Gazing into the Metaboverse: Automated exploration and contextualization of metabolic data

This manuscript was automatically generated on March 3, 2020.

Authors

ⓘ Jordan A. Berg[1], ⓘ Youjia Zhou[2], ⓘ T. Cameron Waller[1,3], ⓘ Yeyun Ouyang[1], ⓘ Ian George[1], ⓘ Tyler Van Ry[1,4], ⓘ James Cox[1,4], ⓘ Bei Wang[2], ⓘ Jared Rutter[1,5]

1. Department of Biochemistry, University of Utah
2. School of Computing, University of Utah
3. Division of Medical Genetics, Department of Medicine, School of Medicine, University of California San Diego
4. Metabolomics Core Facility, University of Utah
5. Howard Hughes Medical Institute, University of Utah

[†] To whom correspondence should be addressed:

[*] Current author order and contributions may change as manuscript in finalized.

Abstract

Metabolism is a complex network of metabolic reaction; however, scientists have had to use a largely reductionist approach to understanding these systems. While previously necessary due to technological limitations, current computer age technological advances allow us to survey, model, and explore the biological details of individual cells and populations of cells. Yet, our ability to contextualize and extract the full extent of these enormous datasets continues to lag and often results in focusing on only a handful of entities from a dataset. We present Metaboverse, a multi-omic computational analysis framework and application for the automated extraction of regulatory events, patterns, and trends from user data within the metabolic network and interactive visualization of the data. This framework will be foundational in increasing our ability to holistically understand static and temporal metabolic events and shifts and gene-metabolite intra-cooperativity, as well as ensure we obtain the maximum amount of information from our data. Metaboverse if freely available under a GPL-3.0 license at https://github.com/Metaboverse/.

Introduction

Metabolism is a complex network of reactions and interactions between genes, enzymes, complexes, and metabolites. Traditionally, to understand these complex components, scientists have adopted a reductionist approach to teasing apart the characteristics and mechanics of these processes and how they fit into the larger picture of biology and disease. By doing so, many interesting properties of metabolism can be missed. For example, in differential gene expression analysis, researchers rely on thresholds of magnitude and statistical significance to prioritize genes for follow-up study. However, doing so can inadvertently limit the scope of study of metabolism when in fact metabolism is a highly interconnected system where distal components and their modulation can have rippling effects across the network. The current approach is analogous to telling the story of Little Red Riding Hood, but only by reading the 20 most frequent words used in the study. Certainly doing so highlights key words like "wolf" and "little red riding hood," but doing so prevents a coherent story from being told and would make it difficult for someone who had never read the story of Little Red Riding Hood from comprehending the story.

Over the past decade, several computational tools have emerged and become popular for their focus on trying to solve these issues in data contextualization. We will highlight four, and while others exist, we focus on tools representative and most popular for their respective properties. First is MetaboAnalyst, which relies largely on set enrichment methods, or looking at the belongingness of sets of significantly changes analytes, for extracting interesting information. While network visualization is available, its ability to extract regulatory information is limited if non-existent. Second is Cytoscape, which focuses on network representations of metabolism and other systems. While a variety of plug-ins are available for customizing analyses, again, pattern recognition features are lacking. MetExplore focuses on the curation of networks, and is particularly useful for collaborative annotation of emerging organisms. It additionally can layer experimental data on the network for visualization. Reactome, which Metaboverse uses for the curation of biological networks, also offers analytical tools for user data, but again relies on set enrichment or manual methods for identifying patterns. While all have their respective utility, there is still a need for tools that automate pattern and trend detection, especially when data is sparse, across metabolic networks in order to extract regulatory and other features from data.

In order to address these limitations in current conventions of metabolic data analysis, contextualization, and interpretation, we created the software application, *Metaboverse*, to aid users in filling in the details of their model's metabolic story. *Metaboverse* is an interactive tool for

exploratory data analysis that searches user data in the context of the metabolic network to identify interesting patterns and trends in the data. *Metaboverse* will aid scientists in formulating new hypotheses from their data and aid them in designing follow-up experiments for the characterization of their model. *Metaboverse* operates across the entire metabolic network to quickly and automatically detect patterns and trends from a pre-designed pattern library, or can accept interactive input from the user where they can define certain patterns or trends they would like to identify across the global metabolic network. Figure 1 provides a graphical abstract to illustrate *Metaboverse's* role in the exploratory data analysis of biological data in the context of metabolism.

Traditional (reductionist)
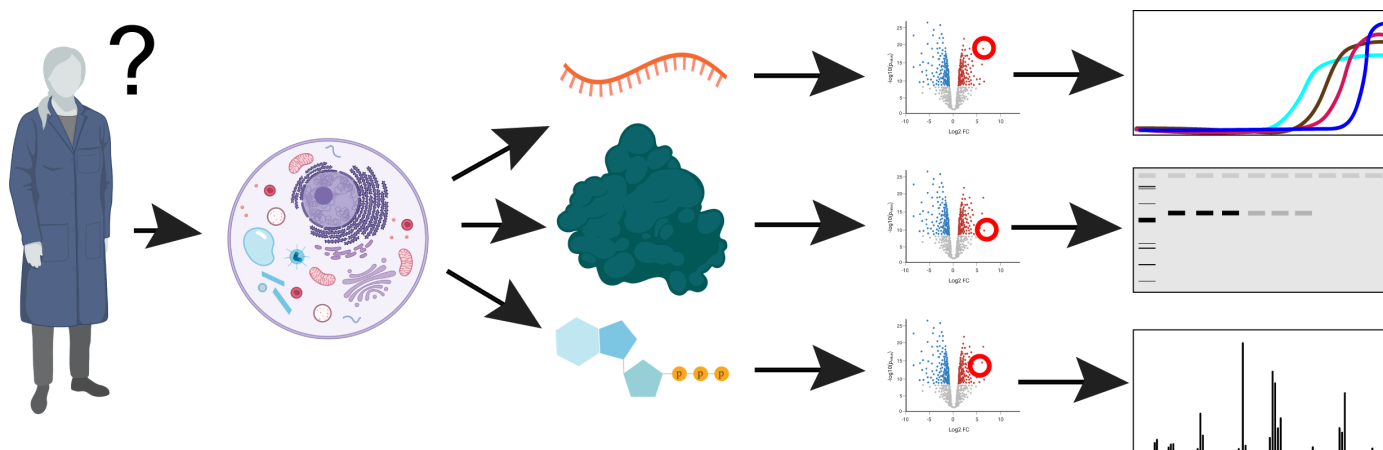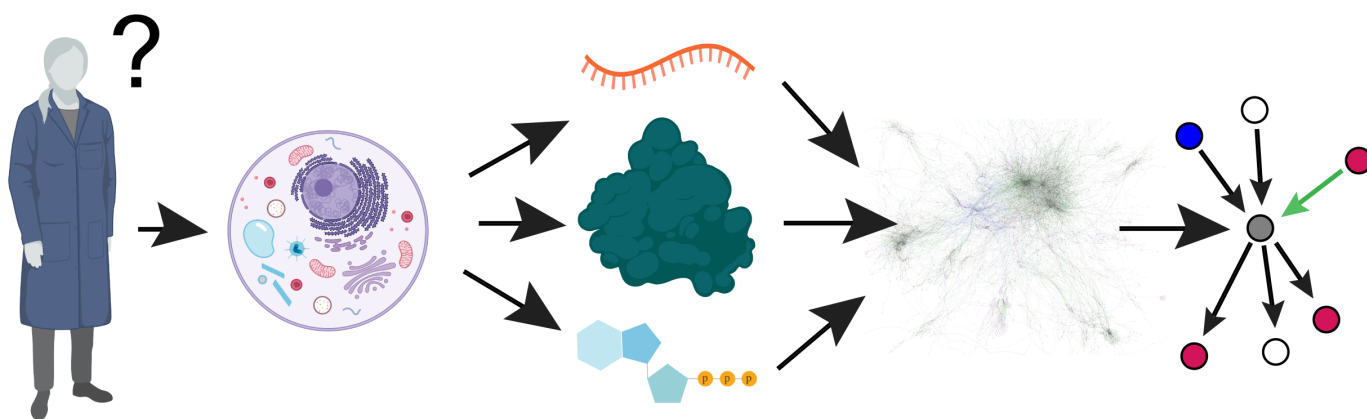

Metaboverse (holistic)

Figure 1: Metaboverse conceptual overview. Illustration comparing traditional metabolic data analysis methods and the holistic approaches that Metaboverse offers.

In order to provide a platform for the exploration of single or multi-omic metabolic data, we developed several new computational features to aid in the aims discussed above. First, we developed a pattern search engine for the rapid and automated identification of patterns and trends in -omic data on the metabolic network. Conceptually, this search engine borrows principles of topological motif searching from graph theory. In the computational science context, a motif is simply a reoccuring pattern in network structure, or the organization of network entities and their relationships to one another. However, with -omic data, we are more interested in identifying patterns in expression or abundance of genes, proteins, and metabolites. We therefore adapted this methodology to search the global metabolic patterns for interesting patterns in the network. For

example, at a reaction the input may be high and the output low, indicating some sort of regulatory event occuring in the model. *Metaboverse* will search the global network from a predefined library of regulatory patterns and return an ordered graphical table of conserved patterns. Users will also be able to design their own patterns through an interactive pattern drawing tool, and even design specific scenarios that are cognizant of feature type. For example, one might be interested in a pattern where a protein displays higher expression, but the resulting metabolite is decreased. The user can also define multi-step patterns that may occur over two or more reactions.

Another feature introduced in *Metaboverse* allows for the interactive exploration of specific reactions or reaction entities with on-the-fly pattern search analysis. Users can explore specific pathways of interest and look for other interesting patterns and trends in pathways of interest. Users can also select a specific metabolite, protein, complex, or gene and explore patterns across pathways in a feature neighborhood search. For example, a user might identify a change in one metabolite in a particular pathway and want to explore what distal effects this change has in other pathways that use this metabolite. This functionality moves the analysis away from our traditional, strictly defined pathway approach to analysis, and helps contextualize the far-reaching effects changes in metabolism can have across the classical pathways in metabolism.

One challenge in metabolomics data analysis in sparsity of data points. While thousands of metabolites exist in human metabolism, the current state of the technology for determining which mass spectra belong to which metabolite can be challenging and often results in a limited number of data points being quantified. These can lead to gaps in the metabolic network which can be challenging to explore and analyze. We therefore introduce a reaction collapse feature that allows for summary of reactions in pathways for which data is missing. This methodology can use RNA-seq data to inform prioritization of particular paths. For example, one metabolite may be converted to a downstream product in two different manners, but by using the gene expression data from a model, one can determine that one path is active while the other in inactive.

Metaboverse is designed to handle standard two-condition experiments, flux metabolomics, and time-course experiments. Time-course inputs can be single-omics, or static RNA-seq and/or proteomics with multiple time-point metabolomics. Users input fold change and statistical measures from their respective -omics, and *Metaboverse* reconciles the inputs for layering on the metabolic network. *Metaboverse* can handle data from a variety of model organisms, including humans, mouse, yeast, zebrafish, and more. The foundational curation of *Metaboverse* is built on the Reactome curations of metabolism, so any of the 90+ species available on that platform are also available within the *Metaboverse* environment. In order to validate these methodologies available in *Metaboverse* we analyzed two-condition, flux metabolomics, and time-course datasets and provide vignettes that highlight *Metaboverse's* reliability in extracting canonical features, as well as novel features and patterns, from well-defined biological models. We outline the technical specs for computational biologists in the methods section, and the biological utilities in the main text for wet bench biologists. We intend that *Metaboverse* will be foundational in our ability to more deeply and holistically explore metabolism and aid in our ability to provide more context within metabolic models.

Results

Metaboverse is a dynamic, user-friendly tool for the exploration of high-throughput biological data in organism-specific pathways

Figure 2. Overview

Data vignettes

What does Metaboverse help find or speed up compared to standard analysis?
i.e. Volcano plot vs motif search

1. Static (Ian)

Figure 3. Data
Supp Table 1. Motif results

2. Time-course (Yeyun)

Figure 4. Data
Supp Table 2. Motif results

3. Flux data (Cameron)

Figure 5. Data
Supp Table 3. Motif results

Performance

Table 1. Performance break-down
Table 2. Comparison to existing tools

Discussion

We hope that this tool will bring a new perspective to users' data and help draw the connections needed to aid them in extracting new and exciting hypotheses from their data that would be difficult to do without this tool.

Methods

1. Network Curation

2. Data overlay and broadcasting for missing entities

3. Collapsing reactions with missing expression or abundance user data

4. Regulatory pattern (motif) searches

5. Nearest neighborhood searches

6. Network visualization and exploration

6.1 Visualizing pathways and super-pathways

6.2 Visualizing compartments

6.3 Data scaling and interpretation

6.4 Dynamic network plotting

6.5 Annotation

6.6 Additional features

7. Packaging

7.1 Framework

7.2 Documentation and Test Cases

8. Validation using biological data

8.1 Curation of existing datasets

8.2 New datasets

8.3 Analysis and data visualization

Acknowledgements

Contributions

- Writing - Review & Editing: J.A.B., Y.Z., T.C.W., Y.O., T.V., J.C., B.W., J.R.
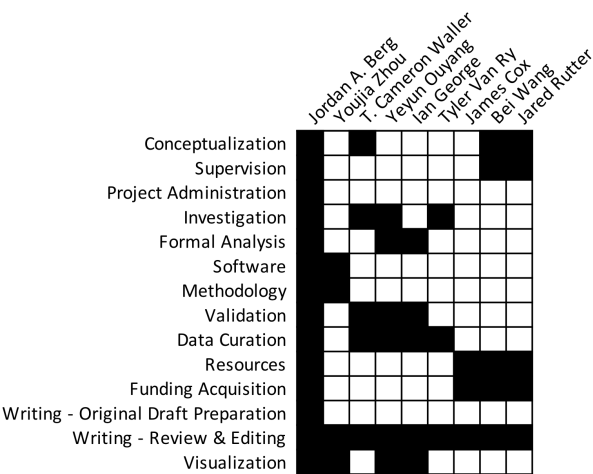- Visualization: J.A.B., Y.Z., Y.O., I.G.

Figure 2: Author contributions. Table visualizing respective contributions of each author.

Supplementary Material

References