Benchmarking DNA Foundation Models on Binning Human Gut Microbial Strains

Anders Havbro Hjulmand, Eisuke Okuda, Andreas Flensted Olsen
{ahju,eiok,frao}@itu.dk

Supervisor: Veronika Cheplygina

***Abstract*—Genomic DNA sequences encode vast amounts of information that governs the function of biological processes. Decoding the "language" of DNA is essential in understanding the affiliated organisms and their functions.**

Metagenomics is an advancing field of research that studies the interactions between thousands of microorganisms in an environment. One central environment is the human gut microbiome, where diseases have been associated with specific organisms. Metagenomic binning is a key approach in identifying present organisms in such environments.

The success of Language Models in NLP has inspired the development of DNA foundation models that learn general-purpose representations of DNA. Previous studies have benchmarked DNA foundation models on metagenomics binning using data from marine and plant environments.

In this study, we benchmarked six DNA foundation models on metagenomics binning using data from the human gut microbiome at strain resolution. To conduct a systematic model comparison, we used the same clustering algorithm and a distinct model-calibrated similarity threshold. Additionally, the strain lineages were utilized to conduct a genus level clustering analysis.

The models generally struggled to construct strain level clusters, but performed better at genus level resolution. At strain resolution, only one DNA foundation model performed better than the baselines. The suboptimal performance of DNA foundation models may be attributed to their input length constraints and their pre-training objectives.

The code and data is available at github.com/eisuke119/Research-Project.

***Index Terms*—DNA foundation models, Metagenome, Metagenomics binning**

I. Introduction

FOUNDATION models in Natural Language Processing (NLP) are characterized by their ability to learn general-purpose representations of text, through self-supervised training on massive datasets, that can be applied to a wide range of downstream tasks. The success of foundation models like GPT-4 [1], LLaMA [2], and BERT [3] has inspired the application of similar ideas to other domains such as protein language models used for protein generation and protein folding prediction [4], [5].

Next-generation sequencing (NGS) has emerged in the past two decades as a fast and inexpensive technology to sequence whole genome DNA, leading to an extraordinary amount of unlabeled DNA sequences [6]. This abundance of data combined with a lack of annotation, has recently inspired the development of DNA foundation models to learn general-purpose representations of DNA. These models are pre-trained on large genomic datasets such as the human reference genome which contains $3,2$ billion pairs of DNA nucleotides (base-pairs) [7]. The goal of these models is to capture complex relationships between base-pairs, useful for decoding the biological information in DNA sequences such as gene expressions.

A large number of microorganisms such as bacteria, virus, and fungi inhabit a wide range of environments, including the human gut. The combination of the microorganisms and their environment is referred to as the microbiome, and the complete genetic material of the organisms in the environment is referenced as the metagenome [8].

Advances in NGS, particularly shotgun sequencing, has made it possible to characterize the entire metagenome within an environmental sample. This technique produces millions of unlabeled and overlapping reads each being a fragment of the DNA in the sample. Through a process called assembly, the set of reads are reduced to a smaller number of contiguous sequences known as contigs [9].

Metagenomics binning involves reconstructing genomes from the contigs into metagenomic assembled genomes (MAGs). Contigs are grouped by features such as tetranucleotide frequencies (TNF). Metagenomics binning is essentially a clustering problem, where the goal is to identify clusters of contigs, such that each cluster corresponds to a genome.

Metagenomics binning is an important step in metagenomic studies because it reduces the complexity contained in hundred-thousands of contigs into often hundreds of microbial bins. These bins are then taxonomically profiled by aligning the contigs to known reference genomes, to identify present organisms in a sample [9]–[12].

Previous studies have benchmarked DNA foundation models across various biological tasks showcasing state-of-the-art performance [13]–[16]. The foundation models have similarly been benchmarked in metagenomics binning using the CAMI2 dataset, containing samples from marine and plant environments [17], [18]. However, the models have not yet been evaluated on metagenomics binning using data from the human gut microbiome. Exploring this gap is crucial, as the human gut microbiome and its connection to health and disease is a more recent and advancing field of research [19].

Specifically, metagenomics binning can help identify associations between host phenotypes and microbial organisms. Recent studies focusing on profiling the human gut microbiome from different cohorts also reveal novel species, genera and phyla [20]. Other studies also highlighted the importance

of identifying species at strain level resolution (sub-species), since strains from the same species can be beneficial, benign or harmful to a microbial environment. This is vital in the analysis of the human gut microbiome where some particular strains are pathogens related to diseases such as colectoral cancer, diabetes, and inflammatory bowel disease [21]. With the large increase in unannotated metagenomics data, DNA-foundation models could help to identify previously unknown correlations between organisms through metagenomics binning.

In this study, we benchmark six DNA foundation models along with three baseline models on metagenomics binning using strain resolution data from the human gut microbiome. Additionally, we perform a clustering analysis on a higher taxonomic level than the original data, thereby exploring the models' capabilities to distinguish between genomes that are less biologically related. Our contribution is summarized as follows:

- **A systematic evaluation of performance across models** using the same clustering algorithm and a distinct similarity threshold calibrated to each model.
- **A framework for clustering on higher taxonomic levels** using the Hausdorff distance, a metric commonly applied in medical image segmentation [22].
- **Result:** One DNA foundation model outperforms the baselines [17], whilst the remaining perform worse than the baselines.
- **Result:** The models generally struggled to construct strain level clusters, but performed better at genus level resolution.

II. Background and Related Work

A. Deep learning in metagenomics

DNA sequence analysis in metagenomics relies on aligning assembled contigs to reference databases to indicate presence or absence of organisms within the sample environment [10]. With the advances in Deep Learning (DL), contigs have been analyzed using various architectures designed to solve specific genomic tasks.

Deep CNN architectures have been used to analyze local structures in DNA sequences. Wu. et. al. [23] uses a CNN architecture to classify virulent contigs using binary supervision. Other CNN architectures focus within specific regions of the contigs to locate key regulatory elements associated with specific phenotypes [24].

Other deep learning architectures have been used to globally analyze the microbiome by characterizing species and their relative abundances using metagenomics binning. Traditional methods rely on the descriptive non-learned features such as tetranucleotide frequencies and probabilistic distances of the contigs to known reference genomes to cluster contigs into MAGs [10]–[12], [25].

VAMB [9] uses a variational autoencoder to map the tetranucleotide frequencies and species abundances into a latent space that learns the binning and distance thresholds between contigs. VAMB and extensions of the autoencoder architecture have shown competitive results to state of the art metagenomic binners such as **MetaBat2** [25], [26].

B. Language modeling for DNA sequences

The advances in NLP has led to treating DNA as language consisting of 4 characters **'A', 'G', 'T', 'C'** where a contiguous DNA sequence can be considered a whole text or sentence depending on length and objective. Similar to **N-grams**, "words" can be constructed from contiguous k characters - called **k-mers**. With inspiration from Word2Vec [27], a key step was to learn static k-mer embeddings as done in **DNA2Vec** [28] and **Metagenome2vec** [29].

However, static embeddings fail to capture polysemantic relationships. K-mers can have varying meaning depending on their surrounding context resulting in polysemy. Such polysemy is especially prevalent in the non-coding regions of DNA, which play a key role in regulating gene expression [30].

To encode the inherent structures of DNA, attention mechanisms have been applied to learn contextualized k-mer embeddings. One of the early attempts in using attention was Deep Microbes' LSTM architecture [31], that showed promising performance in taxonomic classification, paving the way DNA foundation models using transformer architectures [30], [32]–[35].

A key strength of transformer models is the attention mechanism that attends to all tokens within a fixed context window and learn semantic relationships between tokens. The context window size is often constrained to be less than 2048, as the attention mechanism scales quadratically with the input size. This input constraint is too strict for DNA sequences of up to billions of base-pairs.

This poses a critical challenge for adopting transformer based architectures to building successful foundation models for DNA [36]. The same challenge in NLP has led to alternatives such as State Space Models [37], [38], Hyena [39], and Flash Attention [40] which bypass the quadratic complexity of the attention mechanism. These alternatives have become competitors to transformer based architectures in NLP tasks using much fewer parameters and training time. These architectures have also been deployed to DNA sequences to handler longer sequences up to 1M base-pairs [41], [42].

C. Related Benchmark Studies

Numerous DNA foundation models have been benchmarked, in both original studies and benchmark studies, on a range of different tasks e.g. read classification, gene function prediction, functional identification, and metagenomics binning [13]–[15], [17].

In the original studies, the authors all showcase state of the art performance of their contributions, but often leave out competing models in their comparisons, due to the specificity of the downstream tasks.

In the lack of a standard evaluation, DNABERT-2 [43] presented a benchmark dataset called Genome Understanding Evaluation (GUE), designed to benchmark DNA foundation models on multi-species genome classification [43]. The GUE benchmark dataset has so forth not been used elsewhere.

In BEND [13], DNA foundation models are benchmarked on a range of tasks such as gene finding, non-coding variant effects, and enhancer annotation on the human genome. The central result from the paper shows that none of the foundation models generally perform better than the others in all tasks. Additionally, the paper finds that the representation of some models focuses on local gene structures whilst others latches onto non-coding regions. Lastly, the paper finds that none of the foundation models have successfully captured the long range dependencies existing in the human genome. BEND is limited to perform its evaluation tasks on specific regions of the human genome.

Similar benchmark studies also highlights that no single model perform best on short and long range tasks across the human genome [14], [15], while another study finds that foundation models perform similar to conventional supervised machine learning methods [16].

In DNABERT-S [17], foundation models are benchmarked on metagenomics binning and species classification using data from marine and plant environments, as well as two synthetic datasets [18]. They find superior performance of their own species-aware model, DNABERT-S, compared to other foundation models and baselines.

DNA foundation models have not yet been benchmarked on data from the human gut microbiome. This paper investigates the foundation models' performances on metagenomics binning using human gut microbiome samples. This is done by clustering contig embeddings into bins and compare those to the true strain labels. Additionally, we explore the models' understandings of evolutionary lineage by grouping strains into their shared ancestor. We then analyze the models' capabilities to distinguish between less biologically related strains.

III. Literature Review

We conducted a systematic literature review to identify unknown DNA foundation models to include in the benchmark.

1) Study eligibility criteria: We considered a study eligible if it met one of the following criteria: (1) it was an original paper presenting a DNA foundation model, or (2) it was a benchmark study evaluating DNA foundation models. Since foundation models only became relevant after the introduction of transformers in 2017, we only included studies published later than 2017 [47].

2) Search strategy: The literature search was conducted on October 5th 2024 using the OpenAlex API [48], which compiles data from sources such as ORCID, ROR, DOAJ, Unpaywall, Pubmed, ISSN, web crawls, arXiv, and Zenodo.

We used the search query *("DNA" AND "foundation model") OR ("DNABERT")*, including the term *DNABERT* based on prior knowledge of the model's existence. All phrases were enclosed in double quotation marks to perform an exact match, as a fuzzy query yielded too many irrelevant results. We refer to the results of the OpenAlex search as *reports*. These cover many types of documents such as journal articles, pre-prints, conference abstracts, and unpublished manuscripts [49].

The OpenAlex API offers several query options, such as searching within *abstract* or *full text*. We used the *search* parameter which matches across titles, abstracts, and full texts, thus providing a more comprehensive retrieval of reports.

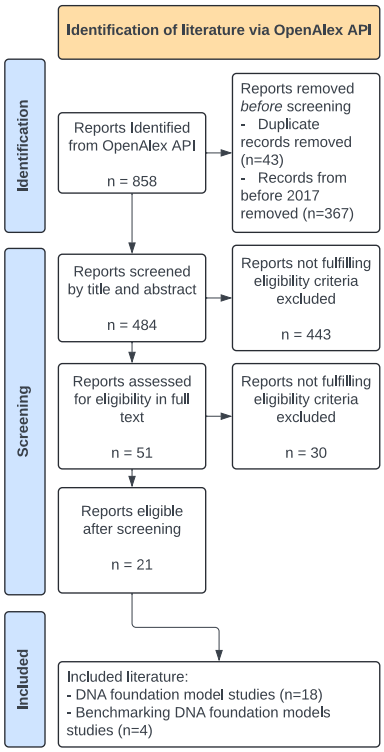

Fig. 1: Literature review on DNA foundation models using the OpenAlexAPI. From the initial 858 reports, we identified 18 original studies on DNA foundation models and 4 studies benchmarking DNA foundation models.

3) Identification of relevant studies: We screened the titles and abstracts of all reports retrieved from the OpenAlex search, based on the eligibility criteria. For the reports that passed the initial screening, we conducted a full-text review.

4) Results of the literature review: The results of the OpenAlex literature review on DNA foundation models are outlined in Figure 1. The search returned 858 results, of which 43 were duplicate records and 367 were published before 2017. After title and abstract screening, 443 additional reports were excluded, leaving 51 eligible for full-text screening. From these, 30 were excluded during full-text review, resulting in the identification of 18 studies presenting DNA foundation models and 4 studies benchmarking DNA foundation models.

IV. Methods

A. Models

We identified 18 DNA foundation models in the literature review, but due to limited accessibility of pre-trained models

TABLE I: Overview of the DNA foundation models benchmarked in our study.

Model	Architecture	Pre-Training strategy	Tokenizer	Window size	Sequence length	Pretraining data	Source	Benchmarked in this study
DNABERT	BERT	MLM	Overlapping k-mer	512	512	Human genome	[30]	✗
DNABERT-2	BERT	MLM	BPE	512	10,000[a]	Multispecies	[43]	✓
DNABERT-S	BERT	C^2LR, MI-Mix	BPE	512	10,000[a]	Multispecies	[17]	✓
NT	BERT	MLM	Non-overlapping 6-mer	1000	6,000	Multispecies	[32]	✗
NT V2	BERT	MLM	Non-overlapping 6-mer	2,048	12,000	Multispecies	[44]	✓
GROVER	BERT	MLM	BPE	512	8,192[b]	Human genome	[34]	✓
GENA-LM	BERT	MLM	BPE	512	4,500	Multispecies	[45]	✓
Hyena-DNA	Hyena	NTP	Single nucleotide	1,000,000	1,000,000	Human genome	[42]	✓

[a] DNABERT-2 and DNABERT-S adopted Attention with Linear Biases ALiBi [46] which was showcased to handle sequences of length up to 10,000 at inference. This context length was also used in the evaluation of DNABERT-2 and DNABERT-S. In addition, the sequence length limit of DNABERT-2 and DNABERT-S is considered to be 3,000 in [44], and 1,000-4,000 in [45].

[b] No explicit length was reported in GROVER [34]. We derived the sequence length by assuming that the 512 BPE tokens had an average length of 16, following a previous benchmark study BEND [13].

and time constraints, only six were benchmarked in this study. These six models are outlined in Table I, including two that are earlier versions of others. The remaining 12 models can be found in Appendix A.

1) DNA foundation models: The first DNA foundation model, **DNABERT** [30], was released in 2021. It uses a BERT based architecture with a masked-language-modeling (MLM) pre-training objective. In the model, DNA sequences are tokenized using a sliding window of size $k \in 3, 4, 5, 6$ and stride 1, corresponding to overlapping k-mers. The model was trained solely on the human genome with a maximum context window size of 512. DNABERT achieved state-of-the-art performance when fine-tuned on downstream tasks, and demonstrated that pre-training only on the human genome could transfer to other organisms with good performance. Another version from the same paper, DNABERT-XL, bypasses the input size constraints by splitting sequences into pieces with max token-length of 512, independently feeds them into the model, and concatenates the last layer representations into a final output.

DNABERT-2, the successor to DNABERT, also uses a BERT based architecture, but adopts Sentence Piece and Byte-Pair Encoding tokenization (BPE) to replace the fixed length vocabulary of k-mers. The pre-training data was extended to include 135 different species across 6 categories, e.g. fungi and viral species. The attention blocks in DNABERT-2 was modified to handle longer input sequences during inference by adopting Attention with Linear Biases (ALiBi) [46]. To increase computational complexity, DNABERT-2 uses Flash Attention [40].

DNABERT-S [17] uses the architecture of DNABERT-2, but implements a different pre-training objective called Curriculum Contrastive Learning (C^2LR) [50]. In this approach, positive pairs are non-overlapping sequences from the same genome and negative instances are sequences from a different genome. This pre-training objective should encourage the model to cluster and separate different species within the embedding space. Manifold Instance Mixup (MI-Mix) [51] was used as a regularization strategy. These design choices of DNABERT-S are intended to construct species-aware embeddings, catered to tasks such as metagenomics binning [17].

The Nucleotide Transformer (**NT**) refers to a family of models that share the original BERT architecture, but vary in the number of training genomes and number of model parameters from 500M up to 2,5B [32]. This is up to ~ 25 times larger than DNABERT which uses 86M parameters and DNABERT-2/DNABERT-S which uses 117M. The NT uses a sliding window size of $k = 6$ and stride $= 6$, such that sequences are tokenized using non-overlapping 6-mers. The NT models were extended in a second version **NT V2** [44]. We used the NT V2 multispecies model with 100M parameters.

GROVER [34] and **GENA-LM** [45] also use a BERT-based architecture and a BPE tokenizer. GROVER is solely pre-trained on the human genome, whereas GENA-LM constitutes a family of models trained on genomes from multiple species including the human genome.

Shared by all the transformer based architectures is their token-context windows that limits the ability to handle long DNA sequences (see Table I).

In contrast, **Hyena-DNA** uses the Hyena architecture [39] with nucleotide level tokens and a context window size of up to 1M base-pairs [42]. This greatly exceeds the window size of transformers, allowing Hyena-DNA to capture long-range dependencies between nucleotides. The model is pre-trained using a next-token-prediction (NTP) objective and is up to 160 times faster than transformer based architectures [42].

2) Baseline Models: We compared the DNA foundation models to three baseline models.

Tetra-Nucleotide Frequencies, **TNF**, is a widely used representation of DNA sequences [25]. A tetra-nucleotide is a DNA sequence of 4 letters (4-mer), resulting in a vocabulary of 256 unique tetra-nucleotides. TNF counts the relative frequency of each unique tetranucleotide in a sequence and uses it as a 256-dimensional embedding.

VAMB uses a variational autoencoder to map TNF's and species abundance into a latent space. We transform the TNF embeddings into the learned latent space of VAMB by a dot-product operation, such that each sequence is represented with a 103-dimensional embedding.

DNA2Vec uses a Word2Vec approach to learn static embeddings for varying length k-mers [28]. We used the 4-mer DNA2Vec embeddings that were pre-trained on the human genome. We take the dot-product of the TNF and the pre-trained 4-mer embeddings to represent each DNA sequence

with a 256-dimensional embedding.

B. Dataset

We used the MetaHIT "error-free" contig dataset produced in MetaBAT [25]. The dataset was derived from reads obtained by the MetaHIT consortium [52], coming from 264 human gut microbiome samples [53]. MetaBAT mapped these reads to known reference genomes at strain level resolution from the NCBI-database [54], and selected 290 genomes with the highest coverage, ranging in length from 1.9M to 10.5M base pairs. To create the dataset, the aligned genomes were shredded into contigs, where each have 31 overlapping base pairs at both ends. These contigs serve as the data representation of each strain. Appendix C shows the name of the 290 reference strain genomes and their number of contigs, ranging from 41 to $1,617$ contigs.

The contigs are considered "error-free" as they were obtained directly from known reference genomes rather than being constructed from reads by an assembly algorithm. Importantly, all the contigs have a ground-truth genome label.

We removed contigs with less than $2,5$k base pairs, resulting in a total of $177,146$ contigs. The contigs lengths span $2,5$K to $64,5$K base pairs with a heavy right-tailed distribution (see Figure 2).

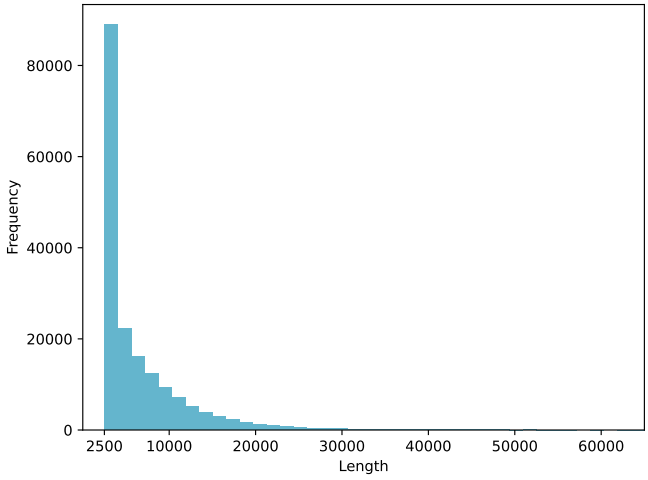

Fig. 2: Histogram of contig lengths in the MetaHit "error-free" dataset. Many DNA foundation models used in this study have sequence length constraints, preventing them from processing the full input of numerous contigs.

C. Metagenomics binning task

Metagenomics binning can essentially be formulated as a clustering problem where the number of clusters correspond to the number of strains in the dataset. The goal is to identify non-overlapping clusters such that each strain is reconstructed from its embedded contigs. We used the same clustering algorithm across models to conduct a systematic comparison between the six DNA foundation models and the three baseline models.

1) K-medoid clustering algorithm: We used the modified K-medoid clustering algorithm implemented in [17], [25], where the number of clusters is unknown (See Appendix B for the algorithmic implementation). The algorithm operates on the N $\times$ N similarity matrix, where N is the number of contigs. Each entry in the matrix corresponds to the cosine similarity between two contigs.

For Z steps, the algorithm picks the contig with the highest similarity to all other contigs as the seed centroid, and considers contigs within a similarity threshold γ to be the neighboring set of contigs $\mathcal{I}$.

An average similarity of the set of neighbors $\mathcal{I}$ is used to update the seed centroid to traverse the embedding space for a total of T iterations. The final set of neighboring contigs $\mathcal{I}$ at the last iteration, are considered to belong to the same cluster, and removed from the embedding space before continuing to the next step. The algorithm will find clusters in the densest area of the embedding space in the beginning, and then move into the less dense areas of the embedding space at later iterations.

After Z steps, clusters containing less contigs than the minimum bin size, m, are removed, and contigs from these clusters are not assigned a label. We set the minimum bin size m $= 10$, as the smallest number of contigs belonging to a genome was 41. We let the K-medoid algorithm pick Z $= 1000$ seed centroids, allowing it to find more clusters than the 290 ground truth genomes. Lastly, we set the number of neighborhood seed updates T $= 3$, following the setup in [17].

We considered two strategies for dealing with contigs that were not assigned a cluster in the K-mediod algorithm. The first strategy was simply to remove the unassigned contigs as done in [17]. The second strategy was to assign contigs to their nearest cluster-centroid .

Each cluster was given a numeric id according to their step z of assignment. These id's were aligned to the ground truth labels via a linear sum assignment algorithm, that matches predicted and ground truth labels by treating entries in the confusion matrix as weights. The algorithm then picks pairs of rows and columns, such that the sum of weights are maximized under the constrains. Importantly, the algorithm requires a certain amount of data points in each predicted cluster, making it necessary to set the minimum bin size *m* in the modified K-medoid algorithm.

2) Selection of threshold γ: The threshold parameter γ defines the decision boundary for including contigs in the set of neighbors, significantly influencing the clustering results. Intuitively, a high threshold results in small clusters that may be pure, but does not contain all contigs belonging to that strain. Contrarily, a low threshold results in larger and impure clusters where contigs are not from the same strain.

We set aside all contigs from 10% of the strains (29) by random sampling, and use this seperate dataset to calibrate the threshold parameter. From the total of $177,146$ contigs, $18,666$ were set aside as a threshold dataset and the remaining $158,480$ were used for the metagenomics binning task.

For each of the models, the threshold was calibrated by computing the similarity of each contig to its strain centroid. The

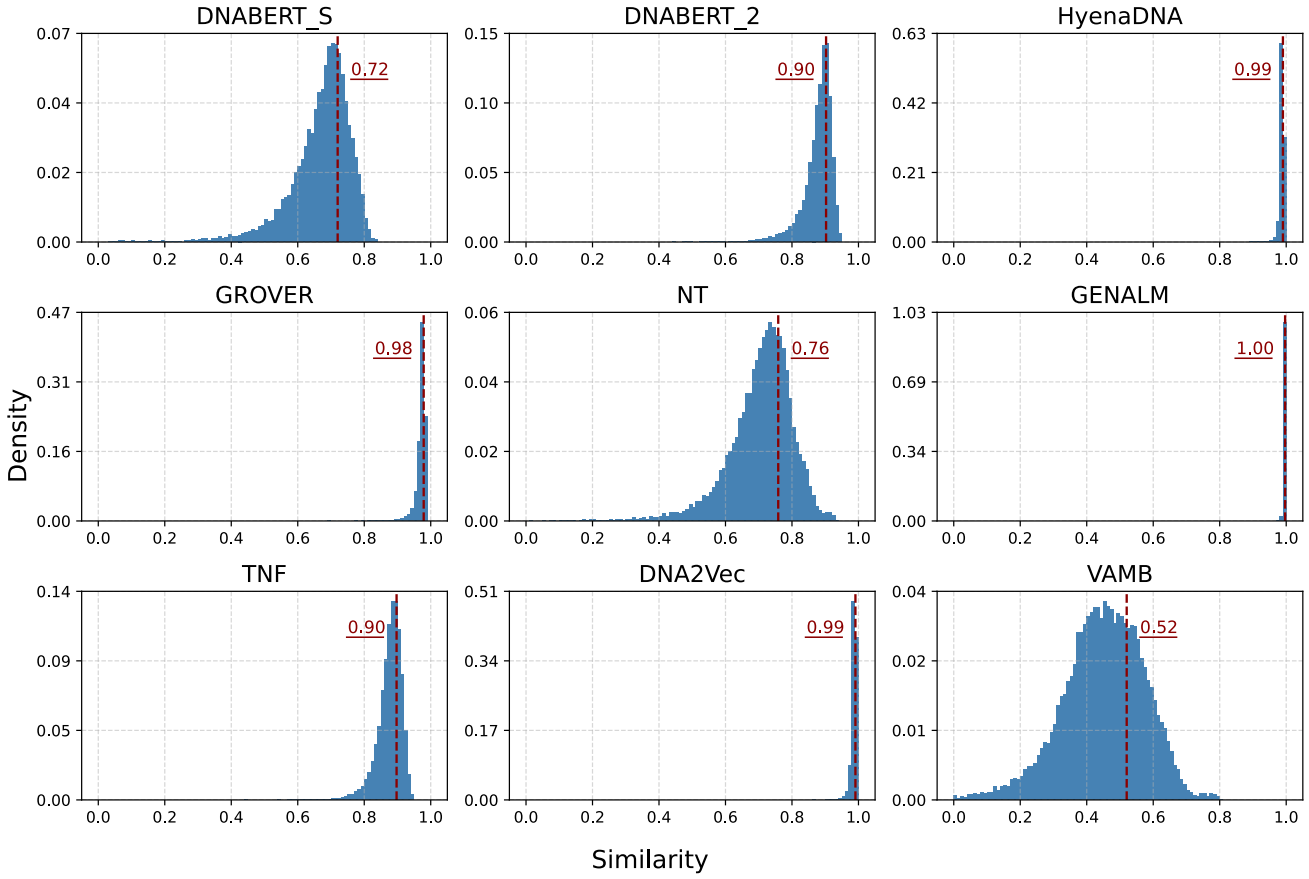

Fig. 3: Distributions of similarities of each contig to its strain centroid across models. The red dotted line correspond to the 70th percentile which is used as the threshold parameter γ in the K-medoid algorithm. To conduct a fair comparison across models, we use a distinct calibrated threshold for each model.

threshold was defined as the 70th percentile of all similarities, following the setup in [17].

The similarity distributions for each of the models is seen in Figure 3. The distributions are different across models, resulting in a model distinct threshold. We also note that similarities and variances are different across models. HyenaDNA, Gena-LM, and DNA2Vec have high similarities with low variance, whereas DNABERT-S, DNABERT-2, NT, and VAMB has lower similarities and higher variance.

This means that using a global threshold across models, would result in biased clustering results as the threshold parameter is influenced by model-distinct embedding spaces. To ensure a fair comparison across models, we used a model-calibrated threshold determined from the threshold dataset as outlined in Figure 3.

3) Evaluation: To evaluate model performance, we compared the predicted clusters with the ground truth labels of contigs, by counting the number of strains that were successfully identified at the following F1-score thresholds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. We consider a genome to be weakly identified if its F1 score exceeds 0.1, moderately identified at an F1 score above 0.5, and precisely identified at an F1 score above 0.9. In contrast, a random learner achieves an F1-score of 0.0032.

We used the silhouette score to measure the quality of the predicted clusters. A silhouette score of 1 indicates great separation between clusters, -1 is the poorest score, where data points are assigned to incorrect clusters, and 0 indicates that clusters are bundled together in the embedding space.

D. Genus level clustering

The phylogenetic tree that describes the hierarchical relationships between all living organisms, was utilized for further analysis of the resulting clusters. All strains in the dataset was mapped to the closest documented taxonomic level, to enable aggregation of strains into their shared origins. Appendix C shows the closest available taxonomy level that was mapped to each strain. Of the documented taxonomy levels associated with the strains, 271 was labeled at the genus level, 14 at the family level, and only 6 was identified at the order level [55]. Genus, family and order are two, three, and four levels up in the phylogenetic tree from strains.

In the further analysis, genus was selected as the taxonomic level because it was the super set containing the majority of the strains. Strains that was mapped to higher taxonomic levels i.e. family and order were removed, decreasing the number of contigs in the evaluation dataset from $158,480$ to $147,312$.

TABLE II: Number of strains contained in the 10 largest genera.

Genus name	Number of strains
Bacteroides	66
Clostridium	22
Eubacterium	13
Streptococcus	11
Bifidobacterium	10
Escherichia	10
Ruminococcus	9
Parabacteroides	8
Alistipes	7
Lactobacillus	7
Other[a]	74

[a] Other comprises 48 genera.

The genus labels revealed a class imbalance where the largest genus, *Bacteroides*, contained 74 unique strains while others only comprised a single strain (see Appendix C). To ensure a sufficient number of strains within the aggregated genus levels, we identified the 10 genera containing the most strains and used in the further analysis as outlined in Table II.

The ground truth strain labels were used to conduct a hierarchical clustering. The goal of this analysis was to explore the models' abilities to distinguish between groups of strains from distinct genera. We expect strains originating from the same genus to be more similar in the embedding space, whilst strains from different genera are more distant.

To determine the similarity between strains, we adopted the Hausdorff distance, commonly used in medical image segmentation, which determines the similarity between two sets of points [22]. The sets of points were defined as the contigs belonging to each strain.

In the vanilla implementation, Hausdorff distance uses the maximum distance of all minimum pairwise distances between the two sets of points. To avoid excessive influence of outlying contigs in the embedding space, we used the fractional Hausdorff distance with a 95'th percentile instead of the maximum distance.

This resulted in a 243×243 similarity matrix, where each entry corresponds to the Hausdorff distance between two strains. Subsequently, we used an agglomerative clustering algorithm where distances between clusters were assigned using Ward's method.

We evaluated the agglomerative clustering results with the ground truth genus labels, using the same F1 score criteria as in the K-medoid evaluation, with thresholds above: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9.

V. Results

A. Metagenomics binning on strains

1) Overall model results: The results in Figure 4 show that none of the models were able to classify all 261 strains in the data. The best performing model, DNABERT-S, identified 123, 44, and 5 strains at F1 score thresholds of 0.1, 0.5, 0.9, using the removal strategy. This corresponds to weakly identifying 47% the strains, moderately identifying 17%, and precisely identifying 2%. The second best performing model was the simple baseline TNF, that identified 66, 6, and 0 strains at F1 score thresholds of 0.1, 0.5, 0.9, while the third best model DNABERT-2 identified 43, 0, and 0 strains.

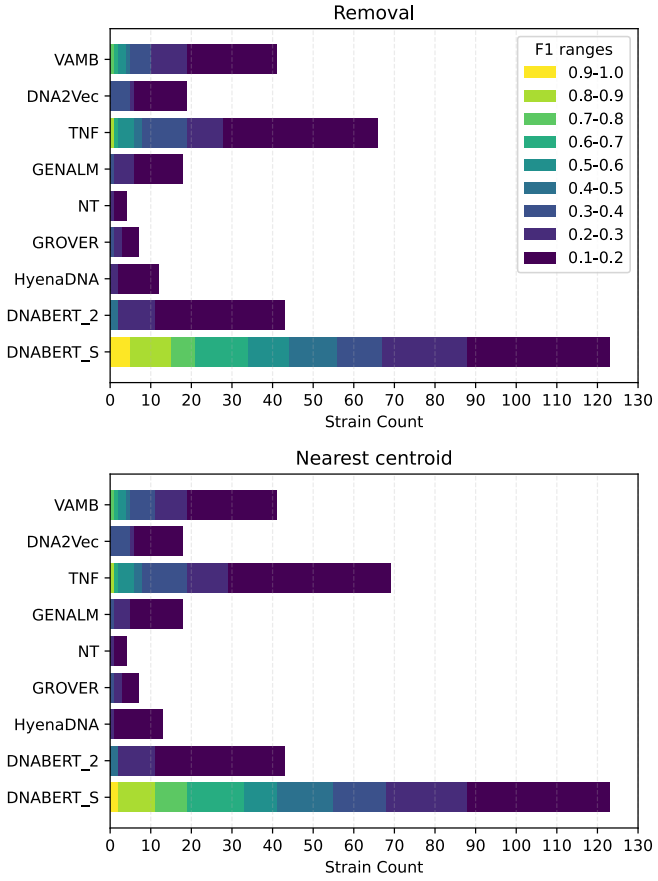

Fig. 4: **Strain level** metagenomics binning results based on removing unpredicted contigs (top) and assigning unpredicted contigs to the nearest centroid (bottom). DNABERT-S outperforms the baselines, but the remaining DNA foundation models perform worse than baselines.

Based on the overall metagenomics binning results in Figure 4, we were unable to correlate groups of model architectures with higher performance. Despite the best performing model being a transformer based architecture, the baseline TNF based on k-mer distributions, along with VAMB that uses a variational autoencoder showed similar or better performance than the rest of the benchmarked foundation models. Notably, Hyena-DNA performed poorly, despite it being the only model able to exploit the complete DNA sequences.

2) Assignment strategy results: The results across models seemed agnostic to the nearest centroid strategy compared to removing unassigned contigs (see Figure 4). This is an interesting finding as the four best performing models (DNABERT-S, TNF, VAMB, and DNABERT-2) also had the highest number of unassigned contigs, as shown in Table III (when disregarding GENA-LM). The unassigned contigs constitute a set of distant points in the embedding space, that were not assigned to any cluster in the K-medoid algorithm. Using a different percentile for calibrating the threshold parameter

TABLE III: Unclassified contigs and silhouette scores across models.

Strategy	Removal		Nearest centroid
	Unassigned contigs	Silhouette score	Silhouette score
DNABERT 2	990	-0.091	-0.091
DNABERT S	7494	-0.005	-0.007
NT	258	-0.110	-0.110
GROVER	286	-0.172	-0.171
GENA-LM	2745	-0.123	-0.112
Hyena-DNA	327	-0.133	-0.171
TNF	3981	-0.053	-0.053
DNA2VEC	536	-0.136	-0.134
VAMB	2759	-0.083	-0.083

would likely have changed the number of unassigned contigs.

3) Silhouette score results: The reported silhouette scores in Table III all show negative values, indicating poor clustering quality across all models. The poorest silhouette score was GROVER with -0.172.

The silhouette score results aligns with the overall F1 score binning results in Figure 4. DNABERT-S with the best performance had a silhouette score of -0.05, and the second best, TNF, had a score of -0.053 (using the removal strategy). Across assignment strategies in the K-medoid algorithm, the silhouette scores arbitrarily decreases, increases or remains intact. This underlines that the model performance was agnostic to the proposed assignment strategy.

B. Genus level clustering

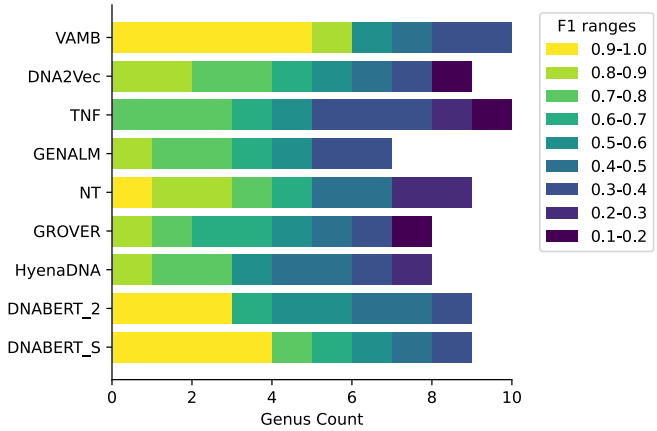

Fig. 5: **Genus level** clustering results, only including the top-10 genera containing the highest number of genomes. The best performing model, VAMB, identified five genera with an F1 score threshold above 0.9. Notably, the model rankings differ compared to those observed during clustering at strain level resolution.

1) Overall genus level results: The genus level results reveal that similarities between organisms with the same genus label can be exploited across the models. Figure 5 shows the F1 scores on the genus level agglomerative clustering using Hausdorff distances. The results demonstrate that models were generally able to successfully separate genera in the embedding space.

The figure illustrates that the model ranking performance has changed compared to the metagenomics binning. VAMB is now the best performing model, identifying 10 and 5 genera at F1 score thresholds greater than 0.3 and 0.9, while the second best model DNABERT-S, classified 9 and 4 genera at the same thresholds. TNF identified all 10 genera at a 0.1 threshold, but failed to achieve this with a threshold above 0.7. This suggests that more complex embeddings, such as the latent space of VAMB, and learned embeddings of DNA foundation models, successfully captured key distinctions between genera, while simple tetra nucleotide frequencies failed to capture these.

The genus level clustering results may be inflated, because we limited the analysis to classes only comprising the top 10 genera with the highest number of strains, potentially making the task easier for the models.

2) Further genus level analysis: We show a detailed outline of the DNABERT-S genus clustering results in Figure 6. The figure visualizes the clustered strain similarities with ground truth genus labels, accompanied by the confusion matrix and a two-dimensional embedding space.

Subfigure (a) shows the Hausdorff distances between clustered strains. Most notable is the large cluster of similar strains belonging to the genus *Bactoreides*, marked by blue in the dendrogram. Other salient clusters include *Escherichia, Bifidobacterium*, and *Streptococcus*, which appear visually distinct in the dendrogram. Genera with less than seven strains, marked in light grey, are mostly scattered in-between the top 10 genus clusters, but some form larger coherent groups.

An interesting result is how the genera *Escherichia* and *Bifidobacterium* present themselves as dense and well separated clusters in the t-SNE plot in Figure 6(c). This partition is supported by the confusion matrix in Figure 6(b) where all strains belonging to *Escherichia* or *Bifidobacterium* were correctly classified to their corresponding genus. These results align with the structure of the phylogenetic tree in Figure 7 where *Escherichia* and *Bifidobacterium* are more biologically distinct than any other strain, sharing common ancestors with other genera only five levels up the tree.

VI. Discussion and conclusions

Metagenomics binning on human gut microbiome data is an important step towards finding associations between microbial genomes and their human host. We evaluated six DNA foundation models on metagenomics binning using human gut microbiome assembled contigs with strain and genus level ground truth labels.

A. Metagenomics binning

DNABERT-S achieves superior performance over the other models, which could be attributed to its pre-training objective that focuses on differentiating species. The remaining DNA foundation models, pre-trained using an MLM objective, did not show any improvement over the baseline model TNF, and most performed worse than the other baselines.

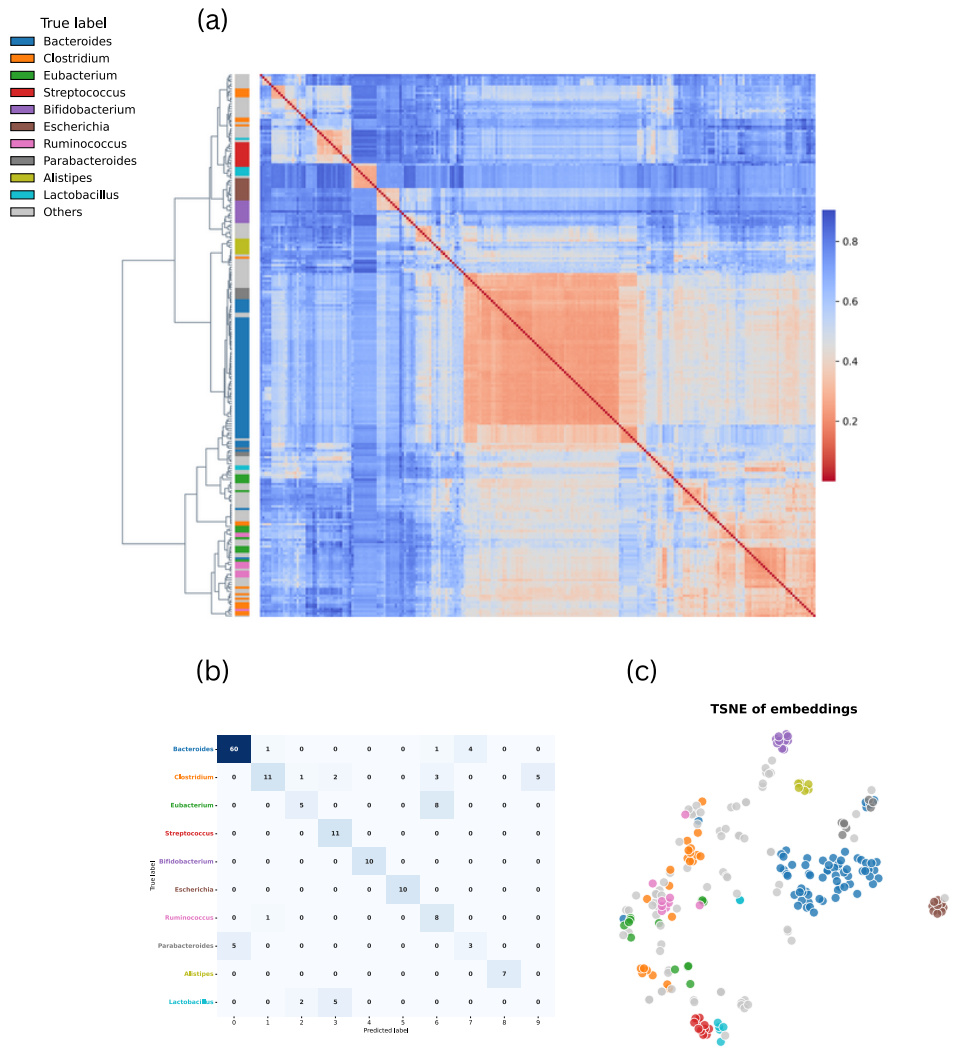

Fig. 6: Further analysis of the genus level clustering results from DNABERT-S. (a) shows a heatmap of the 243×243 Hausdorff similarity matrix with the dendrogram of the agglomorative clustering algorithm on the left colored by ground truth genera. (b) displays the 10×10 confusion matrix with the ground truth genus labels on the y-axis and the clustering labels on the x-axis. (c) shows a 2D t-SNE plot of the same Hausdorff similarity matrix visualized in (a). Some clusters visually identified in (a) align with the confusion matrix and the t-SNE embeddings. For example, *Bacteroides* (blue) and *Parabacteroides* (dark grey) are overlapping across (a), (b), and (c).

The foundation models' poor performance may be partially due to their inability to process long sequences (see Table I). For instance, GROVER has a maximum sequence length of $8,192$, which prevents it from handling the full input of $\sim 42,000$ contigs (see Figure 2), whereas the baseline models do not have such constraints. Surprisingly, Hyena-DNA, which is capable of processing all contigs in our dataset, still performs worse than all the baseline models.

The poor performance of Hyena-DNA contrasts with the results reported in the original paper, where it achieves state-of-the-art performance on 12 out of 18 downstream tasks adopted from the NT paper [32], [42]. This highlights that no single foundation model consistently outperforms its competitors across a wide range of tasks.

In the DNABERT-S paper [17] they compare their proposed model with competing foundation models and baselines on a metagenomics binning task using the CAMI2 dataset covering microorganisms from marine and plant environments [18]. The model ranking in our results is consistent with the ranking in DNABERT-S, showing that DNABERT-S achieves the best performance followed by TNF and VAMB across our dataset and the CAMI2 dataset. This highlights a coherent performance across microbiome environments. The results could be biased towards DNABERT-S, as we adopted the K-medoid algorithm from their paper.

While the performance ranking of models in our study

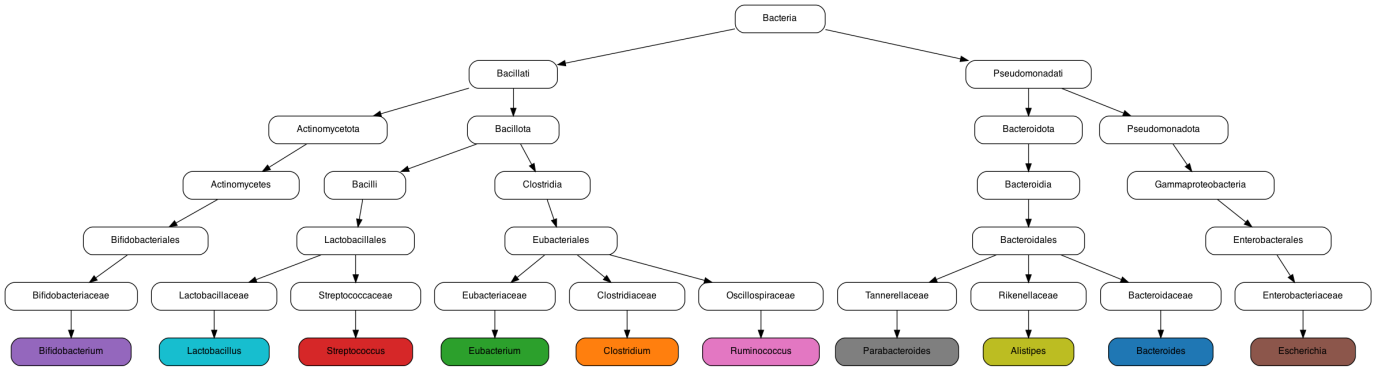

Fig. 7: The lineage of the 10 most frequent genera visualized in a phylogenetic tree according to their taxonomy. The root node represents the *Bacteria* domain and the leaf nodes correspond to the genus level taxonomy. *Bifidobacterium* (purple) and *Escherichia* (brown) are structurally distinct from other genera in the tree. This is also captured in the t-SNE embeddings where distinct purple and brown clusters appear in the 2-dimensional space, see Figure 6(c).

aligns with DNABERT-S, the absolute performance, measured by the number of genomes identified, is considerably lower. Using 527 species from a marine environment dataset, DNABERT-S, TNF, and VAMB identified approximately 200 (38%), 125 (24%), and 75 (14%) species at an F1 higher than 0.5. In our dataset of 261 strains, these models only identified 44 (17%), 6 (2%), and 4 (1%) strains at an F1 threshold higher than 0.5. We believe that the decrease in performance is affected by binning on strain resolution data as opposed to species resolution. In our dataset, 1.3 strains on average originate from the same species. Strains from the same species overlap in large regions of their DNA, and may be distinct only in a few base-pair positions, whereas different species have less overlap. The small mutations in DNA sequences at strain resolution may be overlooked by the model representations, resulting in an embedding space where strains are overlapping.

Importantly, none of the DNA foundation models were pre-trained on strain level data. This suggests that pre-training on species resolution data, might not generalize to strains. In contrast, our results indicate that merging lower taxonomic levels into their common ancestors is a more straightforward task.

Our results are in contrast to a previous benchmark study BEND [13], which found that DNA foundation models performed well in most tasks focusing on local regions of the human genome. This highlights that representations of DNA learned by current foundation models lacks generalisability across a wide range of tasks. While BEND benchmarks the foundation models on one particular genome, metagenomics binning relies on learning the representations and associations of multiple genomes. Future foundation models could focus on developing representations capable of solving inherently different tasks in genomics and metagenomics.

B. Genus level clustering

In contrast to the difficulty of clustering at strain resolution, we found that it was a more straightforward task for the models to distinguish between aggregated genus levels. These results were based on distances between sets of points, in contrast to the K-medoid algorithm that relies on pairwise distances. Our results demonstrate that the Hausdorff distance, was effective in separating sets of strains to distinct genera. This suggests that the Hausdorff distance could be applied in other parts of metagenomic research, e.g. when comparing sets of identified and unidentified genomes in a metagenomics sample.

C. Limitations and future work

In the literature review, we identified 18 DNA foundation models, but due to limited availability of pre-trained models and resource constraints, only 6 were benchmarked. It is important to note that our study is not a comprehensive benchmark, as the remaining 12 models could potentially influence the results. Future work should focus on including these additional models to provide a more complete evaluation of the field.

Our results rely on the linear sum assignment algorithm which aligns predictions and ground truth labels, such that the F1 scores are maximized. Future work could explore the distribution of contigs within a labeled cluster to investigate the assignments of the algorithm.

We used the ”error-free” dataset produced in MetaBAT [25], where contigs are obtained directly from reference genomes. This approach bypasses a critical step in a realistic metagenomics pipeline, where reads are assembled into contigs by an algorithm.

Future work could apply metagenomics binning on contigs obtained from an unsupervised assembly algorithm, that enables clustering without knowing the number of genomes. To evaluate this realistic metagenomics pipeline, the empirically formed clusters should be aligned to reference genomes. This alignment could reveal currently unknown genomes as some group of contigs might not correspond to a known reference.

D. Concluding remarks

In this study, we benchmarked six DNA foundation models on metagenomics binning using strain resolution data from the human gut microbiome.

We found that only one foundation model identified more strains than the baselines. The suboptimal performance of

foundation models may be attributed to their input length constraints and pre-training objectives.

In addition, we constructed a framework for conducting a genus level analysis using the Hausdorff distance. This analysis revealed that the DNA foundation models were able to capture the biological distinctions between genera.

REFERENCES

[1] OpenAI, J. Achiam, S. Adler, *et al.*, *GPT-4 Technical Report*, arXiv:2303.08774, Mar. 2024. DOI: 10.48550/arXiv.2303.08774. [Online]. Available: http://arxiv.org/abs/2303.08774 (visited on 11/25/2024).

[2] H. Touvron, T. Lavril, G. Izacard, *et al.*, *LLaMA: Open and Efficient Foundation Language Models*, arXiv:2302.13971, Feb. 2023. DOI: 10.48550/arXiv.2302.13971. [Online]. Available: http://arxiv.org/abs/2302.13971 (visited on 11/25/2024).

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 [cs], May 2019. DOI: 10.48550/arXiv.1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805 (visited on 12/12/2024).

[4] A. Madani, B. Krause, E. R. Greene, *et al.*, "Large language models generate functional protein sequences across diverse families," en, *Nature Biotechnology*, vol. 41, no. 8, pp. 1099–1106, Aug. 2023, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-022-01618-2. [Online]. Available: https://www.nature.com/articles/s41587-022-01618-2 (visited on 11/25/2024).

[5] J. Abramson, J. Adler, J. Dunger, *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," en, *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w. [Online]. Available: https://www.nature.com/articles/s41586-024-07487-w (visited on 12/12/2024).

[6] D. Qin, "Next-generation sequencing and its clinical application," *Cancer Biology & Medicine*, vol. 16, no. 1, pp. 4–10, Feb. 2019, ISSN: 2095-3941. DOI: 10.20892/j.issn.2095-3941.2018.0055. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6528456/ (visited on 11/25/2024).

[7] *Homo sapiens genome assembly GRCh38.p14*, en. [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/ (visited on 11/25/2024).

[8] J. R. Marchesi and J. Ravel, "The vocabulary of microbiome research: A proposal," en, *Microbiome*, vol. 3, no. 1, p. 31, Jul. 2015, ISSN: 2049-2618. DOI: 10.1186/s40168-015-0094-5. [Online]. Available: https://doi.org/10.1186/s40168-015-0094-5 (visited on 11/28/2024).

[9] J. N. Nissen, J. Johansen, R. L. Allesøe, *et al.*, "Improved metagenome binning and assembly using deep variational autoencoders," en, *Nature Biotechnology*, vol. 39, no. 5, pp. 555–560, May 2021, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-020-00777-4. [Online]. Available: https://www.nature.com/articles/s41587-020-00777-4 (visited on 11/29/2024).

[10] G. Roy, E. Prifti, E. Belda, and J.-D. Zucker, "Deep learning methods in metagenomics: A review," *Microbial Genomics*, vol. 10, no. 4, p. 001 231, 2024, Publisher: Microbiology Society, ISSN: 2057-5858. DOI: 10.1099/mgen.0.001231. [Online]. Available: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001231 (visited on 11/29/2024).

[11] A. Blanco-Míguez, F. Beghini, F. Cumbo, *et al.*, "Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4," en, *Nature Biotechnology*, vol. 41, no. 11, pp. 1633–1644, Nov. 2023, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-023-01688-w. [Online]. Available: https://www.nature.com/articles/s41587-023-01688-w (visited on 11/29/2024).

[12] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biology*, vol. 20, no. 1, p. 257, Nov. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0. [Online]. Available: https://doi.org/10.1186/s13059-019-1891-0 (visited on 11/29/2024).

[13] F. I. Marin, F. Teufel, M. Horlacher, *et al.*, "BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks," en, Oct. 2023. [Online]. Available: https://openreview.net/forum?id=uKB4cFNQFg (visited on 11/25/2024).

[14] Z. Liu, J. Li, S. Li, *et al.*, *GenBench: A Benchmarking Suite for Systematic Evaluation of Genomic Foundation Models*, arXiv:2406.01627, Jun. 2024. DOI: 10.48550/arXiv.2406.01627. [Online]. Available: http://arxiv.org/abs/2406.01627 (visited on 11/25/2024).

[15] H. Feng, L. Wu, B. Zhao, *et al.*, "Benchmarking DNA Foundation Models for Genomic Sequence Classification," *bioRxiv*, p. 2024.08.16.608288, Aug. 2024, ISSN: 2692-8205. DOI: 10.1101/2024.08.16.608288. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11343214/ (visited on 11/25/2024).

[16] Z. Tang, N. Somia, Y. Yu, and P. K. Koo, "Evaluating the representational power of pre-trained DNA language models for regulatory genomics," *bioRxiv*, p. 2024.02.29.582810, Sep. 2024, ISSN: 2692-8205. DOI: 10.1101/2024.02.29.582810. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10925287/ (visited on 11/25/2024).

[17] Z. Zhou, W. Wu, H. Ho, *et al.*, *DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings*, arXiv:2402.08777, Oct. 2024. DOI: 10.48550/arXiv.2402.08777. [Online]. Available: http://arxiv.org/abs/2402.08777 (visited on 11/29/2024).

[18] F. Meyer, A. Fritz, Z.-L. Deng, *et al.*, "Critical Assessment of Metagenome Interpretation: The second round of challenges," en, *Nature Methods*, vol. 19, no. 4, pp. 429–440, Apr. 2022, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-022-

01431-4. [Online]. Available: https://www.nature.com/articles/s41592-022-01431-4 (visited on 12/01/2024).

[19] I. Cho and M. J. Blaser, "The Human Microbiome: At the interface of health and disease," *Nature reviews. Genetics*, vol. 13, no. 4, pp. 260–270, Mar. 2012, ISSN: 1471-0056. DOI: 10.1038/nrg3182. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418802/ (visited on 11/29/2024).

[20] S. Leviatan, S. Shoer, D. Rothschild, M. Gorodetski, and E. Segal, "An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species," en, *Nature Communications*, vol. 13, no. 1, p. 3863, Jul. 2022, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-022-31502-1. [Online]. Available: https://www.nature.com/articles/s41467-022-31502-1 (visited on 12/10/2024).

[21] J. Gilbert, M. J. Blaser, J. G. Caporaso, J. Jansson, S. V. Lynch, and R. Knight, "Current understanding of the human microbiome," *Nature medicine*, vol. 24, no. 4, pp. 392–400, Apr. 2018, ISSN: 1078-8956. DOI: 10.1038/nm.4517. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7043356/ (visited on 12/13/2024).

[22] M. Nawaz, A. Uvaliyev, K. Bibi, *et al.*, "Unraveling the complexity of Optical Coherence Tomography image segmentation using machine and deep learning techniques: A review," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102 269, Sep. 2023, ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2023.102269. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611123000873 (visited on 12/11/2024).

[23] W. S, F. Z, T. J, *et al.*, "DeePhage: Distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach," en, *GigaScience*, vol. 10, no. 9, Sep. 2021, Publisher: Gigascience, ISSN: 2047-217X. DOI: 10.1093/gigascience/giab056. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34498685/ (visited on 11/29/2024).

[24] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," en, *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, Aug. 2015, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.3300. [Online]. Available: https://www.nature.com/articles/nbt.3300 (visited on 11/29/2024).

[25] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," en, *PeerJ*, vol. 3, no. 8, e1165, 2015, ISSN: 2167-8359. DOI: 10.7717/peerj.1165. [Online]. Available: https://escholarship.org/uc/item/58v471ws (visited on 11/29/2024).

[26] S. Pan, X.-M. Zhao, and L. P. Coelho, "SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing," *Bioinformatics*, vol. 39, no. Supplement_1, pp. i21–i29, Jun. 2023, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad209. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad209 (visited on 11/29/2024).

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, Sep. 2013. DOI: 10.48550/arXiv.1301.3781. [Online]. Available: http://arxiv.org/abs/1301.3781 (visited on 11/29/2024).

[28] P. Ng, *Dna2vec: Consistent vector representations of variable-length k-mers*, arXiv:1701.06279, Jan. 2017. DOI: 10.48550/arXiv.1701.06279. [Online]. Available: http://arxiv.org/abs/1701.06279 (visited on 11/30/2024).

[29] S. N. Aakur, V. Indla, V. Indla, *et al.*, *Metagenome2Vec: Building Contextualized Representations for Scalable Metagenome Analysis*, arXiv:2111.08001, Nov. 2021. DOI: 10.48550/arXiv.2111.08001. [Online]. Available: http://arxiv.org/abs/2111.08001 (visited on 11/29/2024).

[30] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab083. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab083 (visited on 12/01/2024).

[31] Q. Liang, P. W. Bible, Y. Liu, B. Zou, and L. Wei, "DeepMicrobes: Taxonomic classification for metagenomics with deep learning," *NAR Genomics and Bioinformatics*, vol. 2, no. 1, lqaa009, Mar. 2020, ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa009. [Online]. Available: https://doi.org/10.1093/nargab/lqaa009 (visited on 11/29/2024).

[32] H. Dalla-Torre, L. Gonzalez, J. M. Revilla, *et al.*, *The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics*, en, Pages: 2023.01.11.523679 Section: New Results, Jan. 2023. DOI: 10.1101/2023.01.11.523679. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1 (visited on 12/01/2024).

[33] B. Ligeti, I. Szepesi-Nagy, B. Bodnár, N. Ligeti-Nagy, and J. Juhász, "ProkBERT family: Genomic language models for microbiome applications," English, *Frontiers in Microbiology*, vol. 14, Jan. 2024, Publisher: Frontiers, ISSN: 1664-302X. DOI: 10.3389/fmicb.2023.1331233. [Online]. Available: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1331233/full (visited on 12/02/2024).

[34] M. Sanabria, J. Hirsch, P. M. Joubert, and A. R. Poetsch, "DNA language model GROVER learns sequence context in the human genome," en, *Nature Machine Intelligence*, vol. 6, no. 8, pp. 911–923, Aug. 2024, Publisher: Nature Publishing Group, ISSN: 2522-5839. DOI: 10.1038/s42256-024-00872-0. [Online]. Available: https://www.nature.com/articles/s42256-024-00872-0 (visited on 12/01/2024).

[35] E. Nguyen, M. Poli, M. G. Durrant, *et al.*, "Sequence modeling and design from molecular to genome scale with Evo," *Science*, vol. 386, no. 6723, eado9336, Nov. 2024, Publisher: American Association for the Ad-

vancement of Science. DOI: 10.1126/science.ado9336. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.ado9336 (visited on 12/05/2024).

[36] A. Wichmann, E. Buschong, A. Müller, *et al.*, "Meta-Transformer: Deep metagenomic sequencing read classification using self-attention models," *NAR Genomics and Bioinformatics*, vol. 5, no. 3, lqad082, Sep. 2023, ISSN: 2631-9268. DOI: 10.1093/nargab/lqad082. [Online]. Available: https://doi.org/10.1093/nargab/lqad082 (visited on 11/29/2024).

[37] A. Gu and T. Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, arXiv:2312.00752, May 2024. DOI: 10.48550/arXiv.2312.00752. [Online]. Available: http://arxiv.org/abs/2312.00752 (visited on 12/01/2024).

[38] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, *Hungry Hungry Hippos: Towards Language Modeling with State Space Models*, arXiv:2212.14052, Apr. 2023. DOI: 10.48550/arXiv.2212.14052. [Online]. Available: http://arxiv.org/abs/2212.14052 (visited on 12/05/2024).

[39] M. Poli, S. Massaroli, E. Nguyen, *et al.*, *Hyena Hierarchy: Towards Larger Convolutional Language Models*, arXiv:2302.10866, Apr. 2023. DOI: 10.48550/arXiv.2302.10866. [Online]. Available: http://arxiv.org/abs/2302.10866 (visited on 12/05/2024).

[40] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*, arXiv:2205.14135, Jun. 2022. DOI: 10.48550/arXiv.2205.14135. [Online]. Available: http://arxiv.org/abs/2205.14135 (visited on 12/05/2024).

[41] Y. Schiff, C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov, *Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling*, arXiv:2403.03234, Jun. 2024. DOI: 10.48550/arXiv.2403.03234. [Online]. Available: http://arxiv.org/abs/2403.03234 (visited on 12/02/2024).

[42] E. Nguyen, M. Poli, M. Faizi, *et al.*, *HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution*, arXiv:2306.15794, Nov. 2023. DOI: 10.48550/arXiv.2306.15794. [Online]. Available: http://arxiv.org/abs/2306.15794 (visited on 12/05/2024).

[43] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, *DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome*, arXiv:2306.15006, Mar. 2024. DOI: 10.48550/arXiv.2306.15006. [Online]. Available: http://arxiv.org/abs/2306.15006 (visited on 12/01/2024).

[44] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, *et al.*, "Nucleotide Transformer: Building and evaluating robust foundation models for human genomics," en, *Nature Methods*, pp. 1–11, Nov. 2024, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-024-02523-z. [Online]. Available: https://www.nature.com/articles/s41592-024-02523-z (visited on 12/01/2024).

[45] V. Fishman, Y. Kuratov, A. Shmelev, *et al.*, *GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences*, en, Pages: 2023.06.12.544594 Section: New Results, Aug. 2024. DOI: 10.1101/2023.06.12.544594. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.06.12.544594v3 (visited on 12/01/2024).

[46] O. Press, N. A. Smith, and M. Lewis, *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*, arXiv:2108.12409 [cs], Apr. 2022. DOI: 10.48550/arXiv.2108.12409. [Online]. Available: http://arxiv.org/abs/2108.12409 (visited on 12/11/2024).

[47] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762, Aug. 2023. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762 (visited on 10/10/2024).

[48] J. Priem, H. Piwowar, and R. Orr, *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*, arXiv:2205.01833, Jun. 2022. DOI: 10.48550/arXiv.2205.01833. [Online]. Available: http://arxiv.org/abs/2205.01833 (visited on 11/29/2024).

[49] M. J. Page, D. Moher, P. M. Bossuyt, *et al.*, "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," en, *BMJ*, vol. 372, n160, Mar. 2021, Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, ISSN: 1756-1833. DOI: 10.1136/bmj.n160. [Online]. Available: https://www.bmj.com/content/372/bmj.n160 (visited on 10/10/2024).

[50] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A Simple Framework for Contrastive Learning of Visual Representations*, arXiv:2002.05709, Jul. 2020. DOI: 10.48550/arXiv.2002.05709. [Online]. Available: http://arxiv.org/abs/2002.05709 (visited on 12/05/2024).

[51] V. Verma, A. Lamb, C. Beckham, *et al.*, *Manifold Mixup: Better Representations by Interpolating Hidden States*, arXiv:1806.05236, May 2019. DOI: 10.48550/arXiv.1806.05236. [Online]. Available: http://arxiv.org/abs/1806.05236 (visited on 12/05/2024).

[52] S. D. Ehrlich, "MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract," en, in *Metagenomics of the Human Body*, K. E. Nelson, Ed., New York, NY: Springer, 2011, pp. 307–316, ISBN: 978-1-4419-7089-3. DOI: 10.1007/978-1-4419-7089-3_15. [Online]. Available: https://doi.org/10.1007/978-1-4419-7089-3_15 (visited on 11/29/2024).

[53] *ERP000108 : Study : SRA Archive : NCBI*. [Online]. Available: https://www.ncbi.nlm.nih.gov/Traces/index.html?view=study&acc=ERP000108 (visited on 11/29/2024).

[54] E. W. Sayers, E. E. Bolton, J. R. Brister, *et al.*, "Database resources of the national center for biotechnology information," eng, *Nucleic Acids Research*, vol. 50, no. D1, pp. D20–D26, Jan. 2022, ISSN: 1362-4962. DOI: 10.1093/nar/gkab1112.

[55] C. L. Schoch, S. Ciufo, M. Domrachev, *et al.*, "NCBI Taxonomy: A comprehensive update on curation, resources and tools," eng, *Database: The Journal of Biological Databases and Curation*, vol. 2020, baaa062,

Jan. 2020, ISSN: 1758-0463. DOI: 10.1093/database/baaa062.

[56] X. Zhang, M. Yang, X. Yin, Y. Qian, and F. Sun, *DeepGene: An Efficient Foundation Model for Genomics based on Pan-genome Graph Transformer*, en, Pages: 2024.04.24.590879 Section: New Results, May 2024. DOI: 10.1101/2024.04.24.590879. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.04.24.590879v2 (visited on 12/02/2024).

[57] X. Shen and X. Li, *OmniNA: A foundation model for nucleotide sequences*, en, Pages: 2024.01.14.575543 Section: New Results, Jan. 2024. DOI: 10.1101/2024.01.14.575543. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.01.14.575543v1 (visited on 12/02/2024).

[58] A. Malusare, H. Kothandaraman, D. Tamboli, N. A. Lanman, and V. Aggarwal, *Understanding the Natural Language of DNA using Encoder-Decoder Foundation Models with Byte-level Precision*, arXiv:2311.02333, Aug. 2024. DOI: 10.48550/arXiv.2311.02333. [Online]. Available: http://arxiv.org/abs/2311.02333 (visited on 12/02/2024).

[59] S. Roy, J. Wallat, S. S. Sundaram, W. Nejdl, and N. Ganguly, "GENEMASK: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning," in *ECAI 2023*, IOS Press, 2023, pp. 2002–2009. DOI: 10.3233/FAIA230492. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/FAIA230492 (visited on 12/02/2024).

[60] G. Benegas, S. S. Batra, and Y. S. Song, *DNA language models are powerful predictors of genome-wide variant effects*, en, Pages: 2022.08.22.504706 Section: New Results, Aug. 2023. DOI: 10.1101/2022.08.22.504706. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.08.22.504706v3 (visited on 12/02/2024).

[61] S. Roy, S. Sural, and N. Ganguly, *Unlocking Efficiency: Adaptive Masking for Gene Transformer Models*, arXiv:2408.07180, Aug. 2024. DOI: 10.48550/arXiv.2408.07180. [Online]. Available: http://arxiv.org/abs/2408.07180 (visited on 12/02/2024).

[62] W. An, Y. Guo, Y. Bian, *et al.*, "MoDNA: Motif-oriented pre-training for DNA language model," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, ser. BCB '22, New York, NY, USA: Association for Computing Machinery, Aug. 2022, pp. 1–5, ISBN: 978-1-4503-9386-7. DOI: 10.1145/3535508.3545512. [Online]. Available: https://dl.acm.org/doi/10.1145/3535508.3545512 (visited on 12/02/2024).

APPENDIX A
THE 18 DNA FOUNDATION MODELS FROM THE LITERATURE REVIEW

Model	**Source**	**Benchmarked in this study**
DNABERT	[30]	✗
DNABERT-2	[43]	✓
DNABERT-S	[17]	✓
NT	[32]	✗
NT V2	[44]	✓
GROVER	[34]	✓
GENA-LM	[45]	✓
Hyena-DNA	[42]	✓
Caduceus	[41]	✗
Evo	[35]	✗
ProkBERT	[33]	✗
DeepGene	[56]	✗
OmniNA	[57]	✗
ENBED	[58]	✗
GENEMASK	[59]	✗
GPN	[60]	✗
CM-GEMS	[61]	✗
MoDNA	[62]	✗

APPENDIX B
MODIFIED K-MEDOID ALGORITHM FOR METAGENOMICS BINNING

Algorithm 1 describes the modified K-medoid clustering algorithm as implemented in [17], [25]. The algorithm outline is adopted from [17].

Algorithm 1 Modified K-Medoid Clustering

Require: Threshold γ, minimum bin size m, embeddings $E \in \mathbb{R}^{N \times d}$, number of steps Z, number of iterations T
Ensure: Predictions $p \in \mathbb{R}^N$

Initialize: $p_i = -1$ for $i = 1, \ldots, N$, similarity matrix $S = EE^\top$ with $S_{ij} = 0$ if $S_{ij} < \gamma$, density vector $d \in \mathbb{R}^N$ with $d_i = \sum_{j=1}^{N} S_{ij}$
for step $z = 1$ to Z **do**
    Select seed index $s = \arg\max_{s'} d_{s'}$ and corresponding seed E_s
    for iteration $t = 1$ to T **do**
        Find neighborhood indices $\mathcal{I}$ of E_s where $s(E_i, E_s) > \gamma$ and $p_i = -1$ for each $i \in \mathcal{I}$
        Update seed: $E_s \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} E_i$
    end for
    Set $p_i \leftarrow z$, $d_i \leftarrow 0$ for each $i \in \mathcal{I}$
    Update density: $d_x \leftarrow d_x - \sum_{i \in \mathcal{I}} S_{xi}$ for each $x \in \{1, 2, \ldots, N\}$
end for
for step $z = 1$ to Z **do**
    Find indices $\mathcal{I}$ where $p_i = z$ for each $i \in \mathcal{I}$
    if $|\mathcal{I}| < m$ **then**
        Set $p_i \leftarrow -1$ for each $i \in \mathcal{I}$
    end if
end for
Return: Predictions p

APPENDIX C
DETAILED DATASET DESCRIPTION: TAXONOMY LEVEL, GENOME & CONTIGS

Taxonomy level	Taxonomy name	Genome	Number of Contigs
genus	Acidaminococcus	Acidaminococcus_D21_uid55871	385
genus	Acidaminococcus	Acidaminococcus_fermentans_DSM_20731_uid43471	367
genus	Acidaminococcus	Acidaminococcus_intestini_RyC_MR95_uid74445	382
genus	Actinomyces	Actinomyces_ICM47_uid170984	448
genus	Adlercreutzia	Adlercreutzia_equolifaciens_DSM_19450_uid223286	472
genus	Aeromicrobium	Aeromicrobium_JC14_uid199535	731
genus	Akkermansia	Akkermansia_muciniphila_ATCC_BAA_835_uid58985	430
genus	Alcanivorax	Alcanivorax_hongdengensis_A_11_3_uid176602	42
genus	Alistipes	Alistipes_AP11_uid199714	466
genus	Alistipes	Alistipes_HGB5_uid67587	555
genus	Alistipes	Alistipes_JC50_uid199660	640
genus	Alistipes	Alistipes_finegoldii_DSM_17242_uid168180	602
genus	Alistipes	Alistipes_indistinctus_YIT_12060_uid75115	455
genus	Alistipes	Alistipes_onderdonkii_DSM_19147_uid199292	649
genus	Alistipes	Alistipes_putredinis_DSM_17216_uid54803	403
genus	Alistipes	Alistipes_shahii_WAL_8301_uid197175	525
genus	Anaerotruncus	Anaerotruncus_colihominis_DSM_17241_uid54807	626
order	Bacteroidales	Bacteroidales_bacterium_ph8_uid199652	514
genus	Bacteroides	Bacteroides_1_1_14_uid49709	1062
genus	Bacteroides	Bacteroides_1_1_30_uid68191	1017
genus	Bacteroides	Bacteroides_1_1_6_uid55577	1165
genus	Bacteroides	Bacteroides_20_3_uid50765	978
genus	Bacteroides	Bacteroides_2_1_16_uid41707	837
genus	Bacteroides	Bacteroides_2_1_22_uid41621	923
genus	Bacteroides	Bacteroides_2_1_33B_uid41591	783
genus	Bacteroides	Bacteroides_2_1_56FAA_uid68193	926
genus	Bacteroides	Bacteroides_2_1_7_uid55579	853
genus	Bacteroides	Bacteroides_2_2_4_uid55581	1142
genus	Bacteroides	Bacteroides_3_1_19_uid49705	824
genus	Bacteroides	Bacteroides_3_1_23_uid49963	1053
genus	Bacteroides	Bacteroides_3_1_33FAA_uid41705	852
genus	Bacteroides	Bacteroides_3_1_40A_uid62053	864
genus	Bacteroides	Bacteroides_3_2_5_uid55583	858
genus	Bacteroides	Bacteroides_4_1_36_uid61871	731
genus	Bacteroides	Bacteroides_4_3_47FAA_uid55585	854
genus	Bacteroides	Bacteroides_9_1_42FAA_uid55587	894
genus	Bacteroides	Bacteroides_D20_uid42369	709
genus	Bacteroides	Bacteroides_D22_uid49721	1036
genus	Bacteroides	Bacteroides_D2_uid55591	1154
genus	Bacteroides	Bacteroides_barnesiae_DSM_18169_uid199296	561
genus	Bacteroides	Bacteroides_caccae_ATCC_43185_uid54521	692
genus	Bacteroides	Bacteroides_caccae_CL03T12C61_uid181623	893
genus	Bacteroides	Bacteroides_capillosus_ATCC_29799_uid54531	709
genus	Bacteroides	Bacteroides_cellulosilyticus_CL02T12C19_uid181624	1206
genus	Bacteroides	Bacteroides_cellulosilyticus_DSM_14838_uid55279	1039
genus	Bacteroides	Bacteroides_clarus_YIT_12056_uid66155	629
genus	Bacteroides	Bacteroides_coprocola_DSM_17136_uid54879	722
genus	Bacteroides	Bacteroides_coprophilus_DSM_18228_uid55301	595
genus	Bacteroides	Bacteroides_dorei_5_1_36_D4_uid55593	893
genus	Bacteroides	Bacteroides_dorei_CL02T00C15_uid181625	998
genus	Bacteroides	Bacteroides_dorei_CL03T12C01_uid181626	875
genus	Bacteroides	Bacteroides_dorei_DSM_17855_uid54993	826
genus	Bacteroides	Bacteroides_eggerthii_1_2_48FAA_uid61869	1463
genus	Bacteroides	Bacteroides_eggerthii_DSM_20697_uid54989	668
genus	Bacteroides	Bacteroides_faecis_MAJ27_uid86875	1008
genus	Bacteroides	Bacteroides_finegoldii_CL09T03C10_uid181638	1556
genus	Bacteroides	Bacteroides_finegoldii_DSM_17565_uid54985	836
genus	Bacteroides	Bacteroides_fluxus_YIT_12057_uid66157	417
genus	Bacteroides	Bacteroides_fragilis_3_1_12_uid55575	938
genus	Bacteroides	Bacteroides_fragilis_638R_uid84217	788
genus	Bacteroides	Bacteroides_fragilis_CL03T00C08_uid181628	830
genus	Bacteroides	Bacteroides_fragilis_CL05T00C42_uid181632	1617
genus	Bacteroides	Bacteroides_fragilis_CL07T00C01_uid181630	875

Taxonomy level	Taxonomy name	Genome	Number of Contigs
genus	Bacteroides	Bacteroides_fragilis_HMW_610_uid181634	919
genus	Bacteroides	Bacteroides_fragilis_HMW_615_uid181635	909
genus	Bacteroides	Bacteroides_fragilis_HMW_616_uid181809	840
genus	Bacteroides	Bacteroides_fragilis_NCTC_9343_uid57639	839
genus	Bacteroides	Bacteroides_fragilis_YCH46_uid58195	834
genus	Bacteroides	Bacteroides_gallinarum_DSM_18171_uid199285	619
genus	Bacteroides	Bacteroides_intestinalis_DSM_17393_uid54881	1005
genus	Bacteroides	Bacteroides_massiliensis_B84634___DSM_17679_uid199226	748
genus	Bacteroides	Bacteroides_nordii_CL02T12C05_uid170043	968
genus	Bacteroides	Bacteroides_oleiciplenus_YIT_12058_uid182882	907
genus	Bacteroides	Bacteroides_ovatus_3_8_47FAA_uid68195	1069
genus	Bacteroides	Bacteroides_ovatus_ATCC_8483_uid54543	1082
genus	Bacteroides	Bacteroides_ovatus_CL02T12C04_uid181636	1223
genus	Bacteroides	Bacteroides_ovatus_CL03T12C18_uid181637	1119
genus	Bacteroides	Bacteroides_ovatus_SD_CMC_3f_uid46973	1096
genus	Bacteroides	Bacteroides_pectinophilus_ATCC_43243_uid54987	501
genus	Bacteroides	Bacteroides_plebeius_DSM_17135_uid54991	688
genus	Bacteroides	Bacteroides_salanitronis_DSM_18170_uid63269	686
genus	Bacteroides	Bacteroides_salyersiae_CL02T12C01_uid170041	919
genus	Bacteroides	Bacteroides_stercoris_ATCC_43183_uid54825	615
genus	Bacteroides	Bacteroides_thetaiotaomicron_VPI_5482_uid62913	958
genus	Bacteroides	Bacteroides_uniformis_ATCC_8492_uid54547	749
genus	Bacteroides	Bacteroides_uniformis_CL03T00C23_uid181639	789
genus	Bacteroides	Bacteroides_uniformis_CL03T12C37_uid181640	792
genus	Bacteroides	Bacteroides_vulgatus_ATCC_8482_uid58253	843
genus	Bacteroides	Bacteroides_vulgatus_CL09T03C04_uid181641	795
genus	Bacteroides	Bacteroides_vulgatus_PC510_uid47771	784
genus	Bacteroides	Bacteroides_xylanisolvens_CL03T12C04_uid181622	987
genus	Bacteroides	Bacteroides_xylanisolvens_XB1A_uid197168	884
genus	Barnesiella	Barnesiella_intestinihominis_YIT_11860_uid175259	557
genus	Bifidobacterium	Bifidobacterium_12_1_47BFAA_uid61873	421
genus	Bifidobacterium	Bifidobacterium_adolescentis_ATCC_15703_uid58559	368
genus	Bifidobacterium	Bifidobacterium_adolescentis_L2_32_uid54549	363
genus	Bifidobacterium	Bifidobacterium_bifidum_BGN4_uid167988	370
genus	Bifidobacterium	Bifidobacterium_bifidum_IPLA_20015_uid180937	410
genus	Bifidobacterium	Bifidobacterium_catenulatum_DSM_16992_uid55369	345
genus	Bifidobacterium	Bifidobacterium_longum_1_6B_uid180435	452
genus	Bifidobacterium	Bifidobacterium_longum_2_2B_uid180437	452
genus	Bifidobacterium	Bifidobacterium_longum_BBMN68_uid60163	332
genus	Bifidobacterium	Bifidobacterium_longum_infantis_ATCC_55813_uid55465	383
genus	Bifidobacterium	Bifidobacterium_pseudocatenulatum_DSM_20438_uid55303	371
genus	Bilophila	Bilophila_4_1_30_uid72973	660
genus	Bilophila	Bilophila_wadsworthia_3_1_6_uid61875	708
genus	Blautia	Blautia_hansenii_DSM_20583_uid55275	464
genus	Blautia	Blautia_hydrogenotrophica_DSM_10507_uid54939	573
genus	Bryantella	Bryantella_formatexigens_DSM_14469_uid54943	396
order	Burkholderiales	Burkholderiales_bacterium_1_1_47_uid51545	420
genus	Butyrivibrio	Butyrivibrio_crossotus_DSM_2876_uid55091	438
genus	Catenibacterium	Catenibacterium_mitsuokai_DSM_15897_uid54829	437
order	Clostridiales	Clostridiales_bacterium_1_7_47FAA_uid55287	1043
order	Clostridiales	Clostridiales_bacterium_OBRC5_5_uid175258	357
genus	Clostridium	Clostridium_D5_uid63427	875
genus	Clostridium	Clostridium_HGF2_uid61051	678
genus	Clostridium	Clostridium_L2_50_uid54559	475
genus	Clostridium	Clostridium_M62_1_uid54557	661
genus	Clostridium	Clostridium_SS2_1_uid54553	521
genus	Clostridium	Clostridium_SY8519_uid68705	154
genus	Clostridium	Clostridium_asparagiforme_DSM_15981_uid55115	955
genus	Clostridium	Clostridium_bartlettii_DSM_16795_uid54809	491
genus	Clostridium	Clostridium_bolteae_ATCC_BAA_613_uid54523	1116
genus	Clostridium	Clostridium_cf_saccharolyticum_K10_uid197201	585
genus	Clostridium	Clostridium_citroniae_WAL_17108_uid76581	1051

Taxonomy level	Taxonomy name	Genome	Number of Contigs
genus	Clostridium	Clostridium_clostridioforme_2_1_49FAA_uid76955	896
genus	Clostridium	Clostridium_difficile_ATCC_43255_uid67589	144
genus	Clostridium	Clostridium_difficile_NAP08_uid49121	93
genus	Clostridium	Clostridium_difficile_QCD_63q42_uid54949	59
genus	Clostridium	Clostridium_difficile_R20291_uid40921	43
genus	Clostridium	Clostridium_hathewayi_DSM_13479_uid55373	1135
genus	Clostridium	Clostridium_leptum_DSM_753_uid54605	578
genus	Clostridium	Clostridium_methylpentosum_DSM_5476_uid55281	554
genus	Clostridium	Clostridium_nexile_DSM_1787_uid55077	600
genus	Clostridium	Clostridium_ramosum_DSM_1402_uid54811	499
genus	Clostridium	Clostridium_saccharolyticum_WM1_uid51419	178
genus	Clostridium	Clostridium_spiroforme_DSM_1552_uid54607	404
genus	Clostridium	Clostridium_symbiosum_WAL_14163_uid63097	861
genus	Clostridium	Clostridium_symbiosum_WAL_14673_uid63157	770
genus	Collinsella	Collinsella_aerofaciens_ATCC_25986_uid54525	392
genus	Coprobacillus	Coprobacillus_29_1_uid62161	629
genus	Coprobacillus	Coprobacillus_8_2_54BFAA_uid82733	594
genus	Coprococcus	Coprococcus_ART55_1_uid197176	442
genus	Coprococcus	Coprococcus_catus_GD_7_uid197174	569
genus	Coprococcus	Coprococcus_comes_ATCC_27758_uid54883	534
genus	Coprococcus	Coprococcus_eutactus_ATCC_27759_uid54541	534
genus	Desulfovibrio	Desulfovibrio_piger_ATCC_29098_uid54519	454
genus	Dialister	Dialister_invisus_DSM_15470_uid55761	313
genus	Dialister	Dialister_succinatiphilus_YIT_11850_uid81763	359
genus	Dorea	Dorea_formicigenerans_4_6_53AFAA_uid73033	604
genus	Dorea	Dorea_formicigenerans_ATCC_27755_uid54513	519
genus	Dorea	Dorea_longicatena_DSM_13814_uid54515	494
genus	Eggerthella	Eggerthella_1_3_56FAA_uid61877	539
genus	Enterococcus	Enterococcus_faecium_1_230_933_uid55701	418
genus	Enterococcus	Enterococcus_faecium_Aus0085_uid214432	272
family	Erysipelotrichaceae	Erysipelotrichaceae_bacterium_2_2_44A_uid73037	825
family	Erysipelotrichaceae	Erysipelotrichaceae_bacterium_3_1_53_uid59459	742
family	Erysipelotrichaceae	Erysipelotrichaceae_bacterium_5_2_54FAA_uid46995	503
family	Erysipelotrichaceae	Erysipelotrichaceae_bacterium_6_1_45_uid181401	780
genus	Escherichia	Escherichia_1_1_43_uid55599	745
genus	Escherichia	Escherichia_3_2_53FAA_uid55601	839
genus	Escherichia	Escherichia_4_1_40B_uid55603	764
genus	Escherichia	Escherichia_coli_042_uid161985	874
genus	Escherichia	Escherichia_coli_07798_uid181911	778
genus	Escherichia	Escherichia_coli_0_1288_uid181931	844
genus	Escherichia	Escherichia_coli_2362_75_uid60613	843
genus	Escherichia	Escherichia_coli_AA86_uid179782	794
genus	Escherichia	Escherichia_coli_JJ1886_uid226103	862
genus	Escherichia	Escherichia_coli_KTE150_uid184560	801
genus	Eubacterium	Eubacterium_3_1_31_uid81761	884
genus	Eubacterium	Eubacterium_biforme_DSM_3989_uid55117	365
genus	Eubacterium	Eubacterium_cylindroides_T2_87_uid197177	234
genus	Eubacterium	Eubacterium_dolichum_DSM_3991_uid54609	357
genus	Eubacterium	Eubacterium_eligens_ATCC_27750_uid59171	457
genus	Eubacterium	Eubacterium_hadrum_DSM_3319_uid183778	460
genus	Eubacterium	Eubacterium_hallii_DSM_3353_uid54535	580
genus	Eubacterium	Eubacterium_rectale_ATCC_33656_uid59169	614
genus	Eubacterium	Eubacterium_rectale_uid197161	509
genus	Eubacterium	Eubacterium_rectale_uid197162	594
genus	Eubacterium	Eubacterium_siraeum_DSM_15702_uid54603	467
genus	Eubacterium	Eubacterium_siraeum_V10Sc8a_uid197178	454
genus	Eubacterium	Eubacterium_siraeum_uid197160	424
genus	Eubacterium	Eubacterium_ventriosum_ATCC_27560_uid54517	479
genus	Faecalibacterium	Faecalibacterium_cf__prausnitzii_KLE1255_uid60645	494
genus	Faecalibacterium	Faecalibacterium_prausnitzii_A2_165_uid54551	536

Taxonomy level	Taxonomy name	Genome	Number of Contigs
genus	Faecalibacterium	Faecalibacterium_prausnitzii_L2_6_uid197183	518
genus	Faecalibacterium	Faecalibacterium_prausnitzii_M21_2_uid54555	509
genus	Faecalibacterium	Faecalibacterium_prausnitzii_uid197157	501
genus	Flavonifractor	Flavonifractor_plautii_ATCC_29863_uid80691	687
genus	Fusobacterium	Fusobacterium_gonidiaformans_ATCC_25563_uid55569	268
genus	Fusobacterium	Fusobacterium_mortiferum_ATCC_9817_uid55571	393
genus	Gordonibacter	Gordonibacter_pamelaeae_7_10_1_b_uid197167	462
genus	Haemophilus	Haemophilus_parainfluenzae_ATCC_33392_uid63561	323
genus	Haemophilus	Haemophilus_parainfluenzae_HK2019_uid180434	365
genus	Haemophilus	Haemophilus_parainfluenzae_T3T1_uid72801	337
genus	Holdemania	Holdemania_filiformis_DSM_12042_uid55297	572
genus	Johnsonella	Johnsonella_ignava_ATCC_51276_uid77897	123
genus	Klebsiella	Klebsiella_4_1_44FAA_uid80417	863
family	Lachnospiraceae	Lachnospiraceae_bacterium_1_1_57FAA_uid68209	502
family	Lachnospiraceae	Lachnospiraceae_bacterium_1_4_56FAA_uid68205	521
family	Lachnospiraceae	Lachnospiraceae_bacterium_2_1_58FAA_uid68203	502
family	Lachnospiraceae	Lachnospiraceae_bacterium_3_1_46FAA_uid66427	456
family	Lachnospiraceae	Lachnospiraceae_bacterium_3_1_57FAA_CT1_uid68201	1191
family	Lachnospiraceae	Lachnospiraceae_bacterium_4_1_37FAA_uid63581	461
family	Lachnospiraceae	Lachnospiraceae_bacterium_5_1_63FAA_uid61883	521
family	Lachnospiraceae	Lachnospiraceae_bacterium_7_1_58FAA_uid81607	951
family	Lachnospiraceae	Lachnospiraceae_bacterium_9_1_43BFAA_uid66425	433
genus	Lactobacillus	Lactobacillus_acidophilus_30SC_uid63605	343
genus	Lactobacillus	Lactobacillus_amylovorus_GRL1118_uid160233	331
genus	Lactobacillus	Lactobacillus_johnsonii_DPC_6026_uid162057	329
genus	Lactobacillus	Lactobacillus_reuteri_DSM_20016_uid58471	295
genus	Lactobacillus	Lactobacillus_rossiae_DSM_15814_uid199472	41
genus	Lactobacillus	Lactobacillus_ruminis_ATCC_25644_uid179878	396
genus	Lactobacillus	Lactobacillus_ruminis_SPM0211_uid67955	337
genus	Lactococcus	Lactococcus_raffinolactis_4877_uid178066	623
genus	Megamonas	Megamonas_funiformis_YIT_11815_uid82999	415
genus	Megamonas	Megamonas_hypermegale_uid197163	335
genus	Megasphaera	Megasphaera_elsdenii_DSM_20460_uid71135	397
genus	Methanobrevibacter	Methanobrevibacter_smithii_ATCC_35061_uid58827	281
genus	Methanobrevibacter	Methanobrevibacter_smithii_DSM_2374_uid55123	277
genus	Methanobrevibacter	Methanobrevibacter_smithii_DSM_2375_uid54983	247
genus	Methanomassiliicoccus	Methanomassiliicoccus_Mx1_Issoire_uid207287	318
genus	Mitsuokella	Mitsuokella_multacida_DSM_20544_uid55073	399
genus	Odoribacter	Odoribacter_splanchnicus_DSM_20712_uid63397	693
genus	Oscillibacter	Oscillibacter_ruminantium_GH1_uid199475	299
genus	Parabacteroides	Parabacteroides_D13_uid55997	853
genus	Parabacteroides	Parabacteroides_distasonis_ATCC_8503_uid58301	783
genus	Parabacteroides	Parabacteroides_distasonis_CL03T12C09_uid181647	805
genus	Parabacteroides	Parabacteroides_goldsteinii_CL02T12C30_uid178274	1064
genus	Parabacteroides	Parabacteroides_johnsonii_CL02T12C29_uid181649	701
genus	Parabacteroides	Parabacteroides_johnsonii_DSM_18315_uid55269	673
genus	Parabacteroides	Parabacteroides_merdae_ATCC_43184_uid54545	681
genus	Parabacteroides	Parabacteroides_merdae_CL03T12C32_uid181650	768
genus	Parabacteroides	Parabacteroides_merdae_CL09T00C40_uid181721	682
genus	Paraprevotella	Paraprevotella_clara_YIT_11840_uid76949	673
genus	Paraprevotella	Paraprevotella_xylaniphila_YIT_11841_uid66381	617
genus	Parasutterella	Parasutterella_excrementihominis_YIT_11859_uid66383	495

Taxonomy level	Taxonomy name	Genome	Number of Contigs
genus	Phascolarctobacterium	Phascolarctobacterium_YIT_12067_uid62745	383
genus	Porphyromonas	Porphyromonas_somerae_DSM_23386_uid199036	411
genus	Prevotella	Prevotella_bivia_DSM_20514_uid182041	396
genus	Prevotella	Prevotella_copri_DSM_18205_uid55277	612
genus	Prevotella	Prevotella_disiens_FB035_09AN_uid51531	512
genus	Prevotella	Prevotella_stercorea_DSM_18206_uid78321	489
genus	Roseburia	Roseburia_hominis_A2_183_uid73419	588
genus	Roseburia	Roseburia_intestinalis_L1_82_uid55267	704
genus	Roseburia	Roseburia_intestinalis_XB6B4_uid197179	638
genus	Roseburia	Roseburia_intestinalis_uid197164	669
genus	Roseburia	Roseburia_inulinivorans_DSM_16841_uid55375	746
family	Ruminococcaceae	Ruminococcaceae_bacterium_D16_uid52825	539
genus	Ruminococcus	Ruminococcus_5_1_39BFAA_uid55629	595
genus	Ruminococcus	Ruminococcus_JC304_uid199662	506
genus	Ruminococcus	Ruminococcus_bromii_uid197158	399
genus	Ruminococcus	Ruminococcus_champanellensis_18P13_uid197169	421
genus	Ruminococcus	Ruminococcus_gnavus_ATCC_29149_uid54537	553
genus	Ruminococcus	Ruminococcus_lactaris_ATCC_29176_uid54903	489
genus	Ruminococcus	Ruminococcus_obeum_ATCC_29174_uid54509	612
genus	Ruminococcus	Ruminococcus_obeum_uid197165	610
genus	Ruminococcus	Ruminococcus_torques_ATCC_27756_uid54511	435
genus	Ruminococcus	Ruminococcus_torques_uid197166	527
genus	Ruminococcus	Ruminococcus_uid197156	560
genus	Streptococcus	Streptococcus_parasanguinis_ATCC_15912_uid49313	356
genus	Streptococcus	Streptococcus_parasanguinis_ATCC_903_uid62541	345
genus	Streptococcus	Streptococcus_parasanguinis_F0405_uid60563	364
genus	Streptococcus	Streptococcus_parasanguinis_F0449_uid180419	389
genus	Streptococcus	Streptococcus_salivarius_CCHSS3_uid70481	393
genus	Streptococcus	Streptococcus_salivarius_JIM8777_uid162145	362
genus	Streptococcus	Streptococcus_salivarius_K12_uid180858	413
genus	Streptococcus	Streptococcus_salivarius_M18_uid179893	378
genus	Streptococcus	Streptococcus_salivarius_SK126_uid55863	335
genus	Streptococcus	Streptococcus_thermophilus_CNCM_I_1630_uid179896	307
genus	Streptococcus	Streptococcus_thermophilus_CNRZ1066_uid58221	280
genus	Streptococcus	Streptococcus_vestibularis_ATCC_49124_uid62665	279
genus	Streptococcus	Streptococcus_vestibularis_F0396_uid60559	314
genus	Subdoligranulum	Subdoligranulum_4_3_54A2FAA_uid80415	702
genus	Subdoligranulum	Subdoligranulum_variabile_DSM_15176_uid54539	486
genus	Sutterella	Sutterella_wadsworthensis_2_1_59BFAA_uid181408	462
genus	Sutterella	Sutterella_wadsworthensis_3_1_45B_uid62165	543
genus	Synergistes	Synergistes_3_1_syn1_uid80429	506
genus	Tannerella	Tannerella_6_1_58FAA_CT1_uid80413	616
genus	Veillonella	Veillonella_3_1_44_uid47845	340
genus	Veillonella	Veillonella_6_1_27_uid47835	350
genus	Veillonella	Veillonella_ACP1_uid172974	354
genus	Veillonella	Veillonella_atypica_ACS_049_V_Sch6_uid51525	363
genus	Veillonella	Veillonella_dispar_ATCC_17748_uid55331	338
genus	Veillonella	Veillonella_oral_taxon_158_F0412_uid61047	354
genus	Weissella	Weissella_confusa_LBAE_C39_2_uid168254	393
genus	Methanomethylophilus	archaeon_Mx1201_uid196597	264
order	Clostridiales	butyrate_producing_bacterium_SM4_1_uid197180	327
order	Clostridiales	butyrate_producing_bacterium_SS3_4_uid197159	525
genus	Anaerostipes	butyrate_producing_bacterium_SSC_2_uid197181	432