16S rRNA k-mer composition encodes microbial functional potential

Liu J[1,*], De Paolis Kaluza MC[2], Bromberg Y[1,3,*]

1 Department of Biology, Emory University, 1510 Clifton Road NE, Atlanta, GA 30322, USA
2 Institute for Experiential AI, Northeastern University, 177 Huntington Avenue, Boston, MA 02115, USA
3 Department of Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA

* Corresponding author: yana.bromberg@emory.edu, jliu80@emory.edu
Tel: +1-732-932-9763 Ext. 218; Fax +1-732-932-8965

Running Title: embeRNA

Abstract

16S rRNA amplicon sequencing is widely used to profile microbiome taxonomic composition and functional potential. Most 16S rRNA-based analysis methods depend on comparing sequenced reads against reference marker genes from previously characterized organisms. Thus, method accuracy declines in environments dominated by uncharacterized microbes.

We uncovered a direct link between 16S rRNA and genome-encoded functions. Using fully sequenced bacterial genomes, we show that (i) whole-genome k-mer composition is predictive of functions encoded in the genome and (ii) 16S rRNA k-mer profiles reflect their source genome k-mer compositions. Leveraging these relationships, we developed embeRNA, a neural network-based framework that predicts functions directly from 16S rRNA k-mer embeddings, without taxonomy assignment or phylogenetic placement. Furthermore, by producing per-function probability scores rather than categorical assignments, embeRNA allows users to adapt decision thresholds to match study goals and sample characteristics, e.g. balancing precision vs. recall or accounting for community novelty.

We trained embeRNA on a large collection of bacterial function-omes and evaluated it using a stringent "novel microbes" benchmark, where all test 16S rRNA sequences were dissimilar to those seen in training (all <97% identical). On this test set of phylogenetically novel organisms, embeRNA outperformed reference-based methods overall and achieved significantly better performance for the “hard to label” set of functions.

In testing on soil metagenomes with paired 16S rRNA amplicon and whole metagenome shotgun (WMS) sequencing data, embeRNA recovered most WMS-inferred functions and yielded abundance profiles strongly correlated with WMS results. Together, our results indicate that 16S rRNA k-mer composition carries substantial functional signal and that 16S amplicon data can be used to complement WMS -based inference to broaden functional characterization of microbiomes, particularly in understudied environments.

Key words: 16S rRNA, k-mer, functional profiling, microbiome, metagenome analysis

Introduction

Microbial communities inhabit every ecosystem and host-associated environment[1-3], where they collectively drive core biochemical reactions and shape ecological processes[4-6]. Advances in whole metagenome shotgun (WMS) sequencing and targeted 16S rRNA amplicon sequencing have enabled broad surveys of community compositions and, increasingly, functional potentials.

A wide range of computational pipelines (at least 49 by our count) are routinely used to translate sequencing data into taxonomic profiles and/or functional potentials[7-55] (**Figure 1**). Whole metagenome shotgun (WMS) sequencing -inferred functions are generally treated as being high-confidence/precision. However, WMS sequencing depth limits coverage of individual, especially rare, genes and organisms, making it easy to miss low-abundance or infrequent functionality (low recall) and limiting the resolution of abundance estimates[56-58].

16S rRNA amplicon sequencing is a cost-effective approach for ensuring both high precision and recall in profiling microbiome taxonomic composition, typically at the genus level. Existing annotation methods link 16S rRNA sequences to the corresponding microbial isolates via aligning input sequences[18] or corresponding k-mer vectors[39, 59] to reference trees or databases[59-61]. Once microbiome taxonomic/phylogenetic profiles are established, corresponding functional profiles can be inferred by combining functions of member taxa, e.g. as is done by PICRUSt / PICRUSt2 or Tax4Fun[40, 62, 63]. While this approach is effective for well-characterized environments such as the human gut[64, 65], it is constrained by reference data availability, yielding biased or incomplete function predictions for poorly characterized microbiomes.

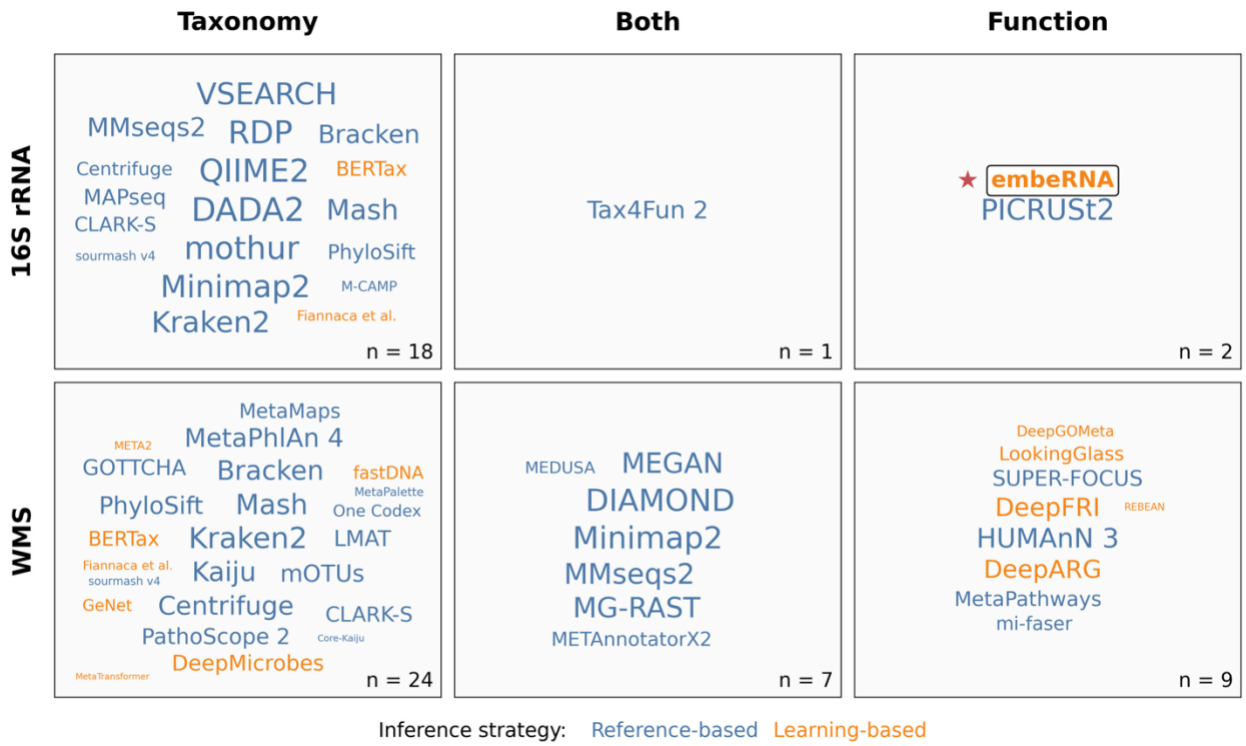

Figure 1. Microbiome analysis tools vary.
49 routinely used microbiome analysis tools are organized by input sequence type (16S rRNA or whole metagenome shotgun sequencing, WMS) and primary prediction target (taxonomy, both taxonomy and function, or function). For existing tools, name size is proportional to log2-transformed Google Scholar citation count; embeRNA is highlighted as the method introduced in this study and is not citation-scaled. Color indicates inference strategy (blue: reference-based tools or tools commonly embedded in workflow requiring reference databases; orange: learning-based tools). Reference-based methods infer outputs by anchoring input data to known references, whereas learning-based methods infer outputs from patterns learned during training without explicit reference matching at inference time. Because some tools, such as Diamond or MMseqs2, are used broadly beyond microbiome research, citation count/name size should be interpreted as a rough indicator of overall adoption rather than microbiome-specific usage.

To summarize, most extant approaches focus on resolving taxonomy, whether from 16S or WMS data (**Figure 1**). Many methods treat taxonomic composition as an end goal, answering the question “Who is there?” Some determine the taxonomic composition as the first step in functional inference, asking “What are they doing?” Notably, most of these tools rely on reference-based inference, or are commonly embedded in workflows that depend on reference databases, whereas learning-based approaches infer outputs from learned sequence patterns without explicit reference matching during inference. In contrast, some methods profile functions from WMS data directly, where both reference-based and learning-based approaches are available. Overall, however, accurately and comprehensively characterizing microbiome functional potential remains a challenge. Here we ask, for the first time, can microbiome functions be inferred from 16S rRNA sequence directly, i.e. without seeking a reference?

Sequence is key to understanding function. Short k-mer composition vectors provide compact, fixed-size representations of DNA sequences of any length, capturing reproducible genomic signatures such as GC content and oligonucleotide usage biases[66-69]. Because the 16S rRNA gene is embedded within the genome, its k-mer patterns may reflect aspects of genome-wide sequence composition, resulting from exposure to the same evolutionary constraints and, possibly, carrying information beyond taxonomy alone. Indeed, prior studies have shown that 16S rRNA k-mer features can predict microbiome sample characteristics and microbial phenotypes, e.g. habitat or host-associated traits[70], suggesting that marker-gene k-mers encode informative biological signal.

Here, we show that (i) whole genome k-mer representations are predictive of microbial functions and that (ii) 16S rRNA k-mer representations reflect corresponding whole genome k-mer representations. Together, these observations support a direct 16S rRNA-to-function link.

To make further use of this link, we built embeRNA (embedding RNA), a neural network-based method that predicts functional profiles directly from 16S rRNA k-mer embeddings without explicit phylogenetic placement, or taxonomic assignments.

Results

Whole genome k-mer representations reflect genome-encoded functions. We first investigated whether bacterial whole genome k-mer representations (1to5-mers; **Methods**) reflect the corresponding encoded functions. Using 1,369 bacterial genomes (1,349 unique species) from the *Balanced set*, we trained classifiers to predict enzymatic function presence from whole genome k-mer vectors (**Methods**). These classifiers substantially outperformed simple (random and majority) and permuted-label control baselines (Table 1; **Methods**), confirming that genome k-mer composition carries predictive information about the encoded functional repertoires.

Table 1. Whole genome k-mers predict functions.

Model	per-function F1 (mean ± SE)*	Hamming loss (mean ± SE)**
Logistic Regression (LR)	**0.750 ± 0.008**	0.136 ± 0.001
Extra Trees (ET)	0.714 ± 0.010	**0.108 ± 0.002**

Random baseline	0.573 ± 0.011	0.237 ± 0.001
LR permuted-label control	0.562 ± 0.010	0.276 ± 0.001
ET permuted-label control	0.538 ± 0.013	0.183 ± 0.002
Majority baseline	0.511 ± 0.014	0.173 ± 0.001

Bolded values indicate best performance per column. *Values are mean ± SE F1 (Eqn. 1) across function labels and cross-validation (CV) folds. F1 scores weigh all functions equally regardless of prevalence; **Values are mean ± SE across CV folds; Hamming loss (Eqn. 2)

16S rRNA k-mers reflect whole genome k-mer composition. To evaluate if 16S rRNA sequences can predict whole genomes in the k-mer space, we trained regression models using the genomes in the *Balanced set* (**Methods**). Both of our models substantially outperformed the dummy mean baseline (**Table 2**) and the two permutation controls: (i) shuffled genome/16S rRNA pairings and (ii) shuffled within-genome k-mer values, confirming that predictive signal is specific to real sequence relationships.

Table 2. Predicting whole genome k-mers from 16S rRNA k-mers.

Condition	***Model***	***R^2 (mean ± SE)****
Real data	Ridge regression	**0.785 ± 0.008**
	Random forest	**0.784 ± 0.006**
	Dummy mean baseline	-0.004 ± 0.002
Permuted pairs	Ridge regression	-0.007 ± 0.010
	Random forest	0.005 ± 0.012
	Dummy mean baseline	-0.004 ± 0.002
Genome permutation	Ridge regression	-9.992 ± 0.288
	Random forest	-9.834 ± 0.282
	Dummy mean baseline	-9.831 ± 0.282

*Values are mean ± SE across CV folds.

embeRNA predicts microbial functions from 16S rRNA k-mers. We further asked whether 16S rRNA-derived k-mers could directly predict genome-encoded functions.

Note that complete 16S rRNA gene sequences, comprising ten semi-conserved regions interspersed with nine hypervariable regions, are primarily inferred from complete genome sequencing. However, targeted/amplicon sequencing used for microbiome analysis relies on a collection of available primers. The longest sequence that can thus theoretically (if not practically) be extracted contains all nucleotides from just before the start of variable region 1 (V1) through the end of variable region 9 (V9), i.e. effectively

dropping most of the first and last conserved regions. We thus used the 16S rRNA V1-V9, extracted from the complete genome data in our *embeRNA set* (24,585 genomes), as input and binary EC three-digit profiles as output to develop the first version of embeRNA - a neural network for predicting functions from 16S rRNA k-mer representations (**Methods**).

To rigorously evaluate model generalization, we used a genus-stratified strategy, where all genomes from the same genus were assigned to the same split (train, validation, or test). Across ten independent repetition models with different random splits, embeRNA achieved a median F1=0.91 (IQR: 0.89-0.91) on held-out test sets; best-performing replicate reached F1=0.92 (precision 0.93, recall 0.91) and precision-recall area under the curve (PR AUC) of 0.98. A shuffled-label baseline, randomly permuting labels across all samples in each test set, performed substantially worse (PR AUC = 0.62; **Methods**), confirming that embeRNA captured genuine sequence k-mer-function associations.

The final (V1-V9) model, retrained on all available data in *embeRNA set*, achieved F1=0.96 (precision 0.96, recall 0.96; **Eqn. 1**) when evaluated on the full *embeRNA set*, representing best-case performance on previously seen sequences (thresholds derived as the median optimal F1 cutoff across replicates, **SOM Table 1; Methods**).

embeRNA generalizes across amplicon-sequenced regions. The first embeRNA model used the nearly complete 16S rRNA gene sequence (V1-V9) as input. However, most amplicon sequencing datasets focus on shorter, mostly variable, regions. V3-V4 is most frequently targeted in modern 16S rRNA amplicon sequencing studies, followed closely by the V4 region alone. V3-V4 has well-validated, nearly universal primers and spans ~460bp, which is ideal for paired-end Illumina sequencing (e.g. 2×250bp or 2×300bp). It thus allows for high quality, cost-effective sequencing while maintaining strong taxonomic classification accuracy. The V4 region alone (~290 bp) is commonly selected in large-scale projects for cost efficiency. Historically, V6-V8 was popular during the 454 pyrosequencing era but fell out of favor with global switch to Illumina and problematic GC-regions[71]. The choice of gene regions to sequence often depends on the sequencing platform constraints, target taxonomic groups, and plans for comparison to existing datasets.

We thus evaluated embeRNA across six additional regions (**Figure 2**). All region-specific models performed comparably to V1-V9: for example, the V6-V8 model achieved a median F1 of 0.89 [IQR: 0.88-0.90], with the best replicate reaching F1 of 0.93 and PR AUC of 0.98 (shuffled-label baseline PR AUC = 0.60). Performance was robust across all tested regions (**Figure 2**).

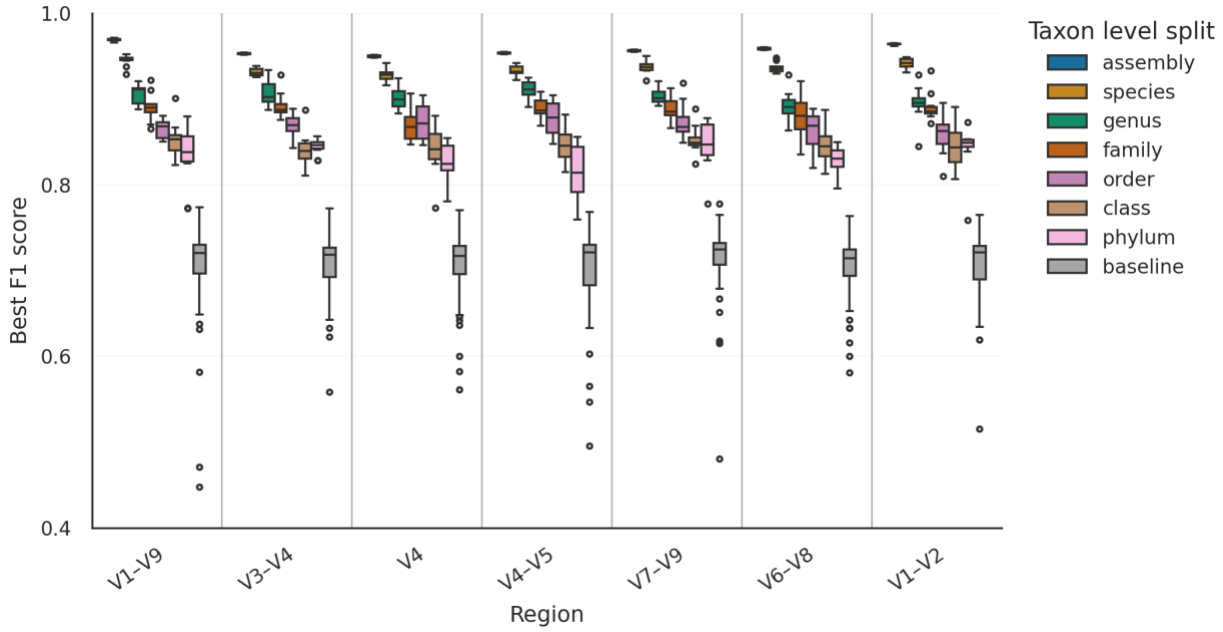

Figure 2. embeRNA maintains strong performance across 16S rRNA regions and stringent taxonomic holdout splits.
Boxplots show the distribution of best F1 scores for embeRNA-EC models trained on different 16S rRNA regions (x-axis) and evaluated under increasingly stringent taxon-level holdout splits (colours: assembly to phylum; "baseline" indicates a random/shuffled control). For each region × split level, results are summarized across ten independent replicate models (different random seeds for splitting training/validation/testing), each evaluated on its corresponding held-out test set. Vertical separators delineate regions.

embeRNA outperforms reference-based methods on phylogenetically novel 16S rRNAs. To evaluate functional inference on previously unseen taxa, we assembled a *Novel Microbes set* (**Methods**): 5,542 16S rRNAs from genomes released after the embeRNA training set, with <97% similarity to any training reference (median maximum similarity: 95%, IQR: 93-96%).

We asked: how well could embeRNA inform functions for these novel microbes? From the *Novel Microbes set*, V3-V4 regions were extracted (5,473 sequences) and EC three-digit functions predicted using four strategies: (i) embeRNA V3-V4; (ii) PICRUSt2 (phylogenetic placement); and (iii-iv) taxonomy-to-core approaches using Kraken2[39] and RDP Classifier[59] k-mer based taxonomic classifiers; taxonomy-to-core methods assign

each V3-V4 sequence to a taxonomic group and use the "core" EC set of functions present in every member of the assigned group (**Methods**).

Prediction coverage differs across tools. By nature of the neural networks, embeRNA generates predictions for every input sequence and therefore was able to infer presence or absence of functions for all 5,473 V3-V4 sequences. For the taxonomy-to-core function approaches, both RDP Classifier and Kraken2 assigned all sequences to a taxonomic group. However, for 511 sequences (9%), the Kraken2-assigned taxa (and their ancestors) were absent from the *embeRNA set*, preventing function inference. For PICRUSt2, one sequence was too distant from the reference tree to yield predictions, and four additional sequences had NSTI > 2 (nearest sequenced taxon index), a range suggested to generate unreliable predictions due to lack of closely related references[40]; otherwise, median NSTI=0.06 [IQR: 0.03-0.11].

For standardizing the comparison data, analysis was restricted to the 4,961 V3-V4 sequences where all four tools generated predictions. The downstream evaluation functional space was defined as the union of 255 functions spanning (i) the 237 functions predictable by all four 16S rRNA-based tools and (ii) the 241 actual functions across *Novel Microbes set* assemblies. The microbes in this set contain a median of 137 unique EC three-digit functions each. embeRNA and PICRUSt2 predicted comparable numbers of functions per sample - a median of 137 and 140, respectively, versus 86-89 for taxonomy-to-core approaches.

embeRNA attained best overall performance. embeRNA achieved the highest F1 of 0.851 (95% bootstrap CI: [0.851-0.852]), followed by PICRUSt2 (0.835; CI: [0.834-0.835]), Kraken2 (0.751; CI: [0.749-0.753]), and RDP (0.749; CI: [0.748-0.751]). The taxonomy-to-core methods achieved highest precision (~0.93) but substantially lower recall (~0.63), consistent with their conservative "core functions only" design. embeRNA and PICRUSt2 attained relatively lower precision (embeRNA precision=0.845; PICRUSt2 precision=0.816), and higher recall (embeRNA recall=0.857 and PICRUSt2 recall=0.854).

embeRNA enables tunable precision-recall trade-offs. Unlike the binary outputs of other tools, embeRNA outputs continuous probabilities (**Figure 3**), enabling threshold adjustment to prioritize precision or recall. For example, at RDP-precision-matched

operating point (precision = 0.927), embeRNA achieved substantially higher recall (0.716, 95% CI: [0.713-0.720]) vs. RDP (0.629).

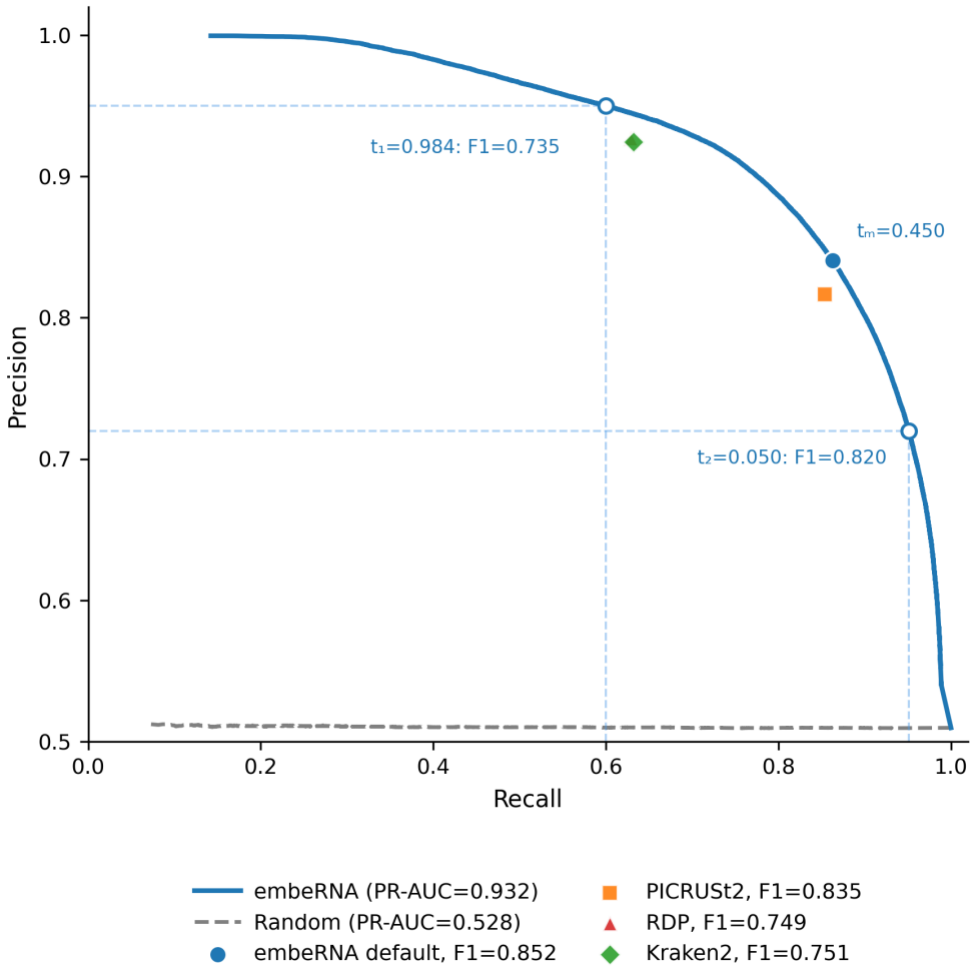

Figure 3. embeRNA outperforms reference-dependent 16S rRNA-based approaches on the *Novel Microbes set* and permits threshold-dependent tuning.
Precision-recall (PR) curve for embeRNA-EC V3-V4 model evaluated on the *Novel Microbes set* (PR-AUC = 0.932). The grey dashed line shows the random shuffled baseline (PR-AUC = 0.528). Filled markers denote the default operating points for embeRNA (t_m=0.450: F1=0.852, precision=0.841, recall=0.863), PICRUSt2 (F1=0.835), RDP taxonomy-to-core (F1=0.749), and Kraken2 taxonomy-to-core (F1=0.751) on the same V3-V4 sequences. Open blue circles indicate two additional embeRNA operating points obtained by adjusting the classification threshold to favour either higher precision (with threshold t_1=0.984: F1=0.735, precision=0.950, recall=0.600) or higher recall (with threshold t_2=0.050: F1=0.820, precision=0.720, recall=0.951); dashed guidelines project these points to the axes to illustrate the associated precision-recall trade-offs.

For hard to classify functions embeRNA has fewer false positives than PICRUSt2. Some functions such as transcription and translation are present everywhere and therefore easy to predict. Others are absent across the vast swaths of the community and are also fairly easy to label as such. We asked about the difficult ones, i.e. when embeRNA and PICRUSt2 give conflicting calls for the same function, which is more likely to be correct?

As taxonomy-to-core approaches called less than two thirds of the actual functions, we restricted this analysis to embeRNA vs PICRUSt2 to avoid conflating disagreement with coverage differences. Here, we used each genome's encoded functions as ground truth for each prediction.

Among the 27 (median, [IQR: 21-35]) functions per 16S rRNA, where embeRNA and PICRUSt2 disagreed (159,618 total disagreement calls across all *Novel Microbes set*), embeRNA was more frequently correct than PICRUSt2 (58.6% of the time vs. 41.4%). Note, a random baseline was correct 50.7% of the time (**Methods**). This advantage was most pronounced for truly absent functions (true negatives, n = 89,547; embeRNA correct 63.9% of the time vs. PICRUSt2 36.1%; random baseline 56.0%), indicating fewer false-positive calls in the hard to predict functional space (**Figure 4**, **SOM Figure 1**). Among truly present functions (true positives, n = 70,071), embeRNA remained more accurate than PICRUSt2 (52.0% vs. 48.0%; 43.8% for random), although slightly less obviously so. This true-negative advantage is orthogonal to F1, which by definition does not capture absence-calling (negative) accuracy.

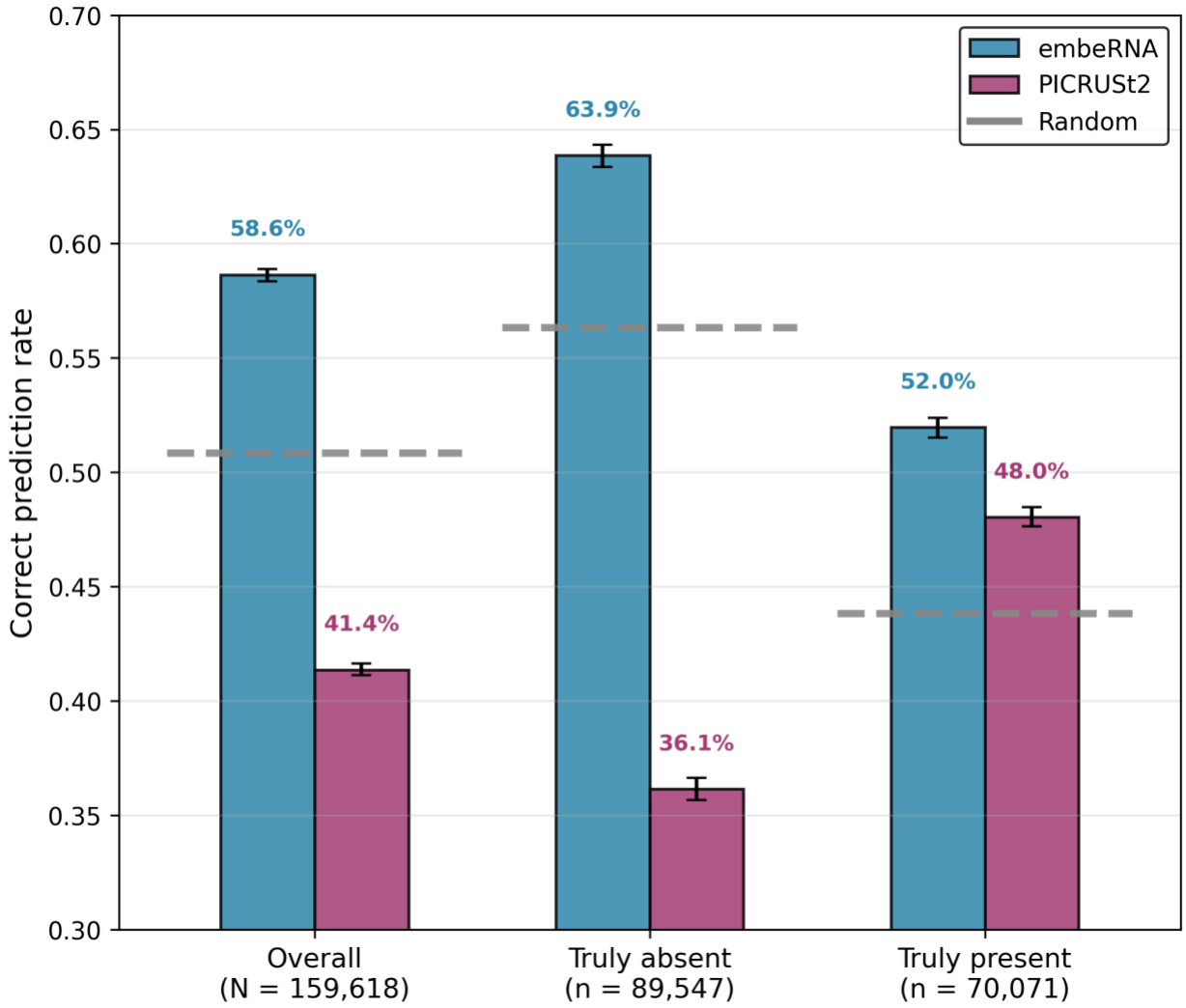

Figure 4. embeRNA outperforms PICRUSt2 in head-to-head prediction of hard to predict functions in the *Novel Microbes set*.
Bar plots show the proportion of correct predictions made by each tool (embeRNA, blue; PICRUSt2, purple) among hard to predict functions for which the two tools disagreed on presence or absence. Error bars indicate 95% bootstrap confidence intervals. The random baseline (grey dashed lines) was generated by assigning function presence or absence according to each function's empirical prevalence.

embeRNA predicts WMS-inferred functions in soil samples well. We compared functional profiles from embeRNA and PICRUSt2 against HUMAnN 3[44] processed WMS data for 22 blueberry rhizosphere and bulk soil samples with paired 16S rRNA V6-V8 amplicon and WMS sequencing data[72] (**Methods**). Both 16S rRNA-based tools recovered 95% of WMS-detected functions per sample, as expected given the limited depth of WMS vs. amplicon sequencing[57].

Additionally, embeRNA and PICRUSt2 predicted a quarter more functions than WMS (median embeRNA 234, PICRUSt2 235, versus HUMAnN 3 186 EC functions per sample). Most (56 per sample) of these functions were predicted by both 16S tools.

Note that in microbiome studies, abundance profiles are more common than presence/absence profiles. We computed pseudo-abundance profiles by scaling embeRNA's and PICRUSt2's per-16S presence/absence vectors according to their observed frequencies within a sample (**Methods**). embeRNA's abundance profiles were better correlated with HUMAnN 3 (Spearman correlation median ρ = 0.74, IQR: 0.72-0.75) than PICRUSt2's profiles (median ρ = 0.70, IQR: 0.69-0.71).

To understand performance differences, we further compared the tools' function abundance levels by (i) determining each function's relative abundance within each tool's profile, and (ii) splitting functions in each sample into ten decile bins based on their abundance rank. Across the compared functions per sample (median 251), for functions with large abundance bin discrepancies between the two tools (≥3 bins apart; ~4% of compared functions per sample), embeRNA's estimates aligned more closely with WMS in a median of 72% of cases per sample (IQR: 64-82%; for example, sample Bact200 is displayed in **Figure 5**).

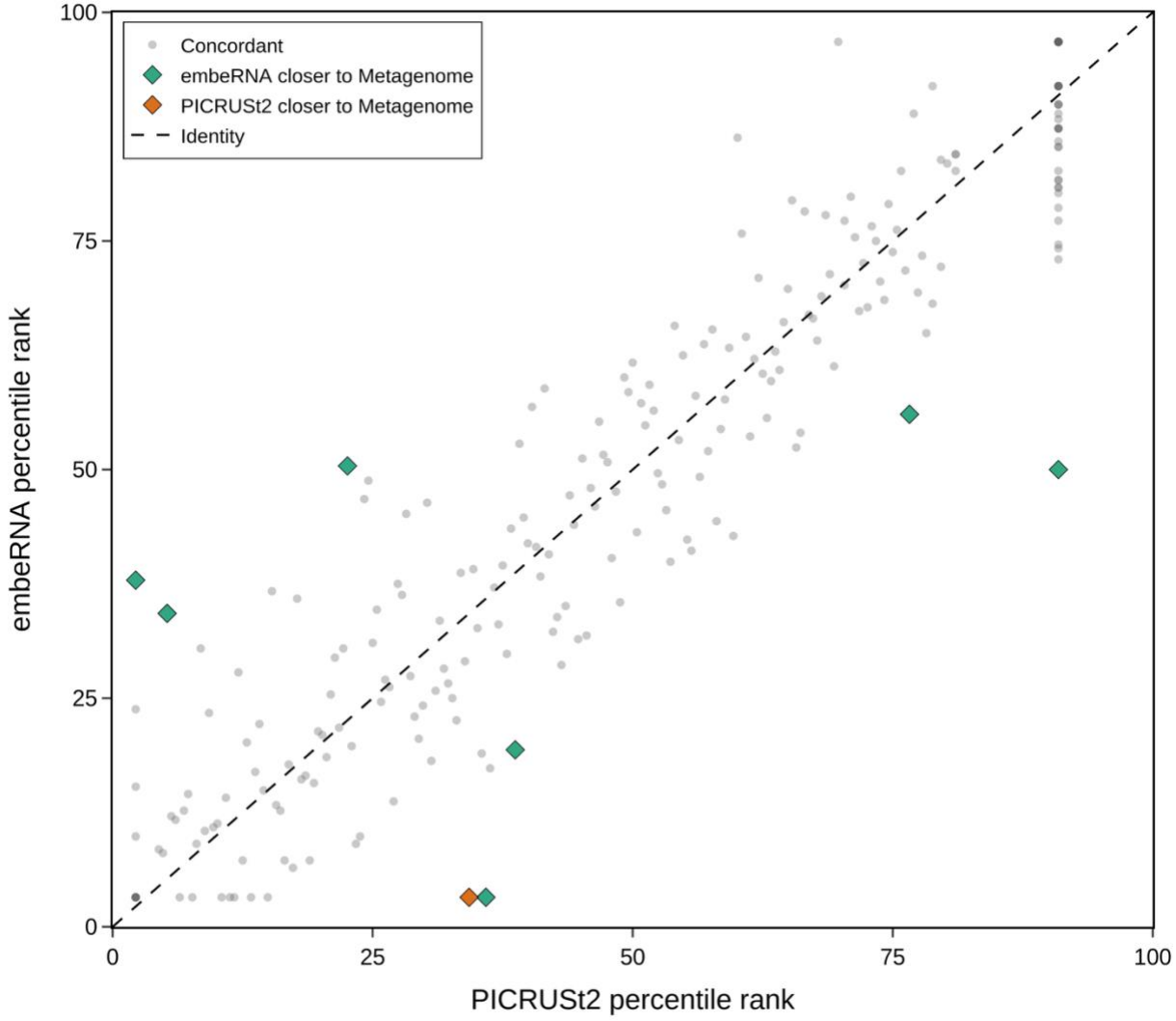

Figure 5. embeRNA more accurately reflects WMS-informed functional abundance when predictions disagree with PICRUSt2.
Shown is an example soil sample (Bact200) with paired 16S rRNA amplicon and whole-metagenome shotgun (WMS) data. Functional profiles were inferred from 16S data using embeRNA and PICRUSt2, and from WMS using HUMAnN3. Within each method, functions were ranked and grouped into decile-based abundance bins. Each point represents a function, positioned by its percentile rank in the PICRUSt2 profile (x-axis) and embeRNA profile (y-axis). The dashed line indicates identity between methods. Functions with substantial disagreement (≥3-bin difference) were evaluated against the WMS-derived abundance: points are colored green if embeRNA is closer to WMS, orange if PICRUSt2 is closer, and blue if both are equally close.

embeRNA framework is flexible with respect to the choice of functional space. Although until this point we focused on EC three-digit functions, embeRNA can be trained to predict any user-defined functions. The Fusion database, for example, catalogs 433,891 microbial functions, spanning both known and uncharacterized functionality[73]. We trained embeRNA-Fusion models using 16S rRNA sequences from the Fusion-specific assemblies paired with curated Fusion functional profiles. To reduce sparsity and

focus on broadly supported functions, we trained embeRNA with, and predicted, only the Fusion functions present in at least 10% of Fusion assemblies (4,482 functions total).

Across ten repeated random genus-level split evaluations using full-length (V1-V9) 16S rRNA, embeRNA-Fusion achieved a median F1 of 0.82 on held-out test sets (precision 0.81; recall 0.84; **SOM Figure 2**). Models trained on shorter regions performed comparably (e.g. V6-V8 F1 0.80; precision 0.80; recall 0.81), demonstrating that the same 16S rRNA k-mer-to-function framework readily adapts to alternative ontologies and customized function sets.

Discussion

In this work, we describe a link between 16S rRNA and organismal function that has not previously been observed: 16S rRNA k-mer composition reflects that of its source genome, which, in turn, tracks the history of and organism's evolutionary and environmental constraints - the two factors that also shape its function potential. Based on this observation, we built embeRNA, a framework that infers functional potential from 16S rRNA k-mer composition modelling rather than from taxonomy or phylogenetic placement, enabling functional characterization for microbial "dark matter".

Mechanistic basis for inferring function from 16S rRNA k-mer composition. A central question underlying embeRNA is whether (and why) the k-mer composition of a single marker gene carries sufficient information to infer an organism's broader functional potential. Our results support a mechanistic bridge built on two principles of genomic evolution.

First, whole genome k-mer composition is predictive of genome-encoded functional profiles. This relationship is governed by two interacting forces: ancestral lineage and environmental selection. Lineage history provides the primary template for genomic signatures, as inherited biases in replication and DNA repair mechanisms drive characteristic oligonucleotide usage patterns[66-69]. These 'phylogenetic signals' ensure functional conservation within lineages[74, 75]. However, prolonged exposure to specific ecological constraints further refines these patterns, driving functional convergence across disparate lineages[76, 77]. Habitat-linked shifts in GC content and higher-order oligonucleotide properties demonstrate how environmental filtering imposes readable,

niche-specific signatures on the genome[69, 78-80]. Because lineage history and niche adaptation simultaneously shape functional repertoire and genome k-mer composition, the latter serves as a robust proxy for the former.

Second, 16S rRNA k-mer composition closely mirrors that of its source genome, consistent with the established cohesion of compositional signatures across genomic regions[66-68, 79]. Together, these two observations close a logical chain: genome k-mers predict functions, and 16S rRNA k-mers reflect genome k-mers, so direct prediction from 16S rRNA k-mers to functions is mechanistically grounded. This chain supports generalisation beyond taxonomic boundaries: even when a query sequence has no close phylogenetic neighbour in the training data, the signal embedded in its k-mer composition shaped by lineage and environmental adaptation history may still point toward the correct functional archetype.

embeRNA's compositional inference reduces false positives for distant relatives. embeRNA's small advantage over PICRUSt2 on novel microbes (F1: 0.851 vs. 0.835) understates the most consequential method difference, which lies in the true-negative calls that the F1 metric fails to capture. Among functions where the two tools disagree, embeRNA correctly identifies truly absent functions twice as often as PICRUSt2; that is embeRNA says these are absent, while PICRUSt2 claims their presence.

This pattern extends to community-level analysis: in soil samples, among functions where embeRNA and PICRUSt2 disagree, embeRNA's abundance estimates aligned more closely with WMS 72% (median, IQR: 64%-82%) of the time. The convergence of this pattern across the set of individual novel microbes and complex soil communities supports the robustness of embeRNA's compositional inference.

The mechanistic basis for this reduced false-positive rate lies in how the two tools handle novelty. PICRUSt2 transfers functional annotations from the nearest reference(s) in a pre-built phylogenetic tree. For divergent sequences, the nearest reference may be taxonomically distant, and imputation produces dense predictions reflecting the reference's broad functional repertoire rather than the query organism's specific capabilities, systematically overcalling reference-specific traits. embeRNA, by contrast, assigns per-function probability scores anchored to the query's own k-mer signature:

when a functional signal is absent from the k-mer profile, the model assigns low support regardless of what phylogenetic neighbours encode.

A practical consequence is the tunability of embeRNA's output. Because predictions are continuous probabilities rather than binary calls, users can adjust the decision threshold to match study-specific precision-recall requirements. At precision-matched operating points (precision = 0.927, matching RDP), embeRNA achieved higher recall (0.716 vs. 0.629), demonstrating genuine flexibility rather than a fixed trade-off.

Complementarity between WMS and 16S rRNA-based function inference. Both embeRNA and PICRUSt2 recovered WMS-inferred soil functional profiles well, but both also predicted additional functions per sample (a median of 58 for embeRNA and 64 for PICRUSt2). Some of these are ecologically central to soil biogeochemical cycles: nitrogenases (EC 1.18.6 and EC 1.19.1, predicted by both tools in all samples) and complex carbohydrate degraders (EC 4.2.2, predicted by both tools in 13 samples) are well-documented components of soil microbial communities[81, 82]. We thus suspect that WMS sequencing depth may have excluded these likely true positives[57]. However, as embeRNA and PICRUSt2 share overlapping reference genome collections, we cannot exclude the chance that these agreements reflect shared systematic over-predictions. Resolving the true source of these observations would require deeper WMS sequencing or targeted experimental validation.

Whereas embeRNA and PICRUSt2 are expected to perform similarly well for well-characterized organisms, our results suggest that embeRNA provides a greater advantage for predicting functions in novel microbes with limited reference coverage.

Expanding functional inference into unannotated space. The embeRNA architecture is inherently flexible: to illustrate this, we trained embeRNA-Fusion on the Fusion database and applied it to the same 22 blueberry soil samples[72], for which no Fusion-based profiling had previously been reported.

Predicted functional profiles clustered by sampling site (PERMANOVA: R^2=0.42; p=0.0003) and, within the Collingwood site, by soil compartment (rhizosphere vs. bulk; R^2=0.33, p=0.01; **SOM Note 1**). Differentially abundant functions between rhizosphere and bulk soil (FDR < 0.05) included enrichment of functions describing chemical defence

and physical colonization machinery in the rhizosphere which involve functions not described in the Enzyme Commission space, validating the framework against known root-interface biology[83-86].

More notably, 43% of the differentially abundant functions corresponded to Fusion clusters with no annotated protein sequences, demonstrating that embeRNA-Fusion detects ecologically structured variation in functionally dark space, a capability that EC-based approaches cannot provide. Experimental validation of these uncharacterised functions, for example, via metatranscriptomics or targeted biochemical assays, represents an important direction for future work. We expect embeRNA to provide novel understanding of microbial functions in other underrepresented ecological microbiomes.

Conclusions

We demonstrate that 16S rRNA k-mer composition preserves genomic signatures predictive of bacterial genome-encoded functional potential. By leveraging this link, embeRNA achieves robust performance on phylogenetically novel organisms, with a particular advantage in reducing false-positive functional assignments compared to reference-based approaches. Our results reveal complementarity between 16S rRNA- and WMS-based functional profiling: the two approaches recover largely overlapping functional landscapes, but 16S-based inference extends visibility into rare and low-abundance taxa that WMS routinely under-samples. The generalisation to uncharacterised functional space via Fusion further enables detection of ecologically structured variation in regions inaccessible to annotation-dependent approaches.

While embeRNA treats individual 16S rRNA sequences as independent inputs, microbial ecosystems are defined by emergent community-level activities. Integrating all 16S rRNA sequences from a sample into a unified input to model community-scale metabolic networks represents a natural next step. By shifting from taxonomic "guilt-by-association" to intrinsic sequence-to-function mapping, embeRNA provides a scalable framework for deciphering the functional architecture of complex microbiomes.

Methods

Datasets description

This work used four datasets as described below (***Balanced, embeRNA, Novel Microbes,*** and ***Microbiome sets***).

1. ***Balanced set:*** To assess the 16S rRNA k-mers to genome k-mers to function relationships, we assembled a phylogenetically diverse set of bacterial genomes. Our group previously constructed the Fusion database[73], comprising 433,891 function clusters derived from 8,906 complete bacterial genome assemblies from NCBI RefSeq database. From this resource, a species-balanced subset of 1,502 assemblies was created to capture ⩾90% of the genetic diversity present across the full Fusion genome collection.

   Here, we use this subset, further filtering it to retain assemblies with (i) at least one full-length 16S rRNA gene and (ii) an annotated species-level taxonomy label. This *Balanced set* comprises 1,369 organisms/assemblies spanning 1,349 unique species and containing 4,744 full-length 16S rRNA sequences (median 3 copies per organism; Interquartile range / IQR 2-4).

   From this resource, we constructed two datasets: (i) 16S rRNA/genome pairs, where each of the 16S rRNA sequences and its source genome were independently represented as concatenated 1, 2, 3, 4 and 5-mer (1to5-mer) vectors; and (ii) genome/function pairs, where each of the 1,369 organisms' 1to5-mer genome vector was paired with its Enzyme Commission (EC) number[87] three-digit functional profile.

   Across all 1369 organisms in the *Balanced set*, 208 unique EC three-digit functions were observed. Individual organisms carried a median of 124 [IQR: 106-137] functions, while each function was present in 806 [IQR: 311-1,281] assemblies.

2. ***embeRNA set:*** To develop models that infer microbial functions directly from 16S rRNA k-mer composition, we compiled *embeRNA set*, a dataset of 24,585 unique complete prokaryotic genome assemblies derived from (i) PICRUSt2[40] (n = 16,136) and (ii) Fusion[73] (n = 8,449) references. Note that while all genomes from Fusion database are bacteria, 364 of PICRUSt2 genomes are archaea. Assemblies that appeared in both resources (n = 2,620) were removed from the PICRUSt2 set to avoid duplication. All retained assemblies were complete genomes containing at least one 16S rRNA sequence and assigned taxonomic labels (species to phylum).

Overall, the *embeRNA set* spans 10,359 unique species across 2,262 genera (**SOM Figure 3**) and contains 56,734 16S rRNA sequences (median 1 [IQR: 1-3] copy per genome). Each of these 16S rRNAs was represented to 1to5-mer, and was paired with the EC (three-digit) function profile of its source genome.

Because the original PICRUSt2 EC profiles do not include EC numbers from enzymatic class seven[40], we re-extracted full-digit EC annotations for all PICRUSt2 reference genomes from the JGI IMG database[88] using the same procedure employed in PICRUSt2, and subsequently summarized these annotations to EC three-digit profiles. When identical 16S rRNA sequences were associated with different annotated function sets across assemblies, we retained each sequence with its original function profile.

Across the 8,449 Fusion assemblies, 209 unique EC three-digit functions were observed, spanning all seven enzymatic classes. The 16,136 PICRUSt2 assemblies contained 260 EC three-digit functions, also spanning all seven classes. We defined the functional space of the *embeRNA set* as the union of the above two, yielding 261 EC three-digit functions, corresponding to 81% of all EC three-digit classes recognized at the time of writing (accessed 01/16/2026)[89]. Within this functional space, a typical genome has a median of 141 [IQR: 122-155] EC three-digit functions, and each function is present in 14,140 [IQR: 2,506-23,385] genomes.

For embeRNA model training, we further restricted the target space to EC functions present in at least 1% of all assemblies in *embeRNA set*, resulting in 238 EC three-digit functions used in the embeRNA models. Functions underrepresented in current reference collections fall outside this scope regardless of their true environmental frequency. This threshold is user-definable, allowing users to adjust the precision/recall tradeoff for their specific application.

3. ***Novel Microbes set:*** To evaluate model generalizability, we additionally assembled an evaluation set of 2,533 complete bacterial genomes comprising 5,542 unique full-length 16S rRNA genes deposited into the RefSeq database after (02/28/2018) the construction of both components of the *embeRNA set* (Fusion/PICRUSt2). In this set, each 16S rRNA had <97% sequence identity (typical same species threshold) to any

of the sequences in the *embeRNA set* (**SOM Note 2**). Each pair of the 5,542 16S rRNA sequence' 1to5-mer embedding with the EC three-digit function profile of its source genome was generated and served as one sample pair of input and output. The 2,533 assemblies contain a total of 241 unique EC three-digit functions, spanning all seven enzyme classes.

4. ***Microbiome set***: To evaluate model performance on real amplicon sequencing data, we used a set of 22 bulk and blueberry rhizosphere soil samples with both 16S rRNA V6-V8 amplicon sequencing and WMS sequencing data from a previous study[72].

Data preparation

Here we describe sequence featurization, function profile construction, model training/evaluation, and dataset-specific processing steps required to reproduce each analysis.

1to5-mer embeddings. We represented DNA sequences (full-length 16S rRNA genes, extracted 16S rRNA regions, or whole genomes) using k-mer frequency embeddings constructed by concatenating normalized frequency vectors for k = 1-5. For each k, counts of all possible k-mers were computed across the sequence and normalized by the total number of observed k-mers of that order. The resulting five normalized vectors were concatenated into a 1,364-dimensional feature vector. This representation was used for all sequence-based analyses and models in this study.

EC three-digit functional profiles. We represented genome enzymatic potential using EC three-digit presence/absence profiles, derived from four-digit EC annotations for each genome assembly in the *Balanced*, *embeRNA*, and *Novel Microbes* datasets.

Four-digit EC annotations for assemblies in the above-mentioned sets were first collected. From the latest Fusion database, we obtained the four-digit ECs for the Fusion-sourced assemblies in *embeRNA set* and all assemblies in the *Balanced set*[73]. From JGI IMG databases, we collected the latest four-digit EC annotations for the remaining PICRUSt2-sourced assemblies in *embeRNA set*[88]. For assemblies in the *Novel Microbes set*, EC annotations were generated de novo. Protein sequences for each genome were retrieved from NCBI RefSeq and annotated with four-digit EC numbers following the procedure as

described in the Fusion study[73]. Protein-level EC annotations were then aggregated at the genome level.

For each genome, four-digit EC numbers were then truncated to the first three digits (e.g., 1.1.1.2 → 1.1.1). An EC three-digit function was considered present if at least one protein in the genome carried the corresponding EC annotation and absent otherwise.

Predicting genome k-mer composition from 16S rRNA k-mers

We trained regression models to predict a genome's 1to5-mers from those of its 16S rRNAs, on the *Balanced set* of assemblies.

Cross-validation. Because genomes can encode multiple 16S copies, evaluation used grouped 5-fold cross-validation by assembly, assigning all 16S sequences from the same assembly to the same fold. The same folds were reused across models and controls to enable paired comparisons.

Regression models.

1. Dummy mean baseline: predicted the average whole genome 1to5-mer profile across all training-fold samples for all test samples, computed per k-mer label.
2. Ridge regression + PCA: a linear baseline in which input features were standardized using training-fold statistics, reduced with PCA (up to 256 components), and mapped to genome 1to5-mers using ridge regression. Regularization strength was selected by inner cross-validation on a log-spaced alpha grid (10^{-4}-10^{4}).
3. Random forest regressor: a nonlinear baseline (300 trees; max depth 30; min samples per leaf 10).

Evaluation metrics. We evaluated models' performance on held-out folds using variance-weighted multi-output R^2, weighting per-feature R^2 by the variance of that feature across genomes.

Permutation controls. To verify that models learned a genuine 16S rRNA to genome mapping, we repeated training under two training-label destructive regimes (within each training fold) while evaluating against the true test labels by: (i) randomly shuffling the 16S rRNA - genome pairings, and (ii) randomly permuting the 1to5-mer feature positions within each genome vector.

Predicting EC functional profiles from genome k-mers

We trained models to predict a genome's EC three-digit functions' presence/absence from its whole genome 1to5-mers.

Cross-validation. To avoid evaluating on closely related genomes, we performed grouped 5-fold cross-validation by species, assigning all assemblies of the same species to the same fold. The same folds were reused across models/baselines/controls.

Classification models. We trained one binary classifier per EC label:

1. Logistic regression (LR): L2-regularized logistic regression with class-balanced loss. Features were standardized using training-fold mean and standard deviation, with the same transform applied to test folds.
2. Extra Trees (ET): tree-based classifier with 300 trees per label

Baselines and permutation control. We included baseline models and permutation controls as following: (i) majority baseline: predicted an EC as present for all test assemblies if its training prevalence ≥0.5; (ii) random-prevalence baseline: predicted presence for each EC by sampling Bernoulli(p) with p equal to the label's training prevalence; and (iii) label-permutation control: independently permuted each EC across training samples before model training (preserving the prevalence while destroying feature label connections), and evaluated the predictions against true test labels.

Evaluation metrics. We evaluated the predicted presence/absence for each function against its ground truth on test folds. We define true positives (TP) as EC labels correctly predicted as present, false positives (FP) as EC labels predicted present but absent in the ground truth, and false negatives (FN) as EC labels present in the ground truth but predicted absent. Precision, recall, and F1 for each function were computed as:

(Eqn. 1)

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2TP}{2TP+FP+FN}$$

We additionally reported Hamming loss, the fraction of incorrectly predicted labels across all samples and labels within a test fold:

(Eqn. 2) $$HL = \frac{1}{n \cdot L} \sum_{i=1}^{n} \sum_{l=1}^{L} \mathbb{1}(y_{i,l} \neq \hat{y}_{il}),$$

where n is the number of test samples, L is the number of EC labels, $y_{i,l} \in \{0|1\}$ is the ground-truth presence/absence for label l in sample i, and $\hat{y}_{il} \in \{0|1\}$ is the corresponding prediction.

embeRNA: predicting functions from 16S rRNA k-mer

Next, we developed embeRNA framework to inform functions from 16S rRNA k-mer using the *embeRNA set*.

16S rRNA region extraction and embeddings. From each full-length 16S rRNA sequence in the *embeRNA set*, we extracted commonly sequenced regions (V1-V2, V3-V4, V4, V4-V5, V6-V8, V7-V9, and V1-V9) using V-Xtractor[90] for training models with different region as input. Each region was converted to a 1to5-mer embedding as described above, and inherited its assembly's EC three-digit profile.

Model architecture and training. embeRNA is a shallow fully connected neural network that maps a 16S rRNA k-mer embedding to EC probabilities. The architecture comprised two hidden layers (512 units each; ReLU activations) and a final linear layer producing logits for each EC label. Models were implemented in PyTorch[91] (v2.1.2) and trained with Adam optimizer (learning rate 10^{-3}; batch size 1,024) to minimize binary cross-entropy on logits (independently per label). At inference, logits were converted to probabilities using the sigmoid function.

Taxonomy-stratified train/validation/test splits. To evaluate generalization to unseen taxa, we performed taxonomy-stratified splits at a user selected rank (including assembly, species, genus, family, order, class, phylum). Assemblies were grouped by their lineage at that rank and randomly partitioned into train/validation/test (80/10/10); all 16S rRNA sequences of given region from assemblies in a group were restricted to the same split. Unless noted, results use genus-level stratification.

Model selection and binarization threshold. We trained multiple (default: 10) replicate models using different random seeds for train/validation/test split. For each replicate, the checkpoint with the lowest validation loss was selected. On each replicate's test set, we computed precision-recall curves (on all predictions across labels and samples) and the corresponding PR-AUC.

We built the final embeRNA model by initializing from the best-performing replicate and retraining on all available 16S rRNA-function pairs in the *embeRNA set*. We used the median of threshold values where each of the ten replicate models achieved their best F1 score as the final threshold (t_m; **SOM Table 1**) to convert embeRNA predicted values to binary (>t_m as presence, otherwise absence).

The entire front-to-end process of training and validating procedure, test-set evaluation, and final model selection of an embeRNA model for selected 16S rRNA region (e.g., V6-V8), at given taxonomy split (e.g., genus), can be found in **SOM Note 3**.

embeRNA-Fusion models. To generalize beyond EC space, we trained analogous embeRNA models to predict Fusion function from 16S rRNA 1to5-mers, using the same workflow. To reduce sparsity, the Fusion output space was restricted to functions present in ≥10% of Fusion assemblies).

Evaluation of embeRNA on Novel Microbes set

On the *Novel Microbes set* (sequences with <97% similarity to any 16S rRNA in *embeRNA set*), we evaluated how well embeRNA can inform EC three-digit functions if the V3-V4 region of these novel 16S rRNA sequences were captured.

Compared 16S rRNA-based prediction strategies. Using V3-V4 as input, we compared:

- embeRNA: We applied the finalized V3-V4 genus-split model to predict functions for each sequence, binarizing the predicted values with t_m **(SOM Table 1)**.
- PICRUSt2: We customized the PICRUSt2 reference assembly function profiles by: (i) updating the four-digit Enzyme Commission (EC) profiles using the latest annotations from the JGI IMG database (successfully updating 19,717 of the 20,000 references); and (ii) generating a three-digit EC presence/absence profile from these updated four-digit data to serve as the PICRUSt2 reference function

table. Each V3-V4 sequence was then placed into the reference tree by PICRUSt2, and functions were predicted using these customized profiles.

- Taxonomy-to-core approaches: for each sequence, we (i) first assigned taxonomy using Kraken2[39] and RDP Classifier[59] separately, then (ii) predicted the "core" set of EC three-digit functions present in all *embeRNA set* genomes within the assigned group; if an assigned taxon is absent from *embeRNA set*, back off to higher ranks; if no assignment/match exists, return all-zero.

Evaluation. Predictions across all functions and tested samples were compared to ground-truth EC profiles of *Novel Microbes set* restricted to the intersection of functional spaces supported by all tools, using precision, recall, and F1 (**Eqn. 1**).

We next compared the two best-performing tools, embeRNA and PICRUSt2, on the hard to predict functions, i.e. those for which the two approaches made opposite predictions. Using the true function presence/absence as the reference, we calculated the proportion of correct predictions for each tool across three scenarios: (i) all disagreement cases, (ii) disagreement cases in which the function was truly present, and (iii) disagreement cases in which the function was truly absent. As a baseline, we generated random presence/absence predictions by Bernoulli sampling with probability equal to the observed prevalence of truly present functions in the disagreement set ($p = 0.44$), and evaluated this baseline across the same three scenarios.

Evaluation of embeRNA on real world microbiome data (*Microbiome set*)

On a set of 22 soil samples with both whole metagenome shotgun sequencing (WMS) and V6-V8 16S rRNA amplicon sequencing data, we further evaluated embeRNA's EC three-digit function prediction against those informed from WMS by HUMAnN 3[44].

Predicting functions from 16S rRNA using embeRNA. We first processed the input sequences by (i) merging paired-end reads using bbmerge[92] (v39.06), (ii) extracting the exact V6-V8 regions using V-Xtractor (v2.0), and (iii) embedding reads to 1to5-mers. The 1to5-mers were then passed through the finalized V6-V8 embeRNA model for function prediction. Per-16S EC probabilities were binarized using t_m **(SOM Table 1)**. Presence for each function were then summed through all 16S sequences to obtain within sample function counts / abundance.

Predicting functions from 16S rRNA using PICRUSt2. We first generated amplicon sequencing variants (ASV)s and an ASV abundance table from the V6-V8 data using DADA2[18] (v3.11). The resulting ASVs were then processed with PICRUSt2 to predict functions, using the customized EC three-digit reference function table described previously. Notably, when constructing sample-level function pseudo-abundance profiles, we omitted PICRUSt2's default normalization by 16S rRNA gene copy number so that the resulting profiles would be directly comparable to those generated by embeRNA.

Predicting functions from WMS using HUMAnN 3. We quality-filtered WMS reads with fastp[93] (v0.23.4; parameters: -e 25 -l 70) and removed host contamination by (i) aligning reads to the human reference genome (GRCh38[94]) with BWA-MEM[95] (v0.7.17), and (ii) processing alignments with SAMtools[96] (v1.19.2). We then used HUMAnN 3[44] (v3.9) with the UniRef50 database[97] to infer EC four-digit function profiles, which we summarized to EC three-digit profiles by summing abundances across all EC four-digit entries within each EC three-digit category.

Comparison to WMS profiles. We compared the per-sample function categories predicted by embeRNA and PICRUSt2 with those inferred from HUMAnN 3. For each 16S rRNA-based method, we quantified (i) the proportion of WMS-inferred function categories recovered in each sample and (ii) the number and proportion of additional function categories predicted by the method but not detected by HUMAnN 3.

For abundance-based comparisons, we converted each sample's function counts to relative abundances by normalizing to the total function count in that sample. We then calculated Spearman correlations between each 16S rRNA-based function profile and its corresponding WMS-based profile.

Comparison of discordant function abundance bins between embeRNA and PICRUSt2. To assess differences in abundance estimation between embeRNA and PICRUSt2, we compared functions whose predicted abundance levels differed between the two tools within the same sample, and evaluated those discordant functions against the corresponding HUMAnN 3-derived abundance levels. For each sample and method (embeRNA, PICRUSt2, and HUMAnN 3), we converted function abundances to relative abundances and assigned functions to ten decile bins based on within-sample abundance

rank. We then identified functions with substantial disagreement between embeRNA and PICRUSt2, defined as a difference of at least three abundance bins, and evaluated which tool assigned the abundance bin closer to that of the corresponding HUMAnN 3-derived profile. For each sample, we summarized the proportion of discordant functions for which each 16S rRNA-based tool assigned an abundance bin closer to the corresponding HUMAnN 3 bin.

Statistics.

Where applicable, we estimated the 95% confidence intervals (CIs) for observed values via bootstrapping, using 80% subsamples of the data without replacement over 1000 iterations.

Data availability.

Assembly accessions for the *Balanced set*, assembly accessions and 16S rRNA identifiers for the *Novel Microbes set*, EC abundance percentile rank plots between embeRNA and PICRUSt2 for all soil samples, together with the lists of embeRNA predicted Fusion functions significantly different between soil compartments (rhizosphere vs. bulk), are available on Figshare (DOI: 10.6084/m9.figshare.31297531). Blueberry soil microbiome sequencing data are available from the NCBI Sequence Read Archive under BioProjects PRJNA434066 (16S rRNA amplicon) and PRJNA484230 (whole-metagenome shotgun)[72].

Software availability.

The embeRNA source code is available on Bitbucket at https://bitbucket.org/bromberglab/emberna.

Acknowledgements

We are grateful to R. Prabakaran for all helpful discussions and to Sergio Gramacho and Aiden Maloney-Bertelli for technical support (all Emory University). We also thank all those who deposit their experimental results into public databases and those who maintain the databases. The work of Y.B. and J.L. was supported by National Institute of Dental and Craniofacial Research, R01-DE032216 and the National Science Foundation award #2310114.

References

1. Schmartz GP, Rehner J, Schuff MJ, Molano LG, Becker SL, Krawczyk M, Tagirdzhanov A, Gurevich A, Francke R, Muller R, Keller V, Keller A. Exploring microbial diversity and biosynthetic potential in zoo and wildlife animal microbiomes. Nat Commun. 2024;15(1):8263. Epub 20240926. doi: 10.1038/s41467-024-52669-9. PubMed PMID: 39327429; PMCID: PMC11427580.

2. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vazquez-Baeza Y, Gonzalez A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project C. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457-63. Epub 20171101. doi: 10.1038/nature24621. PubMed PMID: 29088705; PMCID: PMC6192678.

3. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804-10. doi: 10.1038/nature06244. PubMed PMID: 17943116; PMCID: PMC3709439.

4. Henry LP, Bruijning M, Forsberg SKG, Ayroles JF. The microbiome extends host evolutionary potential. Nat Commun. 2021;12(1):5141. Epub 20210826. doi: 10.1038/s41467-021-25315-x. PubMed PMID: 34446709; PMCID: PMC8390463.

5. Li Z, Riley WJ, Marschmann GL, Karaoz U, Shirley IA, Wu Q, Bouskill NJ, Chang KY, Crill PM, Grant RF, King E, Saleska SR, Sullivan MB, Tang J, Varner RK, Woodcroft BJ, Wrighton KC, Coordinators EBII, Brodie EL. A framework for integrating genomics, microbial traits, and ecosystem biogeochemistry. Nat Commun. 2025;16(1):2186. Epub 20250304. doi: 10.1038/s41467-025-57386-5. PubMed PMID: 40038282; PMCID: PMC11880341.

6. Zhao M, Xue K, Wang F, Liu S, Bai S, Sun B, Zhou J, Yang Y. Microbial mediation of biogeochemical cycles revealed by simulation of global changes with soil transplant and cropping. ISME J. 2014;8(10):2045-55. Epub 20140403. doi: 10.1038/ismej.2014.46. PubMed PMID: 24694714; PMCID: PMC4184012.

7. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377-86. Epub 20070125. doi: 10.1101/gr.5969107. PubMed PMID: 17255551; PMCID: PMC1800929.

8. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261-7. Epub 20070622. doi: 10.1128/AEM.00062-07. PubMed PMID: 17586664; PMCID: PMC1950982.

9. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386. Epub 20080919. doi: 10.1186/1471-2105-9-386. PubMed PMID: 18803844; PMCID: PMC2563014.

10. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF.

Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537-41. Epub 20091002. doi: 10.1128/AEM.01541-09. PubMed PMID: 19801464; PMCID: PMC2786419.

11. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013;29(18):2253-60. Epub 20130704. doi: 10.1093/bioinformatics/btt389. PubMed PMID: 23828782; PMCID: PMC3753567.

12. Konwar KM, Hanson NW, Page AP, Hallam SJ. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. BMC Bioinformatics. 2013;14:202. Epub 20130621. doi: 10.1186/1471-2105-14-202. PubMed PMID: 23800136; PMCID: PMC3695837.

13. Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243. Epub 20140109. doi: 10.7717/peerj.243. PubMed PMID: 24482762; PMCID: PMC3897386.

14. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2:33. Epub 20140905. doi: 10.1186/2049-2618-2-33. PubMed PMID: 25225611; PMCID: PMC4164323.

15. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. Epub 20141117. doi: 10.1038/nmeth.3176. PubMed PMID: 25402007.

16. Freitas TA, Li PE, Scholz MB, Chain PS. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015;43(10):e69. Epub 20150312. doi: 10.1093/nar/gkv180. PubMed PMID: 25765641; PMCID: PMC4446416.

17. Minot SS, Krumm N, Greenfield NB. One codex: a sensitive and accurate data platform for genomic microbial identification. BioRxiv. 2015:027607.

18. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581-3. Epub 20160523. doi: 10.1038/nmeth.3869. PubMed PMID: 27214047; PMCID: PMC4927377.

19. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721-9. Epub 20161017. doi: 10.1101/gr.210641.116. PubMed PMID: 27852649; PMCID: PMC5131823.

20. Koslicki D, Falush D. MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. mSystems. 2016;1(3). Epub 20160607. doi: 10.1128/mSystems.00020-16. PubMed PMID: 27822531; PMCID: PMC5069763.

21. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257. Epub 20160413. doi: 10.1038/ncomms11257. PubMed PMID: 27071849; PMCID: PMC4833860.

22. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132. Epub 20160620. doi: 10.1186/s13059-016-0997-x. PubMed PMID: 27323842; PMCID: PMC4915045.

23. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32(24):3823-5. Epub 20160818. doi: 10.1093/bioinformatics/btw542. PubMed PMID: 27540266.

24. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. Epub 20161018. doi: 10.7717/peerj.2584. PubMed PMID: 27781170; PMCID: PMC5075697.

25. Silva GG, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics. 2016;32(3):354-61. Epub 20151009. doi: 10.1093/bioinformatics/btv584. PubMed PMID: 26454280; PMCID: PMC4734042.

26. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3. Epub 20170102. doi: 10.7717/peerj-cs.104. PubMed PMID: 40271438; PMCID: PMC12016282.

27. Matias Rodrigues JF, Schmidt TSB, Tackmann J, von Mering C. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics. 2017;33(23):3808-10. doi: 10.1093/bioinformatics/btx517. PubMed PMID: 28961926; PMCID: PMC5860325.

28. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-8. Epub 20171016. doi: 10.1038/nbt.3988. PubMed PMID: 29035372.

29. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6(1):23. Epub 20180201. doi: 10.1186/s40168-018-0401-z. PubMed PMID: 29391044; PMCID: PMC5796597.

30. Fiannaca A, La Paglia L, La Rosa M, Lo Bosco G, Renda G, Rizzo R, Gaglio S, Urso A. Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinformatics. 2018;19(Suppl 7):198. Epub 20180709. doi: 10.1186/s12859-018-2182-6. PubMed PMID: 30066629; PMCID: PMC6069770.

31. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100. doi: 10.1093/bioinformatics/bty191. PubMed PMID: 29750242; PMCID: PMC6137996.

32. Zhu C, Miller M, Marpaka S, Vaysberg P, Ruhlemann MC, Wu G, Heinsen FA, Tempel M, Zhao L, Lieb W, Franke A, Bromberg Y. Functional sequencing read annotation for high precision microbiome analysis. Nucleic Acids Res. 2018;46(4):e23. doi: 10.1093/nar/gkx1209. PubMed PMID: 29194524; PMCID: PMC5829635.

33. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodriguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ,

Vargas F, Vazquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852-7. doi: 10.1038/s41587-019-0209-9. PubMed PMID: 31341288; PMCID: PMC7015180.

34. Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun. 2019;10(1):3066. Epub 20190711. doi: 10.1038/s41467-019-10934-2. PubMed PMID: 31296857; PMCID: PMC6624308.

35. Georgiou A, Fortuin V, Mustafa H, Rätsch G. Meta2: Memory-efficient taxonomic classification and abundance estimation for metagenomics with deep learning (2019). URL https://arxiv org/abs. 2019.

36. Menegaux R, Vert JP. Continuous Embeddings of DNA Sequencing Reads and Application to Metagenomics. J Comput Biol. 2019;26(6):509-18. Epub 20190219. doi: 10.1089/cmb.2018.0174. PubMed PMID: 30785347.

37. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10(1):1014. Epub 20190304. doi: 10.1038/s41467-019-08844-4. PubMed PMID: 30833550; PMCID: PMC6399450.

38. Rojas-Carulla M, Tolstikhin I, Luque G, Youngblut N, Ley R, Schölkopf B. Genet: Deep representations for metagenomics. arXiv preprint arXiv:190111015. 2019.

39. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. Epub 20191128. doi: 10.1186/s13059-019-1891-0. PubMed PMID: 31779668; PMCID: PMC6883579.

40. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38(6):685-8. doi: 10.1038/s41587-020-0548-6. PubMed PMID: 32483366; PMCID: PMC7365738.

41. Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genom Bioinform. 2020;2(1):lqaa009. Epub 20200219. doi: 10.1093/nargab/lqaa009. PubMed PMID: 33575556; PMCID: PMC7671387.

42. Tovo A, Menzel P, Krogh A, Cosentino Lagomarsino M, Suweis S. Taxonomic classification method for metagenomics based on core protein families with Core-Kaiju. Nucleic Acids Res. 2020;48(16):e93. doi: 10.1093/nar/gkaa568. PubMed PMID: 32633756; PMCID: PMC7498351.

43. Wemheuer F, Taylor JA, Daniel R, Johnston E, Meinicke P, Thomas T, Wemheuer B. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. Environ Microbiome. 2020;15(1):11. Epub 20200518. doi: 10.1186/s40793-020-00358-7. PubMed PMID: 33902725; PMCID: PMC8067651.

44. Beghini F, McIver LJ, Blanco-Miguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife. 2021;10. Epub 20210504. doi: 10.7554/eLife.65088. PubMed PMID: 33944776; PMCID: PMC8096432.

45. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. Structure-based

protein function prediction using graph convolutional networks. Nat Commun. 2021;12(1):3168. Epub 20210526. doi: 10.1038/s41467-021-23303-9. PubMed PMID: 34039967; PMCID: PMC8155034.

46. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. mSystems. 2021;6(3):101128msystems0058321. Epub 20210629. doi: 10.1128/mSystems.00583-21. PubMed PMID: 34184911; PMCID: PMC8269244.

47. Hoarfrost A, Aptekmann A, Farfanuk G, Bromberg Y. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. Nat Commun. 2022;13(1):2606. Epub 20220511. doi: 10.1038/s41467-022-30070-8. PubMed PMID: 35545619; PMCID: PMC9095714.

48. Mock F, Kretschmer F, Kriese A, Bocker S, Marz M. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proc Natl Acad Sci U S A. 2022;119(35):e2122636119. Epub 20220826. doi: 10.1073/pnas.2122636119. PubMed PMID: 36018838; PMCID: PMC9436379.

49. Morais DAA, Cavalcante JVF, Monteiro SS, Pasquali MAB, Dalmolin RJS. MEDUSA: A Pipeline for Sensitive Taxonomic Classification and Flexible Functional Annotation of Metagenomic Shotgun Sequences. Front Genet. 2022;13:814437. Epub 20220307. doi: 10.3389/fgene.2022.814437. PubMed PMID: 35330728; PMCID: PMC8940201.

50. Blanco-Miguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, Nickols WA, Piccinno G, Piperni E, Puncochar M, Valles-Colomer M, Tett A, Giordano F, Davies R, Wolf J, Berry SE, Spector TD, Franzosa EA, Pasolli E, Asnicar F, Huttenhower C, Segata N. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41(11):1633-44. Epub 20230223. doi: 10.1038/s41587-023-01688-w. PubMed PMID: 36823356; PMCID: PMC10635831.

51. Schriefer AE, Kumar B, Zolty A, Didier A, Nirmal M, Greeshma G, Nadiv N, Perez M, Preetam R, Mahankuda SK. M-CAMPTM: a cloud-based web platform with a novel approach for species-level classification of 16S rRNA microbiome sequences. Current Bioinformatics. 2023;18(1):21-39.

52. Wichmann A, Buschong E, Muller A, Junger D, Hildebrandt A, Hankeln T, Schmidt B. MetaTransformer: deep metagenomic sequencing read classification using self-attention models. NAR Genom Bioinform. 2023;5(3):lqad082. Epub 20230911. doi: 10.1093/nargab/lqad082. PubMed PMID: 37705831; PMCID: PMC10495543.

53. Irber L, Pierce-Ward NT, Abuelanin M, Alexander H, Anant A, Barve K, Baumler C, Botvinnik O, Brooks P, Dsouza D. sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. Journal of Open Source Software. 2024;9(98):6830.

54. Tawfiq R, Niu K, Hoehndorf R, Kulmanov M. DeepGOMeta for functional insights into microbial communities using deep learning-based protein function prediction. Sci Rep. 2024;14(1):31813. Epub 20241230. doi: 10.1038/s41598-024-82956-w. PubMed PMID: 39738309; PMCID: PMC11685674.

55. Prabakaran R, Bromberg Y. Deciphering enzymatic potential in metagenomic reads through DNA language models. Nucleic Acids Res. 2025;53(16). doi: 10.1093/nar/gkaf836. PubMed PMID: 40901999; PMCID: PMC12407101.

56. Bagci C, Negri T, Buena-Atienza E, Gross C, Ossowski S, Ziemert N. Ultra-deep long-read metagenomics captures diverse taxonomic and biosynthetic potential of soil microbes. Gigascience. 2025;14. doi: 10.1093/gigascience/giaf135. PubMed PMID: 41133998; PMCID: PMC12690461.

57. Liu J, Coker MO, Osazuwa-Peters N, Peter O, Idemudia NL, Schlecht NF, Obuekwe O, Eki-Udoko FE, Bromberg Y. Whole Metagenome Sequencing: not Deep Enough for Complete Microbial Function Recovery. bioRxiv. 2025. Epub 20251106. doi: 10.1101/2025.11.04.685665. PubMed PMID: 41278862; PMCID: PMC12637550.

58. Zielińska K, Pantiukh K, Łabaj PP, Kosciolek T, Org E. A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenges. bioRxiv. 2025:2025.07.06.662369. doi: 10.1101/2025.07.06.662369.

59. Wang Q, Cole JR. Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. Microbiol Resour Announc. 2024;13(4):e0106323. Epub 20240304. doi: 10.1128/mra.01063-23. PubMed PMID: 38436268; PMCID: PMC11008197.

60. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72. doi: 10.1128/AEM.03006-05. PubMed PMID: 16820507; PMCID: PMC1489311.

61. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6. Epub 20121128. doi: 10.1093/nar/gks1219. PubMed PMID: 23193283; PMCID: PMC3531112.

62. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814-21. Epub 20130825. doi: 10.1038/nbt.2676. PubMed PMID: 23975157; PMCID: PMC3819121.

63. Asshauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 2015;31(17):2882-4. Epub 20150507. doi: 10.1093/bioinformatics/btv287. PubMed PMID: 25957349; PMCID: PMC4547618.

64. Aurora R, Sanford T. The Microbiome: From the Beginning to the End. Mo Med. 2024;121(4):310-6. PubMed PMID: 39575080; PMCID: PMC11578570.

65. Salvadori M, Rosso G. Update on the gut microbiome in health and diseases. World J Methodol. 2024;14(1):89196. Epub 20240320. doi: 10.5662/wjm.v14.i1.89196. PubMed PMID: 38577200; PMCID: PMC10989414.

66. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995;11(7):283-90. doi: 10.1016/s0168-9525(00)89076-9. PubMed PMID: 7482779.

67. Karlin S, Mrazek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 1997;179(12):3899-913. doi: 10.1128/jb.179.12.3899-3913.1997. PubMed PMID: 9190805; PMCID: PMC179198.

68. Campbell A, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci U S A. 1999;96(16):9184-9. doi: 10.1073/pnas.96.16.9184. PubMed PMID: 10430917; PMCID: PMC17754.

69. Bize A, Midoux C, Mariadassou M, Schbath S, Forterre P, Da Cunha V. Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. BMC Genomics. 2021;22(1):186. Epub 20210316. doi: 10.1186/s12864-021-07471-y. PubMed PMID: 33726663; PMCID: PMC7962313.

70. Asgari E, Garakani K, McHardy AC, Mofrad MRK. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. Bioinformatics. 2019;35(6):1082. doi: 10.1093/bioinformatics/bty652. PubMed PMID: 30099528; PMCID: PMC6419898.

71. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol. 2015;6:771. Epub 20150804. doi: 10.3389/fmicb.2015.00771. PubMed PMID: 26300854; PMCID: PMC4523815.

72. Yurgel SN, Nearing JT, Douglas GM, Langille MGI. Metagenomic Functional Shifts to Plant Induced Environmental Changes. Front Microbiol. 2019;10:1682. Epub 20190726. doi: 10.3389/fmicb.2019.01682. PubMed PMID: 31404278; PMCID: PMC6676915.

73. Mahlich Y, Zhu C, Chung H, Velaga PK, De Paolis Kaluza MC, Radivojac P, Friedberg I, Bromberg Y. Learning from the unknown: exploring the range of bacterial functionality. Nucleic Acids Res. 2023;51(19):10162-75. doi: 10.1093/nar/gkad757. PubMed PMID: 37739408; PMCID: PMC10602916.

74. Allison SD, Martiny JB. Colloquium paper: resistance, resilience, and redundancy in microbial communities. Proc Natl Acad Sci U S A. 2008;105 Suppl 1(Suppl 1):11512-9. Epub 20080811. doi: 10.1073/pnas.0801925105. PubMed PMID: 18695234; PMCID: PMC2556421.

75. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 2013;7(4):830-8. Epub 20121213. doi: 10.1038/ismej.2012.160. PubMed PMID: 23235290; PMCID: PMC3603392.

76. Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107(43):18634-9. Epub 20101011. doi: 10.1073/pnas.1009480107. PubMed PMID: 20937887; PMCID: PMC2972931.

77. Zaneveld JR, Parfrey LW, Van Treuren W, Lozupone C, Clemente JC, Knights D, Stombaugh J, Kuczynski J, Knight R. Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation. Trends Microbiol. 2011;19(10):472-82. Epub 20110825. doi: 10.1016/j.tim.2011.07.006. PubMed PMID: 21872475; PMCID: PMC3184378.

78. Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO Rep. 2005;6(12):1208-13. doi: 10.1038/sj.embor.7400538. PubMed PMID: 16200051; PMCID: PMC1369203.

79. Arias PM, Butler J, Randhawa GS, Soltysiak MPM, Hill KA, Kari L. Environment and taxonomy shape the genomic signature of prokaryotic extremophiles. Sci Rep. 2023;13(1):16105. Epub 20230926. doi: 10.1038/s41598-023-42518-y. PubMed PMID: 37752120; PMCID: PMC10522608.

80. Kellom M, Berg M, Chen IM, Chu K, Clum A, Huntemann M, Ivanova NN, Kyrpides NC, Mukherjee S, Reddy TBK, Roux S, Seshadri R, Szabo G, Varghese NJ, Woyke T, Eloe-Fadrosh EA. Tetranucleotide frequencies differentiate genomic boundaries and metabolic strategies across environmental microbiomes. mSystems. 2025;10(8):e0174424. Epub 20250708. doi: 10.1128/msystems.01744-24. PubMed PMID: 40626735; PMCID: PMC12363243.

81. Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, Anantharaman K, Lane KR, Thomas BC, Pan C, Northen TR, Banfield JF. Mediterranean grassland soil C-N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. Nat Microbiol. 2019;4(8):1356-67. Epub 20190520. doi: 10.1038/s41564-019-0449-y. PubMed PMID: 31110364; PMCID: PMC6784897.

82. Masuda Y, Mise K, Xu Z, Zhang Z, Shiratori Y, Senoo K, Itoh H. Global soil metagenomics reveals distribution and predominance of Deltaproteobacteria in nitrogen-fixing microbiome. Microbiome. 2024;12(1):95. Epub 20240524. doi: 10.1186/s40168-024-01812-1. PubMed PMID: 38790049; PMCID: PMC11127431.

83. Bhattacharyya A, Mavrodi O, Bhowmik N, Weller D, Thomashow L, Mavrodi D. Bacterial biofilms as an essential component of rhizosphere plant-microbe interactions. Methods Microbiol. 2023;53:3-48. Epub 20230622. doi: 10.1016/bs.mim.2023.05.006. PubMed PMID: 38415193; PMCID: PMC10898258.

84. Yang L, Qian X, Zhao Z, Wang Y, Ding G, Xing X. Mechanisms of rhizosphere plant-microbe interactions: molecular insights into microbial colonization. Front Plant Sci. 2024;15:1491495. Epub 20241106. doi: 10.3389/fpls.2024.1491495. PubMed PMID: 39606666; PMCID: PMC11600982.

85. Upadhyay SK, Srivastava AK, Rajput VD, Chauhan PK, Bhojiya AA, Jain D, Chaubey G, Dwivedi P, Sharma B, Minkina T. Root Exudates: Mechanistic Insight of Plant Growth Promoting Rhizobacteria for Sustainable Crop Production. Front Microbiol. 2022;13:916488. Epub 20220714. doi: 10.3389/fmicb.2022.916488. PubMed PMID: 35910633; PMCID: PMC9329127.

86. Yang CX, Chen SJ, Hong XY, Wang LZ, Wu HM, Tang YY, Gao YY, Hao GF. Plant exudates-driven microbiome recruitment and assembly facilitates plant health management. FEMS Microbiol Rev. 2025;49. doi: 10.1093/femsre/fuaf008. PubMed PMID: 40158196; PMCID: PMC12007450.

87. Biology IUoBaM. Enzyme Nomenclature: Academic Press, San Diego, California; 1992.

88. Chen IA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter SJ, Webb C, Wu D, Varghese NJ, Reddy TBK, Mukherjee S, Ovchinnikova G, Nolan M, Seshadri R, Roux S, Visel A, Woyke T, Eloe-Fadrosh EA, Kyrpides NC, Ivanova NN. The IMG/M data management and analysis system v.7: content updates and new features. Nucleic Acids Res. 2023;51(D1):D723-D32. doi: 10.1093/nar/gkac976. PubMed PMID: 36382399; PMCID: PMC9825475.

89. McDonald AG, Boyce S, Tipton KF. ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res. 2009;37(Database issue):D593-7. Epub 20080906. doi: 10.1093/nar/gkn582. PubMed PMID: 18776214; PMCID: PMC2686581.

90. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. J Microbiol Methods. 2010;83(2):250-3. Epub 20100827. doi: 10.1016/j.mimet.2010.08.008. PubMed PMID: 20804791.

91. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin ZM, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai JJ, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neur In. 2019;32. PubMed PMID: WOS:000534424308009.

92. Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PLoS One. 2017;12(10):e0185056. Epub 20171026. doi: 10.1371/journal.pone.0185056. PubMed PMID: 29073143; PMCID: PMC5657622.

93. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90. doi: 10.1093/bioinformatics/bty560. PubMed PMID: 30423086; PMCID: PMC6129281.

94. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood JM, Clarke L, Koren S, Boitano M, Peluso P, Li H, Chin CS, Phillippy AM, Durbin R, Wilson RK, Flicek P, Eichler EE, Church DM. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849-64. Epub 20170410. doi: 10.1101/gr.213611.116. PubMed PMID: 28396521; PMCID: PMC5411779.

95. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM2013. doi: https://doi.org/10.48550/arXiv.1303.3997.

96. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2). doi: 10.1093/gigascience/giab008. PubMed PMID: 33590861; PMCID: PMC7931819.

97. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926-32. Epub 20141113. doi: 10.1093/bioinformatics/btu739. PubMed PMID: 25398609; PMCID: PMC4375400.