Sample selection

Sample selection for proteomics largely followed sample selection for metabolomics. In total, 447 stool samples were selected for profiling.

LC-MS/MS

Proteins were proteolytically digested using trypsin, and each digest subjected to automated offline high pH reversed-phase fractionation with fraction concatenation. in order to reduce overall sample complexity and increase coverage of the proteome. LC-MS/MS analysis for each fraction was performed using a Thermo Scientific Q-Exactive Orbitrap mass spectrometer resident at UCLA, outfitted with a custom-built nano-ESI interface. Samples were loaded onto an in-house packed capillary LC column (70 cm x 75 μm, 3 μm particle size), and data was acquired for 120 min. Precursor MS spectra were collected over 400–2000 m/z, followed by data-dependent MS/MS spectra of the twelve most abundant ions, using a collision energy of 30%. A dynamic exclusion time of 30 s was used to discriminate against previously analyzed ions.

Peptide identification and protein data roll-up

Mass spectra from the resulting analyzes were evaluated using the MSGF+ software v10072 using the HMP 1 gut reference genomes (HMP_Refgenome-gut_2015-06-18). Briefly, after conversion of the metagenomic assemblies into predicted open reading frames (e.g., predicted proteins), libraries were created using the forward and reverse direction to allow determination of False Discovery Rate (FDR). The reverse decoy database allows measurement of the rate of detection of false hits, which in turn allows for FDR calculation and appropriate filtering of the data to maximize the real peptide identifications while minimizing spurious ones. MSGF+ was then used to search the experimental mass spectra data against both the forward/reverse decoy databases. Cut-offs for data included: MSGF+ spectra probability (>1×10^{10}, equivalent to a BLAST e-value), mass accuracy (±20 ppm), protein level FDR of 1% and one unique peptides per protein identification.

Sample and data tracking

Sample processing for mass spectrometry analytical data collection and primary processing of raw instrument data, was coordinated and tracked using the PNNL PRISM system. The Pan-omics Research Information Storage and Management System acquires data from mass spectrometers and other instruments, collects laboratory information, and tracks and controls the intermediate data processing.