analytical chemistry

Article

pubs.acs.org/ac

Data Streaming for Metabolomics: Accelerating Data Processing and Analysis from Days to Minutes

J. Rafael Montenegro-Burke,[†,△] Aries E. Aisporna,[†,△] H. Paul Benton,[†] Duane Rinehart,[†] Mingliang Fang,[†] Tao Huan,[†] Benedikt Warth,[†] Erica Forsberg,[†] Brian T. Abe,[‡] Julijana Ivanisevic,[#] Dennis W. Wolan,[∥] Luc Teyton,[‡] Luke Lairson,[§] and Gary Siuzdak*[,†,⊥]

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

[‡]Department of Immunology and Microbial Science, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[§]Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

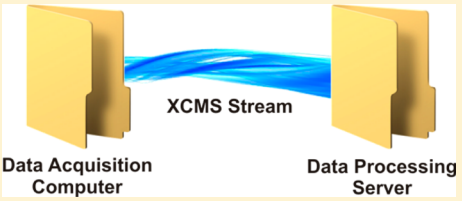
[#]Metabolomics Research Platform, Faculty of Biology and Medicine, University of Lausanne, Rue du Bugnon 19, 1005 Lausanne, Switzerland

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

[∥]Departments of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

S *Supporting Information*

ABSTRACT: The speed and throughput of analytical platforms has been a driving force in recent years in the "omics" technologies and while great strides have been accomplished in both chromatography and mass spectrometry, data analysis times have not benefited at the same pace. Even though personal computers have become more powerful, data transfer times still represent a bottleneck in data processing because of the increasingly complex data files and studies with a greater number of samples. To meet the demand of analyzing hundreds to thousands of samples within a given experiment, we have developed a data streaming platform, XCMS Stream, which capitalizes on the acquisition time to compress and stream recently acquired data files to data processing servers, mimicking just-in-time production strategies from the manufacturing industry. The utility of this XCMS Online-based technology is demonstrated here in the analysis of T cell metabolism and other large-scale metabolomic studies. A large scale example on a 1000 sample data set demonstrated a 10 000-fold time savings, reducing data analysis time from days to minutes. Further, XCMS Stream has the capability to increase the efficiency of downstream biochemical dependent data acquisition (BDDA) analysis by initiating data conversion and data processing on subsets of data acquired, expanding its application beyond data transfer to smart preliminary data decision-making prior to full acquisition.



Data streaming has been adopted by mobile device applications especially in the entertainment and electronic gaming industries; however, its broader application to science has been largely overlooked. For example, while data acquisition technologies, computing power and software have significantly evolved over the last 10 years, especially in the area of mass spectrometry,[1−3] the ability to interrogate multidimensional data during acquisition typically requires manual evaluation of a few parameters, resulting in incomprehensive and subjective assessments, not only of the data but also of the quality of the analysis. Consequently, waiting until a full data set has been acquired to proceed with data analysis is usually the common practice. Nevertheless, the ability to survey recently acquired discrete data packets from large-scale data sets pertaining to LC-MS, NMR, or omic data in general can be limited by the location of the generated data files because of the use of centralized data analysis servers and in recent years, the more common situation of sharing laboratory equipment in interdisciplinary research.[4] We previously introduced the concept of data streaming applications that were focused on cloud computing capabilities,[5] and while cloud-based processing and storage of data offers several distinct advantages, the utility streaming can have even broader applications to the scientific community when interconnected to downstream data processing and analysis.

Highly complex multidimensional data and an increase in data file sizes have spurred the transition from instrumentation computer workstations to both on-site servers and workstations for data analysis. This typically requires the manual transfer of

Received: October 3, 2016
Accepted: December 16, 2016
Published: December 16, 2016

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259

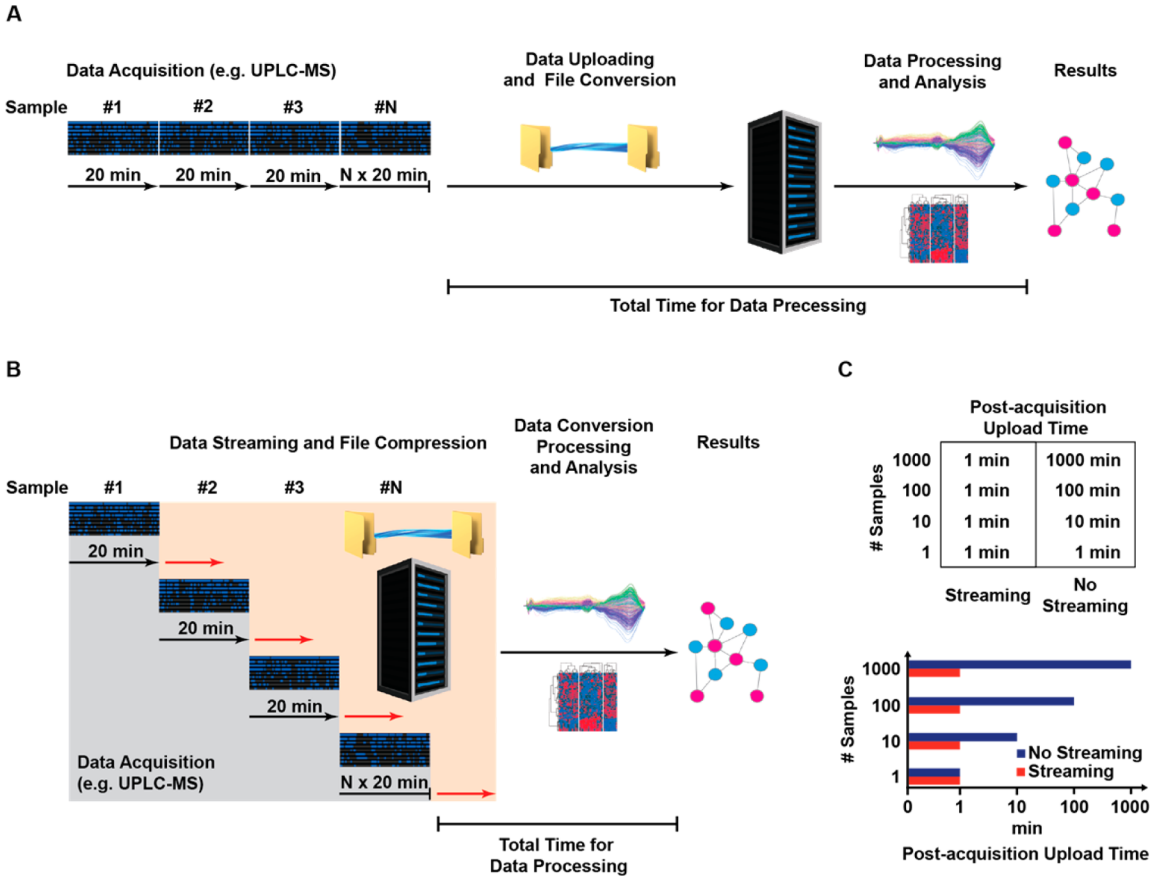


Figure 1. Theoretical time comparison of analytical process with data streaming capabilities. (A) Traditional process of data acquisition followed by data conversion and uploading to server for data processing and analysis before obtaining results. (B) Alternative process utilizing real-time data streaming. Files are compressed and streamed to server after acquisition while other data files are being acquired, reducing the time needed for obtaining results. (C) Direct data upload time comparison for different number of samples between with and without streaming capabilities (assuming 1 min upload time for each data file for both streaming and no streaming scenarios).

data files via portable media (e.g., external hard drives), not only decreasing the efficiency of the process but also increasing the wait time for results. For example, the transfer time of 50 GB of data (a common size for metabolomic studies) with an external hard drive at a copy speed of 30 MB/s would take 1 h (30 min to transfer the data from an instrument computer to an external hard drive and another 30 min to copy to the destination computer's hard drive). Furthermore, data file conversion is a heavy computational task, which ultimately slows down personal computers. For instance, the conversion of 1 GB raw LC-MS data file from a proprietary format to an open format used by data processing and data analysis bioinformatic tools can take up to 30 min for dedicated multicore servers using software such as ProteoWizard.[6] Thus, it is not uncommon to perform these tasks overnight to avoid hampering other functions performed during normal working hours. However, these time limitations can be overcome by simultaneously compressing and streaming data files to processing servers while data acquisition (LC-MS) is underway. This capability, depicted in Figure 1A and B, exploits the time after the first sample has been analyzed until the completion of the entire sample list, maximizing the efficiency of the analysis and decreasing the time needed for the generation of results. Furthermore, manually uploading different files as data sets can be a long and tedious task, which is drastically increased with a larger number of data files (Figure 1C). Another advantage of a streaming platform is coupling data processing and analysis capabilities. This removes the necessity to physically start the process (submit "Job" in XCMS Online) once all data files have been streamed and assigned to their respective data sets, which reduces the total time needed for the generation of results after data acquisition.

Herein, we describe XCMS Stream, a data streaming platform for real-time data processing and data analysis and compare this platform to alternative manual uploading using a LC-MS metabolic study of human CD4 and CD8 T cells, as well as recently reported data from a 1000 sample data set.

EXPERIMENTAL SECTION

Development of XCMS Stream. XCMS Stream was developed using the programming language C# and it has been tested on Windows 7 with.Net 4.5 framework (Redmond, WA). HTTPS are used for the communication with XCMS Online,[7] which has been written using the programming languages JavaScript, HTML5, CSS3, and PHP and it connects to MySQL database. MD5 checksum was added for file integrity corroboration throughout the streaming process.

Sample Preparation and LC-MS Analysis. *T Cell Isolation.* Human CD4 and CD8 T cells were purified from the peripheral blood of healthy donors by Ficoll-Paque density gradient centrifugation. Fresh blood was diluted 1:1 in PBS, layered onto Ficoll-Paque Premium 1.084 (GE Healthcare, Chicago, IL) and centrifuged at 400 × *g* for 40 min without braking. Peripheral blood mononuclear cells were isolated, and

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259

subsequently washed 3× in Isolation Buffer (PBS, 0.5% BSA, 2 mM EDTA). CD4 or CD8 T cells were purified using negative isolation kits (Miltenyi, Bergisch Gladbach, Germany) on an AutoMACS Pro Separator (Miltenyi) according to the manufacturer's protocol. Cell purity was assessed by flow cytometry on a MACSQuant Analyzer 10 (Miltenyi) and cell counting performed using a hemocytometer. Cells were washed in PBS, pelleted, and flash frozen in liquid nitrogen.

Metabolite Extraction and LC-MS Analysis. CD4 and CD8 T cells samples were prepared as previously described.[8] Briefly, cell pellets (~10^6 cells) were extracted with 1 mL of cold MeOH/ACN/H_2O (2:2:1, v/v) solvent mixture, vortexed for 30 s, and incubated in liquid nitrogen for 1 min. After thawing at room temperature, the samples were sonicated for 10 min. This cell lysis procedure of freezing and sonication was repeated three times followed by incubation at −20 °C for 1 h and centrifugation at 13 000 rpm and 4 °C for protein precipitation. The supernatant was then evaporated to dryness and reconstituted in 100 μL ACN/H_2O (1:1, v/v) for LC-MS analysis.

Analysis were performed with an HPLC system (1200 series, Agilent Technologies, Santa Clara, CA) coupled to an Impact II Q-TOF (Bruker, Billerica, MA). Samples were analyzed using a Luna Aminopropyl, 3 μm, 150 mm × 1.0 mm I.D. column (Phenomenex, Torrance, CA). The mobile phase consisted of A = 10 mM ammonium acetate and 10 mM ammonium hydroxide in 95% water and B = 95% acetonitrile. A 60 min linear gradient was used with an injection volume of 8 μL.

XCMS Stream Workflow. The data streaming strategy described here combines technology used mainly in the entertainment industry. However, XCMS Stream adds to this existing technology by coupling two data processing platforms and improves the efficiency of untargeted biomolecular interrogation. A general overview of this strategy and process is shown in Figure 2. Upon sample acquisition completion, the data file is compressed (zipped) in order to reduce the streaming time by reducing the size of the data file. The data file is first streamed to an auxiliary server (XO Cloud), which acts as a buffer between XCMS Online[7,9,10] and the acquisition computer (instrumentation computer). XO Cloud temporarily stores all class data files pertaining to the different data sets (e.g., triplicates of WT and KO for pairwise analysis, grouped using masks) needed for the selected data processing job. Subsequently, the data sets are then uploaded to XCMS Online for automatic data conversion, data sets creation and processing as selected in the XCMS Stream interface (HPLC, qTOF, *Homo sapiens*, etc.). It should be noted that the file compression, streaming and data sets collection in XO Cloud occurs while data acquisition is still underway. Furthermore, as long as all files needed for the selected job type (e.g., single, pairwise, multigroup comparison, etc.) have been acquired and streamed to XO Cloud, XCMS Online will automatically start with data processing and data analysis while data is still being acquired for other data sets.

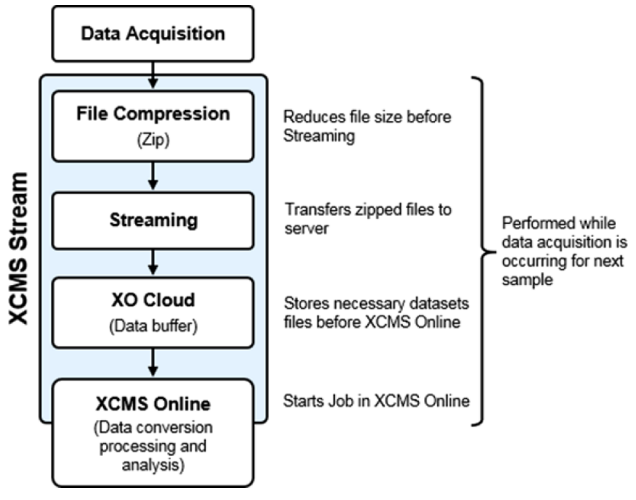


Figure 2. XCMS Stream flowchart showing the general strategy for data streaming. After the acquisition of each individual LC-MS run, the data file is compressed before being streamed to reduce its size. XO Cloud serves the purpose of "Data buffer" between streaming and "Job" submission to XCMS Online. This is necessary because the data processing and data analysis in XCMS Online cannot start without all files necessary for the requested "Job". Upon data upload completion to XCMS Online, data processing and analysis can then take place.

XCMS Stream Interface. XCMS Stream allows for a simple and intuitive solution to streaming large data files and performs data processing and analysis automatically concurrent with data acquisition. In its current version, XCMS Stream has four main settings areas (Figure 3) and for correct connection to the respective servers, XCMS Online user's email address and password are needed to login. In "Directory", the folder where data files are being saved during acquisition as well as the virtual location of the XO Cloud can be selected. Inputting meta information about the run, such as column details, chromatography or sample information is optional. In "Job Information", settings regarding the type of streaming and XCMS Online job are selected. "Run Type" allows for two types of streaming capabilities: *i)* "Online", which streams data files as soon as the acquisition for each file is completed regardless of other current acquisition jobs and *ii)* "Batch", which allows the automatic uploading of data files already generated. The instrument manufacturer and the type of preionization separation are settings selected in "Machine Type" and "Chromatography" respectively. For XCMS Online data processing and analysis, the "Job Type" option indicates which type of analysis should be performed (single job, pairwise comparison and multigroup). Additionally, "Job Parameters" such as type of LC (HPLC or UPLC) as well as type of mass spectrometer (Q, q-TOF and trap instruments) and ionization mode can be selected for already optimized data processing settings. "Bio Sources" can be selected to look at metabolites endogenous only to that specific species and its default setting is *Homo sapiens*.

In the "Data Information" section, the number of files to be streamed, the length of each chromatogram and the post run times (column equilibration) are selected accordingly to configure streaming times and protocols. Lastly, "Masks" (4 user selected characters followed by "*") are used to indicate the relationship between data sets and data files. These same masks must be added at the beginning of the filenames selected in the sample list. Furthermore, quality control samples can be added to a job by clicking and selecting them (highlighted in blue) and will not be used for statistical analysis (Ctrl + click to unselect). Finally, the intuitive green and red buttons start and stop the streaming process, respectively.

Applications. To test the performance and demonstrate its capabilities, we selected an LC-MS metabolic study comparing CD4 and CD8 T cells isolated from different human donors. For this comparison, 10 samples were streamed during the

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259

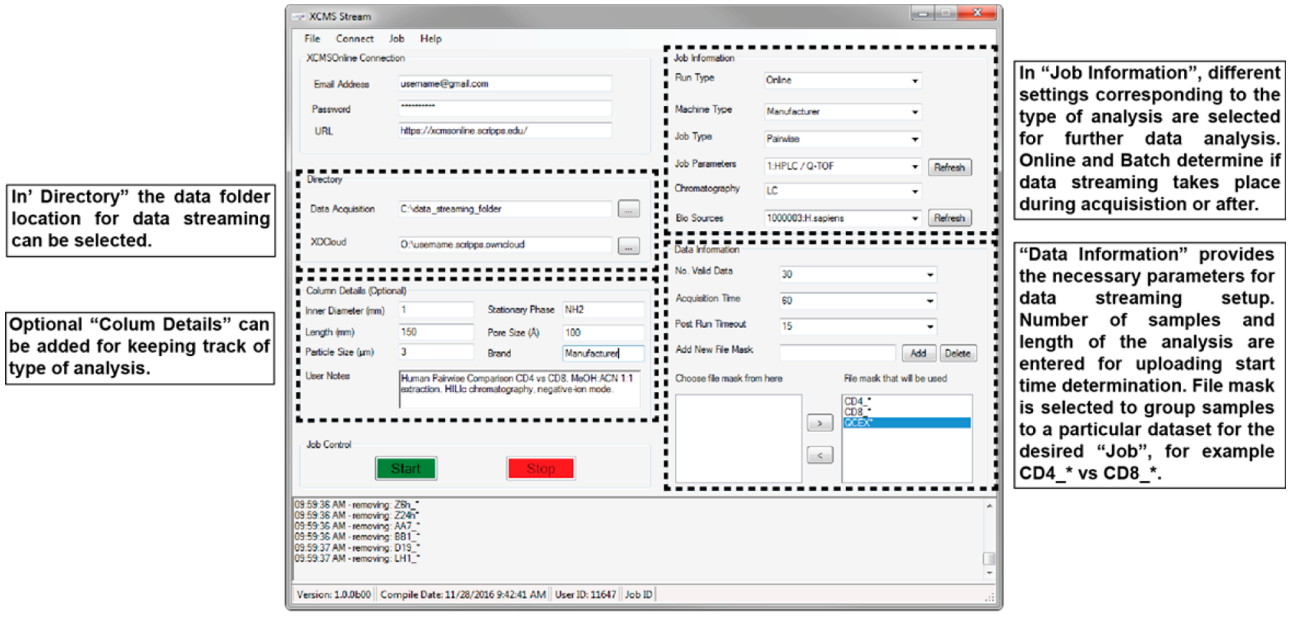


Figure 3. XCMS Stream screenshot of user-friendly interface. The "Directory" section indicates where the files are stored. As an optional section, the "Column Details" allows the data entry of specific stationary phase information for the particular analysis. In "Job Information", the possibility to stream the data online or offline can be selected, where online refers to data streaming while other samples are being acquired and offline refers to data streaming after all samples have been acquired. Furthermore, the selection of single, pairwise and multigroup jobs are available as well as a "Bio Source" option (*H. sapiens* is default). "Data Information" is utilized for the determination of the number of samples and when each data file is complete to start the streaming process. Moreover, the file mask allows for correctly assignment of each sample to a particular data set for the "Job" in XCMSonline data processing and analysis.

acquisition of central carbon metabolic profiling (3 GB each). In Figure 4A, the shorter time needed to obtain results between online streaming and manual uploading can be appreciated. The real-time data streaming during acquisition allowed for quick turnaround of results only 4 h after completion of data acquisition. In contrast, 18 h were needed with manual uploading to obtain results (4.5 times longer). The differential time of 10 h corresponds to the time between the completion of the LC-MS experiment and when the files are manually uploaded to XCMS Online. In our study, data acquisition was completed after working hours (~8:00 PM), as it often is the case. This immediately decreases the efficiency of the analysis, due to the lag time before the files can be uploaded. The manual uploading step (3.5 h) is considerably long and it can be attributed to the lack of file compression and the direct uploading to XCMS Online, which at the same time is converting the data files to mzML format. This highlights the advantage of XO Cloud, which buffers the high volume of data being received (streamed) with the XCMS Online processing functions. Additionally, XCMS Stream's capability to not only transfer data but also start data processing represents a huge advantage due to the fact that this can be done without having to physically start the job in XCMS Online. This is especially beneficial in the case of manual uploading being completed when users are performing other tasks or during nonworking hours.

XCMS Stream's real-time data processing and data analysis capabilities can be applied to smart preliminary data decision -making prior to full sample list acquisition by analyzing the data from the first samples. For example, paired analyses of the pooled samples (small aliquots from each sample are pooled to condition the column) can provide information about reproducibility and stability of the system by monitoring retention times shifts and ion intensity changes. Additionally, in the case of a large number of samples, the sample list can be setup so that the first samples can be compared against each other and aid determine if enough differences are observed or if the type of analysis is appropriate for the particular experiment. Such information is extremely valuable before spending several days acquiring fruitless data with expensive laboratory equipment.

The second streaming modality available in XCMS Stream, batch streaming, was also compared in this study and for that we used the experimental design of a previously published study by Lewis et al., where 1000 urine samples were analyzed using LC-MS (Figure 4B).[11] In this modality, in contrast to online streaming, data files are streamed to XO Cloud after all data files have been acquired. These files can be located at computers both on the instrument or personal computers used by users for data analysis. The distinct differences between batch streaming and manual uploading are files are compressed prior to streaming and all files are uploaded in a batch, independently of which data sets they correspond to. Furthermore, the file conversion takes place in the server after data streaming and, which drastically reduces the burden of the acquisition computer. These differences, though small, represent big advantages. For instance, using our T cells comparison in the metabolic LC-MS experiments we determined streaming and uploading times of 2.3 and 10.5 min per sample for streaming and manual upload, respectively. These estimated times translate to huge efficiency improvement by reducing the data transfer time from 7.3 to 1.6 days when batch streaming is used instead of traditional manual uploading. Further, as mentioned above, it automatically started data processing and analysis.

It should be noted that a conservative number of samples was selected in this comparison and the advantages of streaming grow with increasing number of samples and time of analysis.

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259

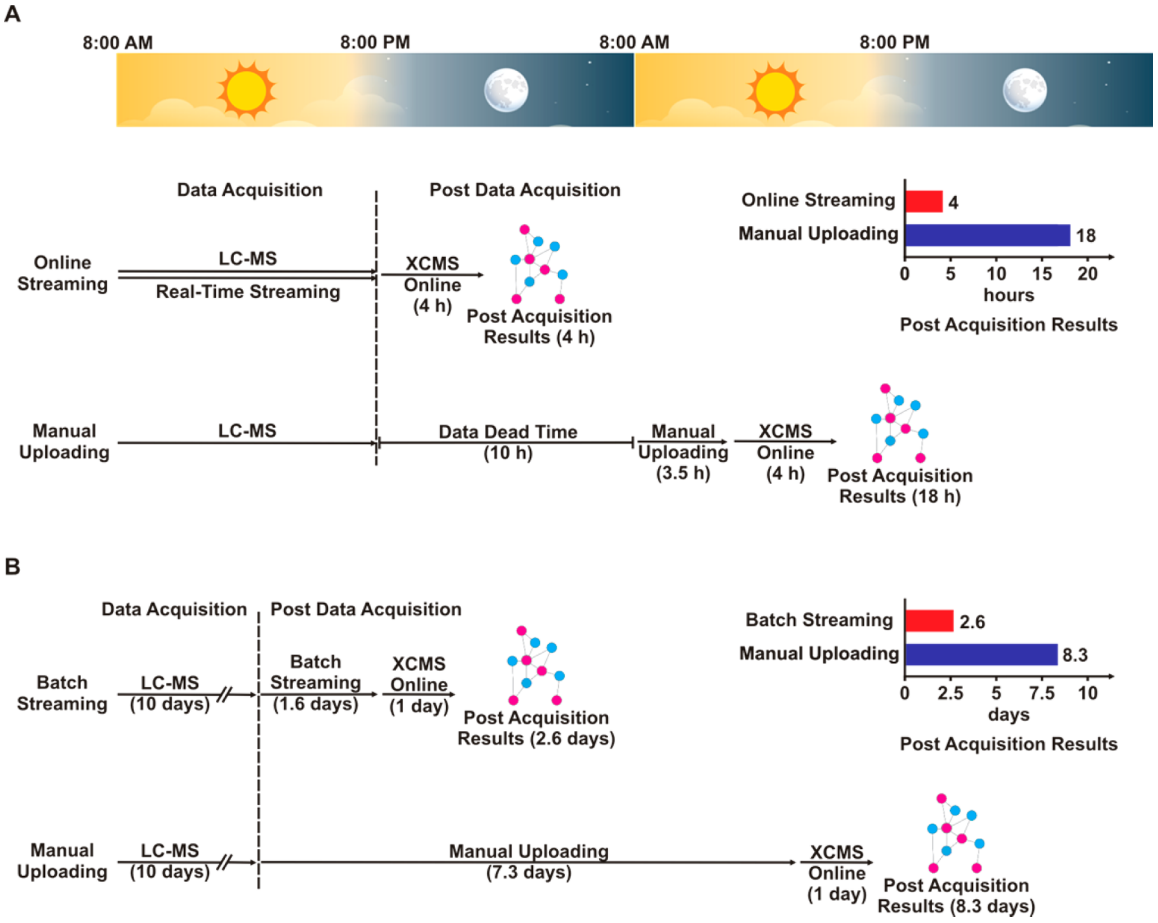


Figure 4. Time comparison between XCMS Stream and manual data uploading for the pairwise analysis of CD4 and CD8 human T cells and 1000 urine samples from ref.[11] (A) A Large time savings are gained by "Online Streaming" with results being generated only 4 h after data acquisition compared to "Manual Uploading" (18h). Data dead time is the time after the completion of data acquisition and data uploading for processing. (B) In "Batch Streaming", the data files are automatically uploaded to user specific data sets and an XCMS Online job is generated. This is performed after data acquisition is completed. The time savings of "Batch Streaming" compared to "Manual Uploading" for 1000 urine samples is 5.7 days.

For example, in the study by Lewis et al., where 1000 urine samples were analyzed in a LC-MS metabolomics profiling experiment.[11] However, other studies have analyzed considerable larger number of samples (5000−10 000).[12,13] To estimate time savings, we extrapolated the different streaming and uploading times experimentally determined, 2.3 and 10.5 min, respectively (Figure 5). These time savings cover a wide range depending on the modality spanning a few minutes to 73 days and it should be noted that a logarithmic scale was needed. The streaming time in the online modality was as expected the same independently of the number of samples, 2.5 min (streaming time for one sample). Additionally, in batch streaming, large time savings are observed for larger number of samples. For larger studies (1000 and 10 000),[12,13] the data transfer times differences can represent time savings in the range of 6−60 days, and given the direction of personalized or precision medicine,[14−16] it is safe to say the tendency of studies with large sets of samples will remain. Moreover, the complexity of samples is dramatically increasing, as metabolomics is surging to the forefront of methodologies routinely employed to analyze highly complex biological samples comprised of many hundreds of bacterial species, including the metabolites from distal gut, skin, and oral microbiomes. While the examples provided here focus on large sample numbers and the corresponding savings in time, we likewise envision that the duration of analyses of microbiome-derived metabolomic datasets will be vastly shortened with the application of XCMS Stream.

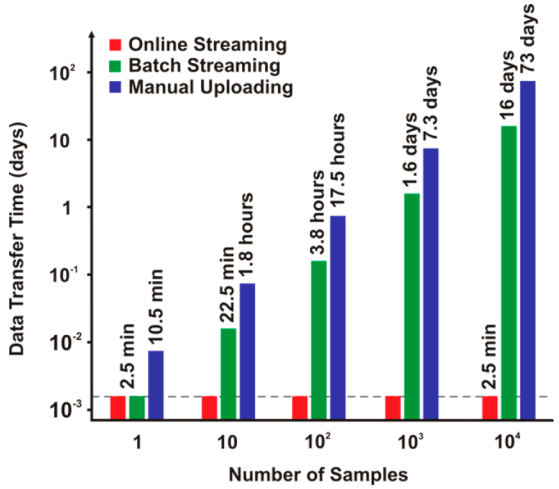


Figure 5. Extrapolation of time comparison between XCMS Stream and manual data uploading for large data sets. Data transfer time comparison in days (logarithmic scale) for different number of samples between online, batch streaming and manual uploading.

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259

CONCLUSION

In summary, the development of XCMS Stream allows users to increase the efficiency of their analytical workflow by utilizing streaming technology to reduce data transfer times. Previously, data acquisition time was considered "useless" in terms of data processing and analysis time. However, we demonstrate with both experimental examples conducted in our laboratories and with previously published large-scale studies, that streaming data files during acquisition can provide dramatic times savings (hours to days). Furthermore, the compression and streaming of files in batch mode significantly reduced the total time in the analytical workflow of metabolomics studies. We expect the XCMS Stream approach, freely available at https://xcmsonline.scripps.edu, to not only to be used in metabolomics studies but also to stimulate researchers in other fields to implement streaming technology to their workflows.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b03890.

Prerequisites for XCMS Stream, installing ownCloud, installing XCMS Stream, how to use XCMS Stream, batch run example, and troubleshooting guide (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: 858-784-9415. E-mail: siuzdak@scripps.edu.

ORCID

J. Rafael Montenegro-Burke: 0000-0001-7787-3414

Benedikt Warth: 0000-0002-6104-0706

Erica Forsberg: 0000-0001-5190-1501

Author Contributions

△J.R.M.-B. and A.E.A. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the following for funding assistance: Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-AC02-05CH11231, and the National Institutes of Health (NIH) grant R01 GMH4368.

REFERENCES

(1) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinf.* **2010**, *11*, 395.

(2) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Res.* **2015**, *43*, W251.

(3) Montenegro-Burke, J. R.; Phommavongsay, T.; Aisporna, A. E.; Huan, T.; Rinehart, D.; Forsberg, E.; Poole, F. L.; Thorgersen, M. P.; Adams, M. W. W.; Krantz, G.; Fields, M. W.; Northen, T. R.; Robbins, P. D.; Niedernhofer, L. J.; Lairson, L.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2016**, *88*, 9753−9758.

(4) Schiermeier, Q. *Nature* **2012**, *492*, 299−300.

(5) Rinehart, D.; Johnson, C. H.; Nguyen, T.; Ivanisevic, J.; Benton, H. P.; Lloyd, J.; Deutschbauer, A.; Arkin, A.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2015**, *32*, 534−527.

(6) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534−2536.

(7) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035−5039.

(8) Ivanisevic, J.; Zhu, Z.-J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876−6884.

(9) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(10) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *86*, 6931−6939.

(11) Lewis, M. R.; Pearce, J. T. M.; Spagou, K.; Green, M.; Dona, A. C.; Yuen, A. H. Y.; David, M.; Berry, D. J.; Chappell, K.; Horneffer-van der Sluis, V.; Shaw, R.; Lovestone, S.; Elliott, P.; Shockcor, J.; Lindon, J. C.; Cloarec, O.; Takats, Z.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2016**, *88*, 9004.

(12) Begley, P.; Francis-McIntyre, S.; Dunn, W. B.; Broadhurst, D. I.; Halsall, A.; Tseng, A.; Knowles, J.; Goodacre, R.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 7038−7046.

(13) Fredriksen, Å.; Meyer, K.; Ueland, P. M.; Vollset, S. E.; Grotmol, T.; Schneede, J. *Hum. Mutat.* **2007**, *28*, 856−865.

(14) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. *Nat. Protoc.* **2011**, *6*, 1060−1083.

(15) Trifonova, O.; Knight, R. A.; Lisitsa, A.; Melino, G.; Antonov, A. V. *Drug Discovery Today* **2016**, *21*, 103−110.

(16) Wishart, D. S. *Nat. Rev. Drug Discovery* **2016**, *15*, 473−484.

DOI: 10.1021/acs.analchem.6b03890
Anal. Chem. 2017, 89, 1254−1259