with a strong record of innovation in semiconductors and automobiles, also has very few biopharmaceutical inventors. Overall, the figures show that the locations of biopharmaceutical innovation have remained largely the same since 2000; clearly, most inventors reside in the United States, the legacy biopharmaceutical nations in Europe and in Japan.

The near-absence of patent inventorship (12 patent inventors in China and 9 in India versus >4,000 in the United States since 2000) raises questions about the level of pharmaceutical innovation in these emerging economies. What's more, the low level of R&D investment in these countries—according to the Pharmaceutical Researchers and Manufacturing Association, in 2010 US pharmaceutical R&D investments were $40.7 billion (80.2% of the global total), Chinese investments were $142 million (0.3% of the global total) and Indian investments were $43.9 million (0.1% of the global total)[5]—adds further doubt to the innovative capacity of these countries. For these emerging economies to achieve future growth in innovation similar to that seen in developed economies, clearly a larger corps of inventors and much increased R&D investment will be necessary.

Another explanation for the low output is that innovation in emerging economies is fundamentally different from the proprietary model of Western biopharmaceutical innovation, which requires patenting and high levels of R&D investment upfront. If innovation in emerging economies really is different, neither inventorship of patents nor R&D investment would be a good proxy for innovation.

In this respect, a report from the Massachusetts Institute of Technology (MIT) Taskforce on Innovation and Production, contrasting product development in China and Germany, may be informative[6]. The task force made the following two observations. First, new business creation in Germany was not done through the start-up model familiar in the United States, but rather occurred "through the transformation of old capabilities and their reapplication, repurposing and commercialization." Thus, the strength of German firms was in repurposing existing assets.

Second, the task force further observed that Chinese firms, by contrast, excelled in scale-up to mass-manufacturing "not because of low-cost labor, but because of their ability to move complex advanced product designs into production and commercialization." This sentiment is echoed in *Run of the Red Queen*[7], where the authors combined over 200 interviews and industry analysis to conclude that in China process innovation, rather than product innovation, is central to economic growth.

Accordingly, one must consider that the contributions of China, and of other countries outside regions that have an already established biopharmaceutical sector, will be in areas beyond inventing new molecules; rather, their strengths may be in developing better tools and methods for research or in developing better methods to refine the patented inventions. China is a world leader in scientific publishing[8] and in patent filings[9], neither of which seem to be delivering proportional outputs. Chinese papers have low citations rates, both domestically and internationally[7], and (as shown here) Chinese inventors appear on few patents covering globally marketed biopharmaceuticals. In other words, China's current strategy appears to be based on promoting traditional outputs, which do not support its core strengths.

With the above factors in mind, governments in countries such as China and India that have a low rate of drug inventorship and a proven ability to reduce costs beyond simply providing low-cost manual labor may be directing their resources inappropriately. Rather than focusing policy and funding on historical drivers of Western innovation such as patents and publications, their economic goals may be better achieved by directing policy and translational funding to maximize drivers of non-Western innovation that are unique to their own territories.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

Yali Friedman

DrugPatentWatch.com and Thinkbiotech, Washington, DC, USA.
e-mail: info@thinkbiotech.com

1. Jelinek, M. *et al. Res. Technol. Manag.* **55**, 16–6 (2012).
2. Code of Federal Regulations, Title 21 C.F.R. § 314.53.
3. Friedman, Y. *Nat. Rev. Drug Discov.* **9**, 835–836 (2010).
4. Light, D.W. *Health Aff.* **28**, w969–w977 (2009).
5. Pharmaceutical Research and Manufacturers of America, Pharmaceutical Industry Profile 2012. http://www.phrma.org/sites/default/files/159/phrma_industry_profile.pdf (PhRMA; Washington, DC, April 2012)
6. Locke, R.M. & Wellhausen, R. *A Preview of the MIT Production in the Innovation Economy Report.* http://web.mit.edu/pie/news/PIE_Preview.pdf (Massachusetts Institute of Technology, Cambridge, MA, 2013).
7. Breznitz, D. & Murphree, M. *Run of the Red Queen Government, Innovation, Globalization, and Economic Growth in China* (Yale University Press, New Haven CT, 2011)
8. National Science Board. *Science and Engineering Indicators 2012.* NSB 12–01 (National Science Foundation, Washington, DC, 2012)
9. World Intellectual Property Organization. *2012 World Intellectual Property Indicators* (WIPO, Geneva, Switzerland, December 2012)

Metabolomic data streaming for biology-dependent data acquisition

To the Editor:

Over the past 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 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 in the R programming language, to a web-based platform with a graphical user interface[2]. This web-based platform, called 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 to analyze terabytes of data with XCMS Online (**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 is a robust platform for nonexperts and experts to perform metabolomic data processing. Despite the advantages of cloud-based data processing, however, a 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. 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 tandem mass spectrometry (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 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 occurs in parallel to the data acquisition. If each liquid chromatography–mass spectrometry (LC-MS) run is considered as a discrete 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 2 months by hundreds of unique users. (Note: we accessed data only from users who 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 four to 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 h to 72 h, but on average was 20 h, depending 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 determined that by using a streaming approach most of the data upload could occur in parallel with LC-MS data acquisition and be completed before the last LC-MS sample is analyzed. 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 h to fewer than 3 h, a reduction of sevenfold.

In the current version of the 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-min run (60-min run plus wash and re-equilibration), was 57 min. The total time saving would be the number of runs multiplied by the average upload time per run. When analyzing large data sets, 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. We installed a script (**Supplementary Software 1**) on the computer workstation of an Agilent (Santa Clara, CA) 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 archiving and retrieving of data as well as defining sample groups. Heightened security is achieved by encryption, and file checksums are compared upon completion of transfer to reduce the risk of file corruption. These scripts are available to all XCMS Online users via the website (https://xcmsonline.scripps.edu/nbt); each script will have slight modifications depending on the type of mass spectrometer.

As described above, in untargeted metabolomics, users typically acquire profiling data first. After the data are uploaded to XCMS Online and processed, an XCMS user can inspect the results and select for additional analysis, metabolomic features that have statistical values above a defined threshold (e.g., $P \leq 0.01$ and fold change > 2) as well as METLIN database hits. 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 data-analysis time. 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 practical; 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 of interest.

The introduction of data streaming offers an improvement upon the autonomous metabolomic workflow. Instead of acquiring MS^2 data for every metabolomic feature, we suggest that investigators acquire MS^2 data

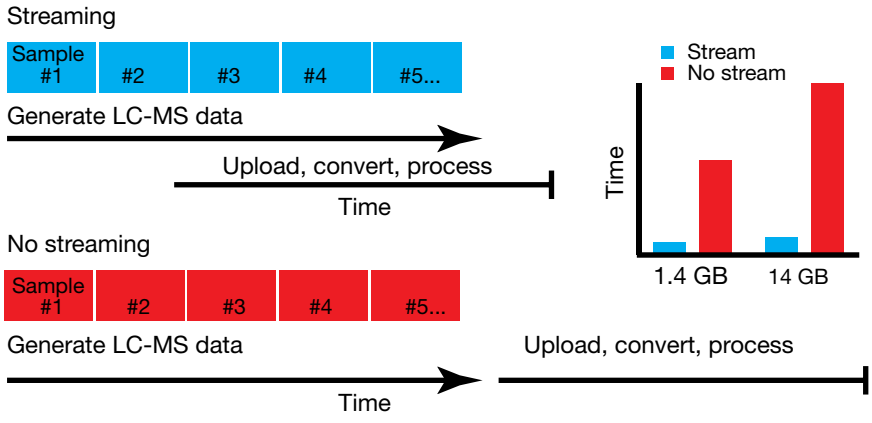

Figure 1 Time saved by metabolomic data streaming. XCMS-based data streaming workflow (top left) allows data upload and processing after each LC-MS run is performed, dramatically reducing the processing time after the data are acquired for the final sample (top right). A thousand XCMS Online data sets were examined for their average processing time without streaming. For low-resolution data (~1.4 GB) and high-resolution data (~14.0 GB) over 10 h and 20 h was required after the final LC-MS analysis was performed, respectively. Streaming allowed a sevenfold decrease in average processing time after data acquisition.

only for the features of interest based on predefined 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 biology-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 biology-dependent data acquisition.

To demonstrate XCMS Online–based streaming and the utility of biology-dependent data acquisition, we performed experiments on tumor samples and normal tissues using our existing XCMS Online platform (**Fig. 2** and **Supplementary Fig. 3**). For this comparison, we prepared 28 normal and tumor samples for LC-MS analysis. In brief, metabolites from 10 mg of tissue were isolated as described previously by using an acetone and methanol extraction, and analyzed by an Agilent QTOF instrument[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 biology-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 a mass-to-charge ratio (m/z) of the most statistically meaningful biology-dependent 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 biology-dependent data acquisition were set at $P \leq 0.001$, fold change ≥ 1.5 and intensity > 10,000 ion counts. Those ions that had accurate mass matches (<15 p.p.m.) to the METLIN metabolite database were 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 biology-dependent data acquisition, we wrote a script that enables automated metabolic pathway analysis (**Supplementary Software 2**). This script finds putatively identified metabolites (based on accurate mass) in the same metabolic pathway as those that are dysregulated and then prioritizes 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 (http://www.reactome.org/)[9], The Small Molecule Pathway Database (http://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**). Although we only demonstrated our approach by using Agilent instrumentation, data streaming and biology-dependent data acquisition can be performed on instruments from any vendor (Agilent (Santa Clara, CA), AB SCIEX (Framingham, MA), Thermo Fisher Scientific (Waltham, MA), Bruker (Billerica, MA) and Waters (Manchester, UK)) (**Supplementary Fig. 4**). Also, although our biology-dependent MS^2 acquisition is designed to generate data on 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 insights.

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 biology-dependent data acquisition. Here we demonstrated the benefits of data streaming for mass spectrometry–based metabolomics. We

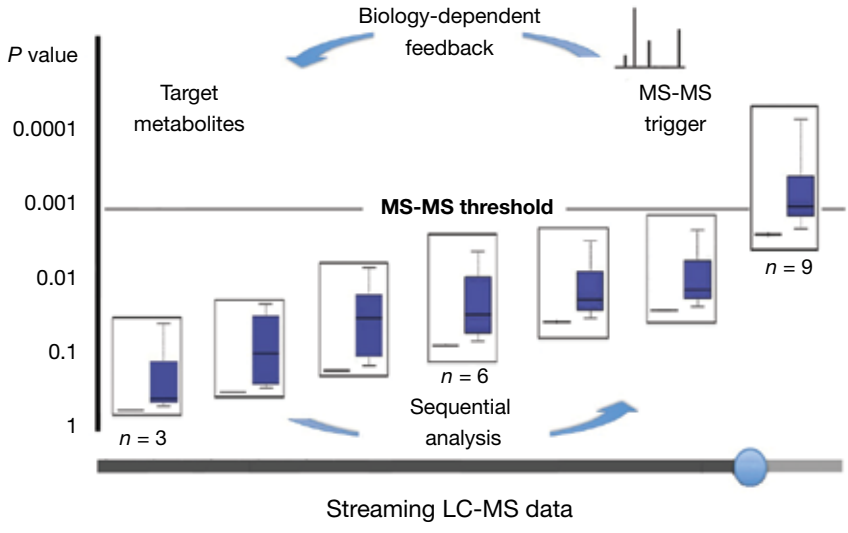

Figure 2 Biology-dependent data acquisition from tumor samples. Instead of using data-driven acquisition of MS^2 data that relies on intensity, signal-to-noise ratio (S/N) or prior acquisition of precursor ions, biology-dependent data acquisition relies on statistics generated after each sample run for 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 performed. 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 minimum–maximum; boxes, interquartile range).

expect that this concept could be extended to any experimental analysis requiring data upload and real-time feedback from cloud-based processing.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nbt.2927)

ACKNOWLEDGMENTS

This work was supported by the US National Institutes of Health grants 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 US National Institutes of Health National Institute on Aging grant L30 AG0 038036 (G.J.P.). Financial support was also received from the US Department of Energy grants FG02-07ER64325 and DE-AC0205CH11231 (G.S.).

AUTHOR CONTRIBUTIONS

D.R., T.N., C.H.J., J.I., H.P.B., J.L., A.P.A. and A.M.D. performed experimental work. D.R., T.N., C.H.J., J.I., J.L., G.J.P. and G.S. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

***Duane Rinehart*[1–4,9]*, Caroline H Johnson*[1–4,9]*, Thomas Nguyen*[1–4]*, Julijana Ivanisevic*[1–4]*, H Paul Benton*[1–4]*, Jessica Lloyd*[5–7]*, Adam P Arkin*[8]*, Adam M Deutschbauer*[8]*, Gary J Patti*[5–7] *& Gary Siuzdak*[1–4]

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, California, USA. [2]Department of Cell and Molecular Biology, The Scripps Research Institute, La Jolla, California, USA. [3]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA. [4]Center for Metabolomics, The Scripps Research Institute, La Jolla, California, USA. [5]Department of Chemistry, Washington University, St. Louis, Missouri, USA. [6]Department of Genetics, Washington University, St. Louis, Missouri, USA. [7]Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, USA. [8]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [9]These authors contributed equally to this work.

e-mail: gjpattij@washu.edu or siuzdak@scripps.edu

1. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. *Anal. Chem.* **78**, 779–787 (2006).
2. Tautenhahn, R., Patti, G.J., Rinehart, D. & Siuzdak, G. *Anal. Chem.* **84**, 5035–5039 (2012).
3. Kienzler, R., Buruggmann, R., Rangananthan, A. & Tatbul, N. in *Euro-Par 2011: Parallel Processing Workshops, Part II, LNCS 7156* (eds., Alexander, M. *et al.*) 467–476 (Springer, Berlin, Heidelberg, 2012).
4. Tautenhahn, R. *et al. Nat. Biotechnol.* **30**, 826–828 (2012).
5. Nikolskiy, I., Mahieu, N.G., Chen, Y.J., Tautenhahn, R. & Patti, G.J. *Anal. Chem.* **85**, 7713–7719 (2013).
6. Liu, H., Sadygov, R.G. & Yates, J.R. III. *Anal. Chem.* **76**, 4193–4201 (2004).
7. Neumann, S., Thum, A. & Bottcher, C. *Metabolomics* **9**, S84–S91 (2013).
8. Yanes, O., Tautenhahn, R., Patti, G.J. & Siuzdak, G. *Anal. Chem.* **83**, 2152–2161 (2011).
9. Matthews, L. *et al. Nucleic Acids Res.* **37**, D619–D622 (2009).
10. Frolkis, *A. et al. Nucleic Acids Res.* **38**, D480–D487 (2010).
11. Kamburov, A., Cavill, R., Ebbels, T.M., Herwig, R. & Keun, H.C. *Bioinformatics* **27**, 2917–2918 (2011).