Comparison of Machine Learning Algorithms for Tissue Metabolomic Analysis in Hepatocellular Carcinoma (HCC)

Erin B. Evangelista[1], Sandi A. Kwee[1,2], Miles M. Sato[1], Lu Wang[2], Guoxiang Xie[2], Wei Jia[2], & Linda L. Wong[2]

[1]The Queen's Medical Center, [2]University of Hawaii Cancer Center, Cancer Biology Program

Introduction

- Hepatocellular carcinoma (HCC), which comprises the majority of liver cancers, is also the fifth-most common cancer and is the third leading cause of cancer-related deaths worldwide [1].
- Metabolomics is the systematic and large-scale study of small molecules, known as *metabolites*, within living systems such as, cells, biofluids, tissues, and micro-organisms [2].
- Tumors may differ significantly from adjacent normal tissue with regards to chemical and structural composition, resulting in pathobiologically significant alterations in their metabolomic profiles.
- Machine learning (ML) can be applied to distinguish patterns in metabolomics data to allow for more accurate classification performance than traditional statistical models [3].
- The goal of this research study is to identify metabolite signatures which may distinguish HCC from non-tumor liver tissue. In addition to their potential use as diagnostic classifiers, such signatures may also aid in identifying biochemical alterations associated with tumor features.
- As a first step of investigation, we evaluated three ML algorithms – support vector machine (SVM), partial least squares discriminant analysis (PLS-DA), and random forest (RF) – for metabolite signature discovery, using receiver operating characteristic (ROC) analysis to compare the classification performance of these algorithms across different classes of metabolites.

Materials and Methods

Patient Cohort

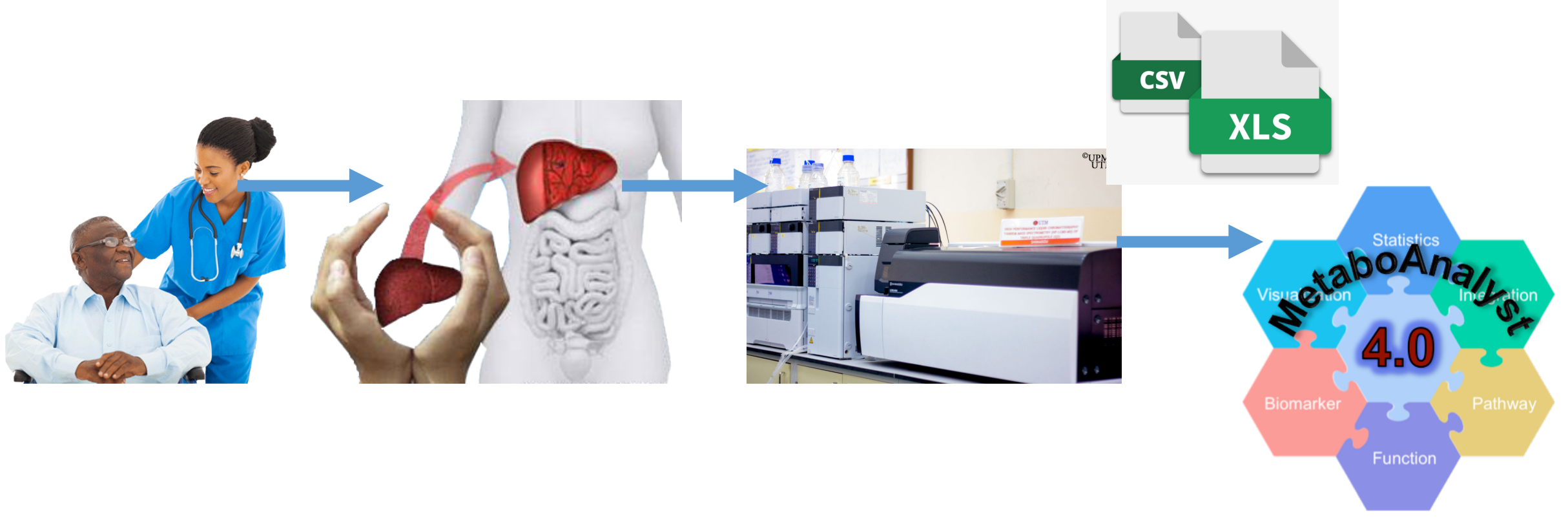
- Between February 2012 and March 2017, 53 patients diagnosed with HCC gave written informed consent to provide liver tissue samples for this research study. All samples were stored in liquid nitrogen following surgical tumor resection.
- Patients over 350 pounds, pregnant or lactating, had a serious underlying medical condition, or received chemotherapeutic, molecularly targeted, biological, or radiotherapeutic treatment for HCC were excluded from participating in the research study [2].

Metabolomic Analyses

- Targeted metabolomics was carried out using both ultra-performance liquid chromatography coupled to tandem mass spectrometry (UPLC-MS/MS) and gas chromatography time-of-flight mass spectrometry (GC-TOFMS). Quantification using authentic standards resulted in profiles of the following classes of metabolites: bile acids (BA, 42 metabolites), small molecules (SM) and free fatty acids (FFA) (128 metabolites in total), and phospholipids (lipids, 109 metabolites). Samples and compounds that were not successfully profiled with significant loss of data (10% missing data) were not included in the analysis.

MetaboAnalyst

- Biomarker discovery was carried out using Metaboanalyst 4.0 (McGill University). Software packages for biomarker discovery and evaluation were accessed via the 'metaboanalyst.ca' web-portal and also implemented locally using the R package MetaboanalystR 2.0. Missing-value imputation was performed by K-nearest neighbor (KNN). Metabolite concentration values were quantile normalized, log transformed, and mean-centered. ROC curves and areas under the ROC curves (AUC) were calculated to evaluate the classification performance of each ML algorithm.



Results

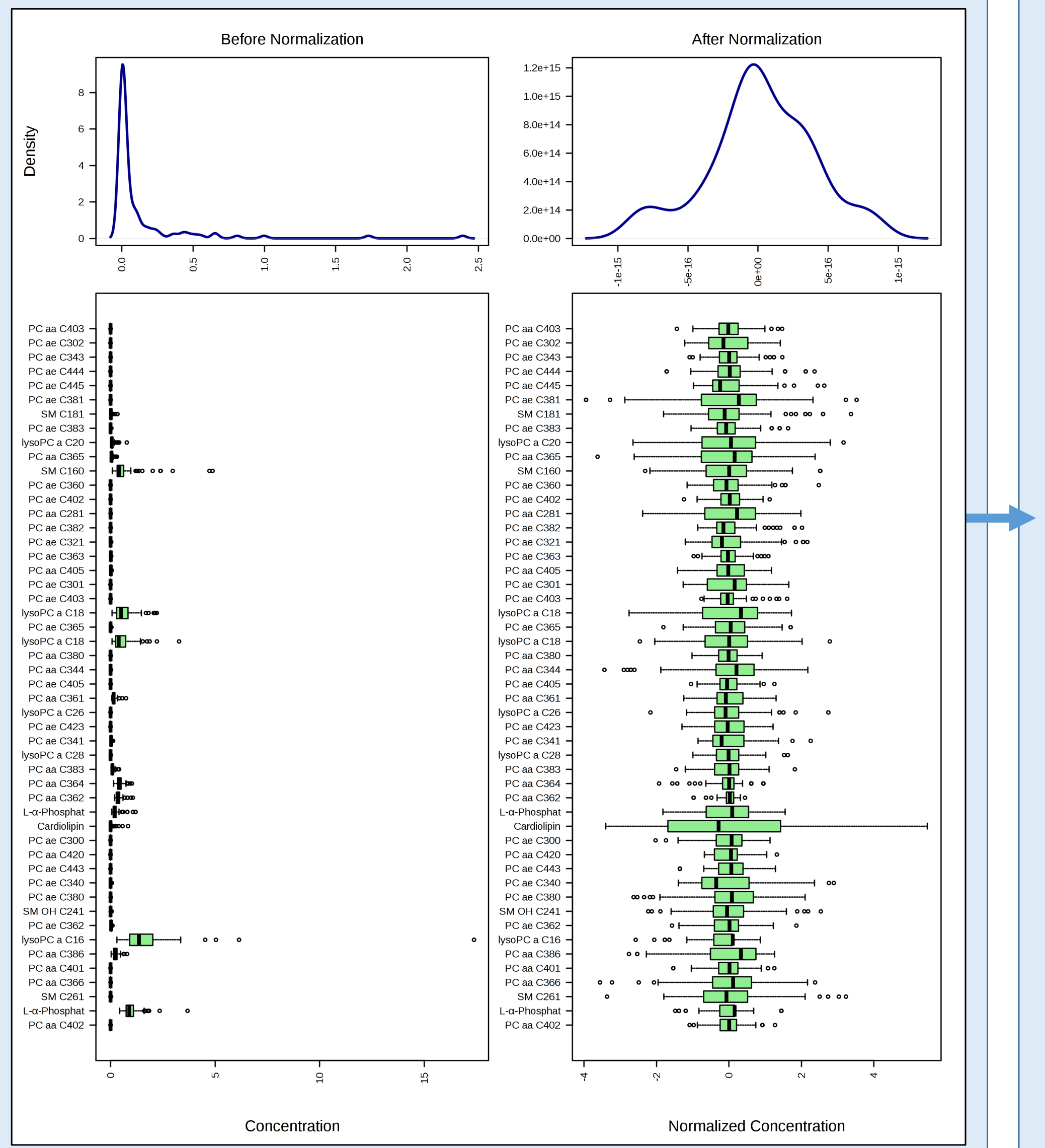


Figure 1. Example of box plots and kernel density plots before and after normalization. Normalization is essential to optimize and eventually generalize the performance of ML classifiers. Successful normalization of the datasets for all metabolite classes was achieved. The normalization results of a profile comprised of 107 phospholipids are shown above as an example.

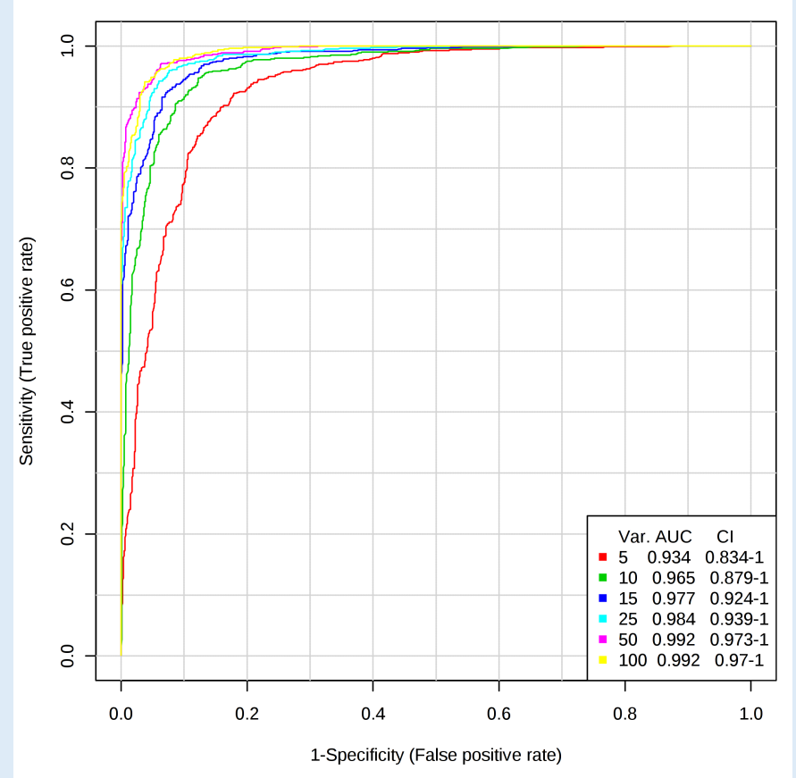


Figure 2. ROC curve based on lipid signatures derived from SVM.

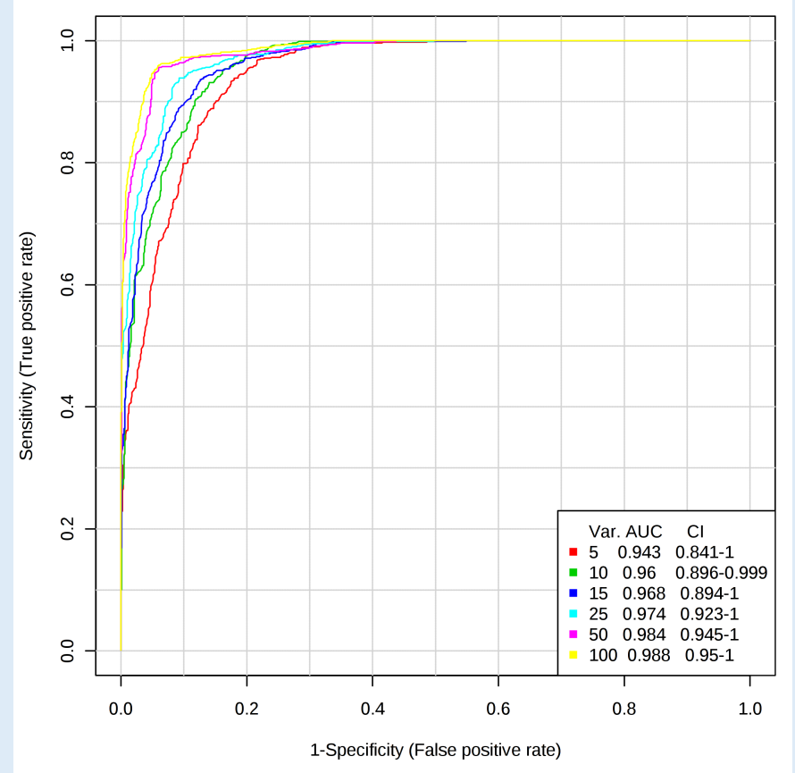


Figure 3. ROC curve based on lipid signatures derived from PLS-DA.

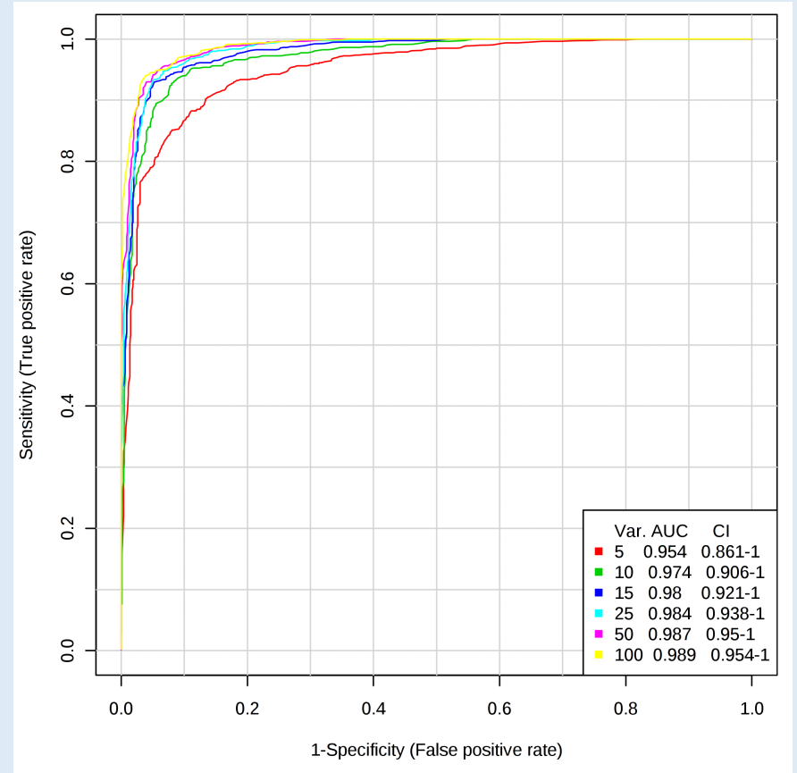


Figure 4. ROC curve based on lipid signatures derived from RF ML.

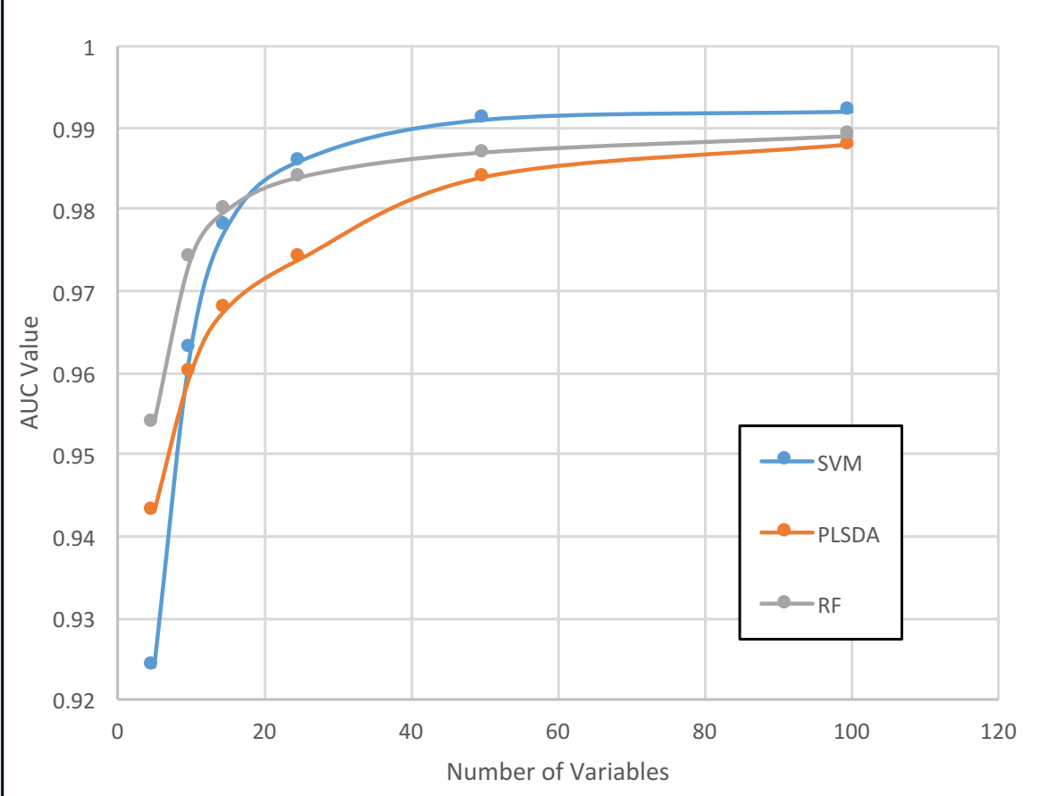


Figure 5. An example of how we compared the ML algorithms based on their AUC values. AUC reflects overall classification performance for distinguishing HCC from liver tissue. Performance is shown as a function of the number of variables (i.e. metabolites) in the signature. This example is from our analysis of phospholipid signatures, which performed the best from among signatures that included SM, BA, and FFA metabolites (see Table 1).

Table 1

Areas under the ROC curve values (AUC) for the first 50 variables of each metabolite class using SVM ML algorithm. Signatures using 10 metabolites shown in bold to point out their relative performance (BA < FFA < SM < Phospholipids).

Metabolite class	Number of Metabolites	AUC	Lower bound 95% CI	Upper bound 95% CI
Small molecules (including amino acids and energy metabolites)	3	0.807	0.669	0.922
	5	0.89	0.752	0.989
	10	**0.934**	**0.843**	**0.991**
	20	0.947	0.838	1
	38	0.943	0.839	1
Bile Acids	2	0.681	0.44	0.869
	3	0.678	0.448	0.855
	5	0.688	0.446	0.854
	10	**0.695**	**0.51**	**0.83**
	20	0.678	0.498	0.838
	42	0.682	0.514	0.8
Free Fatty Acids	2	0.839	0.63	0.975
	3	0.861	0.593	0.979
	5	0.892	0.676	0.977
	10	**0.895**	**0.779**	**0.969**
	20	0.858	0.731	0.953
	47	0.829	0.712	0.927
Phospholipids	5	0.924	0.818	0.979
	10	**0.963**	**0.891**	**0.999**
	15	0.978	0.896	1
	25	0.986	0.94	1
	50	0.991	0.969	1

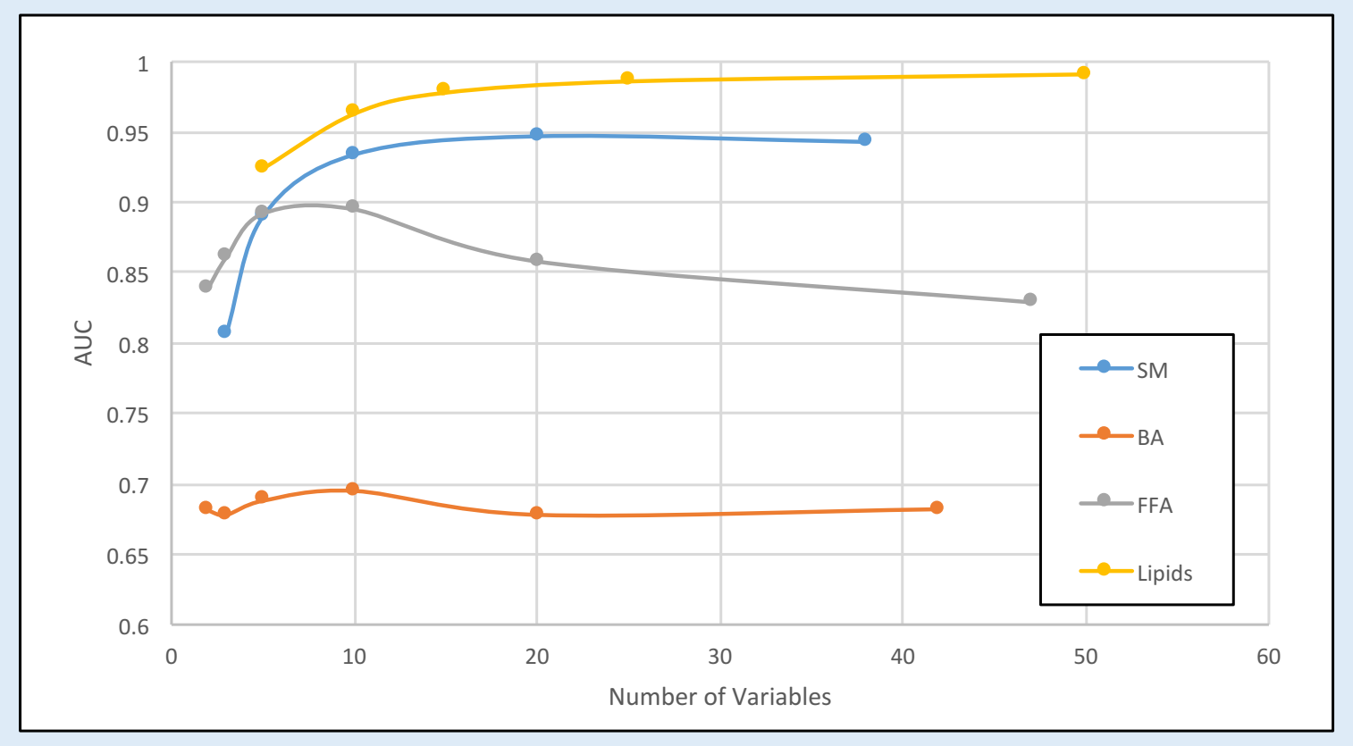


Figure 6. After we identified that the SVM ML algorithm consistently performed better than PLS-DA and RF in the lipid data, we compared signatures derived by SVM ML for the four metabolite classes (SM, BA, FFA, and Lipids). The lipid signatures consistently outperformed other signatures derived from SM, BA, and FFA metabolites regardless of the signature size (shown as number of variables)

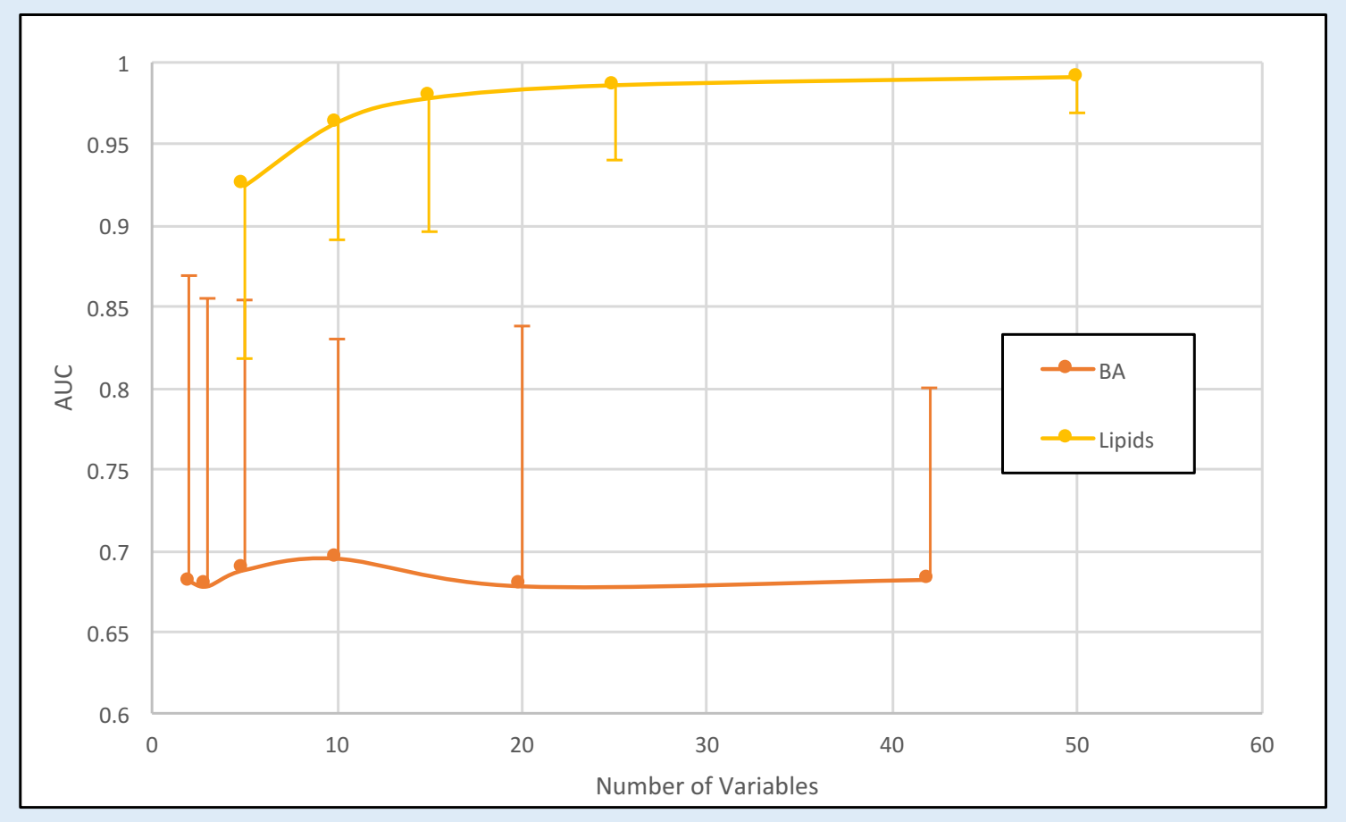


Figure 7. Comparison of AUC values (with 95% confidence interval bands) for signatures derived from bile acids versus phospholipids. With the SVM ML algorithm, the AUC values corresponding to lipid signatures were significantly higher than those corresponding to BA as signature size increased to > 10 metabolites. Therefore, lipid signatures are expected to outperform bile acid signatures for distinguishing HCC from normal liver tissue. As this study comprised only a discovery phase, further testing and validation in other datasets will be necessary to confirm the diagnostic performance of lipid signatures for HCC.

Conclusions and Future Directions

- For the FFA and Lipids, SVM was the best at discriminating between tumor and non-tumor sample (Figure 2), and between true negatives and positives in cross validation. However, for SM and BA, the RF algorithm performed better.
- At times, one algorithm worked better with fewer variables while another at a higher number of variables (Figure 5), but all worked generally well.

Among the three ML algorithms, the support vector machine learning algorithm was associated with the highest AUC values for the free fatty acids and the phospholipids (Figure 5); BUT the random forest machine learning was associated with the highest AUC values for the small molecules and bile acids.

- Nonetheless, there was no statistically significant difference between any of the three algorithms; null hypothesis accepted.
- Using the RF algorithm, AUC values of signatures derived from each metabolite class was compared as part of the DISCOVERY phase for a metabolomic signature that could potentially discriminate HCC. Signatures derived from phospholipid metabolites were found to consistently outperform signatures derived from other metabolite classes, and in particular those derived from bile acid metabolites.

Based on the results, we now hypothesize that phospholipid signatures have the potential to accurately distinguish between HCC and normal liver tissue (Figure 6, Table 1).

- Future directions will include identifying the cellular pathways associated with the metabolomic signatures identified in this study. Specifically, molecular pathway analysis will be performed based on the differentially expressed metabolites identified by ML in this study. Cellular and molecular pathways identified in this manner could potentially be exploited for therapeutic gain in HCC.
- We will also pursue further TESTING and VALIDATION to determine whether phospholipid signatures can serve as robust biomarkers for clinical diagnosis or molecular sub-classification of HCC.

Acknowledgments

We are extremely grateful to Lori Tsue and Sherry Chan for coordinating the Queen's Summer Research Internship 2019 and for all their help in preparing and guiding us towards the completion of this research study.

This research study was supported by Queen's Pilot Research Study Grant SE-2018-002 and National Institute of Health/National Cancer Institute grant R01CA161209-06.

References

[1] Marrero, J.A., Kulik, L.M., Sirlin, C.B., Zhu, A.X., et al. (2018). Diagnosis, Staging, and Management or Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology 2018; 68:723-50* Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29624699

[2] Guo, W., Tan, H., Wang, N., Wang, X., & Feng, Y. (2018). Deciphering hepatocellular carcinoma through metabolomics: from biomarker discovery to therapy evaluation. *Cancer Management and Research 2018; 10:715-34.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5903488/pdf/cmar-10-715.pdf

[3] Wei, R., Wang, J., Wang, X., Xie, G., Wang, Y., Zhang, H., Peng, CY., Rajani, C., Kwee, S., Liu, P., & Jia, W. (2018). Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning.