Science & Society

Metabolizing Data in the Cloud

Benedikt Warth,[1,2]
Nadine Levin,[1,2]
Duane Rinehart,[1]
John Teijaro,[1]
H. Paul Benton,[1] and
Gary Siuzdak[1,*]

Cloud-based bioinformatic platforms address the fundamental demands of creating a flexible scientific environment, facilitating data processing and general accessibility independent of a countries' affluence. These platforms have a multitude of advantages as demonstrated by omics technologies, helping to support both government and scientific mandates of a more open environment.

Moving to the Cloud

We are increasingly surrounded by cloud computing, whether it is private or scientific in nature; thousands of computers and servers are used to process and handle our information [1]. The benefits of cloud-based computing over downloadable desktop-based software include straightforward data sharing, transfer, management and archiving, standardized data formats, distributed data processing, and global access that is independent of local high-end hardware [2,3]. Moreover, the cloud provides a universal platform for data analysis across communities and locations, enabling outside researchers to perform modified analyses and ask new questions utilizing the same data sets. These cloud-based platforms also make computational power available to academic institutions with limited resources, which otherwise may not be able to afford the infrastructure required to analyze complex data sets. This was recently demonstrated by the establishment of a low-cost and scalable experimentation platform to address professional training needs in basic biology [4]. Even high-end labs are starting to purchase time from cloud providers. Furthermore, these advantages coincide with recent government mandates seeking to make science more accessible and transparent [5].

The advantages of cloud computing are numerous, though it is not without its challenges; it requires reliable and fast Internet connections, creates a time lag between uploading and processing data, lacks flexibility and control over resources, and requires high-level security to ensure that proprietary information is safe [6,7]. Particularly, in multiple applications, data protection is an ongoing and controversial issue. The user does not know who might gain access to patient data and, once uploaded, cannot control its protection nor track its usage. Moreover, the missing availability of the source code can make cloud services a black box. Yet, even with these challenges, cloud computing has recently emerged as a cost-effective and more 'open' way to perform research in the field of life sciences. In fact, biology and cloud computing have become irrevocably intertwined, particularly for the omics disciplines, which deal with data sets reaching the terabyte range [2,7], and driven by the increasing performance of high-throughput methods, such as next-generation sequencing or high-resolution mass spectrometry-based metabolomics and proteomics [8]. Responding to the massive amounts of data generated, National Institutes of Health initiatives, including 'Big Data to Knowledge', and bioinformatics hubs are now exploring cloud-computing options.

Here, we describe cloud-based computing in metabolomics, which is currently facing the key challenge of integrating within the broader context of systems biology (Figure 1).

Metabolomics: Leading from Behind

Metabolomics, a multidisciplinary technology and the newest and fastest growing omic domain, aims to decipher metabolic changes in a given biological system. Despite overcoming challenges in its early stages of development and establishing community standards in the

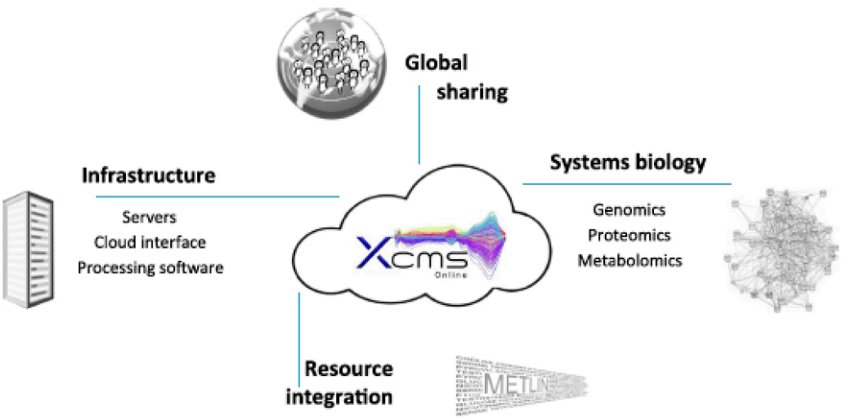

Figure 1. Cloud-Based Computing Provides Interconnectedness with Infrastructure, Data Sharing, and the Ability to Integrate Multiple Resources. The multiomics XCMS Online platform with over 12 500 registered users represents a freely available, cloud-based resource offering data processing, archiving and sharing, easy-to-use statistical tools, intuitive visualization, pathway analysis, and – in combination with the METLIN database – metabolite identification.

last decade [9], the field of metabolomics may still be regarded as an emerging scientific discipline. Given the scope of these efforts, significant time and funding resources are required to establish a comprehensive metabolomics lab, making the technology inaccessible to a large share of biological scientists without the required instrumentation or substantial bioinformatic support. To address the informatic challenge, XCMS Online [10] and the associated METLIN metabolite database [11] represent a model of how community-oriented cloud computing can be implemented on a global scale. This free, cloud-based platform for untargeted metabolomics data provides users with data processing tools, data streaming, statistical analysis, metabolite identification via high-resolution tandem mass spectrometry library searching, pathway-based multiomic analysis, and downloadable results for any alternative analyses. Originally developed as a command-line driven software (XCMS) [12], the online version requires neither programming skills nor expensive hardware. The challenges of cloud computing can be overcome by streamlining data uploads, using redundant grid-based processing and redundant secure data storage, and providing a user-friendly platform with enterprise-grade security that is compliant with established regulations [6]. Because data transfer still represents a major bottleneck in the metabolomics arena, we recently launched XCMS Stream [13]. This data streaming solution holds the potential to reduce data analysis time from days to minutes.

The most significant value of the universal access options provided by cloud computing is the fast and simple sharing of resources (Figure 2A). In the case of XCMS Online, uploaded liquid chromatography-mass spectrometry raw data files, as well as the results of any specific data processing job including the experimental parameters and settings, can be shared either privately with any collaborator or publicly with the entire community. Both options require only a few mouse clicks. Thus, this type of cloud-based platform also increases the field's ability to compare results with those of publicly shared jobs. Alongside XCMS Online, similar initiatives underpin the expanding role of cloud-based workflows in metabolomics and beyond. For example, MetaboAnalyst is a valuable collection of online tools for data analysis and interpretation [14], and Galaxy, a platform with roots in genomics and transcriptomics, is currently diversifying into metabolomics [15]. The recently funded US$8 million Horizon2020 project 'PhenoMeNal' aims to build a community-supported e-infrastructure based on cloud resources for medical metabolomics applications [Edmunds, S. (2016) https://blogs.biomedcentral.com/gigablog/2016/07/19/guest-posting-building-phenomenal-metabolomics-e-infrastructure]. It will support data processing and analysis for molecular phenotype data and leverage existing cloud infrastructures and data repositories. Although this platform is so far dedicated to the European biomedical community, these efforts again confirm the potential of cloud computing in the omics area in general and the primary role of metabolomics in particular.

Common alternatives to cloud computing are downloadable, vendor-specific software tools to process and evaluate metabolomic data sets. While these software solutions offer advantages, they are typically very costly and sometimes are not compatible with data formats acquired on competitor instruments. With an XCMS Online user population of over 12 500 scientists in 120+ countries, we estimate the cost savings of the scientific community to be over US$70 million since its launch in 2011 (Figure 2B), while the cost for development and maintenance of XCMS Online has been below US$2 million, resulting in a community return on investment of approximately 97%. These savings are expected to increase to over US$250 million by 2021 by extrapolating across the current user base. These calculations are based on the average cost from four different vendors over the course of the last 5 years and the assumption that there would be one license for every three XCMS Online

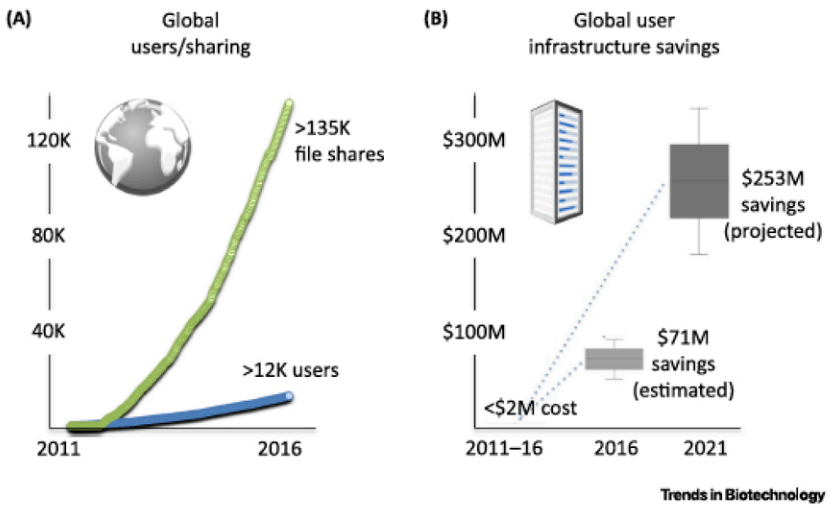

Figure 2. XCMS Online Exemplifies the Value of Cloud-Based Computing. This resource has demonstrated an ability to analyze biological systems using multiomic technologies, with its growing usage a result of (A) ease of data sharing and (B) significant cost savings. A conservative estimated cost savings (B) to the community is about US$70 million, as calculated based on the average cost of comparable downloadable software from different vendors over the last 5 years and the number of XCMS Online users (assuming three users per purchased program). The increasing importance of data sharing (A) is demonstrated by the ratio of registered users to shared data sets.

users. Importantly, these estimates are conservative, as additional costs for software updates or the need for the acquisition of more than one program were not taken into account.

It is worth noting that in an alternative scenario, in which a lab invests in trained informatics personnel applying open-source software or creating in-house scripts, the anticipated budget far exceeds that of purchased software solutions. In addition, in-house software further increases the cost to the field because many academic developers release software with little to no documentation. Besides the financial burden, exporting and sharing data and results with colleagues and the public require additional efforts and are less convenient or even impossible when using vendor programs or in-house scripts. Often, side-step analyses and crucial parameters can be hidden or forgotten, causing issues for replicating published findings. Hence, sharing standardized data sets including underlying raw data utilizing cloud-based solutions clearly supports the increasing demand for reproducibility of scientific data and the request of publishers and grant bodies to make data and software publicly accessible.

The Path Upward

Overall, the value of cloud-based bioinformatic platforms goes beyond convenience and addresses the fundamental demands of putting analytical raw data into a biological context and creating a more open scientific environment for rapidly translating science. As stated by Schadt *et al.* [2], the future success of the biomedical and life sciences will depend on the ability of investigators to properly interpret large-scale, multidimensional data sets that are generated by high-throughput technologies. From this perspective, freely accessible, cloud-based platforms are invaluable to facilitate appropriate data processing and handling. Making more metabolomics data accessible will trigger integration with other high-content data sets such as those in genomics and proteomics. New software to promote the integration and analysis of these data sets could generate more in-depth insights. Having a central 'hub' for these large data sets and a way to integrate the information contained would be an invaluable asset to the scientific community.

The success of metabolomic platforms will likely evolve and expand into other fields, such as clinical diagnostics, pharmacokinetics, imaging, molecular modeling, education, and the social sciences, to name only a few though challenges in data protection will certainly remain to some extent. Because cloud-based tools are inherently easier to access, newly launched platforms should also stimulate synergies through smartly connecting resources. One particularly interesting possibility is the adoption of these technologies in less-affluent countries, where the necessary resources are unavailable. Given these unique advantages, it is inevitable that cloud-based computing will become an integral part of scientific communities.

Acknowledgments

The authors would like to thank their colleagues from the Siuzdak lab for valuable discussions and acknowledge the following for funding assistance: The Austrian Science Fund (FWF): J-3808 (Erwin Schrödinger Fellowship awarded to B.W.), 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 grants R01 GMH4368 and PO1 A1043376-02S1.

[1]Center for Metabolomics and Departments of Chemistry, Molecular and Computational Biology, Immunology and Microbial Science and Chemical Physiology, The Scripps Research Institute, La Jolla, CA, USA
[2]These authors contributed equally

*Correspondence: siuzdak@scripps.edu (G. Siuzdak).

http://dx.doi.org/10.1016/j.tibtech.2016.12.010

References

1. Marx, V. (2013) Biology: the big challenges of big data. *Nature* 498, 255–260
2. Schadt, E.E. *et al.* (2010) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 11, 647–657
3. Buyya, R. *et al.* (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. *Future Gener. Comput. Syst.* 25, 599–616
4. Hossain, Z. *et al.* (2016) Interactive and scalable biology cloud experimentation for scientific inquiry and education. *Nat. Biotech.* 34, 1293–1298
5. Levin, N. and Leonelli, S. (2016) How does one "open" science? Questions of value in biological research. *Science, Technology and Human Values* Published online October 3, 2016. http://dx.doi.org/10.1177/0162243916672071
6. Datta, S. *et al.* (2016) Secure cloud computing for genomic data. *Nat. Biotech.* 34, 588–591
7. Rinehart, D. *et al.* (2014) Metabolomic data streaming for biology-dependent data acquisition. *Nat. Biotech.* 32, 524–527
8. Johnson, C.H. *et al.* (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell. Biol.* 17, 451–459
9. Fiehn, O. *et al.* (2007) The metabolomics standards initiative (MSI). *Metabolomics* 3, 175–178
10. Tautenhahn, R. *et al.* (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* 84, 5035–5039
11. Tautenhahn, R. *et al.* (2012) An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotech.* 30, 826–828
12. Smith, C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787
13. Montenegro-Burke, J.M. *et al.* (2016) Data streaming for metabolomics: accelerating data processing and analysis from days to minutes. *Anal. Chem.* Published online December 16, 2016. http://dx.doi.org/10.1021/acs.analchem.6b03890.
14. Xia, J. *et al.* (2015) MetaboAnalyst 3.0 making metabolomics more meaningful. *Nucleic Acids Res.* 43, W251–W257
15. Boekel, J. *et al.* (2015) Multi-omic data analysis using Galaxy. *Nat. Biotech.* 33, 137–139