Supplementary Material

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

Contents

1 Comparison between 16S rRNA gene sequencing and shotgun metagenomic sequencing

Two primary methods for taxonomic profiling of microbial communities exist: 16S gene sequencing and whole genome shotgun sequencing (WGS). The advantages and disadvantages of 16S gene sequencing versus WGS methods for the purpose of taxonomic profiling are listed in the following.

1.1 Advantages of 16S rRNA sequencing

Cost: currently, 16S gene sequencing is significantly more cost-effective than WGS and requires less starting material than WGS [5]. The low cost of 16S gene sequencing is one of the primary reasons for its widespread use.

1.2 Disadvantages of 16S rRNA sequencing

- **No gold standard hyper-variable region:** The full 16S gene sequence length cannot be sequenced by all sequencing technologies (short-read second-generation platforms). Thus, the hyper-variable regions need to be sequenced instead. However, there is no consensus about choice of hyper-variable regions influencing the phylogenetic resolution [14].
- **Precision of taxonomic assignment:** 16S rRNA sequencing has a lower precision compared to WGS in taxonomic assignment [15]. While WGS may be used to assign species-level classifications to many sequences with confidence, 16S rRNA sequencing classification is typically restricted to the genus-level, with species-level predictions based on OTUs being of generally lower precision [16]. In addition, fungal and viral genomes, which do not contain 16S rRNA genes and thus are not picked up with 16S rRNA sequencing, are sequenced during WGS.
- **Bias:** 16S rRNA sequencing suffers from significant PCR amplification biases. This is a result of differences in PCR primer binding to certain 16S rRNA gene sequences, which may cause entire groups of bacteria to be underrepresented in the sample [5]. In addition to PCR amplification biases, over estimation of α diversity and profiling pipeline can also introduce biases in the 16S rRNA data [14]. WGS is a comparatively unbiased method for taxonomic profiling.

2 Evaluation metrics

2.1 Precision, Recall, F1

Precision, recall (sensitivity), and F1 are among the most popular metrics for evaluation of classification performance in machine learning. Precision and recall are defined as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}, Recall = \frac{TruePositive}{TruePositive + FalseNegative},$$

F1 score ($0 \leq F1 \leq 1$) is the harmonic average of the precision and recall maintaining a balance between both precision and recall:

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$$

2.2 Micro versus Macro averaged metrics

In the evaluation, we distinguish between micro and macro metrics. Micro-averaging is average of metrics over instances (samples), while macro averaging computes the metrics for each class separately and then simply average over classes. Macro-metrics are useful when we want to give equal importance to all categories, although the categories are imbalanced.

In this paper in order to provide detailed information on how our classifiers perform in both micro and macro cases, we have reported both micro- and macro- averaging precision and recall. In particular we optimized our classifiers for macro-F1 to make a trade of between precision and recall as well as giving equal importance to all categories.

3 Visualization Methods

Principal component analysis (PCA): is one of the most widely used techniques for dimensionality reduction. PCA provides us with the principal directions (Vs) in which the data exhibits maximum variation. These principal directions are the eigenvectors of the data covariance matrix and the eigenvalue corresponding to each eigenvector indicates the variance of the data in that particular direction. Thus, by a linear transformation we can project the data to a low-dimensional space with principal directions as the axes. Despite its widespread use in modern data analysis, PCA has some limitations. Since PCA involves covariance matrix estimation and eigenvalue decomposition, the computational complexity of PCA for M samples, each represented with m features, is of the order $O(m^2M + m^3)$. In addition, in order to have a reliable estimation of the covariance matrix, PCA requires a large number of samples and typically underperforms with limited numbers of data points. Finally, PCA is a linear transformation with several constraints, and thus cannot address the non-linearity in the data [18, 1].

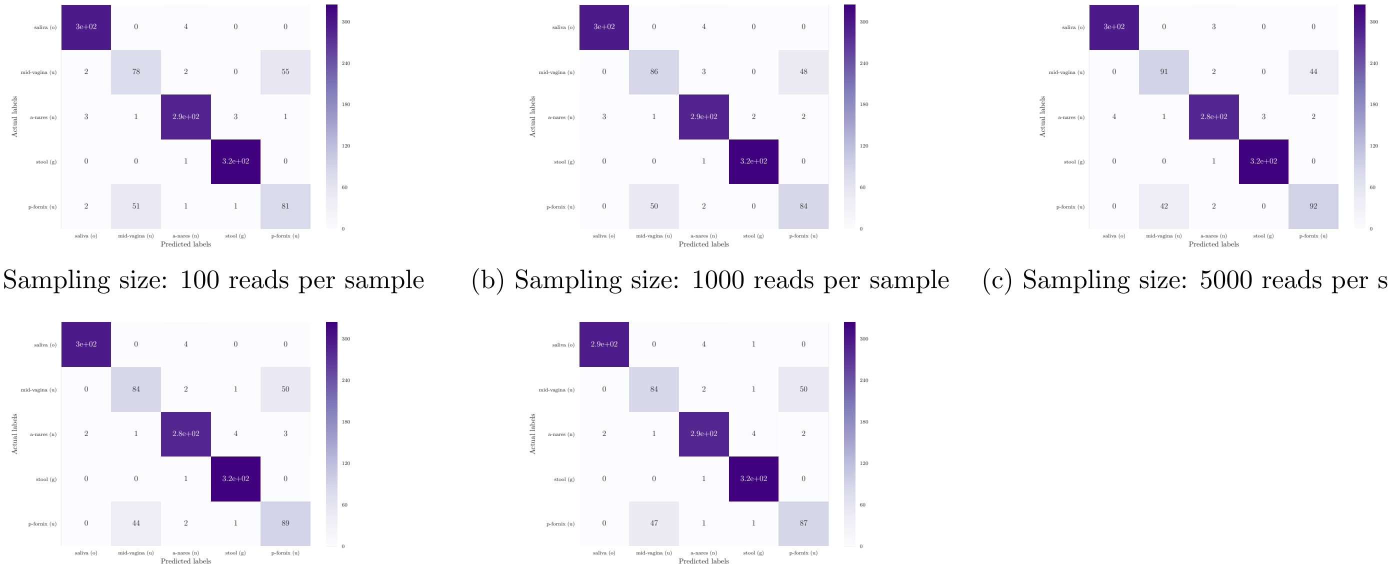
t-Distributed Stochastic Neighbor Embedding (t-SNE): is a dimensionality reduction technique which has been successfully applied to visualization problems, since it attempts to preserve pairwise distance distribution of points in lower dimensions [19]. Since projection to lower dimensions involves the distribution of relative distances, it requires large amounts of data points in order to find a meaningful representation. t-SNE is a nonlinear transformation with a computational complexity of $O(M^2)$. However, a recent implementation of t-SNE has been reported to have a computational complexity of $O(M\log(M))$ [20]. t-SNE is a relatively new technique in machine learning, and only recently has been used for the analysis of genomic data [2, 13, 3, 12].

Supervised representation learning: both PCA and t-SNE are unsupervised methods, meaning that the output representation is trained independently from the instances' categories. Deep neural models provide a powerful framework for training multiple levels of representation for the input data in both unsupervised and supervised manners. In supervised training, like multi-class classification task, specialized representations of the input data are trained using the back-propagation algorithm [17]. In such framework, the activation function of the last hidden-layer for the input data is a specialized representation of data for that particular task or even other tasks [4]. Such a representation trained in a supervised manner can also be visualized using t-SNE or other visualization techniques.

4 Body-site classification results

4.1 Confusion matrix for body-site classification using different sampling sizes

As shown in the Figure 1, increase in the sampling size from 100 reads per sample helps in discrimination between similar body-sites (mid-vagina and urogenital). However, further increases (more then 5000 reads per sample) results in over-fitting.



(a) Sampling size: 100 reads per sample

(b) Sampling size: 1000 reads per sample

(c) Sampling size: 5000 reads per sample

(d) Sampling size: 10000 reads per sample

(e) Using all reads

Figure 1: **The confusion matrix for the classification of 5 major body-sites for different sampling sizes**, 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).

5 Crohn's disease results

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 [9]. In contrast to Ulcerative Colitis, which exhibits shallow inflammation of the tissue, Crohns disease is characterized by ulcerous transmural inflammation that extends deep into the tissue of the gut [10]. Previous genome-wide association studies have demonstrated a link between intestinal microbe response pathways and IBD development [7]. This evidence, along with observations of significantly altered microbiome compositions in IBD patients in humans, points towards a role for the microbiome in the regulation of IBD. While the mechanism for this regulation is not well understood, it is thought that dysbiosis of the intestinal microbiome may impact the immune system within the intestine, leading to the progression of Crohns disease.

5.1 Bootstrapping result for Crohn's disease dataset

Similar to bootstrapping for the body-site dataset, $\bar{D_S}$and $\bar{D_R}$ for different values of k with respect to sampling sizes are shown in Figure 2. The structure of the curve is similar to the body-site dataset. For each k, the interval that $\bar{D_S}$ and $\bar{D_R}$ converge to their minimum values show a propoer range for picking a sampling size resulting in self-consistent and representative representations.

6 Ecological environments classification results

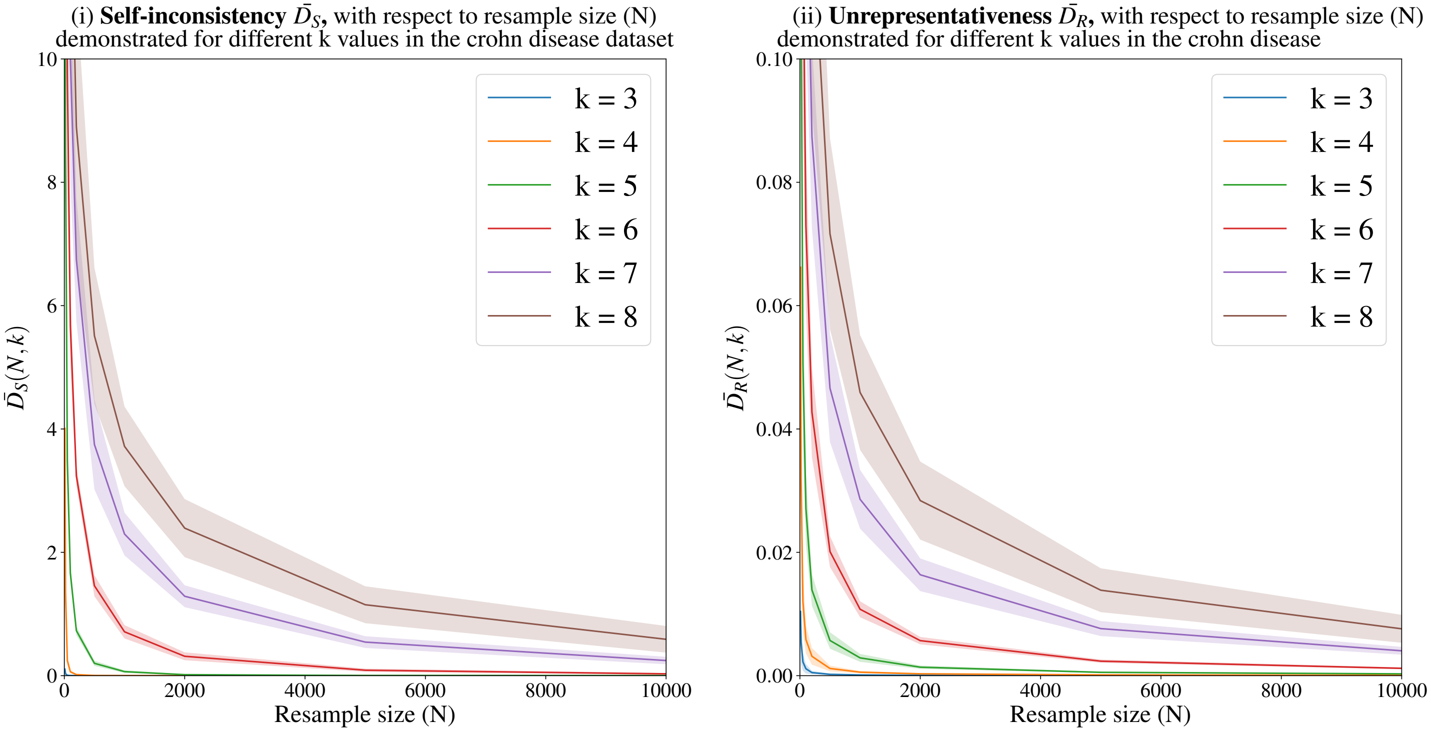
The confusion matrix for 18-way classification of ecological environments is shown in Figure 3. The confusion matrix shows that most of the major miss-classifications are due to similarities between the environment pairs, examples are (marine sediment, sediment), (food, food fermentation), and (marine, aquatic).

7 Organismal environment classification results

The confusion matrix for the task of classifying the 5 organisms' gut environments is presented in Figure 4. In addition, visualizations of representative 16S rRNA gene sequences in 5 organisms' guts obtained through using PCA, t-SNE, and t-SNE on the activation function of the last layer of the trained Neural Network is depicted in Figure 5.

8 Representation creation run-time comparison

A comparison between run-time for k-mer representation creation versus OTU picking for NIH Human Microbiome Project samples [11, 8] and Crohn's disease samples [6] are summarized in Table 1.



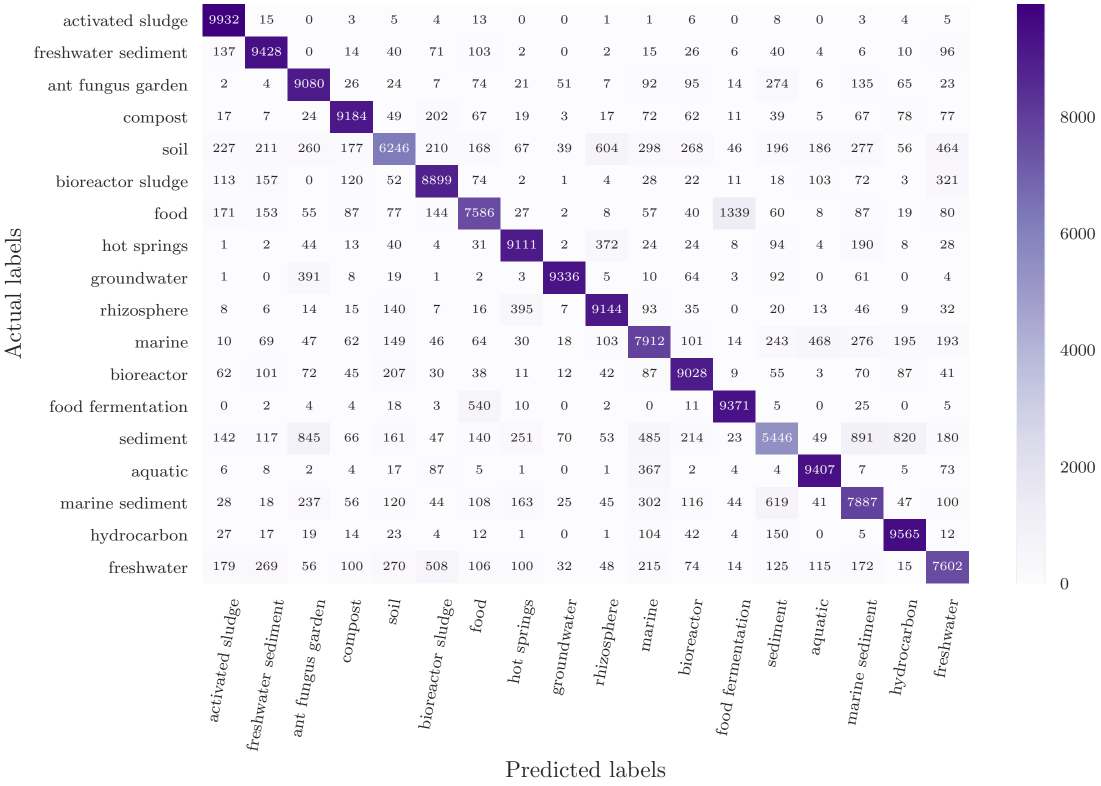
(i) **Self-inconsistency $\bar{D_S}$,** with respect to resample size (N) demonstrated for different k values in the crohn disease dataset

(ii) **Unrepresentativeness $\bar{D_R}$,** with respect to resample size (N) demonstrated for different k values in the crohn disease

Figure 2: Measuring self-inconsistency ($\bar{D_S}$) and unrepresentativeness ($\bar{D_R}$) for the Crohn's disease dataset. Each point presents the 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.

Database	# of samples	Run-time for k-mers	Run-time for OTU
Crohn's disease	1359	381 sec	5125 sec
Body-site	12075	3339 sec	66996 sec

Table 1: Run-time comparsion between OTU versus k-mer representation creation.



	activated sludge	freshwater sediment	ant fungus garden	compost	soil	bioreactor sludge	food	hot springs	groundwater	rhizosphere	marine	bioreactor	food fermentation	sediment	aquatic	marine sediment	hydrocarbon	freshwater
activated sludge	9932	15	0	3	5	4	13	0	0	1	1	6	0	8	0	3	4	5
freshwater sediment	137	9428	0	14	40	71	103	2	0	2	15	26	6	40	4	6	10	96
ant fungus garden	2	4	9080	26	24	7	74	21	51	7	92	95	14	274	6	135	65	23
compost	17	7	24	9184	49	202	67	19	3	17	72	62	11	39	5	67	78	77
soil	227	211	260	177	6246	210	168	67	39	604	298	268	46	196	186	277	56	464
bioreactor sludge	113	157	0	120	52	8899	74	2	1	4	28	22	11	18	103	72	3	321
food	171	153	55	87	77	144	7586	27	2	8	57	40	1339	60	8	87	19	80
hot springs	1	2	44	13	40	4	31	9111	2	372	24	24	8	94	4	190	8	28
groundwater	1	0	391	8	19	1	2	3	9336	5	10	64	3	92	0	61	0	4
rhizosphere	8	6	14	15	140	7	16	395	7	9144	93	35	0	20	13	46	9	32
marine	10	69	47	62	149	46	64	30	18	103	7912	101	14	243	468	276	195	193
bioreactor	62	101	72	45	207	30	38	11	12	42	87	9028	9	55	3	70	87	41
food fermentation	0	2	4	4	18	3	540	10	0	2	0	11	9371	5	0	25	0	5
sediment	142	117	845	66	161	47	140	251	70	53	485	214	23	5446	49	891	820	180
aquatic	6	8	2	4	17	87	5	1	0	1	367	2	4	4	9407	7	5	73
marine sediment	28	18	237	56	120	44	108	163	25	45	302	116	44	619	41	7887	47	100
hydrocarbon	27	17	19	14	23	4	12	1	0	1	104	42	4	150	0	5	9565	12
freshwater	179	269	56	100	270	508	106	100	32	48	215	74	14	125	115	172	15	7602

Figure 3: Confusion matrix for classification of 18 ecological environments using SVM classifier.

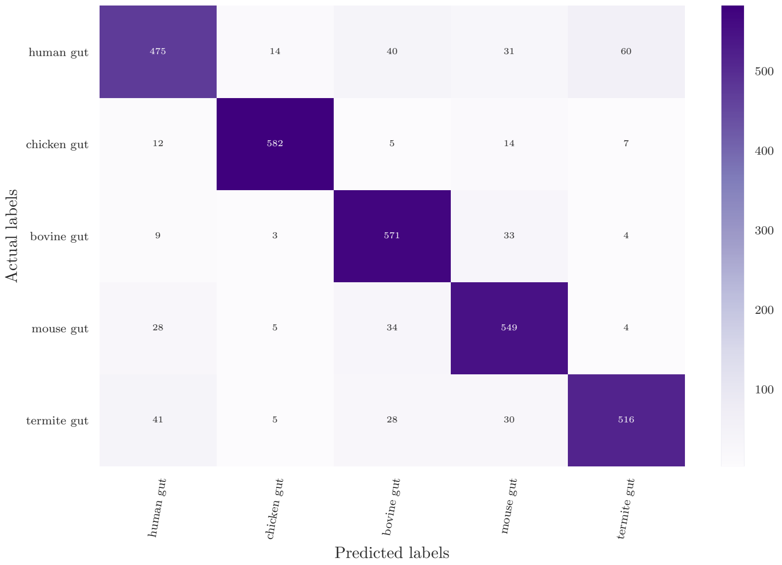


Figure 4: Confusion matrix for classification of 5 organismal environments using SVM classifier.

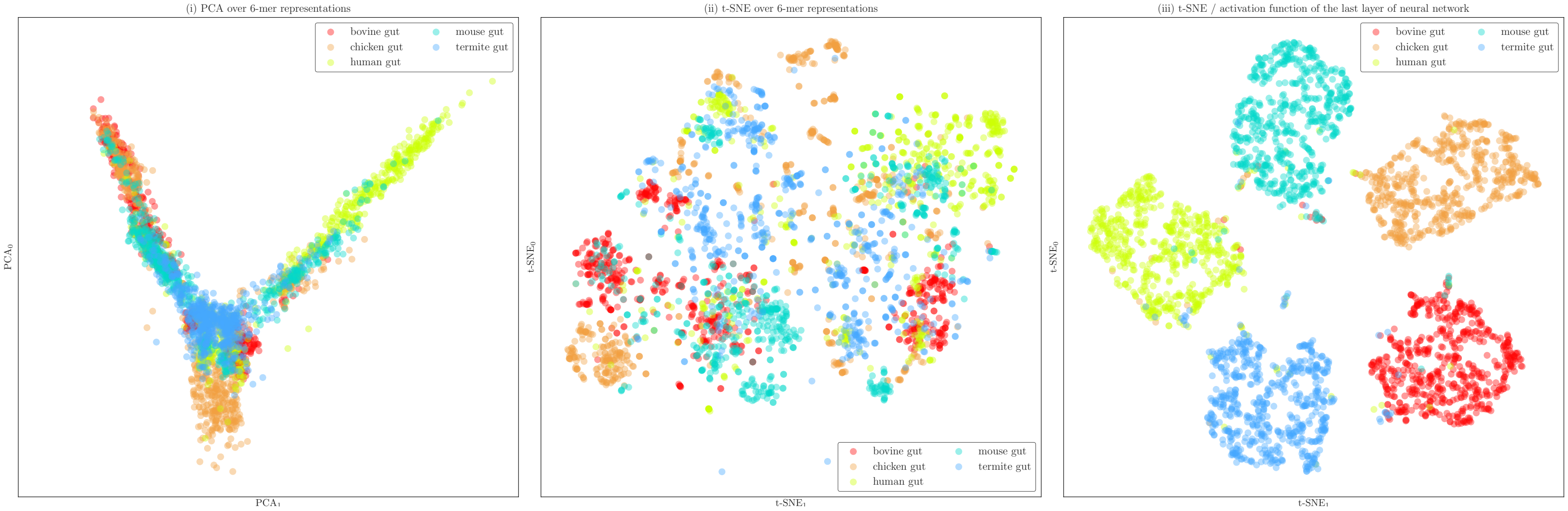


Figure 5: Visualization of 5 organismal environments using different projection methods: (i) PCA over 6-mer distributions with unsupervised training, (ii) t-SNE over 6-mer distributions with unsupervised training, (iii) visualization of the activation function of the last layer of the trained Neural Network (projected to 2D using t-SNE).

References

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis, 2010. ISSN 19395108.

[2] El Ad David Amir, Kara L. Davis, Michelle D. Tadmor, Erin F. Simonds, Jacob H. Levine, Sean C. Bendall, Daniel K. Shenfeld, Smita Krishnaswamy, Garry P. Nolan, and Dana Pe'Er. ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, 2013. ISSN 10870156. doi: 10.1038/nbt.2594.

[3] Ehsaneddin Asgari and Mohammad R.K. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE*, 10(11), 2015. ISSN 19326203. doi: 10.1371/journal.pone.0141287.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013. ISSN 01628828. doi: 10.1109/TPAMI.2013.50.

[5] Fabien Cottier, Kandhadayar Gopalan Srinivasan, Marina Yurieva, Webber Liao, Michael Poidinger, Francesca Zolezzi, and Norman Pavelka. Advantages of meta-total rna sequencing (metrs) over shotgun metagenomics and amplicon-based sequencing in the profiling of complex microbial communities. *npj Biofilms and Microbiomes*, 4(1):2, 2018.

[6] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host and Microbe*, 15(3):382–392, 2014. ISSN 19346069. doi: 10.1016/j.chom.2014.02.005.

[7] Dirk Gevers, Subra Kugathasan, Dan Knights, Aleksandar D. Kostic, Rob Knight, and Ramnik J. Xavier. A Microbiome Foundation for the Study of Crohn's Disease, 2017. ISSN 19346069.

[8] Curtis Huttenhower and Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 2012. ISSN 1476-4687. doi: 10.1038/nature11234. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564958&tool=pmcentrez&rendertype=abstract`.

[9] Michael D. Kappelman, Sheryl L. Rifas-Shiman, Ken Kleinman, Dan Ollendorf, Athos Bousvaros, Richard J. Grand, and Jonathan A. Finkelstein. The Prevalence and Geographic Distribution of Crohn's Disease and Ulcerative Colitis in the United States. *Clinical Gastroenterology and Hepatology*, 5(12):1424–1429, 2007. ISSN 15423565. doi: 10.1016/j.cgh.2007.07.012.

[10] Martin W. Laass, Dirk Roggenbuck, and Karsten Conrad. Diagnosis and classification of Crohn's disease, 2014. ISSN 18730183.

[11] Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, Vivien Bonazzi, Jean E. McEwen, Kris A. Wetterstrand, Carolyn Deal, Carl C. Baker, Valentina Di Francesco, T. Kevin Howcroft, Robert W. Karp, R. Dwayne Lunsford, Christopher R. Wellington, Tsegahiwot Belachew, Michael Wright, Christina Giblin, Hagit David, Melody Mills, Rachelle Salomon, Christopher Mullins, Beena Akolkar, Lisa Begg, Cindy Davis, Lindsey Grandison, Michael Humble, Jag Khalsa, A. Roger Little, Hannah Peavy, Carol Pontzer, Matthew Portnoy, Michael H. Sayre, Pamela Starke-Reed, Samir Zakhari, Jennifer Read, Bracie Watson, and Mark Guyer. The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323, 2009. ISSN 10889051. doi: 10.1101/gr.096651.109.

[12] Vanessa M. Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C. Wilson, Renee Moore, Terrill K. Mcclanahan, Svetlana Sadekova, and Joel A. Klappenbach. Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939, 2017. ISSN 15461696. doi: 10.1038/nbt.3973.

[13] Alexander Platzer. Visualization of SNPs with t-SNE. *PLoS ONE*, 8(2), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0056883.

[14] Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The madness of microbiome: Attempting to find consensus best practice for 16S microbiome studies. *Applied and Environmental Microbiology*, pages AEM.02627–17, 2018. ISSN 0099-2240. doi: 10.1128/AEM.02627-17. URL `http://aem.asm.org/lookup/doi/10.1128/AEM.02627-17`.

[15] Rachel Poretsky, Luis M. Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T. Konstantinidis. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE*, 9(4), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0093827.

[16] Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.

[17] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.

[18] Jonathon Shlens. A tutorial on principal component analysis. *Internet Article*, pages 1–13, 2005. ISSN 00219991. doi: 10.1.1.115.3503. URL [PDF].

[19] L J P Van Der Maaten and G E Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. ISSN 1532-4435. doi: 10.1007/s10479-011-0841-3. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=7911431479148734548related:VOiAgwMNy20J.

[20] Laurens van der Maaten. Barnes-Hut-SNE. *CoRR*, 1301.3342:1–11, 2013. URL http://arxiv.org/abs/1301.3342.