Synthetic and Systems Biotechnology 7 (2022) 574–585

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Increasing prediction performance of colorectal cancer disease status using random forests classification based on metagenomic shotgun sequencing data

Yilin Gao [1], Zifan Zhu [1], Fengzhu Sun [*]

Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, RRI 201, Los Angeles, CA, United States

ARTICLE INFO

Keywords:
Microbiome
Colorectal cancer
Metagenomic shotgun sequencing
Random forests

ABSTRACT

Dysfunction of microbial communities in various human body sites has been shown to be associated with a variety of diseases raising the possibility of predicting diseases based on metagenomic samples. Although many studies have investigated this problem, there are no consensus on the optimal approaches for predicting disease status based on metagenomic samples. Using six human gut metagenomic datasets consisting of large numbers of colorectal cancer patients and healthy controls from different countries, we investigated different software packages for extracting relative abundances of known microbial genomes and for integrating mapping and assembly approaches to obtain the relative abundance profiles of both known and novel genomes. The random forests (RF) classification algorithm was then used to predict colorectal cancer status based on the microbial relative abundance profiles. Based on within data cross-validation and cross-dataset prediction, we show that the RF prediction performance using the microbial relative abundance profiles estimated by Centrifuge is generally higher than that using the microbial relative abundance profiles estimated by MetaPhlAn2 and Bracken. We also develop a novel method to integrate the relative abundance profiles of both known and novel microbial organisms to further increase the prediction performance for colorectal cancer from metagenomes.

1. Introduction

Human microbial community is an aggregate of bacteria, archaea, viruses, plasmids, eukaryotes, etc. that reside on or within human body sites. According to the Human Microbiome Project, these microorganisms are ten times more than the number of human cells, but only take up 1–3% of human body mass due to their tiny sizes [1]. Human microbiome has been demonstrated to play important roles in metabolic functions, immune system processes and other physiological activities [2]. Previous studies found strong associations between the abundance levels of some microbial organisms with various diseases, such as rheumatoid arthritis [2], diabetes [3,4], inflammatory bowel disease [5,6], and colorectal cancer [7]. These findings provide critical information towards understanding the potential roles of the human microbiome in disease developments.

Colorectal cancer (CRC) is the third most common cancer worldwide. In USA alone, close to 137,000 people are diagnosed with CRC and 50,000 people die from it annually [8]. Several studies have shown that hereditary, family medical history [9,10] of diseases like inflammatory bowel disease [11], diabetes [12], and behavior factors including alcohol consumption [13], smoking [14] and obesity [15] are associated with CRC development. In addition to these risk factors, other studies have established associations of the human gut microbiome with CRC [7,16–21]. Compositional alterations of bacteria genera like *Fusobacterium* [22] and species like *Escherichia coli* [23] and *Bacteroides fragilis* [24] were shown to be associated with the development of CRC. In addition to independent investigations, several cross-study analyses were conducted and many reproducible microbiome biomarkers were found through comparisons of various CRC datasets collected by different research groups [20,21].

; CRC, colorectoral cancer; LOSO, leave-one-sample-out; AUC, area under the operating characteristic curve; LODO, leave-one-dataset-out; LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; RF, random forests; AUPRC, area under the precision and recall curve.

Peer review under responsibility of KeAi Communications Co., Ltd.

* Corresponding author.
E-mail addresses: yilingao@usc.edu (Y. Gao), zifanzhu@usc.edu (Z. Zhu), fsun@usc.edu (F. Sun).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.synbio.2022.01.005
Received 1 November 2021; Received in revised form 14 December 2021; Accepted 19 January 2022

An important scientific task in CRC-microbiome association studies is prediction, namely predicting a sample's disease status based on its microbiome profile. A general workflow is to first separate the data set into training set and testing set, then build a machine learning predictive model based on input features using the training set, and finally evaluate the model performance on the testing set. In the context of CRC-microbiome association study, the input features are usually microbial species abundances obtained from sequence alignments against a microbial reference database and the outcome to be predicted is disease or normal. In previous investigations, several machine learning models were successfully applied to the CRC-microbiome association study, including LASSO [7,21], random forests [17,18,20], neural network [25], etc.

However, these studies were mostly focused on revealing the link between CRC and known microbial organisms having reference genomes in the database such as NCBI RefSeq [26]. The role of unknown (novel) microbial organisms is mostly neglected in these studies due to lack of reference genomes. Studies have shown that about 40–50% of metagenomic shotgun reads cannot be mapped to known genomes, and thus current reference-based analyses usually do not take these unmapped reads into consideration [27,28]. Even though this number has been greatly improved by large-scale metagenomic assembly efforts in recent years with the possibility of the mapping rate to be over 80% [29], a sizable fraction of reads still cannot be mapped to these databases. Meta-analysis of clinical gut microbiome have shown the associations of novel microbial organisms with numerous diseases indicating the potential of improving the prediction accuracy by including novel microbial organisms [30].

Zhu et al. [31] developed a metagenomic predictive pipeline, MicroPro, that used information from both the known and novel microbial organisms. After the characterization of known microbial organisms by sequence alignment to known genomes, MicroPro assembled the unmapped reads pooled from all the samples into contigs. Then, it clustered all the assembled contigs into bins so that each bin was considered as a novel organism. In this way, the bins could serve as the reference database for the novel organisms and contribute in further predictive analysis. However, the weight assignment of known and novel organisms in the classification model remains a challenge, since they may not contribute equally to the prediction for different datasets.

To deal with the challenges mentioned above, we developed a new colorectal cancer predictive pipeline that incorporates both known and novel microbial organisms by a leave-one-sample-out (LOSO) model stacking method. We demonstrated that our pipeline was able to achieve significantly higher prediction performance when compared with an existing study [20].

2. Results

2.1. Workflow of the CRC-microbiome prediction analysis

To predict the CRC disease status, we built a workflow as shown in Fig. 1. First, we ran MicroPro [31] on each of the input datasets and generated abundance tables for both known and novel microbial organisms. We then renormalized known and novel microbial species abundance tables so that the relative abundance levels of each sample summed up to 1 for both known and novel species. After the renormalization, we separated the dataset into training and testing sets, trained two random forests models on the training set using the relative abundance levels of known and novel species, respectively, and applied the trained models to the testing set to derive predictive probabilities. We then used a leave-one-sample-out (LOSO) model stacking method to incorporate both predictive probabilities from known and novel models to obtain the overall AUC score. We applied the workflow to six publicly available metagenomic CRC datasets [7,16–20] from different populations as described in Table 1.

2.2. Known microbial relative abundance profiles extracted from Centrifuge yields higher prediction performance than other metagenomic taxonomic profiling tools

Thomas et al. [20] conducted a CRC-microbiome predictive analysis on the same set of six public CRC datasets described in Table 1. In their study, they used MetaPhlAn2 [32] to generate known microbial abundances, and then trained a random forests model using these abundances to predict the CRC disease status. In MicroPro, however, Centrifuge [33] was used for known microbial abundance extraction. Both MetaPhlAn2 and Centrifuge have been commonly used in metagenomic data analyses for metagenomic taxonomic profiling. Centrifuge is an alignment-based method that counts *k*-mer frequency of raw reads and compares them with a compressed composite reference genomes. MetaPhlAn2 relies on the identification of clade-specific marker genes in raw reads to estimate the abundance of each taxa. Centrifuge uses Full-text index in Minute space (FM-index) to compress the reference genomes and removes redundancies to save the storage space. Consequently, it allows users to use NCBI RefSeq, which contains over 36.5 million sequences with a total of 109 billion base pairs [33], as the reference database. On the other hand, MetaPhlAn2 profiles species-level composition of microbial communities by using a reference genome database that contains over one million unique clade-specific marker genes identified from around 17,000 reference genomes [32]. Besides these two profiling tools, according to a study conducted by Ye et al. [34], Bracken [35], an add-on

Table 1
Six metagenomic datasets related to colorectal cancer.

Dataset	Country	No. of cases	No. of controls	Reference
Zeller	France	91	93	[7]
Yu	China	74	54	[16]
Hannigan	USA/Canada	27	28	[17]
Feng	Austria	46	63	[18]
Vogtmann	USA/Canada	52	52	[19]
Thomas	Italy	61	52	[20]

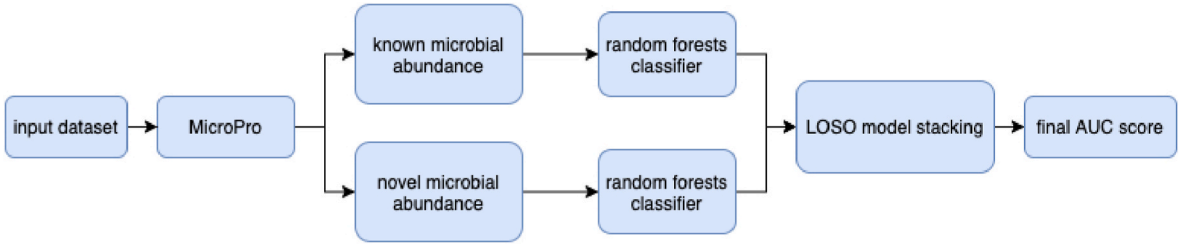

Fig. 1. Workflow of the CRC-microbiome prediction analysis. MicroPro is applied to the input metagenomic dataset to characterize known and novel abundance tables. Two random forests classifiers are trained using known and novel abundances, respectively. LOSO model stacking is used to incorporate predictive probabilities from both random forests classifiers and applied to the independent test metagenomic dataset to derive the prediction performance in terms of the AUC score. LOSO: leave-one-sample-out; AUC: area under the operating characteristic curve.

tool based on Kraken [36] with more accurate relative abundance quantification, is the top performance profiling tool compared with other tools including MetaPhlAn2 and Centrifuge, in terms of a more accurate abundance at species level as well as time complexity. Both Kraken and Centrifuge are k-mer matching algorithms, while Kraken relies on a probabilistic hash table for k-mers, Centrifuge uses an FM-index and within-species compression. Also, Kraken assigns each read to exactly one taxonomy, while Centrifuge can provide multiple taxonomic assignments per read. As an add-on tool for Kraken, Bracken further improves the classification by reassigning the unclassified reads from Kraken results based on a probabilistic estimate of the true abundance profile. Tamames et al. [37] showed that different metagenomic analysis methods usually generate different taxonomic annotation results, which can impact the follow-up predictive analysis. Therefore, we investigated the difference of Centrifuge, MetaPhlAn2 and Bracken in terms of the prediction performance of colorectal cancer based on the generated known microbial abundance profiles.

We compared the performance of these three profiling tools in

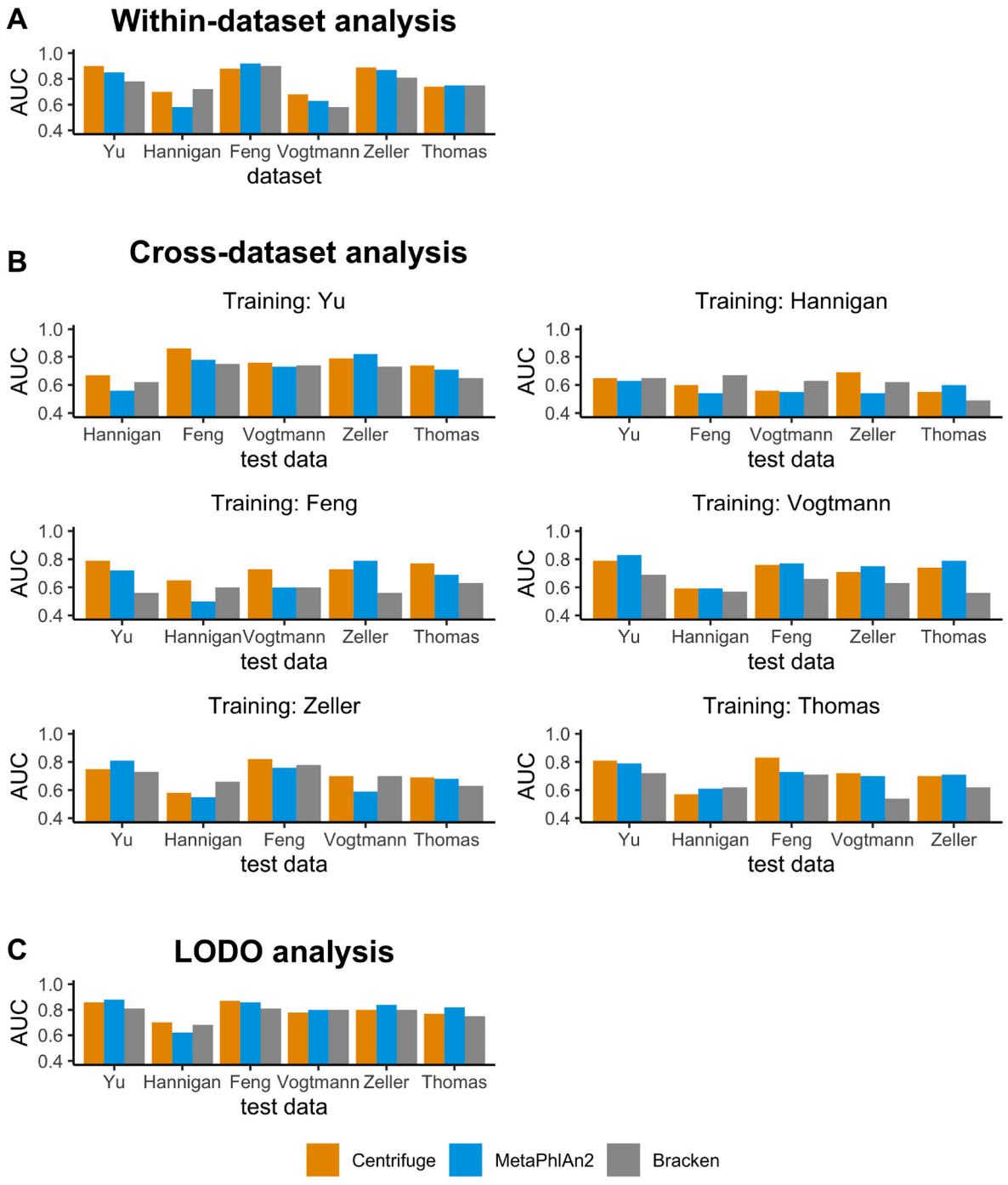

Fig. 2. Barplots of AUC scores for predicting disease status (case/control) using random forests with 1000 decision trees on known species abundance profiles characterized by Centrifuge (orange), MetaPhlAn2 (blue), and Bracken (grey), in three experimental designs: within-dataset (subfigure A), cross-dataset (subfigure B), and leave-one-dataset-out (LODO) (subfigure C). AUC scores of within-dataset and cross-dataset analyses are averages from 30 independent repetitions, while AUC scores of LODO analysis are averages from 10 independent repetitions. MetaPhlAn2 AUC scores are directly taken from study by Thomas et al. [20]. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

within-dataset, cross-dataset, and leave-one-dataset-out (LODO) analyses. The detailed steps for conducting these analyses can be found in subsection 4.4. We first used the same parameters in RF classification as in Thomas et al. [20] with 1000 decision trees. The AUC scores for the three tools in within-dataset, cross-dataset and LODO settings are shown in Fig. 2. Among all the comparisons, AUCs based on Centrifuge are higher than that based on MetaPhlAn2 for 25/42 cases, the same in 1/42 case and lower in 16/42 cases. AUCs based on Centrifuge are higher than that based on Bracken for 31/42 cases, the same in 3/42 cases and lower in 8/42 cases. Overall, Centrifuge has the best AUC scores among the three tools in the three settings.

We particularly focus on the comparison between Centrifuge and MetaPhlAn2. For within-dataset, Centrifuge yields higher AUCs than MetaPhlAn2 in 4 out of 6 cases. For cross-dataset, Centrifuge outperforms MetaPhlAn2 in 19 out of 30 cases, performs similarly in 1 case and underperforms in 10 cases. For LODO, Centrifuge outperforms MetaPhlAn2 in 2 out of 6 cases. A two-sided paired Mann-Whitney *U* test [38] on these 42 comparisons gives a p-value of 0.040. The average

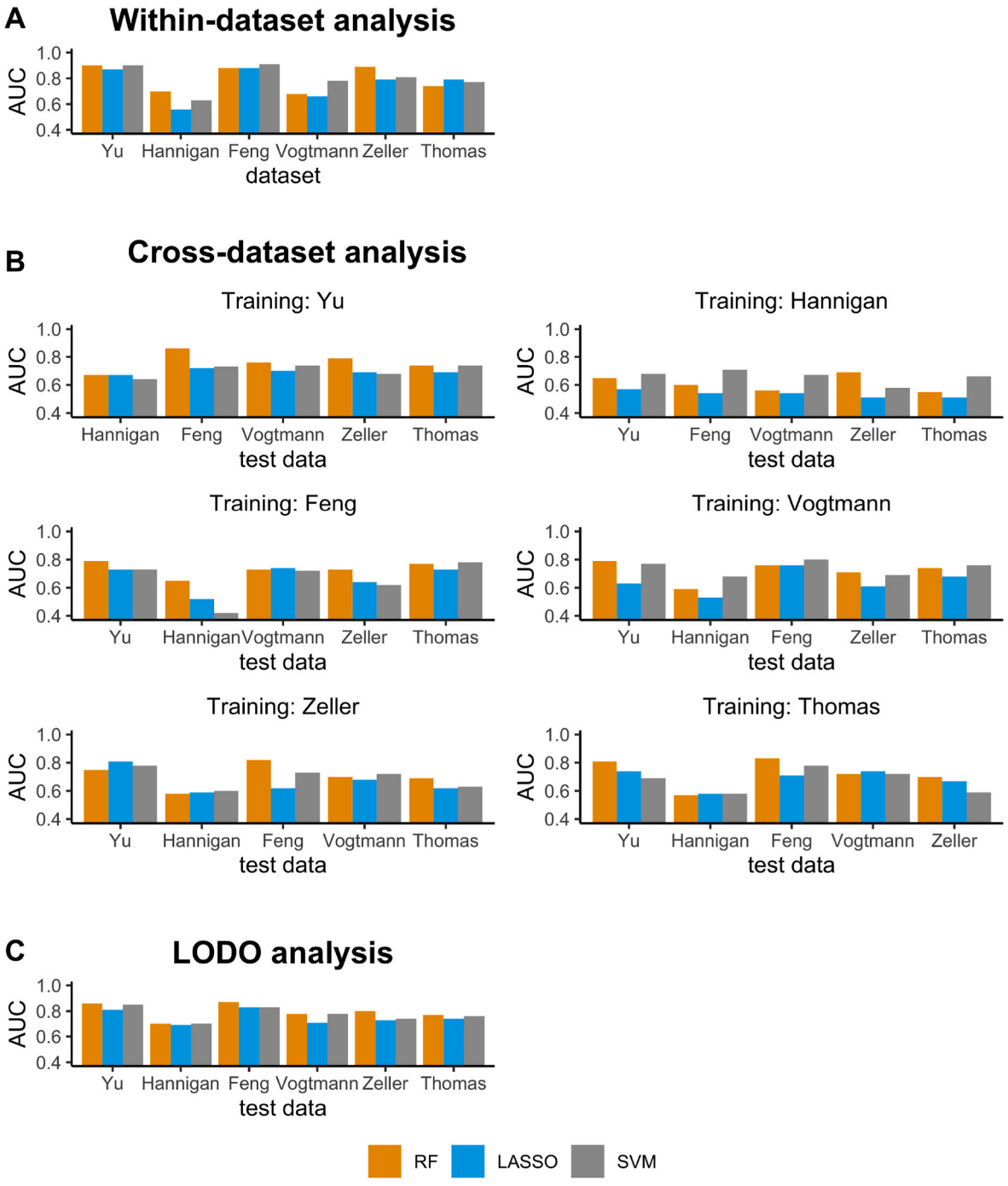

Fig. 3. Barplots of AUC scores for predicting CRC status (case/control) using known microbial species abundance profiles trained by RF (blue), LASSO (orange) and SVM (grey) in three experimental designs: within-dataset (subfigure A), cross-dataset (subfigure B), and leave-one-dataset-out (LODO) (subfigure C). AUC scores of within-dataset and cross-dataset analyses are averages from 30 independent repetitions, while AUC scores of LODO analysis are averages from 10 independent repetitions. Random forests classifiers use 1000 decision trees. The regularization coefficients in LASSO and SVM are the ones that maximize AUCs of 10-fold cross validation. The microbial species abundance profiles were log10-transformed (a small constant, 1e − 10, was added to 0 abundance) before model training. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

AUCs across the 42 comparisons using Centrifuge and MetaPhlAn2 are 0.735 and 0.710, respectively. These results demonstrate that the known microbial abundances extracted by Centrifuge has better predictive power of colorectal cancer disease status compared with that extracted by MetaPhlAn2. Similar results hold for the comparison between Centrifuge and Bracken. Therefore, we used Centrifuge in the follow-up analysis.

2.3. Random forests classifier outperforms LASSO and SVM in terms of AUC scores

In order to show the differences in AUC scores among different machine learning methods, we trained two additional machine learning classifiers: support vector machine (SVM) [39] and least absolute shrinkage and selection operator (LASSO) [40], under the three experimental settings. The details about training these two classifiers can be found in subsection 4.5. The results are shown in Fig. 3. Among all the three experimental settings, RF outperforms SVM and LASSO for 25 out of 42 cases. The average AUC scores for the 42 experiments by RF, LASSO and SVM are 0.735, 0.679 and 0.716, respectively. The two-sided paired Mann-Whitney *U* test [38] gives a p-value of 2.626e − 6 between the 42 comparisons of RF and LASSO indicating significant better performance of RF over LASSO. The corresponding p-value for the comparison of RF and SVM is 0.13 indicating that RF and SVM perform similarly. These results show that the linear model in LASSO cannot fully capture the relationship between microbial abundance and the probability of having CRC.

We further explored the trend of AUC scores of different experiments by three different classifiers. As shown in Fig. 4, both LASSO and SVM show very similar trends of AUC scores to RF among all 42 experiments (Spearman correlation of 0.79 and 0.75, respectively). This indicates the marked differences in AUC scores by training and testing on different datasets are not classifier-specific. In addition, we demonstrate that when dealing with classification tasks with high feature dimensions, RF can yield higher prediction accuracy than SVM and LASSO. Therefore, in the remaining part of the paper, we will only concentrate on the results based on RF.

2.4. Increasing the number of decision trees in a random forests model significantly improves the prediction performance

Choice of the number of decision trees in a random forests model affects its prediction performance. When using a small number of decision trees for a dataset with a huge number of features, the random forests model usually has relatively poor prediction performance. On the other hand, increasing the number of decision trees can potentially improve the prediction performance, but would need more computational resources. Friedman et al. [41] cautioned the potential risk of overfitting when increasing the number of decision trees in a random forests model. On the other hand, Oshiro et al. [42] showed that the mean and median AUCs tend to converge when increasing the number of decision trees in a random forests model. In this section, we investigated whether increasing the number of decision trees could boost the prediction performance.

We used the cross-dataset random forests setting for the experiment. We took the Hannigan dataset as the training data, while the other five datasets as the testing data in turn. We investigated the prediction performance for the number of 500, and 1000 to 10,000 trees by step of 1000 while keeping all other parameters unchanged. Each experiment was repeated 30 times and the mean AUCs were reported in Fig. 5. The figure shows that the mean AUCs have an increasing trend for all datasets when the number of decision trees is increased from 500 to 3000, and then stabilize after 3000. Considering the computational cost of a large number of decision trees, choosing the number of decision trees between 3000 and 6000 would be reasonable. In our following analysis we used 5000 decision tress in all random forests models.

We compared the prediction performance using random forests model with 5000 decision trees trained on abundance profiles extracted from Centrifuge in our analysis with that in Thomas et al. [20] where they trained random forests with 1000 decision trees on abundance profiles extracted from MetaPhlAn2. Both models used only known microbial abundances. Our model outperforms that of Thomas et al.

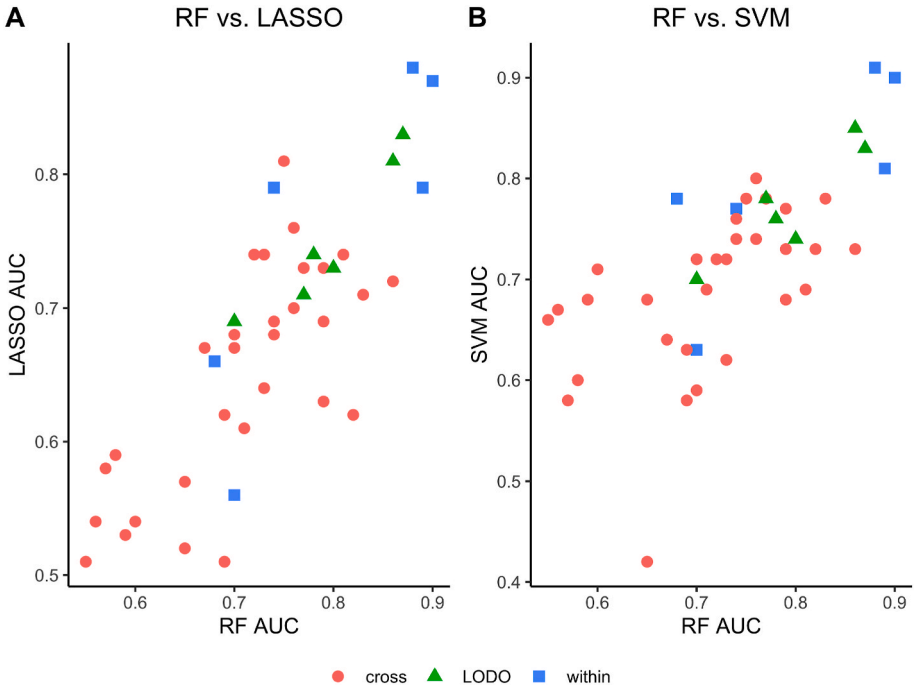

Fig. 4. Scatter plots of AUC scores by three different classifiers: RF, LASSO and SVM in three experimental settings. X-axis shows the RF AUC scores, and y-axis shows AUC scores obtained by LASSO (subfigure A) and SVM (subfigure B). The microbial species abundance profiles were log10-transformed (a small constant, 1e − 10, was added to 0 abundance) before model training. Spearman correlation for RF vs. LASSO (subfigure A) is 0.79, and for RF vs. SVM (subfigure B) is 0.75.

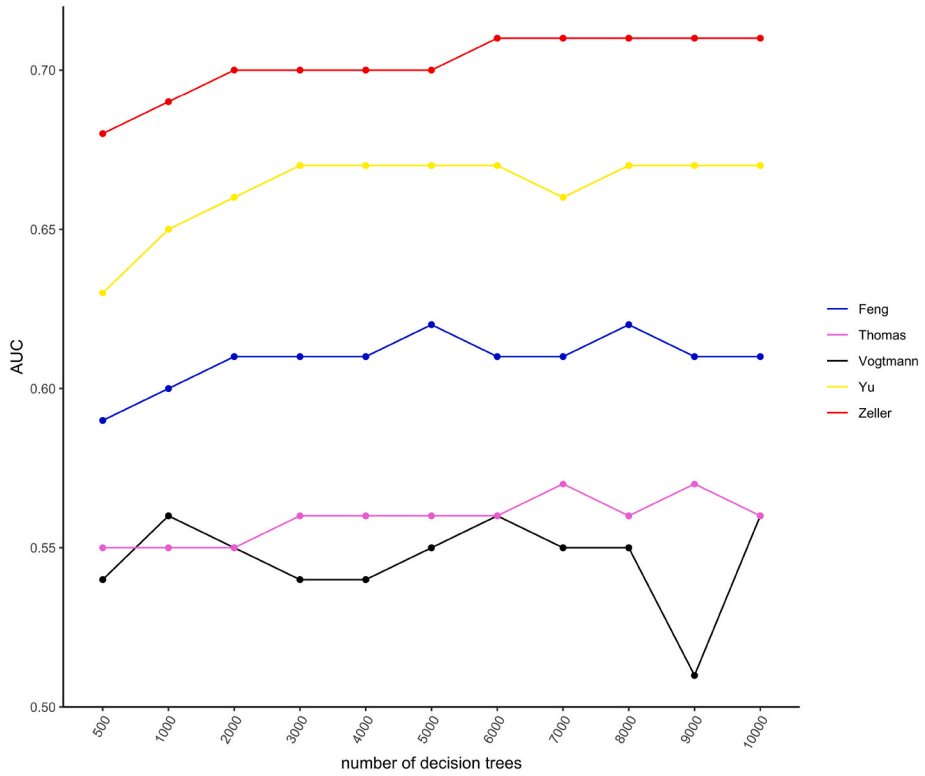

Fig. 5. The AUC scores generally increase with the number of decision trees in RF in cross-dataset analysis. The Hannigan dataset is used as the training set. Each line refers to the testing set used in the analysis. All AUC scores are averages from 30 independent repetitions.

[20] in 5 out of 6 cases in within-dataset analysis (Fig. 6 A), 20 out of 30 cases in cross-dataset analysis (Fig. 6 B), and 2 out of 5 cases in LODO analysis (Fig. 6C). A paired Mann-Whitney *U* test of two models' AUC results gives a p-value of 0.004. We also compared the results of using 1000 trees with 5000 trees based on the microbial abundances calculated using Centrifuge. While the mean AUC of 5000-tree model improves slightly over 1000 trees (0.741 vs. 0.735), the Mann-Whitney *U* test p-value of 0.000 8 showed these results are significantly different. This indicates that increasing the number of decision trees from 1000 to 5000 significantly improves the prediction performance.

2.5. Including novel microbial organisms slightly improves the performance of CRC prediction

We next investigated whether including novel microbial organisms in the model could further improve the prediction performance. We compared random forests models with and without abundance information of the novel microbial organisms in within-dataset and cross-dataset settings. For all the analyses, novel microbial abundances was incorporated using the leave-one-set-out (LOSO) model stacking method described in subsection 4.6.

In the setting of within-dataset (Fig. 6 A), the AUCs are increased for 3/6 datasets when incorporating novel microbial organisms by the LOSO model stacking, while the AUCs of the other 3 datasets do not change. For the cross-dataset setting (Fig. 6 B), the AUCs of 15/30 are increased, 4/30 are the same, and 11/30 are decreased when novel microbial organisms are incorporated using LOSO. For both settings, LOSO AUCs are significantly higher than those reported in Thomas et al. [20] with a Mann-Whitney *U* test p-value of 0.001 5. Compared to the p-value of 0.002 in the previous subsection, incorporating the information from novel microbial organisms decreases the p-value slightly in CRC prediction. The average AUCs across the 36 comparisons integrating both known and novel microbial organisms, with only the known microbial organisms and that reported in Thomas et al. [20] are 0.735, 0.731, and 0.695, respectively.

We also measured the performance of the predictive models using the area under the precision and recall curves (AUPRC). The precision and recall curve is able to capture more information then AUC when the input data used in the predictive model is highly imbalanced, especially when disease prevalence is very low among the study population. In our study, as shown in Table 1, most of the CRC datasets we used were balanced. However, we still investigated the AUPRC results to see if it was consistent with the previous AUC results. Similar to the comparison of AUC results, we compared the AUPRC results using only known microbial species with LOSO model stacking results in within-dataset and cross-dataset settings, and the predictive probabilities were taken from random forests models with 5000 decision trees. As shown in Fig. 7, the LOSO model stacking AUPRC scores were higher than known species AUPRC scores in 14/36 cases, the same in 8 cases and lower in 14 cases. We did not see significant differences in their performance. The AUPRC results didn't show as much improvements as the AUC results, probably because the datasets we used were well-balanced enough.

2.6. Removing cross-dataset batch effect by ComBat before training did not improve the performance

In cross-dataset analysis, heterogeneity in different datasets is a big challenge for machine learning prediction analyses. More specifically, the batch effects of different datasets may affect the predictions and result in low AUC scores. We investigated if normalization across the datasets can improve prediction performance in cross-dataset analyses. Therefore, we first removed the cross-dataset batch effects from the abundance profiles by the ComBat [43] function in R's 'sva' package [44]. We then trained RF models to see if the AUC scores can be improved and the results are shown in Fig. 8. It can be seen that AUC scores were only improved in 6/30 cases after removing batch effect using ComBat. The average scores for with and without ComBat normalization were 0.686 and 0.710, respectively, with a p-value of

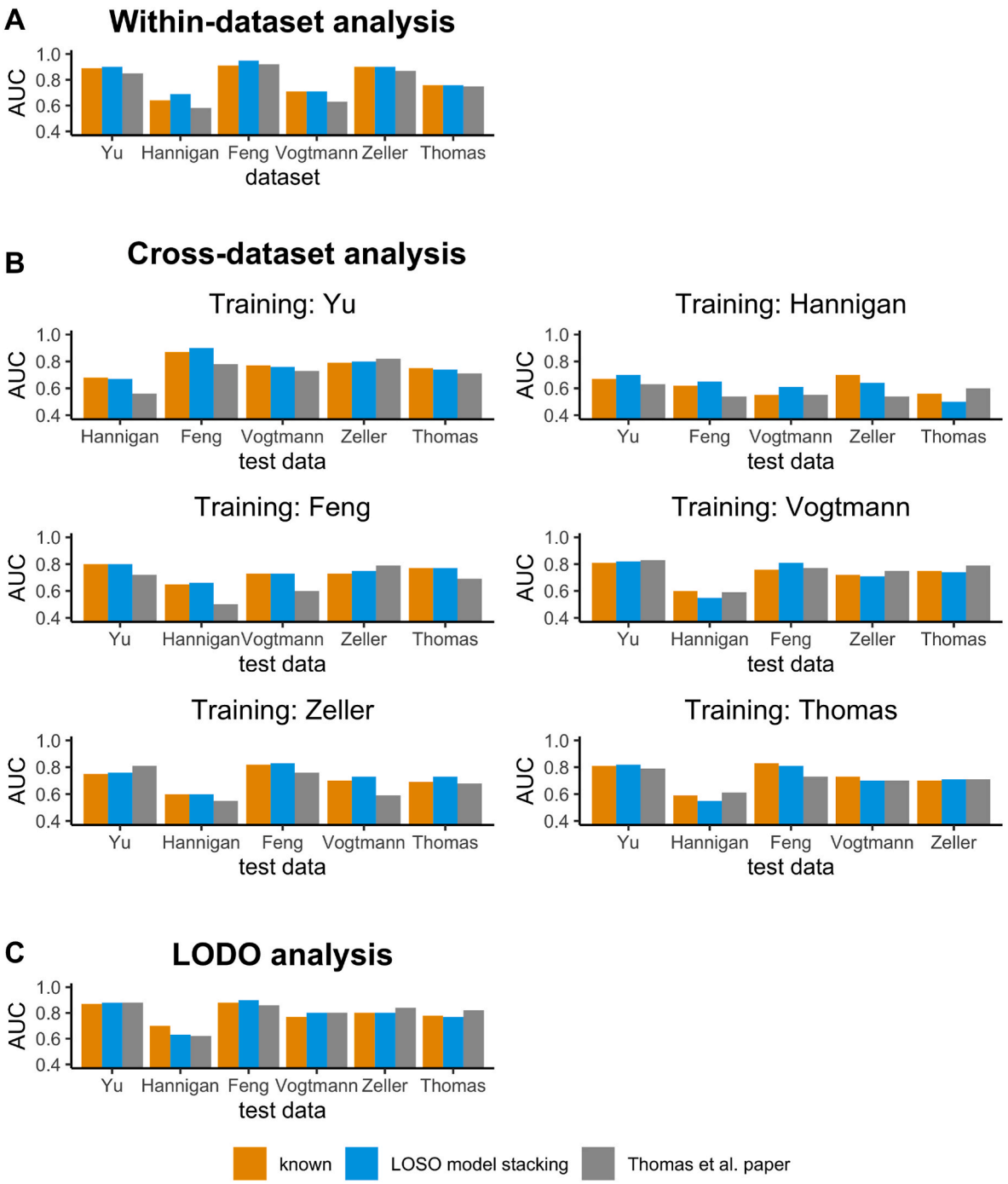

Fig. 6. Barplots of random forests AUC scores using known microbial abundances (orange), LOSO model stacking (blue), and results from Thomas et al. [20] in three experimental designs: within-dataset (subfigure A), cross-dataset (subfigure B) and leave-one-dataset-out (LODO) (subfigure C). AUC scores of within-dataset and cross-dataset analyses are averages from 30 independent repetitions, while AUC scores of LODO analysis are averages from 10 independent repetitions. Random forests models in 'known' and 'LOSO model stacking' use 5000 decision trees while those in 'Thomas et al. paper' use 1000 decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

0.003 by the two-sided paired Mann-Whitney *U* test. These results demonstrate that ComBat is not helpful in increasing the AUC score for our setting. In the future we may consider other methods to deal with heterogeneity, or investigate other underlying reasons that may cause low AUC scores when dealing with different datasets.

3. Discussion

In this study, we performed comprehensive CRC-microbiome predictive analyses on six publicly available metagenomic datasets. We showed that the microbial relative abundance profiles extracted from Centrifuge gives higher prediction AUCs compared to other metagenomic taxonomic profiling tools based on random forests classification results. We also showed that increasing the number of decision trees to 5000 from 1000 as used in Thomas et al. [20] and including novel microbial abundances can further slightly improve the prediction performance compared with the results from Thomas et al. [20].

The random forests algorithm is a widely used machine learning

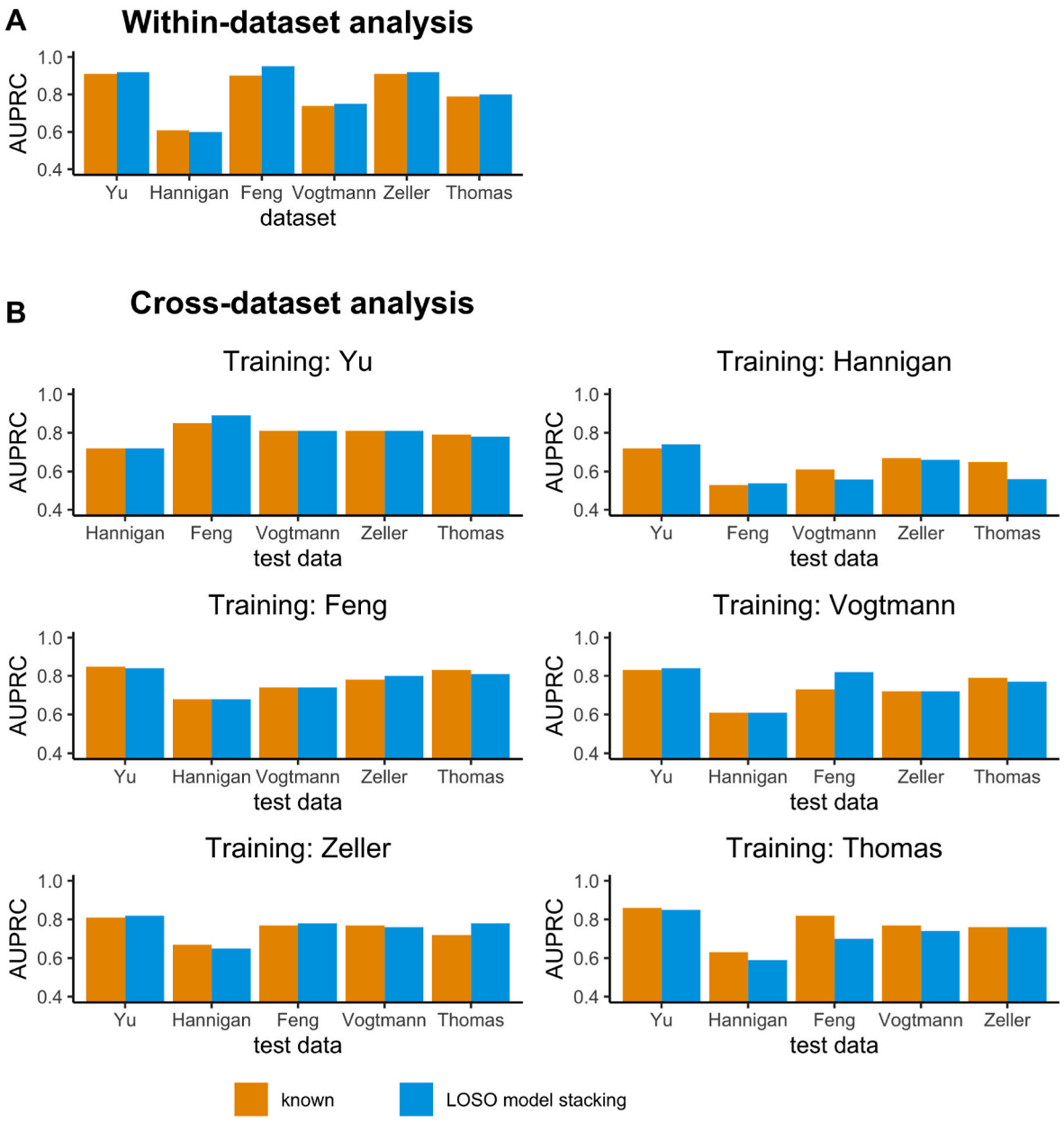

Fig. 7. Barplots of random forests AUPRC scores using known microbial abundances (orange) and LOSO model stacking (blue) in two experimental designs: within-dataset (subfigure A) and cross-dataset (subfigure B). All AUPRC scores are averages from 30 independent repetitions. Random forests models in 'known' and 'LOSO model stacking' use 5000 decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

algorithm for classification tasks in metagenomic analyses. It is based on ensemble learning and bagging algorithms, and can reduce overfitting compared to the decision tree algorithm. In our study, we showed that instead of sticking to a relative small number of decision trees, increasing the number of decision trees in case of a large-scale input data benefited the prediction. In addition, We also trained two other machine learning classifiers, LASSO and SVM, to compare the prediction performance in the three experimental settings. Random forests outperforms LASSO and performs similarly as SVM in all three settings (Fig. 3). LASSO and SVM present similar trends as random forests in terms of AUC scores (Fig. 4). This further indicates the differences in AUC scores by training and testing on different datasets are not classifier-specific.

Our study also showed that different metagenomic taxonomic profiling tools demonstrated different predictive abilities of the disease status. We compared the AUC results generated by Centrifuge [33] to the results by MetaPhlAn2 [32] used in Thomas et al. [20], as well as Bracken [35], and found that Centrifuge has better metagenomic taxonomic profiling ability than other tools in terms of the CRC disease status prediction. Centrifuge computes abundances by sequence alignments against all the reference genomes including those with low abundance. This could be a possible reason for Centrifuge to outperform MetaPhlAn2 in terms of disease status prediction.

It is not surprising that the cross-dataset analyses have much lower prediction accuracies than within-dataset analyses, considering that the heterogeneity always gives variability of outcomes when dealing with cross-study analysis. In particular, the Hannigan dataset always has a poor prediction performance whether it is treated as a training or a testing dataset. Thomas et al. [20] showed that the Hannigan dataset differs markedly from other datasets when performing principal coordinate analysis with the Bray–Curtis distance using MetaPhlAn2 microbial species abundances. The heterogeneity between different

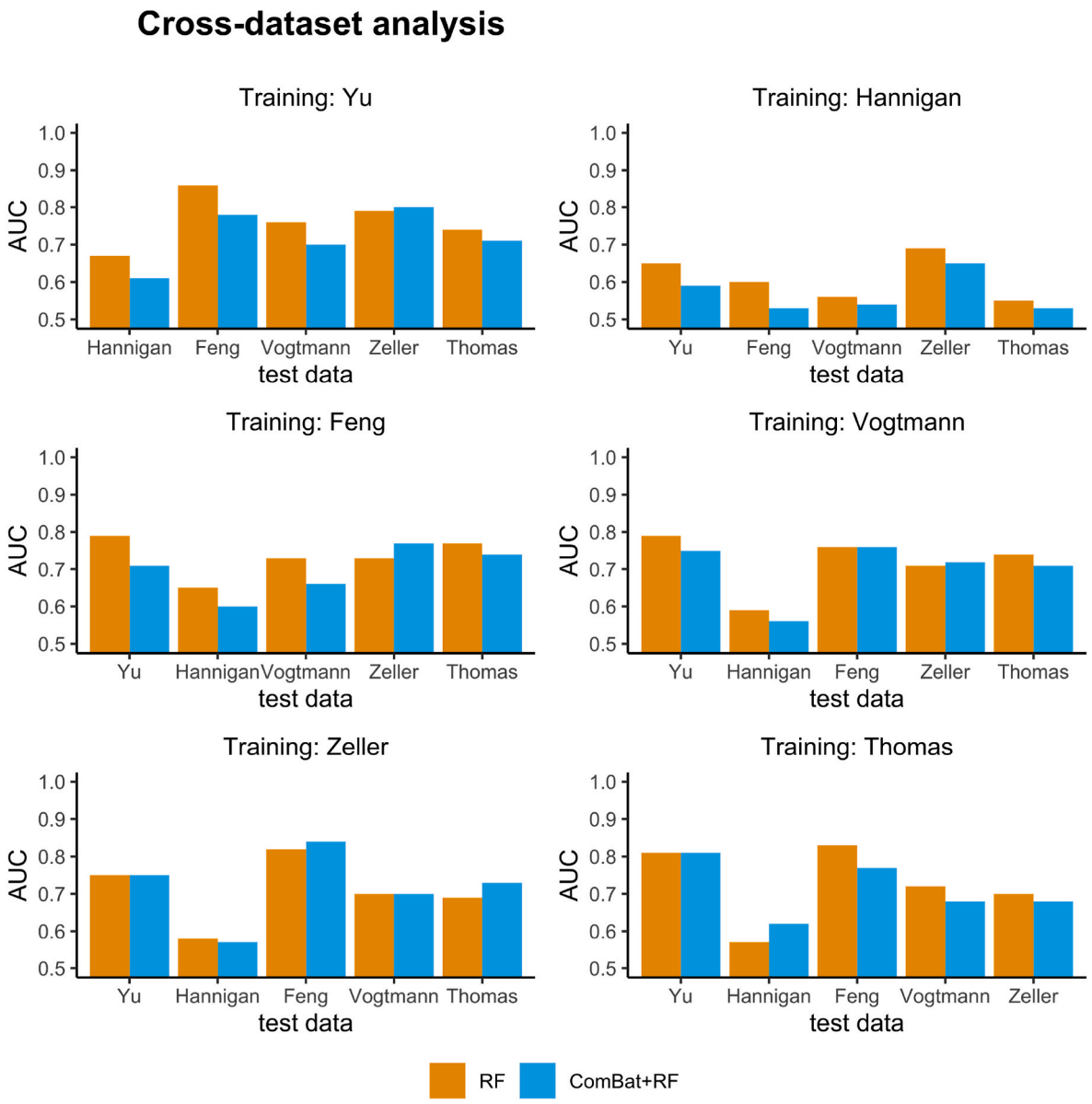

Fig. 8. Barplots of cross-dataset AUC scores for predicting disease status (case/control) using known species abundance profiles trained by random forests directly (blue) and removing batch effect by ComBat before training random forests models (orange). AUC scores are averages from 30 independent repetitions. Random forests classifiers use 1000 decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

datasets results in non-biological experimental variations, which is commonly known as 'batch effect', and this can lead to difficulty in cross-dataset predictions. Many methods have been developed for data normalization and batch effect reduction. For example, the ComBat function in R's 'sva' package is an algorithm to remove batch effects using either parametric or non-parametric empirical Bayes frameworks [43]. In our study, we also explored the potential of removing batch effects before training machine learning models. However, the cross-dataset AUC results did not improve (Fig. 8). How to normalize the metagenomic data across different datasets to improve prediction accuracy is a topic for future studies.

We also explored different ways to combine known and novel microbial abundances in constructing disease predictive models. A straightforward method is to pool the known and novel features together and train a single random forests model using the combined abundance table. This method is often referred to as feature stacking. We compared the LOSO model stacking with the feature stacking in terms of AUC scores and the results are shown in Supplementary Fig. S1. For within-dataset results (Figure S1 A), 5 out of 6 datasets have better mean AUC scores when using the LOSO model stacking method. For cross-dataset results (Figure S1 B), the LOSO model stacking method is even more outstanding; it outperforms feature stacking in 24 out of 30 cases. For leave-one-dataset-out LOSO model (Figure S3 A), the AUC for 3 out of 6 datasets increased when incorporating the novel microbial abundances, while the AUC for one dataset did not change. Consistent with the AUC results, the AUPRC results showed similar trend: LOSO model stacking outperforms feature stacking in 6/6 cases in within-dataset settings (Figure S2 A), 22/30 cases in cross-dataset settings (Figure S2 B), and 3/6 cases in LODO settings (Figure S3 B). All these results indicate that the LOSO model stacking method outperforms the feature stacking in terms of the prediction performance. Also, the LOSO model stacking method is computationally more efficient than the feature stacking method since it does not need to train a random forests model using the combined abundance profiles with much larger feature dimensions. For future analyses, researchers can apply the LOSO model stacking method to other machine learning prediction tasks.

NCBI RefSeq database is one of the most widely used databases for microbial research. However, it is vastly incomplete and usually only about half of reads in human gut can be mapped to the NCBI RefSeq database leaving a large fraction of the reads unused in most studies. To overcome this issue, several groups constructed more complete human gut microbial genome databases using cross assembly and binning, for example, the Unified Human Gastrointestinal Genome (UHGG) collection [29]. The fraction of reads that can be mapped to UHGG from the human gut was markedly increased to over 80% [29]. However, it is not known whether the microbial abundance profiles using UHGG can increase the prediction accuracy of complex diseases such as CRC. In order to answer this question, we mapped the reads to UHGG, derived the microbial abundance profiles for each sample, and predicted the disease status using RF. The results are provided in supplementary material. The mapping rates for the datasets were markedly increased from around around 50% to close to 90% as shown in Fig. S4, consistent with previous studies [29]. Despite the marked increase in mapping rates, the prediction accuracy for CRC measured by AUC based on UHGG did not improve as shown in Fig. S5. One potential explanation is that the microbial organisms associated with CRC belong to or are strongly associated with the NCBI known microbial species. Therefore, the inclusion of more microbial genomes in UHGG compared to the NCBI RefSeq database did not increase the prediction accuracy of CRC. This result is consistent with that in subsection 2.5 that including novel microbial organisms did not markedly increase the prediction accuracy. The usefulness of UHGG for the prediction of other disease status such as inflammatory diseases, diabetes, obesity, etc. compared to NCBI RefSeq needs to be further studied.

4. Materials and methods

4.1. Colorectal cancer metagenomic datasets

We analyzed six publicly available and geographically diverse colorectal cancer metagenomic datasets with download links from their original papers [7,16–20]. We excluded the samples from patients with adenoma so that only samples from patients diagnosed with CRC and healthy controls were used. The numbers of cases and controls for each dataset are shown in Table 1. In total, there are 351 CRC samples and 342 healthy controls. The percentage of CRC samples in each dataset is close to 0.5 according to Table 1, which indicates the datasets used in our study are balanced.

4.2. Generating known and novel microbial relative abundance profiles by MicroPro [31]

We used MicroPro [31] to generate known and novel microbial abundance profiles for the six CRC datasets. There are three main steps in the MicroPro pipeline: (1) characterization of the known microbial species abundance through sequence alignments against reference genomes, (2) extraction of novel microbial abundances by an assembly-binning-based algorithm, and (3) machine learning predictive analysis.

In our study, we need to generate microbial abundances for both within-dataset and cross-dataset analysis. For the within-dataset analysis, we ran the MicroPro pipeline directly and generated both known and novel abundance tables. For cross-datasets, the known microbial abundances were generated in the same way as the within-dataset analysis. To derive abundances of novel organisms for the testing dataset in the cross-dataset setting, we treated the novel metagenomic bins of the training dataset as the reference database, mapped the unmapped reads of the testing dataset back to this reference, and calculated the novel abundances based on the mapping results. We then renormalized known and novel species abundance table so that the relative abundance levels of each sample summed up to 1 for both known and novel species.

4.3. Generating known microbial relative abundance profiles by other metagenomic taxonomic profiling tools

In this study, we compared the performance of three different metagenomic taxonomic profiling tools in terms of prediction performance of colorectal cancer based on the generated known microbial profiles. The known microbial abundance profiles generated by MicroPro described in subsection 4.2 were considered as known abundance generated by Centrifuge [33] since MicroPro directly uses Centrifuge for characterizing known abundance profiles. MetaPhlAn2 results were directly taken from Thomas et al. [20]. For generating known abundance profiles from Bracken [35], we ran Bracken pipeline on all FASTQ sequence files for the six datasets. The number of microbial species identified by Bracken were two to three times more than the number generated by MicroPro. So in order to decrease the running time, we filtered out microbial species that had nonzero abundances in less than 10% of all samples for each dataset.

4.4. Random forests CRC predictive models

In our study, we built random forests classification models to test the predictive power of the generated known and novel microbial abundances on the CRC disease status. All the random forests analyses were carried out with the ‘caret’ package in R [45]. We used two parameter settings for different purposes of analysis. For comparison with the study conducted by Thomas et al. [20] as well as comparing the performance of three different metagenomic taxonomic profiling tools, we used 1000 decision trees and the ‘mtry’ parameter was tuned by a 10 fold cross-validation. We then performed the predictive analysis three settings: within-dataset, cross-dataset and leave-one-dataset-out (LODO). For the rest of the analysis, we used 5000 decision trees and again the ‘mtry’ parameter was tuned by a 10 fold cross-validation, then performed the predictive analysis for two settings: within-dataset and cross-dataset. We didn’t conduct analysis for LODO settings with 5000 decision trees due to the extremely long computing time.

For the within-dataset setting, we randomly split the samples in each dataset into 80% training set and 20% testing set. Then we trained a random forests model on the training data and applied it to the testing data to derive the predictive probabilities for each testing sample. The process was repeated for 30 times to obtain the average AUCs (area under the receiver operating characteristic curve).

For the setting of cross-dataset, one of the six datasets was treated as the training set and another in the remaining five was treated as the testing set in turn. A random forests model was trained on the training set and then applied to the testing set for prediction. When applying the trained model to the testing set, if there was a missing feature in the testing matrix, we added this feature with 0 abundance in all samples, so that all features in the training data were presented in the testing data as well.

For LODO setting, one of the six datasets was selected as the testing set while the other five datasets were treated as the training set together. All the microbial species in the five training sets were used as features in the random forests model, and we added a 0 abundance column to the testing set if any feature in the training sets is missing in the testing set.

In terms of model evaluation, for random forests model using only known abundances, AUC was used as the performance measurement. An AUC score of 1 indicated a perfect prediction and 0.5 represented a random guess. For the random forests model using both known and novel microbial abundances, we first used a LOSO model stacking method to obtain weighted predictive probabilities and then derived the final AUC scores. The details of the LOSO model stacking method are described in subsection 4.6.

4.5. LASSO and SVM CRC predictive models

Microbial species abundance profiles were first log10-transformed

and a small constant (1e − 10) was added to 0 abundance to avoid the indefinite value of $\log_{10}(0)$ before training LASSO and SVM. In terms of hyperparameters, we chose the regularization parameter from the range of 0–0.5 with step 0.001 that maximized AUC by 10-fold cross validation in the training set in LASSO. For SVM, we used the Gaussian radial basis kernel and chose the regularization coefficient C that maximized AUC by 10-fold cross validation from default values of 0.25, 0.5 and 1. We used the 'caret' package [45] in R to implement both LASSO and SVM algorithms. The training process was repeated for 30 times for within-dataset and cross-dataset settings, and 10 times for LODO setting. The AUC scores reported were the average for the 30 or 10 repetitions.

4.6. Leave-one-sample-out (LOSO) model stacking method

In machine learning, ensemble methods that integrate predictions from multiple models are commonly used to boost the prediction performance. According to previous studies [46,47], ensemble methods are often more accurate than the component methods that the ensemble methods contain. Among various ensemble methods, model stacking is an efficient method that uses the predictions generated by other machine learning algorithms as the inputs of a second layer algorithm, which then combines the input predictions to form a new set of predictions. If we choose weighted linear combinations of predictors as the second layer algorithm, then we only need to determine the weights of each input predictor and combine them to form an optimal predictor. Consider two studies $S1$ and $S2$. The disease statuses of all individuals in study $S1$ and part of the individuals in study $S2$ are known. We have several predictors, $p_1, p_2, \ldots, p_K$, based on study $S1$ and we are interested in linear combinations of these predictors to maximize the AUC for the integrated predictor in study $S2$. We can use the information of the individuals with known disease statuses in study $S2$ to decide the weight for each predictor. Since we want to maximize the AUC of the integrated predictor, we should give higher weights to predictors with high AUCs and lower weights to predictors with low AUCs according to the prediction results for individuals with known disease statuses. Therefore, if the AUC for the k-th predictor is AUC_k, the weight for the k-th predictor is proportional to $\max(AUC_k - 0.5, 0)$ since AUC for random guess is 0.5.

Based on this idea, we developed a leave-one-sample-out (LOSO) model stacking method. The high level idea of LOSO model stacking is to weight predictive probabilities based on known and novel microbial abundances by their respective prediction accuracy without bringing any inference by the testing data.

For each sample i in the testing set, we obtained two predictive probabilities p_i^{known} and p_i^{novel} by applying the trained random forests model using only known or novel microbial abundances to the testing set, respectively. We then excluded i from the testing set and computed AUC scores AUC_i^{known} and AUC_i^{novel} by comparing the remaining predictive probabilities to their true disease statuses. The LOSO model stacking method weighted p_i^{known} and p_i^{novel} by AUC_i^{known} and AUC_i^{novel} subtracting the background AUC score of 0.5. The detailed formula of LOSO model stacking is shown in equation (1).

$$p_i = \frac{\max(AUC_i^{known} - 0.5, 0) \times p_i^{known} + \max(AUC_i^{novel} - 0.5, 0) \times p_i^{novel}}{\max(AUC_i^{known} - 0.5, 0) + \max(AUC_i^{novel} - 0.5, 0)} \quad (1)$$

4.7. Prediction performance measured by the area under the precision-recall curve (AUPRC)

When comparing the prediction performance of using only known species incorporating both known and novel species, we used an additional measure AUPRC (area under the precision-recall curve). To calculate AUPRC scores by the R package PRROC [48], we took the predictive probabilities from random forests classifier with 5000 decision trees using known microbial species and novel microbial species, respectively. In order to calculate the LOSO model stacking results, we first followed the same procedure as described in subsection 4.6 to generate the LOSO predictive probabilities using equation (1), and then used these probabilities to calculate AUPRC scores.

5. Conclusions

- We showed that microbial relative abundance profiles extracted from Centrifuge yielded high prediction performance of colorectal cancer status than that extracted from other tools.
- We showed that random forests model had the best classification performance in CRC disease status prediction, compared with LASSO and SVM models.
- We also found that increasing the number of decision trees in the random forests model significantly improved the classification performance for a large-scale input data matrix like the microbial abundance table.
- We developed a novel CRC predictive pipeline that incorporated both known and novel microbial organisms by a LOSO model stacking method.
- We applied our pipeline to six public colorectal cancer datasets and found it improved the prediction performance compared with an existing study.

Data availability

All the datasets used in this study are publicly available in the European Nucleotide Archive (ENA) database (https://www.ebi.ac.uk/ena). Accession number for Yu is PRJEB10878 [16], for Hannigan is PRJNA389927 [17], for Feng is ERP008729 [18], for Vogtmann is PRJEB12449 [19], for Zeller is ERP005534 [7], and for Thomas is SRP136711 [20].

All the codes used in analysis can be found at https://github.com/lynngao/CRC_analysis.

Author contributions statement

Yilin Gao: Methodology, Investigation, Formal analysis, Data Curation, Visualization, Software, Writing - Original Draft

Zifan Zhu: Methodology, Investigation, Data Curation, Writing - Review & Editing

Fengzhu Sun: Conceptualization, Methodology, Writing - Review & Editing, Supervision

Funding

This research was partially supported by US National Institutes of Health (NIH) [R01GM120624, 1R01GM131407]. Z.Z. was also supported by an Andrew J. Viterbi Fellowship.

Declaration of competing interest

There is NO Competing Interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2022.01.005.

References

[1] Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;486(7402):215–21. https://doi.org/10.1038/nature11209.
[2] Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol 2015;31(1):69–75. https://doi.org/10.1097/MOG.0000000000000139.

[3] Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013;498(7452):99–103. https://doi.org/10.1038/nature12198.

[4] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55–60. https://doi.org/10.1038/nature11450.

[5] Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2014;15(3):382–92. https://doi.org/10.1016/j.chom.2014.02.005.

[6] Haberman Y, Tickle TL, Dexheimer PJ, et al. Corrigendum. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest 2015;125(3):1363. https://doi.org/10.1172/JCI79657.

[7] Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10:766. https://doi.org/10.15252/msb.20145645.

[8] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA A Cancer J. Clin 2019;69(1):7–34. https://doi.org/10.3322/caac.21551.

[9] Butterworth AS, Higgins JP, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. Eur J Cancer 2006;42(2):216–27. https://doi.org/10.1016/j.ejca.2005.09.023.

[10] Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001;96(10):2992–3003. https://doi.org/10.1111/j.1572-0241.2001.04677.x.

[11] Lutgens MW, van Oijen MG, van der Heijden GJ, et al. Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies. Inflamm Bowel Dis 2013;19(4):789–99. https://doi.org/10.1097/MIB.0b013e31828029c0.

[12] Tsilidis KK, Kasimis JC, Lopez DS, et al. Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. BMJ 2015;350:g7607. https://doi.org/10.1136/bmj.g7607.

[13] Bagnardi V, Rota M, Botteri E, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. Br J Cancer 2015;112(3):580–93. https://doi.org/10.1038/bjc.2014.579.

[14] Botteri E, Iodice S, Bagnardi V, et al. Smoking and colorectal cancer: a meta-analysis. JAMA 2008;300(23):2765–78. https://doi.org/10.1001/jama.2008.839.

[15] Ma Y, Yang Y, Wang F, et al. Obesity and risk of colorectal cancer: a systematic review of prospective studies. PLoS One 2013;8(1):e53916. https://doi.org/10.1371/journal.pone.0053916.

[16] Yu J, Feng Q, Wong SH, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66(1):70–8. https://doi.org/10.1136/gutjnl-2015-309800.

[17] Hannigan GD, Duhaime MB, Ruffin MT, et al. Diagnostic potential and interactive dynamics of the colorectal cancer virome. mBio 2018;9(6). https://doi.org/10.1128/mBio.02248-18. e02248–18.

[18] Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 2015;6:6528. https://doi.org/10.1038/ncomms7528.

[19] Vogtmann E, Hua X, Zeller G, et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. PLoS One 2016;11:1–13. https://doi.org/10.1371/journal.pone.0155362.

[20] Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25(4):667–78. https://doi.org/10.1038/s41591-019-0405-7.

[21] Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25(4):679–89. https://doi.org/10.1038/s41591-019-0406-6.

[22] Zhou Z, Chen J, Yao H, et al. Fusobacterium and colorectal cancer. Front Oncol 2018;8:371. https://doi.org/10.3389/fonc.2018.00371.

[23] Cougnoux A, Dalmasso G, Martinez R, et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. Gut 2014;63(12):1932–42. https://doi.org/10.1136/gutjnl-2013-305257.

[24] Haghi F, Goli E, Mirzaei B, et al. The association between fecal enterotoxigenic B. fragilis with colorectal cancer. BMC Cancer 2019;19(1):879. https://doi.org/10.1186/s12885-019-6115-1.

[25] Reiman D, Metwally A, Dai Y. Using convolutional neural networks to explore the microbiome, annual international conference of the IEEE engineering in medicine and biology society. In: IEEE engineering in medicine and biology society. Annual international conference 2017; 2017. p. 4269–72. https://doi.org/10.1109/EMBC.2017.8037799.

[26] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35(Database issue):D61–5. https://doi.org/10.1093/nar/gkl842.

[27] Nayfach S, Rodriguez-Mueller B, Garud N, et al. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res 2016;26(11):1612–25. https://doi.org/10.1101/gr.201863.115.

[28] Sunagawa S, Mende DR, Zeller G, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 2013;10(12):1196–9. https://doi.org/10.1038/nmeth.2693.

[29] Almeida A, Nayfach S, Boland M, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2021;39:105–14. https://doi.org/10.1038/s41587-020-0603-3.

[30] Nayfach S, Shi ZJ, Seshadri R, et al. New insights from uncultivated genomes of the global human gut microbiome. Nature 2019;568(7753):505–10. https://doi.org/10.1038/s41586-019-1058-x.

[31] Zhu Z, Ren J, Michail S, et al. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biol 2019;20(1):154. https://doi.org/10.1186/s13059-019-1773-5.

[32] Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9(8):811–4. https://doi.org/10.1038/nmeth.2066.

[33] Kim D, Song L, Breitwieser FP, et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016;26(12):1721–9. https://doi.org/10.1101/gr.210641.116.

[34] Ye S, Siddle K, Park D, et al. Benchmarking metagenomics tools for taxonomic classification. Cell 2019;178:779–94. https://doi.org/10.1016/j.cell.2019.07.010.

[35] Lu J, Breitwieser F, Thielen P, et al. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science 2017;3:e104. https://doi.org/10.7717/peerj-cs.104.

[36] Wood D, Salzberg S. Wood de, salzberg sl.. kraken: ultrafast metagenomic sequence classification using exact alignment. Genome Biol 2014;15:R46. https://doi.org/10.1186/gb-2014-15-3-r46. genome biol 15: R46.

[37] Tamames J, Cobo-Simón M, Puente-Sánchez F. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. BMC Genom 2019;20(1):960. https://doi.org/10.1186/s12864-019-6289-6.

[38] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull 1945;1(6):80–3. https://doi.org/10.2307/3001968.

[39] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.

[40] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[41] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.

[42] Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest?, machine learning and data mining in pattern recognition. MLDM 2012. Lect Notes Comput Sci 2012;7376:154–68. https://doi.org/10.1007/978-3-642-31537-4_13.

[43] Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics (Oxford, England) 2007;8:118–27. https://doi.org/10.1093/biostatistics/kxj037.

[44] Leek JT, Johnson WE, Parker HS, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28(6):882–3. https://doi.org/10.1093/bioinformatics/bts034.

[45] Kuhn M. Building predictive models in r using the caret package. J Stat Software 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05.

[46] Dietterich TG. Machine-learning research: four current directions. AI Mag 1997;18(4):97–136.

[47] Džeroski S, Ženko B. Is combining classifiers with stacking better than selecting the best one? Mach Learn 2004;54:255–73. https://doi.org/10.1023/B:MACH.0000015881.36452.6e.

[48] Grau J, Grosse I, Keilwagen J. Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. Bioinformatics 2015;31(15):2595–7. https://doi.org/10.1093/bioinformatics/btv153.