bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MicroPheno: Predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples

Ehsaneddin Asgari[1,3], Kiavash Garakani[1], Alice C. McHardy[3], and Mohammad R.K. Mofrad[1,2,*]

[1]Department of Bioengineering, University of California, Berkeley, CA 94720, USA

[2]Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA

[3]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, BS 38124, Germany

[*]{mofrad@berkeley.edu}

Motivation: Microbial communities play important roles in the function and maintenance of various biosystems, ranging from the human body to the environment. A major challenge in microbiome research is the classification of microbial communities of different environments or host phenotypes. The most common and cost-effective approach for such studies to date is 16S rRNA gene sequencing. Recent falls in sequencing costs have increased the demand for simple, efficient, and accurate methods for rapid detection or diagnosis with proved applications in medicine, agriculture, and forensic science. We describe a reference- and alignment-free approach for predicting environments and host phenotypes from 16S rRNA gene sequencing based on k-mer representations that benefits from a bootstrapping framework for investigating the sufficiency of shallow sub-samples. Deep learning methods as well as classical approaches were explored for predicting environments and host phenotypes.

Results: k-mer distribution of shallow sub-samples outperformed the computationally costly Operational Taxonomic Unit (OTU) features in the tasks of body-site identification and Crohn's disease prediction. Aside from being more accurate, using k-mer features in shallow sub-samples allows (i) skipping computationally costly sequence alignments required in OTU-picking, and (ii) provided a proof of concept for the sufficiency of shallow and short-length 16S rRNA sequencing for phenotype prediction. In addition, k-mer features predicted representative 16S rRNA gene sequences of 18 ecological environments, and 5 organismal environments with high macro-F1 scores of 0.88 and 0.87. For large datasets, deep learning outperformed classical methods such as Random Forest and SVM.

Availability: The software and datasets are available at https://llp.berkeley.edu/micropheno.

1 Introduction

Microbial communities have important functions relevant to supporting, regulating, and in some cases causing unwanted conditions (e.g. diseases) in their hosts/environments, ranging from organismal environments, like the human body, to ecological environments, like soil and water. These communities typically consist of a variety of microorganisms, including eukaryotes, archaea, bacteria and viruses. The human microbiome is the set of all microorganisms that live in close association with the human body. It is now widely believed that changes in the composition of our microbiomes correlate with numerous disease states, raising the possibility that manipulation of these communities may be used to treat diseases. Normally, the microbiota (particularly the intestinal microbiota) of the microbiome play several important roles in humans, which include: (i) prevention of pathogen growth, (ii) education and regulation of the host immune system, and (iii) providing energy substrates to the host [1]. Consequently, dysbiosis of the human microbiome can promote several diseases, including asthma [2, 3], irritable bowel syndrome [4, 5], Clostridium difficile infection [6], chronic periodontitis [7, 8], cutaneous leishmaniasis [9], obesity [10, 11], chronic kidney disease [12], Ulcerative colitis [13], and Crohns disease [14, 15]. The human microbiome appears to play a particularly important role in the development of Crohns disease. Crohns disease is an inflammatory bowel disease (IBD) with a prevalence of approximately 40 per 100,000 and 200 per 100,000 in children and adults, respectively [16].

Microbiomes present in the environment also serve important functions, such as nutrient cycling [17]. Due to differences in nutrient availability and environmental conditions, microbiomes in different environments are each characterized by unique structures and microbial compositions [18–21]. Furthermore, each microbiome plays a unique role in the environment. For instance, the ocean microbiome generates half of the primary production on Earth [19]. The soil microbiome surrounding the root of plants has a great impact on plant fertility and growth [22]. Additionally, analyzing microbial communities in drinking water is one of the primary concerns of the environmental sciences [18].

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The starting point in data collection for many of the above mentioned projects is 16S rRNA gene sequencing of microbial samples [23]. There are a number of features that make the 16S rRNA gene ideal for use as a taxonomic 'fingerprint' for microbiome composition characterization. First, the 16S rRNA gene is highly conserved across bacteria and archaea. Secondly, the gene consists of both conserved regions, for which universal species-independent PCR primers may be directed against, and nine hypervariable regions (V1-V9), along which species-specific sequence differences may accumulate to allow for differential identification of species [24]. Sequencing of specific hypervariable regions of the 16S rRNA gene may be performed for determination of microbial community composition [25]. After the 16S rRNA hypervariable regions are sequenced from a microbial sample, the obtained sequences are then processed using bioinformatics software (such as QIIME [26, 27], Mothur [28], or Usearch [29]]) and clustered into groups of closely related sequences referred to as Operational Taxonomic Units (OTUs), which may then be used to assist in the functional profiling of microbial samples. Later in§1.2 we discuss the pros and cons of OTU features in details.

1.1 Machine learning for environments or host phenotypes classification

Several recent studies predicted the environment or host phenotypes using 16S gene sequencing data for body-sites [30, 31], disease state [32–35], ecological environment quality status prediction [36], and subject prediction for forensic science [37, 38]. In all, OTUs served as the main input feature for the down stream machine learning algorithms. Random Forest and then, ranking second, linear Support Vector Machine (SVM) classifiers were reported as the most effective classification approaches in these studies [31, 39, 34].

Related prior work on body-site classification [30, 31] used the following datasets: Costello Body Habitat (CBH - 6 classes), Costello Skin Sites (CSS - 12 classes) [40], and Pei Body Site (PBS - 4 classes) [31]. An extensive comparison of classifiers for body-site classification over CBH, CSS, and PBS on top of OTU features has been performed by Statnikov et al [31]. The best accuracy levels measured by relative classifier information (RCI) achieved by using OTU features are reported as 0.784, 0.681, and 0.647 for CBH, CSS, and RCI respectively. Due to the insufficiency of the number of samples (on average 57 samples per class for CBH, CSS, and PBS) as well as the unavailability of raw sequences for some of the datasets mentioned above, instead of using the same dataset we replicate the state-of-the-art approach suggested in [31], i.e. Random Forest and SVM over OTU features for a larger dataset (Human Microbiome Project dataset). We then compare OTU features with k-mer representations. Working on a larger dataset allows for a more meaningful investigation and better training for deep learning approaches.

Detecting disease status based on 16S gene sequencing is becoming more and more popular, with applications in the prediction of Type 2 Diabetes [35] (patients: 53 samples, healthy:43 - Best accuracy: 0.67), Psoriasis (151 samples for 3 classes - Best accuracy: 0.225), IBD (patients: 49 samples, healthy:59 - Best AUC:0.95) [32], (patients: 91 samples, healthy: 58 samples - Best AUC:0.92) [33]. Similar to body-site classification datasets, the datasets used for disease prediction were also relatively small. In this paper, we use the Crohn's disease dataset [14] with 1359 samples (patients: 731 samples, negative class: 628 samples) for evaluating our proposed method and then compare it with the use of OTU features.

We focus on machine learning approaches for classification of environments or host phenotypes of 16S rRNA gene sequencing data, which is the most popular and cost-effective sequencing method for the characterization of microbiome to date [39]. Studies on the use of machine learning for predicting microbial phenotype instead of environments/host phenotype [41, 42], as well as predictions based on shotgun metagenomics and whole-genome microbial sequencing are beyond the scope of this paper, although we believe that one may easily adapt the proposed approach to shotgun metagenomics, similar to the study by Cui et al. on IBD prediction [43].

Recently, deep learning methods became popular in various applications of machine learning in bioinformatics [44, 45] and in particular in metagenomics [46]. However, to the best of our knowledge this paper is the first study exploring environment and host phenotype prediction from 16S rRNA gene sequencing data using deep learning approaches.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1.2 16S rRNA gene sequence representations

OTU representation

As reviewed in§1.1, prior machine learning works on environment/host phenotype prediction have been mainly using OTU representations as the input features to the learning algorithm. Almost all popular 16S rRNA gene sequence processing pipelines cluster sequences into OTUs based on their sequence similarities utilizing a variety of algorithms [47, 27]. QIIME allows OTU-picking using three different strategies: **(i) closed-reference OTU-picking:** sequences are compared against a marker gene database (e.g. Greengenes [48] or SILVA [49]) to be clustered into OTUs and then the sequences different from the reference genomes beyond a certain sequence identity threshold are discarded. **(ii) open-reference OTU-picking:** the remaining sequences after a closed-reference calling go through a de novo clustering. This allows for using the whole sequences as well as capturing sequences belonging to new communities which are absent in the reference databases [50]. **(iii) pure de novo OTU-picking:** sequences (or reads) are only compared among themselves and no reference database is used. The third strategy is more appropriate for novel species absent in the current reference. Although OTU clustering reduces the analysis of millions of reads to working with only thousands of OTUs and simplifies the subsequent phylogeny estimation and multiple sequence alignment, OTU representations have several shortcomings: **(i)** All three OTU-picking strategies involve massive amounts of sequence alignments either to the reference genomes (in closed/opened-reference strategies) or to the sequences present in the sample (in open-reference and de novo strategies) which makes them very expensive [51] in comparison with reference-free/alignment-free representations. **(ii)** Overall sequence similarity is not a proper condition for grouping sequences and OTUs can be phylogenetically incoherent. For instance, a single mutation between two sequences is mostly ignored by OTU-picking algorithms. However, if the mutation does not occur within the sample, it might be a signal for assigning a new group. In addition, several mutations within a group most likely are not going to be tolerated by OTU-picking algorithms. However, having the same ratio across samples may suggest that the mutated sequences belong to the same group [52, 47]. **(iii)** The number of OTUs and even their contents are very sensitive to the pipeline and parameters, and this makes them difficult to reproduce [53].

k-mer representations

k-mer count vectors have been shown to be successful input features for performing machine learning on biological sequences for a variety of bioinformatics tasks [54]. In particular, recently k-mer count features have been largely used for taxonomic classifications of microbial metagenomics sequences [55–60]. However, to the best of our knowledge, k-mer features have not been explored for phenotypical and environmental characterizations of 16S rRNA gene sequencing. The advantages of using k-mer features over OTUs are later discussed in the discussion section.

In this paper we propose using shallow sub-samples for predicting environments and host phenotypes from 16S rRNA sequencing. Our approach is fast, reference-free and alignment-free, while contributing in building accurate classifiers outperforming conventional OTU features in body-site identification and Crohn's disease classification. We propose a bootstrapping framework to investigate the sufficiency of shallow sub-samples for the prediction of the phenotype of interest, which proves the sufficiency of short-length and shallow sequencing of 16S rRNA. In addition, we explore deep learning methods as well as classical approaches for the classification and show that for large datasets, deep learning can outperform classical methods. Furthermore, we explore PCA, t-SNE, and supervised deep representation learning for visualization of microbial samples/sequences of different phenotypes. We also show that k-mer features can be used to predict representative 16S rRNA gene sequences belonging to 18 ecological environments and 5 organismal environments with high macro-F1s.

2 Material and Methods

2.1 Datasets

Body-site identification

We employ the metagenomic 16S rRNA gene sequence dataset provided by the NIH Human Microbiome Project (HMP) [61, 62][1]. In particular, we use processed, annotated 16S rRNA gene sequences of up

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to 300 healthy individuals, each sampled at 4 major body-sites (oral, airways, gut, vagina) and up to three time points. For each major body-site, a number of sub-sites were sampled. We focus on 5 body sub-sites: anterior nares (nasal) with 295 samples, saliva (oral) with 299 samples, stool (gut) with 325 samples, posterior fornix (urogenital) with 136 samples, and mid vagina (urogenital) with 137 samples, in total 1192 samples. These body-sites are selected to represent differing levels of spatial and biological proximity to one another, based on relevance to pertinent human health conditions potentially influenced by the human microbiome. In order to compare k-mer based approach with state-of-the-art OTU features we collect the closed-reference OTU representations of the same samples in HMP [62] [2] obtained using the QIIME pipeline [50].

Crohn's disease prediction

For the classification of Crohns disease, we use the metagenomics 16S rRNA gene sequence dataset described in [14][3], which is currently the largest pediatric Crohns disease dataset available. This dataset included annotated 16S rRNA gene sequence data for 731 pediatric ($\leq$ 17 years old) patients with Crohns disease and 628 samples verified as healthy or diagnosed with other diseases, making a total of 1359 samples. The 16S rRNA sequencing dataset was targeted towards the V4 hypervariable region of the 16S rRNA gene. Similar to the body-site dataset, in order to compare the k-mer based approach with the approach based on OTU features we collect the OTU representations of the same samples from Qiita repository [4] obtained using QIIME pipeline [50].

Prediction of the environment for representative 16S rRNA gene sequences

MetaMetaDB provides a comprehensive dataset of representative 16S rRNA gene sequences of various ecological and organismal environments collected from existing 16S rRNA databases spanning almost 181 million raw sequences. In the MetaMetaDB pipeline, low-quality nucleotides, adaptors, ambiguous sequences, homopolymers, duplicates, and reads shorter than 200bp, as well as chimera have been removed and finally 16S rRNA sequences are clustered with 97% identity generating 1,241,213 representative 16S rRNA sequences marked by their environment [63]. MetaMetaDB divides its ecological environments into 34 categories and its organismal environments into 28 categories. We create three datasets which are subsets of MetaMetaDB to investigate the discriminative power of k-mers in predicting microbial habitability. Since the sequences in MetaMetaDB were already filtered and semi-identical sequences were removed, OTU-picking would not be relevant as it would result in an almost one-to-one mapping between the sequences and OTUs (we verified this using QIIME).

Ecological environment prediction: MetaMetaDB is imbalanced in terms of the number of representative sequences per environment. For this study, we pick the ecological environments with more than 10,000 samples, ending up with 18 classes of ecological environments: activated sludge, ant fungus garden, aquatic, bioreactor, bioreactor sludge, compost, food, food fermentation, freshwater, freshwater sediment, groundwater, hot springs, hydrocarbon, marine, marine sediment, rhizosphere, sediment, and soil [5]. We make two datasets out of the sequences in these environments: **ECO-18K** containing 1000 randomly selected instances per class (a total of 18K sequences) and **ECO-180K**, which is 10 times larger than **ECO-18K**, i.e. contains 10,000 randomly selected instances per class (a total of 180K sequences).

Organismal environment prediction: from the organismal environments in MetaMetaDB we select a subset containing gut microbiomes of 5 different organisms (bovine gut, chicken gut, human gut, mouse gut, termite gut) and down-sample each class to the size of the smallest class ending up having 620 sample per class (in total containing 3100 sequences) we call this dataset **5GUTS-3100**.

2.2 MicroPheno computational workflow

We describe using deep learning and classical methods for classification of the environments or host phenotypes of microbial communities using k-mer frequency representations obtained from shallow sub-sampling of 16S rRNA gene sequences. We propose a bootstrapping framework to confirm the sufficiency of using a small portion of the sequences within a 16S rRNA sample for determining the underlying phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

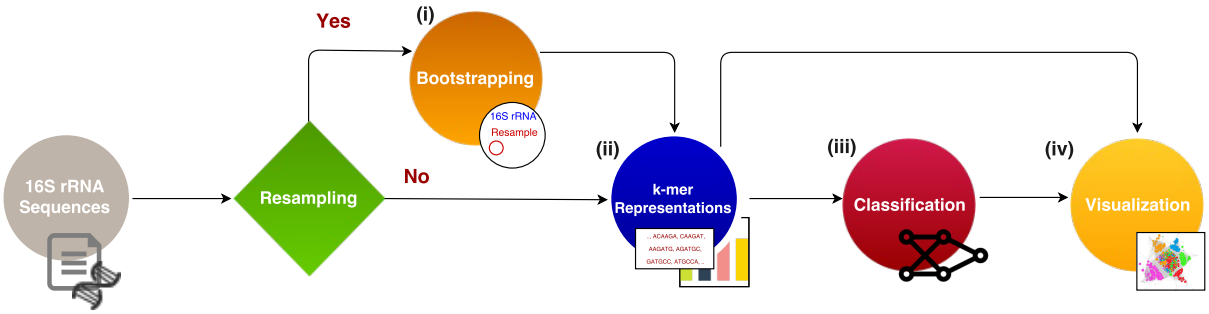

Figure 1: **The components and the data flow in the MicroPheno computational workflow**

The MicroPheno computational workflow has the following steps (Figure 1): (i) to find the proper size N for the sample, such that it stays representative of the data and produces a stable k-mer profile, the 16S rRNA sequences go through a bootstrapping phase. (ii) Afterwards, the sub-sampled sequences are used to find the best value for k for classification, to produce the k-mer representations of the samples. (iii) The k-mer representations are used for classification with Deep Neural Networks (DNN), Random Forest (RF), and Linear SVM. (iv) Finally, the k-mer representations as well as the supervised representations trained using DNNs are used for visualization of the 16S rRNA gene sequences or samples. In what follows, these steps are explained in detail.

Bootstrapping: Confirming the sufficiency of only a small portion of 16S rRNA sequences for environment or host phenotype classification is important because (i) sub-sampling reduces the preprocessing run-time, and (ii) more importantly, it proves that even a shallow 16S rRNA sequencing is enough. For this purpose we propose a resampling framework to give us quantitative measures for finding the proper sampling size. Let $\theta_k(X_i)$ be the normalized k-mer distribution of X_i, a set of sequences in the ith 16S rRNA sample. We investigate whether only a portion of X_i, which we represent as $\widetilde{x}_{ij}$, i.e. j^{th} resample of X_i with sample size N, would be sufficient for producing a proper representation of X_i. To quantitatively find a sufficient sample size for X_i we propose the following criteria in a resampling scheme. **(i) Self-consistency**: resamples for a given size N from X_i produce consistent $\theta_k(\widetilde{x}_{ij})$'s, i.e. resamples should have similar representations. **(ii) Representativeness**: resamples for a given size N from X_i produce $\theta_k(\widetilde{x}_{ij})$'s similar to $\theta_k(X_i)$, i.e. similar to the case where all sequences are used.

We quantitatively define self-inconsistency and unrepresentativeness and seek parameter values that minimize them. We measure the **self-inconsistency** ($\bar{D}_S$) of the resamples' representations by calculating the average Kullback Leibler divergence among normalized k-mer distributions for N_R resamples (here N_R=10) with sequences of size N from the ith 16S rRNA sample:

$$\bar{D}_{Si}(N,k,N_R) = \frac{1}{N_R(N_R-1)} \sum_{\substack{\forall p,q \\ (p \neq q)}} \in \{1,2,\cdots,N_R\} D_{KL}(\theta_k(\widetilde{x}_{ip}), \theta_k(\widetilde{x}_{iq})),$$

where $|\widetilde{x}_{il}| = N$; $\forall l \in \{1,2,\cdots,N_R\}$. We calculate the average of the values of $\bar{D}_{Si}(N,k,N_R)$ over the M different 16S rRNA samples:

$$\bar{D}_S(N,k,N_R) = \frac{1}{M} \sum_{i=1}^{M} \bar{D}_{Si}(N,k,N_R).$$

We measure the **unrepresentativeness** ($\bar{D}_R$) of the resamples by calculating the average Kullback Leibler divergence between normalized k-mer distributions for N_R resamples (N_R=10) with size N and using all the sequences in X_i for the ith 16S rRNA sample:

$$\bar{D}_{Ri}(N,k,N_R) = \frac{1}{N_R} \sum_{\forall p \in \{1,2,\cdots,N_R\}} D_{KL}(\theta_k(\widetilde{x}_{ip}), \theta_k(X_i)),$$

where $|\widetilde{x}_{il}| = N$; $\forall l \in \{1,2,\cdots,N_R\}$. We calculate the average over $\bar{D}_{Ri}(N,k)$'s for the M 16S rRNA samples:

$$\bar{D}_R(N,k,N_R) = \frac{1}{M} \sum_{i=1}^{M} \bar{D}_{Ri}(N,k,N_R).$$

For the experiments on body-site and the dataset for Crohn's disease, we measure self-inconsistency

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$\bar{D}_S$ and unrepresentativeness $\bar{D}_R$ for $N_R = 10$ and $M = 10$ for any $8 \geq k \geq 3$ with sampling sizes ranging from 20 to 10000. Each point in Figure 4 represents the average of 100 ($M \times N_R$) resamples belonging to M randomly selected 16S rRNA samples, each of which is resampled $N_R = 10$ times. Since in the ecological and organismal datasets each sample is a single sequence, the bootstrapping step is skipped.

k-mer representation: We propose $l1$ normalized k-mer distribution of 16S rRNA gene sequences as input features for the down-stream machine learning classification algorithms as well as visualization. Normalizing the representation allows for having a consistent representation even when the sampling size is changed. For each k-value we pick a sampling size that gives us a self-consistent and representative representation measured by $\bar{D}_S(N,k,N_R)$ and $\bar{D}_R(N,k,N_R)$, respectively, as explained above.

Classification: Random Forests and linear SVM are the state-of-the-art classical approaches for categorical prediction on 16S rRNA sequencing [31, 39, 34] and in general for many machine learning problems in bioinformatics [64]. These two approaches, which are respectively instances of non-linear and linear classifiers, are both adopted in this study. In addition to these classical approaches, we also evaluate the performance of deep Neural Network classifiers in predicting environments and host phenotypes.

We evaluate and tune the model parameter in a stratified 10 fold cross-validation scheme. To ensure optimizing for both precision and recall, we optimize the classifiers for the harmonic mean of precision and recall, i.e. F1. In particular, to give equal importance to the classification categories, specifically when we have imbalanced classes, we use macro-F1, which is the average of F1's over categories. Finally the evaluation metrics are averaged over the folds and the standard deviation is also reported.

Classical learning algorithms: We use a one-versus-rest strategy for multi-class linear SVM [65] and tune parameter C, the penalty term for regularization. Random Forest [66] classifiers are tuned for (i) the number of decision trees in the ensemble, (ii) the number of features for computing the best node split, and (iii) the function to measure the quality of a split.

Deep learning: We use the Multi-Layer-Perceptrons (MLP) Neural Network architecture with several hidden layers using Rectified Linear Unit (ReLU) as the nonlinear activation function. We use the softmax activation function at the last layer to produce the probability vector that can be regarded as representing posterior probabilities [67]. To avoid overfitting, we perform early stopping and also use dropout at hidden layers [68]. A schematic visualization of our Neural Networks is depicted in Figure 2. Our objective is minimizing the loss, i.e. cross entropy between output and the one-hot vector representation of the target class. The error (the distance between the output and the target) is used to update the network parameters via a Back-propagation algorithm using Adaptive Moment Estimation (Adam) as the optimizer [69]. We start with a single hidden layer and incrementally increase the number of layers with systematic exploration of the number of hidden units and dropout rates to find a proper architecture. We stop adding layers when increasing the number of layers does not result in achieving a higher macro-F1 anymore. In addition, for the visualization of samples we use the output of the $(n-1)^{th}$ hidden layer. Later in the results DNN-nL is a short form for a MLP Neural Network with n layers.

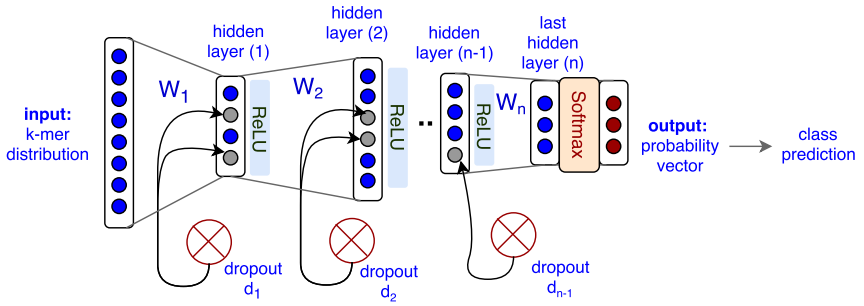

Figure 2: General architecture of the MLP Neural Networks that have been used in this study for multi-class classification of environment and host phenotypes.

Visualization: To project 16S rRNA sequencing samples to 2D for visualization purposes, we explore Principal Component Analysis (PCA) [70] as well as t-Distributed Stochastic Neighbor Embedding (t-SNE) [71] as instances of respectively linear and non-linear dimensionality reduction methods. In addition, we explore the use of supervised deep representation learning in visualization of data [72], i.e. we visualize the activation function of the last hidden layer of the Neural Network trained for prediction of environments or host phenotypes to be compared with unsupervised methods. More details on visualization methods are provided as supplementary materials.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Implementations: MicroPheno uses implementations of Random Forest, SVM, t-SNE, and PCA in the Python library scikit-learn [73] and Deep Neural Networks are implemented in the Keras[6] deep learning framework using TensorFlow back-end.

3 Results

In this section, the results are organized based on datasets. As discussed in Section 2.2 we have several choices in each step in the computational workflow: choosing the value of k in k-mer, the sampling rate, and the classifiers. To explore the parameter space more systematically we followed the steps demonstrated in Figure 3. (i) In the first step, for each value of $8 \geq k \geq 3$, we pick a stable sample size based on the output of bootstrapping. (ii) We perform the classification task using tuned Random Forest for different k values and their selected sampling sizes based on boostrapping. We selected Random Forest because we found it easy to tune and because it oftentimes outperforms linear SVM [31, 64]. (iii) As the third step, for a selected k we investigate the role of sampling size (N) in classification. (iv) Finally, we compare different classifiers for the selected k and N. We also compare the performance of our proposed k-mer features with that of OTU features in classification tasks.

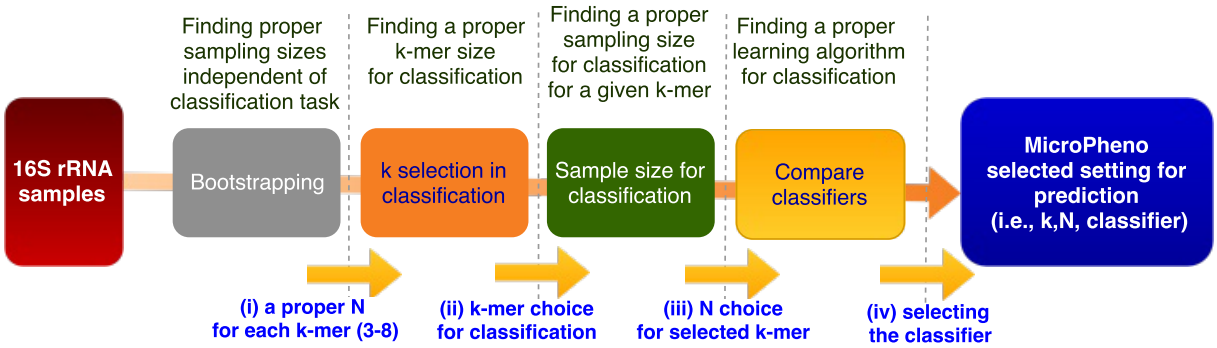

Figure 3: Steps we take to explore parameters for the representations and how we choose the classifier for prediction of the phenotype of interest in this study

3.1 Body-site identification

(i) Bootstrapping for sampling rate selection for k-mers: Higher k values require higher sampling rates to produce self-consistent and representative representations (Figure 4). For each k, we consider a certain threshold on $\bar{D_S}$ and $\bar{D_R}$ to ensure picking a sampling size resulting in self-consistent and representative representations.

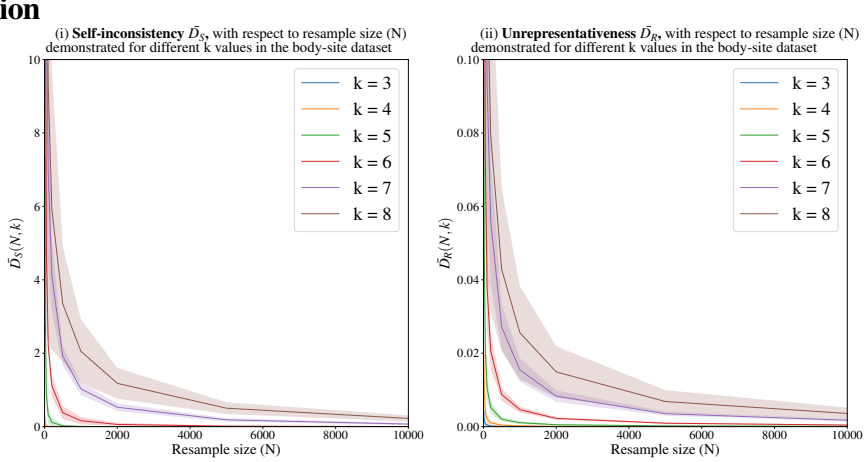

Figure 4: **Measuring (i) self-inconsistency ($\bar{D_S}$), and unrepresentativeness ($\bar{D_R}$)** for the body-site dataset. Each point represents an average of 100 resamples belonging to 10 randomly selected 16S rRNA samples. Higher k values require higher sampling rates to produce self-consistent and representative samples.

(ii) Classification for different values of k with a sampling size selected based on the output of bootstrapping: Interestingly, using only 3-mer features with a very low sampling rate (≈20/15000=0.0013) provides a relatively high performance for 5-way classification. The value of macro-F1 increases with the value of k from 3 to 6, but increasing k further than that does not have any additional effect on macro-F1 (Table 1, step (ii)).

(iii) Exploring the sampling size (N) for a selected k-mer: For a selected k-value (k=6), using the Random Forest classifier for different sampling sizes are presented in Table 1, step (iii). The results

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Step	Representation	Resample size/≈15000/fasta	Classifier	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
				Precision	Recall	F1	Precision	Recall	F1
(ii)	3-mers	20	RF	0.84 ± 0.02	0.84 ± 0.02	0.84 ± 0.02	0.75 ± 0.03	0.75 ± 0.03	0.74 ± 0.03
	4-mers	100		0.86 ± 0.03	0.86 ± 0.03	0.86 ± 0.03	0.77 ± 0.03	0.77 ± 0.03	0.77 ± 0.03
	5-mers	500		0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.82 ± 0.03	0.82 ± 0.03	0.82 ± 0.03
	6-mers	2000		$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.85 \pm 0.04}$	0.84 ± 0.05
	7-mers	5000		$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.85 \pm 0.05}$
	8-mers	8000		0.9 ± 0.03	0.9 ± 0.03	0.9 ± 0.03	0.85 ± 0.05	0.84 ± 0.05	0.84 ± 0.05
(iii)	6-mers	100	RF	0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.82 ± 0.04	0.82 ± 0.03	0.81 ± 0.03
		1000		0.9 ± 0.03	0.9 ± 0.03	0.9 ± 0.03	0.83 ± 0.04	0.83 ± 0.04	0.83 ± 0.04
		2000		$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.84 \pm 0.05}$
		5000		0.9 ± 0.03	0.9 ± 0.03	0.9 ± 0.03	$\mathbf{0.85 \pm 0.04}$	0.84 ± 0.04	$\mathbf{0.84 \pm 0.04}$
		10000		0.9 ± 0.03	0.9 ± 0.03	0.9 ± 0.03	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.05
		All sequences		0.9 ± 0.03	0.9 ± 0.03	0.9 ± 0.03	0.84 ± 0.05	0.84 ± 0.04	0.84 ± 0.05
(iv)	6-mers	5000	RF	$\mathbf{0.9 \pm 0.03}$	$\mathbf{0.9 \pm 0.03}$	$\mathbf{0.9 \pm 0.03}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.84 \pm 0.04}$	$\mathbf{0.84 \pm 0.04}$
			SVM	0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.76 ± 0.06	0.76 ± 0.03	0.74 ± 0.04
			DNN-5L	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.79 ± 0.02	0.79 ± 0.03	0.79 ± 0.02
-	6-mers	5000	DNN-4L (4 classes)	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$

Table 1: **The results for classification of major body-sites using k-mer representations.** The set of rows matches the steps (ii to iv) mentioned in Figure 3, i.e k-mer selection, N (sample size) selection, and finally selection of the classifier. The classifiers (Random Forest, Support Vector Machine and Neural Network classifiers) are tuned and evaluated in a stratified 10xfold cross-validation setting. The step column refers to the steps in Figure 3. The last row shows the Neural Network's performance in the classification of body-sites when the urogenital body-sites are combined.

suggest that changing the sampling size from 0.6% to 100% of the sequences will not change the classification results substantially, suggesting that in body-site identification, a very shallow sub-sampling of the sequences is sufficient for a reliable prediction. Using more sequences does not necessarily increase the discriminative power and may even result in over-fitting. We selected a sampling size of 5000 for 6-mers (the sampling size with the highest macro-F1 and the minimum standard deviation) for comparison between classifiers in the next step.

(iv) Comparison of classifiers for the selected N, k: In the body-site prediction task, the Random Forest classifier obtained the top macro-F1 (0.84) for this 5-way classification (Table 1, step (iv)). The confusion matrix in Figure 5 shows that the most difficult decision for the classifier is to distinguish between mid vagina and posterior fornix, both of which are urogenital body-sites. The visualizations of body-site samples obtained through using PCA, t-SNE, and t-SNE on the activation function of the last layer of the trained 5-layered Neural Network are presented in Figure 6. These results suggest that supervised training of representations using Neural Networks provides a non-linear transformation of data that can discriminate between dissimilar body-sites with reasonably accuracy. As shown in the last row of Table 1, combining the urogenital body-sites increases the macro-F1 to 0.99 ± 0.01 using the Neural Network.

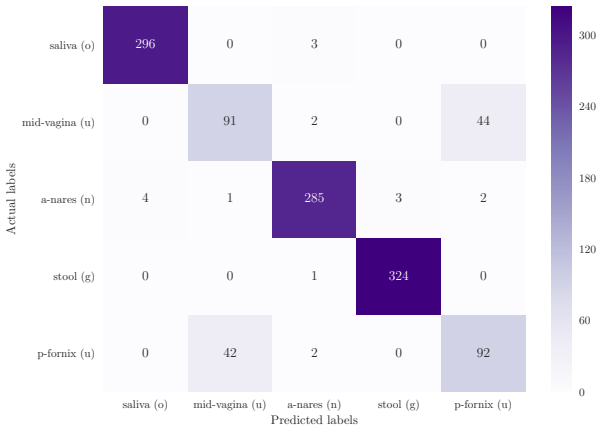

Figure 5: **The confusion matrix for the classification of 5 major body-sites**, using Random Forest classifier in a 10xfold cross-validation scheme. The presented body-sites are saliva (**o:** oral), mid-vagina (**u:** urogenital), anterior nares (**n:** nasal), stool (**g:** gut), and posterior fornix (**u:** urogenital).

Comparison of k-mer and OTU features in body-site identification: For a comparison between OTU features and k-mer representations in body-site identification, the Random Forest classifier (as an instance of non-linear classifier) and linear SVM (as an instance of linear classifier) were tuned and evaluated in a stratified 10xfold cross-validation setting. Our results suggest that for both k-mer features and OTUs, Random Forest is the best choice (Table 2). In addition, with almost $\frac{1}{5}$ of the size of OTU features and in spite of being considerably less expensive to calculate, k-mers marginally outperforms OTU features in body-site identification.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

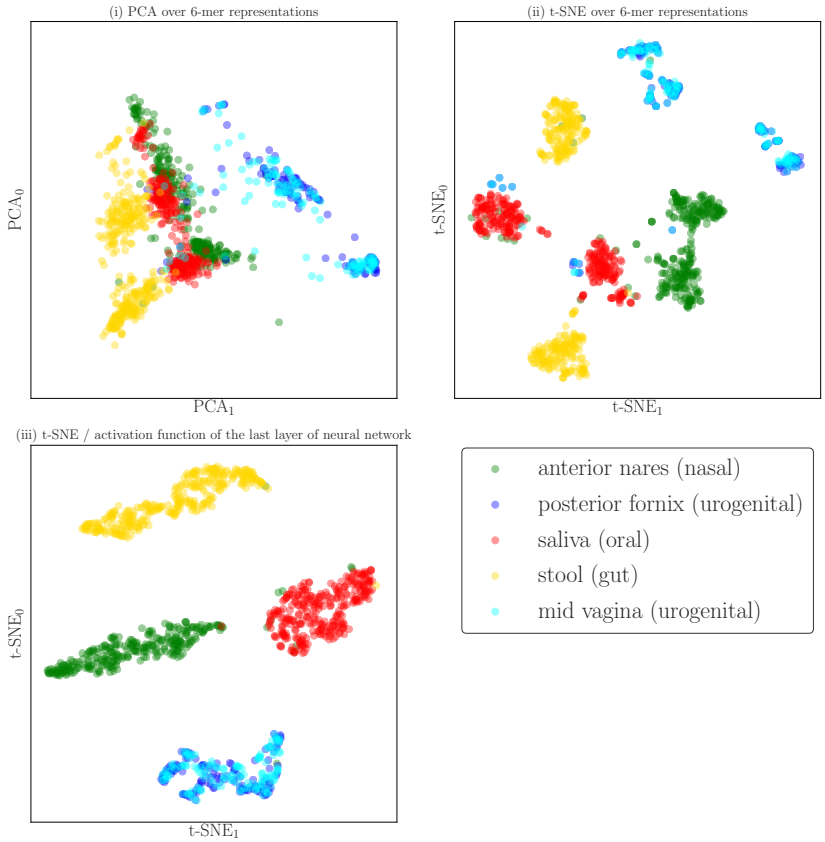

Figure 6: **Visualization of the body-site dataset using different projection methods:** (i) PCA over 6-mer distributions with unsupervised training, (ii) t-SNE over 6-mer distributions with unsupervised training, (iii) visualization of the activation function of the last layer of the trained Neural Network (projected to 2D using t-SNE).

Features	Classifiers	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
		Precision	Recall	F1	Precision	Recall	F1
6-mer features (size: 4096)	RF	**0.9 ± 0.03**	**0.9 ± 0.03**	**0.9 ± 0.03**	**0.85 ± 0.04**	**0.84 ± 0.04**	**0.84 ± 0.04**
	SVM	0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.76 ± 0.06	0.76 ± 0.03	0.74 ± 0.04
OTU features (size: 20589)	RF	0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.83 ± 0.03	0.83 ± 0.03	0.83 ± 0.03
	SVM	0.85 ± 0.03	0.85 ± 0.03	0.85 ± 0.03	0.77 ± 0.05	0.78 ± 0.04	0.76 ± 0.04

Table 2: Comparison of k-mer and OTU features in body-site identification.

3.2 Crohn's disease prediction

(i) Bootstrapping for sampling rate selection for k-mers: Similar to bootstrapping for the body-site dataset (Figure 4), $\bar{D}_S$ and $\bar{D}_R$ for different values of k with respect to sampling sizes were calculated. As the structure of the curve is similar to the body-site dataset, to avoid redundancy, the figure is provided as supplementary material.

(ii) Classification for different values of k with a sampling size selected based on the output of bootstrapping: Choosing k=6 with a sampling size of 2000 (≈2,000/38,000 = 0.05) provided a macro-F1 of 0.75 which is the minimum k with top performance (Table 3, step (ii)).

(iii) Exploring the sampling size (N) for a selected k-mer: For a selected k-value (k=6), using the Random Forest classifier for different sampling sizes while ncreasing the sampling size from 100 (100/38000=0.003) to 5000 (5000/38000=0.13) increased the macro-F1 from 0.7 to 0.75 (Table 3, step (iii)). However, using all sequences instead of 0.13 of them in each sample, did not increase the discriminative power.

(iv) Comparison of classifiers for the selected N, k: For selected values of k and N, the results of the Crohn's disease prediction using Random Forest, SVM, and Neural Network classifiers are provided (Table 3, step (iv)). The Random Forest classifier obtained the top macro-F1 (0.75) for this binary classification. The projection of the Crohn's disease dataset using with PCA and t-SNE over raw k-mer representations, as well as t-SNE over the representation learned though the supervised learning of the Neural Network are visualized in Figure 7. The trained Neural Network provides

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

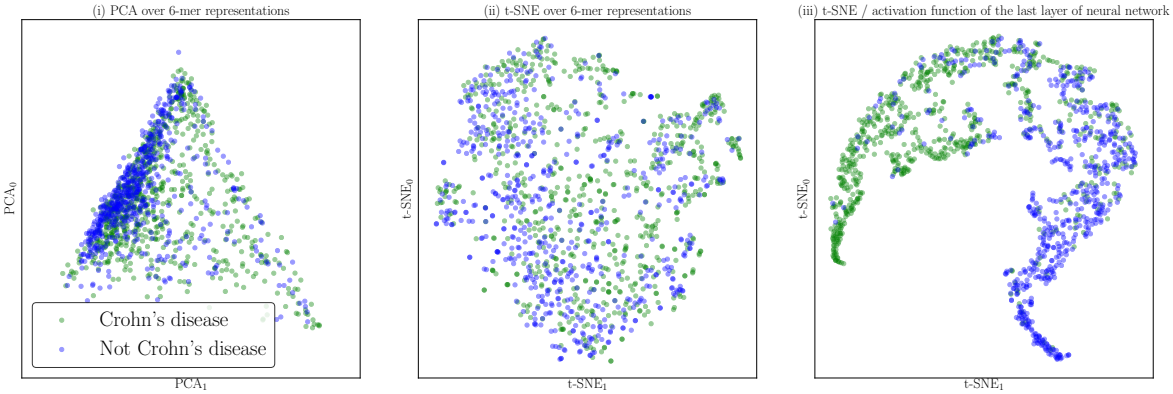

Figure 7: **Visualization of the Crohn's disease dataset using different projection methods**: (i) PCA over 6-mer distributions with unsupervised training, (ii) t-SNE over 6-mer distributions with unsupervised training, (iii) visualization of the activation function of the last layer of the trained Neural Network (projected to 2D using t-SNE).

Step	Representation	Resample size/≈38000/fasta	Classifier	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
				Precision	Recall	F1	Precision	Recall	F1
(ii)	3-mers	20	RF	0.62 ± 0.05	0.62 ± 0.05	0.62 ± 0.05	0.62 ± 0.05	0.61 ± 0.05	0.61 ± 0.05
	4-mers	100		0.7 ± 0.05	0.7 ± 0.05	0.7 ± 0.05	0.69 ± 0.05	0.69 ± 0.05	0.69 ± 0.05
	5-mers	500		0.74 ± 0.05	0.74 ± 0.05	0.74 ± 0.05	0.74 ± 0.05	0.74 ± 0.05	0.74 ± 0.05
	6-mers	2000		**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
	7-mers	5000		**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
	8-mers	8000		**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
(iii)	6-mers	100	RF	0.71 ± 0.04	0.71 ± 0.04	0.71 ± 0.04	0.71 ± 0.04	0.7 ± 0.04	0.7 ± 0.04
		1000		0.75 ± 0.04	0.75 ± 0.04	0.75 ± 0.04	0.76 ± 0.04	0.75 ± 0.04	0.75 ± 0.04
		2000		0.75 ± 0.04	0.75 ± 0.04	0.75 ± 0.04	0.75 ± 0.04	0.75 ± 0.04	0.75 ± 0.04
		5000		**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
		10000		**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.05**	**0.75 ± 0.04**	**0.75 ± 0.05**
		All sequences		**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.05**	**0.75 ± 0.04**	**0.75 ± 0.05**
(iv)	6-mers	5000	RF	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
			SVM	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.67 ± 0.04	0.67 ± 0.04
			DNN-7L	0.7 ± 0.02	0.7 ± 0.02	0.7 ± 0.02	0.7 ± 0.03	0.7 ± 0.02	0.7 ± 0.03

Table 3: **The results for classification of Crohn's disease prediction using k-mer representations.** The set of rows matches the steps (ii to iv) mentioned in Figure 3, i.e k-mer selection, N (sample size) selection, and finally selection of the classifier. The classifiers (Random Forest, Support Vector Machine and Neural Network classifiers) are tuned and evaluated in a stratified 10xfold cross-validation setting. The step column refers to the steps in Figure 3.

Features	Classifiers	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
		Precision	Recall	F1	Precision	Recall	F1
6-mer features (size: 4096)	RF	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
	SVM	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.67 ± 0.04	0.67 ± 0.04
OTU features (size: 9511)	RF	0.74 ± 0.04	0.74 ± 0.04	0.74 ± 0.04	0.74 ± 0.04	0.74 ± 0.04	0.74 ± 0.04
	SVM	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04

Table 4: **Comparison of k-mers and OTU features in the detection of the Crohn's disease phenotype**. For this comparison, Random Forest classifier (as an instance of non-linear classifiers) and linear SVM (as an instance of linear classifiers) have been used. The classifiers are tuned and evaluated in a stratified 10xfold cross-validation setting.

a non-linear transformation of data that can identify subjects diagnosed with Crohn's disease with reasonable performance.

Comparison of k-mer and OTU features in Crohn's disease classification: For a comparison between OTU features and our proposed k-mer representations in detecting Crohns disease from metagenomic samples, the Random Forest classifier (as an instance of non-linear classifiers) and linear SVM (as an instance of linear classifiers) were tuned and evaluated in a stratified 10xfold cross-validation. For both k-mer features and OTUs, the Random Forest classifier performed best (Table 4). In addition, even though only half of the number of features were used and in spite of being considerably less expensive to calculate, k-mers marginally outperformed OTU features in the detection of Crohns disease.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.3 Ecological environment prediction

(ii) Classification for different values of k: As stated before, for the ecological and organismal datasets we do not need to perform resampling as we classify single 16S rRNA gene representative sequences. We thus can skip steps (i) and (iii) (Figure 3). Step (ii) in Table 5 shows the effect of k in the performance of the classification of the 18 environments for the ECO-18K dataset. Using k=6 provides the best classification performance, with a macro-F1 of 0.73 which is relatively high for a 18-way classification (has a mere 0.06 chance of randomly being assigned correctly for balanced dataset).

(iv) Comparison of classifiers for the selected k: For selected values of k, the results of the environment prediction with the Random Forest, SVM, and Neural Network classifiers are provided (Table 5, step (iv), ECO-18K dataset). SVM classifier obtained the top macro-F1 (0.79) for 18-way classification. To see the effect of increasing the number of data points in classification performance we repeat the classifier comparison (step iv) for the ECO-180K dataset. The results are summarized in Table 5-ECO-18K dataset, showing that feeding more training instances results in better training for the deep learning approach, which outperformed the SVM and achieved a macro-F1 of 0.88, which is very high for a 18-way classification task. In training the Neural Networks for the ECO-18K dataset, increasing the number of hidden layers from 3 to more did not help result in improvements. However, using the ECO-180K dataset, which is 10 times larger, allowed us to train a deeper network and increased the macro-F1 by 5 percent going from 3 layers to 5 layers. Increasing the number of layers further did not result in any improvements.

Neural Network Visualization: Visualizations of representative 16S rRNA gene sequences in 18 ecological environments obtained through using PCA, t-SNE, and t-SNE on the activation function of the last layer of the trained Neural Network are presented in Figure 8. For ease of visualization, we randomly picked 100 samples per class. These results suggest that supervised training of representations using Neural Networks provides a non-linear transformation of data containing information about high-level similarities between environments in the sub-plot on the right (scatter plot (iii)), where such structures appeared in the visualization only when more hidden layers were used: **(i)** On the left, the environments containing water are clustered in a dense neighborhood: marine, aquatic, freshwater, hot springs, bioreactor sludge (described in the source: "Bioreactor sludge is usually the sludge inside the bioreactor that treats waste-water."), groundwater, and, surprisingly, rhizosphere (an environment where plants, soil, water, microorganisms, and nutrients meet and interact). **(ii)** In the middle, environments labeled related to sediment are found: sediment, freshwater sediment, marine sediment, and soil. **(iii)** Environments containing food, like food, food fermentation, and compost are at the bottom of the plot.

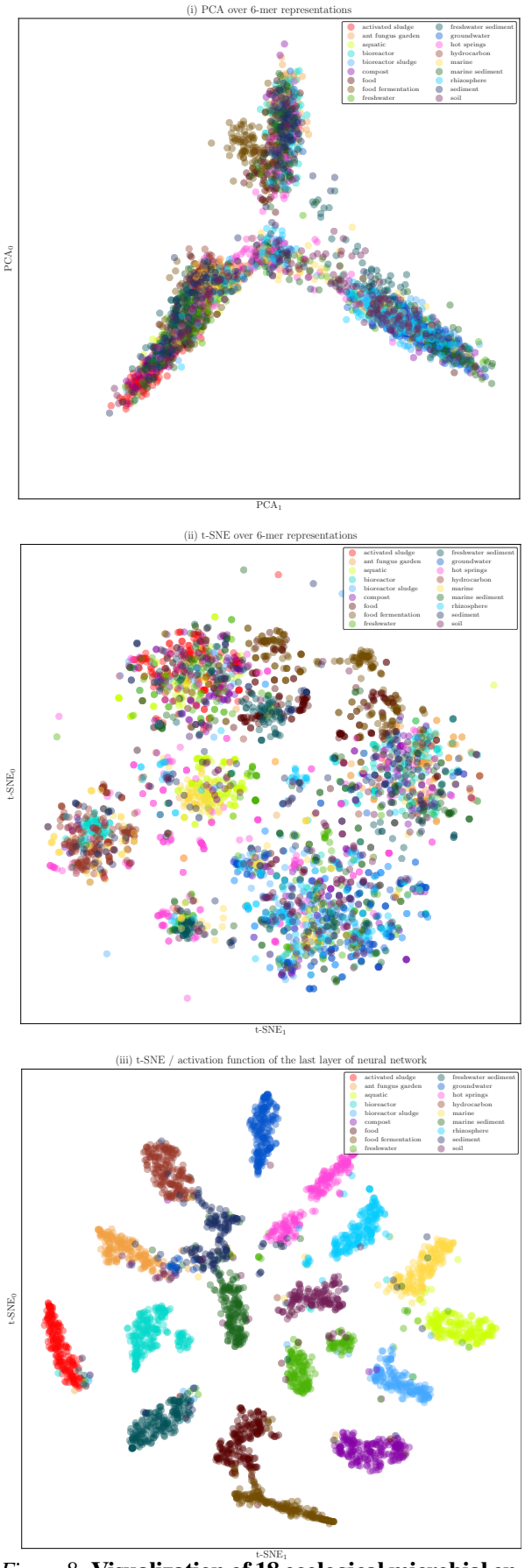

Figure 8: **Visualization of 18 ecological microbial environments using different projection methods**: (i) PCA over 6-mer distributions with unsupervised training, (ii) t-SNE over 6-mer distributions with unsupervised training, (iii) visualization of the activation function of the last layer of the trained Neural Network (projected to 2D using t-SNE).

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Step	Representation	Dataset	Classifier	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
				Precision	Recall	F1	Precision	Recall	F1
(ii)	3-mers	ECO-18K	RF	0.6 ± 0.01	0.6 ± 0.01	0.6 ± 0.01	0.63 ± 0.02	0.6 ± 0.01	0.57 ± 0.01
	4-mers			0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.7 ± 0.01	0.67 ± 0.01	0.65 ± 0.01
	5-mers			0.72 ± 0.01	0.72 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.72 ± 0.01	0.71 ± 0.01
	6-mers			$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.73 \pm 0.01}$
	7-mers			0.74 ± 0.01	0.74 ± 0.01	0.74 ± 0.01	$\mathbf{0.76 \pm 0.01}$	0.74 ± 0.01	$\mathbf{0.73 \pm 0.01}$
	8-mers			0.72 ± 0.01	0.72 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.72 ± 0.01	0.71 ± 0.01
(iv)	6-mers	ECO-18K	RF	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	0.75 ± 0.01	0.73 ± 0.01
			SVM	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$
			DNN-3L	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01
(iv)	6-mers	ECO-180K (10x larger)	RF	0.83 ± 0.0	0.83 ± 0.0	0.83 ± 0.0	0.84 ± 0.0	0.83 ± 0.0	0.83 ± 0.0
			SVM	0.86 ± 0.0	0.86 ± 0.0	0.86 ± 0.0	0.87 ± 0.01	0.86 ± 0.0	0.86 ± 0.0
			DNN-5L	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$

Table 5: **The results for the task of selecting between 18 ecological environments**. The classifiers (Random Forest, Support Vector Machine and Neural Network classifiers) are tuned and evaluated in a stratified 10xfold cross-validation setting over both ECO-18K and ECO-180K datasets. The step column refers to the steps in Figure 3.

Step	Representation	Dataset	Classifier	Micro-metrics (averaged over samples)			Macro-metrics (averaged over classes)		
				Precision	Recall	F1	Precision	Recall	F1
(ii)	3-mers	5GUTS-3100	RF	0.8 ± 0.02	0.8 ± 0.02	0.8 ± 0.02	0.8 ± 0.02	0.8 ± 0.02	0.79 ± 0.02
	4-mers			0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.01
	5-mers			0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.85 ± 0.02
	6-mers			$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	0.87 ± 0.01	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$
	7-mers			$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$
	8-mers			0.86 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	0.87 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
(iv)	6-mers	5GUTS-3100	RF	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01
			SVM	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$
			DNN-5L	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01

Table 6: **The results for the task of classifying 5 organismal environments belonging to 5 organisms' gut**. The classifiers (Random Forest, Support Vector Machine and Neural Network classifiers) are tuned and evaluated in a stratified 10xfold cross-validation setting over 5GUTS-3100 dataset. The step column refers to the steps in Figure 3.

3.4 Organismal environment identification

(ii) Classification for different values of k: The results show that k=6 and 7 provide a high classification macro-F1 of 0.87 for 5 classes (0.2 chance of randomly occurring), step (ii) in Table 6.

(iv) Comparison of classifiers for the selected k: For selected values of k the results of the organism prediction using Random Forest, SVM, and Neural Network classifiers are provided (Table 6, step (iv)). The SVM classifier obtained the top macro-F1 (0.88) for 5-way classification.

4 Conclusion

In this work, MicroPheno, a new approach for predicting environments and host phenotypes on 16S rRNA gene sequencing has been presented based on k-mer representations of shallow sub-samples. We conclude with discussing the results of this study in three parts: (1) the use of k-mers versus OTUs, (2) the benefits of shallow sub-sampling, and (3) classical methods versus the deep learning approach.

K-mers versus OTUs: To evaluate MicroPheno, we compared k-mer representations with OTU features in two tasks of body-site identification and Crohn's disease classification. We replicated the state-of-the-art approach, i.e. Random Forest over OTU features, on datasets larger than those that were previously explored. We showed that k-mer features outperform conventional OTUs, while having several advantages over OTUs, as listed below:

- The k-mer representations are easy to compute at no computational cost for any type of alignment to references or tasks of finding pair-wise sequence similarity within samples as in OTU-picking pipelines. Just to get an idea of the computational efficiency of k-mer calculation in comparison with OTUs, note that OTU picking for the Crohn's disease dataset of 1359 samples required more than 5 hours using 5 threads, while 6-mer distribution calculation took around 5 minutes using the optimal sampling size (N) for classification using the same number of threads.
- Taxonomy-independent analysis is often the preferred approach for amplicon sequencing when the samples contains unknown taxa. k-mer features can be used without making any assumptions about the taxonomy. However, OTU-picking pipelines make assumptions about the taxonomy as discussed in §1.2;

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

therefore they can even be phylogenetically incoherent.

- k-mer distribution is a well-defined representation, while OTUs are sensitive to the pipeline and the parameters.
- Sequence similarities are naturally incorporated in the k-mer representations for the downstream learning algorithm, while with grouping sequences into certain categories, the sequence similarities between OTUs are ignored.

The main disadvantage of k-mer features over OTUs is that using short k-mers makes it more difficult to trace the relevant taxa to the phenotype of interest. When such an analysis is needed, using OTUs or increasing the size of k would be an alternative solution. However, as long as prediction is concerned, using a k-mer representation seems to be the best choice for an accurate and rapid detection/diagnosis over 16S rRNA sequencing samples.

Shallow sub-sampling: We proposed a bootstrapping framework to investigate the consistency and representativeness of k-mer distributions for different sampling rates. Our results suggest that, depending on the k-mer size, even very low sub-sampling rates (0.001 to 0.1) (for k between 3 to 7) not only can provide a consistent representation, but can also result in better predictions while possibly avoiding overfitting. Setting aside the save in preprocessing time as a natural benefit of sampling, this result also suggest that at least for similar phenotypes of interest, shallow sequencing of the microbial community would be sufficient for an accurate prediction.

Classical classifiers versus deep learning: We studied the use of deep learning methods as well as classic machine learning approaches for distinguishing among human body-sites, diagnosis of Crohn's disease, and predicting the environments from representative 16S gene sequences. Studying the role of dataset size in the classification of ecological environments showed that for large datasets using deep learning provides us with more accurate predictions. However, when the number of samples are not large enough, Random Forests performs better on both OTUs and k-mer features. In addition, we observed that for classification over representative sequences as opposed to samples (pool of sequences), the SVM outperformed the Random Forest classifier. Another advantage of using deep learning in classification was that supervised training of a proper representation of data results in a more discriminative representation for the downstream visualization compared to the unsupervised methods (PCA and t-SNE on the raw k-mer distributions). For body-site identification and even more clearly in ecological environment classification, the model was able to extract more high-level similarities between the environments.

References

[1] S. V. Lynch and O. Pedersen. "The Human Intestinal Microbiome in Health and Disease". In: *New England Journal of Medicine* 375.24 (2016), pp. 2369–2379. ISSN: 0028-4793.

[2] B. J. Marsland, K. Yadava, and L. P. Nicod. "The airway microbiome and disease". In: *Chest* 144.2 (2013), pp. 632–637. ISSN: 19313543.

[3] M.-C. Arrieta et al. "Early infancy microbial and metabolic alterations affect risk of childhood asthma". In: *Science Translational Medicine* 7.307 (2015). ISSN: 1946-6242.

[4] D. M. Saulnier et al. "Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome". In: *Gastroenterology* 141.5 (2011), pp. 1782–1791. ISSN: 00165085.

[5] I. Cho and M. J. Blaser. *The human microbiome: At the interface of health and disease*. 2012.

[6] G. Cammarota, G. Ianiro, and A. Gasbarrini. *Fecal microbiota transplantation for the treatment of clostridium difficile infection: A systematic review*. 2014.

[7] Z. L. Deng et al. "Dysbiosis in chronic periodontitis: Key microbial players and interactions with the human host". In: *Scientific Reports* 7.1 (2017), pp. 1–13. ISSN: 20452322.

[8] P Jorth et al. "Metatranscriptomics of the Human Oral Microbiome during Health and Disease". In: *mBio* 5.2 (2014), e01012–14–e01012–14. ISSN: 2150-7511.

[9] C. Gimblet et al. "Cutaneous Leishmaniasis Induces a Transmissible Dysbiotic Skin Microbiota that Promotes Skin Inflammation". In: *Cell Host and Microbe* 22.1 (2017), 13–24.e4. ISSN: 19346069.

[10] P. J. Turnbaugh, F. Bäckhed, L. Fulton, and J. I. Gordon. "Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome". In: *Cell Host and Microbe* 3.4 (2008). ISSN: 19313128.

[11] V. K. Ridaura et al. "Gut microbiota from twins discordant for obesity modulate metabolism in mice". In: *Science* 341.6150 (2013). ISSN: 10959203.

[12] A. Ramezani and D. S. Raj. "The Gut Microbiome, Kidney Disease, and Targeted Interventions". In: *Journal of the American Society of Nephrology* 25.4 (2014). ISSN: 1046-6673.

[13] S. Michail et al. "Alterations in the gut microbiome of children with severe ulcerative colitis". In: *Inflammatory Bowel Diseases* 18.10 (2012),

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pp. 1799–1808. ISSN: 10780998.

[14] D. Gevers et al. "The treatment-naive microbiome in new-onset Crohn's disease". In: *Cell Host and Microbe* 15.3 (2014), pp. 382–392. ISSN: 19346069.

[15] V. Pascal et al. "A microbial signature for Crohn's disease". In: *Gut* 66.5 (2017), pp. 813–822. ISSN: 14683288.

[16] M. D. Kappelman et al. "The Prevalence and Geographic Distribution of Crohn's Disease and Ulcerative Colitis in the United States". In: *Clinical Gastroenterology and Hepatology* 5.12 (2007), pp. 1424–1429. ISSN: 15423565.

[17] J. A. Gilbert and J. D. Neufeld. "Life in a World without Microbes". In: *PLoS Biology* 12.12 (2014). ISSN: 15457885.

[18] A. J. Pinto, C. Xi, and L. Raskin. "Bacterial community structure in the drinking water microbiome is governed by filtration processes". In: *Environmental Science and Technology* 46.16 (2012), pp. 8851–8859. ISSN: 0013936X.

[19] M. A. Moran. *The global ocean microbiome*. 2015.

[20] S. Sunagawa et al. "Structure and function of the global ocean microbiome". In: *Science* 348.6237 (2015). ISSN: 10959203.

[21] N. Fierer. *Embracing the unknown: Disentangling the complexities of the soil microbiome*. 2017.

[22] J. M. Chaparro, A. M. Sheflin, D. K. Manter, and J. M. Vivanco. *Manipulating the soil microbiome to increase soil health and plant fertility*. 2012.

[23] M. Hamady and R. Knight. *Microbial community profiling for human microbiome projects: Tools, techniques, and challenges*. 2009.

[24] J. M. Janda and S. L. Abbott. *16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls*. 2007.

[25] S. Chakravorty et al. "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria". In: *Journal of Microbiological Methods* 69.2 (2007), pp. 330–339. ISSN: 01677012.

[26] J. G. Caporaso et al. *QIIME allows analysis of high-throughput community sequencing data*. 2010.

[27] B. Lawley and G. W. Tannock. *Analysis of 16S rRNA Gene Amplicon Sequences Using the QIIME Software Package*. Vol. 1537. Springer, 2017, pp. 153–163. ISBN: 978-1-60761-819-5.

[28] P. D. Schloss et al. "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities". In: *Applied and Environmental Microbiology* 75.23 (2009), pp. 7537–7541. ISSN: 00992240.

[29] R. C. Edgar et al. "UCHIME improves sensitivity and speed of chimera detection". In: *Bioinformatics* 27.16 (2011), pp. 2194–2200. ISSN: 13674803.

[30] D. Knights, E. K. Costello, and R. Knight. *Supervised classification of human microbiota*. 2011.

[31] A. Statnikov et al. "A comprehensive evaluation of multicategory classification methods for microbiomic data". In: *Microbiome* 1.1 (2013), p. 11. ISSN: 2049-2618.

[32] X. Xu et al. "MetaDP: a comprehensive web server for disease prediction of 16S rRNA metagenomic datasets". In: *Biophysics Reports* 2.5-6 (2016), pp. 106–115.

[33] A. Eck et al. "Robust microbiota-based diagnostics for inflammatory bowel disease". In: *Journal of Clinical Microbiology* 55.6 (2017), pp. 1720–1732. ISSN: 1098660X.

[34] C. Duvallet et al. "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses". In: *Nature Communications* 8.1 (2017), p. 1784. ISSN: 2041-1723.

[35] A. P. Carrieri et al. "Host Phenotype Prediction from Differentially Abundant Microbes Using RoDEO". In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Cham: Springer International Publishing, 2017, pp. 27–41. ISBN: 978-3-319-67834-4.

[36] T. Cordier et al. "Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning". In: *Environmental Science & Technology* 51.16 (2017). ISSN: 0013-936X.

[37] N. Fierer et al. "Forensic identification using skin bacterial communities." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.14 (2010), pp. 6477–81. ISSN: 1091-6490.

[38] S. E. Schmedes et al. "Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification". In: *Forensic Science International: Genetics* 32 (2018), pp. 50–61.

[39] E. Pasolli et al. "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights". In: *PLoS Computational Biology* 12.7 (2016). ISSN: 15537358.

[40] E. K. Costello et al. "Bacterial community variation in human body habitats across space and time." In: *Science (New York, N.Y.)* 326.5960 (2009), pp. 1694–7. ISSN: 1095-9203.

[41] B. E. Dutilh et al. "Explaining microbial phenotypes on a genomic scale: Gwas for microbes". In: *Briefings in Functional Genomics* 12.4 (2013), pp. 366–0380. ISSN: 20412649.

[42] E. M. Ross et al. "Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle". In: *PLoS ONE* 8.9 (2013). ISSN: 19326203.

[43] H. Cui and X. Zhang. "Alignment-free supervised classification of metagenomes by recursive SVM". In: *BMC Genomics* 14.1 (2013). ISSN: 14712164.

bioRxiv preprint doi: https://doi.org/10.1101/255018; this version posted January 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[44] E. Asgari and M. R. Mofrad. "Continuous distributed representation of biological sequences for deep proteomics and genomics". In: *PLoS ONE* 10.11 (2015). ISSN: 19326203.

[45] S. Min, B. Lee, and S. Yoon. "Deep learning in bioinformatics". In: *Briefings in Bioinformatics* (2016), bbw068. ISSN: 1467-5463.

[46] G. Ditzler, R. Polikar, and G. Rosen. "Multi-Layer and Recursive Neural Networks for Metagenomic Classification". In: *IEEE Transactions on Nanobioscience* 14.6 (2015). ISSN: 15361241.

[47] N.-P. Nguyen, T. Warnow, M. Pop, and B. White. "A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity". In: *npj Biofilms and Microbiomes* 2.1 (2016). ISSN: 2055-5008.

[48] D. McDonald et al. "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea". In: *ISME Journal* 6.3 (2012), pp. 610–618. ISSN: 17517362.

[49] C. Quast et al. "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools". In: *Nucleic Acids Research* 41.D1 (2013). ISSN: 03051048.

[50] J. R. Rideout et al. "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences". In: *PeerJ* 2 (2014), e545. ISSN: 2167-8359.

[51] Y. Cai et al. "ESPRIT-Forest: Parallel clustering of massive amplicon sequence data in subquadratic time". In: *PLoS Computational Biology* 13.4 (2017). ISSN: 15537358.

[52] A. F. Koeppel and M. Wu. "Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units". In: *Nucleic Acids Research* 41.10 (2013), pp. 5175–5188. ISSN: 03051048.

[53] Y. He et al. "Erratum to: Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity". In: *Microbiome* 3.1 (2015). ISSN: 2049-2618.

[54] G. Marçais and C. Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6 (2011), pp. 764–770. ISSN: 13674803.

[55] A. C. McHardy et al. "Accurate phylogenetic classification of variable-length DNA fragments". In: *Nature Methods* 4.1 (2007), pp. 63–72. ISSN: 15487091.

[56] K. R. Patil et al. *Taxonomic metagenome sequence assignment with structured output models*. 2011.

[57] D. E. Wood and S. L. Salzberg. "Kraken: Ultrafast metagenomic sequence classification using exact alignments". In: *Genome Biology* 15.3 (2014). ISSN: 1474760X.

[58] J. Kawulok and S. Deorowicz. "CoMeta: Classification of metagenomes using k-mers". In: *PLoS ONE* 10.4 (2015). ISSN: 19326203.

[59] P. Menzel and A. Krogh. "Kaiju : Fast and sensitive taxonomic classification for metagenomics". In: *bioRxiv* 7 (2015), pp. 1–9. ISSN: 2041-1723.

[60] K. Vervier et al. "Large-scale machine learning for metagenomics sequence classification". In: *Bioinformatics* 32.7 (2016), pp. 1023–1032. ISSN: 14602059.

[61] J. Peterson et al. "The NIH Human Microbiome Project". In: *Genome Research* 19.12 (2009), pp. 2317–2323. ISSN: 10889051.

[62] C. Huttenhower and Human Microbiome Project Consortium. "Structure, function and diversity of the healthy human microbiome." In: *Nature* 486.7402 (2012). ISSN: 1476-4687.

[63] C. C. Yang and W. Iwasaki. "MetaMetaDB: A database and analytic system for investigating microbial habitability". In: *PLoS ONE* 9.1 (2014). ISSN: 19326203.

[64] R. S. Olson et al. "Data-driven advice for applying machine learning to bioinformatics problems". In: *Biocomputing 2018*. WORLD SCIENTIFIC, 2017.

[65] J. A. K. Suykens and J Vandewalle. "Least Squares Support Vector Machine Classifiers". In: *Neural Processing Letters* 9.3 (1999), pp. 293–300. ISSN: 1573-773X.

[66] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125.

[67] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[68] N. Srivastava et al. "Dropout: prevent NN from overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. ISSN: 15337928.

[69] D. P. Kingma and J. L. Ba. "Adam: a Method for Stochastic Optimization". In: *International Conference on Learning Representations 2015* (2015), pp. 1–15. ISSN: 09252312.

[70] I. T. Jolliffe and I. T. Jolliffe. *Principal Component Analysis*. 1986.

[71] L. J. P. Van Der Maaten and G. E. Hinton. "Visualizing high-dimensional data using t-sne". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. ISSN: 1532-4435.

[72] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013). ISSN: 01628828.

[73] F. Pedregosa and G Varoquaux. *Scikit-learn: Machine learning in Python*. Vol. 12. ACM, 2011, pp. 2825–2830. ISBN: 9781783281930.