Polly El-MAVEN and multi-omics data analysis

Cloud Processing of Metabolomics Data

John Janiszewski and Rick Schneider, National Center for Advancing Translational Sciences (NCATS), NIH

Maheswari Karthikeyan, Sunil Dhakad, and Manmeet Dayal, Elucidata Corporation

Overview

High-resolution mass spectrometry (HRMS) based metabolomics studies can provide a holistic analysis of endogenous metabolites and establish biochemical impact. A metabolomics experiment using contemporary instrumentation can produce data files that may be hundreds of gigabytes (GBs) in size. Subsequent processing requires computing power not found on conventional personal workstations, hence the requirement for a robust computing environment that encompasses all aspects of data analysis - from peak curation to pathway analysis, and enables fast and efficient processing of the data.

Introduction

Often, no single platform can serve the above-mentioned steps, and the user is required to explore multiple tools/software with varying functionalities to generate comprehensive and understandable data. The content, format, and variety of associated metadata from processing need to be stored and accessed on-demand to speed downstream analysis, such that experimental insights will be meaningful and can be tested. Polly, a cloud-based platform, enables high-speed LC/MS data processing and metabolomic analysis in linear workflows that can be customized depending on experimental design (Figure 1).

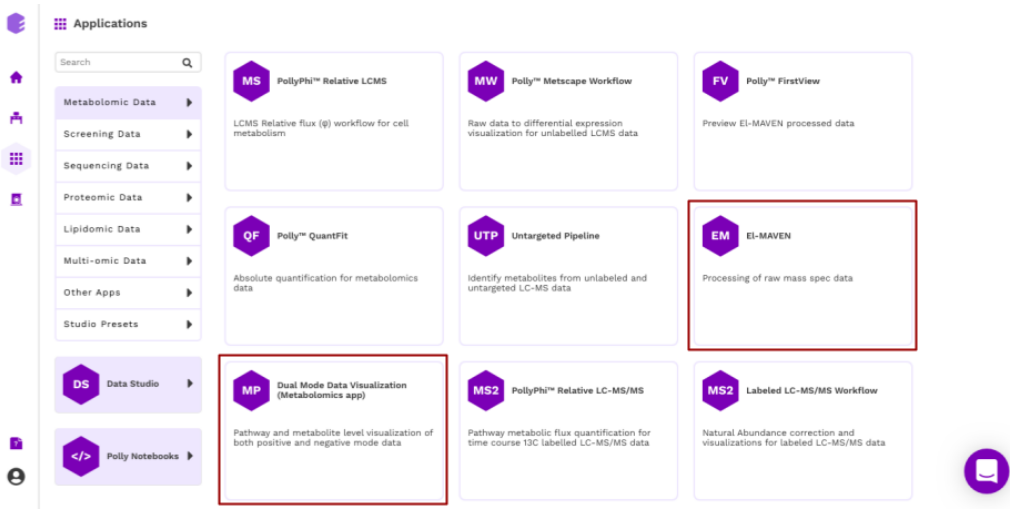

Figure 1: Polly comprises a multitude of ready-to-use workflows and applications categorized on the basis of data they cater to, encompassing all omics. The highlighted apps, Polly El-MAVEN and Polly Dual Mode Data Visualization App, enable the end-to-end analysis of metabolomics data

This application note will cover the use of El-MAVEN and the Dual Mode Data Visualization app for liquid chromatography-high resolution mass spectrometry (LC-HRMS) metabolomics data processing using a combined targeted-untargeted approach to deliver high-quality datasets in screening applications.

Finally, an example of Polly's IntOmix optimized network analysis will be provided, demonstrating the ability to generate a multi-omics understanding of perturbed pathways and network modules.

Product details

Polly

Elucidata presents Polly - an end-to-end analytics platform for LC/MS metabolomics data that efficiently streamlines integration and processing of raw data, comparisons to reference libraries, quality checks, and pathway analysis including user and third-party apps. The cloud-based platform enables seamless communication among project stakeholders and permits them to securely access, share, and simultaneously work with stored data, thereby accelerating the understanding of metabolic alterations.

El-MAVEN

El-MAVEN is an open-source mass spectrometric data analysis platform that provides fast, powerful, and interactive analysis capabilities that speeds the processing (peak integration and metabolite identification) of huge (100 GBs) datasets. El-MAVEN resides as an app in the Polly framework. The resultant data is accessible to multiple apps and customizable pipelines within Polly.

Dual-Mode Data Visualization (Metabolomics App)

Polly Dual-Mode Data Visualization (Metabolomics App) was used to perform the downstream analysis that involves normalization and scaling of the data, quality checks, statistical analysis, comparative analysis, and initial interrogations of pathway-level insights.

Polly IntOmix

Polly IntOmix is an application specifically geared towards integrated network analysis of multi-omics datasets. The focus is on active metabolic networks and submodules across experimental conditions. IntOmix analysis can reveal how metabolism, pathways and associated biomarkers shift in response to perturbation (e.g. in cell-based models of disease or target biology).

Pipeline for end-to-end metabolomic data processing on cloud

1. Custom computing environment as per user's requirement

Local workstations may be underpowered in terms of processing hefty volumes of data efficiently. Polly El-MAVEN allows the user to seamlessly process large, unwieldy data.Based on the requirements for computing power for a given set of study data, users choose a machine configuration best suited for

launching the Polly El-MAVEN application, thus eliminating the need for extensive RAM and processing requirements at a local CPU (Figure 2).

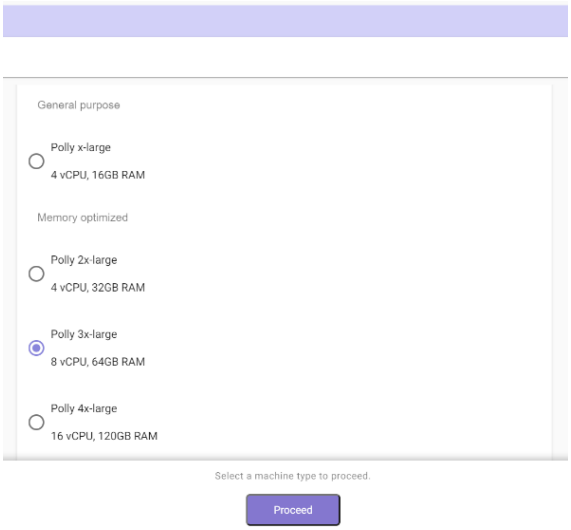

Figure 2: Machine configuration options prior to the initiation of El-MAVEN for processing and integration of metabolomics study data

2. Robust data processing on Polly El-MAVEN

Processing of raw data using a targeted approach in Polly El-MAVEN requires a compound database or spectral libraries such as the NIST library, MoNA library, IROA library, or any custom in-house spectral library with metabolites of interest for performing feature detection and curation. The availability of associated parameters for peak integration, retention alignment, exact mass, and MS/MS properties with spectral matching are helpful to gain confidence in molecular identity, specific to the data in hand (Figure 3).

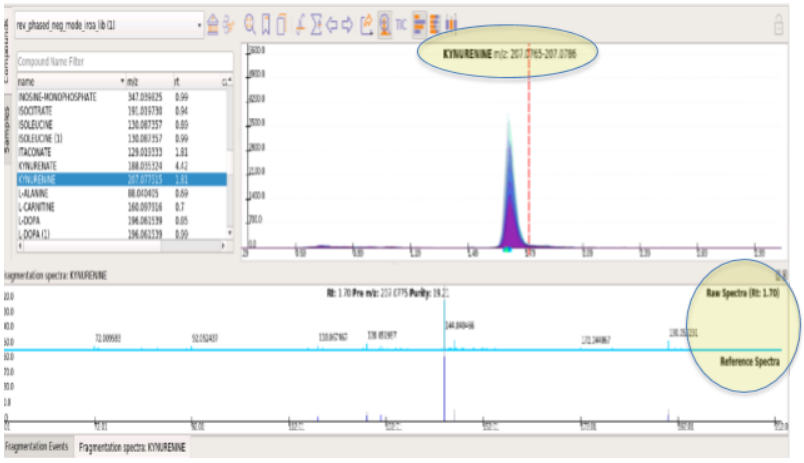

Figure 3: Spectral Matching in El-MAVEN; overlaid extracted ion chromatograms (top) of kynurenine are presented for all injections made. The vertical red line shows the average elution time recorded in the reference library, while the experimental and reference MS/MS spectral comparisons are displayed in the lower traces of the figure, demonstrating the spectral alignment of product ions and confirmation of identity

We used a library of 630 authentic metabolite standards provided by Mass Spectrometry Metabolite Library of Standards (MSMLS), IROA Technologies, Sea Girt, NJ to construct several targeted libraries. The targeted metabolite libraries contain LC retention time for each metabolite using specific column chemistry along with ppm accuracy of the molecular ion and MS/MS spectra for corresponding metabolites. Targeted library assignments may rely on retention time, ppm mass accuracy, and MS/MS spectral match for metabolite identification.

The practicality of maintaining a standard reference library is clearly realized when reproducible bioanalytical conditions have been developed and employed. Targeted analyses of study samples provide a rapid assessment with high confidence in the curated sets of metabolites for each study cohort. In this illustration, the IROA library was utilized for the reference standards, and following injection in four modes of LC/MS analysis (+/- ion C18 reverse phase, and +/- ion Hypercarb for polars), the coverage of nearly 600 metabolites across the KEGG pathway map represented excellent coverage of a wide domain of biochemical space (Figure 4).

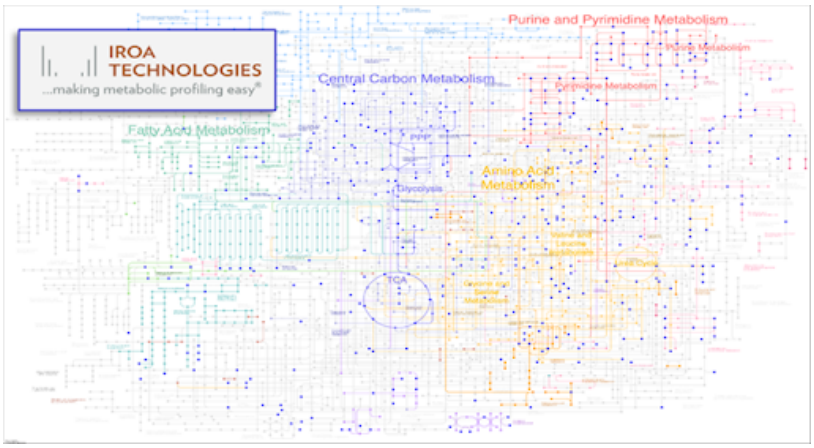

Figure 4: IROA library of targeted metabolites (blue dots) overlaid on the KEGG map, demonstrating excellent coverage of a large proportion of biochemical pathways. Molecular ion and retention time data were acquired and libraries created from four modes of analysis.
We thank Raghav Sehgal for helping us create this KEGG map.

The high-resolution mass spectral extracted ion chromatogram shows peak shape and response across study cohorts (Figure 5). Color coding of study cohorts (bar graph, right side) allows the user to quickly identify relative responses for a chosen metabolite relative to study design (e.g. dose, time, genetic status, etc.). El-MAVEN integration software was also used for all functions of peak finding, integration parameters, and peak curation for downstream analysis.

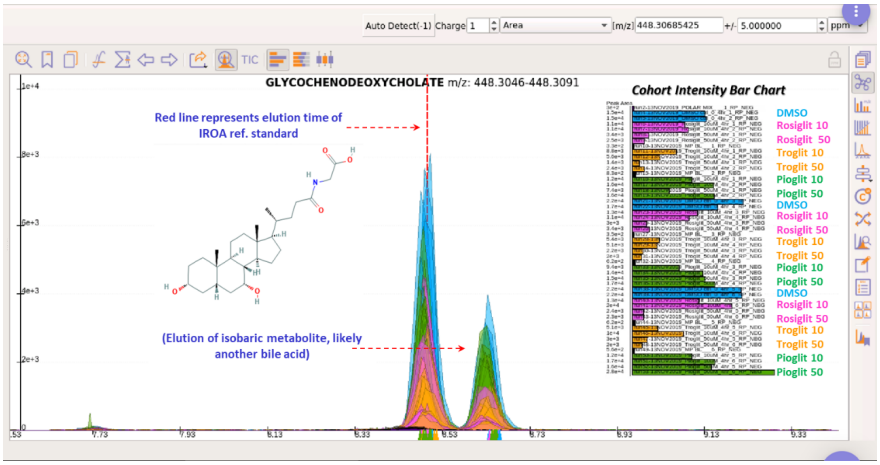

Figure 5: Features of El-MAVEN: Overlaid XIC's of m/z 448.300-448.313, representing glycochenodeoxycholate in RP neg ion mode. Bar chart at right shows summary integrated values of selected m/z for each sample injected, allowing a rapid estimate of changes across cohorts of the study. Here, metabolite levels are significantly reduced in Troglitazone cohorts (tan) when compared to DMSO controls (blue)

3. Downstream analysis of Metabolomics Data using Polly Apps

Once EI-MAVEN review of the LCMS metabolomics data is complete, the user can choose from a number of production-ready workflows, or build their own custom processing pipeline, to assess and describe results. Figures 6 and 7 provide a snapshot of some of these characteristics that are vital to understanding study data and the subsequent interpretation of metabolomics results.

The Polly Dual-Mode Data Visualization application allowed us to perform the general downstream analysis where we normalized and scaled the data. It contains methods for assessment of data quality, along with 2-D and 3-D PCA and differential expression analysis that comprise a composite of functionality used to quickly compare and describe metabolomics experimental results.

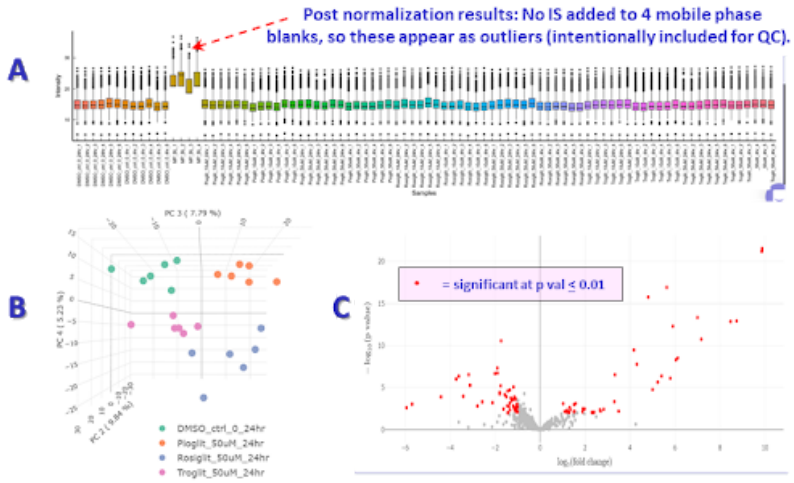

Figure 6: Assessments available to check data quality. A) Visualization of log transformed, normalized results to an internal standard. B) 3-D PCA of 50 µM substrates (24 hr) analyzed in reverse phase neg. ion mode. C) Volcano Plot of DMSO vs. troglitazone (4 hr samples), each dot representing a metabolite

Polly IntOmix aided in the pathway-level interpretation of the data and identifying the most connected network. It lists out the perturbed pathways for the cohort comparison under investigation as a result of enrichment and topology analysis. It also allows comparing the insights with existing literature in a matter of a few clicks.

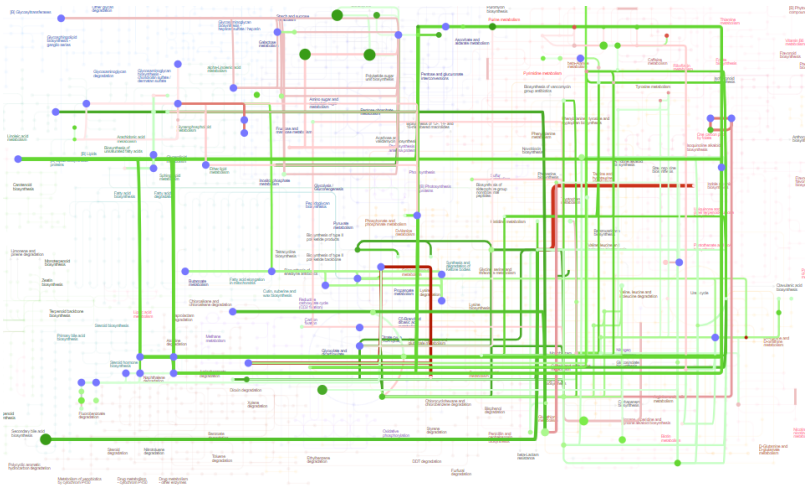

Figure 7: Polly IntOmix showing the optimized network for the Rosiglitazone 50µM with respect to DMSO at 4 hours comparison

The restore functionality in the applications allows the user to restore any analysis to the last step. Also, this enables an effective collaboration on the study findings by sharing the analyses with peers.

Summary

Polly is an enterprise solution for end-to-end processing of large multi-omics datasets that enables the generation of actionable insights from raw data with an intuitive and interactive interface.

References

- Agrawal S. et al. (2019) El-MAVEN: A Fast, Robust, and User-Friendly Mass Spectrometry Data Processing Engine for Metabolomics. In: D'Alessandro A. (eds) High-Throughput Metabolomics. Methods in Molecular Biology, vol 1978. Humana, New York, NY
- Clasquin, Michelle F et al. (2012) “LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine.” Current protocols in bioinformatics vol. Chapter 14 doi:10.1002/0471250953.bi1411s37

