Application of zero-inflated negative binomial mixed model to human microbiota sequence data

Rui Fang *[,1], Brandie D. Wagner[1,2,3], J. Kirk Harris[2,3], Sophie A. Fillon[4]

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Aurora, Colorado, USA
[2] Department of Pediatrics, Division of Pulmonology, University of Colorado Denver, School of Medicine, Aurora, Colorado, USA
[3] University of Colorado Microbiome Research Consortium (MiRC), Aurora, Colorado, USA
[4] Department of Pediatrics, Section of Gastroenterology, Hepatology and Nutrition, Digestive Health Institute, Gastrointestinal Eosinophilic Diseases Program, Mucosal Inflammation Program, Children's Hospital Colorado, University of Colorado Denver, School of Medicine, Aurora, Colorado, USA

Abstract

Identification of the majority of organisms present in human-associated microbial communities is feasible with the advent of high throughput sequencing technology. However, these data consist of non-negative, highly skewed sequence counts with a large proportion of zeros. Zero-inflated models are useful for analyzing such data. Moreover, the non-zero observations may be over-dispersed in relation to the Poisson distribution, biasing parameter estimates and underestimating standard errors. In such a circumstance, a zero-inflated negative binomial (ZINB) model better accounts for these characteristics compared to a zero-inflated Poisson (ZIP). In addition, complex study designs are possible with repeated measurements or multiple samples collected from the same subject, thus random effects are introduced to account for the within subject variation. A zero-inflated negative binomial mixed model contains components to model the probability of excess zero values and the negative binomial parameters, allowing for repeated measures using independent random effects between these two components. The objective of this study is to examine the application of a zero-inflated negative binomial mixed model to human microbiota sequence data.

Key words: microbiota, negative binomial, zero-inflation

* Corresponding author: Rui Fang, Address: 12477 E. 19th Avenue, Aurora, CO 80045, Phone: +01-303-724-458, E-mail: rui.fang@ucdenver.edu.

1. Introduction

The human microbiota consists of communities of microorganisms that inhabit the human body. These communities can significantly affect many aspects of human physiology. For example, in healthy individuals the microbiota provides a wide range of metabolic functions that humans lack, making their presence advantageous (Gill et al., 2006; Sommer and Backhed, 2013). In addition, altered microbiotas are associated with a number of chronic inflammatory disorders including autoimmunity and allergic disorders (Aas, Gessert and Bakken, 2003), obesity and diabetes (Devaraj, Hemarajata and Versalovic, 2013). One analytic goal of microbiota studies is to compare the bacterial communities across groups. The human microbiome project endeavors to apply this to human associated communities in order to identify bacteria that either adversely affect or promote health (Group et al., 2009).

Bacteria are generally identified using culturing methods, which assume prior knowledge of the growth condition required for isolation. With the advent of DNA-based sequencing technology, identification of organisms present in the community can now be performed in parallel, which results in significant efficiency compared to culture. The process starts with the collection of human-associated samples for DNA extraction. The DNA is used to amplify 16S PCR gene sequences that are taxonomically informative, and data is collected using next generation sequencing technologies. These data are compared to reference databases to determine organism identity (taxonomic category). The number of sequences for a single taxon is then counted for each sample for comparison within a study.

Microbiota sequence data are high-dimensional with added complexity. They consist of non-negative, highly skewed sequence counts with a large number of zeros. The number of zeros in the dataset is a result of combining samples with different bacterial composition (e.g. disease versus controls or different locations in one subject). Samples collected from different groups can result in unique organisms, and if an organism is detected in one but not another sample, insertion of a zero count is performed. The absence of a count for an organism can be due to the fact that the organism simply isn't present in the sample (true zeros) or that the organism is present but sufficiently rare such that it does not appear in the sequence collection (false zeros). In addition, the number of total sequences varies from sample to sample. This is a result of an inability to specify exactly the number of sequences to be measured on a sample using currently available technology. Note the number of sequences for a given sample is not associated with any biological feature of the sample, and thus should have a random distribution across samples. A common approach to account for the variation in the total number of sequences, is the conversion of the sequence counts to relative abundance (taxon counts/total counts) within a particular sample (Wagner, Robertson and Harris, 2011).

The zero-inflated negative binomial (ZINB) distribution is a mixture of a binary distribution that is degenerate at zero and an ordinary count distribution such as negative binomial. The negative binomial regression can be written as an extension of Poisson regression and it enables the model to have greater flexibility in modeling the relationship

between the conditional variance and the conditional mean compared to the Poisson model. The binary distribution captures the excess number of zeros, which exceed those predicted by the negative binomial distribution.

Often because of a hierarchical study design or data collection where the observations are either clustered or outcomes are collected repeatedly from individual subjects, zero-inflated regression models are extended to include random effects. The random-effects model accounts for the correlation among the repeated measures within a subject.

Few microbiota studies address the additional source of variability attributed to a repeated measures design, however, more recently, authors have begun to utilize methods appropriate for this study design (Smith et al., 2012; Wu et al., 2013). In this work, we apply a generalized mixed model approach to taxa of interest to directly estimate the within subject correlation in a microbiota study with a repeated measures design. Moreover, the application of a zero-inflated distribution to microbiota data is novel.

2. Method

2.1 Motivating example

The dataset is from a study in which pediatric individuals with normal esophageal mucosa provided samples to capture esophageal microbiota. The different sample types include the "gold standard" mucosal biopsy and the minimally invasive capsule-based string collection, the Enterotest™ named Esophageal String Test in that study (EST). Additionally, an oral string segment and nasal cavity swabs were collected for comparison. All of the 15 subjects enrolled in this study had normal histological biopsy findings. Most of the samples had adequate bacterial load for data generation, and only two nasal swabs did not amplify (i.e., 13 nasal swabs and 15 oral strings, ESTs and biopsies). Bacterial ribosomal RNA gene amplification products from mucosal biopsies and from the nasal cavity, oral cavity and EST were produced and sequenced. Additional details of the study and the data generation process have been previously published (Fillon et al., 2012). The aim of the study was to compare the esophageal microbiota identified from biopsies and ESTs, and to show if there are highly similar profiles between the EST and biopsy samples that were different from samples collected from the nasal and oral cavity (Fillon et al., 2012).

2.2 Ethics statement

All human species were collected under approval of the Colorado Multiple Institutional Review Board (COMIRB). Written informed consent and HIPAA authorization were obtained from all participants or from parents or legal guardians of participants younger than 18 years. Assent was obtained from all participants under 18 years.

2.3 Zero-inflated negative binomial mixed model

The zero-inflated negative binomial (ZINB) (WH, 1994; Yau, 2003) model assumes there are two distinct data generation processes, which is determined with the use of a Bernoulli trial. With probability π, the response of the first process is a zero count, and with probability of $(1-\pi)$ the response of the second process is governed by a negative binomial with mean λ and can also generate zero counts. The overall probability of zero counts is the combined probability of zeros from the two processes. Thus, a ZINB model for the response Y can be written as:

$$P(Y=0) = \pi + (1-\pi)(1+k\lambda)^{-1/k}$$

$$P(Y=y) = (1-\pi)\Gamma(y+1/k)(k\lambda)^{y}/[\Gamma(y+1)\Gamma(1/k)(1+k\lambda)^{y+1/k}], \quad y=1,2,\ldots$$

Moghimbeigi *et al.* (Moghimbeigi. A, 2008) developed multi-level ZINB regression for modeling over-dispersed count data with extra zeros. Let Y_{ij} ($i=1,2,\ldots m$; $j=1,2,\ldots n_i$ and $\sum_{i=1}^{m} n_i = n$ gives the total number of observations) be the response variable for the i-th individual subject with j-th repeated measurement, a ZINB mixed model is defined as follows:

$$\log(\lambda_{ij}) = \mathbf{X_{ij}'\beta} + u_i$$

$$\text{logit}(\pi_{ij}) = \mathbf{Z_{ij}'\gamma} + v_i$$

where $\mathbf{X_{ij}}$ and $\mathbf{Z_{ij}}$ are vectors of covariates for the negative binomial and the logistic components, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding vectors of regression coefficients.

An offset, the natural logarithm of the total sequence counts, $\log(\text{Total}_{ij})$, was added into the linear predictor function for the negative binomial component to account for the variable number of sequences per sample inherent in microbiota sequence data. That is, $\log(E(Y_{ij})) = \mathbf{X'_{ij}} + u_i + \log(\text{Total}_{ij})$. This can be simplified to show that $\log(E(Y_{ij})/\text{Total}_{ij}) = \mathbf{X'_{ij}} + u_i$. The left side of this equation is, therefore, modeling the log of the relative abundance as the outcome, assuming the total sequence count is considered a fixed value rather than a random variable. Note that the parameter π_{ij} is not affected by the total sequence count.

Here, u_i and v_i are the random intercepts and they are assumed to be independent and follow the bivariate normal distribution as

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix}\right).$$

For simplicity, we assume the independence of the two random effects. Although this is not a necessary assumption, it is commonly used in the previous literature regarding ZIP/ZINB with random effects (Hur K, 2002; Yau and Lee, 2001). Besides, the process

that generates the false zeros (dependent on sequencing depth) is independent of the process that generates the sequence counts.

A ZINB mixed model was applied to each taxa individually to compare the esophageal microbiota to the other three sample types from the motivating dataset. The expected relative abundances are estimated by calculating the overall mean $E(Y) = (1-\pi)\lambda = \exp(\mathbf{X'\beta})/[\exp(\mathbf{Z'\gamma})+1]$. Point estimates and p-values for the difference between sample types were calculated using linear contrasts of the regression parameters. One hundred and eighty-seven different taxa were identified. Four of these taxa, *Gemella*, *Leptotrichia*, *Aggregatibacter* and *Streptobacillus*, were used as examples to represent the range of the proportion of zero counts. All analyses were performed via the NLMIXED procedure using `SAS 9.3 software (SAS Institute Inc.: Cary, NC, 2011)`. All corresponding code is included in the Appendix.

3. Results

The ZINB mixed model fit was graphically inspected and reasonable describes the empirical data distribution for the four example taxa (**Figure 1**). The model fit for *Aggregatibacter* resulted in a non-positive definite Hessian matrix; the parameter estimates for this organism is therefore not presented. The parameter estimates for the remaining three organisms are given in **Table 1**. The expected relative abundance in the biopsy samples for *Gemella* and *Leptotrichia* is around 1%, whereas *Streptobacillus* is close to 0. In the EST samples, the relative abundance for *Streptobacillus* is slightly larger at 0.3% and significantly smaller for *Leptotrichia* (0.9% versus 0.3%, p-value = 0.05). *Leptotrichia* also differed between EST and oral samples (p-value = 0.05), and between nasal and oral samples (p-value = 0.04) but not between EST and nasal (p-value = 0.68). No other differences were observed across sample types.

The sigmas in **Table 1** correspond to the estimated standard deviations for the normally distributed random subject effects. The variances of the random effect for the zero-inflated part of the model, v_i, was significant, indicating that the probability of a false zero count was different among the subjects. The random effect variance for the count distribution, u_i, was also significant, meaning that some subjects had higher sequence counts than others. Also, as a sensitivity analysis, a model that included correlation between the random effects was estimated. This correlation was not significant, thus providing evidence that the two processes (false zeros and the count process) are independent.

Examination of the full dataset (187 taxa) yielded estimates for 86 taxa where the mixed ZINB models successfully converged. However, the final Hessian matrix was not positive definite for 64 of the models. For those models that could not be estimated, the majority of the taxa had a large percentage of zero counts with either extremely small or large non-zero counts. Comparisons across the sample types were similarly performed as described above across all taxa. Manhattan plots, commonly used in genetic studies, were used here to display the magnitude of the p-values for each comparison ordered by taxonomy line, and color-coded by phylum. Organisms close together, within a phylum,

denote closer phylogenetic relationship. As shown in the Manhattan plots (**Figure 2**), few differences were observed in microbiota composition between from ESTs and biopsies. These results support the use of the EST to sample the microbiota as compared to the "gold standard", the mucosal biopsy. Microbiota captured in the nasal cavity samples revealed differences from EST and oral samples. These results suggest that each microenvironment harbors specific taxa that distinguish the nasal and oral sites from EST and biopsy.

4. Discussion

The distributions of the microbial sequence counts are highly skewed, non-negative and have a large proportion of zeros, for which commonly used statistical approaches may not be appropriate. The large proportion of zeros is intrinsic to the creation of the dataset rather than the data generating process itself, where the dataset contains sequence counts for organisms that were observed in at least one sample, if a particular organism was not observed in a sample it is given a zero value. Therefore, when comparing sequence counts across groups with diverse communities, a large numbers of zero counts are expected. Our working hypothesis is two underlying processes explain the absence of a count for an organism (true and false zeros).

In this paper, the ZINB mixed model was described. This model is useful for analysis of over-dispersed count data with an excess of zeros and repeated measures. This model based approach can additionally be easily extended to include potential confounders as covariates and to test association with continuous variables. The application of the ZINB to the three selected organisms from the microbiota data demonstrated the usefulness of this approach when applied to organisms of interest. However, given the complexity of the model, we are not able to easily apply it to all organisms and it requires adaption and guidelines for high-dimensional applications. The majority of models that did not converge were due to an inability to estimate the relatively large number of parameters with the available data. It is more likely that this model will address more focused questions related to a small subset of organisms of clinical interest.

To assess the effects of misspecification of random effect distributions in the two parts of ZINB regression model, other distributional assumptions apart from normality could be considered in future research. In our study, we separately fit the models to the organisms identified thus ignoring potential correlation among organisms. We are interested in extending the modeling to pairs of organisms multivariately or implementation of a multi-level (two-fold random effects) zero-inflated model.

5. Summery

We have illustrated the novel application of a ZINB model with random effects to a microbiota dataset with a repeated measures design. The range of distributions present for the individual taxa in a microbiota dataset additionally provides insight into when the use of a zero-inflated approach is appropriate.

Table 1 Parameter estimates (standard errors) from ZINB regression model with random effects for three organisms selected from the motivating dataset.

		Gemella	*Leptotrichia*	*Streptobacillus*
Intercept	β_0	-4.68 (0.25)	-4.48 (0.27)	-4.98 (0.62)
String	β_1	0.15 (0.33)	-1.29 (0.38)	1.15 (0.58)
Nasal	β_2	-0.03 (0.40)	-0.89 (0.43)	-3.86 (1.02)
Oral	β_3	0.50 (0.33)	0.002 (0.35)	-0.74 (0.81)
Var (u)	σ_u	-0.39 (0.18)	-0.30 (0.40)	0.66 (0.49)
ZI intercept	γ_0	-17.17 (1540.76)	-1.24 (0.70)	3.87 (2.09)
ZI string	γ_1	-4.75 (16061)	-1.34 (2.20)	-1.87 (1.69)
ZI nasal	γ_2	16.03 (1540.66)	1.23 (0.92)	-7.92 (6.35)
ZI oral	γ_3	-4.29 (12174)	0.27 (0.94)	-2.28 (1.93)
Var (v)	σ_v	0.39 (61.12)	2.15E-9 (0.69)	3.10 (1.73)
Over-dispersion	k	0.58 (0.16)	0.36 (0.28)	0.22 (0.67)

Figure 1 Empirical and fitted ZINB distributions of the human microbiota sequence data for each of four organisms.

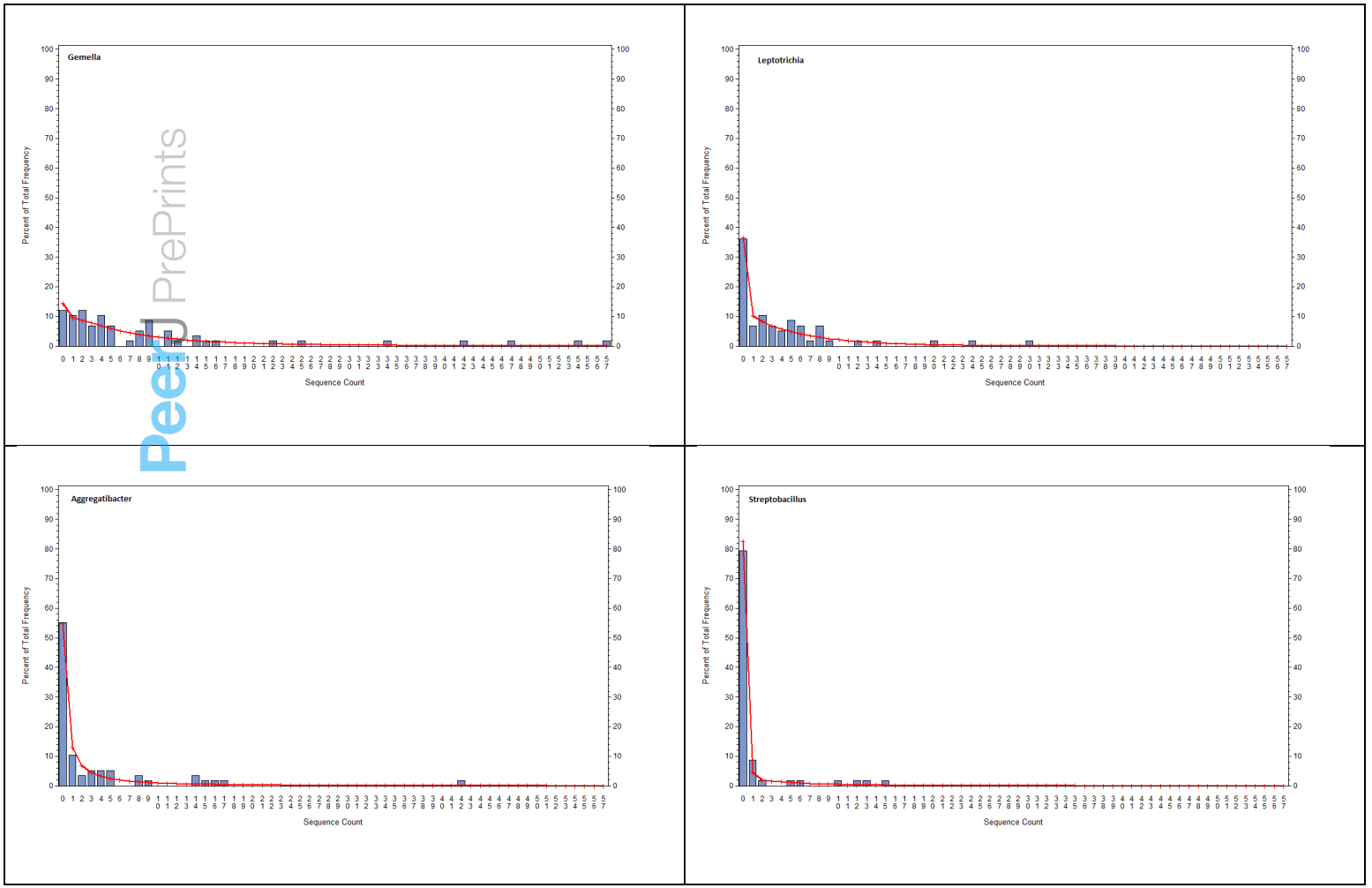
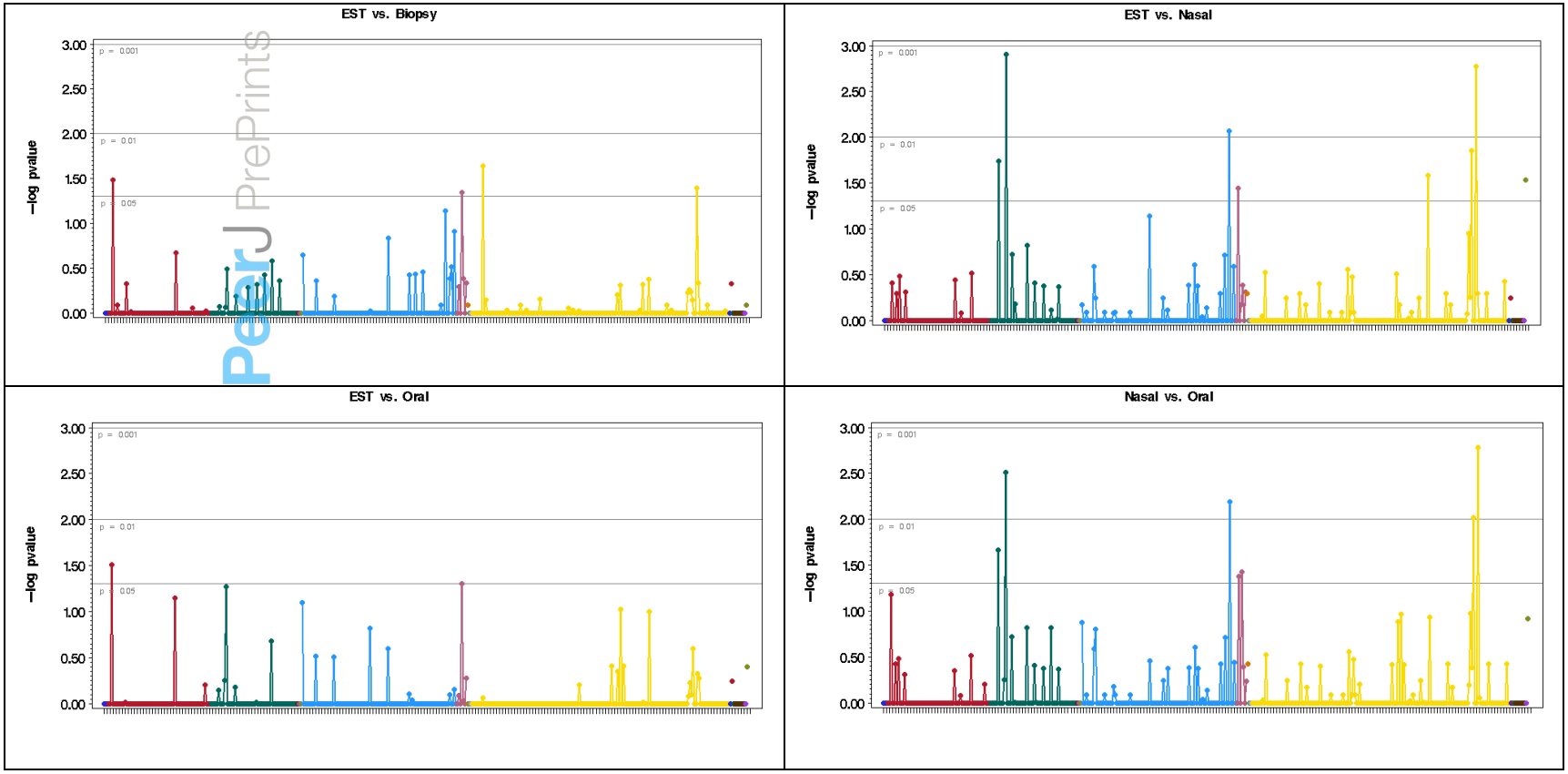


Figure 2 Manhattan plots for the comparisons across all taxa. The y-axis displays the negative log of the p-value; hence higher values indicate increased statistical significance. The reference lines in gray are included to designate the usual critical values. The Manhattan plot is ordered by taxonomy line and the colors correspond to different phyla. For the models that did not converge, the p-values were set to 1.00.



Acknowledgements

The authors recognize the support of REDCap for subjects' database supported by the National Center for Research Resources, the National Center for Advancing Translational Sciences, National Institutes of Health, and Colorado CTSI Grant Number UL1 RR025780. This work was supported by the American Partnership for Eosinophilic Disorders (APFED) Junior Faculty HOPE Research Grant (SF); National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, Colorado CTSI Grant Number KL2 TR000156 (SF) and (RF).

References

Aas, J., Gessert, C. E., and Bakken, J. S. (2003). Recurrent Clostridium difficile colitis: case series involving 18 patients treated with donor stool administered via a nasogastric tube. *Clinical Infectious Disease* **36**, 580-585.

Devaraj, S., Hemarajata, P., and Versalovic, J. (2013). The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical Chemistry* **59**, 617-628.

Fillon, S. A., Harris, J. K., Wagner, B. D., *et al.* (2012). Novel device to sample the esophageal microbiome - the esophageal string test. *PLoS One* **7**, e42938.

Gill, S. R., Pop, M., Deboy, R. T., *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355-1359.

Group, N. H. W., Peterson, J., Garges, S., *et al.* (2009). The NIH Human Microbiome Project. *Genome Research* **19**, 2317-2323.

Hur K, H. D., Henderson W, Khuri S, Daley L (2002). Modeling clustered count data with excess zeros in health care outcome research. *Health Services and Outcomes Research Methodology* **3**, 5-20.

Moghimbeigi. A, Eshraghian ME., Mohammad. K, Mcardle. B (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics* **35**, 1193-1202.

Smith, B. C., McAndrew, T., Chen, Z., *et al.* (2012). The cervical microbiome over 7 years and a comparison of methodologies for its characterization. *PLoS One* **7**, e40425.

Sommer, F., and Backhed, F. (2013). The gut microbiota--masters of host development and physiology. *Nature Reviews Microbiology* **11**, 227-238.

Wagner, B. D., Robertson, C. E., and Harris, J. K. (2011). Application of two-part statistics for comparison of sequence variant counts. *PLoS One* **6**, e20296.

Wu, X., Berkow, K., Frank, D. N., Li, E., Gulati, A. S., and Zhu, W. (2013). Comparative analysis of microbiome measurement platforms using latent variable structural equation modeling. *BMC Bioinformatics* **14**, 79.

Yau, K. K., and Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine* **20**, 2907-2920.

Yau, K. K., Wang, K. and Lee A. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal* **45**, 437-452.

Appendix

SAS code

```
%macro ZINB;
/* start values */
proc countreg data=rui.seqdata;
where seq=&j;
model seq_count=string nasal oral/dist=zinb offset=ltotal;
zeromodel seq_count ~ string nasal oral/link=logistic;
ods output ParameterEstimates=pe;
run;

proc sql;
select estimate as b0 into: b0
     from pe where Parameter='Intercept';
select estimate as b1 into: b1
     from pe where Parameter='string';
select estimate as b2 into: b2
     from pe where Parameter='nasal';
select estimate as b3 into: b3
     from pe where Parameter='oral';
select estimate as c0 into: c0
     from pe where Parameter='Inf_Intercept';
select estimate as c1 into: c1
     from pe where Parameter='Inf_string';
select estimate as c2 into: c2
     from pe where Parameter='Inf_nasal';
select estimate as c3 into: c3
     from pe where Parameter='Inf_oral';
select estimate as k into: k
     from pe where Parameter='_Alpha';
quit;

/* independent random effects */
```

```
proc nlmixed data=rui.seqdata tech=newrap;
where seq=&j;
parms b0=&b0. b1=&b1. b2=&b2. b3=&b3. c0=&c0. c1=&c1.
c2=&c2. c3=&c3. k=&k. su=1 sv=1;
eta = b0 + b1*string + b2*nasal + b3*oral + ltotal + ui;
lambda = exp(eta);
eta_p = c0 + c1*string + c2*nasal + c3*oral + vi;
p0 = 1/(1+exp(-eta_p));

/* define ZINB log likelihood */
if seq_count=0 then ll = log( p0 + (1-
p0)/(1+k*lambda)**(1/k) );
else ll = log((1-p0)) + seq_count*log(k*lambda) -
(seq_count+(1/k))*log(1+k*lambda) + lgamma(seq_count+(1/k))
- lgamma(1/k) - lgamma(seq_count+1);
model seq_count ~ general(ll);
random ui vi  ~ normal ([0,0], [su*su, 0, sv*sv])
subject=Subject;
run;
%mend;

%macro driver ();
%do j=1 %to 187;
%ZINB;
%end;
%mend;
```