Metabolomics Analysis Pipeline

Overview

The Metabolomics Analysis Pipeline was designed using R packages to support the downstream analysis of semi-quantitative mass spectrometry (MS) data output such as that from Metabolon Company. The pipeline filters and analyzes the data to identify metabolites/analytes that are significantly differentially expressed between user identified experimental groups.

Note: This pipeline can be used for analysis of any semi-quantitative mass spectrometry data, so long as the data has been processed upstream so that the peaks have been annotated.

Contents

I. Format of Input Data

This section provides information on the format of data files required to successfully run the pipeline.

1. ***Raw Data*** File

    - This is NOT the raw mass spectrometry data file. This file should be the output from analysis software that has translated the mass spectra data into analytes with abundance levels.
    - Format:
        - Comma-separated value (.csv) file
        - Column headers should be peptide identifications (i.e. metabolites or analytes); only ONE analyte per row mappable to KEGG or Uniprot ID.
        - Row headers should be *Sample Names* and column headers should be the Annotated Analytes.

          In the example below, *5096_0* is the name of the sample and *2924450* is the value for analyte *1_2-propanediol*.

	1_2-propanediol	1_5-anhydroglucitol	1-arachidonoylglycero phosphocholine*
5096_0	29244650	345296.3	713389.3

2. ***Sample Groupings*** File

    - This file provides a breakdown of experimental information i.e. sample groups for comparison.
    - Format
        i. Tab-separated value (.tsv) file
        ii. Three data columns in the following order (illustrated below): 1) *Sample* Name (the sample names listed here should match those in the ***Raw Data*** File), 2) *Group,* and 3) *Order* to display groupings with controls set at 0.

Sample	Group	Order
5096_T0	Initial	0
5096_T42	Late	2

II. Uploading Input Data

Data files must be uploaded to the **Analysis Pipeline** portal, prior to selecting the pipeline. Note that files on the SysBioCube cannot be accessed through this window, instead, a user must download the files from the SysBioCube onto their computer or local drive, then upload them through the **Analysis Pipeline** portal.

To upload files, when on the ***Upload Files*** tab, click the ***Add Files*** button (Figure 1, circled in red below) then navigate through the pop-up windows and select the appropriate files.

After the files have been uploaded through the portal, ***Load*** the **Metabolomics** pipeline (Figure 2). The data files uploaded will appear under the ***User Files*** header for ***Available Input Files*** (See red arrow in the figure below)**,** and can be uploaded to the pipeline for analysis through a drag/drop mechanism. Although optional, it is recommended that users submit a descriptive ***Analysis Name*** for recall at a later date. If no ***Analysis Name*** is provided by the user, the system will assign the analysis a random 8-digit identifier. To start the pipeline, select ***Run***.

Once a job is initiated, a window (Figure 3) will appear providing additional instructions. To proceed, click **OK**.

The window (Figure 4) that follows will illustrate the status of the job, defining *Submission*, initiation (*Started*) and *Completion* time and dates. The status bars will change from uncolored to green as each step is completed.

Once completed, each individual output data file will appear as a downloadable hyperlink below the window (Figure 5).

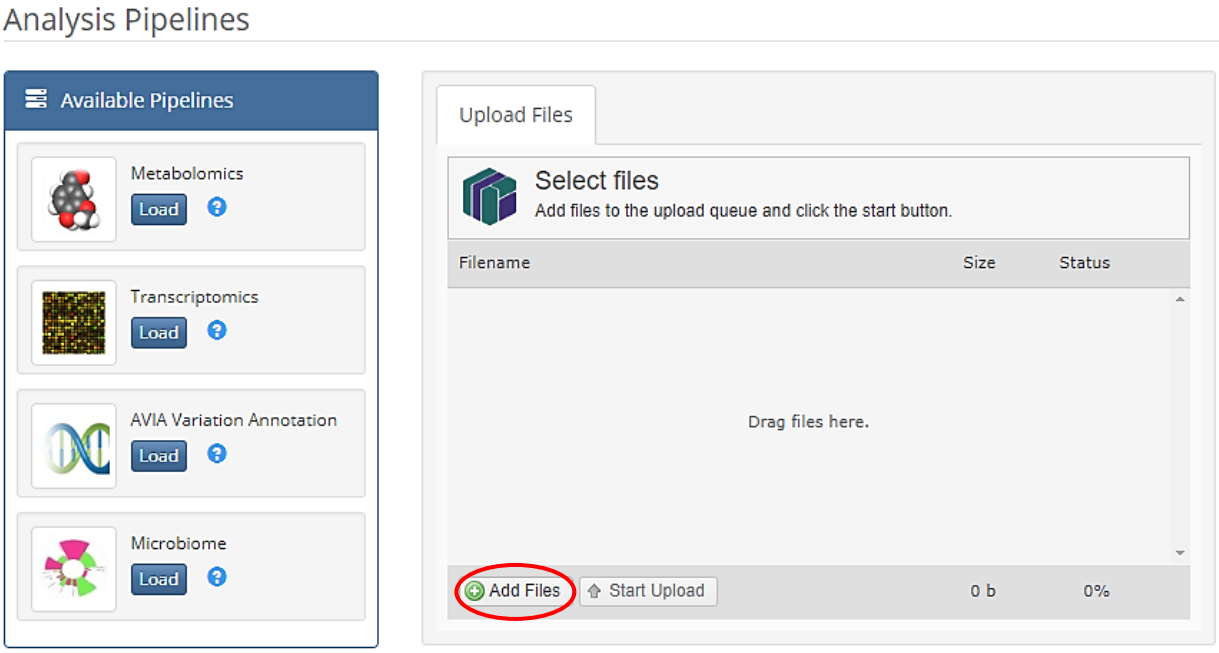

Figure 1. Default Homepage of Analysis Pipelines Tool

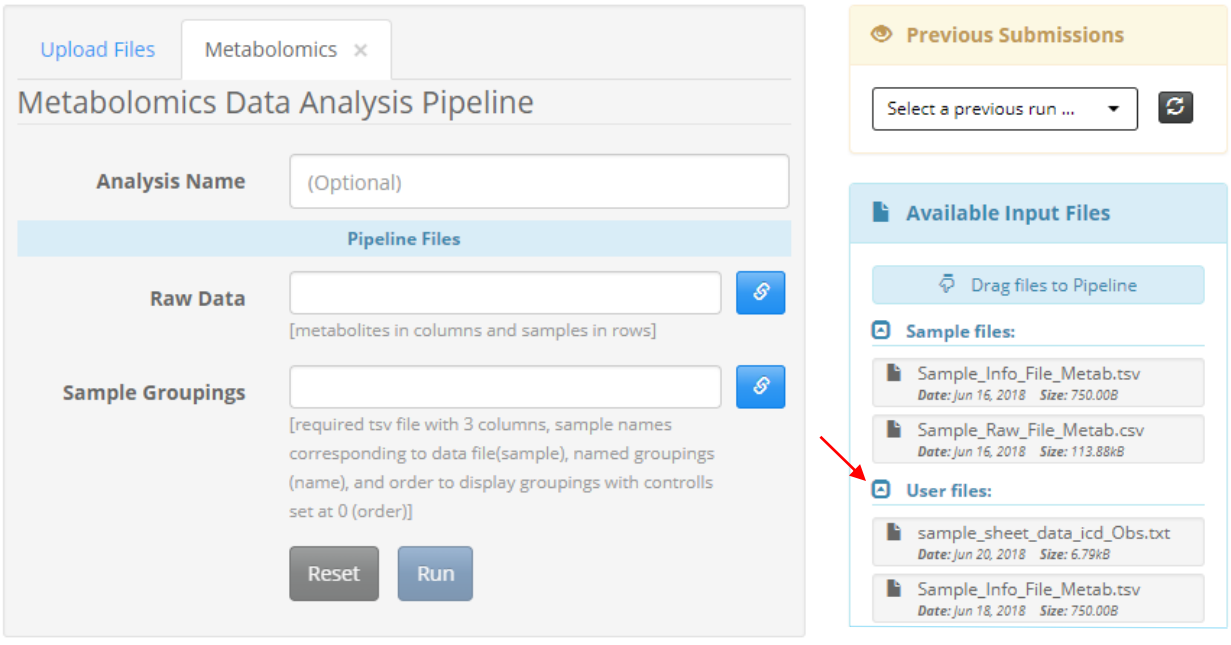

Figure 2. Tool View after Loading Metabolomics Pipeline

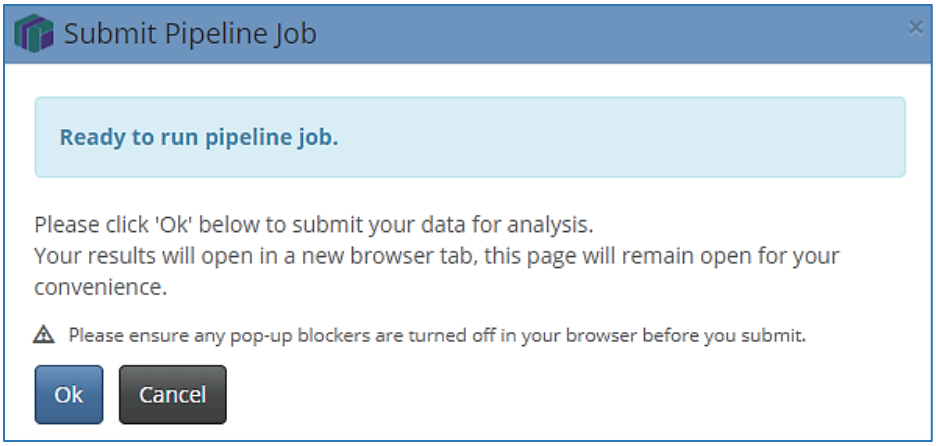

Figure 3. Pop Up Window Generated when User 'Runs' a Pipeline

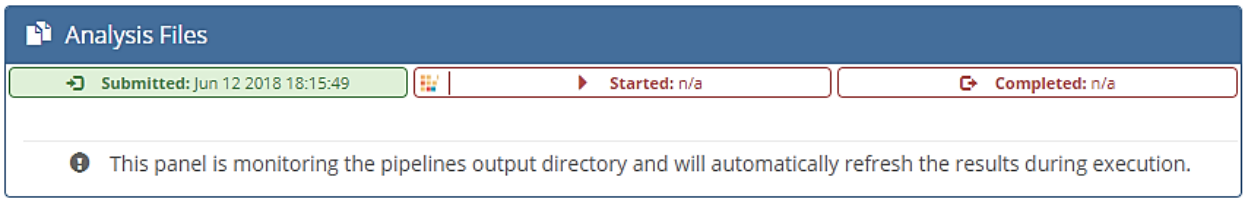

Figure 4. Real-time Status of the most recently Submitted Analysis Run

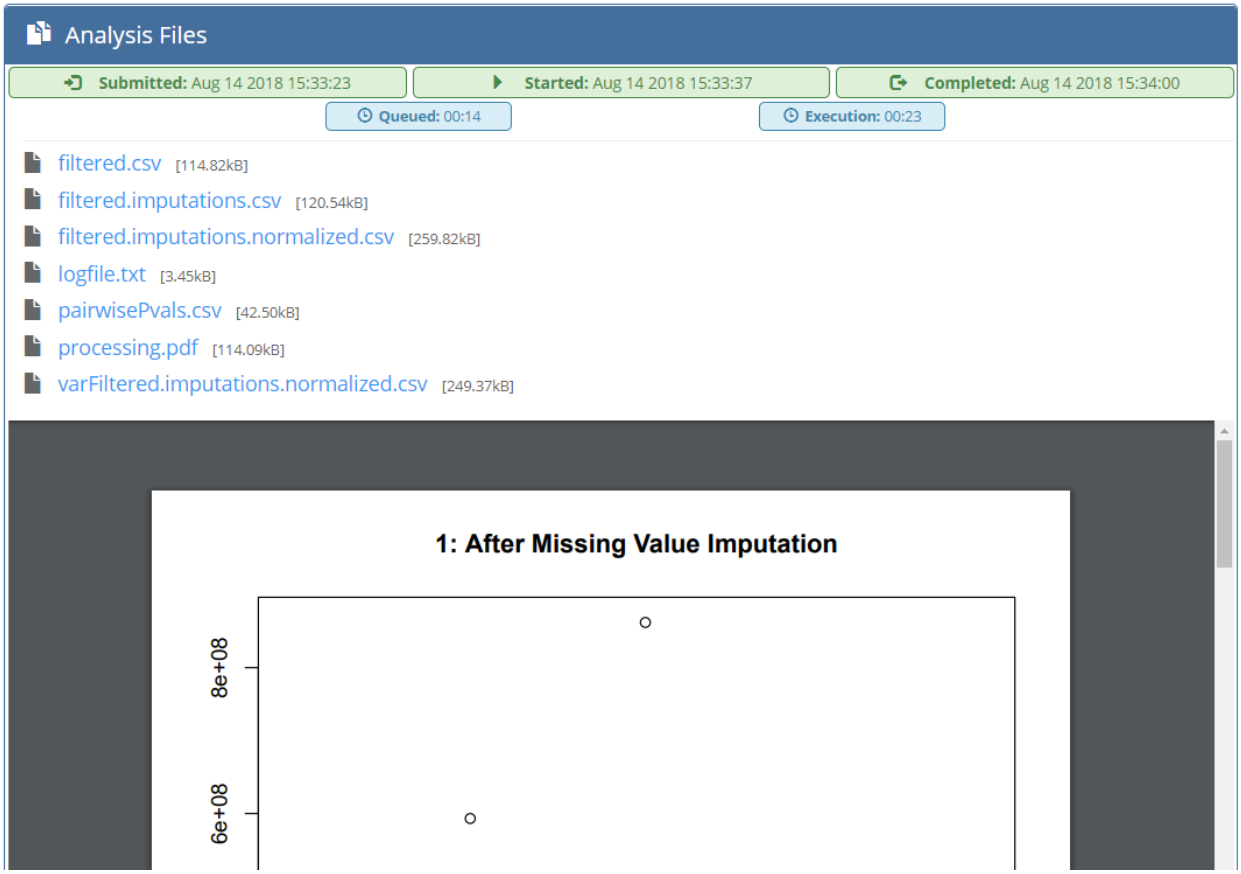

Figure 5. Display of Results from a Completed Analysis Run

III. Retrieving Results

A user can retrieve output data from a current run by clicking on the downloadable hyperlinks (Figure 4) that appear when the analysis is completed.

Data can also be retrieved from previous runs by selecting from the dropdown menu under ***Previous Submissions***, which appears on the right side of the **Analysis Pipeline** portal (Figure 2). The dropdown menu is populated with the ***Analysis Names***.

IV. Overview of Results

The results produced from this pipeline include filtered and normalized versions of the ***Raw Data*** file. The following provide descriptions of what changes were executed on the file, organized based on the file names and/or figure titles.

1. **Filtered.csv** –
    1) Missing Data: Those *Sample*s without a conditional match between *Groups* are removed.
    2) Presence Cut Off: Analytes which are not present in at least 75% of the samples are removed.
    3) Conditional Cut Off: Analytes which are not present in at least 2 samples for each experimental condition are removed. *Datasets are required to have, at a minimum, 2 samples per experimental condition in order to do the statistical analyses that is in the later result files.
2. **Filtered.Imputations.csv** – The data in this file has been both filtered, as described above, and the remaining missing values were replaced using the minimum observed value for each particular analyte from the remaining values.

    Figure #1 in the **Processing.pdf** output file is a scatter plot illustrating the distribution of the data values for each sample post filtering and imputation.
3. **Filtered.Imputations.Normalized.csv** – The data in this file has been filtered and imputed, as described above, and then autoscaled i.e. normalized. For a detailed description of the autoscaling, you can reference Berg et al. (2006)[1]. The following descriptions provide brief overviews:
    1) Row mean scaled: Each value in a row (sample) divided by the sum of that row.
    2) Column mean centered: Each value in a column (analyte) has the mean of that column subtracted from it.
    3) Column SD scaled: Each value in a column (analyte) is divided by the standard deviation of that column.

    Figures #2-3 in the **Processing.pdf** output file are scatter plots illustrating the distribution of the data values for each sample post filtering, imputation and row

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/

and column normalization, respectively.

4. **varFiltered.Imputations.Normalized.csv** – The data in this file has been filtered, imputed and normalized as described in #1-3. The data is then filtered using interquartile range (IQR) and those with IQR of <0.5, indicating the difference between 75th and 25th quartiles is less than 2 fold, were removed.

   The figure on page 4 in the **Processing.pdf** output file is illustrating IQR values with the cutoff. Figure #4 (page 5) is a scatter plot illustrating the distribution of the data values for each sample post filtering, imputation, normalization and variance filtering.

The remaining results files are the actual outcome of the analyses.

5. **PairwisePvals.csv** – The data in this file includes the P-values for a non-moderated t-test comparing each metabolite between conditions; the conditions compared are specified in the column headers. Each comparison is done in duplicate with and without FDR correction; FDR correction is specified in the column titles.
6. **Processing** – This PDF file offers visualization of the data as it was normalized then analyzed.
   1) Figures #1-5 (pages 1-5) are described above in #1-4.
   2) The figure on page 6 is a principle component analysis (PCA) of a group wise comparison across all analytes. This analysis illustrates the overall separation in groups after a linear transformation of the multivariate data.
   3) The figure on page 7 is a heat map illustrating the relative signal level of each analyte across a sample. The samples are labeled on the bottom of the figure; the analytes would be too cluttered for labeling on the right side, vertical axis. The sample *Group* is illustrated at the top of the heat map using a multi-colored horizontal bar (Red infers low levels; White infers high levels).
   4) The figure on page 8 is a Cluster Dendrogram; the height of the arrow, correlates to the distance (dissimilarity) between the samples based on analyte levels.
   5) The figure on page 9 is a Sammon map. Similar to a PCA, this plot is used to transform multidimensional data but, unlike PCA, the transformation is non-linear.
   6) The figures on pages 10-11 are volcano plots. The log2 of fold change is plotted along the x-axis and –log10 of the p-value is plotted on the y-axis. These scatter plots help to illustrate when analytes are exhibiting high magnitude fold changes and high statistical significance. The horizontal line separates those with p-values <0.05 from those >0.05 and the vertical lines mark fold changes >3-fold (positive or negative).