Sphingolipid, fatty acid and phospholipid metabolites are associated with disease severity and mTOR inhibition in lymphangioleiomyomatosis

Leonardo Bottolo, Suzanne Miller, Simon R. Johnson

SUPPLEMENTARY METHODS

Bottolo L, *et al. Thorax* 2020; 75:679–688. doi: 10.1136/thoraxjnl-2019-214241

SAMPLE PROCESSING AND ANALYSIS

All samples were maintained at -80°C until processed. Quality control recovery standards were added prior to extraction for QC purposes. Samples were prepared using the MicroLab STAR® system (Hamilton Company). Proteins were precipitated with methanol by shaking for 2 minutes followed by centrifugation. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI, and one backup. The organic solvent was removed using a TurboVap® (Zymark). Sample extracts were stored overnight under nitrogen before preparation for analysis. Controls analysed with the experimental samples comprised a pooled matrix sample generated by taking a small volume of each experimental sample as a technical replicate throughout the data set, extracted water samples as process blanks and various QC standards selected not to interfere with the endogenous compounds were spiked into every analysed sample, allowed instrument performance monitoring and aided chromatographic alignment. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites present in all pooled matrix samples. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections.

ULTRAHIGH PERFORMANCE LIQUID CHROMATOGRAPHY-TANDEM MASS SPECTROSCOPY (UPLC-MS/MS)

All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyser operated at 35,000 mass resolution. Sample extracts were dried and reconstituted in solvents compatible with each of the four methods. Reconstitution solvents contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency.

1. Using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. Extracts were gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA).
2. Using acidic positive ion conditions, optimised for more hydrophobic compounds at a higher organic content.

Bottolo L, *et al. Thorax* 2020; 75:679–688. doi: 10.1136/thoraxjnl-2019-214241

3. Using basic negative ion optimised conditions on a separate C18 column. The basic extracts were gradient eluted using methanol and water, with 6.5mM Ammonium Bicarbonate at pH 8.
4. Via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a water/acetonitrile with 10mM Ammonium Formate, pH 10.8 gradient. The MS analysis alternated between MS and data-dependent MS^n scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z.

DATA EXTRACTION AND COMPOUND IDENTIFICATION

Raw data were extracted, peak-identified and QC processed at Metabolon© as described previously (1). The platform and compound identification algorithm uses biochemical identifications based on three criteria: 1) the retention index within a narrow RI window of the proposed identification, 2) mass match to the library +/- 10 ppm, and MS/MS forward and reverse scores between experimental data and standards. The MS/MS scores are based on a comparison of ions present in the experimental spectrum to those present in the library spectrum. The combined use of all three data points is used to distinguish and differentiate compounds. Over 3300 purified standards are registered in LIMS for analysis for determination of their identity.

CONTROL GROUP

To increase study power, 21 control samples from the current study were merged with female controls from two companion metabolites studies available from the NIHR BioResource Rare Diseases, University of Cambridge, resulting in 43 controls subjects. All were healthy women over the age of 18 with no prior history of lung disease (supplementary table 1).

LAM GROUP

The LAM subjects comprised 79 women recruited from the National Centre for LAM in Nottingham UK between 2011 and 2018. All subjects had LAM defined by current ATS/JRS criteria (2). Subjects had a clinical assessment, comprising CT of the chest, abdomen and pelvis, screening for TSC, full lung function. At follow up visits, lung function FEV_1 and DL_{CO} were measured. The study was approved by the East Midlands Research Ethics Committee (13/EM/0264) and all participants gave written informed consent.

Prospective change in FEV_1 was calculated by the regression slope of all FEV_1 values (ΔFEV_1) and expressed as change in ml/year. Only subjects with greater than one year of observations were included for calculation of ΔFEV_1.

Bottolo L, *et al. Thorax* 2020; 75:679–688. doi: 10.1136/thoraxjnl-2019-214241

Serum VEGF-D was determined using Quantikine ELISA DVED00, (R&D Systems, Abingdon, UK).

For exploratory analyses, subjects were categorised into those with mild and more severe disease based upon lung function and disease activity defined by ΔFEV_1. Subjects were also segregated by menopausal status and treatment with rapamycin.

DATA PRE-PROCESSING

Normalisation and imputation of case and control samples serum metabolites were performed following the workflow presented in (1) and (3). Relevant normalisation (N) steps can be summarised as follows: (N.1) Untargeted metabolites and metabolites belonging to 'Xenobiotics' biochemical class were removed from the analysis, reducing the number of targeted metabolites to 820; (N.2) After checking the proportion of missing values across samples and metabolites, no metabolites were removed. (N.3) Each metabolite raw value was rescaled to have median 1 to adjust for variation due to instrument run-day tuning differences; (N.4) A log transformation with base 10 was applied to all the metabolites; (N.4) After transformation, data points lying more than 4 standard deviations from the mean of each metabolite concentration were excluded. For the imputation (I) of missing values, we employed the KNN-TN method of (4) which consists of the following steps: (I.1) Estimation of the detection level (DL) of the machine to be the minimum observed value for the whole dataset; (I.2) Maximum Likelihood Estimation (MLE) of μ_m and σ_m, assuming that each metabolite m (m = 1,…,820) follows a left-truncated (on the DL) Gaussian distribution with mean μ_m and standard deviation σ_m; (I.3) Standardisation of each metabolite using the MLEs of μ_m and σ_m; (I.4) For each metabolite m with a missing value in sample i, detection of its K = 10 closest metabolites (which have an observed value for their ith sample) using the k-nearest neighbours algorithm; (I.5) Imputation of the missing value with a weighted average of the K values found in (I.4). The weights are functions of the Pearson correlations between the metabolite with missing values and its K closest metabolites; (I.6) Transformation of each metabolite back to the original scale as it was before step (I.3).

DIFFERENTIAL ANALYSIS

Differential analysis of 820 targeted serum metabolites was performed by using Limma (5) after correcting for BMI, ethnicity and run day (recording which samples were run on which days relative to each other) using a linear mixed model (6) with study and run day as crossed random effects. We also corrected for age when the hypothesis to test did require the assessment of its effect on the metabolites' levels, and 'study' covariate as a third crossed random effect when we tested differences between LAM women not treated with rapamycin and healthy controls (since we added two extra control groups from the companion metabolites studies). Significant differential metabolites were declared at 10% FDR (7).

Bottolo L, *et al. Thorax* 2020; 75:679–688. doi: 10.1136/thoraxjnl-2019-214241

DIFFERENTIAL NETWORK ANALYSIS

Based on the WGCNA package (8), Differential Network Analysis (9) allows the detection of differential networks (modules) between conditions. Relevant steps of this method include: (S.1) Build correlation matrix *C* within each condition. We used robust Spearman's rank correlation coefficient to calculate the correlation between any pair of metabolites in each condition. (S.2) Compute matrix of adjacent (powered correlation) differences. The soft-threshold power parameter β is chosen such that it is the lowest power for which the scale-free topology R^2 fit between the degree of connectivity *k* and the proportion of metabolites that have connectivity *k* exceeds 0.85. In the real data analysis, this automatic procedure leads to estimated value of β ranging between 5 and 7. (S.3) Derive the Topological Overlap Measure (TOM distance) based on the dissimilarity matrix *T* in order to identify metabolites that share the same metabolites' neighbours in the graph obtained from the matrix of adjacency differences (S.2); (S.4) Hierarchical clustering of dissimilarity matrix *T* based on TOM distance allows partitioning the metabolites into modules that share similar metabolites' neighbours. Thresholding of hierarchical clustering is obtained by using the Dynamic Tree Cut R package (10). (S.5) The permutation-based procedure is employed to assess the statistical significance of the modules detected in (S.4), with the number of permutations = 1,000. The permutation consists in shuffling observations between conditions and, for a given partition obtained in (S.4), the empirical p-value is obtained by calculating how many time the observed average powered correlation difference in a module is greater than the one obtained by shuffling the observations. Finally, (S.5) for each identified module, principal component analysis is performed and the first eigenvalue ('eigen-metabolites') is correlated (Spearman correlation) with selected clinical traits.

BIOINFORMATIC ANALYSIS

Metabolomic pathway analysis was performed by using MetaboAnalyst 4.0 (11) with both significant differential metabolites and metabolites modules detected in the differential network analysis mapped in KEGG pathways. Given the lack of pathways annotation for a large fraction of metabolites, significant metabolomic pathways were declared at a conservative 10% Holm–Bonferroni correction. Finally, topology pathway analysis was performed by selecting the relative-betweenness centrality measure (ranging between 0 and 1) which quantifies the importance of a subgroup of metabolites in a given metabolomic pathway.

REFERENCES

1. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46(6):543–50.
2. McCormack FX, Gupta N, Finlay GR, Young LR, Taveira-Da Silva AM, Glasgow CG, et al.

Official American thoracic society/Japanese respiratory society clinical practice guidelines: Lymphangioleiomyomatosis diagnosis and management. Am J Respir Crit Care Med. 2016;194(6):748–61.

3. Krumsiek J, Mittelstrass K, Do KT, Stückler F, Ried J, Adamski J, et al. Gender-specific pathway differences in the human serum metabolome. Metabolomics. 2015;11(6):1815–33.
4. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. BMC Bioinformatics. 2017;18(1).
5. 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(7):e47.
6. Pinheiro J, Bates D, editors. Linear Mixed-Effects Models: Basic Concepts and Examples. In: Mixed-Effects Models in S and S-PLUS. Springer-Verlag; 2006. p. 3–56.
7. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.
8. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9.
9. Tesson BM, Breitling R, Jansen RC. DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. 2010;11.
10. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20.
11. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46(W1):W486–94.