A fast small-sample kernel independence test for microbiome community-level association analysis

Xiang Zhan[1], Anna Plantinga[2], Ni Zhao[3], and Michael C. Wu[1]

[1]*Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA*

[2]*Department of Biostatistics, University of Washington, Seattle, WA 98195, USA*

[3]*Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA*

February 22, 2017

Abstract

To fully understand the role of microbiome in human health and diseases, researchers are increasingly interested in assessing the relationship between microbiome composition and host genomic data. The dimensionality of the data as well as complex relationships between microbiota and host genomics pose considerable challenges for analysis. In this paper, we apply a kernel RV coefficient (KRV) test to evaluate the overall association between host gene expression and microbiome composition. The KRV statistic can capture non-linear correlations and complex relationships among the individual data types and between gene expression and microbiome composition through measuring general dependency. Testing proceeds via a similar route as existing tests of the generalized RV coefficients and allows for rapid p-value calculation. Strategies to allow adjustment for confounding effects, which is crucial for avoiding misleading results, and to alleviate the problem of selecting the most favorable kernel are considered. Simulation studies show that KRV is useful in testing statistical independence with finite samples given the kernels are appropriately chosen, and can powerfully identify existing associations between microbiome composition and host genomic data while protecting type I error. We apply

the KRV to a microbiome study examining the relationship between host transcriptome and microbiome composition within the context of inflammatory bowel disease and are able to derive new biological insights and provide formal inference on prior qualitative observations.

Keywords: Kernel, Microbiome composition, Multivariate association test, Omnibus test, RV coefficient

1 Introduction

The human body is inhabited by many complex communities of microorganisms and their composition (defined as the microbiome) have been increasingly recognized to play an important role in many human disease conditions, including obesity (Turnbaugh et al., 2009), type 2 diabetes (Qin et al., 2012), and inflammatory bowel disease (Morgan et al., 2015). Recent advances in next-generation sequencing technologies now allow investigators to quantify the composition of the microbiome using direct DNA sequencing of the 16S ribosomal RNA gene (Lasken, 2012). Based on their sequence similarity, the raw 16S sequence reads are often clustered into Operational Taxonomic Units (OTUs), which is a commonly used microbial diversity unit and can be considered as surrogate of a bacterial taxon when clustered at 97% similarity level (Stackebrandt and Goebel, 1994). Many downstream analyses are performed based on the OTU abundances, among which a powerful mode of analysis is the community level analysis, wherein overall microbiome composition of multiple OTUs is assessed for identifying overall shifts among different conditions (Li, 2015). Community level analysis can be more powerful than examination of individual taxa when there are systematic, modest changes in abundance but individual taxa do not have a strong effect (Plantinga et al., 2017, Zhao et al., 2015).

Recently, there is considerable interest in understanding the relationship between overall microbiome composition and profiles of other types of genomic data. For example, Morgan et al. (2015) was interested in determining whether host gene expression profiles, overall and within specific candidate pathways, are globally related to microbiome composition in patients with inflammatory bowel disease. Unfortunately, how to systematically examine the relationship between high-dimensional microbiome compositional profiles and other high-dimensional gene expression data remains unclear. The

authors resorted to associating individual gene expression and individual OTUs by using the top principal components, as well as making qualitative observations regarding relationships, in which no formal inference was conducted. It would be of considerable practical interest to devise a means for formal inference of hypothesis testing and for conducting more systematic association analysis.

Assessing overall association relationships between two sets of variables can be accomplished using a range of different methods. For example, the RV coefficient (Escoufier, 1973) provides insight into the global correlation between the two random vectors (e.g., a vector of microbiome profiles and a vector of gene expression values). However, as a generalization of the Pearson correlation coefficient, RV coefficient can only measure linear dependency. The high dimensionality of the data, the complexity of the relationships between data types, and inherent structure (e.g., phylogenetic relationships) among the taxa pose grand challenges for the RV coefficient. To accommodate general dependency patterns beyond linearity, one strategy is to incorporate distance metrics as in the GRV statistic (Minas, Curry, and Montana, 2013). Motivated by GRV, we map the original vector spaces to reproducing kernel Hilbert spaces (RKHSs) and consider kernel RV (KRV) coefficient as the RV coefficient between the RKHS-images of the two random vectors. It turns out that this KRV statistic is closely related to existing statistics that measure multivariate statistical independence, including the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005, 2008) and distance covariance (Székely, Rizzo, and Bakirov, 2007).

Despite the correspondences of KRV with many existing multivariate dependency metrics, the testing design of these existing statistics do not fit the current microbiome association analysis. This is because current microbiome studies often have a relatively small sample size, while most existing multivariate dependency tests depend on asymptotic results (e.g., the HSIC test). Thus, a more accurate finite-sample null distribution is desired for a microbiome association test (Chen et al., 2016, Plantinga et al., 2017). To evaluate significance based on the KRV statistic, we adopt the GRV testing strategy (Minas et al., 2013), which approximates the empirical null distribution of all KRV permutations to a Pearson type III distribution by matching the first three moments. Since the empirical moments of the null KRV permutation distribution are easy to calculate based on previous

results on RV-type statistics (Josse, Pagè, and Husson, 2008, Kazi-Aoual et al., 1995), parameters of the Pearson type III distribution can be explicitly expressed in closed form. Finally, the p-value of a KRV test can be calculated analytically using this approximated Pearson type III density. The new test design is well-suited for small-sample microbiome studies without using any asymptotic results.

Although we follow the GRV testing framework to examine the association between two vectors, there are key differences. The most important difference is that the proposed KRV test has been applied to a different domain. GRV tests for association between SNPs and gene expressions, where specific distance metrics for SNPs and gene expressions have been explored. In this paper, our major focus is kernel metrics for microbiome composition data. Beyond that, the KRV test also extends the GRV test in the following two aspects. First, the KRV test allows adjustment of confounding effect. Environmental exposures, clinical outcomes and treatment groups (all termed as covariates) are important in assessing the association between microbiome composition and host gene expression. It is possible that some covariates affect both the microbiome composition and gene expression. Under such a scenario, failure to account for these covariates can produce misleading bias of association or affect the testing power. Second, we propose an omnibus KRV test which can accommodate multiple candidate kernels, which is much more efficient than the permutation and meta analysis-based approach used in GRV to accommodate multiple distances. The choice of kernels in KRV is crucial for the success of the test. The optimal kernels with powerful KRV tests depends on both the specific data structures and the underlying association patterns, which however, are often unknown in practice. Without hacking p-values by selecting the most favorable kernels, we incorporate an omnibus procedure in KRV to accommodate multiple candidate kernels. The KRV test with this omnibus kernel is more robust in that it can always have adequate power under different scenarios. Finally, by approaching the problem from the perspective of kernels rather the distances, we are able to related the KRV to existing metrics of generalized statistical dependence to better understand properties.

The rest of the paper is organized as follows. In Section 2, we first introduce the KRV statistic and explore its connection with many existing statistics for multivariate association analysis. Then, we utilize existing testing

strategy in RV-type statistics to evaluate significance based on KRV statistic. Next, we carefully adapt the KRV test to microbiome association analysis by enabling covariates adjustment as well as accommodating multiple OTU kernels in Section 3. The finite sample performance of the proposed KRV test both in testing statistical independence and microbiome association is assessed through numerical studies in Section 4. In Section 5, we apply the KRV test to the dataset of Morgan et al. (2015) examining the relation between host transcriptome and microbiome composition in samples taken from inflammatory bowel disease patients. Our analysis is able to provide additional insights. The paper concludes with a brief discussion in Section 6.

2 A KRV-based Fast Small-sample Kernel Independence Test

RV coefficient (Escoufier, 1973) was developed as a measure of linear correlation between sets of multivariate measurements collected on the same individuals. In particular, let X be an $n \times p$ matrix (of variables $X^1, \ldots, X^p$) and Y be an $n \times q$ matrix (of variables $Y^1, \ldots, Y^q$), corresponding to two sets of variables, such as gene expression values and OTU counts observed from the same n individuals. Then, RV coefficient between X and Y is defined as

$$RV(X, Y) := \frac{tr(S_{XY}S_{YX})}{\sqrt{tr(S_{XX}^2)tr(S_{YY}^2)}}, \tag{1}$$

where $S_{XX} = X'X/(n-1)$, $S_{YY} = Y'Y/(n-1)$, $S_{XY} = X'Y/(n-1)$, $S_{YX} = Y'X/(n-1)$ are sample covariance matrices, given that X and Y are centered by columns.

A notable feature of RV coefficient is that it is only able to capture the linear dependency between two random vectors and does not accommodate nonlinearity or other more general dependencies (Robert and Escoufier, 1976). In practice, complex data such as microbiome and host genome data, often require general methods to detect more general dependencies that are of interest. Motivated by this, we propose the KRV coefficient to measure more general relationship between microbiome composition and host genome expression. Specifically, we kernelize the RV coefficient by embed-

ding the original spaces $\mathcal{X}$ and $\mathcal{Y}$ to some functional spaces spanned by kernels (Hofmann, Schölkopf, and Smola, 2008). Let $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto R$ and $l(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \mapsto R$. be two kernel functions. Then, the KRV coefficient is proposed as

$$KRV(X,Y) := \frac{tr(\tilde{K}\tilde{L})}{\sqrt{tr(\tilde{K}\tilde{K})}\sqrt{tr(\tilde{L}\tilde{L})}}, \tag{2}$$

where $\tilde{K} = HKH$ and $\tilde{L} = HLH$. K and L are two $n \times n$ kernel matrices, where $K_{ij} = k(X_i, X_j)$, $L_{ij} = l(Y_i, Y_j)$, $i, j = 1, \ldots, n$, $H = I - \mathbf{1}\mathbf{1}'/n$ is a centering matrix, I is an identity matrix of order n, and $\mathbf{1}$ is a $n \times 1$ vector of all ones. A sketch of calculating the KRV coefficient is included in Section A.1 of Appendix A.

If the kernel matrices are selected as $K = XX'$ and $L = YY'$, then the KRV coefficient reduces to the RV coefficient. If we replace the two kernel matrices K and L by two distance matrices, then KRV reduces to a GRV coefficient. Beyond its close connection with RV-type statistics, KRV is also similar to some other statistics. In particular, the numerator of KRV is simply the HSIC statistic $tr(\tilde{K}\tilde{L})$ (Gretton et al., 2005, 2008), which has been widely used to characterize statistical independence. Thus, given the kernels being appropriately chosen (Gretton et al., 2005), the KRV statistic can also be used to characterize independence. Such a property, however, has never been studied for other RV-type statistics (Josse et al., 2008, Minas et al., 2013). Similar to the HSIC statistic, distance covariance/correlation (Székely et al., 2007) is also widely used for measuring and testing independence between two groups of variables. It has been shown that distance covariance is equivalent to HSIC (Sejdinovic et al., 2013). In this spirit, KRV is equivalent to distance correlation.

Besides the HSIC statistic and distance covariance statistic, many other statistics have been proposed to measure generalized dependency. Readers are referred to Josse and Holmes (2013) and references therein for further details. Finally, it turns out that our KRV statistic coincides with some existing statistics including the RV for kernels (Purdom, 2006) and the centered kernel alignment statistic (Cortes, Mohri, and Rostamizadeh, 2012). However, the RV for kernels is used for kernel principal component analysis and kernel canonical correlation analysis, and the centered kernel alignment statistic is used to develop algorithms for learning kernels for classification

and regression. Both RV for kernels and centered kernel alignment statistic have not been used for hypothesis testing, which is the focus of the current paper.

Despite the correspondences of KRV with HSIC and distance covariance, the design of the HSIC test (based on asymptotic results) and the distance covariance test (permutations) are often limited. In particular, asymptotic null distribution-based HSIC test is not appropriate for studies with small sample size, such as the micriobiome study considered in this paper. On the other hand, a permutation test of distance covariance can be computationally expensive when the nominal significance level is stringent. Thus, a new fast small-sample independence test based on the KRV statistic is necessary.

The distribution of the KRV statistic is generally unknown due to its complex form. A reasonable strategy is to use permutations. Unlike the permutation-based distance covariance test, we utilize permutations differently. To avoid the computational burden of explicitly resampling and recalculating permuted KRV statistics, we follow testing strategy of existing RV-type statistics (Josse et al., 2008, Minas et al., 2013), to approximate the empirical null distribution of KRV permutations by moment-matching. Specifically, let $Q_i, i = 1, \ldots, n!$ denote the KRV statistics calculated from all $n!$ potential permutations by shuffling rows and columns of one kernel matrix simultaneously. The first three sample moments of $\{Q_1, \ldots, Q_{n!}\}$ are calculated and a Pearson type III density with the same first three moments is obtained. The final p-value is calculated from this approximated Pearson type III density. More details of the Pearson type III approximation can be found in Section A.2 and A.3 of Appendix A.

3 Adapting KRV for Microbiome Association Analysis

In this section, we tailor the KRV framework to facilitate the microbiome association analysis with host gene expression data mainly considered in this paper.

3.1 Kernel Choice

To evaluate the association between microbiome composition and host gene expressions via the KRV test, we first need to select kernels in KRV for both microbiome composition data and gene expression data. In many kernel-based genetic association tests, kernels are used as similarity measures, and concordance between genotype similarity and phenotype similarity is suggestive of association (Broadaway et al., 2016, Wu et al., 2011). Similarly, we treat K_{ij} and L_{ij} in KRV as similarity measures of sample i and j in terms of their microbiome composition profiles and host genomic expression profiles, respectively. The KRV statistic tends to be large if one similarity matrix resembles to the other. That is, concordance in microbiome similarity and host genome similarity is suggestive of association.

More rigorously, kernel matrices K and L need to be positive semi-definite so that the KRV statistic (6) is well-defined. Constructing positive semi-definite kernels for association analysis is a common practice for many different omics data types (Wu et al., 2011, Zhan et al., 2015, 2016, Zhao et al., 2015). For the microbiome composition data considered in this paper, the UniFrac kernels are ecologically meaningful similarity metrics and can accommodate important features of OTU data, e.g. the phylogenetic structure (Chen et al., 2012, Lozupone and Knight, 2005, Lozupone et al., 2007). The UniFrac-type kernels quantify the similarity of two OTU profiles by incorporating both their abundance (or presence/absence) information and phylogenetic relationship. Besides the UniFrac kernels, the Bray-Curtis kernel is also widely used, which quantifies similarity of two microbial communities based on the OTU counts and can be useful when the phylogenetic tree information is unavailable and unreliable. For host gene expression data, some popular choice are the Gaussian kernel ($K_{ij} = k(x_i, x_j) = \exp(-||x_i - x_j||^2/\sigma^2)$) and linear kernel ($K = XX'$) (Liu, Lin, and Ghosh, 2007). To account for correlation among gene expressions, the weighted linear kernel ($K = X\Sigma_{XX}^{-1}X'$) can also be used (Broadaway et al., 2016).

3.2 Accommodating Multiple Kernels

The choice of kernels in KRV is crucial for the success of the test. Different kernels measure different aspects of data nature and assume different association patterns. Unfortunately, selecting the most powerful OTU (or gene

expression) kernel requires both knowledge of the microbiome community structure and how the microbiome influences gene expression. Without such prior knowledge, it is necessary to develop an omnibus test which incorporates multiple candidate kernels. In GRV (Minas et al., 2013), a similar multiple candidate distances issue is solved by meta-analysis for different combinations of distances. P-values from all possible distance combinations are used to calculate the Fisher summary statistic, and permutations are used to establish the significance based on the Fisher summary statistic. The adjustment of multiple distances in GRV is often computationally inefficient due to the need of extra datasets for meta-analysis and also permutations for final p-value calculation.

To avoid potential limitations of GRV, we propose to combine the multiple candidates at the kernel level in KRV rather than the test p-value level as in GRV. Without loss of generality, suppose $k_i, i = 1, \ldots, m$ are candidate OTU kernels, with corresponding kernel matrices $K_i, i = 1, \ldots, m$, and we fix the gene expression kernel l or L. The same omnibus OTU kernel strategy can be applied to accommodate multiple gene expression kernels. Motivated by existing literature in multiple kernel learning (Cortes et al., 2012) and genetic association studies (Wu et al., 2013), we propose to use an omnibus OTU kernel of the form $K_{om} = \sum_{i=1}^{m} \omega_i K_i$ with $\omega_i \geq 0$ and $\sum_{i=1}^{m} \omega_i = C$. Since the KRV statistic is scale invariant, constant C in the constraint $\sum_{i=1}^{m} \omega_i = C$ does not make a real difference. There are many methods to determine the weights $\omega_i, i = 1, \ldots, m$. The simplest strategy is to use unsupervised weights such as $K_{om1} = \sum_{i=1}^{m} K_i/m$ and $K_{om2} = \sum_{i=1}^{m} K_i/tr(K_i)$. An advantage of K_{om1} and K_{om2} is that a direct KRV test between K_{om} and L can be used to establish the final significance. Another more complicated way to select the weights in a supervised way. For example, Cortes et al. (2012) suggest to select the weights that maximize the KRV statistic between the omnibus OTU kernel and gene expression kernel:

$$KRV(K_{om}, L) = \frac{\sum_{i=1}^{m} tr(\omega_i K_i L)}{\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} tr(\omega_i K_i \omega_j K_j)} \sqrt{tr(L^2)}}, \tag{3}$$

subjected to $\omega_i \geq 0$ and $\sum_{i=1}^{m} \omega_i = 1$. The optimal weights $\boldsymbol{\omega}^* = (\omega_1^*, \ldots, \omega_m^*)'$ can be calculated by a Quadratic Programming (QP) algorithm (Cortes et al., 2012). As a consequence of supervised weights learning, p-value of the

test $KRV(K_{om3}, L)$, where $K_{om3} = \sum_{i=1}^{m} \omega_i^* K_i$, is no longer a genuine p-value. Permutations are needed for establish the significance of the test based on K_{om3}. Finally, Wu et al. (2013) suggest to select the individual kernel with the minimum p-value. That is, $K_{om4} = K_i$, where K_i has the smallest KRV p-value among $K_1, \ldots K_m$. Like K_{om3}, a permutation-based procedure is needed to establish the significance between K_{om4} and L. More details on K_{om3} and K_{om4} including the permutation-based p-value calculation procedures along with comprehensive numerical studies comparing K_{om1}, K_{om2}, K_{om3} and K_{om4} are presented in Section B.1 of Appendix B. Based on our numerical studies, it turns out that the omnibus kernel K_{om2} with unsupervised weights $\omega_i = 1/tr(K_i)$ tends to have the best overall performance under most scenarios, and thus is used as the omnibus kernel in the rest of this paper.

3.3 Adjusting for Confounders

It is important to adjust for the effect of confounding variables when testing association. Let Y and Z denote host gene expression and microbiome composition respectively, X denote some covariates, such as age, gender, smoking status and other clinical or environmental variables, which may influence both host gene expression and microbial community diversity. Without adjusting for covariate effects, the association testing results between Y and Z can be misleading, sometimes leads to excessive false positive discoveries. To adjust for the potential confounding effects of X in KRV framework, we utilize the residual-based strategy as widely used in many kernel machine association tests (Broadaway et al., 2016, Hua and Ghosh, 2015, Liu et al., 2007). Let $P_X = X(X'X)^{-1}X'$ denote the projection matrix of the column space of X, and denote the residuals $\widetilde{Y} = (I - P_X)Y$. Then we can calculate the residual kernel as $\widetilde{L}^r_{ij} = l(\widetilde{Y}_i, \widetilde{Y}_j)$. Finally, we replace $\widetilde{L}$ in equation (6) by $\widetilde{L}^r$ to calculate the statistic and conduct the test after adjusting for X. In the univariate scenario (dim(Y)=1) of kernel machine regression, the above procedure is equivalent to testing the association using a restricted maximum likelihood (REML)-based score test (Liu et al., 2007).

4 Simulation Studies

4.1 Statistical Independence Simulation

We first conducted simulations to evaluate the performance of the proposed KRV test in testing statistical independence. We compared our KRV test to the HSIC test and distance covariance (dcov) test, both of which have been widely used for testing statistical independence between two random vectors. As a benchmark, we also compared the GRV test, which has the same test design as the KRV test but uses distance metrics rather than kernels. The setup of this simulation was exactly the same as that in the dcov test paper (Székely et al., 2007). Two continuous random vectors X and Y were simulated, where $p = dim(X) = dim(Y) = 5$, and the marginal distribution of each dimension of X and Y was standard normal. The following four scenarios (A) – (D) were used to simulate the data:
(A) $Cov(X_i, Y_j) = 0$, for $i, j = 1, \ldots, p$, and $Cov(X_i, X_j) = 0$, $Cov(Y_i, Y_j) = 0$ for any $i \neq j$.
(B) $Cov(X_i, Y_j) = 0.1$, for $i, j = 1, \ldots, p$, and $Cov(X_i, X_j) = 0.1$, $Cov(Y_i, Y_j) = 0.1$ for any $i \neq j$.
(C) $Y_{ij} = X_{ij}\epsilon_{ij}$, $i = 1, \ldots n; j = 1, \ldots p$, where ϵ_{ij} are independent standard normal random variables independent of X.
(D) $Y_{ij} = \log(X_{ij}^2)$, $i = 1, \ldots n; j = 1, \ldots p$.
The empirical type I error rates were evaluated when generating data under scenario (A), and the empirical powers were assessed under scenarios (B), (C) and (D). Under each scenario, $N = 10000$ datasets were simulated with varied sample sizes $n = \{20, 40, 60, 80, 100\}$. For the KRV test and HSIC test, we applied the Gaussian kernel to both X and Y to test independence (Gretton et al., 2008). That is, $k(X_1, X_2) = l(X_1, X_2) = \exp\{-||X_1 - X_2||^2/\sigma^2\}$, where $||X_1 - X_2||^2$ is the Euclidean distance between X_1 and X_2, σ^2 is the shape parameter which was selected as the median of the Euclidean distance between each sample pair. The design of the HSIC test is different from the KRV test. The asymptotic null distribution of HSIC statistic is characterized as $\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\mu_j\chi_1^2$, where λ_i, μ_j are eigenvalues of kernel matrices K and L respectively. More details can be found in Sejdinovic et al. (2013). For the GRV test, Euclidean, Manhattan and Mahalanobis distance have been proposed for continuous variables (Minas et al., 2013). For simplicity, we selected both Euclidean distances for X and Y in GRV test (GRV re-

sults with Manhattan and Mahalanobis distance are qualitatively similar). Finally, $B = 10000$ permutations were used in the dcov test (Székely et al., 2007). The nominal significance level was set at $\alpha = 0.05$ and the testing results are reported in Figure 1.

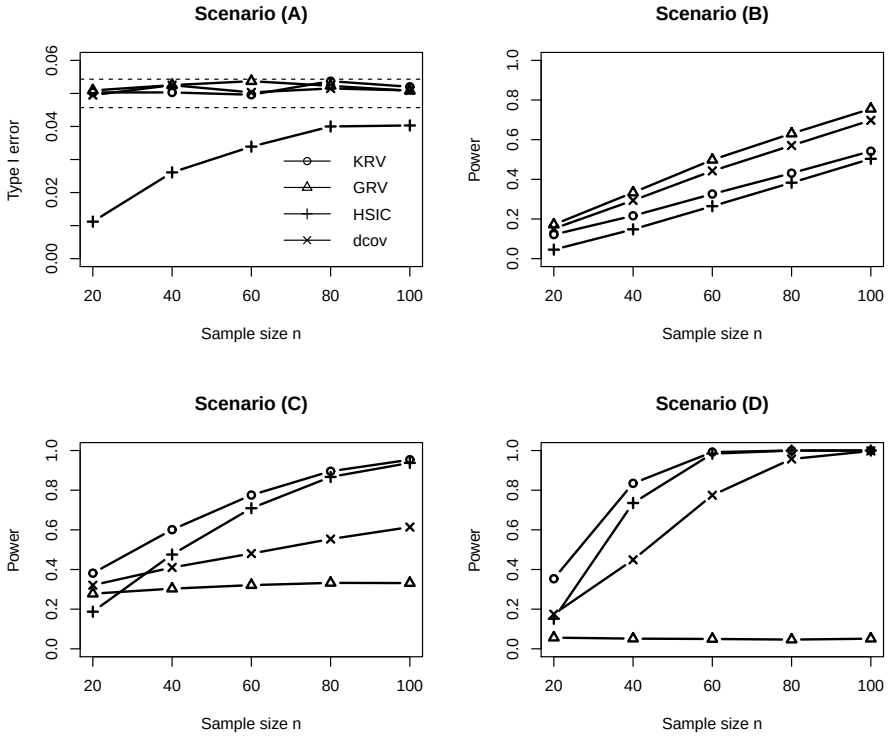

Figure 1: Empirical type I error/power of KRV, GRV, HSIC and dcov test. Scenario (A) is for type I error and Scenario (B)–(D) are for powers under different alternative models. Symbols $\circ$, $\triangle$, $+$ and $\times$ represent KRV, GRV, HSIC and dcov respectively.

Under scenario (A), KRV, GRV and dcov test have correct type I error. The 95% CI of type I error is $0.05 \pm 1.96\sqrt{0.05 \cdot 0.95/10000} = [0.0457, 0.0543]$, which are represented as dash lines in the top-left panel of Figure 1. Clearly, the HSIC test is outside this CI and is extremely conservative especially when sample size is small. This small-sample conservativeness has been observed for other kernel-based association test statistics (Chen et al., 2016). Under scenario (B), GRV and dcov are more powerful than KRV and HSIC. The dependence between X and Y under Scenario (B) is fully described by the Pearson correlation ($Cov(X_i, Y_j) = 0.1$, $i, j = 1, \ldots 5$), and the Gaussian

kernels as applied in KRV and HSIC are less sensitive to such a linear dependency pattern than the Euclidean distances implemented in GRV. The dependency between X and Y under scenario (C) is linear but with random coefficient. KRV and HSIC are more powerful than GRV under this scenario. Finally, there is a nonlinear dependency between X and Y under scenario (D). Since the dependency is purely deterministic, KRV, HSIC and docv is extremely powerful under this scenario. On the other hand, GRV with Euclidean distances fails to detect such a nonlinear dependency in the sense that it has a power close to the nominal type I error rate. GRV tests with other distances (such as Manhattan and Mahalanobis distance) can have improved power, which however, is still less powerful than KRV (data not shown).

To summarize, KRV test is powerful in detecting any kind of departure from statistical independence under each scenario given the kernels are appropriately chosen, such as Gaussian kernels (Gretton et al., 2008). Depending on the distances being used, GRV test can be powerful in detecting certain kind of dependency patterns. However, it is not clear, under what conditions/distances, GRV is able to capture any general dependency patterns among two random vectors. HSIC seems to be as powerful as KRV when the sample size is large. However, it is clear that HSIC is conservative when sample size is relatively small. The permutation-based dcov test tends to be slightly less powerful than KRV (except for Scenario (B)) and always has adequate power to detect any dependencies. However, the computational cost of dcov can be expensive if required number of permutations is large (e.g., for stringent significance levels).

4.2 Microbiome Association Simulation

We also conducted simulation studies to evaluate the performance of KRV in testing microbiome association. We first generated the microbiome composition data which was reflective of real OTU counts in a upper-respiratory-tract microbiome dataset (Charlson et al., 2010). A total of 856 OTUs were simulated and were further partitioned into 20 clusters using the partitioning around medoids algorithm. Finally, we selected a relatively abundant cluster (denoted by $\mathcal{A}$) as the one which affected the outcomes. After the OTU counts $Z_{ij}, i = 1, \ldots n, j = 1, \ldots 856$ were generated, we simulated q host

gene expressions from

$$y_{it} = 0.5X_{i1} + 0.5X_{i2} + \beta_t \cdot scale(\sum_{j \in \mathcal{A}} Z_{ij}) + \epsilon_{it},\ i = 1, \ldots n,\ t = 1, \ldots, q, \quad (4)$$

where X_{i1}, X_{i2} are covariates such as age, gender and smoking status, which may also be related to the microbiome composition. In particular, two different ways of simulating covariates were considered. In the first scenario, the covariates were independent of OTUs, and simulated as $X_{i1} \sim Bernoulli(0.5)$, $X_{i2} \sim N(0,1)$. In the second scenario, we simulated X_{i2} as $N(0,1) + 0.4 \cdot scale(\sum_{j \in \mathcal{A}} Z_{ij})$, which was related to the microbiome composition. The $scale(\cdot)$ function standardized the sum of OTU counts in cluster $\mathcal{A}$. The error terms ϵ_{ik} are independent and identically distributed as normal with mean zero and covariance matrix $\Sigma(\rho)$, where $\Sigma(\rho)$ is compound symmetry covariance matrix with $\rho = 0.2, 0.8$ representing low and high correlation among gene expressions respectively. We simulated $n = 200$ samples and $p = 30$ gene expressions to mimic a mid-size pathway as analyzed in a real data example later in this paper. Under the null model, all $\beta_t = 0$ and 10000 datasets were simulated to evaluate type I error. Two different alternative models were considered. One was the sparse-association model, where only 20% of the gene expressions are related to OTUs. In particular, we set $\beta_t = 0.5$ for $t = 1, \ldots q^*(= 0.2q)$, and zero elsewhere. The other is the dense-association model, where $\beta_t = 0.5$ for $t = 1, \ldots q^*(= 0.5q)$, and zero elsewhere. Under both alternative models, we generated 1000 datasets to assess the power.

To test the association between the simulated microbiome composition and gene expressions data, six different methods were applied including KRV test, GRV test, Gene Association with Multiple Traits (GAMuT) test (Broadaway et al., 2016), Multi-trait Sequence Kernel Association Test (MSKAT) (Wu and Pankow, 2016), Multivariate MiRKAT (MMiRKAT) (Zhan et al., 2017) and the marginal MiRKAT (Zhao et al., 2015). GAMuT uses the same design of HSIC test in previous simulation (Broadaway et al., 2016). MSKAT combine multiple marginal score test statistic through the covariance matrix of all scores and also calculates its p-value asymptotically (Wu and Pankow, 2016). MMiRKAT incorporates a small-sample adjustment to a MSKAT-type test so that the test has a better finite-sample behavior (Zhan et al., 2017). Finally, the marginal MiRKAT tests the association

between one gene expression and OTUs each time followed by Bonferroni correction to the minimum p-value, and we term it as minP for simplicity in the rest of this paper.

We first selected the OTU kernels as used in all six tests. For a little abuse of notation, in this section, we simply use the term kernels for distances when the test is GRV. The weighted UniFrac kernel, unweighted UniFrac kernel, generalized UniFrac kernel with parameter $\theta = 0.5$ and the Bray-Curtis kernel were considered (Zhao et al., 2015). We denote these kernels as K_w, K_u, $K_{0.5}$ and K_{BC} respectively. Then, the omnibus OTU kernel $K_{om} = K_w/tr(K_w)+K_u/tr(K_u)+K_{0.5}/tr(K_{0.5})+K_{BC}/tr(K_{BC})$ was also calculated and applied in all six tests. For the gene expression data, the Gaussian kernel-based KRV/GAMuT is shown to be robust in the previous continuous variables simulation in Section 4.1. To capture the correlation among gene expression, the weighted linear kernel $L = Y\hat{\Sigma}_{YY}^{-1}Y'$ is often shown to be useful (Broadaway et al., 2016). Based on the results of Section B.1 in Appendix B, we selected the gene expression kernel in KRV and GAMuT as $G/tr(G) + L/tr(L)$. On the other hand, the Euclidean distance, Manhattan distance and Mahalanobis distance are recommended in the GRV test (Minas et al., 2013). The Mahalanobis distance tends to be powerful when outcome correlation is high while the other two distances are more powerful with weakly correlated outcomes. An omnibus distance to accommodate three distances was used. Since the trace of a distance matrix is zero, we simply used an average distance matrix of the three in the GRV test.

The empirical type I errors are reported in Table 1. Based on the table, KRV and GRV always have correct type I error under each scenario. GAMuT and MSKAT tend to be very conservative under each scenario, which is also observed in Section 4.1 and other studies (Zhan et al., 2017). This is because the asymptotic p-value calculation in GAMuT and MSKAT work for large-sample genetic association studies, and tends to be conservative with small samples due to estimation error in variance terms (Chen et al., 2016). The small-sample adjustment incorporated in MMiRKAT usually works well with low-dimensional outcomes (Zhan et al., 2017). However, MMiRKAT seems to be a little conservative in this simulation with $p = 30$ outcomes. Finally, minP has correct type I error when outcomes are weakly correlated ($\rho = 0.2$) and is very conservative when outcomes are highly correlated ($\rho = 0.8$). This is due to the conservativeness of the Bonferroni correction when individual

Table 1: Empirical type I error of KRV, GRV, GAMuT, MSKAT, MMiRKAT and minP at nominal level $\alpha = 0.05$.

	$\rho = 0.2$					$\rho = 0.8$				
Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}
KRV	0.0493	0.0512	0.0499	0.0483	0.0500	0.0498	0.0504	0.0527	0.0510	0.0534
GRV	0.0455	0.0475	0.0476	0.0471	0.0452	0.0492	0.0497	0.0489	0.0545	0.0519
GAMuT	0.0271	0.0182	0.0158	0.0207	0.0130	0.0330	0.0200	0.0186	0.0229	0.0176
MSKAT	0.0349	0.0238	0.0254	0.0256	0.0258	0.0341	0.0227	0.0254	0.0278	0.0257
MMiRKAT	0.0383	0.0367	0.0381	0.0350	0.0360	0.0360	0.0367	0.0379	0.0390	0.0380
minP	0.0479	0.0454	0.0491	0.0434	0.0485	0.0188	0.0212	0.0220	0.0206	0.0201

tests are highly correlated. The type I errors of all tests with dependent (X,Z) scenario are similar and reported in Table S2 in Section B.2 of Appendix B.

The empirical powers are reported in Table 2. We first compare the performance of each test with different OTU kernels. Data generated in this simulation have two features. First, the simulated OTUs are phylogenetically related, and reflect a real upper-respiratory-tract microbiome data. Second, based on simulation model (4), the outcomes are affected by the abundance of OTUs (i.e. Z_{ij}), rather than the presence/absence of OTU (i.e. $I[Z_{ij} > 0]$). Given these facts, K_w and $K_{0.5}$ consider both phylogeny and abundance information, and hence are more powerful. On the other hand, K_u ignores the abundance information and K_{BC} ignores the phylogeny information, hence are less powerful. Finally, one can see that tests based on omnibus OTU kernel are quite robust. Under each scenario, the omnibus tests are slightly less powerful than the best test but much more powerful than the worst one.

Next, we compare the power of different tests. We first compare four kernel-based multivariate association tests: KRV, GAMuT, MSKAT and MMiRKAT. Both KRV and GAMuT gain additional power by utilizing an additional kernel to model the structures in gene expression data. Also, as observed in Table 1, GAMuT, MSKAT and MMiRKAT are more or less conservative under small sample size. These two facts explain that KRV is consistently more powerful than GAMuT, MSKAT and MMiRKAT in Table 2. Next, we compare KRV and GRV. Under $\rho = 0.2$, GRV is slightly more powerful than KRV. However, KRV is much more powerful than GRV under $\rho = 0.8$ especially when $q^* = 6$, where the power of KRV and GRV are 0.856 and 0.166 respectively. We also tried other GRV tests. For example, Mahalanobis distance-based GRV has improved power under $\rho = 0.8$ but has much lower power than KRV under $\rho = 0.2$. Similar to previous simulations in Section 4.1, the Gaussian kernel in KRV is often robust to capture general relationship while it is not clear which distance in GRV can achieve such goals. Finally, the comparison between KRV and minP is simple. Under low correlation and sparse signal, minP is slightly more powerful. However, under other scenarios, the association signal can be largely amplified by collectively analyzing multiple outcomes and thus KRV can be much more powerful than minP. The powers of all tests with dependent (X,Z) scenario are similar and reported in Table S3 in Section B.2 of Appendix B.

Table 2: Empirical power of KRV, GRV, GAMuT, MSKAT, MMiRKAT and minP at nominal level $\alpha = 0.05$.

		$\rho = 0.2$					$\rho = 0.8$				
q^*	Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}
6	KRV	0.784	0.084	0.718	0.345	0.677	0.856	0.066	0.804	0.403	0.759
	GRV	0.809	0.080	0.767	0.387	0.614	0.166	0.063	0.157	0.102	0.133
	GAMuT	0.706	0.032	0.532	0.203	0.475	0.803	0.031	0.637	0.275	0.550
	MSKAT	0.277	0.037	0.420	0.116	0.307	0.474	0.037	0.672	0.173	0.522
	MMiRKAT	0.546	0.063	0.458	0.185	0.424	0.799	0.061	0.688	0.333	0.651
	minP	0.834	0.068	0.913	0.381	0.828	0.610	0.036	0.684	0.234	0.601
15	KRV	0.978	0.086	0.946	0.579	0.935	0.969	0.096	0.951	0.603	0.925
	GRV	1.000	0.123	0.999	0.878	0.991	0.574	0.059	0.531	0.261	0.413
	GAMuT	0.963	0.038	0.886	0.439	0.817	0.960	0.034	0.870	0.441	0.827
	MSKAT	0.336	0.027	0.525	0.129	0.357	0.488	0.038	0.737	0.212	0.579
	MMiRKAT	0.662	0.054	0.548	0.250	0.510	0.844	0.066	0.772	0.357	0.729
	minP	0.971	0.081	0.991	0.593	0.971	0.697	0.033	0.772	0.313	0.684

To conclude, there is no uniform most powerful multivariate association test in our simulations. Unlike other methods, which suffer from huge power loss under certain scenarios, the proposed KRV test is always one of the most powerful method in testing the association between OTUs and gene expressions, and always has an adequate power under each scenario.

5 Analysis of host transcriptome and microbiome data

We further applied the KRV test to a dataset from an inflammatory bowel disease (IBD) study (Morgan et al., 2015), which examines how host transcriptome interacts with microbiome in the pathogenesis of IBD. Paired host transcriptome and microbial metagenome data were collected from 255 samples, among which 196 were pre-pouch ileum (PPI) samples and 59 were pouch samples. For each sample, 19908 host transcript expressions and 7000 OTU counts were measured by microarray and 16S rRNA analysis respectively (Morgan et al., 2015). Besides host gene expression and microbiome composition, three additional covariates are available: antibiotic use (yes/no), inflammation score (0-13), and disease outcome (familial adenomatous polyposis or not). Due to heterogeneity reasons, only the 196 PPI samples were used to test the association between host transcriptome and microbiome (Morgan et al., 2015). In particular, a linear model was applied to test the association between each individual transcript and each individual OTU after accounting for the covariates. To reduce multiple testing burden and improve statistical power, principal component analysis (PCA) was applied to the 19908 host transcripts and 7000 OTUs for dimension reduction. The top 9 host PCs (which explain 50% variance in host transcripts) and the top 9 clade PCs (which explain 50% variance in OTUs) were included in individual association analysis, where one host PC and one clade PC is tested for association each time. Finally, after multiple testing adjustment, significant associations between host PCs and clade PCs can be detected at a false discovery rate (FDR) of 0.25. The authors also noted enrichment of microbiome-associated host transcript patterns within the interleukin-12 (IL12) pathway, but no formal statistical testing results were reported (Morgan et al., 2015).

Alternatively to the individual PC based association analysis implemented

in the original study, we jointly tested the association between host gene expressions (either the whole transcriptome or within a certain pathway as IL12) and all 7000 OTUs using all six methods as illustrated in simulation studies. Besides the whole transcriptome and IL12 pathway, we also analyzed two additional pathways. One is Inflammatory mediator regulation of TRP channels pathway (KEGG: hsa04750), and the other is IBD pathway (KEGG:hsa05321). These two pathways are either related to the underlying biological process or related to the disease itself, hence can be of interest. To be consistent with the original studies (Morgan et al., 2015), only the 196 PPI samples were used in our analysis.

For the OTU data, the Bray-Curtis kernel can be directly calculated from the counts, and the phylogenetic tree needs to be first trained for calculating UniFrac-type kernels. Specifically, PyNAST (Caporaso et al., 2010) was used to generate a multiple sequence alignment from the representative OTU sequences identified in the original study. Of the 7000 available OTU sequences, 1646 could not be aligned and were excluded from the phylogenetic tree. A phylogenetic tree relating the remaining 5354 OTUs was produced using FastTree (Price, Dehal, and DehalArkin, 2009). The unweighted, weighted, and generalized UniFrac distances/kernels were calculated using this tree. The same kernel/distance for gene expression data as in Section 4.2 were used in this real data application. For the whole transcriptome, which contains too many genes ($p = 19908 > n = 196$) such that $\hat{\Sigma}$ is not invertible. Thus we simply used the Gaussian kernel in KRV, GRV and GAMuT, and $\hat{\Sigma}^{-1}$-based MSKAT and MMiRKAT are not evaluated under this scenario.

The testing results are reported in Table 3. For the overall association between microbiome composition and all 19908 genes in the whole transcriptome, KRV, GRV and GAMuT are all highly significant while minP is not, probably due to the heavy multiple testing correction burden. Compared with the claimed significance at FDR=0.25 of the original individual analysis, our KRV test is much more powerful detecting associations since it can amplify the marginal association signal by analyzing both OTUs and gene expressions collectively.

For the IL12 pathway, KRV, GRV, GAMuT and minP (except for K_w) are significant at $\alpha = 0.05$ level, which are consistent with findings of the original study stating that microbiome-associated host genome PCs were

Table 3: P-values of different tests examining the host-microbiome association in the real data. The whole transcriptome (whole) contains all 19908 genes, IL12 pathway contains 21 genes, Inflammatory pathway (IF) contains 96 genes, and IBD pathway has 62 genes.

Pathway	Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}
Whole	KRV	0.0011	0.0002	0.0003	0.0014	0.0002
	GRV	0.0055	0.0003	0.0015	0.0024	0.0012
	GAMuT	0.0015	0.0006	0.0026	0.0029	0.0005
	minP	1.0000	1.0000	1.0000	1.0000	1.0000
IL12	KRV	0.0010	0.0004	0.0004	0.0014	0.0003
	GRV	0.0040	0.0003	0.0011	0.0021	0.0009
	GAMuT	0.0017	0.0011	0.0009	0.0029	0.0007
	MSKAT	0.1931	0.5000	0.3024	0.1739	0.2105
	MMiRKAT	0.1744	0.4376	0.3295	0.1674	0.2184
	minP	0.0759	0.0060	0.0237	0.0448	0.0164
IF	KRV	0.0013	0.0003	0.0003	0.0015	0.0003
	GRV	0.0042	0.0002	0.0011	0.0020	0.0009
	GAMuT	0.0018	0.0008	0.0007	0.0029	0.0007
	MSKAT	0.6772	0.5859	0.8096	0.3337	0.6921
	MMiRKAT	0.6288	0.7016	0.7127	0.4383	0.6475
	minP	0.3236	0.0189	0.0974	0.1207	0.0602
IBD	KRV	0.0015	0.0002	0.0004	0.0016	0.0003
	GRV	0.0041	0.0002	0.0011	0.0021	0.0009
	GAMuT	0.0022	0.0007	0.0008	0.0032	0.0007
	MSKAT	0.8046	0.3658	0.6958	0.4789	0.6711
	MMiRKAT	0.7286	0.4402	0.6199	0.4788	0.6248
	minP	0.2090	0.0079	0.0502	0.0698	0.0357

enriched in IL12 pathway (Morgan et al., 2015). Thus, formal statistical inference by KRV and other methods provides support for previous scientific observations. Compared with MSKAT and MMiRKAT, the additional gene expression kernel in KRV boosts its power of detecting associations. For the other two pathways (Inflammatory and IBD), KRV, GRV, and GAMuT are significant while MSKAT, MMiRKAT and minP mostly fail to detect any significance at $\alpha = 0.05$ level except for K_u-based minP. Among all tests, KRV seems to be most powerful in that it always has the smallest p-value under each scenario.

To summarize, the association between individual host transcript and microbiome seems to be weak and complicated. KRV can amplify the association signal by collectively analyzing multiple OTUs and multiple genes, which is more powerful than the original PC-based individual association analysis. The usage of an additional kernel modeling structures and capturing general relationship, along with the fast and robust p-value calculation make KRV more powerful than other methods.

6 Discussion

In this paper, we consider the problem of associating overall microbiome composition with host genomics and propose the KRV test, which can both adjust for confounder effect and accommodate multiple candidate kernels reflecting different data structures or association patterns. As shown in the simulation studies, the proposed KRV test has correct size and can have substantially higher power than existing similar tests in many scenarios. Moreover, KRV testing results on the host-microbiome data not only provides formal statistical inference to support original conclusion (Morgan et al., 2015), but also is able to facilitate microbiome community level analysis and provide additional insights on some other related pathways.

One major contribution of this paper is that we largely adapted the existing GRV test in the microbiome association analysis framework, making it better suited to the host genome-microbiome association problem considered in this paper. KRV extends GRV in the following aspects. First, by applying kernels, KRV is able to capture both more complicated data structure (i.e., the phylogenetic structure inherent to microbiome data) and more general dependencies between two sets of variables. Second, we further extend the

GRV test in a comprehensive association testing framework. KRV can adjust for confounder effect, which is important yet has never been discussed in the GRV test. Furthermore, we propose an omnibus KRV test based on a linear combinations of multiple candidate kernels, which is computationally much more efficient than the way GRV accommodates multiple distances. The omnibus KRV test is robust against the underlying data structures and association patterns. Due to these differences, we think that KRV not only can coexist with the existing GRV test but also can provide beneficial complements to GRV. Another contribution of this paper is that the KRV test provides an important complement to existing statistical independence tests (Gretton et al., 2008, Székely et al., 2007) by providing an efficient test design which neither relies on large samples nor requires permutations. The approximated Pearson type III distribution of the KRV statistic may also shed light on the finite-sample distribution of other statistics such as HSIC and distance covariance.

The proposed KRV in this paper is mainly aimed at microbiome association analysis, however, application of KRV can be beyond this aim. The proposed KRV test can also be useful in other domains due to the following reasons. First, KRV is extremely flexible. X or Y considered in KRV can be either a single variable or a high-dimensional vector. Moreover, its good finite-sample performance makes it an ideal tool for those studies with relatively small sample size, such as metabolomics and proteomics (Zhan et al., 2015). Second, the application of kernels enables KRV to capture structured data types, such as networks, shapes and images as long as appropriate kernels are designed. We leave these to future investigations.

7 Appendices

Appendix A: KRV coefficient and its approximated Pearson type III distribution

A.1 Kernel trick and KRV coefficient

RV coefficient is only able to capture the linear dependency between two random vectors and does not accommodate nonlinearity or other more general dependencies. In practice, complex data such as microbiome and host

genome data, often require general methods to detect more general dependencies that are of interest (Hofmann et al., 2008). Motivated by this, we propose the KRV coefficient to measure more general relationship between microbiome composition and host genome expression.

A symmetric bivariate function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto R$ is a kernel if

$$\int_x \int_y k(x,y)f(x)f(y)dxdy \geq 0,$$

for all functions $f \in L_2(\mathcal{X})$. We always assume that $\mathcal{X}$ is a compact subset of $\mathcal{R}^p$ in this paper. A nice property of kernel is the so called "kernel trick", which states that

$$k(x,y) =< \phi(x), \phi(y) >_{\mathcal{K}}, \tag{5}$$

for some $\phi : \mathcal{X} \mapsto \mathcal{K}$, where $\mathcal{K}$ is some (possibly high or even infinite dimensional) space with inner-product $< \cdot,\cdot >_{\mathcal{K}}$. $\mathcal{K}$ is called the feature space, and ϕ is called kernel (feature) map associated with $k(\cdot,\cdot)$. If we complete $\mathcal{K}$ in the norm induced by the inner-product, then $\mathcal{K}$ is called reproducing kernel Hilbert space (RKHS) (Hofmann et al., 2008).

In the spirit of the kernel trick, we develop the KRV coefficient by calculating RV coefficient in RKHSs. Let $\phi : \mathcal{X} \mapsto \mathcal{K}$ and $\psi : \mathcal{Y} \mapsto \mathcal{L}$ denote two kernel maps associated with kernels $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ respectively. Then, we can define the RV coefficient between the RKHS-images $\phi(X)$ and $\psi(Y)$ as the KRV coefficient between X and Y, that is, $KRV(X,Y) := RV(\phi(X),\psi(Y))$. To calculate the KRV coefficient, we replace the original inner product $< X_i, X_j >_{\mathcal{X}} = X_i'X_j$ in the input space $\mathcal{X}$ with the inner-product in RKHS $\mathcal{K}$, that is $< \phi(X_i), \phi(X_j) >_{\mathcal{K}} = k(X_i, X_j)$. In other words, matrix XX' in the original RV coefficient should be replaced by kernel matrix K with $K_{ij} = k(X_i, X_j)$ in the KRV coefficient. Correspondingly, matrix YY' should be replaced by kernel matrix L, where $L_{ij} = l(Y_i, Y_j)$. Considering X and Y are centered by columns, we correspondingly use the centralized kernel matrix $\tilde{K} = HKH$ and $\tilde{L} = HLH$, where $H = I - \mathbf{1}\mathbf{1}'/n$ is a centering matrix, I is an identity matrix of order n, and $\mathbf{1}$ is a $n \times 1$ vector of all ones. After plugging all the these results into expression of the RV coefficient in the main text, some simple calculations give

$$KRV(X,Y) := RV(\phi(X),\psi(Y)) = \frac{tr(\tilde{K}\tilde{L})}{\sqrt{tr(\tilde{K}\tilde{K})}\sqrt{tr(\tilde{L}\tilde{L})}}. \tag{6}$$

A.2 Derivation of Pearson type III distribution approximation

Let Q denote the observed KRV statistic and $\{Q_1, \ldots, Q_{n!}\}$ denote all $n!$ possible permutations of the KRV statistic. To derive the Pearson type III approximation of the KRV permutation distribution, we first calculate $E_{H_0}(Q)$, $Var_{H_0}(Q)$ and $E_{H_0}(Q^3)$ (or skewness $\gamma_{H_0}(Q)$), where the expectation is with respect to the empirical distribution of $\{Q_1, \ldots, Q_{n!}\}$ under the null model. By applying the kernel trick to existing results on moments of the RV-type statistics (Josse et al., 2008, Kazi-Aoual et al., 1995, Minas et al., 2013), we obtain moments for the KRV statistic. In particular,

$$E_{H_0}(Q) = \frac{\sqrt{\beta_X \beta_Y}}{n-1}, \ \textit{where} \ \beta_X = [tr(\tilde{K})]^2/tr(\tilde{K}^2), \beta_Y = [tr(\tilde{L})]^2/tr(\tilde{L}^2)$$

$$Var_{H_0}(Q) = \frac{2(n-1-\beta_X)(n-1-\beta_Y)}{(n+1)(n-1)^2(n-2)}\left(1 + \frac{n-3}{2n(n-1)}\tau_X\tau_Y\right),$$

$$\textit{where} \ \tau_X = \frac{n-1}{(n-3)(n-1-\beta_X)}\left(n(n+1)\frac{\sum_i(\tilde{K}_{ii})^2}{tr[\tilde{K}^2]} - (n-1)(\beta_X+2)\right),$$

and τ_Y is defined correspondingly. For the third moment, we have

$$
\begin{aligned}
&n(n-1)(n-2)(n-3)(n-4)(n-5)E_{H_0}(Q^3)\\
&= n^2(n+1)(n^2+15n-4)S_3^K S_3^L + 4(n^4-8n^3+19n^2-4n-16)U^K U^L\\
&+ 24(n^2-n-4)(U^K B^L + U^L B^K) + 6(n^4-8n^3+21n^2-6n-24)B^K B^L\\
&+ 12(n^4-n^3-8n^2+36n-48)R^K R^L + 12(n^3-2n^2+9n-12)(T^K S_2^K R^L + T^L S_2^L R^K)\\
&+ 3(n^4-4n^3-2n^2+9n-12)T^K T^L S_2^K S_2^L + 24(n^3-3n^2-2n+8)(R^K U^L + R^L U^K)\\
&+ 24(n^3-2n^2-3n+12)(R^K B^L + B^K R^L) + 12(n^2-n+4)(T^K S_2^K U^L + T^L S_2^L U^K)\\
&+ 6(2n^3-7n^2-3n+12)(T^K S_2^K B^L + T^L S_2^L B^K)\\
&- 2n(n-1)(n^2-n+4)\{(2U^K+3B^K)S_3^L + (2U^L+3B^L)S_3^K\}\\
&- 3n(n-1)^2(n+4)\{(T^K S_2^K + 4R^K)S_3^L + (T^L S_2^L + 4R^L)S_3^K\}\\
&+ 2n(n-1)(n-2)\{[(T^K)^3 + 6T^K T_2^K + 8T_3^K]S_3^L + [(T^L)^3 + 6T^L T_2^L + 8T_3^L]S_3^K\}\\
&+ (T^K)^3[(n^3-9n^2+23n-14)(T^L)^3 + 6(n-4)T^L T_2^L + 8T_3^L]\\
&+ 6T^K T_2^K[(n-4)(T^L)^3 + (n^3-9n^2+24n-14)T^L T_2^L + 4(n-3)T_3^L]\\
&+ 8T_3^K[(T^L)^3 + 3(n-3)T^L T_2^L + (n^3-9n^2+26n-22)T_3^L] - 16[(T^K)^3 U^L + U^K (T^L)^3]\\
&- 6(2n^2-10n+16)(T^K T_2^K U^L + U^K T^L T_2^L) - 8(3n^2-15n+16)(T_3^K U^L + U^K T_3^L)\\
&- (6n^2-30n+24)[(T^K)^3 B^L + B^K (T^L)^3] - 6(4n^2-20n+24)(T^K T_2^K B^L + B^K T^L T_2^L)\\
&- 8(3n^2-15n+24)(T_3^K B^L + B^K T_3^L) - 24(n-2)[(T^K)^3 R^L + R^K (T^L)^3]\\
&+ 6(n-2)(2n^2-10n+24)(T^K T_2^K R^L + R^K T^L T_2^L) + 8(n-2)(3n^2-15n+24)(T_3^K R^L + R^K T_3^L)\\
&+ (n-2)(3n^2-15n+6)[(T^K)^3 T^L S_2^L + (T^L)^3 T^K S_2^K] + 48(n-2)(T_3^K T^L S_2^L + T_3^L T^K S_2^K)\\
&+ 6(n-2)(n^2-5n+6)(T^K T_2^K T^L S_2^L + T^K S_2^K T^L T_2^L),
\end{aligned}
$$

where $T^K = tr(\tilde{K})$, $T_2^K = tr(\tilde{K}^2)$, $T_3^K = tr(\tilde{K}^3)$, $S_2^K = \sum_i (\tilde{K}_{ii})^2$, $S_3^K = \sum_i (\tilde{K}_{ii})^3$, $U^K = \sum_i \sum_j (\tilde{K}_{ij})^3$, $B^K = [diag(\tilde{K})']' \tilde{K} diag(\tilde{K})$, $R^K = [diag(\tilde{K})]' diag(\tilde{K}^2)$ are all scalars. Correspondingly, $T^L, T_2^L, T_3^L, S_2^L, S_3^L, U^L, B^L, R^L$ are the values calculated from kernel matrix $\tilde{L}$. Using results of first three moments, the skewness is calculated as

$$\gamma_{H_0}(Q) = \frac{E_{H_0}(Q^3) - 3E_{H_0}(Q)Var_{H_0}(Q) - E_{H_0}^3(Q)}{Var_{H_0}^{3/2}(Q)}.$$

For simplicity, we use μ, σ^2 and γ to represent $E_{H_0}(Q)$, $Var_{H_0}(Q)$ and $\gamma_{H_0}(Q)$ respectively. Then the Pearson type III density with exact the same

three moments are given by

$$f(x) = \frac{1}{|s|^a \Gamma(a)} |x - \lambda|^{a-1} \exp\left\{-\frac{x-\lambda}{s}\right\},$$

where $a = 4/\gamma^2$, $s = \sigma\gamma/2$ and $\lambda = \mu - 2\sigma/\gamma$. Finally, the p-value of the KRV test can be analytically computed based on this approximated Pearson type III probability density.

The approach we present in this section closely follows the testing strategy used in existing RV-type statistics (Josse et al., 2008, Minas et al., 2013). However, since two kernel matrices are used (rather than two outer product matrices in RV (Josse et al., 2008) and distances matrices in GRV (Minas et al., 2013)), like RV and GRV, we also conduct our numerical studies to evaluate the approximation performance of the Pearson type III probability to the empirical null distribution of KRV permutations. Results of these numerical studies are presented in the next section.

A.3 Evaluation of Pearson type III approximation

In this section, we evaluate the approximation of Pearson type III density to the empirical null distribution of KRV permutations. A subset of the host transcriptome and microbiome data was used. In particular, the expressions of 21 genes in the IL12 pathway were taken as host genomic data, where both the Gaussian kernel and the linear kernel were calculated. All 7000 OTU counts were used to calculate the Bray-Curtis kernel. Then, KRV statistic were calculated separately using samples from each tissue location ($n = 196$ samples from PPI and $n = 59$ samples from pouch). Four different KRV statistic were evaluated: Gaussian and Bray-Curtis kernel with 196 PPI samples, Gaussian and Bray-Curtis kernel with 59 pouch samples, linear and Bray-Curtis kernel with 196 PPI samples, linear and Bray-Curtis kernel with 59 pouch samples. For each KRV statistic, we first calculated the approximated Pearson type III density based on description in the previous section. Then, we permuted the Bray-Curtis kernel one million times and calculated the corresponding KRV statistic using the permuted kernel. Finally, the Pearson type III density was compared to the sampling distribution of KRV permutations. The results are reported in Figure S1, where one can see that the Pearson type III density provides a good approximation to the sampling distribution of KRV permutation under each of the four

scenarios.

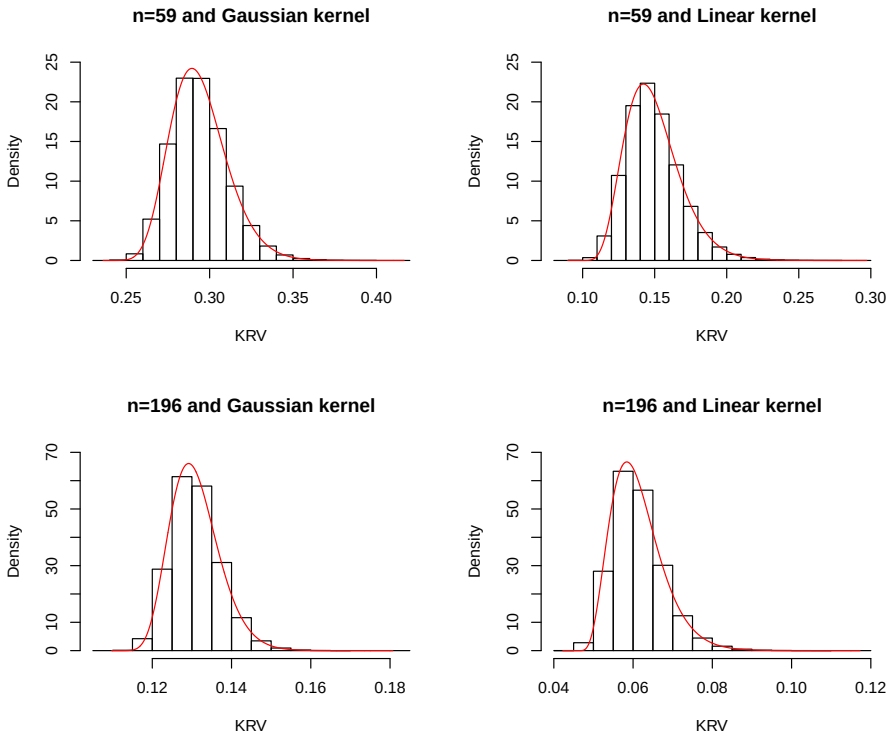

Figure S1: Sampling distribution of KRV statistic based on one million permutations (histogram) and the approximated Pearson type III distribution (curve).

Appendix B: Additional Simulations Studies

B.1 Comparison of different omnibus KRV tests

In this section, we compared different omnibus KRV tests as mentioned in the main text. Without loss of generality, we fix kernel gene expression kernel l and discuss the scenario of accommodating multiple candidate OTU kernels $k_1, \ldots k_m$, such as different UniFrac-type kernels and the Bray-Curtis kernel used in the main text. We denote the corresponding kernel matrices as L, $K_1, \ldots, K_m$ respectively.

Four different omnibus kernels have been proposed in the maintext to

accommodate multiple candidate kernels:

$$K_{om1} = \sum_{i=1}^{m} \frac{K_i}{m}; \quad K_{om2} = \sum_{i=1}^{m} \frac{K_i}{tr(K_i)}; \quad K_{om3} = \sum_{i=1}^{m} \omega_i^* K_i;$$

$$K_{om4} = K_i \textit{ with the minimum p-value.}$$

Among those omnibus kernels, K_{om1} and K_{om2} are trained unsupervised and a KRV test between the omnibus kernel and L can be directly used to calculate its p-value. On the other hand, K_{om3} and K_{om4} are trained supervised. Thus resampling procedures are needed to establish the final significance, which are introduced in the following.

Recall that the optimal weights $\boldsymbol{\omega}^* = (\omega_1^*, \ldots, \omega_m^*)'$ in K_{om3} were trained by maximizing the following target function:

$$KRV(K_{om}, L) = \frac{\sum_{i=1}^{m} tr(\omega_i K_i L)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} tr(\omega_i K_i \omega_j K_j)}\sqrt{tr(L^2)}}, \tag{7}$$

subjected to $\omega_i \geq 0$ and $\sum_{i=1}^{m} \omega_i = 1$ (Cortes et al., 2012). Since the weights ω_i^* depend on L, it is no longer valid to use a KRV test between $\sum_{i=1}^{m} \omega_i^* K_i$ and L to calculate the p-value. Alternatively, denote the p-value of $KRV(\sum_{i=1}^{m} \omega_i^* K_i, L)$ as the observed p-value p_o and let $L^b, b = 1, \ldots, B$ be a permutation of the kernel matrix L. For each permutation, one recalculates the weights ω_{ib}^* using L^b and the p-value p_b of $KRV(\sum_{i=1}^{m} \omega_{ib}^* K_i, L^b)$. The final p-value is calculated as $\sum_{b=1}^{B} I(p_o \geq p_b)/B$.

For K_{om4}, let $p_i, i = 1, \ldots, m$ denote the p-value of the KRV test between K_i and L, and $p_o = min_i\, p_i$, which is no longer a genuine p-value. Permutations are used to obtain a final p-value. In particular, let $L^b, b = 1, \ldots B$ be a permuted kernel matrix and $p_i^b, i = 1, \ldots, m$ denote the KRV p-value between K_i and L^b. Let the final p-value can be obtained by comparing p_o to $p^b = min_i\, p_i^b$, $b = 1, \ldots, B$ and calculated as $\sum_{b=1}^{B} I(p_o \geq p^b)/B$.

Comprehensive simulation studies have been conducted to evaluate the performance of K_{om1}, K_{om2}, K_{om3} and K_{om4}-based KRV tests. The setup of the simulation was the same as the one used in Section 4.2 of the main text. $B = 1000$ permutations were used to establish significance of K_{om3} and K_{om4}. For ease of presenting, we only report the KRV test results with unrelated OTU and covariates in Table S1. Simulation results with other tests and under related OTU and covariates scenario are qualitatively

similar and hence not reported.

Table S1: Comparison of different omnibus OTU-kernels. The first two rows ($q^* = 0$) are type I error, the next four rows ($q^* = 6, 15$) are power, and the last row is average computing time (in seconds) over 10,000 runs.

ρ	q^*	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om1}	K_{om2}	K_{om3}	K_{om4}
0.2	0	0.049	0.051	0.047	0.051	0.052	0.051	0.051	0.049
0.8	0	0.050	0.049	0.046	0.052	0.052	0.051	0.050	0.052
0.2	6	0.790	0.079	0.695	0.342	0.478	0.668	0.384	0.677
	15	0.984	0.081	0.957	0.625	0.801	0.939	0.734	0.958
0.8	6	0.859	0.096	0.789	0.409	0.573	0.753	0.473	0.742
	15	0.973	0.089	0.942	0.567	0.760	0.929	0.693	0.939
T		0.079	0.079	0.078	0.078	0.083	0.087	551.2	310.1

Based on Table S1, all tests have correct type I errors. For omnibus kernels, it seems that K_{om2} and K_{om4} are the most powerful tests in this simulation, followed by K_{om1} and then K_{om3}. Both K_{om2} and K_{om4} are slightly less powerful than the best individual KRV test (K_w) but much more powerful than the worst individual KRV test (K_u) under each scenario. K_{om1} is less powerful than K_{om2} and K_{om4}, which is reasonable in this simulation because the data simulating scheme favors K_w and $K_{0.5}$ (as analyzed in the main text). Thus, K_{om1} suffers from power loss by treating four candidates equally. Finally, K_{om3} is the least powerful test, which is not surprising under the current conditional-type test design: the underlying Pearson type III distribution depends on the kernels being used and the maximization of test statistic as in K_{om3} does not guarantee an optimal power due to the uncertainty in null distribution.

On the other hand, the computing time of K_{om1} and K_{om2} are basically the same as that of each individual kernel test. K_{om3} needs to solve a QP for each permutation and thus is much more computational expensive. Finally, the computational cost of a K_{om4}-based test is about mB ($m = 4, B = 1000$) times that of a individual kernel-based test. Considering both the power performance and computational efficiency, kernel K_{om2} is overall the best and is used as the default omnibus KRV test in the paper. For the GRV test, since the trace of a distance matrix is always zero, we simply use the computational

efficient simple averaging K_{om1} test as a fast way to incorporate multiple distances.

B.2 Simulation results with dependent covariates

In this section, we present the results of simulation data with dependent covariates as mentioned in Section 4.2 of the main text. In particular, we simulated X_{i2} as $N(0,1) + 0.4 \cdot scale(\sum_{j \in \mathcal{A}} Z_{ij})$ to introduce dependence between OTU and covariates. The type I errors and powers under this scenario are reported in Table S2 and Table S3 respectively.

Table S2: Empirical type I error of KRV, GRV, GAMuT, MSKAT, MMiRKAT and minP at nominal level $\alpha = 0.05$ with dependent OTUs and covariates.

ρ	Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}
0.2	KRV	0.0425	0.0521	0.0433	0.0457	0.0420
	GRV	0.0441	0.0479	0.0452	0.0471	0.0464
	GAMuT	0.0237	0.0166	0.0151	0.0189	0.0136
	MSKAT	0.0279	0.0227	0.0214	0.0248	0.0230
	MMiRKAT	0.0317	0.0376	0.0346	0.0343	0.0332
	minP	0.0483	0.0456	0.0464	0.0461	0.0469
0.8	KRV	0.0437	0.0502	0.0451	0.0473	0.0456
	GRV	0.0456	0.0494	0.0473	0.0500	0.0483
	GAMuT	0.0282	0.0201	0.0154	0.0222	0.0133
	MSKAT	0.0327	0.0255	0.0219	0.0243	0.0223
	MMiRKAT	0.0320	0.0370	0.0330	0.0354	0.0319
	minP	0.0166	0.0208	0.0202	0.0198	0.0195

Table S3: Empirical power of KRV, GRV, GAMuT, MSKAT, MMiRKAT and minP at nominal level $\alpha = 0.05$ with dependent OTUs and covariates.

ρ	q^*	Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_{om}
		KRV	0.582	0.072	0.501	0.251	0.474
		GRV	0.590	0.075	0.529	0.289	0.446
	6	GAMuT	0.488	0.028	0.313	0.157	0.275
		MSKAT	0.190	0.028	0.251	0.087	0.200
		MMiRKAT	0.367	0.043	0.294	0.152	0.276
		minP	0.588	0.062	0.684	0.240	0.605
0.2							
		KRV	0.898	0.099	0.831	0.457	0.788
		GRV	0.987	0.125	0.975	0.728	0.932
	15	GAMuT	0.839	0.039	0.670	0.301	0.621
		MSKAT	0.236	0.044	0.377	0.117	0.258
		MMiRKAT	0.489	0.060	0.426	0.184	0.388
		minP	0.816	0.088	0.902	0.414	0.835
		KRV	0.703	0.071	0.609	0.252	0.563
		GRV	0.118	0.049	0.115	0.067	0.092
	6	GAMuT	0.611	0.019	0.416	0.161	0.356
		MSKAT	0.336	0.030	0.510	0.122	0.359
		MMiRKAT	0.646	0.045	0.533	0.223	0.480
		minP	0.379	0.020	0.457	0.117	0.371
0.8							
		KRV	0.883	0.081	0.817	0.415	0.777
		GRV	0.410	0.066	0.386	0.185	0.297
	15	GAMuT	0.831	0.028	0.670	0.298	0.590
		MSKAT	0.386	0.034	0.563	0.158	0.417
		MMiRKAT	0.711	0.058	0.604	0.266	0.559
		minP	0.477	0.035	0.576	0.173	0.480

References

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., et al. (2016) A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *American Journal of Human Genetics* **98**, 525–540.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS One* **5**, e15216.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**, 2106–2113.

Chen, J., Chen, W., Zhao, N., Wu, M. C., and Schaid, D. J. (2016). Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. *Genetic Epidemiology* **40**, 5–19.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research* **13**, 795–828.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29**, 751–760.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *In Algorithmic learning theory* (pp. 63–77). Springer, Berlin Heidelberg.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. *In Advances in neural information processing systems* (pp. 585-592). MIT Press, Cambridge MA.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics* **36**, 1171–1220.

Hua, W. Y., and Ghosh, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics* **71**, 812–820.

Josse, J., Pagès, J., and Husson, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis* **53**, 82–91.

Josse, J., and Holmes, S. (2013). Measures of dependence between random vectors and tests of independence. Literature review. arXiv preprint arXiv:1307.7383

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J. D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational statistics & data analysis* **20**, 643–656.

Lasken, R.S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology* **10**, 631-640.

Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application* **2**, 73–94.

Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics* **63**, 1079–1088.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228–8235.

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative ß diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**, 1576-1585.

Minas, C., Curry, E., and Montana, G. (2013). A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics* **29**, 2555–2563.

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology* **16**, 67.

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R. and Wu, M. C. (2017). MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome* **5**, 17.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**, 1641–1650.

Purdom, E. (2006). Multivariate kernel methods in the analysis of graphical structures. PhD thesis, University of Standford.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **25**, 257–265.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* **41**, 2263–2291.

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* **44**, 846–849.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35**, 2769–2794.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**, 82–93.

Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., et al. (2013). Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. *Genetic Epidemiology* **37**, 267-275.

Wu, B., and Pankow, J. S. (2016). Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology* **40**, 91–100.

Zhan, X., Patterson, A. D., and Ghosh, D. (2015). Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinformatics* **16**, 77.

Zhan, X., Girirajan, S., Zhao, N., Wu, M. C., and Ghosh, D. (2016). A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics* **32**, 3603–3610.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017). A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology* In press, DOI: 10.1002/gepi.22030

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *American Journal of Human Genetics* **96**, 797–807.