The Serum Lipidome Unravels a Diagnostic Potential in Bile Acid Diarrhea

Monika Lewinska[1,*], Martin L. Kårhus[2,*], Anne-Marie Ellegaard[2], Manuel Romero-Gomez[3], Rocio I.R. Macias[4,5], Jesper B. Andersen[1,†,#], Filip K. Knop[2,6,7,†,#]

[1]University of Copenhagen, Department of Health and Medical Sciences, Biotech Research and Innovation Center (BRIC), Copenhagen, Denmark; [2]Center for Clinical Metabolic Research, Copenhagen University Hospital - Herlev and Gentofte, Hellerup, Denmark; [3]UCM Digestive Diseases, Virgen del Rocío University Hospital, SeLiver group at Institute of Biomedicine of Seville (HUVR/CSIC/US), University of Seville, Sevilla, Spain; [4]National Institute for the Study of Liver and Gastrointestinal Diseases (CIBERehd, "Instituto de Salud Carlos III"), Madrid, Spain; [5]Experimental Hepatology and Drug Targeting (HEVEPHARM), IBSAL, University of Salamanca, Salamanca, Spain; [6]Steno Diabetes Center Copenhagen, Herlev, Denmark; [7]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[*/†]Authors contributed equally

[#]**Corresponding authors:** Jesper B. Andersen (Ole Maaløes Vej 5, DK-2200, Copenhagen N Denmark; phone: +45 35325834; e-mail: jesper.andersen@bric.ku.dk) and Filip K. Knop (Gentofte Hospitalsvej 7, DK-2900 Hellerup, Denmark; phone: +45 38674266; e-mail: filip.krag.knop.01@regionh.dk)

Supplemental Material and Methods

Comprehensive metabolomics including 1,295 metabolites was performed on serum samples with three platforms as described previously [1]. Additionally, quantitative measurement of 39 oxylipins was performed with Acquity UPLC H-Class system coupled to a QTRAP® 6500 mass spectrometry (MS) detector (AB Sciex, Concord, ON, Canada) working in electrospray ionization negative mode and multiple reaction monitoring. Metabolite extraction was accomplished by fractionating the samples into pools of species with similar physicochemical properties, using appropriate combinations of organic solvents. Two ultra-high-performance liquid-chromatography (UHPLC)-time of flight- MS-based platforms analyzing methanol and chloroform/methanol serum extracts were combined with the amino acid measurements using an UHPLC-single quadrupole-MS-based analysis. The platform used for the analysis of methanol extraction was optimized for the profiling of fatty acids, oxidized fatty acids, acylcarnitines, and 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 MS detection conditions have been previously described [1].

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

omitted from further statistical analyses. Intra- and inter-batch data normalization was performed following the procedure described by Martinez-Arranz *et al.* [1].

Statistical analyses:

Statistical analysis

Differences in biochemical parameters were established using Mann-Whitney for non-normally distributed values, *t* test for normally distributed values, and Fisher exact test for categorical data (Prism 9.3.0, GraphPad software, USA). We detected 0.1% of missing values in the metabolomics data. The metabolites with more than 5% of missing values were excluded from the analysis. Otherwise, missing values were imputed with the k-nearest neighbor method (feature-wise) and log-transformed. The covariate-testing and correction were performed with linear regression before analysis. Covariates included age, sex, and BMI, as well as batch, which were adjusted in differential expression analysis (R 4.0.4, R Core Team (2022)).

Data sets per metabolic class were calculated as the sum of the normalized areas of all the metabolites with the same chemical characteristics. Outlier analysis was performed before multiple testing (Prism 9.3.0, GraphPad software, USA) and followed by false discovery rate (FDR) correction. Receiver operating characteristic (ROC) curve analysis was used to identify and evaluate the performance of individual metabolites (Metaboanalyst 5.0) [2]. The post-hoc power analysis established the minimal samples size of 24 samples per group at false discovery rate corrected p value = 0.05, and statistical power of 90%. The multivariate ROC analysis was performed by generating ROC curves by Monte-Carlo cross-validation using balanced sub-sampling. In each Monte-Carlo cross-validation, two-thirds of the samples were used to evaluate the feature importance. The classification model was validated using one-third of the remaining samples. The procedure was repeated 50 times to calculate the performance and confidence interval of each model, and to detect the optimal number of features for best

accuracy. The linear support vector machine (SVM) method was used for sample classification. Feature selection was performed with SVM Mean Importance Measure. The average accuracy of the model was based on 100 cross-validations. Model performance was measured using 1,000 permutations and both (AUROC) and the predictive accuracy were with a statistical accuracy of $p < 0.001$ (Metaboanalyst 5.0) [2].

A logistic regression model was generated to calculate the BAD Diagnostic Score (BDS). Metabolite selection was performed based on the SVM Mean Importance Measure and LASSO Frequencies (Metaboanlyst 5.0) [2], which were used to generate the logistic regression equation: $\mathbf{logit(P) = ln(P / (1 - P)) = \alpha + \beta X}$, where α is the intercept term, β is the regression coefficient estimated from the sample dataset, X_i is the set of covariate (concentration) values, and $P = Pr(y=1|x)$ is the probability of the disease (i.e., BAD). Differentially abundant metabolites were detected using a linear model (limma, Bioconductor, R 4.0.4) [3] including age, sex, BMI, and batch as covariates and with FDR correction. Hierarchical cluster analysis was performed with Ward's algorithm using Euclidean distances. Finally, a general linear model was used to establish the relationship between metabolite concentrations and the treatment performance [4] (Prism 9.3.0, GraphPad software, USA).

References:

1 Martinez-Arranz I, Mayo R, Perez-Cormenzana M, Minchole I, Salazar L, Alonso C, *et al.* Enhancing metabolomics research through data mining. J Proteomics 2015;**127**:275-88.

2 Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res 2021;**49**:W388-W96.

3 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;**43**:e47.

4 Kårhus ML, Brønden A, Forman JL, Haaber A, Knudsen E, Langholz E*, et al.* Safety and efficacy of liraglutide versus colesevelam for the treatment of bile acid diarrhoea: a randomised, double-blind, active-comparator, non-inferiority clinical trial. Lancet Gastroenterol Hepatol 2022.