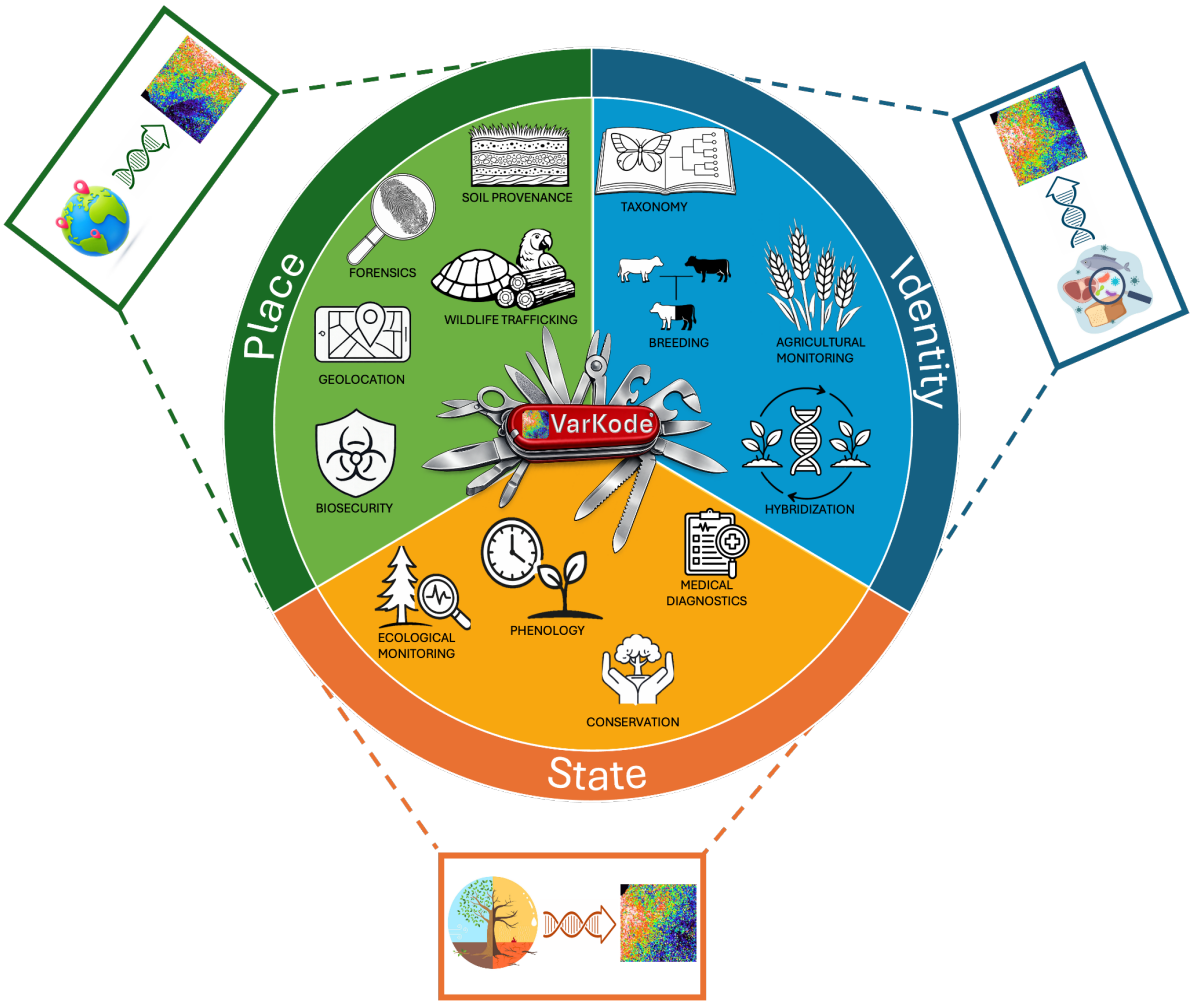


Figure 1. varKodes as an expansion of traditional DNA barcoding. Sequence data from diverse organisms and sample types are transformed into image-based genomic signatures (varKodes), as illustrated in the peripheral insets. The central Swiss army knife represents the flexibility of varKode for multiple forms of inference. These applications fall into three broad categories: Identity (e.g., taxonomy, breeding, agricultural monitoring, and hybridization), Place (e.g., soil provenance, forensics, wildlife trafficking, geolocation, and biosecurity), and State (e.g., ecological monitoring, phenology, conservation, and medical diagnostics). The figure highlights how varKode expands barcoding beyond conventional taxonomic assignment to support fine-scale genomic inference across diverse use cases.

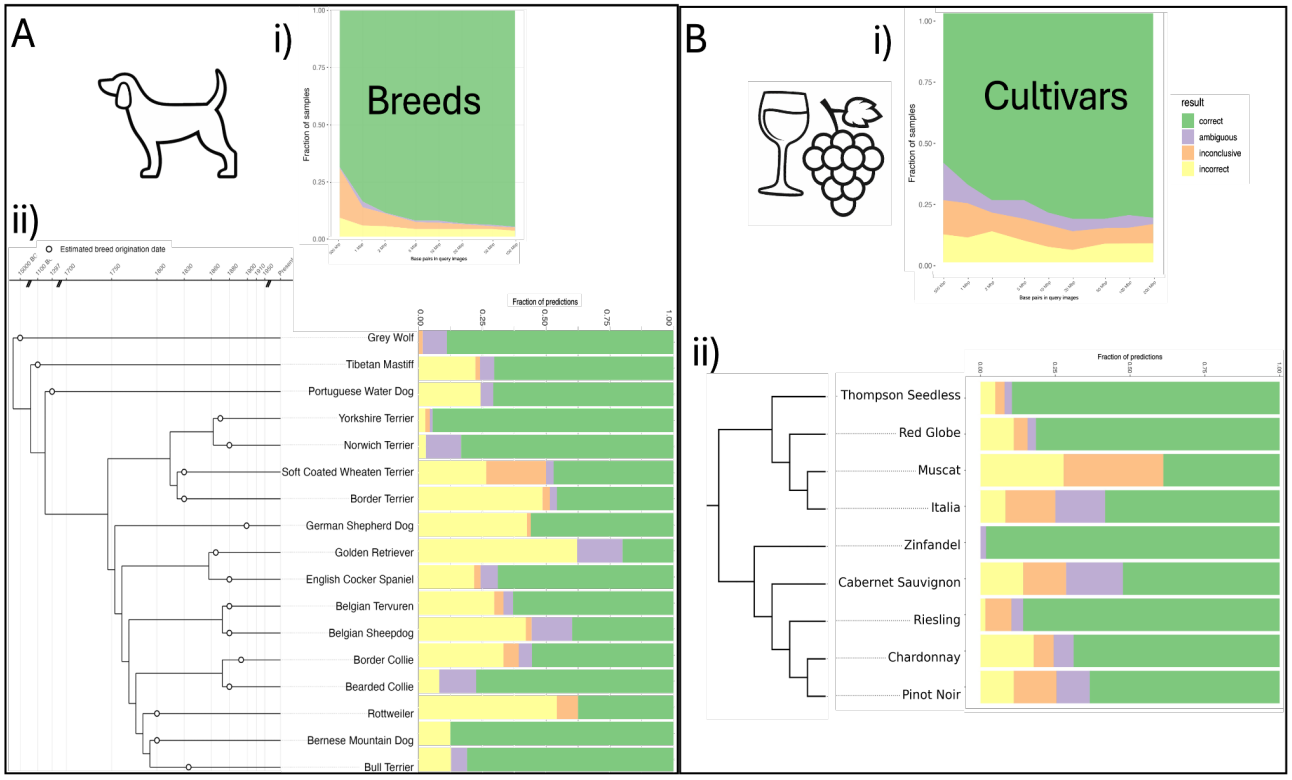


Figure 2. Prediction accuracy of dog and grape genomic identity across classification levels. (**A**) Dog classifications: (**Ai**) prediction outcomes by breed across increasing query image size; and (**Aii**) breed-level prediction outcomes displayed alongside a dated dog-breed cladogram modified from Parker et al.[25]. In **Aii**, terminal branches extend to the present, and open circles indicate estimated breed origination dates on a broken time scale, with the 1700–1910 interval expanded to show recent breed formation. (**B**) Grape classifications: (**Bi**) prediction outcomes by cultivar across increasing query image size; and (**Bii**) cultivar-level prediction outcomes displayed alongside a varietal cladogram modified from Zhou et al.[26]. Percentages in **Ai** and **Bi** indicate the proportion of correct predictions at the largest query size. Predictions are categorized as correct (green), ambiguous (purple), inconclusive (orange), or incorrect (yellow).

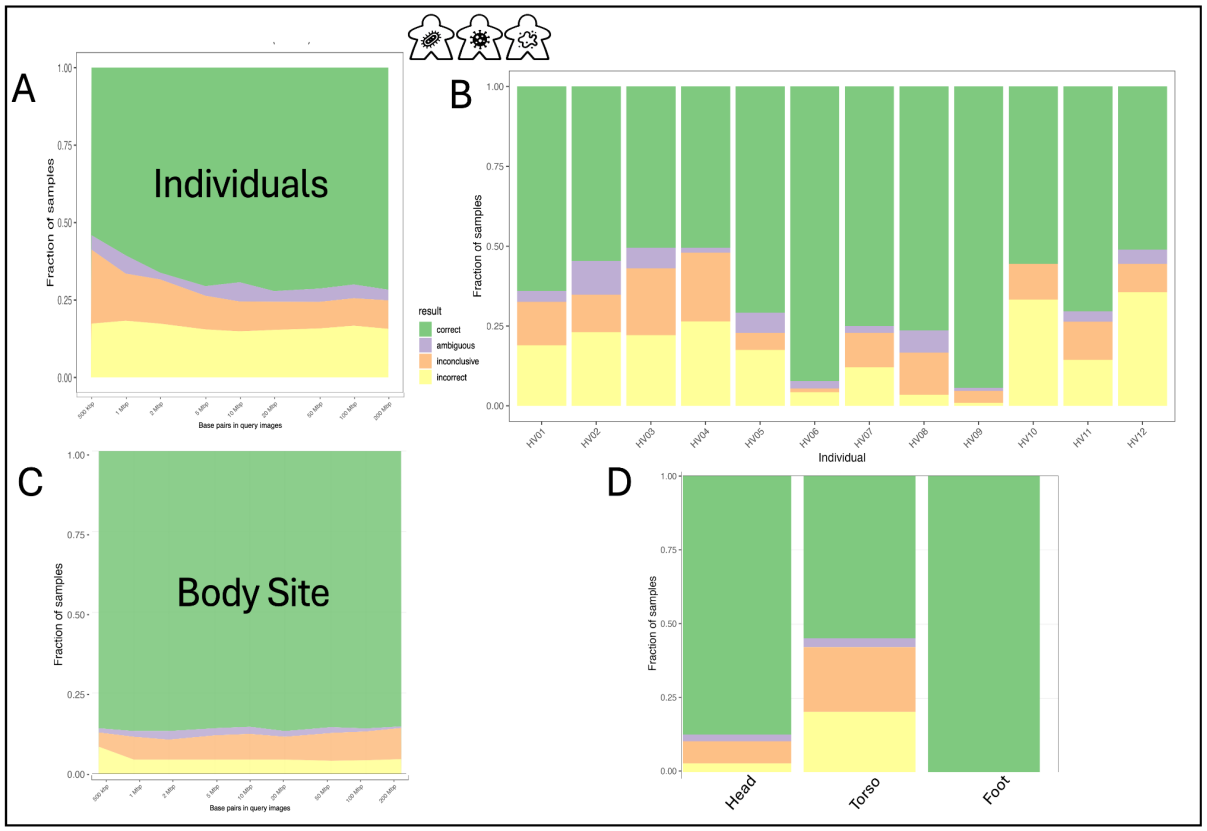


Figure 3. Prediction accuracy of human skin microbiomes for individual host identity and body site. (A) Continuous stacked area plot showing prediction outcomes for skin microbiome samples from Oh et al.[14] classified by host individual across increasing query image size. **(B)** Prediction outcomes summarized for each individual host. **(C)** Continuous stacked area plot showing prediction outcomes for the same skin microbiome samples classified by body site across increasing query image size. **(D)** Prediction outcomes summarized for each body site. In all panels, predictions are categorized as correct (green), ambiguous (purple), inconclusive (orange), or incorrect (yellow).

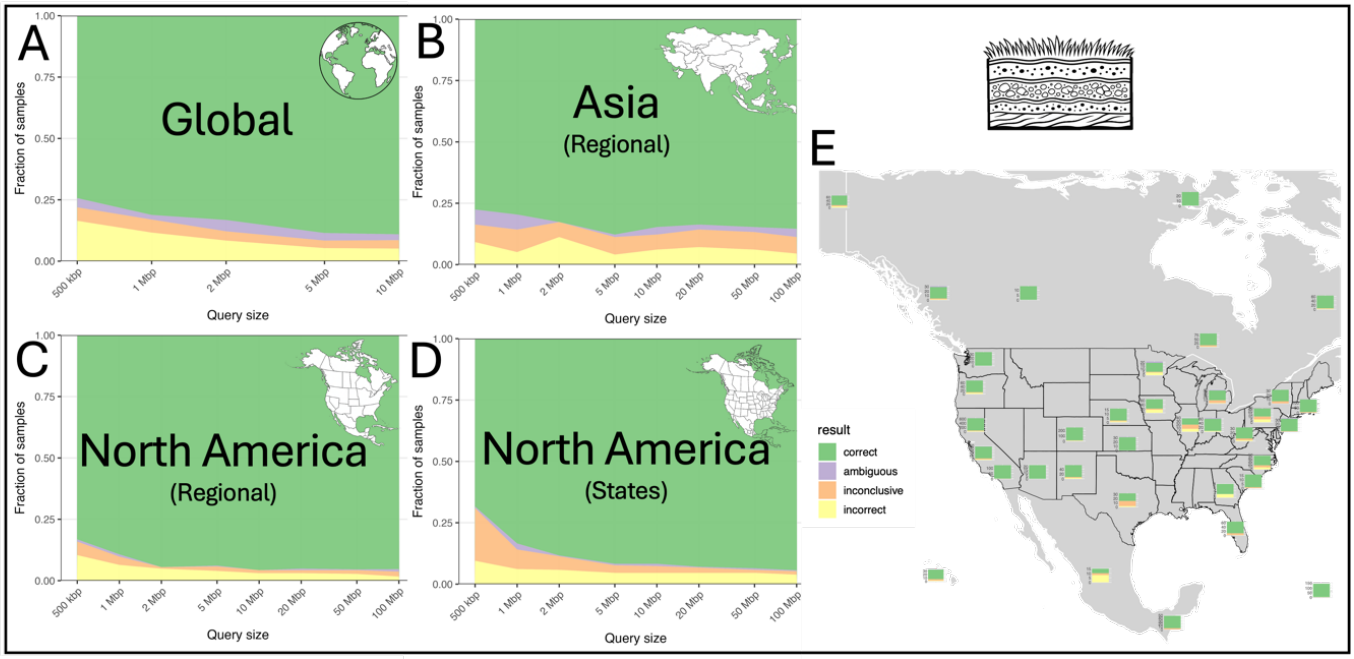


Figure 4. Prediction accuracy of soil metagenome geographic origin across spatial scales. Continuous stacked bar plots **(A–D)** illustrate the fraction of samples predicted at four geographic levels: **(A)** global, **(B)** regional – Asia, **(C)** regional – North America, and **(D)** state-level within North America. Samples are ordered from left to right by increasing total base pairs of query images. Percent accuracy displayed in each panel reflects the proportion of correct predictions at the largest query size for that sample. **(E)** Map of North America displaying prediction outcomes by state. In all panels, predictions are categorized as correct, ambiguous, inconclusive, or incorrect, and consistently color-coded: green (correct), purple (ambiguous), orange (inconclusive), and yellow (incorrect). **A** is modified from de Medeiros et al.[5] to include comparable global-level data and reflects the largest query size available from that analysis, 10 Mbp; panels B–D extend to 100 Mbp.

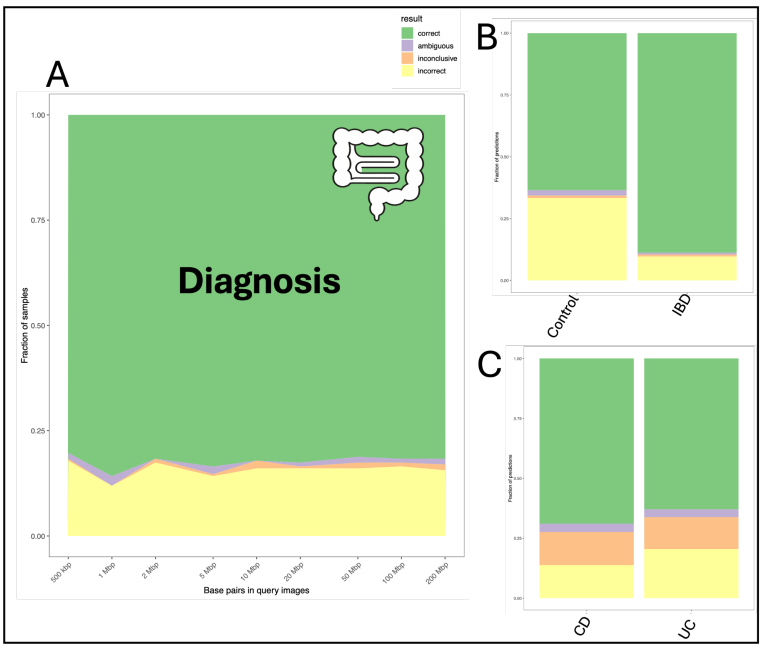
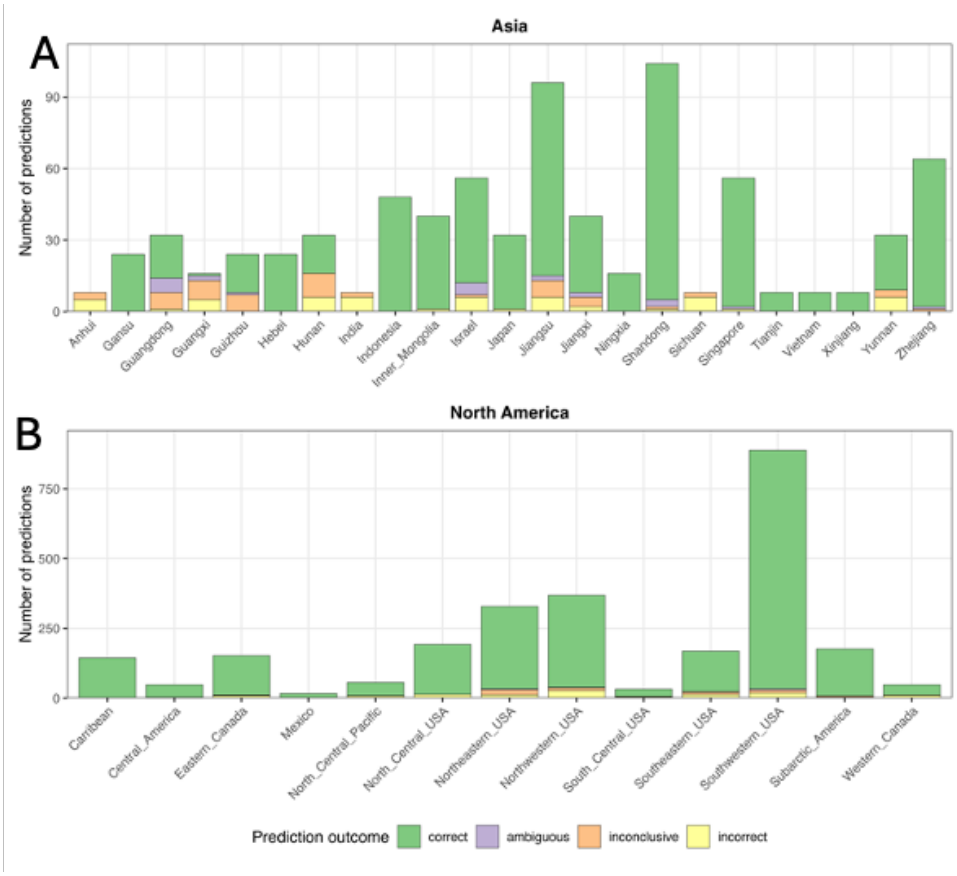


Figure 5. Case study of stool metagenome classification for inflammatory bowel disease. (A) Continuous stacked area plot showing prediction outcomes for stool metagenome samples from Franzosa et al.[19] classified by diagnosis across increasing query image size. The percentage shown indicates the proportion of correct predictions at the largest query size. **(B)** Prediction outcomes summarized for control versus inflammatory bowel disease (IBD) samples. **(C)** Prediction outcomes summarized for ulcerative colitis (UC) versus Crohn's disease (CD). In all panels, predictions are categorized as correct (green), ambiguous (purple), inconclusive (orange), or incorrect (yellow).

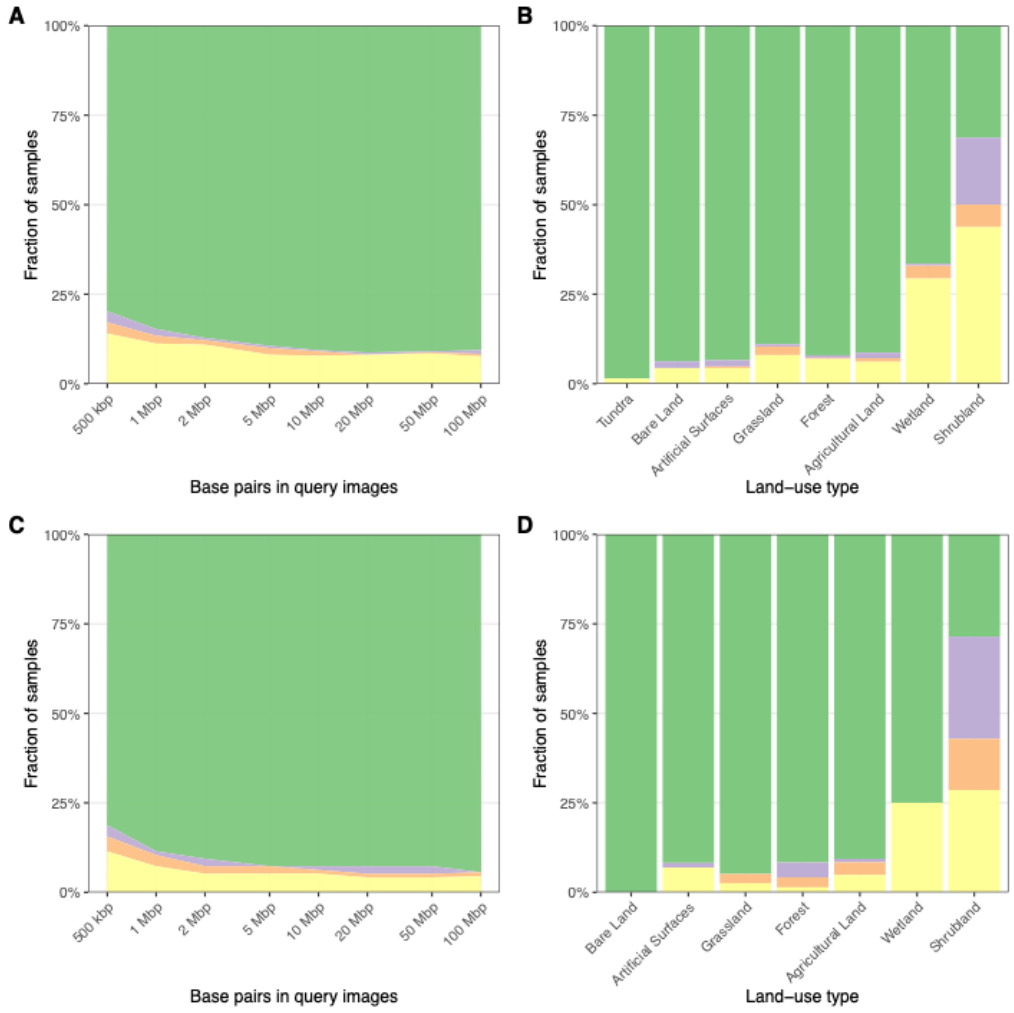
Extended Data

Analysis	Test	Predictor or statistic	Estimate	P-value	Interpretation
Breed age	Binomial logistic regression, decisive predictions only	Age effect	β = 0.217; OR = 1.24; 95% CI = 0.86–1.80	0.246	No clear age effect
Breed age	Binomial logistic regression, ambiguous/inconclusive as failures	Age effect	β = 0.270	0.111	Weak, unsupported age trend
Breed age	Spearman rank correlation	ρ	0.037	0.891	No age association
Genome coverage	Binomial logistic regression, 200-Mbp query size	Estimated coverage effect	β = −0.0238; OR = 0.98; 95% CI = 0.95–1.00	0.091	No positive coverage effect detected

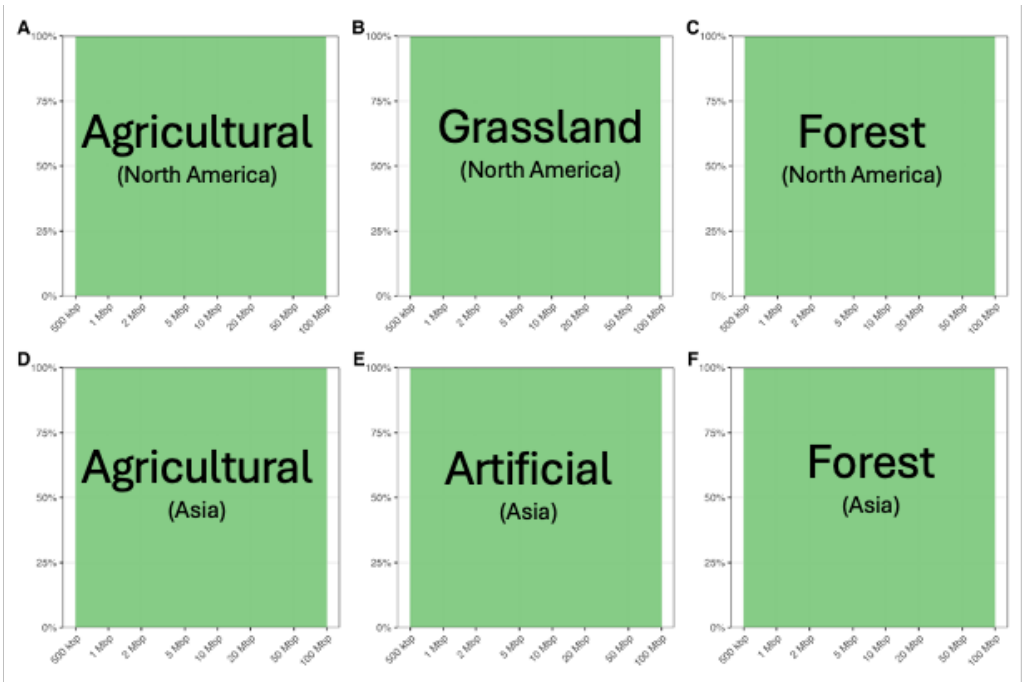
Extended Data Table 1. Summary of breed-age and genome-coverage analyses for dog breed prediction accuracy. Breed-age analyses were restricted to dogs only, with wolves excluded, and included binomial logistic regression models and a Spearman rank correlation. Genome coverage was tested using estimated sequencing coverage from **Supplementary Data 1** as a predictor of correct breed assignment at the largest query size. Across analyses, neither historical breed age nor estimated genome coverage significantly explained variation in varKode classification accuracy.



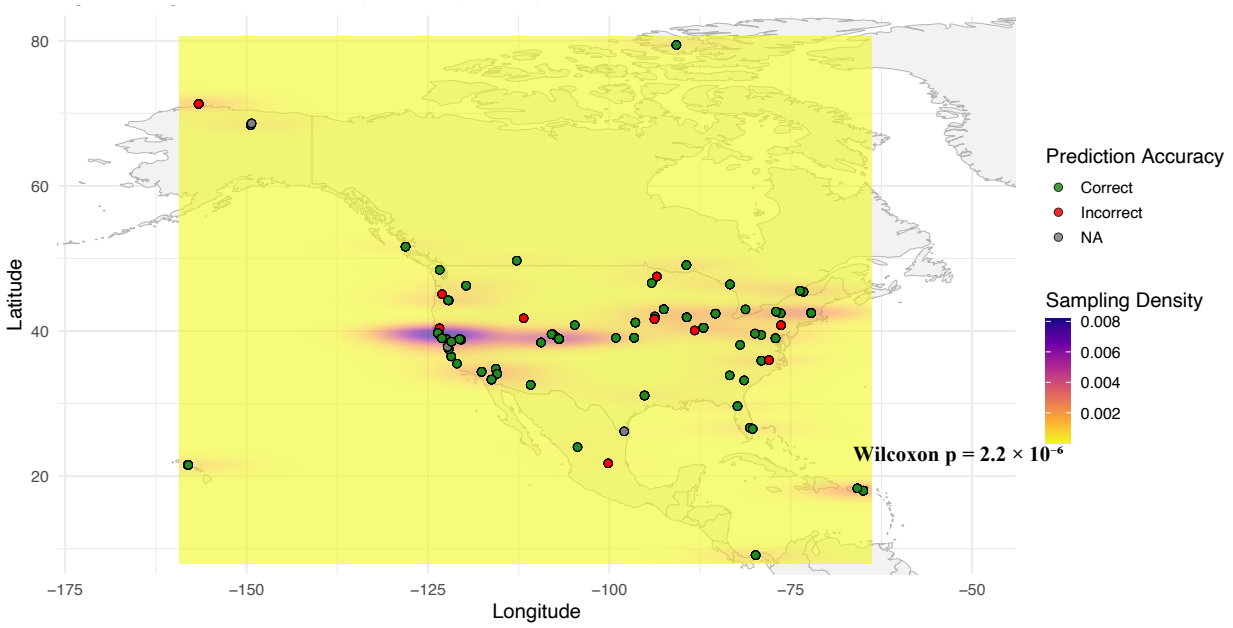
Extended Data Fig. 1. Regional prediction outcomes for soil metagenomes from Asia and North America. (A) Stacked bar plot showing prediction outcomes for Asian provinces, with bars ordered alphabetically. (**B)** Stacked bar plot showing prediction outcomes for North American regions, with bars ordered alphabetically. Counts represent predictions aggregated across all query sizes. Predictions are categorized as correct, ambiguous, inconclusive, or incorrect. Colors indicate prediction outcome: green, correct; purple, ambiguous; orange, inconclusive; yellow, incorrect.



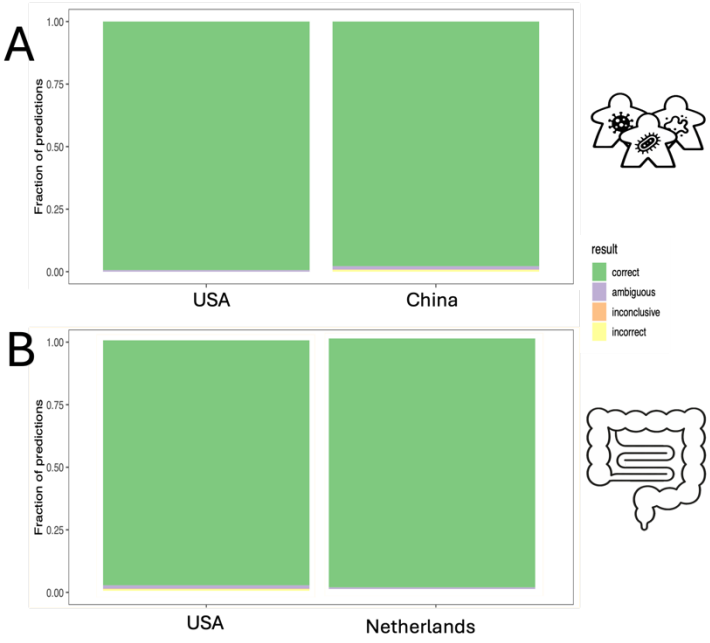
Extended Data Fig. 2. Prediction accuracy for soil metagenome samples from North America and Asia, labeled by land-use type. (A) Overall prediction accuracy across all North American samples. Samples are ordered from left to right by increasing total base pairs of query images. **(B)** Prediction accuracy for North American samples separated by land-use type. **(C)** Overall prediction accuracy across all Asian samples. Samples are ordered from left to right by increasing total base pairs of query images. **(D)** Prediction accuracy for Asian samples separated by land-use type. Predictions are categorized as correct, ambiguous, inconclusive, or incorrect. Colors indicate prediction outcome: green, correct; purple, ambiguous; orange, inconclusive; yellow, incorrect.



Extended Data Fig. 3. Prediction accuracy for a subset of soil metagenome samples, grouped by land type and continent. Continuous stacked bar plots show prediction outcomes for individual samples from: **(A)** Agricultural – North America, **(B)** Grassland – North America, **(C)** Forest – North America, **(D)** Agricultural – Asia, **(E)** Artificial – Asia, and **(F)** Forest – Asia. In all plots, samples are ordered from left to right by increasing total base pairs of query images. Predictions are categorized as correct (green), ambiguous (purple), inconclusive (orange), or incorrect (yellow). Percent accuracy displayed in each panel reflects the proportion of correct predictions at the largest query size for that sample. North American samples are labeled at the U.S. state level, while Asian samples use regional labeling.



Extended Data Fig. 4. Sampling density and prediction accuracy for state-level soil metagenomes across North America. Kernel density estimation heatmap shows sampling density, with background color ranging from yellow (low density) to dark purple (high density). Prediction outcomes are overlaid: green dots indicate correct predictions, red dots indicate incorrect predictions, and grey dots indicate samples with no prediction available (NA). Wilcoxon p = 2.2 × 10^{-6}.



Extended Data Fig. 5. Geographic classification of host-associated metagenomes. **(A)** Prediction outcomes for human skin metagenome samples classified by country of origin, comparing samples from the United States and China. **(B)** Prediction outcomes for stool metagenome samples classified by country of origin, comparing samples from the United States and the Netherlands. In both panels, predictions are categorized as correct (green), ambiguous (purple), inconclusive (orange), or incorrect (yellow).

Supplementary Data

Supplementary Data 1. Standardized sample metadata, analysis labels, and sequence-depth summaries for all varKoder analyses.
This dataset provides standardized sample-level metadata for the five classification datasets analyzed in this study: domestic dog genomes, Vitis vinifera genome skims, human skin metagenomes, soil metagenomes, and human stool metagenomes. For the genome-derived datasets, the table includes accession metadata, taxon labels used for classification, run-level sequence-depth metrics, reference genome size, and estimated sequencing coverage. For the metagenomic datasets, the table includes accession metadata, metagenome labels used for classification, and SRA-derived sequence-depth metrics such as total bases, spots or reads, average read length, and file size. Genome coverage was not calculated for metagenomic datasets because no single reference genome applies. Redundant source-specific metadata fields were consolidated into harmonized columns to make sample inclusion, label construction, and coverage or sequencing-depth variation transparent across analyses.

Supplementary Data 2. varKoder performance summaries across classification tasks.
This dataset summarizes varKoder classification performance for all analyses included in the study. The Analyses sheet defines each task, the By_Query_Size sheet reports aggregate performance across query sizes or sequence-depth thresholds, the Per_Label sheet reports class-specific performance, and the Column_Guide sheet provides definitions for each metric. Performance is reported using exact-match accuracy, precision, recall, F1 score, and AUC where applicable. Macro-averaged metrics were calculated as the unweighted mean of per-label one-versus-rest metrics and therefore treat all labels equally. Support-weighted metrics were calculated by weighting each per-label metric by the number of true samples for that label and therefore reflect performance under the observed sample distribution. Reporting both summaries makes it possible to distinguish label-balanced performance from performance shaped by uneven sampling, particularly in analyses such as dog breed classification, soil geographic classification, and soil habitat classification.