Untargeted Lipidomics of NSCLC Shows Differentially Abundant Lipid Classes in Cancer vs Non-Cancer Tissue

Joshua M. Mitchell[1,2,3] , Robert M. Flight [2,3], Hunter N.B. Moseley [1,2,3,4]

[1]Department of Molecular & Cellular Biochemistry, [2]Markey Cancer Center, [3]Resource Center for Stable Isotope-Resolved Metabolomics, [4]Institute for Biomedical Informatics, University of Kentucky, Lexington KY, United States

♦ Abstract

Lung cancer is the leading cause of cancer death worldwide and non-small cell lung cancer (NSCLC) represents 85% of newly diagnosed lung cancers. The high mortality rate of lung cancer is due in part to the lack of effective treatment options for advanced disease. A major limitation in the development of effective treatment options is our incomplete understanding of NSCLC metabolism at a molecular level. Improvements in mass spectrometry combined with our untargeted assignment tool SMIRFE enables systematic and less biased examinations of NSCLC metabolism.

The molecular formula assignments provided by SMIRFE were then classified to lipid category and class using machine learning models trained on known examples of lipid molecular formulas. Subsequent differential abundance analysis of these classified formulas revealed significant and consistent differences in lipid profiles at the lipid category level between disease and non-disease samples. Both sterols and glycerolipids were consistently and significantly upchanged in disease versus non-disease. This pattern was also observed in the set of samples confirmed by pathology to be primary NSCLC.

The significant upchange in sterols in primary NSCLC suggests a possible therapeutic role for statins and nitrogenous bisphosphonates, pharmaceuticals that inhibit endogenous sterol biosynthesis, in the treatment of primary NSCLC. This hypothesis is consistent with previous epidemiological studies that have identified a therapeutic role for statins in the treatment of NSCLC, but were unable to identify a molecular mechanism for this effect. Additionally, several sterols belonging to the sterol ester subcategory are consistently and significantly upchanged, suggesting increased SCD1 activity. SCD1 expression is known to be a negative prognostic indicator for survival in NSCLC. In our study, a large fraction of the NSCLC samples displayed this phenotype; however, SCD1 mutants are unexpected in all of these samples. This suggests that this metabolic phenotype may be shared across multiple genetic subtypes of NSCLC. Thus, inhibitors of SCD1 and other enzymes involved in the production of this metabolic phenotype could have utility in the treatment of many genetic subtypes of NSCLC.

♦ Materials and Methods

From 86 patients at the University of Louisville and the University of Kentucky with suspected resectable stage I or IIa primary NSCLC, matching disease and non-disease samples were collected. Lipid extracts were prepared from paired disease and non-disease tissue samples and analyzed using ultra-high resolution Fourier transform Mass Spectrometry (FT-MS). Samples were prepared, processed, and mass spectrometry performed by the Center for Environmental and Systems Biochemistry at the University of Kentucky. Mass spectrometry was performed using two different Thermo Tribrid Fusion instruments (Fusion 1 and Fusion 2). Fusion 1 samples were only collected at the University of Louisville, Fusion 2 samples were collected at both University of Louisville and University of Kentucky.

Spectra were characterized using our scan-level peak correspondence algorithm to generate high quality peaklists. These peaklists were assigned using our in-house assignment tool SMIRFE. SMIRFE assigns spectra in an untargeted manner without the use of a metabolite database. The molecular formula assignments generated by SMIRFE were then classified into lipid category and class using a machine learning model trained using examples of known lipid formulas and lipid category and class from the LIPIDMAPS structure database and non-lipid formulas form the Human Metabolome Database. The performance of these models on the training data is described in Table 1.

Using a recursive intersection union algorithm, consistently assigned spectral features across all samples were mapped to isotopologue-resolved molecular formula assignments. Peaks present in less than 25% of disease or non-disease samples were dropped from further analysis. Differential abundance analysis was perfumed using both LIMMA and SDAMS. Log2 fold changes for each categorized corresponded-peak between disease vs non-disease was then calculated.

Table 1: Performance of RF Machine Learning Models on Lipid Training Data

Category	Precision	Out of Bag Accuracy	Number of Examples
Fatty Acyls [FA]	0.841	0.901	2031
Glycerolipids [GL]	0.996	0.995	532
Glycerophospholipids [GP]	0.995	0.996	1886
Polyketides [PK]	0.767	0.885	1376
Prenol Lipids [PR]	0.989	0.971	473
Saccharolipids [SL]	1.000	0.998	102
Sphingolipids [SP]	0.999	0.993	1404
Sterol Lipids [ST]	0.934	0.972	824
not_lipid	0.928	0.799	7587

Table 1: Performance of Random Forest Machine Learning Models on LIPIDMAPS Training Data – For most lipid categories, the models demonstrate excellent out-of-bag accuracy and precision. Polyketides had relatively poorer precision and out-of-bag accuracy but also represent a diverse set of structures and formulas that may not be well represented by a single decision boundary. High precision and out-of-bag accuracy imply that the predicted lipid categories from these models are likely correct provided that the assigned elemental molecular formulas are similar to the population of training elemental molecular formulas.

♦ Results

Figure 1 – Principal Component Analyses

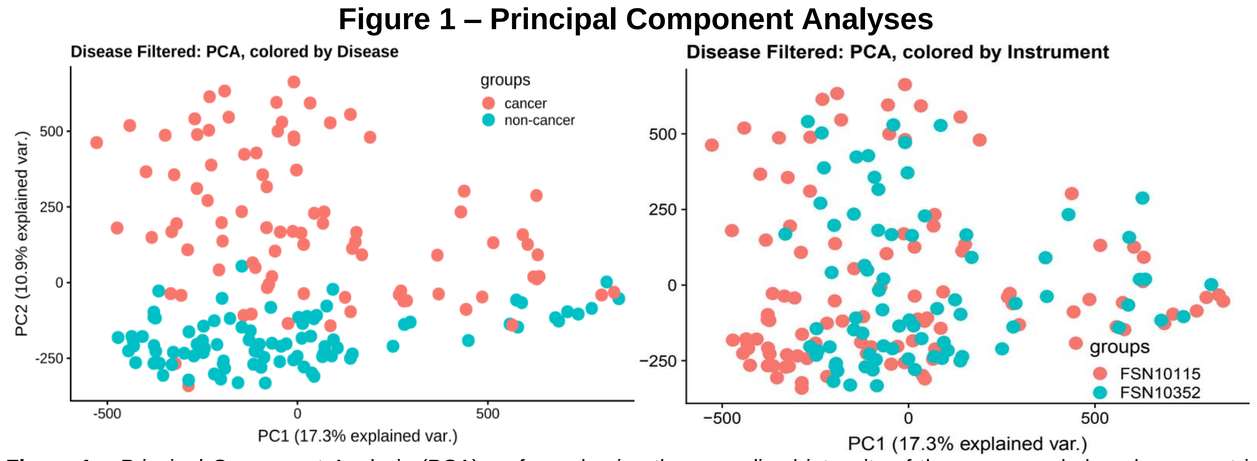

Figure 1 – Principal Component Analysis (PCA) performed using the normalized intensity of the corresponded peaks present in >25% of the disease or non-disease samples reveals a clear, if imperfect, decision boundary between sample classes along principal component 2. The grouping of disease and non-disease samples in PCA space implies that differences in the normalized intensities of the corresponded peaks captures some of the biological variance between disease and non-disease. Also instrument does not show any separation in PC1 and PC2.

Figure 2 – Disease and Non-Disease Sample Correlation

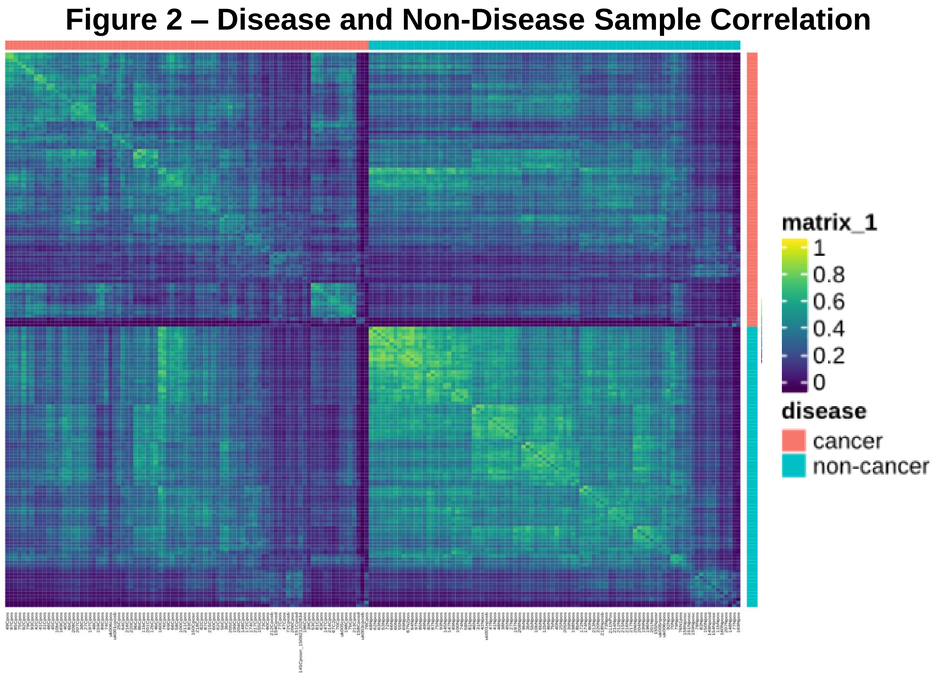

Figure 2 – Correlation heatmap using the normalized intensity of the corresponded peaks present in >25% in disease or non-disease (red and blue respectively) within all samples shows patterns of high correlation within each sample class and less correlation between sample classes. These findings along with Figure 1 support our claim that our assignments reflect differences between the sample classes.

♦ Results continued

Figure 3 – Differentially Abundant Lipid Categories are Observed in Disease

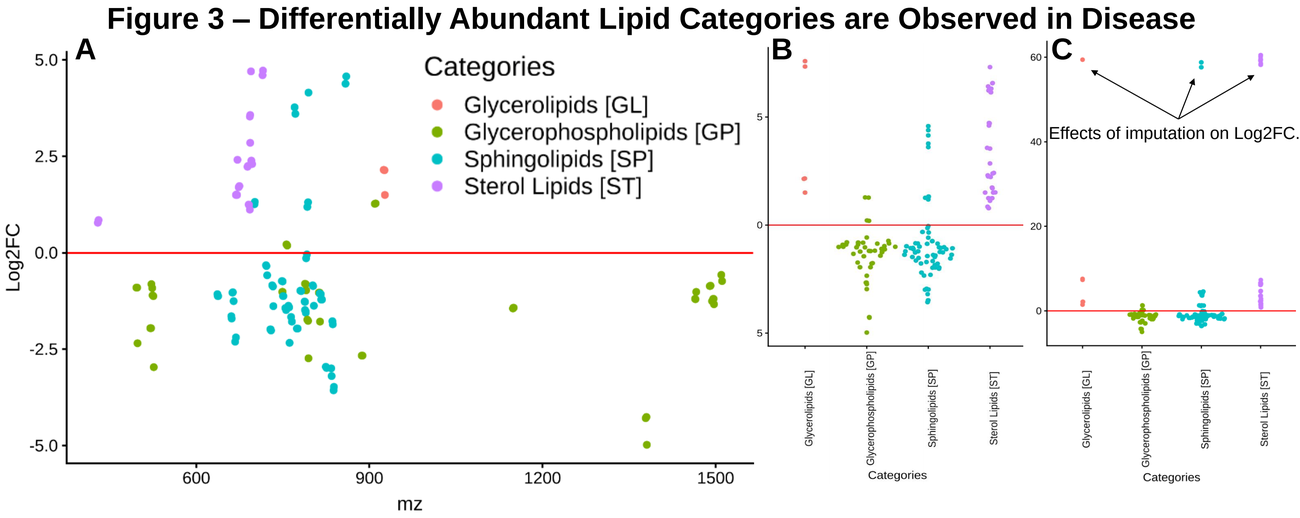

Figure 3 – Differential abundance analysis identified that sterols are upchanged in disease as compared to non-disease. As a category, 38 of 52 sterol features are up-changed for an adjusted p-value of 1.2×10^{-24} with an odds ratio of 39. This consistent and significant pattern of upchanged sterols may represent clinically relevant metabolic reprogramming in NSCLC. The graphs shown are limited to differentially abundant features with adjusted p-values <= 0.001 . Graphs A and B have differential features with imputed values removed, while graph C has the differential features with imputed values left in.

Figure 4 – Differentially Abundant Feature Correlation and Co-occurrence

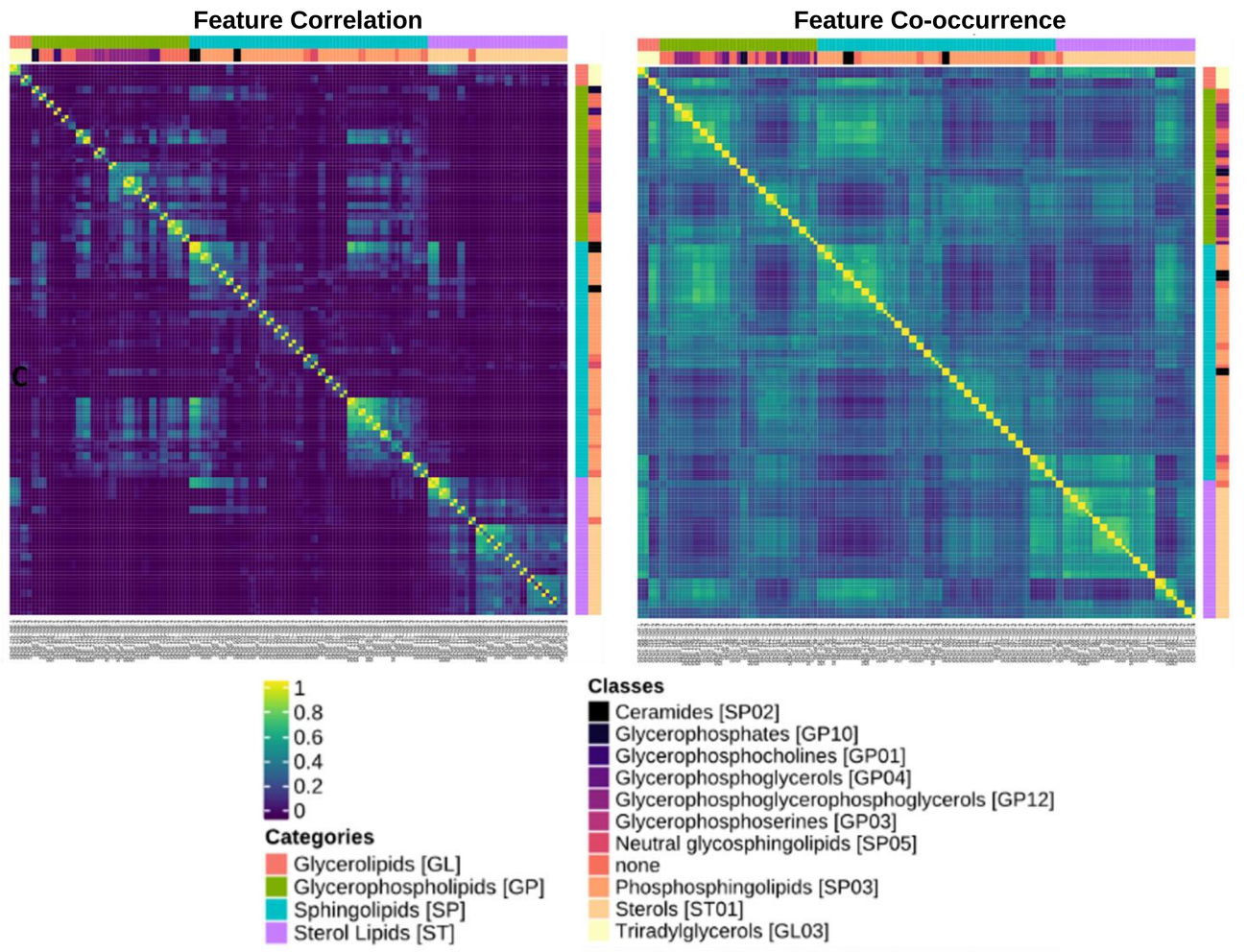

Figure 4 – (Left) Correlation heatmap using the normalized intensity of differentially abundant peaks was calculated using all samples. Two distinct populations of sterols (bottom right corners) were observed. (Right) Co-occurrence heatmap showing how often lipid features were observed together across samples. Additionally, one population of sterols co-occurs with glycerolipids and has higher correlation with the glycerlipids. This suggest a shared regulatory mechanism for these two categories of lipids.

♦ Conclusions and Future Directions

- **SMIRFE assignments combined with machine learning-based lipid category prediction provide untargeted assignments to suspected lipids in human NSCLC and non-disease samples.**
- **These assignments enable the separation of disease from non-disease using PCA, suggesting that differences in the intensities of these assignments across samples reflects biological variance between these classes of samples.**
- **Significant and substantial lipid profile differences in sterols are observed.**
- **The correlation patterns between and within differentially abundant features suggest a coregulation of glycerolipid and sterol metabolism.**
  - **The source of this coregulation is unclear from these results but could involve steroid response element binding proteins (SREBPs).**
- **Many of the upchanged sterols have molecular formulas that correspond to known unsaturated sterol esters (not shown).**
  - **Sterol ester production could simply be a side-effect of metabolic reprogramming in NSCLC or could contribute to the formation of NSCLC. Unsaturation may result from SCD1 activity.**
- **Significantly upchanged sterols in NSCLC could suggest a potential therapeutic role for statins and other mevalonate pathway inhibiting drugs for the treatment of NSCLC.**

Figure 5 – Primary NSCLC Separates Well in PCA Space

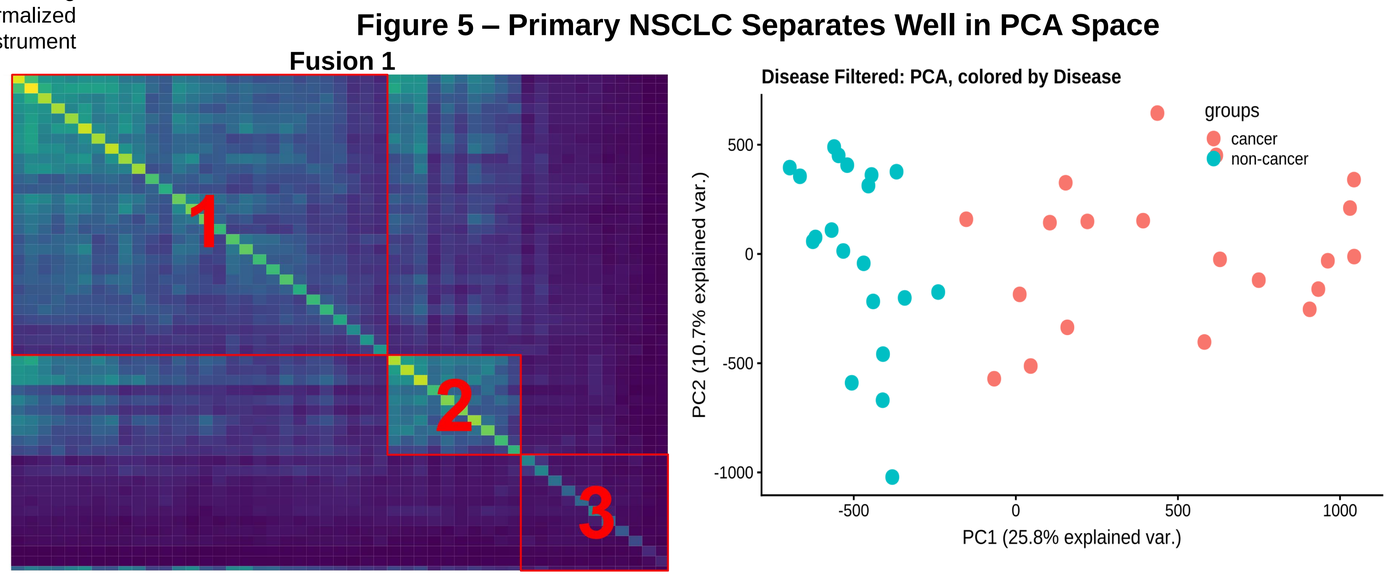

Figure 5 – In Fusion 1, three clusters of disease samples were observed in the correlation heatmap (see Figure 2, left panel, top left corner). Group 1 represented primary NSCLC. Group 2 was a mixture of primary and secondary NSCLC and group 3 was a mixture of primary and secondary with poor spectral data quality. The PCA analysis from Figure 1 was performed only on the group 1 samples and a clear separation was observed.

♦ Funding and Acknowledgements

This project was supported in part by NSF1419282 (PI Moseley), NIH P01CA163223-01A1 (Pis Andrew N. Lane and Teresa W.-M. Fan), and NIH UL1TR001998-01 (PI Kern). We would like to thank Drs. Teresa Fan, Andrew Lane, Richard Higashi and the Center for Environmental and Systems Biochemistry who collected, prepared, and analyzed the paired patient samples used in this analysis.