JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 10, 2015

Pp. 930–939
DOI: 10.1089/cmb.2015.0021

A Method for Determining Taxonomical Contributions to Group Differences in Microbiomic Investigations

ALEXA PRAGMAN[1] RICHARD ISSACSON[2] CHRISTINE WENDT[3] and CAVAN REILLY[4]

ABSTRACT

Here we show how one can decompose the contribution of different levels of taxonomic classification in terms of their impact on differences in the microbiota when comparing two groups. First we demonstrate a difficulty in trying to estimate taxonomic effects at multiple levels simultaneously and demonstrate an approach to determining which taxa have differences in means that are identified. We then develop a model based on an approach that is popular in the RNA-Seq analysis literature and apply it to the problem of determining which taxa differ between two patient groups. This model provides a more powerful method than simpler alternatives. A Bayesian computational strategy is used to obtain exact inference. Simulation studies indicate that the procedure works as intended, and an application to the study of COPD demonstrates the method's practical utility. Software is provided for implementing the method.

Key words: Bayesian modeling, COPD, microbiota, overdispersed counts.

1. INTRODUCTION

The use of 16S rRNA sequencing for characterizing the microbiota in an environment is now widespread. This technology allows one to determine the quantity of many bacterial types in a sample, including those that are currently uncultivable, at least down to the genus level. One question that arises when examining these datasets is what role do the various taxonomical levels play in creating differences between groups. For example, given a data set with counts for a collection of genera, a natural question is do the data indicate differences between groups at other taxonomic levels, such as phyla? For example, Ley et al. (2006) found differences among obese patients and controls at the phylum level, while obviously many human disease processes are driven by specific species; hence, there is a need for methods that partition differences among patient groups across taxonomic levels. The ability to answer such questions could facilitate the interpretation of metagenomic data sets by focusing attention on larger groups of microbes whose shared properties are more easily discerned. The primary challenge here is to rigorously deal with the nested structure of the random taxonomies to avoid reusing the same data to test for differences at distinct taxonomic levels: For example, if we obtain a single genus from each of the observed phyla then our tests at all levels use

[1]Department of Medicine, University of Minnesota, Minneapolis, Minnesota.
[2]Department of Veterinary and Biomedical Sciences, University of Minnesota, St. Paul, Minnesota.
[3]Department of Medicine, VA Medical Center Minneapolis, Minnesota.
[4]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota.

the same data set and return exactly the same *p*-value. Hence, first, we show how to determine which taxa means are actually identifiable. Then, to improve the power to resolve these differences we also propose a fully Bayesian approach for hierarchical smoothing of critical nuisance parameters. This model is similar to techniques that have been developed for the analysis of RNA-Seq data (e.g., Robinson and Smyth, 2007; Robinson and Smyth, 2008; Robinson, McCarthy, et al., 2009 and Anders and Huber, 2010). These models are natural to use as metagenomic data sets, just like RNA-Seq data sets, are composed of counts of reads that map to certain genomic features. Throughout we assume that we have metagenomic data for two groups with replicates within each group; we describe our application in section 4.

2. METHODS

2.1. Identifiability

Due to a lack of identifiability one can not estimate all taxon level means simultaneously unless one has multiple genera within every family, multiple families within every order, and so on. Hence we need to constrain some of the taxon means to be zero. While there were several phyla in our data set with a single genus, the taxonomic structure of our data set was far more complicated. Figures 1–6 display the structures we observed in our application. Given the nested structure of taxonomies, here we suppose that the nonzero parameters are those that are at the coarsest level in the hierarchy. So for a phylum with only one genus observed in our data set, we estimate a phylum effect for that genus and declare the genus effect to be zero (along with all of the intermediate effects). This makes sense because we can still talk about differences between the phyla even when we only have a single representative of some phylum. This principal of defining the identifiable parameter as that which is at the coarsest level will be used to define which parameters are treated as estimable.

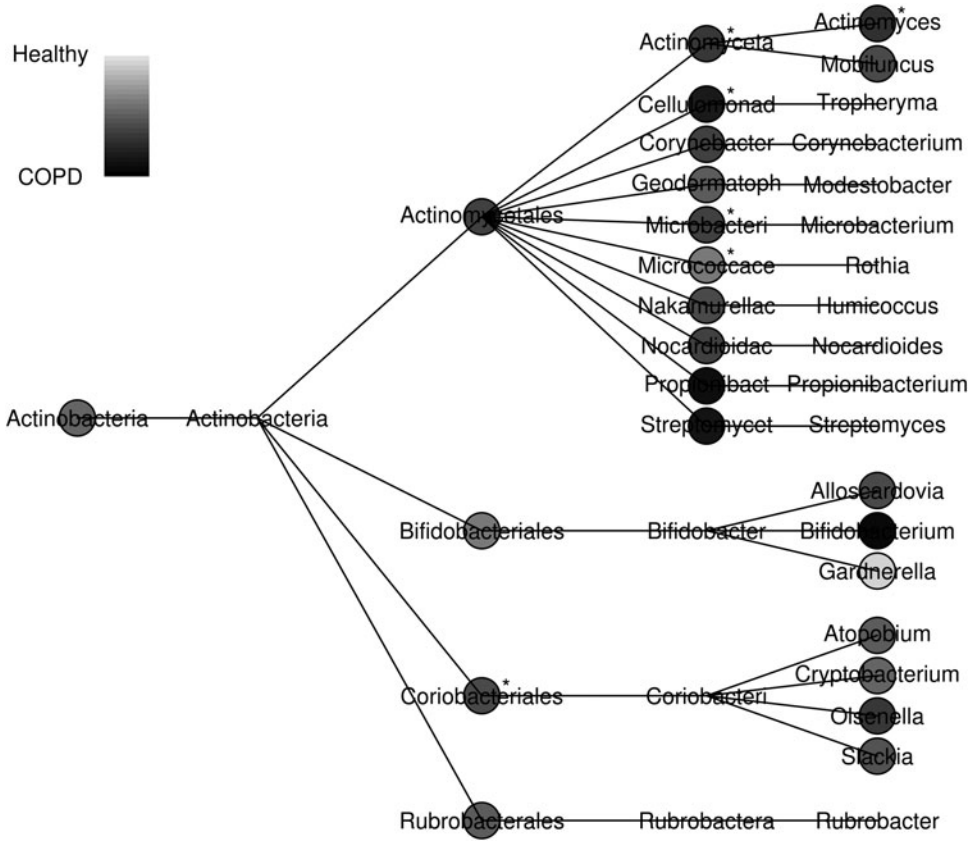

FIG. 1. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for the phylum Actinobacteria (lighter values indicate higher levels in the healthy group). Some family names have been truncated. COPD, chronic obstructive pulmonary disease.

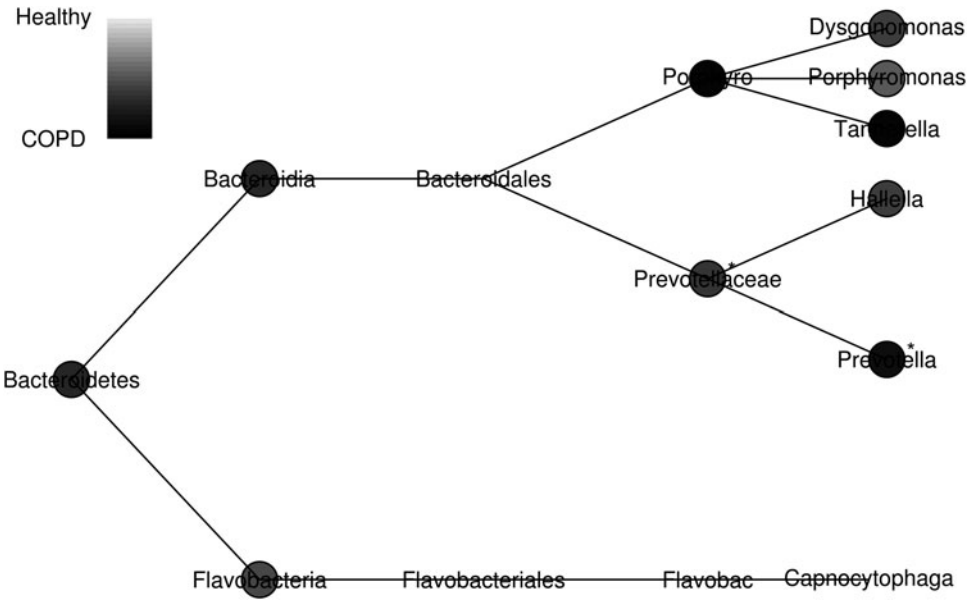

FIG. 2. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for the phylum Bacteroidetes (lighter values indicate higher levels in the healthy group). Some family names have been truncated.

Given the complexity of our observed taxonomic structures we need a method for determining the number of parameters and computational methods for manipulating these structures. We achieve both of these goals by working with a sequence of tables. Let n_g represent the number of distinct genera, and define n_f, n_o, n_c, and n_p analogously (for family, order, class, and phylum). Next let P_g be the number of estimable genus level effects and define P_f, P_o, P_c, and P_p analogously (note that $P_p = n_p$). Next define an $n_p \times n_c$ matrix, $\mathcal{A}_c$ so that the i, j^{th} element is 1 if class j is in phylum i. Now define another $n_p \times n_c$ matrix $\mathcal{B}_c$ so that the i^{th} row of this matrix is all zeros if $|\{j : \mathcal{A}_c[i,j] > 0\}| = 1$ where $|S|$ is the number of elements in set S and $A[i, j]$ represents the i, j^{th} element of the matrix A. If $|\{j : \mathcal{A}_c[i,j] > 0\}| > 1$ then the i^{th} row of $\mathcal{B}_c$ holds $\{j : \mathcal{A}_c[i,j] > 0\}$ and all other elements of this row are zero. Then we find that

$$P_c = |\{\mathcal{B}_c : \mathcal{B}_c > 0\}|.$$

If we define $\mathcal{A}_o, \mathcal{B}_o, \mathcal{A}_f, \ldots, \mathcal{B}_g$ we can also compute P_o, P_f, and P_g analogously.

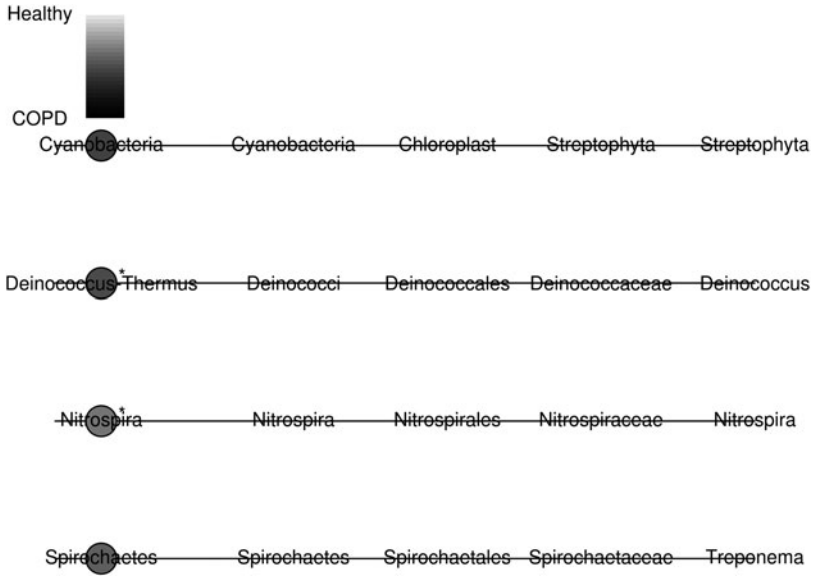

FIG. 3. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for four phyla with only a single genus (lighter values indicate higher levels in the healthy group).

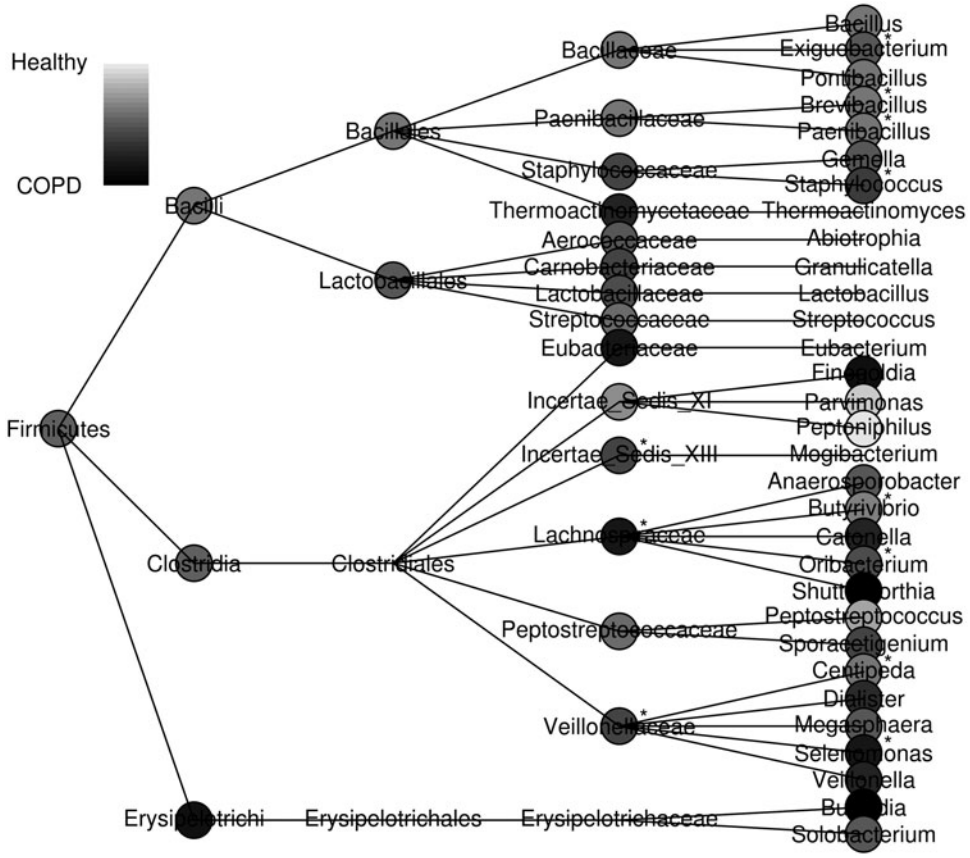

FIG. 4. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for the phylum Firmicutes (lighter values indicate higher levels in the healthy group).

2.2. Relating mean structures across taxonomic lineages

Here we propose the simplest relationship possible for the mean counts between different levels of a taxonomic structure. Let β_f represent the mean count for some family of bacteria and let $\beta_{gi} i = 1, \ldots, I$ represent the mean counts for a collection of genera that are in this family. The model proposed here assumes simply that

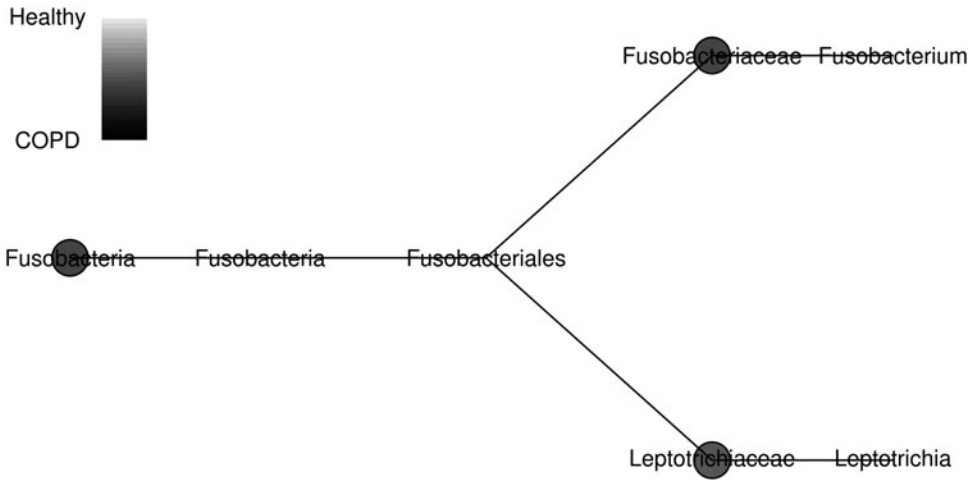

FIG. 5. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for the phylum Fusobacteria (lighter values indicate higher levels in the healthy group).

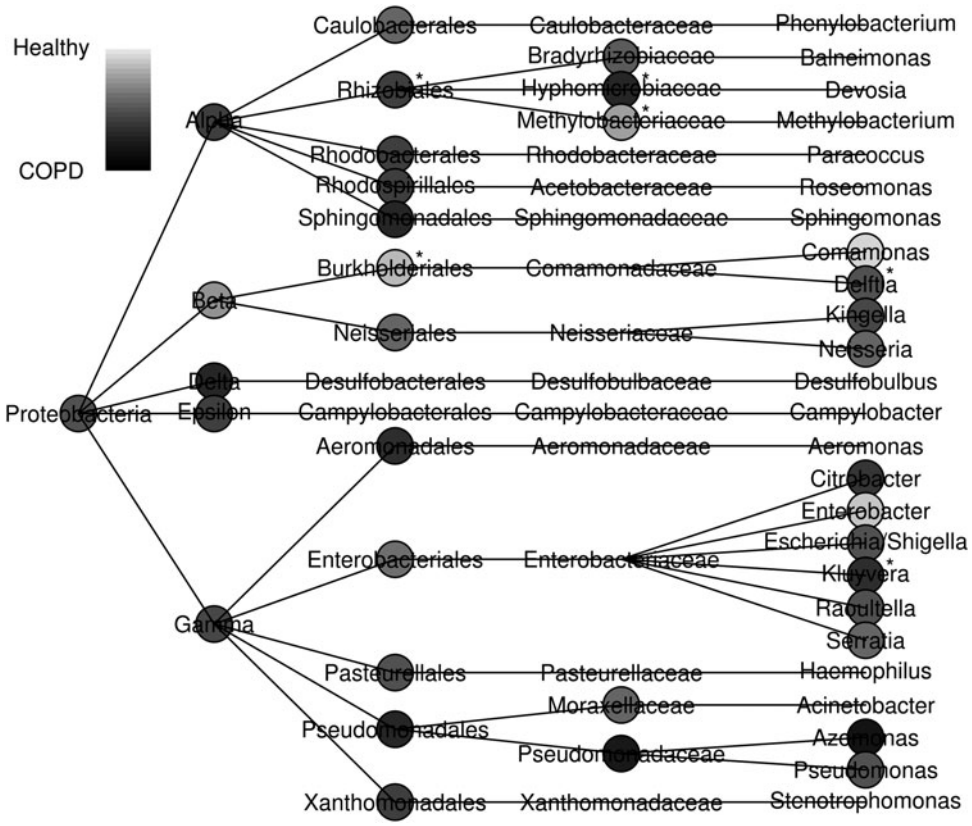

FIG. 6. Posterior medians of the ratio of the mean in the healthy group to the mean in the COPD group for the phylum Proteobacteria (lighter values indicate higher levels in the healthy group). The class names have been truncated (removing ''proteobacteria'').

$$\beta_f = \frac{1}{I}\sum_i \beta_{gi}.$$

We assume the same sort of relationship holds at all levels of our taxonomical structure. If θ_f represents a vector of family level means and θ_g represents a vector of genus level means then we can write

$$\theta_f = M_g \theta_g.$$

Our analysis pipeline provides genus level data for each subject, so we can use the sample mean in conjunction with M_g to obtain an estimate of all family level mean counts. Note that M_g can be recovered from the observed taxonomy in the data set, hence conditional on our observed taxonomy this matrix is known. In fact, using the notation of section 2.1, we find that

$$M_g[i,j] = \frac{\mathcal{A}_g[i,j]}{\sum_j \mathcal{A}_g[i,j]}.$$

While strictly speaking M_g is stochastic, from the perspective of testing for differences between multiple patient groups treating it as known is much like conditioning on a set of covariates as is commonly done in regression modeling (i.e., the taxonomy is ancillary).

Using this same approach we can obtain unbiased estimates of the means at any level of our taxonomic structure. Some of these means will not be identifiable when we consider multiple levels in our taxonomy simultaneously, as we saw in section 2.1, hence we need to select those from our vector of linearly transformed genus level means. However, this selection process can be accomplished by applying another linear transformation to θ_f, which we denote S_f. Suppose that we have data from multiple subjects and let $\bar{y}$ represent the n_g-vector of sample means. Conditional on M_g we can then obtain

unbiased estimates of the family level means, and then use these to obtain means at other levels using the matrices M_f, M_o, and M_c. We then select the identifiable taxa using a set of matrices that select the identifiable components: S_g, S_f, S_0, and S_c. We can then write our estimate of all identifiable taxa means as

$$\hat{\theta} = A\bar{y}$$

where

$$A = \begin{pmatrix} S_g \\ S_f M_g \\ S_o M_f M_g \\ S_c M_o M_f M_g \\ M_c M_o M_f M_g \end{pmatrix},$$

and $\hat{\theta}$ consists of first the identified genera, then the identified families and so on.

2.3. Testing for differences in the means of the identified taxa

If there are two groups of patients, we can use the previous expression to develop a test statistic for testing for a difference between groups at the level of each identified taxon. To this end, let $\bar{y}_i$ for $i = 1, 2$ represent the n_g-vectors of sample means for all genera in our two independent groups. We can then base a test on $A(\bar{y}_1 - \bar{y}_2)$ noting that

$$\text{Var } A(\bar{y}_1 - \bar{y}_2) = A(\text{Var}\,\bar{y}_1 + \text{Var}\,\bar{y}_2)A'.$$

While this can form the basis for a procedure for testing for differences between groups, this approach is not very powerful in our experience.

2.4. A Bayesian model-based approach

While the method from the previous section can be used to compute test statistics, we can develop more powerful tests by taking advantage of recent advances in the analysis of RNA-Seq data. There are two features of these approaches that offer an opportunity to develop improved estimates: likelihood-based inference and hierarchical modeling. We adopt a similar strategy here, except we utilize a fully Bayesian approach to inference. The advantage of this is that we don't need to substitute point estimates for unknown parameters and our inference fully averages over the uncertainty associated with estimation of nuisance parameters. Moreover there is no need to assume that the sample size is sufficiently large for an approximation to hold as Bayesian inference is exact.

2.4.1. The probability model and computation. Let y_{ij} denote the number of reads that map to the 16S rRNA DNA sequence for subject i and genus j. We assume that y_{ij} are independently distributed according to the negative binomial distribution with the mass function

$$p(y_{ij}|\mu_j, \eta_j) = \frac{\Gamma(\eta_j + y_{ij})}{\Gamma(\eta_j)\Gamma(y_{ij} + 1)} \left(\frac{\eta_j}{\eta_j + \mu_j}\right)^{\eta_j} \left(\frac{\mu_j}{\mu_j + \eta_j}\right)^{y_{ij}}$$

parameterized so that the mean of y_{ij} is μ_j and its variance is $\mu_j(1 + \mu_j/\eta_j)$. This differs from the parameterization used by Robinson and Smyth (2007) in that what they call the dispersion we have parameterized as $1/\eta_j$; however, we will refer to this as the dispersion parameter.

To develop a more powerful procedure we now consider methods that utilize shrinkage estimators of the dispersion parameters. To this end we assume that the dispersion parameters are all distributed according to a gamma distribution with parameters α and λ. For comparing two groups we allow for group level mean parameters but assume that the dispersions do not depend on group identity (although this is easy to modify).

In addition to specifying the probability model for the observed data we must also specify prior distributions for all parameters. Here we use prior distributions that are all proper. We specify conditionally conjugate priors (Gelman, 2006) for the means with the following form

$$p(\mu|\eta) = \frac{\Gamma(a+b-2)}{\Gamma(a-2)\Gamma(b)}\eta^{-1}\left(\frac{\eta}{\mu+\eta}\right)^{a-1}\left(\frac{\mu}{\mu+\eta}\right)^{b-1}$$

for $a>2$ and $b>0$. We then find that $\mathrm{E}[\mu|\eta] = \eta\frac{b}{a-3}$ provided $a>3$ and $\mathrm{Var}[\mu|\eta] = \eta^2\frac{b^2-3b+ab}{(a-3)^2(a-4)}$ provided $a>4$. The results are not very sensitive to the values of a and b provided that a is not too much larger than 2; we use the value 3.1 here. With $a=3$ the prior is so vague that it doesn't even have a finite mean. We set $b=100$ although the same results are obtained with a wide range of values for b. We can interpret these prior specifications as assuming the mean is 1000 times the size of the dispersion parameters. Priors for α and λ were constructed using data from the saliva samples from the human microbiome project (the prior mean for the dispersion parameters is in the interval $[5.0\times10^{-5}, 5]$ with 99% probability).

To conduct inference for the model we adopted a Bayesian approach and used the Metropolis algorithm. Sample proportions from the Markov chain were used to conduct inference. To assess convergence of the algorithm we ran multiple chains and monitored the $\sqrt{R}$ statistic of Gelman and Rubin (1992). Using the simulated genera level μ we use the transformation from section 2.2 to obtain simulated values for all estimable effects: each sampled vector of group specific means is premultiplied by the matrix A to obtain sampled values of all identifiable taxa means. Software is online under ''Metagenomic taxonomic decomposition software.''

3. SIMULATIONS

Next we examine some simulations to investigate the operating characteristics of the proposed method. We use the taxonomies and the group sizes we observed in our motivating data as the basis for these simulations; we just simulate data so that there are differences in the means across groups at some taxonomic level. For each scenario we simulate 1000 data sets and use our MCMC algorithm on each of these datasets. For each identified taxon we then compute how frequently we find that the posterior probability of no difference is less than 0.05. For situations with simulated differences between groups we used a relatively large difference in the means to make the results clear in the figures. We present the results for six simulation scenarios:

1. no difference at any taxonomic level
2. a difference at the genus level for the situation where there is a single genus observed from some phylum (Deinococcus-Thermus, Fig. 3)
3. a difference at the genus level with an identified genus level effect with all intermediates identified (Pseudomonas, Fig. 6)
4. a difference at the genus level with an identified genus level effect without all intermediates identified (Tannerela, Fig. 2)
5. a difference at the family level with an identified family level effect and with all intermediates identified but no identified offspring (Lactobacillaceae, Fig. 4)
6. a difference at the family level with an identified family level effect with all intermediates identified and identified offspring (Bacillaceae, Fig. 4)

The results are displayed in Figure 7 (the horizontal lines are at 0.05). The taxa are ordered from left to right in terms of decreasing courseness (i.e., phyla are on the left and genera are on the right). Note that in the null case (upper left panel) the method does not identify any taxa as differing based on the posterior probability of a difference—all taxa have small values for this posterior probability (the mean across all taxa was 0.04 with a maximum of 0.055). In the middle upper panel the results are shown for case 2: the method notes that there is substantial posterior probability for a difference at the phylum level corresponding to this genus. In case 3 there is substantial posterior probability that the genus differs between groups, and there is also evidence for differences at the family, class, order, and phylum level for those taxa that include this genus (although the posterior probability trails off as we move higher up the taxonomy and is very hard to detect at the phylum level). Case 4 shows that when intermediaries are not identifiable one can still detect differences at the other courser taxonomic levels. Case 5 illustrates that we do detect the family difference even though there are no genus differences, while case 6 shows that we can detect differences at the genus level for the family where there is a difference, but the posterior probability of a difference at the family level

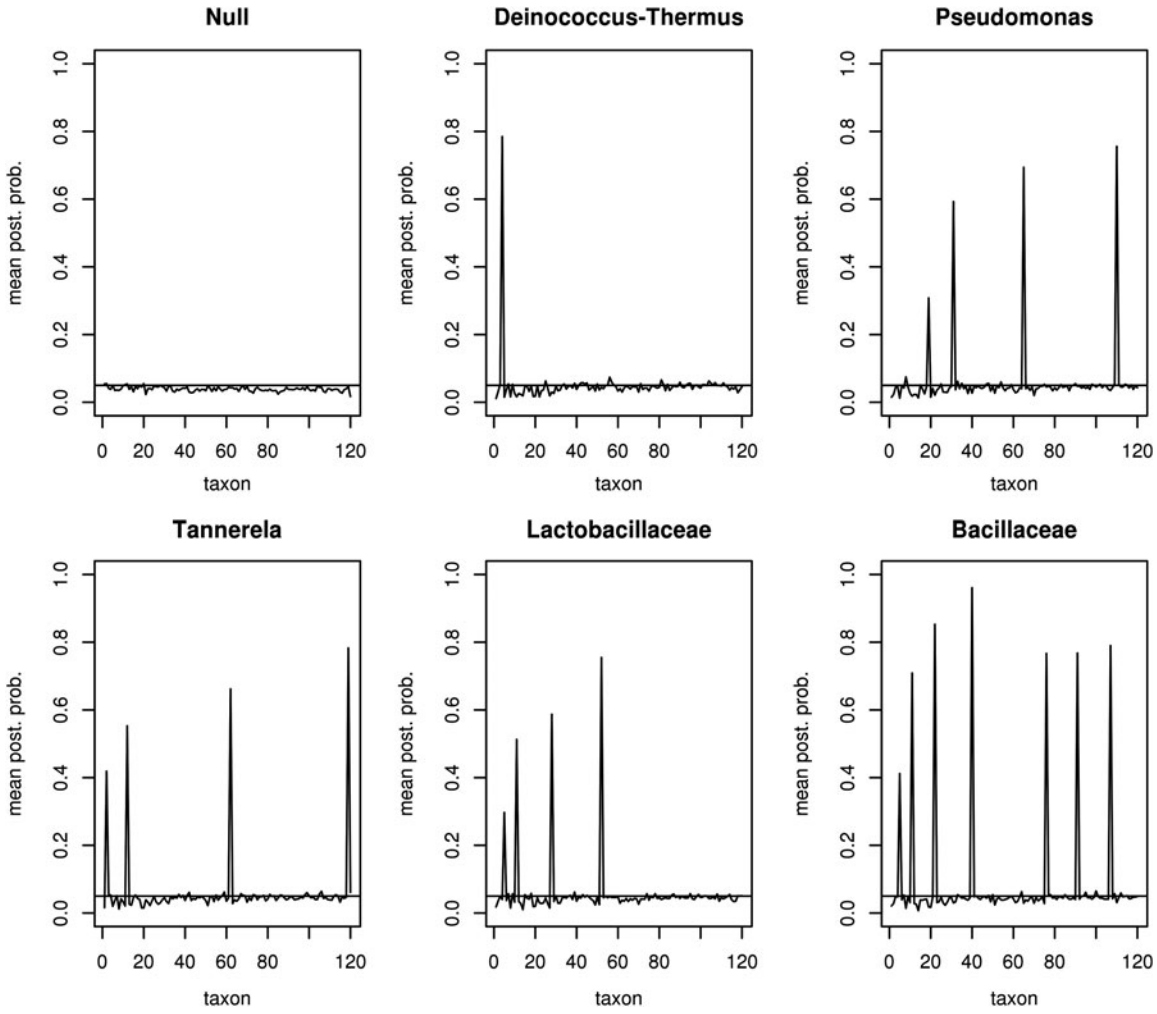

FIG. 7. Simulation results for six scenarios. Each plot shows the frequency with which the posterior probability of a difference between two groups exceeds 95% (the horizontal line is at 0.05). Taxa are ordered from left to right by taxonomical courseness (i.e., phyla are on the left and genera on the right).

is higher. Moreover, this last case demonstrates a situation where the difference can readily be seen to be at the family level since the genera all have approximately equal posterior probabilities.

4. RESULTS

For illustration we examined a data set that used bronchoalveolar lavage fluid samples from patients participating in a clinical trial for chronic obstructive pulmonary disease (COPD) (Roth et al., (2006). As described in Pragman et al. (2012), 142 distinct genera were present in at least 1 subject. As many of these genera were only identified in 1 or 2 subjects, and frequently there was just a single read from that genus, we first conducted some nonspecific filtering to increase power (e.g., Bourgon et al., 2010). Here we filtered on the total read count across all subjects. This ranged from 1 (observed for 21 genera) to 82189, and we required this read count to be at least 10. After this filtering step there were 87 genera, 54 families, 29 orders, 16 classes, and 9 phyla.

To test for differences between the two groups using the proposed Bayesian approach we computed the posterior probability that the genus level means were higher in one group than the other and compared these probabilities to the values 0.025 and 0.975. This identifies 51 genera that were found at different levels in the two patient groups (35 of these were also detected using the R package edgeR). Here we found that the joint probability there were differences among the top 29 genera that show differences between groups exceeds 99%.

In an article by La Rosa et al. (2012) the authors proposed an overdispersed multinomial model for microbiomic data. An important feature of the model they proposed was that there was a single parameter that governs the extent of overdispersion for all genera. In contrast our model has an overdispersion parameter for each genus but smooths them all toward their mean in a data dependent manner. We find substantial evidence for different amounts of overdispersion in our data set. To get a sense of the extent to which overdispersion differs among genera, we looked at every pair of genera in our data set and computed the posterior probability that the dispersion parameter for one genus was less than another genus for every pair of genera (with 87 genera there are 3741 pairs). We then examined how many pairs are such that this posterior probability was less than 5% or exceeded 95%. By this metric we found that 42% of the pairs had different levels of overdispersion. Thus, at least for this dataset, we found substantial evidence for differences in the extent of overdispersion among genera.

4.1. Differences in the identified taxa

We then used the algorithm described in section 2.1 to determine the number of estimable taxon effects in our data set. Using the algorithm described there we found that there were 50 estimable genera, 33 estimable families, 18 estimable orders, 10 estimable classes, and 9 estimable phyla. Using the linear transformation of the observed sample means presented in section 2.2 we then tested for differences between groups using the method of section 2.3; however, this failed to detect any differences. The results from using the Bayesian approach provided strong evidence for many differences among groups at all taxonomic levels. Of the 120 estimable effects we found that 63 were lower in the healthy samples while 11 were higher in the healthy samples if we again determined there was a difference by examining the posterior probability of a difference in the means. We graphically displayed the results as dendograms with circles that represent the relative mean levels across the groups in Figures 2–7 with darker values indicating higher levels in the COPD population (and asterisk for taxa that differ). These figures indicate that the largest differences between the patient populations are largely restricted to points rather far down the trees without large differences among the phyla.

5. DISCUSSION

There is a need for methods that allow one to formally assess the differences in the microbiota of patient populations at multiple taxonomic levels simultaneously. Here we show how to determine the total number of estimable taxon level effects and how to obtain unbiased estimates of these effects. We then demonstrated how to use fully Bayesian methods to obtain superior inference by using a hierarchical model for the dispersion parameters. This model detected many differences between the COPD patients and the healthy controls at multiple taxonomic levels. This approach allows for exact inference in a setting where competing frequentist approaches must rely on large sample approximations that typically are questionable in contemporary microbiomic data sets.

ACKNOWLEDGMENTS

Thanks to John Connett, PhD, and Robert Wise, MD, along with the FORTE study, for supplying the samples. We thank the University of Minnesota Biomedical Genomics Center and the W.M. Keck Center at the University of Illinois Urbana-Champaign for assistance with the sequencing.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Anders, S., and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.

Bourgon, R., Gentleman, R., and Huber, W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* 107, 9546–9551.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* 1, 515–533.
Gelman, A., and Rubin, D. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Stat. Sci.* 7, 457–511.
La Rosa, P.S., Brooks, J.P., Deych, E., et al. 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* 7, e52078.
Ley, R.E., Turnbaugh, P.J., Klein, S., et al. 2006. Human gut microbes associated with obesity. *Nature* 444, 1022–1023.
Pragman, A.A., Kim, H.B., Reilly, C.S., et al. 2012. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS One* 10, e47305.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. 2009. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.
Robinson, M.D., and Smyth, G.K. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887.
Robinson, M.D., and Smyth, G.K. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332.
Roth, M.D., Connett, J.E., D'Armiento, J.M., et al. 2006. Feasibility of retinoids for the treatment of emphysema study. *Chest* 130, 1334–1345.

Address correspondence to:
Dr. Cavan Reilly
Division of Biostatistics
University of Minnesota
A448 Mayo Building, MMC 303
420 Delaware Street SE
Minneapolis, MN 55455

E-mail: cavanr@biostat.umn.edu