**s\_adhoc\_data\_processing\_master.m**

This script contains a modular mass spectrometry imaging (MSI) data analysis pipeline.

By Teresa Murta, June 2021

This script can be used to perform the following steps:

# DESI, MALSI, SIMS or REIMS data pre-processing (using SpectralAnalysis functions)

# Peak detection on a representative spectrum (using SpectralAnalysis functions)

# Matching of the peaks detected in the representative spectrum against HMDB and/or a group of lists of molecules of interest defined by the user

# Creating and saving a datacube (SpectralAnalysis DataRepresentation struct) for each imzml of interest (using SpectralAnalysis functions)

# Creating and saving a data matrix for each normalisation algorithm of interest

# Defining a new “dataset” by combining the original imzmls files. This steps involves the specification of group of original imzmls files that need to be combined, of which masks (define in 10) are to be used reduce each original imzml file to smaller group of pixels of interest, the geometric position of each small group of pixels of interest in the new “dataset” (its position in a 2D grid that will contain all the groups of pixels of interest from all the imzmls combined)

# Saving single ion images for:

## One or more lists of molecules of interest defined by the user

## One or more superclass, class, or subclass of molecules (as defined by HMDB)

## A lists of m/z values

# Running PCA, NMF, k-means, t-SNE, NN-SNE using:

## N most intense peaks detected in the representative spectrum

## Percentile P of all peaks detected in the representative spectrum

## One or more lists of molecules of interest defined by the user

## One or more superclass, class, or subclass of molecules (as defined by HMDB)

## A lists of m/z values

# Saving pictures (matlab figures and pngs) with the main outputs of PCA, NMF, k-means, t-SNE, NN-SNE (e.g.: principal components, representative spectra, single ion images of top drivers, cluster maps)

# Saving user defined masks for regions of interest (SpectralAnalysis RegionsOfInterest struct), by combining the results of k-means (intersected or united) with regions of the image manually defined by the user (using Matlab).

# Running the univariate analyses: t-test, ranksum test, and ROC analysis, which relate the (mean) ion intensities of user defined groups of regions of interest (defined in 10)

# Running ANOVAs to define groups of ions that relate to particular “conditions” such as acquisition date, glass slide number, sample ID, tissue type, etc. Each “condition” has to be defined as a combination of regions of interest (defined in 10)

# Discarding groups of ions defined using the ANOVA results (12) before running any of the multivariate analyses describe in 8.

# Saving the k-means, t-SNE or NN-t-SNE clustering maps as regions of interest (SpectralAnalysis RegionsOfInterest struct). These regions of interest can be used in any subsequent analyses together with or in place of those defined in 10.

# Saving the data from an original imzml or a new “dataset” (defined in 6) in a csv file, which contains the intensity of all pixels of interest, the “main mask” and “small mask” of each pixel, etc.

The requirements to run this script successfully are:

# The most recent version of SpectralAnalysis available at <https://github.com/AlanRace/SpectralAnalysis> added to the Matlab path.

* The most recent version of “adhoc-data-processing-pipeline” available at <https://github.com/NICE-MSI/adhoc-data-processing-pipeline> added to the Matlab path.
* The location of (i.e. the path to) the SpectralAnalysis pre-processing file (extension “.sap”) to be used. An example can be found in “required-files” within the git repository “adhoc-data-processing-pipeline” specified above. The parameters of the pre-processing need to be adequate to the data. The pre-processing file can be edited in Matlab.
* The location of (i.e. the path to) the imzML and ibd data files, which have to be saved in modality and polarity specific folders.
* An excel file named “inputs\_file” saved in the folder that contains the imzml and ibd data files. An example can be found in “required-files” within the git repository “adhoc-data-processing-pipeline” specified above. The inputs file needs to be adjusted to the particular requirements of the analysis, dataset, study goal, etc.

The outputs of running this script are:

# A spectral details folder containing:

## totalSpectrum\_intensities - total spectrum counts (per imzml and mask)

## totalSpectrum\_mzvalues - total spectrum mass channels (per imzml and mask)

## pixels\_num - number of pixels of interest (per imzml and mask)