Cloud-based archived metabolomics data: A resource for in-source fragmentation/annotation, meta-analysis and systems biology

Amelia Palermo[1], **Tao Huan**[1,2], **Duane Rinehart**[1], **Markus M. Rinschen**[1], **Shuzhao Li**[5], **Valerie B. O'Donnell**[6], **Eoin Fahy**[3], **Jingchuan Xue**[1], **Shankar Subramaniam**[3], **H. Paul Benton**[1,*], **Gary Siuzdak**[1,4,*]

[1]Scripps Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Rd., La Jolla, CA, 92037, USA

[2]Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, BC, V67 1z1, Canada

[3]Department of Bioengineering, University of California San Diego, 9500 Gilman Dr, La Jolla, CA, 92093-0412, USA

[4]Department of Chemistry, Molecular and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Rd., La Jolla, CA, 92037, USA

[5]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA

[6]Systems Immunity Research Institute, Cardiff University, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, UK

Abstract

Archived metabolomics data represent a broad resource for the scientific community. However, the absence of tools for the meta-analysis of heterogeneous data types makes it challenging to perform direct comparisons in a single and cohesive workflow. Here we present a framework for the meta-analysis of metabolic pathways and interpretation with proteomic and transcriptomic data. This framework facilitates the comparison of heterogeneous types of metabolomics data from online repositories (*e.g.*, XCMS Online, Metabolomics Workbench, GNPS, and MetaboLights) representing tens of thousands of studies, as well as locally acquired data. As a proof of concept, we apply the workflow for the meta-analysis of i) independent colon cancer studies, further interpreted with proteomics and transcriptomics data, ii) multimodal data from Alzheimer's disease and mild cognitive impairment studies, demonstrating its high-throughput capability for the systems level interpretation of metabolic pathways. Moreover, the platform has been modified for improved knowledge dissemination through a collaboration with Metabolomics Workbench and LIPID MAPS. We envision that this meta-analysis tool will help overcome the primary bottleneck in analyzing diverse datasets and facilitate the full exploitation of archival

---

*Authors to whom correspondence should be addressed: H. Paul Benton, Tel: (858) 784-9415, hpbenton@scripps.edu, Gary Siuzdak, Tel: (858) 784-9113, siuzdak@scripps.edu, webpage: https://masspec.scripps.edu/.

metabolomics data for addressing a broad array of questions in metabolism research and systems biology.

Keywords

Archived data; meta-analysis; metabolic pathways; metabolomics; proteomics; transcriptomics; systems biology

Introduction

Metabolites are the prime drivers of biological activity as they regulate enzyme reactions[1], protein activation and gene/protein expression[2]. Ultimately, metabolites provide an accessible functional readout for the activity of the system and in themselves modulate the phenotype[3]. In line with this, the meta-analysis of untargeted high-resolution mass spectrometry (MS) metabolomic data obtained from distinct studies can be used to obtain a better understanding of the altered metabolic processes and active endogenous metabolites affecting the system over a broad population of samples. This type of analysis requires the generation and/or recollection of multiple metabolomic data sets across several independent studies, to provide a more comprehensive picture than an individual study. In some cases, the data sets required for the meta-analysis have already been generated and made available on public databases. In this regard, several data storage infrastructures have been recently developed to address the raising call for metabolomics data sharing and currently encompass more than 500 untargeted high-resolution data sets. Emerging open-access ecosystems include MetaboLights[4], MetabolomicsWorkbench[5], Metabolonote[6], Global Natural Products Social Molecular Networking (GNPS)[7] and metabolomic data aggregation services, such as metabolomeXchange[8] (http://www.metabolomexchange.org/site/) and Omics Discovery Index (http://www.omicsDI.org)[9]. In addition, the LIPID MAPS service provides a link into MetabolomicsWorkbench to support the direct deposition of lipidomics data (www.lipidmaps.org).[10,11]

These publicly available datasets can reduce the workload for data re-collection as well as foster transparency and collaboration between researchers. However, owing to the absence of tools for their cohesive meta-analysis and to the heterogeneity of the stored data, that are often obtained by different types of mass spectrometry-based metabolome profiling workflows, each study remains only partially utilized for comparative analyses.

Currently, the meta-analysis of metabolomic pathways is carried out by comparing and analyzing the results reported in published papers (*e.g.*, fold change comparison, absolute concentrations from targeted studies), thus ignoring the total content of information on metabolites contained in the raw profiling data. Moreover, the interpretation of the meta-analysis findings in the context of proteomic and transcriptomic dysregulations remains a manual task as no systems level data interpretation tool currently provides this functionality. For example, depending on data type, there are many tools for integration of multi-omics data available including correlation analysis, multivariate comparison, regression/machine learning for sample classification.

Here we report MetaXCMS framework, to enable the meta-analysis of heterogeneous types of archived untargeted, high-resolution MS data across metabolomics, proteomics, transcriptomics, and genomics. The XCMS Online metabolomics platform[12,13] is an environment for the direct re-analysis and comparison of data from transcriptomics and metabolomics repositories and/or acquired locally, to gather insights into the dysregulated active metabolites and pathways over independent studies and populations of subjects/samples. We deploy this workflow by integrating and interpreting at systems level archival data sets from two independent colon cancer studies obtained from the XCMS Online Public repository.[14] In addition, we tested this framework in the meta-analysis of archival multimodal metabolomics data acquired from plasma samples from patients with Alzheimer's disease, mild cognitive impairment and cognitive normal patients, from the MetabolomicsWorkbench.[15]

Results

Workflow for the meta-analysis of archival metabolomic data and systems level integration.

We developed Meta XCMS framework for the meta-analysis and interpretation of archival metabolomics data by developing and combining different bioinformatic modules to be facilitated with the XCMS Online platform (Figure 1).

In the meta-analysis workflow, the raw MS data sets from individual studies can be uploaded in XCMS Online to perform data processing and analysis, including peak detection, retention time alignment, putative annotation and statistical significance testing, to a final list of detected and dysregulated metabolic features.[13] At this level, the user can set the processing and statistical parameters depending on the analytical platform employed for metabolome profiling and on the statistical test needed for that study. The processed jobs can then be selected and downloaded to be inputted into the Meta XCMS framework for further analysis. Multiple metabolomic analysis (analytical modalities) for a given study can be combined together for comprehensive coverage, for example, on lipid and central carbon metabolism *(e.g.*, combining data obtained by reversed phase chromatography coupled with positive electrospray ionization (ESI) MS, and hydrophilic interaction liquid chromatography in negative ESI-MS, *etc.*).[16,17] Moreover, metabolomics data sets obtained from high-resolution untargeted studies archived in the Metabolomics Workbench can be directly uploaded for meta-analysis, and the user interface also supports the upload of metabolomics data-sets in text/tsv format, obtained through alternative data pre-processing workflows.

The Meta XCMS framework code is based on mummichog version 2.0.7[17,18] and leverages itself on this open source platform. Briefly, the mummichog algorithm performs a Fisher's exact test on the number of metabolites jointly dysregulated in the studies as opposed to the total metabolites in the pathway, to predict active pathways directly from putative metabolic features. To allow for a multi-file input the code and algorithm was adjusted. Using either data from the Metabolomics Workbench or XCMS Online data is read into the system via convenient tsv/csv peak list file formats. Data read into the system is first parsed, each feature is tagged to its input file to allow for tracing throughout the system.

Several possible adducts combinations are calculated onto the feature masses. These features are queried against the pathway database, each file is processed separately and ranked on their corresponding p-value score to statistically eliminate false positives. Once a list of possible hits is obtained, features that match the neutral mass in the pathway, they are merged between the different files. This is done such that any compound that is seen in more than one file is merged together and the best p-value score is taken. This method allows for the expanded coverage and keeps the statistical validation. Next, the regular *mummichog* process continues with the statistical validation of the matched pathways. Finally, the output is processed to simply future analyses.

For each individual study included in the meta-analysis, the user can set a specific significance threshold (p-value), *m/z* tolerance for metabolite putative annotation, and filter the metabolic feature list according to a specified intensity threshold. We recommend to carefully choose these parameters considering the size of each individual study and the type of metabolomics platform used.

Notably, lipids comprise around a third of all metabolites, but they require distinct processing approaches for accurate annotation and pathway prediction. In particular, removing spurious MS signals is critical to improve statistical power, especially for lipids where multiple forms of the same molecule can be detected/exist. LipidFinder[19] has been recently developed at LIPID MAPS[11] to alleviate these artefacts. Here, we suggest using LipidFinder post processing of XCMS outputs. This helps to further the broaden the output of lists of putative structures and their categories for more accurate meta-analysis of lipidomics data.

As part of the Meta XCMS framework we suggest performing multi-omic data integration by superimposing user-uploaded transcriptomic and proteomic data sets onto the dysregulated pathways. Using a list of dysregulated genes (as gene symbols or loci) and proteins (as UniProt accession IDs or gene symbols) obtained from studies targeting a given biological question users can generate improved confidence on the pathway hits. The integration with proteomics and transcriptomics results offers the possibility to gauge a systems level mechanistic understanding of pathways dysregulation and metabolite activity in the investigated biological system.[20]

The downloaded "metabolite results" table reports the list of all the overlapping dysregulated metabolites detected in the studies included in the meta-analysis and used for pathway prediction.

The "pathway results" table showcases the output for the metabolic pathways jointly dysregulated in the studies. For each metabolic pathway, it reports the number of overlapping dysregulated metabolites detected in different studies with respect to the total number of metabolites in the pathway. By clicking on the number of "shared metabolites" the complete list of metabolic features is shown. Entries can be further filtered based on the adduct type, study group or pairwise job group. Moreover, this table reports the overlapping dysregulated genes and proteins from the uploaded proteomics and transcriptomics data for

each metabolic pathway, thus providing a rapid glance on the biological process from a system-wide perspective.

Expanding the capability of XCMS Online for the meta-analysis of archived data.

To allow the meta-analysis of metabolomics data sets obtained from disparate sources, we enabled easy parsing file options of tsv/csv in Meta XCMS framework (Figure 1 and Figure 2). First, the user can select studies processed in the XCMS Online private space or in the XCMS Online Public. These data are then easily downloaded to be parsed into the Meta-XCMS framework. Alternatively, the Metabolomics Workbench data can also be used. On many of the studies there are already existing outputs of identified metabolites or feature lists. In these instances where a metabolite is already identified it will be read into the system and used as a confirmed metabolite of the pathway analysis.

This strategy is aimed at fostering data dissemination and at actively promoting the full exploitation of archived metabolomics data through stimulating further meta-analysis for results validation or for generating novel hypothesis.

Analysis of archived metabolomics data.

We tested the workflow in the systems level meta-analysis of two independent colon cancer tissue metabolomics data sets by leveraging archival data from the XCMS Online Public repository, and for the meta-analysis of Alzheimer's disease and mild cognitive impairment studies in plasma, including heterogeneous profiling data from the Metabolomics Workbench.

Colon Cancer.—Several previous studies have pinpointed the multifaced metabolic reprogramming underlining colon cancer.[14,24,25] Here we performed the meta-analysis of archived untargeted metabolomics data from a study investigating the role of bacterial biofilms in colon cancer[14] (Study A) and a second colon cancer study, recently performed in our laboratory (Study B) (Figure 3A). Study A involved 30 subjects diagnosed with colon cancer from stage 3 to 4 (18 females and 12 males, 61–88 years old) and was available in the XCMS Online Public repository[14], while Study B involved 19 subjects diagnosed with colon cancer (13 females and 6 males, 62–92 years old). More details on study design and samples collection are available in the Supporting Information, Table S1. Both studies used similar platforms for metabolome profiling (reversed phase chromatography coupled with ESI(+)-quadrupole time-of-flight (Q-TOF) mass spectrometry) and sample preparation protocols, therefore we expected comparable metabolome coverage and overlapping dysregulations (excluding inter-population heterogeneity).

First, raw data from study B were uploaded in XCMS Online and processed as a paired job. This job and the archival job (Study A) were then selected as input for meta-analysis in meta XCMS framework (Figure 3A). The results unveiled the presence of 30 metabolic pathways with at least ten dysregulated putative metabolic features across study A and B (Table S2, Figure 4A), among these glycerophospholipids metabolism, aspartate and asparagine metabolism, glycine, serine and alanine metabolism, carnitine shuttle, tyrosine metabolism, steroidal hormones and bile acids. The dysregulation of the glycerophospholipid metabolism

has been previously confirmed correlating with altered viability, proliferation and colorectal cancer development.[26] The meta-analysis also highlighted the dysregulation of the aspartate and asparagine pathway which includes spermine/spermidine biosynthesis and degradation (polyamine pathway) where N1-acetylspermidine, N1-acetylspermine, spermidine and N1,N12-diacetylsperimine, spermidine dialdehyde, spermic acid were found jointly up-regulated, a finding consistent with previous work (Figure 4C).[14,30] Of note, in the bile acids biosynthesis pathway taurine and taurochenodeoxycholate were upregulated (Figure 4B). Increased levels of conjugated bile acids have been previously reported to highly associate with colon cancer.[27] In particular, taurochenodeoxycholate can be hydrolyzed releasing taurine, a sulfur amino acid further transformed by the gut microbiota to form compounds with genotoxic activity (e.g., H2S), and colon tumor promoters (deoxycholic acid).[28,29]

To Interpret this evidence in light of the variations occurring at proteomic and transcriptomic level, colon cancer data sets obtained from The Cancer Genome Atlas[24] and The CPTAC Proteomics Data Portal[32] were uploaded and processed in meta XCMS framework. Approximately 90% of the up-regulated metabolic pathways were further supported by dysregulated proteins and gene transcripts (Table S3). For instance, both polyamines and bile acids pathway dysregulations were confirmed (Figure 3A).

Alzheimer's disease.—Alzheimer's disease (AD) is a progressive neurodegenerative disorder of unknown etiology.[33] AD and dementia patients are usually subject to a long pre-AD period known as mild cognitive impairment (MCI). Here we used public available metabolomics data obtained from previous longitudinal studies performed at the Mayo Clinic Study of Aging (MCSA) and Mayo Clinic Alzheimer Disease Research Center (ADRC).[15] Plasma samples were from AD, MCI and cognitive normal (CN) subjects (15 individuals/group). The metabolomics data sets and meta-data were publicly available in the Metabolomics Workbench and formerly generated by LC-Q-TOF MS in four analytical modalities (hydrophilic interaction liquid chromatography (HILIC) and reversed phase liquid chromatography (RP), in both positive and negative ESI modes) for comprehensive metabolome coverage. We downloaded the raw data sets from the Metabolomics Workbench repository and uploaded them in the XCMS Online for processing and statistical analysis to extract significant metabolic variations in AD *vs* CN and MCI *vs* CN. The resulting XCMS Online jobs were then used as input for the Meta XCMS framework to detect shared metabolic changes at different disease stages.

Meta XCMS framework predicted 24 dysregulated metabolic pathways with at least ten metabolic features dysregulated in the AD vs CN and the MCI vs CN groups, over a total of 101 paths (Table S4). We manually compared the output with the dysregulated pathways reported in the original publication.[15] Our development predicted the dysregulation of tyrosine, glycerophospholipid, aspartate and asparagine, glycine, serine and alanine metabolism, urea cycle, tryptophan, purine metabolism, together with the other pathways reported in Table S4. The original work reported 50 total dysregulated pathways, of which nine were consistently predicted across AD vs CN and MCI vs CN (Figure 5). Our approach predicted a total of 101 dysregulated pathways, of which 24 pathways were reported in the original publication, demonstrating the efficiency of the workflow in identifying jointly dysregulated metabolic pathways from heterogeneous archived metabolomic data.

Discussion

Archived metabolomics data are a rich source of information for second-order analysis by the scientific community. However, the heterogeneity of the data and the lack of tools for their cohesive re-analysis and interpretation hinders their full utilization. To address this, we developed a framework for archival data re-processing, analysis, integration and interpretation at systems level. The workflow moves from the metabolomics tools available in the XCMS Online, further combining them with a bioinformatic development specifically designed for the meta-analysis of heterogeneous metabolomics data.

A key aspect of the workflow is the use of a pathway-centric approach to the meta-analysis, which allows the direct prediction of the dysregulated metabolic pathways from putative metabolic features jointly detected in the archived data / studies. This is performed through the embedment of a recently developed tool for metabolic pathway prediction from putative annotations of metabolic features extracted from different types of metabolomics data sets.[17,18] This tool allows for higher confidence in the putative pathway enrichment results by estimating the probability of a pathway being dysregulated on the basis of the total number of dysregulated metabolites detected. It is also worth noting that, when attempting to re-analyze archival metabolomics data, the physical samples are not directly accessible and often no longer available. This makes it unfeasible to perform further mass spectrometry fragmentation experiments and metabolite identity confirmation. In this scenario, performing direct pathway prediction analysis represents a practical strategy to bypass this limitation and directly formulate biological hypothesis from the archived data, to be later tested through independent targeted studies or biochemical assays.

The use of a pathway-centric approach also streamlines the interpretation of the metabolomic data by superimposing the dysregulated proteins and transcripts to each metabolic pathway. This provides a rapid glance on the system in the light of other omics regulatory levels, introducing the possibility for the orthogonal confirmation of the insights extracted from the archived metabolomics data.

Despite recent efforts aimed at standardizing metabolome profiling and reporting,[21,35,36] a widely-adopted consensus on untargeted MS-based workflows in the perspective of meta-analysis is still missing.[37,38] The metabolic profiles are indeed usually acquired by a variety of different analytical solutions,[16,39–41] thus introducing heterogeneity in the type of available data. For example, Metabolomics Workbench[5] currently stores > 190 untargeted high-resolution MS studies, for a total of ~ 400 different analyses (i.e., different analytical modalities including ESI positive and negative acquisition modes), while Metabolights[4] stores ~350 among GC- and LC-MS based studies. This heterogeneity complicates the development of bioinformatics solutions for the automated meta-analysis, as each different metabolomics platform calls for specific data processing and metabolite annotation pipelines. To circumvent this limitation, we designed the workflow in a modular fashion: each study can be processed as pairwise XCMS Online job using different processing and statistical settings and the resulting jobs can be used as input for further meta-analysis. This allows the use of raw data acquired by different metabolomics platforms and modalities. For example, the user can upload both lipidomics and metabolomics data obtained by

different chromatographic or ionization modes for comprehensive metabolome coverage and improved pathway prediction.[17] This, together with the ability to perform multi-omics data integration, represents a fundamental advantage over our previous development for the meta-analysis of metabolomics data.[42,43]

Besides raw data sets, the workflow also supports the direct comparison of untargeted studies already processed and available in the XCMS Online Public cloud and in the Metabolomics Workbench. Of note, several data sets currently archived in public databases are not compliant with the ISA guidelines for meta-data reporting, unearthing the need of harmonized and more pragmatic guidelines for metabolomics data sharing in public repositories.[37]

The meta-analysis of metabolomics data sets and their interpretation at systems level has the potential of streamlining different types of study comparisons. For example, a meta-analytical approach can be used for i) providing further validation of metabolite dysregulations in the context of independent set of samples (*e.g.*, in biomarker studies); ii) stimulating the generation of biological hypothesis from the re-analysis of archived untargeted studies; iii) streamlining the exclusion of experimental artifacts to reduce the list of dysregulated metabolites before performing time consuming structural elucidation[42]; iv) excluding metabolic dysregulations due to physiologic heterogeneity in different populations of subjects, therefore taking a step towards the identification of therapeutic targets and biomarkers of broad applicability. In particular, in biomarker discovery the automatic integration of multiple archival metabolomics studies can be a cost-effective strategy to minimize the inter-study bias introduced by genetic and environmental factors.[44–46] This strategy is not limited to archived data, since the difficulty in cross-laboratory comparison has impeded the biomedical applications of metabolomics. In an emergency situation like the current COVID-19 pandemic, the pathway-centric meta-analysis can be important for identifying scientific consensus in a timely manner.

As proof of concept, we demonstrated the utility of the workflow in two meta-analytical studies. First, we leveraged archival data sets from a previous study available in the XCMS Online Public,[14] for the autonomous comparison with a colon cancer study recently performed in our laboratory. The workflow allowed a rapid glance on metabolic pathways jointly dysregulated and validation at systems level (*e.g.*, the bile acid and the polyamine pathways). In the second example, we applied the workflow for the re-analysis of archival biomarker studies obtained from the Metabolomics Workbench database. The workflow permits the streamlined and autonomous prediction of metabolic pathways changed in both AD and MCI patients in plasma (pre-AD) in agreement and beyond the results previously obtained by manual meta-analysis.[15]

In summary, there are many challenges in the analysis of diverse datasets including variability in experimental designs as well as information types that are largely platform dependent. However, by combining a fully automated workflow with an enhanced strategy for data storage and direct connection to the Metabolomics Workbench data repository, the described approach can provide a solution for meta-analysis, with the ultimate goal of maximizing the usage and dissemination of information-rich archival metabolomics data.

With the growing number of metabolomics, proteomics and genomics data generated to cover a wide range of biological questions, this workflow paves the way to unlock biological insights in the era of "big data" and "open science".

Materials and Methods

Meta analysis framework.

The system has been built on a local based Flask system with code based on mummichog version 2.0.7 running on python 2.7.[18] Several files were altered to allow for a multi-file input and hosting on a web frontend. Metabolomics Workbench data is processed using either csv/tsv formats directly into the software or via XCMS processing to csv/tsv files. Once the data is read each feature is processed against all possible adducts for search masses. Using the mummichog algorithm, the search masses are searched against the pathway database, each file is processed separately and ranked on their corresponding p-value score. Once a list of possible hits is obtained, they are merged between the different files such that any compound that is seen in more than one file is merged together and the best p-value score is taken. Now, the regular mummichog process continues with the statistical validation of the altered pathways. Finally, the output is processed to make further analysis simpler and result are downloaded by the user. Framework code has been made available via github at - https://github.com/hpbenton/archive-mummi

Metabolomics profiling data.

The colon cancer study A was available in the XCMS Online Public Space as a processed job. The raw MS data from high-resolution metabolome profiling for the colon cancer study B were archived in our laboratory, and previously obtained as part of a pilot colon cancer study (Study B). Both profiling studies were performed in RP-ESI(+)-Q-TOF profiling. More details on study A and B experimental setup can be found in previous published work.[14,17] Metabolome profiles for AD, MCI, and CN plasma samples were downloaded from the Metabolomics Workbench repository,[5] uploaded and re-processed in the XCMS Online. These studies were performed at the Mayo Clinic Study of Aging ((MCSA) and Mayo Clinic Alzheimer Disease Research Center (ADRC) and more details on study design and experimental procedures are available in previous published work.[15]

Data processing and re-analysis.

Raw archival data sets were uploaded as *.mzXML* files in the XCMS Online and processed as pairwise jobs. Before processing, the profiling data were manually examined for assessing the quality and the parameters for further processing and analysis. Colon cancer study A was already processed and available in the XCMS Online Public. The XCMS jobs were used for further meta-analysis in the Meta XCMS framework. P-value, intensity and ppm error settings are reported in the Supporting Information, Text S1.

Proteomics and Transcriptomics data.

Transcriptomics data were obtained from The Cancer Genome Atlas (TCGA)[31] in the frame of a previous colon cancer study involving 22 subjects (22 colon cancer tissue samples *vs* 22 paired normal tissues). Dysregulated genes were selected based on a p-value cut off of 0.01

and fold change > 4. A total of 7138 dysregulated transcripts were included in the final data set for upload in the XCMS Online as gene symbols. Proteomics data were obtained by the Clinical Proteomic Tumor Analysis Consortium (CPTAC), involving 90 patients affected by colon cancer (90 colon cancer tissue samples and 90 paired normal tissues). Dysregulated proteins were filtrated by p-value < 0.01 and fold change >2, obtaining a total of 2545 dysregulated proteins uploaded in the XCMS Online as UniProt accession IDs for multiomic analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Hackett SR et al. Systems-level analysis of mechanisms regulating yeast metabolic flux. Science (80-.). (2016). doi:10.1126/science.aaf2786

2. Johnson CH, Ivanisevic J & Siuzdak G Metabolomics: Beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology (2016). doi:10.1038/nrm.2016.25

3. Joyce AR & Palsson BO The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol 7, 198–210 (2006). [PubMed: 16496022]

4. Haug K et al. MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. (2013). doi:10.1093/nar/gks1004

5. Sud M et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. (2016). doi:10.1093/nar/gkv1042

6. Ara T et al. Metabolonote: A Wiki-Based Database for Managing Hierarchical Metadata of Metabolome Analyses. Front. Bioeng. Biotechnol (2015). doi:10.3389/fbioe.2015.00038

7. Wang M et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology (2016). doi:10.1038/nbt.3597

8. Haug K, Salek RM & Steinbeck C Global open data management in metabolomics. Current Opinion in Chemical Biology (2017). doi:10.1016/j.cbpa.2016.12.024

9. Perez-Riverol Y et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nature Biotechnology (2017). doi:10.1038/nbt.3790

10. O'Donnell VB, Dennis EA, Wakelam MJO & Subramaniam S LIPID MAPS: Serving the next generation of lipid researchers with tools, resources, data, and training. Sci. Signal (2019). doi:10.1126/scisignal.aaw2964

11. Fahy E et al. LipidFinder on LIPID MAPS: Peak filtering, MS searching and statistical analysis for lipidomics. Bioinformatics (2019). doi:10.1093/bioinformatics/bty679

12. Tautenhahn R, Patti GJ, Rinehart D & Siuzdak G XCMS online: A web-based platform to process untargeted metabolomic data. Anal. Chem (2012). doi:10.1021/ac300698c

13. Forsberg EM et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat. Protoc (2018). doi:10.1038/nprot.2017.151

14. Johnson CH et al. Metabolism links bacterial biofilms and colon carcinogenesis. Cell Metab. 21, 891–897 (2015). [PubMed: 25959674]

15. Trushina E, Dutta T, Persson XMT, Mielke MM & Petersen RC Identification of Altered Metabolic Pathways in Plasma and CSF in Mild Cognitive Impairment and Alzheimer's Disease Using Metabolomics. PLoS One (2013). doi:10.1371/journal.pone.0063644

16. Ivanisevic J et al. Toward Omic scale metabolite profiling: a dual separation mass spectrometry approach for coverage of lipids and central carbon metabolism. Anal. Chem (2013). doi:10.1021/ac401140h

17. Huan T et al. Autonomous Multimodal Metabolomics Data Integration for Comprehensive Pathway Analysis and Systems Biology. Anal. Chem 90, 8396–8403 (2018). [PubMed: 29893550]

18. Li S et al. Predicting network activity from high throughput metabolomics. PLoS Comput Biol (2013). doi:10.1371/journal.pcbi.1003123

19. O'Connor A et al. LipidFinder: A computational workflow for discovery of lipids identifies eicosanoid-phosphoinositides in platelets. JCI Insight (2017). doi:10.1172/jci.insight.91634

20. Guijas C, Montenegro-Burke JR, Warth B, Spilker ME & Siuzdak G Metabolomics activity screening for identifying metabolites that modulate phenotype. Nature Biotechnology (2018). doi:10.1038/nbt.4101

21. Rocca-Serra P et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics (2016). doi:10.1007/s11306-015-0879-3

22. Côté R et al. The Ontology Lookup Service: Bigger and better. Nucleic Acids Res. (2010). doi:10.1093/nar/gkq331

23. Larralde M et al. mzML2ISA & nmrML2ISA: generating enriched ISA-Tab metadata files from metabolomics XML data. Bioinformatics (2017). doi:10.1093/bioinformatics/btx169

24. Brown DG et al. Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. Cancer Metab. (2016). doi:10.1186/s40170-016-0151-y

25. Denkert C et al. Metabolite profiling of human colon carcinoma-deregulation of TCA cycle and amino acid turnover. Mol. Cancer (2008). doi:10.1186/1476-4598-7-72

26. Yan G, Li L, Zhu B & Li Y Lipidome in colorectal cancer. Oncotarget (2016). doi:10.18632/oncotarget.7960

27. Degirolamo C, Modica S, Palasciano G & Moschetta A Bile acids and colon cancer: Solving the puzzle with nuclear receptors. Trends in Molecular Medicine (2011). doi:10.1016/j.molmed.2011.05.010

28. Ridlon JM, Wolf PG & Gaskins HR Taurocholic acid metabolism by gut microbes and colon cancer. Gut Microbes (2016). doi:10.1080/19490976.2016.1150414

29. Devkota S et al. Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10–/–mice. Nature (2012). doi:10.1038/nature11225

30. Wallace HM & Caslake R Polyamines and colon cancer. European Journal of Gastroenterology and Hepatology (2001). doi:10.1097/00042737-200109000-00006

31. Tomczak K, Czerwi ska P & Wiznerowicz M The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkologia (2015). doi:10.5114/wo.2014.47136

32. Edwards NJ et al. The CPTAC data portal: A resource for cancer proteomics research. J. Proteome Res (2015). doi:10.1021/pr501254j

33. Selkoe DJ Alzheimer's disease: genes, proteins, and therapy. Physiol. Rev (2001). doi:10.1152/physrev.2001.81.2.741

34. Caspi R et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. (2018). doi:10.1093/nar/gkx935

35. Fiehn O et al. The metabolomics standards initiative (MSI). Metabolomics (2007). doi:10.1007/s11306-007-0070-6

36. Peng J, Chen YT, Chen CL & Li L Development of a universal metabolome-standard method for long-term LC-MS metabolome profiling and its application for bladder cancer urine-metabolite-biomarker discovery. Anal. Chem (2014). doi:10.1021/ac5011684

37. Spicer RA, Salek R & Steinbeck C Comment: A decade after the metabolomics standards initiative it's time for a revision. Scientific Data (2017). doi:10.1038/sdata.2017.138

38. Spicer RA, Salek R & Steinbeck C Compliance with minimum information guidelines in public metabolomics repositories. Sci. Data (2017). doi:10.1038/sdata.2017.137

39. Martin JC et al. Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. Metabolomics (2015). doi:10.1007/s11306-014-0740-0

40. Cajka T, Smilowitz JT & Fiehn O Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography-High-Resolution Mass Spectrometry Platforms. Anal. Chem (2017). doi:10.1021/acs.analchem.7b03404

41. Telu KH, Yan X, Wallace WE, Stein SE & Simõn-Manso Y Analysis of human plasma metabolites across different liquid chromatography/mass spectrometry platforms: Cross-platform transferable chemical signatures. Rapid Commun. Mass Spectrom (2016). doi:10.1002/rcm.7475

42. Patti GJ, Tautenhahn R & Siuzdak G Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nature Protocols (2012). doi:10.1038/nprot.2011.454

43. Cajka T & Fiehn O Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. Trends Analyt. Chem (2014). doi:10.1016/j.trac.2014.04.017

44. Drucker E & Krapfenbauer K Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA Journal (2013). doi:10.1186/1878-5085-4-7

45. Mamas M, Dunn WB, Neyses L & Goodacre R The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Arch. Toxicol (2011). doi:10.1007/s00204-010-0609-6

46. Xia J, Broadhurst DI, Wilson M & Wishart DS Translational biomarker discovery in clinical metabolomics: An introductory tutorial. Metabolomics (2013). doi:10.1007/s11306-012-0482-9
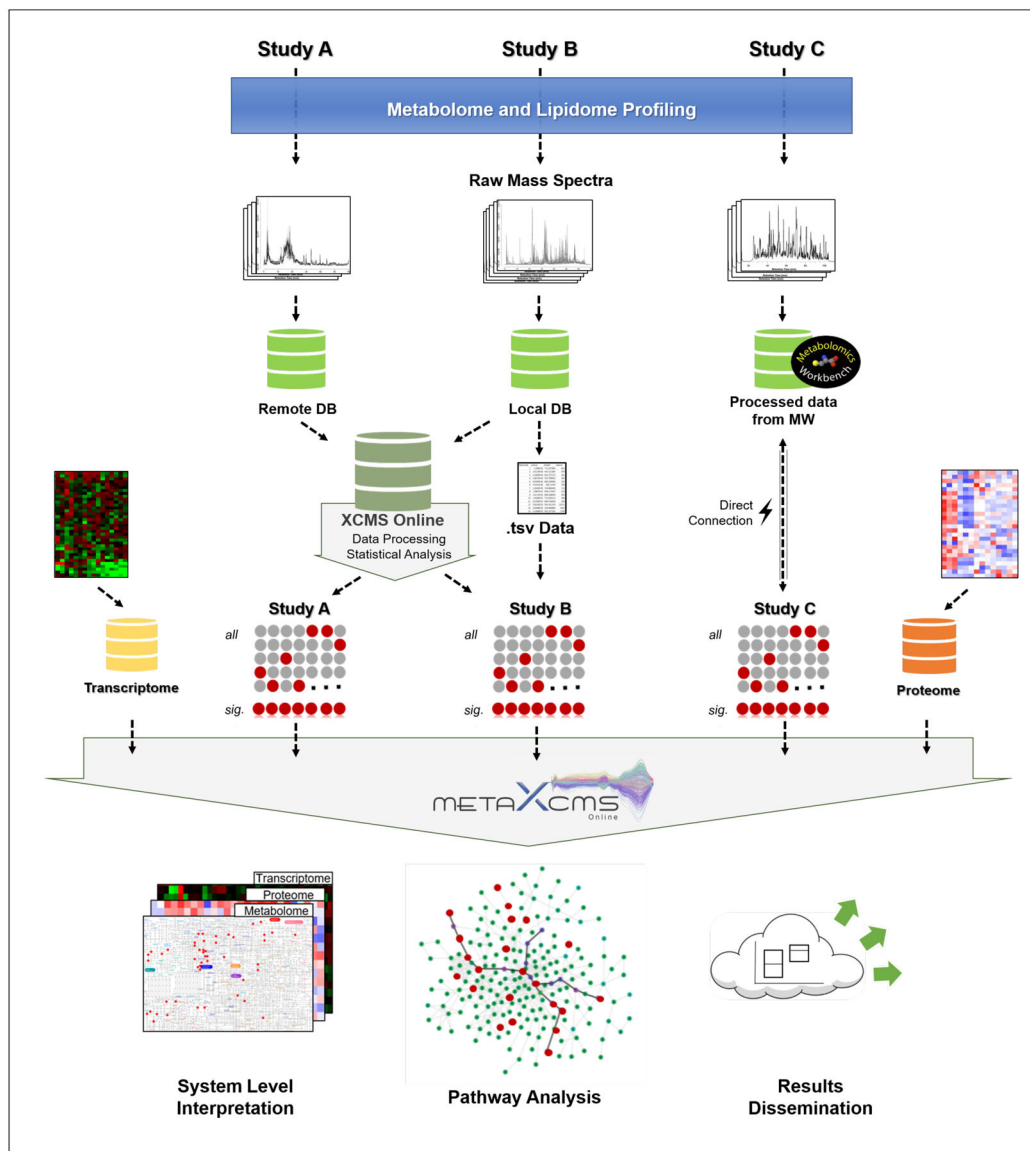
Figure 1.
Workflow for the meta-analysis of heterogenous metabolomics data in XCMS Online: metabolomics data sets from public repositories are uploaded in the XCMS Online database and processed for metabolic features detection and statistical analysis. The jobs are then used as input for meta-analysis and integration with proteomics and transcriptomics data in Meta XCMS framework. Results can be shared in the XCMS Online cloud for knowledge dissemination.
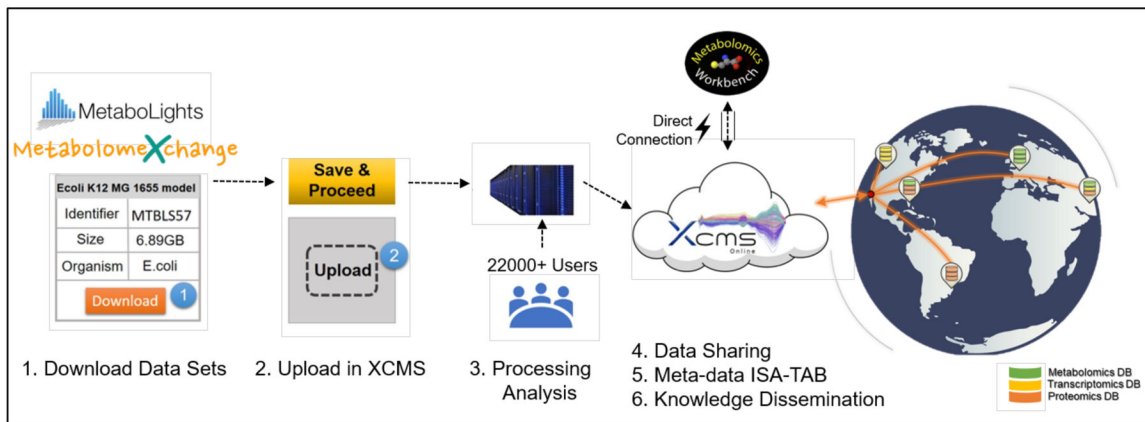
Figure 2.
Developments of the meta XCMS framework to enhance archival metabolomics data processing, archiving and sharing for meta-analysis and systems level interpretation.
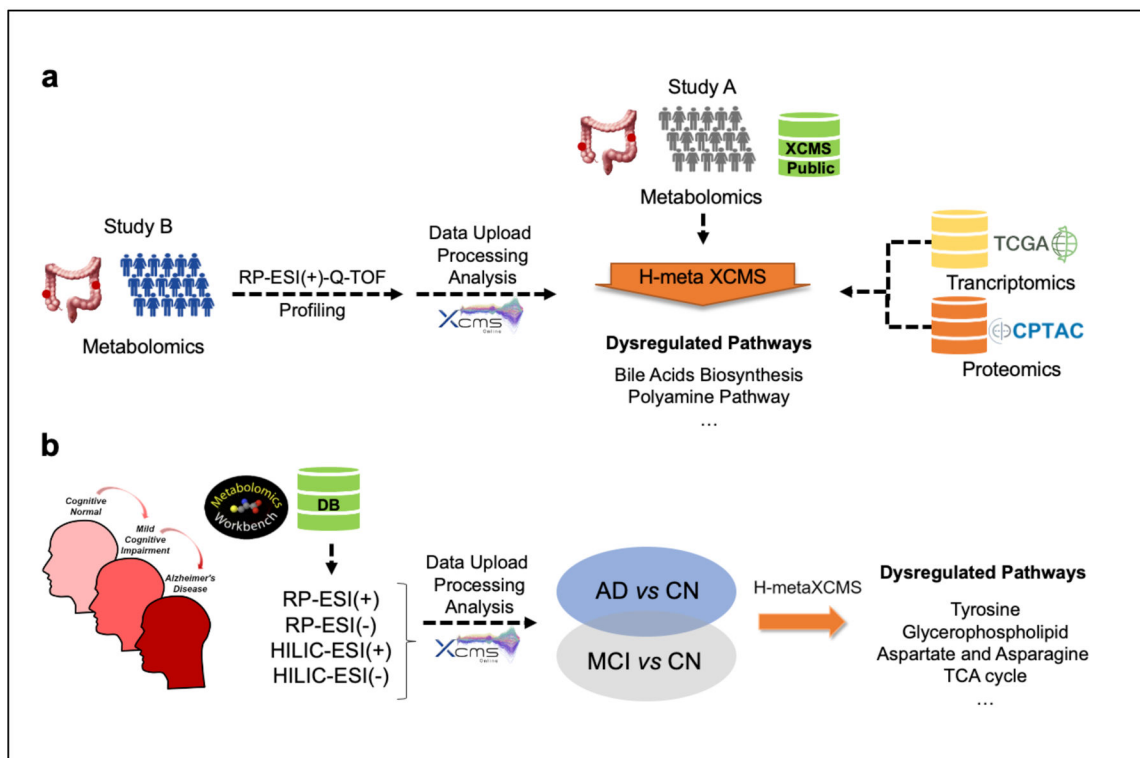
Figure 3.
a) Meta-analysis of colon cancer metabolomics studies, and systems level interpretation with proteomics and transcriptomics data; b) cohesive re-analysis of heterogeneous archival metabolomics data from AD and MCI, compared with CN.
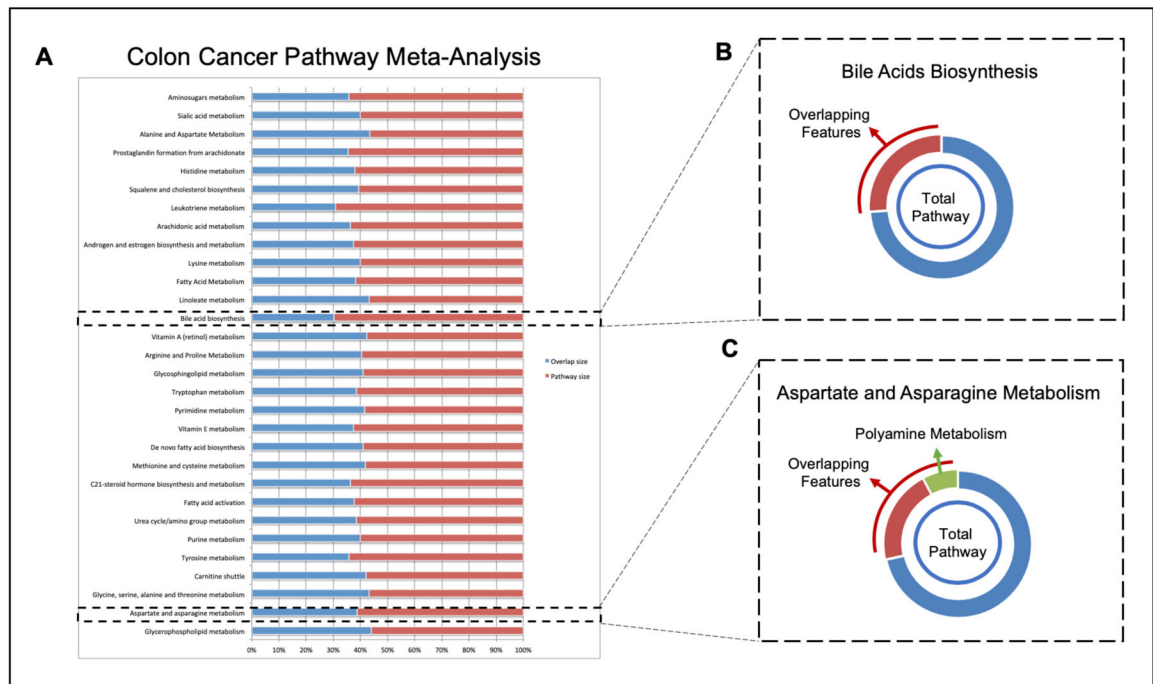
Figure 4.
Meta-analysis across independent colon-cancer studies predicts 30 dysregulated metabolic pathways (A), including bile acids biosynthesis (B) and polyamine metabolism (C).
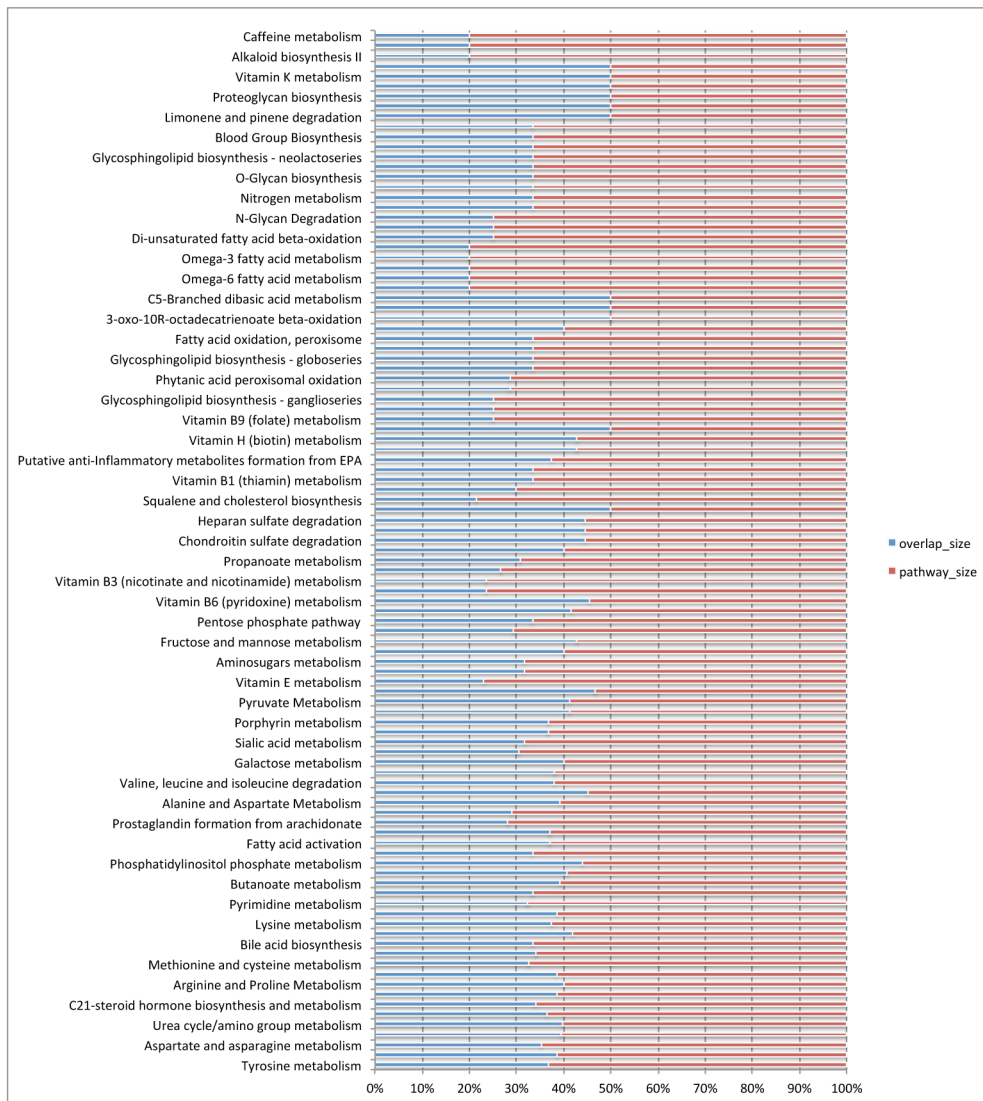
Figure 5.
Meta-analysis across AD vs CN and MCI vs CN studies in plasma predicts 101 dysregulated metabolic pathways.