MassExplorer: A Tool to Help Guide Analysis of Mass Spectrometry Samples

Vishnu Shankar (vishnus1@stanford.edu)
Zare Lab (Dept. of Chemistry)

Last Updated: July 10, 2020

I. Motivation: In the last five years, several projects (1-3) have leveraged the sensitivity of desorption electrospray ionization mass spectrometry (DESI-MS) and other mass-spectrometry based methods to detect metabolites and other molecular species in a sample before using statistical learning methods to find important peaks that correlate with the disease state. Fig. 1 summarizes this methodology, consisting of sample collection, measurement of spectra, data processing, and statistical learning methods.

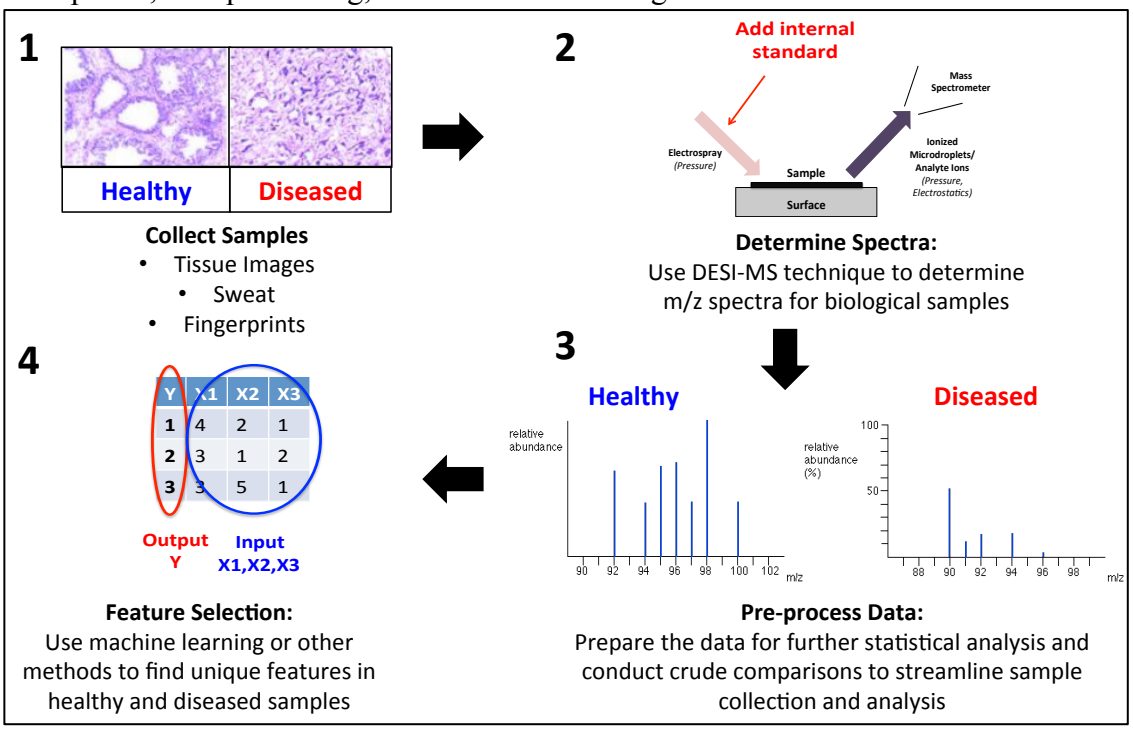


Fig. 1: Summary of experimental methodology and analysis

Given the wide applicability of this methodology, we motivate the development of an online aid *MassExplorer* (https://massexplore.shinyapps.io/massexplorer4/) with two aims:

- Automate repetitive statistical pre-processing
- Guide measurements by providing statistical insights and visualizations that are accessible to biologists and chemists

II. Modules: MassExplorer consists of 6 modules that are summarized below.
The modules below correspond to the online version of the application, which is suited for visualization and smaller jobs.

1. **Input Module**: The user can input *multiple* files corresponding to a "healthy" set and "disease" set. The application accepts files in either typical mass spectra distribution or in imaging format (as shown in the application). Additionally, the user can specify an internal standard peak, which is used in normalizing the intensities, a tolerance, which determines how much a peak can vary between the samples in order to be considered the same, and a threshold, which allows the user to exclude metabolites that are detected in only few samples.
2. **Pairwise Comparison**: The intensities are processed and normalized according to the chosen settings in the first panel. One can then visualize the processed data, selecting a spectrum of interest from each set. Also, the user can specify the range of interest and also look up the peaks in a table, which can be searched and sorted according to intensities.
3. **Get Training Matrix:** This module shows the average overlayed spectra between groups. Additionally, the user can download the formatted data to a .csv file, where the headers consist of the processed peaks and the rest of the table is populated by the intensities. The downloaded csv also consists of the file number in normal file formats or Scan, X, Y, Patient ID, Disease.State for imaging formats.

4. **Find Specific Differences:** Both datasets are processed according to the user inputted tolerance in the "Input Module." To determine the aggregate statistically significant differences between both sets, significance analysis of microarrays (SAM) (4) is used to calculate a modified t-statistic (i.e. q-statistic) in each cluster. The plot shows the fold change (log2) between sets for clusters with statistically significant differences, where the calculated q-statistic accounts for the false discovery rate (FDR). SAM is able to account for FDR by assigning a score to each detected analyte, before using permutations to estimate the percentage of analytes that are identified by chance. Compared to other procedures that calculate statistically significance, SAM is preferable, as it does not assume the data is normally distributed and can work equally well at both small and large sample sizes. The user here can also download the SAM plots and outputs to their local machine.
5. **Build Model:** To find the most informative peaks that can distinguish both groups, a binary logistic classifier is built using the LASSO (5). If the number of samples are too few, the model uses a 10-fold cross-validation to train the model and select the tuning parameter (how much shrinkage should be imposed on the selected co-efficients). Otherwise, the model is trained via a leave one patient out cross-validation. This module includes plots to indicate the cross-validation performance, the selected peaks, and the model performance.
6. **Visualize Peak Variability:** To further understand the peaks that can distinguish groups, one can type in any peak and visualize the distribution of a metabolite between both groups via a boxplot. Based on the entered peak, this module also helps visualize the spatial distribution of the metabolite.

III. References

1. Eberlin LS, Tibshirani RJ, Zhang J, Longacre TA, Berry GJ, Bingham DB, et al. Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *PNAS*. 2014;111:2436–2441. doi: 10.1073/pnas.1400274111

2. Eberlin LS, Margulis K, Planell-Mendez I, Zare RN, Tibshirani R, Longacre TA, et al. (2016) Pancreatic Cancer Surgical Resection Margins: Molecular Assessment by Mass Spectrometry Imaging. *PLoS Med* 13(8): e1002108. doi:10.1371/ journal.pmed.1002108

3. Eberlin, L. S., Norton, I., Orringer, D., Dunn, I. F., Liu, X., Ide, J. L., … Agar, N. Y. R. (2012). Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *PNAS*, (18), 1–6. http://doi.org/10.1073/pnas.1215687110

4. V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98:5116–5121, April 2001.

5. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc Ser B 1996;58: 267–88.

IV. Frequently Asked Questions (FAQs)

1. Why did the online application shutdown?

The application is built to shutdown under the following conditions:

- **The file format selected is incompatible with the files provided. See sample formats below.**

Select file format:

(•) Normal
( ) Imaging

Sample File:

Mass:Charge	Intensity
50.00	0.00
51.00	2.57
52.00	5.06
53.00	5.68
54.00	5.52
55.00	4.35

Select file format:

( ) Normal
(•) Imaging

Sample File:

Scan	360	Scan	361
X	3.8	X	4
Y	2.4	Y	2.4
Z	1	Z	1
m/z	Intensity	m/z	Intensity
50.13077	443.581421	50.120516	553.307312
50.130837	1167.427856	50.120583	1084.705078

Often, users will input imaging files in the *transpose* format (where each spectra is across the page), causing the application to exit.

2. Are there any best practices for using the application?

- **File naming**
  - The program looks for a number in the filename to uniquely map a set of mass spectra data to a patient or sample. For example, if the name is "file_1.csv", the program looks for 1 to assign a sample identifier to all the mass spectra in "file_1.csv".
  - In the current implementation, no two objects in the same group can have the same number. For instance, there can only be a single file corresponding to "healthy_1.csv". If there is both "healthy_1.csv" and "file_1.csv", the program will not know which set of spectra belong to the sample. This limitation does not apply across datasets, where one can have "healthy_1.csv" and "disease_1.csv".
  - Best practice: Simply name the files in a way where the numbers are unique in the same group and can be easily processed (e.g. healthy_1.csv, healthy_2.csv, healthy_3.csv, etc.)
- **What order to use the modules?**
  - Input Module: Check to see that the 'upload is complete' after inputting the files.
  - Pairwise Comparison: If this is a new batch of data, it is best to visualize the data in 'pairwise comparison' to get a sense as to whether the data has any contaminants or large electronic noise.
  - Get Training Matrix: To conduct any of the further processing in the application, a training matrix needs to be initially constructed. Therefore, this module produces a plot comprising the average spectra from both sample groups and allows the user to download a csv of the training matrix to the directory. The user can track the progress in building the training matrix in log.txt (see Q3).
  - Find Specific Differences: After building the training matrix in the previous module, if the sample size is small (e.g. couple 1000 spectra, 2 or 3 patients/mice tissue), it is appropriate to select this module. If the sample size is large, do not select this module. It will take time to process and lag the computation.
  - Build Model: After building the training matrix, if the sample size is large, it is appropriate to select this module.
  - Visualize Peak Variability: This module is highly complementary to "Find Specific Differences" and "Build Model". Specifically, one can type the m/z of any metabolite in this module and the program helps autocomplete the search. Subsequently, one can compare the variability in the intensity distribution across both groups, thereby helping understand why the statistical modules selected these m/z as relevant for classifying the groups. The spatial plots represent the average intensity at that particular X, Y pixel from all scans (0 intensity if

none available). This plot can help get a sense as to how the pixel intensities are spatially distributed.

3. How can I get a better sense of what the application is doing 'under the hood'?

To give users greater transparency, the newest version (released on July 12, 2020) now documents every pre-processing step that the application completes in a log.txt file found in the MassExplorer_Output directory in (~/Downloads/). This log provides a more detailed account of the application's progress. In addition, this directory, created by this application, consists of several R data objects that can be read using readRDS in the programming language R and processed further. The summary of saved objects is described below:

- fileA_data_raw, fileB_data_raw: The raw unprocessed mass spectrometry data from each file (A corresponds to first sample set, B the second). These objects are saved again, after the centroids are calculated (see below).
- 1_centroid_pks: The program calculates centroids to tolerate variance in the detected peaks within the specified tolerance in the input module. For example, it is likely that 89.023 and 89.025 both correspond to lactate. This part of the calculation ensures that these peaks are considered the same.
- 1_centroid_assignment: This is the most expensive part of the calculation, where every detected peak in every sample is assigned to the nearest centroid.
- 2_fileA/fileB_training_dat: After calculating the centroids, the maximum intensity peak is collected corresponding to each centroid. These objects correspond to a dataframe where each row is a unique pixel and each column is a different metabolite.
- 3_descriptor.table: This holds the meta-data for identifying each pixel, including the Scan, X, Y, Disease State, and Patient information.
- 3_training_table: This object combines both 2_fileA_training_dat and 2_fileB_training_dat to produce an overall training matrix.
- Glmnet_cv_model: This object corresponds to the lasso model (if the user chooses to use this module) that was constructed.
- SAM_results: This object holds the results from conducting the significant analysis of microarrays.

IV. Appendix (Sample Plots)

Figure S1: Input Module Screenshot

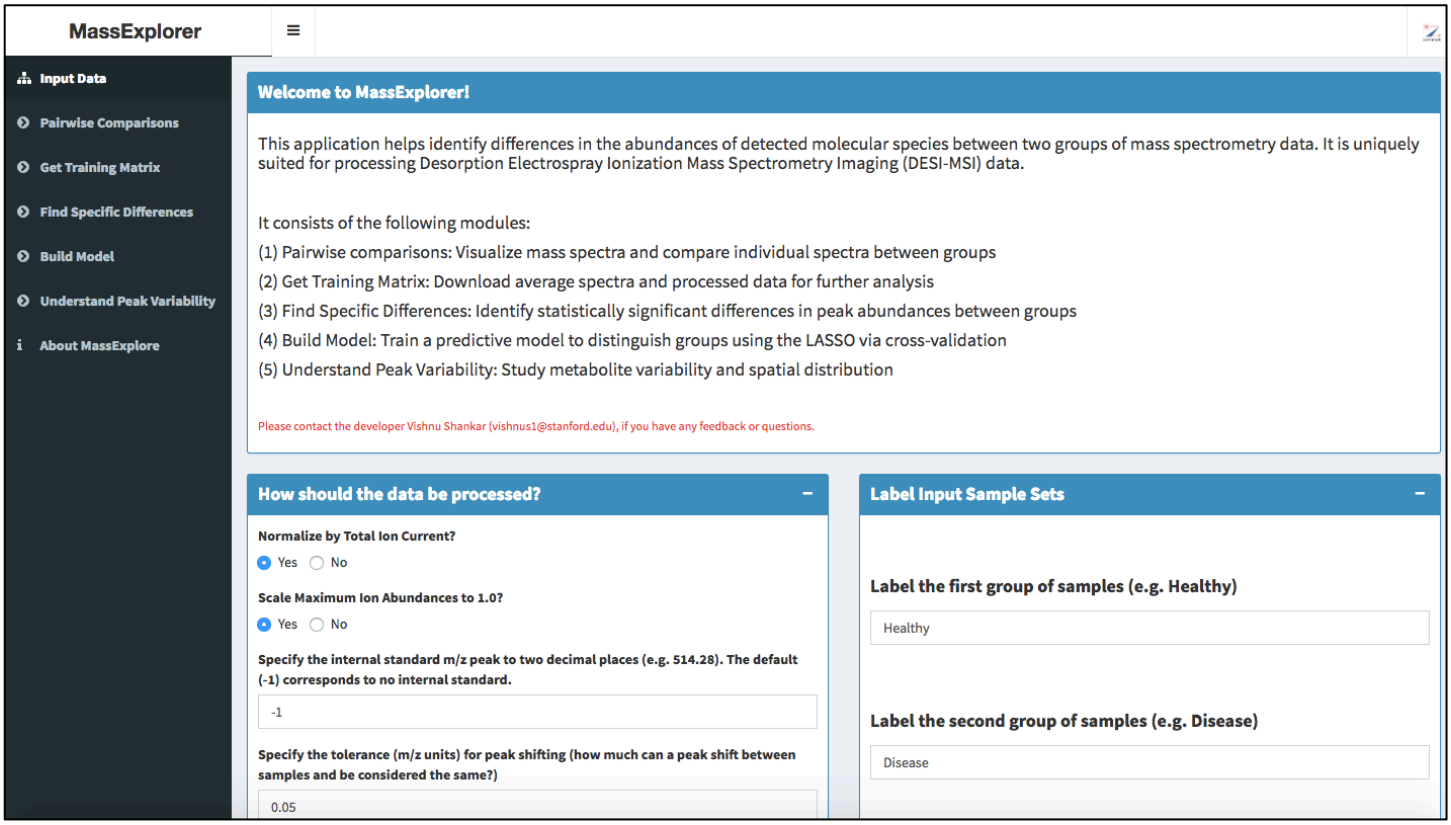


Figure S2: Pairwise Comparisons Module Screenshot

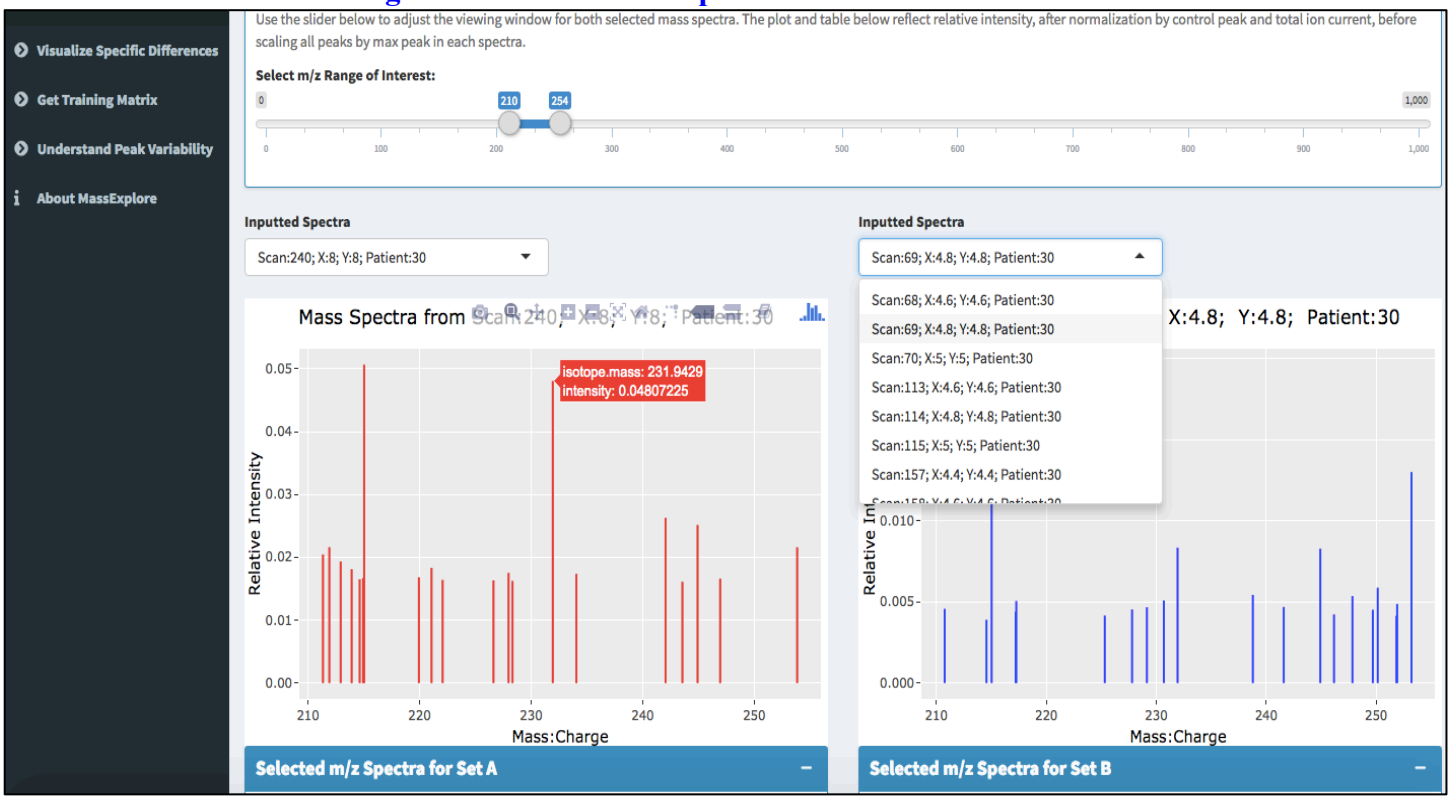


Figure S3. Average Spectra comparison between groups
Get Training Matrix Module Screenshot

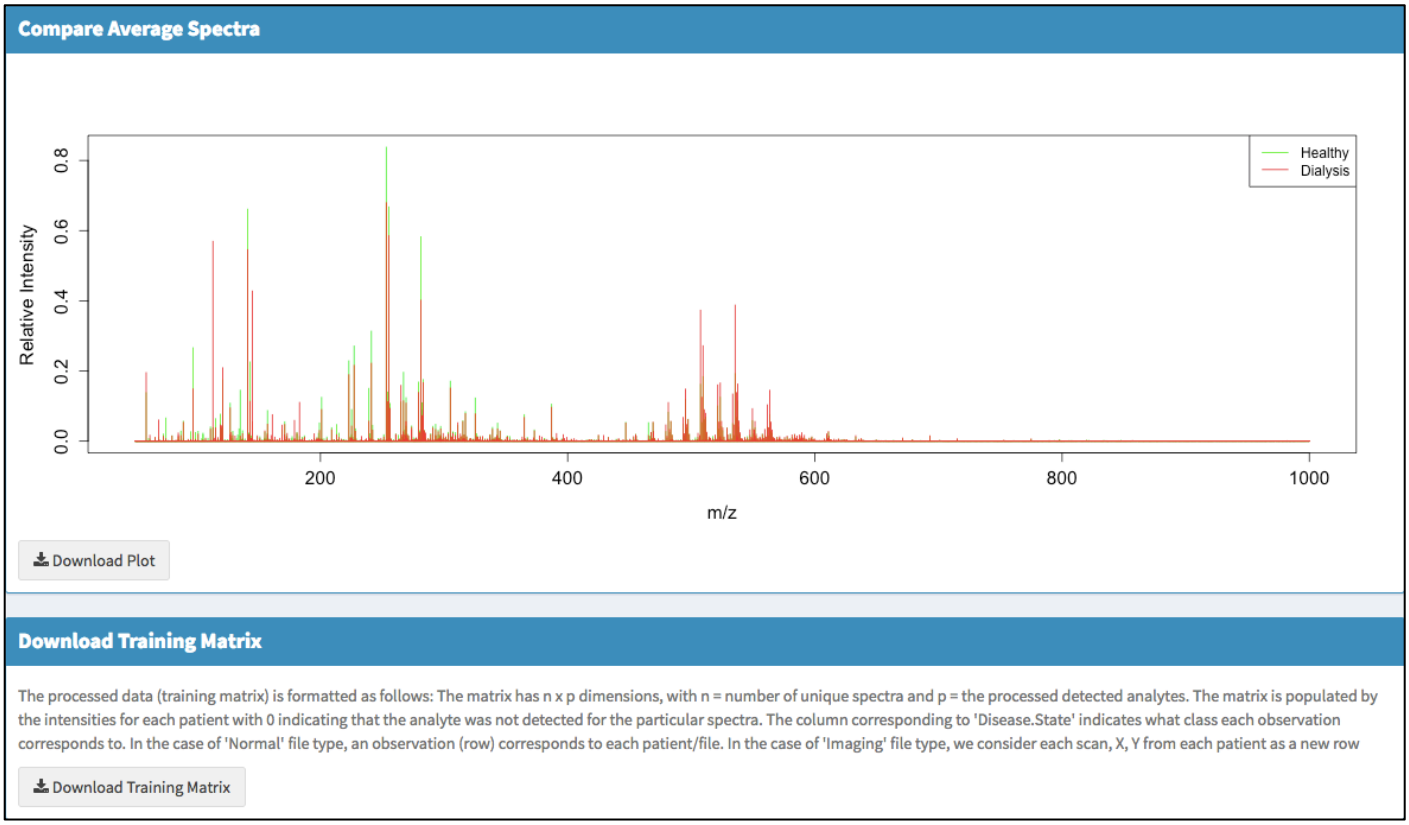


Figure S4: Find Specific Differences Module Screenshot

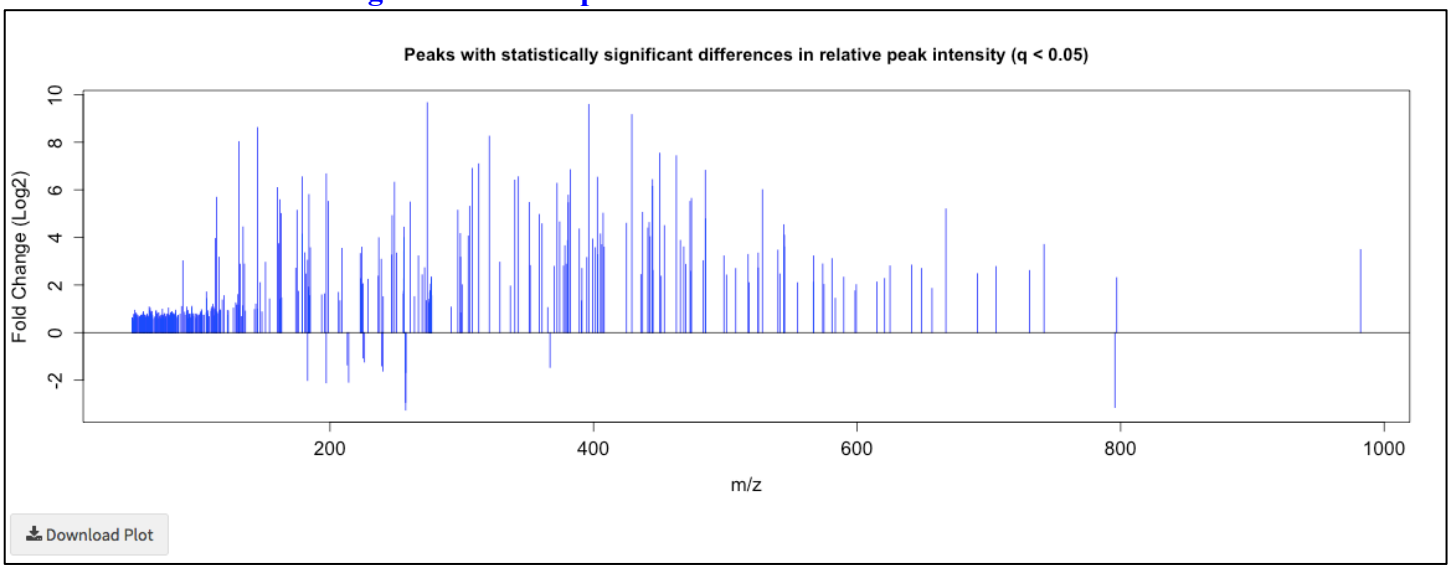


Table of Statistically Significant Peaks

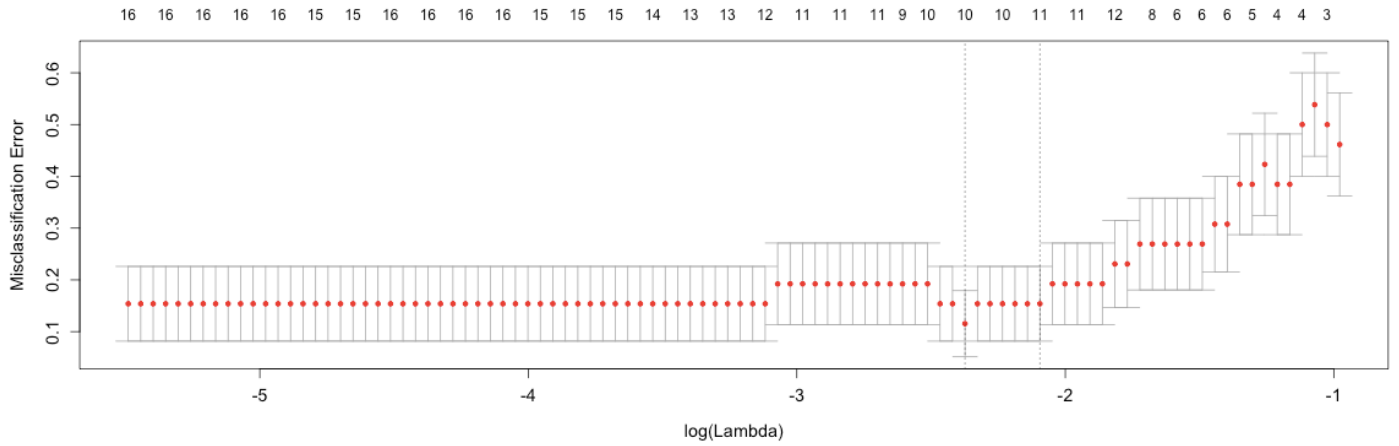
Show 25 entries Search:

Peak	Direction of Change	Fold Change
273.8211	Up regulated	809.204
396.3551	Up regulated	770.098
429.0496	Up regulated	572.079
144.9502	Up regulated	392.892

Figure S5a. Build Model Screenshot (Cross-validation performance)

Cross-validation Performance: Selection of Tuning Parameter

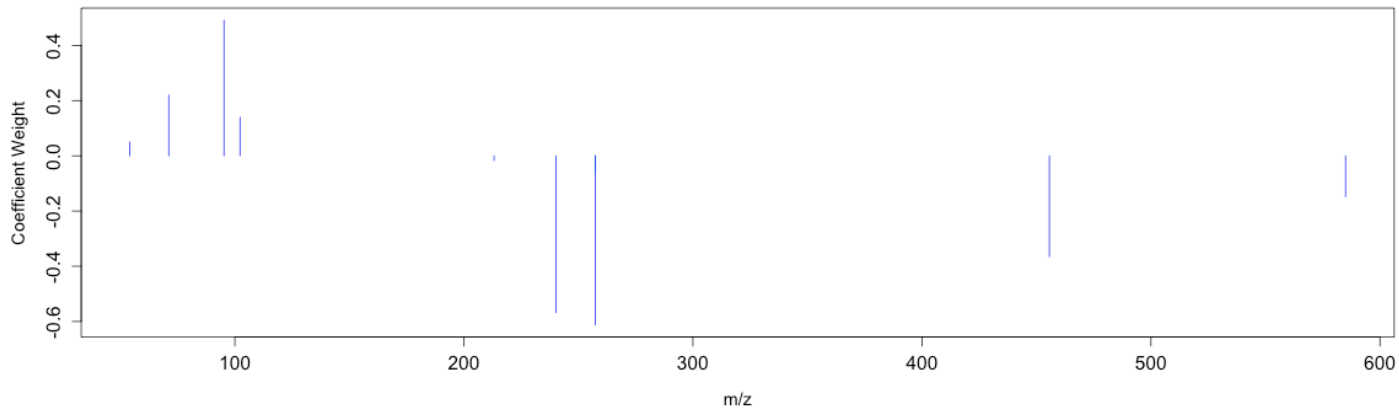
The plot below shows the cross-validation classification performance as a function of the tuning parameter (lambda), selected via cross-validation. The vertical dotted line corresponds to the selected model. The numbers on top of the plot indicate the number of non-zero peaks that correspond to the model with the particular tuning parameter and classification error.



Download Cross Validation Plot

Figure S5b. Selected LASSO Peaks

Selected LASSO Peaks



Download Plot Download Model

Show 25 entries Search:

name	coefficient
257.441469809524	-0.61224109
240.199153073171	-0.56719536
95.2475089672132	0.49096662
455.6442075	-0.36527133
71.1380233225806	0.21985984
(Intercept)	0.21633793

Figure S5c. LASSO Cross-validation Model Performance

Model Performance

Show 25 entries Search:

Description	Output
Number of Unique Detected Metabolites	22343
Number of Observations	26
Number of Selected LASSO Peaks	10
Method of Model Selection	Leave one patient out CV
Selected Model Tuning Parameter	0.115384615384615
Accuracy	0.85
True Positive Rate	0.79
False Positive Rate	0.21
True Negative Rate	0.92
False Negative Rate	0.08
Description	Output

Showing 1 to 10 of 10 entries Previous 1 Next

Download Performance Statistics

Figure S6a. Understand Peak Variability Module
Visualize variability of selected metabolite m/z 117.01 through boxplot

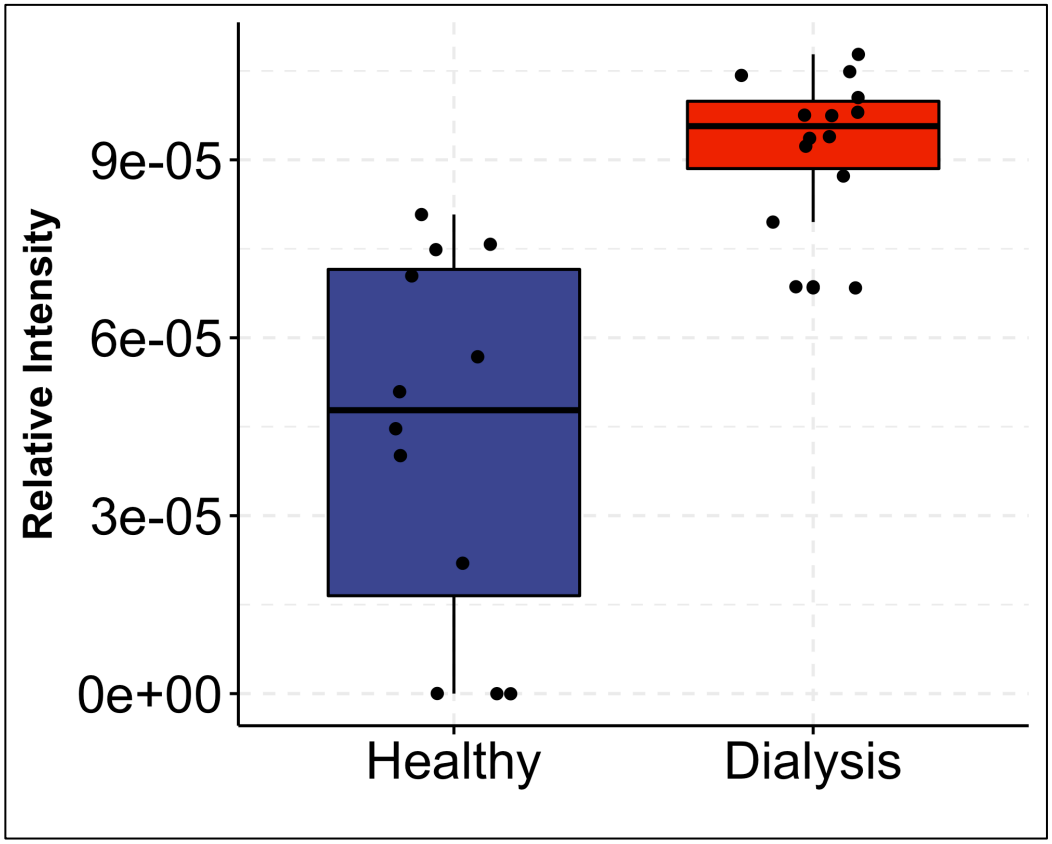


Figure S6b. Visualize spatial distribution of selected metabolite