Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011)
P. Qian, Y. Zhou, C. Rudin, eds.

Methods for the Analysis of Human Microbiome Data

Paul Brooks
Department of Statistical Sciences and Operations Research, and
Center for the Study of Biological Complexity
Virginia Commonwealth University
Richmond, VA
jpbrooks@vcu.edu

Maria Rivera
Department of Biology, and
Center for the Study of Biological Complexity
Virginia Commonwealth University
Richmond, VA

Ed Boone, David Edwards, Qin Wang
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University
Richmond, VA

Patricio LaRosa, Bill Shannon
Washington University in St. Louis School of Medicine
St. Louis, MO

Abstract

High-throughput sequencing technology is currently being applied to study the human *microbiome*. The human microbiome is the community of microbes that populate habitats on and around various body sites. A sample from a subject is processed to create profiles that characterize the population of microbes; the ultimate goal is to find relationships between these microbiome profiles and subject physiology and health. In this paper, we address some of the challenges that arise in the analysis of data generated in a microbiome experiment. First, we discuss methods for visualizing microbiome data. Second, we describe procedures for assessing bias introduced in the sample processing pipeline so that proper estimates of true abundances may be obtained. Finally, we describe a statistical hypothesis test for confirming that groups of samples arise from different distributions.

Keywords: human microbiome; L1-norm principal component analysis; designed mixture experiment; Dirichlet distribution

Introduction

Microbiome experiments are among the newest in generating large biological data sets by leveraging advances in sequencing technology. In human experiments, a sample is collected from a habitat such as the skin or saliva, DNA is extracted from organisms in the sample, the bacterial sequences are amplified and distinguished from human DNA, and a profile is generated that gives an indication of what organisms are present. The microbiome is known to play an active role in disease and physiological condition, but the mechanisms of this role are not well understood. The potential knowledge derived from these experiments has wide-ranging impact in the practice of medicine. Large experiments characterizing "healthy" microbiome profiles [5] and those corresponding to specific disease states [11, 4] are underway. At the end of each experimental pipeline, high-dimensional data sets demand sophisticated analysis techniques.

Data with large numbers of variables and small numbers of observations is now commonplace due to microarray and genome-wide association study (GWAS) experiments. Many techniques have been proposed to overcome the challenges associated with such short-and-wide data sets. In this paper, we address some of the unique challenges presented by microbiome data. We focus on a particular type of sequence data generated based on the 16S gene. This gene is ubiquitous in prokaryotic organisms and yet species usually have unique sequences. Often microbiome experiments focus on extracting a particular region of 16S rRNA and apply a taxonomic classification of the sequences based on a reference database. The result is a data set with one row for each sample, one column for each taxonomic unit, and entries indicating the number of sequences discovered.

Visualizing high-dimensional data such as microbiome data is essential to confirming non-random effects, and is a step in analysis that is often tempting to skip in favor of more sophisticated methods. If a result cannot be visualized, regardless of the method of discovery, then its presence is likely due to random noise rather than a true association. Ecologists have long placed value on two-dimensional representations of high-dimensional data and have contributed to the development of methods for principal coordinates analysis (PCoA). Microbial ecologists have followed suit with microbiome data (see, e.g., [7]). Caporaso et al. [3] develop movies for a three-dimensional representation of microbiome data over time. Segata et al. [10] have contributed other valuable visualizations of effects in microbiome data, including the ability to view data in a cladistic context.

Another basic aspect of analysis of microbiome data that can be overlooked is a proper appreciation of the systems that generate the data. Analysts have a tendency to view microbiome data as a fixed and final data set rather than the outcome of an experiment. In a 16S rRNA experiment, bacterial DNA is extracted and subjected to a process called PCR amplification. This process, and other factors including unknown and varying DNA copy numbers, contributes to results that do not reflect the true quantities in original samples. These nonlinear transformations of the original quantities in the final data mask the true abundances of organisms and complicate analyzing variation. Kanagawa [6] and others have tried to address bias introduced by PCR. Alm [1] highlights problems with applying traditional methods for analyzing variation in microbiome data. To date, a proper accounting of bias introduced at each step of a microbiome experiment has not been reported.

After a microbiome data set has been properly curated, often the goal is to then find patterns that correlate with the physiological and medical conditions of the human subjects. Confirming differences between groups of samples is an essential task in understanding how the microbiome changes with conditions such as the presence and absence of disease. Until recently, only non-parametric approaches for this task have been used. Schloss [9] reviews and compares several of these techniques. La Rosa et al. [8] establish that microbiome profiles do not arise from multinomial distributions and propose a parametric test for confirming differences based on the Dirichlet multinomial distribution. They provide power calculations for a variety of scenarios; these calculations are not easily computed for the non-parametric approaches.

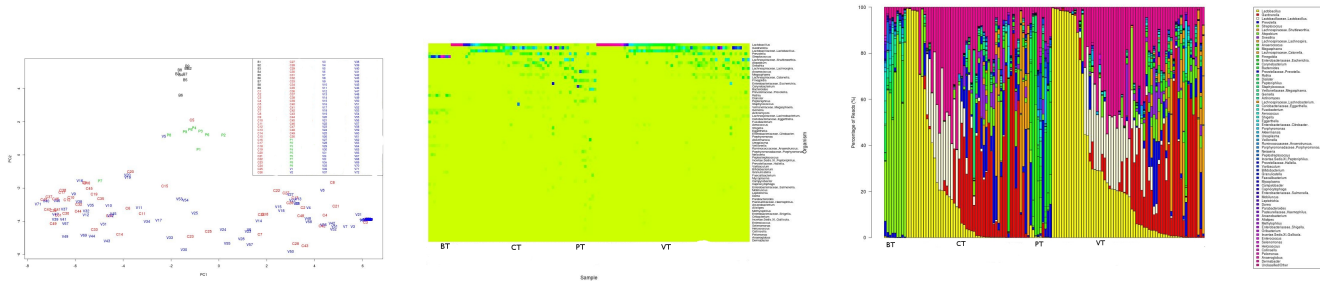
In the remainder of the paper, we give brief overviews of methods of visualization of microbiome data, a proposed experiment for uncovering bias in a microbiome experiment, and a method for confirming that groups of microbiome profiles arise from distinct distributions.

Visualizing Microbiome Data

In this section, we suggest some traditional visualization methods that are well-suited to microbiome data, and describe a recent method for PCoA. In a typical sample, between 5,000 and 30,000 sequences are assigned to taxonomical units. We assume throughout the remainder of the paper that samples have been normalized by dividing by the total number of sequences. A sample profile is therefore a vector of proportions of organisms present.

Figure 1 contains three visualizations of the same data generated at Virginia Commonwealth University [4], (unpublished data, The Vaginal Microbiome Consortium at VCU). The samples were collected from the buccal mucosa, perineum, mid-vaginal wall, and cervix. In Figure 1(a), a two-dimensional visualization of the data is presented that distinguishes the vaginal from non-vaginal samples. Among the vaginal samples, a continuum is observed between those exhibiting profiles commonly-associated with healthy and unhealthy microbiomes. The plot is created by applying a method for PCoA based on a new method for principal component analysis called L1-PCA* [2].

PCoA is the name for a collection of methods for finding the best two-dimensional representation of data. In its simplest form, PCoA is a plot of the scores of observations on the first two principal components of a principal component analysis (PCA). By using PCA, the best-fitting two-dimensional subspace is determined by minimizing the sum of squared distances of points to the fitted subspace. The scores on the first two principal components represent the coordinates of points in the best-fitting two-dimensional subspace. In contrast to PCA, methods for PCoA typically begin with a distance matrix that capture the distances between observations. Other methods can be derived by changing the distance metric. L1-PCA* is a pure L1-norm PCA based on finding successive hyperplanes of best L1 fit [2]. The method exploits a unique relationship between L1-regression and L1-projection. The starting point for L1-PCA* is not a distance matrix but is rather similar to PCA in that it seeks best-fitting subspaces. The use of the L1-norm rather than the L2-norm as in PCA provides a robustness to outlier observations. This robustness is desirable



(a) (b) (c)

Figure 1: Three visualizations of the same data: (a) plot of scores on the first two principal components calculated using L1-PCA* [2], (b) heatmap, and (c) stacked bar plot.

in analysis of biomedical data.

Figures 1(b) and 1(c) are created using traditional visualization methods. Figure 1(b) is a heatmap where each sample is represented by a column. The figure reveals that the buccal mucosa samples are dominated by *Streptococcus* and the vaginal samples are dominated by *Lactobacillus* and *Gardnerella*. The stacked bar plot in 1(c) contains the same information, but relative abundances of organisms are represented by area rather than colors. Each sample is represented by a bar. Presence and absence of organisms is easier to assess using a heatmap, while relative abundance of organisms is easier to observe in a stacked bar or stacked area plot. While these visualization methods are well-established, choosing color palettes and methods for sorting observations to convey information properly can be nontrivial tasks.

The Truth about Metagenomics

Microbiome data sets are the result of a multi-step experiment involving sample collection, DNA extraction, PCR amplification, sequencing, and taxonomic classification. Technical variation is introduced at each step that clouds the signal present in the data. Further, several of the steps introduce bias that transforms the signal. Bias may be due to differential lysability in DNA extraction, primer design, PCR amplification, and sequencing protocol. In this section, we propose a mixture experiment for modeling the bias introduced at each step in microbiome sample processing.

Figure 2 is a stacked bar plot of 39 replicate profiles of a *mock community*; that is, a sample created by mixing known quantities of bacteria. The last bar presents the known proportions of organisms present in the samples. Variation is relatively low among replicate samples, but bias in the processing pipeline masks the true proportions of taxa present in the mock community.

The mixture experiment that we propose involves a subset of a simplex centroid design, where $\sim$50-100 mock communities are created with known concentrations of subsets of 6-10 organisms. Then we fit a model for mixtures (either full or with a subset of interaction terms) that (1) tells us whether organisms work together to amplify their observed abundance, compete to suppress other organisms' observed abundances, or act independently and (2) gives us a model that can be used to recover the "true" proportions in both mock and clinical samples. A model can be fit at each step in the pipeline to determine how much bias is contributed by each procedure.

There are a number of challenges with implementing the results of such a mixture experiment. Some of the organisms in clinical samples have not yet been discovered or their genomes have not been sequenced and their 16S copy number is unknown and therefore they cannot be included in constructing the mock communities. For a particular body habitat, there are usually 200-400 genera expected to be recovered from a population of human subjects. The mixture experiment can only address the bias affecting a small subset of these organisms. Despite these drawbacks, applying the mixture experiment to the "major players" first would go a long way to provide more detailed information about true abundances and elucidate the nature of the effects of the processing steps on the data at the end of the pipeline.

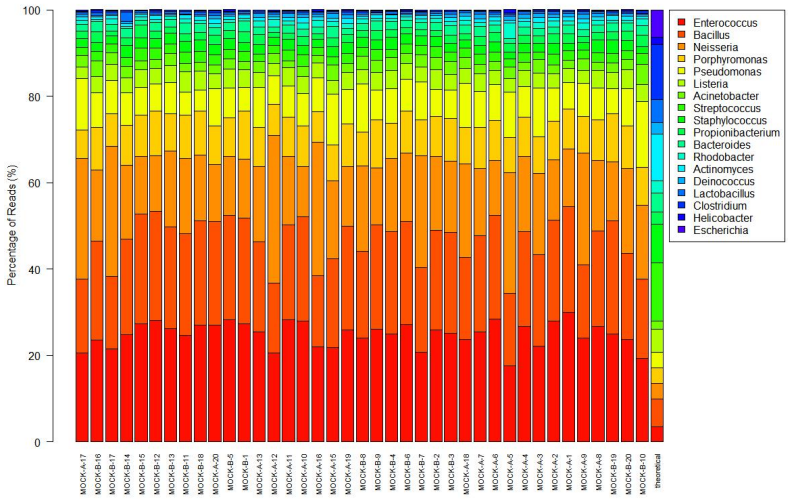


Figure 2: Stacked bar plot presenting the proportion of genera present in replicate samples of a mock community, together with the expected/theoretical yield.

Confirming Differences between Microbiome Profiles

The ultimate goal of any microbiome experiment is to find associations of profiles with health and physiology. Therefore, techniques and tests for comparing groups of samples are essential. LaRosa et al. [8] present power calculations for such tests involving the Dirichlet multinomial distribution, where the counts data are analyzed directly. In the presentation here, we assume that the counts have been normalized to proportions and develop a log-likelihood ratio test based on the continuous Dirichlet distribution.

The log-likelihood Dirichlet test is an alternative to multivariate analysis of variance (MANOVA), where data are assumed multivariate normal. The Dirichlet distribution can take a variety of shapes and reflect more variation than the multivariate normal. The Dirichlet distribution is a multivariate distribution with one shape parameter for each component. If a vector of random variables has components that individually follow a Gamma distribution and are independent except that their sum is one (as with proportions), then the vector follows a Dirichlet distribution.

The basic idea of the test is to fit a Dirichlet distribution to each of two groups of samples independently and then to the pooled samples. If the fits of the individual distributions are determined to be significantly tighter than the pooled distribution, then we may conclude that the two groups arise from different distributions. The log-likelihood ratio serves as a test statistic that asymptotically follows a χ^2 distribution with degrees of freedom equal to the number of components in the Dirichlet distribution. For microbiome data, the components are comprised of a subset of taxa and an "Other" component so that each observation is a vector that sums to one.

We conduct a type I error study and power analysis to compare the proposed test to MANOVA. For two distributions, we define the *discrepancy* Δ to be the Euclidean distance between the vectors of shape parameters in d-dimensional space. The data used in the simulation are based on samples from the Human Microbiome Project - Core Microbiome Sampling Protocol A [5]. There are 139 buccal mucosa samples and 148 stool samples. When $\Delta = 0$, all samples in the simulation are drawn from the distribution fit to the buccal mucosa sample profiles. As Δ is increased, half of the samples are drawn from the buccal mucosa sample distribution and half are drawn from a distribution whose parameters are modified along the vector in d-dimensional space connecting the parameters for the buccal mucosa sample fit distribution and the parameters for the stool sample fit distribution. All tests are conducted at significance level $\alpha = 0.05$. We selected the ten most prevalent taxa (by median) from each group plus an "Other" group for the test.

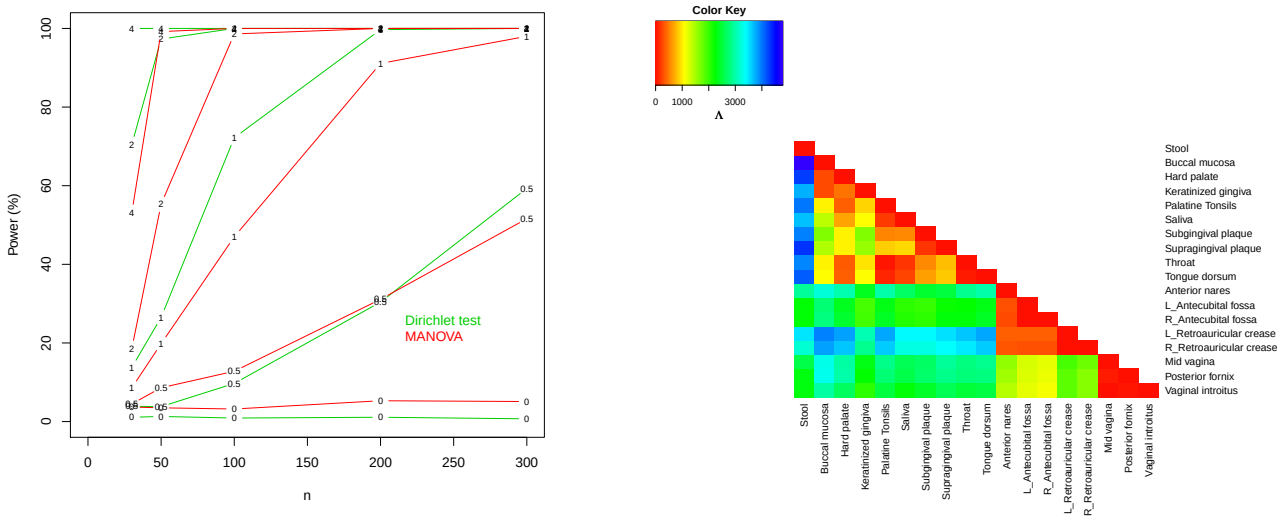
Figure 3(a) contains the results of the type I error study and power analysis. The curves for $\Delta = 0$ indicate the

tendency of the tests to confirm differences between groups of data when no difference actually exists; i.e, the type I error. For MANOVA, this rate is usually between 4 and 5% which is expected given the level of significance used for the tests. The Dirichlet test controls type I error more effectively, with rates around 1.0%. For $\Delta = 0.5$, the power for the Dirichlet test and MANOVA are comparable. For $\Delta \geq 1.0$, the Dirichlet test has at least as much power as MANOVA for nearly all n, a difference that is particularly pronounced for small n.

The proposed test is applied to the problem of discriminating between groups of clinical samples taken from different body sites. The null hypothesis, that the samples from different body sites arise from the same distribution, is rejected for nearly every pair of body sites with p-values less than 0.00001 with few exceptions. Most of these exceptions are for pairs of skin sites taken from symmetric sites. The success in discriminating between body sites confirms the observed differences, and is an indication that the test is useful for discriminating based on other metadata.

To consider the magnitude of differences between microbiome profiles at the body sites, a heat map of the test statistics (Λ) is plotted in Figure 3(b). As indicated by the orange triangles, the oral sites, skin sites, and vaginal sites have greater within-group similarity than similarity with sites from other groups. Of the oral sites, the palatine tonsils, throat, and tongue dorsum appear to have the most similar distributions. The distribution fit to the buccal mucosa samples is most different from the other oral sites and from the distribution fit to the stool samples. The stool samples and buccal mucosa samples seem to warrant their own categories of microbiome distributions, as indicated by the high discrepancies with all other body sites.

In summary, the proposed log-likelihood Dirichlet test controls type I error better than MANOVA and provides greater power at smaller sample sizes and discrepancies than MANOVA. The test is applied to discriminating between body sites and confirms the known differences between different body sites while not confirming differences between samples taken from symmetric body sites.



(a) (b)

Figure 3: (a) Results of a simulation study of type I error and power (%) as a function of sample size ($n/2$ from each group) for the proposed log-likelihood Dirichlet test (green) and MANOVA (red) for $m = 10$ (the number of taxa) and for the simulation based sequences from the V1-V3 region of the 16S gene. The numbers plotted with each curve indicate the discrepancy (Δ) between the two distributions sampled during each iteration of the simulation. Curves are plotted for $\Delta = 0, 0.5, 1, 2, 4$. (b) Heatmap of the test statistic Λ for V1-V3 samples from each pair of 18 body habitats on human subjects.

Discussion

This paper addresses three important considerations when analyzing microbiome data: visualization, assessment of bias, and testing for differences between groups of samples. A vitally important analysis step that is not covered here involves assessing variation in microbiome data. Before reaching biological or clinical conclusions, one must establish that the discovered effects properly account for technical variation, intraperson temporal variation, and interperson variation.

Acknowledgments

We would like to thank the entire Human Microbiome Project consortium, including the four sequencing centers (the Broad Institute, Washington University, Baylor College of Medicine, and the J. Craig Venter Institute), associated investigators from many additional institutions, and the NIH Office of the Director in Roadmap Initiatives. We would also like to thank the Vaginal Microbiome Consortium at VCU which includes Jennifer M. Fettweis, Joao P. Alves, Joseph F. Borzelleca, James P. Brooks, Christopher J. Friedline, Philippe H. Girerd, Abigail L. Glascock, Michael D. Harwich, Stephanie L. Hendricks, Kimberly K. Jefferson, Vladimir Lee, Michael C. Neale, Melissa A. Prestosa, Federico A. Puma, Mark A. Reimers, Maria C. Rivera, Myrna G. Serrano, Nihar U. Sheth, Judy L. Silberg, Logan J. Voegtly, Timothy P. York, Cynthia N. Cornelissen, Jerome F. Strauss III, Lindon J. Eaves, and Gregory A. Buck.

This research is supported in part by NIH-NIAID Award UH3AI08326-01, NIH-NIAID Award UH3AI08326-01, and NIH grants U54 HG004968, UH3 AI03265, UH2 AI083266, R01 DE02154401.

References

[1] Alm, E., 2011, "On the statistical analysis of genomic surveys," submitted.

[2] Brooks, J.P., Dulá, J.H., Boone, E.L., 2011, "A pure L1-norm principal component analysis," submitted.

[3] Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J.I., Knight, R., 2011, "Moving pictures of the human microbiome," Genome Biology, 12, R50.

[4] J.M. Fettweis, J.P. Alves, J.F. Borzelleca, J.P. Brooks, C.J. Friedline, Y. Guan, X. Gao, P. Girerd, M.D. Harwich, S.L. Hendricks, K.K. Jefferson, V. Lee, H. Mo, M.C. Neale, F.A. Puma, M.A. Reimers, M.C. Rivera, S.B. Roberts, M.G. Serrano, N.U. Sheth, J.L. Silbert, L. Voegtly, E.C. Prom-Wormley, B. Xie, T.P. York, C.N. Cornelissen, JL Strauss, LJ Eaves, and GA Buck, "The vaginal microbiome: disease, genetics and the environment," Nature Precedings, `http://dx.doi.org/10.1038/npre.2010.5150.1`, 2010.

[5] The Human Microbiome Consortium, 2011, "Structure, function and diversity of the human microbiome in an adult reference population," submitted.

[6] Kanagawa, T., "Bias and artifacts in multitemplate polymerase chain reactions (PCR)," Journal of Bioscience and Bioengineering, 96(4), 317-323.

[7] Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., Knight, R., 2010, "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns," Nature Methods, 7, 813-819.

[8] La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., Shannon, W., 2011, "Power calculations for taxonomical-based analysis of human microbiome data," submitted.

[9] Schloss, P.D., 2008, "Evaluating different approaches that test whether microbial communities have the same structure," The ISME Journal, 2, 265-275.

[10] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011, "Metagenomic biomarker discovery and explanation," Genome Biology, 12, R60.

[11] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I., 2007, "The human microbiome project," Nature, 449, 804-810.