

ELECTRONIC SUPPLEMENTARY MATERIAL

RESEARCH DESIGN AND METHODS

Study Population

The prospective SWIFT cohort enrolled a racially and ethnically diverse group of 1,035 women (age 20 – 45 years) who were diagnosed with GDM (via a 3-h 100g OGTT, in accordance with Carpenter and Coustan criteria) and delivered singleton pregnancies ≥ 35 weeks of gestation at KPNC hospitals from 2008 to 2011 [30, 38]. Study recruitment, selection criteria, methodologies, and all other detailed information have been described previously [38-40]. Each SWIFT participant provided written consent to attend three in-person study exams that included 2-hour 75 grams OGTTs under research protocols; the initial exam was at 6 to 9 weeks postpartum, with follow-up exams at one and two years postpartum. At each study exam, trained research staff collected blood samples at both fasting and 2-h time points during 75 grams OGTT and completed anthropometric and other assessments. At the same time, plasma samples were processed from all blood samples and kept at -80°C for future studies.

Study Design

For this study, we selected the incident diabetes cases among Hispanic and Asian groups, and pair-matched them to control women without diabetes by age, race and ethnicity, pre-pregnancy BMI, and glucose tolerance at 6-9 weeks postpartum (normal or impaired). The aim of this pair-match study design was to minimize the influence of clinical variables from the results of both predictive signature as well as pathophysiology. We designed a 1:1.5 pair-match study by selecting 55 cases [Hispanic (n=30) and Asian (n=25) women] who were pair-matched with 85 controls [Hispanic (n=60) and Asian (n=25) women]. The inclusion of multiple races in the study was to ensure homogeneity of race and ethnic groups. In this study, 25 Asian cases were pair-matched with 25 Asian controls in 1:1 pair-matching, whereas 30 Hispanic cases were pair-matched with 60 Hispanic controls in 1:2 pair-matching. Overall, the study had 1:1.5 pair-matched design. The greater number of participants of Hispanic background was due to their pair-match ability within SWIFT cohort, and does not represent the ethnicity-related T2D risk in this study.

The fasting plasma samples were collected from all women at the “baseline” exam (6 to 9 weeks postpartum), from among the 1,010 participants confirmed not to have T2D at the baseline exam via the 2-hr 75 g OGTT. Details of the SWIFT prospective cohort design and follow up are published elsewhere. (53, 54). If women later progressed to T2D during the two-year follow-up period (termed here as the “follow-up” time-point), these newly diagnosed incident T2D cases (n=55) are referred to as “case” (also termed as “T2D” interchangeably). On the other hand, women who did not develop T2D during the follow-up period (n=85) are referred to as “control” (also termed as “non-T2D” interchangeably) (**Figure-1**).

Targeted Lipid Profiling (Targeted-Lipidomics)

Targeted-lipidomics Design

The fasting plasma samples collected at 6-9 weeks postpartum from the SWIFT Study were sent to Metabolon, Inc. (Morrisville, NC) for analysis via a targeted approach combining gas (GC) and liquid phase (LC) chromatography. A single-blind targeted-lipidomics analysis (LC-MS) of 1100 lipid species (from natural lipids, phospholipids, and sphingolipids) was performed on each plasma sample at Metabolon Inc (NC, USA). In this study, the natural lipid group consisted of 26 species of cholesterol esters (CE), 59 species of diacylglycerol (DAG), 493 species of triacylglycerol (TAG), and 26 species of free fatty acids (FFA). The phospholipid group consisted of 140 species of phosphatidylcholine (PC), 216 species of phosphatidylethanolamine (PE), 28 species of phosphatidylinositol (PI), 26 species of lysophosphatidylcholine (LPC), and 26 species of lysophosphatidylethanolamine (LPE). The sphingolipid group consisted of 12 species of dihydroceramide (DCER), 12 species of ceramide (CER), 12 species of hexosylceramide (HCER), 12 species of lactosylceramide (LCER), and 12 species of sphingomyelin (SM).

Complex Lipids Extraction

Complex lipid analysis was performed at Metabolon, Inc. Briefly, 75 µl of each plasma sample was extracted using either a modified Bligh-Dyer protocol or an automated BUMÉ extraction according to the method described elsewhere. Samples from both extraction modes were then dried down under nitrogen and reconstituted in 0.25mL of dichloromethane: methanol (50:50) containing 10 mM ammonium acetate.

Mass-spectroscopy Analysis

Samples were placed into vials or a glass-lined 96-well plate for flow injection and mass spectrometry (FIA-MS). FIA-MS was performed on a SCIEX 5500 QTRAP equipped with a SelexION Differential Mobility Separation (DMS) cell, and was operated in Multiple Reaction Monitoring (MRM) mode using both positive and negative mode electrospray in a Turbo V ion source. Each sample was subjected to two analyses, with 50 μ L of sample injected at a flow rate of 7 μ L/min for each analysis. In the first analysis, 472 MRM pairs corresponding to 448 endogenous lipids and 24 internal standards were monitored. This includes PCs, PEs, LPCs, LPEs, and PIs (negative ion mode), as well as SMs (positive ion mode), and the SelexION was used to apply Compensation Voltages (CoV) optimized for each lipid class, using *n*-propanol as the DMS modifier and the Separation Voltage set to 3500 V. In analysis 2, 706 MRM pairs corresponding to 676 endogenous lipids and 30 internal standards were monitored, comprising free fatty acids (negative ion mode) as well as CEs, DAGs, TAGs, CERs, DCERs, GCERs, and LCERs (positive ion mode), and the SelexION was not used. Both analyses included 20 MRM cycles with 20 msec per MRM pair, a settling time of 50 msec, and a pause between mass ranges of 5 msec. For quantitation, the concentration of each target compound was calculated using the ratio of signal intensity for the target compound to that of the assigned deuterated internal standard of known concentration. At least one deuterated standard was present for each lipid class, with the exception of PIs (which were quantified using the deuterated PE standard).

Data Preparation and Statistical Analysis of the Quality of the Final Dataset

A stringent data acquiring option was applied. Lipid metabolites with missing values (below the detection limit) were removed from dataset. The dataset was subjected to quantile normalization followed by log₂ transformation in the Metaboanalyst 3.0 platform. The normalized dataset showed semi-normal distribution for which non-parametric analysis is preferable. Batch effects, due to several runs and imbalance in glucose tolerance (e.g., control had a relatively higher number of impaired glucose tolerance members in the study population), were adjusted for by using 6 QC samples and the standard protocols of combat algorithm within Metaboanalyst 3.0. These 6 QC samples which were taken from the first shipment were rerun with the samples of the second shipment. All of these 6 QC samples were pair-matched Hispanic. It constructed three pairs of QC

samples. Out of these three pairs, two pairs were NGT (normal glucose tolerance) and one pair was IGT (impaired glucose tolerance). The final dataset was scrutinized through a Principal Component Analysis (PCA), and a Partial Least Squares- Discriminant Analysis (PLS-DA).

PCA is an unsupervised statistical method where multidimensional data (e.g., data of variables) are orthogonally clustered to determine the contributing factors (which are known as principal components). Since it is an unsupervised method, it does not name the factors but rather designates them as principal component 1 (PC1), principal component 2 (PC2), and so on based on their percentage of contributions. In PCA, we had attempted to find out the existence of influencing factors with their contribution percentage.

On the other hand, PLS-DA is a supervised method where multidimensional data are orthogonally clustered based on known factors such as case and control. In PLS-DA, we wanted to know how much class separation existed between case and control. PLS-DA can easily overfit the data since it is a supervised method and thus warrants rigorous validation. This validation however is far from a straightforward, single-step approach. The predictivity of the PLS-DA model is estimated by Q^2 -calculation. Metaboanalyst 3.0 platform has a default 10-fold CV based Q^2 -calculation option for the PLS-DA model. However, the Q^2 depends not only on the between-class separation but also on the within-class variability. This makes it difficult to give a general Q^2 value that corresponds to a good classification. That is why it is very important to do permutation analysis to prove that the observed class separation was not a random event. In this study, we did this permutation analysis through application of the empirical Bayes algorithm which is an integral part of PLS-DA modelling in Metaboanalyst 3.0.

Differential Expression Analysis and Pathway Analysis

A non-parametric test (Wilcoxon-Mann-Whitney test, α -value set at $p < 0.05$) followed by multiple comparisons with false discovery rate analysis (FDR, α -value set at $p < 0.05$) was carried out to identify the differentially expressed lipid metabolites between case and control using Metaboanalyst 3.0. These differentially expressed lipid metabolites were deployed for the pathway analysis by adopting two approaches; (1) a direct approach where differentially expressed lipid metabolites were used, and (2) an *in-silico* approach where the interacting proteins with the differentially expressed lipid metabolites were used.

In the direct approach, the differentially expressed metabolites were used for both KEGG pathway analysis and metabolite set enrichment analyses (MSEA) by using a “reference metabolome based on your analysis platform” module in Metaboanalyst 3.0 under a *Homo sapiens* background. The hypergeometric test algorithm for over-representation analysis and the relative-betweenness centrality algorithm for the pathway topology analysis were set for the KEGG pathway analysis. The metabolite set enrichment pathways (MSEP) were identified in Metaboanalyst 3.0 by using the pathway-associated metabolite set library.

Biologically, enzymes (proteins) interact with metabolites to transform them into new metabolites. Therefore, the concentration change of certain metabolites involves the action of certain enzymes. These enzymes have been designated here as “interacting proteins”. MBrole 2.0 is a bioinformatic platform which allows us to identify these interacting proteins. In the *in-silico* approach, the interacting proteins from the human background were identified through the MBrole2.0 platform by setting the FDR <0.005. These proteins were deployed into the STRING 10.5 platform under k-means clustering to identify the involved KEGG pathways. The acceptance of the pathway was strictly scrutinized by putting FDR < 0.05.

Predictive Analytics

The biomarker analysis module of Metaboanalyst 3.0 was used for univariate ROC analysis. The univariate ROC analysis calculated the AUC-ROC for an individual lipid metabolite using the default setting of Metaboanalyst 3.0. The 95% confidence interval was calculated using 500 bootstrapping. In the multivariate ROC analysis, the stepwise (both ways) multiple logistic regression (MLR) was carried out in R-studio using the “glm” function under significant contributor calculation. The final ROC curve was constructed under 95% confidence intervals (R-script is available in the supplementary materials section).

Machine learning analyses were carried out through WEKA 3.8 (University of Waikato NZ). Application of the filtered classifier (FC) algorithm exhibited a promising AUC-ROC at the initial screening stage where the suitability of different machine learning algorithms with our dataset was checked. FC is a supervised multiple classifier algorithm in Weka 3.8. It uses the “J48” algorithm and the “Discretize -R first-last -precision 6” algorithm together as a classifier and filter, respectively. It transforms the numeric attributes to nominal (e.g., yes and no), and considers the

values of classes (e.g., here case was “1” and control was “0”) in the splitting and bagging process. The FC algorithm constructs a decision tree consisting of nodes and leaves. A node is a branch-point that can lead to other nodes or to a leaf. A leaf is an endpoint with no further branching, as a decision or prediction has been achieved for these sets of samples. Other samples which are still undecided travel further down the nodes and branch until they reach a leaf.

The final classifier (here FC) was further optimized for balancing between the chance of data overfitting, higher ROC possibility, and F-measurement. The optimization was carried out by K-fold cross-validation (CV) with confident threshold 1.0 and binary output selection. The K-fold cross-validation (K-fold CV) is the validation method of choice for small population sizes (e.g., 140 participants in this case). Due to the small number of samples, care must be taken to optimize the value of K. Higher-folds (K) can lead to an overfit-model while low values fit poorly or produce highly variable and biased models. Our predictive model was rigorously scrutinized for its tolerability over a large range of K (up to 100-fold CV)., We chose a K of 45-fold CV as optimized by using the “one standard error rule” since we found that 45-fold CV had a relatively low error rate with an insignificant loss of predictability from the highest possible discrimination. Moreover, this 45-fold CV was not at the saturation point of accuracy (i.e., 55-fold to 90-fold CV was the saturation point in this case), suggesting our predictive model is not overfitted.

The confident threshold indicates the degree of pruning. In general, pruning means trimming by cutting away dead or overgrown branches or stems of a tree, especially to increase fruitfulness and growth. In predictive analytics, a smaller confidence threshold incurs less pruning of the decision tree. It is prone to result in data-overfitting due to not cleaning up the bias properly. Thus, the confident threshold 1.0 indicates its highest degree of pruning to guard against data-overfitting and bias selection. The binary output also provides a degree of protection against overfitting and bias selection. It allows the variables/attributes to be divided into two decisions (e.g., yes or no) with a single definitive line.

R-scripts for stepwise multiple logistic regression with example of lipids

```
> View(Top_50)
> myData = Top_50
> gr.glm<-glm(CLASS ~ ., data = myData, family = binomial)
> #this applies the model to generate predictions on the data
```

```

> gr.pred<-predict(gr.glm, type="response")
> gr.pred.p<-predict(gr.glm, type="response")
> gr.pred<-factor(gr.pred.p>0.5, labels=c("0", "1"))
> table(myData$CLASS, gr.pred)
  gr.pred
    0  1
0 75 10
1 16 39
> class.summary <- function(y, pred, pred.prob){
+   tab <- table(y=y, pred=pred)           # Confusion matrix
+   acc <- (sum(diag(tab))) / sum(tab)     # Accuracy rate
+   mis <- (sum(tab) - sum(diag(tab)))/sum(tab) # Misclassification Rate
+   # Deviance
+   if(any(pred.prob == 0))
+     pred.prob[pred.prob == 0] <- 0.001
+   if(any(pred.prob == 1))
+     pred.prob[pred.prob == 1] <- 1 - 0.001
+   dev <- -2*sum((y=="S")*log(pred.prob) + (1-(y=="S"))*log(1-pred.prob))
+   return(list(conf.tab=tab, acc=acc, mis=mis, dev=dev))
+ }
> class.summary(myData$CLASS, gr.pred, gr.pred.p)
$conf.tab
  pred
y    0  1
0 75 10
1 16 39

$acc
[1] 0.8142857

$mis
[1] 0.1857143

$dev
[1] 276.7928

> summary(gr.glm)

Call:
glm(formula = CLASS ~ ., family = binomial, data = myData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max

```

-2.1536 -0.6386 -0.1692 0.4728 2.0507

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	49.0102	52.4290	0.935	0.3499
A1	-2.5833	1.3871	-1.862	0.0625 .
A2	-1.5381	1.2213	-1.259	0.2079
A3	1.1694	1.3942	0.839	0.4016
A4	-0.2614	0.1408	-1.857	0.0634 .
A5	-0.1648	0.1899	-0.868	0.3855
A6	-0.2372	1.3703	-0.173	0.8625
A7	0.9530	1.2205	0.781	0.4349
A8	1.1919	1.7206	0.693	0.4885
A9	-3.0808	2.2823	-1.350	0.1771
A10	-1.9902	1.4350	-1.387	0.1655
A11	4.2602	2.1987	1.938	0.0527 .
A12	-2.6629	1.2163	-2.189	0.0286 *
A13	-0.8558	1.4773	-0.579	0.5624
A14	4.1653	2.3776	1.752	0.0798 .
A15	-1.7590	1.6615	-1.059	0.2898
A16	-0.7433	1.7736	-0.419	0.6752
A17	-0.2782	1.3647	-0.204	0.8385
A18	2.2807	4.2245	0.540	0.5893
A19	-1.0462	4.1127	-0.254	0.7992
A20	-4.7120	2.8567	-1.649	0.0991 .
A21	7.7800	4.3613	1.784	0.0744 .
A22	1.1988	3.4069	0.352	0.7249
A23	-3.0053	3.1378	-0.958	0.3382
A24	-4.9202	3.1815	-1.547	0.1220
A25	-0.5574	3.4620	-0.161	0.8721
A26	5.0094	3.5916	1.395	0.1631
A27	0.7099	2.5666	0.277	0.7821
A28	1.5095	1.9979	0.756	0.4499
A29	2.4170	4.3592	0.554	0.5793
A30	-1.9669	3.7579	-0.523	0.6007
A31	-3.4791	2.6244	-1.326	0.1850
A32	2.3707	3.2388	0.732	0.4642
A33	-0.1195	3.4725	-0.034	0.9725
A34	0.5942	2.4654	0.241	0.8095
A35	-0.5071	2.9560	-0.172	0.8638
A36	-5.8998	4.5546	-1.295	0.1952
A37	1.6478	3.5302	0.467	0.6407
A38	2.1195	4.1502	0.511	0.6096

A39	-6.8852	4.8569	-1.418	0.1563
A40	-2.6694	2.9520	-0.904	0.3658
A41	1.3444	0.9765	1.377	0.1686
A42	0.9490	1.5847	0.599	0.5493
A43	2.2931	2.5512	0.899	0.3687
A44	-1.4826	1.6530	-0.897	0.3698
A45	-0.1120	1.2867	-0.087	0.9306
A46	5.3134	2.4019	2.212	0.0270 *
A47	2.4145	2.2166	1.089	0.2760
A48	3.6136	1.7794	2.031	0.0423 *
A49	-2.1703	1.8337	-1.184	0.2366
A50	0.4947	1.4096	0.351	0.7256

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 187.6 on 139 degrees of freedom
 Residual deviance: 108.0 on 89 degrees of freedom
 AIC: 210

Number of Fisher Scoring iterations: 6

```
> class.summary(myData$CLASS, gr.pred, gr.pred.p)
```

```
$conf.tab
```

```
  pred
y  0  1
0  75 10
1  16 39
```

```
$acc
```

```
[1] 0.8142857
```

```
$mis
```

```
[1] 0.1857143
```

```
$dev
```

```
[1] 276.7928
```

```
> #here is the step optimization of the model generated above
```

```
> stepGR<-step(gr.glm, direction = "both", k=2, trace=0)
```

```
> step.gr.pred<-predict(stepGR, type="response")
```

```
> step.gr.pred.p<-predict(stepGR, type="response")
```

```

> step.gr.pred<-factor(step.gr.pred.p>0.5, labels=c("0", "1"))
> #class summary of the optimized model
> class.summary(myData$CLASS, step.gr.pred, step.gr.pred.p)
$conf.tab
  pred
y   0  1
0  73 12
1  20 35

$acc
[1] 0.7714286

$mis
[1] 0.2285714

$dev
[1] 202.3963

> summary(stepGR)

```

```

Call:
glm(formula = CLASS ~ A1 + A4 + A12 + A20 + A24 + A26 + A39 +
     A41 + A42 + A46 + A48 + A49, family = binomial, data = myData)

```

```

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1399 -0.7712 -0.3709  0.7496  2.4520

```

```

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 26.45346   17.17344   1.540  0.12347
A1           -1.33257    0.66823  -1.994  0.04613 *
A4           -0.17442    0.08941  -1.951  0.05108 .
A12          -1.41523    0.61921  -2.286  0.02228 *
A20          -2.65083    1.55071  -1.709  0.08737 .
A24          -1.17480    0.76614  -1.533  0.12518
A26           4.50447    1.88444   2.390  0.01683 *
A39          -3.96193    1.50825  -2.627  0.00862 **
A41           1.36087    0.49478   2.750  0.00595 **
A42           0.90311    0.45713   1.976  0.04820 *
A46           3.39550    1.12921   3.007  0.00264 **
A48           2.16710    1.05897   2.046  0.04072 *
A49          -1.52842    0.99100  -1.542  0.12300

```

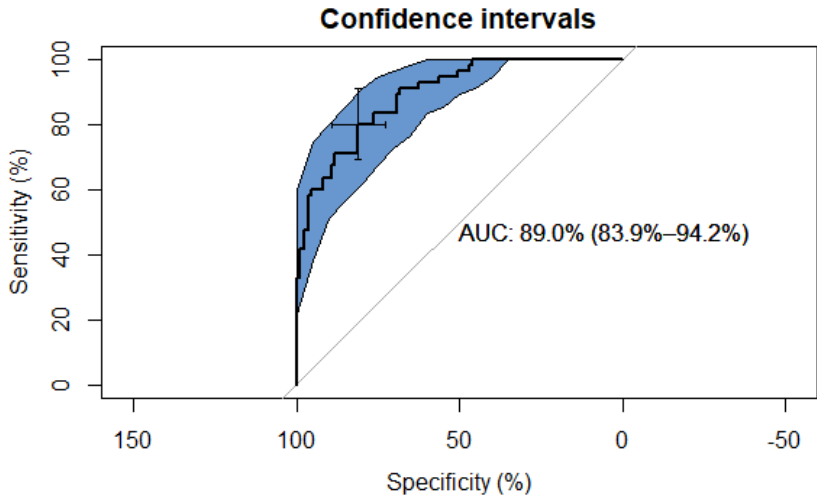
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

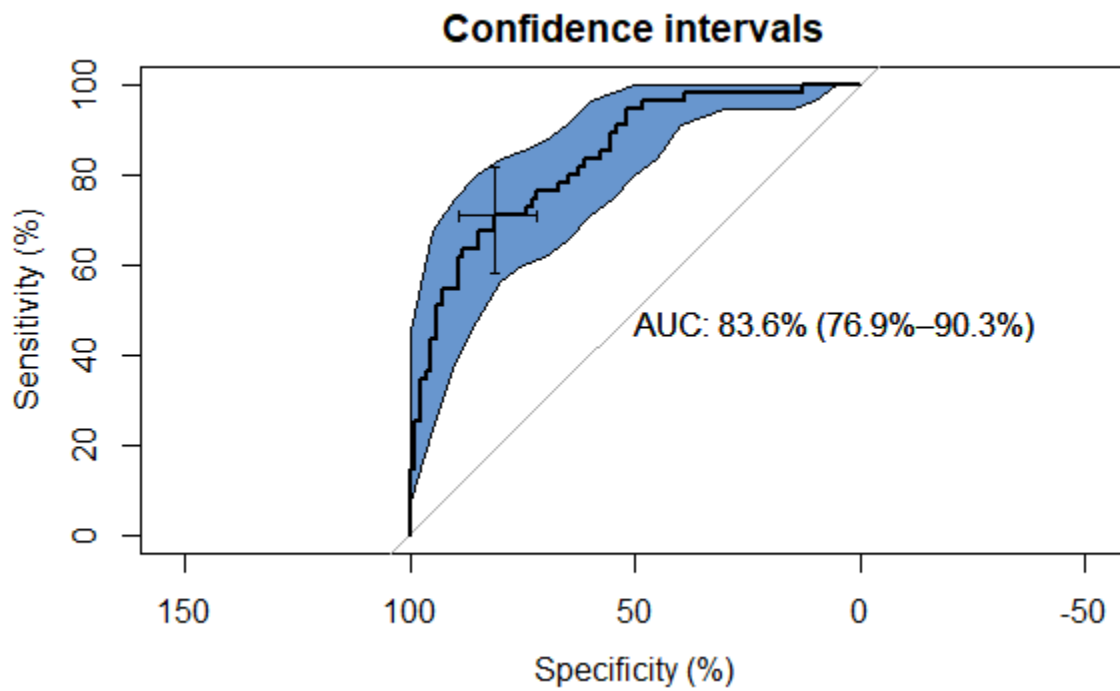
Null deviance: 187.60 on 139 degrees of freedom
Residual deviance: 134.15 on 127 degrees of freedom
AIC: 160.15

Number of Fisher Scoring iterations: 5

```
> # the AUC curve plots with confidence intervals, these are from the home ma
de function below and using library(pROC)
> library(pROC)
> plotROCconf <- function(outcome, prob) {
+   rocobj <- plot.roc(
+     outcome,
+     prob,
+     main = "Confidence intervals",
+     percent = TRUE,
+     ci = TRUE,
+     # compute AUC (of AUC by default)
+     print.auc = TRUE
+   ) # print the AUC (will contain the CI)
+   ciobj <- ci.se(rocobj, # CI of sensitivity
+                 specificities = seq(0, 100, 5)) # over a select set of spe
cificities
+   plot(ciobj, type = "shape", col = "#1c61b6AA") # plot as a blue shape
+   plot(ci(rocobj, of = "thresholds", thresholds = "best")) # add one thresh
old
+   cex.axis=5
+ }
> plotROCconf(myData$CLASS, gr.pred.p)
```



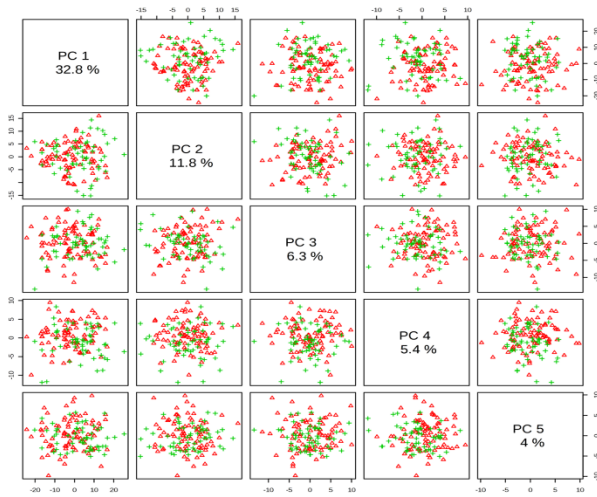
```
> plotROCconf(myData$CLASS, step.gr.pred.p)
```



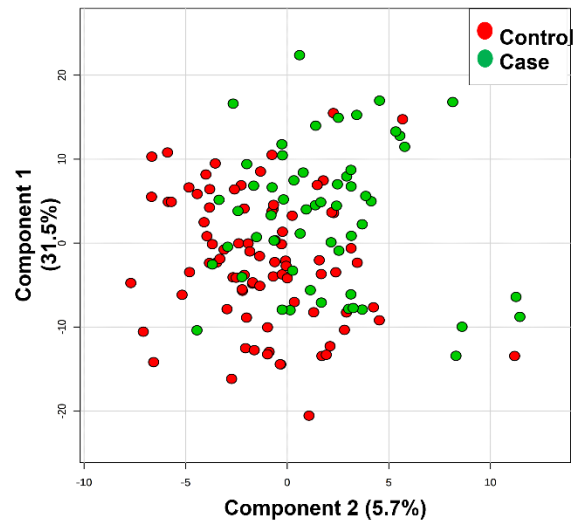
ELECTRONIC SUPPLEMENTARY FIGURES

ESM Fig. 1: The quality control of final dataset

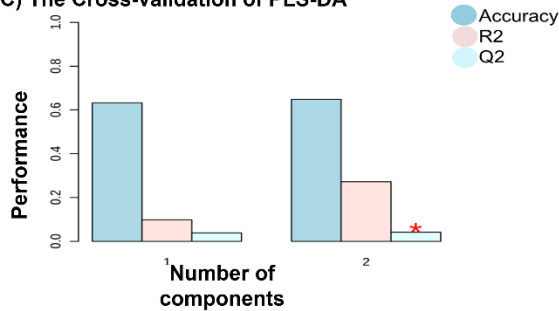
(A) The principal component analysis (PCA)



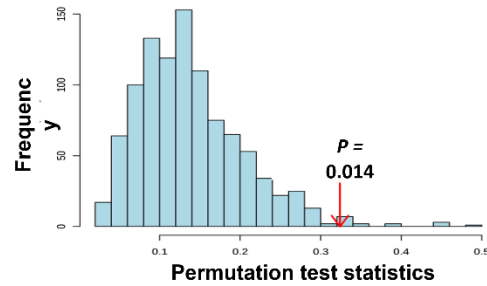
(B) 2D score plot of PLS-DA



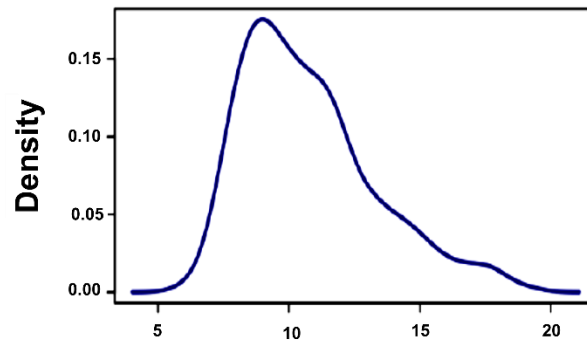
(C) The Cross-validation of PLS-DA



(D) The Empirical analysis



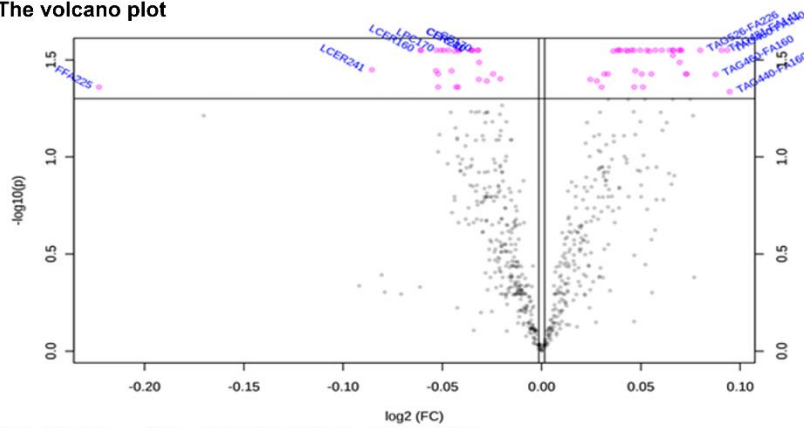
(E) Data distribution after normalization and transformation



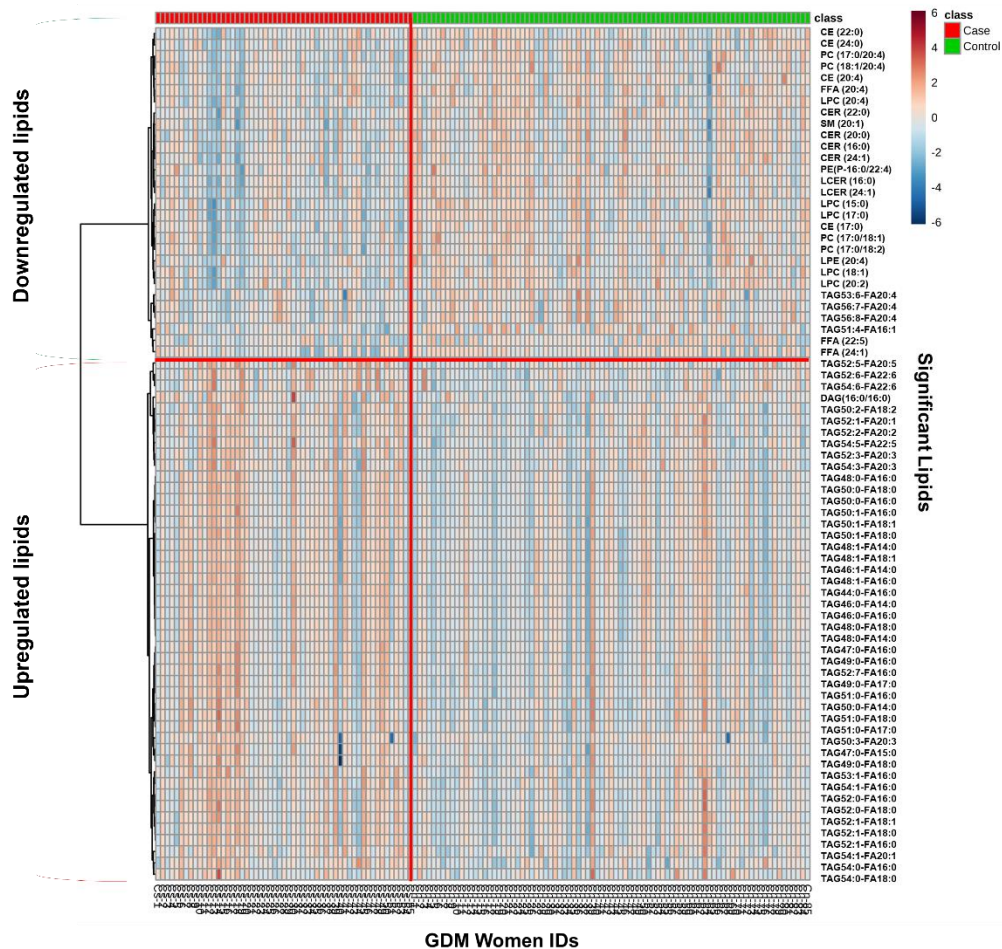
ESM Fig.1: The quality control of final data. (A) The contribution of different components in PCA has been shown here. (B) The distribution of two groups (Case and Control) in PLS-DA. (C) A cross-validation analysis of the performance of PLS-DA based on R^2 and Q^2 . (D) A permutation analysis for PLS-DA. (E) The distribution of (quantile) normalized and log2 transformed data.

ESM Fig.2: The volcano plot and heat map of the final dataset

(A) The volcano plot

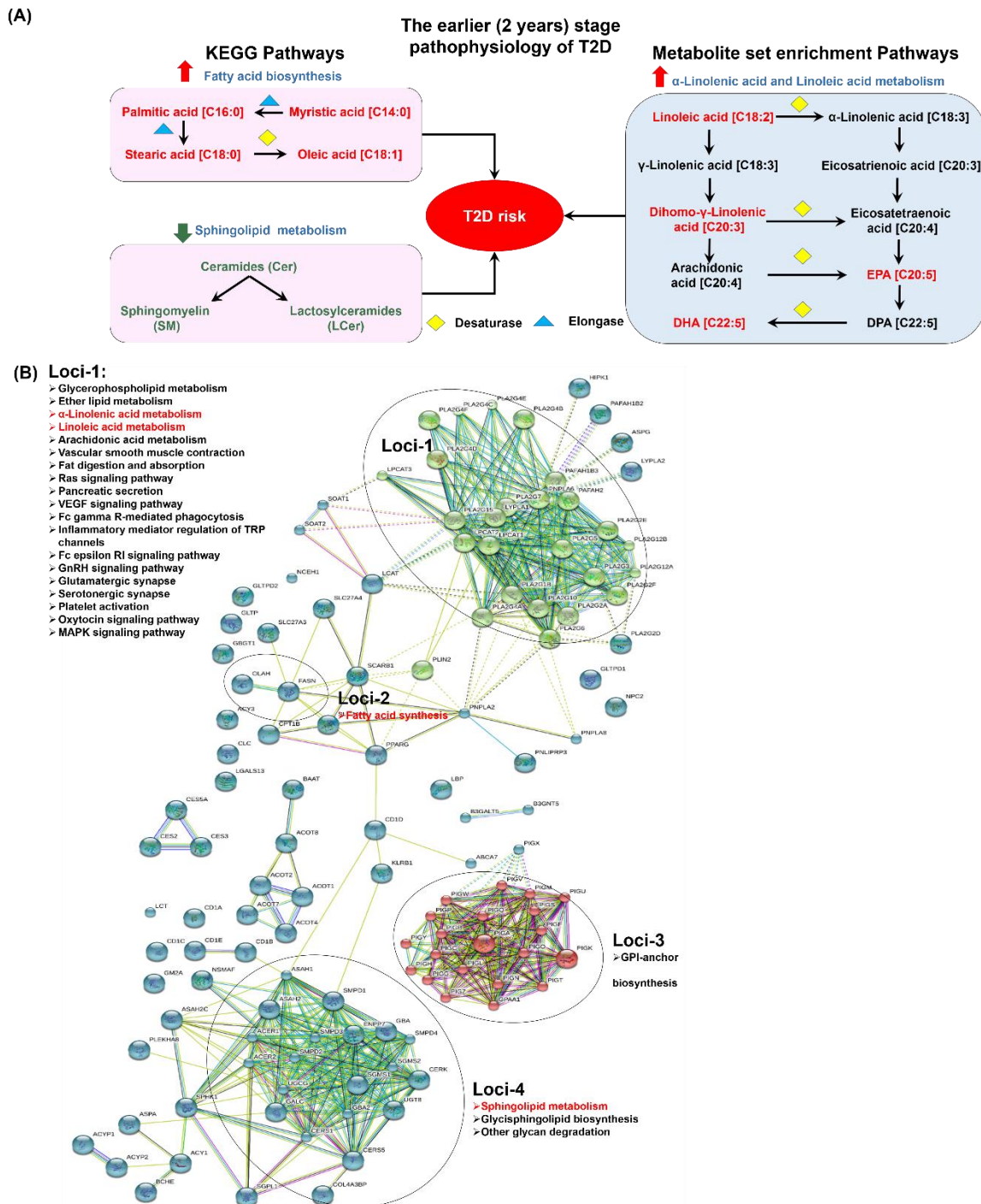


(B) The heatmap of the significant lipid metabolites



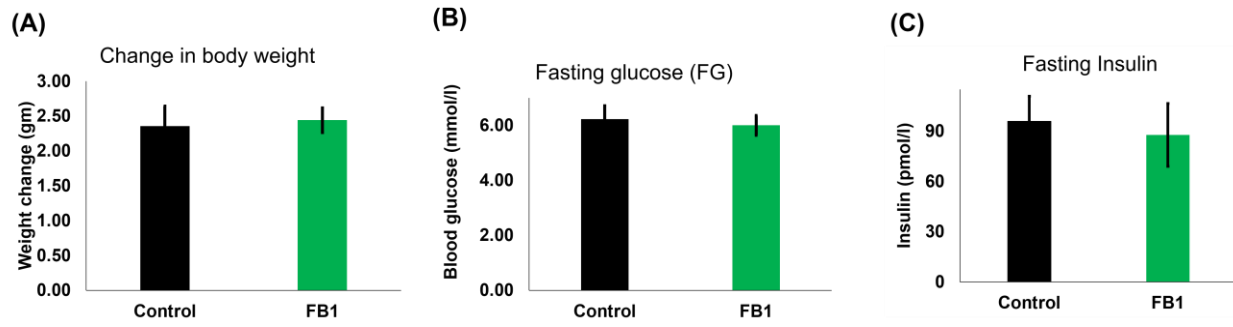
ESM Fig.2: The volcano plot and heatmap. (A) The volcano plot of the final dataset. (B) The heatmap of the differently expressed significant lipid metabolites.

ESM Fig.3: The summary of the early stage T2D pathophysiology



ESM Fig.3: The summary of the early stage T2D pathophysiology. (A) The putative pathways and their (putative) interactions found in direct approach. (B) The putative pathways and their (putative) interactions found in indirect approach.

ESM Fig.4



ESM Fig.4: (A) The change in body weight after 3 weeks of treatment. (B) The comparison of fasting glucose after 3 weeks of treatment. (C) The comparison of fasting insulin after 3 weeks of treatment.