

A faecal metabolite signature of impaired fasting glucose: results from two independent population-based cohorts

Ana Nogal^{1*}, Francesca Tettamanzi^{1,2*}, Qiuling Dong³, Panayiotis Louca¹, Alessia Visconti¹, Colette Christiansen¹, Taylor Breuninger⁴, Jakob Linseisen^{4,5,6}, Harald Grallert^{3,7}, Nina Wawro^{3a,4}, Francesco Asnicar⁸, Kari Wong⁹, Andrei-Florin Baleanu¹, Gregory A. Michelotti⁹, Nicola Segata⁸, Mario Falchi¹, Annette Peters^{3a,7,10}, Paul W. Franks^{11,12}, Vincenzo Bagnardi¹³, Tim D Spector¹, Jordana T Bell¹, Christian Gieger^{3,7}, Ana M Valdes¹⁴, Cristina Menni¹

Supplementary material

Metabolomics profiling

Metabolite concentrations were measured from faecal samples by Metabolon Inc. (Durham, USA) using an untargeted LC-MS platform. All samples were maintained at -80°C until processing. As a means of quality control, several recovery standards were added prior to the first step in the extraction process. Briefly, to remove protein, dissociate small molecules bound to proteins or trapped within the precipitated protein matrix, and to recover chemically diverse metabolites, proteins were precipitated in methanol and vigorously shaken for 2 minutes (Glen Mills GenoGrinder 2000), then centrifuged. The resulting extract was divided into five fractions; both aliquots (i) and (ii) were analysed using acidic positive ion conditions and chromatographically optimised for hydrophilic and hydrophobic compounds respectively, aliquot (iii) was analysed using a basic negative ion optimised conditions using a dedicated separate dedicated C18 column, aliquot (iv) was analysed using negative ionisation following

elution from a hydrophilic interaction liquid chromatography column, while aliquot (v) was reserved as a back-up.

Several controls were analysed in concert with experimental samples. (i) a pooled sample generated from a small volume of each experimental sample of interest served as a technical replicate throughout the platform run; (ii) extracted water samples served as process blanks; (iii) and a cocktail of standards, known not to interfere with measurements, spiked into every analysed sample facilitated instrument performance monitoring and aided chromatographic alignment. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% or more of the pooled technical replicate samples. Experimental samples and controls were randomised across the platform run.

Compound identification

Metabolites were identified by comparison of the ion features in the experimental samples to a reference library of chemical standard entries that included retention time/index, molecular weight (m/z), and MS spectra. Identification of known chemical entities is based on comparison across all 3 features to metabolomic library entries of purified standards. More than 3300 commercially available purified standard compounds have been acquired and registered into the library, while additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass

spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

Metabolite quantification and normalisation

Peaks were quantified using area-under-the-curve. Raw area counts for each metabolite in each sample were normalised to correct for variation resulting from instrument inter-day tuning differences by the median value for each run-day, therefore, setting the medians to 1.0 for each run. This preserved variation between samples but allowed metabolites of widely different raw peak areas to be compared on a similar graphical scale.

Metagenomic assessment in TwinsUK

Faecal sample collection

Participants collected stool samples at home in pre-labelled kits (containing 2 x 25ml tube or 1 x 25ml tube and 1 x 10ml Zymo buffer), which were posted to them before their clinic visit date and brought with them to the visit. In the laboratory, samples were homogenised, aliquoted into 4 bijoux tubes, and stored at -80°C , within 2 hours of receipt.

DNA extraction, library preparation, and sequencing

To isolate genomic DNA from faecal material, bijoux tubes were removed from the freezer and grounded with glass beads and 5-6ml distilled water (Spex Grinder, 10 seconds, 800 strokes per minute). The supernatant was centrifuged and further

grounded (5 minutes, 1000 strokes per minute) before 200-300µl of the sample was mixed with 10µl PK solution and 720µl of Lysis/Bind Master Mix). Proteins were degraded by the binding solution and subsequently extracted by KingFisher Flex robot. DNA was washed in 2 steps using washing solutions and eluted in MagMax Core Elution Buffer in 100µl. Library preparation and sequencing was performed by GenomeScan.

Metagenome quality control and preprocessing

Sequenced metagenomes were processed using the YAMP pipeline (v. 0.9.5.3). Briefly, identical reads were removed. Reads were filtered to remove adapters, known artefacts, phix174, and then quality trimmed (PhRED quality score < 10). Reads that became too short after trimming (N < 60 bp) were discarded. We retained singleton reads (i.e., reads whose mate has been discarded) to retain as much information as possible. Contaminant reads belonging to the host genome were removed (build: GRCh37), and low-quality samples (i.e., samples with <10M reads after QC) were discarded.

Microbiome taxonomic profiling

The metagenomic analysis was conducted following the general guidelines and based on the bioBakery computational environment. High resolution taxonomic profiling of the metagenomes was performed using MetaPhlAn 4.beta.2 with the January 2021 database and default parameters.

Statistical analysis

We run random forest regression (1000 trees and a third of features number as number of variables randomly sampled as candidates at each split) and classification models (1000 trees and square root of features number as number of variables randomly sampled as candidates at each split) with compositional data using 5-folds cross-validation. Before running the models, gut microbiota variables with variance zero or near to zero were excluded using the `nearZeroVar` function implemented in R in the `caret` package (the included/excluded SGBs are shown in **Supplementary Table 6**). For the classifiers, the continuous response was converted into two classes based on the top and bottom quartiles. The features were ranked based on the node purity.