

TECHNICAL NOTE

Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks

Pajau Vangay ¹, Benjamin M. Hillmann ² and Dan Knights ^{1,2,*}

¹Bioinformatics and Computational Biology, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455; and ²Department of Computer Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455

*Correspondence address. Dan Knights, Bioinformatics and Computational Biology, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. E-mail: dknights@umn.edu  <http://orcid.org/0000-0002-8205-2511>

Abstract

The use of machine learning in high-dimensional biological applications, such as the human microbiome, has grown exponentially in recent years, but algorithm developers often lack the domain expertise required for interpretation and curation of the heterogeneous microbiome datasets. We present Microbiome Learning Repo (ML Repo, available at <https://knights-lab.github.io/MLRepo/>), a public, web-based repository of 33 curated classification and regression tasks from 15 published human microbiome datasets. We highlight the use of ML Repo in several use cases to demonstrate its wide application, and we expect it to be an important resource for algorithm developers.

Keywords: microbiome; machine learning; repository; database

Findings

Background

Machine learning is widely used as a method for classification and prediction, with a growing number of applications in human health [1]. The use of machine learning in biological fields [2, 3], and more specifically the microbiome research field [4–7], has grown exponentially owing to the robustness of these algorithms to high-dimensional data. However, challenges exist for large-scale meta-analysis because they often require manual curation of metadata and standardized processing of raw sequence data, resulting in variation in the results derived from chosen datasets across studies [8, 9]. In addition, microbiome research data can be challenging to access and analyze for expert machine learning algorithm developers, who often do not have the domain expertise required to interpret the data and metadata in complex microbiome studies. There exist general resources with curated classification tasks from a variety of domains. The

University of California Irvine Machine Learning Repository [10] revolutionized machine learning methods development by giving developers access to many curated datasets; its widespread usage and impact can be seen from its thousands of resulting citations. Currently, we are unaware of any machine learning repository dedicated to microbiome classification tasks. We constructed a complementary database to address this deficiency, in order to promote the development of and use of improved machine learning methods for the microbiome community.

Workflow

We present the Microbiome Learning Repo (ML Repo), a repository of 33 curated classification and regression tasks involving human microbiome data. Our 33 tasks are derived from 15 publicly available human microbiome datasets, which include 12 amplicon-based and 3 shotgun sequencing datasets (Table 1). These datasets vary across sequencing technology platforms,

Received: 21 August 2018; Revised: 24 February 2019; Accepted: 26 March 2019

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1: Microbiome datasets with available classification tasks in ML Repo

Project name	V Region	Target size	No. samples	No. subjects	Area	Description	Sequencing technology	Study design
Cho 2012	V3	177	95	47	Antibiotics	Mouse fecal and cecal samples, control vs 4 kinds of antibiotics	454	Cross-sectional
Claesson 2012	V4	221	168	168	Age	Elderly and young adults	454	Cross-sectional
David 2014	V4	282	235	11	Diet	Plant-based vs animal-based diet, cross-over study	Illumina MiSeq	Longitudinal
Gevers 2014	V4	173	1,321	668	IBD	Biopsies from patients with IBD prior to treatment	Illumina MiSeq	Cross-sectional
HMP 2012	V35	527	6,407	242	Body habitat, sex	Up to 18 body sites across 242 healthy subjects at 1–2 time points	454	Cross-sectional
Kostic 2012	V35	569	190	95	Colorectal cancer	Adjacent healthy vs tumor colon biopsy tissues	454	Paired
Montassier 2016	V56	280	28	28	Bacteremia	Patients prior to chemotherapy who did or did not develop bacteremia	454	Cross-sectional
Morgan 2012	V35	569	231	231	IBD	Healthy controls, patients with Crohn’s disease or ulcerative colitis	454	Cross-sectional
Turnbaugh 2009	V2	230	281	154	Obesity	Monozygotic or dizygotic twin pairs concordant for body mass index class, and their mothers	454	Cross-sectional
Wu 2011	V12	244	95	10	Diet	Controlled high-fat or low-fat feeding on 10 subjects over 10 days	454	Longitudinal
Yatsunenko 2012	V4	282	531	531	Geography, age, sex	Humans of varying ages from the USA, Malawi, and Venezuela	Illumina MiSeq	Cross-sectional
Ravel 2011	V12	240	396	396	Bacterial vaginosis	Vaginal samples from 4 ethnic groups; Nugent scores for bacterial vaginosis	454	Cross-sectional
Karlsson 2013	NA	NA	144	144	Diabetes	Patients with normal, impaired, or type 2 diabetes glucose tolerance categories	Illumina HiSeq	Cross-sectional
Qin 2012	NA	NA	134	134	Diabetes	Chinese healthy controls vs patients with type 2 diabetes	Illumina HiSeq	Cross-sectional
Qin 2014	NA	NA	130	130	Cirrhosis	Healthy controls vs patients with cirrhosis	Illumina HiSeq	Cross-sectional

ML Repo contains 33 classification and regression tasks from 15 publicly available human microbiome datasets shown here. IBD: inflammatory bowel disease; NA: not applicable.

16s hypervariable regions, and study design, in order to help developers ensure robustness of algorithms across data types. We streamlined the microbiome data using a single post-processing workflow (Fig. 1A). We downloaded trimmed and quality-filtered sequencing reads for 8 datasets from QIITA [11], and raw sequences for 7 datasets from public repositories. Raw sequences were trimmed and quality filtered using SHI7 [12] or QIIME [13]. We picked operational taxonomic units (OTUs) from all quality-filtered sequences using a closed-reference method with the BURST [14] aligner against both the National Center for Biotechnology Information (NCBI) RefSeq 16S ribosomal RNA project [15] and the Greengenes 97 database [16]. Samples with <1,000 sequencing reads were dropped for 10 datasets, while we applied a lower threshold of 100 sequencing reads per sample for 5 datasets that had lower expected bacterial load. Full details regarding the data preprocessing are provided for each data set in the *mlrepo-source* branch of the GitHub repository, under *preprocessing/make.mappings.r*. As a result, for each dataset we generated RefSeq-based OTU and taxa abundance counts, and Greengenes-based OTU and taxa abundance counts. These counts are presented in tables that are organized as follows: OTUs or taxa as rows, and samples as columns. OTUs are represented as either NCBI genome identifiers or Greengenes identifiers. Taxa are represented as “kingdom; phylum; class; order;

family; genus; species; strain,” with highest taxonomic specificity where possible. We excluded additional post-processing filtering and normalization steps so that these parameters can be included in future benchmarking use cases as needed. We also limit our data to OTU and taxa tables because other metrics such as α and β diversity can be subsequently generated as needed.

Sample metadata from individual studies were manually curated to generate viable prediction tasks. When available, published study exclusion criteria, such as reported use of antibiotics, were applied accordingly and confounders were removed by dropping samples or stratification. Well-known confounders were accounted for when constructing prediction tasks for other human-associated conditions; e.g., predicting age using the Yatsunenko 2012 dataset is restricted to samples from the USA owing to the known variation in gut microbiomes across different geographical locations. Details of how samples were subset for each prediction task can be found in the *mlrepo-source* branch of the GitHub repository, under *preprocessing/make.mappings.r*. Studies that were cross-sectional by design but contained several samples per subject were filtered to contain 1 sample per subject. In study designs with paired diseased-healthy or pre- and post-intervention samples, samples were reduced to 2 samples per subject with subject identifiers provided as confounder

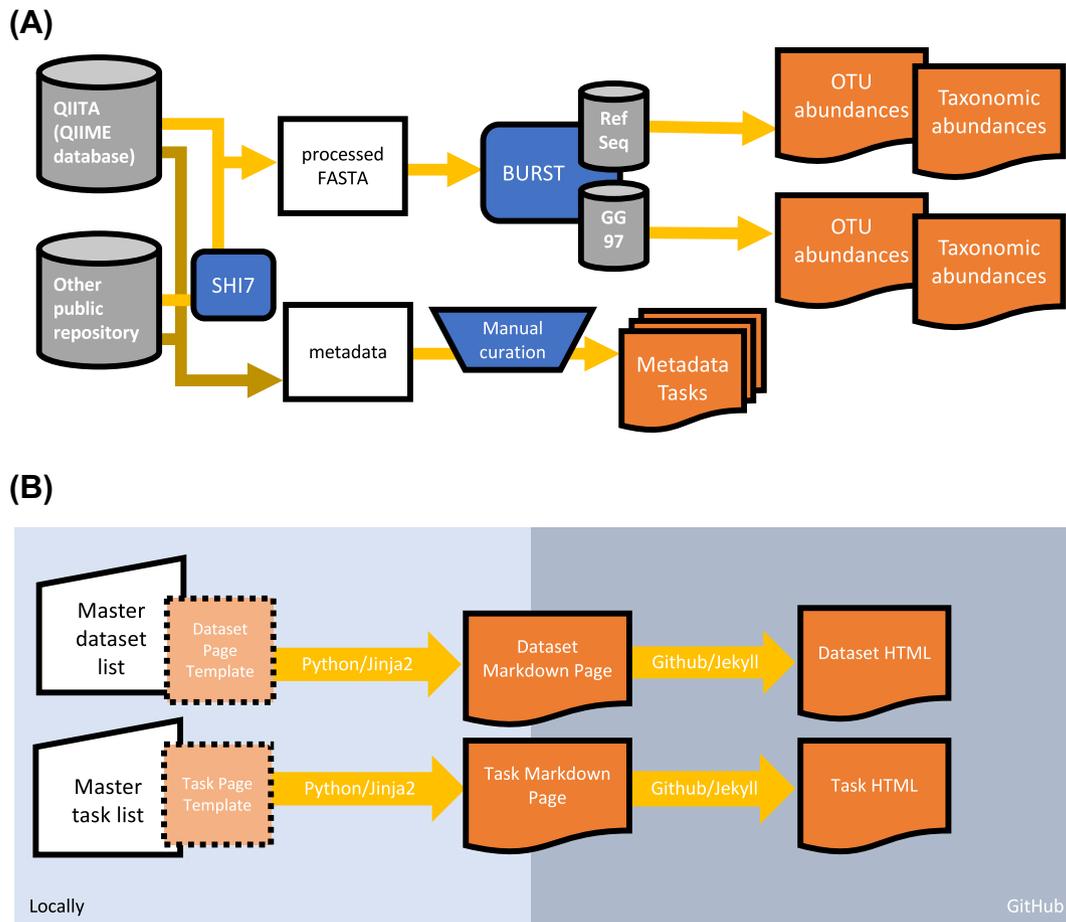


Figure 1: Data processing workflow and website generation. (A) Quality-filtered sequences were obtained from either the QITA or from another public repository and trimmed and filtered using SHI7. Reference-based OTUs were picked using BURST with the NCBI RefSeq and Greengenes 97 (GG 97) databases. (B) Individual GitHub Markdown pages were generated from dataset and task lists with a custom Python script and Jinja2 template, then uploaded to GitHub to be hosted.

variables. Hence, each prediction task is made available as an individual, compartmentalized metadata file that contains sample identifiers, responses to predict, and optionally, confounder variables that are inherent to the research study design such as paired healthy and diseased samples from the same subject (see Methods for more details). As a result, we generated 33 distinct tasks for predicting human-associated responses.

Publicly available web-based interface

We expect 2 types of users: (i) machine learning algorithm developers with limited knowledge of microbiome study designs and (ii) microbiome researchers interested in obtaining additional datasets for meta-analysis. Generally, we expect that method developers will be most interested in sweeping through the full set of prediction tasks for benchmarking, and hence would prefer to download a single compressed file containing all tasks and data. On the other hand, we expect microbiome researchers to be more selective in downloading specific datasets and tasks depending on their research domain. Hence, researchers may prefer to browse specific details about tasks and datasets prior to downloading.

On the basis of these expected use cases, we created a publicly available web-based interface for ML Repo hosted by GitHub Pages [17]. Tasks are organized by relevant response categories (Fig. 2A). Task pages contain descriptive details such as sam-

ple size and response type that are specific to the selected prediction task, as well as links for downloading OTU tables, taxa tables, and sample metadata (Fig. 2B). Dataset pages contain important details about the entire dataset, including links to the original research study, as well as original metadata files and quality-filtered sequences (Fig. 2C). We also provide a single compressed file containing the entire set of available tasks (OTU tables, taxa tables, and relevant metadata) for download from the main home page.

Benefits of curated microbiome-based prediction tasks

We expect ML Repo to be beneficial for both the machine learning community as well as the microbiome research community. ML Repo will be a powerful complement to the University of California Irvine's Machine Learning Repository because it will allow for benchmarking curated classification tasks with high-dimensional data and hence enable the subsequent development of novel algorithms for these complex datasets. Our streamlined approach in generating OTU and taxa tables offers a rich set of 15 datasets that microbiome researchers can use directly for further comparison with their own studies, for teaching and learning purposes, or for large meta-analyses. We expect that our provided OTU and taxa tables will also be beneficial for researchers with limited access to high-performance computing resources or bioinformatics skills necessary for processing

(A)

MLRepo
Machine learning repository for microbiome datasets

[View On GitHub](#)

Available Tasks

- Bacteremia
 - Download a single file containing all available tasks
 - Bacteremia**
 - › bacteremia vs no bacteremia
- Diet
 - Diet**
 - › high fat vs low fat diet
 - › animal vs plant diet, last diet day
- Antibiotics
 - Antibiotics**
 - › chlortetracycline vs control, cecal
 - › chlortetracycline vs control, fecal

(B)

Task: bacteremia vs no bacteremia
Patients prior to chemotherapy who did or did not develop bacteremia

Project	Montassier 2016
Topic area	Bacteremia
Sample type	human stool
Number of samples	28
Response type	binary
Additional task details	
Multiple samples per subject?	No
Task mapping file	task.txt
OTU file gg97	otutable.txt
Taxa file gg97	taxatable.txt
OTU file RefSeq	otutable.txt
Taxa file RefSeq	taxatable.txt

[back to task index](#)

(C)

Montassier 2016
Patients prior to chemotherapy who did or did not develop bacteremia

Overview	
Description	Patients prior to chemotherapy who did or did not develop bacteremia
Study design	Cross-Sectional
Topic area	Bacteremia
Attributes	Treatment: NObact, bact
Dataset notes	
Number of samples	28
Number of subjects	28

Other Details	
16s hypervariable region	V56
Targeted amplicon size	280
Sequencing technology	454
Fraction of sequences mapped to database	
Processed sequences	montassier2016.fasta.gz
Raw metadata file	mapping-orig.txt
Raw sequence source	https://www.ncbi.nlm.nih.gov/sra/SRX733464
Literature source	https://www.ncbi.nlm.nih.gov/pubmed/27121964

[back to task index](#)

Figure 2: Screenshots of ML Repo web interface. (A) Available classification and regression tasks are listed by high-level phenotype categories for browsing. (B) Individual task webpages contain links to files for classifying a specific task, as well as relevant task-specific metadata. (C) Individual dataset webpages contain relevant metadata pertaining to the entire dataset, as well as links to raw metadata files and sequencing data.

raw sequencing data. In addition, we also expect microbiome-specific methods development to benefit from our repository. The subset of samples found in each prediction task metadata file replace the work of rigorously deciphering metadata and understanding the subtle differences of individual research studies. New methods, such as OTU-picking algorithms, can be eval-

uated not only on metrics such as speed and accuracy but also based on overall impact to study findings.

Comparison to similar databases

Although a number of microbiome repositories exist, many are

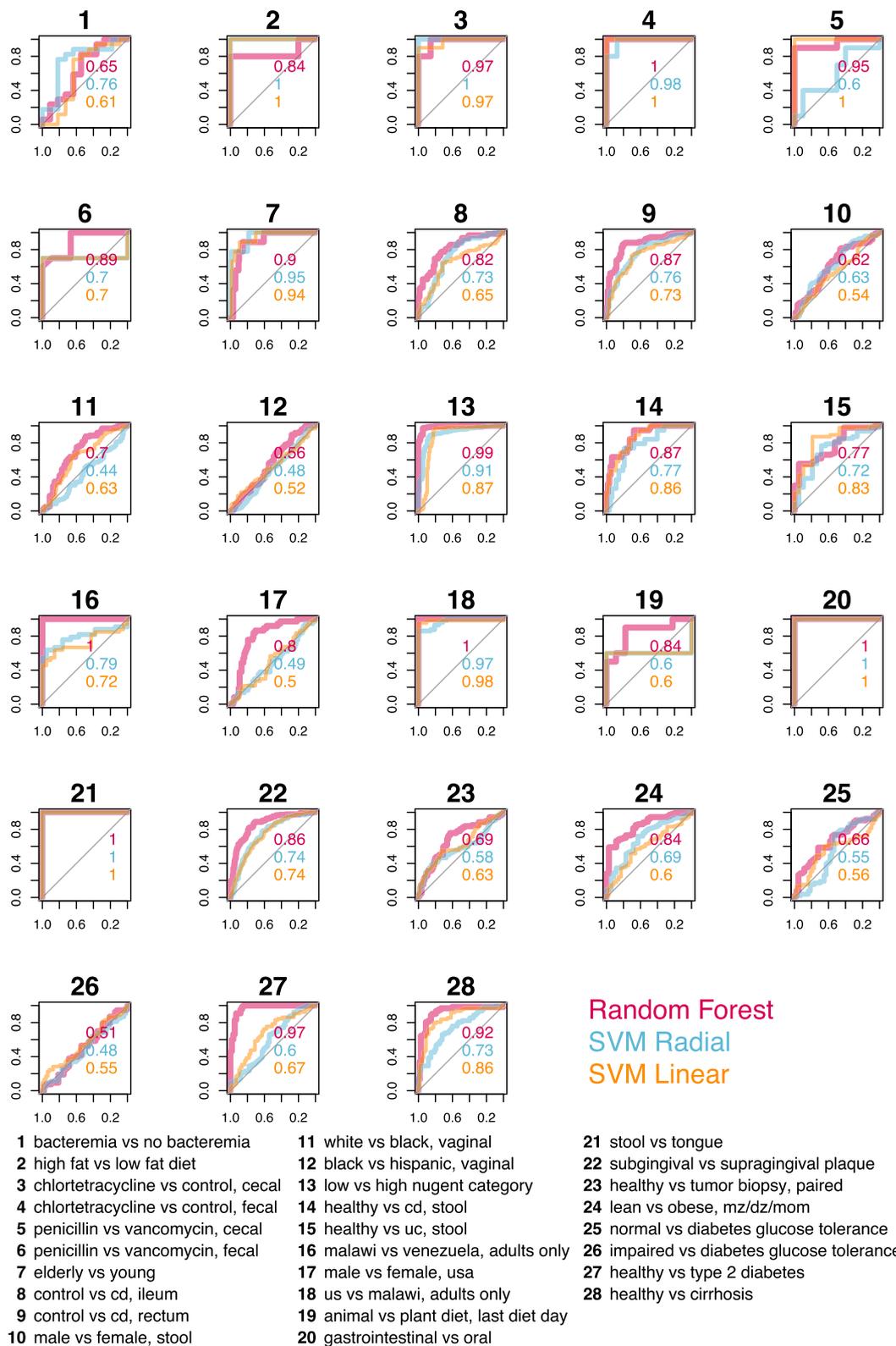


Figure 3: ROCs comparing random forest and SVM with different kernels. Sweeping across all binary classification tasks available in ML Repo (28), we compare ROCs of random forest, SVM with a radial kernel, and SVM with a linear kernel. AUCs are listed within plots and are colored respective to each model. cd: Crohn's disease; dz: dizygotic; mz: monozygotic; uc: ulcerative colitis.

intended as data archival repositories [18, 19] or function as resources for aggregating across studies [20]. Resources such as

QIITA [11] offer an extensive collection of datasets, and mock-community-based Mockrobiota [21] is well-suited for bench-

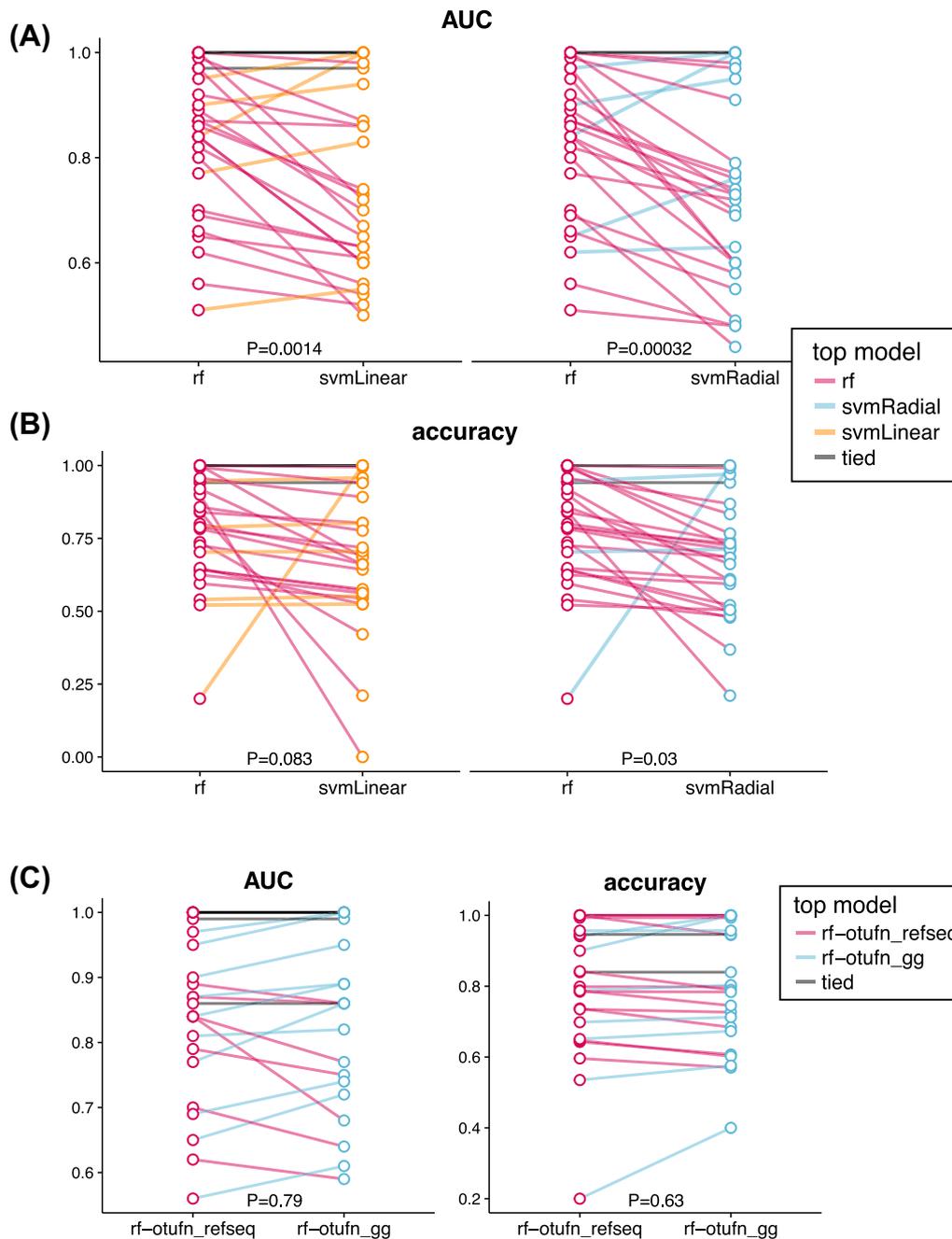
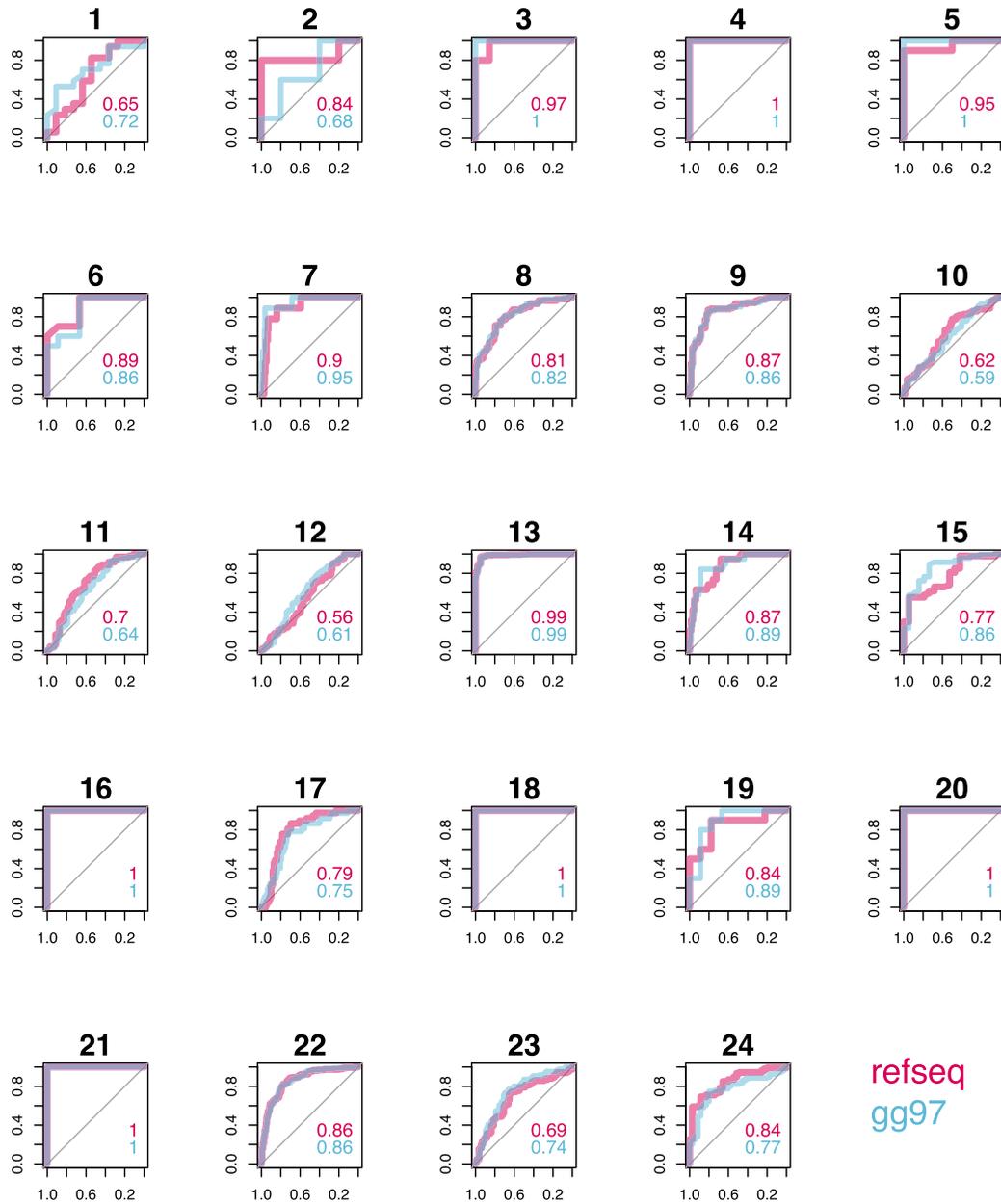


Figure 4: Summary statistics of framework and database comparisons. (A) AUCs of random forest (rf) to SVM-Linear (left) and random forest to SVM-Radial (right). Paired t-tests reveal that random forest results in significantly higher AUC than both SVM-Linear ($P = 0.0014$) and SVM-Radial ($P = 0.00032$). (B) Accuracies of random forest to SVM-Linear (left) and random forest to SVM-Radial (right). Paired t-tests reveal that random forest results in significantly better accuracy than SVM-Radial ($P = 0.03$), but not SVM-Linear ($P = 0.083$). (C) AUCs (left) and accuracies (right) of random forest classifications of 24 tasks using OTUs picked with NCBI RefSeq database or Greengenes (gg) database as predictors. Student t-test reveals that reference database choice has limited impact on classification AUC or accuracy. Lines are colored by the top model for each classification task.

marking upstream methods, but neither offers support for the metadata interpretation necessary for predicting high-level phenotypes. Microbiome-based repositories that do provide manually curated metadata include `curatedMetagenomicData` [22] and `MicrobiomeHD` [23]. Although `curatedMetagenomicData` offers a collection of shotgun-metagenomics datasets with varying human sample types with gene, pathway, and taxonomic abundance tables, its data are accessible only via `Bioconductor` [24] and are stored as `ExpressionSet` objects, which inte-

grates metadata and abundance data. Although `curatedMetagenomicData` is an impressive repository with many features, it is most suitable for advanced bioinformaticians because its interface may hinder use by beginner data analysts and in teaching environments. `MicrobiomeHD` offers easily accessible taxonomic abundance tables with curated metadata but is limited only to amplicon-based sequencing data, human stool samples, and case-control responses. And although both `curatedMetagenomicData` and `MicrobiomeHD` provide manually curated meta-



- | | |
|----------------------------------------------|-----------------------------------------------|
| 1 bacteremia vs no bacteremia | 13 low vs high nugent category |
| 2 high fat vs low fat diet | 14 healthy vs cd, stool |
| 3 chlortetracycline vs control, cecal | 15 healthy vs uc, stool |
| 4 chlortetracycline vs control, fecal | 16 malawi vs venezuela, adults only |
| 5 penicillin vs vancomycin, cecal | 17 male vs female, usa |
| 6 penicillin vs vancomycin, fecal | 18 us vs malawi, adults only |
| 7 elderly vs young | 19 animal vs plant diet, last diet day |
| 8 control vs cd, ileum | 20 gastrointestinal vs oral |
| 9 control vs cd, rectum | 21 stool vs tongue |
| 10 male vs female, stool | 22 subgingival vs supragingival plaque |
| 11 white vs black, vaginal | 23 healthy vs tumor biopsy, paired |
| 12 black vs hispanic, vaginal | 24 lean vs obese, mz/dz/mom |

Figure 5: ROCs comparing NCBI RefSeq and Greengenes 97 (gg97) databases. Sweeping across 16s-based binary classification tasks available in ML Repo (24), we compare ROCs of random forest with genus-level taxonomic summaries as predictors from OTU-picking strategies with the NCBI RefSeq prokaryote reference database and the Greengenes 97 reference database. AUCs are listed within plots and are colored respective to each database. cd: Crohn's disease; dz: dizygotic; mz: monozygotic; uc: ulcerative colitis.

data, biological interpretation is still required because other sample metadata, e.g., antibiotic use, may have biological relevance in predicting responses. This poses a potential problem for machine learning developers with limited biological and microbiome domain expertise. ML Repo resolves this issue by explicitly defining classification and regressions tasks for predicting responses that have either been manually curated to remove confounders or been specifically annotated with biological confounders that must be controlled for. Metadata files in ML Repo are task-specific and, hence, are simplified to contain only (i) sample identifiers indicating samples that should be used for the prediction task, (ii) corresponding high-level phenotypes or responses, and optionally, (iii) a confounder that should be accounted for owing to its biological relevance. In addition, datasets in ML Repo include both amplicon-based and shotgun-metagenomics datasets covering a variety of human sample types, and are easily accessible via a web-based interface.

Case studies

We compare the performance of 3 machine learning models: a random forest [25], and a support vector machine (SVM) [26] with either a radial or linear kernel. Sweeping through available tasks with binary responses, we compare our models by examining receiver operating curves (ROCs) and areas under the curve (AUC), considered the standard method for machine learning model evaluation [27, 28] (Fig. 3). Through comparison of ROCs, we can see that random forest outperforms or ties the other 2 models in 21 of the 28 tasks. The choice of kernels for SVM seems to have limited impact on overall mean accuracy, yet a linear kernel was able to perfectly classify penicillin-treated and vancomycin-treated mouse cecal contents when the other models could not; further examination of the microbial features in these samples may be warranted to better elucidate the strengths of this kernel. We also performed pairwise comparisons of random forest against the other models across all tasks. When evaluated by AUC, random forest performed significantly better than both SVM with a linear kernel ($P = 0.0014$) and with a radial kernel ($P = 0.00032$) (Fig. 4A). We found that random forest accuracy improvements were moderate when compared with SVM-Linear ($P = 0.083$) and SVM-Radial ($P = 0.03$) (Fig. 4B), which may be explained by the fact that, unlike AUC, accuracy ignores class prediction probability estimates. Our results support the broad usage [4, 5, 8, 29] and acceptance of random forest as a robust classifier [6] with high-dimensional microbiome data.

To assess the impact of reference database choice on classification accuracies, we also used the classification tasks to compare random forest using OTUs picked with the Greengenes 97 database or the NCBI RefSeq Targeted Loci Project 16s project. We found that there was limited impact of database choice on overall classification accuracies (Figs 4C and 5). This may be due to (i) large effect sizes that are driven mainly by several well-characterized bacterial taxa present in both databases (e.g., stool vs tongue samples), or (ii) small effect sizes such that classification is difficult regardless of the database (e.g., male vs female stool). Note that OTU-picking with the Greengenes database resulted in more OTU features in every dataset (Table 2); hence, these findings further highlight how the smaller, higher-quality NCBI RefSeq database can recover the same signal from the larger Greengenes database.

Future work

We expect and hope that the broader microbiome research community will add new datasets and prediction tasks to ML Repo. We provide instructions [30] on our GitHub repository to guide users to create a fork from our repository, add the appropriate data and files, and update the master task and dataset lists. Researchers can then submit a pull request for our review, and requests that are properly formatted will be accepted and merged into the repository. We expect that data submissions will come from either the original researchers or those well acquainted with the datasets, and hence will expect that sample selection and subsetting will have undergone rigorous review for prediction tasks.

Methods

Pre-processing of sequencing reads

When available, preprocessed FASTA files were downloaded from QIITA (or previously, the QIIME database). For all other datasets, raw FASTQ files were downloaded from sources listed in Supplemental Table 1. Adaptors and barcodes were removed and sequences were quality filtered (at Phred score $\geq Q20$) using SHI7 [12] or QIIME [13]. OTUs were picked from processed FASTA files using BURST [14] with Greengenes [16] 97 or the NCBI RefSeq Targeted Loci Project 16s project [15] (accessed on 4 July 2017). Samples with sequencing depth $< 1,000$ sequences per sample were dropped for all studies, except for 5 datasets [31–35], where the minimum threshold was 100 sequences per sample.

Selection of classification tasks

Classification tasks were selected on the basis of reported study results, biologically relevant high-level phenotypes, and sufficient sample sizes. Original metadata files and research methods were rigorously and manually curated in order to subset samples with minimal confounders. For confounders that were inherent to the study, we include an additional variable to control for in the task metadata files. The presence of control variables can be found by examining “control.vars” in the Tasks table.

Website generation

Website templating was developed using Jinja2 [36] and custom Python scripts. Individual webpages were generated by iterating through items in the Tasks and Datasets tables, and dynamically populating templates to generate individual Markdown [37] pages. The resulting Markdown pages are hosted as GitHub Pages.

Case study benchmarking

Case study results were generated with custom R [38] scripts, which can be found in the */example* folder in the ML Repo Github repository. To compare machine learning models, we iterated through tasks with binary responses. OTU counts were converted to relative abundances, filtered at a minimum of 10% prevalence across samples, and collapsed at a complete-linkage correlation of 95% (which is done by calculating the Pearson correlation between each pair of OTUs using all complete pairs of observations, hierarchically clustering the results, and cutting the resulting dendrogram at a height of 0.05). We then constructed a 5-fold cross-validation for tasks containing > 100 sam-

Table 2: Description of available prediction tasks

Dataset	Attributes	Description	Area	Regression?	Sample size	No. of Features				Control variable
						OTU, refseq	OTU, gg	Taxa, refseq	Taxa, gg	
Cho 2012	Abx: Control, Chlortetracycline	5 Groups of mice treated with 4 different antibiotics or no antibiotics	Antibiotics		47	293	1,144	299	141	N
	Abx: Control, Chlortetracycline	5 Groups of mice treated with 4 different antibiotics or no antibiotics	Antibiotics		45	293	1,144	299	141	N
	Abx: Penicillin, Vancomycin	5 Groups of mice treated with 4 different antibiotics or no antibiotics	Antibiotics		47	293	1,144	299	141	N
	Abx: Penicillin, Vancomycin	5 Groups of mice treated with 4 different antibiotics or no antibiotics	Antibiotics		45	293	1,144	299	141	N
Claesson 2012 David 2014	AGE: Elderly, Young Diet: Plant, Animal	Elderly or young adults Individuals on the last day of an animal or plant diet intervention	Age Diet		167 18	569 1,747	3,763 6,293	662 1,535	279 695	N Y
	DIAGNOSIS: no, CD	Healthy controls and patients with CD	IBD		140	943	3,547	992	446	N
Gevers 2014	DIAGNOSIS: no, CD	Healthy controls and patients with CD	IBD		160	943	3,547	992	446	N
	PCDAI	PCDAI scores of patients with CD at 6 months after sampling	IBD	X	68	943	3,547	992	446	N
HMP 2012	PCDAI	PCDAI scores of patients with CD at 6 months after sampling	IBD	X	51	943	3,547	992	446	N
	HMPBODYSUPERSITE: Oral, Gastrointestinal.tract, HOST.SUBJECT.ID SEX: male, female	Gastrointestinal tract and oral cavity of healthy adults	Body habitat		2,070	3,121	9,383	3,090	1,218	Y
	HMPBODYSUBSITE: Stool, Tongue.dorsum; HOST.SUBJECT.ID HMPBODYSUBSITE: Subgingival_plaque, Supragingival_plaque; HOST.SUBJECT.ID Classification: IGT, T2D	Healthy male and female adults Stool and tongue of healthy adults Subgingival and supragingival plaque of healthy adults	Sex Body habitat		180	3,121	9,383	3,090	1,218	N
	Classification: NGT, T2D	Impaired or type 2 diabetes glucose tolerance categories Normal or type 2 diabetes glucose tolerance categories	Diabetes		101	12,845	NA	3,758	NA	N
Karlsson 2013	Classification: NGT, T2D	Diabetes		96	12,845	NA	3,758	NA	NA	N

Table 2: Continued

Dataset	Attributes	Description	Area	Regression?	Sample size	No. of Features			Control variable	
						OTU, refseq	OTU, gg	Taxa, refseq		
Kostic 2012	DIAGNOSIS: Healthy, Tumor; HOST_SUBJECT.ID	Colorectal carcinoma tumors and adjacent nonaffected tissues	Cancer		172	908	3,228	980	409	Y
Montassier 2016	Treatment: bact, NObact	Patients prior to chemotherapy who did or did not develop bacteremia	Bacteremia		28	541	1,852	640	228	N
Morgan 2012	ULCERATIVE_COLITIS_OR_CROHNS.DIS: Crohn's disease, Healthy	Healthy controls or patients with CD or ulcerative colitis	IBD		128	829	3,677	877	367	N
	ULCERATIVE_COLITIS_OR_CROHNS.DIS: Ulcerative Colitis, Healthy	Healthy controls or patients with CD or ulcerative colitis	IBD		128	829	3,677	877	367	N
Qin 2012	Diabetic: Y, N	Healthy controls or patients with type 2 diabetes	Diabetes		124	11,880	NA	2,526	NA	N
Qin 2014	Cirrhotic: Cirrhosis, Healthy	Healthy controls or patients with cirrhosis	Cirrhosis		130	8,483	NA	2,579	NA	N
Ravel 2011	Ethnic.Group: Black, Hispanic	Vaginal microbiomes of black and Hispanic women	Vaginal		199	586	1,093	660	305	N
	Nugent.score.category: low, high	Predict Nugent score	Vaginal		342	586	1,093	660	305	N
	Nugent.score	category (low, high) from vaginal microbiome	Vaginal	X	388	586	1,093	660	305	N
	pH	Predict Nugent score from vaginal microbiome	Vaginal	X	388	586	1,093	660	305	N
	Ethnic.Group: White, Black	Predict pH from vaginal microbiome	Vaginal		200	586	1,093	660	305	N
Turnbaugh 2009	OBESITYCAT: Lean, Obese; ZYGOSITY: MZ, DZ, Mom	Vaginal microbiomes of white and black women	Obesity		142	557	4,051	680	232	Y
Wu 2011	DIET: HighFat, LowFat	Lean or obese individuals (monozygotic or dizygotic twins or their mothers) Individuals after completing a high-fat or low-fat diet intervention	Diet		10	292	1,769	361	136	N
Yatsunenکو 2012	AGE	Infants (up to age 3 years) from the USA	Age	X	49	4,660	15,783	4,021	1,544	N
	COUNTRY: GAZ: Venezuela, GAZ: Malawi	Individuals living in Malawi or Venezuela	Geography		54	4,660	15,783	4,021	1,544	N
	SEX: male, female	Males and females from the USA	Sex		129	4,660	15,783	4,021	1,544	N
	COUNTRY: GAZ: United States of America, GAZ: Malawi	Individuals living in the USA or Malawi	Geography		150	4,660	15,783	4,021	1,544	N

Abx: antibiotics; CD: Crohn's disease; DZ: dizygotic; GG: Greenegenes; IBD: inflammatory bowel disease; IGT: impaired glucose tolerance; MZ: monozygotic; NGT: normal glucose tolerance; PCDAI: Pediatric Crohn's Disease Activity Index; T2D: type 2 diabetes; GAZ: Gazeteer, an ontology of place names.

ples, or a leave-one-out cross-validation for tasks with fewer samples. For n-fold cross validation, samples were assigned to folds such that classes were equally balanced within each fold (e.g., if our task contained 40% healthy and 60% diseased samples, our folds would also be selected to represent this distribution). For tasks that contained control variables, we selected folds such that samples with the same control variable value were contained within the same fold. For example, for a task dataset containing matching stool and oral samples from subjects, the Subject Identifier would be listed as the control variable and we should assign samples to folds such that all samples from a specific subject were contained within a fold. This step is crucial to avoid biasing or overfitting the training model; test folds should contain not only new samples but also samples that are independent from those in the training set. Models were constructed using the “caret” package [39]. Control parameters were set using the function `trainControl` with parameter `method = “none”` and default parameters. Default settings for all models are as follows: SVM radial basis σ is set to 0.1, all SVMs [40]C is set to 1, and randomForest number of trees is set to 500 and number of variables to split is \sqrt{p} , where p is the number of features. This entire process was bootstrapped 100 times, and the mean class probabilities were used to calculate the resulting AUCs and ROCs. To compare classification accuracies using different reference databases, we used a similar procedure but held the model constant and predicted using different base OTU tables. This framework enables comparison of a myriad of machine learning models available in the “caret” package and can be easily expanded to compare different OTU-picking algorithms, or normalization and filtering techniques.

Availability of supporting data and materials

All test datasets are available in the Microbiome Learning Repo site [17]. Snapshots of our code and other supporting data are available in the GigaScience database, GigaDB [41].

Availability of supporting source code and requirements

Project name: Microbiome Learning Repo
 Project home page: <https://knights-lab.github.io/MLRepo/>
 Operating system: platform independent
 Programming language: Python, R
 License: MIT License
 Restrictions: None
 RRID:SCR_017079

Supporting information Abbreviations

AUC: area under the curve; IBD: inflammatory bowel disease; ML Repo: Microbiome Learning Repo; NCBI: National Center for Biotechnology Information; OTU: operational taxonomic unit; ROC: receiver operating curve; SVM: support vector machine.

Competing interests

D.K. serves as CEO and holds equity in CoreBiome, a company involved in the commercialization of microbiome analysis. The University of Minnesota also has financial interests in CoreBiome under the terms of a license agreement with CoreBiome. These interests have been reviewed and managed by the Uni-

Table 3 Glossary

Term	Definition
OTU	Operational taxonomic unit, group of closely related organisms based on DNA sequence similarity
16S	16S ribosomal RNA gene, component of the prokaryotic ribosome, used to reconstruct phylogenies
FASTA	Text-based format for representing nucleotide sequences with single-letter codes
FASTQ	Text-based format for representing nucleotide sequences and corresponding quality scores, with single-letter codes for nucleotides and quality
Taxa	Groups of ≥ 1 populations of organisms. Usually summarized at phylum, class, order, family, genus, or species levels
Metadata	Descriptive data pertaining to samples within a study
Shotgun	Shotgun metagenomics sequencing breaks up all available DNA into random small segments and uses chain termination to sequence reads. Reads can be aligned directly to a reference database, or overlapping reads can be assembled into contiguous sequences

versity of Minnesota in accordance with its Conflict-of-Interest policies.

Funding

This work is supported by funds from National Institutes of Health grant R01AI121383.

Authors' contributions

Conceptualization: P.V. and D.K.; data curation: P.V.; formal analyses: P.V.; methodology: P.V., B.M.H., D.K.; software: P.V.; writing—original draft: P.V.; writing—review and editing: B.M.H. and D.K.

References

- Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349:255–60.
- Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002;8:68–74.
- Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000;16:906–14.
- Aagaard K, Riehle K, Ma J, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One* 2012;7:e36466.
- Yatsunenkov T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature* 2012;486:222–7.
- Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS Microbiol Rev* 2011;35:343–59.
- Smith MI, Yatsunenkov T, Manary MJ, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 2013;339:548–54.
- Pasolli E, Truong DT, Malik F, et al. Machine learning meta-

- analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol* 2016;12:e1004977.
9. Sze MA, Schloss PD. Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio* 2016;7(4); doi: 10.1128/mBio.01018-16.
 10. Asuncion A, Newman D. UCI machine learning repository. 2007. <https://ergodicity.net/2013/07/>. Accessed August 1, 2017.
 11. Qiita Development Team. Qiita. <http://qiita.microbio.me/>. Accessed August 1, 2017.
 12. Al-Ghalith GA, Hillmann B, Ang K, et al. SHI7 is a self-learning pipeline for multipurpose short-read DNA quality control. *mSystems* 2018;3(3);doi:10.1128/mSystems.00202-17.
 13. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7:335–6.
 14. Al-Ghalith G, Knights D. BURST enables optimal exhaustive DNA alignment for big data. Zenodo 2017 doi:10.5281/zenodo.806850.
 15. O’Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–45.
 16. McDonald D, Price MN, Goodrich J, et al. An improved GreenGenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 2012;6:610–8.
 17. Vangay P, Hillmann B, Knights D. MLRepo Github Page. <https://knights-lab.github.io/MLRepo>.
 18. Hunter S, Corbett M, Denise H, et al. EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* 2014;42:D600–6.
 19. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res* 2011;39:D19–21.
 20. Forster SC, Browne HP, Kumar N, et al. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res* 2016;44:D604–9.
 21. Bokulich NA, Rideout JR, Mercurio WG, et al. Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 2016;1(5); doi:10.1128/mSystems.00062-16.
 22. Pasolli E, Schiffer L, Manghi P, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;14:1023–4.
 23. Duvall C, Gibbons SM, Gurry T, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017;8:1784.
 24. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5:R80.
 25. Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
 26. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.
 27. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 2005;17:299–310.
 28. Ling CX, Huang J, Zhang H, AUC: a statistically consistent and more discriminating measure than accuracy. In: *IJCAI’03 Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003*. San Francisco, CA: Morgan Kaufmann; 2003:519–24.
 29. Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;498:99–103.
 30. Vangay P, Hillmann B, Knights D. Instructions for adding new datasets. <https://github.com/knights-lab/MLRepo/blob/master/add-datasets-readme.md>.
 31. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn’s disease. *Cell Host Microbe* 2014;15:382–92.
 32. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* 2012;486:215–21.
 33. Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res* 2012;22:292–8.
 34. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014;505:559–63.
 35. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature* 2009;457:480–4.
 36. Ronacher A. Jinja2. 2017. <http://jinja.pocoo.org/>.
 37. Gruber J, Markdown. 2004. <https://daringfireball.net/projects/markdown>. Accessed August 1, 2017.
 38. Bunn A, Korpela M. Crossdating in dplR. 2013. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.0.5851&rep=rep1&type=pdf>. Accessed September 1, 2017.
 39. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;28(5):1–26.
 40. Gonzalez A, Navas-Molina JA, Kosciolk T, et al., Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* 2018;15:796–798.
 41. Vangay P, Hillmann BM, Knights D. Supporting data for “Microbiome Learning Repo (ML Repo): a public repository of microbiome regression and classification tasks.” *GigaScience Database* 2019. <http://dx.doi.org/10.5524/100581>.