Supporting Information:

Aggregated Molecular Phenotype (AMP) Scores: Enhancing Assessment and Visualization of Mass Spectrometry Imaging Data for Tissue-Based Diagnostics

Jessie R. Chappel[1], Mary E. King[2], Jonathon Fleming[1], Livia S. Eberlin[2*], David M. Reif[3*], Erin S. Baker[4*]

[1] Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, NC 27606, USA

[2] Department of Surgery, Baylor College of Medicine, Houston, TX 77030, USA

[3] Predictive Toxicology Branch, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Durham, NC 27709, USA

[4] Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

*** Correspondence:**
Co-corresponding Authors
eberlin@IDCm.edu; david.reif@nih.gov; erinmsb@unc.edu

Table of Contents

Methods:

MSI Datasets

AMP scores were constructed and validated using metabolomic MSI data from three previously published studies. These studies evaluated frozen human tissue sections using desorption electrospray ionization (DESI) MSI to characterize the molecular profile of the different phenotypes. [1-3] In this work, we considered a subset of features from negative ion mode data from each study, with each feature having an *m/z* value, associated signal intensity (abundance), and x and y coordinates, associated with the MS image. In total, 210 tissue samples were analyzed including: 20 follicular thyroid adenoma (FTA) (15,088 pixels), 20 papillary thyroid carcinoma (PTC) (13,554 pixels), 41 normal breast (NB) (1,674 pixels), 81 invasive ductal carcinoma breast cancer (IDC) (31,639 pixels), 1 heterogeneous NB/IDC (1,910 pixels), 15 normal ovarian (NO) (11,126 pixels), 14 borderline ovarian tumor (BOT) (3,759 pixels), 14 high grade serous carcinoma (HGSC) (9,965 pixels), and 2 heterogeneous NO/HGSC (7,304 pixels) samples. Five pairwise phenotype comparisons were then assessed in the study: (1) FTA vs. PTC samples from DeHoog *et al.*, (2) NB vs. IDC samples from Porcari *et al*., and (3) NO vs. BOT, (4) NO vs. HGSC, and (5) BOT vs. HGSC from Sans *et al*. Further information on data collection can be found in their respective publications.[1-3]

Data Pre-Processing

For each pairwise comparison, data were filtered by first binning each *m/z* value to the nearest hundredth and removing *m/z* values that appear in fewer than 10% of pixels. For all thyroid and ovarian tissue samples, the *m/z* range was restricted to 100-1200. Breast tissue data was collected across multiple laboratories, and cosine similarity analyses of the full mass spectra obtained from

the different labs revealed that a higher m/z range was necessary to achieve high spectral similarity. The *m/z* value range for the breast tissue samples was thus restricted to 700-1200 as done in the original manuscript to maintain consistency.[2] Once all unique features were identified, missing values were imputed using the minimum detected value in the dataset. Following filtering, normality was assessed using density plots, which revealed a strong right skew. Therefore, all abundances were $\log_2$ transformed and then normalized at the sample level by dividing by the feature with the highest signal intensity, or abundance, value. The data were then assessed for sample outliers and batch effects were then assessed using principal components analysis (PCA).

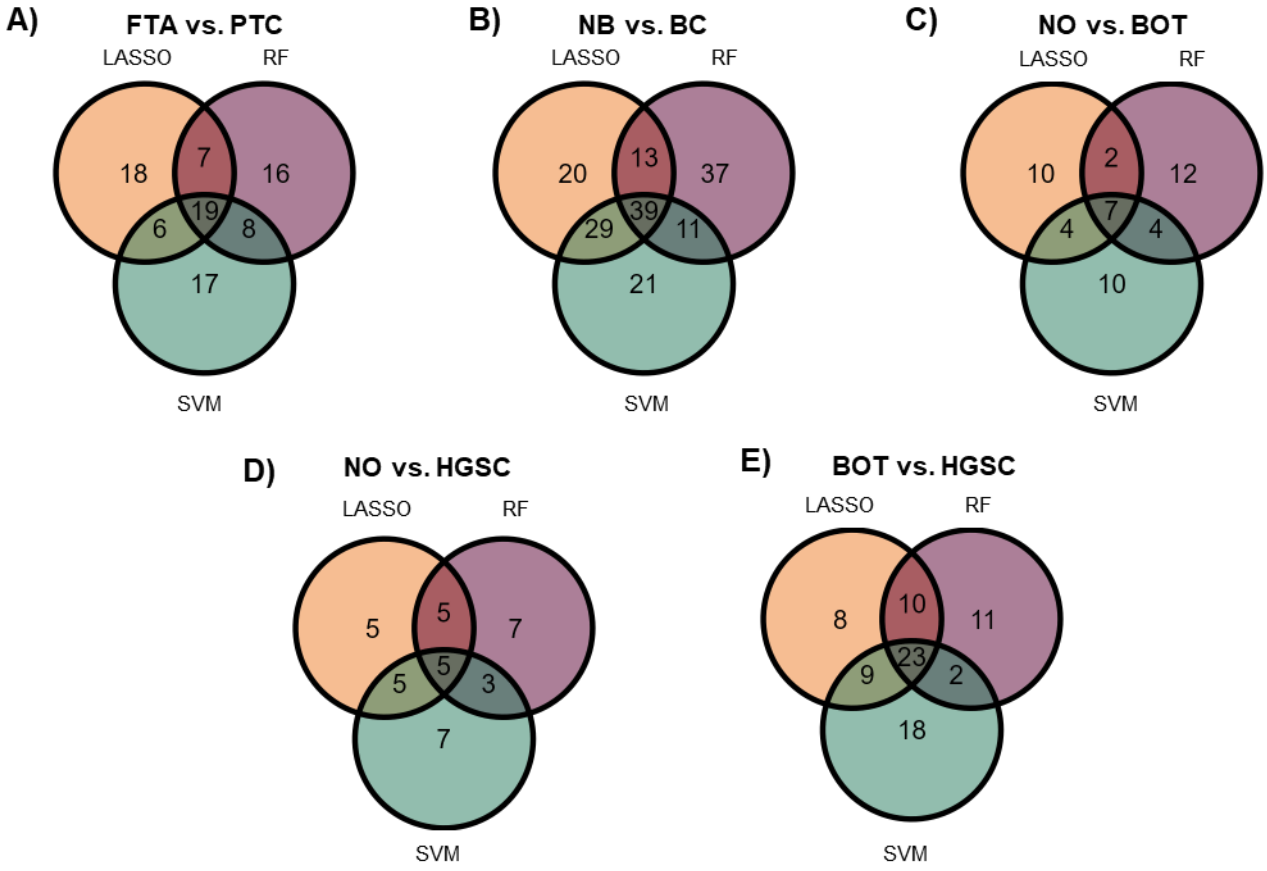

Figure S1: Venn diagrams showing the overlap of features selected by Lasso, random forest (RF) and support vector machine (SVM) for A) FTA vs. PTC samples, B) NB vs. BC samples, C) NO vs. BOT samples, D) NO vs. HGSC, and E) BOT vs. HGSC samples.

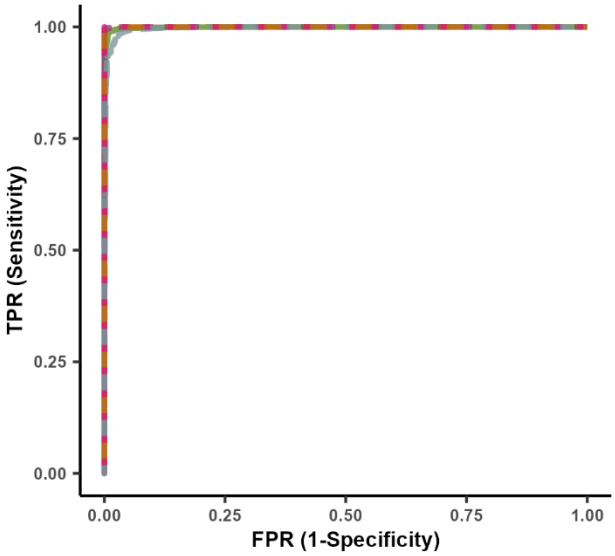

Figure S2: Receiver operator characteristic (ROC) curves for phenotype prediction performance of AMP scores for each pairwise comparison.

LDA Comparison:

Introduction:

Linear discriminant analysis (LDA) is a widely used statistical technique for classification tasks. It seeks to find an optimal linear combination of input features that maximizes the separation between classes of interest, enabling the prediction of class labels and estimation of associated posterior probabilities. In mass spectrometry imaging (MSI) analyses, LDA models are initially constructed using labeled pixels, and subsequently, unlabeled testing sets are projected onto the discriminant axes to facilitate classification.[4] Analogous to AMP scores, the resulting posterior probabilities can be leveraged not only for pixel class prediction but also for the creation of probability heatmaps, enabling the visualization of tissue sections based on the likelihoods of different classes. Based on these similarities, we aim to assess how AMP scores compare to LDA output for both the classification of pixels and the visualization of tissue sections.

Methods:

Data were pre-processed and split into training and testing sets using the same steps used for AMP score analysis. LDA models were built using each of the respective training sets using the function 'lda' function from the MASS package in R (v4.2.1). These models were then used to make predictions on the testing data. A posterior probability threshold value of 0.5 was used, so that pixels with a probability less than 0.5 were predicted to be class 1, while pixels with probabilities greater than or equal to 0.5 were predicted to be class 2. Predicted phenotype labels were then compared to true pixel labels from pathology, and accuracy, sensitivity, and specificity were calculated using functions from the R package caret. Boxplots and AMP score heatmaps were then made in R using the package ggplot2 to compare results.

Results and Discussion:

In general, pixel classification with posterior probabilities resulting from LDA demonstrated high accuracy, sensitivity, and specificity. This trend was best seen in the healthy or benign versus diseased tissue sections, where performance measurements were above 99%, except for the sensitivity in the NB versus IDC comparison, which was 92.6%. These results were comparable to classification with AMP scores, which had all performance metrics above 97% for healthy or benign versus diseased tissue comparison, except for the sensitivity in the NB versus IDC comparison, which was 92.7%. The most notable disparity in performance was seen for the BOT versus HGSC comparison, where LDA had 91.5% accuracy, while AMP scores had 96.1% accuracy. This discrepancy is primarily due to LDA having difficulties classifying HGSC pixels, which is reflected in the lowered sensitivity of the model (79.1%). These results suggest that AMP scores may have an advantage when trying to distinguish pairwise disease phenotypes.

Comparison	Accuracy	Sensitivity	Specificity
FTA vs. PTC	99.9%	99.9%	99.9%
NB vs. IDC	99.1%	92.6%	99.5%
NO vs. BOT	99.8%	99.7%	100%
NO vs. HGSC	99.9%	99.9%	100%
BOT vs. HGSC	91.5%	79.1%	96.2%

Table S1: Summary of LDA performance for each pairwise comparison. From top to bottom: Follicular thyroid adenoma (FTA) vs. papillary thyroid carcinoma (PTC), normal breast (NB) vs. invasive ductal carcinoma (IDC), normal ovarian (NO) vs. borderline ovarian tumor (BOT), NO vs. high grade serous carcinoma (HGSC), and BOT vs. HGSC.

While prediction performance was similar between LDA and AMP scores, there were notable differences in the distribution of posterior probabilities and AMP scores. Boxplots depicting the LDA posterior probabilities revealed narrow interquartile ranges, indicating that the majority of predictions were clustered near 0 or 1. This is in contrast with the distribution of AMP scores, which were more normal in

shape and typically had a greater interquartile range. However, despite probabilities having a narrower interquartile range, they often resulted in more extreme misclassification. For example, while the highest class 1 score assigned to a pixel with AMP scores was 0.645, LDA assigned a probability of 1 to several class 1 pixels. This shows that LDA was more prone to assigning absolute probabilities for misclassifications, whereas AMP scores provided a wider range of scores for misclassified pixels.

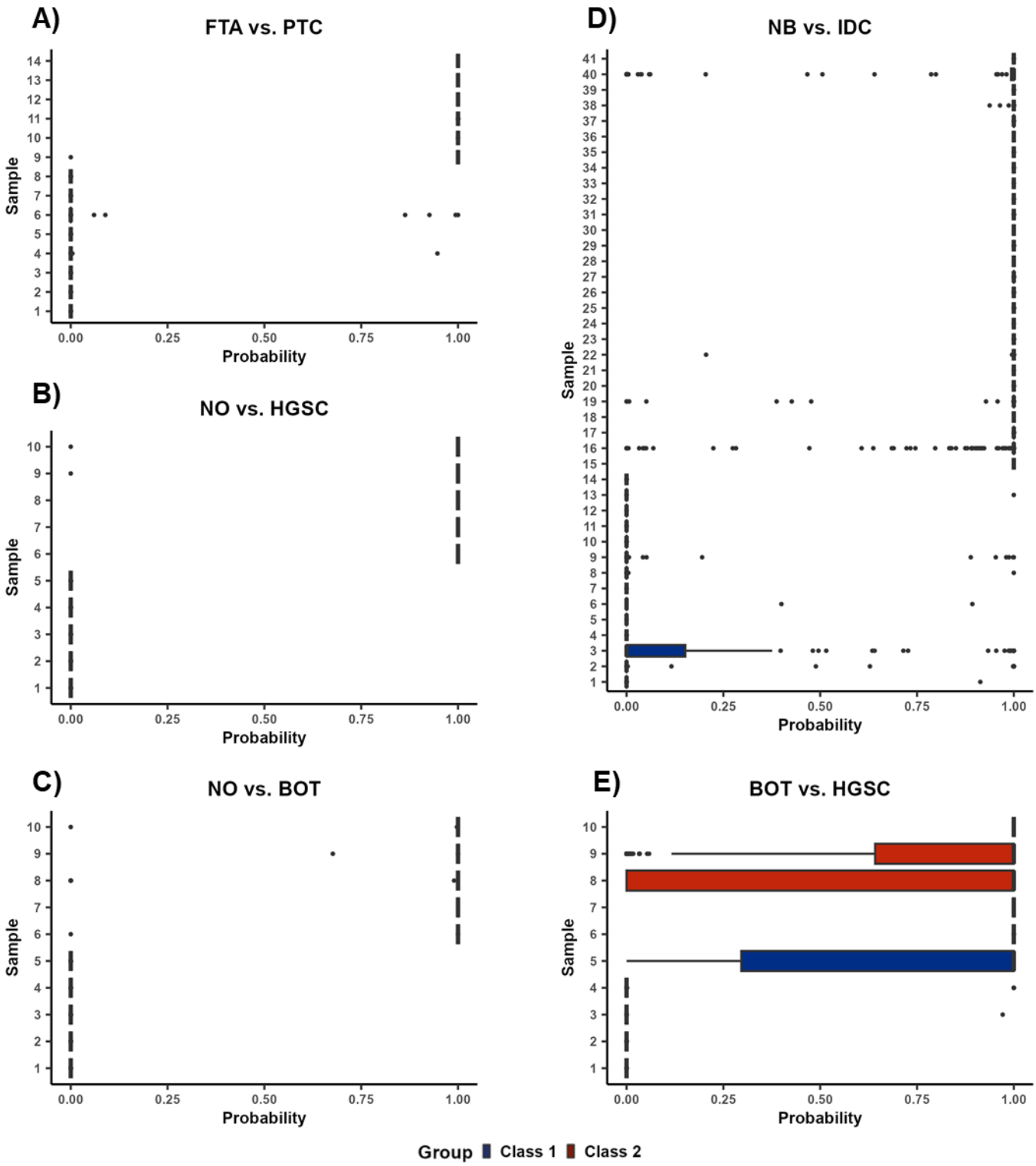

Figure S3: LDA poster probabilities for A) FTA vs. PTC, B) NO vs. BOT, C) NO vs. HGSC, D) NB vs. BC, and E) BOT vs. HGSC. Each individual box and whiskers plot shows the distribution of probabilities across all pixels within a sample. Outlier points are defined as observations more than 1.5 times the interquartile range away from the upper or lower quartile.

The tendency to only predict extreme values was further demonstrated when visualizing heterogeneous tissue sections. Rather than highlighting a transition region between phenotypes, plotted probabilities showed rigid borders between phenotypes. Further, several extreme misclassifications can be seen, particularly in panel B and C. These heatmaps also failed to line up as closely to pathology annotations as the AMP score heatmaps, which can be best seen in panel A where the probability heatmap classifies the bottom of the tissue as mostly cancerous, failing to capture that there are small normal regions throughout. In all, these results suggest that while LDA usually accurately classifies pixels, posterior probabilities do not identify subtle changes. This highlights the utility of AMP scores, which not only preserve high classification performance on homogeneous tissue sections, but also are able to capture gradual changes in heterogeneous tissues.

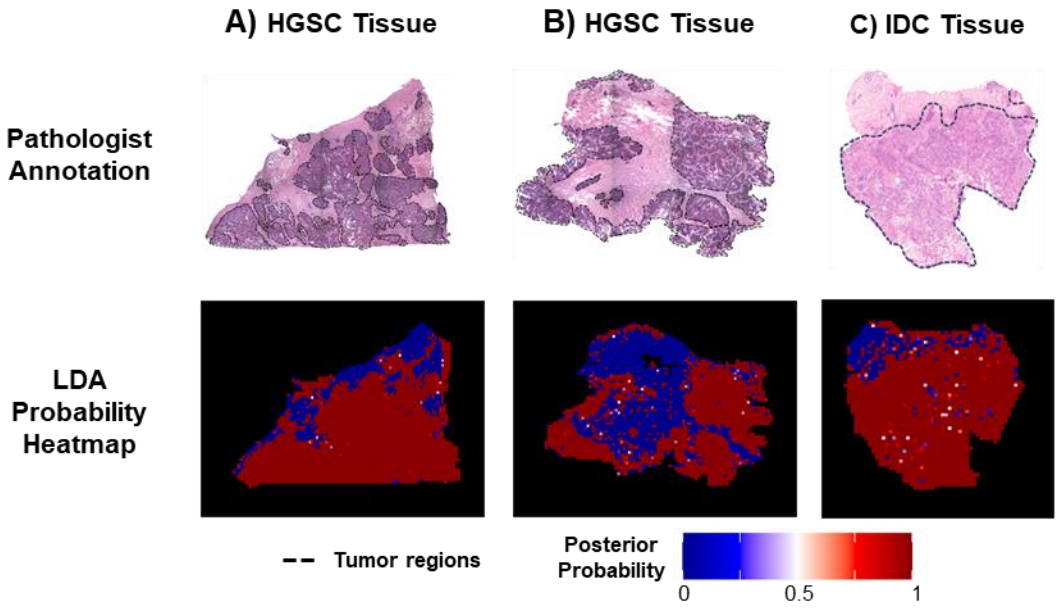

Figure S4: Comparison of LDA probability heatmaps for heterogeneous tissues to annotated H&E slides for example HGSC tissue shown in A) and B), IDC tissue in C). On the H&E slides, tissue outlined in black corresponds to tumorous areas with healthy tissue in the other regions.

Phenotype Border Analysis:

To investigate the molecular composition of the transition region between phenotypes, we present six mass spectra obtained from six adjacent pixels in a heterogeneous tissue section. These pixels originate from a normal region and transition into cancerous tissue. By examining the AMP scores of these pixels, it becomes evident that rather than a sharp change at the border, AMP scores progress gradually, with increases ranging from 0.048 to 0.123 between adjacent pixels (**Figure S5)**. The differences in these scores can be traced back to both the raw mass spectra for these pixels, as well as filtered mass spectra, which present only the features that were selected and used in AMP score calculation. From the raw mass spectra, it can be seen that one characteristic that delineates the AMP scores is the number of features detected in the low mass range (100-500) (**Figure S5)**. Pixels with lower AMP scores exhibit a high number of features in this range, with many of them having a considerable relative abundance. As we transition to pixels with intermediate AMP scores, the quantity of features in this range begins to decrease, accompanied by a decline in their relative abundance. This trend persists as the AMP scores assigned to a pixel increases, indicating that AMP scores correspond to distinct molecular profiles. In the presented filtered spectra, only features selected for AMP score modeling were used, and the color of these features in the presented spectra is based on the weight resulting from logistic regression in the AMP score pipeline (**Figure S6)**. Features with a dark red color represent features whose increased abundance is strongly associated with cancerous tissue, and those in darker blue represent features whose increased abundance is strongly associated with normal tissue. In these plots, it can be seen that pixels with low AMP scores tend to have mainly normal associated features detected, while those with higher AMP scores have primarily cancer associated features detected. In all, we believe these spectra demonstrate that pixels with intermediate AMP scores share molecular characteristics with both normal and cancerous phenotypes. Thus, the localization of pixels with intermediate scores at phenotype borders suggests that biological differences between phenotypes occurs gradually.

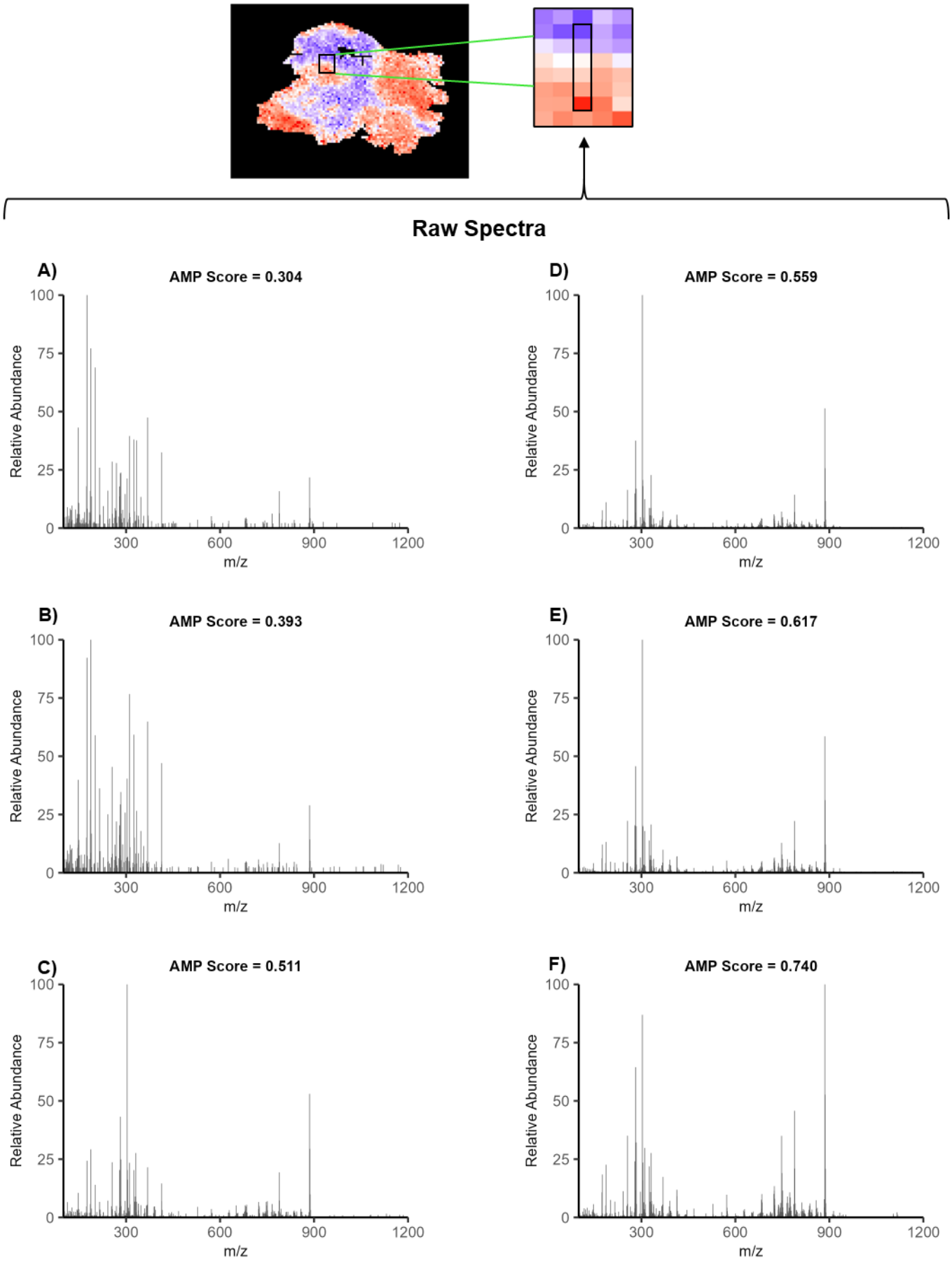
Raw Spectra
A)
AMP Score = 0.304
B)
AMP Score = 0.393
C)
AMP Score = 0.511
D)
AMP Score = 0.559
E)
AMP Score = 0.617
F)
AMP Score = 0.740
Relative Abundance
m/z
0
25
50
75
100
300
600
900
1200

Figure S5: Raw mass spectra of six adjacent pixels in a heterogeneous HGSC tissue section **(Figure 7B)** that span the transition region between normal and diseased tissue.

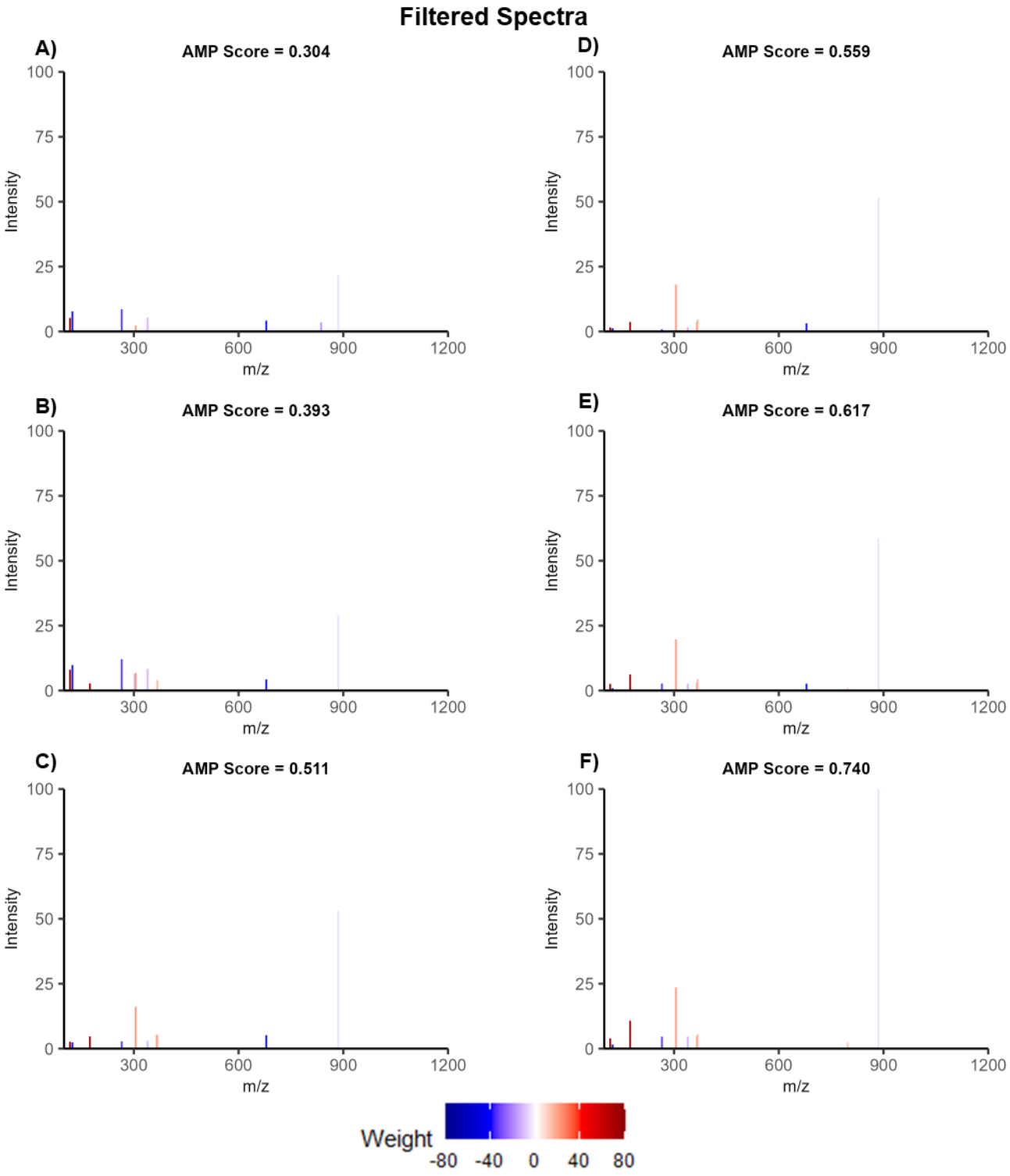

Figure S6: Filtered mass spectra of six adjacent pixels in a heterogeneous HGSC tissue section (**Figure 7B**) that span the transition region between normal and diseased tissue. This figure matches **Figure S5**, but only includes selected features. Intensities are colored according to the weight that feature had in AMP score calculation, with blue values indicating an association with normal tissue and red colors having an association with diseased tissue.

(1) DeHoog, R. J.; Zhang, J.; Alore, E.; Lin, J. Q.; Yu, W.; Woody, S.; Almendariz, C.; Lin, M.; Engelsman, A. F.; Sidhu, S. B.; et al. Preoperative metabolic classification of thyroid nodules using mass spectrometry imaging of fine-needle aspiration biopsies. *Proc Natl Acad Sci U S A* **2019**, *116* (43), 21401-21408. DOI: 10.1073/pnas.1911333116 From NLM Medline.

(2) Porcari, A. M.; Zhang, J.; Garza, K. Y.; Rodrigues-Peres, R. M.; Lin, J. Q.; Young, J. H.; Tibshirani, R.; Nagi, C.; Paiva, G. R.; Carter, S. A.; et al. Multicenter Study Using Desorption-Electrospray-Ionization-Mass-Spectrometry Imaging for Breast-Cancer Diagnosis. *Anal Chem* **2018**, *90* (19), 11324-11332. DOI: 10.1021/acs.analchem.8b01961 From NLM Medline.

(3) Sans, M.; Gharpure, K.; Tibshirani, R.; Zhang, J.; Liang, L.; Liu, J.; Young, J. H.; Dood, R. L.; Sood, A. K.; Eberlin, L. S. Metabolic Markers and Statistical Prediction of Serous Ovarian Cancer Aggressiveness by Ambient Ionization Mass Spectrometry Imaging. *Cancer Res* **2017**, *77* (11), 2903-2913. DOI: 10.1158/0008-5472.CAN-16-3044 From NLM Medline.

(4) Liang, S.; Singh, M.; Dharmaraj, S.; Gam, L. H. The PCA and LDA analysis on the differential expression of proteins in breast cancer. *Dis Markers* **2010**, *29* (5), 231-242. DOI: 10.3233/DMA-2010-0753 From NLM Medline.