Supplemental Digital Content 2.

Introduction to basic concepts of microbial analyses relevant to this study

16S-rRNA microbiome analysis

The microbiome analysis in this study is based on a common method sequencing 16S ribosomal RNA (rRNA) from fecal specimens. The 16S rRNA is the most established genetic marker, as it contains regions with species-specific sequences that allow for discrimination between different bacteria. Sequence reads are clustered together with other related sequences at a pre-defined level of identity and then quantified. A level of 97% sequence identity is often chosen as representative of a species and 95% for a genus. These clusters of similar sequencing reads are referred to as operational taxonomic units (OTUs). OTU counts are summarised and can then undergo further processing (e.g. determination of alpha diversity, as described below). Microbial identification and taxonomic classification is accomplished by comparison of sequences with genetic libraries/ reference databases. Counts of sequences can be assigned to taxa on phylogenetic levels with increasing order/decreasing resolution: species, genera, families, orders, classes, and phyla from the kingdom of bacteria. Summarization of relative abundances (relative to the total count of sequences per sample) provides insights into microbial composition and allows to compare the findings.

Alpha diversity

Alpha diversity is the microbial biodiversity within a habitat unit (here: fecal sample). There are several metrics taking into account different aspects of diversity such as bacterial richness, abundance, evenness, and phylogenetic information, or combinations of these. Observed species is the count of unique OTUs in a sample. Chao1 is the predicted number of taxa in a sample by extrapolating out the number of rare organisms that may have been missed due to undersampling. Observed species and Chao1 are measures of richness (counting/estimating the number of species), but do not take the abundances or phylogenetic diversity of the types into account. Phylogenetic diversity can be determined by locating each OTU on the phylogenetic tree of life, thus taking into account the evolutionary relationship between bacteria. Faith's Phylogenetic Diversity is the minimum total branch length of the phylogenetic tree that incorporates all OTUs in a sample. Rarefaction allows comparing of alpha diversity observed in different samples. It is necessary because the total number of reads (the bacterial 16S-nucleotid sequences) differ between samples. Rarefaction generates the expected number of species by (statistically) drawing the same amount of sequences at random from each faecal sample. Rarefaction curves generally grow rapidly at first, as the most common species are found, but the curves plateau as only the rarest species remain to be sampled.

Beta diversity

Beta diversity is the relative diversity between two habitats/samples. Beta diversity analyses in this study were all based on a metric called weighted UniFrac (1). UniFrac is the sum of phylogenetic tree branch lengths that is unique to one environment or the other. These lengths are then weighted by the relative abundance of respective bacteria. Weighted UniFrac distances for each pair of samples yield a distance matrix. The distance matrix can then undergo further testing, e.g. for associations with other variables (in this study Adonis, a type of non-parametric analysis of variance was calculated), or clustering. Clustering by the Unweighted Pair Group Method with Arithmetic Mean (UGPMA) is used to construct a dendrogram (tree diagram) reflecting the higher-order structure present in the pairwise distance matrix. The nearest two samples are merged into a new higher-level cluster, and the distance between the new cluster and the remaining samples is calculated. This is repeated until all samples are clustered.

Machine learning

Machine learning methods make use of nondeterministic algorithms allowing for pattern recognition and computerised learning from large amounts of data to make predictions from one class of variables to another (2). In the realm of microbiome research they are increasingly used to identify patterns of bacteria associated with clinical phenotypes, e.g. machine learning models have predicted severity of irritable bowel syndrome (3) or weight regain after a restrictive diet in mice from gut bacterial signatures (4). Machine learning models are usually built by identification of relevant (bacterial) features. In a second step, these are used for training and building of a comprehensive model. This model can then make predictions about other aspects of the data, as to classify the presence of psychological distress in IBS patients in this study. As machine learning is susceptible to overfitting, external validation is usually required. Without validation under real-world conditions (as in this study) their value is limited to a demonstration of the existence of systematic associations between different classes of data.

1. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71:8228-35.
2. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. Cell. 2018.
3. Tap J, Derrien M, Törnblom H, Brazeilles R, Cools-Portier S, Doré J, Störsrud S, Le Nevé B, Öhman L, Simrén M. Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. Gastroenterology. 2016;152:111-23.
4. Thaiss CA, Itav S, Rothschild D, Meijer MT, Levy M, Moresi C, Dohnalová L, Braverman S, Rozin S, Malitsky S. Persistent microbiome alterations modulate the rate of post-dieting weight regain. Nature. 2016;540:544-51.