Subgroup identification using virtual twins for human microbiome studies

Hyunwook Koh (hyunwook.koh@stonybrook.edu)
The State University of New York, Korea

Method Article

Keywords: Human microbiome, Subgroup identification, Virtual twins, Cancer immunotherapy, Personalized medicine, Precision medicine

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1548419/v1

Subgroup identification using virtual twins for human microbiome studies

Hyunwook Koh[1,*]

[1]Department of Applied Mathematics and Statistics, The State University of New York, Korea, Incheon, South Korea

Correspondence:

Hyunwook Koh

Address: B521, 119-2 Songdo Moonhwa-Ro, Yeonsu-Gu, Incheon, 21985, South Korea

Email: hyunwook.koh@stonybrook.edu; Phone: +82-032-626-1918

Abstract

Background: Even when the same treatment is employed, some patients are cured, while others are not. The patients that are cured may have beneficial microbes in their body that can boost treatment effects, but it is vice versa for the patients that are not cured. That is, treatment effects can vary depending on the patient's microbiome. If the effects of candidate treatments are well-predicted based on the patient's microbiome, we can select a treatment that is suited to the patient's microbiome or can alter the patient's microbiome to improve treatment effects.

Methods: Here, I introduce a streamlined analytic method, named microbiome virtual twins (MiVT), to evaluate the interplay between microbiome and treatment. MiVT is based on the subgroup identification framework, called virtual twins, that involves a two-step algorithm, 1) treatment effect prediction through machine learning and 2) subgroup identification using a decision tree. MiVT, however, employs a new prediction method, named distance-based machine learning (dML), to improve prediction accuracy in microbiome studies and a new significance test, named bootstrap-based test for regression tree (BoRT), to test if each subgroup's treatment effect is the same with the overall treatment effect.

Results: I demonstrate *in silico* that dML robustly reaches a high prediction accuracy and BoRT is a valid significance test with correctly controlled type I error rates. I also demonstrate the use of MiVT *in praxis* through the gut microbiome study on the effects of cancer immunotherapies on melanoma patients.

Conclusions: The results from MiVT can serve as a useful guideline in microbiome-based personalized medicine to select the therapy that is most suited to the patient's microbiome or to use dietary supplements or therapeutics to tune the patient's microbiome to be suited to the

treatment. MiVT can be implemented using an R package, MiVT, freely available at https://github.com/hk1785/MiVT.

Keywords: Human microbiome, Subgroup identification, Virtual twins, Cancer immunotherapy, Personalized medicine, Precision medicine

Background

The human microbiome is the entire ecosystem of all microbes that inhabit different organs (e.g., gut, mouth, nose, skin, etc) of the human body. The roles of the microbiome on human health or disease have been increasingly studied due to the recent advances in high-throughput sequencing technologies. The key underlying channels through which the microbes can influence human health or disease have been found as immunologic or metabolic regulations and digestive processes [1-3]. The microbiome industry has been rapidly growing, and microbiome-based dietary supplements (e.g., prebiotics, probiotics, dietary fiber), therapeutics (e.g., antibiotics, pharmabiotics, fecal microbiota transplant, phage therapy) and diagnostics are currently flooded.

The two major sequencing platforms for microbiome profiling are 16S rRNA-based amplicon sequencing [4, 5] and shotgun metagenomics [6]. Either of these sequencing platforms can produce various types of metagenomic information, yet the type of the microbiome data on which I focus here is the typical microbiome data that are on microbial abundance and phylogenetic tree information. The data are high-dimensional including numerous microbial features, such as operational taxonomic units (OTUs) or amplicon sequence variants (ASVs), that are characterized by their relative abundances, taxonomic annotations, and phylogenetic tree relationships. The data

are also sparse with excessive zeros, and highly skewed with few microbial features that occupy most of the total abundance; hence, most of the other microbial features are rare variants. The underlying etiological mechanisms can be multifactorial. That is, many microbial features can jointly influence human health or disease, especially the complex disease like cancer, diabetes, obesity, asthma, atopy, brain disorder and so forth. However, it is also likely that only few microbial features solely influence human health or disease [7]. The high complexity of the microbiome data and underlying etiological mechanism makes the downstream data analysis challenging. Hence, more delicate analytic methods and protocols are needed.

Here, I especially pay attention to research question on if the microbiome can improve (or lower) treatment effects. Even when the same medical treatment is employed, some patients are cured, while others are not. For example, the melanoma of the former U.S. president, Jimmy Carter, has been cured by the cancer immunotherapy, called Pembrolizumab, but the same treatment effect does not apply to all the patients for all different types of cancer [8-10]. There are also various cancer immunotherapies, and their treatment effects can all vary. We can suspect that the patients that are cured may have beneficial microbes in their body that can boost treatment effects, but it is vice versa for the patients that are not cured.

Matson et al. showed that the microbiome can improve treatment effects through randomized control trials using genetically similar mice, germ-free mice and gut microbiota transplant [11]. The researchers report that the anti-carcinogenicity of the cancer immunotherapy can be doubled (or halve) depending on the microbiome the mouse had [11]. This indicates that the treatment effect can be improved by altering the microbiome, and it is also the reason why the coadministration of microbiome-based dietary supplements and therapeutics along with a primary treatment like the cancer immunotherapy has been intensely studied.

However, the limitation of Matson et al. is that it was an animal (not human) microbiome study through mouse trials [11]. We have different genetic traits and surrounding environments from mice or any other animals. Hence, the human microbiome should be different from any other animals' microbiome, and it is hard to apply the results from animal trials to the personalized medicine or precision medicine for the humans. The human microbiome studies are therefore deemed to be observational studies that are on the humans, to which we need analytic methods and protocols that are suited.

In this paper, I introduce a streamlined analytic method, named microbiome virtual twins (MiVT), that can predict treatment effects, and then identify subgroups by treatment effects based on microbiome composition to evaluate the interplay between microbiome and treatment. MiVT is based on the subgroup identification framework, called virtual twins [12], that involves a two-step algorithm, 1) treatment effect prediction through machine learning and 2) subgroup identification using a decision tree. MiVT, however, employs a new prediction method, named distance-based machine learning (dML). dML is based on a dimension reduction technique using ecological distance measures and a series of machine learning methods, elastic net (EN) [13], random forest (RF) [14] and deep feedforward network (DFN), to improve prediction accuracy in microbiome studies. MiVT also employs a new significance test, named bootstrap-based test for regression tree (BoRT), to test if each subgroup's treatment effect is the same with or different from the overall treatment effect. Thereby, we can, for example, interpret the final results from MiVT as the patients that have *Mogibacterium* $< 0.00511\%$ and *Akkermansia* $\geq 0.0126\%$ in their gut have significantly a higher treatment effect, 0.100%, compared with the overall treatment effect, 0.0526% (*P*-value (BoRT): <.001). Such results can serve as a useful guideline in microbiome-based personalized medicine to select the best therapy among multiple candidates

depending on the patient's microbiome composition or to use dietary supplements or therapeutics to tune the patient's microbiome composition to improve treatment effect.

The rest of this paper is organized as follows. In the *Methods and materials* section, the methodological details of dML and BoRT are dissected. In the *Results* section, dML and BoRT are evaluated *in silico* (see *Simulations*), and then the use of MiVT is demonstrated *in praxis* through the gut microbiome study on the effects of cancer immunotherapies on melanoma patients (see *Real data applications*). Finally, in the *Discussion and conclusions* section, potential extensions and implementations of MiVT are discussed.

Methods and materials

General settings

Suppose that there are N patients. Let Y_i be a binary cure outcome (0: uncured, 1: cured), T_i be a binary treatment status (0: control, 1: treatment), X_i be a vector of p microbial features (e.g., OTUs, ASVs), where $p >> N$, and Z_i be a treatment effect for $i = 1, \ldots\ldots, N$. While I can describe the binary treatment status as 0 (control) vs. 1 (treatment), it can be more generally 0 (placebo) vs. 1 (treatment), 0 (old treatment) vs. 1 (new treatment), 0 (treatment level 1) vs. 1 (treatment level 2), and so forth. The microbial features can be correlated with each other, and they tend to be phylogenetically related in microbiome studies. The treatment and microbial features can have marginal effects on the cure outcome. It is assumed that the treatment has no effect on microbial features [12], which is simply satisfied if the microbiome is profiled before the treatment.

Step 1 (treatment effect prediction)

To see which microbiome composition improves treatments effects, we first need to predict treatment effects. The treatment effect can be measured by comparing the cure rate of the patient who received the treatment (say, treatment) and the cure rate of the same patient who did not receive the treatment (say, control) as in (Eq. 1).

$$Z_i = P(Y_i = 1|T_i = 1, X_i) - P(Y_i = 1|T_i = 0, X_i) \quad (1)$$

However, the same patient cannot be assigned to both of the treatment and control groups at the same time. Therefore, one of the cure rates is always missing, and thus patient-level treatment effects Z_i's are not measurable. This dilemma is called the "fundamental problem of causal inference" [15].

An alternative approach can be to employ genetically identical twins while assigning one of the twins to the treatment group and the other to the control group, and then we can compare their cure rates. For example, we can first transfer cancer cells into the twins making both of them cancer patients, and then give a medication to one of the twins and a placebo to the other, and then compare their cure rates. However, this approach can be limitedly permitted in an animal trial like the mouse trials of [11]. It is, of course, unethical to conduct such trials on the humans. Again, it is hard to apply the results from animal trials to the personalized medicine or precision medicine for the humans.

Therefore, to predict treatments effects in an observational study where the human trials on twins are not permitted, I employ the method, called virtual twins [12]. The virtual twins mimic the animal trials on twins, and its key ideas are as follows. We can first predict the cure rate of the patient in the treatment group, and then predict the cure rate of "virtually" the same patient (say, virtual twin) in the control group. Then, we can finally predict the treatment effect by subtracting the former cure rate of the real patient from the latter cure rate of the virtual twin. Here, "virtual"

means to switch only the data of the variable on the treatment status from 1 to 0, while "twin" means to fix all the other data on the other variables. That is, more formally, the virtual twins calculate the cure rates for both treatment and control groups by flipping the treatment status on the data while fixing all the other data in a prediction model, and then the treatment effects are calculated as in (Eq. 2).

$$\hat{Z}_i = f(Y_i = 1|T_i = 1, X_i) - f(Y_i = 1|T_i = 0, X_i) \quad (2)$$

where $f(.)$ is a prediction model, $f(Y_i = 1|T_i = 1, X_i)$ is the predicted cure rate of the treatment, and $f(Y_i = 1|T_i = 0, X_i)$ is the predicted cure rate of the control.

We can notice here that the success of virtual twins primarily hinges on the accuracy of the prediction model $f(.)$ in (Eq. 2). The original virtual twins paper [12] suggests to use RF [14], but here I introduce dML for higher accuracy in microbiome studies. dML tailors machine learning methods, such as EN [13], RF [14] and DFN, to account for the unique features of the microbiome data, such as the high-dimensionality, sparsity and phylogenetic relationships. For this, dML first extracts the lower-dimensional representations of the microbial features (say, coordinates) through multidimensional scaling [16] based on an ecological distance measure, such as Euclidean distance (Euclidean), Jaccard dissimilarity (Jaccard) [17], Bray-Curtis dissimilarity (BC) [18], unweighted UniFrac distance (UUniFrac) [19], generalized UniFrac distance (GUniFrac) [20] or weighted UniFrac distance (WUniFrac) [21]. The coordinate matrix can be derived by the eigen-decomposition of the kernel matrix (denoted as, $K_{(h)}$) in (Eq. 3).

$$K_{(h)} = -\frac{1}{2}\left(I_N - \frac{1_N 1_N^T}{N}\right) D_{(h)}^2 \left(I_N - \frac{1_N 1_N^T}{N}\right), \quad (3)$$

where h is an index for a distance measure in a set of candidate measures (e.g., $h \in$ {Euclidean, Jaccard, BC, UUniFrac, GUniFrac, WUniFrac}), $D_{(h)}$ is the $N \times N$ pairwise distance matrix and

$D_{(h)}^2$ is its element-wise square matrix, I_N is the $N \times N$ identity matrix, and 1_N is the $N \times 1$ vector of 1's. Let $\lambda_{(h)1}$, …, $\lambda_{(h)M}$ be positive eigenvalues and $q_{(h)1}$, …, $q_{(h)M}$ be their corresponding eigenvectors obtained by the eigen-decomposition of the kernel matrix $K_{(h)}$ in (Eq. 3). Then, the $N \times M$ coordinate matrix (denoted as, $V_{(h)}$) can be derived as in (Eq. 4).

$$V_{(h)} = Q_{(h)} \Lambda_{(h)} \tag{4}$$

where $Q_{(h)}$ is the $N \times M$ matrix of $q_{(h)1}$, …, $q_{(h)M}$, $\Lambda_{(h)}$ is the $M \times M$ diagonal matrix of $\lambda_{(h)1}^{1/2}$, …, $\lambda_{(h)M}^{1/2}$, and $M \leq N$. Then, the coordinates (i.e., the lower-dimensional representations of the microbial features) are used as inputs in a machine learning method, such as EN [13], RF [14] or DFN, which is to relax the high-dimensionality and sparsity of the microbiome data and modulate phylogenetic tree information using phylogenetic and non-phylogenetic distance measures.

The distance measures are well-designed by properly reflecting the microbial abundance and phylogenetic tree information in their formula; hence, they have been widely used in many prior statistical methods [22-28]. However, they are distinct distance measures. For example, Euclidean, Jaccard [17] and BC [18] are non-phylogenetic, while the UniFrac distances [19-21] are phylogenetic. In addition, Jaccard [17] and UUniFrac [19] are based on incidence (i.e., presence/absence) information, while Euclidean, BC [18], GUniFrac [20] and WUniFrac [21] are based on abundance information. In practice, we do not know which distance measure makes the best prediction accuracy in advance due to the varying and unknown nature of the underlying prediction patterns.

Furthermore, the performance of machine learning methods also varies depending on underlying prediction patterns. The EN fine-tunes the extent of variable selection and shrinkage in a linear model through the regularization that linearly combines the L_1 and L_2 penalties [13].

Thereby, the EN can suit the prediction patterns that are linear with varying sparsity levels. The RF [14] is a bootstrap aggregation method that averages the predictions resulting from the collection of bagged decision trees built with randomly selected inputs. Thereby, the RF can suit the prediction patterns that are non-linear with varying sparsity levels. The DFN (a.k.a. multi-layer perceptron) extracts various linear combinations of inputs, and then nonlinearly maps them to the outputs through a large number of artificial neurons and hidden layers. Thereby, the DFN can suit various prediction patterns that are multifactorial or not, linear or nonlinear, and so forth. However, the DFN may require a huge sample size because of its high model complexity. Similarly, in practice, we do not know which machine learning method makes the best prediction accuracy in advance due to the varying and unknown nature of the underlying prediction patterns.

Therefore, dML employs the *k*-fold cross-validation (CV) to select the optimal combination of distance measure and machine learning method that results in the smallest cross-entropy of (Eq. 5).

$$L_{(h)(l)} = -\frac{1}{N_\Phi}\sum_{i \in \Phi}[y_i \log\left(f_{(h)(l)}(X_i)\right) + (1 - y_i)\log\left(1 - f_{(h)(l)}(X_i)\right)], \quad (5)$$

where Φ is the validation set of patients, N_Φ is the number of patients in the validation set, h is an index for a distance measure in a set of candidate measures (e.g., $h \in$ {Euclidean, Jaccard, BC, UUniFrac, GUniFrac, WUniFrac}), and l is an index for a machine learning method (e.g., $l \in$ {EN, RF, DFN}. As a result, dML can robustly adapt to various prediction patterns (e.g., linear or nonlinear, sparse or dense, rare or common, phylogenetically related or independent, and so forth) through the extensive search in distance measure and machine learning method.

Let $f_{dML}(.)$ denote the prediction model with the optimal combination of distance measure and machine learning method that results in the smallest cross-entropy. Then, the treatment effects are predicted as in (Eq. 6).

$$\hat{Z}_i = f_{dML}(Y_i = 1|T_i = 1, X_i) - f_{dML}(Y_i = 1|T_i = 0, X_i) \quad (6)$$

where $f_{dML}(Y_i = 1|T_i = 1, X_i)$ is the predicted cure rate of the treatment, and $f_{dML}(Y_i = 1|T_i = 0, X_i)$ is the predicted cure rate of the control.

Step 2 (subgroup identification)

To see which microbiome composition improves treatments effects, after the first step of treatment effect prediction, we need to classify patients into subgroups by treatment effects based on patients' microbiome composition. For this, the virtual twins paper [12] suggests to use a regression tree [29] that involves stratifying or segmenting the predictor space (i.e., the microbial feature space or upper-level taxonomic space in microbiome studies) into a number of simple regions by treatment effects. Thereby, we can make simple and useful interpretations using a nice graphical representation of the top-down tree structure [29].

Let R_j's be distinct and non-overlapping regions and $\hat{Z}_{R_j}$'s be their corresponding treatment effects estimated by the mean treatment effects for the patients in each region for $j = 1, \ldots, J$. Then, if we let A be the group of all N patients, $\hat{Z}_{\mathrm{A}}$ becomes the overall mean treatment effect. R_j's. $\hat{Z}_{R_j}$'s can be efficiently found through recursive binary partitioning [29]. As a result, we can identify subgroups (R_j's) and estimate their treatment effects ($\hat{Z}_{R_j}$'s) that can be compared with the overall treatment effect ($\hat{Z}_{\mathrm{A}}$). R_j is, for example, the subgroup of patients that have *Mogibacterium* $<$ 0.00511% and *Akkermansia* $\geq$ 0.0126% in their gut, and $\hat{Z}_{R_j}$ is their treatment effect estimated as, 0.100%, that can be compared with the overall treatment effect $\hat{Z}_{\mathrm{A}}$ of 0.0526%.

However, it has all been so far about parameter estimation with no facility for hypothesis testing. The problem is that we do not know if the estimated difference between each subgroup's treatment effect and the overall treatment effect is statistically significant, which is on the null and alternative hypotheses in (Eq. 7).

$$H_0: \hat{Z}_{R_j} = \hat{Z}_{\mathrm{A}} \text{ vs. } H_1: \hat{Z}_{R_j} \neq \hat{Z}_{\mathrm{A}}, \tag{7}$$

Thus, I introduce a significance test, BoRT, that is based on the test statistic (denoted as, *U*) in (Eq. 8).

$$U = \hat{Z}_{R_j} - \hat{Z}_{\mathrm{A}} \tag{8}$$

The test statistic *U* is simply the difference in mean between the overall and subgroup treatment effects. If *U* is positive, the overall treatment effect is greater than the subgroup treatment effect, but it is vice versa if *U* is negative. A large absolute value of *U* tends to lend credence to H_1.

The distribution of $\hat{Z}_{\mathrm{A}}$ can be approximated using the bootstrap method [30] by random sampling with replacement of the patient-level treatment effects $\hat{Z}_i$'s. Let $\hat{Z}_i^b$'s be a bootstrap resample of the patient-level treatment effects $\hat{Z}_i$'s. Then, the bootstrap overall treatment effect (denoted as, $\hat{Z}_A^b$) can be calculated as in (Eq. 9).

$$\hat{Z}_A^b = \frac{1}{N}\sum_{i=1}^{N} \hat{Z}_i^b \tag{9}$$

Under H_0, all *N* patient in *A* are equally likely to belong to R_j, which indicates a random relocation of the selected region (denoted as, R_j^r). Hence, the null test statistic value can be calculated as in (Eq. 10).

$$U_{Null}^b = \hat{Z}_{R_j}^b - \hat{Z}_A^b, \tag{10}$$

where $\hat{Z}_{R_j}^b = \frac{1}{N_{R_j}}\sum_{i \in R_j^r} \hat{Z}_i^b$ and N_{R_j} is the number of patients that belong to R_j. If we repeat it many times (say, *B* times), *B* null test statistic values (U_{Null}^b for $b = 1, \ldots B$) are generated. Then, a *P*-

value is calculated as the proportion of the null test statistic values that are equal to or greater than the observed test statistic value as in (Eq. 11).

$$\sum_{b=1}^{B} I(|U_{Null}^{b}| \geq |U_{Obs}|)/B, \quad (11)$$

where $I(.)$ is an indicator function and U_{Obs} is the observed test statistic value that is calculated using the original data.

Results

Simulations

To reflect real microbiome composition, I first estimated parameters (i.e., proportions and dispersion) of the Dirichlet-multinomial distribution [31] based on 755 microbial features (that have the mean proportion > 10^{-5}) of the gut microbiome for 39 melanoma patients prior to immunotherapy in [32]. Then, I randomly generated counts for 200 patients from the Dirichlet-multinomial distribution using the estimated parameters and total counts randomly generated from the uniform distribution from 10,000 to 100,000 to reflect varying total read counts. A half of patients (i.e., 100 patients among 200 patients) was assigned to the test set, while the other half of patients was assigned to training set. I generated binary cure outcomes (Y_i's) based on the logistic regression model (Eq. 12).

$$\text{logit P}(Y_i = 1) = \beta_0 + \beta_1 T_i + \beta_2 \sum_{j\in\Omega} X_{ij} + \beta_3 \sum_{j\in\Omega} T_i X_{ij}, \quad (12)$$

where i is the patient (i = 1, ……, 200), j is the microbial feature (j = 1, ……, 755), Y_i is the binary cure outcome, $T_i = 1$ (treatment) for a half of patients, $T_i = 0$ (placebo) for the other half of patients, X_{ij} is the proportion, $\beta_0 = 0.1$ (marginal placebo effect), $\beta_1 = 0.5$ (marginal treatment effect), Ω is the set of microbial features that influence treatment effects, β_2 = is the marginal effect of the

282 microbial features and β_3 is the interaction effect (treatment effect influenced by microbial
283 features).

284 I surveyed two sets of the marginal (β_2) and interaction (β_3) effects of the microbial features
285 in (Eq. 12) as $\beta_2 = 0.0005$ and $\beta_3 = 0.001$ for relatively small effects and $\beta_2 = 0.001$ and $\beta_3 =$
286 0.0015 for relatively large effects, respectively. I also surveyed Ω in (Eq. 12) (i.e., the set of
287 microbial features that influence treatment effects) using two different scenarios, respectively.
288 First, I randomly selected 10 % of the microbial features (denoted as, $\Omega = \{$randomly selected
289 features$\}$). Second, I partitioned microbial features into 10 clusters using the partitioning-around-
290 medoids (PAM) algorithm [33] based on phylogenetic distances, and then randomly selected one
291 cluster (i.e., $\Omega = \{$phylogenetically related features$\}$). This mimics a situation when
292 phylogenetically related microbial features jointly influence treatment effects. I repeated each
293 scenario 300 times, and report average estimates.

294 I evaluated the proposed method, dML, compared with other existing methods, EN [13], RF
295 [14] and DFN addressing compositional issues using the centered log-ratio transformation [34],
296 with respect to test classification error and test area under the curve (AUC). I observed that as the
297 marginal (β_2) and interaction (β_3) effects of the microbial features in (Eq. 12) increase, the
298 prediction accuracy increases for all surveyed methods, but their relative ranks are equally retained
299 [Fig. 1 and Fig. 2]. We can observe that dML reaches the smallest test classification error and the
300 highest test AUC (i.e., the highest prediction accuracy) for both scenarios of randomly selected
301 features [Fig. 1A,C and Fig. 2A,C] and phylogenetically related features [Fig. 1B, D and Fig.
302 2B,D]. This indicates that dML robustly reaches the highest prediction accuracy through the
303 extensive search in distance measure and machine learning method.

I also evaluated BoRT with respect to type I error rate and power. The empirical type I error was calculated as the proportion of the P-values for the randomly relocated regions (R_j^r's) that are smaller than 0.05, and the empirical power was calculated as the proportion of the P-values for the selected regions (R_j's), as they are, that are smaller than 0.05. We can observe that the empirical type I error rates are close to 5% [Table 1]. Hence, BoRT is a valid significance test with the correct control of type I error rate. We can also observe that the empirical powers for the relatively large effects of $\beta_2 = 0.001$ and $\beta_3 = 0.0015$ in (Eq. 12) are greater than the empirical powers for the relatively small effects of $\beta_2 = 0.0005$ and $\beta_3 = 0.001$ in (Eq. 12) [Table 1].

Real data applications

Here, I demonstrate the use of MiVT through the gut microbiome study on the effects of cancer immunotherapies on melanoma patients in [32]. The researchers collected fecal samples from metastatic melanoma patients prior to immunotherapy, and processed them via shotgun metagenomics [32]. Then, the researchers processed raw sequence data using NGS-QC and NCBI BMTagger Human Contamination Screening Tool for quality controls, and then constructed feature tables, taxonomic annotations and phylogenetic tree using MetaPhlAn [35]. More details on metagenomic sequencing and profiling procedures can be found in [32].

The microbiome data contain 39 metastatic melanoma patients treated by immune checkpoint inhibitors targeting the programmed cell death 1 protein (PD-1), cytotoxic T lymphocyte–associated antigen 4 (CTLA-4) or both PD-1 and CTLA-4, and 755 microbial features that have the mean proportion > 10^{-5}. I dropped one patient treated by Anti-CTLA-4 only, and compared 14 patients treated by Anti-PD-1 only with 24 patients treated by both Anti-PD-1 and Anti-CTLA-4 [Table 2]. Of the 14 patients treated by Anti-PD-1 only ($T_i = 0$), 10 (71.4%) were non-responders

($T_i = 0$) and 4 (28.6%) were responders ($Y_i = 1$), while of the 24 patients treated by both Anti-PD-1 and Anti-CTLA-4 ($T_i = 1$), 10 (41.7%) were non-responders ($Y_i = 0$) and 14 (58.3%) were responders ($Y_i = 1$) [Table 2]. Hence, it seems more likely to be a responder if the patient is treated by both Anti-PD-1 and Anti-CTLA-4 than by Anti-PD-1 only. However, the Fisher's exact test gives a non-significant result on the association between treatment and outcome status (*P*-value: 0.101).

I surveyed if the gut microbiome improves (or lowers) the effect of Anti-CTLA-4 over Anti-PD-1 using MiVT. I performed the treatment prediction on the feature-level, but the subgroup identification on the lower-dimensional genus and species levels, respectively, while removing unknown and unclassified genera and species in taxonomic annotation. For reference, genera and species are also better perceived by microbiome researchers than OTUs or ASVs. I interpret the results on genera [Fig. 3 and Table 3] and species [Fig. 4 and Table 4] as follows.

On genera [Fig. 3 and Table 3]**.** The melanoma patients that have the genus *Mogibacterium* < 0.00511% and the genus *Akkermansia* ≥ 0.0126% in their gut have significantly a higher effect of Anti-CTLA-4, 0.100%, compared with the overall effect of Anti-CTLA-4, 0.0526% (*P*-value (BoRT): <.001) [Fig. 3 and Table 3]. On the contrary, the melanoma patients that have the genus *Mogibacterium* ≥ 0.00511%, the genus *Erysipelatoclostridium* ≥ 0.0036% and the genus *Roseburia* ≥ 0.165% in their gut have significantly a lower effect of Anti-CTLA-4, 0.000%, compared with the overall effect of Anti-CTLA-4, 0.0526% (*P*-value (BoRT): <.001) [Fig. 3 and Table 3]. This indicates that the genera, *Mogibacterium*, *Erysipelatoclostridium* and *Roseburia*, might be harmful in the administration of Anti-CTLA-4 over Anti-PD-1, while the genus *Akkermansia* might be beneficial.

On species [Fig. 4 and Table 4]. The melanoma patients that have the species *Faecalibacterium prausnitzii* $\geq$ 0.0498%, the species *Erysipelotrichaceae bacterium 3_1_53* < 0.0136% and the species *Streptococcus infantis/mitis* < 0.00567 in their gut have significantly a higher effect of Anti-CTLA-4, 0.100%, compared with the overall effect of Anti-CTLA-4, 0.0526% (*P*-value (BoRT): <.001) [Fig. 4 and Table 4]. On the contrary, the melanoma patients that have the species *Faecalibacterium prausnitzii* < 0.0498% and the species *Eubacterium sp. 3_1_31* < 0.014% in their gut have significantly a lower effect of Anti-CTLA-4, 0.000%, compared with the overall effect of Anti-CTLA-4, 0.0526% (*P*-value (BoRT): <.001) [Fig. 4 and Table 4]. This indicates that the species, *Erysipelotrichaceae bacterium 3_1_53* and *Streptococcus infantis/mitis*, might be harmful in the administration of Anti-CTLA-4 over Anti-PD-1, while the species, *Faecalibacterium prausnitzii* and *Eubacterium sp. 3_1_31*, might be beneficial.

For additional reference, the RF with UUniFrac was the optimal combination that resulted in the smallest CV cross-entropy of 0.656 [Table 5].

Discussion and conclusions

In this paper, I introduced a streamlined analytic method, MiVT, that predicts treatment effects and identifies subgroups by treatment effects based on the patient's microbiome composition to evaluate the interplay between microbiome and treatment. As parts of MiVT, I introduced a new prediction method, dML, to improve prediction accuracy in microbiome studies and a new significance test, BoRT, to test if each subgroup's treatment effect is the same with or different from the overall treatment effect. I demonstrated *in silico* that dML robustly reaches a high prediction accuracy and BoRT is a valid significance test correctly controlling type I error rates. I

also demonstrated the use of MiVT *in praxis* through the gut microbiome study on the effects of cancer immunotherapies on melanoma patients [32]. This example study was equipped with the binary cure outcome, binary treatment status, microbial features and phylogenetic tree that are required to use MiVT. Moreover, the assumption that the treatment has no effect on microbial features was satisfied because the fecal samples were collected prior to immunotherapy. I performed the subgroup identification on the lower-dimensional genus and species levels for better interpretability, but any other taxonomic levels can also be surveyed. The results from MiVT can be a useful guideline in microbiome-based personalized medicine or precision medicine to select the therapy that is most suited to the patient's microbiome or to use dietary supplements or therapeutics to tune the patient's microbiome to be suited to the treatment.

I described MiVT only for a binary cure outcome, yet in practice, there are many different types of cure outcomes, such as continuous, survival and repeated measures outcomes. Hence, further extensions of MiVT are needed to make it more practical. The candidate distance measures, machine learning methods and implementation procedures that I described were sufficient to reach the robust performance in my simulations and real data applications. However, researchers may believe that they are less sufficient, and thus, for example, they want to consider some more candidate parameter values, repeat 10-fold CV more times, and so forth. Hence, I added various user options in the R package, MiVT, for different model specifications, implementation procedures, and so forth. It would be better to make it overly sufficient than less sufficient. If it is less sufficient, MiVT may not have enough flexibility to make it robust.

Abbreviations

ASV: amplicon sequence variant

AUC: Area under the curve

BC: Bray-Curtis dissimilarity

BoRT: Bootstrap-based test for regression tree

CTLA-4: Cytotoxic T lymphocyte-associated antigen 4

CV: Cross-validation

DFN: Deep feedforward network

dML: Distance-based machine learning

EN: Elastic net

Euclidean: Euclidean distance

GUniFrac: Generalized UniFrac distance

Jaccard: Jaccard dissimilarity

MiVT: Microbiome virtual twins

OTU: Operational taxonomic units

PAM: Partitioning-around-medoids

PD-1: Programmed cell death 1 protein

RF: Random forest

UUniFrac: Unweighted UniFrac distance

WUniFrac: Weighted UniFrac distance

Acknowledgements

The author is grateful to the reviewers for their insightful observations and comments.

Author's contributions

H.K. is the only author who contributes to every aspect of this work.

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2021R1C1C1013861).

Availability of data and materials

H.K. used public metagenomic sequencing data for the gut microbiome study on the effects of cancer immunotherapies on melanoma patients that are available at the European Nucleotide Archive under the accession number, PRJNA397906. MiVT can be implemented using an R package, MiVT, that is freely available at https://github.com/hk1785/MiVT.

Ethics approval and consent to participate

All utilized microbiome datasets are publicly available. No ethics approval or consent to participate was required for this study.

Consent for publication

All utilized microbiome datasets are publicly available. No consent for publication was required for this study.

Competing interests

The author declares no competing interest.

References

1. Garrett WS, Gordon JI, Glimcher LH. Homeostasis and inflammation in the intestine. *Cell* 2010:140:859-870.
2. Cox LM, Yamanishi S, Sohn J, Alekseyenko AV, Leung JM, Cho I, Kim SG, Li H, Gao Z, Mahana D et al. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* 2014:158(4):705-721.
3. Cox LM, Blaser MJ Antibiotics in early life and obesity. *Nat Rev Endocrinol.* 2014:11(3):182-190.
4. Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques. *Genome Res.* 2009:19(7):1141-1152.
5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010:7(5):335-336.
6. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp.* 2012:2(3).
7. Koh H, Zhao N. A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. *Microbiome* 2020:8(63).
8. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC. et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med.* 2010:363(8):711-723.

9. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, Daud A, Carlino MS, McNeil C, Lotem M et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N Engl J Med.* 2015:372(26):2521-2532.
10. Weber JS, D'Angelo SP, Minor D, Hodi FS, Gutzmer R, Neyns B, Hoeller C, Khushalani NI, Miller Jr, WH, Lao CD et al. Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *Lancet Oncol.* 2015:16(4):375-384.
11. Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre M, Luke JJ, Gajewski TF. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* 2018:359(6371):104-108.
12. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med.* 2011:30(24):2867-2880.
13. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B* 2005:67(2):301-320.
14. Breiman L. Random Forests. *Machine Learning* 2001:45:5-32.
15. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol.* 1974:66(5):688-701.
16. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika* 1952:17:401-419.
17. Jaccard P. The distribution of the flora in the alpine zone. *New Phytol.* 1912:11:37-50.
18. Bray JR, Curtis JT. An ordination of the upland forest communities of Southern Wisconsin. *Ecol Monogr.* 1957:27(32549).

19. Lozupone CA, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol.* 2005:71:8228-8235.
20. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 2016:32:2611-2617.
21. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol.* 2007:73:1576–1585.
22. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* 2001:26:32–46.
23. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 2001:82:290–297.
24. Tang Z, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* 2016:32:2618-2625.
25. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Hum Genet.* 2015:96(5):797-807.
26. Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome* 2017:5(45).
27. Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome* 2017:5(17).

28. Koh H, Li Y, Zhan X, Chen J, Zhao N. A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Front Genet.* 2019:458(10).

29. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees, New York: Routledge; 1984.

30. Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann Statist.* 1979:7(1):1-26.

31. Sanders HL. Marine Benthic Diversity: A Comparative Study. *Am Nat.* 1968:102(925):243-282.

32. Aitchison J. The statistical analysis of compositional data. *J R Statist Soc B.* 1982:44(2):139-177.

33. Mosimann JE. On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions. *Biometrika* 1962:49(1/2):65-82.

34. Frankel AE, Coughlin LA, Kim J, Froehlich TW, Xie Y, Frenkel EP, Koh AY. Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients. *Neoplasia* 2017:19(10):848-855.

35. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J Math Model Algor.* 2006:5:475-504.

Tables and Figure Legends

Table 1. Empirical type I error rates and powers for BoRT (unit: %).

	Type I error (%)	Power (%)	
		Randomly selected features	Phylogenetically related features
Small effects	5.027	45.035	46.688
Large effects	4.860	49.312	47.368

Table 2. The contingency table on the treatment and outcome status.

	Anti-PD-1 ($T_i = 0$)	Anti-PD-1 & Anti-CTLA-4 ($T_i = 1$)
NR ($Y_i = 0$)	10 (71.4%)	10 (41.7%)
R ($Y_i = 1$)	4 (28.6%)	14 (58.3%)
Sum	14 (100%)	24 (100 %)

Table 3. The results of BoRT from MiVT on the microbial genera in the gut of melanoma patients that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1. *R_1, R_2, R_3, R_4, R_5 and R_6 are the identified subgroups that correspond with the terminal nodes from left to right in Fig. 3. N_{R_j} is the sample size for each subgroup j = 1, ……, 6. Overall TE represents the overall treatment effect, and Subgroup TE represents the subgroup treatment effect.

	R_1	R_2	R_3	R_4	R_5	R_6
N_{R_j}	8	5	5	5	5	10
Overall TE	0.053%	0.053%	0.053%	0.053%	0.053%	0.053%
Subgroup TE	0.000%	0.020%	0.080%	0.020%	0.080%	0.100%
Subgroup TE – Overall TE	-0.053%	-0.033%	0.027%	-0.033%	0.027%	0.047%
P-value (BoRT)	<.001	0.126	0.188	0.128	0.196	<.001

Table 4. The results of BoRT from MiVT on the microbial species in the gut of melanoma patients that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1. *R_1, R_2, R_3, R_4 and R_5 are the identified subgroups that correspond with the terminal nodes from left to right in Fig. 4. N_{R_j} is the sample size for each subgroup j = 1, ……, 5. Overall TE represents the overall treatment effect, and Subgroup TE represents the subgroup treatment effect.

	R_1	R_2	R_3	R_4	R_5
N_{R_j}	5	5	7	5	16
Overall TE	0.053%	0.053%	0.053%	0.053%	0.053%
Subgroup TE	0.000%	0.020%	0.014%	0.040%	0.100%
Subgroup TE – Overall TE	-0.053%	-0.033%	-0.038%	-0.013%	0.047%
P-value (BoRT)	0.007	0.129	0.024	0.580	<.001

Table 5. The CV cross-entropy values for each combination of distance measure and machine learning method from the real data application of the gut microbiome study on the effects of cancer immunotherapies on melanoma patients.

	Euclidean	Jaccard	BC	UUniFrac	GUniFrac	WUniFrac
EN	0.994	0.903	0.813	0.761	0.745	0.897
RF	0.686	0.672	0.704	0.656	0.714	0.677
DFN	3.322	2.733	2.655	3.862	3.113	2.674

588 **Fig. 1.** Empirical test classification errors (see Error (%)) and test AUC (see AUC (%)) for the
589 relatively small effects of $\beta_2 = 0.0005$ and $\beta_3 = 0.001$ (Eq. 12). **A** and **C**. For a situation when
590 randomly selected features influence treatment effects (i.e., Ω = {randomly selected features}. **B**
591 and **D**. For a situation when phylogenetically related microbial features influence treatment effects
592 Ω = {i.e., phylogenetically related features}.
593
594 **Fig. 2.** Empirical test classification errors (see Error (%)) and test AUC (see AUC (%)) for the
595 relatively large effects of $\beta_2 = 0.001$ and $\beta_3 = 0.0015$ (Eq. 12). **A** and **C**. For a situation when
596 randomly selected features influence treatment effects (i.e., Ω = {randomly selected features}. **B**
597 and **D**. For a situation when phylogenetically related microbial features influence treatment effects
598 Ω = {i.e., phylogenetically related features}.
599
600 **Fig. 3.** The fitted regression tree by MiVT on the microbial genera in the gut of melanoma patients
601 that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1 (Unit: %). * G46:
602 *Mogibacterium*, G10: *Erysipelatoclostridium*, G33: *Roseburia*, G103: *Akkermansia*, G13:
603 *Massiliomicrobiota*.
604
605 **Fig. 4.** The fitted regression tree by MiVT on the microbial species in the gut of melanoma patients
606 that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1 (Unit: %). * S40:
607 *Faecalibacterium prausnitzii*, S6: *Eubacterium sp. 3_1_31*, S8: *Erysipelotrichaceae bacterium*
608 *3_1_53*, S183: *Streptococcus infantis/mitis*.

Figures

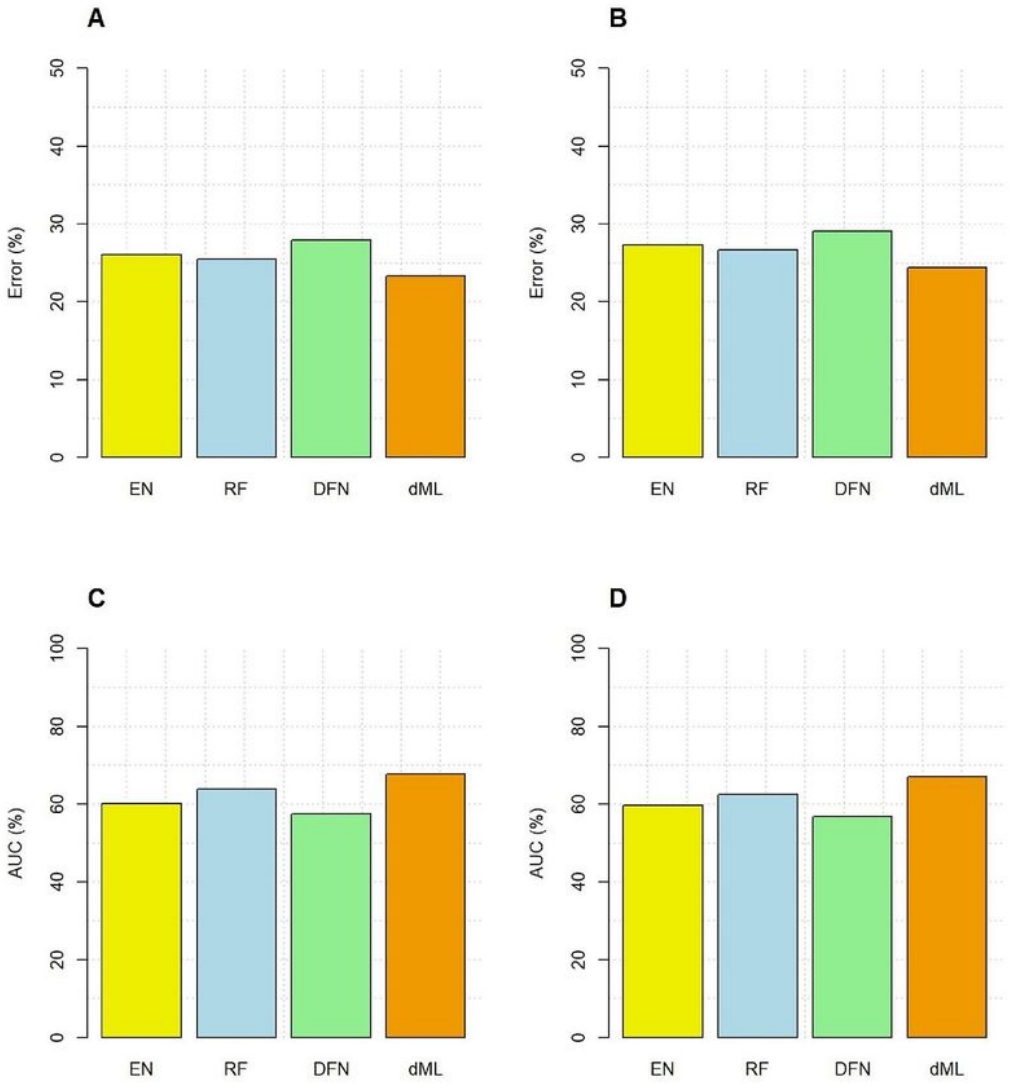

Fig. 1. Empirical test classification errors (see Error (%)) and test AUC (see AUC (%)) for the relatively small effects of $\beta_2 = 0.0005$ and $\beta_3 = 0.001$ (Eq. 12). **A** and **C**. For a situation when randomly selected features influence treatment effects (i.e., Ω = {randomly selected features}. **B** and **D**. For a situation when phylogenetically related microbial features influence treatment effects Ω = {i.e., phylogenetically related features}.

Figure 1

See image above for figure legend

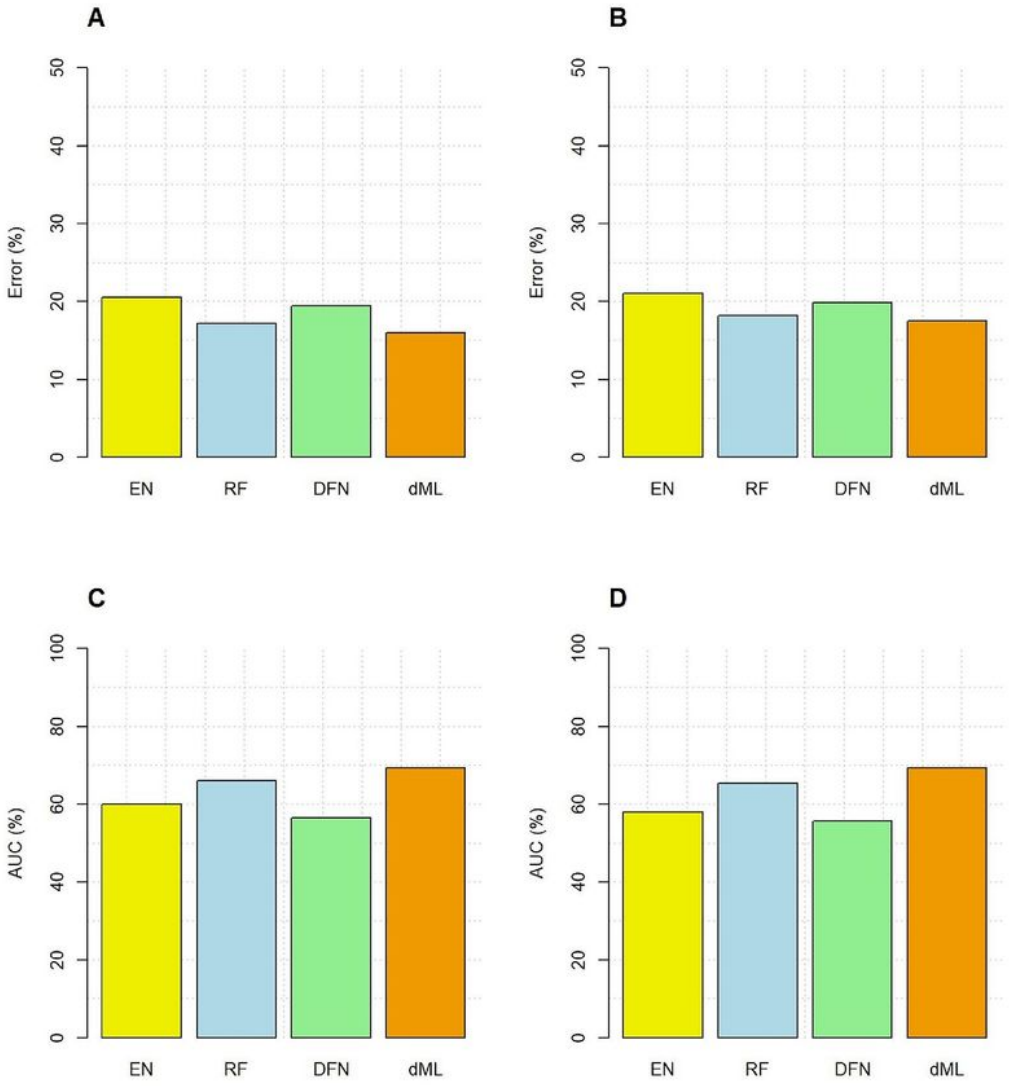

Fig. 2. Empirical test classification errors (see Error (%)) and test AUC (see AUC (%)) for the relatively large effects of $\beta_2 = 0.001$ and $\beta_3 = 0.0015$ (Eq. 12). **A** and **C**. For a situation when randomly selected features influence treatment effects (i.e., Ω = {randomly selected features}. **B** and **D**. For a situation when phylogenetically related microbial features influence treatment effects Ω = {i.e., phylogenetically related features}.

Figure 2

See image above for figure legend

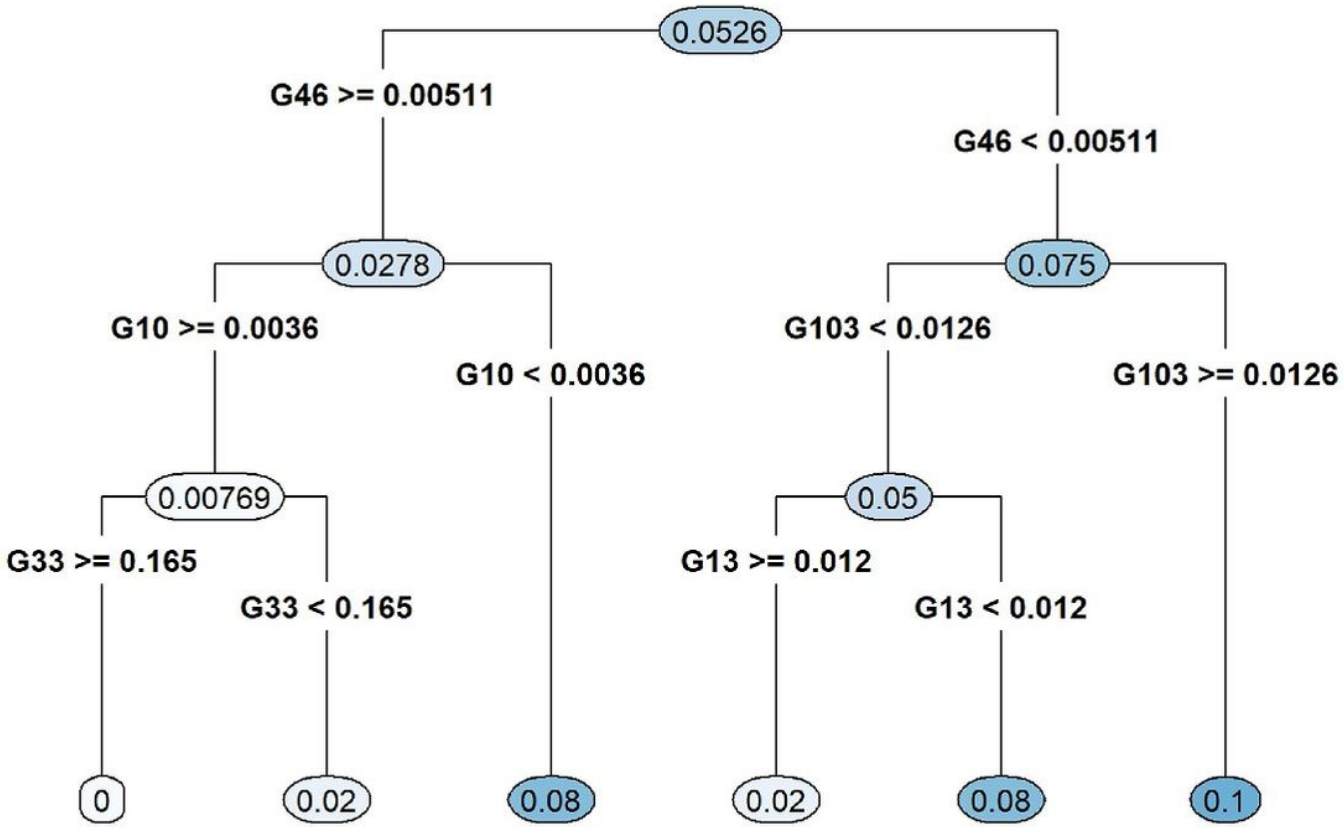

Fig. 3. The fitted regression tree by MiVT on the microbial genera in the gut of melanoma patients that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1 (Unit: %). * G46: *Mogibacterium*, G10: *Erysipelatoclostridium*, G33: *Roseburia*, G103: *Akkermansia*, G13: *Massiliomicrobiota*.

Figure 3

See image above for figure legend

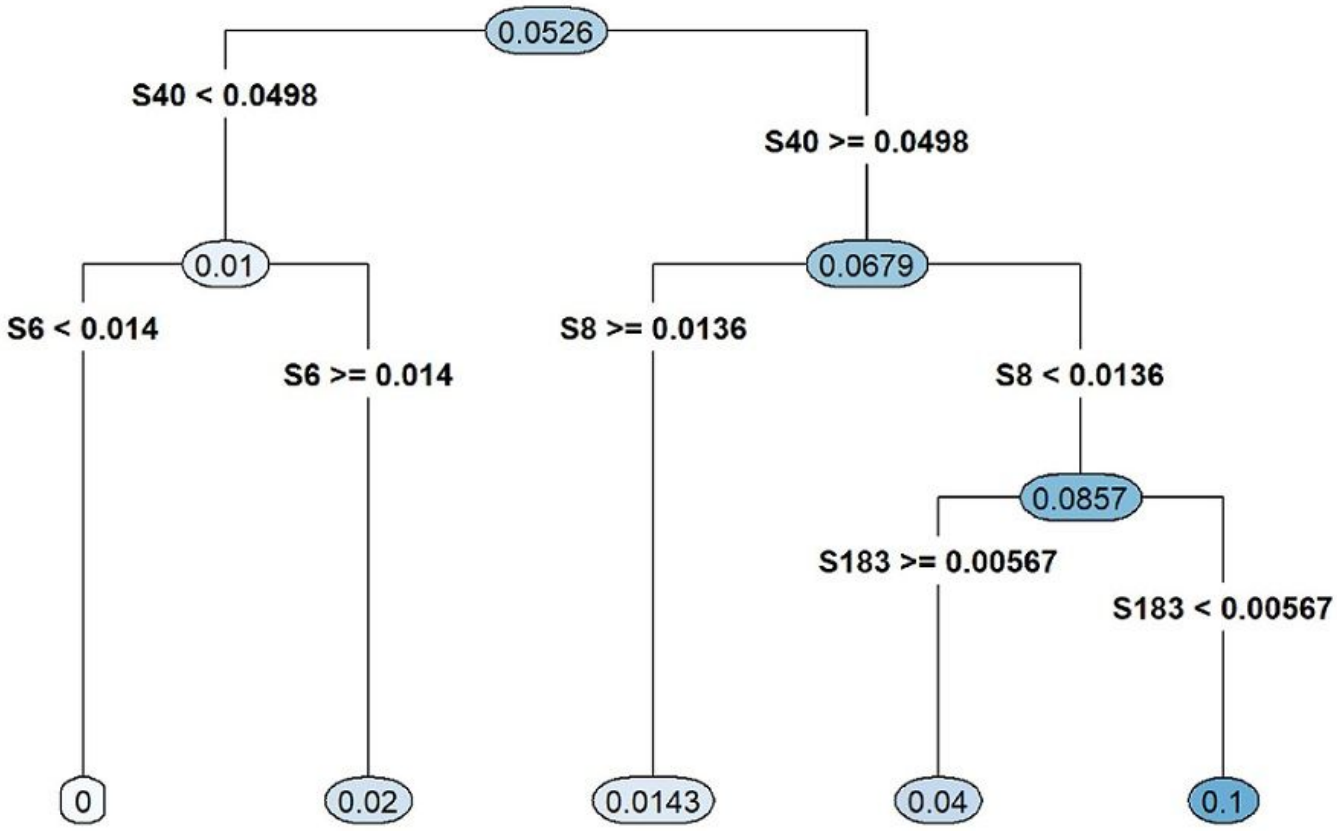

Fig. 4. The fitted regression tree by MiVT on the microbial species in the gut of melanoma patients that improve (or lower) the effect of Anti-CTLA-4 over Anti-PD-1 (Unit: %). * S40: *Faecalibacterium prausnitzii*, S6: *Eubacterium sp. 3_1_31*, S8: *Erysipelotrichaceae bacterium 3_1_53*, S183: *Streptococcus infantis/mitis*.

Figure 4

See image above for figure legend