

S1 Text.

Transcriptomics

For the study of transcriptomic profiles of participants at high risk of diabetes (WP2.1) and new onset of diabetes (WP2.2) 3,094 samples of mRNA from whole blood samples were processed by University of Oxford and shipped to University of Geneva for RNA-sequencing. Concentration of mRNA per samples was assessed using the Qubit2.0 from Invitrogen. The quality of the samples was then assessed using the TapeStation Software (A.01.04) with an RNA Screen Tape from Agilent to check the mRNA quality on gel. 29 samples were discarded at this point due to low mRNA quality. The remaining samples were processed and sequencing libraries were prepared. Quality of the libraries was evaluated using Qubit and TapeStation using DNA1000 Screen Tape. One sample was discarded after library preparation due to low quality. The remaining samples (3,064) were placed in Flow cell PE using the cBOT system from Illumina. The samples were then sequenced on the Illumina HiSeq2000 platform using 49 bp paired-end reads. Sequenced reads were mapped to the GRCh37 reference genome (2001) with GEM (Marco-Sola et al. 2012). Identification of samples mix-ups and labeling errors is possible when genotypes are available (tHoen et al. 2013). For each sample, we tested the heterozygous sites in DNA genotypes for expression of both alleles in the RNAseq data. Samples mix-ups or mislabeled show lower levels of expression of both alleles. Using the package matchASE from the suite QTL-tools (manuscript in preparation), we tested each expression profile (BAM files) against all imputed genotypes from DIRECT WP2 (release version October 2015) to identify the best matching expression-genotype pair. The analysis identified 201 samples with mismatch between expression and genotypes. Of those samples 137 could be corrected as we identified another genotype with a good match within the data

set. One sample matched a genotype registered as not having mRNA sample. In these samples we are not certain whether genotype or RNAseq samples were swap with another individual. Use of the samples is not recommended in association with other phenotypes. Other 4 samples had low quality or were mixes of RNAseq samples and would not be identified with confidence. For 60 samples we would not find a suitable match between expression and the available genotypes. Gender identification in RNAseq compare expression levels of genes in the autosomal region for the chromosome Y and the expression of the XIST gene in the chromosome X. To confirm gender information, we compare the gender provided by clinical reports and the gender identified by genotype data with the gender identified from RNAseq data.

Exploratory proteomics

We performed antibody bead array assays for profiling of plasma proteins (Drobin K et al., 2013) using a selected set of 779 antibodies (targeting 385 proteins) availability from the Human Protein Atlas (HPA) (Uhlen et al., 2015). EDTA plasma samples from the four different DIRECT study centers were distributed across microtiter plates via a supervised randomization procedure. After sample randomization, the 96-well microtiter plates were stored at -80°C until further use. Plasma samples were thawed at 4°C and centrifuged for 10 min at $3,000 \times g$. Three microliters of each sample were diluted in 22 μ l of 1x PBS using a liquid handler (SELMA, CyBio). As described earlier (Drobin K et al., 2013), samples were then biotinylated with NHS-biotin, the reaction was quenched using TRIS-HCl. Biotinylated samples were diluted in assay buffer, heated treated at 56°C for 30 min and then added the bead arrays prepared above. The magnetic beads were washed on the next day using a plate washer (EL406, Biotek). Fluorescent streptavidin was added for the detection of biotinylated proteins captured by bead-immobilized antibodies. The beads were analysed using the FlexMap 3D instrument (Luminex Corp.) operated by the xPONENT software version 4.2. The median fluorescence intensity (MFI) was

used to represent the relative amount of target protein binding to each of the antibody-coupled bead identity.

The obtained data was evaluated based on intensity levels and three antibodies were excluded from further analysis eight samples were flagged that seemingly failed. Such samples were those 1) that had median values of MFIs ± 2 SD or below the median of control measurement without any sample (buffer only), and 2) that were identified as outliers using Robust PCA using 'rrcov' R package (version 1.4-3) (Hubert et al., 2005). The cutoff probability values in an outlier diagnostic plot were set to 0.001 for both score and orthogonal distances. The samples deviating beyond the cutoffs in both distance coordinates were classified as outliers, setting alpha, the proportional tolerance, to 0.9. The remaining data set was denoted as annotated.

Targeted proteomics

Samples from DIRECT study centers were manually randomization by a mix-shake-distribute procedure and placed into 96-well plates. All samples were analyzed at SciLifeLab in Stockholm using several different immunoassay platforms.

Proteins were measured in EDTA plasma using the Cardiometabolic, Cardiovascular II, Cardiovascular III, Development and Metabolism panels from Olink Proteomics AB (Uppsala, Sweden) according to the instructions for the Proximity Extension Assays (PEA) (PMID: 24755770). The obtained normalized expression values (NPX) values were obtained from Olink's NPX manager software version 0.0.85.0. Magnetic bead-based assays were used for the analysis of FGF21 (SPRCUS627, MerckMillipore) and a panel consisting of CXCL10, ICAM-1, IL1R-alpha, and RETN (LXSAHM, R&D Systems). The assays were performed according to the instructions and the instrumentation for liquid handling as introduced above. The beads were analyzed using the FlexMap 3D instrument (Luminex Corp.) operated by the xPONENT software version 4.2. The obtained MFI values were converted into concentration

values using 5-parametric fitting. Plasma levels of IL1-beta and TNFR1-alpha were quantified in accordance with the instructions for the microfluidic ELISA assays (PMID: 27170460) from ProteinSimple.

Additional proteins were analyzed in randomized samples using the services from Myriad RBM (Myriad GmbH, Germany) and for hsCRP (MLM Medical Labs GmbH, Germany).

Targeted metabolomics

Plasma concentrations of 163 metabolites were determined using a targeted metabolomic approach with the Absolute*IDQ*TM p150 kit (BIOCRATES Life Sciences AG, Innsbruck, Austria). After data export from Met*IDQ*, a first technical QC comprising analysis of peak shapes, retention times, and compound identity was performed. In a second QC step possible batch effects, study centre effects and effects of different phenotypes were investigated by using principal component analysis (PCA). The PCA of the data showed centre specific clusters. Further analysis showed phenotypic and technical (sample handling/ preparation/ collection/ storage and transport) differences between the study centres. Therefore, data were corrected for batches and study centres. Lower outliers were defined as samples with >33% of metabolite concentrations below 25% quantile – 1.5*IQR. Upper outliers were defined as samples with >33% of metabolite concentrations above 25% quantile + 1.5*IQR. Metabolite traits with too many zero concentration samples and NAs (>50%) were excluded (none). The Coefficient of Variation (CV) was calculated in reference samples for each metabolite over all plates. Metabolite traits with CV>0.25 were excluded. Metabolite traits with > 95% of samples below LOD were marked.

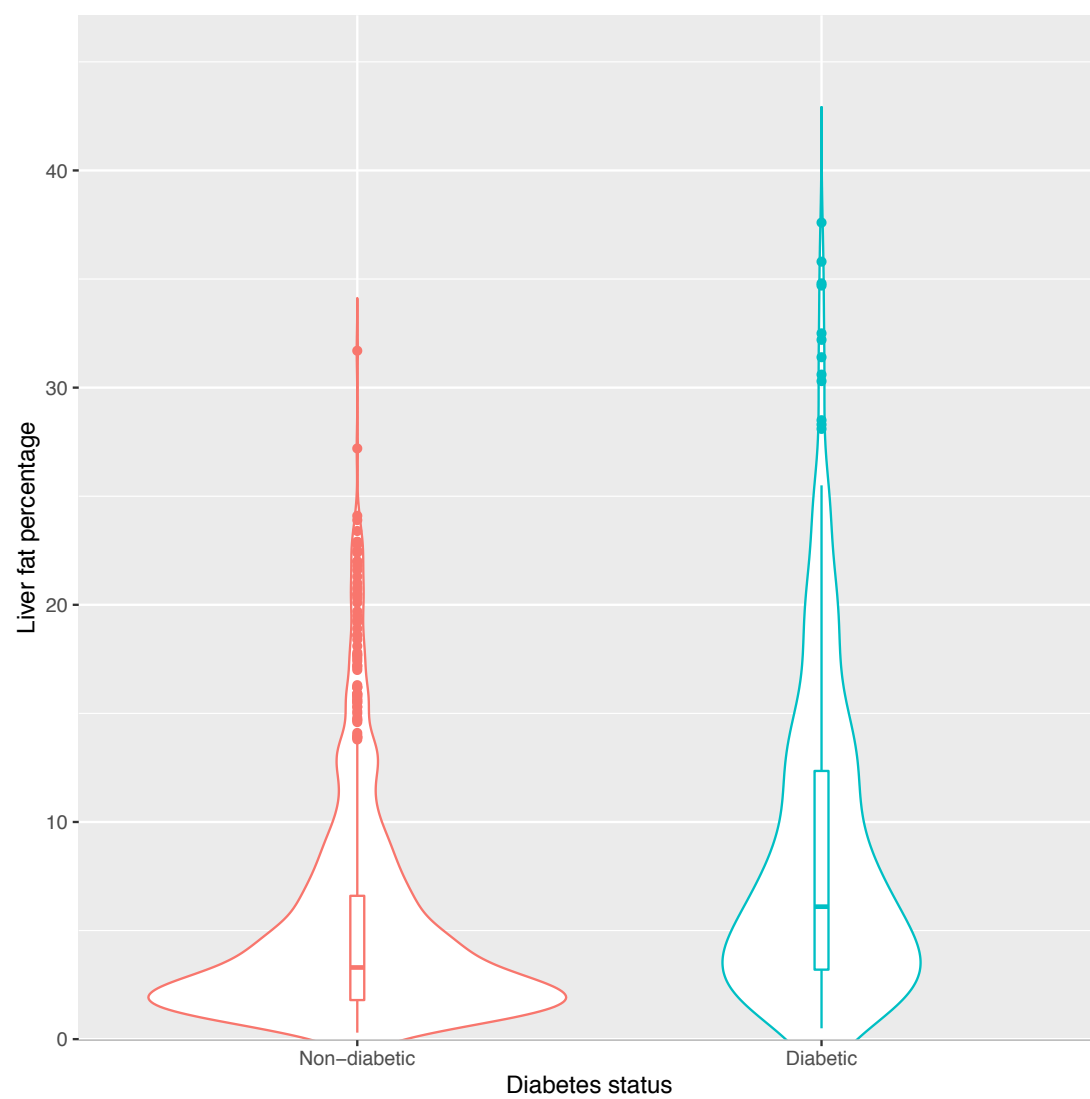
Untargeted metabolomics

Untargeted LC/MS-based techniques cover a broader spectrum of metabolites, in contrast to the targeted techniques where metabolites are limited to a predefined set of molecules. Non-targeted metabolomics data sets typically contain approx. 20-30% missing values, affecting more than 80% of the measured compounds. Missingness in the data can occur for several reasons. Firstly, a molecule might not be present at all in the sample. Secondly, missing values can be due to technical effects like, (i) metabolite concentrations are below the limit of detection (LOD), (ii) matrix or contamination effects hamper the quantification of a metabolite in a sample through co-eluting compounds due to altered ionization properties (signal suppression or enhancement), and (iii) challenges in computational processing of spectra, such peak picking and peak alignment (e.g. overlapping peaks). In order to understand how missingness seems within metabolon data, we compared the missing values of a metabolite between different run days and investigated the standard deviations of the run day missingness proportions across the metabolites. We further split the missing value distribution of metabolites as per the pathways, so that we know which pathways are informative or most relevant. We considered a sample to have a “low measurement” for a given metabolite if the sample's measurement was in the lower tail of the distribution of measurements for the metabolite. In such a case we called the metabolite "low". An outlier is a metabolite-sample pair where the distance of the log₁₀-transformed metabolite measurement from the mean of the metabolite was greater than 4 times the standard deviation of the metabolite. Outliers were removed from the data set.

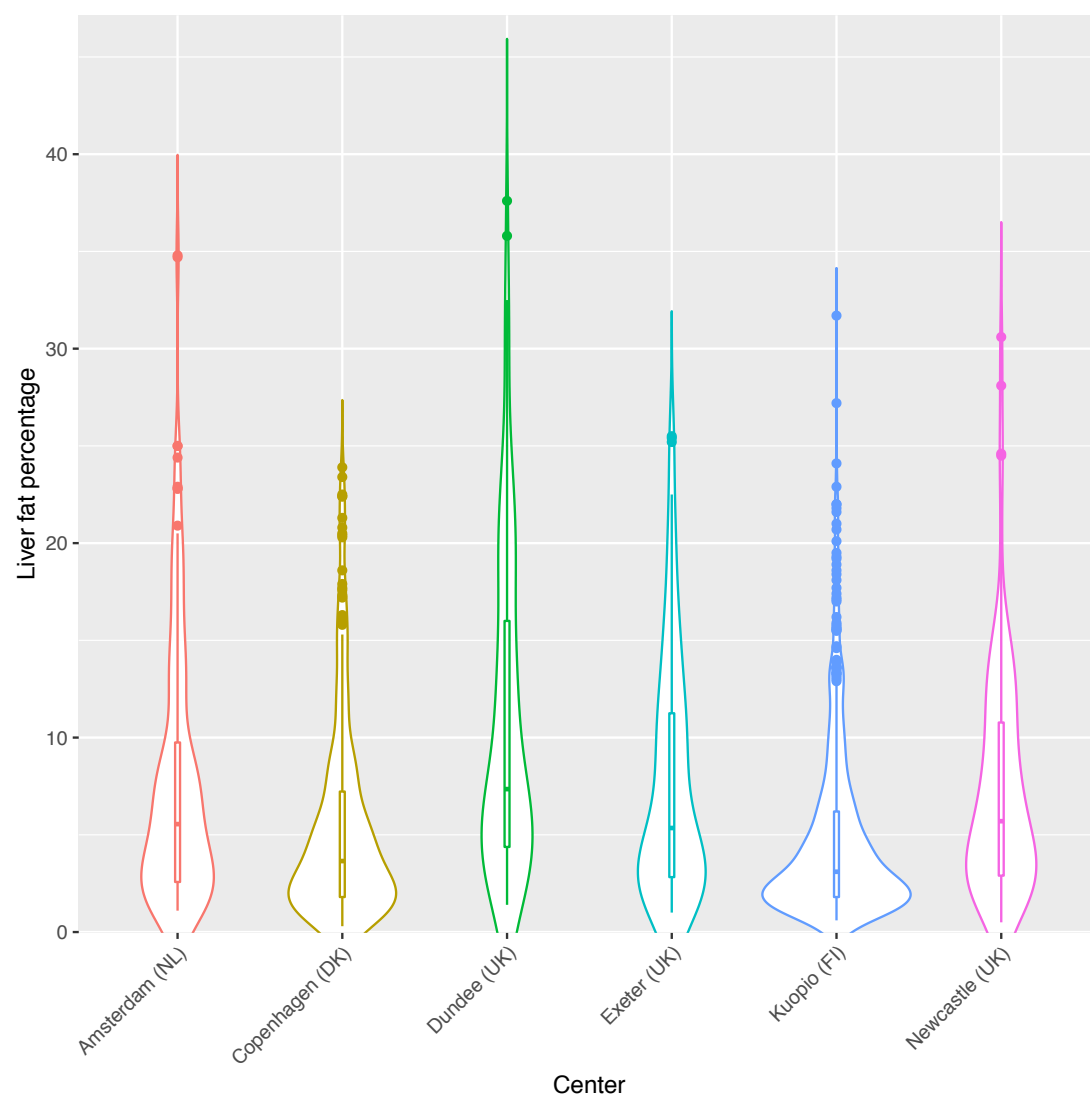
In the next QC step, we computed the coefficient of variation (CV) of measurements by run day for every metabolite. We then calculated the median CV over run days and used this as a measure of the variability of the measurement process. A median CV greater than 0.25 was considered as evidence that the measurement process was unable to yield reliable measurements for the given metabolite. In such cases we excluded the metabolite due to low quality. We also

excluded a metabolite unless the CV could be computed for at least two run days. In particular, metabolites that had nothing but missing values in the reference plasma data were excluded. The number of metabolites were excluded later on due to a median CV over run days that is too high (> 0.25) or because there were so few non-missing measurements for a metabolite that its CV could only be computed for a single run day or not at all. In other words, we require a minimum of 2 run day CVs before we consider computing the median.

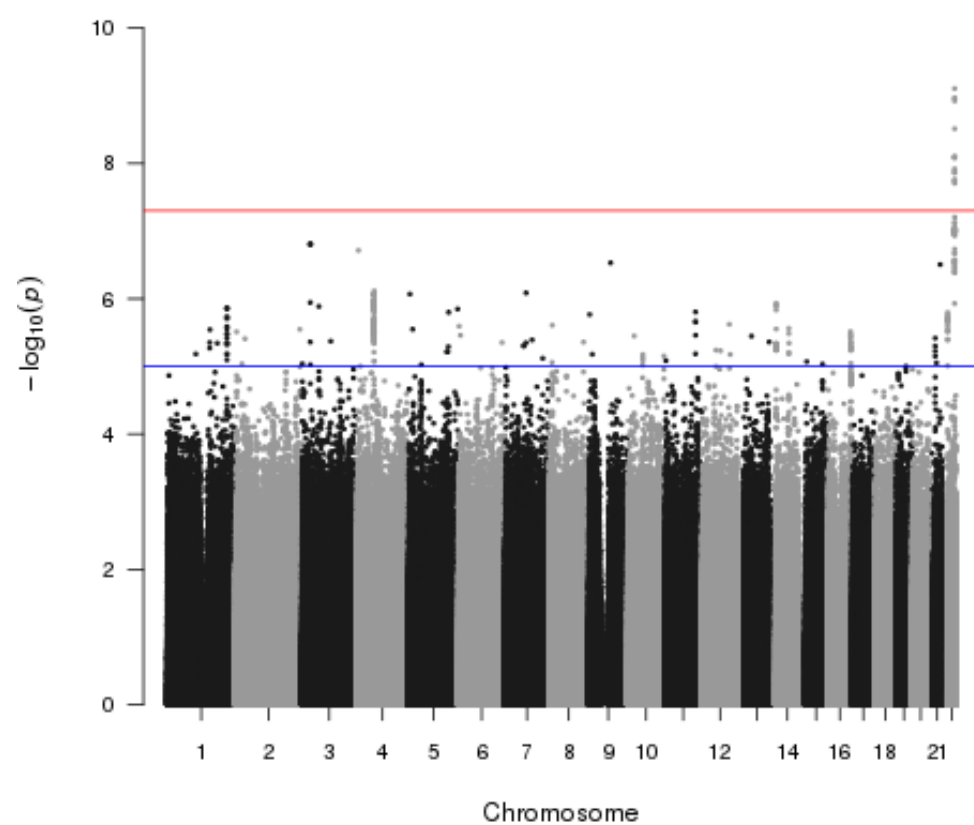
After these QC steps, missing values were imputed with the “multivariate imputation by chained equations” (MICE) method by using R package mice (version 2.2.5). We set the number of imputations to $m=20$ for all methods, assuming this to be a sufficient number to assure accuracy of the obtained estimates in a simulation study.



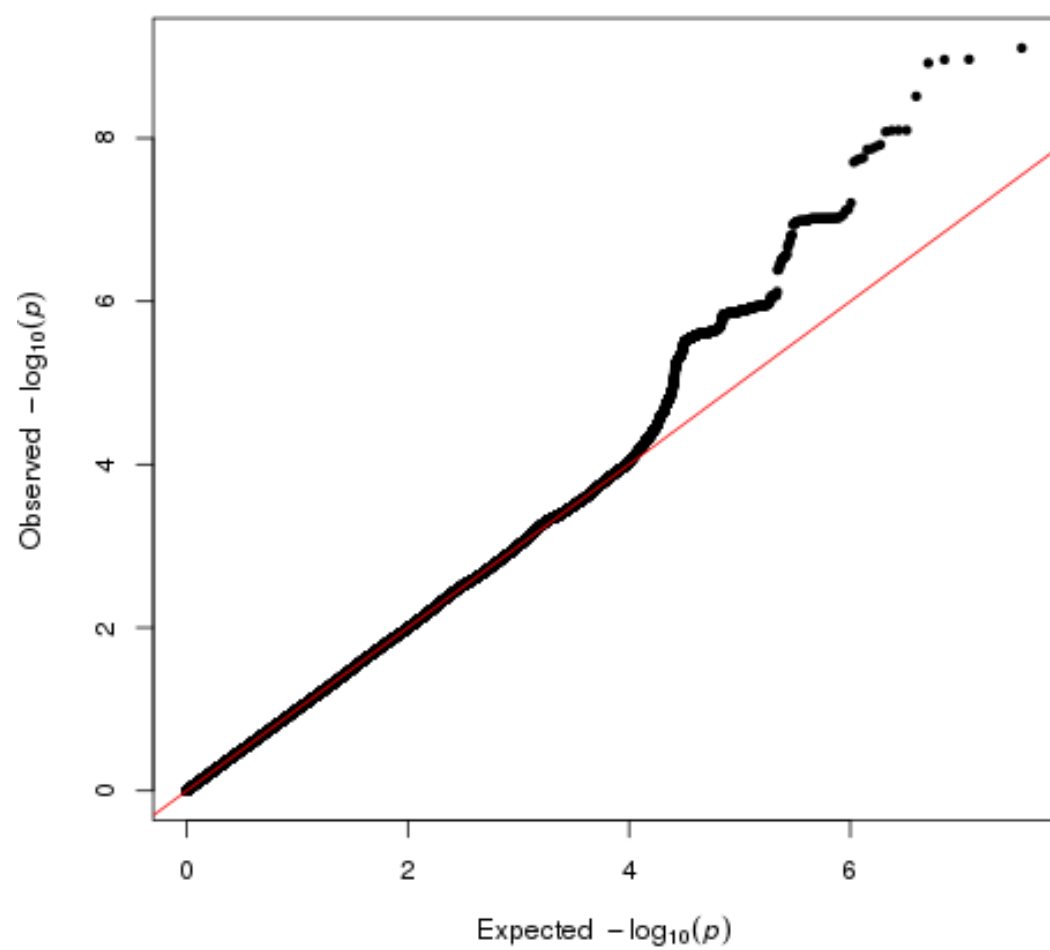
S1 Fig.



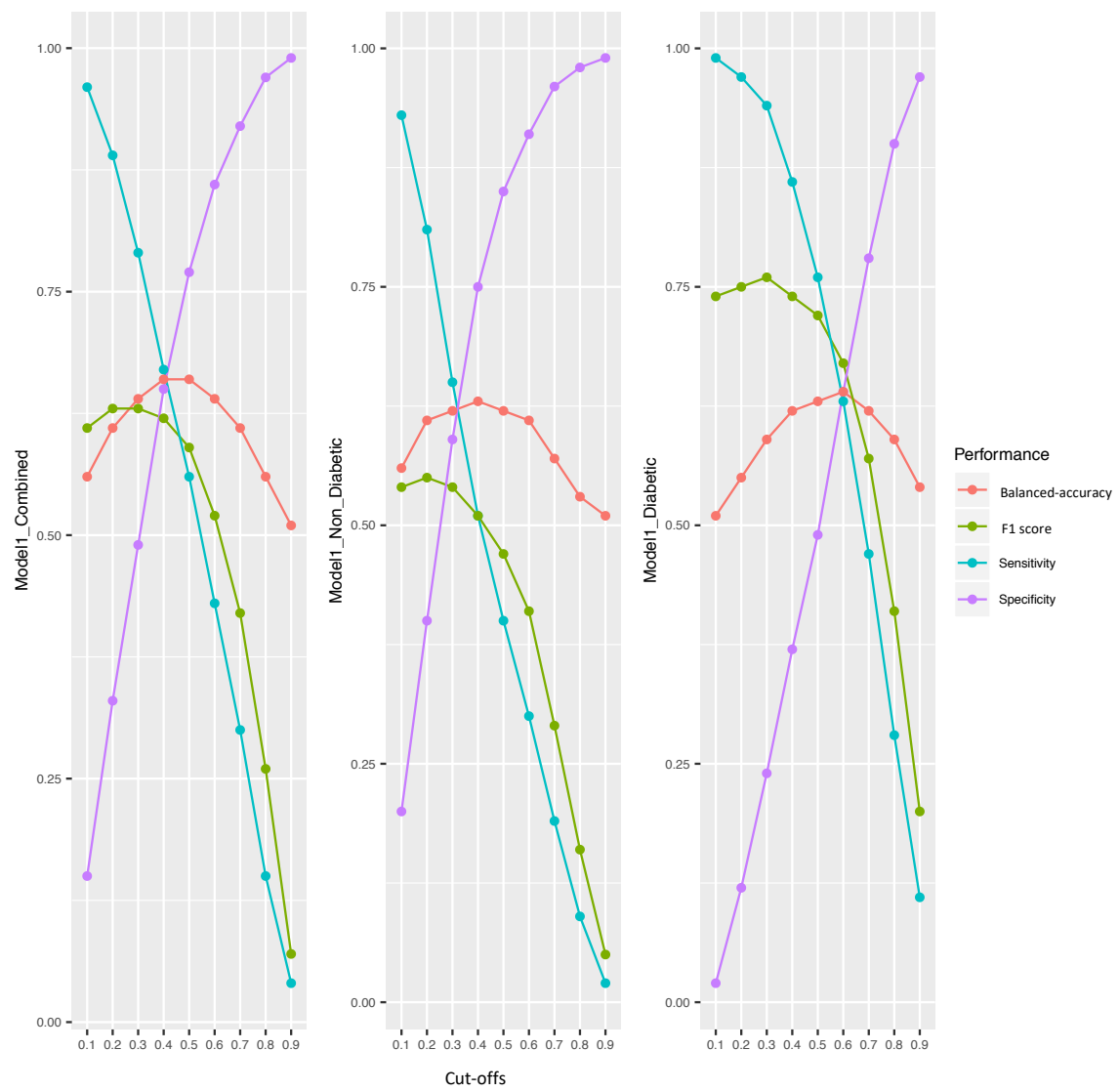
S2 Fig.



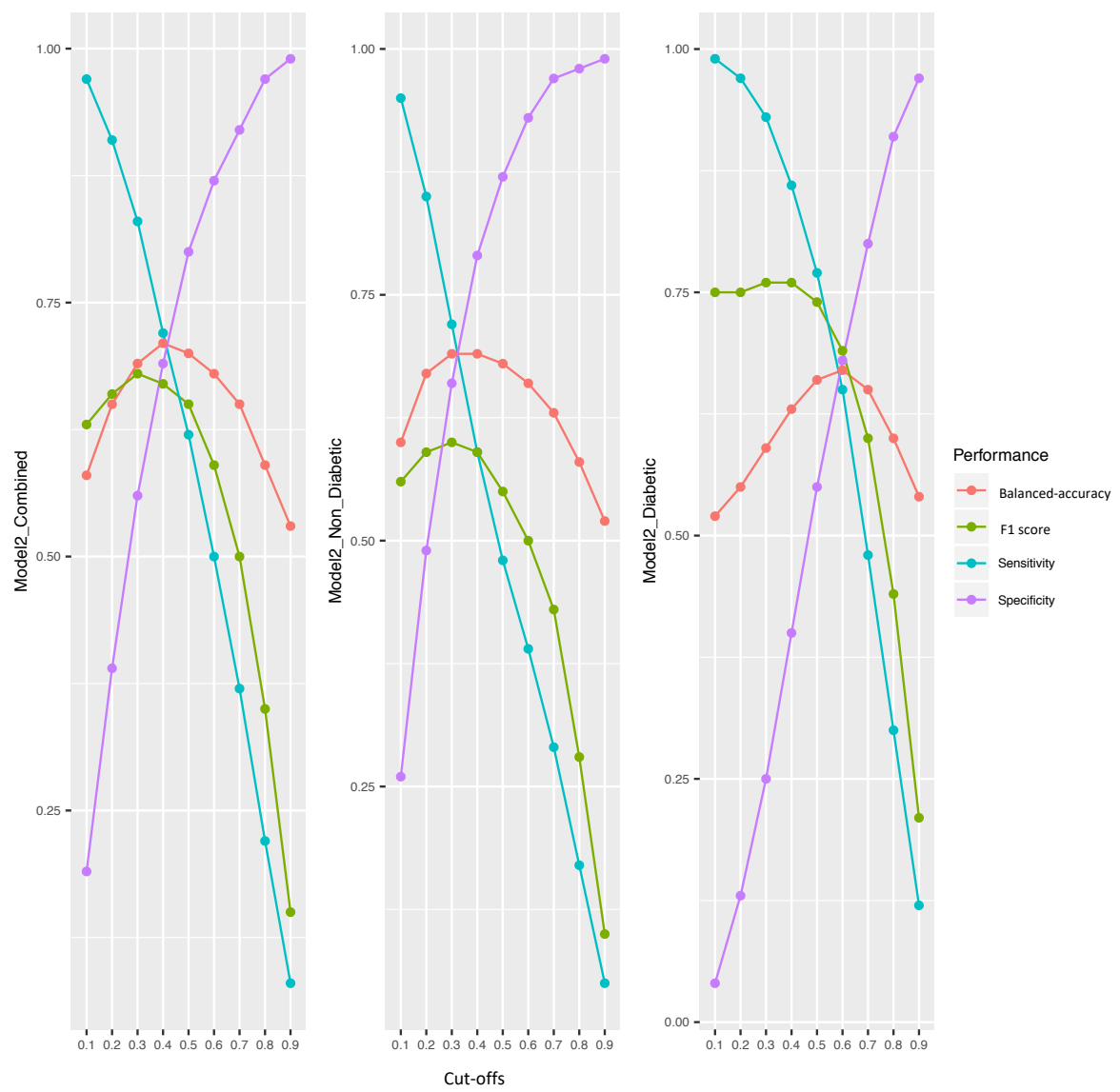
S3 Fig.



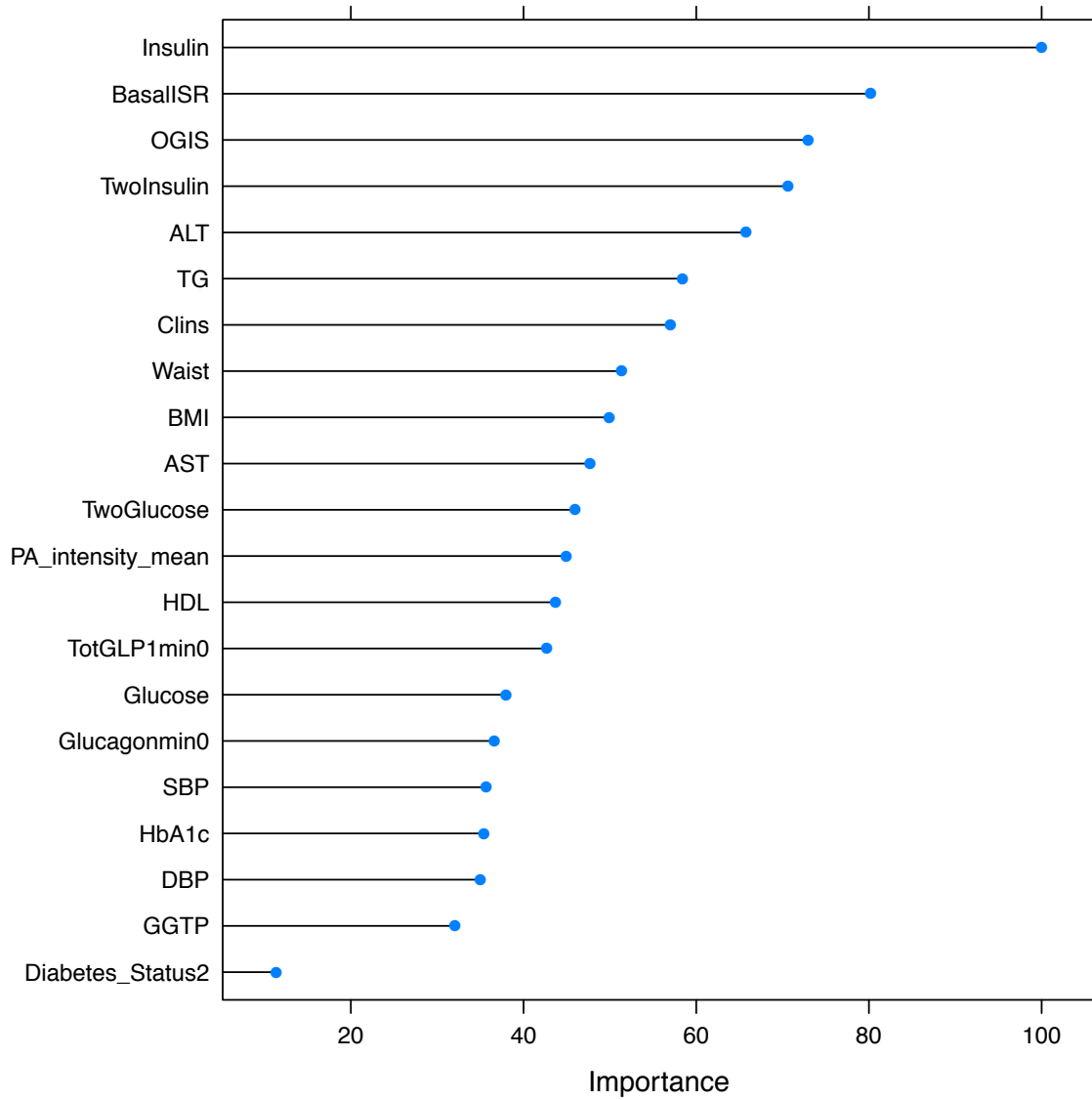
S4 Fig.



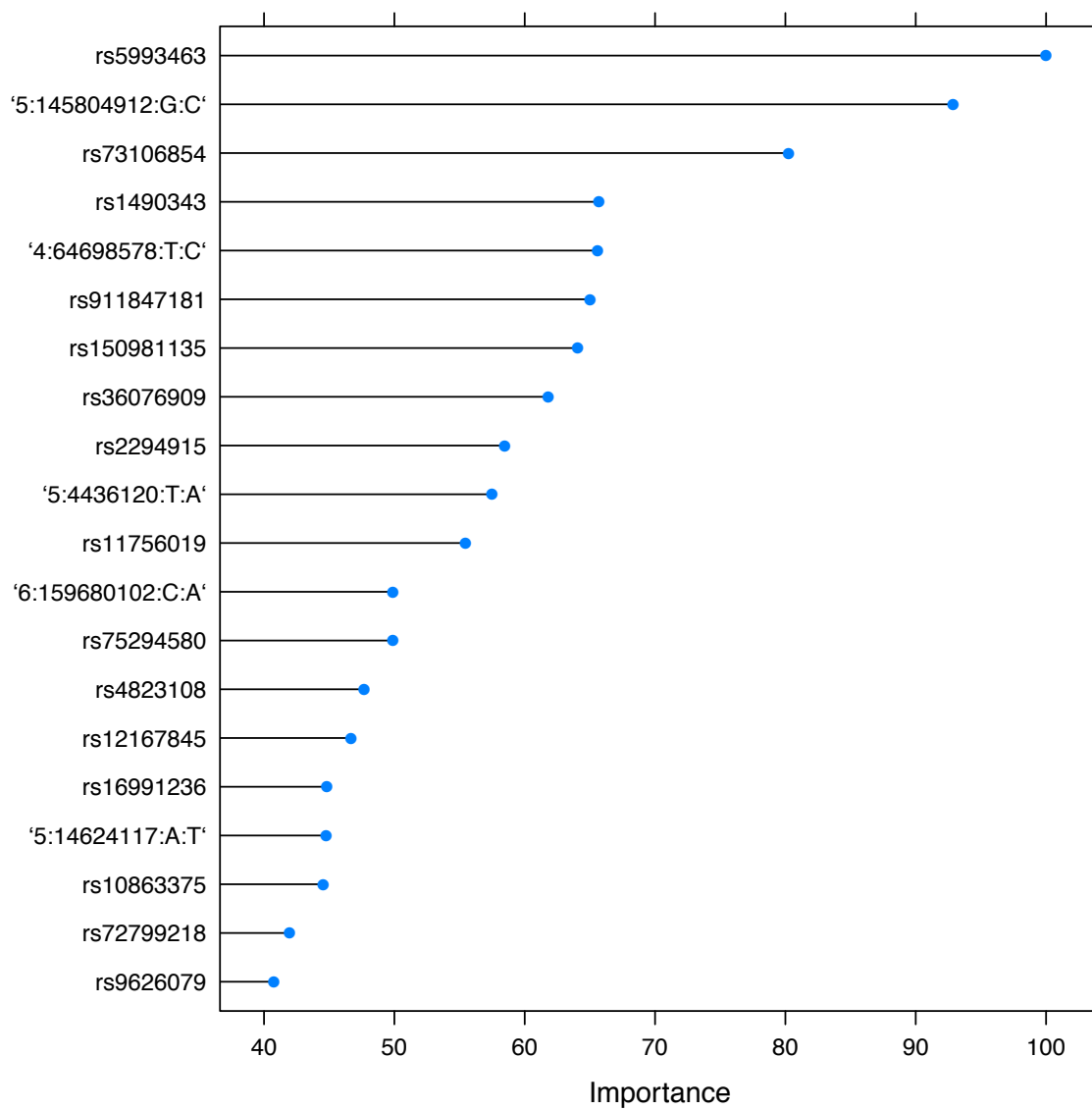
S5 Fig.



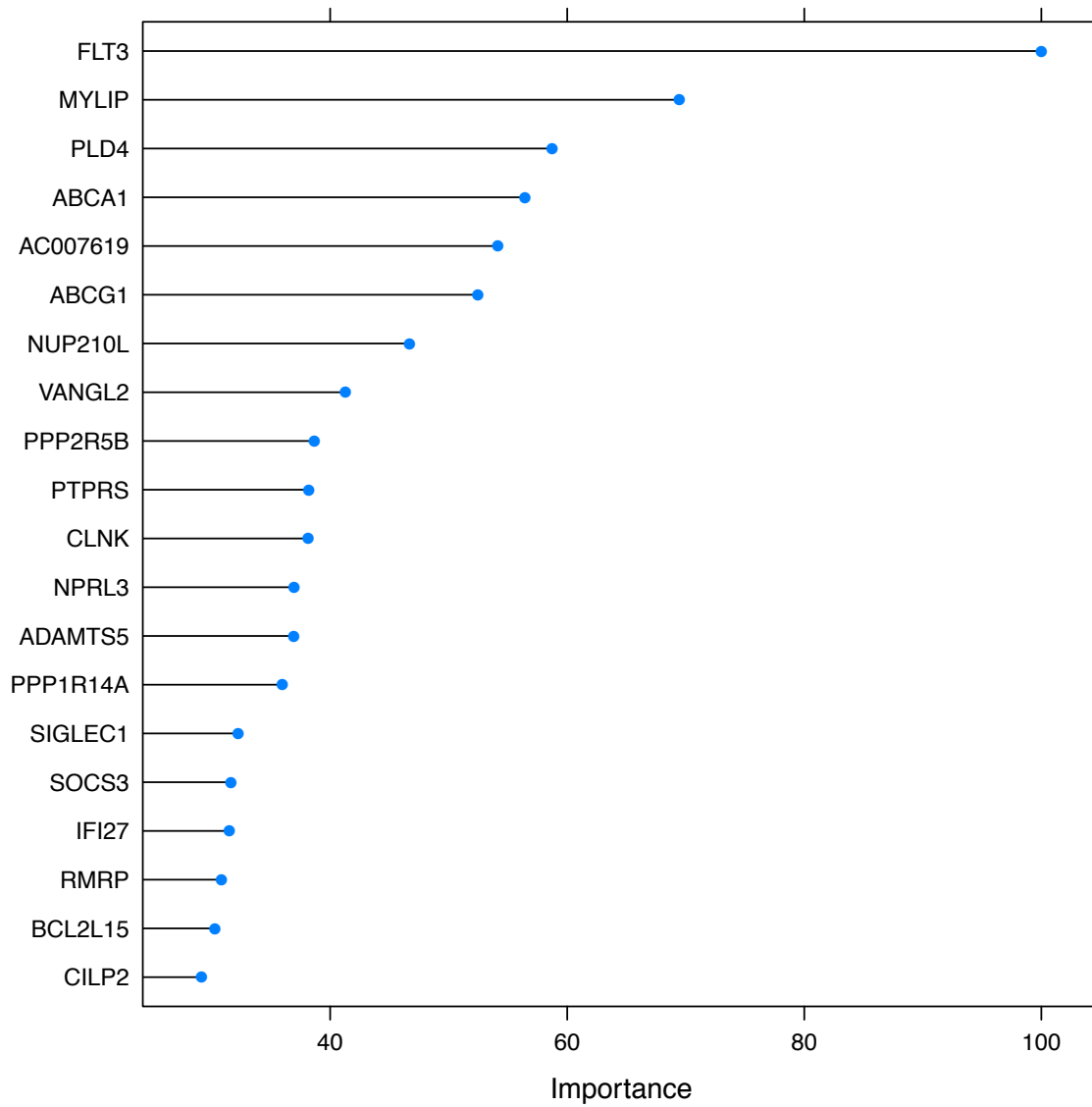
S6 Fig.



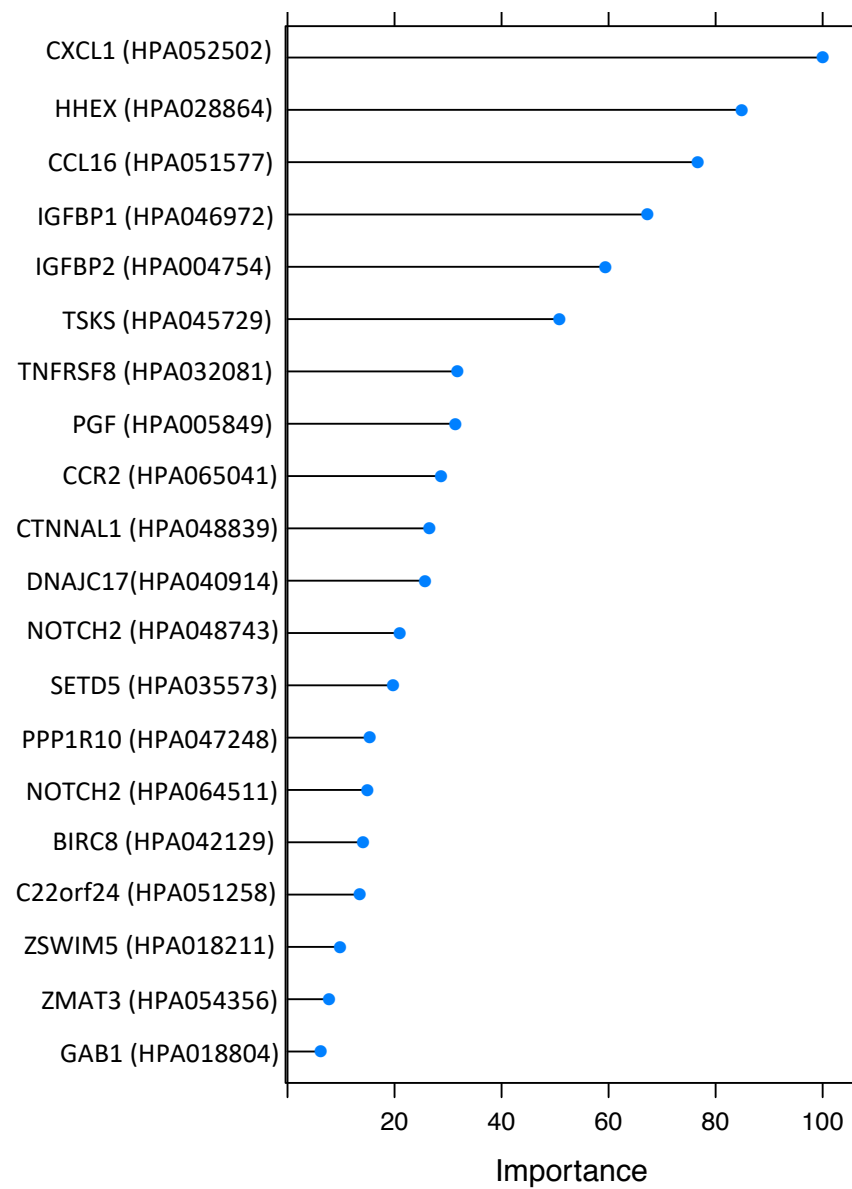
S7 Fig.



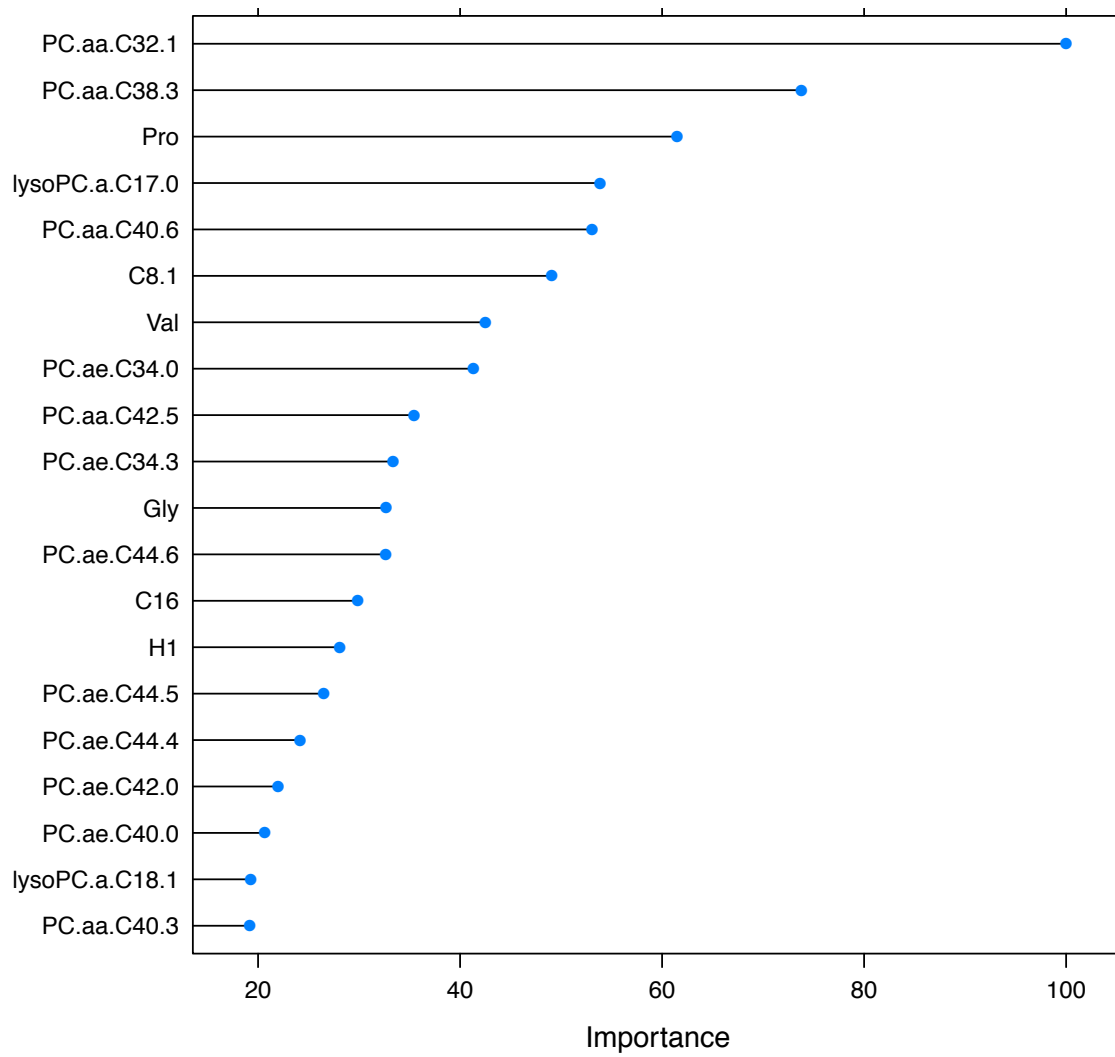
S8 Fig.



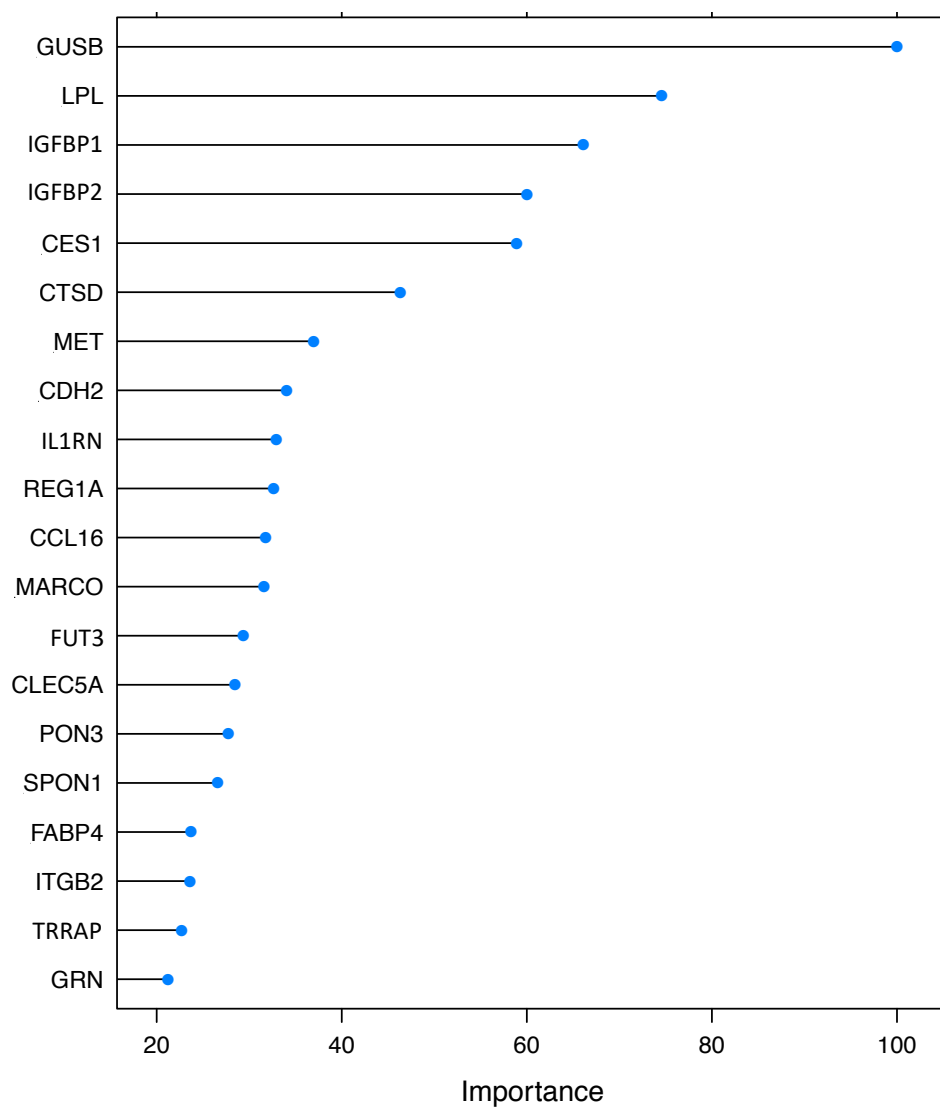
S9 Fig.



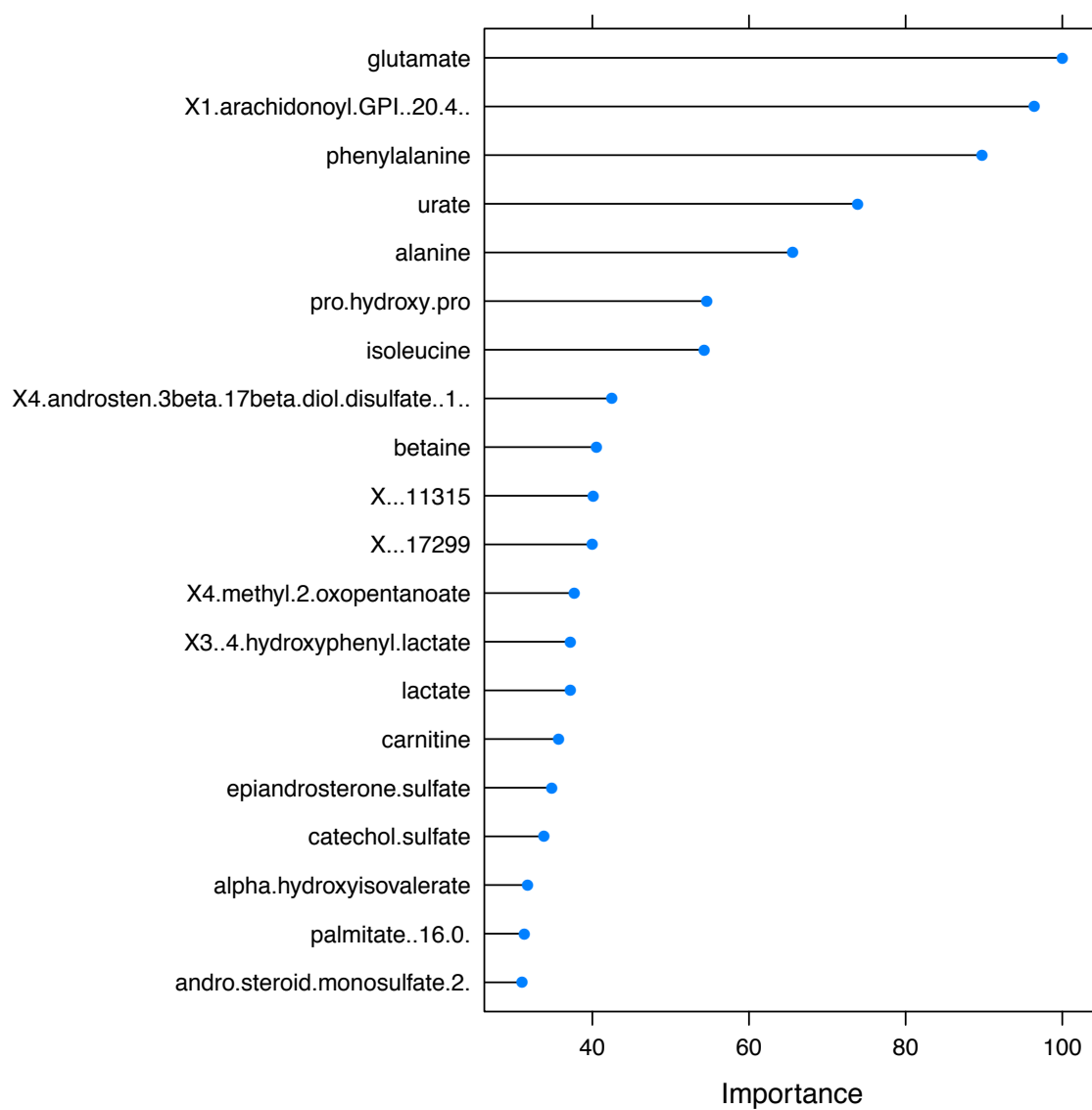
S10 Fig.



S11 Fig.



S12 Fig. Variable importance for the targeted proteomic model (only top 20)



S13 Fig. Variable importance for the untargeted metabolomic model (only top 20)

S1 Table.

Abbreviation	Meaning
Age	Age (yrs)
Sex	"male", "female"
Weight	Weight (nearest 0.1 kg)
Height	Height (nearest cm)
BMI	Body mass index - (kg/m ²)
Waist	Waist circumference (nearest cm)
Waist_Hip	Waist to hip ratio [Waist(cm)/Hip(cm)]
Diabetes_Status	"Non-diabetic", "diabetic"
Hx	Family history of T2D (parents/siblings/children) - yes, no
SBP	Mean systolic blood pressure (mm Hg)
DBP	Mean diastolic blood pressure (mm Hg)
alcohol_status	Consume alcohol "never", "occasionally", "regularly"
smoking_status	"never", "ex-smoker", "current smoker"
Hypercholesterolaemia	"yes", "no"
ALT	Alanine transaminase (U/L)
AST	Aspartate transaminase (U/L)
AST_ALT	AST to ALT ratio - (AST/ALT)
GGTP	Gamma-glutamyl transpeptidase (U/L)
HDL	Fasting high-density lipoprotein cholesterol (mmol/L)
LDL	Fasting low-density lipoprotein cholesterol (mmol/L)
Chol	Total cholesterol (mmol/L)
TG	Fasting triglyceride (mmol/L)

Glucose	Fasting glucose from venous plasma samples (mmol/L)
TwoGlucose	2-hour glucose (mmol/L) Frequently-sampled 75g oral glucose tolerance tests (OGTT) in pre-diabetics and mixed meal tolerance tests (MMTT) in diabetics
MeanGlucose	mean glucose during the OGTT/MMTT (mmol/L) calculated using trapezoidal integration (area/time)
HbA1c	Hemoglobin A1C (mmol/mol)
Insulin	Fasting insulin from venous plasma samples (pmol/L)
TwoInsulin	2-hour insulin (pmol/L)
MeanInsulin	mean insulin during the OGTT/MMTT (pmol/L) calculated using trapezoidal integration (area/time)
BasalISR	insulin secretion at the beginning of the OGTT/MMTT
TotalISR	integral of insulin secretion during the whole OGTT/MMTT
OGIS	Insulin sensitivity 2-hour OGIS ($\text{ml} \times \text{min}^{-1} \times \text{m}^{-2}$) according to the method of Mari et al
Stumvoll	insulin sensitivity index ($\text{ml} \times \text{min}^{-1} \times \text{kg}^{-1}$) according to the method of Stumvoll et al
Matsuda	insulin sensitivity index (arbitrary) according to the method of Matsuda et al
Clinsb	insulin clearance, calculated from basal values as (insulin secretion)/(insulin concentration)
Clins	mean insulin clearance during the OGTT/MMTT calculated as (mean insulin secretion)/(mean insulin concentration)
GlucoseSens	Glucose sensitivity, slope of the dose-response relating insulin secretion to glucose concentration

	($\text{pmol} \times \text{min}^{-1} \times \text{m}^{-2} \times \text{mmol}^{-1} \times \text{L}^{-1}$)
RateSens	Rate sensitivity, parameter characterizing early insulin secretion ($\text{pmol} \times \text{m}^{-2} \times \text{mmol}^{-1} \times \text{L}^{-1}$)
PFR	Potential Factor Ratio (dimensionless)
Glucagonmin0	Fasting Glucagon concentration (pg/ml)
IncGlucagonmin60	1-hour glucagon increment (pg/ml)
IncGLP1min60	1-hour GLP-1 increment (pg/ml)
ActGLP1min0	Concentration of fasting active GLP-1 in plasma (pg/ml)
TotGLP1min0	Concentration of fasting total GLP-1 in plasma (pg/ml)
PA_intensity_0_48f	Number of values in VM-HPF at ≥ 0 & ≤ 48 Physical activity – Accelerometry
PA_intensity_48_154f	Number of values in VM-HPF at ≥ 48 & ≤ 154 Physical activity – Accelerometry
PA_intensity_154_389f	Number of values in VM-HPF at ≥ 154 & ≤ 389 Physical activity – Accelerometry
PA_intensity_389_9999f	Number of values in VM-HPF at ≥ 389 & ≤ 9999 Physical activity – Accelerometry
PA_intensity_mean	Mean high-pass filtered vector magnitude mean physical activity intensity of the wear time of Accelerometry
TEI	total daily energy intake (kcal/day) based on validated multipass food habit questionnaire
ProteinI	total daily intake of dietary proteins (g/day) based on validated multipass food habit questionnaire
FatI	total daily intake of dietary fats (g/day)

	based on validated multipass food habit questionnaire
MUFatI	daily intake of dietary monounsaturated fats (g/day)
	based on validated multipass food habit questionnaire
PUFatI	daily intake of dietary polyunsaturated fats (g/day)
	based on validated multipass food habit questionnaire
SatFatI	daily intake of dietary saturated fats (g/day)
	based on validated multipass food habit questionnaire
CHOI	total daily intake of dietary carbohydrates (mg/day)
	based on validated multipass food habit questionnaire
SugarI	total daily intake of dietary sugar (mg/day)
	based on validated multipass food habit questionnaire
FibreI	total daily intake of dietary AOAC fibre (g/day)
	based on validated multipass food habit questionnaire

S2 Table.

Characteristics	Non-diabetic cohort		Diabetic cohort		Combined cohorts	
	MRI	Non-MRI	MRI	Non-MRI	MRI	Non-MRI
N (%)	1011 (45.30)	1223 (54.70)	503 (63.30)	292 (36.70)	1514 (49.99)	1515 (50.01)
Age (yr)	62 (56, 66)	63 (57.5, 67)	62 (56, 68)	64 (58, 68)	62 (56, 67)	63 (58, 67)
Sex, <i>n</i> (% female)	196 (19.40)	352 (28.80)	216 (42.90)	123 (42)	412 (27.20)	475 (31.40)
Weight (kg)	84.20 (76.25, 93.15)	84.20 (75.65, 93.80)	88.20 (77, 100)	89 (77.65, 101.88)	85.30 (76.5, 95.30)	84.90 (76, 95)
Waist circ. (cm)	99 (93, 106)	99 (92, 106)	103 (93, 112)	103 (94.25, 111)	100 (93, 109)	100 (93, 107)
BMI (kg/m ²)	27.44 (25.45, 29.90)	27.65 (25.65, 30.35)	29.89 (26.81, 34.08)	29.89 (26.74, 33.88)	28.02 (25.82, 31.16)	27.94 (25.47, 31.16)
SBP	131.33 (121.33, 142)	128 (119, 140)	129.17 (120.33, 139)	132.33 (121.33, 141)	130.67 (121.33, 140.67)	128.67 (119.33, 140.33)
DBP	81.67 (77, 87.33)	79.67 (73, 85)	76 (69.33, 82.67)	74.33 (68.33, 82.67)	80.33 (74, 86)	79.33 (72.25, 84.67)
HbA1c (mmol/mol)	37 (36, 39)	37 (35, 39)	46 (43, 50)	46 (43, 50)	39 (36, 44)	38 (36, 41)
Fasting glucose (mmol/L)	5.80 (5.5, 6.1)	5.70 (5.3, 6.1)	6.90 (6.1, 7.8)	7.30 (6.6, 8)	6 (5.6, 6.6)	5.80 (5.4, 6.4)
Fasting insulin (pmol/L)	55.20 (32.40, 78.60)	59.40 (34.20, 91.20)	86 (55.80, 135.80)	90.20 (62.08, 131)	62.40 (37.80, 94.90)	64.20 (37.80, 100.20)
2hr glucose (mmol/L)	5.90 (4.90, 7.15)	5.80 (4.80, 7)	8.60 (6.60, 10.38)	8.40 (6.70, 10.40)	6.50 (5.20, 8.40)	6.10 (5, 7.70)
2hr insulin (pmol/L)	210.60 (118.20, 366)	207 (115.80, 376.80)	386.60 (219.60, 610.50)	346.10 (212.20, 519.20)	256.20 (137.40, 455.20)	231.60 (127.80, 422.40)
Triglycerides (mmol/L)	1.24 (0.95, 1.67)	1.23 (0.935, 1.675)	1.30 (0.95, 1.81)	1.42 (1.05, 2.05)	1.25 (0.95, 1.72)	1.27 (0.95, 1.75)
ALT (U/L)	16 (12, 23)	14 (10, 21)	22 (18, 29)	25 (19, 33)	19 (13, 26)	16 (10, 24)
AST (U/L)	26 (22, 32)	25 (21, 31)	23 (20, 28)	23 (19, 29)	25 (21, 31)	25 (21, 30)
Alcohol intake, <i>n</i> ("never", "occasionally", "regularly")	111, 201, 698 (11,20,69)	160, 240, 820 (13,20,67)	90, 126, 287 (18,25,57)	44, 73, 174 (15,25,60)	202, 327, 985 (13,22,65)	204, 313, 994 (13,21,66)
Liver fat	3.30 (1.80, 6.60)	NA	6.10 (3.20, 12.35)	NA	4.20 (78.50, 123.20)	NA
Fatty liver, <i>n</i> (% yes)	344 (34)	NA	296 (58.80)	NA	640 (42.30)	N

S3 Table.

[illegible]

S4 Table.

UK Biobank Field number	Description
22402	Liver fat percentage
31	Sex
21001	BMI (instance 2, at imaging visit)
48	Waist circumference (instance 2, at imaging visit)
4080	Systolic blood pressure, automated reading (mean value if 2 values are available)
93	Systolic blood pressure, manual reading (if Automated values are not available)
4079	Diastolic blood pressure, automated reading
94	Diastolic blood pressure, manual reading (if Automated values are not available)
2443	Diabetes diagnosed by doctor (instance 2, at imaging visit)
30620	Alanine aminotransferase (ALT)
30650	Aspartate aminotransferase (AST)
30870	Triglycerides
30730	Gamma glutamyltransferase (GGT)
30740	Glucose
30750	Glycated haemoglobin (HbA1c)

S5 Table.

Models	Diabetic cohort	Non-diabetic cohort
1	0.70 (95% CI=0.68, 0.72)	0.70 (95% CI=0.69, 0.72)
2 (fasting glucose)	0.74 (95% CI=0.71, 0.75)	0.78 (95% CI=0.76, 0.79)
2 (HbA1c)	0.73 (95% CI=0.71, 0.75)	0.78 (95% CI=0.76, 0.79)
3	0.79 (95% CI=0.77, 0.81)	0.80 (95% CI=0.78, 0.81)
FLI	0.76 (95% CI=0.71, 0.80)	0.76 (95% CI=0.73, 0.79)
HSI	0.75 (95% CI=0.70, 0.79)	0.72 (95% CI=0.68, 0.75)
NAFLD-LFS	0.80 (95% CI=0.77, 0.85)	0.77 (95% CI=0.74, 0.80)