Discovering Patterns in Human Microbiome Data (HMD)
March 16-18, 2015

SPEAKER TITLES/ABSTRACTS

Vanni Bucci
University of Massachusetts Dartmouth

"Predictive Models of Microbiome Dynamics: Designing Bacterial Cocktails to Ameliorate Enteric Infection and to Stimulate Immune Systems"

DNA sequencing technologies have unveiled the fundamental role of the intestinal microbiota in the maintenance of gut homeostasis. However, as the analysis of this data has mostly been descriptive it has been difficult to transfer this knowledge to the rational design of therapies aimed at modulating this community. In this lecture I will present a set of mathematical and computational methods to predict temporal microbiome dynamics and to perform combinatorial stability analysis constrained on temporal microbiome surveys. I will present how these methods were applied (1) to determine intestinal microbiota states that are compatible and refractory to antibiotic perturbation and enteric infections, (2) to suggest sets of intestinal bacteria that provide resistance to *C. difficile* infection and (3) to design stable probiotic cocktails with capability in stimulating immunity.

Hector Corrada Bravo
University of Maryland College Park

"Statistical and Visualization Methods for Metagenomic Analysis**"

In this talk we will give an overview of development efforts in our group for statistical and computational methods for analysis of metagenomics data. We will describe methods based on mixture models to perform differential abundance analysis that address pervasive sparsity in metagenomic assays, both targeted and whole metagenome. We will describe our modeling strategy for detecting intervals of differential abundance in longitudinal datasets based on smoothing spline anova methods. We will conclude with a preview of our newly designed methods for interactive exploratory visualization and computational analysis of large population metagenomic studies.

Susan Holmes
Stanford University

"Generalizing PCA to Accommodate for the Multiple Sources of Data and Constraints in the Human Microbiome"

Multitable analyses involving PCAIV regression, multi table extension of correlation coefficients and redundancy analysis with constraints allow the visualization and interpretation of complex data from

the microbiome. In this context, multivariate responses include protein abundances and transcriptomic signals that can be explained by sparse combinations of particular bacterial abundance signatures.

This talk contains joint work with David Relman, Julie Josse, Julia Fukuyama, Joey McMurdie, Ben Callahan and Elisabeth Purdom.

Bill Shannon
Washington University St. Louis

"Microbiome Power/Sample Size Calculations (plus a bit of formal hypothesis testing)"

We are writing software for formal statistical power/sample size calculations and hypothesis testing based on the Dirichlet-multinomial model. As you know, it is necessary to have a formal hypothesis, test statistic, and effect size in order to use classical decision theory in science. We have found the Dirichlet-multinomial distribution allows us to do this based on the vector of taxa proportions pi, and the overdispersion parameter theta, and a defined effect size which is a modified version of Cramer's phi which increases as the taxa proportions become more different.

In this talk I will argue that it is time to move microbiome research from the exploratory and discovery phase to a more formal statistical hypothesis testing framework. This is needed as companies begin to develop microbiome treatment strategies that will meet FDA approval processes. I will present the mathematics of our approach, examples from prior work where it has been applied, and a detailed example for a one sample power calculation and how we considered effect sizes for that problem. During this talk I will present software we have developed, including a new cloud based power calculation tool.

CONTRIBUTED TALKS - TITLES/ABSTRACTS

Giseon Heo
University of Alberta

"Comparing Clostridium Difficile Infected Patients before and after a Treatment Using Loops in DNA Sequences"

To analyze the data that is presented as a measure of similarity or dissimilarity, one of the appropriate statistical methods is cluster analysis. Computational topologists recently developed a new method for such situations, called persistent homology. Persistent homology studies the evolution of the topological features found in a space as some parameter increases. Persistent homology captures higher order features which clustering does not. There are three descriptive statistics in persistent homology, namely barcodes, persistence diagrams and persistence landscape. Persistence landscapes are useful for statistical inference as they belong to a space of p-integrable functions, which forms a separable Banach space. Under some regularity conditions, the random variable from a separable Banach space and its continuous linear functionals satisfy the Strong Law of Large Numbers and the Central Limit Theorem. We apply tools from both statistics and computational topology to the DNA sequences taken from clostridium difficile infected patients and their donors. Our statistical and topological data analysis is able to detect interesting patterns among the patients and donors; it

provides a visualization of DNA sequences in the form of clusters and loops. We also apply support vector machines to classify the persistence landscapes we obtain. This is a joint work with Pavel Petrov.

Gholamali Rahnavard,
The Broad Institute of Harvard and MIT

"High-Sensitivity Pattern Discovery in High-Dimensional Heterogeneous Datasets"

Recent advances in technology, engineering, bioinformatics, and microbiome studies have generated tremendous amounts of data on a previously unrealized scale.One of the many goals of the statistical sciences is to discover latent structures in data generated by a tremendously noisy world. We present a novel hierarchical framework, HAllA (Hierarchical All-against-All significance testing), for well-powered association discovery in high-dimensional heterogeneous datasets. HAllA is able to achieve high statistical power, reproducibility in the face of high-dimensional data, and discovering various patterns (not limited to linear) by combining mutual-information-based hierarchical hypothesis testing with false discovery rate (FDR) correction. We use (i) dimension reduction within both predictor and response variables for filtering possible noise in data and (ii) the optimization of hierarchical testing to reduce computation time and limit redundant or uninformative tests. We further validated HAllA's performance using simulated data and in comparison with 10 combination of state-of-the-art association discovery methodologies such as canonical-correlation analysis (CCA) with Pearson correlation coefficient and independent component analysis (ICA) with maximal information coefficient (MIC). Finally, we discuss several real-world applications of HAllA on published datasets. HAllA is implemented with document at http://huttenhower.sph.harvard.edu/halla.

Co-authors: Yo Sup Moon, Levi Waldron, and Curtis Huttenhower

Ayshwarya Subramanian
Harvard University

"Multivariate Association of Microbial Communities with Rich Metadata in High-Dimensional Studies"

With the rapidly increasing availability of massive, human microbiome datasets, there is much interest in uncovering relationships among organisms in these communities as well as between organisms and clinical covariates of their hosts (e.g. disease status, diet, and lifestyle variables). High-throughput metagenomic datasets have a number of properties that complicate their analysis: these include high-dimensionality, sparsity, non-normality, and compositional structure. To combat these challenges, we have developed MaAsLin (Multivariate Association with Linear Models): a multivariate statistical framework that uses boosted, additive, general linear models to find associations between clinical metadata and high-dimensional experimental data. MaAslin incorporates (i) boosting to tackle high-dimensionality, (ii) zero-inflated models to accommodate sparsity, and (iii) multiple choices of link function to manage different types of metadata. The individual components of MaAsLin have been benchmarked against other, similar methods; we have also compared the method as a whole to other univariate (e.g. LefSe [1]) and multivariate (e.g. DirMulti [2]) association detection methods. MaAsLin has been successfully applied to the inflammatory bowel disease [3] microbiome, where it

was used to show that microbial community function is (IBD) more significantly perturbed in the disease condition that community membership. MaAsLin is currently available [4] for use as a downloadable software package and as a Galaxy web server [5].

References:
[1] https://bitbucket.org/biobakery/biobakery/wiki/lefse
[2] http://statgene.med.upenn.edu/softprog.html
[3] Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012 Apr 16;13(9):R79.
[4] URL: http://huttenhower.sph.harvard.edu/maaslin
[5] https://bitbucket.org/biobakery/biobakery/wiki/maaslin

Co-authors: Timothy Tickle2, Levi Waldron3, Yiren Lu4, Lauren McIver1, George Weingart1, Curtis Huttenhower1

1 Department of Biostatistics, Harvard School of Public Health, Boston, MA
2 The Broad Institute, Cambridge, MA
3 CUNY School of Public Health at Hunter College, New York, NY
4 Department of Computer Science, Columbia University

Rochelle E Tractenberg
Georgetown University

"Diversity vs. Frequency: Integrating Mechanistic Knowledge into Algorithmic Analyses of the Urinary Microbiome"

Background: Geman et al. (2014) argue persuasively for integrating mechanistic information into algorithmic omics discovery processes used in cancer. This poster describes how this integration can also work in the analysis of the urinary microbiome. Groah et al. (in review) found that the algorithmically-derived diversity index values describing the urine microbiome for samples from adults with and without neuropathic bladder (NB; a disorder arising in spinal cord injury or disease) did not differ. However, when mechanistic information (from other microbiomes) about specific genera and species was applied, clinically-relevant differences were observed that have implications for decision-making around interventions as well as detection or prevention of urinary tract infections and other urinary symptoms.

Data collection methods:
Urine Acquisition: A 50-100 ml urine sample was collected from each of 47 patients with and without NB. All subjects were "healthy", experiencing no urinary symptoms at the time of urine sampling.

SMRT Sequencing Methods: Urine was sequenced within 6 hours of collection using standard methods. Thawed samples were clarified by low-speed centrifugation and bacterial genomic DNA was isolated from pelleted bacteria using a DNeasy tissue extraction kit (Qiagen) according to manufacturer's instruction. 16S rRNA genes were amplified using universal primers 28F 5'-AGAGTTTGATCMTGGCTCAG-3' and 1492R 5'-ACCTTGTTACGACTT-3'. These amplification products were used for sequencing template preparation using DNA Template Prep Kit 2.0 (Pacific Biosciences) followed by DNA Polymerase binding (DNA Polymerase Binding Kit P4, Pacific

Biosciences). Ternary complexes were loaded onto SMRT cells and sequenced on the Pacific Biosciences RSII instrument using DNA Sequencing Reagent 2.0 (Pacific Biosciences) and one 90 min movie. Resulting circular consensus sequence reads were used for taxonomic classification and diversity measurements using PathoScope, to characterize the biological sample to the species-level, generating an excel file containing proportions of each classified organism.

Data analysis methods:
Shannon, Simpson, Simpson Inverse, and Fisher diversity Index values were computed for the samples. Abundance was also computed for each genus, class and species. Benjamini-Hochberg FDR corrections were applied to t-tests comparing individuals with/without NB and males and females. In an analysis of the NB sample, abundance and diversity indices in those who could and could not void were also compared.

Results: Diversity index values did not differ significantly in any comparison, but clinically significant differences in abundance ("frequency") of several specific classes and species were observed in many of the subgroup comparisons.

Conclusions: Tractenberg et al. (2001) demonstrated that single value summaries of complex clinical phenomenology (behavior disturbance in Alzheimer's disease) can obscure clinically meaningful group differences, and the analyses of Groah et al. show that the same is true for analyses of the urine microbiome.

Co-authors: Marcos Perez-Losada, Ljubica Caldovic, Suzanne L Groah

Duncan Wadsworth
Rice University

"Bayesian Variable Selection for Multinomial-Dirichlet Regression with an Application to Microbiome Data Integration"

High-dimensional count data, especially that found in metagenomic studies, display several characteristics which make it difficult to model. Zero-inflation, skewness, and overdispersion all cause difficulty for standard probability models such as the Poisson and Negative Binomial distributions. The Dirichlet-Multinomial distribution has been suggested as a more appropriate way to model data with these characteristics and can be extended, in a straight-forward way, to a regression framework via log-linear models. In the microbiome setting this extension allows the integration of taxonomic count data with other information taken on the same samples, e.g. demographics, diet logs, etc. However, the number of variables can quickly exceed the number of samples and, assuming that many of the parameters are superfluous, variable selection methods may be used to determine which variables represent important associations between the two data types. We propose a Bayesian approach to variable selection using spike-and-slab priors which simultaneously estimates variable inclusion/exclusion and the log-linear regression parameters. Our method compares favorably in simulations to one recently proposed by Chen and Li (2013) and to a simple L_1 regularized multinomial probit regression approach. Finally, we present data analysis results from two recent microbiome studies. This is joint work with Raffaele Argiento, Michele Guindani, and Marina Vannucci.

Brandie D. Wagner
University of Colorado

"Analysis of Longitudinal Microbiota Data"

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, for which commonly used statistical approaches may not be appropriate. In addition, complex study designs are of interest with longitudinal sample collection or multiple samples collected from the same subject, thus requiring approaches to account for this additional source of variability. Many advances with respect to applied statistical methods for next-generation sequencing have been made which include the use of the negative binomial distribution that includes an offset to account for variable sequencing effort and overdispersion, the application of zero-inflated models to deal with excess zero counts, and methods appropriate for application to studies with repeated measures. The zero-inflated negative binomial mixed model encapsulates all three aspects. This model approach is useful for evaluating individual organisms but ignores the interactions and multivariate structure of the bacterial communities. Conversely, ecological methods can be used to evaluate turnover in community composition over time. Here, a beta diversity measure is regressed on a time lag variable using a time series model. With this approach, a rate of change in composition is provided accounting for the multivariate characteristics of the data but does not provide information for changes in specific organisms over time and is currently used to investigate temporal changes in a single community. The use of both the ecological and organism specific approaches provide complementary information and when combined may prove to be a useful methodology to analyze longitudinal microbiota data.

Co-authors: Rui Fang, Miranda Kroehl, J Kirk Harris, Joshua Miller, Marci Sontag and Peter M. Mourani

Michelle Wright
Virginia Commonwealth University

"Alpha Diversity of the Vaginal Microbiome Clusters within Families: a Twin Study"

Background: Differences in the diversity of vaginal microbiome profiles have been reported in women of different ethnic backgrounds suggesting the influence of host and microbiome genetic factors. Yet relatively little is known whether this clustering is due to genetic or shared environmental factors.

Purpose: The purpose of this study was to examine the vaginal microbiomes of monozygotic (MZ) and dizygotic (DZ) twin pairs to determine if host genetics were associated with alpha diversity scores.

Methods: V1-V3 regions of bacterial 16S rRNA were amplified and sequenced from 112 MZ and 61 DZ female twin pairs. Alpha diversity, diversity of bacterial species within individuals, was measured using the inverse Simpson's Index. Variance in alpha diversity within groups of MZ and DZ pairs separately was determined by analysis of variance techniques. Evidence for genetic influences on alpha diversity would be observed if the mean squares within DZ pairs was significantly greater than that of the MZ pairs.

Results: Vaginal microbial alpha diversity was significantly different between both MZ twin pairs, ($F_{111, 112} = 1.978$, $p < 0.001$); and DZ twin pairs, ($F_{60, 61} = 2.569$, $p < 0.001$) which suggests familial clustering. However, a test between MZ and DZ pairs did not detect significant genetic influence in differences in alpha diversity ($F_{50,107}$, $p = 0.598$).

Conclusions: Alpha diversity of bacterial species were similar within families, regardless of zygosity type, and for this data attributable to shared environment factors and not genetic influences. These preliminary analyses suggest that while there were not strong global genetic effects present, future analyses may identify genetic differences that are species specific.