A Streamlined Workflow for Untargeted Metabolomics

Employing XCMS^{plus}, a Simultaneous Data Processing and Metabolite Identification Software Package for Rapid Untargeted Metabolite Screening

Baljit K. Ubhi[1], H. Paul Benton[2], Duane Rinehart[2] and Gary Siuzdak[2]
[1]SCIEX, Redwood City, CA, [2]Scripps Center for Metabolomics and Mass Spectrometry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, United States.

The conventional mass spectrometry metabolomics workflow tends to be a two-step process. Data is collected at MS1 and then processed to find any differential features of interest. Data is then re-acquired and any MS/MS information collected is used for metabolite identification and manual database searching. XCMS software is the "World's most cited metabolomics software" as it has over 1000+ citations in literature and very trusted in the metabolomics community.

A streamlined workflow using the TripleTOF® System and XCMS^{plus} Software combines both MS and MS/MS data collection into a single-injection workflow, and provides simultaneous data processing and metabolite identification. This brings the overall process down from ***weeks to days***, allowing allows for a more efficient and streamlined approach to move the researcher from sample to biology in a timely fashion.

XCMS^{plus} provides a host of features including multivariate statistical analyses and global visualization approaches such as the interactive cloud plot. With improved multi-group analysis capabilities, faster on-site data processing and unlimited storage and customization to their local environment, metabolomic researchers will be able to accelerate their discovery workflows and shorten time translating data into biological information. Data

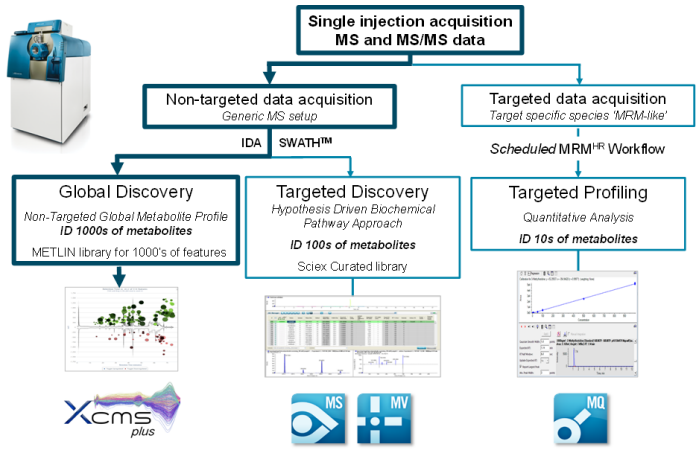

Figure 1. Untargeted Metabolite Workflow using the TripleTOF® System. Collect data using an untargeted approach and process the data using the XCMS^{plus} software. XCMS^{plus} is optimized for untargeted metabolite screening. By combining raw data processing and retention time correction with statistical analysis, the software identifies and quantifies endogenous metabolites that vary between samples.

from a previously acquired Zucker rat study (on the TripleTOF® system) was processed as proof of concept, through this streamlined workflow using XCMS^{plus} software.

Key Features of Untargeted Metabolite Screening using XCMS^{plus} Software

- Fast acquisition of high resolution, accurate mass MS and MS/MS data on the TripleTOF® Systems enables a single injection data acquisition strategy
- Data can be loaded, processed and reviewed in a single interactive workspace
- Streamlined and simplified data extraction parameters
- Submit and monitor the status of multiple jobs simultaneously
- Statistical analysis including paired/unpaired t-tests, parametric and non-parametric testing (including FDR), and multivariate data analysis techniques.
- Simplified metabolite identification by linking your data to a composite database
- Unlimited data storage capabilities

Material and Methods

Data Generation: Sample preparation, LC conditions and data acquisition parameters have been previously described for the Zucker Rat project[3].

Data Processing: Data was processed in XCMS^{plus} Software in both pair-wise and multi-group job mode. The data files are automatically loaded and read by XCMS^{plus} Software and converted for peak finding and RT correction/alignment. A ***pairwise*** job allows the comparison of two groups whereas ***multigroup*** job allows 3 or more groups of samples to be analyzed. Any MS data was compared versus the composite database for metabolite identification, MS/MS data were used for confirmation online available METLIN database[2].

Workflow Basics

The TripleTOF® System is a highly flexible and powerful MS system that can be used for both untargeted and targeted metabolomics as outlined in Figure 1. The workflow for untargeted metabolite screening using XCMS^{plus} Software will be the focus of this work.

XCMS^{plus} Software directly processes SCIEX LC/MS and LC/MS/MS *.wiff data files for feature detection and comparison. With the ability to process multiple jobs all at once, the software provides a single interactive workspace for monitoring job status and for reviewing results. Data can be processed from multiple vendors including Waters, Agilent, Thermo, LECO and Bruker.

Data can be processed in a number of ways which are referred to as “jobs”. A ***pairwise*** job allows the comparison of two groups and such paired and un-paired test analysis can be conducted. A ***multigroup*** job allows 3 or more groups of samples to be analyzed. As the number of groups is different so is the availability of statistical tests, here a user cannot conduct a one-way test. Both approaches allow principal component analysis (PCA) which is a multivariate analysis technique used to visualize data from multiple groups of samples and multiple variables (or features) in n-dimensional space. A t-test compares one variable at a time in two groups, i.e. tyrosine in the control sample versus tyrosine in the diseased sample.

Efficiency of Data Processing

Forty IDA data files from a TripleTOF 5600+ system were loaded into the XCMS^{plus} Software as well as uploaded to XCMS Online for comparison. The data files took around 24 hours to upload and process using the online version – this combines data upload which for these amount of data (~3.23GB) took 4-5 hours, then processing this data took a further 16-18 hours. The same 40 IDA data files were loaded and processed in XCMS^{plus} Software and data loading took around one minute and processing took around 50-60 minutes. Therefore a comparison between the two approaches is as follows:

- Processing time for XCMS Online: ~24 hours
- Processing time for XCMS^{plus}: ~1 hour

For cases where results are reviewed and data needs to be re-processed then using the Online approach means allowing another 24 hours. With XCMS^{plus}, this time is far reduced where a user can review reprocessed results just after an hour.

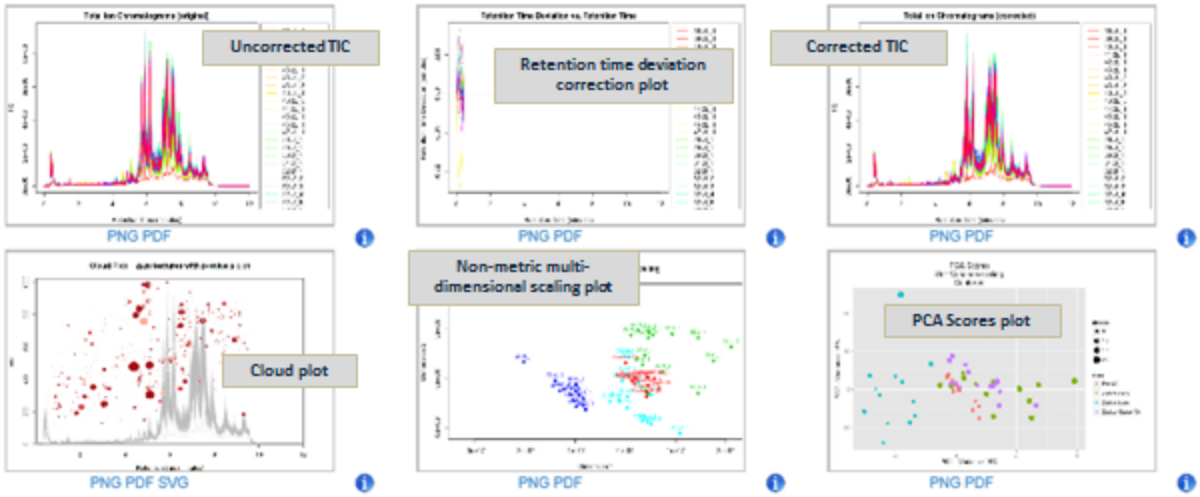

Figure 2. XCMS^{plus} Software - Interactive Workspace. View your data both pre- and post-alignment/correction (uncorrected TIC/corrected TIC) and where any of the alignment /correction was applied (retention time deviation correction plot).

Interactive Workspace with Powerful Statistics

XCMS[plus] outputs a results panel with an interactive workspace (Figure 2) where each extracted ion chromatogram can be reviewed for each significant feature. There is also an array of multivariate reports, from principal component analysis to an interactive heat map with hierarchal cluster analysis. The principal component analysis allows users to define which scaling method is best for the data and see both scores and loadings plots with metabolite annotations. In the interactive heat map, both extracted ion chromatograms (XICs) and spectral plots are given for each feature. This allows for fast, efficient interaction with the data and understanding of the experiment.

Interrogating the LC/MS data from the Zucker rat study, we could see that the different phenotypes clustered together in the PCA Scores plot (Figure 3). The different phenotypes (fatty, lean and obese) are colored to show the sample groups.

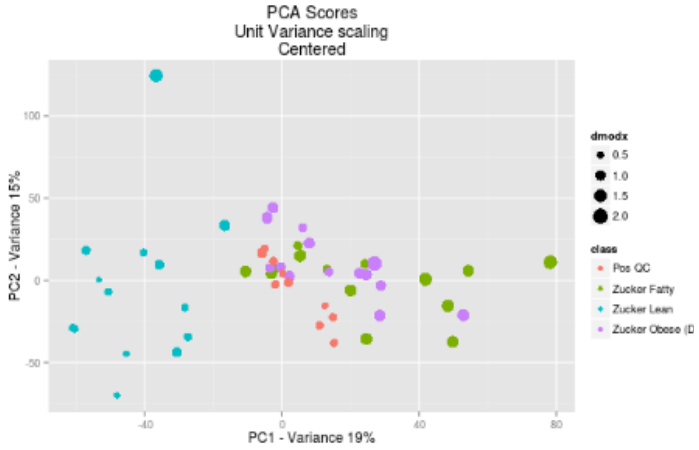

Figure 3. Principal Component Analysis (PCA) Scores Plot from XCMS[plus] Results. Three phenotypes of rat were analyzed from the Zucker rat study, including a pooled sample (QC). Here the groups of samples can be seen to cluster together on the Scores plot based on that phenotype. In this case, a non-discriminant analysis was performed, meaning no prior knowledge of the groups was used for this visualization.

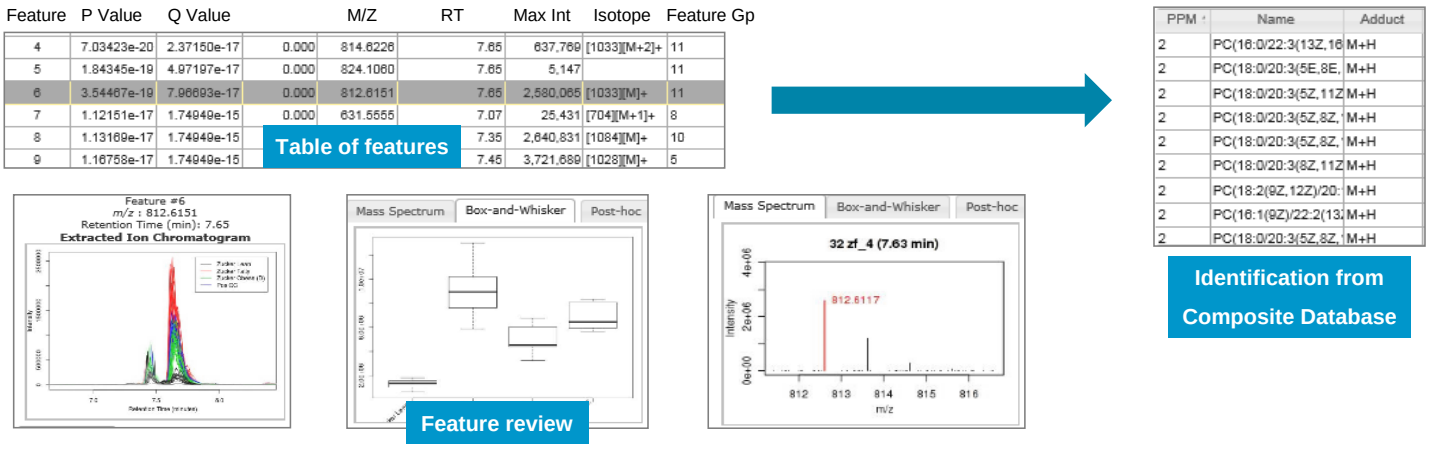

Figure 4. View Results Table Page. This page includes the feature table (top) which outlines all features picked in the peaking finding step during processing. Features can be ordered by any column such as p-value or q-value, etc. A feature can be highlighted and the XIC information across samples, the area differences between samples (Box and Whisker plot) and the m/z information can be viewed (middle). Finally, identification information (bottom) is shown.

Working with the Results Table

From the scores plot, the user can “view results table”. The table can be sorted based on p-value significance or q-value significance (Figure 4, top). The q-value is the false discovery rate adjusted p-value. The feature table identifies the retention time at which a particular feature was found as well as the intensity. Other observations (as columns) of interest can be added by the user and used for sorting (such as m/z value or retention time (RT), Corr Var = correlation variation, Max Int = maximum intensity, Feature Gp = feature group). Any isotopes and adducts are also identified as well as feature grouping which groups together any related ions.

From the feature table, a feature can be highlighted and the XIC can be visualized as an overlay from all the samples processed.

One can then review the raw data and see if this is an actual feature or just noise or an irrelevant peak eluting in the solvent front. Retention time information as well as the accurate mass is displayed in this XIC plot (Figure 4, middle). The mass spectrum for that XIC can be viewed as well as a simple box and whisker plot highlighting the differences between the groups. Finally each feature is matched for identification to the composite database. If MS/MS spectra are available then the column labelled “MS/MS” is populated with “y” meaning MS/MS is available. The database identifications are listed in a table (Figure 4) and ranked by m/z error (ppm) and then alpha/numerically. Then using METLIN website (https://metlin.scripps.edu) any MS/MS confirmation could be made.

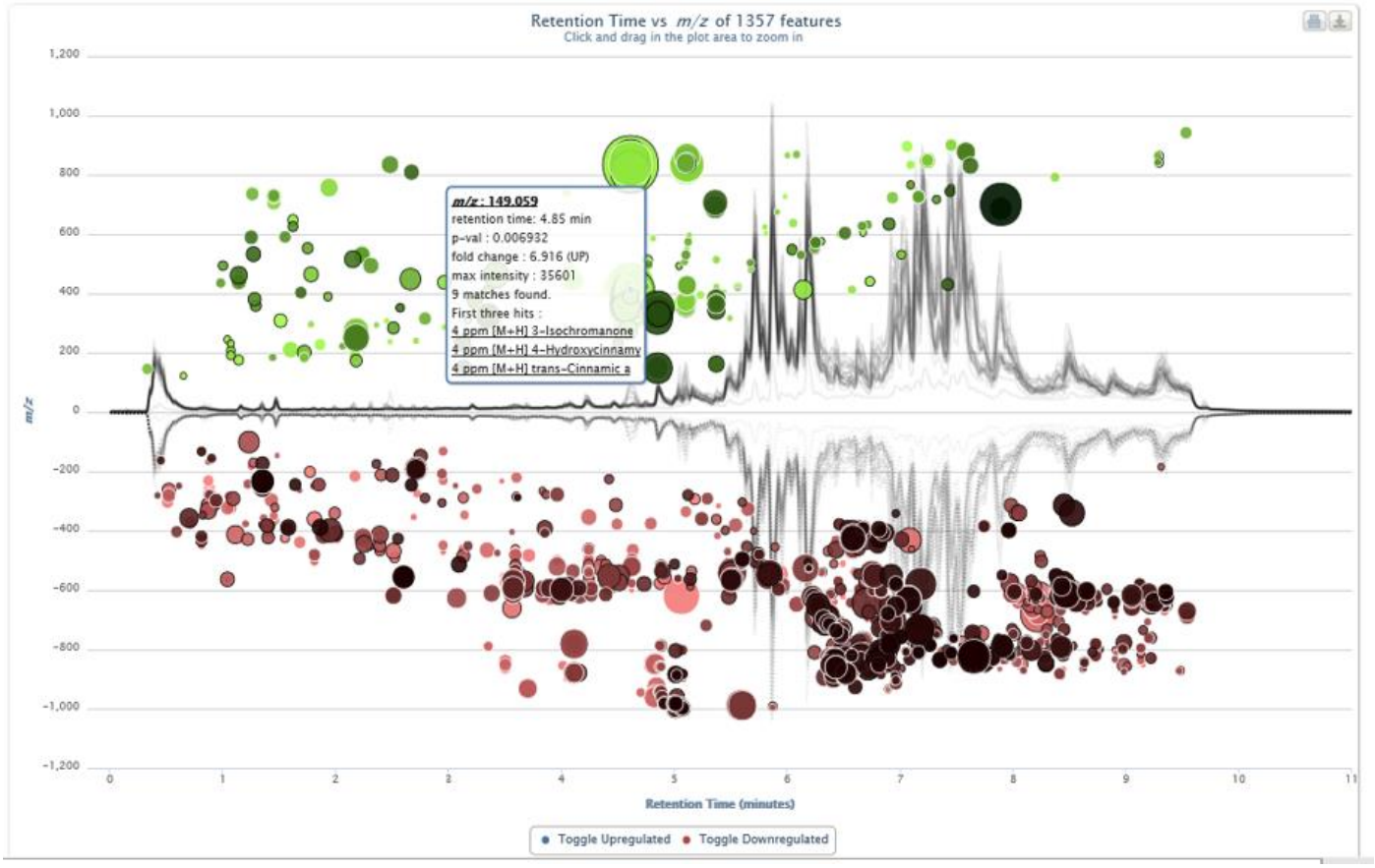

Figure 5. The Interactive Cloud Plot. The cloud plot displays features whose intensities are altered between sample groups. In this case 1357 features where highlighted as having a p-value of less than 0.01. Up-regulated features are represented as circles on the top of the plot and down-regulated features are represented as circles on the bottom of the plot, where the size and the degree of color saturation corresponds to the (log) fold change of the feature. Circles with black outlines indicate hits in the database. The lighter or darker the color of the circle relates directly to the significance of the p-value. The cloud plot is completely interactive, where you can filter p value, m/z value, as well as retention time to allow the viewer to see more/less features based on the filtering criteria.

Interactive Cloud Plot

The most compelling visualization tool in XCMS^{plus} software is the interactive cloud plot (Figure 5). Key visualization features include:

- P-value is represented by how dark or light the color is.
- Fold change is represented by the radius of each feature.
- Retention time is represented by position on the x-axis.
- Mass-to-Charge ratio is represented by position on y-axis.
- Sliders for p-value and fold change are in the Main Panel.
- Sliders for intensity, retention time, and mass-to-charge are in the Advanced Tab. A link to the table representation of the graph is displayed at the bottom along with the settings used to generate the graph.

The cloud plot is completely interactive and can be filtered on any function listed above. So you can be more stringent for the p-value and also the mean fold change allowing for only the more highly significant features to be displayed.

The end results from XCMS^{plus} mean a confirmed list of significant metabolites. From the Zucker rat study it was observed that there were many lipids changes amongst the three phenotypes studied (lean, fatty and obese) as well as other small metabolites including bile acids.

Conclusions

XCMS^{plus} software can accelerate project completion time from ***weeks to days*** with local processing and batch analysis. With unlimited data storage capacity and the security of a local desktop package, XCMS^{plus} offers the complete solution for untargeted metabolomics research. Beyond the robust data processing and visualization features of XCMS^{plus}, virtually the entire metabolomics community can privately (and publicly) share their data and results within the XCMS^{plus}.

References

1. Benton H.P. and Ivanisevic J. *et al*. Autonomous metabolomics for rapid metabolite identification in global profiling. Analytical Chemistry, 2015 87(2):884-91
2. METLIN database (https://metlin.scripps.edu)
3. Complementary LC- and GC-Mass Spectrometry Techniques Provide Broader Coverage of the Metabolome, SCIEX technical note 10620114-01.

Document number: RUO-MKT-02-2326-A

Headquarters
500 Old Connecticut Path | Framingham, MA 01701 USA
Phone 508-383-7700
www.sciex.com

International Sales
For our office locations please call the division headquarters or refer to our website at
www.sciex.com/offices