Supplemental Material

Analysis of Composition of Microbiomes (ANCOM): A novel method for studying microbial composition

S1 Data and the statistical parameters

For simplicity of exposition and notation, throughout this section we shall use the phrase "microbial ecosystem" or simply "ecosystem" to describe the source of the "specimen" which is interrogated to obtain the OTUs of various taxa. For example, when comparing the composition of gut microbiome of babies born vaginally with those born through C-section the ecosystem of interest is the gut and the fecal sample is the specimen. Hence it is important to note that the observed OTUs are the taxa abundance in the specimen and not the abundance in the entire ecosystem where the specimen was derived from. Using the specimen level OTU abundance data, in this paper we develop methodology for comparing total taxa abundance in the ecosystem between two or more populations. Throughout this section K denotes the number of populations and n_k denotes the number of subjects randomly selected from the k-th population, $k = 1, 2, \ldots, K$. For example, to compare the gut microbiome composition of vaginally delivered babies with C-section

delivered babies, we obtain a sample of 100 vaginally delivered babies and a sample of 50 C-section delivered babies. Here $K = 2$ and $n_1 = 100$ and $n_2 = 50$. Throughout the paper, in the main text as well as in this Supplementary text, the terms "population" and "test group" (or simply "group") are used interchangeably.

I. Observable data: For a biological specimen obtained from the j-th subject, $j = 1, \ldots, n_k$ from the k-th population, let $\mathbf{Y}_j^{(k)} = (Y_{1j}^{(k)}, Y_{2j}^{(k)} \ldots, Y_{pj}^{(k)})'$ denote the vector of OTUs representing p taxa. Note that $\mathbf{Y}_j^{(k)}$ represents the abundance in the biological specimen and not the total microbial abundance in the ecosystem at the time of sampling. Secondly, $\mathbf{Y}_j^{(k)}$ is one random realization from the j-th subject, which will vary from specimen to specimen from the same subject. Furthermore, subjects themselves are a random sample from the given population. Thus implying that $\mathbf{Y}_j$ is a random variable which has two components of variation, namely, variability between specimens within the same subject and variability between subjects. Typically, in most microbiome studies researchers do not obtain more than one specimen at a given time, consequently, variability between specimens within subject at a given time point is not measured. Hence, similar to existing methods, in this paper we therefore do not account for specimen to specimen variability within a given subject at a given time point. However, it is straightforward to modify the methodology to account for that variance component.

II. Statistical parameters: In the following we describe all the parameters governing the OTU data obtained from a random sample of subjects from a population. Throughout this section we denote the average value of a random observation Y over the suitable population (which will be clear from the context) by $E(Y)$, the expected value of Y. The terms mean, average value, expected value, and expectation of abundance are all equivalent and will be used interchangeably. In the remainder of the article, we use i, j and k to denote taxon, individual subjects and population respectively.

A. Subject specific parameters

Expected abundance of a taxon in a specimen within a subject: For a randomly chosen biological specimen (e.g. fecal sample) obtained from the j-th subject in the k-th population, let the expected OTU count of the i-th taxon (e.g. *Bifidobacterium*) be denoted by $E(Y_{ij}^{(k)}|\theta_{ij}^{(k)}) = \theta_{ij}^{(k)}$. It is important to note that θ_{ij} represents the expected abundance within the specimen from a subject and not the abundance of the taxon in the ecosystem of the subject.

Expected abundance of a taxon in the ecosystem within a subject: For the j-th subject in the k-th population, we denote the expected abundance of the i-th taxon in the ecosystem of interest by $\mu_{ij}^{(k)}$. Since a specimen obtained from the j-th subject is a small fraction of the total in the ecosystem, it is reasonable to assume that for the i-th taxon $\mu_{ij}^{(k)} = c_j\theta_{ij}^{(k)}$ for some positive constant c_j that is specific to subject j at the time of sampling. For example, c_j may represent the total volume of the ecosystem where the biological specimen was derived from.

Expected relative abundance of a taxon in a specimen within a subject: For a randomly selected biological specimen obtained from the j-th subject in the k-th population, the expected relative abundance of the i-th taxon is given by $\lambda_{ij}^{(k)} = \frac{\theta_{ij}^{(k)}}{\sum_{r=1}^{p}\theta_{rj}^{(k)}}$.

Expected relative abundance of a taxon in the ecosystem of interest within a subject: The expected relative abundance of the i-th taxon in the ecosystem of interest in the j-th subject from the k-th population is parameterized as $\gamma_{ij}^{(k)} = \frac{\mu_{ij}^{(k)}}{\sum_{r=1}^{p}\mu_{rj}^{(k)}}$. If all taxa are randomly distributed in the ecosystem where the specimen is derived from, and if one assumes that the biological specimen is a reasonable representation of the true mix for a given subject, then one may assume $\gamma_{ij}^{(k)} = \lambda_{ij}^{(k)}$.

B. Population specific parameters

Expected abundance of a taxon in a specimen obtained from a random subject in the k-th population: Note that $\theta_{ij}^{(k)}$ (defined above) is a random variable and changes from subject to subject within the k-th population. The expected abundance of the taxon in a random specimen obtained from a random subject in the k-th population is given by $\eta_i^{(k)} = E(\theta_{ij}^{(k)})$. It is important to note that $\eta_i^{(k)}$ represents mean abundance in a specimen and NOT the mean abundance of the taxon in the ecosystem where the specimen was derived from.

Expected abundance of a taxon in the ecosystem of interest in the k-th population: Similar to $\theta_{ij}^{(k)}$, $\mu_{ij}^{(k)}$ is a random variable and changes from subject to subject in the k-th population. Hence the mean abundance of a taxon in the ecosystem of interest in the k-th population is given by $\nu_i^{(k)} = E(\mu_{ij}^{(k)})$. This is the primary parameter of interest for biologists.

Expected relative abundance of a taxon in a specimen obtained from a random subject in the k-th population: For the i-th taxon in the k-th population we define $\delta_i^{(k)} = \frac{\eta_i^{(k)}}{\sum_{r=1}^{p} \eta_r^{(k)}}$.

Expected relative abundance of a taxon in the ecosystem of interest in the k-th population: For the i-th taxon, the expected relative abundance in the ecosystem of interest for the population is given by:

$$\rho_i^{(k)} = E\Big[\frac{\theta_{ij}^{(k)}}{\sum_{r=1}^{p} \theta_{rj}^{(k)}}\Big] = E\Big[\frac{c_j \theta_{ij}^{(k)}}{\sum_{r=1}^{p} c_j \theta_{rj}^{(k)}}\Big] = E\Big[\frac{\mu_{ij}^{(k)}}{\sum_{r=1}^{p} \mu_{rj}^{(k)}}\Big]$$

Note that the above expression does not require knowledge of the distribution of c_j's. The parameters described above, along with their estimators, are summarized in Table S1.

Table S1: Summary of various parameters and corresponding estimators

Analysis	Parameter Description	Unknown Parameter	Estimator
Subject specific	Expected abundance of i-th taxon in a random specimen from the j-th subject in the k-th population	$E(Y_{ij}^{(k)}\|\theta_{ij}^{(k)}) = \theta_{ij}^{(k)}$	$Y_{ij}^{(k)}$ (OTU for i-th taxon)
	Expected relative abundance of i-th taxon in a random specimen from the j-th subject in the k-th population	$\lambda_{ij}^{(k)} = \frac{\theta_{ij}^{(k)}}{\sum_{r=1}^{p}\theta_{rj}^{(k)}}$	$\hat{\lambda}_{ij} = \frac{Y_{ij}}{\sum_{r=1}^{p} Y_{rj}}$
	Expected total abundance of i-th taxon in j-th subject in the k-th population	$\mu_{ij}^{(k)} = c_j E(Y_{ij}^{(k)}\|\theta_{ij}^{(k)})$	Not estimable unless c_j is known
	Expected relative abundance of i-th taxon in j-th subject in the k-th population	$\gamma_{ij}^{(k)} = \frac{\mu_{ij}^{(k)}}{\sum_{r=1}^{p}\mu_{rj}^{(k)}}$	$\hat{\gamma}_{ij}^{(k)} = \frac{Y_{ij}^{(k)}}{\sum r=1^{p} Y_{rj}}$
Population specific	Expected abundance of i-th taxon in a random specimen from the k-th population	$\eta_i^{(k)} = E(\theta_{ij}^{(k)})$	$\hat{\eta}_i^{(k)} = \frac{1}{n_k}\sum_{r=1}^{n_k} Y_{ir}^{(k)}$
	Expected abundance of i-th taxon in the k-th population	$\nu_i^{(k)} = E(\mu_{ij}^{(k)})$	Not estimable unless c_j's are known
	Relative abundance of i-th taxon in the k-th population	$\rho_i^{(k)} = E(\gamma i j^{(k)})$	$\hat{\rho}_i^{(k)} = \frac{1}{n_k}\sum_{r=1}^{n} \hat{\gamma}_{ir}^{(k)}$

S2 Analysis of Composition of Microbiomes (ANCOM) using relative abundance

Statistical Hypotheses

As noted earlier, for $i = 1, 2, \ldots, p$, the comparison among the K populations in terms of $\eta_i^{(k)}$ is not equivalent to comparing $\nu_i^{(k)}$. However, the relative abundance $\rho_i^{(k)}$'s can be compared using the specimen level relative abundance estimates obtained from each subject. More precisely, $\hat{\lambda}_{ij}^{(k)}$ can be used for drawing inferences on $\rho_i^{(k)}$ among the K populations. For each subject j, $j = 1, 2, \ldots, n_k$, $k = 1, 2, \ldots, K$, $\sum_{i=1}^{n_k} \hat{\lambda}_{ij}^{(k)} = 1$, we view these as compositional data and apply the general ideas developed by Aitchison[1] to analyze microbiome data. Following Aitchison we log-transform the ratios after adding

a small constant ω to $Y_{ij}^{(k)}$ to avoid logarithms for zero values. In all numerical work reported in this paper we took $\omega = 0.001$, although some may prefer to take $\omega = 1$. Note that log-transformation of data is inspired by the Box-Cox family of transformations which are routinely used in data analysis [2]. Thus, along the lines of Aitchison's compositional data analysis, we perform all our inferences on the expectation of the log-transformed ratios rather than the ratios themselves.

As demonstrated in the following propositions the above formulation allows us to draw inferences regarding the mean abundances $\nu_i^{(k)}$, the main parameter of interest, under the following assumption which may be reasonable in the context of microbiome data. Since within each population k and for each taxon i, the random variables $\mu_{ij}^{(k)}$ as well as $\gamma_{ij}^{(k)}$ are identically and independently distributed for all subjects j, $j = 1, 2, \ldots, n_k$, we shall drop the index j from $\mu_{ij}^{(k)}$ as well as from $\gamma_{ij}^{(k)}$ in the following propositions. For simplicity of exposition the rest of this section will be devoted to the case $K = 2$, although the methodology is applicable more generally for $K > 2$.

Assumption A: The mean abundance (in log scale) of at most $p - 2$ taxa are different between two populations. More precisely, suppose $E[\log(\mu_i^{(1)}/\mu_i^{(2)})] = d_i$, $i = 1, 2, ..., p$. Then among $d_1, d_2, \ldots, d_p$, at most $p - 2$ are non-zero.

Assumption B: Mean abundance (in log scale) of all p taxa do not differ by the same amount between two populations. But if they do, then the difference is zero. More precisely, suppose $E[\log(\mu_i^{(1)}/\mu_i^{(2)})] = d_i$, $i = 1, 2, ..., p$, and if $d_i = d$, for all i. Then $d = 0$.

Note: For notational simplicity we shall drop the phrase "(in log scale)" from "mean abundance (in log scale)" in rest of this text.

Recall that $E[\log(\gamma_i^{(1)}/\gamma_i^{(2)})] = E[\log(\mu_i^{(1)}/\mu_i^{(2)})]$. Therefore the above assumptions apply

at the specimen level as well.

Proposition 1: For $j = 1, 2, \ldots, p$ (with $p > 2$), suppose either Assumption A or Assumption B is true. Furthermore, suppose for all j and r ($r \neq j$), $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] = E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]$. Then, for all j, $E[\log(\mu_j^{(1)})] = E[\log(\mu_j^{(2)})]$.

Proof: Since for all j and r, ($r \neq j$), $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] = E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]$, therefore $E[\log(\mu_j^{(1)})] - E[\log(\mu_r^{(1)})] = E[\log(\mu_j^{(2)})] - E[\log(\mu_r^{(2)})]$. Hence

$$E[\log(\mu_j^{(1)})] - E[\log(\mu_j^{(2)})] = E[\log(\mu_r^{(1)})] - E[\log(\mu_r^{(2)})],\ \forall j, r, (r \neq j). \quad (1)$$

Assumption A implies that there exist at least 2 taxa whose mean abundances are same between the two populations. Let r be the index of one such taxon, i.e., $E[\log(\mu_r^{(1)})] = E[\log(\mu_r^{(2)})]$. This, together with (1), implies that $E[\log(\mu_j^{(1)})] - E[\log(\mu_j^{(2)})] = 0$ for all j.

Instead of Assumption A, suppose Assumption B is true. Again from (1) we note that for every j and for every r ($r \neq j$), $E[\log(\mu_j^{(1)})] - E[\log(\mu_j^{(2)})] = E[\log(\mu_r^{(1)})] - E[\log(\mu_r^{(2)})]$. Thus appealing to Assumption B we have $E[\log(\mu_j^{(1)})] - E[\log(\mu_j^{(2)})] = 0$.

Proposition 2: Suppose there are p taxa ($p > 2$) and suppose Assumption B is true.
(a) If there exists a j such that for every r, $r \neq j$,

$$E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] \neq E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]. \quad (2)$$

Then $E[\log(\mu_j^{(1)})] \neq E[\log(\mu_j^{(2)})]$.
(b) Suppose for some j, $j = 1, 2, \ldots, p$, $E[\log(\mu_j^{(1)})] = E[\log(\mu_j^{(2)})]$, then there exists at

least one $r(\neq j)$ such that (2) does not hold.

Proof: Recall that for all $j, r = 1, 2, \ldots, p$, $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] = E[\log(\mu_j^{(1)}/\mu_r^{(1)})]$ and $E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})] = E[\log(\mu_j^{(2)}/\mu_r^{(2)})]$. According to the assumption in (a) there exists a j such that for all $r \neq j$, $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] \neq E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]$. Assumption A implies that there exist at least 2 taxa whose mean abundances are same between the two populations. Let r be the index of one such taxon, i.e., $E[\log(\mu_r^{(1)})] = E[\log(\mu_r^{(2)})]$. Equivalently we have $E[\log(\gamma_r^{(1)})] = E[\log(\gamma_r^{(2)})]$. Combining this with (2) we have $E[\log(\gamma_j^{(1)})] \neq E[\log(\gamma_j^{(2)})]$. Equivalently we have $E[\log(\mu_j^{(1)})] \neq E[\log(\mu_j^{(2)})]$. Hence we prove (a).

To prove (b), suppose for some j, $j = 1, 2, \ldots, p$, $E[\log(\mu_j^{(1)})] = E[\log(\mu_j^{(2)})]$. Then we know from Assumption A that there exists at least 1 more taxon that has same mean abundance in the two populations. Denote the index of this taxon by r. Then $E[\log(\mu_r^{(1)})] = E[\log(\mu_r^{(2)})]$ and consequently $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] = E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]$. Hence we prove (b).

Remark 1: If for all r and j $(r \neq j)$ $E[\log(\gamma_j^{(1)}/\gamma_r^{(1)})] \neq E[\log(\gamma_j^{(2)}/\gamma_r^{(2)})]$, then there exists at least $p-1$ taxa which have differentially abundant population means. If exactly $p-1$ taxa are differently abundant, then it is not possible to identify taxa with differentially abundant population means using the log ratios. In most applications, it is unlikely that there will be at least $p-1$ taxa that have differentially abundant population means.

We now provide two examples to illustrate the above propositions.

Example 1: To illustrate Proposition 1, suppose we have two groups each consisting of three taxa. Suppose $(E(\log(\mu_1^{(1)}), E(\log(\mu_2^{(1)}), E(\log(\mu_3^{(1)})) = (3, 4, 2)$ and

$(E(\log(\mu_1^{(2)}), E(\log(\mu_2^{(2)}), E(\log(\mu_3^{(2)})) = (a, b, c)$. Then under the assumptions of Proposition 1 we have

$$a - b = -1, a - c = 1, b - c = 2. \quad (4)$$

For some c_1, c_2 and c_3, let

$$a - 3 = c_1, b - 4 = c_2, c - 2 = c_3. \quad (5)$$

Substituting the values of a, b and c from (5) into (4) we obtain

$$c_1 = c_2 = c_3.$$

The above equality together with (5) and Assumption B imply that

$$c_1 = c_2 = c_3 = 0.$$

Hence from (5) we have $a = 3, b = 4, c = 2$. Thus satisfying proposition 1.

Example 2: Similar to Example 1, suppose we have two groups consisting of three taxa each. Suppose $(E(\log(\mu_1^{(1)})),\ E(\log(\mu_2^{(1)})),\ E(\log(\mu_3^{(1)}))) = (3, 4, 2)$ and $(E(\log(\mu_1^{(2)})),$ $E(\log(\mu_2^{(2)})), E(\log(\mu_3^{(2)}))) = (1, 4, 2)$. Thus in this example, only $E(\log(\mu_1^{(1)})) \neq E(\log(\mu_1^{(2)}))$ but the rest are equal. Trivially, for all $r \neq 1$, we have $E(\log(\mu_1^{(1)}/\mu_r^{(1)})) \neq E(\log(\mu_1^{(2)}/\mu_r^{(2)}))$, thus verifying part (a) of Proposition 2. Similarly, it is trivial to verify part (b) of Proposition 2.

Since data on taxa abundance in the ecosystem is not available, therefore, for $i =$

$1, 2, \ldots, p$, it is not possible to test the following hypotheses directly

$$H_{0i} : E(\log(\mu_i^{(1)})) = E(\log(\mu_i^{(2)})),$$
$$\text{against } H_{ai} : E(\log(\mu_i^{(1)})) \neq E(\log(\mu_i^{(2)})). \quad (6)$$

However, by virtue of Propositions 1 and 2, for each i, the above hypotheses can be tested by testing the following $(p-1)$ hypotheses regarding the abundance of the i-th taxon relative to the r-th taxon for every $r \neq i$.

$$H_{0ri} : E[\log(\mu_i^{(1)}/\mu_r^{(1)})] = E[\log(\mu_i^{(2)}/\mu_r^{(2)})],$$
$$\text{against } H_{ari} : E[\log(\mu_i^{(1)}/\mu_r^{(1)})] \neq E[\log(\mu_i^{(2)}/\mu_r^{(2)})]. \quad (7)$$

Statistical Decision Rule:

For each taxon i, $i = 1, 2, \ldots, p$, we test the hypotheses (7) for all $r \neq i$ using the log-ratios $\log(\frac{\hat{\gamma}_{ij}^{(k)}}{\hat{\gamma}_{rj}^{(k)}})$, $j = 1, 2, \ldots, n_k$, $k = 1, 2, \ldots, K$. The testing problem may be formulated using standard ANOVA model:

$$\log\left(\frac{\hat{\gamma}_{ij}^{(k)}}{\hat{\gamma}_{rj}^{(k)}}\right) = \alpha_{ir} + \beta_{irk} + \epsilon_{irjk}, \quad (8)$$

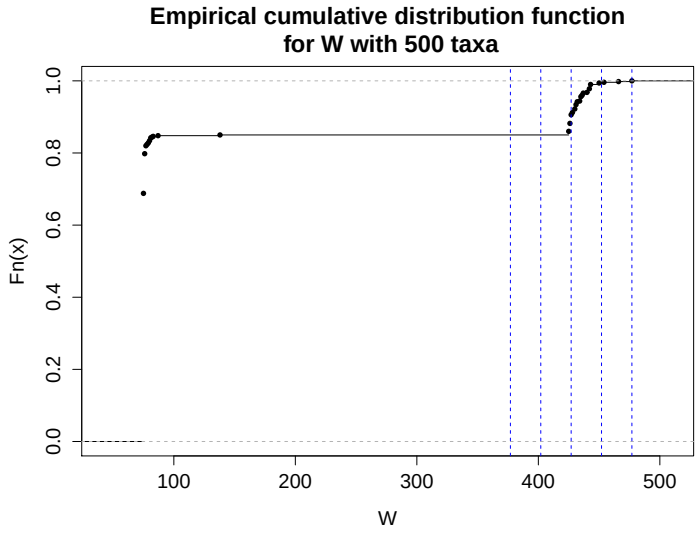
where, for a given pair i, r, α_{ir} is the overall common mean and β_{irk} is the effect of the k-th group (or k-th level of the factor). We may assume ϵ_{irjk} are identically and independently distributed across samples $j = 1, 2, \ldots, n_k$ and groups $k = 1, 2, \ldots, K$, with $\epsilon_{irjk} \sim N(0, \sigma_{ir}^2)$. Of course, as in standard ANOVA, one may allow heteroscedasticity, where the variance σ_{ir}^2 may vary with group. Then the null hypothesis, for the taxa pair i and r reduces to the standard ANOVA hypothesis $H_{0ir} : \beta_{ir1} = \beta_{ir2} = \ldots = \beta_{irK} = 0$.

For each i and for each $r \neq i$, as with classical statistical inference, depending upon the validity of the distributional assumptions regarding the random error ϵ, and sample sizes, one may use either the standard parametric t-test (or F-test if $K > 2$) or use a nonparametric procedure such as Wilcoxon rank sum test (or the Kruskal-Wallis test for $K > 2$) or a resampling procedure such as the permutation or bootstrap to compute the p-values.

Altogether we are testing $p(p-1)/2$ distinct hypotheses H_{0ir}, $r \neq i$. Consequently, when decisions regarding the significance of each taxa is being made, we need to apply a multiple testing correction such as the Benjamini-Hochberg (BH) procedure when rejecting the sub-hypothesis H_{0ir}, $r \neq i$. However, if p is small (say, less than 10) then one may apply the Bonferroni procedure rather than the BH procedure. For each taxon i, let W_i denote the number of sub-hypotheses H_{0ir}, $r \neq i$, that are rejected. In the ideal setting, by virtue of Propositions 1 and 2, we would reject the null hypothesis (1) if $W_i = p - 1$. However, such a decision rule is potentially conservative. That is, the cut-off $p - 1$ may be too stringent. If p is small (e.g. less than 10) then we may arbitrarily choose the cut-off to be $p - 2$ otherwise we shall make use of the empirical distribution of $\{W_1, W_2, \ldots, W_p\}$ do determine the cut-off. Similar to gene expression data, the empirical distribution of $\{W_1, W_2, \ldots, W_p\}$ is bimodal with mode on the right of distribution corresponding to taxa with differentially abundant means and the mode on the left corresponding to the taxa whose means are not differentially abundant. Typical empirical CDF of $\{W_1, W_2, \ldots, W_p\}$ is provided in Figure S1 the long flat region corresponds to the flat region between the two modes. To arrive at suitable critical value, we search for changes in $F_W(W_i)$, where F_W is the empirical cumulative distribution function of W_i's. Starting from $W_{(p)}$, which is the largest order statistic of W_i, we search intervals of length 0.05 to detect drops in $F_W(W_i)$. Let these intervals be

denoted by $I_j = \big(W_{(p)} - 0.05(j)p, W_{(p)} - 0.05(j-1)p\big]$, $j = 1, 2, \ldots$. Let $\delta_j = \Delta F_W(I_j)$ denote the change in F_W in I_j. We choose $w_0 = min\{W_i : W_i \in I_j\}$ as the cutoff if $\{\delta_j \geq \tau, \delta_{j+1} < \tau, \delta_{j+2} < \tau\}$, where τ is a fraction (taken as 0.02 in this case). A value of τ closer to 0 is equivalent to w_0 close to $p-1$, thereby indicating a more conservative cutoff. Thus, for the i-th taxon, we reject the null hypothesis (1) if $W_i > w_0$. Although the choice of w_0 is fairly flexible, for purposes of this article we choose $w_0 = 0.75$.

Figure S1: Example plot of the empirical cumulative distribution function of W_i. The vertical dotted lines show the intervals for determining the cutoff.



Remark 2: Whether Assumption A and Assumption B are valid or not, one can compare relative abundances of taxa (at the ecosystem level) in two or more groups by testing hypotheses (7) against pre-specified reference taxon i. In such a case only $p-1$ hypotheses are tested instead of $p(p-1)/2$ hypotheses as described above. Standard BH procedure (or Bonferroni) is applied on the $p-1$ p-values. Thus in this case all inferences will be performed relative to the $i-th$ taxon.

Remark 3: A multiple testing correction applied on the complete set of p-values corresponding to H_{0ir}, $r \neq i$ is conservative due to the large number of hypotheses. Note that each decision rule $W_i > w_0$ depends only on the set of sub-hypotheses H_{0ir} for a given i. Hence an alternative correction strategy would be to apply the Benjamini-Hochberg (or Bonferroni) procedure on H_{0ir} for each taxon (i) separately, for determining W_i. We recommend using this strategy in large microbiome datasets.

Remark 4: Note that the p-values corresponding to the sub-hypotheses H_{0ir} are potentially dependent and the BH procedure is not necessarily robust for arbitrary dependence structures. However, our extensive simulation studies (see Figure 2 in the main paper) reveal that our above strategy always controls the FDR at 0.05, never exceeds it.

Remark 5: Suppose a researcher conducts a multifactorial study and is interested in comparing the taxa abundance across various levels of each factor. For example, an investigator interested in studying taxa abundance according to various levels of factors such as gender, race, mode of delivery, antibiotic use. Model (8) can be easily extended as in classical factorial analysis model consisting of f factors:

$$\log\left(\frac{\hat{\gamma}_{ij}^{(k_1,k_2,\ldots,k_f)}}{\hat{\gamma}_{rj}^{(k_1,k_2,\ldots,k_f)}}\right) = \alpha_{ir} + \beta_{irk_1} + \beta_{irk_2} + \ldots + \beta_{irk_f} + \epsilon_{irjk_1k_2\ldots k_f}, \quad (9)$$

where β_{irk_s} denotes the effect of k_s-th level of the s-th factor. As in classical multifactorial ANOVA, one can even introduce interactions into the model.

Remark 6: As in the case of standard linear fixed and mixed models, the above method-

ology can be easily extended to account for covariates and random effects by suitably modifying (9). Thus the standard machinery available for linear fixed and mixed effects models can be exploited, the only difference being the outcome or response variables are suitable log-ratios of observed abundances of taxa in the sample. One can invoke PROC GLM or PROC MIXED in SAS or use packages such as lme4 or nlme in R.

Remark 7: In view of Remark 6, when there are several covariates present in the model, a researcher can apply his/her favorite model selection procedure for selecting variables/factors to arrive at a parsimonious model for further analysis.

Table S2: Comparison of various methods available in literature

Article	Model	Parameter studied	Assumptions
La Rosa et al. 2012 [3]	Overdispersed Dirichlet-Multinomial (ODM)	δ_i	Taxa level OTU counts within each individual are distributed as ODM.
Holmes et al. 2013 [4]	Dirichlet-Multinomial mixtures	δ_i	Matrix of occupancies follow multinomial distribution.
Chen and Li 2013 [5]	Overdispersed Dirichlet-Multinomial (ODM)	η_i	Same model as La Rosa et al. 2012. Used for variable selection.
Paulson et al. 2013 [6]	Zero Inflated Gaussian	η_i	1. Logarithm of OTUs follow normal distribution. 2. Implicitly assume that the sum of the OTU counts within a subject is constant.
This paper	Aitchison's log-ratio	ν_i	1. Flexible. If distributional assumptions needed for standard ANOVA are not satisfied then non-parametric methods including resampling procedures can be used. 2. Either (a) The mean abundance of at most $p-2$ taxa are different between two populations, or (b) If all p taxa are differentially abundant then the mean abundance of all p taxa do not differ by the same amount between two populations. But if they do, then the difference is zero.

S3 An illustration

We re-analyzed a microbial dataset from a recently published study [7] to illustrate ANCOM. LaRosa et al. [7] conducted a multifactorial study consisting of factors such as, gender, mode of delivery, breast milk consumption, gestational age (categorized into three levels, < 26 weeks, $26 - 28$ weeks and > 28 weeks). Other variables included, day of life of the infant when the fecal sample was obtained (continuous variable), amount of antibiotics used (continuous variable). Data were collected on 58 infants (922 observations) in two batches. Authors were specifically interested in comparing the various levels of the above factors and the effect of above variables on the relative abundance of 3 microbial classes, namely, Bacilli, Clostridia and Gammaproteobacteria. Since repeated fecal samples were collected on each infant, the authors used linear mixed effects models (with AR(1) covariance structure) on the relative abundance data (the outcome variable) for each class of bacteria separately. Furthermore they performed stratified analysis according to the three gestational age categories separately since they were interested in comparing the various levels of the above factors and the effect of above variables on the relative abundance of 3 microbial classes, namely, Bacilli, Clostridia and Gammaproteobacteria according to the gestational age of the infant.

We implemented ANCOM on the log-ratios of the 4 bacterial classes (Bacilli, Clostridia, Gammaproteobacteria and Others) using linear mixed effects modeling with AR(1) correlation structure within each individual. Rather than performing a stratified analysis by each gestational age category, we included gestational age as a 3 level factor in each of our mixed effects models and, in addition to all the main effects, we included interactions of each of the above factors and continuous variables with gestational age in order to discover if the effects of any of the factors or variables changed with the gestational age. Our methodology can be described in the following steps.

Analysis of composition of microbiomes

(1) Step 1: For a given bacterial class, denoted by i, we fitted a log-ratio linear mixed effects model consisting of all main effects, gestational age, gender, mode of delivery, amount of breast milk, day of life of sample, amount of antibiotics used and batch. Additionally, we included the interaction of each of these variables with gestational age. Thus corresponding to each bacterial class i, we fitted 3 log-ratio mixed effects models involving the above terms in the model. As in [7], correlations due to repeated measurements were modeled using the AR(1) structure. Thus, for each i and each interaction term, we obtain 3 p-values due to the three log-ratio linear mixed effects models. Since 3 is small, we applied our statistical decision rule described above using the Bonferroni correction rather than the BH correction for multiple testing and the threshold $w_0 = p - 2$ to declare if an interaction (with gestational age) for the bacterial class i is significant at a false positive rate of 0.05.

(2) Step 2: Re-analyze the parsimonious model including the main effects and those interaction terms which were significant in the previous step. Again we use our statistical decision rule as described above.

Select all factors and variables (and interactions) that are significant in the above step. Note that if an interaction is significant then we automatically report the corresponding main effects as well.

Results of our analysis using ANCOM are summarized in Table S3, S4 and S5. Note that for each class i, among all the interaction terms, only the interaction between C-section and gestational age (GA$*$C-Section) is significant in at least 2 log-ratio models, i.e. exceed $w_0 = 3-2 = 1$. Hence for all bacterial classes this interaction term is retained for Step 2 of our analysis and all other interaction terms are dropped.

From Table S3 we observe that breast milk, day of life and GA$*$C-Section are significantly associated with the abundance of Bacilli. Similarly day of life, days on antibiotics

and GA∗C-Section are significantly associated with the abundance of Clostridia (Table S4), while only day of life and GA∗C-Section are significantly associated with the abundance of Gammaproteobacteria (Table S5).

All analyses reported in this section were using PROC MIXED procedure in SAS version 9.0. For illustration purposes, in Figure 3 of the main text we provided the unadjusted average OTU abundance of the three bacterial classes according to the significant factors using PROC MIXED.

Table S3: ANCOM analysis of Bacilli. Models in the second step include all main effects and the interaction of C-Section and gestational age categories which was significant in the first step. The bacterial classes Bacilli, Clostridia, Gammaproteobacteria and Others are denoted by B, C, G and O, respectively. The gestational age is denoted by GA. Significant variables at the end of second step are highlighted in gray.

	Step 1			Step 2		
Factor	$\frac{B}{C}$	$\frac{B}{G}$	$\frac{B}{O}$	$\frac{B}{C}$	$\frac{B}{G}$	$\frac{B}{O}$
GA	0.235	0.005	0.001	0.011	0.001	<.0001
Gender	0.181	0.451	0.057	0.085	0.426	0.093
C-Section	0.151	0.223	<.0001	0.088	0.145	<.0001
Breast milk	0.138	<.0001	0.098	0.132	0.001	0.005
Day of life	<.0001	0.000	0.012	<.0001	0.001	0.026
Days on antibiotics	0.083	0.232	0.273	0.004	0.412	0.231
Sampling period	0.639	0.003	0.015	0.226	0.006	0.004
GA∗Gender	0.596	0.031	0.167			
GA∗C-Section	0.003	<.0001	<.0001	0.001	<.0001	<.0001
GA∗Breast milk	0.341	0.084	0.418			
GA∗Day of life	0.442	0.713	0.222			
GA∗Days on antibiotics	0.591	0.505	0.908			

Table S4: ANCOM analysis of Clostridia. Models in the second step include all main effects and the interaction of C-Section and gestational age categories which was significant in the first step. The bacterial classes Bacilli, Clostridia, Gammaproteobacteria and Others are denoted by B, C, G and O, respectively. The gestational age is denoted by GA. Significant variables at the end of second step are highlighted in gray.

	Step 1			Step 2		
Factor	$\frac{C}{B}$	$\frac{C}{G}$	$\frac{C}{O}$	$\frac{C}{B}$	$\frac{C}{G}$	$\frac{C}{O}$
GA	0.235	0.023	0.521	0.011	0.053	0.051
Gender	0.181	0.057	0.980	0.088	0.656	0.144
C-Section	0.151	0.719	0.077	0.085	0.310	0.599
Breast milk	0.138	0.041	0.846	0.132	0.176	0.457
Day of life	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Days on antibiotics	0.083	0.014	0.397	0.004	0.001	0.061
Sampling period	0.639	0.045	0.031	0.226	0.263	0.002
GA∗Gender	0.596	0.137	0.116			
GA∗C-Section	0.003	0.012	0.068	0.001	0.002	0.019
GA∗Breast milk	0.341	0.141	0.864			
GA∗Day of life	0.442	0.562	0.812			
GA∗Days on antibiotics	0.591	0.311	0.562			

S4 Simulation Study Design

We performed extensive simulation studies to investigate the performance of ANCOM, t-test and ZIG [6] for comparing two populations. We generated a random sample $n_1 = 20$ subjects from population 1 and $n_2 = 30$ subjects from population 2. We considered two different patterns of number of taxa, $p = 500$ or $p = 1000$. Since the differences among the three statistical procedures in terms of false discovery rate (FDR) and power did not change with p, in this study we considered only these 2 patterns of p.

For the j-th subject from the k-th population we generated an ecosystem with OTU count $A_{ij}^{(k)}$ for the i-th taxon using a Poisson distribution. For subjects from the first population $A_{ij}^{(1)}|\mu_{ij} \sim^{ind} Poisson(\mu_{ij})$ and for subjects from the second populations $A_{ij}^{(1)}|\mu_{ij}, u_{ij} \sim^{ind} Poisson(\mu_{ij} + u_{ij})$. In both populations $\mu_{ij}^{(k)}$ were generated indepen-

Table S5: ANCOM analysis of Gammaproteobacteria. Models in the second step include all main effects and the interaction of C-Section and gestational age categories which was significant in the first step. The bacterial classes Bacilli, Clostridia, Gammaproteobacteria and Others are denoted by B, C, G and O, respectively. The gestational age is denoted by GA. Significant variables at the end of second step are highlighted in gray.

	Step 1			Step 2		
Factor	$\frac{G}{B}$	$\frac{G}{C}$	$\frac{G}{O}$	$\frac{G}{B}$	$\frac{G}{C}$	$\frac{G}{O}$
GA	0.005	0.023	0.020	0.001	0.053	0.210
Gender	0.451	0.057	0.055	0.426	0.310	0.581
C-Section	0.223	0.719	0.043	0.145	0.656	0.077
Breast milk	<.0001	0.041	0.019	0.001	0.176	0.407
Day of life	0.000	<.0001	0.195	0.001	<.0001	0.180
Days on antibiotics	0.232	0.014	0.063	0.412	0.001	0.086
Sampling period	0.003	0.045	0.000	0.006	0.263	<.0001
GA∗Gender	0.031	0.137	0.004			
GA∗C-Section	<.0001	0.012	0.578	<.0001	0.002	0.677
GA∗Breast milk	0.084	0.141	0.181			
GA∗Day of life	0.713	0.562	0.754			
GA∗Days on antibiotics	0.505	0.311	0.824			

dently from a gamma distribution $Gamma(a, 1)$. We chose a to take values of 50, 200 and 10000 to represent low, medium and high abundance taxa. In the second population u_{ij} were generated independently from a uniform distribution $U(l_1, u_1)$. Thus the mean abundance of the i-th taxon in the first population is $\gamma_i^{(1)} = a$ and in the second population it is $\gamma_i^{(2)} = a + (l_1 + u_1)/2$. To represent low, medium and high abundance, we chose (l_1, u_1) to be $(100, 150)$, $(200, 400)$ and $(10000, 15000)$, respectively.

To generate abundance of taxa at the specimen level for the j-th subject, we generated $c_j \sim 1/Uniform(l_2, u_2)$, so that $Y_{ij}^{(k)} = [A_{ij}^{(k)} * c_j]$, where $[x]$ denotes the integer part of x. We considered two patterns of (l_2, u_2), namely, $(100, 200)$ and $(200, 500)$. Lastly, we considered 5 different patterns of π the proportion of taxa with differentially abundant means, namely, 0.05, 0.1, 0.15, 0.2 and 0.25. 100 simulated datasets were generated

for each combination of total number of taxa and proportion of differentially abundant taxa. We show the results for larger values of c_j ($(l_2, u_2) = (100, 200)$) in the main text. Results for smaller values of c_j using 500 taxa are shown in Figure S2. We observe that the FDR of t-test and ZIG further increased while ANCOM controlled FDR below the nominal level.

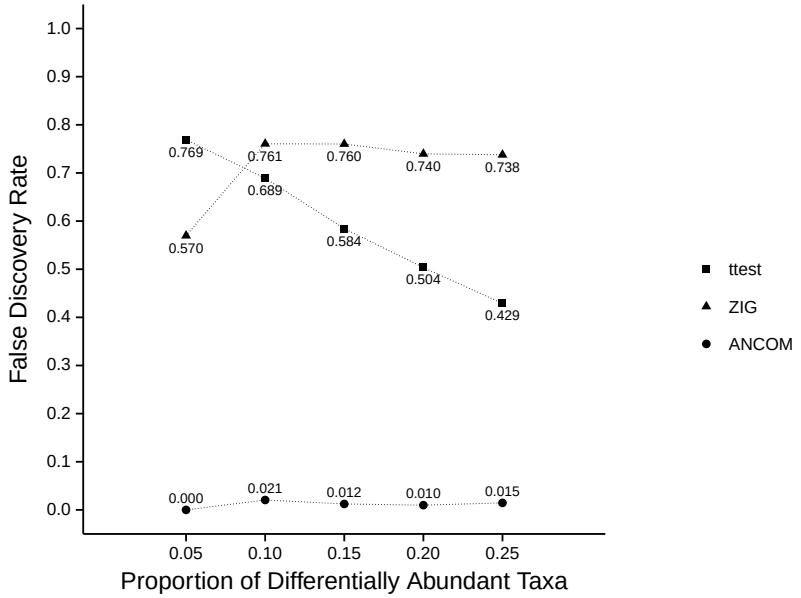
Since the ultimate problem of interest for a biologist is to test the following hypotheses:

$$H_0 : \nu_i^{(1)} = \nu_i^{(2)} \;\; H_a : \nu_i^{(1)} \neq \nu_i^{(2)},$$

in this simulation study we estimated the FDR and power for the above hypotheses regarding the mean abundance of taxa in the ecosystem and not regarding the mean abundance of taxa at the specimen level. Although all analyses were performed in the log scale, the FDR and power were computed for the original hypotheses in terms of ν-s. Since only $Y_{ij}^{(k)}$ are observable therefore all three tests used only $Y_{ij}^{(k)}$ and not $A_{ij}^{(k)}$, which are never observable.

Since tables of microbial count are usually sparse, often researchers perform a simple dimension reduction to focus on a restricted group of taxa. One can summarize the OTU tables to a taxa level and perform ANCOM on the resulting table. We applied ANCOM to analyze the published global gut data [8] consisting of 11,905 OTUs. As commonly done [5], we restricted the analysis to taxa that are present in at least 25% of the samples. This is done because low frequency OTUs are often thought to be difficult to interpret statistically. After filtering out such OTUs we discovered that ANCOM took less than 25 minutes to process the data on a Macbook Pro (Intel Core i7, 2.4 GHz, 16GB RAM). Although it is not a common practice to analyze all OTUs without applying any such filters, to demonstrate the computation speed of ANCOM, we conducted additional simulation studies using a wide range of total OTUs. For 100

Figure S2: Comparison of False Discovery Rate (FDR) to detect differentially abundant microbial taxa by ZIG and ANCOM, based on 100 simulated datasets consisting of 500 taxa for smaller values of c_j. Values of π ranges from 0.05 to 0.25.



simulated OTUs, ANCOM took only 4 seconds, for 1000 OTUs it took 7 minutes, for 10000 OTUs it took 3 hours. The R code to execute ANCOM can be accessed from *http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm.*

References

1. J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:139–177, 1982.

2. D.C. Montgomery, E.A. Peck, and G.G.Vining. *Introduction to Linear Regression Analysis*. Wiley, New York, 5th edition, 2012.

3. P.S. La Rosa, J.P. Brooks, E. Deych, E.L. Boone, and D.J.et.al. Edwards. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *PLoS ONE*, 7(12), 2012.

4. I. Holmes, K. Harris, and C. Quince. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS ONE*, 7(2), 2013.

5. J. Chen and H. Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7:418–442, 2013.

6. J.N. Paulson, O.C. Stine, H.C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10:1200–1202, 2013.

7. P.S. LaRosa, B.B. Warner, Y. Zhou, G.M. Weinstock, E. Sodergren, and C.M.H.et.al. Moore. Patterned progression of bacterial populations in the premature infant gut. *Proceedings of National Academy of Sciences*, 111(34):12522–12527, 2014.

8. T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, and M.et.al. Contreras. Human gut microbiome viewed across age and geography. *Nature*, 486:222–227, 2012.