Tree Variable Selection for Paired Case–Control Studies with Application to Microbiome Data

Min Lu and Hemant Ishwaran

1 Introduction

Paired samples occur in microbiome studies when they are collected from different locations of the same individual or from paired individuals with familial ties. Human microbiome can be shared among family members with variations in each individual's microbial community [16, 18]. Suppose an identifiable "core microbiome" exists at the microbial gene level and deviations from this core are associated with different physiologic states. It is of interest to study how family ties play a role in these deviations. For example, if deviations from a core gut microbiome are associated with body mass index (BMI), we can define "individual" and "family" outcomes with labels obese/lean, where for example obese family means the individual comes from a family containing at least one member who is obese, and lean family means the individual comes from a family whose members are all lean. By analyzing such outcomes, we can examine how each array of microbial genes is associated with obesity both at the family and individual levels.

There are methods available for paired case–control studies, but they have limitations for analyzing the type of data considered here. For example, McNemar's test [11] is a statistical test used for paired data. However, it is primarily intended only for dichotomous features. Multilevel models with binary-dependent variables [1, 5] are another class of methods used. However, microbiome data is often high-dimensional, which makes it challenging to implement these approaches. Another challenge is that these models assume linearity but the association between obesity and microbiome features is likely to be nonlinear.

M. Lu · H. Ishwaran (✉)
Division of Biostatistics, University of Miami, Coral Gables, FL, USA

© Springer Nature Switzerland AG 2021

S. Datta, S. Guha (eds.), *Statistical Analysis of Microbiome Data*,
Frontiers in Probability and the Statistical Sciences,
https://doi.org/10.1007/978-3-030-73351-3_12

To illustrate our proposed methodology, we will use data from a cross-sectional study focusing on obesity in twins [9, 16, 17]. Data was collected from human stools of adult female monozygotic or dyzygotic twins or their mothers. We utilize 142 of these samples with 54 pairs where pair is defined as family members including mother and daughters. The bacterial lineages present in the fecal microbiota of these individuals were characterized by 16S rRNA sequencing, targeting the full-length gene with an ABI 3730xl capillary sequencer. Sequences were identified by assignment to taxonomic groups using operational taxonomic units (OTUs). Specific details of how data was processed can be found in [16].

The original analysis found that obesity is associated with phylum-level changes in the microbiota and reduced bacterial diversity using linear approaches, such as PCA (principal components analysis). Here, we will focus on detecting which taxonomic groups are the most informative for obesity risk at both the family and individual levels using a novel approach that draws upon tree-based concepts.

2 Gini Index

Consider Y a categorical (factor) outcome such that $Y \in \{1, \ldots, J\}$ for $J \geq 2$. Given a p-dimensional feature $\mathbf{X}$, the goal is to classify $\mathbf{X}$ into one of the J classes. We call this J-class problem and $\{1, \ldots, J\}$ the J-class labels for Y. Gini index is widely used for constructing classification trees that are nonparametric estimators used for the J-class problem. Given an input feature, classification trees work by identifying the unique leaf (terminal node) of the tree that $\mathbf{X}$ resides within. Each leaf of the tree is labeled with a class among $\{1, \ldots, J\}$ or a probability distribution over the classes, signifying that the leaf has been classified into either a specific class or a particular probability distribution, and this information is used for making decisions about $\mathbf{X}$.

A classification tree is built by splitting the data, constituting the root node of the tree, into subsets that constitute the successor children. The splitting is applied to features and is based on a pre-chosen splitting rule, which in addition to the Gini index includes splitting methods utilizing AUC (area under the ROC curve) and entropy metrics [3, 12]. This process is repeated on each derived subset in a recursive manner. The recursion is stopped when the subset at a node has all the same values of the outcome, or when a prespecified criterion is reached (such as minimal size of a node). The final nodes of the tree are referred to as the leaves. Classification trees can be combined to form ensemble estimators. An example is random forests [2], a popular tree-based learning method capable of handling a large number of predictors. In order to handle big data, rather than using a classification tree as just described, random forests is constructed by using random trees where each tree is constructed from subsampled data and where tree splitting employs random feature selection [2].

Classification tree splitting based on the Gini index splitting rule can be formally described as follows. If $\mathbf{p} = (p_1, \ldots, p_J)$ are the data class proportions of Y for classes 1 through J, respectively, the Gini index of impurity is defined as

$$\phi(\mathbf{p}) = \sum_{j=1}^{J} p_j(1 - p_j) = 1 - \sum_{j=1}^{J} p_j^2.$$

As mentioned, classification trees are grown using the Gini index by splitting features recursively into left and right daughter nodes. In particular, tree splits are obtained by minimizing tree impurity. The Gini index split statistic for a split s on a continuous feature x_m at a given tree node is

$$\theta(Y, x_m, s) = \frac{n_l}{n}\phi(\mathbf{p}_l) + \frac{n_r}{n}\phi(\mathbf{p}_r),$$

where the subscripts $l = \{x_m \leq s\}$ and $r = \{x_m > s\}$ denote the left and right daughter nodes formed by the split on x_m at s (n_l and n_r are the sample sizes of the two daughter nodes where $n = n_l + n_r$ is the parent sample size). To reduce tree impurity, the goal is to find x_m and s to *minimize*

$$\theta(Y, x_m, s) = \frac{n_l}{n}\left(1 - \sum_{j=1}^{J} \frac{n_{j,l}^2}{n_l^2}\right) + \frac{n_r}{n}\left(1 - \sum_{j=1}^{J} \frac{n_{j,r}^2}{n_r^2}\right),$$

where $n_{j,l}$ and $n_{j,r}$ are the number of cases of class j in the left and right daughters, respectively, and $n_j = n_{j,l} + n_{j,r}$ are the number of cases of class j and $n = \sum_{j=1}^{J} n_j$. With some algebra, it can be shown this is equivalent to *maximizing* the split statistic

$$g(Y, x_m, s) = \frac{1}{n}\sum_{j=1}^{J} \frac{n_{j,l}^2}{n_l} + \frac{1}{n}\sum_{j=1}^{J} \frac{(n_j - n_{j,l})^2}{n - n_l}.$$

Although the Gini index is primarily used as a splitting rule, we observe that it can be used as a fast preliminary variable ranking method. This is because variables that are used to split a tree are often those variables that have highest variable importance as measured by prediction error, especially if these splits occur high up in the tree, i.e., near the root node that comprised all the data [8]. Thus, it is reasonable to rank variables in terms of size of their Gini index values calculated using the full data as this will generally rank variables by predictive power. For each of the p predictors $x_1, \ldots, x_p$, define

$$G(Y, x_m) = g(Y, x_m, s_{\max}),$$

where

$$s_{\max} = \arg\max_s g(Y, x_m, s)$$

and $g(Y, x_m, s)$ is the split statistic calculated from the root node data (i.e., using the full data; thus, n is the sample size). Variables are ranked in order of importance

by size of $G(Y, x_m)$. This variable selection procedure is fully nonparametric and can be computed quickly even in big data settings. The following section provides a demonstration of how this approach works for our problem.

2.1 Simulation Analysis

Consider a binary class setting and denote the outcome as $Y^I \in \{0, 1\}$, where $Y^I = 0$ represents a lean individual and $Y^I = 1$ an obese individual. Family outcome is denoted as $Y^F \in \{0, 1\}$, where $Y^F = 0$ signifies an individual from a family with all lean members and $Y^F = 1$ indicates an individual from a family where at least one member is obese. Association with $Y^I = 1$ reflects how host adiposity influences the gut microbiome, whereas association with $Y^F = 1$ reflects environmental exposure influences. How the host genotype affects the gut microbiome under environmental exposure is reflected by an association with both $Y^I = 1$ and $Y^F = 1$.

We use the following simulation where Y^F is specified according to

$$\mathbb{P}\{Y^F = 1|\mathbf{X} = \mathbf{x}\} = \text{logistic}\big(-2 + x_1 + x_2 + x_3 + 2 \times \mathbf{1}_{\{x_1<0.5\}}\big) \tag{1}$$

and Y^I is specified by

$$\mathbb{P}\{Y^I = 1|Y^F = 1, \mathbf{X} = \mathbf{x}\} = \text{logistic}\big(-2 + x_4 + x_5 + x_6 + 2 \times \mathbf{1}_{\{x_4<0.5\}}\big), \tag{2}$$

where $\text{logistic}(\alpha) = 1/(1 + e^{-\alpha})$. In this scenario, x_1, x_2, and x_3 are associated with environmental exposures that cause the presence of obesity, while x_4, x_5, and x_6 are associated with host adiposity, given that the host is under these types of environmental exposures.

The feature space dimension was set to $p = 10$. Features were independently drawn from a uniform distribution U(0, 1). Variables unrelated to outcome, representing noise variables, were also added to the design matrix. For Y^F, noise variables were $x_4, \ldots, x_{10}$. For Y^I, noise variables were x_1, x_2, x_3 and $x_7, \ldots, x_{10}$. Split statistics, $g(Y, x_m, s)$, are plotted in Fig. 1 for features x_1, x_4, and x_{10} and for both outcomes $Y = Y^F$ and $Y = Y^I$. Red color represents the family-level outcome Y^F, and blue is used for the individual-level outcome Y^I. Variable x_1 in (a) predicts obesity at the family level and is associated with Y^F, and the true optimal split point occurs at 0.5. We can see that the split statistic of x_1 is high for both Y^F and Y^I and both peak at around 0.5. Variable x_4 in (b) is associated with $\mathbb{P}\{Y^I = 1|Y^F = 1\}$, therefore is associated with Y^I, and has a true optimal split point of 0.5. We can see that the split statistic $g(Y^I, x_4, s)$ is high for Y^I and reaches its peak near 0.5 (although not exactly at the true value—we will come back to this point later). In contrast, the split statistic $g(Y^F, x_4, s)$ for Y^F does not at all have an optimized value near 0.5 and its peak value occurs near its edge. This edge effect is typical of noisy variables and is a property of the Gini splitting rule called end-cut preference, ECP [6]. Variable x_{10} in (c) is a noise variable, and its split statistic is low for

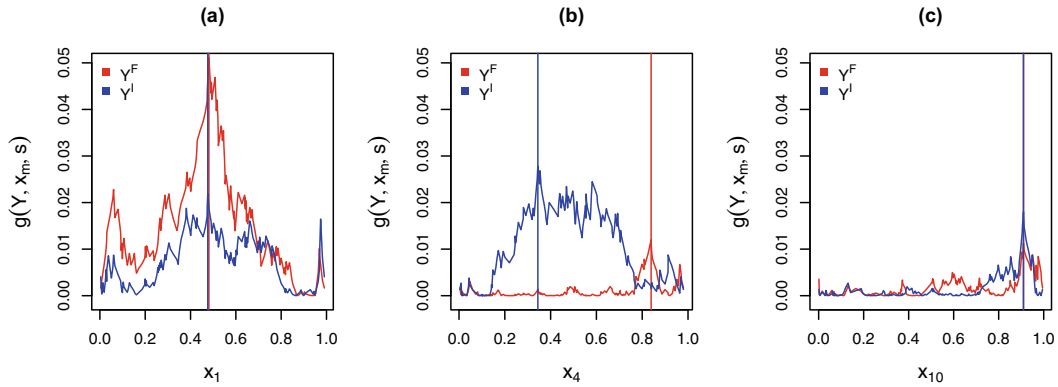

Fig. 1 Univariate split statistics for x_1, x_4, and x_{10} from simulation (1)–(2). Values $g(Y, x_m, s)$ are shown across different split values s. Red and blue display family-level outcome Y^F and individual-level outcome Y^I, respectively. Vertical lines mark the optimal split statistic $G(Y, x_m)$. Variable x_1 is associated with Y^F with true optimal split point of 0.5. Variable x_4 is associated with $\mathbb{P}\{Y^I = 1|Y^F = 1\}$ with true optimal split point of 0.5. Variable x_{10} is a noise variable

both Y^F and Y^I. Observe that its optimal split points are close to the edge for both outcomes, which as stated is typical behavior of a noisy variable.

Comparing the results across Fig. 1, it is clear that $G(Y, x_m)$, which is the highest point of $g(Y, x_m, s)$, is useful for variable ranking. However, focusing only on family-level outcomes (red color) will ignore features like x_4 that are related to the individual-level outcome (blue color). Checking both split statistics clearly helps better understand the underlying associations.

3 Multivariate Gini Index

Tang and Ishwaran [13] defined a multivariate Gini index split statistic obtained by averaging univariate Gini split statistics. For the bivariate outcome problem, this can be described as

$$\overline{g}_u(Y^F, Y^I, x_m, s) = \frac{1}{2}\big[g(Y^F, x_m, s) + g(Y^I, x_m, s)\big].$$

The subscript "u" is used to emphasize that the split statistic is unweighted. We can define

$$\overline{G}_u(Y^F, Y^I, x_m) = \overline{g}_u(Y^F, Y^I, x_m, s_{u_{\max}})$$

for ranking variables, where

$$s_{u_{\max}} = \arg\max_s \overline{g}_u(Y^F, Y^I, x_m, s).$$

Larger values of $\overline{G}_u(Y^F, Y^I, x_m)$ identify informative variables and smaller values indicate noise variables.

3.1 Conditional Gini Index

The problem with the split statistic $\overline{g}_u(Y^F, Y^I, x_m, s)$ is that by averaging across the outcomes it ignores the correlation between Y^F and Y^I. To resolve this issue, we introduce the following conditional Gini split statistic.

Let $\pi_c = \mathbb{P}\{Y^I = 1 | Y^F = 1\}$ be the population proportion of obese cases among individuals with at least one obese family member. The subscript "c" is used to emphasize this is a conditional probability. Because there are only two classes, we have $\mathbf{p}_c = (p_c, 1 - p_c)$ and $\phi(\mathbf{p}_c) = 2p_c(1 - p_c)$, where p_c is the sample estimator of π_c. For a split s on variable x_m, the conditional Gini split statistic is defined as

$$\theta_c(Y^F, Y^I, x_m, s) = \frac{\tilde{n}_l}{\tilde{n}}\phi(\mathbf{p}_c) + \frac{\tilde{n}_r}{\tilde{n}}\phi(\mathbf{p}_c),$$

where as before subscripts l and r denote the left and right daughter nodes formed by the split. The numbers of cases $Y^F = 1$ in the daughters are $\tilde{n}_l$ and $\tilde{n}_r$, where $\tilde{n} = \tilde{n}_l + \tilde{n}_r$. The numbers of these cases where $Y^I = 1$ in the left and right daughters are denoted by $\tilde{n}_{1,l}$ and $\tilde{n}_{1,r}$ respectively. It can be shown that minimizing $\theta_c(Y^F, Y^I, x_m, s)$ is equivalent to maximizing

$$g_c(Y^F, Y^I, x_m, s) = \frac{\tilde{n}_{1,l}^2}{\tilde{n}\tilde{n}_l} + \frac{\tilde{n}_{1,r}^2}{\tilde{n}\tilde{n}_r}.$$

We can define

$$G_c(Y^F, Y^I, x_m) = g_c(Y^F, Y^I, x_m, s_{c_{\max}})$$

for ranking variables, where $s_{c_{\max}} = \arg\max_s g_c$.

Now because $g_c(Y^F, Y^I, x_m, s)$ conditions on $Y^F = 1$, it is not designed to identify signal affecting Y^F. To resolve this, define the conditional weighted split statistic

$$\overline{g}_{cw}(Y^F, Y^I, x_m, s) = \frac{1}{w_F + w_I}\big[w_F \cdot g(Y^F, x_m, s) + w_I \cdot g_c(Y^F, Y^I, x_m, s)\big]$$

for detecting features that affect both Y^F and Y^I. Observe that when $w_F = w_I = 1$, this becomes an unweighted split statistic and will be denoted by $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$.

Weighted indices can be calculated as $w_F = \sum_i^n \mathbf{1}_{\{Y_i^F=1\}}$ and $w_I = \sum_i^n \mathbf{1}_{\{Y_i^I=1\}}$, which adjust for the fact that there are always more obese cases for Y^F than Y^I. The maximum value for the conditional weighted split statistic is

$$\overline{G}_{cw}(Y^F, Y^I, x_m) = \overline{g}_{cw}(Y^F, Y^I, x_m, s_{cw_{\max}}),$$

where $s_{cw_{\max}} = \arg\max_s \overline{g}_{cw}$. In a likewise fashion, define the maximum conditional unweighted split statistic $\overline{G}_{cu}(Y^F, Y^I, x_m)$.

Figure 2 displays: (a) $g_c(Y^F, Y^I, x_m, s)$, (b) $\overline{g}_u(Y^F, Y^I, x_m, s)$, (c) $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$, and (d) $\overline{g}_{cw}(Y^F, Y^I, x_m, s)$ for variables x_1, x_4, and x_{10} from the simulation (1)–(2). Variable x_4 affects the conditional probability $\mathbb{P}(Y^I = 1|Y^F = 1)$, which is plotted in purple color. Returning to the point made earlier regarding Fig. 1b, when comparing Fig. 2a to Fig. 1b, we find $g_c(Y^F, Y^I, x_4, s)$ characterizes x_4 better than $g(Y^I, x_4, s)$ as the maximum value is closer to the true splitting point 0.5. Another point to observe is that the goal of $\overline{g}_u(Y^F, Y^I, x_m, s)$ and $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$ is to detect features associated with Y^F and/or Y^I. However, $\overline{g}_u(Y^F, Y^I, x_m, c)$ in (b) is less effective than $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$ in (c) because it ranks x_4 similarly to noise variable x_{10} (shown in orange). In contrast, $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$ in (c) and the weighted $\overline{g}_{cw}(Y^F, Y^I, x_m, s)$ in (d) properly rank x_4 as more informative than x_{10}. In fact, the weighted split statistic tends to do an even better job.

Figure 3 displays maximum Gini split statistics for all $p = 10$ variables averaged over 100 independent replications. For convenient calibration, the averaged split statistic for the noise variable x_7 is used as a selection cutoff. When comparing subfigures (c) with (b), we see that $G_c(Y^F, Y^I, x_m)$ performs better in terms of selecting the true signals, x_4, x_5, and x_6, than $G(Y^I, x_m)$. When comparing subfigures (f) with (d), we observe that the weighted Gini split statistic utilizing the conditional Gini index, $\overline{G}_{cw}(Y^F, Y^I, x_m)$, outperforms the simple averaged Gini split statistic, $\overline{G}_u(Y^F, Y^I, x_m)$, in selecting the true signal variables $x_1, \ldots, x_6$ (in (d) the informative variable x_6 is not selected, whereas the noise variable x_{10} is selected). The performances of $\overline{G}_{cu}(Y^F, Y^I, x_m)$ and $\overline{G}_{cw}(Y^F, Y^I, x_m)$ are roughly similar except that noise variable x_{10} is less likely to be chosen using $\overline{G}_{cw}(Y^F, Y^I, x_m)$. Thus as before, the weighted split statistic tends to do a better job. Finally, when comparing subpanel (f) to (a) notice that $\overline{G}_{cw}(Y^F, Y^I, x_m)$ is as good as $G(Y^F, x_m)$ in identifying variables x_1, x_2, x_3 related to Y^F. However, this does not mean $G(Y^F, x_m)$ is not useful, since when combined with $\overline{G}_{cw}(Y^F, Y^I, x_m)$ it allows one to detangle variable relationships with the two outcomes.

4 Variable Importance

Another effective tool for variable selection is variable importance (VIMP). The permutation VIMP for a variable x_m is the prediction error for the model sub-

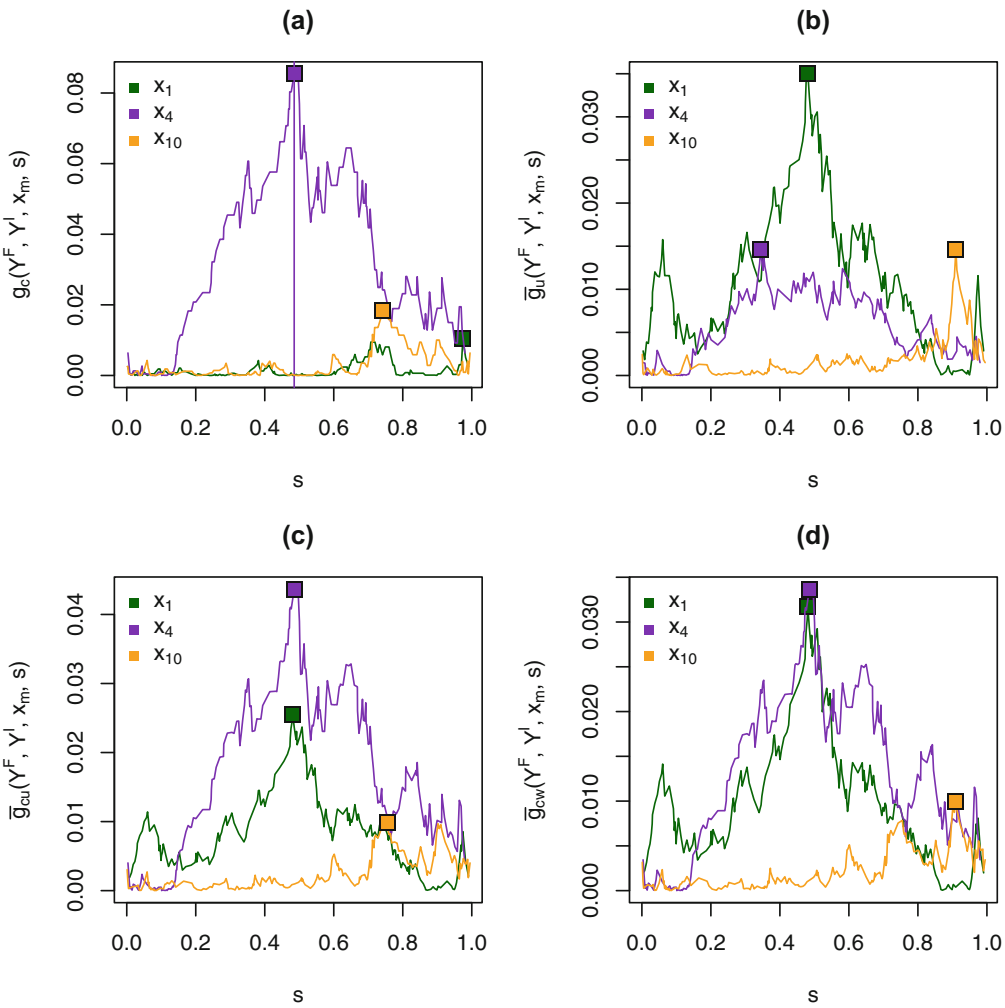

Fig. 2 Multivariate split statistics for x_1, x_4, and x_{10} from simulation (1)–(2). Curves displayed are: (**a**) $g_c(Y^F, Y^I, x_m, s)$, (**b**) $\overline{g}_u(Y^F, Y^I, x_m, s)$, (**c**) $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$, and (**d**) $\overline{g}_{cw}(Y^F, Y^I, x_m, s)$ with maximum statistic marked by a square point

tracted from the prediction error for the model using data that randomly permutes x_m [7]. This procedure can be implemented over independent bootstrap samples and the value averaged to obtain a more stable estimator [7]. More formally, let $\hat{\mathrm{PE}}(Y)$ be the averaged out-of-sample (called out-of-bag and abbreviated as OOB) misclassification error for the original model. Let $\hat{\mathrm{PE}}(Y, x_m^*)$ be the averaged OOB misclassification error when x_m is randomly permuted. The VIMP for x_m is

$$I(Y, x_m) = \hat{\mathrm{PE}}(Y, x_m^*) - \hat{\mathrm{PE}}(Y).$$

To determine if variables affect the conditional probability $\mathbb{P}(Y^I = 1 | Y^F = 1)$, we define a conditional VIMP analogous to the conditional Gini index. Conditional VIMP is calculated by restricting the data to those cases where $Y^F = 1$. The conditional VIMP index for x_m is

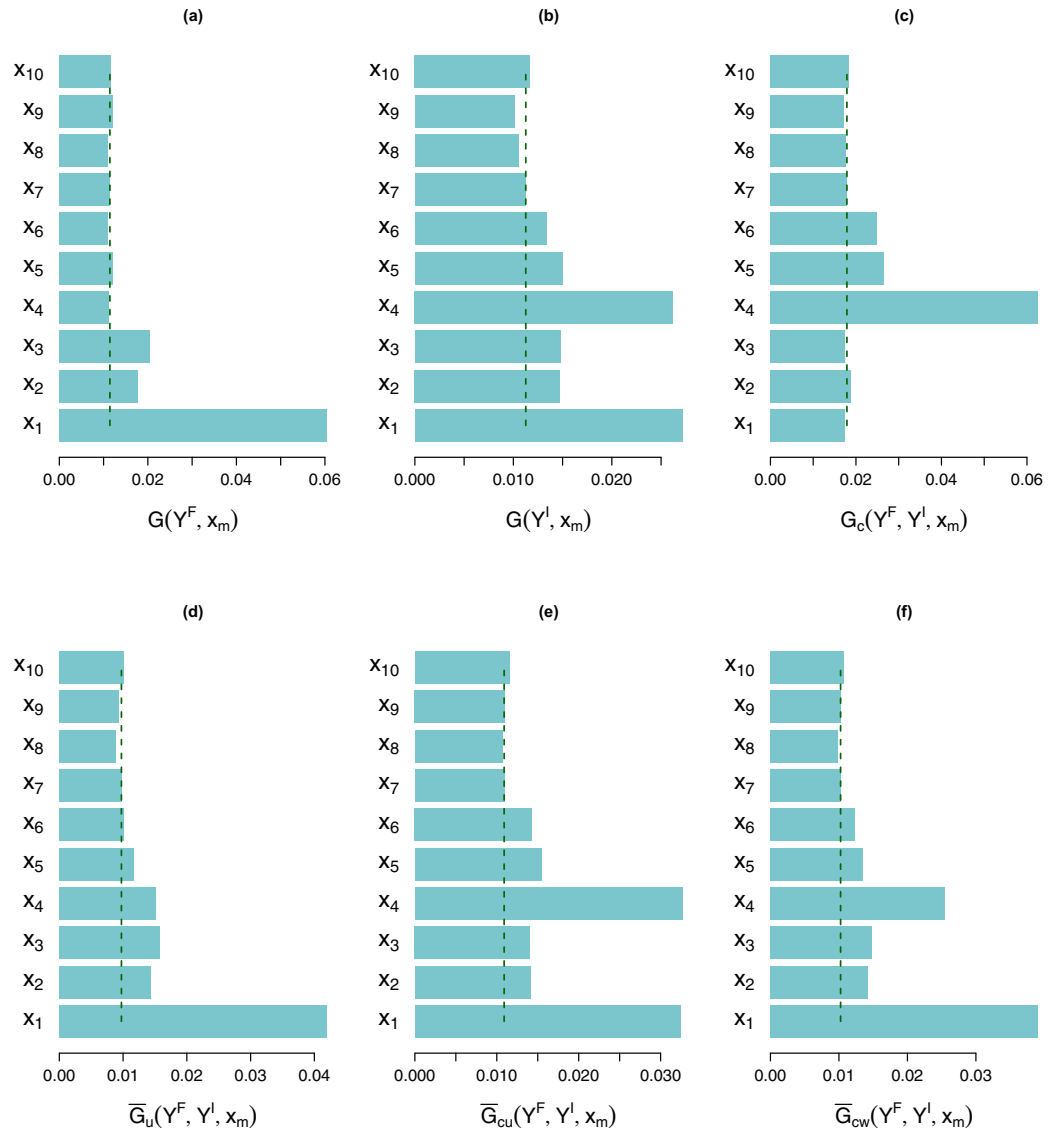

Fig. 3 Variable ranking from maximum split statistics for simulation (1)–(2) repeated 100 times independently. Dashed line is the averaged value of maximum Gini split statistic for noise variable x_7 that represents a convenient cut-off value

$$I_c(Y^I, x_m) = \hat{\mathrm{PE}}_c(Y^I, x_m^*) - \hat{\mathrm{PE}}_c(Y^I).$$

Figure 4 displays VIMP for all p features for our simulation. Values have been averaged over 100 independent replications. Unconditional VIMP, $I(Y^F, x_m)$, for Y^F displayed in subpanel (a) successfully ranks the true signal variables x_1, x_2 and x_3 as the most informative. When comparing subpanel (c) to (b), we see that conditional VIMP, $I_c(Y^I, x_m)$, is better at selecting true signal variables x_4, x_5, and x_6 than unconditional VIMP, $I(Y^I, x_m)$. In subfigure (b), VIMP for x_1 is very large and would lead to incorrect selection compared with (c).

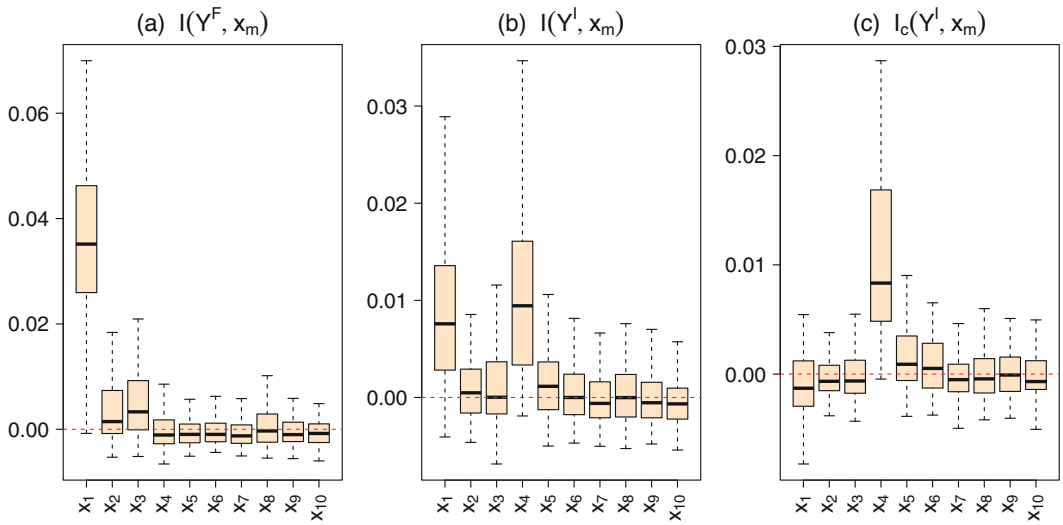

Fig. 4 Variable importance from simulation (1)–(2) averaged over 100 independent replications

5 Analysis of Obesity Using Microbiome Data

Now we return to the microbiome obesity data described earlier ($n = 142$ and $p = 174$). Outcomes were coded as before: $Y^I = 0$ represents a lean individual, $Y^I = 1$ an obese individual, $Y^F = 0$ signifies an individual from a family with all lean members, and $Y^F = 1$ indicates an individual from a family where at least one member is obese. Table 1 of the Appendix provides convenient abbreviated names for features. The features were originally coded using names of kingdom, phylum, class, order, family, genus, and species separated with a dash line and capital letter [9]. In order to make the name shorter for the figures, we use Table 1 to shorten the name for some classification labels. However to avoid duplication, we did not shorten all names. Therefore, some x-labels in Fig. 5 use full classification names.

Figure 5 displays split statistics for 6 representative features, chosen to illustrate how host and environmental factors affect the gut microbiome. Univariate split statistics $g(Y^F, x_m, s)$ for the family outcome Y^F are shown using red lines, and conditional split statistics $g_c(Y^F, Y^I, x_m, s)$ are displayed using orange lines. Bivariate split statistics, $\overline{g}_{cu}(Y^F, Y^I, x_m, s)$ and $\overline{g}_{cw}(Y^F, Y^I, x_m, s)$, lie between these two lines. Recall when optimal split points appear toward the edge of feature's range that this is a sign of a noisy feature (referred to as the ECP property [6]).

Subfigures (a), (b), (c) represent features informative for the environmental outcome Y^F. In all three figures, $g(Y^F, x_m, s)$ takes large values across the range of feature values. However, these three features are not informative for $\mathbb{P}\{Y^I = 1 | Y^F = 1\}$ as $g_c(Y^F, Y^I, x_m, s)$ is near zero in all instances. Thus, they do not reflect how host adiposity influences the gut microbiome under environmental exposure.

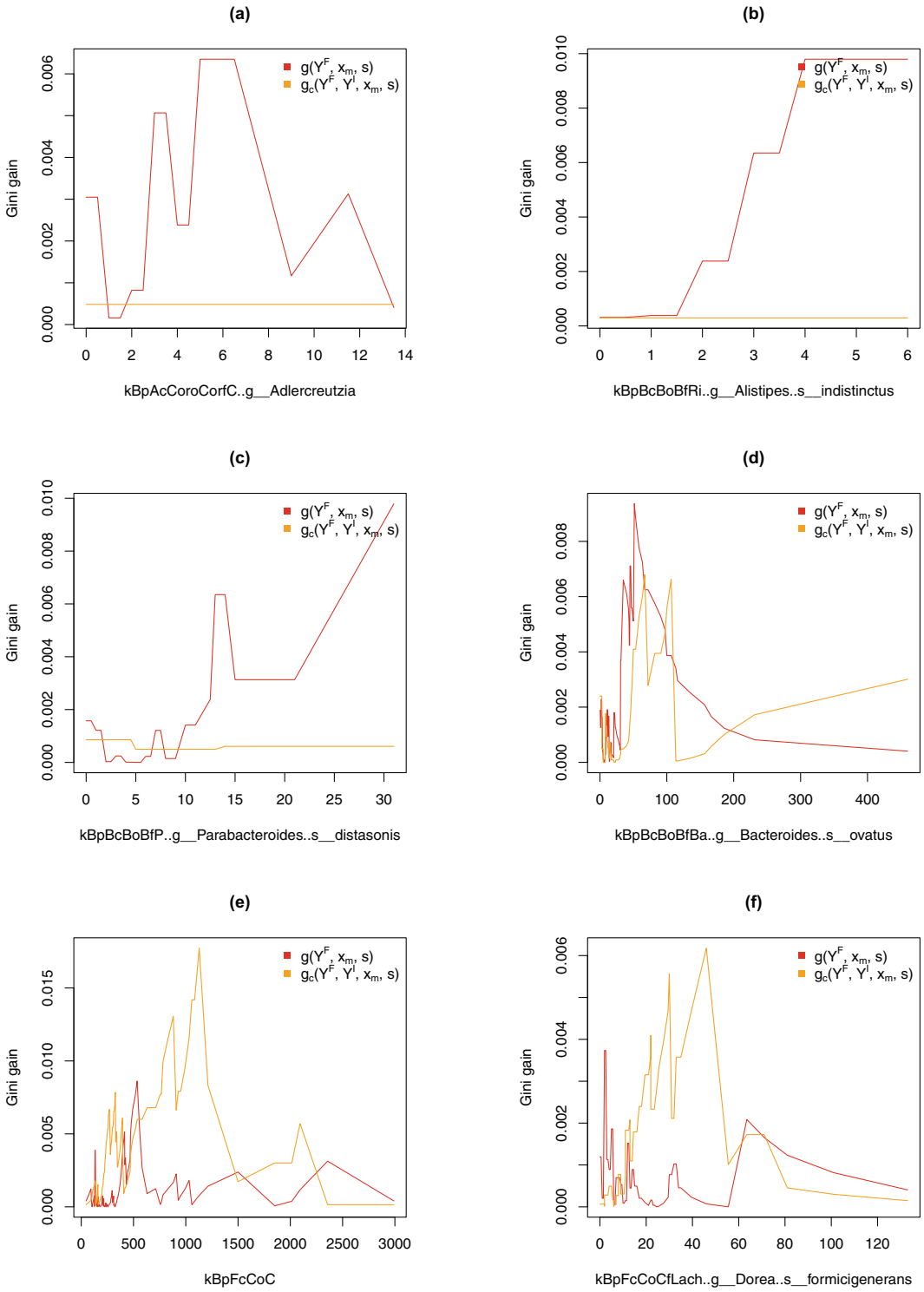

Fig. 5 Split statistics for microbiome obesity data. Shown are 6 representative variables illustrating how taxonomic groups predict obesity risk at the family level (shown using the univariate Gini split statistic on Y^F, $g(Y^F, x_m, c)$, plotted in red) and at the individual level (shown using the univariate conditional Gini split statistic $g_c(Y^F, Y^I, x_m, c)$, plotted in orange). Variable names are abbreviated according to Table 1

Subfigures (d) and (e) represent features that are informative for both $Y^F = 1$ and $\mathbb{P}\{Y^I = 1|Y^F = 1\}$ as both $g(Y^F, x_m, s)$ and $g_c(Y^F, Y^I, x_m, s)$ assume relatively large values. These features identify influences from both environmental exposure and host adiposity. For (d), the two maximum split statistics are nearly the same, which suggests that effect of environmental exposure and host adiposity is roughly the same for this feature. For (e), $g_c(Y^F, Y^I, x_m, s)$ attains a much larger maximum statistic than $g(Y^F, x_m, s)$ at a higher feature value. This suggest the effect of environmental exposure and host adiposity depends on the feature value, for example, whether the feature value is larger than 500 or 1000.

Subfigure (f) is a feature that mainly reflects the influence from host adiposity, rather than environmental exposure. This is because values of $g(Y^F, x_m, s)$ are overall small and its optimal split point is close to the edge of its range, signaling that it is likely a noisy variable for Y^F.

The values of $G(Y^F, x_m)$ and $G_c(Y^F, Y^I, x_m)$ are given in Fig. 6. The sizes of circles are scaled proportional to $\overline{G}_{cw}(Y^F, Y^I, x_m)$. Phylum groups are used to color circles. It is interesting to note that features informative for $Y^F = 1$ and $\mathbb{P}\{Y^I = 1|Y^F = 1\}$ belong primarily to the Fusobacteria phylum. Generally, the values of $G_c(Y^F, Y^I, x_m)$ are smaller than $G(Y^F, x_m)$. However, when they are

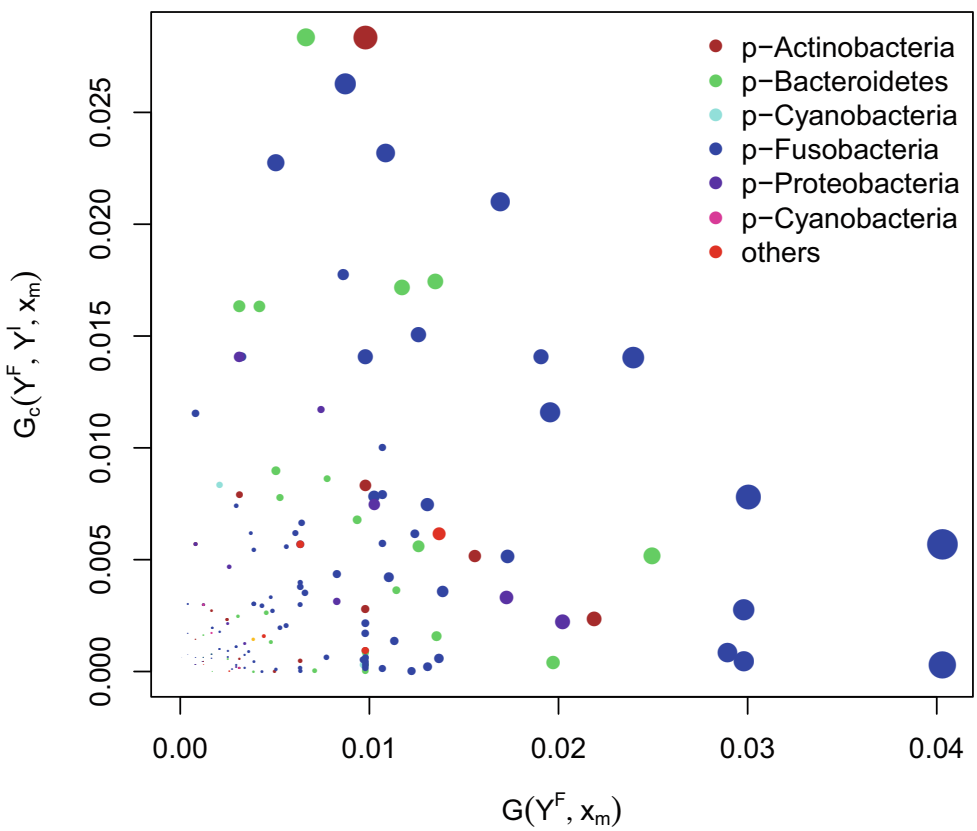

Fig. 6 Comparison between $G(Y^F, x_m)$ and $G_{Y^F=1}(Y^I, x_m)$. The sizes of circles are proportional to $\overline{G}_{cw}(Y^F, Y^I, x_m)$. The color of circles identifies the phylum

weighted to obtain $\overline{G}_{cw}(Y^F, Y^I, x_m)$, we can see that there is a nice balancing of values.

Finally, Fig. 7 displays unconditional VIMP, $I(Y^F, x_m)$, and conditional VIMP, $I_c(Y^I, x_m)$, for all p variables. Many variables have small or negative values, thus showing that VIMP can be used as an effective means to dimension. Note that due to invariance of trees under monotonic transformations of features, split statistics and VIMP are invariant to the magnitude of the feature values. For split statistics, it is the quantile of the split that makes a difference. Thus, we did not normalize independent variables, which were OTU count values [9].

Our results are consistent with previous studies. Turnbaugh et al. [15, 16] found that obesity is associated with phylum-level changes in the microbiota. We found the same as displayed in Fig. 6. Moreover, our findings (see Fig. 5) have shown that changes occur differently at the family and individual levels. A similar finding was observed in a study of high-fat diet-induced obesity and diabetes in mice [4]. This study found a high-fat diet-altered proportion of Bacteroides-related bacteria and reduced Bifidobacteria (Gram-positive, phylum Actinobacteria).

We carefully note that our analysis can only establish an association between microbiome and obesity [14] and not causation. Casual inference is far more demanding, and further studies would be needed to be able to move beyond associative analyses [10].

6 Discussion

Fast nonparametric selection of features that accounts for correlation in paired data is a valuable tool for microbiome data analysis. Variable selection procedures can choose features that reflect influences from external effects (between pairs) and internal effects (within pairs), but without taking into account the paired structure of the data, they will be inefficient in separating the two types of effects. Our proposed conditional Gini split statistic, when used alone or averaged with univariate Gini split statistics, serves two purposes. First, the maximum value of the split statistic can be used for variable ranking and variable selection. Conditional Gini is able to select variables reflecting how the microbiome is affected by host adiposity given the same environmental exposures. Second, how the value of the split statistic varies within a feature provides useful insight into the magnitude of the external and/or internal effects. The optimal split point for conditional Gini represents the threshold that a feature can separate lean and obese individuals given the same environmental exposure. We demonstrated these two aspects in a systematic comparative simulation and through a real data application. We found that the paired structure of the data played a very strong role in performance of our methods. Without controlling for family level of obesity, features that only affect individual level of obesity are often noticeably masked.

There are other variable selection procedures designed for multivariate outcomes. However, in big data settings, computational speed plays a key role. Practically speaking, the best method is not always optimal for the researcher because computational times can be too long. Our Gini split statistics can be rapidly computed for a large number of features in big data settings, and because the calculations are univariate, the procedure could be parallelized to further reduce runtimes. Users can simulate a noise feature to determine the cutoff for screening noise variables. Potentially, our Gini indices can be used as tree splitting rules so that all the features can be taken into consideration together. Moreover, our approach could leverage powerful machine learning methods such as random forests and boosting to provide a direct approach to analyze paired data. Another potential improvement to our work would be to use additional data to study the effect of number of individuals on Y^F. In the analysis we used, all families have 2 or 3 individuals, which made it impossible for us to study the effect of number of individuals.

Acknowledgments This work was supported by the National Institutes of Health (grant numbers R01 CA200987 and R01 HL141892 to H.I.).

Appendix

See Table 1 and Fig. 7.

Table 1 Abbreviated feature names for microbiome obesity data

Abbrev.	Full form	Abbrev.	Full form
kB	k-Bacteria	oBi	..o-Bifidobacteriales
pF	..p-Firmicutes	oE	..o-Erysipelotrichales
pA	..p-Actinobacteria	oL	..o-Lactobacillales
pB	..p-Bacteroidetes	fB	..f-Bifidobacteriaceae
pC	..p-Cyanobacteria	fBa	..f-Bacteroidaceae
pF	..p-Fusobacteria	fC	..f-Coriobacteriaceae
pP	..p-Proteobacteria	fL	..f-Lactobacillaceae
pS	..p-Synergistetes	fLach	..f-Lachnospiraceae
cA	..c-Actinobacteria	fM	..f-Micrococcaceae
cB	..c-Bacteroidia	fP	..f-Porphyromonadaceae
cBci	..c-Bacilli	fPe	..f-Peptostreptococcaceae
cC	..c-Clostridia	fPr	..f-Prevotellaceae
cCor	..c-Coriobacteriia	fR	..f-Ruminococcaceae
cE	..c-Erysipelotrichi	fRi	..f-Rikenellaceae
oA	..o-Actinomycetales	fS	..f-Streptococcaceae
oC	..o-Clostridiales	fV	..f-Veillonellaceae
oCor	..o-Coriobacteriales	gB	..g-Bifidobacterium
oB	..o-Bacteroidales	gC	..g-Corynebacterium

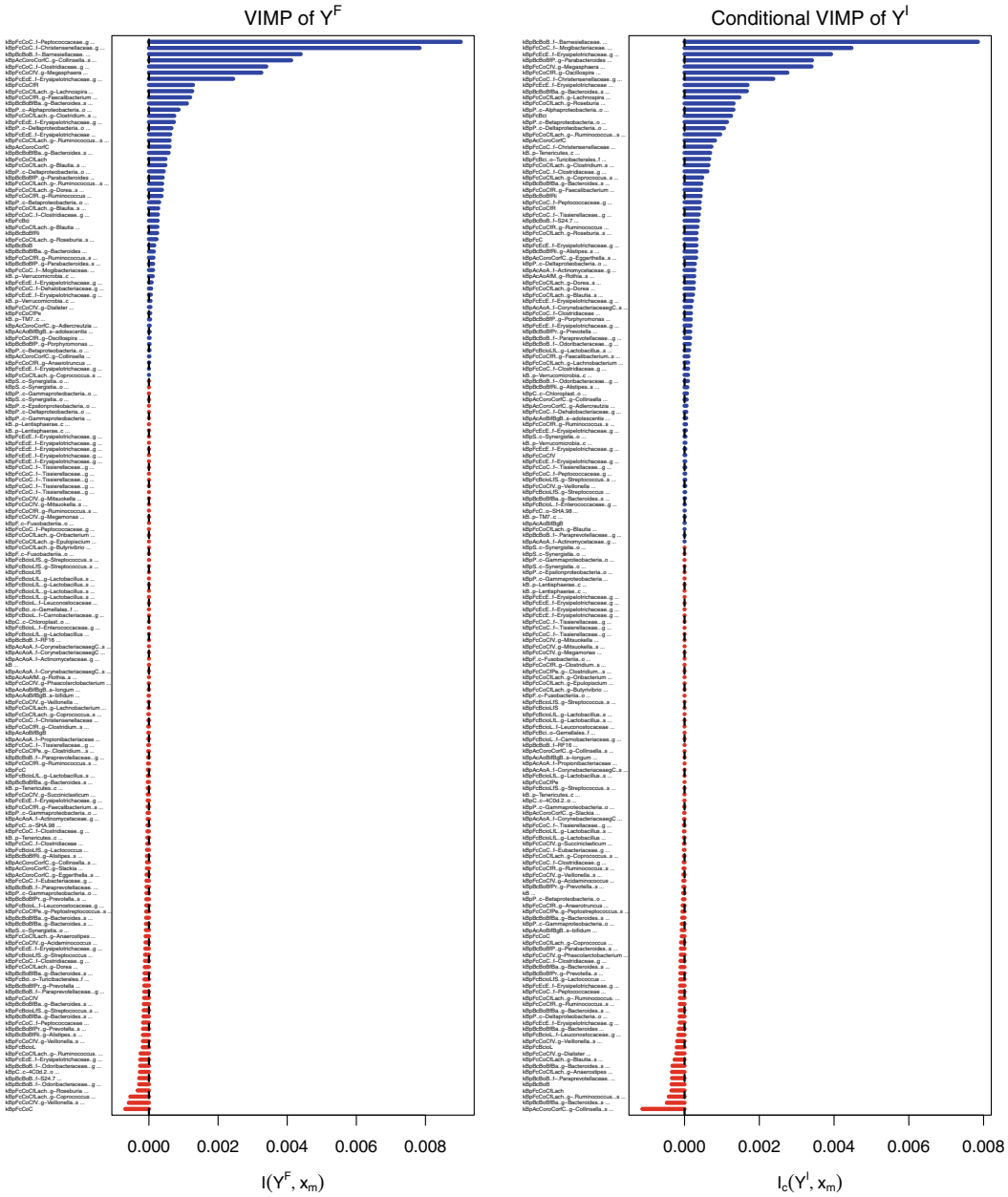

Fig. 7 Variable ranking using VIMP for microbiome data of obesity. Variables with higher value on the left reflect how the gut microbiome is influenced by environmental factors. Variables with higher values in the right reflect how gut microbiome is affected by host adiposity given the environmental exposures

References

1. Benedetti, A., Platt, R., Atherton, J.: Generalized linear mixed models for binary data: are matching results from penalized quasi-likelihood and numerical integration less biased? PLoS One **9**(1), e84601 (2014)
2. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)

3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Belmont, California (1984)
4. Cani, P.D., Bibiloni, R., Knauf, C., Waget, A., Neyrinck, A.M., Delzenne, N.M., Burcelin, R.: Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. Diabetes **57**(6), 1470–1481 (2008)
5. Capanu, M., Gönen, M., Begg, C.B.: An assessment of estimation methods for generalized linear mixed models with binary outcomes. Stat. Med. **32**(26), 4550–4566 (2013)
6. Ishwaran, H.: The effect of splitting on random forests. Machine Learning **99**, 75–118 (2015)
7. Ishwaran, H., Lu, M.: Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat. Med. **38**, 558–582 (2019)
8. Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S.: High-dimensional variable selection for survival data. J. Am. Stat. Assoc. **105**(489), 205–217 (2010)
9. Knights Lab: Monozygotic or dizygotic twin pairs concordant for BMI class, and their mothers (Project Turnbaugh 2009). https://github.com/knights-lab/MLRepo/tree/master/datasets/turnbaugh. Cited 28 Mar (2020)
10. Maruvada, P., Leone, V., Kaplan, L.M., Chang, E.B.: The human microbiome and obesity: moving beyond associations. Cell Host Microbe **22**(5), 589–599 (2017)
11. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
12. Shalev-Shwartz, S., Ben-David, S.: Chapter 18. Decision Trees. Understanding Machine Learning. Cambridge University Press, Cambridge (2014)
13. Tang, F., Ishwaran, H.: Random forest missing data algorithms. Stat. Anal. Data Min. ASA Data Sci. J. **10**(6), 363–377 (2017)
14. Tsai, F., Coyle, W.J.: The microbiome and obesity: is obesity linked to our gut flora? Curr. Gastroenterol. Rep. **11**(4), 307–313 (2009)
15. Turnbaugh, P.J., Gordon, J.I.: The core gut microbiome, energy balance and obesity. J. Physiol. **587**(17), 4153–4158 (2009)
16. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., ... Egholm, M.: A core gut microbiome in obese and lean twins. Nature **457**(7228), 480–484 (2009)
17. Vangay, P., Hillmann, B. M., Knights, D.: Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. GigaScience **8**(5), giz042 (2019)
18. Xia, Y., Sun, J., Chen, D.G.: Statistical Analysis of Microbiome Data with R. Springer, Singapore (2018)