MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition

Xing Hua, Lei Song, Guoqin Yu, James J. Goedert,
Christian C. Abnet, Maria Teresa Landi, Jianxin Shi

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health,
Bethesda, Maryland

Correspondence to:

Jianxin Shi, PhD
Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH
9609 Medical Center Drive, 7E596, Bethesda, MD 20892
Tel: (240) 276-7419
Email: Jianxin.Shi@nih.gov

Abstract

The microbiome is the collection of all microbial genes and can be investigated by sequencing highly variable regions of 16S ribosomal RNA (rRNA) genes. Evidence suggests that environmental factors and host genetics may interact to impact human microbiome composition. Identifying host genetic variants associated with human microbiome composition not only provides clues for characterizing microbiome variation but also helps to elucidate biological mechanisms of genetic associations, prioritize genetic variants, and improve genetic risk prediction. Since a microbiota functions as a community, it is best characterized by beta diversity, that is, a pairwise distance matrix. We develop a statistical framework and a computationally efficient software package, microbiomeGWAS, for identifying host genetic variants associated with microbiome beta diversity with or without interacting with an environmental factor. We show that score statistics have positive skewness and kurtosis due to the dependent nature of the pairwise data, which makes P-value approximations based on asymptotic distributions unacceptably liberal. By correcting for skewness and kurtosis, we develop accurate *P*-value approximations, whose accuracy was verified by extensive simulations. We exemplify our methods by analyzing a set of 147 genotyped subjects with 16S rRNA microbiome profiles from non-malignant lung tissues. Correcting for skewness and kurtosis eliminated the dramatic deviation in the quantile-quantile plots. We provided preliminary evidence that six established lung cancer risk SNPs were collectively associated with microbiome composition for both unweighted (P=0.0032) and weighted (P=0.011) UniFrac distance matrices. In summary, our methods will facilitate analyzing large-scale genome-wide association studies of the human microbiome.

Keywords: microbiome, genome-wide association study, gene-environment interaction, host genetics, tail probabilities, skewness and kurtosis

Introduction

The human body is colonized by bacteria, viruses and other microbes that exceed the number of human cells by at least 10-fold and that exceed the number of human genes by at least 100-fold. The relationship between a person and his or her microbial population, termed the microbiota, is generally mutualistic. The microbiota may promote human health by inhibiting infection by pathogens, conditioning the immune system, synthesizing and digesting nutrients, and maintaining overall homeostasis. The microbiome, which is the collection of all microbial genes, can be investigated through massively parallel, next-generation DNA sequencing technologies. By amplifying and sequencing highly variable regions of 16S ribosomal RNA genes that are present in all eubacteria, cost-effective and informative microbiome profiles down to the genus level are obtained.

The human microbiome has been associated with diseases, including obesity[1], inflammatory bowel disease (IBD)[2], colorectal cancer[3] and breast cancer[4]. Thus, identifying factors that have a sustained impact on the microbiome is fundamental for elucidating its role in health conditions and for developing treatment strategies. Increasing evidence suggests that microbiome composition at a specific site of the human body is impacted by environmental factors[5,6], host genetics[7,8], and possibly by their interactions. In the mouse, quantitative trait loci (QTL) studies have identified loci contributing to the variation of the gut microbiome using linkage analysis[9,10]. Recently, Goodrich et al.[11] systematically investigated the heritability of the human gut microbiome by comparing monozygotic twins to dizygotic twins and found substantial heritability in different microbiome metrics, suggesting the important role of host genetics on gut microbiome diversity. Associations between individual host genetic variants and microbiome taxa abundances have also begun to emerge in other human samples[7,8,12]. These studies suggest that genome-wide association studies (GWAS) have great potential to identify host genetic variants associated with microbiome diversity.

GWAS of complex human diseases have identified many risk SNPs; however, the biological mechanisms are largely unknown for the majority of the risk SNPs. QTL studies of intermediate traits, e.g., gene expression[13,14], DNA methylation[15,16], chromatin structure[17,18], and metabolite production[19,20], have provided useful insights on biological mechanisms of the GWAS findings. The human microbiome at a specific body site is another important and informative intermediate trait for interpreting GWAS signals. Knights et al.[8] reported that a risk SNP for IBD located in *NOD2* was associated with the relative abundance of *Enterobacteriaceae* in the human gut microbiome. Tong et al.[7] show that a loss-of-function allele in *FUT2* that increases the risk of developing Crohn's Disease (CD) may modulate energy metabolism of the gut microbiome. In both examples, the microbiome is a potential intermediate for explaining the association between risk SNPs and disease risks, although a formal mediation analysis is required based on samples with genotype, microbiome, and disease status data. Moreover, identifying microbiome-associated host genetic variants has the potential to prioritize SNPs for discovery and to improve the performance of polygenetic risk prediction.

Three types of microbiome metrics can be derived as phenotypes for GWAS analysis. First, for each taxon at a specified taxonomic level (phylum, class, order, family, genus, and species), we calculate the relative abundance (RA) of the taxon as the ratio of the number of sequencing reads assigned to the taxon to the total number of sequencing reads. In 16S ribosomal RNA sequence profiles, approximately 100-200 taxa with average RAs ≥0.1% (from the phylum level to the genus level) across samples are abundant enough for QTL analysis. One can perform a Poisson regression to examine the association between RA of each taxon and each SNP. Significant associations are identified using Bonferroni correction ($P<5\times10^{-8}/200=2.5\times10^{-10}$) or by controlling FDR at an appropriate level. Second, multiple alpha-diversity metrics[21] can be calculated to reflect the richness (e.g., number of unique taxa) and evenness of each microbiome community after a procedure called rarefication, that eliminates the dependence between estimated alpha diversity and the variable total number of sequencing reads across subjects. Once alpha-diversity metrics are derived, one may perform standard GWAS with alpha diversity as the phenotype using linear regression.

Because a microbiota functions as a community, the most important analysis for a microbiome GWAS may be by assessing the complete structure of the community by using a pairwise microbiome distance matrix (or beta-diversity) of the microbial community. Microbiome distances can be defined in different ways, based on using phylogenetic tree information or each taxon's abundance information. Bray–Curtis dissimilarity[22] quantifies the difference between two microbiome communities using the abundance information of specific taxa. UniFrac[23-25] is another widely used distance metric. Unlike the Bray–Curtis dissimilarity metric, UniFrac compares microbiome communities by using information on the relative relatedness of each taxon, specifically by phylogenetic distance (branch lengths on a phylogenetic tree). UniFrac has two variants: the weighted UniFrac[24] that accounts for the taxa abundance information, and the unweighted UniFrac[23] that only models the information of presence or absence. Recently, a generalized UniFrac distance metric[26] was developed to automatically appreciate the advantages of weighted and unweighted UniFrac metrics and was shown to provide better statistical power to detect associations between human health conditions and microbiome communities. GWAS based on a microbiome distance matrix aims to identify host SNPs associated with microbiome composition. Intuitively, the microbiome distances tend to be smaller for pairs of subjects with similar genotypic values at the associated SNP. In addition, it is also of great interest to identify host SNPs that interact with an environment factor to affect microbiome composition. Importantly, beta diversity is temporally more stable compared with RA of taxa and alpha-diversity metrics based on the data from the Human Microbiome Project[27] (data not shown), suggesting smaller power loss for a GWAS due to temporal variability. To our knowledge, no statistical methods or software packages have been designed to efficiently analyze microbiome GWAS data using distance matrices as phenotypes.

In this paper, we develop a statistical framework and a computationally efficient package, microbiomeGWAS, for analyzing microbiome GWAS data. Our package allows the detection of host SNPs with a main effect or interaction with an environment factor, i.e. host SNPs interacting with an environment factor to affect the microbiome composition. We calculate the variance of the score statistics by appropriately considering the dependence of the pairwise distances. Importantly, we show that the score statistics have positive skewness and kurtosis due to the dependence in pairwise distances, which makes the approximation of small P-values based on the asymptotic distribution too liberal, which easily yields false positive associations. Resampling methods, e.g. bootstrap or permutation, are computationally prohibitive for accurately approximating small P-values. We propose to improve the tail probability approximation by correcting for skewness and kurtosis of the score statistics. Numerical investigations demonstrate that our method provides a very accurate approximation even for P=10^{-7}. MicrobiomeGWAS runs very efficiently, taking 36 minutes for analyzing main effects and 69 minutes for analyzing both main and interaction effects for a study with 2000 subjects and 500,000 SNPs using a single core. MicrobiomeGWAS can be freely downloaded at https://github.com/lsncibb/microbiomeGWAS.

We illustrate our methods by applying microbiomeGWAS to non-malignant lung tissue samples ($N = 147$) in the Environment And Genetics in Lung cancer Etiology (EAGLE) study[28,29]. Because smoking may alter microbiome composition, we tested both main effect and gene-smoking interaction effect. When P-values were calculated based on asymptotic distributions, the quantile-quantile (QQ) plots strongly deviated from the uniform distribution. Also, nine loci achieved genome-wide significance based on asymptotic approximations. Correcting for skewness and kurtosis eliminated the inflation and also the genome-wide significance of these loci. However, we provide evidence that the established lung cancer risk SNPs are associated with lung microbiome composition.

Material and Methods

A score statistic for testing main effect

Suppose that we have a set of N subjects genotyped with SNP arrays. For notational simplicity, we consider only one SNP with minor allele frequency (MAF) denoted as f. Our interest centers on testing whether the genotype of the SNP is associated with microbiome composition. Let $g_n = 0,1,2$ represent the number of the minor alleles for the n^{th} subject. We assume that the 16S rRNA gene of microbiota from a target site (e.g., gut) has been sequenced for these samples. Let d_{ij} be the microbiome distance between the i^{th} and j^{th} subjects and $\boldsymbol{D}$ be the distance matrix.

Intuitively, if the SNP is associated with the microbiome composition, the microbiome distances tend to be smaller for subject pairs with similar genotypic values, as is illustrated in Figure 1. For N subjects, $N(N-1)/2$ pairs can be divided to three groups with genetic distance 0, 1 and 2. For example, a pair of subjects with genotype (AA, AA) or (BB, BB) has genetic distance 0; a pair of subjects with genotype (AA, BB) or (BB, AA) has genetic distance 2; all other pairs have genetic distance 1. Apparently, we expect the microbiome distance to be positively correlated with genetic distance for subject pairs.

We define $G_{ij} = |g_i - g_j|$ as the genetic distance for a pair of subjects (i, j). We assume $d_{ij} = \alpha + \beta_M G_{ij} + \varepsilon_{ij}$ for all pairs of subjects. The score statistic for testing $H_0: \beta_M = 0$ (main effect) vs. $\beta_M > 0$ is derived by maximizing $\sum_{i<j}\left(d_{ij} - \alpha - \beta_M G_{ij}\right)^2$:

$$S_M = \sum_{i<j} d'_{ij} G_{ij} \quad \text{with} \quad d'_{ij} = d_{ij} - \frac{1}{N(N-1)/2}\sum_{k<l} d_{kl}. \tag{1}$$

The variance $Var_0(S_M|\boldsymbol{D})$ under $H_0: \beta_M = 0$ is calculated by considering the dependence in $\left(G_{ij}, G_{kl}\right)$ and conditioning on the distance matrix $\boldsymbol{D}$. Briefly, we have $Var_0(S_M|\boldsymbol{D}) = \sum_{i<j,k<l} d'_{ij} d'_{kl} Cov\left(G_{ij}, G_{kl}\right)$. When (i, j, k, l) are distinct, G_{ij} and G_{kl} are independent, i.e. $Cov\left(G_{ij}, G_{kl}\right) = 0$. Some algebra leads to

$$Var_0(S_M|\boldsymbol{D}) = \frac{N(N-1)}{2} Var\left(G_{ij}\right)\mu_2 + N(N-1)(N-2)Cov\left(G_{ij}, G_{ik}\right)\mu_3 \tag{2}$$

where

$$\mu_2 = \frac{2}{N(N-1)}\sum_{i<j}\left(d'_{ij}\right)^2 \tag{3}$$

and

$$\mu_3 = \frac{2}{N(N-1)(N-2)}\sum_{i<j<k}\left(d'_{ij}d'_{ik} + d'_{ij}d'_{jk} + d'_{ik}d'_{jk}\right). \tag{4}$$

The details for calculating $Var(G_{ij})$ and $Cov(G_{ij}, G_{ik})$ are in **Appendix A**. The variance-normalized score statistic $Z_M = S_M/\sqrt{Var_0(S_M|\boldsymbol{D})} \sim N(0,1)$ under H_0 asymptotically.

In analyses of real data, we typically have to adjust for covariates, including demographic variables and principal component analysis (PCA) scores derived based on genotypes to eliminate potential population stratification. Let $X_i = (x_{i1}, \cdots, x_{iv})$ denote the v covariates for the i^{th} subject. We assume $d_{ij} = \alpha + \beta_M G_{ij} + \sum_{t=1}^{v} w_t |x_{it} - x_{jt}| + \varepsilon_{ij}$. Define $d'_{ij} = d_{ij} - \hat{\alpha} - \sum_{t=1}^{v} \widehat{w}_t |x_{it} - x_{jt}|$ with $(\widehat{w}_1, \cdots, \widehat{w}_v)$ being estimated under H_0: $\beta_M = 0$. It is straightforward to verify that the score equation for β_M evaluated at H_0: $\beta_M = 0$ is $S'_M = \sum_{i<j} d'_{ij} G_{ij}$. We can similarly derive the variance $Var_0(S'_M|\boldsymbol{D}')$ and the normalized score statistic $Z'_M = S'_M/\sqrt{Var_0(S'_M|\boldsymbol{D}')}$. Here, $\boldsymbol{D}'$ denotes the residue distance matrix with $(\boldsymbol{D}')_{ij} = d'_{ij}$.

A score statistic for testing gene-environment interaction

Let E_i denote an environmental variable. Define $\Delta_{ij} = |g_i E_i - g_j E_j|$. We extend the statistical framework to detect the SNP-environment interaction by assuming $d_{ij} = \alpha + \beta_M G_{ij} + \beta_E |E_i - E_j| + \beta_I \Delta_{ij} + \varepsilon_{ij}$, where β_M denotes the main genetic effect, β_I denote the additive gene-environment effect and β_E denotes the main effect of the environmental factor. We consider testing the null hypothesis that the SNP is not associated with microbiome composition either directly or by interacting with E, $i.e.$ H_0: $\beta_M = \beta_I = 0$. The alternative hypothesis is H_1: $\beta_M > 0$ or $\beta_I > 0$.

Again, we estimate β_E and α under H_0 and calculate $d'_{ij} = d_{ij} - \hat{\alpha} - \hat{\beta}_E |E_i - E_j|$. The score equations evaluated under H_0 are $S_M = \sum_{i<j} d'_{ij} G_{ij}$ for β_M and $S_I = \sum_{i<j} d'_{ij} \Delta_{ij}$ for β_I. Similar to (2), we derive the variance $Var_0(S_I|\boldsymbol{D}')$ by accounting for the dependence in $(\Delta_{ij}, \Delta_{kl})$:

$$Var_0(S_I|\boldsymbol{D}') = \frac{N(N-1)}{2} Var(\Delta_{ij})\mu_2 + N(N-1)(N-2) Cov(\Delta_{ij}, \Delta_{ik})\mu_3. \quad (5)$$

Let $Z_M = S_M/\sqrt{Var_0(S_M|\boldsymbol{D}')}$ and $Z_I = S_I/\sqrt{Var_0(S_I|\boldsymbol{D}')}$. Asymptotically, $Z_M \sim N(0,1)$ and $Z_I \sim N(0,1)$ under H_0.

In **Appendix B**, we derive

$$Cov_0(S_M, S_I|\boldsymbol{D}') = \frac{N(N-1)}{2} Cov\left(G_{ij}, \Delta_{ij}\right)\mu_2 + N(N-1)(N-2)Cov\left(\mathrm{G}_{ij}, \Delta_{ik}\right)\mu_3. \tag{6}$$

The correlation $\rho = \ Cor_0(Z_M, Z_I|\boldsymbol{D}')$ is calculated as $\rho = Cov_0(S_M, S_I|\boldsymbol{D}')/\sqrt{Var_0(S_M|\boldsymbol{D}')Var_0(S_I|\boldsymbol{D}')}$.

Asymptotically, (Z_M, Z_I) follows a bivariate normal distribution with a correlation matrix $\boldsymbol{\Omega} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

In **Appendix C**, we derive a statistic for jointly testing $H_0\colon \beta_M = \beta_I = 0$ vs. $H_1\colon \beta_M > 0$ or $\beta_I > 0$. Briefly, the 2D plane is partitioned to four parts (**Figure 2**). The joint statistic is derived as

$$Q = \begin{cases} (Z_M, Z_I)\boldsymbol{\Omega}^{-1}(Z_M, Z_I)^T & (Z_M, Z_I) \in A_1 \\ (w_1 Z_M + w_2 Z_I)^2 & (Z_M, Z_I) \in A_2 \\ (w_2 Z_M + w_1 Z_I)^2 & (Z_M, Z_I) \in A_3 \\ 0 & (Z_M, Z_I) \in A_4 \end{cases} \tag{7}$$

where $w_1 = (\theta - 1/\theta\,)/2,\ w_2 = (\theta + 1/\theta\,)/2$ and $\theta = \sqrt{(1-\rho)/(1+\rho)}$. The asymptotic P-value is calculated as

$$P(Q > b^2) = q_1 P(\chi_2^2 > b^2) + q_2 P(N(0,1) > b) + q_3 P(N(0,1) > b), \tag{8}$$

where $q_i = P\left((Z_M, Z_I) \in A_i\right)$.

Improved *P*-value approximations by correcting for skewness and kurtosis

Theoretic investigation suggests that the score statistics Z_M and Z_I have a positive skewness, which makes the tail probability approximations based on the asymptotic distribution $N(0,1)$ unacceptably liberal (**Figures 3A and 3B**). In a numeric example with skewness $\gamma = 0.2$, $P(Z > 5)$ =2.9×10^{-7} based on $N(0,1)$, which is approximately two orders of magnitude more significant than P=3.9×10^{-5} based on 10^8 permutations. The significance inflation becomes worse for smaller *P*-values and larger skewness γ. Similar but more tedious calculations suggest that both statistics have positive kurtosis, making the approximation based on $N(0,1)$ even worse. One possible solution is to approximate tail probabilities using permutations or bootstrap. However, these resampling methods are computationally prohibitive for testing millions of common SNPs in a large-scale study.

To address this problem, we calculated the skewness γ and kurtosis κ of the score statistics under H_0 (**Appendix D**). We propose to improve the tail probability approximation $P_0(Z > b)$ by correcting for the skewness and kurtosis, following the skewness correction in linkage analysis[30,31]. Technical details are provided in **Appendix E**. Correcting for both skewness and kurtosis leads to an approximation

$$P_0(Z > b) \approx e^{-b\xi_1+(1+\sigma_1^2)\xi_1^2/2+\gamma\xi_1^3/6+\kappa\xi_1^4/24}\,\Phi(-\sigma_1\xi_1), \qquad (9)$$

where ξ_1 satisfies $\xi + \gamma\xi^2/2 \; + \kappa\xi^3/6 = b$, $\sigma_1^2 = 1 + \gamma\xi_1 + \kappa\xi_1^2/2$ and $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Correcting for skewness but ignoring kurtosis (i.e., assuming $\kappa = 0$) leads to an approximation

$$P_0(Z > b) \approx e^{-b\xi_2+(1+\sigma_2^2)\xi_2^2/2+\gamma\xi_2^3/6}\,\Phi(-\sigma_2\xi_2), \qquad (10)$$

where $\xi_2 = (\sqrt{1+2\gamma b} - 1)/\gamma$, $\sigma_2^2 = 1 + \gamma\xi_2$. Numerical results presented in **Figure 3B** demonstrate that (9) works very well.

Given the distance matrix $\boldsymbol{D}$, $\gamma_M \propto 1/N^{1/2}$, $\gamma_I \propto 1/N^{1/2}$, $\kappa_M \propto 1/N$ and $\kappa_I \propto 1/N$ (**Appendix D**). Thus, skewness decays much more slowly with sample size N than kurtosis (**Figures 3C and 3D**). Thus, even for a large study with thousands of samples, correcting for skewness is necessary for accurately evaluating tail probabilities. Importantly, both skewness and kurtosis highly depend on the MAF, suggesting that the impact of skewness and kurtosis is different across SNPs with different MAF. Numerical studies (**Figures 3C and 3D**) show that skewness and kurtosis are minimized when MAF=0.5 and maximized when MAF≈0.2-0.3.

Finally, we discuss how to approximate the tail probability of Q in (7) for testing $H_0: \beta_M = \beta_I = 0$ by correcting for non-normality in Z_M and Z_I. When $(Z_M, Z_I) \in A_2$ (or A_3), we calculate the skewness

$E(w_1Z_M + w_2Z_I)^3$ and the kurtosis $E(w_1Z_M + w_2Z_I)^4 - 3$ and use (9) to approximate $P(w_1Z_M + w_2Z_I > b)$. When $(Z_M, Z_I) \in A_1$, we first approximate their marginal P-values as p_M and p_I by (9), then calculate the normal quantile $z_M = \Phi(1 - p_M)$ and $z_I = \Phi(1 - p_I)$. Because the correction primarily impacts the tails of the distributions, the correlation between the two statistics will remain roughly unchanged, i.e., $cor_0(Z_M, Z_I) \approx cor_0(z_M, z_I)$. Thus, when $(Z_M, Z_I) \in A_1$, the tail probability is approximated as $P(\chi_2^2 > (z_M, z_I)\Omega^{-1}(z_M, z_I)')$.

Results

Simulation results

The main purpose of simulations was to investigate the type-I error of Z_M (for testing main genetic effect), Z_I (for detecting SNP-environment interactions) and Q (for detecting either main genetic effect or SNP-environment effect or both). Simulations were performed under different combinations of sample size, MAF and microbiome distance matrices. To make simulations realistic, we used an unweighted distance matrix of the fecal microbiome samples with the 16S rRNA V4 region sequences from the American Gut Project (AGP). The OTU table rarefied to 10,000 sequence reads per sample, as well as metadata, was downloaded from the AGP website. Samples with less than 10,000 sequence reads were excluded from analysis. The weighted and the unweighted UniFrac distance matrices were generated in the Quantitative Insights Into Microbial Ecology[21] (QIIME) pipeline. Because antibiotics may substantially change microbiome composition to generate outliers that may distort the null distribution, we excluded samples with self-reported history of antibiotic usage within one month. After quality control, 1879 subjects remained for analysis. In simulations, we randomly selected N samples for a given sample size N.

For each setting, the type-I error rates were evaluated based on 10^8 simulations under H_0. For the interaction test and the joint test, the binary environment factor had a frequency of 50% and was simulated independent of the SNP. The type-I error rates are summarized in Table 1 for weighted UniFrac distance matrix. The skewness and kurtosis are reported in Figures 3C and 3D. The statistics adjusted for skewness and kurtosis have accurate type-I error rates while the statistics without adjustment have unacceptably high type-I error rates. As sample size increases, the impact of skewness and kurtosis decreases. However, even for a study with $N = 1000$, the type-I error rates are still seriously inflated. The results for the unweighted UniFrac distance matrix and for MAF=0.5 are reported in **Table S1**.

Software implementation, memory requirement and computational complexity

We implemented our algorithms in a software package, microbiomeGWAS, which is freely available at https://github.com/lsncibb/microbiomeGWAS. MicrobiomeGWAS requires three sets of files: a microbiome distance matrix file, a set of PLINK binary files for GWAS genotypes, and a set of covariates. MicrobiomeGWAS processes one SNP at a time and does not load all genotype data into memory; thus, it requires only memory for storing the distance matrix. Variance, skewness and kurtosis can be partitioned into two parts related with the microbiome distance matrix and the MAF of the SNP separately; thus, we can quickly calculate these quantities for a predefined grid of MAFs. The overall computational complexity is about $O(N^2M)$, where N is sample size and M is the number of SNPs. Figure 4 reports the computation time on a Linux server using a single core.

GWAS of microbiome diversity in adjacent normal lung tissues

We applied our methods to a set of lung cancer patients of Italian ancestry in the EAGLE[28] study. All subjects have germline genome-wide SNPs[29] and 16S rRNA microbiome data (V3-V4 region, Illumina MiSeq, 300 paired-end) in histologically normal lung tissues from these patients. Here, the histologically normal lung tissues were 1~5 cm from the tumor tissue. We performed a series of quality control steps to filter out low quality sequence reads: average quality score <20 over 30bp windows, less than 60% similarity to the Greengenes[32] reference or identified as chimera reads using UCHIME[33]. Sequence reads were then processed by QIIME[21] to produce relative abundances (RA) of taxa, two alpha diversity metrics (observed number of species and Shannon’s index) and beta-diversity metrics (unweighted and weighted

UniFrac distances) rarified to 1000 reads. We included 147 subjects with at least 1000 high quality sequence reads for genetic association analysis.

Out of the 147 subjects, 78 are current smokers, 8 are never smokers and 61 are former smokers. Because of the small number of never smokers, we merged never and former smokers as non-current smokers. All of the genetic association analyses were adjusted for sex, age, smoking status, and the top three PCA scores derived based on genome-wide SNPs. Here, the top three PCA scores were selected controlling population stratification because other PCA scores were unassociated with the distance matrices. We included 383,263 common SNPs with MAF ≥ 10% because rarer SNPs were expected to have no statistical power given the current sample size. We first performed GWAS analysis using PLINK[34] to identify SNPs associated with taxa with average RA greater than 0.1% or two alpha-diversity metrics. We did not detect genome-wide significant associations with either main effects or gene by smoking interactions.

Next, we performed GWAS analysis using unweighted and weighted UniFrac distance matrices as a representation of eubacteria beta-diversity. The results for testing main effects are reported in **Figure 5**. Results for testing joint effects (main effect and SNP by smoking status interaction) are reported in **Figure S1**. Because of the small sample size, we observed large values of skewness and kurtosis with magnitude varying with the MAF of the SNPs (**Figure 5A**). The score statistics based on the weighted UniFrac distance matrix had a much larger skewness and kurtosis than did the unweighted UniFrac matrix. **Figures 5B and 5C** report the quantile-quantile (QQ) plot of the logarithm of the association P-values for the unweighted and weighted UniFrac distance matrices, respectively. For each distance matrix, we produced QQ plots for P-values based on the asymptotic approximation and for P-values adjusted for skewness and kurtosis. For both distance matrices, the QQ plots before adjustment strongly deviated from the expected uniform distribution. Our adjustment eliminated the deviation. In addition, consistent with the observation that the skewness and kurtosis were larger for the weighted UniFrac distance matrix, the QQ plot deviated more for the analysis based on the weighted UniFrac distance. Note that the skewness and kurtosis only affect the tail probabilities; thus, the inflation of the QQ plot is not reflected by the genomic control lambda value[35] calculated as the median of P-values. In fact, lambda ≈ 1 for all four QQ plots.

Without correcting for skewness and kurtosis, we identified three and six loci achieving genome-wide significance ($P < 5 \times 10^{-8}$) for the unweighted and weighted UniFrac distance matrices, respectively (**Figure 5D**). After correcting for skewness and kurtosis, no locus remained genome-wide significant (**Figure 5D**), which was verified by 10^8 permutations. Importantly, skewness and kurtosis had a dramatic effect on tail probabilities. Here, we use SNP rs12785513 as an example, which was identified as the top SNP in both analyses. In the unweighted UniFrac analysis, P= 4.4×10^{-9} without adjustment and P=1.6×10^{-6} after adjustment, a 364-fold inflation. The inflation was even larger for weighted UniFrac analysis because of larger skewness and kurtosis (**Figure 5A**). In fact, P= 3.4×10^{-10} without adjustment and P=3.5×10^{-6} after adjustment, a 1000-fold inflation. Although these SNPs were not significant genome-wide, they were the top SNPs from the current study. Thus, we report box-plots for each of these nine SNPs (**Figure 5E**). As expected, in all box plots, microbiome distances tend to be larger in subject pairs with greater genetic distance at these SNPs. These associations remain to be replicated in studies with larger sample sizes.

Finally, we concentrated on the six common SNPs in four genomic regions reported to be associated with lung cancer risk in GWAS of European subjects: rs2036534 and rs1051730 at 15q25.1[36-39] (*CHRNA5–CHRNA3–CHRNB4*), rs2736100 and rs401681 at locus 5p15.33[29,40] (*TERT*/*CLPTM1L*), rs6489769[41] at 12p13.3 (*RAD52*), and rs1333040 at 9p21.3[42] (*CDKN2A*/*CDKN2B*). The SNPs at 15q25.1 and 5p15.33 have the largest effect sizes for lung cancer risk based on the meta-analysis from the Transdisciplinary Research in Cancer of the Lung (TRICL) consortium[42]: OR=1.32 for rs1051730, OR=1.26 for rs2036534, OR=1.13 for rs2736100, and OR=1.14 for rs401681. Rs3131379 at locus 6p21.33[40] (*BAT3*/*MSH5*) was excluded because MAF=7.5%. No SNPs were significantly associated with taxa RAs or alpha-diversity metrics after correcting for multiple testing (data not shown). However, association analysis based on the

UniFrac distance matrices provided evidence that these SNPs may be associated with the lung microbiota (**Table 2**). Importantly, for both unweighted and weighted UniFrac analyses, all six SNPs had *P*-value < 0.5. These SNPs were independent except that rs2036534 and rs1051730 at 15q25.1 were weakly correlated with R^2=0.15. A test combining six z-scores (Z_M) and adjusting for the weak correlation yielded overall P-values 0.0033 and 0.011 for the unweighted and the weighted UniFrac distance matrices, respectively. These results suggest that lung cancer risk SNPs were enriched for genetic association with the composition of the lung microbiome. The results for testing interactions and joint effects are reported in **Table S2**.

Discussion

We developed a software package, microbiomeGWAS, for identifying host genetic variants associated with microbiome composition. MicrobiomeGWAS can test both main effect and SNP-environment interactions. Importantly, we found that the score statistics had positive skewness and kurtosis and that the tail probabilities evaluated based on asymptotic approximations were very liberal. We addressed this problem by explicitly adjusting for skewness and kurtosis. MicrobiomeGWAS runs very efficiently and takes only 36 minutes for testing main effects and 69 minutes for testing joint effects in a GWAS with 2000 subjects and 500,000 markers. Other statistical methods exist for testing the association of microbiome distance matrices. PERMANOVA[43] is an extension of multivariate analysis of variance to a matrix of pairwise distances and relies on permutations to evaluate significance. MiRKAT[44], a recently proposed method based on kernel regression, takes hours for evaluating one association for 2000 subjects. Neither is computationally feasible for analyzing a large-scale GWAS of microbiome.

Interactions of host genetic susceptibility with the microbiome have been postulated for many conditions, including inflammatory bowel diseases[45,46], autoimmune and rheumatic diseases[47-50], diabetes[51], and cancer especially of the colon[52]. All models of these host-microbiome interactions also note the critical role of environmental factors including diet, smoking, drugs, and antibiotics and other medications[53].

Although based on a very small initial sample set, the suggestive associations that we found between the six known lung cancer risk SNPs and the microbiome of adjacent normal lung tissue samples, including effects of cigarette smoking, provide preliminary evidence that our microbiomeGWAS method is likely to be a useful tool for generating data that will unravel host-microbiome interactions with high confidence.

We are working on two extensions for microbiomeGWAS: (1) jointly testing additive and dominant effects and (2) testing genetic associations using many microbiome distance matrices. We have assumed an additive effect model (**Figure 1**); however, several top SNPs in the EAGLE data suggest a dominant effect (e.g. rs8083714 in **Figure 5E**). Thus, a statistic for jointly testing the additive and dominant effects might be powerful for this scenario. The second extension is motivated by the fact that that the power to detect associations depends heavily on the choice of distance matrix. The recently developed generalized UniFrac[26] (gUniFrac) defines a series of distance matrices to reflect different emphasis of using taxa relative abundance information. gUniFrac has been shown to have a robust power for association studies[26]. Extending microbiomeGWAS to gUniFrac, however, requires solving two problems. First, the computational complexity is proportional to the number of distance matrices analyzed for associations, which can be addressed by implementing the algorithms using multithreading technology. Second, we need to derive accurate analytic approximations to the association *P*-values by correcting for the multiple testing introduced by many distance matrices. MiRKAT[44] has an option for using gUniFrac; however, intensive permutations are required to evaluate P-values.

In summary, GWAS of the microbiome of each body site has a potential to understand microbiome variation, to elucidate biological mechanisms of genetic associations, to improve the power of identifying novel disease-associated genetic variants, and to improve the performance of genetic risk prediction. We expect our methods and software to be useful for large-scale GWAS of human microbiome.

Appendices

Appendix A: $Var(G_{ij})$, $Cov(G_{ij}, G_{ik})$, $Var(\Delta_{ij})$ and $Cov(\Delta_{ij}, \Delta_{ik})$.

We first calculate $E(G_{ij})$, $Var(G_{ij})$ and $Cov(G_{ij}, G_{ik})$. Let $p_t = P(g_i = t)$ with $p_0, p_1, p_2 \geq 0$ and $p_0 + p_1 + p_2 = 1$. We can also assume the Hardy–Weinberg equilibrium and characterize the probabilities as the allele frequency: $p_0 = (1-f)^2, p_1 = 2f(1-f)$ and $p_2 = f^2$. Some algebra leads to

$$E(G_{ij}) = E|g_i - g_j| = \sum_{m,n\in\{0,1,2\}} p_m p_n |m-n| = 2p_0p_1 + 2p_1p_2 + 4p_0p_2; \qquad (A1)$$

$$Var(G_{ij}) = E(G_{ij}^2) - E(G_{ij})^2 = (2p_0p_1 + 2p_1p_2 + 8p_0p_2) - (2p_0p_1 + 2p_1p_2 + 4p_0p_2)^2; \qquad (A2)$$

$$Cov(G_{ij}, G_{ik}) = p_1(1-p_1) + 4p_0p_2(1+p_1) - (2p_0p_1 + 2p_1p_2 + 4p_0p_2)^2. \qquad (A3)$$

Now consider $\Delta_{ij} = |g_iE_i - g_jE_j|$. When E_i is binary, $g_iE_i = 0, 1$ or 2. Let $p_t' = P(g_iE_i = t)$. Then, $E(\Delta_{ij})$, $Var(\Delta_{ij})$ and $Cov(\Delta_{ij}, \Delta_{ik})$ can be calculated similarly using (A1), (A2) and (A3).

Appendix B: Calculating $\rho = Cor_0(Z_M, Z_I|\boldsymbol{D}')$

Let $G_{ij}' = G_{ij} - EG_{ij}$ and $\Delta_{ij}' = \Delta_{ij} - E\Delta_{ij}$. We first calculate the covariance under H_0:

$$Cov_0(S_M, S_I|\boldsymbol{D}') = Cov_0\left(\sum\nolimits_{i<j} d_{ij}' G_{ij}', \sum\nolimits_{m<n} d_{mn}' \Delta_{mn}'\right) = \sum\nolimits_{i<j,m<n} d_{ij}' d_{mn}' Cov(G_{ij}, \Delta_{mn}).$$

When (i, j, m, n) are distinct, $Cov(G_{ij}, \Delta_{mn}) = 0$. Some algebra leads to

$$Cov_0(S_M, S_I|\boldsymbol{D}') = \binom{N}{2} Cov(G_{ij}, \Delta_{ij})\mu_2 + 6\binom{N}{3} Cov(G_{ij}, \Delta_{ik})\mu_3 \qquad (A4)$$

with μ_2 and μ_3 specified in (2) and (3). Combining (2), (5) and (A4), we have

$$\rho = \frac{Cov_0(S_M, S_I|\boldsymbol{D}')}{\sqrt{Var_0(S_M|\boldsymbol{D}')Var_0(S_I|\boldsymbol{D}')}} \xrightarrow{N\to\infty} \frac{Cov(G_{ij}, \Delta_{ik})}{\sqrt{Cov(G_{ij}, G_{ik})Cov(\Delta_{ij}, \Delta_{ik})}}. \qquad (A5)$$

(A5) suggests that the correlation is asymptotically independent of the microbiome distance matrix. In real data analyses, we found that (A5) was very accurate when sample size $N \geq 50$. The details of calculating $Cov(G_{ij}, \Delta_{ij})$ and $Cov(G_{ij}, \Delta_{ik})$ are provided in **Supplemental Data**.

Appendix C: A statistic for jointly testing $H_0\colon \beta_M = \beta_I = 0$ vs $H_1\colon \beta_M > 0$ or $\beta_I > 0$

Denote $\boldsymbol{Z} = (Z_M, Z_I)^T$. Under H_0, $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Let $\xi_M = E_1 Z_M \geq 0$ and $\xi_I = E_1 Z_I \geq 0$ be the non-centrality parameter of the two score statistics. Apparently the original testing problem is equivalent for testing $H_0\colon \xi_M = \xi_I = 0$ vs $H_1\colon \xi_M > 0$ or $\xi_I > 0$. Given the observed values (Z_M, Z_I), the likelihood ratio statistic is simplified as

$$Q = \boldsymbol{Z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Z} - (\boldsymbol{Z} - \boldsymbol{\xi})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \boldsymbol{\xi}) \qquad (A6)$$

where $\boldsymbol{\xi} = (\xi_M, \xi_I)^T = \operatorname{arginf}_{\xi_M \geq 0, \xi_I \geq 0} Q$ (**Figure S2A**).

To simplify the optimization problem in (A6), we perform a linear transformation: $\boldsymbol{Y^T} = \boldsymbol{Z}^T \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\boldsymbol{v^T} = \boldsymbol{\xi}^T \boldsymbol{\Sigma}^{-\frac{1}{2}}$, where

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{1-\rho} & 0 \\ 0 & 1/\sqrt{1+\rho} \end{pmatrix}. \qquad (A7)$$

Under this transformation, $\boldsymbol{Q} = \boldsymbol{Y}^T \boldsymbol{Y} - (\boldsymbol{Y} - \boldsymbol{v})^T (\boldsymbol{Y} - \boldsymbol{v})$ and can be interpreted as the difference of the square of two distances (**Figure S2B**). The original parameter space $\{(\xi_M, \xi_I)\colon \xi_M \geq 0, \xi_I \geq 0\}$ is now transformed to $\{(v_1, v_2)\colon v_2 \geq \theta v_1, v_2 \geq -\theta v_1\}$ with $\theta = \sqrt{(1-\rho)/(1+\rho)}$. Thus, the new parameter space is bounded by two lines represented by $v_2 \geq \theta v_1$ and $v_2 \geq -\theta v_1$. We partition the 2D plane into four parts (see **Figure S2B**), identify $\boldsymbol{v} = \operatorname{arginf}_{v \in A_1} (\boldsymbol{Y} - \boldsymbol{v})^T (\boldsymbol{Y} - \boldsymbol{v})$ and calculate Q:

$$Q = \begin{cases} Y_1^2 + Y_2^2 & (Y_1, Y_2) \in A_1 \\ (Y_2 - Y_1/\theta)^2/(1+\theta^{-2}) & (Y_1, Y_2) \in A_2 \\ (Y_2 + Y_1/\theta)^2/(1+\theta^{-2}) & (Y_1, Y_2) \in A_3 \\ 0 & (Y_1, Y_2) \in A_4 \end{cases} \qquad \text{(A8)}$$

We now perform an inverse transformation using matrix

$$\boldsymbol{\Sigma}^{\frac{1}{2}} = \left[\frac{1}{\sqrt{2}}\begin{pmatrix} \sqrt{1-\rho} & 0 \\ 0 & \sqrt{1+\rho} \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\right] \qquad (A9)$$

to return to the original parameter space. The four areas $\{A_1, A_2, A_3, A_4\}$ under the original space are in **Figure 2** and **Figure S2C**.

Tedious calculations show that $(Y_2 + Y_1/\theta)^2/(1+\theta^{-2}) = (w_2 Z_M + w_1 Z_I)^2$ with $w_1 = (\theta - 1/\theta)/2$ and $w_2 = (\theta + 1/\theta)/2$. Similarly, $(Y_2 - Y_1/\theta)^2/(1+\theta^{-2}) = (w_1 Z_M + w_2 Z_I)^2$. This proves (7). In addition, $w_1 Z_M + w_2 Z_I \geq 0$ and $w_1^2 + 2\rho w_1 w_2 + w_2^2 = 1$; thus, $P\{(w_1 Z_M + w_2 Z_I)^2 > b^2\} = P\{w_1 Z_M + w_2 Z_I > b\} = P\{N(0,1) > b\}$. This proves (8). The probabilities in (8) could also be calculated from **Figure S2B**: $q_1 = 1/2 - (\arctan\theta)/\pi$, $q_2 = q_3 = 1/4$.

Appendix D: Calculating skewness and kurtosis under H_0

By definition, $\gamma = E_0(S_M^3|\boldsymbol{D}')/Var_0^{3/2}(S_M|\boldsymbol{D}')$ and $\kappa = E_0(S_M^4|\boldsymbol{D}')/Var_0^2(S_M|\boldsymbol{D}') - 3$. We first calculate $E_0(S_M^3|\boldsymbol{D}')$. Let $G_{ij}' = G_{ij} - EG_{ij}$. We have

$$E_0(S_M^3|\boldsymbol{D}') = E_0\left(\sum\nolimits_{i<j} d_{ij}' G_{ij}'\right)^3 = \sum_{i<j,m<n,s<t} d_{ij}' d_{mn}' d_{st}'\, E G_{ij}' G_{mn}' G_{st}'.$$

Figure S3A lists all combinations of (i,j,m,n,s,t) with $E G_{ij}' G_{mn}' G_{st}' \neq 0$, which leads to

$$E_0(S_M^3|\boldsymbol{D}') = \binom{N}{2}\mu_4 E G_{ij}'^4 + \binom{N}{3}\left(\mu_5 E G_{ij}'^2 G_{ik}' + \mu_6 E G_{ij}' G_{jk}' G_{ik}'\right) + \binom{N}{4}\left(\mu_7 E G_{ij}' G_{jk}' G_{kl}' + \mu_8 E G_{ij}' G_{ik}' G_{il}'\right),$$

where $(\mu_4, \mu_5, \mu_6, \mu_7, \mu_8)$ are provided in **Supplemental Data**. Similarly,

$$E_0(S_M^4|\boldsymbol{D}') = E_0\left(\sum\nolimits_{i<j} d_{ij}' G_{ij}'\right)^4 = \sum_{i<j,m<n,s<t,x<y} d_{ij}' d_{mn}' d_{st}' d_{xy}'\, E G_{ij}' G_{mn}' G_{st}' G_{xy}'.$$

Figure S3B lists 15 combinations of (i,j,m,n,s,t,x,y) with $E G_{ij}' G_{mn}' G_{st}' G_{xy}' \neq 0$. Thus,

$$E_0(S_M^4|\boldsymbol{D}) = \binom{N}{2}\mu_9 EG_{ij}'^4 + \binom{N}{3}\left(\mu_{10}EG_{ij}'^3G_{ik}' + \mu_{11}EG_{ij}'^2G_{ik}'^2 + \mu_{12}EG_{ij}'^2G_{jk}'G_{ik}'\right)$$

$$+\binom{N}{4}\left(\mu_{13}EG_{ij}'^2G_{jk}'G_{kl}' + \mu_{14}EG_{ij}'G_{jk}'^2G_{kl}' + \mu_{15}EG_{ij}'^2G_{ik}'G_{il}' + \mu_{16}EG_{ij}'G_{jk}'G_{ik}'G_{il}' + \mu_{17}EG_{ij}'G_{jk}'G_{kl}'G_{il}' + \mu_{18}EG_{ij}'^2G_{kl}'^2\right)$$

$$+\binom{N}{5}\left(\mu_{19}EG_{ij}'G_{jk}'G_{kl}'G_{lm}' + \mu_{20}EG_{ij}'G_{ik}'G_{il}'G_{im}' + \mu_{21}EG_{ij}'G_{ik}'G_{il}'G_{lm}' + \mu_{22}EG_{ij}'G_{ik}'G_{lm}'^2\right) + \binom{N}{6}\mu_{23}EG_{ij}'G_{ik}'G_{lm}'G_{ln}'$$

The constants $(\mu_9, \cdots, \mu_{23})$ are dependent on $\boldsymbol{D}$ and are provided in **Supplemental Data**.

Note that $Var_0(S_M|\boldsymbol{D}')\sim O(N^3), E_0(S_M^3|\boldsymbol{D}')\sim O(N^4)$, thus $\gamma\sim O(\frac{1}{\sqrt{N}})$. Similarly, we can prove $\kappa\sim O(\frac{1}{N})$.

Appendix E: Improve *P*-value approximations by adjusting for skewness and kurtosis

We assume that $E_0Z = 0,\ Var_0Z = 1, \gamma = E_0Z^3$ and $\kappa = E_0Z^4 - 3$ under the original probability measure P_0. The tail probability $P_0(Z > b)$ for a large value of b is sensitive to the non-normality of Z, characterized by γ and κ. We define a new probability measure by embedding to the exponential probability density

$$dP_\xi = \exp\big(\xi Z - \phi(\xi)\big)\, dP_0 \qquad (A10)$$

where $\phi(\xi) = \log E_0 \exp(\xi Z)$ is the log moment generating function. Note that $\gamma = \phi'''(0)$ and $\kappa = \phi''''(0)$. Because $E_0(Z) = 0$ and $Var_0(Z) = 1$, the Taylor expansion leads to $\phi(\xi) \approx \xi^2/2 + \gamma\xi^3/6 + \kappa\xi^4/24$. Under P_ξ, we have

$$E_\xi Z = \int Z dP_\xi = \phi'(\xi) \approx \xi + \frac{\gamma}{2}\xi^2 + \frac{\kappa}{6}\xi^3 \qquad (A11)$$

and

$$Var_\xi Z = \phi''(\xi) \approx 1 + \gamma\xi + \frac{\kappa}{2}\xi^2. \qquad (A12)$$

We choose ξ such that $E_\xi Z \approx b$ by numerically solving an equation

$$\xi + \frac{\gamma}{2}\xi^2 + \frac{\kappa}{6}\xi^3 = b. \qquad (A13)$$

Under the probability measure P_ξ, $Z\sim N(b, \sigma^2)$ approximately with $\sigma^2 = 1 + \gamma\xi + k\xi^2/2$ in (A12).

By the likelihood ratio identity and (A10), we have

$$P_0(Z > b) = E_0 I_{Z>b} = E_\xi \frac{dP_0}{dP_\xi} I_{Z>b} = E_\xi e^{\phi(\xi)-\xi Z} I_{Z>b} = e^{\phi(\xi)} E_\xi e^{-\xi Z} I_{Z>b}. \qquad (A14)$$

Note that $e^{-\xi Z}$ decays very fast when Z increases. Thus, the integral $E_\xi e^{-\xi Z} I_{Z>b}$ does not heavily depend on the tail distribution of Z. Assuming $Z \sim N(b, \sigma^2)$ under P_ξ, we can verify that

$$E_\xi e^{-\xi Z} I_{Z>b} = e^{-b\xi + \frac{\sigma^2 \xi^2}{2}} \Phi(-\sigma\xi). \qquad (A15)$$

Combining (A14) and (A15) gives $P_0(Z > b) \approx e^{\phi(\xi) - b\xi + \frac{\sigma^2 \xi^2}{2}} \Phi(-\sigma\xi)$, which is further approximated as $P_0(Z > b) \approx e^{-b\xi + \frac{(1+\sigma^2)}{2}\xi^2 + \frac{\gamma}{6}\xi^3 + \frac{\kappa}{24}\xi^4} \Phi(-\sigma\xi)$, because $\phi(\xi) \approx \xi^2/2 + \gamma\xi^3/6 + \kappa\xi^4/24$ based on the Taylor expansion. This proves (9).

If we correct skewness but assume kurtosis $\kappa = 0$, then $\phi(\xi) \approx \xi^2/2 + \gamma\xi^3/6$. We recalculate ξ by setting $\kappa = 0$ in (A13) to derive $\xi = (\sqrt{1 + 2\gamma b} - 1)/\gamma$. This proves (10).

Supplemental Data

Supplemental Data include 2 tables and 3 figures can be found with this article online at XXX.

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.

Web Resources

The URLs for data provide herein are as follows:

American Gut Project: https://www.microbio.me/americangut/
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/
QIIME: http://qiime.org/
EAGLE study: http://eagle.cancer.gov/

References

1. Turnbaugh, P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-4 (2009).
2. Morgan, X.C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* **13**, R79 (2012).
3. Ahn, J. *et al.* Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* **105**, 1907-11 (2013).
4. Goedert, J.J., Jones, G., Hua, X., Xu, X., Yu, G., Flores, R., Falk, R. T., Gail, M. H., Shi, J., Ravel, J. and Feigelson, S. H. Investigation of the Association Between the Fecal Microbiota and Breast Cancer in Postmenopausal Women: a Population-Based Case-Control Pilot Study. *J Natl Cancer Inst.* **1;107(8)**(2015).
5. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048-1052 (2014).
6. Wu, G.D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105-8 (2011).
7. Tong, M. *et al.* Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *ISME J* **8**, 2193-206 (2014).
8. Knights, D. *et al.* Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med* **6**, 107 (2014).
9. McKnite, A.M. *et al.* Murine Gut Microbiota Is Defined by Host Genetics and Modulates Variation of Metabolic Traits. *Plos One* **7**(2012).
10. Benson, A.K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* **107**, 18933-8 (2010).
11. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789-99 (2014).
12. Davenport, E.R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS One* **10**, e0140301 (2015).
13. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
14. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* **24**, 14-24 (2014).

15. Bell, J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines (vol 12, pg R10, 2011). *Genome Biology* **12**(2011).
16. Shi, J. *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat Commun* **5**, 3365 (2014).
17. McVicker, G. *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science* **342**, 747-749 (2013).
18. Kilpinen, H. *et al.* Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science* **342**, 744-747 (2013).
19. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat Genet* **43**, 565-9 (2011).
20. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**, 35-46 (2009).
21. Caporaso, J.G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-6 (2010).
22. Bray, J.R.a.C., J. T. An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* **27:325-349**(1957).
23. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-35 (2005).
24. Lozupone, C.A., Hamady, M., Kelley, S.T. & Knight, R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* **73**, 1576-85 (2007).
25. Lozupone, C., Hamady, M. & Knight, R. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, 371 (2006).
26. Chen, J. *et al.* Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**, 2106-13 (2012).
27. Gevers, D. *et al.* The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol* **10**, e1001377 (2012).
28. Landi, M.T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008).
29. Landi, M.T. *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* **85**, 679-91 (2009).
30. Tu, I.P. & Siegmund, D. The maximum of a function of a Markov chain and application to linkage analysis. *Advances in Applied Probability* **31**, 510-531 (1999).
31. Siegmund, D. Sequential Analysis: Tests and Confidence Intervals. *New York: Springer.* (1985).
32. DeSantis, T.Z., Dubosarskiy, I., Murray, S.R. & Andersen, G.L. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* **19**, 1461-8 (2003).
33. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-200 (2011).
34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
35. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
36. Hung, R.J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-7 (2008).
37. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404-6 (2008).
38. Thorgeirsson, T.E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-42 (2008).

39. Amos, C.I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**, 616-22 (2008).

40. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**, 1407-9 (2008).

41. Shi, J. *et al.* Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. *Cancer Discov* **2**, 131-9 (2012).

42. Timofeeva, M.N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* **21**, 4980-95 (2012).

43. Anderson, M.J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26:** , 32-46. (2001).

44. Zhao, N. *et al.* Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *Am J Hum Genet* **96**, 797-807 (2015).

45. Leone, V.A., Cham, C.M. & Chang, E.B. Diet, gut microbes, and genetics in immune function: can we leverage our current knowledge to achieve better outcomes in inflammatory bowel diseases? *Current Opinion in Immunology* **31**, 16-23 (2014).

46. Huang, H., Vangay, P., McKinlay, C.E. & Knights, D. Multi-omics analysis of inflammatory bowel disease. *Immunol Lett* **162**, 62-8 (2014).

47. Troncone, R. & Discepolo, V. Celiac disease and autoimmunity. *J Pediatr Gastroenterol Nutr* **59 Suppl 1**, S9-S11 (2014).

48. Yeoh, N., Burton, J.P., Suppiah, P., Reid, G. & Stebbings, S. The role of the microbiome in rheumatic diseases. *Curr Rheumatol Rep* **15**, 314 (2013).

49. Sparks, J.A. & Costenbader, K.H. Genetics, environment, and gene-environment interactions in the development of systemic rheumatic diseases. *Rheum Dis Clin North Am* **40**, 637-57 (2014).

50. Smith, J.A. Update on ankylosing spondylitis: current concepts in pathogenesis. *Curr Allergy Asthma Rep* **15**, 489 (2015).

51. Nielsen, D.S., Krych, L., Buschard, K., Hansen, C.H. & Hansen, A.K. Beyond genetics. Influence of dietary factors and gut microbiota on type 1 diabetes. *FEBS Lett* **588**, 4234-43 (2014).

52. Birt, D.F. & Phillips, G.J. Diet, genes, and microbes: complexities of colon cancer prevention. *Toxicol Pathol* **42**, 182-8 (2014).

53. Marietta, E., Rishi, A. & Taneja, V. Immunogenetic control of the intestinal microbiota. *Immunology* **145**, 313-22 (2015).

Table 1: Type-I error rates estimated based on 10^8 simulations. Minor allele frequency = 20%. Simulations were based on the weighted UniFrac distance matrix of the gut microbiome data from the American Gut Project. Reported are the type-I error inflation factor. A value greater than 1 indicates an inflated type-I error.

		Z_M			Z_I			Q		
	N	$\alpha = 10^{-3}$	10^{-5}	10^{-7}	10^{-3}	10^{-5}	10^{-7}	10^{-3}	10^{-5}	10^{-7}
asymptotic approximation	100	5.5	51.6	610.0	4.7	36.1	342.8	7.3	80.9	1148.0
	200	3.7	23.0	187.3	3.1	15.8	105.5	4.6	33.0	316.7
	500	2.4	9.4	45.2	2.1	6.7	25.5	2.8	11.9	64.1
	1000	2.0	5.7	21.3	1.8	4.4	14.0	2.2	6.9	28.5
adjusted for skewness and kurtosis	100	1.0	1.2	0.7	1.0	1.1	0.6	1.0	1.5	2.0
	200	1.0	1.1	1.0	1.0	1.1	0.7	0.9	1.3	1.8
	500	1.0	1.1	1.3	1.0	1.0	0.9	0.9	1.0	1.7
	1000	1.0	1.0	1.2	1.0	1.0	0.8	0.9	1.0	1.1

Table 2: Association P-values between lung cancer risk SNPs and microbiome composition in the EAGLE data.

SNP	Chr	Annotated genes	unweighted UniFrac	weighted UniFrac
rs2036534	15q25.1	*CHRNA3/4/5*	0.425	0.167
rs1051730	15q25.1	*CHRNA3/4/5*	0.020	0.401
rs2736100	5p15.33	*TERT*	0.089	0.267
rs401681	5p15.33	*CLPTM1L*	0.056	0.005
rs6489769	12p13.3	*RAD52*	0.197	0.329
rs1333040	9p21.3	*CDKN2A/B*	0.249	0.224
Overall test			0.0032	0.011

Figure 1 Microbiome distances are positively correlated with genetic distances at an associated SNP.

Figure 2 Define the joint test for testing $H_0: \beta_M = \beta_I = 0$ vs. $\beta_M > 0\ or\ \beta_I > 0$. We assume that $Z_M \sim N(0,1)$, $Z_I \sim N(0,1)$ and $cor(Z_M, Z_I) = \rho$ under H_0. Details are in **Appendix C**.

Figure 3 Correcting tail probabilities for skewness and kurtosis. (A) The standard normal distribution $N(0,1)$ and an approximately normal distribution with positive skewness. The skewness has big impact when calculating the tail probability $P(Z > b)$ for a large value of b. (B) Numerical evaluation of tail probability approximation for Z_M. We used the unweighted UniFrac distance matrix of 500 samples from the American Gut Project (AGP). For each value of $b(> 0)$, we calculated P-values $P(Z_M > b)$ based on $N(0,1)$, skewness correction, both skewness and kurtosis correction, and 10^8 simulations. (C) Skewness depends on MAF of SNPs and the sample size of the study, calculated based on the weighted UniFrac distance matrix in AGP data. (D) Kurtosis depends on MAF of SNPs and the sample size, calculated based on the weighted UniFrac distance matrix in AGP data.

Figure 4 Computation time for a microbiome GWAS with 500,000 SNPs. “Main”: computation time for testing main effect only. “All”: computation time for testing main effect, interaction and the joint null hypothesis $H_0: \beta_M = 0, \beta_I = 0$.

Figure 5 Results of analyzing the microbiome GWAS data of 147 adjacent normal lung tissues in the EAGLE study. (A) Skewness and kurtosis for the main effect test using the unweighted and the weighted UniFrac distance matrices. (B) Quantile-quantile (QQ) plot for association P-values using the unweighted UniFrac distance matrix. “Adjusted”: P-values were corrected for skewness and kurtosis. “Unadjusted”: P-values were approximated based on the asymptotic distribution $N(0,1)$. (C) Quantile-quantile (QQ) plot for association P-values using the weighted UniFrac distance matrix. (D) Manhattan plots based on the unweighted or the weighted UniFrac distance matrices. (E) Box plots for the top nine loci in microbiome GWAS analysis. Subject pairs are classified into three groups according to the genetic distance $|g_i - g_j|$ at the SNP. The y-coordinate is the microbiome distance.

Figure 1

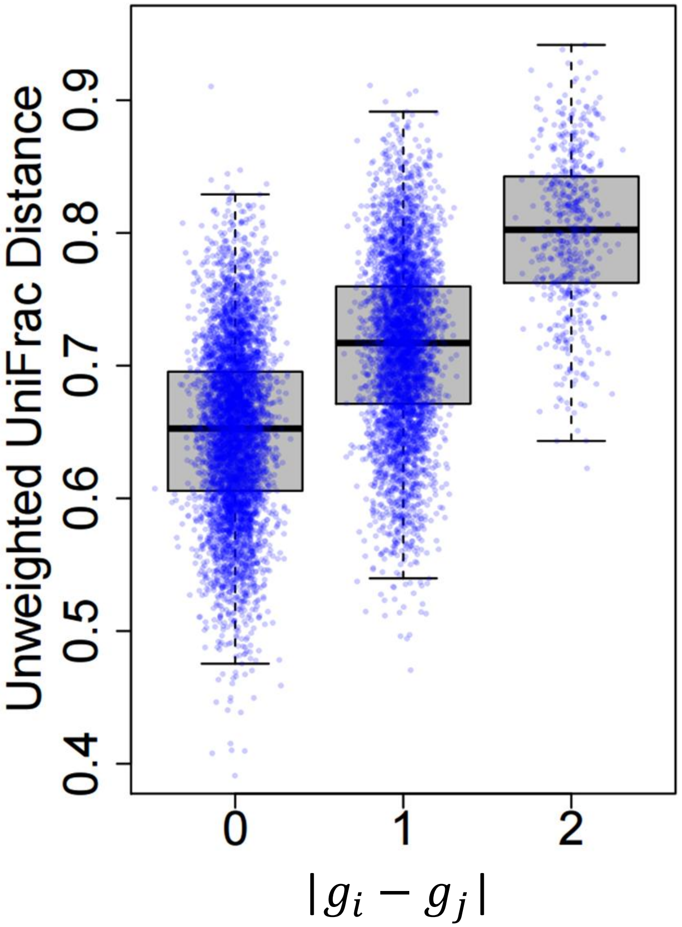


Figure 2

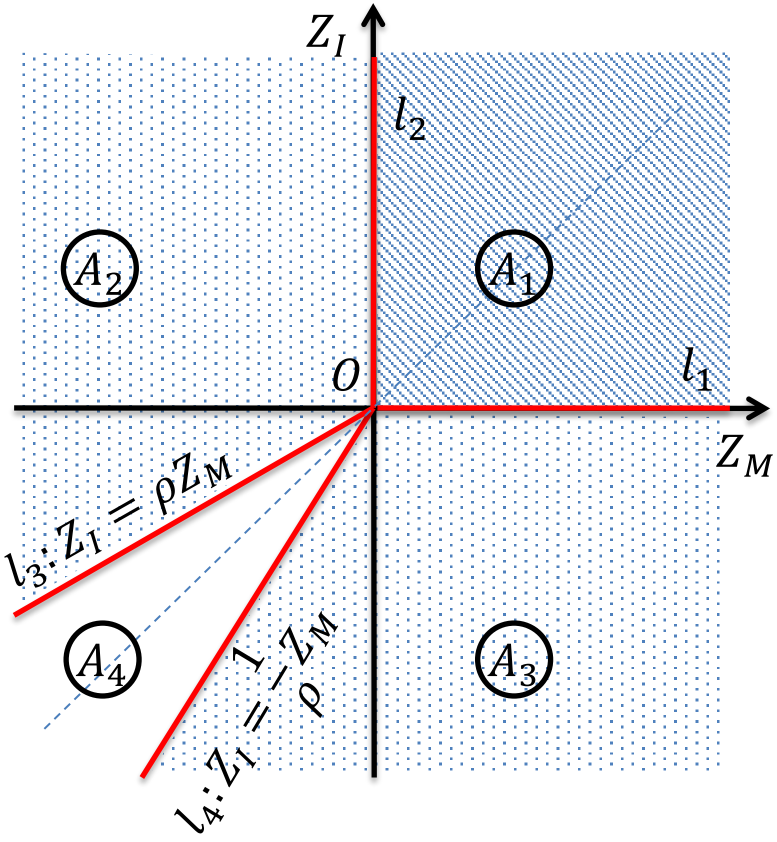


Figure 3

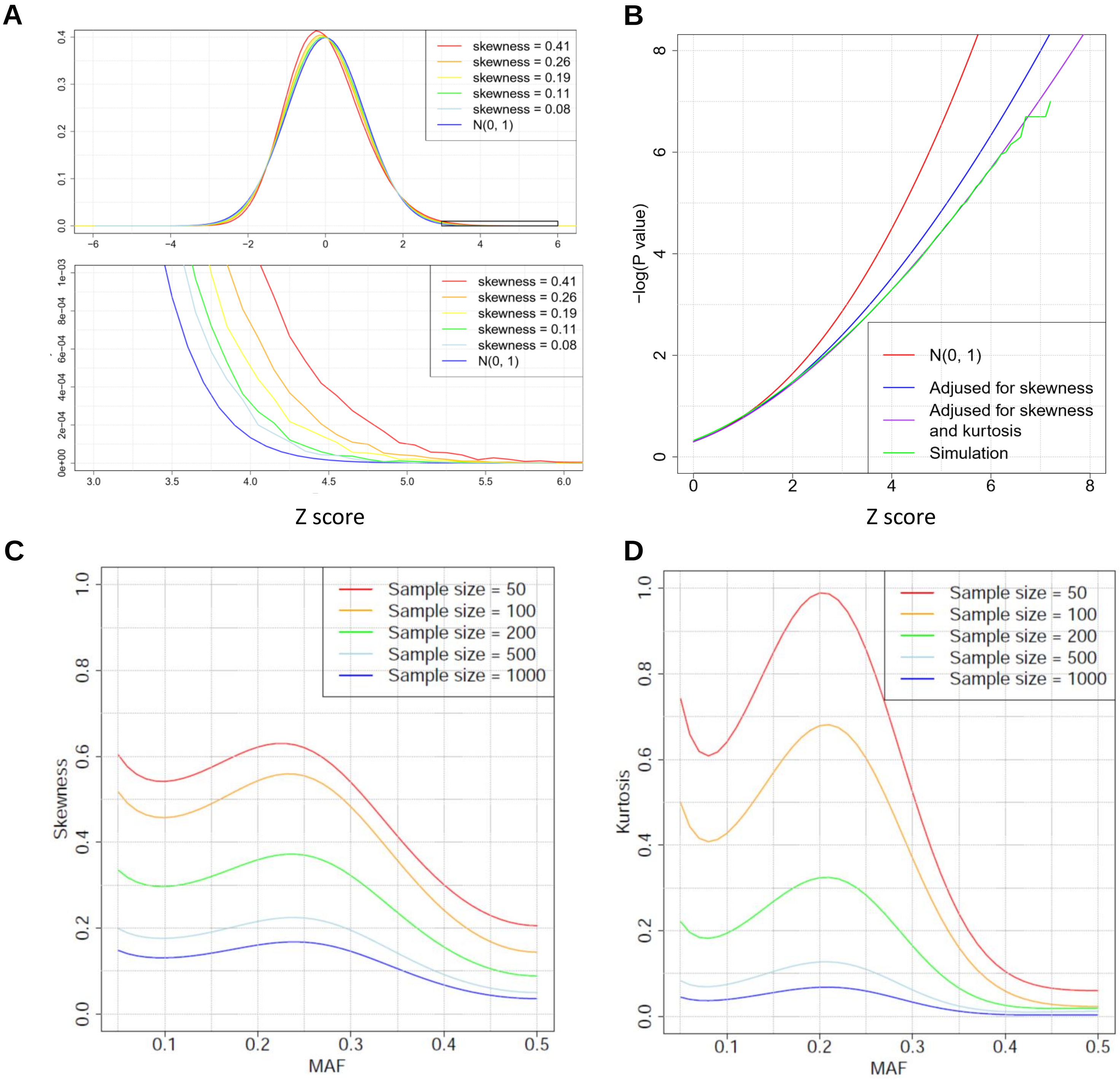
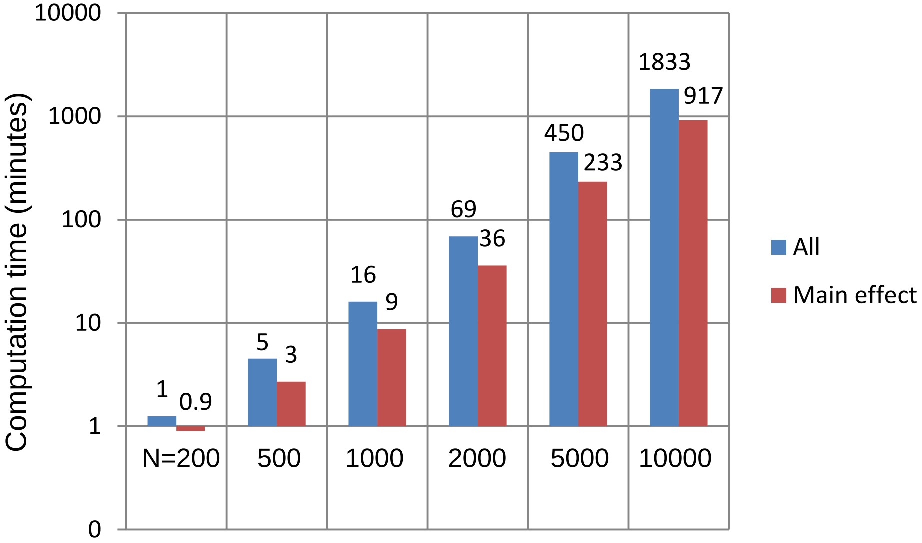

A
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6
skewness = 0.41
skewness = 0.26
skewness = 0.19
skewness = 0.11
skewness = 0.08
N(0, 1)
1e−03
8e−04
6e−04
4e−04
2e−04
0e+00
3.0
3.5
4.0
4.5
5.0
5.5
6.0
skewness = 0.41
skewness = 0.26
skewness = 0.19
skewness = 0.11
skewness = 0.08
N(0, 1)
Z score
B
−log(P value)
0
2
4
6
8
0
2
4
6
8
N(0, 1)
Adjused for skewness
Adjused for skewness and kurtosis
Simulation
Z score
C
Skewness
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
MAF
Sample size = 50
Sample size = 100
Sample size = 200
Sample size = 500
Sample size = 1000
D
Kurtosis
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
MAF
Sample size = 50
Sample size = 100
Sample size = 200
Sample size = 500
Sample size = 1000

Figure 4


10000
1000
100
10
1
0
Computation time (minutes)
1
0.9
5
3
16
9
69
36
450
233
1833
917
N=200
500
1000
2000
5000
10000
All
Main effect

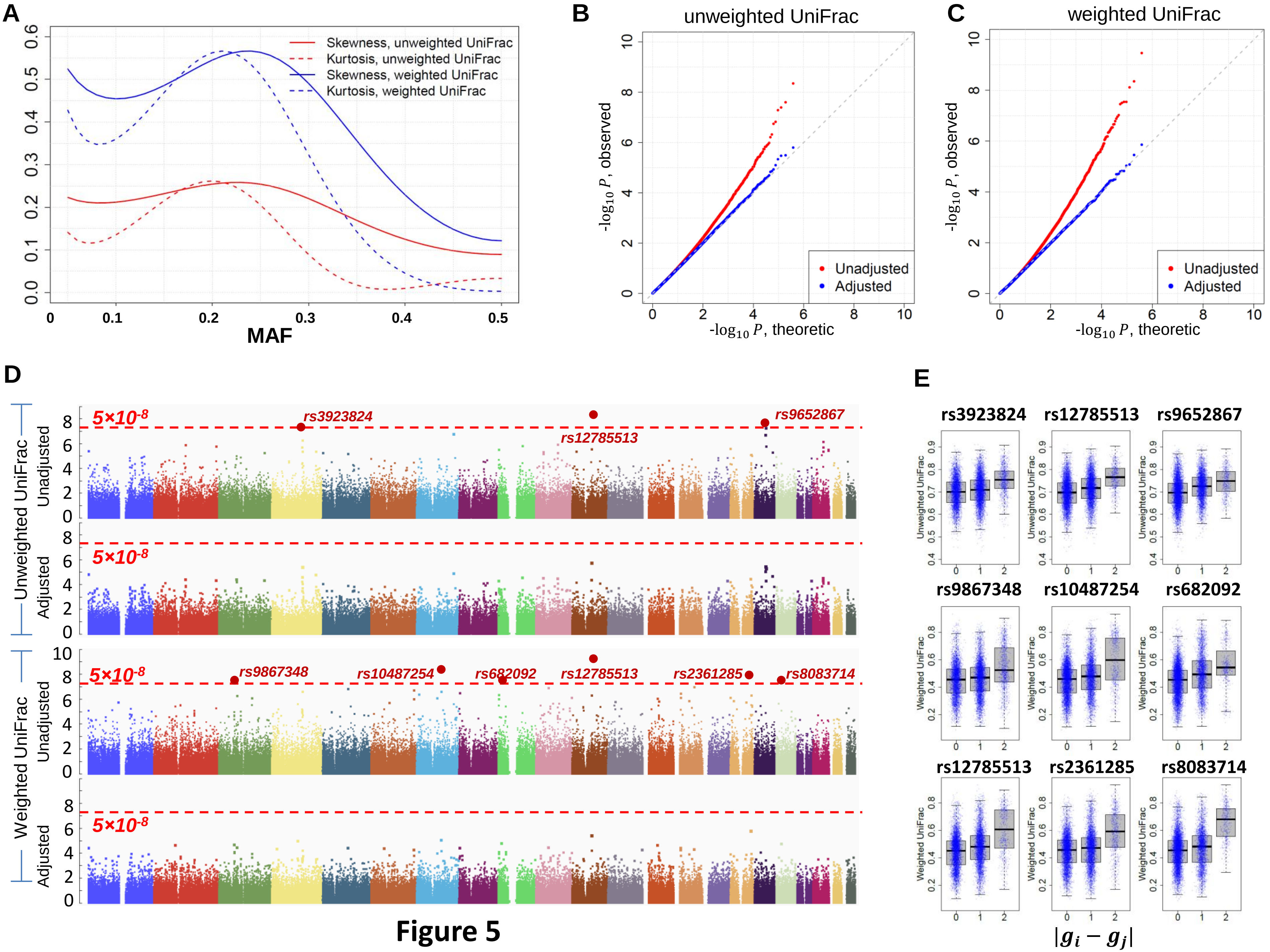

A
Skewness, unweighted UniFrac
Kurtosis, unweighted UniFrac
Skewness, weighted UniFrac
Kurtosis, weighted UniFrac
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
0.1
0.2
0.3
0.4
0.5
MAF
B
unweighted UniFrac
C
weighted UniFrac
$-\log_{10} P$, observed
$-\log_{10} P$, theoretic
0
2
4
6
8
10
Unadjusted
Adjusted
D
Unweighted UniFrac
Weighted UniFrac
Unadjusted
Adjusted
5×10^{-8}
rs3923824
rs12785513
rs9652867
rs9867348
rs10487254
rs682092
rs2361285
rs8083714
E
rs3923824
rs12785513
rs9652867
rs9867348
rs10487254
rs682092
rs12785513
rs2361285
rs8083714
Unweighted UniFrac
Weighted UniFrac
0.2
0.4
0.5
0.6
0.7
0.8
0.9
0
1
2
$|g_i - g_j|$

Figure 5