Mirkat Peeks at Associations between the Microbiome and Disease

June 15, 2015

JM Kocarnik

The microbiome refers to the community of microbes and bacteria that lives in and on each individual. Recent advancements in high-throughput sequencing technology have allowed researchers to measure the presence and composition of the microbiome, and then evaluate these measurements for associations with disease. Generally, the methods utilized for these analyses have relied on comparing phylogenetic distances between the species observed. While useful, these methods have challenges that preclude extending microbiome analyses to alternative outcomes of interest, such as longitudinal or survival data. To address these issues, Drs. Ni Zhao and Michael Wu in the Public Health Sciences Division developed a new statistical method called the Microbiome Regression-based Kernel Association Test (MiRKAT). Presented in *The American Journal of Human Genetics*, this approach improves on current methods and allows researchers to extend microbiome research to investigations of alternative outcomes.

Microbial communities can be profiled by performing targeted sequencing of the 16S rDNA sequences. These sequences can identify the presence and relative abundance of microbial species contained in a sample. After clustering these species into operational taxonomic units based on their 16S rDNA sequence similarity, distance metrics can be constructed to measure the phylogenetic or taxonomic dissimilarity between each sample. These pairwise distances between each pair of samples can then be compared to the distribution of some outcome of interest in order to test for an association between microbiome diversity and disease.

A key limitation to existing approaches is the need to choose an appropriate metric for measuring these pairwise distances. UniFrac distances are the most popular (and can be weighted, unweighted, or generalized), but additional distances can also be used (such as Bray-Curtis or Euclidian). Since these distances prioritize differences differently, choosing a non-optimal distance metric will reduce statistical power to detect a true association. Choosing the optimal metric for a given project is difficult, however, and cherry-picking results from analyzing multiple metrics increases the chance of false positive findings. In addition, current analytic approaches can be difficult to interpret, do not allow for easy covariate adjustment, and are challenging to apply to outcomes such as survival or multivariate data.

"The need for more flexible analysis tools became apparent and motivated our development of

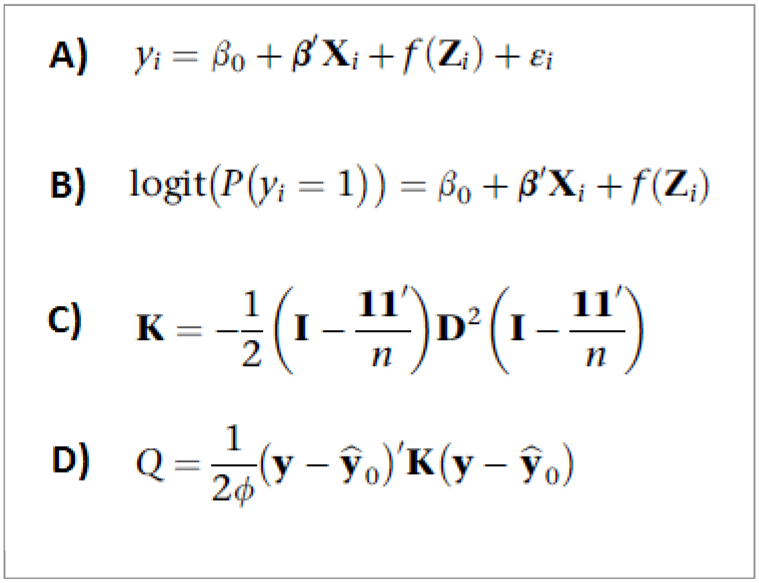
MiRKAT," said senior author Dr. Wu. "MiRKAT arose, in part, because my collaborators wanted to understand the role of the microbiome in a set of clinical samples, but the standard analysis approaches did not easily accommodate aspects of the study design. Our work makes a significant effort to close this gap and to help researchers better analyze their data and obtain more reliable and meaningful results."

In short, the MiRKAT method uses the kernel machine framework to compare pairwise similarity in the outcome variable to pairwise similarity in the microbiome profiles. High correspondence between these measurements suggests an association. Mathematically, this resembles a linear or logistic regression model with an added kernel function (see figure), which also allows controlling for potential confounders. The kernel measures the similarity between different individuals, and different kernels can be chosen to evaluate more advanced underlying models. A score statistic then allows researchers to obtain a p-value for covariate-adjusted associations between microbiome diversity and an outcome of interest.

An extension of their method, "Optimal" MiRKAT, alleviates the issue of having to choose the best distance metric by simultaneously considering several kernel options and correcting for these multiple iterations. Both methods performed well in simulated and real datasets, demonstrating the power, efficiency, and utility of this new method for microbiome-profiling studies. For example, the authors used MiRKAT in existing data to replicate an association between microbiome profiles in the lung and smoking status, but were also able to further control for potential confounders not possible in the original research.

"MiRKAT opens the doors for a lot of new research," said Dr. Wu. "As a generic tool which is more flexible than what is already out there, it will enable the conduct of new studies and facilitate research into the role of microbiome across a wide range of diseases. This is particularly the case as researchers start to investigate the relationship between the microbiome and more complex outcomes such as time to event outcomes and longitudinal measurements. MiRKAT also opens new areas of methodological research which we hope to pursue. The statistical underpinnings of the approach are similar to what is done in genetic association studies; consequently, we hope to exploit these similarities to enable joint analysis of microbiome data with other complex -omics data such as genotyping or metabolomics. This will be a major focus of my group's efforts in the near future."

Citation:
Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. 2015. Testing in microbiome-profiling studies with MiRKAT, the Microbiome Regression-based Kernel Association Test. *Am J Hum Genet*. 96(5):797-807. doi: 10.1016/j.ajhg.2015.04.003.



A) $y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + f(\mathbf{Z}_i) + \varepsilon_i$

B) $\text{logit}(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + f(\mathbf{Z}_i)$

C) $\mathbf{K} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{D}^2\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)$

D) $Q = \frac{1}{2\phi}(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \hat{\mathbf{y}}_0)$

Image provided by Dr. Jonathan Kocarnik

Mathematical representation of the MiRKAT regression framework. The association between the microbiome and an outcome of interest can be calculated through a linear kernel machine model for continuous outcomes (A) or a logistic kernel machine model for binary outcomes (B). The kernel measures the phylogenetic or taxonomic distance (C), and a score statistic (D) is calculated to generate a p-value for the association.