An Application of Variable Selection Methods in Omics Data obtained by GC-MS and LC-MS for the study of Non-Alcoholic Fatty Liver Disease

Sabatini S[1, 2], Saponaro C[3], Gaggini M[1], Carli F[1], Rosso C[4], Bugianesi E[4], Gastaldelli A[1]

[1] Institute of Clinical Physiology, CNR Pisa, Italy. [2] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy. [3] INSERM, U1190, Lille, France.
[4] Division of Gastroenterology and Hepatology and Lab. of Diabetology, Dept. of Medical Sciences, University of Turin, Turin, Italy.

MASSA 2020 - VIP
Virtual International Pre-congress

EPoS
Elucidating Pathways of Steatohepatitis

IFC

UNIVERSITÀ DI SIENA
1240

Introduction:

In recent years, omics technologies have been widely exploited in the study of metabolic diseases for the identification of metabolites (or genes) relevant in the discrimination among different conditions. Typically, the features are highly inter-correlated (fig. 1), and their number greatly exceeds the number of samples. Variable selection methods are necessary to deal with multicollinearity and overfitting. NAFLD is a complex liver disorder, that ranges from liver steatosis to non-alcoholic steatohepatitis (NASH) and cirrhosis [1]. It is associated with alterations of glucose and lipid metabolism and obesity is one of the major risk factors [2]. However, the exact role of obesity (BMI) and fat accumulation in the outbreak and progression of NAFLD remains mostly unclear.

Aims and Objectives:

The aim of this work is to evaluate and compare the use of two sparse classification methods, Elastic Net (EN) [3] and Sparse Partial Least Square Discriminant Analysis (SPLS-DA) [4], in the study of the impact of BMI on glucose and lipid metabolism in NAFLD subjects.

Methods:

We analysed data from 34 non-diabetic biopsy proven NAFLD subjects (11 lean, 12 overweight and 12 obese) and 8 controls (CT). We measured non esterified fatty acids (FFA) and aminoacid plasma composition by GC-MS, and target lipidomics (Cer, PC, LysoPC, PE, DG and TG composition) by LC-QTOF. In vivo fluxomic analysis was performed by stable isotope tracer infusion; tracer enrichment was measured by GC-MS and lipolysis and endogenous glucose production (EGP) were obtained from modelling analysis of tracer turnover. We calculated desaturation index (SCD1=palmitoleic/ palmitic acid) and unsaturated-to-saturated fat ratio (UFA/SFA), IR in the periphery (HOMA), adipose tissue (Lipo-IR, Adipo-IR) and liver (Hep-IR). We applied EN and SPLS-DA in order to make a discrimination based on the metabolic profile between CT and NAFLD subjects, between CT and lean (BMI≤25) NAFLD subjects and between lean and obese (BMI>30) NAFLD subjects. In the latter comparison, variables were expressed as fold-changes with respect to the median of the controls (log2 (-/median CT)). Hyper-parameters were tuned through repeated cross-validation (CV) and permutation tests were used to assess the statistical significance of the classifications.

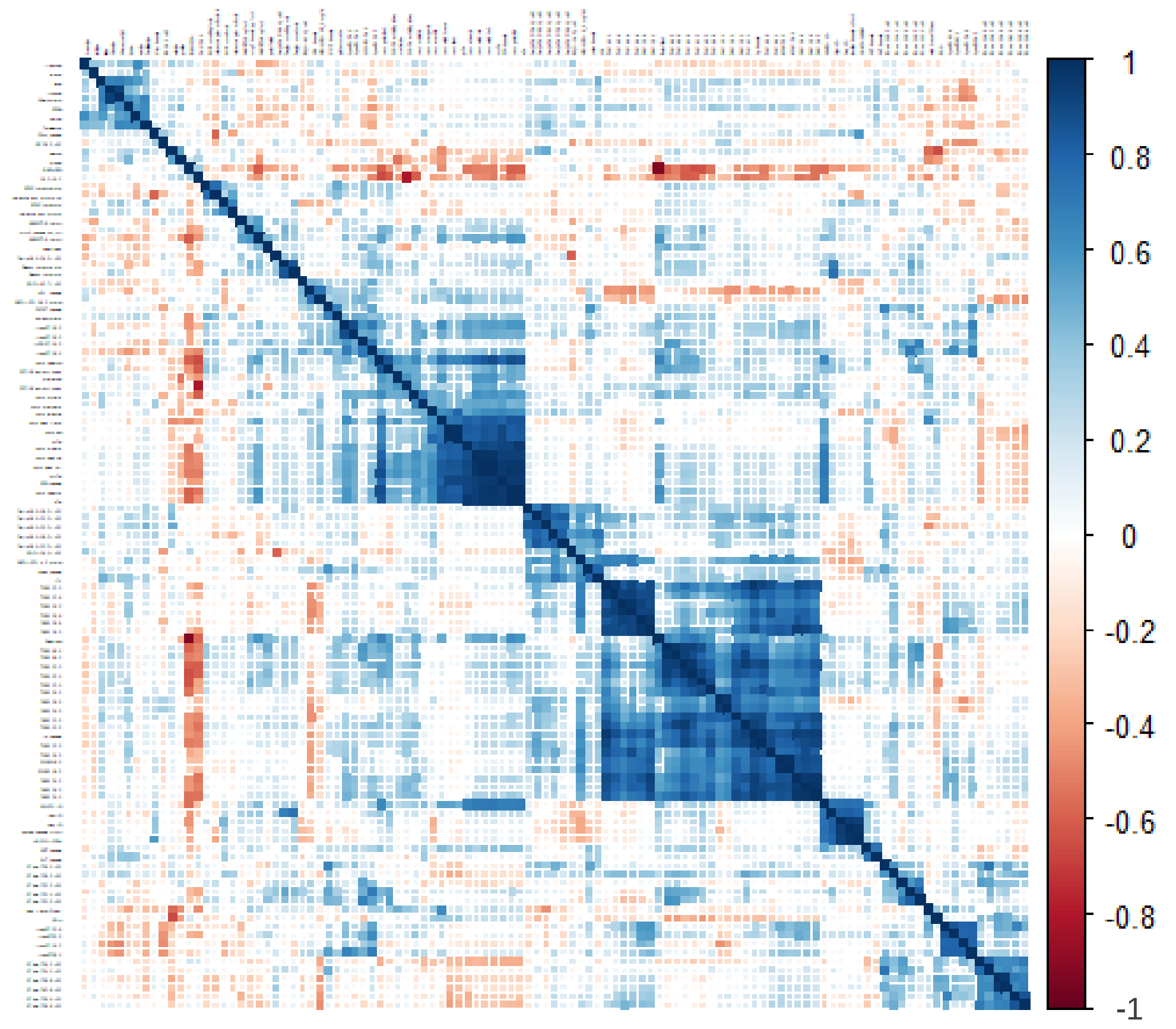
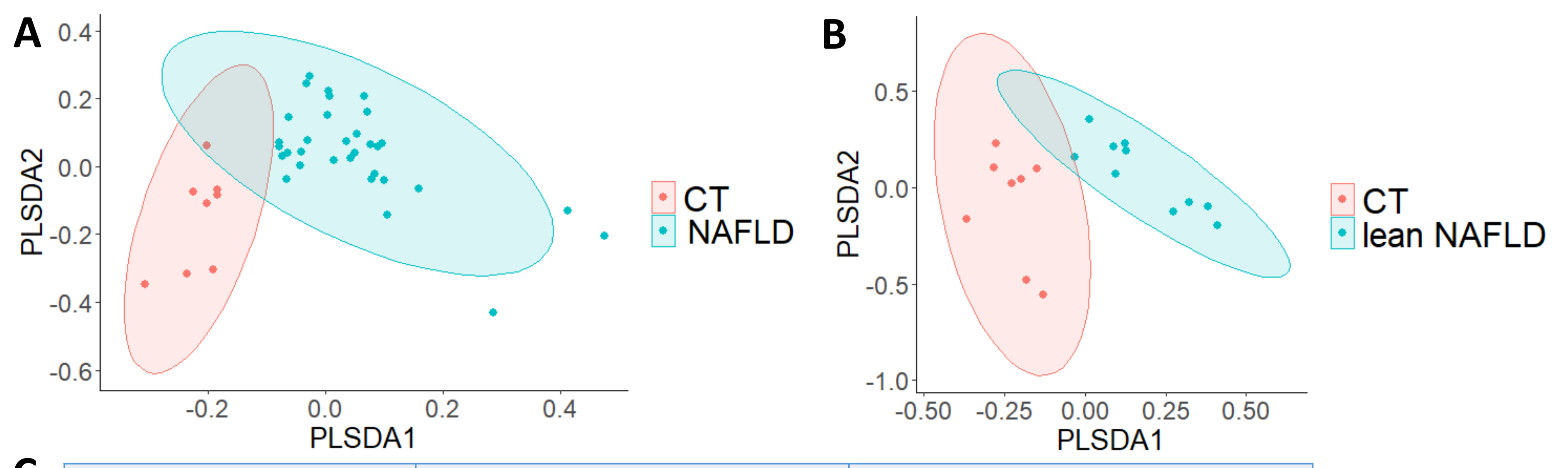


Figure 1: Spearman Rank Correlation matrix of all the features, ordered according hierarchical clustering algorithm. Blue clusters in the matrix point out the multicollinearity between the variables.

Results:

Classification Analysis: Controls vs NAFLD

In the controls versus all NAFLD and controls versus lean NAFLD subjects' discriminations (fig. 2), both EN and SPLS-DA were able to discriminate between the classes with very high accuracy measures (≥89%). The two methods brought out a recurrent group of relevant discriminant factors: AAs (histidine, threonine, methionine and aromatic AA) (fig. 3B), insulin resistance indexes (HOMA, Lipo-IR and Hep-IR),UFA/SFA and several lipidomic species (LysoPC, TG and Cer, mainly) (fig. 3B).



C

	EN			SPLS-DA		
	Accuracy	P-value	Selected Features n.	Accuracy	P-value	Selected Features n.
CT vs NAFLD	97.5%	<0.001	40/100	92%	0.002	85/100
CT vs NAFLD: lean	89%	0.01	45/100	92%	0.003	51/100

Figure 2: Scores plots of SPLS-DA between controls and all NAFLD subjects (panel A) and between controls and lean NAFLD subjects (panel B). In panel C, the accuracy (20 times repeated 5-fold CV), the p-value (permutation tests, n=1000) and the number of selected features performed for both EN and SPLS-DA are given.

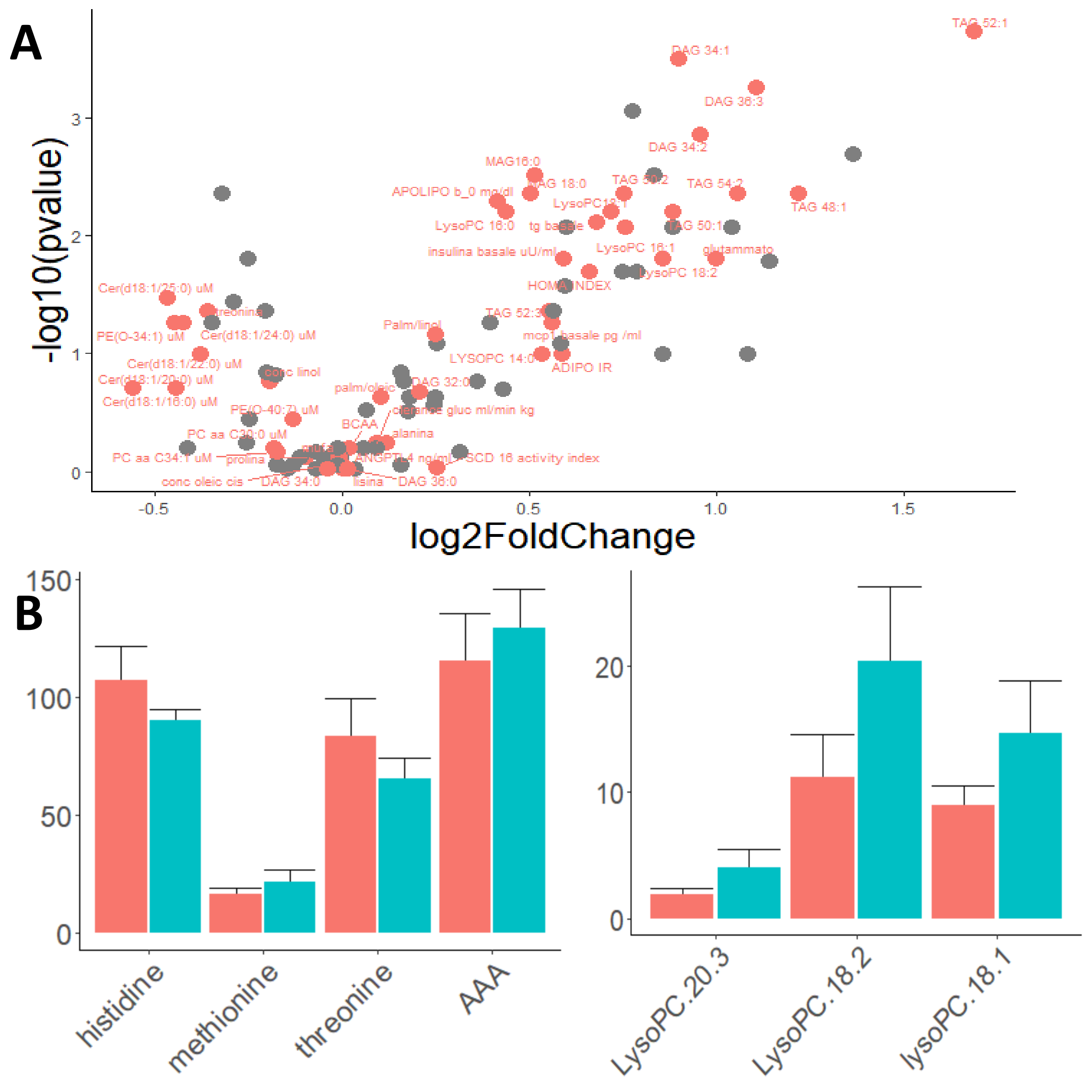
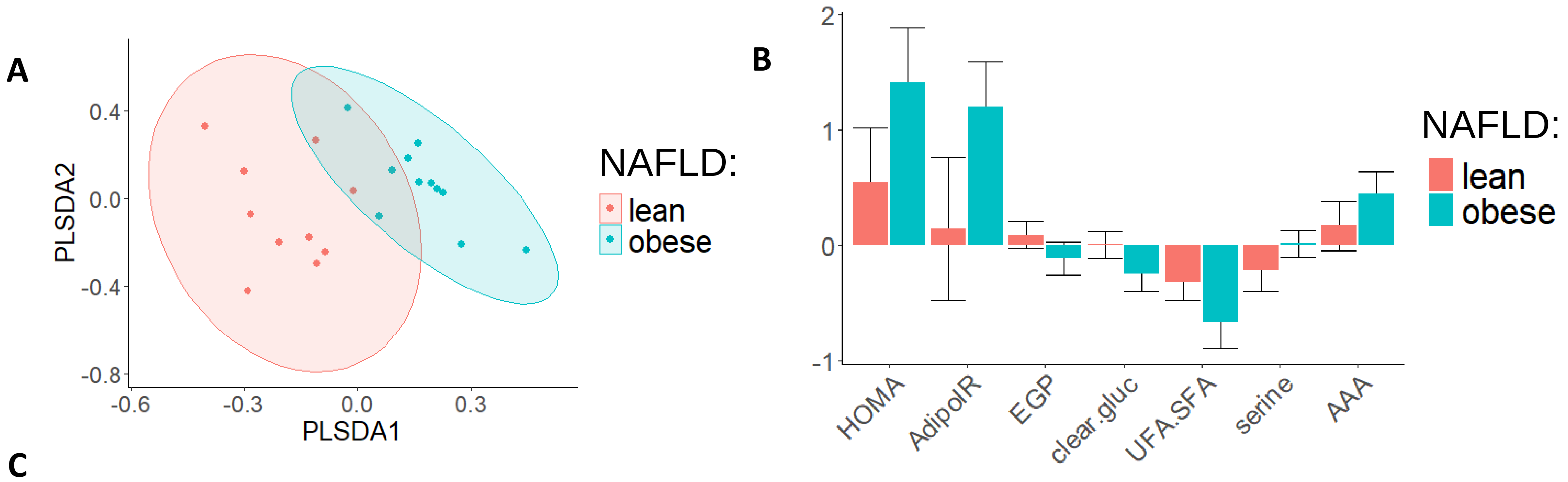


Figure 3: A) Volcano plot of the features. In red, those selected by ENET; as we expect, they are in great concentration in the higher part of the graph. B) Histograms of some of the features selected by both EN and SPLS-DA in the classification between controls and lean NAFLD subjects.

Classification analysis: lean vs obese NAFLD

Unlike SPLS-DA, EN was able to discriminate lean vs obese NAFLD subjects with a good accuracy in predictions and significant p-value (fig. 4C), selecting a few number of variables. Insulin resistance's indexes (HOMA and Adipo IR), amino acids (serine, aromatic AA) and FFAs (UFA/SFA) result as main discriminating factors for both the models (fig 4B).



C

	EN			SPLS-DA		
	Accuracy	P-value	Selected Features n.	Accuracy	P-value	Selected Features n.
NAFLD: lean vs obese	82%	0.03	13/100	70%	0.16	18/100

Figure 4: A) Scores plots of SPLS-DA between lean and obese NAFLD subjects. B) fold-changes of the features selected by both EN and SPLS-DA. C) accuracy (20 times repeated 5-fold CV), p-value (permutation tests, n=1000) and number of selected features performed for both the models are given.

Conclusions:

This study shows that insulin resistance, amino acids, FFAs and lipids (TG, Cer, LysoPC, mainly) are relevant to distinguish non-NAFLD from NAFLD subjects, even when they have similar BMI. On the contrary, when comparing lean versus obese NAFLD subjects the main discriminator factors are insulin resistance in adipose tissue and periphery, amino acids and FFAs. The two classification methods considered here show similar performances, although the sparsity of EN is greater.

Acknowledgment:

Funded by:Horizon2020 under grant agreement: no.634413, EPoS, Elucidating pathways of steatohepatitis.

Bibliography:

1. A. Gastaldelli, K. Cusi; JHEP Reports, 1(4)(2019), p. 312-328.
2. M. Gaggini, M. Morelli, E. Buzzigoli, R.A. DeFronzo, E. Bugianesi, A. Gastaldelli, Nutrients, 5(5)(2013), p. 1544-1560.
3. H. Zou , T. Hastie; J. R.Statist.Soc.Series B, 67(2)(2005), p. 301–320.
4. H. Chun, S. Keleş; J. R.Stat.Soc.Series B, 72(1)(2010), p. 3–25.