HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2023 June 16.

Published in final edited form as:

Anal Chem. 2018 July 17; 90(14): 8396–8403. doi:10.1021/acs.analchem.8b00875.

Autonomous Multi-Modal Metabolomics Data Integration for Comprehensive Pathway Analysis

Tao Huan[†], **Amelia Palermo**[†], **Julijana Ivanisevic**[°], **Duane Rinehart**[†], **David Edler**[§], **Thiery Phommavongsay**[†], **H. Paul Benton**[†], **Carlos Guijas**[†], **Xavi Domingo**[†], **Benedikt Warth**[‖], **Gary Siuzdak**[†,⊥,*]

[†]Scripps Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

[°]Metabolomics Platform, Faculty of Biology and Medicine, University of Lausanne, Switzerland

[§]Department of Molecular Medicine and Surgery, Karolinska Institute, 171 77 Stockholm, Sweden

[‖]Department of Food Chemistry and Toxicology, Faculty of Chemistry, University of Vienna, Währingerstraße 38, 1090 Vienna, Austria

[⊥]Departments of Chemistry, Molecular and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

Abstract

Comprehensive metabolomic data acquisition can be achieved using multiple orthogonal separation and mass spectrometry (MS) analytical techniques. However, drawing biologically relevant conclusions from this data and combining it with additional layers of data collected by other omic technologies present a significant bioinformatic challenge. To address this, a data processing approach was designed to automate the comprehensive prediction of dysregulated metabolic pathways/networks from multiple data sources. The platform autonomously integrates multiple MS-based metabolomics data types without constraints due to different sample preparation/extraction, chromatographic separation, or MS detection method. This multi-modal analysis streamlines the extraction of biological information from the metabolomics data as well as the contextualization within proteomics and transcriptomics datasets. As a proof of concept, this multi-modal analysis approach was applied to a colorectal cancer (CRC) study, in which complementary liquid chromatography mass spectrometry (LC/MS) data were combined with proteomic and transcriptomic data. Our approach provided a highly resolved overview of colon cancer metabolic dysregulation, with an average 17% increase of detected dysregulated metabolites per pathway and an increase in metabolic pathway prediction confidence. Moreover, 95% of the altered metabolic pathways matched with the dysregulated genes and proteins,

[*]Author to whom correspondence should be addressed: Gary Siuzdak, Tel: (858) 784-9415, siuzdak@scripps.edu.

Conflict of Interest Disclosure

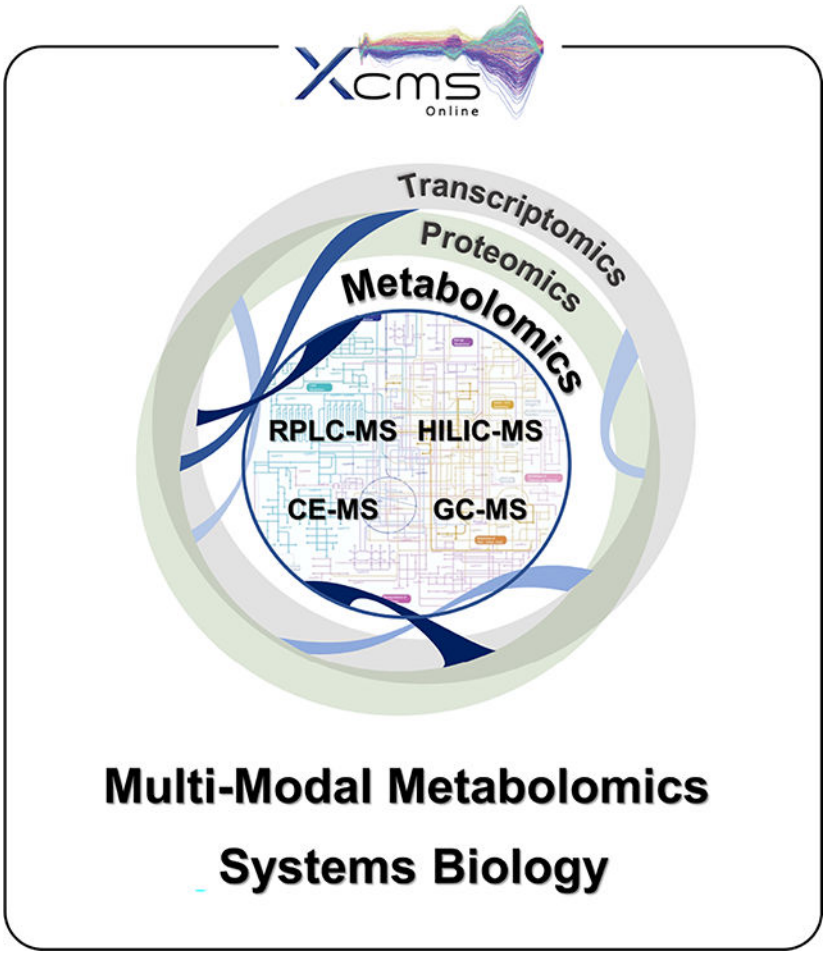
The authors declare no competing financial interest.

Supporting information.

Colon tissue sample preparation; LC-MS data collection; autonomous MS/MS data collection; LC-MS data processing; metabolite identification; detailed clinical features of the colon cancer samples; predicted pathway dysregulation in each analytical mode; predicted pathway dysregulations in the multimodal XCMS analysis; top 10 predicted pathway dysregulation ranked by the number of dysregulated genes.

providing additional validation at a systems level. The analysis platform is currently available via the XCMS Online (XCMSOnline.scripps.edu).

Graphical Abstract



Introduction

Metabolomics is emerging as an indispensable technology in the post-genomic era of biology, correlating 'omic' data analysis towards biomarker discovery[1,2], mechanistic pathway findings[3] and ultimately metabolite activity discovery for phenotype modulation[4]. Among various analytical platforms utilized to perform metabolomic analysis, mass spectrometry (MS) is the most prominent technology owing to its high throughput, sensitivity and specificity. However, metabolomic data analyses remain analytically biased by the technique being used to perform the experiments including the. type of chromatography and ionization. For example, reversed phase liquid chromatography mass spectrometry (RP-LC/MS) typically favors hydrophobic metabolites and ultimately limits comprehensive metabolome analysis. To achieve broad metabolome coverage for comprehensive biological interpretation, the combination of multiple orthogonal or complementary metabolomics platforms has been proposed[5–8]. This "*multi-modal*" strategy

involves the parallel analysis of the same batch of biological samples using several complementary analytical techniques and approaches. For instance, a dual-separation LC-MS strategy has been developed to profile metabolites in both lipid metabolism and central carbon metabolism using reversed phase LC (RPLC) in positive mode electrospray ionization (ESI) or hydrophilic interaction liquid chromatography (HILIC) with ESI[5]. In yet another example, gas chromatography mass spectrometry (GC-MS) and LC-MS metabolic profiling approaches have been combined to provide a systems-level understanding of the pathological metabolic outcomes of aristolochic acid-induced nephrotoxicity[9]. More recently, MS-based hydrophilic and hydrophobic metabolite profiling have been performed to comprehensively elucidate the metabolic differences between a wild-type and a recombinant human cell line[10]. Taken together these examples demonstrated the vast, but so far not routinely employed potential of multi-modal analysis in global metabolomics studies.

The wide metabolome coverage achieved by multi-modal profiling will allow for an in-depth mechanistic understanding through coordinated mapping of detected metabolic changes onto biological pathways. However, performing these integrative pathway analyses following the multi-modal metabolite profiling is not a straightforward task. Prior to pathway analysis metabolomics data require preprocessing (i.e. metabolite feature extraction, alignment, grouping and annotation), statistical analysis and metabolite identification, to prepare a list of confirmed dysregulated metabolites[11–13]. Although the development of bioinformatic tools, such as XCMS Online[14] and MZmine[15], have significantly streamlined the process of metabolic feature extraction, alignment and statistical analysis, metabolite identification still remains time-consuming and labor-intensive part of the workflow[16]. For example, to perform pathway analysis following a multi-modal metabolomic experiment, metabolite identification needs to be repeated for each metabolomic data set, making the data preprocessing workload demanding and time-consuming. In addition, bioinformatic algorithms to automatically deduce the overlap and merge the complementary metabolite information from different analytical modes are lacking. In order to gain a more comprehensive biological understanding from multi-modal metabolomic data, researchers are processing each individual metabolomic data set separately and manually integrating the complementary metabolite information. Since the manual data integration relies heavily on the experience of the analyst and is prone to the bias of the individual researcher, the quality of data integration and pathway interpretation is not always ensured. These afore mentioned challenges in interpreting multi-modal metabolomic data warrant the development of bioinformatic tools to enhance processing efficiency and interpretation of multimodal metabolomics data sets, thereby directly linking the obtained comprehensive metabolomic information to its underlying biological context.

To address these challenges, we designed an autonomous multi-modal analysis approach within the cloud-based XCMS Online[14] platform to integrate the heterogeneous metabolomic data sets. The approach is composed of two modules. The first module embeds a pathway analysis algorithm that can automatically combine metabolomic data generated from multiple analytical platforms including RP(+), RP(−), HILIC(+), and HILIC(−) and perform pathway analysis prior to the confirmation of putative metabolic identities in each data set. Importantly, as long as the accurate masses of the metabolite features are

provided, there are no other restrictions on the design of the analytical platform to perform integrated pathway analysis. The second module is a metabolomics-guided systems biology tool developed to overlay the user-uploaded transcriptomic and/or proteomic data onto the predicted pathways generated from the first module. In addition, this has been implemented to streamline the multi-modal metabolomics workflow, from metabolomic data processing and statistical analysis to integrative pathway prediction and multi-omic integration. The performance of the approach has been demonstrated in a colon cancer study. Following data processing, multi-modal metabolomic analysis was integrated with transcriptomic and proteomic data for a systems-level pathway analysis.

Experimental Section

Multi-modal XCMS web design.

The multi-modal XCMS web interface was constructed using PHP and JavaScript. The pathway analysis program was coded in Python and placed on a dedicated server[17]. Multi-modal XCMS analysis results, including dysregulated metabolic pathways and underlying metabolic information, are zipped and stored in the results folder on the server. The pathway analysis results can also be visualized from the XCMS Online website through the "Systems Biology Results" button on the left side panel of the results summary page. Tables presenting the multi-modal XCMS analysis results, including metabolic pathway results and predictive metabolite results are displayed using a JavaScript package, DataTables (https://datatables.net/). The interactive pathway visualization tool, "Pathway Cloud Plot", was displayed using a JavaScript package, Highcharts (https://www.highcharts.com/).

The back-end multi-omic data integration algorithm is programmed in python; the code was modified to interface with a relational database and run in a clustered environment. The front-end views are a mixture of PHP, HTML5, and JavaScript, optimized for responsive data visualization.

Chemicals and reagents.

Ammonium acetate (NH_4AC) and ammonium hydroxide (NH_4OH) were purchased from Sigma Aldrich. LC-MS grade 0.1% formic acid (FA) in acetonitrile (ACN), 0.1% FA in water (H_2O), and methanol (MeOH) were purchased from Honeywell (Muskegon, MI, USA). LC-MS grade acetonitrile (ACN) was purchased from Fisher Scientific (Morris Plains, NJ, USA), LC-MS grade water was purchased from J.T. Baker (Philipsburg, NJ, USA).

Sample preparation.

Ten colon cancer and paired histologically normal tissues were collected from patients undergoing surgery and were immediately stored at −80 °C (the study was approved by the regional ethical board at the Karolinska Institute). Clinical features of the samples can be found in Supporting Information (SI), Table S-1. For each sample, 10 mg frozen colon tissue was used for metabolite extraction. The detailed extraction protocols can be found in the Text S-1.

Multi-modal LC-MS analysis.

Complementary metabolomic profiling were carried out using a Bruker Impact II QTOF mass spectrometer (Billerica, MA, USA) coupled with an Agilent 1200 series capillary HPLC system (Palo Alto, CA, USA) in four different analytical modes to achieve a comprehensive metabolome coverage. These modes include RPLC-MS in ESI positive and negative modes and HILIC-MS in ESI positive and negative modes. Detailed LC-MS parameters for all the four analyses can be found in the Text S-2. The order of sample sequence was randomized. Autonomous tandem LC-MS experiments were performed on the pooled samples to collect MS/MS data using the same LC-MS instrumentation for metabolite identification.

Metabolomic data processing and multi-modal analysis.

Internal mass calibration was performed using the sodium formate (NaFA) segment programmed in the LC-MS analysis. MS data were converted to mzXML files using Bruker Compass Data Analysis 4.4. These files were uploaded to XCMS Online for data processing including peak detection, retention time correction, profile alignment, and isotope annotation. Data from all four analyses were processed using pairwise comparison (cancer tumor vs. adjacent control tissue) and the parameter settings can be found in the Text S-3. Dysregulated metabolite features were confirmed by matching experimental tandem MS data against standard MS/MS spectra in METLIN MS/MS spectral library[18]. After the completion of metabolomic data processing, a multi-modal XCMS job was created by selecting the “multi-modal” job type in XCMS Online Job tab to include all four processed data sets that were acquired using different LC-MS techniques. . Prior to multi-modal analysis and pathway prediction, the appropriate p-value, fold change, and intensity thresholds for multi-modal analysis were defined in the parameter setting window (values shown in Table S-2). Finally, the integration of heterogenous metabolomic data sets was performed automatically, followed by the prediction of dysregulated metabolic pathways and display the pathway analysis results. The algorithm for multi-modal data integration and pathway prediction is detailed in the Results and Discussion section.

Transcriptomic and proteomic data sets.

Transcriptomic and proteomic data were integrated with the predicted metabolic pathways for the systems-level understanding of the pathway dysregulation. Comprehensive transcriptomic data downloaded from netgestalt[19]. These data were originally generated from The Cancer Genome Atlas (TCGA) in a study of 22 colon cancer tissue samples vs. 22 normal tissue samples[20]. A total of 7,138 genes with p-value ≤ 0.01 and fold change ≥ 4 were selected as dysregulated genes for multi-omics integration. Comprehensive proteomic data was also downloaded from netgestalt[19]. The data was originally generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) in a study of 90 colon cancer tissue samples vs. 30 normal tissue samples[21]. A total of 2,545 proteins with p-value ≤ 0.01 and fold change ≥ 2 were considered as overexpressed and used in multi-omic analysis. The p-value and fold change thresholds for protein and gene data were defined following the common literature settings.

Results and Discussion

Integrating heterogeneous analytical data for pathway analysis.

The advantage of using multiple analytical strategies to gain a comprehensive coverage of metabolite classes is well understood, however a significant challenge remains on how to directly correlate the metabolic information for pathway analysis across different LC/MS analytical platforms. This is largely because metabolites contain high level of chemical diversity yielding different chromatographic retention and ionization behaviors depending on the applied analytical approach, making comparative analysis across disparate data sets difficult. Conventionally, to carry out pathway analysis, dysregulated metabolites need to be confirmed in each analytical mode typically requiring significant manual effort. And while the prediction of dysregulated pathways from metabolite features (mainly based on accurate m/z ratios and putative metabolite matches) has recently become available[17], each data set in a multi-modal metabolomics study still must be interpreted individually (Figure 1). In contrast, the pathway analysis described here makes it possible to automatically integrate multiple metabolomic data sets for pathway analysis allowing not only for more comprehensive but also more time effective data evaluation (Figure 1A). It is also worth noting that this pathway analysis tool not only takes LC/MS data, but also other types of mass spectrometry data, such as capillary electrophoresis (CE)/MS and chemical ionization (CI) GC/MS (Figure 1A).

The pathway analysis algorithm was developed to facilitate the biological interpretation of metabolomic data sets that were generated using heterogeneous experimental approaches (Figure 1A). The algorithm initially processes the metabolite feature tables from the individual XCMS Online jobs. These tables contain accurate *m/z* values, retention times, and statistical *p*-values of all the metabolite features extracted by XCMS, a metabolomics peak picking and statistical analysis package embedded in XCMS Online[22]. A user-defined *p*-value cutoff is applied to divide each feature table into statistically significant and non-significant feature lists. The accurate *m/z* values of the metabolite features from both lists are then matched against background knowledge - BioCyc metabolite database[23] - to assign putative metabolite identities (mass error for matching is defined by user depending on instrument accuracy). This preprocessing step generates two lists of significant and non-significant putative metabolite identities. Following their generation, the output lists for each metabolomics experiment are finally merged into two lists containing significant and reference (non-significant) putative metabolite identities. Shared putative metabolite identities in multiple analyses are deduplicated in the merged list, thus all putative metabolites will be weighted the same in the pathway analysis regardless of the number of techniques or platforms by which they were detected. Once the two merged lists are prepared, pathway enrichment analysis is performed using Fisher's exact test (FET) described in the *mummichog*[24] algorithm, which evaluates the probability of a pathway being dysregulated given the number of dysregulated metabolite identities involved. This *mummichog*-based pathway analysis strategy directly uses putative metabolic identities for accelerated pathway analysis, prioritizing the discovery of biological information from the metabolomic data set and leaving metabolite identification as a means of pathway confirmation in the final step of the metabolomics workflow. It is worth noting that since the

metabolic pathways are predicted using putative metabolite identities, there is a possibility of false positive pathway identification due to false positive identity assignments. Therefore, further validation experiments, such as MS/MS-based metabolite identification, should be performed to confirm the metabolite identities. The last update of the METLIN database provides MS/MS spectra covering over 15,000 metabolites, including lipids, amino acids, peptides, and natural products, among other chemical classes. By using METLIN spectral library for metabolite identify validation, it is possible to unbiasedly and thoroughly confirm the dysregulated metabolites and metabolic pathways.

To analyze multi-modal metabolomics data sets, users need to first process each individual metabolomic analysis (pairwise or multigroup) on XCMS Online to generate metabolite feature tables[14]. Users can then create a multi-modal job to integrate the complementary information from several different metabolomic analysis. Several parameters, including polarity, *p*-value cutoff, intensity threshold and mass tolerance, need to be defined for each metabolomic analysis to accurately extract the significant and the non-significant lists of features in each data set. Parameter settings have been detailed elsewhere[25]. The user also needs to define the proper metabolic model for pathway analysis by selecting samples biosource. Although central carbon metabolism, such as TCA cycle, glycolysis, oxidative phosphorylation, purine and pyrimidine metabolism, involved in energy production and storage are highly conserved and shared across different biological species, there are many species-specific metabolic pathways[23]. To provide precise pathway analysis, multi-modal XCMS archives comprehensive pathway information from over 7,600 metabolic models from BioCyc[23] with more pathways being implemented from Reactome[26] and Wikipathways[27]. After defining the parameters, users can confirm all the settings and submit the job.

The process of multi-modal pathway analysis takes only a few minutes and the user will receive an automated email notification once the job is completed. The user can then view the analysis results in XCMS Online through a web browser and download for further analysis (if desired) and archiving. Firstly, the multi-modal job results summary page presents the total ion chromatogram (TICs), the cloud plot, and the score plots of principal component analysis (PCA) from each individual job. This provides the user a quick glance at all different experiments and datasets involved in a multi-modal analysis. Secondly, the multi-modal integrative pathway analysis results are presented in three visualization modules implemented in multi-modal XCMS, including pathway analysis results, pathway cloud plot, and predicted metabolite results (Figure 2). It is important to note that in the predictive metabolite results table (Figure 2C), metabolite features matching the same metabolite according to their *m/z* values and possible adduct formations are all listed. The fold changes and *p*-values of these metabolite features allow for a rapid similarity comparison to find out the possible existence of the same metabolite in different analytical modes, therefore assisting and facilitating the metabolite identification process. Finally, to gain a detailed understanding on how the dysregulated metabolites contribute to a pathway dysregulation, users can visualize the list of dysregulated metabolite features simply by clicking on the number of overlapping metabolites in a particular dysregulated pathway. (Figure 3).

Incorporating transcriptomic and proteomic data into the metabolomic pathway analysis.

Bringing metabolomics into systems biology is an emerging research direction for the functional understanding of metabolites at a systems level. However, due to the heterogeneous chemical natures of metabolites, the metabolome coverage offered by a single untargeted profiling approach is usually limited. Therefore, integrating the more unbiased transcriptome and proteome profiling results with metabolomic data covering only limited metabolic information does not permit the complete interpretation of the investigated biological process. Fortunately, with the help of multi-modal XCMS, it is now possible to combine metabolic information from complementary metabolomic data sets to generate comprehensive information on pathway dysregulation. Further multi-omic data integration based on such complete pathway information enables a truly comprehensive biological understating at the global scale.

After the completion of pathway analysis, users can upload transcriptomic and/or proteomic data to perform multi-omic data integration. Dysregulated genes should be uploaded as a list of gene symbols and dysregulated proteins should be uploaded as a list of either gene symbols or protein UniProt accession IDs. These omic data lists have to be uploaded in comma separated values (csv) file format. A matching algorithm[17] is embedded to match user-uploaded dysregulated genes and/or proteins against genes and/or proteins implicated in the metabolic pathways that have been identified as enriched following pathway analysis (as described in previous section). After the multi-omic data integration, the overlapping genes and proteins in each pathway are presented in the systems biology results page with detailed overlapping information accessed by clicking the numbers (Figure 2A). Multi-omic data integration ultimately provides yet another level of validation of pathway analyses based on metabolomic data, thus enabling for a rapid and more meaningful analysis at a systems level.

Colon cancer.

The metabolic activities in cancer cells are fundamentally different from those in normal epithelial cells. This dramatic metabolic reprogramming meets the raised demand for nutrients, bioenergetics and biosynthesis, and fuels cancer cell growth and proliferation[28–30]. Thus, a better understanding of the associated metabolic dysregulation may enable the development and optimization of therapeutic strategies to selectively target cancer metabolic vulnerabilities.

We used tissue samples to assess the performance of multi-modal pathway analysis using XCMS. This study was carried out in four different analytical modes to achieve the comprehensive coverage of colon cancer metabolome and lipidome. Using four analytical modes, 29,394, 13,231, 14,610 and 8,565 metabolic features were detected in RP(+), RP(−), HILIC(+) and HILIC(−) analyses, respectively. Among them, 881, 721, 587 and 266 features were statistically significant (p-value ≤ 0.01) with fold changes ≥ 1.5 or ≤ 0.67 and MS signal intensity ≥ 10,000.

In the next step we applied multi-modal analysis to the results from different experiments and to perform integrative pathway analysis. Many dysregulated pathways were uniquely predicted only in one analytical mode, reflecting the unique metabolome coverage of

each analytical mode. For instance, histidine degradation was only observed in HILIC(+) analysis, where histidine and its derivatives were detected (Table S-3). Other dysregulated pathways were commonly predicted in data sets from two (or more) different metabolomic analysis modes. In these cases, the dysregulated pathway information from each analytical mode were complementary to each other. For example, the dysregulation of adenine and adenosine salvage I pathway was observed in HILIC(+) and HILIC(−) analytical modes. In the dysregulation of this pathway, while the upregulation of adenosine was commonly detected in both analytical modes, the upregulation of adenine was uniquely detected by HILIC(+) and the upregulation of 5-phospho-α-D-ribose 1-diphosphate and AMP were uniquely detected by HILIC(−) (Table S-3). These examples demonstrate that by integrating the unique and complementary metabolic information from multiple modes, it is possible to gain an unprecedented view of pathways dysregulation. Compared to the pathway analysis results obtained by using a single metabolomic data set, multi-modal-based pathway analysis has improved performance in two aspects. First, multi-modal XCMS automatically combines the metabolic information from multiple analytical modes, therefore dysregulated metabolites belonging to the same pathways are grouped together, leading to an average of 17% increase in the number of dysregulated metabolites per pathway (Figure 4). Second, the increased number of dysregulated metabolites detected in a given pathway improves the confidence of pathway prediction as indicated by an associated lower statistical *p*-value, which enables the discovery of an additional of 53 dysregulated pathways (Figure 4). If all the predicted pathways were included in the comparison without setting a threshold based on their p-values, we could have observed more pathway overlapping between different data sets (data not shown). However, the pathway p-value is affected by the number of dysregulated metabolites detected in each pathway and the pathways with high p-values (p-value > 0.05) usually don't present enough number of dysregulated metabolites. These pathways show limited statistical confidence and they were excluded in our real comparison. Further, combining the comprehensive metabolic information from the multi-modal XCMS analysis, we were able to extract the colon cancer-related metabolic network as shown in Figure 5. Importantly, all the involved metabolites were positively confirmed using MS/MS spectra matching against the METLIN spectra library (Table S-6).

Since metabolic pathways are predicted using putative metabolite identities, some dysregulated pathways can be false positively discovered due to falsely assigned metabolite identities. We noticed that the largest source of incorrect identity assignment is due to artifact metabolite features generated by low quality peak picking. To facilitate rapid inspection of metabolite annotations, we have implemented the data visualization tool, allowing users to have direct access to the LC chromatogram, MS spectra and box plots of the metabolite features (Figure 3). Using this function, metabolite features not correctly extracted during the peak picking process can be immediately spotted and the pathways predicted based on these features can be excluded from further metabolite identification and biological interpretation. On the other hand, metabolic pathways with more dysregulated metabolites are more likely to be relevant to the biological activity. Therefore, the number of dysregulated metabolites can be used to gauge the false positives in pathway prediction. For instance, in our colon cancer analysis, the top 20 pathways predicted in multi-modal XCMS (Table S-5), with an average of 6.6 dysregulated metabolites involved, are consistent

with previous reports on colon cancer. In addition, users can also confirm the predicted pathway information through metabolite identification based on their biological knowledge. Metabolic pathways known to be relevant to the given biological research topic have higher chance to be correctly predicted as the involved dysregulated metabolites are more likely to be presented in the data set. Therefore, users can first perform metabolite identification for the pathways that are relevant to the biological question. This biological knowledge-based metabolite identification is an effective way to confirm the pathways predicted from our multi-modal XCMS analysis. For instance, we started our pathway confirmation by the metabolites involved in pathways that are relevant to cancer metabolism (Figure 5). As shown in Table S-6, all the putative metabolic identities were confirmed as correct assignments, thus allowing us to further interpret the corresponding dysregulated pathways. Further, colon cancer transcriptomic and proteomic data were uploaded for multi-omic data integration to gain a systems-level understanding of these dysregulated pathways. These colon cancer-associated multi-omic data sets were obtained from public available data repositories[31,32]. While it is generally recommended to use multi-omic data sets generated from the same set of biological samples, obtaining data from publicly available repositories, where experimental conditions are either the same or very similar, can also be highly informative[33,34]. This is especially useful for studies where a large quantity of curated data is available, such as human cancer research[20]. Moreover, given the recent effort of developing infrastructure for data curation and sharing, multi-omic integration using open-access data will be very convenient and practical in the near future[35–38].

By integrating colon cancer transcriptomic and proteomic data, the dysregulated pathways were confirmed on the multi-omic level. Overall, 95% of the dysregulated metabolic pathways were matched by both dysregulated genes and proteins. Table S-7 lists top 10 dysregulated pathways ranked by the number of dysregulated genes. The multi-omic information allows a comprehensive mechanistic understanding of the pathway dysregulation. For instance, the dysregulation of spermine and spermidine metabolism is strongly correlated with colon cancer progress, which has been evidenced in many studies[39]. By using multi-modal XCMS, we were able to integrate multi-modal metabolomic data sets together with colon cancer transcriptomic and proteomic data to understand the dysregulated genes, proteins and metabolites on the global scale (Figure 6).

After multi-modal metabolomics analysis, the biological relevance of these dysregulated pathways require further verification, e.g., whether or not they are causally associated with colon cancer proliferation and how they contribute to colon cancer tumorigenesis. These questions can be addressed by additional biochemical experiments, where the application of multi-modal XCMS can improve the extraction of biological information from multiple complementary metabolomic data sets, therefore leaving more resources for additional biochemical experiments.

Conclusions

In summary, a multi-modal XCMS metabolomics data analysis approach has been developed for comprehensive pathway analysis using data acquired on multiple analytical platforms. This web-based tool automates metabolomic data integration and interpretation a and allows

for its combined analysis with proteomics and transcriptomics datasets. The approach was demonstrated in a colon cancer study to perform automatic data integration, metabolic pathway analysis, and multi-omic pathway interpretation. With the help of this data integration platform it is now possible to automatically perform comprehensive biological analysis and, perhaps more importantly, this approach allows for data processing of archived data sets obtained from different analytical platforms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

This research was partially funded by Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the US Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-AC02-05CH11231 (G.S.); and National Institutes of Health grants R01 GM114368-03, P30 MH062261-17, and P01 DA028555-09 (G.S.)

The authors thank the following for funding assistance: Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the US 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 (grants R01 GM114368 (G.S.) and PO1 A1043376-02S1 (G.S.)). We also thank Linh Truc Hoang for assisting the preparation of colon tissue samples for the metabolomics analysis. The results of colon cancer multi-omic data integration are part based upon data generated by the TCGA Research Network.

Reference

(1). Wen W; Li D; Li X; Gao Y; Li W; Li H; Liu J; Liu H; Chen W; Luo J Nature communications 2014, 5.

(2). Bogdanov M; Matson WR; Wang L; Matson T; Saunders-Pullman R; Bressman SS; Beal MF Brain 2008, 131, 389–396. [PubMed: 18222993]

(3). Johnson CH; Ivanisevic J; Siuzdak G Nature reviews Molecular cell biology 2016, 17, 451.

(4). Guijas C; Montenegro-Burke JR; Warth B; Spilker ME; Siuzdak G Nature Biotechnology 2018, 90, 3156–3164. .

(5). Ivanisevic J; Zhu Z-J; Plate L; Tautenhahn R; Chen S; O'Brien PJ; Johnson CH; Marletta MA; Patti GJ; Siuzdak G Analytical chemistry 2013, 85, 6876–6884. [PubMed: 23781873]

(6). Chen J; Zhou L; Zhang X; Lu X; Cao R; Xu C; Xu G Electrophoresis 2012, 33, 3361–3369. [PubMed: 23109122]

(7). Fei F; Bowdish DM; McCarry BE Analytical and bioanalytical chemistry 2014, 406, 3723–3733. [PubMed: 24714971]

(8). Contrepois K; Jiang L; Snyder M Molecular & Cellular Proteomics 2015, 14, 1684–1695. [PubMed: 25787789]

(9). Ni Y; Su M; Qiu Y; Chen M; Liu Y; Zhao A; Jia W FEBS letters 2007, 581, 707–711. [PubMed: 17274990]

(10). Yusufi FNK; Lakshmanan M; Ho YS; Loo BLW; Ariyaratne P; Yang Y; Ng SK; Tan TRM; Yeo HC; Lim HL Cell Systems 2017, 4, 530–542. e536. [PubMed: 28544881]

(11). Xia J; Sinelnikov IV; Han B; Wishart DS Nucleic acids research 2015, 43, W251–W257. [PubMed: 25897128]

(12). Kamburov A; Cavill R; Ebbels TM; Herwig R; Keun HC Bioinformatics 2011, 27, 2917–2918. [PubMed: 21893519]

(13). Kanehisa M; Goto S; Sato Y; Furumichi M; Tanabe M Nucleic acids research 2011, 40, D109–D114. [PubMed: 22080510]

(14). Tautenhahn R; Patti GJ; Rinehart D; Siuzdak G Analytical chemistry 2012, 84, 5035–5039. [PubMed: 22533540]

(15). Pluskal T; Castillo S; Villar-Briones A; Orešič M BMC bioinformatics 2010, 11, 395. [PubMed: 20650010]

(16). Domingo-Almenara X; Montenegro-Burke JR; Benton HP; Siuzdak G Analytical chemistry 2017, 90, 480–489. [PubMed: 29039932]

(17). Huan T; Forsberg EM; Rinehart D; Johnson CH; Ivanisevic J; Benton HP; Fang M; Aisporna A; Hilmers B; Poole FL Nature methods 2017, 14, 461. [PubMed: 28448069]

(18). Guijas C; Montenegro-Burke JR; Domingo-Almenara X; Palermo A; Warth B; Hermann G; Koellensperger G; Huan T; Uritboonthai W; Aisporna AE Analytical chemistry 2018.

(19). Shi Z; Wang J; Zhang B Nature methods 2013, 10, 597–598. [PubMed: 23807191]

(20). Muzny DM; Bainbridge MN; Chang K; Dinh HH; Drummond JA; Fowler G; Kovar CL; Lewis LR; Morgan MB; Newsham IF; Reid JG; Santibanez J; Shinbrot E; Trevino LR; Wu YQ; Wang M; Gunaratne P; Donehower LA; Creighton CJ; Wheeler DA, et al. Nature 2012, 487, 330–337. [PubMed: 22810696]

(21). Zhang B; Wang J; Wang X; Zhu J; Liu Q; Shi Z; Chambers MC; Zimmerman LJ; Shaddox KF; Kim S Nature 2014, 513, 382–387. [PubMed: 25043054]

(22). Smith CA; Want EJ; O'Maille G; Abagyan R; Siuzdak G Analytical chemistry 2006, 78, 779–787. [PubMed: 16448051]

(23). Caspi R; Billington R; Ferrer L; Foerster H; Fulcher CA; Keseler IM; Kothari A; Krummenacker M; Latendresse M; Mueller LA Nucleic acids research 2016, 44, D471–D480. [PubMed: 26527732]

(24). Li S; Park Y; Duraisingham S; Strobel FH; Khan N; Soltow QA; Jones DP; Pulendran B PLoS Comput Biol 2013, 9, e1003123. [PubMed: 23861661]

(25). Forsberg EM; Huan T; Rinehart D; Benton HP; Warth B; Hilmers B; Siuzdak G Nature protocols 2018, 13, 633–651. [PubMed: 29494574]

(26). Fabregat A; Sidiropoulos K; Garapati P; Gillespie M; Hausmann K; Haw R; Jassal B; Jupe S; Korninger F; McKay S Nucleic acids research 2015, 44, D481–D487. [PubMed: 26656494]

(27). Kutmon M; Riutta A; Nunes N; Hanspers K; Willighagen EL; Bohler A; Mélius J; Waagmeester A; Sinha SR; Miller R Nucleic acids research 2015, 44, D488–D494. [PubMed: 26481357]

(28). Ward PS; Thompson CB Cancer cell 2012, 21, 297–308. [PubMed: 22439925]

(29). Pavlova NN; Thompson CB Cell metabolism 2016, 23, 27–47. [PubMed: 26771115]

(30). Cairns RA; Mak TW Nature Reviews Cancer 2016, 16, 613–614.

(31). Tomczak K; Czerwińska P; Wiznerowicz M Contemporary oncology 2015, 19, A68. [PubMed: 25691825]

(32). Edwards NJ; Oberti M; Thangudu RR; Cai S; McGarvey PB; Jacob S; Madhavan S; Ketchum KA Journal of proteome research 2015, 14, 2707–2713. [PubMed: 25873244]

(33). Field D; Sansone S-A; Collis A; Booth T; Dukes P; Gregurick SK; Kennedy K; Kolar P; Kolker E; Maxon M Science 2009, 326, 234–236. [PubMed: 19815759]

(34). Piwowar HA; Day RS; Fridsma DB PloS one 2007, 2, e308. [PubMed: 17375194]

(35). Kaye J; Heeney C; Hawkins N; De Vries J; Boddington P Nature Reviews Genetics 2009, 10, 331–335.

(36). Savage N Cell 2017, 168, 551–554. [PubMed: 28187273]

(37). Majumder MA; Guerrini CJ; Bollinger JM; Cook-Deegan R; McGuire AL Genetics in Medicine 2017.

(38). Wang M; Carver JJ; Phelan VV; Sanchez LM; Garg N; Peng Y; Nguyen DD; Watrous J; Kapono CA; Luzzatto-Knaan T Nature biotechnology 2016, 34, 828.

(39). Wallace HM; Caslake R European journal of gastroenterology & hepatology 2001, 13, 1033–1039. [PubMed: 11564951]

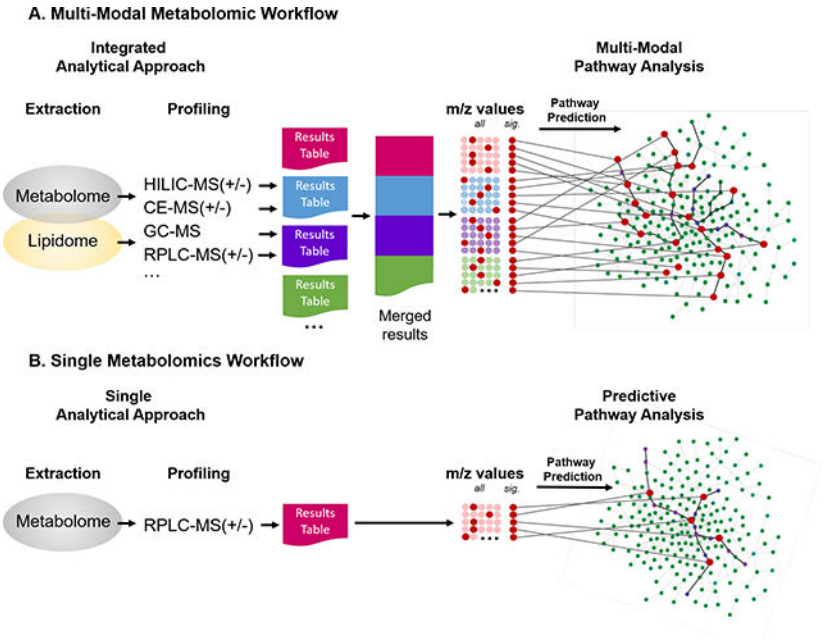


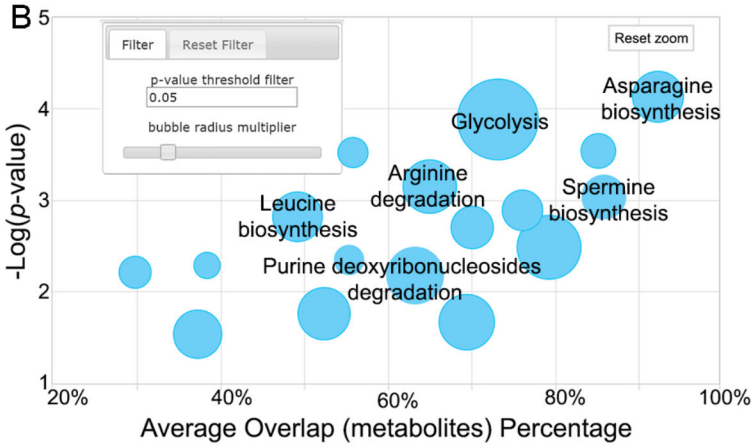
Figure 1.

(A) The schematic of a multi-modal metabolomics workflow where the data processing is integrated from multiple analytical approaches. This is compared to a single analytical approach (RPLC-MS) that is traditionally used for pathway mapping (B).

A

Pathway	Overlap Genes	Total Genes	Overlap Proteins	Total Proteins	Overlap Metabo.	Total Metabo.	Pathway p-values
gluconeogenesis	8	24	5	24	12	13	9.1e-4
sucrose degradation	3	8	1	8	8	8	1.2e-3
TCA cycle	10	17	5	12	10	11	1.6e-3

B



C

Metabolites	Pathway(s) Involved	p-value	Fold Change	Retention Time	Adducts	Source Job	Feature
(R)-lactate	anhydromurpeptides recycling mixed acid fermentation methylglyoxal degradation						
		1.1e-10	3.5	24.12	M+H[1+]	HILIC	5
Spermidine	spermine degradation						
		8.9e-8	4.8	1.5	M+HCOO[-]	RPLC	1124

Figure 2.

Multi-modal pathway analysis results. (A) Summary of pathway analysis results. The numbers of overlapping gene and protein show up after uploading dysregulated gene and protein data for multi-omic data integration (B) Pathway cloud plot. Each metabolic pathway is represented by a bubble. Metabolic pathways with higher statistical significance are located in the top right corner, showing low p-value and high metabolic overlapping. (C) Feature analysis results. Metabolic features matching the same metabolite according to their *m/z* values and possible adduct formations are all listed.

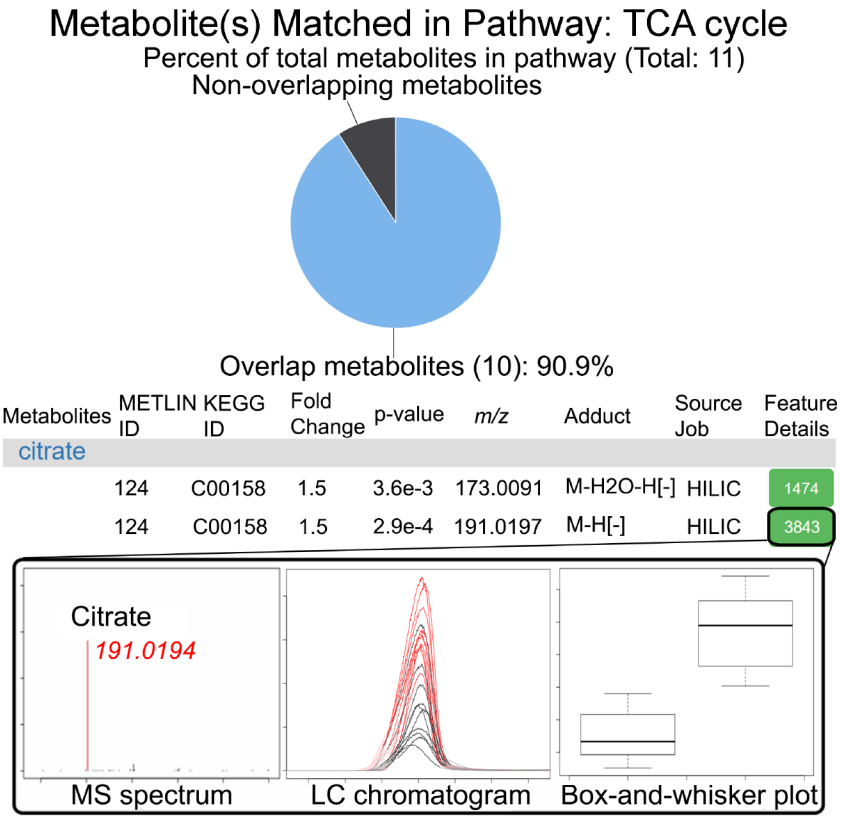


Figure 3.

Detailed metabolic feature information in each dysregulated pathway. Metabolic feature details for each dysregulated pathway can be accessed by clicking on the number of overlapping metabolites in the pathway analysis results table (Figure 2A). The pie chart on the top shows the number percentage of the overlapping and non-overlapping metabolites detected in all analyses. For each metabolic feature, the green feature ID button allows users to get detailed MS information including the LC chromatogram, MS spectrum, and box-and-whisker plot so that visual checking of the feature quality is available to assist the metabolite confirmation. If one dysregulated metabolite is detected in multiple analytical platforms, all the dysregulated metabolic features will be listed, along with their IDs of the associated analytical platforms.

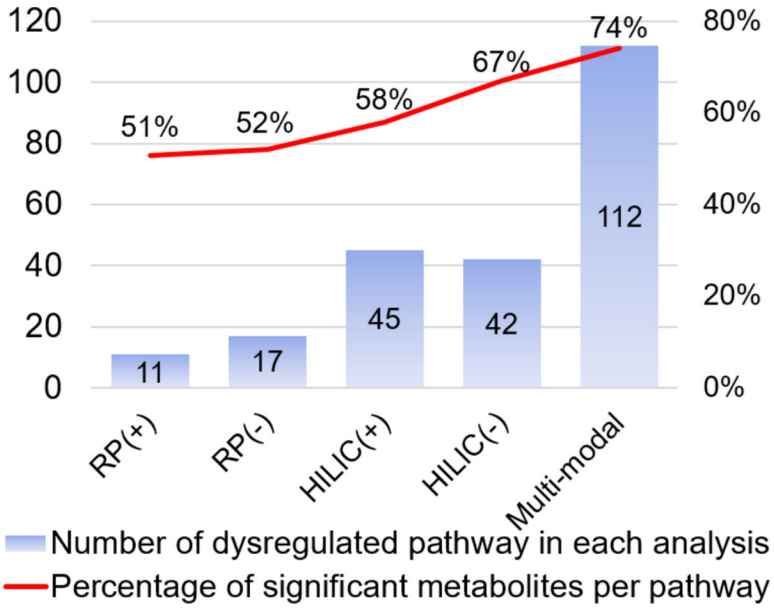


Figure 4.

Number of significantly dysregulated pathways and dysregulated metabolites per pathway from RP(+), RP(−), HILIC(+), HILIC(−) and multi-modal analyses. Blue columns represent the number of statistically significant pathways (p-value ≤ 0.05) observed in each metabolomic analysis. Red line shows the average percentage of significantly dysregulated metabolites involved in dysregulated pathways in each metabolomics analysis. The percentage value is determined by first calculating the percentage of dysregulated metabolites out of all the metabolites involved in each pathway and then averaging the percentages across all the dysregulated pathways.

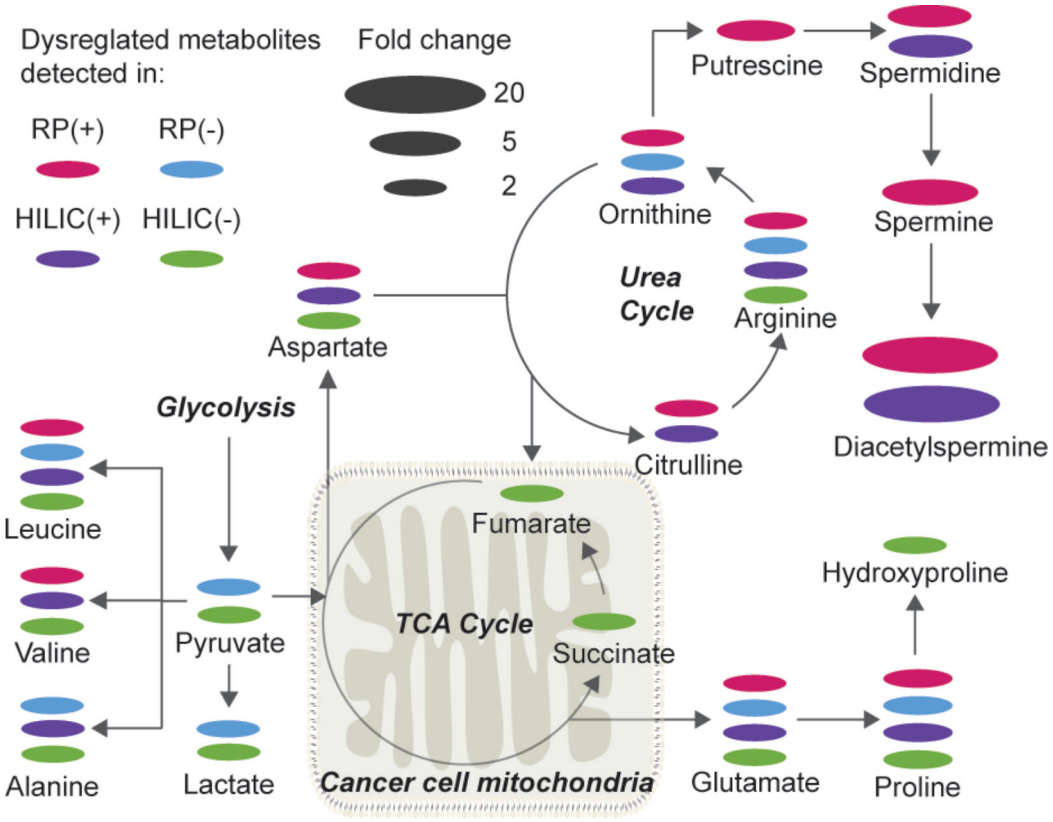


Figure 5.

Colon cancer-associated metabolic dysregulations illustrated by metabolic network developed from multi-modal metabolomics pathway analysis in multi-modal XCMS.

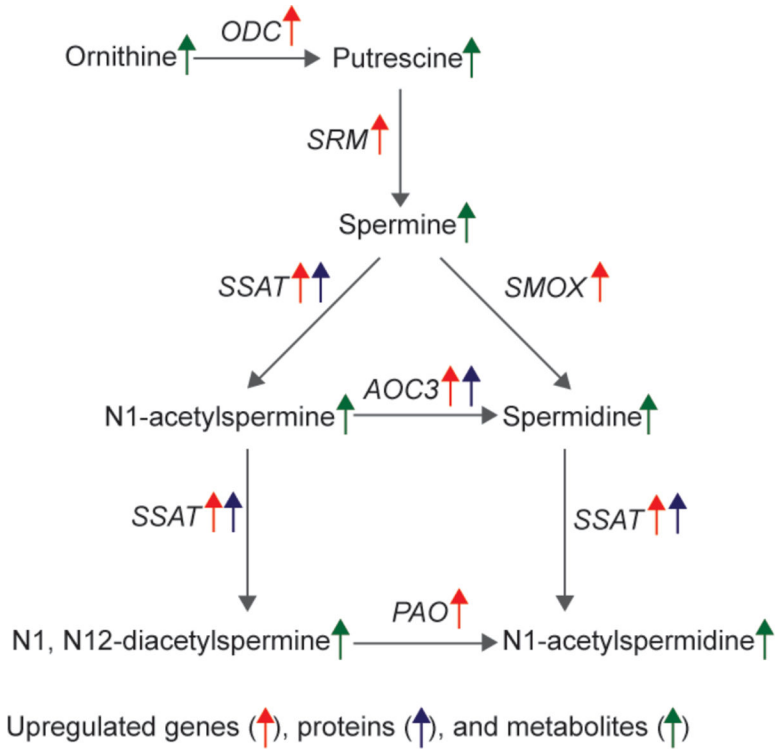


Figure 6.

Systems-level interpretation of the dysregulated spermine and spermidine metabolism pathway. ODC, ornithine decarboxylase; SRM, spermidine synthase; SSAT, spermidine/spermine N^1-acetyltransferase; SMOX, spermine oxidase; AOC3, membrane primary amine oxidase; PAO, polyamine oxidase.