Supplemental Information

Metabolite Profiling

Metabolite profiling data of 600 samples was acquired separately for the two independent sub-studies described in the main text with 300 samples in each study. The two studies were defined as sub-study 1 and sub-study 2. Sample information was blinded to the analysts (WD, DW, MB, RG) until after data acquisition. Serum samples were prepared in a randomised order and according to a protocol described elsewhere (1, 2). Briefly, samples were allowed to defrost on ice at 4°C. For deproteinization, 420 µL of methanol (Sigma-Aldrich, UK) were added to a 140 µL aliquot of serum. The samples were vortex mixed for 15 s followed by centrifugation (15 min, 13 363*g*). Two aliquots of 220 µL were transferred to separate 1.5 mL polypropylene Eppendorf tubes (Fisher Scientific, Loughborough, UK) and lyophilised (HETO VR MAXI vacuum centrifuge attached to a HET CT/DW 60E cooling trap; Thermo Life Sciences, Basingstoke, UK).

One lyophilised aliquot was used for analysis, applying a UPLC-MS platform in positive ion mode and the second aliquot was employed for analysis applying an ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) platform in negative ion mode. The remaining 60 µL of serum from each participant was pooled to create a single pooled Quality Control (QC) sample, of which 140 µL aliquots were deproteinized and lyophilised as described above. These QC samples were applied for conditioning of the analytical system, signal correction and quality assurance as previously described (1, 3).

Samples in sub-study 1 and sub-study 2 were analysed in a random order and in two independent experiments, each experiment consisting of three analytical batches. Batches 1-3 were applied in sub-study 1 and batches 4-6 were applied in sub-study 2. The respective batches were composed of 100 subject samples and 28 intermediate QC injections, which were completed within 44 h after sample reconstitution and batch initiation. All samples were reconstituted in 70 µL 50:50 methanol/water followed by centrifugation for 15 min at

13,363*g*. The supernatants were transferred to 2 mL low-volume chromatography vials sealed with septum containing screw caps and stored at 4 °C in the UPLC autosampler.

Analytical UPLC-MS measurements were carried out on an ACQUITY UPLC system (Waters, Elstree, UK) that was interfaced with a LTQ-Orbitrap XL hybrid mass spectrometer equipped with an electrospray ionisation source (ThermoFisher Scientific, Bremen, Germany). The UPLC method was operated over a 22 minute run time as described previously (4). The same gradient was applied for positive and negative ion mode to enable accurate integration of data for metabolite identification for both ion modes. In total, 10µl of sample was injected on to an ACQUITY BEH C_{18} column (Waters, Elstres, UK; 2.1 mm i.d., 100 mm length, 1.7 µm particle size). The column was eluted at a flow rate of 0.4 $ml \cdot min^{-1}$ with a binary solvent gradient composed of A (99.9% water and 0.1% formic acid) and B (99.9% methanol and 0.1% formic acid) and a column temperature of 50 °C. The column was operated for three analytical batches (separately, 3 positive ion mode for one column and 3 negative ion mode for a separate column). 50% of the column eluent was introduced to the electrospray source of the LTQ-Orbitrap XL mass spectrometer. The mass spectrometer was operated in one ion mode and was tuned using a calibration solution and a single mass (*m/z* 514 positive ion mode and *m/z* 524 negative ion mode) at the start of each set of 300 subject samples. Mass calibration was performed according to the manufacturer's instructions before each analytical batch. Accurate mass data was acquired in the *m/z* range 50-1000 with a scan rate of 0.4 s. After each analytical batch the column was washed for 30 minutes with 100% methanol.

Statistical analyses

Due to the comparison with QC samples, unreliably detected metabolites were deleted from the data sets (see, 'Raw Data Pre-Processing'). Consequently, missing values in the remaining data might reflect ions with a concentration below detection limit or truly absent ions.

Accordingly, missing values in the combined dataset (batches 1-6) for positive and negative ion modes respectively were replaced by a value close to zero; namely, the higher value from two calculations: (a) mean value for the metabolite minus 3 standards deviations or (b) lowest value in distribution x 0.5.

Classification-based selection of metabolites

For feature selection, also known as variable selection as part of the model construction process, we applied multivariate classification-based feature selection using Random Forests (RF) (5) and Partial Least Squares Discriminant Analysis (PLS-DA) (6) in the two sub-studies. Two R packages, *randomForest* and *pls*, were employed. To remove the dependency between metabolite ranking and metabolite concentration (7), autoscaling was applied before feature selection.

RF is a machine learning strategy based on growing an ensemble of many binary decision trees. With RF, random training sets (i.e. bootstrap samples) are repeatedly drawn from all available study participants. For each bootstrap sample, a decision tree is constructed by randomly choosing a variable subset of all predictor variables at each node. Among these predictors, RF selects the variable that best splits data into two daughter nodes. The performance of each tree is tested in *the out of bag* (oob) sample which comprises the study participants not included in the respective bootstrap sample used to construct the tree. In our data, ntree and mtry were fixed at 500 and one third of predictor variables, respectively. In total, 100 bootstrap samples were selected with replacement on the matched case-control pairs. Metabolites were ranked according to the RF importance scores reflecting the difference between oob error of a variable randomly permuted and the original variable. The higher the importance score, the more important a predictor is with regards to classification.

PLS-DA is an extension of the partial least squares method by Wold (8), which extracts successive linear combinations of the original explanatory variables in order to maximize the

covariance between these variables and a set of response variables. In the special case of PLS-DA, the response variable is binary and the latent variables are selected such that a large proportion of variance in the explanatory variables is explained and maximum separation between cases and controls is obtained (6). Metabolites of each sub-study were ranked according to the regression coefficients representing the importance of each metabolite in classification.

The results of the two methods were aggregated to a ranking list by the average of each ranking score. Here, the average ranking score indicates the overall importance of each metabolite or metabolite group with regards to classification of T2D cases and controls. A subset of 60 predictors per ion mode and sub-study were selected for further statistical analyses, because larger subsets would not improve discrimination accuracy of the models (data not shown).

References for Supplementary Information

1. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 2011;6:1060-83.
2. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, et al. Development of a robust and repeatable uplc-ms method for the long-term metabolomic study of human serum. Analytical chemistry 2009;81:1357-64.
3. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and qc samples in large-scale and ms-driven untargeted metabolomic studies of humans. Bioanalysis 2012;4:2249-64.
4. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, et al. Development of a robust and repeatable uplc-ms method for the long-term metabolomic study of human serum. Anal Chem 2009;81:1357-64.
5. Breiman L. Random forests. Machine Learning 2001;45:5-32.
6. Barker M, Rayens W. Partial least squares for discrimination. J Chemometrics 2003:166-73.
7. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC genomics 2006;7:142.
8. Wold H. Estimation of principal components and related models by iterative least squares. New York: Academic Press, 1966.