TRUNCATED RANK-BASED TESTS FOR TWO-PART MODELS WITH EXCESSIVE ZEROS AND APPLICATIONS TO MICROBIOME DATA

BY WANJIE WANG1, ERIC CHEN2 AND HONGZHE LI2,*

1National University of Singapore, staww@nus.edu.sg
2University of Pennsylvania, *hongzhe@upenn.edu

High-throughput sequencing technology allows us to test the compositional difference of bacteria in different populations. One important feature of human microbiome data is that it often includes a large number of zeros. Such data can be treated as being generated from a two-part model that includes a zero point-mass. Motivated by analysis of such non-negative data with excessive zeros, we introduce several truncated rank-based two-group and multi-group tests, including a truncated rank-based Wilcoxon rank-sum test for two-group comparison and two truncated Kruskal-Wallis tests for multi-group comparisons. We show both analytically through asymptotic relative efficiency analysis and by simulations that the proposed tests have higher power than the standard rank-based tests in typical microbiome data settings, especially when the proportion of zeros in the data is high. The tests can also be applied to repeated measurements of compositional data via simple within-subject permutations. In a simple before-and-after treatment experiment, the within-subject permutation is similar to the paired rank test. However, the proposed tests handle the excessive zeros, which leads to a better power. We apply the tests to compare the microbiome compositions of healthy children and pediatric Crohn’s disease patients and to assess the treatment effects on microbiome compositions. We identify several bacterial genera that are missed by the standard rank-based tests.

1. Introduction. The human microbiome includes all microorganisms in various human body sites such as gut, skin and mouth. Gut microbiome has been shown to be associated with many human diseases, including obesity, diabetes and inflammatory bowel disease (Turnbaugh et al., 2006; Qin et al., 2012; Manichanh et al., 2012). Two high-throughput sequencing based approaches, including 16S ribosomal RNA (rRNA) sequencing and shotgun metagenomic sequencing, are commonly used in microbiome studies (Turnbaugh et al., 2007; Qin et al., 2010). Bioinformatics methods are available for quantifying the microbial relative abundances based on such sequencing data, which typically involve aligning the reads to some known database or marker genes (Huson et al., 2007; Segata et al., 2012). Since the DNA yielding materials are different across different samples, the resulting numbers of sequencing reads vary greatly from sample to sample. In order to make the microbial abundance comparable across samples, the abundance in read counts is usually normalized to the relative abundance of the bacteria observed, which results in high dimensional compositional data. Some of the most widely used metagenomic processing software such as MEGAN (Huson et al., 2007) and MetaPhlAn (Segata et al., 2012) only outputs the relative abundances of the bacterial taxa.

arXiv: math.PR/0000000

*This research was supported by NIH grants GM123056 and GM129781. Hongzhe Li is the corresponding author.

Keywords and phrases: Asymptotic relative efficiency; Differential abundance analysis; Two-part model; Truncation;
In microbiome studies, one is often interested in identifying the bacterial taxa such as genera or species that show different distributions between two or more conditions. One important feature of microbiome compositional data is that the data include large clumps of zeros that represent the absence of the bacterial taxa in the samples, especially for those relatively rare taxa. Zero observations can also result from under-sampling of the sequencing reads for rare taxa. As an example, Figure 1 shows the heatmap of zeros and the relative abundances for 26 healthy and normal samples and samples from 85 Crohn’s disease children for each of the 60 bacterial genera. These data were collected at the University of Pennsylvania (Lewis et al., 2015). We are interested in identifying the bacterial genera that have different distributions between healthy and Crohn’s disease patients. We observed over 62.5% of the observations are zero. In addition, the compositional data are often skewed, which makes parametric modeling of such data difficult and tests based on parametric distributional assumptions problematic. We present detailed analysis of this data set in Section 6.

Non-parametric tests, such as the Wilcoxon rank-sum test and Kruskal-Wallis test, can be applied to such data and are commonly used in analysis of microbiome data. However, such rank-based tests tend to have low power because of the large number of ties from zero observations (Lachenbruch, 1976, 2001; Hallstrom, 2010). A two-part test combining the square of a test statistic for comparison of the proportion of zeros and the square of an appropriate normal test such as the Wilcoxon rank-sum to compare the non-zero scores was proposed and evaluated by Lachenbruch (2001). This two-part test was recently applied to analysis of microbiome data (Wagner, Robertson and Harris, 2011). However, the theory developed for Lachenbruch’s 2 degree of freedoms $\chi^2$ test assumes that binomial test statistic and the test statistics for the continuous part to be independent, which only holds under the assumptions of independent errors of the binomial and continuous part of the distribution (Lachenbruch, 2002). However, such an assumption may not hold for the microbiome relative abundance data generated by sequencing since both can depend on the sequencing depth.

Fig 1: Heatmap of zeros and the relative abundances of 60 bacterial genera for 26 healthy children and 85 patients with pediatric Crohn’s disease, showing over 62.5% of the observations being zero (Lewis et al., 2015).
To account for excessive zeros in non-negative distributions, Hallstrom (2010) introduced a truncated Wilcoxon rank-sum test, where the Wilcoxon rank-sum test is performed after removing an equal and maximal number of zeros from each sample. He showed that this test recovers much of the power loss from the standard application of the Wilcoxon test. Compared with a directional modification of the two-part test proposed by Lachenbruch (Lachenbruch, 1976, 2001), the truncated Wilcoxon test has similar power when the non-zero relative abundances are independent of the proportion of zeros. In addition, the truncated Wilcoxon test is relatively unaffected when the error terms of the two distributions are dependent. Hallstrom (2010) however only considered the two-sample test under the setting of equal sample sizes.

In this paper, we assume that the data are generated from two-part models with point-mass at zero as one of the components. However, we do not make any distributional assumption on the continuous non-zero part. We develop several rank-based tests for general two-group comparison with possible unequal sample sizes, and for the multiple-group comparisons with equal or unequal sample sizes. Particularly, we extend the truncated Wilcoxon rank-sum test of Hallstrom (2010) to data with unequal sample sizes, and develop a modified Kruskal-Wallis test to account for clumps of zeros for multiple-group comparisons with equal sample sizes and unequal sample sizes, respectively. These new tests are based on the idea of data truncation and asymptotic calculations and can effectively deal with the clumps of zeros in the data. The asymptotic null distributions of the tests are given. The key difficulty of deriving such truncated rank-based tests is to calculate the variance of the test statistic under the null, which does not have a closed-form expression. We instead use asymptotic analysis to obtain approximations of the variance estimates.

In order to demonstrate the advantages of the proposed truncated rank-based tests, we also derive the asymptotic relative efficiency of the proposed tests compared to commonly used Wilcoxon rank-sum test and Kruskal-Wallis test when the data are generated from two-part models (Lachenbruch, 2001). We observe in our simulations large gains in efficiency, especially when the proportions of zeros in the data are high. These tests are rank-based, easy to calculate and provide new tools for identifying the bacterial taxa with different distributions among different groups in human microbiome studies. We apply and compare the proposed tests by analyzing a microbiome study conducted at the University of Pennsylvania (Lewis et al., 2015), including comparing the gut microbiome difference between healthy and pediatric Crohn’s disease patients and assessing the effects of treatment over time.

2. A truncated Wilcoxon rank-sum test for data with excessive zeros.

2.1. A truncated Wilcoxon rank-sum test. Consider the two-sample setting where we have \(N_1\) non-negative independent observations from population 1, \(X = (x(1), \cdots, x(N_1))\), and \(N_2\) independent observations from population 2, \(Y = (y(1), \cdots, y(N_2))\), where \(x(i)\) (or \(y(i)\)) represents the relative abundance of a bacterium in the \(i\)th sample of group 1 (or 2). For most of the bacteria, the data \(X\) and \(Y\) include many zeros, which represent absence of the bacterium in these samples or below detection limits. We are interested in testing whether these observations \(X\) and \(Y\) are from the same distribution, i.e., the hypothesis testing problem that

\[
H_0 : x \sim F, y \sim F \quad \text{vs} \quad H_1 : x \sim F, y \sim G, F \neq G,
\]

where \(F\) and \(G\) are both probability density functions with a proportion of zeros. We consider the two-part model of Lachenbruch (1976), which assumes that the data are generated from the following distributions,

\[
\begin{align*}
  x(i) &\sim (1 - \theta_1)\delta_0 + \theta_1 f, \\
  y(j) &\sim (1 - \theta_2)\delta_0 + \theta_2 g,
\end{align*}
\]

\(i = 1, \cdots, N_1, j = 1, \cdots, N_2,\)
where $\theta_1$ (or $\theta_2$) is the probability of being non-zero in population 1 (or 2), $\delta_0$ is point mass at 0, and $f$ (or $g$) is the distribution for nonzero element in the population 1 (or 2).

Because of the excessive zeros, the standard non-parametric Wilcoxon rank-sum test statistic is less effective. A truncated Wilcoxon rank-sum test statistic has been proposed by Hallstrom (2010) for the case $N_1 = N_2 = N$. We first extend his test to the general setting where $N_1 \neq N_2$, and examine the asymptotic relative efficiency compared to the standard Wilcoxon test statistic.

Given $X$ and $Y$, denote $n_1$ (or $n_2$) as the number of non-zero observations in $X$ (or $Y$) and let $p_1$ (or $p_2$) be the proportion of non-zero observations in $X$ (or $Y$), where $p_1 = n_1/N_1$ (or $p_2 = n_2/N_2$). Let $p = \max\{p_1, p_2\}$. Let $[a]$ denote the largest integer that is smaller than $a$. We rank the combined observations $X \cup Y$ from the largest to the smallest so that zeros have the highest ranks, where the tied measurements are given the average rank. For $X$ (or $Y$), we keep only $[pN_1]$ (or $[pN_2]$) observations with the smallest ranks, which implies that the observations removed are all zeros. The truncated samples are denoted as $\tilde{X}$ and $\tilde{Y}$, respectively. Let $R$ denote the sum of the ranks of all observations in $X$. The Wilcoxon test statistic can be written as

$$T_W = \frac{S^2}{\text{Var}[S]}, \quad \text{where } S = R - \frac{N_1 + N_2 + 1}{2}N_1,$$

and $\text{Var}[S]$ is the variance of $S$ under the null hypothesis.

The same procedure can be applied to the truncated data $\tilde{X}$ and $\tilde{Y}$. Rank the combined observations $\tilde{X} \cup \tilde{Y}$, from the largest to the smallest, and let $r$ denote the sum of ranks of all observations in $\tilde{X}$. We define the counterpart of $S$ as

$$s = r - \left[\frac{p(N_1 + N_2)}{2}\right] + 1\left[\frac{pN_1}{2}\right] - \frac{1}{4}\left(\frac{p_1 + p_2}{2}\right)^2 \left(1 - \frac{p_1 + p_2}{2}\right)(N_2 - N_1),$$

and define the truncated Wilcoxon test statistic as

$$T_{\tilde{W}} = \frac{s^2}{\text{Var}[s]}.$$

The statistic $s$ is very similar to $S$, except an extra term

$$\frac{1}{4}\left(\frac{p_1 + p_2}{2}\right)^2 \left(1 - \frac{p_1 + p_2}{2}\right)(N_2 - N_1),$$

which is caused by the difference between the variances due to different sample sizes. This term disappears when $N_1 = N_2$. Under the alternative hypothesis, this term is a small order term compared to the other part of $s$. Under the null hypothesis, this extra term is used to eliminate the effect of different sample sizes, leading to an expectation close to 0.

To calculate $\text{Var}[s]$ and to derive the asymptotic distribution of $s$, we show that under the null that both $x$ and $y$ follow the same distribution with a point mass at 0, when $N_1 \to \infty$, $N_2 \to \infty$, we have

$$\text{E}[S] = 0, \quad \text{E}[s] = O(\max\{N_1, N_2\}),$$

where the remainder $O(\max\{N_1, N_2\})$ is caused by the difference between $\left[\frac{pN_1}{2}\right]$ and $pN_1$ (or $\left[\frac{pN_2}{2}\right]$). In addition, we have

$$\text{Var}[S] = \frac{N_1^2N_2^2}{4}\text{E}\{(p_2 - p_1)^2\} + \frac{N_1N_2}{12}\text{E}(N_1^2p_1^2p_2 + N_2^2p_1p_2^2 + p_1p_2),$$

$$\text{Var}[s] = \frac{N_1^2N_2^2}{4}\text{Var}\{(p_2 - p_1)p\} + \frac{N_1N_2}{12}\text{E}(N_1^2p_1^2p_2 + N_2^2p_1p_2^2 + p_1p_2),$$

$$\text{Var}[\tilde{S}] = \frac{N_1^2N_2^2}{4}\text{Var}\{(p_2 - p_1)p\} + \frac{N_1N_2}{12}\text{E}(N_1^2p_1^2p_2 + N_2^2p_1p_2^2 + p_1p_2).$$
(see Supplemental Materials). With these results, under the null hypothesis, when \( N_1 \to \infty, N_2 \to \infty, \) and \( c \leq N_1 / N_2 \leq C \) for some constants \( c \) and \( C \), we have
\[
\operatorname{Var}[S] = N_1 N_2 (N_1 + N_2) \theta (1 - \theta + \theta^2 / 3) / 4 + O(N^{2.5}),
\]
\[
\operatorname{Var}[s] = N_1 N_2 (N_1 + N_2) \theta^3 (1 - \theta + 1 / 3) / 4 + O(N^{2.5}),
\]
where \( \theta \) is the expected value of the proportion of non-zero observations in the data. Hence, \( \sqrt{\operatorname{Var}[s]} = O(\sqrt{N_1 N_2 \max\{N_1, N_2\}}) \), and \( \operatorname{E}[s] = O(\max\{N_1, N_2\}) \) is relatively small when \( N_1 \) and \( N_2 \) are large. Asymptotically, \( s / \sqrt{\operatorname{Var}(s)} \) follows a standard normal distribution, and the truncated Wilcoxon test statistic \( T_{IW} = s^2 / \operatorname{Var}[s] \) has a \( \chi^2 \) null distribution.

The asymptotic analysis above provides a way of approximating \( \operatorname{Var}(s) \) using the main term of \( \operatorname{Var}(s) \), which leads to the final test statistic
\[
T_{IW} = \frac{s^2}{N_1 N_2 (N_1 + N_2)(p_1^2 + p_2^2)^{3/2}(4/3 - p_1 + p_2) / 4}.
\]
This is used in our simulation and real data analysis.

2.2. Pitman’s asymptotic relative efficiency of \( T_{IW} \) and \( T_{IW} \) for two-sample test. To compare the test statistics \( T_{IW} \) and \( T_{IW} \), we evaluate the asymptotic property of the relative efficiency. Since we reject the null hypothesis when the statistic is large, the Pitman’s relative efficiency is defined as
\[
\operatorname{ARE}(T_{IW}, T_{IW}) = \frac{\operatorname{E}_1[T_{IW}]}{\operatorname{E}_1[T_{IW}]} = \frac{\operatorname{E}_1[s^2 / \operatorname{Var}[s]]}{\operatorname{E}_1[s^2 / \operatorname{Var}[s]]}.
\]
Here, \( \operatorname{E}_1 \) denotes the expectation of the statistic under the alternative hypothesis. A value larger than one implies power gain using the truncated Wilcoxon test statistic compared to the standard Wilcoxon statistic.

For the two-part model (1), we need some terms to quantify the difference between the two distributions. Let \( \theta = (\theta_1 + \theta_2) / 2, \theta_m = \max\{\theta_1, \theta_2\} \), and \( \Delta \theta = \theta_2 - \theta_1 \). Define \( \Delta_{f,g} = P(x < y) + P(x = y) / 2 - 1 / 2, \) where \( x \sim f \) and \( y \sim g \). The term \( \Delta_{f,g} \) is used to measure the effect size of the non-zero part. The following theorem provides an explicit expression for the asymptotic relative efficiency (ARE).

**Theorem 1** Under model (1) and that \( N_1 \to \infty, N_2 \to \infty, \) and \( c \leq N_1 / N_2 \leq C \) holds for some constants \( c \) and \( C \), then the ARE can be derived as
\[
\operatorname{ARE}(T_{IW}, T_{IW}) = \frac{1}{\theta^2} \left( \frac{1 - \theta + \theta^2 / 3}{1 - \theta + 1 / 3} \left( \frac{\theta_1 \theta_2 \Delta_{f,g} + \Delta \theta \theta_m / 2}{\theta_1 \theta_2 \Delta_{f,g} + \Delta \theta / 2} \right)^2 \right)^2 + O(N^{-1/2}).
\]
Especially, we have that \( \operatorname{E}(s|n_1, n_2) = n_1 n_2 \Delta_{f,g} \).

To illustrate the gain in efficiency of using the truncated Wilcoxon test, we present the \( \operatorname{ARE}(W_m, W) \) for the following four different parameter settings to assess the effect of \( \Delta_{f,g}, \theta, \Delta \theta, \) and \( N_1 / N_2 \):

(a) Effect of \( \Delta_{f,g} \): \( N_1 = 40, N_2 = 50 \) and \( \theta_1 = 0.3, \theta_2 = 0.8 \). Let \( \Delta_{f,g} = -1 / 2 + 0.01 k, \) where \( k = 0, 1, 2, \ldots, 100 \).
(b) Effect of \( \theta \): \( N_1 = 40, N_2 = 50, \) and \( \Delta_{f,g} = 0.1 \). Let \( \theta = 0.1 + 0.01 k, \) where \( k = 0, 1, 2, \ldots, 80 \). With a given \( \theta \), take \( \theta_1 = \theta - 0.1 \) and \( \theta_2 = \theta + 0.1 \).
(c) Effect of \( \Delta \theta \): \( N_1 = 40, N_2 = 50 \) and \( \Delta_{f,g} = 0.1 \). Let \( \Delta \theta = 0.02 k, \) where \( k = 0, 1, 2, \ldots, 50 \). With a given \( \Delta \theta \), take \( \theta_1 = 0.5 - \Delta \theta / 2 \) and \( \theta_2 = 0.5 + \Delta \theta / 2 \).
(d) Effect of sample size: \( \theta_1 = 0.3, \theta_2 = 0.8, \Delta_{f,g} = 0.1, \) and \( N_2 = 50 \). Choose \( N_1 \in \{20, 25, 30, \ldots, 115, 120\} \).
Figure 2 shows the ARE for these four different settings. We observe that AREs are larger than one for almost all the parameter settings, indicating that the truncated Wilcoxon statistic has greater power than the standard Wilcoxon statistic. In settings (a) and (b), \( \text{ARE}(T_{IW}, T_{W}) \) increases when \( \Delta_{f,g} \) increases or the proportion of zeros increases. Hence, when the nonzero part are away from each other, the truncated test statistic gains more power compared to the standard Wilcoxon statistics. In setting (c), the \( \text{ARE}(T_{IW}, T_{W}) \) function is not monotone due to the cancellation of difference between non-zero part and the non-zero proportions. In setting (d), obviously \( \text{ARE}(T_{IW}, T_{W}) \) does not depend on the value of \( N_1 \) or \( N_2 \), which can be seen in the formula. In Figure 2 (a), when \( \Delta_{f,g} \) has a large negative value \((-0.44)<1\), \( \text{ARE}(T_{IW}, T_{W}) \) is smaller than 1. This is the case when \( \Delta_{f,g} \) and \( \Delta \theta \) show opposite effects and cancel with each other, leading to a loss of efficiency from the truncation of zeros. However, in real microbiome studies, this scenario is very unlikely to occur since this would imply that individuals who do not carry a particular bacterium have a similar risk as the individuals who have a very high abundance of this same bacterium.

We have further verified the theoretical \( \text{ARE} \) using simulations and present the results in Section 1 of the Supplemental Materials. We observe that the simulated \( \text{ARE} \) is close to the theoretical \( \text{ARE} \), in terms of both values and the trends as we change the parameters.

3. A truncated Kruskal-Wallis test for \( K \)-group comparison with equal sample sizes.

3.1. A truncated Kruskal-Wallis test. Consider the setting where we have data from \( K \) groups \( X_1, X_2, \ldots, X_K \), all containing non-negative independent observations from population \( 1, 2, \ldots, K \), respectively. Each group \( X_i = (x_i(1), \ldots, x_i(N_i)) \) contains \( N_i \) i.i.d observations. We assume that the samples from the \( K \) groups are generated from the following two-part model,

\[
(x_i(j) \sim (1 - \theta_i) \delta_0 + \theta_i f_i, \quad i = 1, \ldots, K).
\]

The hypothesis of interest is

\[
H_0 : \theta_1 = \cdots = \theta_K = \theta, \quad f_1 = \cdots = f_K = f \quad \text{vs} \quad H_1 : \text{not all } \theta_i \text{s and } f_i \text{s are equal}.
\]

The Kruskal-Wallis test is a standard nonparametric test for \( K \)-group comparison based on the ranks. We propose to develop a similar rank-based test that accounts for excessive zeros in the two-part model. We first consider the case when the samples sizes from all \( K \) groups are the same, i.e., \( N_1 = N_2 = \cdots = N_K = N \). Let \( r_i(j) \) be the rank of the \( j \)th observation in the \( i \)th group among all \( KN \) observations. The Kruskal-Wallis statistic can be written as

\[
T_{KW} = \frac{12}{KN^2(KN+1)} \sum_{i=1}^{K} S_i^2,
\]

where \( S_i = \sum_{j=1}^{N} r_i(j) - (KN+1)N/2 \). To derive the distribution of \( T_{KW} \) under the null hypothesis, we rewrite \( T_{KW} \) in terms of \( Y_i \), where \( Y_i = \sum_{j=1}^{i} S_j - i S_{i+1}, 1 \leq i \leq K - 1, \)

\[
T_{KW} = \frac{12}{KN^2(KN+1)} \sum_{i=1}^{K-1} \frac{Y_i^2}{i(i+1)} = \sum_{i=1}^{K-1} \frac{Y_i^2}{\text{Var}[Y_i]}.
\]

With this transformation, \( Y_i \)s are asymptotically independent with each other, and under the null hypothesis, the test statistic \( T_{KW} \) is the summation of \( K - 1 \) chi-square random variables with degree of freedom of 1, and therefore is asymptotically distributed as a \( \chi^2_{K-1} \) distribution.

In order to account for excessive zeros in the data sets, we propose to modify the \( T_{KW} \) statistic using the idea of truncation. Let \( n_i \) denote the number of nonzero observations in
Fig 2: Asymptotic relative efficiency $\text{ARE}(T_{T_W}, T_{W})$ comparing the truncated Wilcoxon rank test and the standard Wilcoxon rank test as a function of (a) $\Delta f,g$, (b) $\theta$, (c) $\Delta \theta$, and (d) $N_1$. For each plot, the horizontal dashed line represents $\text{ARE}(T_{T_W}, T_{W}) = 1$.

For each group $i$, so the proportion of nonzero entries is $p_i = n_i / N$. Let $p = \max_{1 \leq i \leq K} p_i$. For each group $i$, keep the $pN$ observations with smallest ranks (or largest values), so that all the removed entries are zeros. Let $n = pN = \max_{1 \leq i \leq K} p_i N = \max_{1 \leq i \leq K} n_i$, then the truncated data set has $Kn$ observations in total. Rank all $Kn$ observations from smallest to largest, and then let $r_i$ denote the rank-sum of the observations in group $i$. Define

$$s_i = r_i - n(Kn + 1)/2, \quad U_i = \sum_{j=1}^{i} s_j - is_{i+1}, \quad 1 \leq i \leq K - 1,$$

where $s_K = -\sum_{i=1}^{K-1} s_i$. The following Lemma show that $U_i$’s are independent.

**Lemma 1** Under the model that all $x_i(j)$ are independently and identically distributed,

$$\mathbb{E}[U_i] = 0, \quad \text{Cov}(U_i, U_{i_2}) = 0.$$
This leads to our definition of the truncated Kruskal-Wallis test statistic as

$$T_{iKW} = \sum_{i=1}^{K-1} \frac{U_i^2}{\text{Var}[U_i]}$$

which has a $\chi^2$ distribution with degree of freedoms $K - 1$ under the null hypothesis. When $K = 2$, this statistic becomes the truncated Wilcoxon rank-sum statistic given $N_1 = N_2$.

The following Lemma 2 provides an approximation of the variance of $Y_i$ and $U_i$ under null hypothesis and alternative hypothesis.

**Lemma 2** Consider the two-part model (3). Under the null hypothesis and suppose that $N \to \infty$, we have

$$\text{Var}[Y_i] = \frac{i(i+1)K^2N^3\theta}{4} \left\{ \frac{\theta^2}{3} + (1 - \theta) \right\} + O(N^{5/2}), \quad 1 \leq i \leq K - 1;$$

$$\text{Var}[U_i] = \frac{i(i+1)K^2N^3\theta^3}{4} \left\{ \frac{1}{3} + (1 - \theta) \right\} + O(N^{5/2}), \quad 1 \leq i \leq K - 1.$$

Under the alternative hypothesis, we have the upper bound of the variances,

$$\text{Var}_1[Y_i] \leq K^3N^3, \quad \text{Var}_1[U_i] \leq K^3N^3, \quad 1 \leq i \leq K.$$

Based on this lemma, the unknown variance $\text{Var}[U_i]$ can be approximated by

$$\frac{i(i+1)K^2N^3(\sum_{k=1}^{K} p_k/K)^3}{4} \left( \frac{4}{3} - \sum_{k=1}^{K} p_k/K \right).$$

3.2. Pitman’s asymptotic relative efficiency of $T_{iKW}$ and $T_{KW}$ under a two-part model.

We evaluate the relative efficiency of the truncated statistic $T_{iKW}$ with the standard statistic $T_{KW}$ defined by

$$RE(T_{iKW}, T_{KW}) = \frac{E_1[T_{iKW}]}{E_1[T_{KW}]} = \frac{\sum_{i=1}^{K-1} E_1[U_i^2]/\text{Var}[U_i]}{\sum_{i=1}^{K-1} E_1[Y_i^2]/\text{Var}[Y_i]}.$$

We assume that the $K$ groups have the same sample size $N$ and the data are generated from the two-part model (3). To find $RE(T_{iKW}, T_{KW})$, there are four terms to calculate: the expectation of $U_i^2$ and $Y_i^2$ under the alternative hypothesis, and the variance of $U_i$ and $Y_i$ under the null hypothesis. To find $E_1[U_i^2]$ and $E_1[Y_i^2]$, we can calculate $\text{Var}[U_i]$ and $\text{Var}[Y_i]$, and the expectation of $U_i$ and $Y_i$ separately, where the variances are needed under either hypothesis. Lemma 2 provides an approximation of the variance of $Y_i$ and $U_i$ under null hypothesis and alternative hypothesis.

In order to calculate the expectation of $Y_i$ and $U_i$ under alternative hypothesis, we note that the basic terms in $Y_i$ and $U_i$ are the rank-sum of the nonzero observations. Let $r_i^0$ denote the rank-sum of the non-zero observations in group $i$, $1 \leq i \leq K$, and that $s_i^0 = r_i^0 - (1 + \sum_{j=1}^{K} n_j)/2$. Define $\Delta_{i,k} = P(x_i < x_k) + P(x_i = x_k)/2 - 1/2$, $x_i \sim f_i$ and $x_k \sim f_k$, for $1 \leq i, k \leq K - 1$, where $\Delta_{i,k}$ can be used to measure the effect sizes. In addition, one can easily check that

$$E\left( \sum_{j=1}^{i} s_j^0 - i s_{i+1}^0 | n_1, \ldots, n_K \right) = \sum_{j=1}^{i} \sum_{k \neq j} n_j n_k \Delta_{j,k} - i n_{i+1} \sum_{k \neq i+1} n_k \Delta_{i+1,k},$$

for $1 \leq i \leq K - 1$. For model (3) and the effect sizes specified above, we have the results on $E_1[Y_i]$ and $E_1[U_i]$, as shown in the following Lemma.
Lemma 3 Under alternative hypothesis, for $1 \leq i \leq K - 1$, the expectation of $Y_i$ is

$$E_1[Y_i] = N^2 \left( \sum_{j=1}^{i} \sum_{k \neq j} \theta_j \theta_k \Delta_{j,k} - i \theta_{i+1} \sum_{k \neq i+1} \theta_k \Delta_{i+1,k} \right) + \frac{KN^2}{2} \left( i \theta_{i+1} - \sum_{j=1}^{i} \theta_j \right),$$

and the expectation of $U_i$ is

$$E_1[U_i] = N^2 \left( \sum_{j=1}^{i} \sum_{k \neq j} \theta_j \theta_k \Delta_{j,k} - i \theta_{i+1} \sum_{k \neq i+1} \theta_k \Delta_{i+1,k} \right) + \frac{KN^2}{2} \theta(K) \left( i \theta_{i+1} - \sum_{j=1}^{i} \theta_j \right) + o(N^2).$$

Plugging $E_1[U_i]$, Var $[U_i]$ and Var $[U_i]$ into the definition of $\text{ARE}(T_{tKW}, T_{KW})$ and using Lemma 2 and Lemma 3, we obtain the $\text{ARE}$ of $T_{tKW}$ versus $T_{KW}$ given by the following theorem.

Theorem 2 Under model (3) and that $N \to \infty$, then the $\text{ARE}$ can be derived as

$$\text{ARE}(T_{tKW}, T_{KW}) = \sum_{i=1}^{K-1} \left\{ \frac{(\sum_{j=1}^{i} \theta_j + i \theta_{i+1}^2) \Delta_{i} + K \theta(K) (i \theta_{i+1} - \sum_{j=1}^{i} \theta_j)/2}{(\theta(K) (i \theta_{i+1} - \sum_{j=1}^{i} \theta_j)/2)} \right\}^2 + o(1).$$

3.3. Asymptotic relative efficiency for zero-Beta distributions. We consider the case where the nonzero functions $f_{kS}$ are Beta distributions with parameters $\alpha_k$ and $\beta = 1$, in which case the effect size $\Delta_{j,k}$ can be calculated and the $\text{ARE}$ can be expressed in terms of $\alpha_k$'s and $\theta_k$'s. Since compositional data are always between 0 and 1, Beta distribution provides a reasonable parametric model for such data.

Given the Beta distribution for each sample, note that $x_i \sim \text{Beta}(\alpha, 1)$, then $x^\alpha \sim \text{Unif}(0, 1)$, therefore

$$P(x_i < x_k) = P(x_i^\alpha < x_k^\alpha) = P(U < (x_k^\alpha)(\alpha_k/\alpha_i)), \quad U \sim \text{Unif}(0, 1).$$

Since $x_k^\alpha \sim \text{Unif}(0, 1)$, we have $(x_k^\alpha)^{\alpha_k/\alpha_i} \sim \text{Beta}(\alpha_k/\alpha_i, 1)$. Further, for any continuous random variable $x$ with range $[0, 1]$, there is

$$P(U < x) = \int_0^x (1 - F_x(u)) \, du = E[x].$$

This leads to

$$\Delta_{i,k} = P(x_i < x_k) - 1/2 = \frac{\alpha_k/\alpha_i}{\alpha_k/\alpha_i + 1} = \frac{\alpha_k}{\alpha_i + \alpha_k} - 1/2, \quad 1 \leq i, k \leq K.$$

When combining with (4), we have

$$E_1[\tau^0_i | n_1, \cdots, n_K] = n_i \left\{ \frac{n_i + 1}{2} + \sum_{k \neq i} n_k \frac{\alpha_k}{\alpha_i + \alpha_k} \right\},$$

$$E_1[\tau^0_i | n_1, \cdots, n_K] = E_1[\tau^0_i] - \frac{1 + \sum_{k=1}^{K} n_k}{2} = n_i \sum_{k \neq i} n_k \left( \frac{\alpha_k}{\alpha_i + \alpha_k} - 1/2 \right).$$

Plugging these equalities to (5) gives the closed-form expression for $\text{ARE}(T_{tKW}, T_{KW})$.

To demonstrate the gain in efficiency, we calculate $\text{ARE}(T_{tKW}, T_{KW})$ for five group comparisons ($K = 5$) in two scenarios: (a) $\alpha_i = i, 1 \leq i \leq 5, \theta_1 = \cdots = \theta_K = \theta$, where $\theta = 0.1 + 0.01k, k = 0, 1, \cdots, 90$. (b) $\alpha_i = i, 1 \leq i \leq 5, \theta = (0.2, 0.15, 0.3, 0.1, 0.25) + d$, where $d = 0.01k, k = 0, 1, 2, \cdots, 70$. The results are shown in Figure 3. In both cases, we observe a high asymptotic relative efficiency using the truncated Kruskal-Wallis test statistic when compared to the original Kruskal-Wallis test.
Fig 3: Asymptotic relative efficiency. $\text{ARE}(T_{tKW}, T_{KW})$ as a function of (a) $\theta$ when $\theta_1 = \cdots = \theta_K = \theta$, and (b) $\theta(K)$ when not all $\theta_i$s are equal. For each plot, the horizontal dashed line represents $\text{ARE}(T_{tKW}, T_{KW}) = 1$.

4. A truncated Kruskal-Wallis multi-group test with unequal sample sizes. We now consider the setting where we have multiple samples $X_1, X_2, \cdots, X_K$, all containing non-negative independent observations from population $1, 2, \cdots, K$, respectively. Each group $X_i = (x_i(1), \cdots, x_i(N_i))$ contains $N_i$ observations sampled from the two-part model (3). We consider the case that $N_1, N_2, \cdots, N_K$ are not necessarily equal.

Consider the standard Kruskal-Wallis statistic first. Let $r_i(j)$ be the rank of the $j$th observation in the $i$th group among all the observations. Define $S_i = \sum_{j=1}^{N_i} r_i(j) - \left(\sum_{i=1}^{K} N_i + 1\right) N_i/2$ and $Y_i = \sum_{j=1}^{i} (N_i+1) S_j - N_j S_{i+1}$, $1 \leq i \leq K - 1$, where $Y_i$ is also related to the sample size. The Kruskal-Wallis statistic is defined as

$$T_{KW} = \frac{12}{\sum_{j=1}^{K} N_j + 1} \sum_{i=1}^{K-1} \frac{Y_i^2}{N_i+1(\sum_{j=1}^{i} N_j)(\sum_{j=1}^{i+1} N_j)(\sum_{j=1}^{K} N_j)}, \quad 1 \leq i \leq K - 1.$$ 

Each term follows a $\chi^2_1$ distribution asymptotically and is independent with each other, so the statistic follows $\chi^2_{K-1}$ distribution under the null hypothesis.

To account for zeros, for group $i$, there are $n_i$ non-zero elements, and the corresponding non-zero ratio is $p_i = n_i/N_i$. Let $p = \max_{1 \leq i \leq K} p_i$, and keep the $\lfloor p N_i \rfloor$ largest observations only for group $i$ as the truncated data. All the removed entries are zeros. Let $r_i$ be the rank-sum for the truncated samples in group $i$, and

$$s_i = r_i - \lfloor p \sum_{i=1}^{K} N_i \rfloor + 1/2\lfloor p N_i \rfloor.$$ 

We define the statistic $U_i$ as the counterpart of $Y_i$ in the standard Kruskal-Wallis statistic,

$$U_i = \sum_{j=1}^{i} (N_{i+1} s_j - N_j s_{i+1}), \quad 1 \leq i \leq K - 1.$$ 

Then, a natural test statistic based on truncation is

$$T_{tKW} = \sum_{i=1}^{K-1} \frac{U_i^2}{\text{Var}[U_i]},$$

where $\text{Var}[U_i]$ is the variance of $U_i$ under the null hypothesis.
where $\text{Var}[U_i]$ is the variance of $U_i$ under the null. This is calculated by noting that $\text{Var}[U_i] = \text{Var}\{E(U_i|p_1, \cdots, p_K)\} + E\{\text{Var}(U_i|p_1, \cdots, p_K)\}$. Lemma D in Supplemental Materials shows that

$$\text{Var}[U_i|p_1, \cdots, p_K] = \frac{1}{12} E\left[ \left( \sum_{k=1}^{K} n_k + 1 \right) \left\{ \sum_{k=1}^{K} n_k \left[ N_{i+1} \sum_{j=1}^{i} n_j + n_{i+1} \sum_{j=1}^{i} N_j \right]^2 \right\} \right],$$

$$- \left( n_{i+1} \sum_{j=1}^{i} N_j - N_{i+1} \sum_{j=1}^{i} n_j \right)^2),$$

$$E[\text{Var}[U_i|p_1, \cdots, p_K]] = \frac{\theta^2}{12} N_{i+1} \sum_{j=1}^{i} N_j \sum_{j=1}^{K} N_j \left[ \sum_{j=1}^{K} (\sum_{j=1}^{K} N_j) \theta + 3 - 2\theta \right].$$

However, when the sample sizes are not equal, it is difficult to evaluate $\text{Var}[U_i|p_1, \cdots, p_K]$. Instead we approximate this under the null hypothesis by assuming that the non-zero probability $\theta$ is the average of the empirical non-zero probability among all samples and by simulations since $\text{Var}[U_i|p_1, \cdots, p_K]$ only depends on the non-zero proportions, but not on $f$.

Finally, in order to prove that $T_{IKW}$ has an asymptotic distribution of $\chi^2_{K-1}$, we show that $(E[U_i])^2$ is much smaller than $\text{Var}[U_i]$ so that $E[U_i]/\sqrt{\text{Var}[U_i]}$ is approximately 0, and the correlation between either two terms is asymptotically 0. Combining the upper bound of $E[U_i]$ given in Lemma C and the lower bound for the variance term given in Lemma D in the Supplemental Material, we show that $E[U_i]/\sqrt{\text{Var}[U_i]} \rightarrow 0$. Using the upper bound for the covariance in Lemma E, we see that when $\max_{1 \leq i \leq K} N_i/N^{(1)} \rightarrow 0$ and $N^{(1)} \rightarrow \infty$, we have

$$|\text{Cor}[U_i,U_j]| = \frac{|\text{Cov}[U_i,U_j]|}{\sqrt{\text{Var}[U_i]}\sqrt{\text{Var}[U_j]}} \leq C \sqrt{\sum_{k=1}^{K} \frac{1}{N_k}} \leq C\sqrt{K}/\sqrt{N^{(1)}} \rightarrow 0, 1 \leq i,j \leq K - 1.$$  

Therefore, as long as the sample sizes are on the same order and go to infinity, $(U_i/\text{Var}[U_i], i = 1, \cdots, K)$ has an asymptotic multivariate normal distribution with mean zero and identity covariance matrix. We leave the details of these lemmas in the Supplemental Materials.

5. Simulations. To further verify the gain in efficiency in using the proposed truncated rank-based tests, we present simulation studies to evaluate the proposed tests and to compare with the standard Wilcoxon rank-sum and Kruskal-Wallis test statistics. In each of the simulation setups, we perform the following steps.

1. Given parameters $(K, N, \theta, \alpha, \beta, s)$, where $N$, $\theta$, $\alpha$ and $\beta$ are all $K \times 1$ arrays, generate $x$ from the distribution
   $$x_i(j) \overset{i.i.d.}{\sim} (1-\theta_i)\delta_0 + \theta_i\text{Beta}(\alpha_i, \beta_i), \quad 1 \leq i \leq K, 1 \leq j \leq N_i.$$

2. Calculate the p-value from usual rank tests, including the Wilcoxon rank-sum test for two-sample test, and Kruskal-Wallis test for more groups, and our proposed tests.

3. Repeat steps 1-2 for $M$ times, and calculate the power or type I error for a given significance level $\alpha$.

5.1. Simulation 1 - evaluation of Type I error. We first evaluate the type I errors of various test statistics proposed in this paper and compare them to Wilcoxon rank-sum test and Kruskal-Wallis test. To simulate data from the null distribution, for each group $i$, we simulate data from a two-part model,

$$x_i(j) \overset{i.i.d.}{\sim} (1-\theta)\delta_0 + \theta\text{Beta}(a, b),$$

where $a = b = 2$, $\theta = 0.5$. We consider three different scenarios:
can obtain more accurate errors, especially when the sample sizes are large. For small sample sizes, the proposed tests are slightly anti-conservative, indicating the asymptotic approximation of the test statistics may require relatively large sample sizes. In practice, when the sample sizes are small, one can obtain more accurate \( p \)-values based on permutations.

(a) \( K = 2 \), sample sizes = \((0.65, 1) \times N\);
(b) \( K = 3 \), equal sample sizes \( N \);
(c) \( K = 3 \), sample sizes = \((0.7, 1, 1.5) \times N\).

For each scenario, take \( N = \{30, 60, 100, 300, 600, 900\} \), and simulate 100,000 test statistics for each choice of \( N \). The empirical Type I errors are summarized in Table 1 for significance level \( \alpha \in \{0.05, 0.01, 0.001\} \). In general, we observe that the proposed tests have correct Type I errors, especially when the sample sizes are large. For small sample sizes, the proposed tests are slightly anti-conservative, indicating the asymptotic approximation of the test statistics may require relatively large sample sizes. In practice, when the sample sizes are small, one can obtain more accurate \( p \)-values based on permutations.

| \( \alpha \)-level | Two-group - unequal sample sizes | Three-group - equal sample sizes | Three-group - unequal sample sizes |
|-------------------|---------------------------------|---------------------------------|---------------------------------|
| 0.05              | \( T_W \) 0.49 0.49 0.50 0.049 0.49 | \( T_{KW} \) 0.049 0.056 0.009 0.014 | \( T_{KW} \) 6.8 \( 10^{-4} \) 9.3 \( 10^{-4} \) 9.1 \( 10^{-4} \) 1.2 \( 10^{-3} \) 9.9 \( 10^{-4} \) |
| 0.01              | \( T_{tW} \) 0.054 0.051 0.010 0.013 | \( T_{tKW} \) 0.009 0.013 0.010 0.011 | \( T_{tKW} \) 2.2 \( 10^{-3} \) 1.9 \( 10^{-3} \) 1.8 \( 10^{-3} \) 1.6 \( 10^{-3} \) 1.2 \( 10^{-3} \) |
| 0.001             | \( T_{tW} \) 6.8 \( 10^{-4} \) 9.3 \( 10^{-4} \) 9.1 \( 10^{-4} \) 1.2 \( 10^{-3} \) 9.9 \( 10^{-4} \) | \( T_{tKW} \) 2.2 \( 10^{-3} \) 1.9 \( 10^{-3} \) 1.8 \( 10^{-3} \) 1.6 \( 10^{-3} \) 1.2 \( 10^{-3} \) | \( T_{tKW} \) 6.4 \( 10^{-4} \) 8.1 \( 10^{-4} \) 1.0 \( 10^{-3} \) 8.9 \( 10^{-4} \) 8.6 \( 10^{-4} \) |
|                   | \( T_{tKW} \) 2.5 \( 10^{-3} \) 2.1 \( 10^{-3} \) 1.9 \( 10^{-3} \) 1.4 \( 10^{-3} \) 1.3 \( 10^{-3} \) | | |

5.2. Simulation 2 - power comparisons. We next evaluate the power of the proposed tests. We consider three different models with \( K = 3 \) groups and repeat the simulation \( M = 10000 \) times. For the first model, we assume that the proportions of zeros are the same across different groups and examine how the distribution of the non-zero observations affects the power of the proposed tests. We set the \( \theta = (0.5, 0.5, 0.5) \) and equal sample sizes for the three groups, chosen from the set \( \{100, 200, 300, \ldots, 1900, 2000\} \). We set \( \alpha \in \{(1.5, 2, 2.5), (1.7, 2, 2.3)\} \), and \( \beta = (2, 2, 2) \) and calculate the power function for different sample sizes \( N \). The resulting power curves are shown in the top row of Figure 4. We observe a substantial gain in power from the truncated Kruskal-Wallis test.

For the second model, we study how different proportions of zeros affect the test power as the sample size \( N \) changes. We set \( \alpha, \beta = \{(1/2, 1/2, 1/2), (2, 2, 2)\} \), which assumes that
Fig 4: Power curves for the truncated Kruskal-Wallis (solid line) and the Kruskal-Wallis test (dashed line) as a function of the sample size $N$. (a)-(b): a two-part model with $\theta = (0.5, 0.5, 0.5)$, $\alpha = (1.5, 2, 2.5)$, $\beta = (2, 2, 2)$ and (b) $\alpha = (1.7, 2, 2.3)$, $\beta = (2, 2)$. (c)-(d): a two-part model with $(\alpha, \beta) = \{(1/2, 1/2, 1/2), (2, 2, 2)\}$, $\theta \in \{(0.10, 0.15, 0.22), (0.15, 0.20, 0.27)\}$. (e)-(f): a two-part model with $\theta = (0.4, 0.5, 0.6)$ and $\alpha \in \{(1.5, 2, 2.5), (1.7, 2, 2.3)\}$, and $\beta = (2, 2, 2)$. For each plot, red lines represent the power curves with reduced sequencing depths.
the nonzero distributions are the same across different groups. We again assume that the sample sizes are the same for all three groups, chosen from the set \{20, 40, 60, \ldots, 280, 300\}. We set \(\theta \in \{(0.1, 0.15, 0.22), (0.15, 0.20, 0.27)\}\), and calculate the power function for different sample sizes \(N\). The resulting power curves are shown in the middle row of Figure 4. Our test has some improvement over the original test. However, in this case, the improvement is not as large as when \(f\)'s are different.

For the last model, we evaluate the proposed tests when the sample sizes are different in different groups. We set \(\theta = (0.4, 0.5, 0.6)\) and the sample size as \(N \times (0.8, 1, 1.5)\), where \(N \in \{100, 200, 300, \ldots, 1900, 2000\}\). We choose \(\alpha \in \{(1.5, 2, 2.5), (1.7, 2, 2.3)\}\) and \(\beta = (2, 2, 2)\), and calculate the power function for different sample sizes \(N\). The power curves are shown in the bottom row of Figure 4, showing that the truncated Kruskal-Wallis test has much higher power than the Kruskal-Wallis test when there are ties.

We finally examine the sensitivity of the proposed tests to reduced sequencing depths. One effect of having low sequencing depth is that some rare bacteria might not be sequenced, which results in zero counts for bacteria with low but non-zero true relative abundance in the final microbial composition. Specifically, after we generate the compositional data for each population, to minic lower sequencing depth, we set the abundance of the bacteria with small relative abundance to zero with a probability of 0.5, resulting about 5% of non-zeros proportions being set to zero. The final data sets include more zeros due to reduced sequencing depths. We obtain the empirical power again based on the final data sets. The new power curves are shown in Figure 4 (red lines). The proposed test still achieve higher power than the standard tests (dashed line). Overall, we see that the proposed tests are not too sensitive to read depths from sequencing.

6. Identifying the Crohn’s disease-associated bacterial genera and the effects of treatment. Crohn’s disease, a chronic inflammatory bowel disease, is characterized by altered composition of the gut microbiota or dysbiosis. The etiology and clinical significance of the dysbiosis is unknown. In a recent study at the University of Pennsylvania, the composition of the gut microbiota among a cohort of 85 children with Crohn’s disease who were initiating therapy with either a defined formula diet \((n=33)\) or an anti-tumor necrosis factor \(\alpha\) (anti-TNF) drug \((n=52)\) was examined in order to better understand the cause of the dysbiosis in Crohn’s disease. Fecal samples were collected at baseline, 1, 4, and 8 weeks and DNA content characterized by shot-gun genomic sequencing with >0.5 tetra-bytes of total sequences (Lewis et al., 2015). The gut microbiota data of 26 normal children with no known gastrointestinal disorders were similarly collected. The MetaPhlAn program (Segata et al., 2012) was applied to first obtain the relative proportions of the bacterial genera for each of the normal and Crohn’s disease samples. The number of bacterial genera called by the MetaPhlAn program varied from sample to sample. Altogether, 60 bacterial genera were observed in both healthy samples and the Crohn’s disease samples. Figure 1 shows the relative abundances of the 60 genera in all the samples, where large proportions of zeros are observed for many of the genera.

6.1. Comparison of gut microbiome between normal and Crohn’s disease patients. We are interested in identifying the bacterial genera that show different distributions between healthy and Crohn’s disease children. At an \(\alpha\) level of 0.05, the truncated Wilcoxon rank-sum test identified 23 genera that showed different distributions between healthy children and patients with Crohn’s disease, while the standard Wilcoxon test identified 20, among these, 19 were identified by both methods. At FDR of 10\%, the truncated Wilcoxon rank-sum test identified 21 genera and the standard Wilcoxon test identified the same 20 genera. Four genera, including *Eggerthella*, *Lactobacillus*, *Gemella*, and *Rothia* were identified
only by the truncated Wilcoxon rank-sum test. Figure 5 shows the the proportions of zeros and boxplots of these four genera that clearly show difference in abundances in healthy and Crohn’s disease children, both in terms of proportion of zeros and the median of non-zero abundances. All four genera had higher abundances in Crohn’s patients than the healthy controls. Among these genera, *Eggerthella, Lactobacillus*, both are anaerobic, non-sporulating, Gram-positive bacilli, have been reported to be associated with clinically significant bacteremia and Crohn’s disease (Lau et al., 2004). In contrast, for genus *Methanobrevibacter*, the standard Wilcoxon test had a sightly smaller p-value. However, 98% of the disease individuals did not carry this genus. In this case, the truncation may lead to a slightly reduced statistical significance. However, this can also be due to random error or the asymptotic approximation error due to relatively small sample sizes. To verify this, we performed 100,000 permutations of group labels and obtained a p-value of 0.016 and 0.017, for the truncated test and the standard Wilcoxon test, respectively.

**Fig 5:** The box plots of the non-zero relative abundances (on log scale) with proportions of zeros in x-axis labels for the four bacterial genera that were identified by the truncated Wilcoxon rank-sum test only for a nominal \( \alpha \) level of 0.05, where \( p \)-values from \( T_W \) and \( T_{tW} \) tests are shown as \( p_S \) and \( p_s \). The genus *Methanobrevibacter* was identified only by the standard Wilcoxon rank-sum test.

6.2. Comparison of gut microbiome across time after treatment. We next aim to identify the bacterial genera that show changes of relative abundances across the four time points during the anti-TNF treatment. We applied our truncated Kruskal-Wallis test statistic with within-subject permutations (100,000 permutations) and identified 7 genera with changes of abundances during the treatment with \( p < 0.05 \). As a comparison, the Friedman’s test identified only three. Figure 6 shows the boxplots of the abundances of the genera across 4 time points for the four genera identified only by our truncated Kruskal-Wallis test, including *Bacteroides, Roseburia, Eubacterium* and *Bilophila*. Among these, *Bacteroides, Eubacterium*
and Bilophila showed increased abundances at 8 weeks after the anti-TNF treatment. Interestingly, all three genera have been shown to have reduced abundance in Crohn’s disease patients (Gevers et al., 2014). Reduction of Roseburia, a well-known butyrate-producing bacterium of the Firmicutes phylum, has been consistently demonstrated to be associated with Crohn’s disease (Machiels et al., 2014). There has been evidence that the gut bacteria in patients with inflammatory bowel disease do not make butyrate, and that they have low levels of the fatty acid in their gut (Sartor and Mazmanian, 2012). The decreased abundance of Bilophila, may translate into a reduction of commensal bacteria-mediated, anti-inflammatory activities in the mucosa, which are relevant to the pathophysiology of Crohn’s disease. The result shows the effect of anti-TNF treatment in increasing the relative abundances of Roseburia, Bacteroides and Bilophila, therefore potentially increasing the level of fatty acid butyrate and anti-inflammatory activities. This partially explains that 50% of the patients showed clinical improvement, as reflected by reduction of fecal calprotectin below 250 mcg/g (Lewis et al., 2015).

Fig 6: Bar plots of proportions of zeros (top row) and box plots of the relative non-zero abundances (bottom row, on log scale) over four time points during anti-TNF treatment for four bacteria genera that were identified only by the truncated Kruskal-Wallis test.

7. Discussion. Motivated by comparing the distributions of taxa composition in different groups in microbiome studies, we have developed several extensions of the popular rank-based tests to account for clumps of zeros, including truncated rank-based Wilcoxon and Kruskal-Wallis tests for two- or multiple-group comparisons. These tests are rank-based and nonparametric and are easy to implement. By using within-sample permutations, such tests can also be applied to paired samples or repeated measurements analysis. We have shown that the proposed tests have better power than the standard rank-based tests, both by asymptotic relative efficiency analysis and by simulations. We observed a large gain in power when the proportions of zeros in the data sets are high as compared to the standard rank-based tests due the high number of tied ranks. Hallstrom (2010) showed that the truncated Wilcoxon rank-sum test has equal or better power than the test of two-degree of freedoms based on two-part model when the sample sizes are equal. Since our two-sample truncated test is an extension
of Hallstrom’s test to unequal sample sizes, we expect that our proposed tests should have similar or better power than the two-part tests that combine binomial test and nonparametric test for the continuous part. Although such two-part tests can be extended to two-part model for multiple group comparisons, it has not been studied in literature. It would be interesting to compare the performance of our proposed truncated Kruskal-Wallis tests with other tests based on the two-part models.

We have demonstrated the applications of the proposed tests in an analysis of real metagenomic data sets. Our results have shown that the truncated rank-based tests are effective and identify more bacterial genera that are associated with the clinical phenotypes and treatment. As observed in other studies (Wagner, Robertson and Harris, 2011), tests that account for clumps of zeros can be more powerful in testing abundance difference in microbiome studies. We have demonstrated that some well-known Crohn’s disease-associated bacterial genera can be missed by using the standard rank-based tests without adjusting for excessive zeros. We expect to see more applications of the proposed tests in microbiome studies.

The proposed tests have several limitations. First, these tests are rank-based and nonparametric, they cannot directly account for covariate effects in observational studies. If there is no severe covariate imbalance, we can perform quantile-stratification based on covariates and apply the proposed tests to each strata and then combine the results using e.g., Fisher’s combination of \( p \)-values. This requires large sample sizes. Second, since the \( p \)-values of our proposed rank-based tests are calculated based on the asymptotic distributions of the test statistics, which require relatively large sample sizes, they may not be accurate enough when the sample sizes are small. As shown in our Table 1, as the sample sizes increase, the Type I error gets closer to the nominal level. For small sample sizes, we would suggest that the users apply the proposed tests and obtain the \( p \)-values based the asymptotic distributions. Once the taxa are identified, one can run a permutation test to further confirm the results. This will save time to run permutation tests for all the taxa.

Acknowledgment. This research was supported by NIH grants GM129781 and GM123056. We thank Dr. Kafadar, the AE and two reviewers for many helpful comments and suggestions.

SUPPLEMENTARY MATERIALS. The online Supplemental Materials include proofs of Theorem 1, Theorem 2 and all the lemmas. It also includes detailed derivations of the proposed test statistics and simulations to verify the theoretical asymptotic relative efficiency of the proposed tests. An R repository of the proposed method and all analyses performed is available at https://github.com/hongzhe88/Truncated-Rank-based-Tests.

REFERENCES

GEVERS, D., KUGATHASAN, S., DENSON, L. A. and ET AL (2014). The treatment-naive microbiome in new-onset Crohn’s disease. Cell Host & Microbe 15 382–392.

HALLSTROM, A. (2010). A modified Wilcoxon test for non-negative distributions with a clump of zeros. Statistics in Medicine 29 391–400.

HUSON, D. H., AUCH, A. F., Qi, J. and SCHUSTER, S. C. (2007). MEGAN analysis of metagenomic data. Genome research 17 377–386.

LACHENBRUCH, P. (1976). Analysis of Data with Clumping at Zero. Biometrische Zeitschrift 18 351–356.

LACHENBRUCH, P. (2001). Comparison of two-part models with competitors. Statistics in Medicine 20 1215–1234.

LACHENBRUCH, P. A. (2002). Analysis of data with excess zeros. Statistical methods in medical research 11 297–302.

LAU, S., WOO, P., FUNG, A., CHAN, K., WOO, G. and YUEN, K. (2004). Journal of Medical Microbiology 53 1247–1253.
LEWIS, J., CHEN, E. Z., BALDASSANO, R. N., OTLEY, A. R., GRIFFITHS, A. M., LEE, D., BITTINGER, K., BAILEY, A., FRIEDMAN, E. S., HOFFMANN, C., ALBENBERG, L., SINHA, R., COMPER, C., GILROY, E., NESSELMAN, J., GRANT, A., CHEHOUJ, C., LI, H., WU, G. D. and BUSHMAN, F. D. (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn’s Disease. *Cell Host & Microbe* **18** 489-500.

MACHIELS, K., JOOSSENS, M., SABINO, J., DE PRETER, V., ARIJS, I., EECbkHOUT, V., BALLET, V., CLAES, K., VAN IMMERSEEL, F., VERBEKE, K., FERRANTE, M., VERHAEGEN, J., RUTGEERTS, P. and VERMEIRE, S. (2014). A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* **63**(8) 1275–1283.

MANICHANH, C., BORUEL, N., CASELLAS, F. and GUARNER, F. (2012). The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology* **9** 599–608.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIelsen, T., PONS, N., LEVENEZ, F., YAMADA, T. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** 59–65.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y., SHEN, D. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55–60.

SARTOR, R. and MAZMANIAN, S. (2012). Intestinal Microbes in Inflammatory Bowel Diseases. *American Journal of Gastroenterology (Supplements)* **1** 15-21.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. and HUTTENHOWER, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9** 811–814.

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. and GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444** 1027–131.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. and GORDON, J. I. (2007). The human microbiome project. *Nature* **449** 804–810.

WAGNER, B., ROBERTSON, C. and HARRIS, J. (2011). Application of Two–Part Statistics for Comparison of Sequence Variant Counts. *PloS One* **6**(5) e20296.