Systems Biology Guided by XCMS Online Metabolomics

Tao Huan[1,†], Erica M. Forsberg[1,†], Duane Rinehart[1], Caroline H. Johnson[1,2], Julijana Ivanisevic[3], H. Paul Benton[1], Mingliang Fang[1,4], Aries Aisporna[1], Brian Hilmers[1], Farris L. Poole[5], Michael P. Thorgersen[5], Michael W. W. Adams[5], Gregory Krantz[6], Matthew W. Fields[6], Paul D. Robbins[7], Laura J. Niedernhofer[7], Trey Ideker[8], Erica L. Majumder[9], Judy D. Wall[9], Nicholas J.W. Rattray[2,10], Royston Goodacre[10], Luke Lairson[11], and Gary Siuzdak[1,12*]

[1]Scripps Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA
[2]Yale School of Public Health, Yale University, 60 College St, New Haven, Connecticut 06510, USA
[3]Metabolomics Platform, Faculty of Biology and Medicine, University of Lausanne, Rue du Bugnon 19, 1011 Lausanne, Switzerland
[4]School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[5]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia 30602, USA
[6]Department of Microbiology and Immunology and Center for Biofilm Engineering, Montana State University, Montana State University, Bozeman, Montana 59717, USA
[7]Departments of Metabolism and Aging, The Scripps Research Institute-Florida, Jupiter, Florida 33458, USA
[8]Department of Medicine, University of California San Diego, La Jolla, California 92093, USA
[9]Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, USA
[10]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Manchester, M1 7DN, United Kingdom
[11]Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA
[12]Departments of Chemistry, Molecular and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

[†]These authors contributed equally to this work

*Author to whom correspondence should be addressed:

Gary Siuzdak
Tel: (858) 784-9415
E-mail: siuzdak@scripps.edu
Internet: https://masspec.scripps.edu/

Contributions

T.H., E.M.F., D.R., H.P.B., A.A., T.I., M.W.W.A., P.D.R., L.J.N., M.W.F., and G.S. contributed to multi-omic platform design and development; T.H., E.M.F., M.P.T., G.K., N.J.W.R., R.G., M.F., and C.H.J contributed to data collection and analysis. T.H. and E.M.F. would like to acknowledge co-first authorship, C.H.J., J.I., T.I., and G.S. also contributed to manuscript writing.

An aim of systems biology is to understand complex interactions between genes, proteins and metabolites by integrating and modeling multiple data sources. We report an integrated-omics approach within XCMS Online[1] that automatically superimposes raw metabolomic data onto metabolic pathways, and integrates it with transcriptomic and proteomic data (XCMSOnline.scripps.edu).

Mapping downstream metabolite changes onto metabolic pathways and biological networks can provide considerable mechanistic insight that can be confirmed by association to multi-omic data. However, pathway analysis using untargeted metabolomics requires intense data curation, including feature filtering, statistical analysis and metabolite identification. Subjectively defined values such as fold change, *p*-value, and signal intensity cut-off are needed to identify significantly dysregulated metabolite features within enormous datasets. Confirming metabolite identities for pathway analysis typically requires additional tandem mass spectrometry (MS/MS) experiments and matching the spectra to standards or MS/MS spectral databases. The magnitude of these datasets makes it impractical to manually interpret and therefore the use of bioinformatic tools at each step is essential. Multiple analysis platforms are often needed to complete the entire workflow, which can take several weeks depending on the size of the sample cohort and the experience of the analyst.

XCMS was originally developed as a metabolomics data processing algorithm to extract metabolic features out of raw MS data and perform statistical analysis. The evolution of XCMS from a command line tool[2] to an intuitive cloud-based online platform[1] facilitated its use by a broader community. However, the community is still in need of user-friendly tools to take metabolomic output and associate it with metabolic pathways to identify aberrant biological processes. To address this demand, we implemented automated predictive pathway analysis[3] that operates directly on the entire metabolic feature table into the XCMS Online workflow (**Fig. 1**), removing the need to transfer data to another application, and enabling quick and efficient pathway analysis. This process involves uploading raw MS data to XCMS Online where the statistically significant features are identified, then using

Fisher's exact test, dysregulated metabolic pathways are identified from the processed accurate mass data.[3] If gene and protein data are available, they are uploaded and overlaid with the results of the metabolomic analysis. Currently there are over 7600 metabolic models available for pathway analysis from BioCyc v19.5 – 20.0 with contents being updated regularly. Further confirmation of dysregulated pathways can be performed by comparing metabolite spectra, obtained via targeted or autonomous MS/MS, with standard fragmentation spectra from METLIN, which contains MS/MS data on over 14,000 molecules[4]. To address instances where a standard spectrum is not available, we have also recently added machine learning *in silico* fragmentation data to METLIN, generating MS/MS spectra on over 220,000 additional molecules. Our workflow enables (1) evaluation of biochemical relevance by mapping high resolution MS data directly onto pathways, (2) cross-integration of genomic and proteomic data and (3) metabolite identity verification via data dependent MS/MS analysis either separately or as part of the autonomous workflow[4].

Our multi-omic analysis tool uses embedded BioCyc[5] and Uniprot[6] databases to map user-uploaded gene and protein data onto the predicted metabolic pathways (**Supplementary Figure 1**). Results can be viewed in table form or using the interactive Pathway Cloud Plot (**Figure 1**). Visualization of dysregulated pathways appear with greater overlap and statistical significance in the upper right-hand quadrant of the cloud plot. Graph features can be clicked to view more information on overlapping gene, protein and metabolite data, with links to BioCyc, KEGG, and METLIN. Important features can be readily identified, helping to decipher underlying biological mechanisms. Details on the pathway analysis and integrated omics workflow can be found in the **Supplementary Methods**. Currently, data sharing is possible between collaborators and the public and we encourage users to share their data in the XCMS Online community.

To demonstrate metabolic pathway analysis and multi-omic integration, we describe a number of representative sample sets in the **Supplementary Note,** including metabolic pathway analysis using progenitor cell proliferation data and a bacterial induced corrosion study (**Supplementary Figure 2**); proteomic integration with an aging study (**Supplementary Figure 3**); transcriptomic and proteomic integration using a human colon cancer study (**Supplementary Figure 4 and Supplementary Table 1**), a sulfate reducing bacteria nitrate stress-response study (**Supplementary Figure 5**) and a microbial media stress-response study (**Supplementary Table 2 and Supplementary Figure 6)**; and a cohort of 1,600 diabetes plasma samples (**Supplementary Figure 7**) which helps illustrate the scalability of the cloud-based XCMS Online.

Other notable tools providing pathway analysis and multi-omic integration include Galaxy-M[7], Open MS from KNIME[8], and MetaboAnalyst[9]. However, many of these tools still require separate preprocessing of LC-MS data and are not fully integrated into a single program. Our workflow automatically maps metabolomic data directly onto pathways and integrates transcriptomics and proteomics for systems-wide interpretation in one cohesive platform. Additionally, metabolic network mapping is available based on the predictive activity network algorithm[3] for analysis of metabolomic data only, with multi-omics networking in development. In the future, we will incorporate unique metabolic pathways and networks from other sources to provide more comprehensive biological resources.

Data Availability

To assist users with the new workflow, we have provided a sample dataset entitled "Ecoli_glucose-vs-adenosine" (Job ID#1133019) that can be found on XCMS Online under XCMS Public (https://xcmsonline.scripps.edu/landing_page.php?pgcontent=listPublicShares), as well as two instructional videos available within the XCMS Institute (https://xcmsonline.scripps.edu/landing_page.php?pgcontent=institute) under the Omics tab and by clicking Integrated Omics or Pathway Cloud Plot.

Acknowledgements

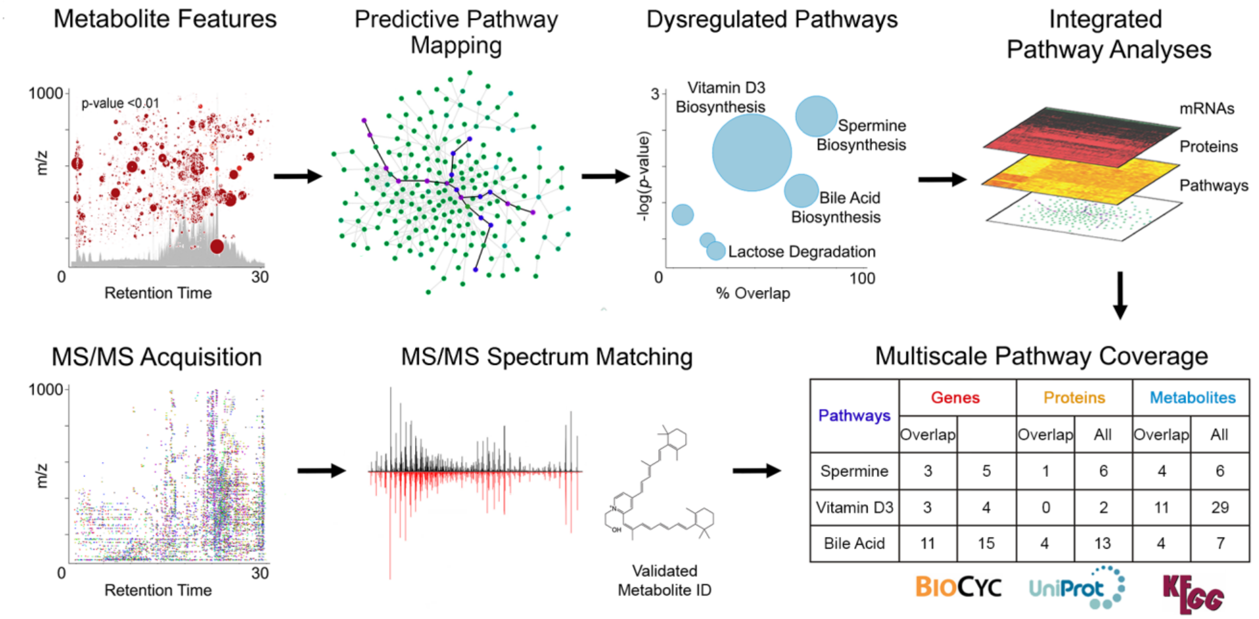
The authors thank the following for funding assistance: Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-AC02-05CH11231 (G.S.), and the National Institutes of Health (NIH) grants R01 GMH4368 (G.S.) and PO1 A1043376-02S1 (G.S.). The authors also thank Profs James Nazroo, Gindo Tampubolon, Neil Pendleton and Fredrick C. W. Wu from the University of Manchester for constructive discussions alongside the Medical Research Council grant MRC G1001375/1 (R.G.) for generous funding.

Competing Financial Interests

The authors have no competing interests as defined by Springer Nature, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Gowda, H. et al. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. *Analytical Chemistry* **86**, 6931-6939 (2014).
2. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry* **78**, 779-787 (2006).
3. Li, S.Z. et al. Predicting Network Activity from High Throughput Metabolomics. *PLoS Comput. Biol.* **9**, 11 (2013).
4. Benton, H.P. et al. Autonomous metabolomics for rapid metabolite identification in global profiling. *Analytical chemistry* **87**, 884-891 (2014).
5. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **42**, D459-D471 (2014).
6. Bateman, A. et al. UniProt: a hub for protein information. *Nucleic Acids Research* **43**, D204-D212 (2015).
7. Davidson, R.L., Weber, R.J.M., Liu, H.Y., Sharma-Oates, A. & Viant, M.R. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience* **5**, 9 (2016).
8. Aiche, S. et al. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics* **15**, 1443-1447 (2015).
9. Xia, J., Sinelnikov, I.V., Han, B. & Wishart, D.S. MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic acids research* **43**, W251-W257 (2015).



Pathways	Genes		Proteins		Metabolites	
	Overlap		Overlap	All	Overlap	All
Spermine	3	5	1	6	4	6
Vitamin D3	3	4	0	2	11	29
Bile Acid	11	15	4	13	4	7

Figure 1. Workflow for metabolomic data and pathway analysis using XCMS Online. A Metabolite Feature Table of statistically significant features is generated from standard XCMS processing; these features automatically undergo Predictive Pathway Mapping using a specified biological model. The pathway cloud plot shows Dysregulated Pathways (blue circles) with increasing statistical significance on the y-axis, metabolite overlap on the x-axis and total number of metabolites in the pathway represented by the circle radius. The Multiscale Pathway Coverage table presents enriched metabolic pathways with overlapped and total metabolites, genes and proteins. MS/MS Data confirms dysregulated pathways by matching metabolite MS/MS spectra with the METLIN database.