doi:10.1038/nature25973

Power Simulations

We performed extensive power simulations to demonstrate that the analyses carried out in our study are well powered. Our simulations indicate very high power for all experiments which test for association between bacterial and ancestral similarity, and between microbiome composition and ancestry proportions (where we define power as the proportion of simulations yielding P<0.05). Specifically, we considered several simulation scenarios, generated multiple synthetic phenotype vectors according to these scenarios, and tested for association between these vectors and genetic/ancestry (depending on the simulation scenario). We then estimated statistical power via the fraction of tests with P value < 0.05.

The simulation settings we considered are as follows:

1. **Simulating a phenotype whose distribution depends on ancestry proportions**. Here, we generated for every individual i a synthetic phenotype y_i according to the formulas:

$$y_i = c\sum_a a_i\beta_a + \boldsymbol{x}_i^T\boldsymbol{\gamma} + \epsilon_i$$
$$\beta_a \sim \mathcal{N}(0,\sigma^2)$$
$$\epsilon_i \sim \mathcal{N}(0,1-\sigma^2)$$
$$\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0},I).$$

Here, the summation is performed over ancestries, a_i is the ancestry proportion of individual i for ancestry a (the fraction of grandparents originating from this ancestry) after centering to obtain a zero mean, β_a is the coefficient of ancestry a, $\boldsymbol{x}_i$ is the vector of covariates of individual i, $\boldsymbol{\gamma}$ is a vector of covariate effects, ϵ_i encodes a residual term, $\sigma^2 \in (0,1)$ is the variance of β_a, and c is constant guaranteeing that y_i has a unit variance on average after regressing out the covariate effects. Hence, under this formulation σ^2 controls the fraction of the variance of y_i explained by ancestry (after regressing out the covariate effects).

We evaluated σ^2 values in the range $[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.999]$ and repeated each experiment 100 times, where in each experiment we (1) randomly sampled β_a, ϵ_i and $\boldsymbol{\gamma}$ values from their distribution; (2) generated y_i values based on the true ancestry proportion values a_i in our data; (3) regressed the covariates x_i out of y_i; and (4) tested if the phenotypes vector y is associated with ancestry via a Mantel test with Spearman correlation, where we tested for association between Euclidean ancestry distances of individuals i and j (given by $\sum_a[a_i - a_j]^2$) and phenotypic differences (given by $[y_i - y_j]^2$).

Steps (3) and (4) in the above procedure mimic the ancestry-microbiome association test described in the paper (**Figure 1e, Extended Table 1 middle column**), with the difference that here we test for association with a single phenotype y_i instead of a vector of bacteria. The proposed formulation is designed to simulate a genome wide association study, which enables evaluating the

model's power in a well-known setting with a minimal set of assumptions. Although the above model could be extended to simulate a multivariate vector instead of a scalar phenotype y_i, this would require introducing additional assumptions into the model which would detract from its generalizability.

The results indicate that our study is very well powered to identify a phenotype whose distribution is given by the formulas above, even for small values of σ^2 (**Supplementary Table 7**).

2. **Simulating a bacterial community based on genetic principal components.** Here, we generated for every individual i, and for each of the $K = 184$ bacterial species present in >5% of genotyped individuals, a vector of relative abundances $t_i^1, \ldots, t_i^K$ defined as follows:

$$t_i^1, \ldots, t_i^K \sim \mathrm{Dir}\left(\alpha_i^1, \ldots, \alpha_i^K\right)$$

$$\alpha_i^j = \begin{cases} \left[1 + \exp\left(-\sum_{m=1}^{5} P_i^m \beta_j^m\right)\right]^{-1} & i \le q \\ 0.5 & i \ge q \end{cases}$$

$$\beta_j^m \sim \mathcal{N}(0, b).$$

Here, α_i^j is the Dirichlet concentration parameter of taxon j in individual i (where larger values indicate a larger tendency to carry taxon j), m iterates over the top five genetic principal components, P_i^m is the m^{th} genetic principal component of individual i, β_j^m is the weight of the m^{th} genetic principal component with respect to taxon j, b is the variance of β_j^m, and q is a tunable parameter controlling the number of species affected by genetic principal components. Hence, larger values of q and of b indicate that a larger fraction of the microbiome composition is affected by genetic ancestry, which should be reflected in the PCos of a microbial β–diversity matrix.

We carried out experiments where we (1) generated bacterial taxa vectors according to the model above, using b values in the grid $[5,10,20]$ and q values corresponding to 0%, 1%, 25%, 50%, 75%, 95% and 100% of K; (2) computed the top PCos of a bacterial Bray-Curtis matrix; and (3) tested for a Spearman correlation between the top genetic PCs and the corresponding top microbiome PCos, with 100 experiments carried out for each evaluated value of q. We then computed P values for association via the standard asymptotic formulas for either a Spearman or a Pearson correlation of two multivariate random variables, and performed a multiple hypothesis correction for testing five different hypotheses via the Benjamini-Hochberg procedure. We measured power as the fraction of experiments with P value < 0.05 (after multiple hypothesis correction).

The results indicate excellent power for finding correlations between genetic PCs and bacterial PCos. (**Supplementary Table 8**).

3. **Simulating a bacterial taxon based on ancestry proportions.** Here, we generated for every individual i a synthetic species t_i whose relative abundance is given by:

$$t_i \sim \mathrm{Beta}(1 + k\boldsymbol{a}_i, 1 + k(1 - a_i)),$$

where a_i is the ancestry proportion of individual i for ancestry a, and $k \geq 0$ is a tunable parameter which controls the association strength. After generating t_i, we scaled all other bacterial species carried by the same individual so that their total relative abundance (including t_i) sums to unity. Hence, when $k = 0$ t_i is distributed uniformly between 0 and 1, whereas larger values of k induce a greater tendency for individuals of ancestry a to have a larger taxon abundance, and for individuals from other ancestries to have a smaller taxon abundance.

As before, we repeated the experiment multiple times, where each experiment is associated with different values of k and of an ancestry a. Specifically, we investigated values of k in the grid $[0,0.25,0.5,1,2,5]$, and repeated each experiment 100 times. As the above formulation simulates an association between a single taxon and a single ancestry, we used our machine learning model to estimate power. Specifically, in each experiment we trained a Ridge regression model to estimate the ancestry proportions of individuals based on their microbiome, and computed the coefficient of determination (R^2) via a 10-fold cross validation. This test mimics the ancestry proportion prediction test described in the results section of the main text. We computed approximate P values by generating a distribution of 1,000 R^2 values obtained under $k = 0$ (corresponding to the null hypothesis) for each ancestry a, and computing the fraction of null R^2 values greater than the one obtained in practice.

The results indicate excellent power in the majority of studied settings, and especially when a corresponds to Ashkenazi ancestry (**Supplementary Table 9**).

4. **Simulating a phenotype based on genetic kinship.** Here, we generated a vector of synthetic phenotype $\boldsymbol{y}$ according to the formula:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\gamma}, \sigma^2\boldsymbol{G} + (1 - \sigma^2)\boldsymbol{I}_n)$$
$$\boldsymbol{\gamma} \sim \mathcal{N}(0, \boldsymbol{I}_c)$$

where $\boldsymbol{X}$ is an $n \times c$ matrix of c covariates for n individuals, using the same covariates defined in the main text, $\boldsymbol{\gamma}$ is a vector of c effect sizes (often denoted as fixed effects), $\boldsymbol{G}$ is an $n \times n$ kinship matrix, $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{I}_c$ is the $c \times c$ identity matrix, and σ^2 controls the fraction of phenotype variance explained by genetic kinship (after regressing out the covariate effects).

We evaluated σ^2 values in the range $[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.999]$ and repeated each experiment 100 times, where in each experiment we (1) randomly

sampled a vector of effect sizes $\boldsymbol{\gamma}$ and a vector of phenotypes $\boldsymbol{y}$ from their distribution; (2) regressed the covariates $\boldsymbol{X}$ out of **y**; and (3) tested if the phenotypes vector $\boldsymbol{y}$ is associated with the genetic kinship matrix **G** via a Mantel test with Spearman correlation.

Steps (2) and (3) in the above procedure mimic the genetic kinship-microbiome association test described in the paper (**Figure 2a, Extended Data Table 1 right column**), with the difference that here we test for association with a phenotypes vector $\boldsymbol{y}$ instead of a matrix of bacterial abundances.

The results indicate that our study has moderate power to identify a phenotype whose distribution is given by the formulas above. (**Supplementary Table 10**). These results are in good agreement with **Figure 4a**, which shows that genetic kinship estimation has very large confidence intervals for samples with less than 2,000 individuals. Nevertheless, our analysis of the well-powered twinsUK data set of Goodrich *et al.* provides an estimate that the average estimated fraction of taxa variance explained by genetic kinship is 1.9%, or at most 8.1% under very liberal assumptions, as explained in the manuscript.

Statistical Aspects of the Microbiome-Association Index

The microbiome-association index (b^2) is a formal measure of the extent to which microbiome composition can predict a phenotype of interest. The value ranges between 0 and 1, with 0 indicating no predictive power and 1 indicating a fully deterministic prediction. b^2 was defined analogously to genetic heritability (commonly defined h^2). Both b^2 and h^2 are officially defined as the proportion of phenotypic variance that can be explained by the microbiome composition or by the genetic contents of an individual, respectively (where the term “explained variance” refers to statistical rather than causal explanatory power).

We now provide a formal description of some of the underlying assumptions behind the microbiome-association index. Denoting Y as a random variable encoding a phenotype of interest, and G, B, E as genetic, microbiome and environmental random variables than can be used to predict Y, respectively, we have:

$$\mathrm{var}(Y) = \mathrm{var}(G) + \mathrm{var}(B) + \mathrm{var}(E) + 2\mathrm{cov}(G,B) + 2\mathrm{cov}(G,E) + 2\mathrm{cov}(B,E).$$

In this work, we used the simpler term:

$$\mathrm{var}(Y) = \mathrm{var}(G) + \mathrm{var}(B) + \mathrm{var}(E),$$

where we estimated G via a polygenic risk score, B via the relative abundance of bacterial genes, and E as a normally distributed random variable. Hence, our derivation implicitly assumes $\mathrm{cov}(G,B) = \mathrm{cov}(G,E) = \mathrm{cov}(B,E) = 0$. We emphasize that G, B, E do not encode the genotypes, microbiome composition and the environment of an individual per se. Rather, they encode variables that are derived from the genotype, the microbiome composition and the environment of an individual, respectively, and can be used to predict Y. Consequently,

$\mathrm{cov}(B, E) = 0$ does not mean that the gut microbiome composition is uncorrelated with the environment. Rather, this equality means that the variables derived from the microbiome and from the environment for phenotype prediction are uncorrelated.

Our assumptions are similar to those commonly made in statistical genetics when estimating the genetic heritability of a phenotype. Specifically, heritability estimation is typically carried out via the equation $\mathrm{var}(Y) = \mathrm{var}(G) + \mathrm{var}(E)$ instead of the more general equation $\mathrm{var}(Y) = \mathrm{var}(G) + \mathrm{var}(E) + 2\mathrm{cov}(G, E)$. If we additionally assume that the contribution of each genetic variant to G is additive, then the heritability estimated from the first formula is typically called "narrow-sense heritability", whereas heritability estimated from the second formula is typically called "broad-sense heritability". Analogously, in this paper we estimate the narrow-sense microbiome-association index, rather than the broad-sense microbiome-association index.

The decision to estimate the narrow-sense microbiome-association index stems from several reasons. First, there is a large body of literature in statistical genetics demonstrating that it is very difficult to identify interactions terms in high-dimensional models, even if such interactions exist[1,2]. This is due to technical reasons – The first order term of the Taylor expansion of the underlying function can accurately approximate the effect of the interaction terms, which are originally encoded in the lower order terms. As our model is high dimensional (due to the ~1,300,000 bacterial genes used in the linear mixed model approach), the overall microbiome-association index estimate will likely be very similar to that of a model that explicitly encodes interactions, regardless of whether such approximations exist in reality.

Second, even if we ignore the above arguments, there is no established method to encode GxE, BxE and GxB interactions in a manner that is well accepted in the statistical genetics community. Hence, any attempt to encode these quantities is likely to be based on subjective considerations and subject to debate. We thus believe that our simplified model will facilitate reproduction of microbiome-association index estimation in different studies with varying study designs.

References

1. Mäki-Tanila, A. & Hill, W. G. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198,** 355–67 (2014).

2. Huang, W. & Mackay, T. F. C. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLOS Genet.* **12,** e1006421 (2016).