**Serum lipidomics as diagnostic potential for metabolic-associated hepatocellular carcinoma**

Monika Lewinska1, Alvaro Santos-Laso2, Enara Arretxe3, Cristina Alonso3, Raul Jimenez-Aguero2, Emma Eizaguirre2, María Jesús Pareja4, Manuel Romero-Gómez5,6, Marco Arrese Jimenez7, Malte P. Suppli8, Filip K. Knop8,9, Stine Karlsen Oversoe10, Gerda Elisabeth Villadsen10, Thomas Decaens11,12, Flair Jose Carrilho13, Claudia P. Oliveira13, Bruno Sangro6,14, Rocio I. R. Macias6,15, Jesus M. Banales2,6,16 and Jesper B. Andersen1

**Supplementary materials and methods:**

**Metabolomic analysis:**

Serum lipidomic profiles were semi-quantified by ultra-high-performance liquid chromatography coupled to mass spectrometry (UHPLC-MS), as previously described1. Briefly, metabolite extraction was accomplished by fractionating the samples into pools of species with similar physicochemical properties, using appropriate combinations of organic solvents. Two UHPLC-time of flight-MS based platforms analyzing methanol and chloroform/methanol serum extracts were combined with the amino acid measurement using an UHPLC-single quadrupole-MS based analysis. Platform used for the analysis methanol extraction was optimized for the profiling of fatty acids, oxidized fatty acids, acyl carnitines, lysoglycerophospholipids (monoacylglycerophospholipids and monoetherglycerophospholipids), free sphingoid bases, bile acids, and steroid sulfates. The chloroform/methanol extract platform provided coverage over glycerolipids (di- and triglycerides), cholesterol esters, sphingolipids (ceramides and sphingomyelins), and glycerophospholipids (diacylglycerophospholipids and 1-ether, 2-acylglycerophospholipids).

Metabolite extraction procedures, chromatographic separation conditions and mass spectrometric detection conditions have been previously described1.

Data pre-processing generated a list of chromatographic peak areas for the metabolites detected in each sample injection. An approximated linear detection range was defined for each identified metabolite, assuming similar detector response levels for all metabolites belonging to a given chemical class represented by a single standard compound. Metabolites for which more than 30% of data points were found outside their corresponding linear detection range were not used for statistical analyses. Intra and inter batch data normalization was performed following the procedure described by Martinez-Arranz et al2.

**Metabolite quantification:**

Absolute quantitation of Linoleyl carnitine, Linoleic acid, Osbond acid, 9(Z),12(Z)-Hexadecadienoic acid, 9(E)-Tetradecanoic acid, Tetradecadienoic acid, Hexadecatrienoic acid, Hydroxyoctadecadienoic acid, 1-Hydroxy-2-Linoleoyl-sn-Glycero-3- Phosphatidylcholine 1-Linoleoyl-2-Hydroxy-sn-Glycero-3-Phosphatidylcholine and 1-Hydroxy-2-Docosapentaenoyl-sn-Glycero-3-Phosphatidylcholine was also performed.

Stock standard solution of Linoleyl carnitine, Linoleic acid, osbond acid, 9(Z),12(Z)-Hexadecadienoic acid, 9(E)-Tetradecanoic acid, (2E, 4E)-2,4-Tetradecadienoic acid, 7(Z),10(Z),13(Z)-Hexadecatrienoic acid, 9-(S)-HODE and 18:0 Lyso PC were prepared individually in methanol at a concentration level of approximately 100, 1000 or 10000 μg/mL, depending on the compound. The working standard was prepared by mixing the appropriate amount of each standard solution in methanol to reach a final concentration of approximately 100 μg/mL. Calibration standards were prepared by consecutive dilutions in methanol, ranged between 0.0005-5 μg/mL for FFAox13 and 0.005-50 μg/mL for the rest of the standards in the calibration curve.

Likewise, internal standards (IS) stock solution of Octadecanoyl (18,18,18-D3)-L-Carnitine, Linoleic-9,10,12,13-D4 acid, Hexadecatrienoic 7(Z),10(Z), (13)14-D6 acid, and 18:1-d7 Lyso PC were prepared at a concentration of 1000 or 5000 μg/mL. The IS working solution was prepared by mixing the appropriate amount of each stock solution in methanol to reach the desire concentration. Quality control (QC) samples (reference sera samples commercially available) were treated according to the same protocol that serum and plasma samples, and were analyzed against the calibration curve. The intra-day precision was determined by analyzing five replicates.

**Estimation of metabolite concentration:**

To estimate the concentration of validated metabolites in the discovery cohort, the measured absolute values of metabolites (normalized to IS and standard curves) were plotted against the normalized chromatographic peak areas for 58 samples. Simple linear regression was applied to generate the equation for each metabolite and estimate R-square (Goodness of fit).

The equations and R square values are presented below:

|  |  |  |
| --- | --- | --- |
| Metabolite | Equation | R-square |
|
| AC(18:2n-6) | Y = -2.071\*X + 1.140 | 0.03627 |
| Linoleic acid | Y = 63.50\*X - 0.7825 | 0.7469 |
| Osbond acid | Y = 0.2222\*X + 0.01720 | 0.8331 |
| Palmitolinoleic acid | Y = 0.09445\*X + 0.008943 | 0.9513 |
| MUFA (14:1n-5trans) | Y = 0.2102\*X + 0.2223 | 0.8924 |
| PUFA (14:2n-x) | Y = 0.4994\*X + 0.9538 | 0.8863 |
| PUFA (16:3n-x) | Y = 0.02936\*X + 0.003510 | 0.8544 |
| Hydroxy-octadecadienoic acid | Y = 0.03322\*X + 0.005443 | 0.8705 |
| PC(18:2/0:0) | Y = 0.3786\*X + 0.008920 | 0.8244 |
| PC(0:0/22:5) | Y = 0.03056\*X + 0.2379 | 0.1593 |

Where Y is the concentration in ug/ml and X is a normalized peak area.

**Statistical analysis:**

Differences in biochemical parameters were established with one-way ANOVA, Kruskal-Wallis test with Tukey’s multiple comparisons, or Fisher exact test for categorical data (Prism 8.2.0).

*A general linear model* was used to identify metabolites different between the groups independently of age, gender, and BMI (SPSS 25.0.0).

Data per metabolic class ware calculated as the sum of the normalized areas of all the metabolites with the same chemical characteristics. The outlier analysis was performed before one-way ANOVA or Kruskal-Wallis test with Tukey’s or Dunn’s multiple testing (Prism 8.2.0) depending on data normality.

*Multivariate analyses* were performed with package mixOmics3 or Metaboanalyst 4.0. The sparse partial least squares discriminant analysis (sPLS-DA) was used to investigate the best separation between the groups for multiple groups and orthogonal partial least squares discriminant analysis (oPLS-DA) for pair-wise compresence. Hierarchical cluster analysis was performed with Ward’s algorithm using Euclidean distances.

*The receiver operating characteristic (ROC) curves* were generated for pair-wise comparison and areas under the curves (AUC) were estimated with Metaboanalyst 4.0.

*The diagnostic model* was constructed using ROC curves generated by Monte-Carlo cross-validation (MCCV) using balanced sub-sampling. In each MCCV, two-thirds (2/3) of the samples were used to evaluate the feature importance and the classification model was validated on 1/3 of the samples that were left out. The procedure was repeated multiple times to calculate the performance and confidence interval of each model detecting the optimal number of features for best accuracy. The linear support vector machine (SVM) method was used for sample classification.

*MAFLD-HCC diagnostic score (MHDS)* was established using simple logistic regression (Prism 8.2.0).

*The odds ratio and relative risk for MHDS*, clinical and biochemical features were established using Prism 8.2.0. We generated confusion matrixes for the features in pulled cohorts (discovery and validation) using a cut-off value of age (50 years), BMI=30, and reference ranges for biochemical features. Statistical significance was measured with Fisher’s exact test (p<0.05), the relative risk was calculated with Koopman asymptotic score and odds ratio with the Baptista-Pike method.

*Pearson correlation* was applied for pattern hunter analysis to detect metabolites significantly correlating with progressive increase or decline. Metabolites were considered as correlated with pattern when (q<0.05) and abs(r)> 0.3.

*Pathway overrepresentation analysis* was performed with the Integrated Molecular Pathway Level Analysis (IMPaLA)4. As the majority of our metabolites belong to lipids, which to a high extend are not included in metabolic pathways, we used all detected serum metabolites with HMDB identifiers as background in the enrichment analysis.

*Lipid pathway enrichment analysis* was performed with BioPAN software and default settings.

**Supporting Tables:**

**Supporting Table 1. Differentially expressed metabolites and pathways between MAFLD-HCC and CTRL.** (A) Significantly different (FDR q>0.05) metabolites between MAFLD-HCC and CTRL, (B) Significantly downregulated pathways (blue), (C) Significantly upregulated pathways

**Supporting Table 2. Differentially expressed metabolites and pathways between MAFLD-HCC and OB-MAFLD.** (A) Significantly different (FDR q>0.05) metabolites between MAFLD-HCC and OB-MAFLD, (B) Significantly downregulated pathways (blue) in MAFLD-HCC, (C) Significantly upregulated pathways in MAFLD-HCC.

**Supporting Table 3. Differentially expressed metabolites and pathways between MAFLD-HCC and AV-HCC.** (A) Significantly different (FDR q>0.05) metabolites between MAFLD-HCC and AV-HCC, (B) Significantly downregulated pathways (blue)in MAFLD-HCC, (C) Significantly upregulated pathways in MAFLD-HCC

**Supporting Table 4. The reference ranges of validated metabolites.** The linear regression analysis was used to generate equations for concentration estimation in discovery and MAFLD(val) cohorts. The range is given as average concentrations ± standard deviation. #concentrations directly measured against the series dilutions of internal standards.

**Supporting Table 5. Pearson correlation of lipids associated with progressive liver damage on the axis CTRL->OB-MAFLD->MAFLD->MAFLD-HCC.** The r>0.3 and r<-0.3, as well as FDR corrected p<0.05 were considered as associated with the pattern of progression.

**Supporting Figures:**

**Supp. Fig. 1** **Clinical and biochemical characteristics of the study population**.

Scatter and violin plots present chemical and biochemical characteristics of discovery (disc) and validation (val) patient data sets. (**A**) Age, (**B**) Body mass index (BMI), (**C**) Gender composition, (**D-I**) liver biochemistry markers (AFP, ALT, AP, GGT, Albumin and Bilirubin), (**J**) Platelets and (**K**) The international normalized ratio (INR) and prothrombin measurements. The statistically significant difference was computed using one-way ANOVA followed by Tukey’s multiple comparison test, Kruskal-Wallis followed by Dunn’s multiple comparison test or chi-square test for categorical data. The reference values for biochemical parameters were marked with red or blue dashed lines. INR was considered normal within the range (0.79-1.2) and prothrombin percentage activity between 70-100%. \*p<0.05, \*\*p<0.01, \*\*\*p<0.001, \*\*\*\*p<0.0001.

**Supp. Fig. 2.** **Patient distribution according to covariates**.

Principal Component Analysis (PCA) plots presenting the patient groups in the discovery set, their distribution and grouping (95% confidence regions) based on (**A**) age, (**B**) gender, (**C**) body mass index (BMI score), (**D**) tobacco smoking, (**E**) liver fibrosis (F score), (**F**) history of type 2 diabetes, (**G**) liver cirrhosis (in HCC patients, only) and (**H**) etiology background of HCC.

**Supp. Fig. 3. Differences in metabolic classes.**

Violin plots presenting major classes of metabolites. The statistically significant differences were computed using one-way ANOVA followed by Tukey’s multiple comparison test or Kruskal-Wallis followed by Dunn’s multiple comparison test. \*p<0.05, \*\*p<0.01, \*\*\*p<0.001, \*\*\*\*p<0.0001.

**Supp. Fig. 4. The significant reactions between major lipid classes.**

Pathway enrichment analysis of lipids (BioPAN). These analyses are presenting the major reactions and predicted genes that change between MAFLD-HCC and (**A**) CTRL, (**B**) OB-MAFLD, and (**C**) Alcohol- and Hepatitis B or C virus related HCC (AV-HCC).

**Supp. Fig. 5. The significant reactions related to fatty acids.**

Pathway enrichment analysis of fatty acids (FA) using (BioPAN). These analyses are presenting the major reaction chains and predicted genes (given in the tables) that differ between MAFLD-HCC and (**A**) CTRL, (**B**) OB-MAFLD, and (**C**) Alcohol- and Hepatitis B or C virus related HCC (AV-HCC).

**Supp. Fig. 6. The diagnostic potential of serum lipidomics.**

(**A**) Calculation of the predictive accuracies using a different number of features in the SVM model (average of the repeated random sub-sampling by cross-validation). i) MAFLD-HCC versus Controls (CRTL), ii) MAFLD-HCC versus obese MAFLD (OB-MAFLD) and iii) MAFLD-HCC versus Alcohol- and Hepatitis B or C virus related HCC (AV-HCC). (**B**) ROC curve for the 5-metabolite model distinguishing MAFLD-HCC patients from MAFLD(val) patients. AUC: Area under the curve, CI: confidence interval.

**Supp. Fig. 7. The differential expressed metabolites (DEM) between MAFLD and OB-MAFLD.**

Volcano plot presenting differentially expressed metabolites (DEMs) between MAFLD and OB-MAFLD. FDR corrected p<0.05 is marked with red dashed line.

**Supp. Fig. 8. Pathway enrichment analysis of lipids.**

Pathway enrichment analysis of lipids (BioPAN) presenting the major reaction chains between (**A**) lipid classes and (**B**) Fatty Acids (FA) species. The predicted genes for FA reaction chains are presented in the table.

**References:**

1 Barr, J. *et al.* Obesity-dependent metabolic signatures associated with nonalcoholic fatty liver disease progression. *J Proteome Res* **11**, 2521-2532, doi:10.1021/pr201223p (2012).

2 Martinez-Arranz, I. *et al.* Enhancing metabolomics research through data mining. *J Proteomics* **127**, 275-288, doi:10.1016/j.jprot.2015.01.019 (2015).

3 Rohart, F., Gautier, B., Singh, A. & Le Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* **13**, e1005752, doi:10.1371/journal.pcbi.1005752 (2017).

4 Cavill, R. *et al.* Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput Biol* **7**, e1001113, doi:10.1371/journal.pcbi.1001113 (2011).