Published in final edited form as:
Nat Biotechnol. 2014 June ; 32(6): 524–527. doi:10.1038/nbt.2927.

Metabolomic data streaming for biological-dependent data acquisition

Duane Rinehart[1,4], **Caroline H. Johnson**[1,4], **Thomas Nguyen**[1], **Julijana Ivanisevic**[1], **Homer Paul Benton**[1], **Jessica Lloyd**[2], **Adam P. Arkin**[3], **Adam M. Deutshbauer**[3], **Gary J. Patti**[2], and **Gary Siuzdak**[1]

Gary J. Patti: gjpattij@washu.edu; Gary Siuzdak: siuzdak@scripps.edu

[1]Department of Chemistry, Molecular, and Computational Biology, and the Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

[2]Departments of Chemistry, Genetics, and Medicine, Washington University School of Medicine, St. Louis, Missouri 6310

[3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

To the Editor

During the last 10 years, metabolomics has emerged as a powerful technology to interrogate cellular biochemistry at the global level. Although much of the success has been driven by advances in mass spectrometry, developments in bioinformatic resources for data processing have been equally important. The widely used metabolomic software XCMS, in particular, has undergone substantial improvements since its first introduction in 2005 (ref. 1). In addition to improved algorithms for peak picking, retention time alignment and data visualization, XCMS has transitioned from a command-line interface requiring expertise with the R programming language, to a web-based platform with a graphical user interface[2]. This web-based platform, which we call XCMS Online, enables thousands of users to upload their metabolomic data and perform cloud-based processing.

Cloud-based processing and storage of metabolomic data with XCMS Online offers several distinct advantages for analyzing metabolomic results. It reduces the need for on-site hardware and software resources for example, and is also easily scalable with computational demands[3]. Indeed, it is now possible with XCMS Online to analyze data on the order of terabytes (Supplementary Fig. 1). Uploading data to XCMS Online requires minimal technical expertise. First-time users can simply chose an appropriate default parameter set for their instrument, whereas advanced users can modify existing parameter sets. Therefore, XCMS Online provides a robust platform for non-experts and experts to perform

[4]These authors contributed equally to this work

AUTHOR CONTRIBUTIONS

DR, TN, CJ, JI, HB, JL, AA and AD performed experimental work. DR, TN, CJ, JI, JL, GP, GS wrote paper.

COMPETING FINANCIAL INTERESTS

The authors have no competing financial interests.

metabolomic data processing. Despite the advantages of cloud-based data processing, however, the major challenge has been the time required to upload metabolomic data files to the XCMS Online server. Depending on file sizes and Internet connection speed, data upload can sometimes take more than a day to complete. Given the cumulative time required to acquire the profiling data, upload the files, inspect the results manually, and then re-run the samples for targeted MS^2 analysis, it can take up to a week to complete the entire untargeted metabolomic workflow.

Here we describe a solution to the time demands of metabolomic data upload to XCMS Online. In brief, we have designed XCMS Online software that enables uploading of metabolomic data files from the instrument computer workstation as they are acquired. Although upload speed is still a function of data size and Internet connection speed, this software introduces improved efficiency to the untargeted metabolomic workflow. That is, much of the data upload time is occurring in parallel to the data acquisition. If each liquid chromatography/mass spectrometry (LC/MS) run is considered as a discreet data packet, the process of uploading these results while simultaneously acquiring data for the next sample can be considered as a type of data ‘streaming.’

To illustrate the time demands of uploading metabolomic data, we analyzed 1,000 jobs processed by XCMS Online over a two month period by hundreds of unique users. (Note: data were only assessed from users that gave permission to perform such comparisons at the time of their XCMS Online registration.) From these 1,000 jobs, we found that the number of samples processed by each user ranged from 4–3,000, with a mean file size of ~14.0 GB for high-resolution data. The upload time using a non-Local Area Network (LAN) connection (Fig. 1) ranged from 15 hours to 3 days, but on average was 20 hours, dependent on the user’s local available speeds. On the basis of each job’s specific LC/MS run time (including column washes when designated) and average Internet connection speed, we calculated how much of this upload time could be reduced in parallel to LC/MS data acquisition by using a streaming approach. We determined that most of the data upload would be accomplished prior to the finishing of the last LC/MS sample analysis. Specifically, for these 1,000 jobs, we determined that streaming would reduce the mean wait time after the last LC/MS run to complete data processing from 20 hours to less than three, a reduction of 7-fold.

In the current version of streaming script file compression is unnecessary as the average data transfer time was less than the time required to complete a single LC/MS run. However, a data compression option is also available to further reduce the data upload time for faster LC/MS experiments, such as ultra performance liquid chromatography (UPLC). As an example, the average time required for uploading data from an 82 minute run (60 minute run plus wash/re-equilibration), was 57 minutes. The total time saving would be the number of runs multiplied by the average upload time per run. When analyzing large datasets, the proposed streaming approach could reduce the upload time of a terabyte of data by three orders of magnitude (Supplementary Figs. 1 and 2).

As a real example to demonstrate the efficacy of streaming in laboratories at different geographical locations, we performed a metabolomic experiment at Washington University in St. Louis, USA. A script (Supplementary data file 1 is downloadable at (https://xcmsonline.scripps.edu/nbt.php) was installed on the computer workstation of an Agilent quadrupole time-of-flight mass spectrometer (QTOF-MS, 6520) at Washington University. The script detects the end of an LC/MS run and initiates the subsequent transfer of the data along with any metadata about the instrument parameters, sample type, etc. to the XCMS Online server. When setting up the streaming, users are presented with options to automatically tag samples based on origin source to facilitate data archival/retrieval as well as sample group definitions. Heightened security is achieved by encryption and file checksums are compared upon completion of transfer to prevent the risk of file corruption. These scripts are available to all XCMS Online users via the website (https://xcmsonline.scripps.edu/nbt.php); each script will have slight modifications depending on the type of mass spectrometer.

As described above, with XCMS Online, users are typically acquiring profiling data first. After the data are uploaded and processed, the results are inspected manually, and metabolomic features that have statistical values above a defined threshold (e.g., P-value ≤0.01 and fold change > 2) as well as METLIN database hits are selected for additional analysis. These features are then re-analyzed and MS^2 data are acquired to structurally support putative database assignments. As an alternative to this type of targeted MS^2 approach, it has been suggested that MS^2 data for structural identification are acquired for every feature at the same time that MS^1 data are acquired for profiling[4]. This untargeted workflow has been referred to as autonomous metabolomics and allows for the immediate generation of MS^2 data, thereby reducing the data-analysis time. Practically, the recent development of mass spectrometers with increasing MS^2 acquisition speeds has made the possibility of acquiring MS^2 data for every metabolomic feature more likely; however, the data quality of the MS^2 spectra obtained at such speeds can still be problematic[5]. Notably, many MS^2 spectra end up being acquired for compounds that are not of interest to the investigator at the expense of decreased data quality for the compounds that are of interest. The introduction of data streaming, however, offers an improvement upon the autonomous metabolomic workflow. Instead of acquiring MS^2 data for every metabolomic feature, we suggest acquiring MS^2 data only for the features of interest to the investigator based on pre-defined statistical thresholds and whether or not the compounds have accurate mass matches in the METLIN metabolite database. Although this is conceptually similar to data-dependent MS^2 acquisition, a workflow that has been used in proteomics[6], this biological dependent data acquisition is unique in that MS^2 is not triggered on the basis of ion intensity. Rather, MS^2 is triggered based on the previously acquired and processed files that have already been uploaded and analyzed by XCMS Online. The data processing involved with automated selection of ions targeted for MS^2 analysis is analogous to that which has already been described[7], but here ion selection and MS^2 acquisition will occur within the same set of experimental runs. In this context, XCMS-based streaming allows for biological-dependent data acquisition.

To demonstrate XCMS Online-based streaming and the utility of biological-dependent data acquisition, we performed a set of experiments on tumor samples and normal tissues using our existing XCMS Online platform (Fig. 2 and Supplementary Fig. 3). For this comparison, 28 normal and tumor samples were prepared for LC/MS analysis. In brief, metabolites from

10 mg of tissue were isolated as described previously by using an acetone/methanol extraction and analyzed by an Agilent QTOF[8]. The experiment was carried out by using the script mentioned above, which communicated with the application programming interface of the mass spectrometry software. For biological dependent data acquisition, instead of processing the data after the final sample upload as shown in Figure 1, the data were uploaded to XCMS Online after each LC/MS run and reprocessed (using a paired Wilcoxon signed-rank test) to identify ions with an *m/z* of the most statistically meaningful biological candidates. The statistical analysis started when the number of samples uploaded per group was equal to three, and the univariate analysis was performed consecutively after each sample was acquired. The thresholds for ions selected by biological dependent data acquisition were set at a P-value ≤ 0.001, a fold change ≥ 1.5, and an intensity of > 10,000. Those that had accurate mass matches (< 15 ppm) to the METLIN metabolite database were further designated for MS^2 analysis. As data streaming progresses, the P-value of the ion shown to be dysregulated between normal and tumor tissues decreases (Fig. 2), and MS^2 is triggered to allow for identification. To augment biological dependent data acquisition, we also introduce a script that enables automated metabolic pathway analysis (Supplementary data file 2 is downloadable at https://xcmsonline.scripps.edu/nbt.php). This script identifies putatively identified metabolites (based on accurate mass) in the same metabolic pathway that are dysregulated and then selects these ions for MS^2 analysis. In short, metabolite identifiers (name, Kyoto Encyclopedia of Genes and Genomes (KEGG) or Chemical Abstracts Service (CAS)) are transmitted via Simple Object Access Protocol (SOAP) or Representational State Transfer (REST) Internet query methods to the three following metabolic pathway databases concurrently Reactome (www.reactome.org)[9], The Small Molecule Pathway Database (www.smpdb.ca)[10] and IMPaLA: Integrated Molecular Pathway Level Analysis (http://impala.molgen.mpg.de)[11]. When two or more putatively assigned metabolites are found to be in the same pathway, the MS^1 data are then searched for the accurate masses of each metabolite in that pathway and putative matches are then targeted for MS^2 analysis (even if they are not dysregulated. In the data shown here, IMPaLA identified four metabolites belonging to the same pathway “urea cycle and metabolism of arginine, proline, glutamate, aspartate, and asparagine”. As a result, this pathway was a target for subsequent analysis and enabled assessment of its role in cancer. In a similar streaming analysis applied to bacterial samples chemically stressed, we found glutamate metabolism to be dysregulated (Supplementary Fig. 3). It should be noted that although we have only demonstrated our approach by using Agilent instrumentation, data streaming and biological dependent date acquisition can be performed on instruments from any vendor (Agilent, AB SCIEX, Thermo, Bruker and Waters) (Supplementary Fig. 4). Also, although our biological dependent MS^2 acquisition is designed to generate data on important peaks of relevance to the investigator, some interesting metabolites may be missed and therefore coupling this platform with standard full-data analysis may provide additional insight.

In summary, cloud-based processing of metabolomic data offers many benefits, but is largely limited by the speed of data transfer over the Internet, a problem reminiscent of online media communications. However, the application of mass spectrometry data streaming will facilitate web-based processing of metabolomic results and additionally offer the possibility of biological-dependent data acquisition. Although here we have only demonstrated the benefits of data streaming for mass spectrometry-based metabolomics, we expect that this concept could be extended to any experimental analysis requiring data upload and real-time feedback from cloud-based processing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the US National Institutes of Health nos. R01 CA170737 (G.S.), R24 EY017540 (G.S.), P30 MH062261 (G.S.), RC1 HL101034(G.S.), P01 DA026146 (G.S.), and R01 ES022181 (G.J.P.), and the U.S. National Institutes of Health-National Institute on Aging no. L30 AG0 038036 (G.J.P.). Financial support was also received from the U.S. Department of Energy grant nos. FG02-07ER64325 and DE-AC0205CH11231 (G.S.)

References

1. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. Analytical chemistry. 2006; 78:779–787. [PubMed: 16448051]
2. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. Analytical chemistry. 2012; 84:5035–5039. [PubMed: 22533540]
3. Kienzler, R.; Buruggmann, R.; Ranganathan, A.; Tatbul, N. Euro-Par 2011: Parallel Processing Workshops, Part II, LNCS 7156. Alexander, M., et al., editors. Springer-Verlag; Berlin Heidelberg: 2012. p. 467-476.
4. Tautenhahn R, et al. Nature biotechnology. 2012; 30:826–828.
5. Nikolskiy I, Mahieu NG, Chen YJ, Tautenhahn R, Patti GJ. Analytical chemistry. 2013; 85:7713–7719. [PubMed: 23829391]
6. Liu H, Sadygov RG, Yates JR 3rd. Analytical chemistry. 2004; 76:4193–4201. [PubMed: 15253663]
7. Neumann S, Thum A, Bottcher C. Metabolomics : Official journal of the Metabolomic Society. 2013; 9:S84–S91.
8. Yanes O, Tautenhahn R, Patti GJ, Siuzdak G. Analytical Chemistry. 2011; 83:2152–2161. [PubMed: 21329365]
9. Matthews L, et al. Nucleic acids research. 2009; 37:D619–622. [PubMed: 18981052]
10. Frolkis A, et al. Nucleic acids research. 2010; 38:D480–487. [PubMed: 19948758]
11. Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC. Bioinformatics. 2011; 27:2917–2918. [PubMed: 21893519]

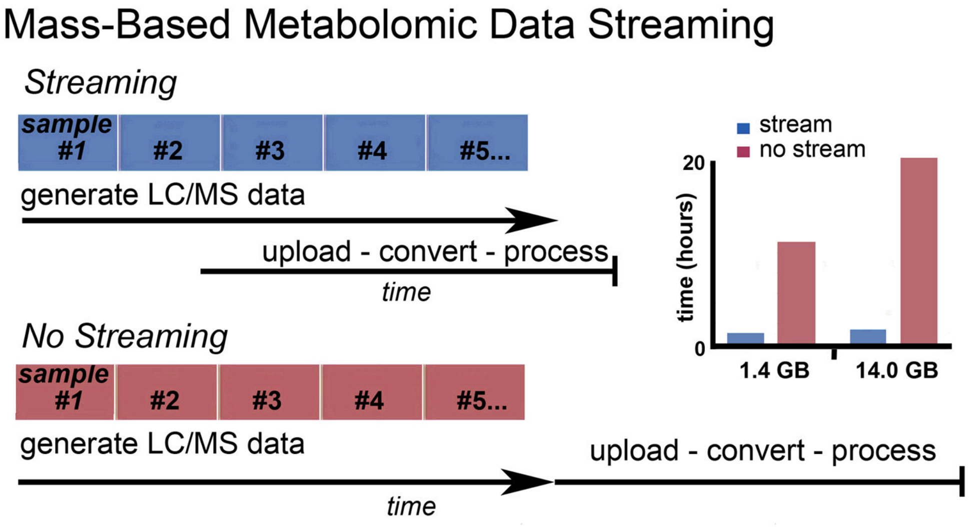

Figure 1.

XCMS-based data streaming workflow (top left) allows for data upload and processing after each LC/MS run is performed, dramatically reducing the processing time after the data are acquired for the final sample (top right). A thousand XCMS Online data sets were examined for their average processing time without streaming. For low-resolution data (~1.4GB) and high-resolution data (~14.0GB) over 10 and 20 hours was required after the final LC/MS analysis was performed, respectively. Streaming allowed for a 7-fold decrease in average processing time after data acquisition.

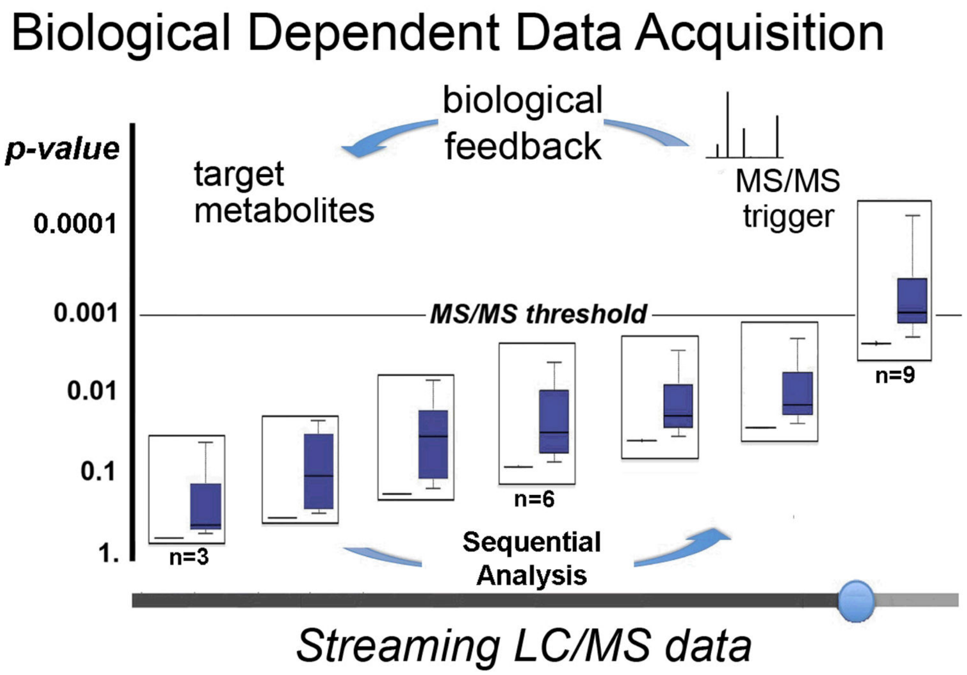

Figure 2.

Biological-dependent data acquisition from tumor samples. Instead of using data driven acquisition of MS^2 data that relies on intensity, signal-to-noise ratio (S/N), or prior acquisition of precursor ions, biological-dependent data acquisition relies on statistics generated after each sample run for mass spectrometry data acquisition decision making. The representative example, generated from cancer tumor samples, shows a decreasing P-value for a feature of interest over the time-course of data streaming. When the P-value for the features reaches 0.001, MS^2 is carried out. A two-tailed Wilcoxon signed-rank test was used to calculate the statistical significance for *n*=28. Box and whisker plots display the full range of variation (whiskers - median with min-max, boxes - the interquartile range).