Emerging Statistical Challenges and Methods for Analysis of Human Microbiome Data

Toby Kenney (Dalhousie Univ.); Glen Satten (Emory Univ.);
Greg Gloor (Univ. of Western Ontario); Rob Beiko (Dalhousie Univ)
Natalie Knox (Bacterial Genomics for Public Health Agency of Canada)
Shyamal Peddada (Univ. of Pitsburg); Michael Wu (Fred Hutch)
Hong Gu (Dalhousie Univ.)

Sep. 15, 2019–Sep. 20, 2019

1 Overview of the Field

Recent developments in high-throughput sequencing have allowed almost all the microbes in a community to be identified through amplicon sequencing of the 16S rRNA ribosomal gene; further, bacterial genes can be identified through shotgun metagenomic sequencing. With many different types of omics data being collected, development of data analysis methods that are tailored towards these different types of data has lagged behind.

This workshop aimed to address some of the main challenges in dealing with the microbiome data. The topics of discussion can mostly be divided into four main categories: modelling, calibrating and correcting for measurement errors in microbiome data; using statistical tools to answer biological questions about the microbiome; integrating multiple types of omics data and applying microbiome data analysis in scientific and clinical contexts.

For the first category, each type of data is subject to its own measurement errors, caused by limitations in the technology. This has contributed to major replicability problems with published results in the field. Recent research has shed some light into some of the measurement bias and variance arising in microbiome data.

For the second category, there are a number of biological questions arising in the study of the microbiome, such as how the microbiome differs between patients and healthy controls, how the microbiome changes over time, and the organisation of microbial communities. There has been a large amount of work in this area from the bioinformatics community, but many of the methods used are ad-hoc, without sound statistical justification, and do not allow statistical inference.

For the third category, the majority of early microbiome studies were based on amplicon sequencing of 16S. However, there are many limitations to this type of data. Therefore, there has been increased interest in collecting more detailed data, such as whole genome sequencing, transcriptomics, metabolomics, etc. These data types could give a more complete picture of how the microbial community functions. However, each data type comes with its own limitations and structure. Therefore new statistical methods are needed to account for these limitations and structures.

For the fourth category, many of the invited researchers work in scientific and clinical settings, and have first-hand experience of the application of microbiome research for answering questions. A number of pre-

sentations at the workshop discussed some of the possible applications of microbiome research, and what statistical problems need to be solved before these applications could be realised.

2 Challenges raised in the workshop

The workshop brought together a number of investigators with different perspectives and different backgrounds that ranged from practitioners and users, to developers of statistical approaches for studying microbial communities. Topics and presentations ranged from highly theoretical to practical applications. The majority of tools presented were used to analyze Operational Taxonomic Unit (OTU) or higher level datasets.

There were several talks on benchmarking different tools for OTU level analysis, with a general conclusion that further work needs to be done. One challenge is that each tool is developed based on very different assumptions about how the data were generated, and what error processes were important or even realistic. There was a spirited discussion about the methodological, biological, and statistical assumptions of various approaches. Participants with deep knowledge of how the data were collected and generated made the point that high throughput sequencing data collection was much more complex and technically challenging than is apparent in many published datasets. One participant made the salient point that the error associated with microbiome datasets was nested, with different steps in the process generating error with different properties. The reality of nested error-generating processes is not currently taken into account appropriately by existing tools and this is a challenge and an opportunity for further development.

One presentation pointed out that microbiome datasets often lack a ground truth, and identified this as a major challenge in the field. This was a very important point, because many tools are developed with particular assumptions in mind. Tool development seems to follow a formula, where novel assumptions are made, tested in silico, and then benchmarked using a publicly-available dataset. The tool is "successful" if it has "more power" on an existing dataset than previous tools. This approach results in tools that identify false positive features, or that identify novel features that fit the particular assumptions, but that may not fit the actual process by which the data is generated. The microbiome field would be well served by datasets with known ground truths that are externally validated by other means.

Less attention was given to methods specifically designed to deal with Amplicon Sequence Variant (ASV), shotgun metagenomic or metatranscriptomic datasets (or the integration of these). Several presentations showed the potential increase in descriptive power by incorporating these approaches. One presentation showed how metagenomic data analysis, particularly the recovery of genomes from metagenome samples, can miss important features including those that relate directly to pathogenicity and antimicrobial resistance. Tools to properly examine these datasets are underdeveloped currently, and it was recommended that there be a major emphasis on tool development in these areas at the expense of OTU-level analysis. Several presentations were made showing that compositional-data based approaches showed promise, although the sentiment was not unanimous that approaches based on log-ratios were always superior.

In the end, the workshop achieved its goals of bringing diverse voices together for a robust discussion. It is rare to get biomedical scientists, practical and theoretical statisticians in a venue where they can have meaningful conversations. We need many more opportunities like this where frank conversations can be had in an environment that can build understanding and trust between all of the fields that need to be present to properly analyze and understand complex high-dimensional data, like that generated in the analysis of microbiomes.

3 Modelling and handling measurement error

3.1 Bias due to the microbiome experiment

Microbiome experiments are still in their infancy. One manifestation of this is that there are no agreed-on methods for standardizing microbiome measurements, or for using positive (external) control samples to adjust, calibrate, or normalize microbial count measurements. A few 'model community' samples with known composition, typically comprising around 10 OTUs, are available either commercially or from various consortia. However, even when these model community samples are run as positive controls in a microbiome experiment, there is no real way to use these results except to qualitatively claim a plate is 'in control.'

An exciting new development in microbiome studies is the development of a model for the biases that can occur in a microbiome experiment. This relatively simple mathematical model may eventually have a profound effect on how microbiome data are analyzed. If this model is correct, then it also provides a language to discuss bias, a framework to imagine how samples might be standardized, and even ideas on types of analysis that could give unbiased results even if individual OTU counts are all biased.

The microbiome bias model was developed by BIRS workshop participant Ben Callahan [15] and presented at the workshop by Ben Callahan. The model states that each OTU (or ASV) is subject to a multiplicative bias. Thus, if the 'true' count for OTU j in sample i was X^*_{ij}, we would instead observe counts X_{ij} given by

$$X_{ij} = X^*_{ij} e^{\beta_j}$$

where e^{β_j} is called the 'bias factor' for the jth OTU. This model further implies that the observed OTU proportions p_{ij} are related to true OTU proportions p^*_{ij} by

$$p_{ij} = \frac{p^*_{ij}}{\sum_{j'} p^*_{ij'}}. \tag{1}$$

Note that changing β_j to $\beta_j + b$, for every OTU j leaves p_{ij} unchanged, so that only relative bias factors can be defined (i.e., a sample is unbiased when $\beta_j = \beta$ for all j).

An important consequence of this model is that the shift in observed probabilities p_{ij} actually depends on the composition of the entire sample because of the denominator in (1). Thus, a sample with many OTUs having the same bias factors may have produce smaller shifts in observed OTU frequencies than a sample with OTUs having highly divergent bias factors. If the bias factors of two phylogenetically-related OTUs are more similar than the bias factors of two OTUs that are not phylogenetically close, this would imply that the sample with phylogenetically-related OTUs would have less biased observations than the sample with phylogenetically-distant OTUs. As a result, this model predicts it is not safe to compare the OTU frequencies of two samples with the hope that any bias 'cancels out.' Note this phenomenon occurs even if the bias factor for a fixed OTU is the same in every sample.

One new development reported at the BIRS workshop for the first time was a relatively simple log-linear modeling framework to estimate bias parameters in (1) using model community data. The original work of Callahan and colleagues did not provide this kind of simple but universal method for estimating bias parameters. BIRS workshop participant Glen Satten presented this model, along with two analyses of model community data.

In the first analysis, the bias factors of a commercially-available (Zymo) model community were compared when extracted under two conditions. In the first condition, the samples were analyzed as provided by Zymo; in the second condition, the samples were mixed with a smokeless tobacco product (Snus) from Sweden that was verified to be free of bacteria. The purpose of this experiment was to see if the biological matrix (i.e., the Snus) affected the bias factors. The experiment was repeated with four different extraction protocols. The analyses showed that the biological matrix had a significant effect on the bias factors for at least 3 of the four protocols (significance of the fourth protocol depends on whether we adjust for multiple comparisons).

The second analysis used the model community data of [2]. These data are unusual in that most samples had only two or three of the seven OTUs studied by Brooks et al [2]. This allowed a test of whether the bias factor of an OTU depended on the composition of the sample; a direct test of the model proposed by Callahan. The analysis showed no evidence of failure of the Callahan model (even while finding significant plate effects).

A final aspect of the bias model of Callahan is that it suggests how samples might be better analyzed in the future. Two approaches were discussed: bias adjustment and selection of models that are impervious to bias. Considering the first approach, it is unreasonable to think that bias factors for every observable OTU or ASV will be measured at some point. However, it is conceivable that the variability in bias factors can be explained by covariates. Some covariates would presumably be related to the physical organization of the bacterium, such as Gram status. Other covariates may be related to PCR, such as primer mismatch or GC content. Residual variability after accounting for important covariates may segregate according to the phylogenetic relationships among OTUs. If this is the case, it is possible to imagine a model-based bias adjustment as a way of normalizing microbiome data. Considering the second approach, it is easy to see that

the denominator in (1) cancels out if a ratio of OTU frequencies within a sample are considered, e.g. if we pick one OTU (say, J) as a reference and only examine ratios $\frac{p_{ij}}{p_{iJ}}$. Similarly, the bias factors would cancel out if we only looked at ratios of ratios, i.e. if we compared the ratio of $\frac{p_{ij}}{p_{iJ}}$ to $\frac{p_{i'j}}{p_{i'J}}$. Compositional data analyses would satisfy these conditions, and so the bias model of Callahan provides an interesting reason to continue development of compositional data analysis methods for microbiome data.

3.2 Correction of Poisson measurement errors

Another important issue with microbiome data is the variance of the microbiome counts. Typically, for count data, the observed variance depends upon the total abundance. For rare OTUs, the relative error is much larger than for common OTUs, while the absolute error is much larger for common OTUs. This discrepancy can impact a number of data analysis methods. For example, principal component analysis (PCA) is a method that identifies the major directions in variation for multivariate data set, such as the microbiome.

A method for correcting PCA for Poisson measurement error was recently developed by workshop participants Hong Gu and Toby Kenney [11] and presented at the workshop by Hong Gu. The method estimates an unbiased variance covariance matrix estimator by correcting Poisson distributed measurement errors. The method can also estimate the principal components of log-transformed data, under Poisson noise. This offers the significant benefit of permitting a log-transformation of sparse data. Handling zero counts has been a challenge for methods based on log-transformation. The semiparametric Poisson model presented offers a method for dealing with these counts.

The talk precipitated in-depth discussion about the relative importance of Poisson error compared to other sources of error in the data, such as OTU bias, and amplification variance. On one hand, it was argued that microbiome data show evidence of large overdispersion, so that the Poisson noise is only a small fraction of the total noise in the data. On the other hand, it was argued that a lot of the overdispersion is in fact signal, so needs to not be removed, and while there is some overdispersion due to the data collection, it is challenging to separate it from the signal until more is know about the data collection. Meanwhile, microbiome data is sparse, and while the Poisson noise may be a small fraction of the noise for abundant OTUs, it is likely to play a significantly larger role for rare OTUs. Work is ongoing on extending the method to correct for overdispersed measurement error.

4 Particular microbiome analyses

4.1 Differential abundance

One of the key questions in microbiome research is how microbial communities from two environments differ, and one of the main approaches to this question is the identification of taxa which are differentially abundant in the two ecosystems. A number of methods for performing differential abundance analysis were discussed. Although simulation studies are often conducted to evaluate the performance of various procedures in terms of false discovery rates (FDR) and power, researchers at this meeting recognized that the following two issues are important to consider: (a) The parameter of interest in the statistical hypothesis. Some methods test for relative abundances and others test for absolute abundance. Both of these parameters are important, depending upon the scientific question of interest. Often researchers are not precise in what exactly they are testing. This leads to wrong analysis and misinterpretation of data. (b) Some statistical tests are specifically designed to test statistical hypotheses regarding relative abundances and others for absolute abundance. The simulation studies are often designed to generate the null data under one or the other type hypothesis, i.e. null hypothesis of no differential abundance or the null hypothesis of null differential relative abundance. However, researchers compare the FDR and power using these same simulated data but different hypotheses. Thus, a data are generated under the null hypothesis of equal relative abundance but are being used to test the null hypothesis of no differential absolute abundance. This will result in inflated FDR for methods designed for no differential abundance and vise-versa. For these reasons simulation studies need to be conducted carefully.

An important issue in differential abundance testing and other microbiome analyses is how to deal with the variation in sequencing depth. The traditional approach in microbiome data analysis is to consider proportions

instead of counts. However, as was demonstrated in one presentation, variations in proportions can be caused by changes in the abundance of OTUs with large bias factors. Log-ratios can be used instead, and have been used in a number of compositional data analyses. However, the microbiome data is very sparse with many zero counts, which are problematic when computing log-ratios. Therefore substantial work is needed on developing methods for approximating log-ratios of sparse counts. Another approach is to treat the count correction as a parameter to be estimated in the model. The relative merits of these methods were presented in several talks. However, it is clear that this is a very challenging problem with no completely satisfactory solutions, and a lot of further work is ongoing in the field.

4.2 Kernel Methods

An alternative to analysis of individual taxa (OTUs) is to conduct community level analysis (also called β-diversity analysis) wherein the entire microbial profiles is collectively assessed for association with an outcome or variable of interest. By focusing on the overall profile, this mode of analysis can be more powerful when taxa show individually modest, yet concerted, shifts. Kernel methods represent a powerful approach for facilitating community level analyses by way of the Microbiome Regression-Based Kernel Association Test (MiRKAT) [25]. Focusing on a quantitative outcome (y_i), MiRKAT relates y_i to covariates C_i and microbiome profiles Z_i through the model

$$y_i = \beta_0 + C_i'\beta + h(Z_i) + \varepsilon_i,$$

where β_0 and β are coefficents for the intercept and covariate effects and ε_i follows a distribution with mean 0 and variance σ^2. $h(\cdot)$ is a generally specified function of the microbiome profiles sitting within a reproducing kernel Hilbert space generated by a positive definite kernel function $K(\cdot,\cdot)$. Then through the connections between kernel methods and linear mixed models, assessing the null hypothesis that $H_0 : h(Z_i) = 0$ can be done by constructing a variance component score statistic

$$Q = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}' K \hat{\varepsilon}$$

where $\hat{\sigma}$ and $\hat{\varepsilon}$ are estimated under H_0 and K is a matrix with $(i,j)^{th}$ element equal to $K(Z_i, Z_j)$. Q asymptotically follows a mixture of χ^2 distributions. Intuitively $K(\cdot,\cdot)$ measures similarity between pairs of individuals based on their microbiome profiles such that K is a similarity matrix. Then this analysis essentially compares similarity in microbiome profiles to similarity in outcomes. By constructing similarities via transformation of ecologically relevant distances and dissimilarities, this enables capture of key structure in microbiome data such as phylogeny and qualitative/quantitative relationships [13, 4, 19].

Kernel methods are a generalization of commonly used distanced-based permutation approaches[1], but allow for improved covariate adjustment and computational efficiency through use of asymptotic distributions. In addition, this allows investigators to harness the rich literature on kernel-based association test procedures, which are widely used within the genetics literature, to facilitate analysis within the context of more complicated microbiome studies such as those with multivariate, longitudinal, cluster correlated, or survival endpoints [23, 18, 24].

In the meeting several talks touched upon kernel methods. One talk focused fully on using kernel methods to conduct integrative analysis of microbiome and other types of genomic data. Specifically, the use of multiple omic technologies (host genetics, epigenetics, proteomics, metataxonomics, etc.) on the same cohort is rapidly increasing. However, investigating associations across complex multivariate outcomes with distinct data structures remains a challenge. One proposed method presented was the use of a dual kernel-based association test (DKAT) to evaluate the similarity between datasets [22]. Specifically, a kernel machine regression model, MiRKAT, may be used as a robust microbiome regression-based kernel association test to circumvent challenges associated with large omics multivariate datasets. This approach is tailored to capture data structure and their inherent characteristics (e.g. high dimensional data, non-normally distributed, zero-abundant). The utility of such a method was applied to identify associations between the gut microbiome and host gene expression of IBD patients [17]. The basic approach of DKAT is to first compute the pairwise similarity in the microbiome profiles and that of the other datasets (e.g. host gene expression profiles). The similarities for each dataset are then compared to each other and assessed via kernels. The test for

multivariate correlation between the kernelized data using a kernel RV coefficient. As multi-disciplinary research continues to expand, the number of outcomes measured per subject will invariably grow and increase in complexity. Sophisticated and computationally efficient statistical approaches that are amenable to large complex datasets will be necessary to enable intricate association testing between microbiome taxa and other measured outcomes.

4.3 Modelling temporal dynamics of microbial communities

Another topic raised was modelling of temporal dynamics of microbial communities. This is a very new topic in microbiome research. While some studies have collected time-series data on the microbiome, the analysis of these studies in terms of efforts to understand the temporal dynamics has been very limited. One talk presented new research into the use of stochastic differential equations for modelling the temporal dynamics. Two projects in this area were presented. The first looked at the dynamics of single genera, focussing on questions of whether there is evidence for temporal continuity and for mean reversion in microbial communities, looking particularly at data from the human gut microbiome from the moving picture dataset [3]. The evidence from the data suggests that there is temporal continuity, and strongly indicates that the system is subject to mean reversion. By analysing the Fisher information matrix for the models in question, it is possible to estimate the most informative sampling scheme for estimating temporal dynamics. For the level of mean reversion estimated from the gut, the best sampling frequency is found to be in the range of one sample every 0.8-3.2 days.

For the temporal interactions between multiple genera, a stochastic generalised Lotka-Volterra equation was used. The theory behind this equation has been developed, and the equation has a unique solution with a stationary ergodic distribution. The parameters of this equation can be estimated from real data using Approximate MLE estimators. These are shown to improve upon previous estimators based on deterministic differential equations with normal error. Using this model, it is possible to estimate the interactions between the most abundant genera. For the gut data from the moving picture data set [3], it was found that there is evidence that *Lachnospiraceae* inhibits both *Ruminococcaceae* and an unspecified genus from family Bacteroidales. *Ruminococcaceae* appears to inhibit *Bacteroidaceae*, which appears to inhibit another unspecified genus from family Bacteroidales. Meanwhile, evidence suggests that *Porphyromonadaceae* stimulates growth of the other genera, particularly *Ruminococcaceae*. This is consistent with what little is known biologically about these organisms, but paves the way for further biological insights into the functioning of microbial communities.

5 Metagenomic data and other Omics Data

5.1 Limitations of Metagenome-Assembled Genomes

Although the analysis of marker genes such as the 16S ribosomal RNA gene can provide insight into the taxonomic composition of microbial communities, these analyses convey little about the functional capabilities of the corresponding microorganisms. Assigning a name such as *Escherichia coli* or *Bacteroides thetaiotaomicron* to an OTU or ASV gives incomplete information due to the potential for considerable functional variation among members of a given species. Genome sequencing from culture can yield further insights, but culturing the entire repertoire of microorganisms from, for example, a stool sample cannot be done. Metagenomics, the shotgun sequencing of DNA extracted directly from a sample without culturing, can overcome this limitation, and we can search these metagenomes for important functions such as pathogenicity, metabolic functions, and antimicrobial resistance. However, the power of metagenomic analysis comes at the cost of tearing genomes into small DNA fragments that lose information about even their closest neighbouring sequence in the genome. Researchers have tried to overcome this last limitation by reassembling entire genomes from metagenomes. This procedure requires the assignment of reads from a potentially complex metagenome to a bin that hopefully corresponds to a single, real genome, and assembly of the reads in that bin to produce a metagenome-assembled genome or MAG. The problem is very challenging, and as with most problems in bioinformatics, multiple approaches have been developed to accomplish this task. A preliminary simulation study of three methods plus one meta-method showed that bin purity (i.e., the percentage of reads in a bin that belonged to the correct genome) and assembly completeness varied substantially by

method and by genome, with closely related genomes often confounding MAG assembly. However, overall accuracy scores fail to address the question of what we are missing in MAG assemblies that are inevitably not perfect. A deeper analysis of the simulated dataset revealed that two types of genetic structures, plasmids and genomic islands, were recovered at extremely low ($< 40\%$ in the best case) levels of sensitivity. These structures are highly mobile and often bear genes that confer pathogenicity and antimicrobial resistance; thus it is important to recognize that MAG reconstruction alone should not be used to infer these traits. Future development should focus on incorporating reference-based assembly, better use of reads that do not assemble in the early stages of the algorithm, and long-read sequencing to augment the traditional short-read approaches.

5.2 Integration of Microbiome and other Omics Data

High throughput microbiome profiling studies have identified associations between microbiome composition and a wide range of human disease and traits including cancer, HIV, menopause, blood pressure, and others [20, 9, 16, 21]. Despite the plethora of findings, the specific mechanisms by which the microbiome influences these conditions and health outcomes remain unclear. To this end, many studies are now interested in integrating other types of genomic data such as metabolomics [14], gene expression [17], and DNA methylation [5] into microbiome studies. These other genomic markers can serve as intermediaries between microbiome composition and outcomes and may provide clues as to the specific manner by which microbes drive subsequent processes and disease. Conversely, guided by recent interest in microbiome studies and the potential impact, many large scale genomic studies, such as large genome wide association study cohorts [10], are also collecting microbiome data. The ability to integrate microbiome data with other types of genomic data promises comprehensive achievement of many biological, clinical and public health questions that have eluded researchers for decades.

Despite the promises of these data, statistical analysis of these data continue to present difficulties for researchers. Some of the central challenges sometimes include the standard problems for analyzing -omics data including high-dimensionality, nonlinear effects, interactions among data features, modest effect sizes, and limited availability of samples. However, in analyzing multiple data types, it is also necessary to accommodate the nature of the individual data types. For example, microbiome data are subject to zero inflation, over-dispersion, compositionality, and structural (e.g. phylogenetic and functional) constraints [12]. Other data types have their own personalities that present challenges. Finally, a major challenge lies in identifying the problem at hand: many researchers are conceptually interested in data integration yet do not have a specific problem that is well formed. The vagueness of the problem often prevents direct translation into mathematical and statistical terms. This last issue poses a particularly grand challenge for statisticians as it hinders development and application of relevant statistical tools.

Particular problems of interest are contextual, but within the context of multi-omics data, we borrow a recently presented taxonomy for describing studies involving multiple omics data types developed by Zhao [26]. In particular, Zhao classifies studies based on *data structure* and *research questions*. Data structure, here, is a simple dichotomization of whether the different types of omics are collected on the same sample or on different individuals (sampling unit). Research questions can be loosely broken down as synthesis questions and transfer questions. Synthesis questions essentially are questions wherein multiple -omics are aggregated to understand an outcome better. Transfer questions are questions wherein one type of -omic data is used to better understand another -omic data type or to better analyze another (possibly with regard to outcome).

Some example studies are given in Table 1. For example, subtyping to identify enterotypes or clusters of microbial communities (individuals) can be done using just microbiome data, but could be improved by also including metabolomics data measured on the same individuals. That the metabolomics improves an already feasible analysis implies that the question is a synthesis question. Similarly, mediation analysis generally requires both data types to be collected on the same individuals. Moreover, mediation would not be feasible without both types of data. Therefore, the question in this case is a transfer question.

	Question Type	
Data Structure	Synthesis	Transfer
Same Samples	Subtyping Prediction models Visualization	Correlation Mediation analysis Effect modification
Different samples	Understanding relationships between disease	Predicted metabolomics

Table 1: *Specific types of multi-omic studies conducted within the context of microbiome profiling fall into categories based on data structure and research question.*

6 Application of microbiome data analysis in scientific and clinical contexts

Decreasing costs for next-generation sequencing have accelerated the availability of this sequencing technology to research labs, hospitals, public health labs, and other regulatory bodies. As such, many are embracing the use of metataxonomic and shotgun metagenomics approaches for a multitude of purposes including infectious disease detection, antimicrobial resistance prediction, and microbial community characterization across various sectors. Though many bioinformatics solutions are available for metataxonomic and shotgun metagenomics data processing, relatively few user-friendly statistical solutions are available. This is a critical gap given that many metagenomic and microbiome studies are being undertaken by biologists, clinicians, or research personnel with limited statistical expertise. In particular, the application of clinical metagenomics – shotgun sequencing of a clinical specimen for unknown infectious disease detection – is rapidly being embraced by frontline laboratories in cases where conventional microbiological testing has failed to identify the infectious agent [7, 8]. Unlike microbial community characterization studies, the goal of a clinical metagenomics approach is to identify and report the causative agent with an interpretation criterion supported by a confidence threshold. To date, no such statistical method or interpretation criterion exists. In this regard, there is a need for infectious disease experts and statisticians to work together towards the development of an approach for reporting of clinical metagenomics findings.

Microbiome methodologies are frequently applied in many studies to identify key taxonomic features that are differentially abundant. As reported during this workshop, many statistical methods have been developed for differential abundance testing, however, random forest classification, a machine learning approach, can also be used to identify key taxa of importance in microbiome studies. In one study presented [6], a random forest classifier approach was used to classify Crohn's disease and healthy control subjects. The model performed well and important features for classification were consistent with taxa identified using differential abundance testing. This approach may be useful to identify biomarkers highly associated with disease.

References

[1] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46, 2001.

[2] J. Paul Brooks, David J. Edwards, Michael D. Harwich, Maria C. Rivera, Jennifer M. Fettweis, Myrna G. Serrano, Robert A. Reris, Nihar U. Sheth, Bernice Huang, Philippe Girerd, Jerome F. Strauss, Kimberly K. Jefferson, Gregory A. Buck, and the Vaginal Microbiome Consortium. The truth about metagenomics: quantifying and counteracting bias in 16s rrna studies. 15:66, 2015.

[3] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.

[4] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[5] Rene Cortese, Lei Lu, Yueyue Yu, Douglas Ruden, and Erika C Claud. Epigenome-microbiome crosstalk: a potential new paradigm influencing neonatal susceptibility to disease. *Epigenetics*, 11(3):205–215, 2016.

[6] Jessica D. Forbes, Chih-yu Chen, Natalie C. Knox, Ruth-Ann Marrie, Hani El-Gabalawy, Teresa de Kievit, Michelle Alfa, Charles N. Bernstein, and Gary Van Domselaar. A comparative study of the gut microbiota in immune-mediated inflammatory diseases–does a common dysbiosis exist? 6(1):221, 2018.

[7] Jessica D. Forbes, Natalie C. Knox, Christy-Lynn Peterson, and Aleisha R. Reimer. Highlighting clinical metagenomics for enhanced diagnostic decision-making: A step towards wider implementation. 16:108–120, 2018.

[8] Jessica D. Forbes, Natalie C. Knox, Jennifer Ronholm, Franco Pagotto, and Aleisha Reimer. Metagenomics: The next culture-independent game changer. 8.

[9] Tiffany Hensley-McBain, Michael C Wu, Jennifer A Manuzak, Ryan K Cheu, Andrew Gustin, Connor B Driscoll, Alexander S Zevin, Charlene J Miller, Ernesto Coronado, Elise Smith, et al. Increased mucosal neutrophil survival is associated with altered microbiota in hiv infection. *PLoS pathogens*, 15(4):e1007672, 2019.

[10] Catherine Igartua, Emily R Davenport, Yoav Gilad, Dan L Nicolae, Jayant Pinto, and Carole Ober. Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome*, 5(1):16, 2017.

[11] Toby Kenney, Tianshu Huang, and Hong Gu. Poisson PCA: Poisson measurement error corrected pca, with application to microbiome data.

[12] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[13] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235, 2005.

[14] Ian H McHardy, Maryam Goudarzi, Maomeng Tong, Paul M Ruegger, Emma Schwager, John R Weger, Thomas G Graeber, Justin L Sonnenburg, Steve Horvath, Curtis Huttenhower, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite interrelationships. *Microbiome*, 1(1):17, 2013.

[15] Michael R McLaren, Amy D Willis, and Benjamin J Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8:e46923, sep 2019.

[16] Caroline M Mitchell, Sujatha Srinivasan, Anna Plantinga, Michael C Wu, Susan D Reed, Katherine A Guthrie, Andrea Z LaCroix, Tina Fiedler, Matthew Munch, Congzhou Liu, et al. Associations between improvement in genitourinary symptoms of menopause and changes in the vaginal ecosystem. *Menopause*, 25(5):500–507, 2018.

[17] Xochitl C. Morgan, Boyko Kabakchiev, Levi Waldron, Andrea D. Tyler, Timothy L. Tickle, Raquel Milgrom, Joanne M. Stempak, Dirk Gevers, Ramnik J. Xavier, Mark S. Silverberg, and Curtis Huttenhower. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. 16(1):67, 2015.

[18] Anna Plantinga, Xiang Zhan, Ni Zhao, Jun Chen, Robert R. Jenq, and Michael C. Wu. MiRKAT-s: a community-level test of association between the microbiota and survival times. 5(1):17, 2017.

[19] Anna M Plantinga, Jun Chen, Robert R Jenq, and Michael C Wu. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics*, 2019.

[20] Robert F Schwabe and Christian Jobin. The microbiome and cancer. *Nature Reviews Cancer*, 13(11):800, 2013.

[21] Shan Sun, Anju Lulla, Michael Sioda, Kathryn Winglee, Michael C Wu, David R Jacobs Jr, James M Shikany, Donald M Lloyd-Jones, Lenore J Launer, Anthony A Fodor, and K Meyer. Gut microbiota composition and blood pressure: The cardia study. *Hypertension*, 73:9981006, 2019.

[22] Xiang Zhan, Anna Plantinga, Ni Zhao, and Michael C Wu. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, 73(4):1453–1463, 2017.

[23] Xiang Zhan, Xingwei Tong, Ni Zhao, Arnab Maity, Michael C Wu, and Jun Chen. A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 2016.

[24] Xiang Zhan, Lingzhou Xue, Haotian Zheng, Anna Plantinga, Michael C Wu, Daniel J Schaid, Ni Zhao, and Jun Chen. A small-sample kernel association test for correlated data with application to microbiome association studies. *Genetic epidemiology*, 42(8):772–782, 2018.

[25] Ni Zhao, Jun Chen, Ian M. Carroll, Tamar Ringel-Kulka, Michael P. Epstein, Hua Zhou, Jin J. Zhou, Yehuda Ringel, Hongzhe Li, and Michael C. Wu. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. 96(5):797–807, 2015.

[26] S. Dave Zhao. Combining multiple genomic data sets. Presented at the 2019 Joint Statistical Meetings (JSM), Denver, CO, 2019.