A Fast Machine Learning Workflow for Rapid Phenotype Prediction from Whole Shotgun Metagenomes

Anna Paola Carrieri[*]
IBM Research UK
Sci-Tech Daresbury
Warrington, UK.

Will PM Rowe[*]
Scientific Computing Dept.
STFC Daresbury Lab.
Warrington, UK.

Martyn Winn
Scientific Computing Dept.
STFC Daresbury Lab.
Warrington, UK.

Edward O. Pyzer-Knapp[†]
IBM Research UK
Sci-Tech Daresbury
Warrington, UK.
epyzerk3@uk.ibm.com

Abstract

Research on the microbiome is an emerging and crucial science that finds many applications in healthcare, food safety, precision agriculture and environmental studies. Huge amounts of DNA from microbial communities are being sequenced and analyzed by scientists interested in extracting meaningful biological information from this big data. Analyzing massive microbiome sequencing datasets, which embed the functions and interactions of thousands of different bacterial, fungal and viral species, is a significant computational challenge. Artificial intelligence has the potential for building predictive models that can provide insights for specific cutting edge applications such as guiding diagnostics and developing personalised treatments, as well as maintaining soil health and fertility. Current machine learning workflows that predict traits of host organisms from their commensal microbiome do not take into account the whole genetic material constituting the microbiome, instead basing the analysis on specific marker genes. In this paper, to the best of our knowledge, we introduce the first machine learning workflow that efficiently performs host phenotype prediction from whole shotgun metagenomes by computing similarity-preserving compact representations of the genetic material. Our workflow enables prediction tasks, such as classification and regression, from Terabytes of raw sequencing data that do not necessitate any pre-prossessing through expensive bioinformatics pipelines. We compare the performance in terms of time, accuracy and uncertainty of predictions for four different classifiers. More precisely, we demonstrate that our ML workflow can efficiently classify real data with high accuracy, using examples from dog and human metagenomic studies, representing a step forward towards real time diagnostics and a potential for cloud applications.

Glossary

Phenotype: observable characteristics of an organism resulting from the interaction of its gene products with the environment.
Microbiome: collective genomes of a microbial community inhabiting a particular environment such as a surface of the human body.
Metagenomics: study of the genetic material of microorganisms constituting the microbiome.
Read: inferred sequence of nucleotides corresponding to all or part of a single DNA fragment, as measured in a sequencing experiment.
Whole metagenome shotgun sequencing: sequencing of the total genetic material of a microbial community.
Taxonomy: the process of naming and classifying organisms into groups within a larger system, according to their similarities.
Marker genes: gene families that can be used to quantify taxonomic diversity.
OTUs - Operational Taxonomy Units: cluster of microorganisms grouped by sequence similarity of a specific marker gene (such as 16S rRNA) and representing a taxon.

Introduction

A research objective of great interest in the last decade is to be able to understand the structure, organization, and functionality of the microbiome and how this affects, and is affected by, the environment that surrounds it. The microbiome is made up of the whole genetic material and interactions of a community of micro-organisms (bacteria, archaea, viruses or fungi) that live in a natural environment. The environment could be an entire organism (e.g. a human being or a mouse), part of it (e.g. the intestine, the skin or the mouth) or a natural habitat (e.g. water or soil).

Trillions of microbes have evolved and continue to live in the human body and can play a positive or negative role in determining the well-being of individuals. An imbalance in the microbial community can lead to the development and progression of various diseases, such as infections, respiratory diseases, metabolic and even psychological illnesses (e.g. depression and anxiety).

Understanding the relationship between microbial communities and the health or disease state of individuals can help with designing effective targeted treatments focused on re-balancing the microbiome. On the other hand, microorganisms residing and interacting in the soil perform important processes such as support of the plant growth and cycling of carbon and other nutrients (Jansson and Hofmockel 2018). Many beneficial functions carried out by the soil microbiome are currently threatened due to changing climate, soil degradation and poor land management practices. Un-

[*] These authors contributed equally to the work
[†] Corresponding author

derstanding which species of micro-organisms are essential for maintaining soil health and fertility is important to comprehend how to manipulate the soil microbiome and restore ecosystem function.

The complex structure in which most of the micro-organisms that make up the microbiome are organized represents an obstacle to traditional in vitro culture. However, recent advances in high-throughput next generation DNA sequencing (NGS) technologies have enabled researchers to characterize and compare microbial communities in diverse natural environments, such as the human gut microbiome, via the analysis of their nucleic acid content. Metagenomics, the study of genetic material of micro-organisms constituting the microbiome, is proving very promising in research studies on the environment, biomedicine and food safety. This is exemplified by several large-scale sequencing initiatives such as the Human Microbiome Project (HMP Consortium 2012), Global Ocean Survey (Rusch and al 2007) and the Earth Microbiome Project (Consortium 2017).

Given the decreasing cost of NGS technologies and the resultant increase in the amount of metagenomic sequencing data being generated, a compelling need for efficient analytic tools able to manage and address the big data challenges of microbiome research arises. Machine learning currently offers some of the most promising tools for building predictive models for classification or regression tasks from biological data, such as metagenomic data. (Libbrecht and Noble 2015) presents an overview of machine learning applications for the analysis of genome sequencing data sets. Data produced in metagenomic studies are both unbalanced and heterogeneous, thus reflecting the current challenges of machine learning in the era of Big Data. (Soueidan and Macha 2018) reviews the contribution of machine learning methods for metagenomics research, focusing on answering several important questions such as microbial species clustering, taxonomic assignment, comparative metagenomics and gene prediction.

In this paper, we focus on phenotype prediction tasks applying data driven machine learning to metagenomic data. The phenotype is an observable characteristic or trait of the microbial community or of the host organism.

The ability to predict the phenotype of a host organism from the measured metagenomes of the microbial community is a powerful analytic tool with many applications. For example, making predictions of whether an individual is healthy, has a condition or a predisposition to a condition from their gut or skin microbiome would help diagnostics and provide valuable insight for the design of personalized treatments focused on re-equilibrating the microbiome. Before presenting our novel approach for phenotype prediction, we set the scene by reviewing previous approaches.

Related works

Current microbiome analytics for phenotype prediction from metagenomic data can be largely split into referenced-based or *de novo* approaches. Reference-based approaches are based on the sequencing of specific marker genes, typically those for the bacterial 16S rRNA. Sequencing generates stretches of DNA bases called reads, which are processed through bioinformatics pipelines such as Quantitative Insights Into Microbial Ecology (QIIME) and clustered into groups called Operational Taxonomy Units (OTUs). Each OTU represents a different microbial species or taxon. OTU tables, summarizing the relative abundance of microbial species for a set of biological samples, are sparse and high dimensional data that have been recently used to train machine learning models such as SVMs, RFs and NNs.

A 2016 machine learning framework (Pasolli et al. 2016) uses quantitative microbiome profiles, including species-level relative abundances and presence/absence of species- and strain-specific markers, as features for ML models. The authors evaluate the use of SVMs, RFs, Lasso and ENet to predict five diseases (liver cirrhosis, colorectal cancer, inflammatory bowel diseases (IBD), obesity, and type 2 diabetes) from six available metagenomic datasets.

In (Carrieri, Haiminen, and Parida 2017) various normalization methods for OTUs table and their impact on phenotype prediction are evaluated for human, mouse, and environmental studies. The authors also address the problem of identifying the most relevant microbial features (OTUs) that could give insight into the structure and function of the differential microbial communities observed between phenotype groups.

The OTU representation of metagenomic data has several disadvantages. Firstly, the construction of the OTU table involves a very large number of sequence alignments, either to the reference genomes (in closed reference strategies) or among sequences present in the sample (de novo strategies), which makes it computationally expensive (Cai et al. 2017). Finally, the resulting OTU abundance tables are biased and sensitive to the specific bioinformatic pipeline used to generate them (He et al. 2015). This could have an impact on accuracy when attempting to predict phenotypes from OTU tables (Ross et al. 2013), (Karlsson et al. 2013).

A recent work (Asgari et al. 2018) presents a reference and alignment-free approach for predicting the phenotype from microbial community samples based on k-mer distributions in 16S rRNA marker gene sequences. K-mers are short overlapping sequences extracted from reads which collectively capture the genetic make-up of a genome or metagenome. The authors apply deep learning methods as well as traditional machine learning approaches for distinguishing among human body-sites, diagnosing Crohn's disease, and predicting the environments from representative 16S gene sequences. The authors demonstrate that k-mer features outperform Operational Taxonomic Unit (OTU) features. However, large amounts of sequencing data are excluded from the analyses as these are based on the sequencing of a single marker gene (16S rRNA).

To our knowledge, current machine learning workflows for phenotype prediction from microbiome data do not allow the prediction of host organism traits from the whole genetic material that is available as the metagenome.

Our approach

In this paper, we present a new fast machine learning workflow that performs phenotype prediction taking into account the whole genetic material in the microbiome and not only specific marker genes. Whole metagenome shotgun sequencing is, in fact, a more informative method which generates huge collections of short sequences called reads. De novo approaches (e.g. k-mer composition) can be applied to whole shotgun metagenomes for their analysis and comparison. For example, the pairwise comparison of k-mer spectra of metagenomes is a de novo analysis method that has been widely used in recent years for efficiently clustering microbiomes using dissimilarity measures (Dubinkina et al. 2016). K-mer spectra could be used as sets of features for training machine learning models and attempting phenotype prediction. However, considering that the size of a single metagenome of one sample typically varies from 1GB to tens of GB (depending on the amount of starting material and sequence effort) the corresponding k-mer spectrum can still take considerable time to compute and can be relatively large in size making the training and prediction task challenging. The first step in our approach is therefore computing a compact representation, or signature, of whole shotgun metagenomes that can be used as a set of features for training machine learning models.

Some very recent studies focused on developing methods to generate representative sketches of metagenomes (Luo et al. 2018), (Ondov et al. 2016). Mash (Ondov et al. 2016) is a method that reduces large sequences to small, representative sketches. Mash extends the MinHash (Broder 1997) dimensionality-reduction technique to include a pairwise mutation distance and P value significance test, enabling the efficient clustering and search of massive sequence collections. However, although MinHash-based tools like Mash can be used to great effect for certain microbiome analytics, there are limitations to standard MinHash techniques; such as the loss of k-mer frequency information and the impact of relative set size on Jaccard similarity estimates (Koslicki and Zabeti 2017). We compute compact representations of metagenomes by applying HULK (Histosketching Using Little K-mers) (Rowe et al. 2018), a fast method that enables the reduction of microbiome sequence data streams to an updateable histosketch of the underlying k-mer spectrum for a metagenome. We then use histosketches generated by HULK to train several machine learning models and perform for the first time, to the best of our knowledge, phenotype prediction from whole shotgun metagenomes. We apply our ML workflow to real public available data demonstrating we are able to efficiently predict phenotypes from whole shotgun metagenomes with high accuracy, making a step forward towards real time diagnostics.

Methods

In this section we describe in more detail the steps of our machine learning workflow, as shown in Fig. 1.

Data

We apply our ML workflow to two different public datasets: whole shotgun human metagenomes from the Human Microbiome Project (HMP Consortium 2012) and whole shotgun dog gut metagenomes (Coelho et al. 2018).

We downloaded the HMP metagenomes from https://www.hmpdacc.org/hmp/HMASM/ randomly selecting 365 samples out of 690 samples, amounting to approximately 1.6 TB of DNA sequence files (or FASTQ files). The phenotype that we predict for this dataset is the body site as 163 are mouth metagenomes, 108 are skin metagenomes and 94 are stool metagenomes. In (HMP Consortium 2012) the authors state that even healthy individuals differ remarkably in the microbes that occupy habitats such as the gut, skin, vagina and mouth, and so these sites should be distinguishable from metagenomic data. This is supported by (Ondov et al. 2016) who show that samples cluster appropriately by body site. We choose this data to demonstrate that our workflow is able to efficiently predict the phenotype from whole metagenomes with high accuracy.

The second dataset we include in our analysis is dog gut microbiome. In this case, we want to be able to predict if dogs have been fed with an altered diet (high protein low carb or high carb low protein). We analyze a total of 4450 metagenome FASTQ files (approximately 1.5 TB); about 50 sequencing runs per dog. We decide to treat each of the 4450 metagenomes as an individual sample in order to have more training data for our ML models.

Histosketch generation

The first step of our machine learning workflow is to generate a compressed representation of each metagenome sample called a histosketch (Yang et al. 2017). To perform this task, we apply a rapid sketching of metagenomic sequence data as implemented in HULK (Histosketching Using Little Kmers). HULK first converts the sequencing reads to overlapping short sub-sequences called k-mers, which are hashed uniformly across a set of bins giving a k-mer spectrum. The frequency value of each bin approximates the observed k-mer frequency, which is an important measure of the overall genetic content. Rather than store the full k-mer spectrum, which would require significant compute time and memory, HULK incrementally updates a fixed-size similarity-preserving histosketch data structure. This compressed representation of the k-mer spectrum utilizes consistent weighting sampling (CWS) to incorporate the k-mer frequency information. The histosketch is updatable as new sequence data is read in, and can be generated from streaming data. The resulting histosketch consists of a set of elements that are k-mer bins selected from the original spectrum according to the CWS scheme, and associated hash values that are used in the selection procedure. The number of elements is the sketch size s, which determines the level of sampling. More details about the histosketching method are provided in (Rowe et al. 2018). We computed histosketches of 365 human metagenomes taken from multiple body sites (HMP dataset) and of 4450 dog gut metagenomes. The sketches were generated using k-mers of length $k = 21$ and a sketch size of $s = 512$. The sketching of a typical 1GB file of sequencing reads, reducing it to a 512 vector, takes about 24 seconds to compute using 4 cores on a standard desktop computer running Mac OSX.

Multi-class and binary classification

In this paper, we focus on machine learning approaches for multi-class and binary classification of microbiome samples

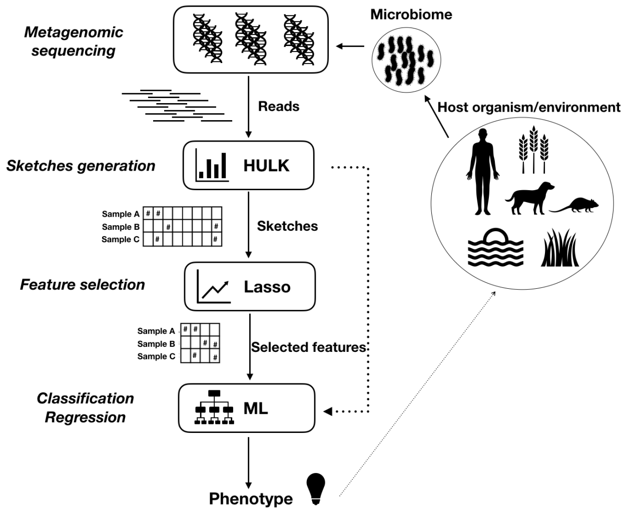

Figure 1: ML workflow for phenotype prediction from whole shotgun metagenomes

from whole shotgun metagenomics data. Metagenomic histosketches, computed by HULK, are feature vectors that can be used to train different machine learning models, and we compare their performance in terms of prediction accuracy and time. More precisely, we use the histosketch elements (bin values) and we discard the associated hash values.

For the human microbiome dataset (HMP), we perform multi-class classification to predict the body site from which the microbial sample has been taken: mouth, stool or skin. For the dog gut microbiome phenotype prediction task, we instead perform binary classification to predict if dogs have been fed with a baseline diet or with an altered diet.

There are a multitude of potential machine learning methods which can be used for these kinds of classification task. Random Forests (RF) and Support Vector Machines (SVM), in particular, have achieved widespread use on classification tasks. They have been applied previously to microbiome data (Statnikov et al. 2013), although using the OTU tables for feature construction rather than whole metagenome histosketches. Here, we use an SVM with a radial basis function (RBF) kernel and RFs.

We also include the Bayesian equivalent of the SVM, the Relevance Vector Machine (RVM). Due to the Bayesian nature of RVMs, they provide inbuilt uncertainty quantification, and whilst more computationally intensive to train, we believe that having error bars on predictions is worth the extra cost for situations in which a false prediction can have severe consequences, such as in healthcare. Other Bayesian methods, such as Gaussian process classifiers, do exist, however they were not investigated as their computational complexity, and subsequent poor scaling, meant that they did not fit our application criteria.

Finally, we investigated the application of the probabilistic classifier Naive Bayes (NB) for the prediction of phenotype, since this technique is well known to be particularly fast to train, and additionally, is able to predict a probability distribution over the set of classes, rather than only the most likely class that the sample belongs to.

For both datasets, we train each model using 80% of the data for and retain 20% as a test set. We then use the training set to perform 10 fold cross validation (CV) to flag problems such as overfitting or underfitting, and to give an indication of out of set performance, and thus model uncertainty. We choose not to perform CV for RVMs as the uncertainty is captured within the model itself.

In order to futher reduce the complexity of the problem, and thus the flexibility of the resulting models, we also apply feature selection, using LASSO (Tibshirani 1996). LASSO removes irrelevant features by placing a constraint on the sum of the absolute values of the model parameters, thus penalizing small coefficients for model parameters, shrinking them to zero. We compare the prediction performance of the different models considering the whole set of features (512 sketch elements) and only the ones selected by LASSO.

To perform our analyses, we use implementations of RFs, SVMs, NBs, Lasso and CV in the Python library scikit-learn (Pedregosa et al. 2011).

Results

In this section, we present the results obtained performing multi-class classification on the HMP dataset and binary classification on the dog gut dataset. The human microbiome dataset has 365 samples of which 163 are taken from mouth, 108 are skin microbiome samples and 94 are stool metagenomes. We then perform three-class classification to predict the body site from which the sample comes from. The dog gut microbiome dataset contains 4450 samples, where 2156 are associated with a baseline diet and the rest with an altered diet. Table 1 shows the F1-scores obtained from four different models on training and test sets for the human and dog datasets. The last two columns for each dataset report the mean and standard deviation of the F1-score over 10-fold cross validation.

For the human microbiome data, all models except Naive Bayes give high F1 scores ($>= 0.89$ when all features used) on training and test sets. The accuracy results for the dog gut microbiome samples are also promising, as all four models give F1-scores $>= 0.81$. The best models, for this dataset, seem again to be SVMs and RVMs with an F1 score of 0.91 on the test set when using all 512 features, followed by RFs with a score of about 0.9. While all the other models lose some accuracy when predicting the kind of dog diet using only the features selected by Lasso, Naive Bayes suffers the least loss from the original feature set, and is significantly faster to train and deploy than any of the other methods.

Speed Accuracy Trade-Off

For a machine learning method to provide insight on a relevant timescale, in addition to being accurate, it must also be fast to train and infer from. Training time will affect the ability to keep the model fresh as new training data is made available, and inference time will affect the ability of the model to be deployed in a real-time diagnostic situation. We graphically demonstrate the tradeoff which takes place be-

Table 1: F1 scores on training and test sets for HMP and dog gut data using the entire set of features (512 sketch elements) and only the ones selected by Lasso (respectively 63 features for the HMP dataset and 43 features for the dog dataset). F1 mean and standard deviation are also reported for 10 fold CV. The best F1 score on the test set is shown in bold.

		HMP data - F1 score				**Dog data - F1 score**			
		Train	Test	Mean CV	Std CV	Train	Test	Mean CV	Std CV
RVMs	All feat.	1	0.89	-	-	1	**0.91**	-	-
	Lasso	1	0.83	-	-	0.89	0.82	-	-
SVMs	All feat.	0.99	**0.98**	0.93	0.04	0.98	**0.91**	0.93	0.013
	Lasso	0.98	0.90	0.90	0.03	0.86	0.82	0.81	0.01
RFs	All feat.	0.99	0.97	0.95	0.03	0.99	0.90	0.89	0.01
	Lasso	0.99	0.90	0.88	0.02	0.89	0.81	0.81	0.006
NB	All feat.	0.79	0.73	0.76	0.08	0.82	0.82	0.81	0.07
	Lasso	0.82	0.73	0.76	0.08	0.82	0.82	0.81	0.01

tween training speed and accuracy by plotting these functions, as well as the model variance across cross validation in Figure 2. In all cases, the speed of prediction by a trained model is even faster than the creation of the histosketch for a sample, and is not an issue in practice.

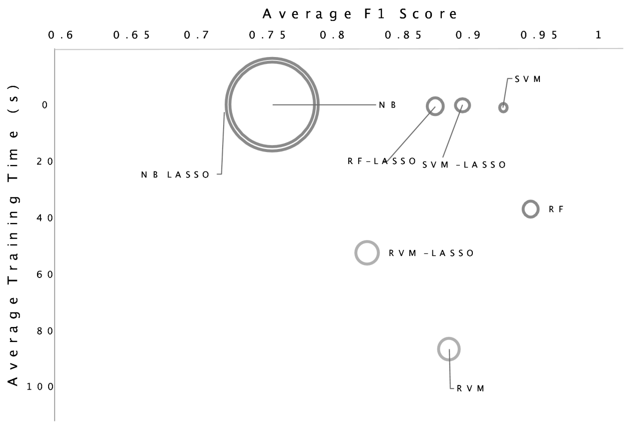

Figure 2: A bubble chart representing the tradeoff between the training speed and accuracy for each model, as calculated for the HMP dataset. The optimal position for a model to be is in the top right. The size of each bubble represents the negative log-likelihood of the true labels given a probabilistic classifier's predictions. For the deterministic methods, this was calculated using Platt's method, whilst for the Bayesian methods the probability estimate was calculated directly.

The uncertainties used to generate Figure 2 were calculated using Wu's generalization of Platt scaling (Wu, Lin, and Weng 2004). This method is known to not scale well, and also has some theoretical concerns (Niculescu-Mizil and Caruana 2005). It may therefore be the case that the slightly worse performance of RVMs is preferred, as they include a more robust definition of the uncertainty in their predictions.

Data Size Compression

An additional benefit of this workflow is the magnitude of the data compression which is achieved through sketching. This is important, as the initial metagenomic datasets are in the GB to TB range, making it impossible to exploit, for example, a cloud infrastructure. The histosketching significantly reduces this size, with the size of a single sketch now sitting in the hundreds to thousands of bytes range (see Table 2). Lasso allows further compression although it comes at the expense of predictive performance (see Table 2).

Data type	Representative size
Whole Metagenome	GB
Histosketch	KB
Lasso-histosketch	0.1 KB

Table 2: Typical data sizes at each stage of the pipeline. Whilst it is impractical to transfer GB of data to the cloud, KB can be managed from even a poor internet connection.

Summary

In order for the promise of AI to move beyond academic endeavour to real applications, it is necessary that the machine learning workload be accurate, fast and scalable. In this work we have demonstrated how, for the first time, we can build an AI workflow which ingests whole metagenomes (at GB scale) and is able to reduce this data into small *histosketches*. The latter compact representations of whole metagenomes are small enough to be amenable to cloud applications, and yet contain enough information for common machine learning methods to produce fast, yet accurate, results. We investigated a range of different types of machine learning algorithms, for their ability to provide accurate solutions at a computational cost small enough to enable the potential for real-time diagnostics. Finally, we consider the cost of adding uncertainty quantification, either through Bayesian methods, or through cross validation-based approaches, since understanding the context of a prediction in terms of its uncertainty is of high importance in many situations in which AI may be deployed.

Acknowledgements

This work was supported by the STFC Hartree Centre's Innovation Return on Research programme, funded by the Department for Business, Energy & Industrial Strategy.

References

Asgari, E.; Münch, P. C.; Lesker, T. R.; McHardy, A. C.; and Mofrad, M. R. 2018. Nucleotide-pair encoding of 16s rrna sequences for host phenotype and biomarker detection. *Bioinformatics* 34(13):i32–i42.

Broder, A. Z. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, 21–29.

Cai, Y.; Zheng, W.; Yao, J.; Yang, Y.; Mai, V.; Mao, Q.; and Sun, Y. 2017. Esprit-forest: Parallel clustering of massive amplicon sequence data in subquadratic time. *PLOS Computational Biology* 13(4):1–16.

Carrieri, A. P.; Haiminen, N.; and Parida, L. 2017. Host phenotype prediction from differentially abundant microbes using rodeo. In Bracciali, A.; Caravagna, G.; Gilbert, D.; and Tagliaferri, R., eds., *Computational Intelligence Methods for Bioinformatics and Biostatistics*, 27–41. Cham: Springer International Publishing.

Coelho, L. P.; Kultima, J. R.; Costea, P. I.; Fournier, C.; Pan, Y.; Czarnecki-Maulden, G.; Hayward, M. R.; Forslund, S. K.; Schmidt, T. S. B.; Descombes, P.; Jackson, J. R.; Li, Q.; and Bork, P. 2018. Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* 6(1):72.

Consortium, T. E. M. P. 2017. A communal catalogue reveals earth's multiscale microbial diversity. *Nature* 551:457 EP –.

Dubinkina, V. B.; Ischenko, D. S.; Ulyantsev, V. I.; Tyakht, A. V.; and Alexeev, D. G. 2016. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics* 17(1):38.

He, Y.; Caporaso, J. G.; Jiang, X.-T.; Sheng, H.-F.; Huse, S. M.; Rideout, J. R.; Edgar, R. C.; Kopylova, E.; Walters, W. A.; Knight, R.; and Zhou, H.-W. 2015. Erratum to: Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3(1):34.

HMP Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.

Jansson, J. K., and Hofmockel, K. S. 2018. The soil microbiome – from metagenomics to metaphenomics. *Current Opinion in Microbiology* 43:162 – 168. Environmental Microbiology - The New Microscopy.

Karlsson, F. H.; Tremaroli, V.; Nookaew, I.; Bergström, G.; Behre, C. J.; Fagerberg, B.; Nielsen, J.; and Bäckhed, F. 2013. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature* 486:207–214.

Koslicki, D., and Zabeti, H. 2017. Improving min hash via the containment index with applications to metagenomic analysis. *bioRxiv*.

Libbrecht, M. W., and Noble, W. S. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16:321–332.

Luo, Y.; Yu, Y. W.; Zeng, J.; Berger, B.; and Peng, J. 2018. Metagenomic binning through low-density hashing. *Bioinformatics* bty611.

Niculescu-Mizil, A., and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632. ACM.

Ondov, B. D.; Treangen, T. J.; Melsted, P.; Mallonee, A. B.; Bergman, N. H.; Koren, S.; and Phillippy, A. M. 2016. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology* 17(1):132.

Pasolli, E.; Truong, D. T.; Malik, F.; Waldron, L.; and Segata, N. 2016. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology* 12(7):1–26.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12:2825–2830.

Ross, E. M.; Moate, P. J.; Marett, L. C.; Cocks, B. G.; and Hayes, B. J. 2013. Metagenomic predictions: From microbiome to complex health and environmental phenotypes in humans and cattle. *PLOS ONE* 8(9):1–8.

Rowe, W. P.; Carrieri, A. P.; Alcon-Giner, C.; Caim, S.; Shaw, A.; Sim, K.; Kroll, J. S.; Hall, L.; Pyzer-Knapp, E. O.; and Winn, M. D. 2018. Streaming histogram sketching for rapid microbiome analytics. *bioRxiv*.

Rusch, D. B., and al. 2007. The sorcerer ii global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLOS Biology* 5(3):1–34.

Soueidan, H., and Macha, N. 2018. Machine learning for metagenomics: methods and tools. *Metagenomics* 1(1).

Statnikov, A.; Henaff, M.; Narendra, V.; Konganti, K.; Li, Z.; Yang, L.; Pei, Z.; Blaser, M. J.; Aliferis, C. F.; and Alekseyenko, A. V. 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1(1):11.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Wu, T.-F.; Lin, C.-J.; and Weng, R. C. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5(Aug):975–1005.

Yang, D.; Li, B.; Rettig, L.; and Cudré-Mauroux, P. 2017. Histosketch: Fast similarity-preserving sketching of streaming histograms with concept drift. In *2017 IEEE International Conference on Data Mining (ICDM)*, 545–554.