

# **SinCHet-MS: Single Cell Heterogeneity analysis toolbox for Mass Spectrometry dataset**

## **Background:**

Single cell mass spectrometry (SCMS) enables the quantification of small molecules such as metabolites at single cell level but the analytics to investigate heterogeneity change is not easily available yet. The SinCHet-MS toolbox is developed in MATLAB and has a graphical user interface (GUI) for cell heterogeneity analysis and visualization. The toolbox provides the following features: 1) enable batch correction; 2) quantify cellular heterogeneity change among different conditions based on D statistics which is defined as the difference of the areas under the Shannon Profiles; 3) Determine the number of subpopulation according to a novel d statistics which is defined as the difference of Shannon index at the certain resolution and the changepoint derived from MARS model; and 4) prioritize markers at the subpopulation levels to explore heterogeneity evolution for further investigation using sGF(Subpopulation Generalized Fisher Product Score).

Renmeng Liu, Jiannong Li, Yunpeng Lana, Tra D. Nguyena, Y. Ann Chen\*, Zhibo Yang\*.

Identifying Changes of Cell Heterogeneity and Subpopulations of Live Cells Using Single Cell Metabolomics.

Corresponding author: [ann.chen@moffitt.org](mailto:ann.chen@moffitt.org)

## **What's included on the webpage:**

1. The pre-compiled standalone version SinCHet-MS for users without MATLAB license.
2. The example dataset consists of log2 transformation of M/Z value for 177 metabolites and 75 single cells including 39 of control human melanoma cancer cell line WM115, and 36 of WM115 cells treated with B-Raf inhibitor Vemurafenib at 1  $\mu$ M for 48 hours (see the detail in our paper).

| <b>Batch</b> | <b>Control</b> | <b>Treatment</b> |
|--------------|----------------|------------------|
| 1            | 13             | 18               |
| 2            | 26             | 18               |
| <b>Total</b> | 39             | 36               |

## **License conditions:**

The SinCHet-MS software is freely available for non-profit academic use. For licensing

opportunities, please contact [ann.chen@moffitt.org](mailto:ann.chen@moffitt.org) or [Haskell.Adler@moffitt.org](mailto:Haskell.Adler@moffitt.org) at Moffitt's Innovation office.

## **Installation:**

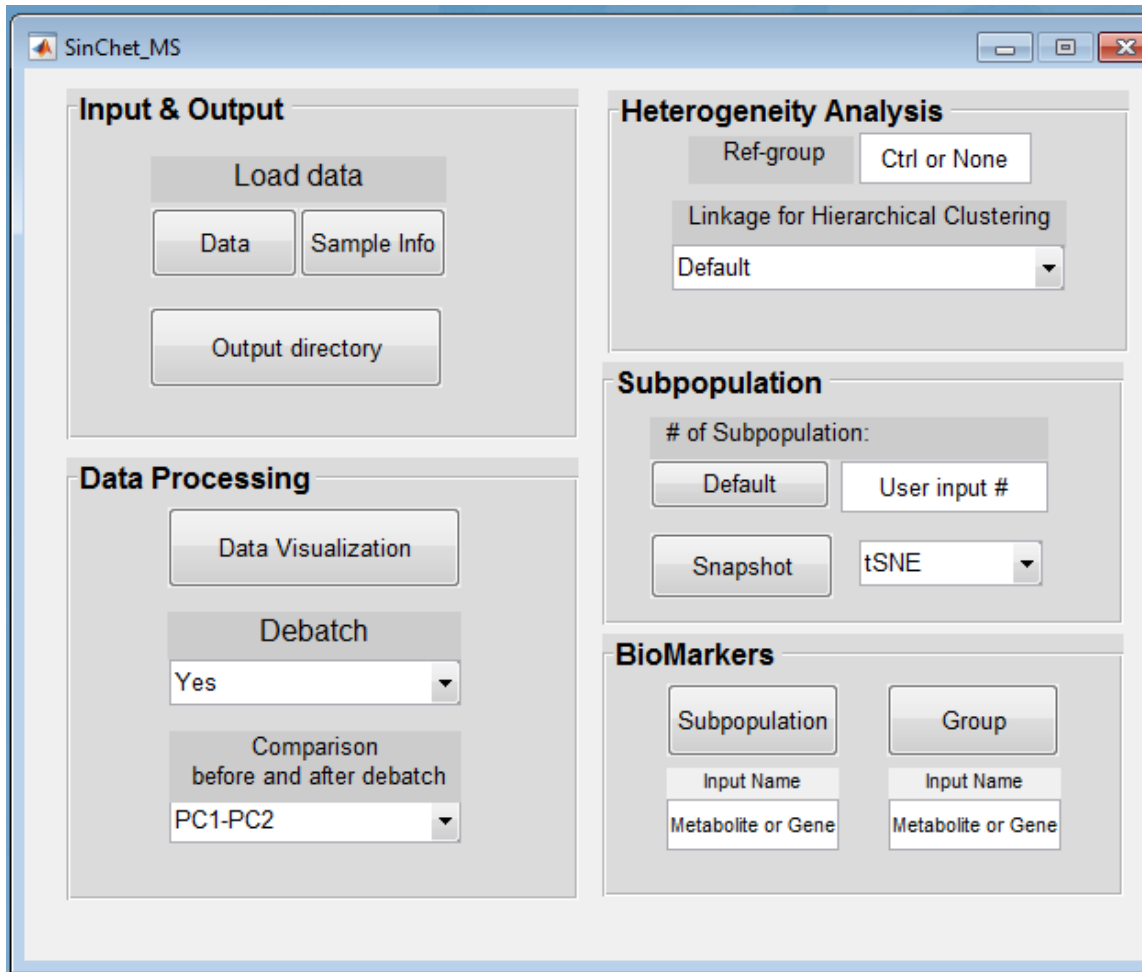
We provide a pre-compiled standalone version SinCHet-MS that can be used in PC without MATLAB.

- (1) Install [Microsoft Windows SDK7.1](#).
- (2) Download pre-compiled version: [SinCHet-MS Compiled.zip](#)
- (3) Unzip the downloaded file, and navigate to the folder "for\_redistribution". Double click the file MyAppInstaller\_mcr.exe to start the installation process of SinCHet-MS and MATLAB Compiler Runtime (MCR).
- (4) Double click the above installed file to start the software.

## **User Manual:**

### **1. Input & Output**

- (1) Using the pre-compiled PC versions above, the main Graphic User Interface of the SinCHet-MS will appear as the following figure.



- (2) Select “Data” and “Sample Info” pushbutton in "Load data" section and a window will display. You can choose the fold in your PC that contains the datasets to be analyzed. The ‘Data’ would be Excel file containing the header with group name at the first row; Metabolite name at the first column and numerical matrix (p x n) for which the p rows are metabolites, and n columns are single cell samples in the first spreadsheet. The ‘Sample Info’ would be contain the header with sample, treatment and batch at the first row; Treatment (column 2) and batch (column 3) should be numeric value in the first spreadsheet in excel file.
- (3) Select the "Output directory" pushbutton and a window will display. You can choose the fold in your PC where all analyzed results and plots will be stored.

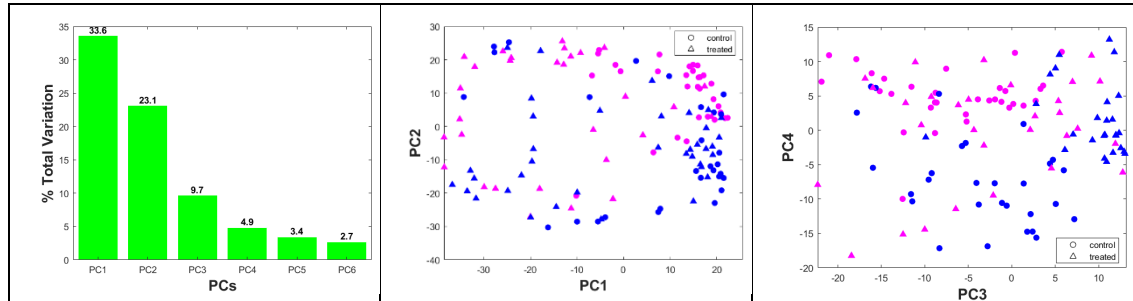
## 2. Data Processing

- (1) Select “Data Visualization” pushbutton to display PCA analysis results for the original dataset in a popup window. The left panel indicated the PC variation From PC1 to PC6. The middle panel and right panel indicated PC1 vs PC2

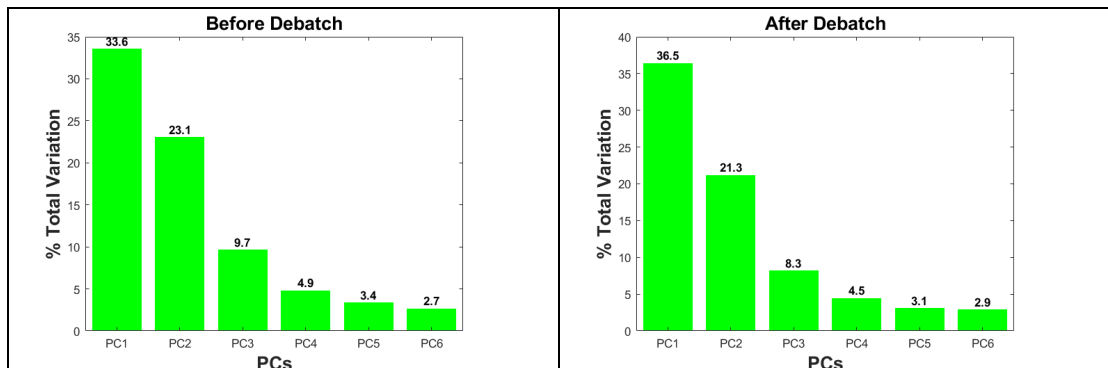
and PC3 vs PC4 scatterplot with colors for the different batches and marker symbols for the different treatments. Those scatterplot will help the user to determine whether or not to perform debatch for the original dataset. The following table lists the current setting for color and marker symbols related to batch and treatment respectively.

Table 1. Current color and marker symbols setting for batch and treatment

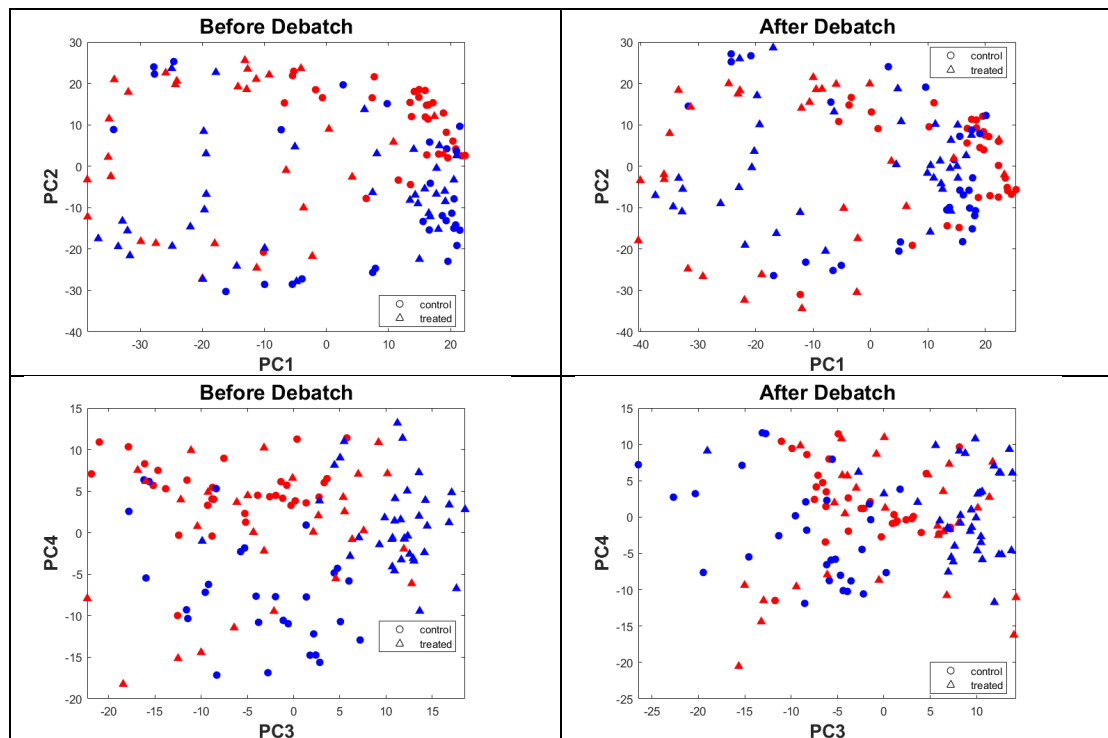
| Batch | Color   | Treatment | Marker symbols           |
|-------|---------|-----------|--------------------------|
| 1     | Red     | 1         | Circle                   |
| 2     | Blue    | 2         | Upward-pointing triangle |
| 3     | Magenta | 3         | Square                   |
| 4     | Cyan    | 4         | Hexagram                 |
| 5     | Green   | 5         | Diamond                  |
| 6     | Yellow  | 6         | Asterisk                 |



(2) Select “Debatch” dropdown menu to display either ‘Yes’ or ‘No’ in a popup window to perform debatch or not for the original dataset. If choose ‘No’, the further analysis will use the original data. If choose ‘Yes’, the toolbox will perform the debatch using Combat method (Fortin et al., 2018). After debatch, the PC variance barplots will be provided and compared between before and after debatch as the following figures. In addition, the debatched data will be output into the spreadsheet ‘data\_ComBat’ in the excel file named “Biomarkers\_Linkage”.

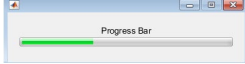


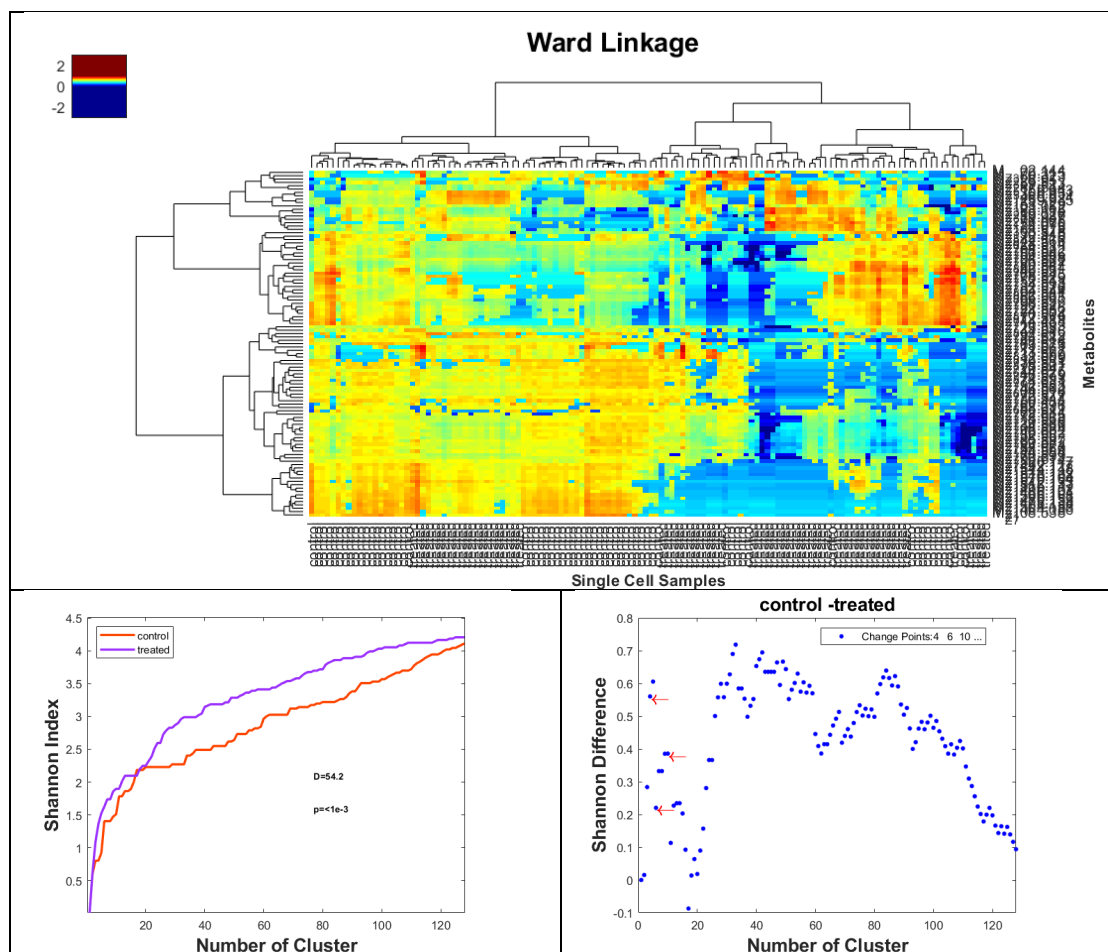
- (3) Select “Comparison before and after debatch” dropdown menu to display either ‘PC1-PC2’ or ‘PC3-PC4’ in a popup window to perform PCA analysis for both original data and debatched data. The toolbox provides PC1 vs PC2 or PC3 vs PC4 scatterplot comparison between before and after debatch with colors for the different batches and marker symbols for the different treatments same as the table 1 described.



### 3. Heterogeneity Analysis

- (1) Input the reference group name if you want to compare the heterogeneity difference between each group with the specified group as the reference (for instance, the example here is ‘Control’ as the reference group). Otherwise, input ‘none’ or any other strings, the toolbox will perform pairwise comparison of heterogeneity difference between groups.
- (2) Select either the default (ward linkage) or user-defined linkage method from the “Linkage for Hierarchical Clustering” dropdown menu for Profile of Shannon Difference (PSD) analysis as our previous publication (Li, Smalley, Schell, Smalley, & Chen, 2017), we performed the hierarchical cluster analyses using Euclidean distance and the selected linkage method (the upper panel hierarchical cluster heatmap). The Shannon

profile for each condition with D statistics and p value evaluated using 1000 permutation (the right panel in bottom) and the detected change points (the left panel in bottom) for PSD of each pair-wised comparison with or without reference group using MARS model will display in a popup window. (Note. This step takes a couple of minutes or longer since permutation step. Therefore the toolbox provides the progressing bar as the following ).

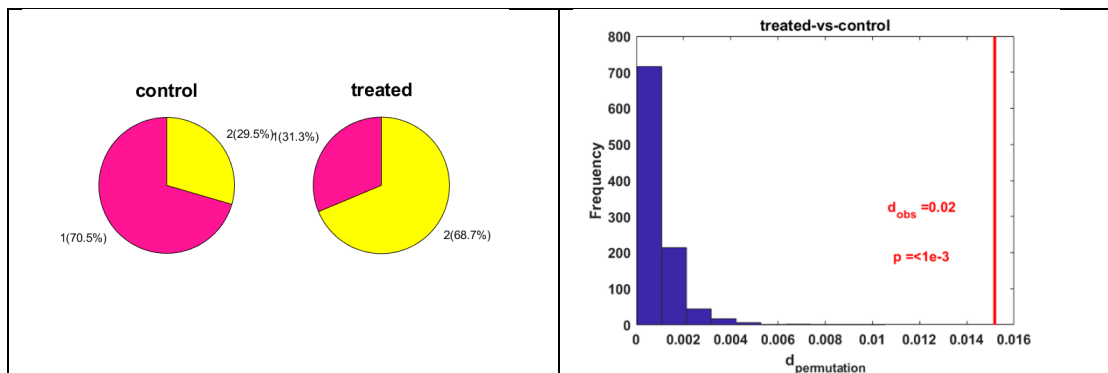


#### 4. Subpopulation

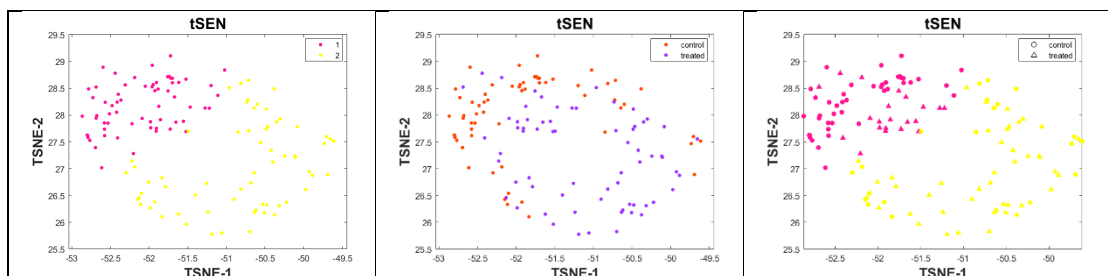
- (1) To determine the number of subpopulation, either select 'Default' or input user-defined the number of subpopulation. 'Default' is defined as 1) if there is no the significant difference of PSD ( $p >= 0.05$ ) among conditions, default value is the minimum value of the change points detected by MARS in all

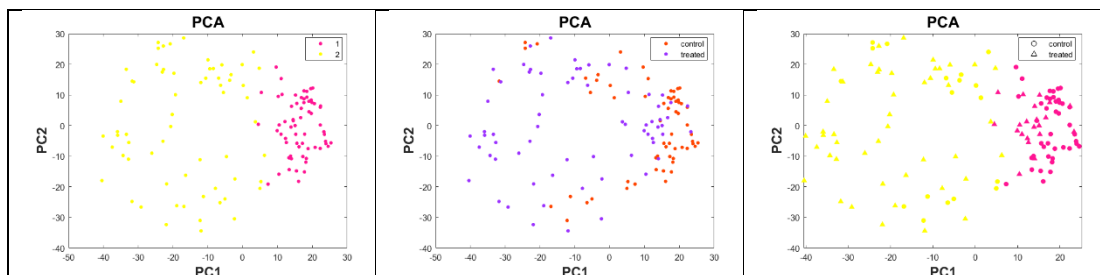
pair-wise comparison; 2) if there is the significant difference of PSD ( $p < 0.05$ ) among conditions, default value is the minimum resolution with the significant difference of  $d$  statistics which is defined as the difference of Shannon index at the certain resolution. The further analysis including snapshot or biomarkers for subpopulation will base on the number of subpopulation determined by default or user input.

- (2) Click on the “Snapshot” button to display subpopulation composition (the left panel) at the selected subpopulation resolution and histogram of the difference of Shannon Index ( $d$ , see the supplemental method as the following section) based on the pair-wised comparison with or without the reference group and  $p$  value was evaluated using permutation procedure ( $n = 1000$  current).



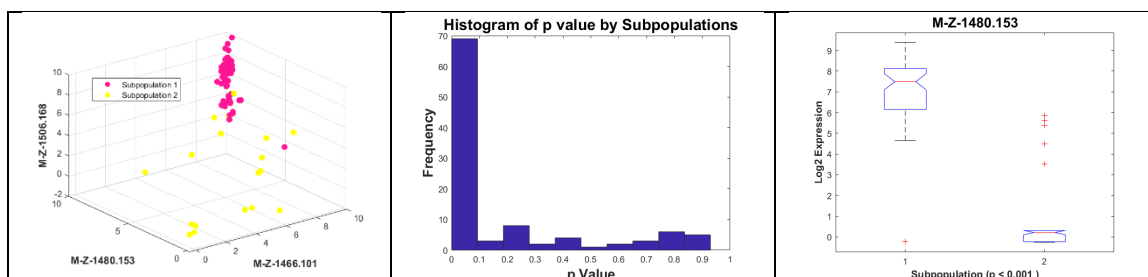
- (3) Select ‘tSNE’ or ‘PCA’ in dropdown menu to display tSNE (the upper panel) or PAC (the bottom panel) analysis results with subpopulation, groups or both labels respectively in a popup window. tSNE analysis was performed using MatLab default setting except the use PCA to reduce the initial dimensions to 50, 'Perplexity' as the sample size if the sample size less than 100, otherwise 100 which make the clusters are tighter than with the default setting and 'Exaggeration' as 20. In general, a larger exaggeration creates more empty space between embedded clusters and the MatLab default value for exaggeration is 1.5.





## 5. Biomarkers

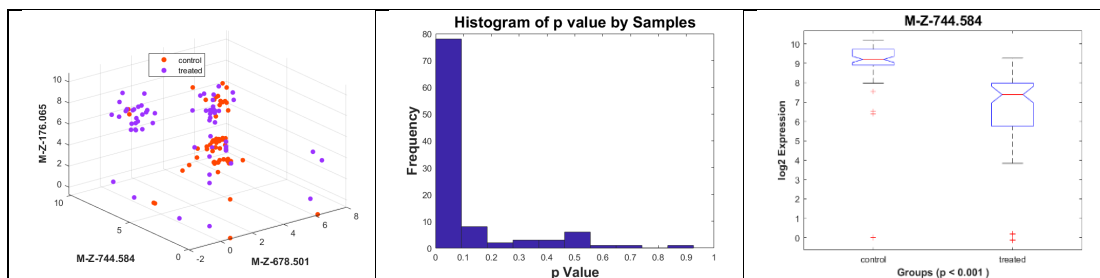
- (1) In the “Biomarkers” section, click the “Subpopulation” button to display the 3D scatter plot (the left panel) for the three metabolites with the top ranks of overall modified Generalized Fisher Product Score (mGF score, see the supplemental method as the following section) across the subpopulations and histogram of overall p value derived from Kruskal-Wallis test (the middle panel). Also output the biomarkers detail results in the excel file named “Biomarkers\_Linkage”. The spreadsheet of ‘Subpopulation-Marker’ in the excel file contains overall p value and pair-wised p values derived from Kruskal-Wallis test, mean expression of each subpopulation, fold change of pair-wise comparison between subpopulations, sum of all paired-wise absolute of fold change, Chisq Fisher product method, p value derived from Chisq Fisher Product Method, FDR estimated by Benjamini and Hochberg (BH) approach for p value derived from Chisq Fisher Product Method, overall GF score and individual subpopulation GF score. “Annotation” spreadsheet provided the each variables annotation in the previous spreadsheet. In addition, the “Sample\_Subpopulation” spreadsheet was output and it provides the information on each sample and subpopulation relationship. In addition, Input metabolite (or genes) name you are interested with either lower or upper case in the “Input Name” section enables to generate the boxplot of this metabolite expression across subpopulations with overall p value (the right panel).



- (2) In the “Biomarkers” section, click the “Group” button to display the 3D scatter plot for the three metabolites with the top ranks of overall mGF score across the groups, histogram of overall p, boxplot of the user-defined metabolites and the



biomarkers detail report with the pair-wise comparison across the groups in “Biomarkers\_Linkage” excel file similar to the previous described. In addition, the results with the GF score which summarized the overall difference between- and within-subpopulation in two groups comparison as our previous description(Li et al., 2017) was saved at “GF\_Score” spreadsheet in the excel file named “Biomarkers\_Linkage”.



Note: All plots generated by the above processing are automatically stored at the output directory with .jpg format. In addition, all plots can be manually saved as a different format such as .png, .tiff, etc. at the desired fold once the figures displayed.

## 6. Supplemental Methods

### 6.1 A novel d statistic to quantify cellular heterogeneity differences at the certain resolution between two conditions

A novel d statistic is developed to quantify cellular heterogeneity differences between two conditions and it is defined as the difference of the Shannon Index (SI) at the certain resolution from two conditions:

$$d = S|_2 - S|_1 \quad (S1)$$

, where  $SH_1$  denotes the SH of the first condition while the  $SH_2$  denotes the SH of the second condition at the certain resolution. Its statistical significance is estimated using a permutation procedure as our previous description (Li et al., 2017).

### 6.2 sGF score to prioritize the biomarkers

sGF is devised to the overall difference among cell subpopulations for each metabolite including p-values from multiple comparison tests and fold change of pairwise comparison among all between interested subpopulation vs any other subpopulation. In general, Fisher's method  $\chi^2$  is aggregation of evidence from each separate rank sum tests for each biomarker (e.g., metabolite).

$$X_i^2 = -2 \sum_j^s \ln(p_{ij}) \quad (\text{S2})$$

where  $p_{ij}$  is the p value derived from rank sum tests for the difference of the expression levels of each biomarker comparison between the interested  $i$ th subpopulation and the other  $j$ th subpopulation ( $j \neq i$ ),  $n$  is the total number of subpopulations. sGF score is incorporated Fisher's method with fold change (FC) for above comparisons.

$$sGF_i = X_i^2 + \sum_j^s |\ln(FC_{ij})| \quad (\text{S3})$$

$FC_{ij}$ , is the FC for biomarker  $i$  at  $j$ th subpopulation of totals of subpopulations pairwise comparison as described above.

#### Reference:

- Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., . . . Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, *167*, 104-120. doi:10.1016/j.neuroimage.2017.11.024
- Li, J., Smalley, I., Schell, M. J., Smalley, K. S. M., & Chen, Y. A. (2017). SinCHet: a MATLAB toolbox for single cell heterogeneity analysis in cancer. *Bioinformatics*, *33*(18), 2951-2953. doi:10.1093/bioinformatics/btx297