



Screening of LC-MS-based metabolic profiling data utilizing statistical modeling to identify candidate biomarkers

Christina M. Sorensen¹, Don S. Daly², Thomas O. Metz¹, Navdeep Jaitly¹, Stephen J. Callister¹, Matthew E. Monroe¹, Jennifer S. Zimmer¹, Ronald J. Moore¹, and Richard D. Smith¹

¹Biological Sciences Division, ²Computer Science and Mathematics Division, Pacific Northwest National Laboratory, Richland, WA 99352

Overview

- The LC-MS-based metabolomics measurements are typically within 100 – 1,000 m/z.
- Chemical noise can be reduced through the use of intensity thresholding or application of matched filtration^{1,2} during feature identification.
- An unknown number of chemical noise features are often included in downstream data processing and analysis.
- Higher confidence can be placed in metabolomic feature detection when statistical modeling is applied.
- A restricted maximum likelihood (REML) model incorporating exploratory data analysis was used in comparative analyses to demonstrate the applicability of this approach for biomarker identification.

Introduction

A comparative LC-MS study of fasting and non-fasting serum and plasma was performed to develop an integrated data analysis approach for metabolic profiling and to identify candidate biomarkers. Samples from the same ten individuals corresponding to fasting and non-fasting states were pooled to create four uniform samples: (1) fasting serum (FS), (2) non-fasting serum (nFS), (3) fasting plasma (FP), and (4) non-fasting plasma (nFP).

These dataset profiles were compared using principle component analysis, t-test, and an integrated approach that includes these steps (Figure 1) integrated with a linear model for candidate biomarker selection. This integrated approach includes statistical testing in a REML model incorporating exploratory data analysis as a single data processing regimen for metabolite profile comparison.

Comparative data analysis of the serum and plasma metabolite profiles was performed to select metabolites for further study (e.g., targeted MS/MS analysis). Relative quantitative estimation utilizing REML offers selection of metabolite profiles within a biological context.

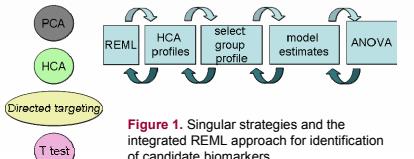


Figure 1. Singular strategies and the integrated REML approach for identification of candidate biomarkers

Results

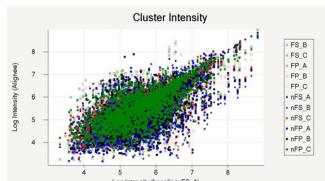


Figure 2. Agreement between intensity measurements for individual features in the aligned and baseline datasets.

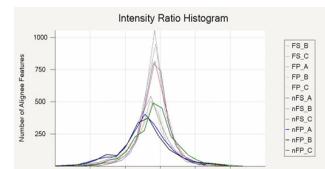


Figure 3. Intensity ratio histograms for aligned datasets.

Table 1. Selection of the number of metabolites after peak picking compared to grouping the metabolites by reproducibility (features within group) or by selecting a metabolite profile (unique features) as one EDA for metabolic profiling.

Sample Type	Number of features (average \pm std dev)	Number of features within group	Unique features to the group
Fasting Plasma	4864 \pm 514	2065	398
Fasting Serum	4456 \pm 354	2290	472
Non-Fasting Serum	4174 \pm 474	1797	185
Non-Fasting Plasma	4550 \pm 211	1787	268

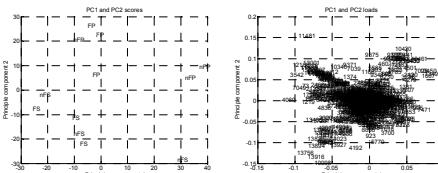


Figure 4. Principle component analysis of metabolites common to all samples prior to application of the linear model. The scores plot (left) illustrates the separation of sample groups. The loads plot (right) shows selection of groups of similar metabolite profiles is difficult when many metabolites are present.

Methods

- Fasting and non-fasting serum and plasma samples were extracted in triplicate using chloroform/methanol (2:1, v/v) followed by analysis of the water-soluble fraction using capillary LC-FTICR MS.
- Raw data were deconvoluted and features determined using in-house software programs, and metabolite feature datasets were chromatographically aligned using the LCMSWARP algorithm.³
- Aligned data were used to compare common biomarker discovery strategies. Principle component analysis and t-test were evaluated as singular strategies together with a more integrated approach incorporating REML.
- Multiplicative model of metabolite abundance

$$A_{\text{dstch}} = \mu_a + D_d + S_s + T_t + F_f + E_{\text{dstch}}$$

where A_{dstch} is the i th abundance measurement of the f th feature measured at time t from a biological sample from the s th sample type (s = serum or plasma) and the d th diet regime (d = fasting or nonfasting) perturbed by the multiplicative random effect E_{dstch} .

• Log transformation forms an additive model

$$A_{\text{dstch}} = \mu_a + D_d + S_s + T_t + F_f + E_{\text{dstch}}$$

- We assume that $E_{\text{dstch}} = \log(E_{\text{dstch}})$ is normally distributed with mean 0 and variance σ^2
- The model was evaluated through inspection of residual plots, interaction plots. Qnom and parameter estimates were made for metabolite groups, ANOVA, confidence intervals, and model error estimates

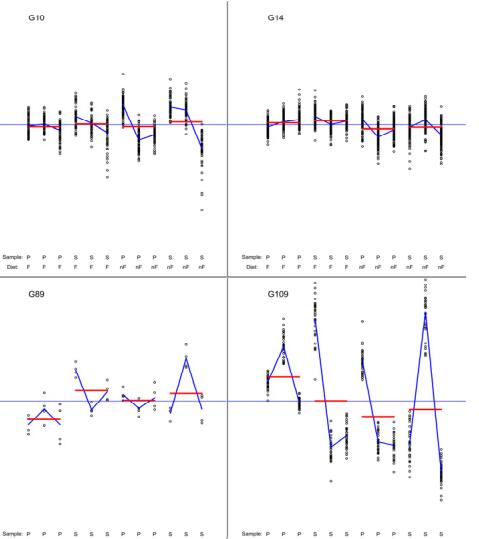


Figure 5. Clustering the individual metabolite abundances (open circles) into groups in conjunction with mixed effect modeling allows data trends across conditions to become more apparent. Mean abundance profiles for the grouped metabolites (G) show: non-normalized (blue) and a mixed effects model (red). Horizontal blue line is the mean abundance across all metabolites in the group. Further exploration is required for interpretation; however those groups with highly correlated biological trends merit further investigation. Interpretation of the influences on the metabolite abundance profiles requires the diagnostic plots shown in Figures 6 and 7.

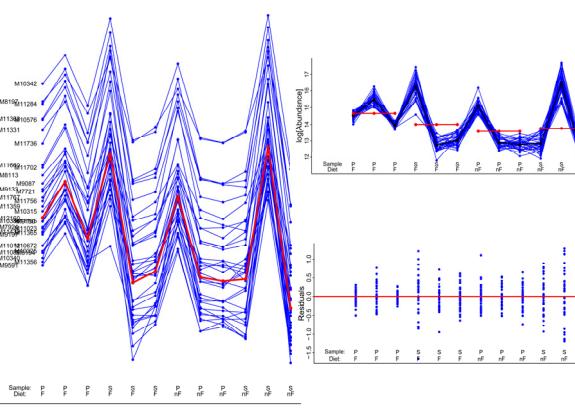
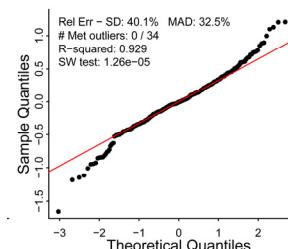


Figure 6. Representative output for G109 metabolites. The observed abundance profile (blue) for each of the G109 metabolites (M) is plotted in the left panel with the group average profile (red). Individual profiles (blue) with metabolite measurability effects removed are shown in the upper right panel with the average profile (black) due to the processing effects alone. In (red) is the mixed effects model for this group of metabolites, minus the processing effects. The residual profiles (lower right panel) shows remaining effects which are represented in the error terms in Figure 7. Metabolite identifications for some members of this group are shown in Table 2.



Term	Est	StdErr	P-val
(Intercept)	14.7	0.121	<.001
TypeS	-0.644	0.0561	<.001
StateN	-1.07	0.0561	<.001
TypeS:StateN	0.83	0.0794	<.001

Figure 7. Qnom plot (left) shows the Group109 metabolite profile fit (red) to the model. The ANOVA results from the G109 REML-fit mixed effects model is summarized by a table of model parameter estimates, their standard errors and P-values of the test that a parameter equals 0.

Metabolite	Molecular Formula	m/z	NET	Metabolite ID	Group
16:0 LPC	C24H51NO7P	496.3398	0.52	M4710	10
18:0 LPC	C26H55NO7P	524.3710	0.57	M5298	10
18:1 LPC	C26H53NO7P	522.3554	0.54	M5240	10
18:2 LPC	C26H51NO7P	520.3398	0.51	M5181	10
16:1 LPC	C24H49NO7P	494.3244	0.49	M4656	14
20:4 LPC	C28H51NO7P	544.3398	0.52	M5662	14
22:6 LPC	C30H51NO7P	568.3401	0.52	M6104	14
20:5 LPC	C28H49NO7P	542.3232	0.50	M5629	89
32:0 PC	C40H81NO8	734.5709	0.84	M9197	109
32:1 PC	C40H79NO8P	732.5545	0.81	M9150	109
34:1 PC	C42H83NO8P	760.5981	0.85	M10576	109
34:2 PC	C42H81NO8P	758.5702	0.82	M10342	109
34:3 PC	C42H79NO8P	756.5541	0.79	M10315	109
36:2 PC	C44H85NO8P	786.6017	0.87	M11368	109
36:3 PC	C44H83NO8P	784.5859	0.84	M11331	109
36:4 PC	C44H81NO8P	782.5699	0.82	M11284	109
36:5 PC	C44H79NO8P	780.5535	0.80	M11185	109
38:3 PC	C46H85NO8P	812.6175	0.89	M11767	109
38:4 PC	C46H83NO8P	810.6019	0.87	M11736	109
38:5 PC	C46H83NO8P	808.5980	0.83	M11702	109
38:6 PC	C46H81NO8P	806.5701	0.81	M11660	109
40:6 PC	C48H85NO8P	834.6025	0.86	M12160	109

Contact Information

Christina M. Sorensen, Ph.D.
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
e-mail: christina.sorensen@pnnl.gov
Phone: (509) 376-6712

Conclusions

- Qualitative and semi-quantitative analyses are applied in a directed approach for selection of markers
- Diagnostic estimates of the processing effects can be made, improving data analysis
- Grouping features can potentially identify metabolites belonging to the same chemical class or pathway
- Validation and identification of selected metabolites benefits from targeted LC-MS/MS and or complementary NMR analyses

Acknowledgements

This research was supported by NIH grants DK070146 and DK071283. Experimental portions of this research were performed in the Environmental Molecular Sciences Laboratory, a U.S. Department of Energy (DOE) national scientific user facility located at the Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract No. DE-AC05-76RLO 1830.

References

- Andreev, V.P., Rejtar, T., Chen, H.-S., Moskovets, E.V., Ivanov, A.R., Karger, B.L. A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal. Chem.* **2003**, *75*, 6314-6326.
- Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779-787.
- Jaitly, N., Monroe, M.E., Pettry, V.A., Clauss, T., Adkins, J.N., Smith, R.D. A robust alignment algorithm for aligning liquid chromatography-mass spectrometry analyses in the AMT tag pipeline. *Anal. Chem.* **2006**, *78*, 7397-7409.