An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents

Daphna Rothschild[1,2,3,4,7], Sigal Leviatan[1,2,7], Ariel Hanemann[5,7], Yossi Cohen[5,7], Omer Weissbrod[6], Eran Segal[1,2]

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, 7610001 Rehovot, Israel

[2]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, 7610001 Rehovot, Israel

[3]Developmental Biology, Stanford University, Stanford, CA 94305, USA

[4]Department of Genetics, Stanford University, Stanford, CA 94305, USA

[5]DayTwo LTD, Tel Aviv Israel

[6]Epidemiology Department, Harvard T.H. Chan School of Public Health Boston, MA 02115, USA

[7]These authors contributed equally

*Correspondence: eran.segal@weizmann.ac.il

Summary

Numerous human conditions are associated with the microbiome, yet studies are inconsistent as to the magnitude of the associations and the bacteria involved, likely reflecting insufficiently employed sample sizes. Here, we collected diverse phenotypes and gut microbiota from 34,057 individuals from Israel and the U.S.. Analyzing these data using a much-expanded microbial genomes set, we derive an atlas of robust and numerous unreported associations between bacteria and numerous human traits, which we show to replicate in cohorts from both continents. Using machine learning models trained on microbiome data, we predict human traits with high accuracy across continents. Subsampling our cohort to smaller cohort sizes yielded highly variable models and thus sensitivity to the selected cohort, underscoring the utility of large cohorts and possibly explaining the source of discrepancies across studies. Finally, many of our prediction models saturate at these numbers of individuals, suggesting that similar analyses on larger cohorts may not further improve these predictions.

Introduction

The human gut microbiota is linked to metabolic disorders such as diabetes and obesity but these links are based on relatively small cohorts of several dozens or hundreds of individuals [1–8]. Although these studies reported many statistically significant

associations, many of these effects are either moderate or do not replicate in other works [9,10]. One such example is alpha diversity, for which there are contradicting reports regarding its association with different phenotypes. While microbiome diversity is mostly regarded as a positive indicator of health [8,11–14], other studies found that increased diversity is associated with microbiome instability [15,16]. Diversity was also shown to increase with age [17,18], but this association was not conclusive in other cohorts [19]. These discrepancies call for studying these questions across larger cohorts from diverse backgrounds. Indeed, in the field of genetics, large cohorts are required since many traits are known to be polygenic and to be affected by small effects from many variants [20,21]. Similarly, in the microbiome we expect that individual bacterial species may have a low abundance or mild associations with human phenotypes, necessitating large sample sizes. In addition, many bacterial species are present in only a relatively small fraction of the population such that the association between their abundance and traits can only be studied in large cohorts that have enough individuals that harbor them.

Apart from cohort size, there are other challenges in finding robust signals from the microbiome. One such challenge stems from the large number of genes that are shared between different bacteria through mechanisms such as horizontal gene transfer [22,23]. Such sharing causes many short metagenomic sequencing reads to map non-uniquely to multiple bacteria, making it difficult to estimate bacterial relative abundance. Several methods were devised to address this issue, .e.g., by mapping to genes that appear in a single copy and are unique to a single species [24]. However, these methods need to be applied anew every time that we wish to use a different reference genome set, as we wished to do here given the much expanded reference of bacterial species groups (SGBs) published in 2019 which added 3,796 new SGBs to the human microbiome catalog [25].

To address the above issues and with the aim of deriving robust microbiome associations, we used metagenomic sequencing to profile the gut microbiome of 34,057 individuals from both Israel and the U.S., for which we also obtained a rich set of phenotypes. We devised a novel algorithm for assessing bacterial relative abundances based on unique genetic elements, and applied it to the recent and much expanded SGB dataset of Pasolli et al. [25]. Using the relative abundances on this expanded genome set and much larger cohort, we identified numerous associations between

microbiome diversity and several human traits. We were also able to develop models that predict these traits using only microbiome data with high accuracy, as in the case of age (R^2=0.31). Notably, these associations replicate across continents, and models derived from the Israeli cohort generalize well to the U.S. cohort, so they are not specific to a certain environment .

By subsampling our cohort to typical cohort sizes used in other studies, we show that associations and predictions derived from smaller cohorts are highly variable and thus sensitive to the selected cohort, underscoring the need for larger cohorts in the microbiome field.

Results

Metagenome samples for 34,057 participants from two continents

We obtained gut metagenomic profiles from 30,083 and 3,974 individuals from Israel and the U.S., respectively, who submitted their sample to a consumer microbiome company and signed an appropriate consent form. Participants also answered questionnaires and provided self-reported phenotypic data and blood tests (e.g., age, gender, BMI and the diabetes marker HbA1C%, **Table 1**).

We randomly selected 90% of the samples from the Israeli cohort (n=27,075 samples) to be our discovery cohort on which we trained predictive models using cross-validation and set aside as independent test sets the remaining 10% of the Israeli cohort (“test1”, n=3,008 samples) and the entire U.S. cohort (“test2”, n=3,974 samples) (**Figure 1a-e, Table 1**). These test sets were only used once to evaluate the performance of the models developed on the discovery cohort.

To compute bacterial relative abundance, we used the representatives of the species-level genome bins (SGBs) classification of Pasolli et al. [25] , as they represent a greatly expanded set of genomes with thousands of new bacterial genomes that increase the number of mapped reads and allow better exploitation of metagenomic samples. We restricted ourselves to 3,127 SGBs that provide a good representation of species diversity (Methods) and used only reads that mapped uniquely to a single SGB representative. We developed a method for estimating the relative abundance of each SGB in every sample (Methods). Our method is based on examining only reads that

map uniquely to a single SGB, since when using unique mappings we expect uniform coverage across SGB genome bins that have the same number of unique positions. This property allows robust estimation of relative abundances, as coverage across the genome bins depends linearly on the SGB's relative abundance. The mean relative abundances of the different species are the same in the two Israeli cohorts but are somewhat different than in the US cohort (**Figure 1f-g**).

Microbiome diversity increases with age and associates with metabolic parameters

We first examined the association of microbiome diversity and human phenotypes since the literature is conflicted even on these basic associations. To this end, we computed alpha diversity using the species level Shannon index and ranked individuals by deciles of alpha diversity (**Figure 2a**). When comparing the top decile and the bottom decile of alpha diversity, we found that HbA1C%, BMI, fasting glucose and fasting triglycerides are significantly higher in the bottom decile while age and HDL cholesterol are significantly lower in the bottom decile (P-value < 1e-16 after FDR correction, Mann Whitney rank-sum test), including a trend across deciles (**Table S1-S6**). Similarly, examining alpha diversity as a function of these traits, we found significant correlations between alpha diversity and each of these traits (**Figure 2b**). Notably, these associations were consistent in both the Israeli and U.S. cohorts (**Figure 2b**), even though the Israeli cohort has significantly higher alpha diversity values (mean 7.3±0.0.77 vs. 7.18±0.67, P-value<10^{-40} ,Mann Whitney rank-sum test). The higher diversity of the Israeli cohort persisted even when subsampling the Isareli cohort to match the U.S. cohort on age and BMI (Methods, **Table S7**).

Microbiome-phenotype associations are consistent across continents

We previously employed linear mixed models to estimate the fraction of phenotypic variance that can be inferred from microbiome composition, termed *microbiome-association-index* (b^2) [26]. Our previous estimates were based on a cohort of 715 individuals and therefore had wide 95% confidence intervals, we revisited these estimates for our two new and larger cohorts. We estimated explained-variance based on alpha-diversity alone (**Figure 3a,** Methods), and based on the full species relative abundances (**Figure 3b,** Methods). Notably, our new estimates agreed well with our previous findings (**Table S8**), but our much larger cohort of 30,083 Israeli individuals

has substantially narrower 95% confidence intervals (**Table S9**). We found that microbiome composition strongly associates with self-reported diabetes (b^2=52%), age (b^2=28%), HbA1C% (b^2=15%), fasting blood glucose (b^2=13%), BMI (b^2=11%), fasting triglyceride (b^2=9%), HDL cholesterol (b^2=6%) and smoking status (b^2=6%). In contrast, the blood levels of thyroid-stimulating hormone (TSH), albumin and clotting (measures by International Normalized Ratio INR) were not significantly associated with the microbiome in our cohort. Notably, b^2 estimates from our U.S. cohort of 3,974 individuals were consistent with those derived from the Israeli cohort (Pearson correlation R=0.75, P-value <0.001) but had wider confidence intervals (**Figure 3b, Table S9,S10**). Although as expected, the variance explained by the full relative abundance matrix (our b^2 estimates) was higher than that explained by alpha diversity alone, these two microbiome features highly agreed (Pearson correlation R=0.52, P-value <0.03).

Different traits are accurately predicted by microbiome composition

We next asked whether various traits can be accurately predicted based only on microbiome composition. We compared two models; a linear model (with ridge regression regularization) and gradient boosted decision trees (GBDT) (Methods). Both models used species relative abundances as input. Our models obtained significant predictions for many traits (**Figure 4a,b**) such as age (R^2=0.35 for linear regression, R^2=0.31 for GBDT, for 10 fold cross validation on train IL samples), gender (AUC=0.64, 0.78), HbA1C% (R^2=0.24, 0.26) and BMI (R^2=0.15, 0.15).

We obtained significant predictions even when performing the analyses separately for each gender, with the exception of height which was significantly predicted (R^2=0.13) in the entire cohort but not in the gender-separated predictions, indicating that its predictions were driven by the prediction of gender. Since metformin, the most common drug used to treat patients with type2 diabetes, is known to affect microbiome composition, we also evaluated the performance of an HbA1C% predictor only on participants who did not report taking metformin and obtained equivalent performance (R^2=0.19 GBDT).

The linear models are attractive since their accuracy was almost similar to boosting decision trees and they are easier to interpret. However, boosting trees performed better across 11 of 12 phenotypes (age being the exception) that had significant

predictions(overall mean R^2 improvement of 0.02+/-0.011, **Figure S1**) suggesting that non-additive interactions between different bacteria are predictive of several traits. As additional evidence for the importance of non-additive interactions among bacteria in predicting traits, for both HbA1C% and BMI the R^2 of the GBDT predictions on held-out subjects was higher than the estimated b^2 for these traits (**Figure 3b**). The b^2 estimation used linear mixed models to estimate the fraction of variation predicted by the microbiome, which does not include any non-linear interaction.

We investigated if the predictive power of the microbiome is mediated through age and sex, since some of the above traits such as HbA1C% are known to increase with age [27] . We found that the microbiome composition predicted age with high accuracy (R^2=0.31), and age and gender alone predict HbA1C% with R^2=0.20 and BMI with R^2=0.02 (GBDT, **Figure 4c-e**). Therefore, we asked whether microbiome-based predictions of HbA1C% and BMI are mediated entirely by its ability to predict age and gender, or whether it carried additional predictive power specific to these traits. We found that adding microbiome to age and gender to the GBDT model significantly improved the predictions of both HbA1C% (from R^2=0.20 to 0.36, **Figure 3c**) and BMI (from R^2=0.03 to 0.18, **Figure 3c**), demonstrating that some of the association between microbiome and these traits is not mediated through age and gender.

We also evaluated the accuracy of our above models, derived only based on the Israeli training set, on our two independent and held out cohorts from Israel and the U.S. and found that they all had highly significant predictions (**Figure 4f**), thereby validating the robustness of our models. We note that prediction accuracy for age and HbA1C% was lower in the U.S cohort, which may be explained by differences in the microbiome composition between the IL and U.S. cohorts and by lower age and HbA1C% levels in the U.S. cohort.

Finally, to examine the importance of cohort size on prediction accuracy, we applied our above prediction pipeline to different random subsamples of our training cohort, ranging from a few hundreds of subjects to 24,000 (Methods). We found that prediction accuracy increases with cohort size (**Figure 4g-i**) and does not saturate even with a cohort of 1,000 individuals. For age, we observed an almost two-fold increase in the R^2 (from 0.18 to 0.30) when increasing the cohort from 1,000 to 12,000 individuals. For

cohorts of hundreds of individuals the standard deviation of the predictions was high, as in the case of HbA1C% for which different subsamples of 200 individuals can reach both R^2=0.4 and R^2=0.0 as likely outcomes (within 2 standard deviations). Together, these results highlight the need for obtaining large cohorts of microbiome as is known to be the case in the field of human genetics.

An atlas of bacterial species that robustly correlate with age, HbA1C% and BMI

We next sought to identify which individual bacterial species are responsible for driving the predictions of our models for age, HbA1C% and BMI, since these traits were predicted with the highest accuracy. We found many bacterial species that exhibited highly significant correlations to these traits (**Figure 5a-c**, e.g., 640, 454, 779 bacteria out of the top 1345 occuring bacteria had significant Spearman correlation, P-value<0.05 after Bonferroni correction for age, HbA1C% and BMI respectively). Moreover, the spearman correlation of the bacterial abundances with these traits was in good agreement between the Israeli and U.S. cohorts (**Figure 5a-c**, R=0.58, 0.57, 0.75 for age, HbA1C% and BMI, P-value<10^{-39}). Notably, the 3 bacteria most strongly associated with BMI in both cohorts included a bacterial species from the Eubacteriaceae family that was only recently assembled and that has no genome in public repositories (unknown SGB, **Figure 5a-c**). Again, we subsample the cohorts to smaller cohort sizes and observe that large cohorts are necessary in order for results to replicate (**Figure 5d-f**).

As a useful resource for the community, we compiled our results into an atlas of summary statistics for all bacterial species and top predicted phenotypes (**Table S11-S16**). For each species we report bacterial associations to human phenotypes based on bacterial log relative abundances. Specifically, we report the Spearman correlation coefficient and P-value, the Pearson correlation coefficient and P-value, the coefficient in the linear model (trained with Ridge regularization), and bacterial feature importance in the GBDT model using the feature attribution framework of SHapley Additive exPlanations[28] (SHAP). In genetics, summary statistics of single nucleotide polymorphisms are widely used to generate polygenic risk scores which were shown to be predictive of disease [29,30]. Similarly, researchers can now use our resource to generate microbiome-based predictions of phenotypes in their datasets by extracting

our reported bacterial regression coefficients and multiplying them by the log of the relative abundances of the corresponding species in their dataset.

Discussion

In this study, we collected the largest cohort to date of metagenomic samples and phenotypic data from two continents, and analyzed it using a much expanded set of reference microbial species. Together, this allowed us to identify highly robust associations between gut microbiome composition and phenotypes, which replicate in both cohorts. We compiled these robust associations into an atlas that can be used by the community to derive trait predictions on smaller datasets, akin to the use of summary statistics in the field of genetics. We show that a large fraction of the variance of several traits such as age, HbA1C% and BMI can be accurately predicted by both linear models and boosting decision trees models. These predictions replicate across continents and there is also high agreement in the set of individual bacterial species that associate with these traits in the Israeli and U.S. cohorts.

When sub-sampling our large cohort into smaller sized cohorts, we found that even cohorts of 1,000 individuals have significantly lower average accuracy of associations between bacteria and phenotypes. Models derived from different sub-samples of smaller cohort sizes display high variability both in the set of bacteria that associate with each trait and in prediction accuracy. These results may explain the relatively low agreement that exists across studies in the set of bacteria associated with different traits and conditions, and they call for employing larger cohort sizes in microbiome studies.

Using an expanded reference set allowed us to study many bacterial species for the first time and to identify novel associations for them. Notably, even among the top associated bacteria we found unnamed bacteria that are prevalent and appear in hundreds and sometimes thousands of individuals from our cohort. These findings emphasize the importance of expanding the reference set of the human microbiome even further, and suggest that such newly identified species may have strong associations with important host phenotypes.

Overall, by combining larger microbiome cohorts and expanded bacterial genome references we robustly characterize bacterial links to many important health parameters, serving an important first step towards unraveling the causal links and mechanisms by which bacteria affect host phenotype.

Methods

Microbiome sample collection, processing and analysis

Participants provided a stool sample using an OMNIgene-Gut stool collection kit (DNA Genotek), and processed according to the methods described in Mendes-Soares et al. [31]: Genomic DNA was purified using PowerMag Soil DNA isolation kit (MoBio) optimized for Tecan automated platform. Illumina compatible libraries were prepared as described in [32], and sequenced on an Illumina Nextera 500 (75bp, single end), or on a NovaSeq 6000 (100bps, single end). Reads were processed with Trimmomatic [33], to remove reads containing Illumina adapters, filter low quality reads and trim low quality regions; version 0.32 (parameters used: -phred33 ILLUMINACLIP:<adapter file>:2:30:10 SLIDINGWINDOW:6:20 CROP:100 MINLEN:90 for 100bps reads, CROP:75 MINLEN:65 for 75bps reads). Reads mapping to host DNA were detected by mapping with bowtie2 [34,34,35] (with default parameters and an index created from hg19) and removed from downstream analysis.

All samples were subsequently downsampled to a depth of 5M reads. Samples with fewer reads were removed from further analysis, leaving us with a reduced sample of participants that was used for downstream microbiome analyses.

Relative abundance estimation of SGBs

The bacterial reference dataset for relative abundance estimation is based on the representative assembly of the species-level genome bins (SGBs) and genus-level genome bins (GGBs) defined by Pasolli et al. [25]. By construction, all assemblies in each SGB are at high average nucleotide identity with one another, and the representative was chosen to be the best quality assembly amongst them.

Out of the 4,930 human SGBs (associated with various body sites), we chose to work with 3,127 SGBs, which were characterized by either belonging to a unique genus or with at least 5 assemblies to justify having a new SGB. We employed this restriction, since we noticed that the cutoff threshold used by Pasolli et. al. to cluster assemblies

into SGBs resulted in small groups with little nucleotide difference from a large nearby SGB is artificially split to a new SGB.

Abundance was calculated by counting reads that best matched to a single SGB of the set. In order to avoid sample reads which may be assigned to more than one SGBs (which might mislead us to believe an SGB appears in sample when it actually does not), we created a mapping of all 100/75-bps reads which are unique to a single of these representatives. We divided each representative genome assembly to consecutive windows such that each window includes 100 unique 100/75-bp reads (unique-100-bins). Since different proportions of reads are unique in different areas of the assembly, these windows are not of constant length, but the number of sample reads expected to uniquely map to them should be constant.

We used bowtie2 [35] to map samples from our cohort versus an index built from the set of representatives of the SGBs (demanding all mappings of length 100/75 to score -40 or above). When analysing the mapping, we looked only at reads whose best map is unique (thus mapped to a location which is unique in the set of representatives). We count the number of reads uniquely mapped to each window of each SGB.

To assess the cover of each SGB, we first choose a window size, which is a multiple of the original unique-100-bins, for which the average cover is at least 20 reads. Next, we sum the number of reads in this enlarged-window, and test the distribution of covers over the windows.

Finally, we take the dense mean of that distribution [36], in order to avoid our cover estimation being biased by a relatively small part of the reference which is highly covered (may come about from plasmids or horizontal transfer which was not identified in the uniqueness process since it did not appear in any other representative) or lowly covered (since this is a representative of an SGB, a strain present in our sample may not include all parts of the representative). When the dense 50% of the cover distribution starts above 0 we conclude the SGBs exists in the sample, and we estimate its relative abundance. The cover estimation for each SGB is the dense mean cover of its representative, normalized by the enlarged-window size.

The relative abundance estimation is the cover divided by the sum of the covers of all representatives we concluded exist in this sample.
Code of the algorithm is provided in github:
https://github.com/erans99/UniqueRelativeAbundance

Cohort matching

We subsampled the IL cohort to match the US cohort on age and BMI using MatchIt package from CRAN repository for r.

Alpha diversity explained variance

We calculated the alpha diversity explained variance by regressing out gender and age from each phenotype, and then using ordinary least squares modeled the phenotype by alpha diversity. To get confidence intervals, we bootstrapped the data 10,000 times.

Microbiome- association- index

We calculated b^2 estimates using linear mixed models as was previously described [26] We used age and gender as fixed effects covariates, and built a microbiome genetic-relationship-matrix, using our developed SGB based relative abundances. The b^2 calculation assumes that the phenotype distributed normally, we removed sample outliers from the IL and US cohorts using the same thresholds (removing less than 5% of individuals Table S17). To account for differences between the population and study prevalence of binary traits, we applied the correction of Lee et al. [37] which has been shown to provide a lower bound on the fraction of explained variance [38]. We also provide uncorrected estimates in Table S18. Phenotype distributions of blood SGPT levels were far from normally distributed and were not estimated.

Phenotypes prediction

We used the gradient boosting trees regressor from Xgboost [39] as the algorithm for the regression predictive model for different phenotypes. We used the gradient boosting trees classifier from Xgboost as the algorithm for the classification predictive model for phenotypes with binary values. All hyperparameters of the xgboost were fitted based only on cross validation of the train set.

The parameters of the predictors when using microbiome features were: colsample_bylevel=0.075, max_depth=6, learning_rate=0.0025, n_estimators=4000, subsample=0.6, min_child_weight=20. These parameters were used for regression as well as classification.

The rest of the parameters had the default values of Xgboost.

For the Ridge linear regression we used the RidgeCV from the scikit-learn package. The parameters used for the regressor were:
alphas=[0.1,1,10,100,1000], normalize=True. The rest of the parameters were the default. The input to the Ridge linear regression was log transformed SGB abundance.

For binary phenotypes SGD classifier from the scikit-learn package was used, with default parameters (L2 normalization).

When using microbiome features for the prediction, only the top 1345 occuring SGBs were used, i.e., the SGBs that were found in the highest number of samples, to avoid overfitting on rare SGBs.

Calculating prediction accuracy as a function of cohort size

For cohort size n (for n=200, 500, 1000, 2000, 3000, 4000, 6000, 8000, 12000, 16000, 20000, 24000; for prediction of HbA1c the maximum size was 16000) we repeated the following process 10 times: we randomly selected a subset of n samples, ran 10 fold cross validation of the prediction and listed the mean and standard deviation of each fold. By repeating the procedure 10 times we received the mean and standard deviation of the prediction accuracy estimate.

Table 1

	Israel Train	Israel Test	U.S. test
Age	55±14	55±14	52±13
Gender M:F	11,019: 16,056	1,129:1,779	1,678:2,296
BMI	28±5 (n=27,018)	28±5 (n=3,003)	27±6 (n=3,951)
Blood HbA1C%	5.8±0.8 (n=19,457)	5.8±0.8 (n=2,181)	5.5±0.66 (n=2,292)
Fasting blood glucose (mg/dl)	103±25 (n=21,613)	104±25 (n=2,405)	98±23 (n=2,291)
Fasting blood triglycerides (mg/dl)	132±78 (n=22,149)	134±82 (n=2,484)	107±71 (n=2,431)
Fasting blood HDL cholesterol (mg/dl)	53±18 (n=22,391)	53±17 (n=2,507)	61±21 (n=2,546)

Figures

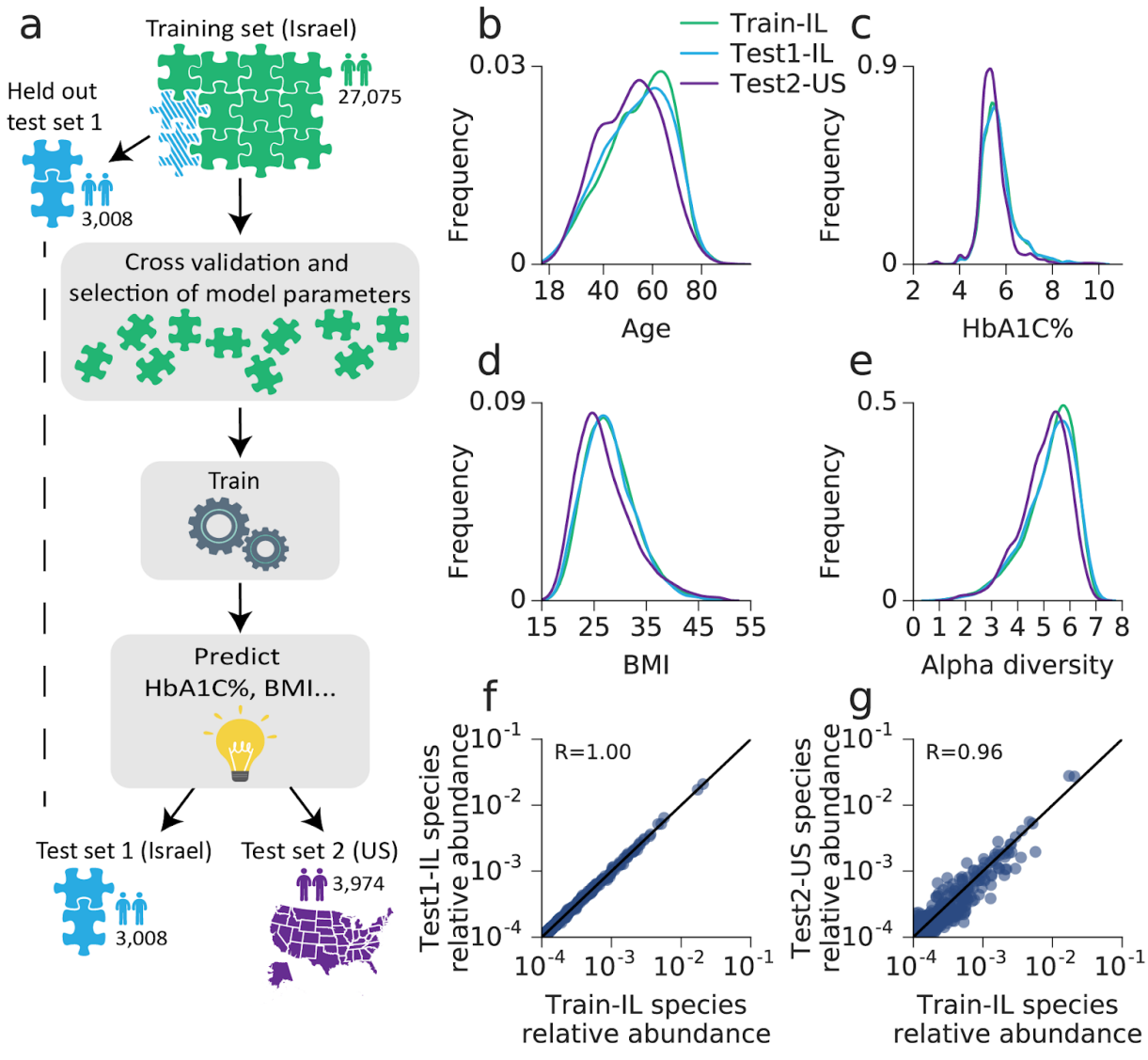

Figure 1: Cohort description and model prediction scheme.

(a) Illustration of cohorts and machine learning process. A training set of 27,075 individuals was randomly selected out of 3,083 Israeli individuals and was used for model parameter selection using 10 fold cross validation and microbiome, age and gender features. For each phenotype the selected model was trained on the 27,075 training samples and then tested on both the held out 3,008 samples of the Israeli population and a separate U.S. test cohort of 3,974 individuals.

(b) Distribution of age in the 3 cohorts, training test1 and test2.
(c) - (e) Same for HbA1C%, BMI and alpha diversity.
(f) A scatter plot comparing the mean log relative abundance of each species, in the Israeli training cohort vs. the Israeli test1 cohort.
(g) Same as (f) in the Israeli training cohort vs. the US test2 cohort.

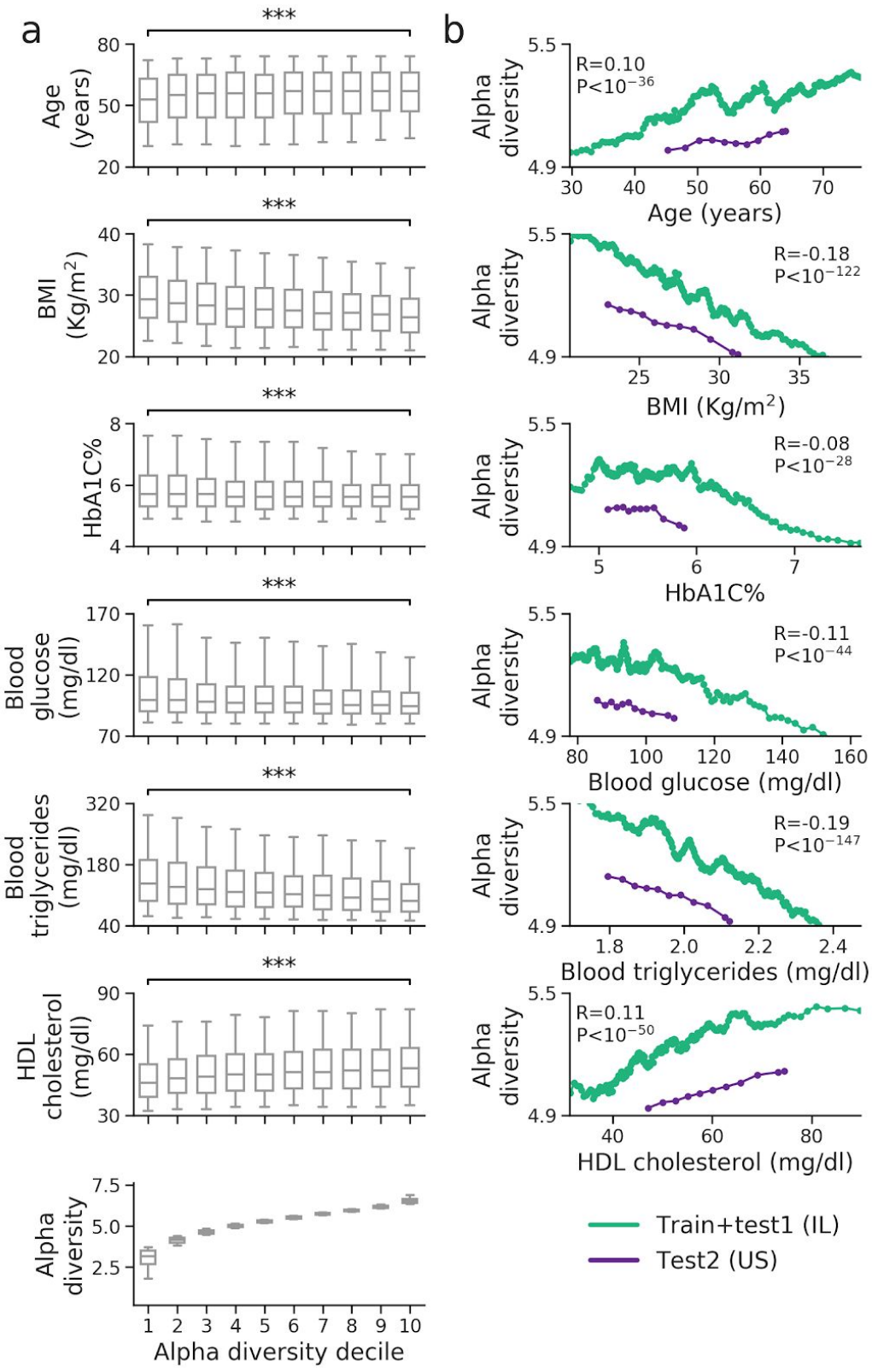

Figure 2: Species level Shannon alpha diversity significantly associates with many phenotypes.

(a) A box-plot of the distribution of phenotype values, for each of 10 deciles of Shannon alpha diversity. Phenotype values in the first and last deciles of alpha diversity are compared using Mann-Whitney rank-sum test where *** signifies P value< 10^{-16} after FDR correction. Boxes correspond to 25-75 percentile of the distribution and whiskers bound percentiles 5-95.

(b) Running average of alpha-diversity (y-axis) for the combined training and test1 cohort (green curve), and for the separate test2 cohort (purple curve), ordered by the phenotype values. For the larger Israeli cohort the average is on 1000 individuals with shift of 100 individuals; for the smaller U.S. cohort the group size and shift were chosen to obtain 10 points and the shift was 10% of the group size. The Pearson correlation and P-value shown are of the Israeli cohort, and are calculated on individual level data.

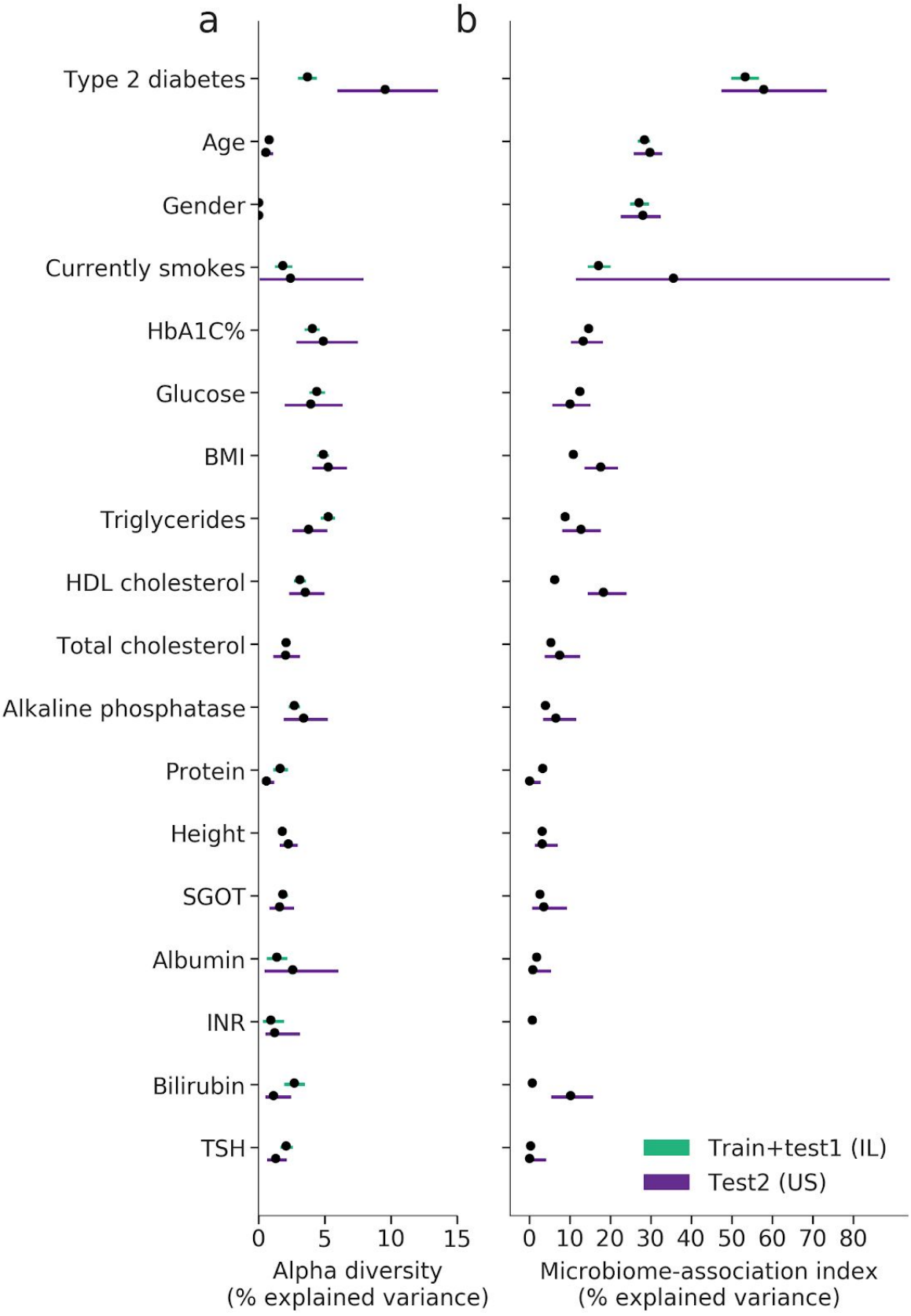

Figure 3: Explained variance of phenotypes based on microbiome features.

(a) The proportion of variance of various phenotypes that can be explained using Shannon alpha diversity in the Israeli (green) and U.S. cohorts (purple) based on a linear model with covariates for age and gender. Also shown is the 95% confidence interval.

(b) The proportion of variance of various phenotypes that can be explained using species-level relative abundances in the Israeli (green) and U.S. (purple) microbiome composition based on a linear mixed model estimation with covariates for age and gender (microbiome association index [26]). Also shown is the 95% confidence interval.

Estimates from the larger cohort have smaller confidence intervals.

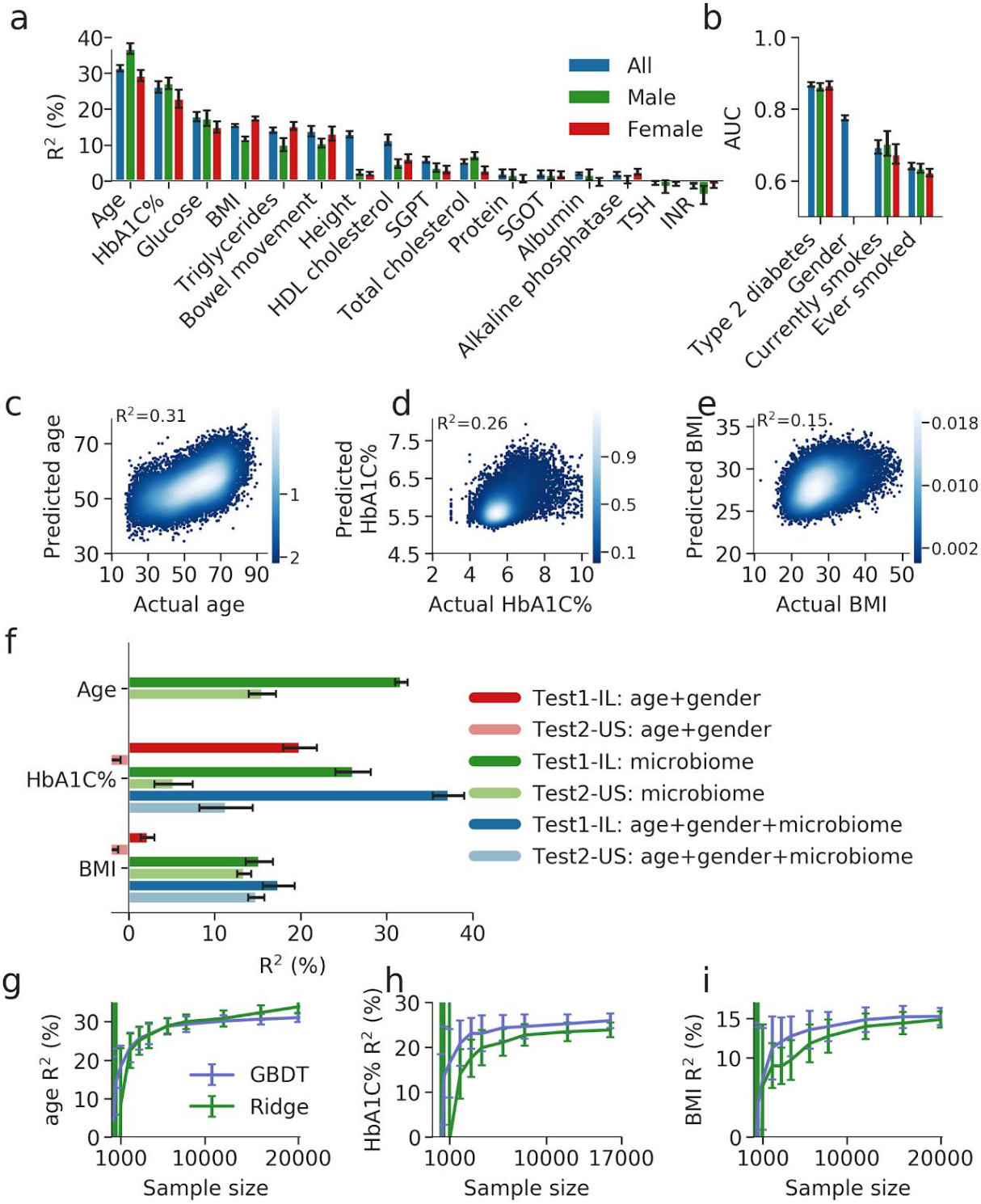

Figure 4: GDBT prediction of phenotypes by the microbiome.

(a) Coefficient of determination (R^2) of prediction of different phenotypes obtained in a 10-fold cross validation scheme on the training set. Predictions are shown for a model trained on the full training set cohort, and two others trained and tested separately on each gender. Each model was trained using only microbiome derived features.

(b) Same as (a), but shown is the area under the curve (AUC) for predicting binary phenotypes.

(c) - (e) Scatter plot of the phenotype and 10-fold cross-validation predicted values of the phenotype, for age, HbA1C% and BMI when training on the Israeli train cohort. R^2 of prediction is reported.

(f) Coefficient of determination (R^2) of predictions of age, HbA1C% and BMI, for models trained with different sets of input features, and tested on the training set in cross validation and on both the held-out Israel and U.S. test sets. Error bars of the test set are from bootstrapping.

(g) - (i) Coefficient of determination (R^2) and standard deviation error bars of predictions of age, HbA1C% and BMI obtained using the same 10-fold cross validation prediction scheme above, across different sub-samples of the cohort of different sizes. For each cohort size *k*, 10 random sub-samples of *k* individuals were obtained and the mean and standard deviation of their predictions are shown.

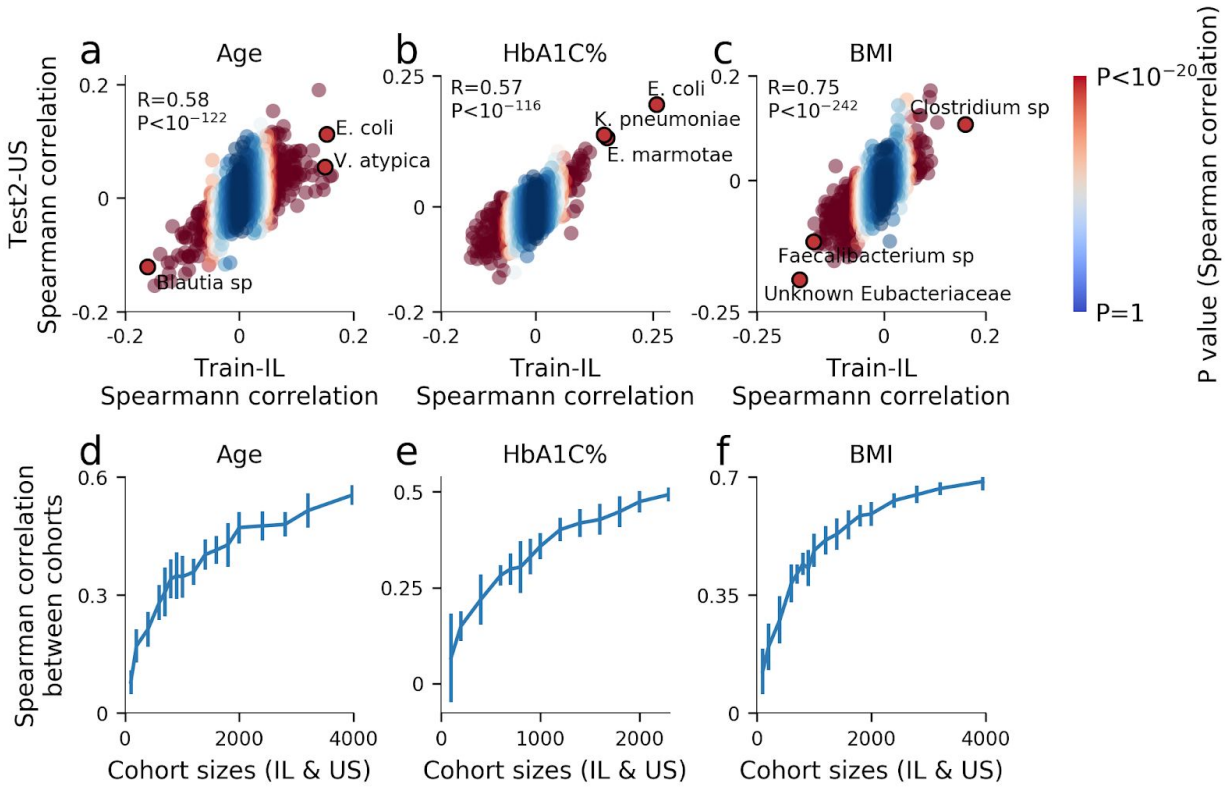

Figure 5: Predictive power of single species.

(a) Spearman correlation of each bacterial species with age in the Israeli training cohort (x-axis, N=27,075) and the U.S. test cohort (y-axis, N=3,974). The correlation and P-value between the correlations of each cohort are shown. Bacteria are colored according to the P-values of the Spearman correlation in the Israeli cohort. The top three bacteria by Israeli P-values that replicate in the U.S. cohort are highlighted.

(b) -(c) Same as (a) for HbA1C% and BMI.

(d) Spearman correlation between the correlations of the Israeli cohort and U.S. cohorts as in (a) but for different sub-samples of cohort sizes. For each cohort size *k*, a sub-sample of *k* individuals was obtained from both the Israeli and U.S. cohorts and this procedure was repeated 10 times to obtain standard deviation error bars.

(e) - (f) Same as (d) for HbA1C% and BMI.

Supplementary Figures

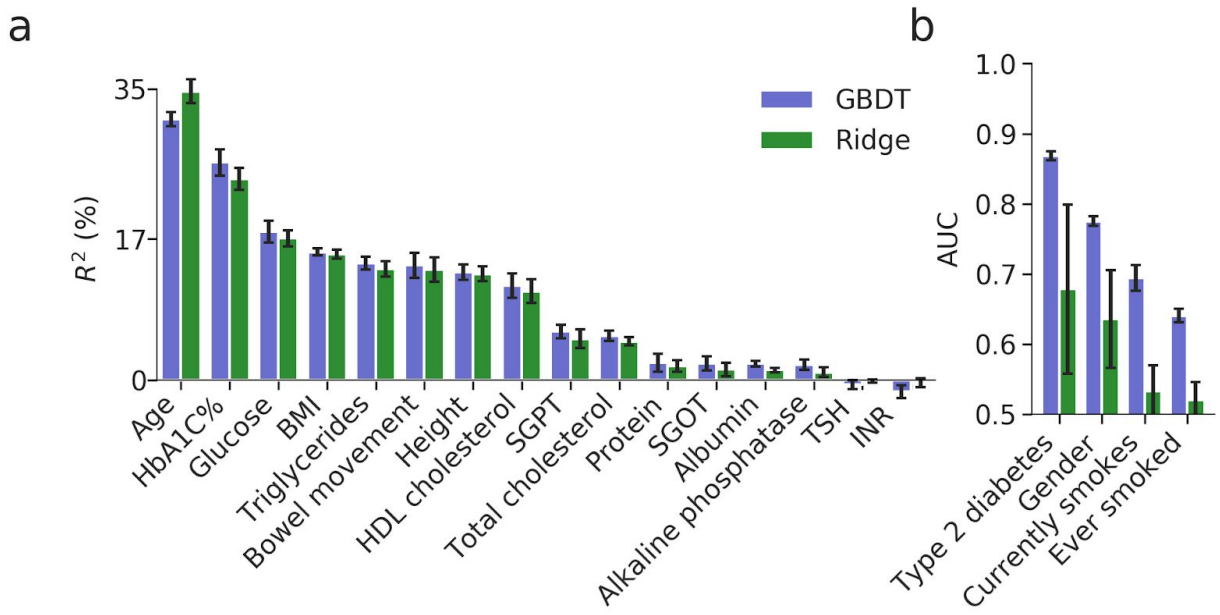

Figure S1: GBDT vs Ridge - prediction of phenotypes by the microbiome.

(a) Coefficient of determination (R^2) of prediction of different phenotypes obtained in a 10-fold cross validation scheme on the training set. Predictions are shown for a model trained on the full training set cohort using GBDT or Ridge regression. Each model was trained using only microbiome derived features.

(b) Same as (a), but shown is the area under the curve (AUC) for predicting binary phenotypes.

Bibliography

1. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).
2. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
3. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
4. Siljander, H., Honkanen, J. & Knip, M. Microbiome and type 1 diabetes. *EBioMedicine* **46,** 512–521 (2019).
5. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51,** 600–605 (2019).
6. Goodrich, J. K. *et al.* Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19,** 731–743 (2016).
7. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5,** e9085 (2010).
8. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352,** 565–569 (2016).
9. Sze, M. A. & Schloss, P. D. Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio* **7,** (2016).

10. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148,** 1258–1270 (2012).
11. Scher, J. U. *et al.* Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. *Arthritis Rheumatol.* **67,** 128–139 (2015).
12. Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500,** 585–588 (2013).
13. Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60,** 631–637 (2011).
14. Tuddenham, S. A. *et al.* The Impact of Human Immunodeficiency Virus Infection on Gut Microbiota α-Diversity: An Individual-level Meta-analysis. *Clin. Infect. Dis.* **70,** 615–627 (2020).
15. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350,** 663–666 (2015).
16. Srinivasan, S. *et al.* Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS ONE* **7,** e37818 (2012).
17. Kong, F. *et al.* Gut microbiota signatures of longevity. *Curr. Biol.* **26,** R832–R833 (2016).
18. Kong, F., Deng, F., Li, Y. & Zhao, J. Identification of gut microbiome signatures

associated with longevity provides a promising modulation target for healthy aging. *Gut Microbes* **10,** 210–215 (2019).

19. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **375,** 2369–2379 (2016).
20. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
21. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12,** e1001779 (2015).
22. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
23. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).
24. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–814 (2012).
25. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176,** 649-662.e20 (2019).
26. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555,** 210–215 (2018).

27. Yang, Y. C., Lu, F. H., Wu, J. S. & Chang, C. J. Age and sex effects on HbA1c. A study in a healthy Chinese population. *Diabetes Care* **20,** 988–991 (1997).

28. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017).

29. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–75, S1 (2012).

30. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18,** 117–127 (2017).

31. Mendes-Soares, H. *et al.* Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. *JAMA Netw. Open* **2,** e188102 (2019).

32. Suez, J. *et al.* Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature* **514,** 181–186 (2014).

33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

34. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9,** 1185–1188 (2012).

35. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.

Genome Biol. **10,** R25 (2009).

36. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568,** 43–48 (2019).
37. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).
38. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci USA* **111,** E5272-81 (2014).
39. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785–794 (ACM Press, 2016). doi:10.1145/2939672.2939785