Pre- and Peri-Processing Steps, Order, and Modifiable Parameters

Introduction

This document will walk users through the pre- and peri-processing workflows used in the MeOW.

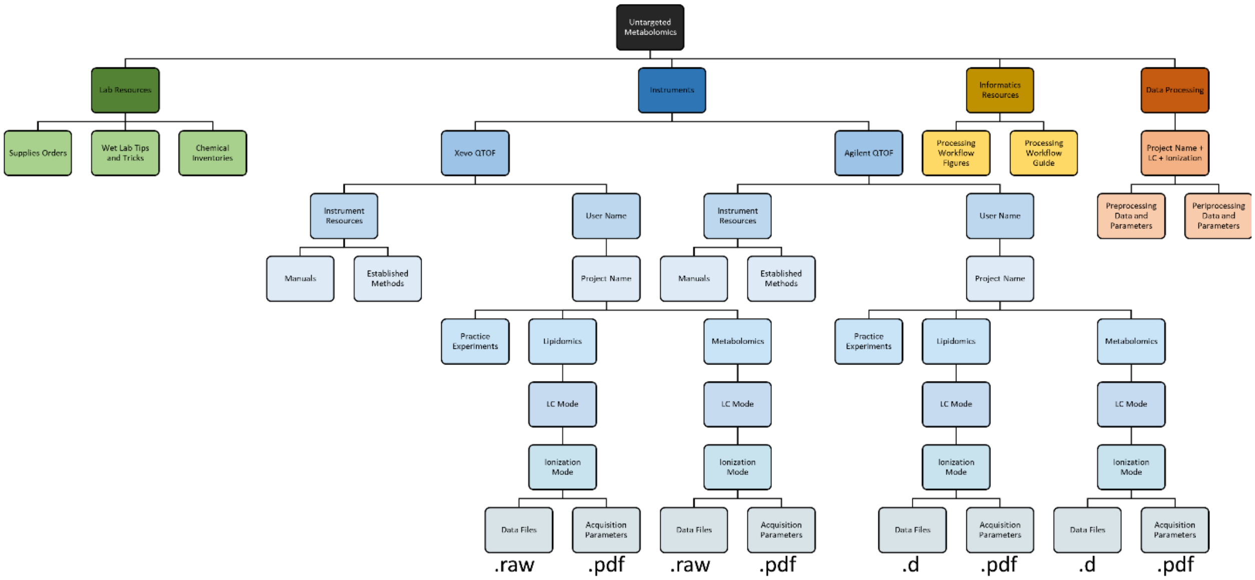
Pre-processing constitutes the conversion of raw spectral data files to a matrix containing metabolite features in rows and samples in columns, with each cell representing the peak intensity or area of a given feature for a given sample.

Peri-processing is the process of filtering, normalizing, and transforming the output matrix from pre-processing to prepare it for use in statistical analysis.

*Both pre- and peri-processing involve quality control checks and plots

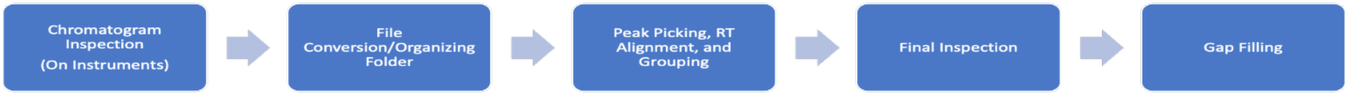
Input Data Expectations

- Folder organization for data collection/data archiving:



- File format:
  - File requirements: lockmass corrected, centroided
  - File types: .raw, .d
  - File pre-requisites: manually inspected for outlier samples, noise level, RT shift window, RT inclusion range

Pre-Processing


Chromatogram Inspection
(On Instruments)
File Conversion/Organizing Folder
Peak Picking, RT Alignment, and Grouping
Final Inspection
Gap Filling

Input/Output:

Input

- A folder with a list of raw filenames (from the instrument)
- Sample Meta-Data sheet (follow POUNCE sheet format)
- Conventions: XYZ

Output

- *Add file path under S:\UntargetedMetabolomics for outputs
- Excel Results File with four sheets:
  1) Original_Feature_Matrix: features as rows, samples in columns, and relative abundances output from the last step of the pre-processing workflow (gap filling).
  2) Bad_Samples: index, name, and notes *include in POUnCE metabolomics sheet
  3) Bad_Features: index, name (e.g. mz_rt), and notes *include in POUnCE metabolomics sheet
  4) Final_Feature_Matrix: features as rows, samples in columns, with features and samples from the Bad_Samples and Bad_Features sheets removed
- Log file of parameters used for pre-processing
- QC/Inspection files:
  - o JPEGs or PNGs of initial chromatogram inspection
  - o JPEGs or PNGs of final inspection (2 separate folders: Fin_Inspection_TIC Fin_Inspection_AllEIC, Fin_Inspection_BadEIC)
- RData file: excel results

Chromatogram inspection

- Inspect sample TICs individually for issues (e.g. break in signal, etc)
- Inspect all together, color-coded by sample types (pooled QC, blanks, samples, etc.) for outliers
- Determine parameters for pre-processing: RT range to keep for peak picking, noise level, RT drift range

File Conversion/Organizing Folders

- Convert to mzML

- Organize files to ensure that 'centerSample' for RT alignment is and select 'center' sample to be used for RT Obiwarp alignment algorithm

Peak Picking and RT alignment (can be iterative)

- Option 1 (Priority): Perform peak picking/alignment on all samples
- Option 2 (other project, needs to be implemented, not a current priority): Perform peak picking/alignment on QC only, then do a targeted search for resulting features on remaining samples
- Modifiable parameters for the config file:
  - Retention time range (after solvent front, before column cleaning)
  - Peak width
  - ppm
  - Noise
  - Prefilter
  - Signal to noise
  - Amount of seconds to merge neighboring peaks (shoulders)
  - Obiwarp bin size for RT correction
  - List of standards and m/z
- Steps to monitor data from standards:
  - Creating chromatogram given a list of m/z and RT range

Final Inspection

- Re-inspect the chromatograms before/after correction
- Inspect known feature EIC before/after correction (e.g. those with standards) + 20 random ones.
  - Option to output all feature EIC (medium res JPEG, use index of features as filenames, one file per name)
- Optional: Notes on 'status' (e.g. removal, good, etc.) for features and samples will be input into QC Excel file. Bad EICs JPEGs dropped in the Bad_Features sheet will automatically be detected for creating the output file (Final_Feature_Matrix).

Gap filling

- Pay attention to the number of gaps filled (should be relatively small)
- Plot before gap filling (x-axis) vs. after gap filling (y-axis), each point is the maximum abundance (log or asinh) of all samples per feature (number of points = number of features)

*Note: All filtering is performed at the peri-processing steps

Peri-Processing

Input/Output

Input

- CSV of grouped and aligned metabolite data from pre-processing with metabolites in rows and samples in columns
- Sample meta-data including *at least*
  - sample name
  - sample type (blank, QC, sample, etc.)
  - external scalar normalization values (weight, volume, creatinine, etc.)
  - batch
  - order of injection

- Metabolite meta-data which at the minimum, would include unique feature ID/name. Could also include identification, adduct information, pathway information, etc.
  - Ex) MeOW pre-processing feature name: mode_m/z_RT_LC
- Input data can include all LC methods and modes, as long as meta-data referring to which LC method and mode a given feature belongs to is included
  - *Note: Peri-process one LC method and ionization mode at a time
- File naming convention should match POUnCE convention

Output

- POUNCE Excel sheet ("MetaboliteStatReadyAbundances") of filtered, normalized, transformed, and scaled metabolite abundances, with features in rows and samples in columns
- Sample meta-data from input (including demarcation of what is filtered in "SampleMeta" sheet)
- Metabolite meta-data from input (including demarcation of what is filtered in "MetabMeta")
- List of parameters used and numbers of metabolites and samples filtered at each step
  - Output as log file (text file created using log4r)
- RData file with metabolite abundance matrix, sample meta-data, and metabolite meta-data
- HTML report of code, Table1, and interactive QC evaluation plots (potentially via knitr)
  - See here for exporting Table1 HTML into a Word doc for publication

Read in data

Format data

- Use .pos or .neg suffixes to keep track of ionization modes processed
- Row names as unique feature names
  - Need to distinguish neutral from ionized masses, if applicable
  - Feature names must match those listed in metabolite metadata from "MetabMeta" sheet
  - Additional metabolite meta-information about features*:
    - m/z or neutral mass (depending on preprocessing output)
    - RT
    - LC method
    - Ionization mode
    - *see MetabMap in POUNCE metabolomics input for how to map these variables
  - Ex) MeOW data: include m/z, RT, LC, and mode (e.g., 205.0977_7.6_HILIC_POS)
- Ensure order of metabolite matrix, metabolite, and sample meta-data data frames is the same
- Ensure metabolites are in rows and samples in columns

Generate basic statistics and QC metrics:

- Produce “"table 1” using the R package ‘table1’ using the sample meta-data file.
- Evaluate feature abundances:
  - If # samples < 30, produce boxplots per sample
  - If # samples >=30, produce a histogram of median abundances (after taking the log or asinh transformation) per sample and a separate histogram of SD or CV of median abundances (if CV, do not take the log).

- Draw a histogram of the number of missing values per sample (use enough breaks to see the granularity)
- Evaluate intensity drifts over time:
    - Plot TIC per sample (ordered).
    - Plot TIC per QC only (ordered)
    - Inspect for any outlier TIC values
        - Ex) flag samples and/or QCs with TICs less that 80% of the mean TIC and use raw data to determine whether they should be removed from the dataset
- Evaluate technical replicate samples:
    1. Calculate correlations between replicate samples (e.g. duplicates, pooled QCs, etc.)
    2. Calculate correlations between random pairs of samples
    3. Draw histograms of the correlation vectors results from 1 and 2, above
- Evaluate consistency of spiked in standards:
    - Create histogram of standard intensity vs. order for each standard
- Evaluate features that are identified and known to be different between groups
    - User input: groups with expected differences, metabolite feature IDs
    - e.g. for smoking metabolites, check boxplot distributions of known smoking-related metabolites between smokers and nonsmokers
- Evaluate technical QC and process QC (if available) variability cutoffs:
    1. Plot the SD of Technical QC vs Process QCs where each dot represents one feature
    2. Determine appropriate threshold level for each, based on the plot
    3. Only retain features with BOTH low technical and low process variability (later step)
- Evaluate variance (CV) of features by sample type
    - Create boxplots of CV for blanks, samples, and QCs
    - QC CV should be lowest
- Produce a PCA, after log transformation, and globally check that there are no major outliers and no major experimental effects (e.g. batch or run order effect)
    - Color PCAs by: sample type (QC vs samples vs blanks), batch (if available), external scalar (if available, ex: weight, osmolarity, etc.), order (color on a scale from light to dark)

Peri-Process data

Filter out unreliable features (features that are likely to be artifacts)

Filter by total missing values:

- Missing values per sample: create histogram, set cutoff, check raw data of samples that exceed cutoff and remove if appropriate (default: 75%)
- Missing values per feature: create histogram, set cutoff, remove features that exceed cutoff
- Report data dimensions before and after each filtering step, as well as cutoff values

Filter by processed blanks:

- Blanks used for filtering should only be those run before any samples
- Remove features that exceed cutoff for presence in blanks

  - Ex) if the mean intensity of a feature in the blanks is more than ½ the mean intensity of the feature in QCs and samples, remove it
- Do not include processed blanks from metabolite and meta-data in subsequent steps after filtering

Filter by (technical) pooled QC:

- User defines which QCs should be used for this step
- Retain features that are present in >80% of QCs
- Remove features that have CV > 30% in QCs (reflects technical variation, CVs should be calculated on non-log transformed data)
  a. Plot distribution of CVs for all QCs and add line where 30% CV cutoff is

Impute missing values

- If missingness is due to low detection limit: impute by ½ minimum
- If missingness is random, then impute by KNN, random forest
- Density plot or boxplot

Normalize

- Options:
  - For urine: evaluate osmolality, creatinine, and MSTUS
  - For blood: evaluate MSTUS, total protein, TIC
  - For cells: evaluate MSTUS, total protein, TIC
  - For tissues: evaluate weight, MSTUS, TIC
- Density plot or boxplot or PCA (with scale and center = TRUE)

Transform

- Options: log2, log10, glog, asinh
  i. Default: log2
  ii. Evaluate asinh
- Density plot or boxplot

QC drift/batch correction

- *Only perform when major batch effects are observed in TICs and/or PCA
- Options: van der Kloet, LOESS, QC-RSC
  i. Default: QC-RSC
- Density plot or boxplot or PCA (with scale and center = TRUE)

Scale and center

i. Options: auto, pareto, range
ii. Density plot or boxplot

Final Examination

- Reproduce plots from step 3, placing the before/after normalization plots next to each other