

1
2
3
4
5
6
7
8 MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition
9 Xing Hua, Lei Song, Guoqin Yu, James J. Goedert,
10 Christian C. Abnet, Maria Teresa Landi, Jianxin Shi
11 Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health,
12 Bethesda, Maryland
13
14 Correspondence to:
15
16 Jianxin Shi, PhD
17 Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH
18 9609 Medical Center Drive, 7E596, Bethesda, MD 20892
19 Tel: (240) 276-7419
20 Email: Jianxin.Shi@nih.gov
21
22
23
24
25

26

Abstract

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

46

47

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

50

51 **Introduction**

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

61 The human microbiome has been associated with diseases, including obesity¹, inflammatory bowel
62 disease (IBD)², colorectal cancer³ and breast cancer⁴. Thus, identifying factors that have a sustained
63 impact on the microbiome is fundamental for elucidating its role in health conditions and for developing
64 treatment strategies. Increasing evidence suggests that microbiome composition at a specific site of the
65 human body is impacted by environmental factors^{5,6}, host genetics^{7,8}, and possibly by their interactions. In
66 the mouse, quantitative trait loci (QTL) studies have identified loci contributing to the variation of the gut
67 microbiome using linkage analysis^{9,10}. Recently, Goodrich et al.¹¹ systematically investigated the
68 heritability of the human gut microbiome by comparing monozygotic twins to dizygotic twins and found
69 substantial heritability in different microbiome metrics, suggesting the important role of host genetics on
70 gut microbiome diversity. Associations between individual host genetic variants and microbiome taxa
71 abundances have also begun to emerge in other human samples^{7,8,12}. These studies suggest that genome-
72 wide association studies (GWAS) have great potential to identify host genetic variants associated with
73 microbiome diversity.

74 GWAS of complex human diseases have identified many risk SNPs; however, the biological mechanisms
75 are largely unknown for the majority of the risk SNPs. QTL studies of intermediate traits, e.g., gene

76 expression^{13,14}, DNA methylation^{15,16}, chromatin structure^{17,18}, and metabolite production^{19,20}, have
77 provided useful insights on biological mechanisms of the GWAS findings. The human microbiome at a
78 specific body site is another important and informative intermediate trait for interpreting GWAS signals.
79 Knights et al.⁸ reported that a risk SNP for IBD located in *NOD2* was associated with the relative
80 abundance of *Enterobacteriaceae* in the human gut microbiome. Tong et al.⁷ show that a loss-of-function
81 allele in *FUT2* that increases the risk of developing Crohn's Disease (CD) may modulate energy
82 metabolism of the gut microbiome. In both examples, the microbiome is a potential intermediate for
83 explaining the association between risk SNPs and disease risks, although a formal mediation analysis is
84 required based on samples with genotype, microbiome, and disease status data. Moreover, identifying
85 microbiome-associated host genetic variants has the potential to prioritize SNPs for discovery and to
86 improve the performance of polygenic risk prediction.

87 Three types of microbiome metrics can be derived as phenotypes for GWAS analysis. First, for each
88 taxon at a specified taxonomic level (phylum, class, order, family, genus, and species), we calculate the
89 relative abundance (RA) of the taxon as the ratio of the number of sequencing reads assigned to the taxon
90 to the total number of sequencing reads. In 16S ribosomal RNA sequence profiles, approximately 100-
91 200 taxa with average RAs $\geq 0.1\%$ (from the phylum level to the genus level) across samples are abundant
92 enough for QTL analysis. One can perform a Poisson regression to examine the association between RA
93 of each taxon and each SNP. Significant associations are identified using Bonferroni correction ($P < 5 \times 10^{-8}/200 = 2.5 \times 10^{-10}$) or by controlling FDR at an appropriate level. Second, multiple alpha-diversity metrics²¹
94 can be calculated to reflect the richness (e.g., number of unique taxa) and evenness of each microbiome
95 community after a procedure called rarefaction, that eliminates the dependence between estimated alpha
96 diversity and the variable total number of sequencing reads across subjects. Once alpha-diversity metrics
97 are derived, one may perform standard GWAS with alpha diversity as the phenotype using linear
98 regression.

100 Because a microbiota functions as a community, the most important analysis for a microbiome GWAS
101 may be by assessing the complete structure of the community by using a pairwise microbiome distance
102 matrix (or beta-diversity) of the microbial community. Microbiome distances can be defined in different
103 ways, based on using phylogenetic tree information or each taxon's abundance information. Bray–Curtis
104 dissimilarity²² quantifies the difference between two microbiome communities using the abundance
105 information of specific taxa. UniFrac^{23–25} is another widely used distance metric. Unlike the Bray–Curtis
106 dissimilarity metric, UniFrac compares microbiome communities by using information on the relative
107 relatedness of each taxon, specifically by phylogenetic distance (branch lengths on a phylogenetic tree).
108 UniFrac has two variants: the weighted UniFrac²⁴ that accounts for the taxa abundance information, and
109 the unweighted UniFrac²³ that only models the information of presence or absence. Recently, a
110 generalized UniFrac distance metric²⁶ was developed to automatically appreciate the advantages of
111 weighted and unweighted UniFrac metrics and was shown to provide better statistical power to detect
112 associations between human health conditions and microbiome communities. GWAS based on a
113 microbiome distance matrix aims to identify host SNPs associated with microbiome composition.
114 Intuitively, the microbiome distances tend to be smaller for pairs of subjects with similar genotypic values
115 at the associated SNP. In addition, it is also of great interest to identify host SNPs that interact with an
116 environment factor to affect microbiome composition. Importantly, beta diversity is temporally more
117 stable compared with RA of taxa and alpha-diversity metrics based on the data from the Human
118 Microbiome Project²⁷ (data not shown), suggesting smaller power loss for a GWAS due to temporal
119 variability. To our knowledge, no statistical methods or software packages have been designed to
120 efficiently analyze microbiome GWAS data using distance matrices as phenotypes.

121 In this paper, we develop a statistical framework and a computationally efficient package,
122 microbiomeGWAS, for analyzing microbiome GWAS data. Our package allows the detection of host
123 SNPs with a main effect or interaction with an environment factor, i.e. host SNPs interacting with an
124 environment factor to affect the microbiome composition. We calculate the variance of the score statistics

125 by appropriately considering the dependence of the pairwise distances. Importantly, we show that the
126 score statistics have positive skewness and kurtosis due to the dependence in pairwise distances, which
127 makes the approximation of small P -values based on the asymptotic distribution too liberal, which easily
128 yields false positive associations. Resampling methods, e.g. bootstrap or permutation, are computationally
129 prohibitive for accurately approximating small P -values. We propose to improve the tail probability
130 approximation by correcting for skewness and kurtosis of the score statistics. Numerical investigations
131 demonstrate that our method provides a very accurate approximation even for $P=10^{-7}$. MicrobiomeGWAS
132 runs very efficiently, taking 36 minutes for analyzing main effects and 69 minutes for analyzing both
133 main and interaction effects for a study with 2000 subjects and 500,000 SNPs using a single core.
134 MicrobiomeGWAS can be freely downloaded at <https://github.com/lscnibb/microbiomeGWAS>.

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

143 Material and Methods

144 A score statistic for testing main effect

145 Suppose that we have a set of N subjects genotyped with SNP arrays. For notational simplicity, we
146 consider only one SNP with minor allele frequency (MAF) denoted as f . Our interest centers on testing
147 whether the genotype of the SNP is associated with microbiome composition. Let $g_n = 0,1,2$ represent
148 the number of the minor alleles for the n^{th} subject. We assume that the 16S rRNA gene of microbiota

149 from a target site (e.g., gut) has been sequenced for these samples. Let d_{ij} be the microbiome distance
 150 between the i^{th} and j^{th} subjects and \mathbf{D} be the distance matrix.

151 Intuitively, if the SNP is associated with the microbiome composition, the microbiome distances tend to
 152 be smaller for subject pairs with similar genotypic values, as is illustrated in Figure 1. For N subjects,
 153 $N(N - 1)/2$ pairs can be divided to three groups with genetic distance 0, 1 and 2. For example, a pair of
 154 subjects with genotype (AA, AA) or (BB, BB) has genetic distance 0; a pair of subjects with genotype
 155 (AA, BB) or (BB, AA) has genetic distance 2; all other pairs have genetic distance 1. Apparently, we
 156 expect the microbiome distance to be positively correlated with genetic distance for subject pairs.

157 We define $G_{ij} = |g_i - g_j|$ as the genetic distance for a pair of subjects (i, j) . We assume $d_{ij} = \alpha +$
 158 $\beta_M G_{ij} + \varepsilon_{ij}$ for all pairs of subjects. The score statistic for testing $H_0: \beta_M = 0$ (main effect) vs. $\beta_M > 0$
 159 is derived by maximizing $\sum_{i < j} (d_{ij} - \alpha - \beta_M G_{ij})^2$:

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

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

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

163 where

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

164 and

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

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

167 In analyses of real data, we typically have to adjust for covariates, including demographic variables and
168 principal component analysis (PCA) scores derived based on genotypes to eliminate potential population
169 stratification. Let $X_i = (x_{i1}, \dots, x_{iv})$ denote the v covariates for the i^{th} subject. We assume $d_{ij} = \alpha +$
170 $\beta_M G_{ij} + \sum_{t=1}^v w_t |x_{it} - x_{jt}| + \varepsilon_{ij}$. Define $d'_{ij} = d_{ij} - \hat{\alpha} - \sum_{t=1}^v \hat{w}_t |x_{it} - x_{jt}|$ with $(\hat{w}_1, \dots, \hat{w}_v)$ being
171 estimated under H_0 : $\beta_M = 0$. It is straightforward to verify that the score equation for β_M evaluated at
172 $H_0: \beta_M = 0$ is $S'_M = \sum_{i < j} d'_{ij} G_{ij}$. We can similarly derive the variance $Var_0(S'_M | \mathbf{D}')$ and the normalized
173 score statistic $Z'_M = S'_M / \sqrt{Var_0(S'_M | \mathbf{D}')}$. Here, \mathbf{D}' denotes the residue distance matrix with $(\mathbf{D}')_{ij} = d'_{ij}$.

174 **A score statistic for testing gene-environment interaction**

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

181 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
182 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
183 the variance $Var_0(S_I | \mathbf{D}')$ by accounting for the dependence in $(\Delta_{ij}, \Delta_{kl})$:

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

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

186 In Appendix B, we derive

$$\text{Cov}_0(S_M, S_I | \mathbf{D}') = \frac{N(N-1)}{2} \text{Cov}(G_{ij}, \Delta_{ij})\mu_2 + N(N-1)(N-2)\text{Cov}(G_{ij}, \Delta_{ik})\mu_3. \quad (6)$$

187 The correlation $\rho = \text{Cor}_0(Z_M, Z_I | \mathbf{D}')$ is calculated as $\rho = \text{Cov}_0(S_M, S_I | \mathbf{D}')/\sqrt{\text{Var}_0(S_M | \mathbf{D}')\text{Var}_0(S_I | \mathbf{D}')}$.

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

189 In Appendix C, we derive a statistic for jointly testing $H_0: \beta_M = \beta_I = 0$ vs. $H_1: \beta_M > 0$ or $\beta_I > 0$.

190 Briefly, the 2D plane is partitioned to four parts (Figure 2). The joint statistic is derived as

$$191 Q = \begin{cases} (Z_M, Z_I) \Omega^{-1} (Z_M, Z_I)^T & (Z_M, Z_I) \in A_1 \\ (w_1 Z_M + w_2 Z_I)^2 & (Z_M, Z_I) \in A_2 \\ (w_2 Z_M + w_1 Z_I)^2 & (Z_M, Z_I) \in A_3 \\ 0 & (Z_M, Z_I) \in A_4 \end{cases} \quad (7)$$

192 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
193 calculated as

$$194 P(Q > b^2) = q_1 P(\chi_2^2 > b^2) + q_2 P(N(0,1) > b) + q_3 P(N(0,1) > b), \quad (8)$$

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

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

197 Theoretic investigation suggests that the score statistics Z_M and Z_I have a positive skewness, which makes
198 the tail probability approximations based on the asymptotic distribution $N(0,1)$ unacceptably liberal
199 (Figures 3A and 3B). In a numeric example with skewness $\gamma = 0.2$, $P(Z > 5) = 2.9 \times 10^{-7}$ based on
200 $N(0,1)$, which is approximately two orders of magnitude more significant than $P=3.9 \times 10^{-5}$ based on 10^8
201 permutations. The significance inflation becomes worse for smaller P -values and larger skewness γ .
202 Similar but more tedious calculations suggest that both statistics have positive kurtosis, making the
203 approximation based on $N(0,1)$ even worse. One possible solution is to approximate tail probabilities

204 using permutations or bootstrap. However, these resampling methods are computationally prohibitive for
205 testing millions of common SNPs in a large-scale study.

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

210
$$P_0(Z > b) \approx e^{-b\xi_1 + (1+\sigma_1^2)\xi_1^2/2 + \gamma\xi_1^3/6 + \kappa\xi_1^4/24} \Phi(-\sigma_1\xi_1), \quad (9)$$

211 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
212 distribution function of $N(0,1)$. Correcting for skewness but ignoring kurtosis (i.e., assuming $\kappa = 0$)
213 leads to an approximation

214
$$P_0(Z > b) \approx e^{-b\xi_2 + (1+\sigma_2^2)\xi_2^2/2 + \gamma\xi_2^3/6} \phi(-\sigma_2\xi_2), \quad (10)$$

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

217 Given the distance matrix 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,
218 skewness decays much more slowly with sample size N than kurtosis (**Figures 3C and 3D**). Thus, even
219 for a large study with thousands of samples, correcting for skewness is necessary for accurately
220 evaluating tail probabilities. Importantly, both skewness and kurtosis highly depend on the MAF,
221 suggesting that the impact of skewness and kurtosis is different across SNPs with different MAF.
222 Numerical studies (**Figures 3C and 3D**) show that skewness and kurtosis are minimized when MAF=0.5
223 and maximized when MAF≈0.2-0.3.

224 Finally, we discuss how to approximate the tail probability of Q in (7) for testing $H_0: \beta_M = \beta_I = 0$ by
225 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

226 $E(w_1 Z_M + w_2 Z_I)^3$ and the kurtosis $E(w_1 Z_M + w_2 Z_I)^4 - 3$ and use (9) to approximate $P(w_1 Z_M +$
227 $w_2 Z_I > b)$. When $(Z_M, Z_I) \in A_1$, we first approximate their marginal P -values as p_M and p_I by (9), then
228 calculate the normal quantile $z_M = \Phi(1 - p_M)$ and $z_I = \Phi(1 - p_I)$. Because the correction primarily
229 impacts the tails of the distributions, the correlation between the two statistics will remain roughly
230 unchanged, i.e., $\text{cor}_0(Z_M, Z_I) \approx \text{cor}_0(z_M, z_I)$. Thus, when $(Z_M, Z_I) \in A_1$, the tail probability is
231 approximated as $P(\chi^2_2 > (z_M, z_I)\Omega^{-1}(z_M, z_I)')$.

232 **Results**

233 **Simulation results**

234 The main purpose of simulations was to investigate the type-I error of Z_M (for testing main genetic effect),
235 Z_I (for detecting SNP-environment interactions) and Q (for detecting either main genetic effect or SNP-
236 environment effect or both). Simulations were performed under different combinations of sample size,
237 MAF and microbiome distance matrices. To make simulations realistic, we used an unweighted distance
238 matrix of the fecal microbiome samples with the 16S rRNA V4 region sequences from the American Gut
239 Project (AGP). The OTU table rarefied to 10,000 sequence reads per sample, as well as metadata, was
240 downloaded from the AGP website. Samples with less than 10,000 sequence reads were excluded from
241 analysis. The weighted and the unweighted UniFrac distance matrices were generated in the Quantitative
242 Insights Into Microbial Ecology²¹ (QIIME) pipeline. Because antibiotics may substantially change
243 microbiome composition to generate outliers that may distort the null distribution, we excluded samples
244 with self-reported history of antibiotic usage within one month. After quality control, 1879 subjects
245 remained for analysis. In simulations, we randomly selected N samples for a given sample size N .

246 For each setting, the type-I error rates were evaluated based on 10^8 simulations under H_0 . For the
247 interaction test and the joint test, the binary environment factor had a frequency of 50% and was
248 simulated independent of the SNP. The type-I error rates are summarized in Table 1 for weighted UniFrac
249 distance matrix. The skewness and kurtosis are reported in Figures 3C and 3D. The statistics adjusted for

250 skewness and kurtosis have accurate type-I error rates while the statistics without adjustment have
251 unacceptably high type-I error rates. As sample size increases, the impact of skewness and kurtosis
252 decreases. However, even for a study with $N = 1000$, the type-I error rates are still seriously inflated.
253 The results for the unweighted UniFrac distance matrix and for MAF=0.5 are reported in **Table S1**.

254 **Software implementation, memory requirement and computational complexity**

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

264 **GWAS of microbiome diversity in adjacent normal lung tissues**

265 We applied our methods to a set of lung cancer patients of Italian ancestry in the EAGLE²⁸ study. All
266 subjects have germline genome-wide SNPs²⁹ and 16S rRNA microbiome data (V3-V4 region, Illumina
267 MiSeq, 300 paired-end) in histologically normal lung tissues from these patients. Here, the histologically
268 normal lung tissues were 1~5 cm from the tumor tissue. We performed a series of quality control steps to
269 filter out low quality sequence reads: average quality score <20 over 30bp windows, less than 60%
270 similarity to the Greengenes³² reference or identified as chimera reads using UCHIME³³. Sequence reads
271 were then processed by QIIME²¹ to produce relative abundances (RA) of taxa, two alpha diversity metrics
272 (observed number of species and Shannon's index) and beta-diversity metrics (unweighted and weighted

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

275 Out of the 147 subjects, 78 are current smokers, 8 are never smokers and 61 are former smokers. Because
276 of the small number of never smokers, we merged never and former smokers as non-current smokers. All
277 of the genetic association analyses were adjusted for sex, age, smoking status, and the top three PCA
278 scores derived based on genome-wide SNPs. Here, the top three PCA scores were selected controlling
279 population stratification because other PCA scores were unassociated with the distance matrices. We
280 included 383,263 common SNPs with $MAF \geq 10\%$ because rarer SNPs were expected to have no
281 statistical power given the current sample size. We first performed GWAS analysis using PLINK³⁴ to
282 identify SNPs associated with taxa with average RA greater than 0.1% or two alpha-diversity metrics. We
283 did not detect genome-wide significant associations with either main effects or gene by smoking
284 interactions.

285 Next, we performed GWAS analysis using unweighted and weighted UniFrac distance matrices as a
286 representation of eubacteria beta-diversity. The results for testing main effects are reported in **Figure 5**.
287 Results for testing joint effects (main effect and SNP by smoking status interaction) are reported in
288 **Figure S1**. Because of the small sample size, we observed large values of skewness and kurtosis with
289 magnitude varying with the MAF of the SNPs (**Figure 5A**). The score statistics based on the weighted
290 UniFrac distance matrix had a much larger skewness and kurtosis than did the unweighted UniFrac matrix.
291 **Figures 5B and 5C** report the quantile-quantile (QQ) plot of the logarithm of the association P-values for
292 the unweighted and weighted UniFrac distance matrices, respectively. For each distance matrix, we
293 produced QQ plots for P-values based on the asymptotic approximation and for P-values adjusted for
294 skewness and kurtosis. For both distance matrices, the QQ plots before adjustment strongly deviated from
295 the expected uniform distribution. Our adjustment eliminated the deviation. In addition, consistent with
296 the observation that the skewness and kurtosis were larger for the weighted UniFrac distance matrix, the
297 QQ plot deviated more for the analysis based on the weighted UniFrac distance. Note that the skewness

298 and kurtosis only affect the tail probabilities; thus, the inflation of the QQ plot is not reflected by the
299 genomic control lambda value³⁵ calculated as the median of P-values. In fact, lambda ≈ 1 for all four QQ
300 plots.

301 Without correcting for skewness and kurtosis, we identified three and six loci achieving genome-wide
302 significance ($P < 5 \times 10^{-8}$) for the unweighted and weighted UniFrac distance matrices, respectively
303 (**Figure 5D**). After correcting for skewness and kurtosis, no locus remained genome-wide significant
304 (**Figure 5D**), which was verified by 10^8 permutations. Importantly, skewness and kurtosis had a dramatic
305 effect on tail probabilities. Here, we use SNP rs12785513 as an example, which was identified as the top
306 SNP in both analyses. In the unweighted UniFrac analysis, $P = 4.4 \times 10^{-9}$ without adjustment and $P = 1.6 \times 10^{-6}$
307 after adjustment, a 364-fold inflation. The inflation was even larger for weighted UniFrac analysis
308 because of larger skewness and kurtosis (**Figure 5A**). In fact, $P = 3.4 \times 10^{-10}$ without adjustment and
309 $P = 3.5 \times 10^{-6}$ after adjustment, a 1000-fold inflation. Although these SNPs were not significant genome-
310 wide, they were the top SNPs from the current study. Thus, we report box-plots for each of these nine
311 SNPs (**Figure 5E**). As expected, in all box plots, microbiome distances tend to be larger in subject pairs
312 with greater genetic distance at these SNPs. These associations remain to be replicated in studies with
313 larger sample sizes.

314 Finally, we concentrated on the six common SNPs in four genomic regions reported to be associated with
315 lung cancer risk in GWAS of European subjects: rs2036534 and rs1051730 at 15q25.1³⁶⁻³⁹ (*CHRNA5-*
316 *CHRNA3-CHRN84*), rs2736100 and rs401681 at locus 5p15.33^{29,40} (*TERT/CLPTM1L*), rs6489769⁴¹ at
317 12p13.3 (*RAD52*), and rs1333040 at 9p21.3⁴² (*CDKN2A/CDKN2B*). The SNPs at 15q25.1 and 5p15.33
318 have the largest effect sizes for lung cancer risk based on the meta-analysis from the Transdisciplinary
319 Research in Cancer of the Lung (TRICL) consortium⁴²: OR=1.32 for rs1051730, OR=1.26 for rs2036534,
320 OR=1.13 for rs2736100, and OR=1.14 for rs401681. Rs3131379 at locus 6p21.33⁴⁰ (*BAT3/MSH5*) was
321 excluded because MAF=7.5%. No SNPs were significantly associated with taxa RAs or alpha-diversity
322 metrics after correcting for multiple testing (data not shown). However, association analysis based on the

323 UniFrac distance matrices provided evidence that these SNPs may be associated with the lung microbiota
324 (**Table 2**). Importantly, for both unweighted and weighted UniFrac analyses, all six SNPs had P -value <
325 0.5. These SNPs were independent except that rs2036534 and rs1051730 at 15q25.1 were weakly
326 correlated with $R^2=0.15$. A test combining six z-scores (Z_M) and adjusting for the weak correlation
327 yielded overall P-values 0.0033 and 0.011 for the unweighted and the weighted UniFrac distance matrices,
328 respectively. These results suggest that lung cancer risk SNPs were enriched for genetic association with
329 the composition of the lung microbiome. The results for testing interactions and joint effects are reported
330 in **Table S2**.

331 Discussion

332 We developed a software package, microbiomeGWAS, for identifying host genetic variants associated
333 with microbiome composition. MicrobiomeGWAS can test both main effect and SNP-environment
334 interactions. Importantly, we found that the score statistics had positive skewness and kurtosis and that
335 the tail probabilities evaluated based on asymptotic approximations were very liberal. We addressed this
336 problem by explicitly adjusting for skewness and kurtosis. MicrobiomeGWAS runs very efficiently and
337 takes only 36 minutes for testing main effects and 69 minutes for testing joint effects in a GWAS with
338 2000 subjects and 500,000 markers. Other statistical methods exist for testing the association of
339 microbiome distance matrices. **PERMANOVA**⁴³ is an extension of multivariate analysis of variance to a
340 matrix of pairwise distances and relies on permutations to evaluate significance. **MiRKAT**⁴⁴, a recently
341 proposed method based on kernel regression, takes hours for evaluating one association for 2000 subjects.
342 Neither is computationally feasible for analyzing a large-scale GWAS of microbiome.

343 Interactions of host genetic susceptibility with the microbiome have been postulated for many conditions,
344 including inflammatory bowel diseases^{45,46}, autoimmune and rheumatic diseases⁴⁷⁻⁵⁰, diabetes⁵¹, and
345 cancer especially of the colon⁵². All models of these host-microbiome interactions also note the critical
346 role of environmental factors including diet, smoking, drugs, and antibiotics and other medications⁵³.

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

351 We are working on two extensions for microbiomeGWAS: (1) jointly testing additive and dominant
352 effects and (2) testing genetic associations using many microbiome distance matrices. We have assumed
353 an additive effect model (**Figure 1**); however, several top SNPs in the EAGLE data suggest a dominant
354 effect (e.g. rs8083714 in **Figure 5E**). Thus, a statistic for jointly testing the additive and dominant effects
355 might be powerful for this scenario. The second extension is motivated by the fact that the power to
356 detect associations depends heavily on the choice of distance matrix. The recently developed generalized
357 UniFrac²⁶ (gUniFrac) defines a series of distance matrices to reflect different emphasis of using taxa
358 relative abundance information. gUniFrac has been shown to have a robust power for association studies²⁶.

359 Extending microbiomeGWAS to gUniFrac, however, requires solving two problems. First, the
360 computational complexity is proportional to the number of distance matrices analyzed for associations,
361 which can be addressed by implementing the algorithms using multithreading technology. Second, we
362 need to derive accurate analytic approximations to the association *P*-values by correcting for the multiple
363 testing introduced by many distance matrices. MiRKAT⁴⁴ has an option for using gUniFrac; however,
364 intensive permutations are required to evaluate *P*-values.

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

369

370

371 **Appendices**

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

373 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

374 $p_0 + p_1 + p_2 = 1$. We can also assume the Hardy–Weinberg equilibrium and characterize the

375 probabilities as the allele frequency: $p_0 = (1 - f)^2$, $p_1 = 2f(1 - f)$ and $p_2 = f^2$. Some algebra leads to

$$E(G_{ij}) = E|g_i - g_j| = \sum_{m,n \in \{0,1,2\}} p_m p_n |m - n| = 2p_0 p_1 + 2p_1 p_2 + 4p_0 p_2; \quad (A1)$$

$$376 \quad Var(G_{ij}) = E(G_{ij}^2) - E(G_{ij})^2 = (2p_0 p_1 + 2p_1 p_2 + 8p_0 p_2) - (2p_0 p_1 + 2p_1 p_2 + 4p_0 p_2)^2; \quad (A2)$$

$$Cov(G_{ij}, G_{ik}) = p_1(1 - p_1) + 4p_0 p_2(1 + p_1) - (2p_0 p_1 + 2p_1 p_2 + 4p_0 p_2)^2. \quad (A3)$$

377 Now consider $\Delta_{ij} = |g_i E_i - g_j E_j|$. When E_i is binary, $g_i E_i = 0, 1$ or 2 . Let $p'_t = P(g_i E_i = t)$. Then,

378 $E(\Delta_{ij})$, $Var(\Delta_{ij})$ and $Cov(\Delta_{ij}, \Delta_{ik})$ can be calculated similarly using (A1), (A2) and (A3).

379 **Appendix B:** Calculating $\rho = Cov_0(Z_M, Z_I | \mathbf{D}')$

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

$$Cov_0(S_M, S_I | \mathbf{D}') = Cov_0\left(\sum_{i < j} d'_{ij} G'_{ij}, \sum_{m < n} d'_{mn} \Delta'_{mn}\right) = \sum_{i < j, m < n} d'_{ij} d'_{mn} Cov(G_{ij}, \Delta_{mn}).$$

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

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

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

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

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

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

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

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

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

392 To simplify the optimization problem in (A6), we perform a linear transformation: $\mathbf{Y}^T = \mathbf{Z}^T \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and
 393 $\mathbf{v}^T = \boldsymbol{\xi}^T \boldsymbol{\Sigma}^{-\frac{1}{2}}$, where

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

394 Under this transformation, $\mathbf{Q} = \mathbf{Y}^T \mathbf{Y} - (\mathbf{Y} - \mathbf{v})^T (\mathbf{Y} - \mathbf{v})$ and can be interpreted as the difference of the
 395 square of two distances (**Figure S2B**). The original parameter space $\{(\xi_M, \xi_I): \xi_M \geq 0, \xi_I \geq 0\}$ is now
 396 transformed to $\{(\nu_1, \nu_2): \nu_2 \geq \theta \nu_1, \nu_2 \geq -\theta \nu_1\}$ with $\theta = \sqrt{(1-\rho)/(1+\rho)}$. Thus, the new parameter
 397 space is bounded by two lines represented by $\nu_2 \geq \theta \nu_1$ and $\nu_2 \geq -\theta \nu_1$. We partition the 2D plane into
 398 four parts (see **Figure S2B**), identify $\mathbf{v} = \text{arginf}_{\mathbf{v} \in A_1} (\mathbf{Y} - \mathbf{v})^T (\mathbf{Y} - \mathbf{v})$ and calculate Q :

$$399 Q = \begin{cases} Y_1^2 + Y_2^2 & (Y_1, Y_2) \in A_1 \\ (Y_2 - Y_1/\theta)^2 / (1 + \theta^{-2}) & (Y_1, Y_2) \in A_2 \\ (Y_2 + Y_1/\theta)^2 / (1 + \theta^{-2}) & (Y_1, Y_2) \in A_3 \\ 0 & (Y_1, Y_2) \in A_4 \end{cases} \quad (A8)$$

400 We now perform an inverse transformation using matrix

$$\boldsymbol{\Sigma}^{\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{\sqrt{1-\rho}}{\sqrt{1+\rho}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad (A9)$$

401 to return to the original parameter space. The four areas $\{A_1, A_2, A_3, A_4\}$ under the original space are in

402 **Figure 2** and **Figure S2C**.

403 Tedious calculations show that $(Y_2 + Y_1/\theta)^2/(1 + \theta^{-2}) = (w_2 Z_M + w_1 Z_I)^2$ with $w_1 = (\theta - 1/\theta)/2$
 404 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
 405 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, $P\{(w_1 Z_M + w_2 Z_I)^2 > b^2\} =$
 406 $P\{w_1 Z_M + w_2 Z_I > b\} = P\{N(0,1) > b\}$. This proves (8). The probabilities in (8) could also be
 407 calculated from **Figure S2B**: $q_1 = 1/2 - (\arctan \theta)/\pi$, $q_2 = q_3 = 1/4$.

408 **Appendix D:** Calculating skewness and kurtosis under H_0

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

$$E_0(S_M^3 | \mathbf{D}') = E_0 \left(\sum_{i < j} d'_{ij} G'_{ij} \right)^3 = \sum_{i < j, m < n, s < t} d'_{ij} d'_{mn} d'_{st} EG'_{ij} G'_{mn} G'_{st}.$$

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

$$412 E_0(S_M^3 | \mathbf{D}') = \binom{N}{2} \mu_4 EG'_{ij}^4 + \binom{N}{3} (\mu_5 EG'_{ij}^2 G'_{ik} + \mu_6 EG'_{ij} G'_{jk} G'_{ik}) + \binom{N}{4} (\mu_7 EG'_{ij} G'_{jk} G'_{kl} + \mu_8 EG'_{ij} G'_{ik} G'_{il}),$$

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

$$E_0(S_M^4 | \mathbf{D}') = E_0 \left(\sum_{i < j} d'_{ij} G'_{ij} \right)^4 = \sum_{i < j, m < n, s < t, x < y} d'_{ij} d'_{mn} d'_{st} d'_{xy} EG'_{ij} G'_{mn} G'_{st} G'_{xy}.$$

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

$$\begin{aligned}
 E_0(S_M^4 | \mathbf{D}) = & \binom{N}{2} \mu_9 EG_{ij}'^4 + \binom{N}{3} (\mu_{10} EG_{ij}'^3 G_{ik}' + \mu_{11} EG_{ij}'^2 G_{ik}'^2 + \mu_{12} EG_{ij}'^2 G_{jk}' G_{ik}') \\
 & + \binom{N}{4} (\mu_{13} EG_{ij}'^2 G_{jk}' G_{kl}' + \mu_{14} EG_{ij}' G_{jk}'^2 G_{kl}' + \mu_{15} EG_{ij}'^2 G_{ik}' G_{il}' + \mu_{16} EG_{ij}' G_{jk}' G_{ik}' G_{il}' + \mu_{17} EG_{ij}' G_{jk}' G_{kl}' G_{il}' + \mu_{18} EG_{ij}'^2 G_{kl}'^2) \\
 & + \binom{N}{5} (\mu_{19} EG_{ij}' G_{jk}' G_{kl}' G_{lm}' + \mu_{20} EG_{ij}' G_{ik}' G_{il}' G_{im}' + \mu_{21} EG_{ij}' G_{ik}' G_{il}' G_{lm}' + \mu_{22} EG_{ij}' G_{ik}' G_{lm}'^2) + \binom{N}{6} \mu_{23} EG_{ij}' G_{ik}' G_{lm}' G_{ln}'
 \end{aligned}$$

415 The constants (μ_9, \dots, μ_{23}) are dependent on \mathbf{D} and are provided in **Supplemental Data**.

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

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

418 We assume that $E_0 Z = 0$, $Var_0 Z = 1$, $\gamma = E_0 Z^3$ and $\kappa = E_0 Z^4 - 3$ under the original probability

419 measure P_0 . The tail probability $P_0(Z > b)$ for a large value of b is sensitive to the non-normality of Z ,
420 characterized by γ and κ . We define a new probability measure by embedding to the exponential
421 probability density

$$dP_\xi = \exp(\xi Z - \phi(\xi)) dP_0 \quad (A10)$$

422 where $\phi(\xi) = \log E_0 \exp(\xi Z)$ is the log moment generating function. Note that $\gamma = \phi'''(0)$ and $\kappa =$
423 $\phi''''(0)$. Because $E_0(Z) = 0$ and $Var_0(Z) = 1$, the Taylor expansion leads to $\phi(\xi) \approx \xi^2/2 + \gamma\xi^3/6 +$
424 $\kappa\xi^4/24$. Under P_ξ , we have

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

425 and

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

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

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

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

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

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

429 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

430 on the tail distribution of Z . Assuming $Z \sim N(b, \sigma^2)$ under P_ξ , we can verify that

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

431 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

432 as $P_0(Z > b) \approx e^{-b\xi + \frac{(1+\sigma^2)\xi^2 + \gamma \xi^3 + \kappa \xi^4}{2}} \Phi(-\sigma \xi)$, because $\phi(\xi) \approx \xi^2/2 + \gamma \xi^3/6 + \kappa \xi^4/24$ based on

433 the Taylor expansion. This proves (9).

434 If we correct skewness but assume kurtosis $\kappa = 0$, then $\phi(\xi) \approx \xi^2/2 + \gamma \xi^3/6$. We recalculate ξ by

435 setting $\kappa = 0$ in (A13) to derive $\xi = (\sqrt{1 + 2\gamma b} - 1)/\gamma$. This proves (10).

436

437 **Supplemental Data**

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

439 **Acknowledgements**

440 This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the
441 National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>). The project was supported by the
442 NIH Intramural Research program.

443

444 **Web Resources**

445 The URLs for data provide herein are as follows:

446

447 American Gut Project: <https://www.microbio.me/american gut/>

448 PLINK: <http://pngu.mgh.harvard.edu/~purcell/plink/>

449 QIIME: <http://qiime.org/>

450 EAGLE study: <http://eagle.cancer.gov/>

451

452

453 **References**

- 454 1. Turnbaugh, P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-4 (2009).
- 455 2. Morgan, X.C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and
456 treatment. *Genome Biol* **13**, R79 (2012).
- 457 3. Ahn, J. *et al.* Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* **105**, 1907-
458 11 (2013).
- 459 4. Goedert, J.J., Jones, G., Hua, X., Xu, X., Yu, G., Flores, R., Falk, R. T., Gail, M. H., Shi, J., Ravel, J.
460 and Feigelson, S. H. Investigation of the Association Between the Fecal Microbiota and Breast
461 Cancer in Postmenopausal Women: a Population-Based Case-Control Pilot Study. *J Natl Cancer
462 Inst.* **107(8)**(2015).
- 463 5. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor
464 environment. *Science* **345**, 1048-1052 (2014).
- 465 6. Wu, G.D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**,
466 105-8 (2011).
- 467 7. Tong, M. *et al.* Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's
468 disease risk polymorphism. *ISME J* **8**, 2193-206 (2014).
- 469 8. Knights, D. *et al.* Complex host genetics influence the microbiome in inflammatory bowel
470 disease. *Genome Med* **6**, 107 (2014).
- 471 9. McKnite, A.M. *et al.* Murine Gut Microbiota Is Defined by Host Genetics and Modulates
472 Variation of Metabolic Traits. *Plos One* **7**(2012).
- 473 10. Benson, A.K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait
474 shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* **107**, 18933-
475 8 (2010).
- 476 11. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789-99 (2014).
- 477 12. Davenport, E.R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS
478 One* **10**, e0140301 (2015).
- 479 13. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:
480 multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
- 481 14. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-
482 sequencing of 922 individuals. *Genome Research* **24**, 14-24 (2014).

- 483 15. Bell, J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in
484 HapMap cell lines (vol 12, pg R10, 2011). *Genome Biology* **12**(2011).
- 485 16. Shi, J. *et al.* Characterizing the genetic basis of methylome diversity in histologically normal
486 human lung tissue. *Nat Commun* **5**, 3365 (2014).
- 487 17. McVicker, G. *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human
488 Cells. *Science* **342**, 747-749 (2013).
- 489 18. Kilpinen, H. *et al.* Coordinated Effects of Sequence Variation on DNA Binding, Chromatin
490 Structure, and Transcription. *Science* **342**, 744-747 (2013).
- 491 19. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat Genet*
492 **43**, 565-9 (2011).
- 493 20. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a
494 founder population. *Nat Genet* **41**, 35-46 (2009).
- 495 21. Caporaso, J.G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat
496 Methods* **7**, 335-6 (2010).
- 497 22. Bray, J.R.A.C., J. T. An ordination of upland forest communities of southern Wisconsin. *Ecological
498 Monographs* **27:325-349**(1957).
- 499 23. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial
500 communities. *Appl Environ Microbiol* **71**, 8228-35 (2005).
- 501 24. Lozupone, C.A., Hamady, M., Kelley, S.T. & Knight, R. Quantitative and qualitative beta diversity
502 measures lead to different insights into factors that structure microbial communities. *Appl
503 Environ Microbiol* **73**, 1576-85 (2007).
- 504 25. Lozupone, C., Hamady, M. & Knight, R. UniFrac--an online tool for comparing microbial
505 community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, 371 (2006).
- 506 26. Chen, J. *et al.* Associating microbiome composition with environmental covariates using
507 generalized UniFrac distances. *Bioinformatics* **28**, 2106-13 (2012).
- 508 27. Gevers, D. *et al.* The Human Microbiome Project: a community resource for the healthy human
509 microbiome. *PLoS Biol* **10**, e1001377 (2012).
- 510 28. Landi, M.T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an
511 integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008).
- 512 29. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of
513 chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**, 679-91 (2009).
- 514 30. Tu, I.P. & Siegmund, D. The maximum of a function of a Markov chain and application to linkage
515 analysis. *Advances in Applied Probability* **31**, 510-531 (1999).
- 516 31. Siegmund, D. Sequential Analysis: Tests and Confidence Intervals. New York: Springer. (1985).
- 517 32. DeSantis, T.Z., Dubosarskiy, I., Murray, S.R. & Andersen, G.L. Comprehensive aligned sequence
518 construction for automated design of effective probes (CASCADE-P) using 16S rDNA.
Bioinformatics **19**, 1461-8 (2003).
- 519 33. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. UCHIME improves sensitivity and
520 speed of chimaera detection. *Bioinformatics* **27**, 2194-200 (2011).
- 521 34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage
522 analyses. *Am J Hum Genet* **81**, 559-75 (2007).
- 523 35. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
- 524 36. Hung, R.J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor
525 subunit genes on 15q25. *Nature* **452**, 633-7 (2008).
- 526 37. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404-6 (2008).
- 527 38. Thorgeirsson, T.E. *et al.* A variant associated with nicotine dependence, lung cancer and
528 peripheral arterial disease. *Nature* **452**, 638-42 (2008).

- 530 39. Amos, C.I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for
531 lung cancer at 15q25.1. *Nat Genet* **40**, 616-22 (2008).
- 532 40. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**,
533 1407-9 (2008).
- 534 41. Shi, J. *et al.* Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of
535 squamous cell lung carcinoma. *Cancer Discov* **2**, 131-9 (2012).
- 536 42. Timofeeva, M.N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis
537 of 14 900 cases and 29 485 controls. *Hum Mol Genet* **21**, 4980-95 (2012).
- 538 43. Anderson, M.J. A new method for non-parametric multivariate analysis of variance. *Austral
539 Ecology* **26**: , 32-46. (2001).
- 540 44. Zhao, N. *et al.* Testing in Microbiome-Proiling Studies with MiRKAT, the Microbiome
541 Regression-Based Kernel Association Test. *Am J Hum Genet* **96**, 797-807 (2015).
- 542 45. Leone, V.A., Cham, C.M. & Chang, E.B. Diet, gut microbes, and genetics in immune function: can
543 we leverage our current knowledge to achieve better outcomes in inflammatory bowel diseases?
544 *Current Opinion in Immunology* **31**, 16-23 (2014).
- 545 46. Huang, H., Vangay, P., McKinlay, C.E. & Knights, D. Multi-omics analysis of inflammatory bowel
546 disease. *Immunol Lett* **162**, 62-8 (2014).
- 547 47. Troncone, R. & Discepolo, V. Celiac disease and autoimmunity. *J Pediatr Gastroenterol Nutr* **59**
548 *Suppl 1*, S9-S11 (2014).
- 549 48. Yeoh, N., Burton, J.P., Suppiah, P., Reid, G. & Stebbings, S. The role of the microbiome in
550 rheumatic diseases. *Curr Rheumatol Rep* **15**, 314 (2013).
- 551 49. Sparks, J.A. & Costenbader, K.H. Genetics, environment, and gene-environment interactions in
552 the development of systemic rheumatic diseases. *Rheum Dis Clin North Am* **40**, 637-57 (2014).
- 553 50. Smith, J.A. Update on ankylosing spondylitis: current concepts in pathogenesis. *Curr Allergy
554 Asthma Rep* **15**, 489 (2015).
- 555 51. Nielsen, D.S., Krych, L., Buschard, K., Hansen, C.H. & Hansen, A.K. Beyond genetics. Influence of
556 dietary factors and gut microbiota on type 1 diabetes. *FEBS Lett* **588**, 4234-43 (2014).
- 557 52. Birt, D.F. & Phillips, G.J. Diet, genes, and microbes: complexities of colon cancer prevention.
558 *Toxicol Pathol* **42**, 182-8 (2014).
- 559 53. Marietta, E., Rishi, A. & Taneja, V. Immunogenetic control of the intestinal microbiota.
560 *Immunology* **145**, 313-22 (2015).

561

562

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

567

N	Z_M			Z_I			Q		
	$\alpha = 10^{-3}$	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
	200	3.7	23.0	187.3	3.1	15.8	105.5	4.6	33.0
	500	2.4	9.4	45.2	2.1	6.7	25.5	2.8	11.9
	1000	2.0	5.7	21.3	1.8	4.4	14.0	2.2	6.9
adjusted for skewness and kurtosis	100	1.0	1.2	0.7	1.0	1.1	0.6	1.0	1.5
	200	1.0	1.1	1.0	1.0	1.1	0.7	0.9	1.3
	500	1.0	1.1	1.3	1.0	1.0	0.9	0.9	1.0
	1000	1.0	1.0	1.2	1.0	1.0	0.8	0.9	1.0

568

569

570

571

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

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

574

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

576 **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

577 $Z_M \sim N(0,1)$, $Z_I \sim N(0,1)$ and $\text{cor}(Z_M, Z_I) = \rho$ under H_0 . Details are in **Appendix C**.

578 **Figure 3** Correcting tail probabilities for skewness and kurtosis. (A) The standard normal distribution
579 $N(0,1)$ and an approximately normal distribution with positive skewness. The skewness has big impact
580 when calculating the tail probability $P(Z > b)$ for a large value of b . (B) Numerical evaluation of tail
581 probability approximation for Z_M . We used the unweighted UniFrac distance matrix of 500 samples from
582 the American Gut Project (AGP). For each value of $b (> 0)$, we calculated P-values $P(Z_M > b)$ based on
583 $N(0,1)$, skewness correction, both skewness and kurtosis correction, and 10^8 simulations. (C) Skewness
584 depends on MAF of SNPs and the sample size of the study, calculated based on the weighted UniFrac
585 distance matrix in AGP data. (D) Kurtosis depends on MAF of SNPs and the sample size, calculated
586 based on the weighted UniFrac distance matrix in AGP data.

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

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

Figure 1

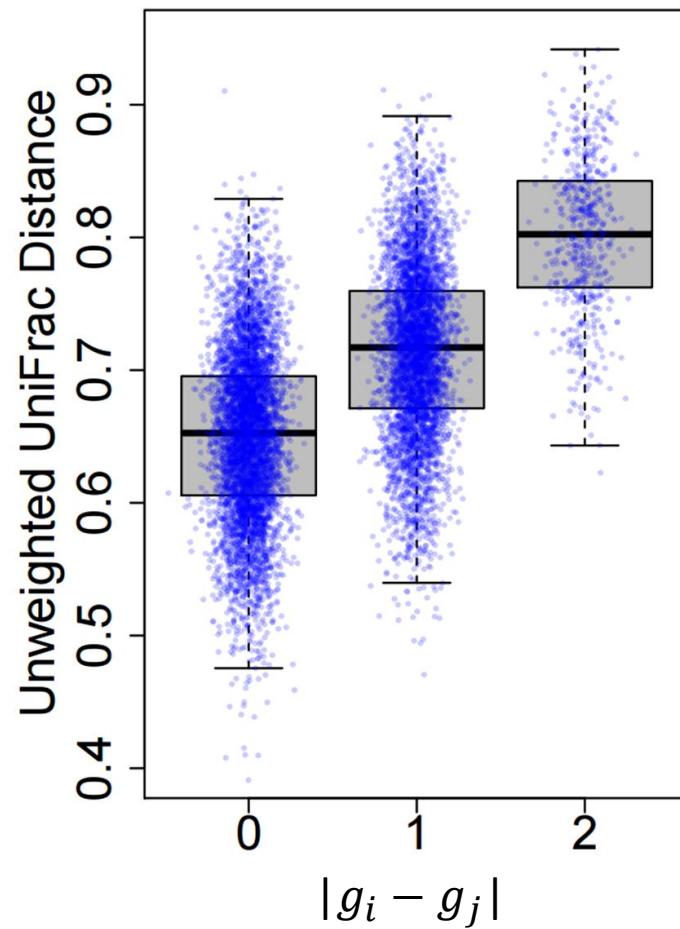


Figure 2

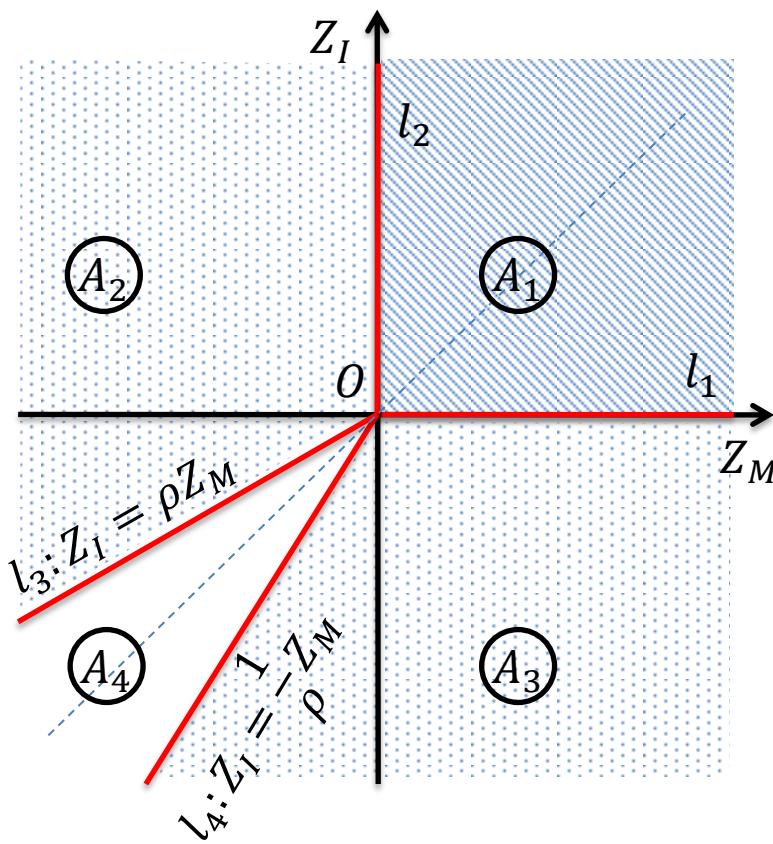


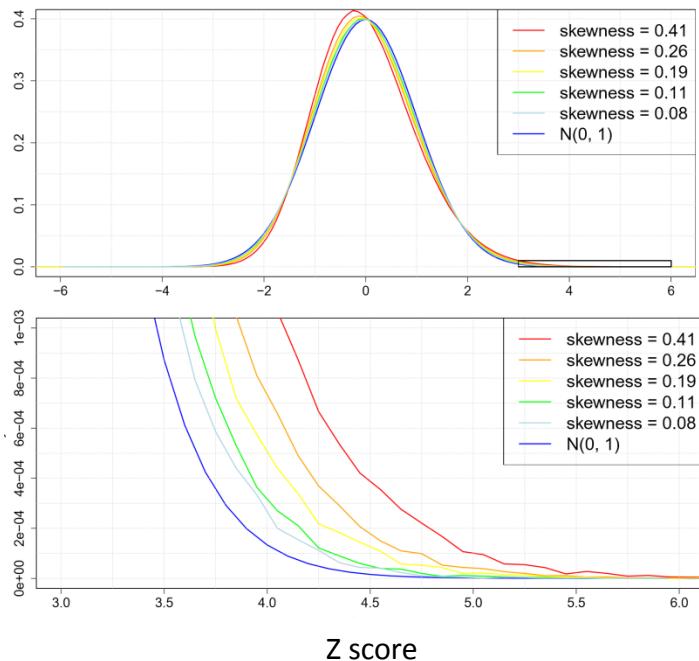
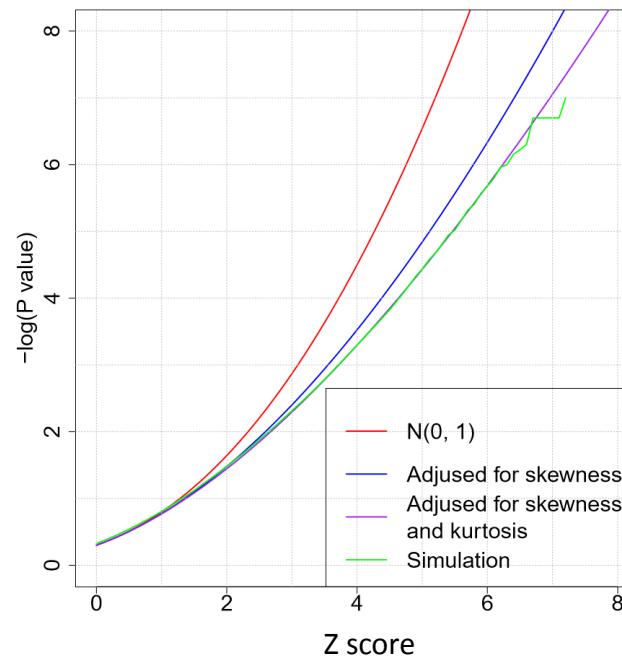
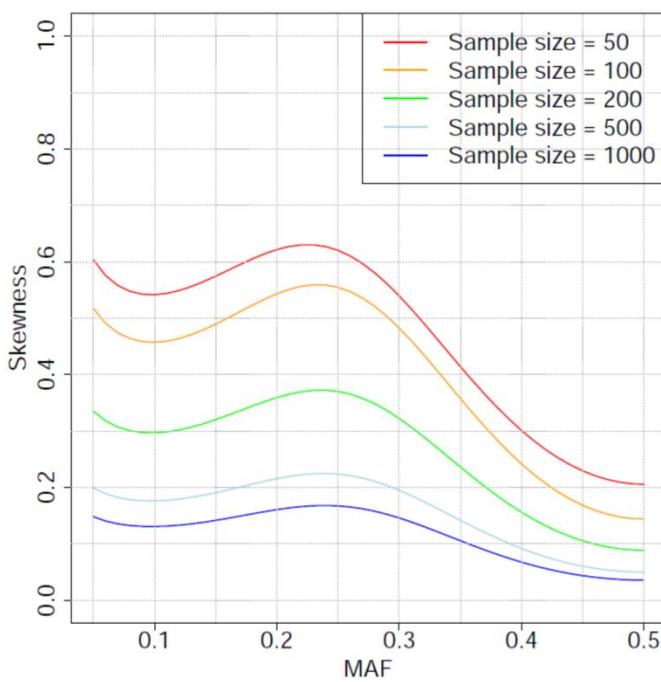
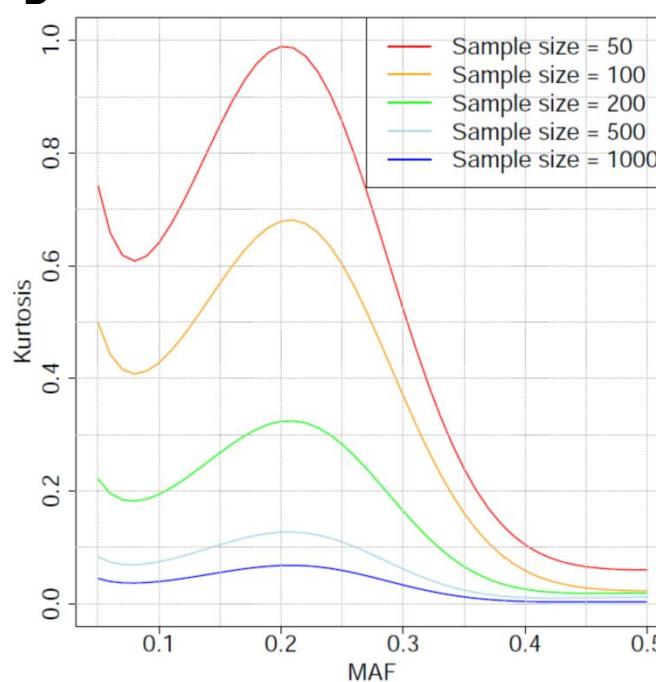
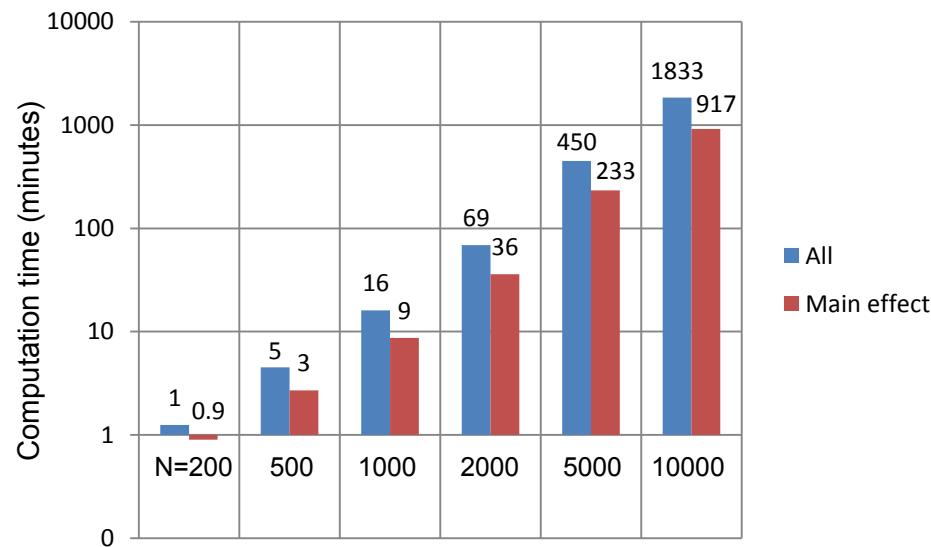
Figure 3**A****B****C****D**

Figure 4



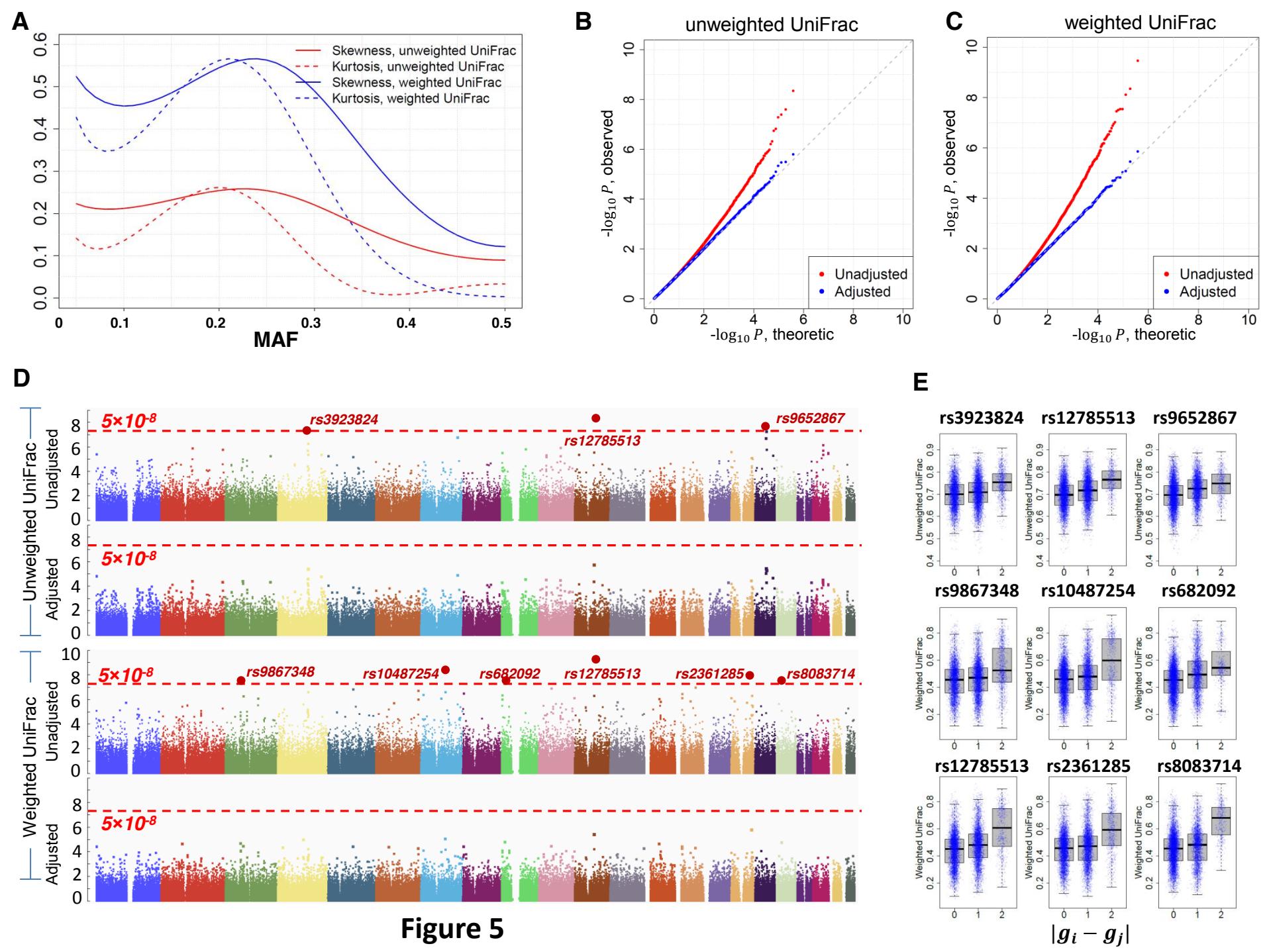


Figure 5