Published in final edited form as:

Hum Hered. 2016 ; 81(2): 117–126. doi:10.1159/000448733.

Identifying Host Genetic Variants Associated with Microbiome Composition by Testing Multiple Beta Diversity Matrices

Xing Hua, James J. Goedert, Maria Teresa Landi, and Jianxin Shi

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Bethesda, Md., USA

Abstract

Objectives: Host genetics have been recently reported to affect human microbiome composition. We previously developed a statistical framework, microbiomeGWAS, to identify host genetic variants associated with microbiome composition by testing a distance matrix. However, statistical power depends on the choice of a microbiome distance matrix. To achieve more robust statistical power, we aim to extend microbiomeGWAS to test the association with many distance matrices, which are defined based on multilevel taxa abundances and phylogenetic information.

Methods: The main challenge is to accurately and rapidly evaluate the significance for millions of SNPs. We propose methods for approximating p values by correcting for the multiple testing introduced by testing many distance matrices and by correcting for the skewness and kurtosis of score statistics.

Results: The accuracy of p value approximation was verified by simulations. We applied our method to a set of 147 lung cancer patients with 16S rRNA microbiome profiles from nonmalignant lung tissues. We show that correcting for skewness and kurtosis eliminated dramatic deviations in the quantile-quantile plot.

Conclusion: We developed computationally efficient methods for identifying host genetic variants associated with microbiome composition by testing many distance matrices. The algorithms are implemented in the package microbiomeGWAS (<https://github.com/lsncibb/microbiomeGWAS>).

Keywords

Microbiome; Genome-wide association study; Host genetics; Tail probabilities; Skewness; Kurtosis

Introduction

The human microbiota is the collection of microbes inhabiting the human body, including bacteria, archaea, and fungi. Advances in high-throughput sequencing [1] and bioinformatic

Jianxin Shi, PhD, Biostatistics Branch, Division of Cancer Epidemiology and Genetics National Cancer Institute, 9609 Medical Center Drive, 7E596, Bethesda, MD 20892 (USA), Jianxin.Shi@nih.gov.

Disclosure Statement

The authors declare no conflict of interest.

analysis [2–4] allow the characterization of the microbiota's collective genome (the microbiome), including composition and diversity. Microbiome profiles can be generated rather efficiently by sequencing 16S rRNA gene amplicons. Differences in such human microbiome profiles have been reported to be associated with health conditions: the fecal microbiome with obesity [5, 6], colorectal cancer [7, 8], breast cancer [9], and inflammatory bowel disease [10]; the oral microbiome with pancreatic [11], oral, and gastrointestinal cancers [12]. Thus, identifying and understanding factors that affect microbiome composition and diversity is crucial for elucidating the etiology of human complex diseases and opportunities for prevention.

Evidence suggests that microbiome composition at a specific body site may be shaped by host genetics. As an example, genetic loci have been reported to be associated with the relative abundance (RA) of specific taxa of the gut microbiome based on linkage studies in mice [13, 14] and based on association studies in humans [15]. Interestingly, Knights et al. [16] reported that a risk single-nucleotide polymorphism (SNP) for inflammatory bowel disease located in *NOD2* was associated with the RA of Enterobacteriaceae in the human gut microbiome, providing an interesting hypothesis that the genetic locus affects the risk of inflammatory bowel disease by affecting the gut microbiome. Blekhman et al. [15] performed genetic studies across body sites in the Human Microbiome Project data and found that host genetic variants associated with microbiome composition were enriched in immunity-related pathways. In addition, a large-scale twin study suggested a strong genetic component for human gut microbiome [17]. These studies imply that identifying host genetic variants associated with microbiome composition has the potential to explain variability of human microbiome composition and to elucidate biological mechanisms for genetic associations with complex diseases.

In a microbiome genome-wide association study (GWAS), one may test the genetic association using alpha diversity metrics or RA of each taxon as phenotype. Because a microbiota functions as a community, one important analysis is testing genetic associations with beta diversity, that is, with a pairwise distance matrix defined based on 16S rRNA sequence data. At a truly associated SNP, microbiome distances are generally smaller for pairs of subjects with more similar genotypic values, i.e., the microbiome distance is positively associated with the genetic distance across subject pairs. Microbiome distances can be defined in different ways, based on phylogenetic tree information or on each taxon's abundance information. Bray-Curtis dissimilarity [18], Kullback-Leibler divergence, UniFrac [19–21], and generalized UniFrac (GUniFrac) distance [22] are widely used beta-diversity metrics.

We have recently developed a statistical framework, microbiomeGWAS [23], for identifying host genetic variants associated with microbiome composition using a single distance matrix. Previous reports and our empirical data [23] suggest that the power to identify associated genetic loci depends on which distance matrix is employed. As an example, based on a lung tissue microbiome GWAS study, a lung cancer risk SNP rs401681 located on chromosome 5p15.33 (*CLPTMIL*) was ambiguously associated with unweighted UniFrac distance matrix with $p = 0.056$ and more clearly with weighted UniFrac distance matrix with $p = 0.005$. This problem has been well realized in microbiome data analysis. Frequently,

associations in nongenetic studies (e.g., case-control studies) are tested using multiple distance matrices, and the most significant association is reported without explicitly accounting for multiple testing, which obviously causes an inflated type-I error rate. MiRKAT [24], a recently developed method for testing associations between microbiome composition and a trait (quantitative or binary), has the option for testing many distance matrices. However, MiRKAT relies on permutations to evaluate significance and thus is not feasible for analyzing millions of SNPs in GWAS.

In this article, we extend our previous analytic framework, microbiomeGWAS [23], to test the association with many microbiome distance matrices. The overall test statistic is defined as the strongest association (measured as the p value of each score statistic) across all tested distance matrices. We calculate the significance by correcting for the multiple testing imposed by testing many distance matrices. Importantly, we show that the skewness and kurtosis of each score statistic make the p-value approximation based on the asymptotic normal assumption too liberal, which is addressed by explicitly correcting for skewness and kurtosis of score statistics. The accuracy of the approximations is verified by extensive simulations. These theoretic investigations allow us to develop algorithms for efficiently analyzing genome-wide SNP markers in a GWAS study without relying on computationally intensive permutations. The algorithms are implemented in the software package microbiomeGWAS (<https://github.com/lscibb/microbiomeGWAS>). The computational complexity depends on the number of subjects, the number of SNPs, and the number of micro-biome distance matrices. It takes microbiomeGWAS about 12 h to analyze data in a GWAS with 2,000 subjects, 500,000 SNPs, and 14 microbiome distance matrices using a single core.

We applied our methods to nonmalignant lung tissue samples in the Environment And Genetics in Lung cancer Etiology (EAGLE) study [25]. We show that correcting for skewness and kurtosis eliminated dramatic deviations in the quantile-quantile (QQ) plot based on asymptotic approximations.

Methods

A Score Statistic for a Single Microbiome Distance Matrix

Suppose that we have a set of N subjects genotyped with SNP arrays. For simplicity, we consider 1 SNP with minor allele frequency (MAF) f . Let g_i represent the number of minor alleles for subject i . We assume that the 16S rRNA gene has been sequenced for all microbiome samples of a specific body site. We first consider one beta diversity distance matrix $\mathbf{D} = (d_{ij})$. If the SNP is associated with the distance matrix, then the microbiome distances tend to be smaller for pairs of subjects when they have identical genotypic values at the SNP (fig.1). Statistically speaking, we need to test whether the pairwise microbiome distance is positively associated with the genetic distance across subject pairs, which was described in our previous paper [23]. Here, we briefly summarize the score statistic for readers' convenience.

Let $G_{ij} = |g_i - g_j|$ be the genetic distance. We assume $d_{ij} = \alpha + \beta G_{ij} + \varepsilon_{ij}$ for all pairs of subjects. The scores static for testing $H_0: \beta = 0$ vs. $\beta > 0$ is delivered by minimizing $\sum_{i < j} (d_{ij} - \alpha - \beta G_{ij})^2$:

$$S = \sum_{i < j} d'_{ij} G_{ij} \text{ with } d'_{ij} = d_{ij} - \frac{2}{N(N-1)} \sum_{m < n} d_{mn}. \quad (1)$$

Under $H_0: \beta = 0$, The variance of S conditioning on $\mathbf{D}$ was derived as

$$\sigma^2 = \text{var}_0(S|\mathbf{D}) = \binom{N}{2} \text{var}(G_{ij}) \mu(\mathbf{D}) + 6 \binom{N}{3} \text{cov}(G_{ij}, G_{im}) \theta(\mathbf{D}) \quad (2)$$

by taking into account the dependence in (G_{ij}, G_{mn}) . Here,

$$\mu(\mathbf{D}) = \frac{2}{N(N-1)} \sum_{i < j} (d'_{ij})^2 \quad (3)$$

and

$$\theta(\mathbf{D}) = \frac{2}{N(N-1)(N-2)} \sum_{i < j < m} (d'_{ij} d'_{im} + d'_{ij} d'_{jm} + d'_{im} d'_{jm}). \quad (4)$$

The variance-scaled score statistic $Z = S/\sigma \sim N(0, 1)$ asymptotically. Note that the variance $\sigma^2 = \text{var}_0(S|\mathbf{D})$ is calculated conditioning on the distance matrix $\mathbf{D}$; thus, it does not depend on the assumption of the distribution for pairwise distances. This makes the null distribution ‘nonparametric’ to the microbiome pairwise distances, an important statistical property previously studied extensively for quantitative trait loci mapping studies based on human families [26].

This framework can also adjust for covariates, e.g., principal component scores derived based on genotypes. Let $\mathbf{X}_i = (x_{i1}, \dots, x_{iv})$ denote the v covariates for the i -th subject. We assume $d_{ij} = \alpha + \beta G_{ij} + \sum_{t=1}^v w_t |x_{it} - x_{jt}| + \varepsilon_{ij}$. We estimate $(\hat{w}_1, \dots, \hat{w}_v)$ under $H_0: \beta = 0$ and define residue $d'_{ij} = d_{ij} - \hat{\alpha} - \sum_{t=1}^v \hat{w}_t |x_{it} - x_{jt}|$. It is easy to verify that the score equation for β evaluated at $H_0: \beta = 0$ is $S' = \sum_{i < j} d'_{ij} G_{ij}$. Let $\mathbf{D}'$ be the residue distance matrix with $(\mathbf{D}')_{ij} = d'_{ij}$. We can similarly derive the conditional variance $\text{Var}_0(\text{and } S'|\mathbf{D}')$ and the normalized score static

$$Z' = S' / \sqrt{\text{Var}_0(S'|\mathbf{D}')}.$$

Testing Association for Multiple Microbiome Distance Matrices

We now consider K distance matrices $(\mathbf{D}_1, \dots, \mathbf{D}_K)$, where $\mathbf{D}_k$ is a $N \times N$ matrix with $(\mathbf{D}_k)_{ij} = d_{k,ij}$. We are interested in testing whether the SNP is associated with any of the K microbiome distance matrices. For each distance matrix $\mathbf{D}_k$, we denote the score equation as S_k , its variance as $\sigma_k^2 = \text{var}_0(S_k | \mathbf{D}_k)$ and the score statistic as $Z_k = S_k / \sigma_k$. The overall statistic for testing the global null hypothesis that none of the K distance matrices is associated with the SNP is defined as

$$T_0 = \max_{1 \leq k \leq K} Z_k.$$

We aim to derive an accurate approximation to the significance for a threshold b :

$$P_0\left(T_0 = \max_{1 \leq k \leq K} Z_k > b\right).$$

To proceed, we first derive the correlation matrix for $(Z_1, \dots, Z_K)$ under H_0 : $\beta_1 = \dots = \beta_K = 0$. For two distance matrices indexed by (s, t) , we derive the correlation $\rho_{st} = \text{cor}_0(Z_s, Z_t | \mathbf{D}_s, \mathbf{D}_t)$ in the Appendix:

$$\rho_{st} = \frac{\binom{N}{2} \text{var}(G_{ij}) \omega(\mathbf{D}_s, \mathbf{D}_t) + 6 \binom{N}{3} \text{cov}(G_{ij}, G_{ik}) \pi(\mathbf{D}_s, \mathbf{D}_t)}{\sqrt{\left(\binom{N}{2} \text{var}(G_{ij}) \mu(\mathbf{D}_s) + 6 \binom{N}{3} \text{cov}(G_{ij}, G_{ik}) \theta(\mathbf{D}_s) \right) \left(\binom{N}{2} \text{var}(G_{ij}) \mu(\mathbf{D}_t) + 6 \binom{N}{3} \text{cov}(G_{ij}, G_{ik}) \theta(\mathbf{D}_t) \right)}}. \quad (6)$$

Here, $\mu(\mathbf{D}_s)$ is specified in equation 3, $\theta(\mathbf{D}_s)$ is specified in equation 4, and $\omega(\mathbf{D}_s, \mathbf{D}_t)$ and $\pi(\mathbf{D}_s, \mathbf{D}_t)$ are specified in the Appendix. It is easy to verify that, as the sample size N is large,

$$\rho_{st} \approx \frac{\pi(\mathbf{D}_s, \mathbf{D}_t)}{\sqrt{\theta(\mathbf{D}_s) \theta(\mathbf{D}_t)}} = \rho_{st}^0. \quad (7)$$

Numerical calculations suggest that the approximation is very accurate when sample size $N \geq 50$. Importantly, ρ_{st}^0 depends only on the distance matrices but not the MAF of the SNP (fig. 2a). Thus, we only need to calculate ρ_{st}^0 once and use it for all SNPs in the genome.

Let

$$\Sigma = \begin{pmatrix} \rho_{11}^0 & \cdots & \rho_{1K}^0 \\ \vdots & \ddots & \vdots \\ \rho_{K1}^0 & \cdots & \rho_{KK}^0 \end{pmatrix}. \quad (8)$$

Assuming that $(Z_1, \dots, Z_K)$ follows a multivariate normal distribution with correlation matrix Σ , the significance $P_0(\max_{1 \leq k \leq K} Z_k > b)$ can be calculated using the function `pmvnorm` in the R package `mvtnorm` [27, 28].

Improve the p-Value Approximation by Correcting for Skewness and Kurtosis

In our recent work [23], we showed that the score statistic Z_t had positive skewness $\gamma_k = E_0(S_k^3 | \mathbf{D}_k) / \sigma_k^3$ and kurtosis $\kappa_k = E_0(S_k^4 | \mathbf{D}_k) / \sigma_k^4 - 3$ because of the dependence in the pairwise distances. The skewness and kurtosis make the score statistic Z_k have a long tail, which leads to a liberal p-value approximation based on the asymptotic distribution $N(0, 1)$, particularly for small p values. We developed an approximation to the tail probability $P_0(Z_k > b)$ as

$$P_0(Z_k > b) \approx \exp\left(-b\xi + \frac{1}{2}(1 + \gamma_k^2)\xi^2 + \frac{1}{6}\gamma_k\xi^3 + \frac{1}{24}\kappa_k\xi^4\right)\Phi(-\tau_k\xi), \quad (9)$$

where ξ satisfies $\xi + \gamma_k\xi^2/2 + \kappa_k\xi^3/6 = b$, $\tau_k^2 = 1 + \gamma_k\xi + \kappa_k\xi^2/2$, and $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$ [23, 26, 29, 30]. The magnitude of skewness and kurtosis depends on the sample size N , the MAF of the SNP, and the microbiome distance matrix (fig.2 b, c). Again, the skewness and kurtosis are calculated conditioning on distance matrices.

For a given SNP in one study, the skewness γ_k and kurtosis κ_k may vary substantially across microbiome distance matrices. As an empirical example, for 100 microbiome samples randomly selected from the American Gut Project (AGP), skewness and kurtosis are quite different (table 1) for Z_u (based on unweighted UniFrac distance matrix) and Z_w (based on weighted UniFrac distance matrix). According to equation 9, the p value $P(Z_u > b)$ may be very different from $P(Z_w > b)$ for a large value of b . As numeric example, $P(Z_u > 8) = 1.2 \times 10^{-9}$ and $P(Z_w > 8) = 5.6 \times 10^{-9}$.

The numerical examples in table 1 suggest that, given $(Z_1, \dots, Z_K)$ corresponding to K distance matrices, defining $T_0 = \max_{1 \leq k \leq K} Z_k$ as the overall statistic is not appropriate because the test values are not directly comparable due to the different null distributions characterized by skewness and kurtosis. Instead, we first calculate the significance p_t by correcting for the specific skewness γ_t and kurtosis κ_t using equation 9 and derive the normal quantile $q_t = \Phi(1 - p_t)$. The overall statistic is then defined as

$$T_1 = \max_{1 \leq t \leq K} q_t. \quad (10)$$

Importantly, the skewness and kurtosis correction in equation 9 has appreciable impact only on the tail of the marginal distribution of each Z_i ; thus, it has negligible impact on the correlation between two statistics, i.e., $\text{corr}(q_s, q_d) \approx \text{corr}(Z_s, Z_d)$, as is demonstrated in numerical examples in figure 2 d. Under the assumption that $(q_1, \dots, q_K)$ follows a multivariate normal distribution with correlation matrix Σ , we can use the function `pvnorm` in the R package `mvtnorm` to calculate the significance $P_0(\max_{1 \leq t \leq K} q_t > b)$.

Efficient Approximation of p Values for Genome-Wide Analysis

Calculating $P_0(\max_{1 \leq t \leq K} q_t > b)$ for millions of SNPs in GWAS using the R package `mvtnorm` is computationally intensive. To address the issue, we propose a hybrid, computationally efficient strategy, which relies on the observation that the correlation matrix defined in equation 8 is approximately constant for SNPs of different MAF. Thus, we simulate $M = 10^6$ random vectors $(z_{m1}, \dots, z_{mK})$ for $m = 1, \dots, M$ according to $N(\mathbf{0}, \Sigma)$ and define $Q_m = \max_{1 \leq t \leq K} z_{mt}$. Then $P_0(\max_{1 \leq t \leq K} q_t > b)$ is approximated by $\hat{P} = \# \{Q_m > b\} / M$. For SNPs with $p < 10^{-4}$, $\hat{P}$ is accurate with relative error

$$\frac{\sqrt{\text{var}(\hat{P})}}{\hat{P}} = \sqrt{\frac{1-P}{PM}} \leq \sqrt{\frac{1-10^{-4}}{10^{-4} \times 10^6}} \approx \frac{1}{10}.$$

For SNPs with $p < 10^{-4}$, we use `mvtnorm` to calculate $P_0(\max_{1 \leq t \leq K} q_t > b)$ to refine the approximation. The number of SNPs in this category may range from 100 to 1,000, depending on how many SNPs are tested in total and the number of truly associated SNPs. Typically, it takes a few minutes to evaluate the significance for all SNPs in the GWAS.

Simulations

We performed simulations to investigate the type-I error rate of the overall statistics T_0 (not adjusted for skewness and kurtosis) and T_1 (adjusted for skewness and kurtosis). We performed simulations using sample size $N = 100, 200, 500$, and $1,000$. The MAF of the SNP was chosen as 20 and 50%. To make simulations realistic, we used the fecal microbiome samples with the 16S rRNA V4 region sequences from the AGP. All data were downloaded from the AGP website (<https://github.com/biocore/American-Gut>). Samples were excluded from our analysis if they had less than 10,000 sequence reads or had a self-reported history of antibiotic usage within 1 month. After quality control, 1,879 subjects remained. All samples were rarefied to have 10,000 reads.

Our simulations used 14 microbiome distance matrices, including weighted and unweighted UniFrac distance matrices calculated using QIIME [2] and 12 distance matrices based on RA of taxa at different levels (species, genus, family, order, class, and phylum). As an example, let $X_i = (x_{i1}, \dots, x_{in})$ denote the RA of n taxa for subject i at the phylum level. For a subject pair (i, j) , we calculate the Euclidean distance

$$\sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}.$$

The Euclidean distance is dominated by the most abundant taxon and is not robust. Thus, we define a rank-based distance metric. For each k , let r_{ik} be the rank of x_{ik} in $\{x_{1k}, \dots, x_{Nk}\}$ across N subjects. The rank-based distance is calculated as $\sum_{k=1}^n |r_{ik} - r_{jk}|$. The 14 distance matrices are denoted as D_u (unweighted UniFrac), D_w (weighted UniFrac), D_s^E (Euclidean distance at species level), D_g^E (genus level), D_f^E (family level), D_o^E (order level), D_c^E (class level), D_p^E (phylum level), D_s^R (rank-based distance at species level), D_g^R (genus level), D_f^R (family level), D_o^R (order level), D_c^R (class level), and D_p^R (phylum level).

For each setting, the type-I error rates were evaluated based on 10^7 simulations under H_0 . The simulations were done by randomly simulating genotypes given MAF while independent of 14 distance matrices. The type-I error rates are summarized in table 2. T_0 , the overall statistic not adjusted for skewness and kurtosis, had unacceptably high type-I error rates, particularly when the level α was small or the sample size was small. On the other hand, T_1 , the overall statistic adjusted for skewness and kurtosis, had an acceptable type-I error. As a numerical example, for $\alpha = 10^{-6}$, MAF = 0.2, and sample size 100, the empirical type-I error rate of T_0 was 3.4×10^{-4} , a 340-fold inflation compared to the specified level $\alpha = 10^{-6}$. As comparison, the empirical type-I error rate of T_1 was 7.0×10^{-7} (s.e. = 3.2×10^{-7}), which is consistent with the specified level $\alpha = 10^{-6}$. As sample size increased, the type-I error rate of T_0 tended to more accurate. However, even with 1,000 samples, T_0 had a type-I error 1.8×10^{-5} , an 18-fold inflation. This set of simulations demonstrates the necessity and accuracy of the p-value approximation based on skewness and kurtosis correction.

GWAS of Microbiome Diversity in Adjacent Normal Lung Tissues

We applied our method to a set of lung cancer patient samples in the EAGLE study [25] with both genome-wide genetic data [31] and microbiome data from adjacent, noncancer tissues. The microbiome data were obtained by sequencing 16S rRNA V3-V4 region amplicons on the Illumina MiSeq sequencing platform. Details for data quality control were presented in previous paper [23]. After quality control, 147 samples with at least 1,000 high-quality sequence reads were left for genetic association analysis. We tested associations for 383,262 common, autosomal SNPs with MAF $\geq 10\%$.

Similar to the simulation study, we used 14 distance matrices for association analysis: unweighted and weighted UniFrac distance matrices, Euclidean distance matrices for six levels (species, genus, family, order, class, and phylum), and rank-based distance matrices for the same six levels. We calculated the correlation for each pair of score statistics $cor(Z_i, Z_j)$, as reported in figure 3. The six score statistics based on Euclidean distance matrices were highly correlated. The six score statistics based on rank distance matrices were also highly correlated. However, the score statistics had low correlation between Euclidean and

rank-based distance matrices. The score statistics based on UniFrac distance matrices were not highly correlated with those based on distance matrices calculated based on taxa RA.

We tested genetic associations for 383,262 SNPs using both T_0 and T_1 . The QQ plot for T_0 dramatically deviated from the global null hypothesis (fig. 4), suggesting a systematic bias when evaluating p values. The QQ plot for T_1 did not deviate from the global null hypothesis (fig. 4), which is expected given the small sample size. These results suggest that our methods for adjusting skewness and kurtosis can produce accurate p values even for studies with small sample sizes. The top SNP rs1334079 had $p < 10^{-9}$ based on T_0 , but its significance dropped to 1.7×10^{-6} based on T_1 , suggesting the huge impact of skewness and kurtosis when evaluating small p values.

We did not identify SNPs achieving genome-wide significance, primarily because of the small sample size. Table 3 presents the results for the top four SNPs (each from four genetic loci) with $p < 10^{-5}$. For each SNP, we report the score statistic values (after adjusting for skewness and kurtosis) for the 14 distance matrices. The strongest association may come from different distance matrices, indicating the importance of testing multiple distance matrices for associations. These results are based on a small study and need to be replicated in future study to establish biological significance.

Discussion

In this paper, we develop a statistical method for identifying host genetic variants associated with human microbiome composition by testing multiple microbiome distance matrices. We define the overall statistic as the strongest association across tested distance matrices, adjusted for multiple testing. The complexity of calculating statistics is $N^2 \times K \times L$, where N is the sample size, K is the number of microbiome distance matrices, and L is the number of SNPs in GWAS. It is computationally infeasible to perform permutations to derive significance for a large-scale GWAS. Thus, the primary challenge is to evaluate the significance. Our method for efficiently and accurately evaluating significance relies on two theoretical investigations. First, we corrected the score statistics for their skewness and kurtosis, and we found that the corrected score statistics follow a multivariate distribution under the null hypothesis. Second, we demonstrated that the correlation matrix of the score statistics is numerically constant for SNPs with different MAF; thus, we only need to calculate the correlation matrix once and use it for all SNPs. The accuracy of the method was verified by large-scale simulations.

We applied the method to the EAGLE lung microbiome GWAS with 147 samples. As expected, we did not identify genome-wide significant SNPs because of limited statistical power. Using this empirical example, we show that correcting for the skewness and kurtosis of score statistics, induced by the dependence of pairwise distances, is crucial for evaluating significance. The p-value approximation based on the asymptotic theory gives a liberal p value, particularly for small p values and when sample size is small. Importantly, our method eliminated the deviation in the QQ plot.

Currently, our method was developed for genotyped SNPs. We are working to extend the method to handle imputed SNPs by modelling the imputation uncertainty, which will allow convenient meta-analysis in the future. Another extension will be to test genetic interaction with microbiome composition for many distance matrices. This extension may be valuable for alterations of the microbiome not only by genetics but also by environmental factors such as smoking, diets, and medications. Finally, we point out that we have used the strongest association $T_1 = \max_{1 \leq t \leq K} q_t$ in equation 10 as the overall statistic, as was also used in MiRKAT [24]. This statistic is optimal only when associations are driven by one single matrix. Where multiple distance matrices are independently associated with a SNP, many other statistics may be used to achieve a robust statistical power when the true alternative hypothesis is unknown. These statistics include weighted sum statistic, rank truncated product method (RTP) [32], adaptive RTP (ARTP) [33, 34], and truncated product method (TPM) [32]. However, computationally expensive permutations are required to evaluate significance for these more complex statistics.

Acknowledgements

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>). The project was supported by the NIH Intramural Research program. The authors thank Dr. Mitchell H. Gail from the National Cancer Institute for valuable suggestions to the project.

Appendix

We calculate $\rho_{12} = \text{cor}_0(Z_1, Z_2 | D_1, D_2)$ under H_0 , where D_k ($k = 1, 2$) is a given distance matrix, Z_k is the score statistic corresponding to D_k . Because

$$Z_k = S_k / \sqrt{\text{var}_0(S_k | D_k)},$$

we have

$$\rho_{12} = \frac{\text{cov}_0(S_1, S_2 | D_1, D_2)}{\sqrt{\text{var}_0(S_1 | D_1) \text{var}_0(S_2 | D_2)}}. \quad (\text{A1})$$

Let

$$d'_{k,ij} = d_{k,ij} - \frac{2}{N(N-1)} \sum_{a < b} d_{k,ab}$$

be the centered distance, then $S_k = \sum_{i < j} d'_{k,ij} G_{ij}$. We have

$$\text{cov}_0(S_1, S_2 | D_1, D_2) = \sum_{i < j, m < n} d'_{1,ij} d'_{2,mn} \text{cov}(G_{ij}, G_{mn}). \quad \text{A2}$$

When (i, j, m, n) are distinct, $\text{cov}(G_{ij}, G_{mn}) = 0$ because $G_{ij} = |g_i - g_j|$ and $G_{mn} = |g_m - g_n|$ are independent. Based on this observation, one can verify that

$$\begin{aligned} \text{cov}_0(S_1, S_2 | D_1, D_2) &= \binom{N}{2} \text{var}(G_{ij}) \omega_{12}(D_1, D_2) \quad A3 \\ &+ 6 \binom{N}{3} \text{cov}(G_{ij}, G_{im}) \pi_{12}(D_1, D_2). \end{aligned}$$

Here,

$$\omega_{12}(D_1, D_2) = \binom{N}{2}^{-1} \sum_{i < j} d'_{1,ij} d'_{2,ij} \quad A4$$

and

$$\begin{aligned} \pi_{12}(D_1, D_2) &= \binom{N}{3}^{-1} \frac{1}{6} \sum_{i < j < k} (d'_{1,ij} d'_{2,ik} + d'_{1,ij} d'_{2,jk} + d'_{1,ik} d'_{2,jk} + d'_{1,ik} d'_{2,ij} + d'_{1,jk} d'_{2,ij} \\ &+ d'_{1,jk} d'_{2,ik}). \quad A5 \end{aligned}$$

Next, we calculate $\text{var}(G_{ij})$ and $\text{cov}(G_{ij}, G_{ik})$. Let $f_i = P(g_i = t)$ with $f_0 + f_1 + f_2 = 1$. Tedious but straightforward calculations lead to the following results:

$$\left\{ \begin{aligned} E(G_{ij}) &= \sum_{i,j \in \{0,1,2\}} f_i f_j |i - j| = 2f_0 f_1 + 2f_1 f_2 + 4f_0 f_2 \\ \text{var}(G_{ij}) &= E(G_{ij}^2) - E(G_{ij})^2 = \sum_{i,j \in \{0,1,2\}} f_i f_j |i - j|^2 - E(G_{ij})^2 \\ &= (2f_0 f_1 + 2f_1 f_2 + 8f_0 f_2) - (2f_0 f_1 + 2f_1 f_2 + 4f_0 f_2)^2 \quad A6 \\ \text{cov}(G_{ij}, G_{ik}) &= E(G_{ij} G_{ik}) - E(G_{ij}) E(G_{ik}) \\ &= \sum_{i,j,k \in \{0,1,2\}} f_i f_j f_k |i - j| |i - k| - E(G_{ij}) E(G_{ik}) \\ &= f_1(1 - f_1) + 4f_0 f_2(1 + f_1) - (2f_0 f_1 + 2f_1 f_2 + 4f_0 f_2)^2 \end{aligned} \right.$$

Combining A1–A6 proves equation 6 in the main text.

References

1. Metzker ML: Sequencing technologies – the next generation. *Nat Rev Genet* 2010; 11: 31–46. [PubMed: 19997069]
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA,

- Widmann J, Yatsunenko T, Zaneveld J, Knight R: QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; 7: 335–336. [PubMed: 20383131]
3. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R: UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; 27: 2194–2200. [PubMed: 21700674]
 4. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG: Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 2013; 10: 57–59. [PubMed: 23202435]
 5. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI: A core gut microbiome in obese and lean twins. *Nature* 2009; 457: 480–484. [PubMed: 19043404]
 6. Flier JS, Mekalanos JJ: Gut check: testing a role for the intestinal microbiome in human obesity. *Sci Transl Med* 2009; 1: 6ps7.
 7. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L: Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* 2013; 105: 1907–1911. [PubMed: 24316595]
 8. Goedert JJ, Hua X, Yu G, Shi J: Diversity and composition of the adult fecal microbiome associated with history of cesarean birth or appendectomy: analysis of the American Gut Project. *EBioMedicine* 2014; 1: 167–172. [PubMed: 25601913]
 9. Goedert JJ, Jones G, Hua X, Xu X, Yu G, Flores R, Falk RT, Gail MH, Shi J, Ravel J, Feigelson SH: Investigation of the association between the fecal microbiota and breast cancer in post-menopausal women: a population-based case-control pilot study. *J Natl Cancer Inst* 2015; 1: 107: djv147.
 10. Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beaugerie L, Cosnes J, Corthier G, Marteau P, Dore J: Low counts of *Faecalibacterium prausnitzii* in colitis microbiota. *Inflamm Bowel Dis* 2009; 15: 1183–1189. [PubMed: 19235886]
 11. Farrell JJ, Zhang L, Zhou H, Chia D, Elashoff D, Akin D, Paster BJ, Joshupura K, Wong DT: Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut* 2012; 61: 582–588. [PubMed: 21994333]
 12. Ahn J, Chen CY, Hayes RB: Oral microbiome and oral and gastrointestinal cancer risk. *Cancer Causes Control* 2012; 23: 399–404. [PubMed: 22271008]
 13. McKnite AM, Perez-Munoz ME, Lu L, Williams EG, Brewer S, Andreux PA, Bastiaansen JWM, Wang XS, Kachman SD, Auwerx J, Williams RW, Benson AK, Peterson DA, Ciobanu DC: Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* 2012; 7: e39191. [PubMed: 22723961]
 14. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, Zhang M, Oh PL, Nehrenberg D, Hua K, Kachman SD, Moriyama EN, Walter J, Peterson DA, Pomp D: Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci USA* 2010; 107: 18933–18938. [PubMed: 20937875]
 15. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG: Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 2015; 16: 191. [PubMed: 26374288]
 16. Knights D, Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, van Sommeren S, Imhann F, Stempak JM, Huang H, Vangay P, Al-Ghalith GA, Russell C, Sauk J, Knight J, Daly MJ, Huttenhower C, Xavier RJ: Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med* 2014; 6: 107. [PubMed: 25587358]
 17. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE: Human genetics shape the gut microbiome. *Cell* 2014; 159: 789–799. [PubMed: 25417156]
 18. Bray JR, Curtis T: An ordination of upland forest communities of southern Wisconsin. *Ecol Monogr* 1957; 27: 325–349.
 19. Lozupone C, Knight R: UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005; 71: 8228–8235. [PubMed: 16332807]
 20. Lozupone CA, Hamady M, Kelley ST, Knight R: Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 2007; 73: 1576–1585. [PubMed: 17220268]

21. Lozupone C, Hamady M, Knight R: UniFrac – an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 2006; 7: 371. [PubMed: 16893466]
22. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H: Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 2012; 28: 2106–2113. [PubMed: 22711789]
23. Hua X, Song L, Yu G, Goedert J, Abnet C, Landi M, Shi J: MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition. *BioRxiv*, 2015, DOI: 10.1101/031187.
24. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC: Testing in microbiome-profiling studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *Am J Hum Genet* 2015; 96: 797–807. [PubMed: 25957468]
25. Landi MT, Consonni D, Rotunno M, Bergen AW, Goldstein AM, Lubin JH, Goldin L, Alavanja M, Morgan G, Subar AF, Linnoila I, Previdi F, Corno M, Rubagotti M, Marinelli B, Albetti B, Colombi A, Tucker M, Wacholder S, Pesatori AC, Caporaso NE, Bertazzi PA: Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* 2008; 8: 203. [PubMed: 18538025]
26. Tang HK, Siegmund D: Mapping quantitative trait loci in oligogenic models. *Biostatistics* 2001; 2: 147–162. [PubMed: 12933546]
27. Genz A: Numerical computation of multivariate normal probabilities. *J Comput Graph Stat* 1992; 1: 10.
28. Genz A: Comparison of methods for the computation of multivariate normal probabilities. *Comput Sci Stat* 1993; 25: 400–405.
29. Tu IP, Siegmund D: The maximum of a function of a Markov chain and application to linkage analysis. *Adv Appl Probab* 1999; 31: 510–531.
30. Siegmund D: *Sequential Analysis: Tests and Confidence Intervals*. New York, Springer, 1985.
31. Landi MT, Chatterjee N, Yu K, et al.: A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* 2009; 85: 679–691. [PubMed: 19836008]
32. Zaykin DV, Zhivotovsky LA, Czika W, Shao S, Wolfinger RD: Combining p values in large-scale genomics experiments. *Pharm Stat* 2007; 6: 217–226. [PubMed: 17879330]
33. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N: Pathway analysis by adaptive combination of P-values. *Genet Epidemiol* 2009; 33: 700–709. [PubMed: 19333968]
34. Dudbridge F, Koeleman BP: Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet* 2004; 75: 424–435. [PubMed: 15266393]

Fig. 1.

If a SNP is associated with microbiome beta diversity, the pairwise microbiome distances are positively correlated with genetic distances across all pairs of subjects. Here, $g_i \in \{0, 1, 2\}$ is the genotypic value for subject i and $|g_i - g_j|$ is the genotypic distance for a subject pair (i, j) . The plot is based on all $N(N-1)/2$ pairs of subjects.

Fig. 2.

a Correlation between score statistics based on two distance matrices is nearly constant for SNPs with different MAF. Correlations were calculated based on the samples randomly selected from the American Gut Project. Here, Z_u and Z_w are score statistics using unweighted and weighted UniFrac distance matrices. Z_p^E is the score statistic using the Euclidean distance matrix based on the RA of the taxa in the phylum level. **b** Skewness of the score statistic depends on the MAF of SNP, sample size, and distance matrix. 'Weighted', the score statistic using weighted UniFrac distance matrix; 'unweighted', the score statistic using unweighted UniFrac distance matrix. **c** Kurtosis of the score statistic depends on the MAF of SNP, sample size, and distance matrix. **d** The $x(y)$ coordinate is the correlation between score statistics before (after) correcting for skewness and kurtosis.

Because we considered score statistics for 14 distance matrices, we have $14 \times 13/2 = 91$ correlation values. The figure shows that $\text{cor}(Z_s, Z_l) \approx \text{cor}(q_s, q_l)$, i.e., adjusting for skewness and kurtosis does not numerically change correlation of score statistics.

Fig. 3.

Correlations between score statistics based on different distance matrices calculated in the EAGLE microbiome data. Here, Z_p^E and Z_p^R represent the score statistics using the Euclidean distance matrix and the rank-based distance matrix for taxa at the phylum level, respectively. 'c' represents the class level; 'o' represents the order level; 'f' represents the family level, 'g' represents the genus level, and 's' represents the species level. Z_u and Z_w represent the score statistics using unweighted and weighted UniFrac distance matrices, respectively.

Fig. 4. QQ plot for $-\log_{10}(p)$ for 383,262 common SNPs in the EAGLE lung tissue microbiome GWAS. The red QQ plot is for T_0 , the statistic not adjusted for skewness and kurtosis. The blue QQ plot is for T_1 , the statistic adjusted for skewness and kurtosis.

Table 1.

Skewness, kurtosis, and p values are different for score statistics based on different distance matrices

	Z_u^a	Z_w^b
Skewness γ	0.295	0.453
Kurtosis κ	0.333	0.415
$P(Z > 5)$	2.5×10^{-5}	4.8×10^{-5}
$P(Z > 6)$	1.2×10^{-6}	2.9×10^{-6}
$P(Z > 7)$	4.3×10^{-8}	1.4×10^{-7}
$P(Z > 8)$	1.2×10^{-9}	5.6×10^{-9}

The calculation was based on 100 samples randomly selected from the American Gut Project.

^aScore statistic based on unweighted UniFrac distance matrix.

^bScore statistic based on weighted UniFrac distance matrix.

Table 2.
Type-I error rate estimates for 2 statistics using multiple distance matrices, 10^7 simulations

	Sample size, n	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-6}$
<i>SNP MAF = 0.2</i>							
T_0 (asymptotic approximation)	100	0.093	0.0347	0.0097	0.00295	9.5E-04	3.4E-04
	200	0.080	0.0264	0.0061	0.00155	4.2E-04	1.2E-04
	500	0.069	0.0198	0.0037	0.00076	1.6E-04	3.6E-05
	1,000	0.065	0.0173	0.0029	0.00051	9.5E-05	1.8E-05
T_1 (adjusted for skewness and kurtosis)	100	0.046	0.0095	0.0010	0.00011	9.7E-06	7.0E-07
	200	0.046	0.0095	0.0010	0.00011	1.2E-05	1.6E-06
	500	0.047	0.0096	0.0010	0.00010	9.3E-06	8.0E-07
	1,000	0.048	0.0096	0.0010	0.00010	9.2E-06	9.0E-07
<i>SNP MAF = 0.5</i>							
T_0 (asymptotic approximation)	100	0.060	0.015	0.0022	0.00034	5.6E-05	1.1E-05
	200	0.057	0.013	0.0017	0.00025	3.6E-05	4.1E-06
	500	0.054	0.012	0.0014	0.00017	1.8E-05	2.0E-06
	1,000	0.053	0.011	0.0012	0.00014	1.6E-05	2.0E-06
T_1 (adjusted for skewness and kurtosis)	100	0.048	0.010	0.0010	0.00010	1.3E-05	1.0E-06
	200	0.049	0.010	0.0010	0.00011	1.1E-05	9.0E-07
	500	0.049	0.010	0.0010	0.00010	8.6E-06	1.0E-06
	1,000	0.050	0.010	0.0010	0.00010	9.9E-06	7.0E-07

Table 3.

The top 4 SNPs with $p < 10^{-5}$ identified in the GWAS of lung tissue microbiome data in the EAGLE study

	SNP			
	rs1334079	rs2291155	rs1572541	rs12574348
MAF	0.28	0.29	0.31	0.14
Euclidean distance				
q_p^E	1.17	0.20	4.74	2.76
q_c^E	5.03	-1.66	3.17	0.48
q_o^E	4.47	-1.84	2.65	0.21
q_f^E	4.24	-1.55	2.02	-0.04
q_g^E	4.42	-1.52	1.91	-0.22
q_s^E	4.47	-1.44	1.91	-0.17
Rank-based distance				
q_p^R	0.82	2.49	1.28	4.70
q_c^R	-0.17	3.46	0.60	4.36
q_o^R	-0.22	4.50	0.24	4.31
q_f^R	-0.94	4.90	0.69	4.10
q_g^R	-1.16	4.69	0.41	3.61
q_s^R	-0.91	4.42	0.55	3.57
UniFrac				
q_u	3.27	0.84	2.18	3.86
q_w	2.24	-0.23	3.53	2.80
$\max q_k$	5.03	4.90	4.74	4.70
p value	1.7E-06	3.1E-06	6.7E-06	8.0E-06

For each SNP, we report the Z-score statistic values for 14 different microbiome distance matrices, the maximum Z-score and the p value for the SNP. The Z-score statistic values reported have been adjusted for skewness and kurtosis. p, phylum; c, class; o, order; f, family; g, genus; s, species.