OXFORD

Briefings in Bioinformatics, 2024, **25(S1)**, bbae478
https://doi.org/10.1093/bib/bbae478
Editorial

Editorial

NIGMS Sandbox: a learning platform toward democratizing cloud computing for biomedical research

Abstract

Biomedical data are growing exponentially in both volume and levels of complexity, due to the rapid advancement of technologies and research methodologies. Analyzing these large datasets, referred to collectively as "big data," has become an integral component of research that guides experimentation-driven discovery and a new engine of discovery itself as it uncovers previously unknown connections through mining of existing data. To fully realize the potential of big data, biomedical researchers need access to high-performance-computing (HPC) resources. However, supporting on-premises infrastructure that keeps up with these consistently expanding research needs presents persistent financial and staffing challenges, even for well-resourced institutions. For other institutions, including primarily undergraduate institutions and minority serving institutions, that educate a large portion of the future workforce in the USA, this challenge presents an insurmountable barrier. Therefore, new approaches are needed to provide broad and equitable access to HPC resources to biomedical researchers and students who will advance biomedical research in the future.

Keywords: NIGMS; Sandbox; learning; platforms; democratizing; cloud computing

Introduction

Cloud computing, i.e. *on-demand and highly scalable computing and storage* on clustered high-power servers managed by service providers such as Google Cloud, Amazon Web Services, and Microsoft Azure, holds the promise of releasing institutions from the burden of owning and continuously upgrading information technology infrastructure to provide high-performance-computing capacity for research. Cloud service providers (CSPs), with resources as well as incentives to invest in updating and innovation, offer the public their services in a pay-as-you-go manner. As consumers, academic and/or research institutions pay for needed resources and computing time based on actual usage. Paying on a per-use-basis for needed cloud capacity and computing time, instead of paying the premium of standing up, maintaining, updating, and operating data centers, is more manageable financially for many institutions. In addition to cost savings, cloud computing has the potential to increase the research productivity of research-intensive institutions. For example, the time needed to complete a project on servers of an on-premises data center can be dramatically reduced by the "near-infinite" parallelization available on commercial cloud infrastructure. Furthermore, user fees incurred by sponsored research activities or training activities are also payable by most research or training grants, respectively, enabling institutions to support research and/or training activities without carrying the financial burden for unnecessary infrastructure or capacity. To strengthen public cloud's service capacity for biomedical research and encourage biomedical researchers to explore use of the cloud, the National Institutes of Health (NIH) hosts more than 250 petabytes of biomedical research data and secured group discounts with major CSPs [1–3].

Despite the vast potential cloud computing holds in providing broad access to big data and scalable data analytic capacity, most biomedical researchers have not been able to take advantage of this technology for their research. Equally if not more concerning is that biomedical students, including graduate students interested in independent research careers, are lacking opportunities or support to learn skills that enable them to use cloud computing. To better understand the biomedical research community's needs and look for ways to overcome barriers to cloud access, NIH's National Institute of General Medical Sciences (NIGMS) issued a Request for Information (RFI) and organized a follow-up workshop for institutional leaders, faculty researchers, postdoctoral fellows, and students of institutions either supported by the NIGMS Institutional Development Award (IDeA) program or that have a historical mission in serving minority students. Through the RFI responses and the workshop, we learned that faculty investigators primarily relied on their institutional servers to access and analyze data, and that, in general, they had not been trained to navigate the cloud. Consequently, biomedical students lacked exposure and mentoring in learning cloud computing skills for research. The consensus among the responses to the RFI and the workshop participants was that new approaches are needed to help democratize cloud computing for biomedical research.

Development of the NIGMS Sandbox

We hypothesized that a cloud-based learning platform may provide a gateway for biomedical investigators and students to start their journey to the cloud. The platform would include a collection of technologies and/or methodologies, each covered in a free-standing learning module, commonly used in biomedical research but not offered through curricula of all institutions. This platform

Received: September 9, 2024. **Revised:** September 10, 2024. **Accepted:** September 12, 2024
Published by Oxford University Press 2024.

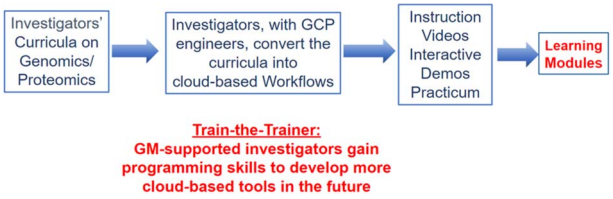

Figure 1. The workflow of building cloud-based learning modules.

would provide access to many users, enabling them to choose the modules that meet their needs and execute them on-demand. The modules would deliver the learning materials in an interactive format and use appropriate cloud resources for data access and analyses so that users would learn the materials at their own pace while gaining experience and skills in navigating the cloud. The platform would also pilot a cloud account management mechanism to assess costs and explore affordable means to sustain the democratization of cloud computing for biomedical research. The platform is named "NIGMS Sandbox for Cloud-based Learning" https://github.com/NIGMS/NIGMS-Sandbox, as it captures the essence of a learning environment that encourages users to learn through "playing."

The Sandbox was built through a collaboration among NIH scientists and technical staff, NIGMS-funded investigators, cloud engineers, bioinformaticians, and project managers from Google Cloud and Deloitte Consulting. To map out a path for the large collaborative team to achieve the aims of this program, two investigator groups, each supported by a supplemental award to an IDeA Network of Biomedical Research Excellence (INBRE) grant, teamed up with cloud engineers and bioinformaticians from Google Cloud and Deloitte to convert an RNA-Seq and a proteomic analysis curriculum, respectively, into cloud-based workflows. They then developed instruction videos to introduce the scientific concepts of these technologies or methods and created interactive demos to walk through the workflow step-by-step and to explain the objectives and cloud tools used for each step. To help users solidify their learning, the team also designed practicums that allow users to practice and solve problems. These efforts yielded two functional modules [4, 5] though which users can learn these technologies and analyze their own research data (Fig. 1). Importantly, the NIGMS-funded biomedical investigators gained software development and cloud engineering skills through working with the software and cloud engineers, which enabled them to develop additional cloud-based tools in the future and to train others to do so (Fig. 1).

Following a similar pathway, NIGMS funded 10 more investigator teams to work with cloud engineers and bioinformaticians from Google Cloud and Deloitte to develop 10 additional learning modules covering a wide range of topics relevant to biomedical research, ranging from fundamentals of bioinformatics to biomarker discovery. The 11 modules in Table 1 collectively represent the first release of the NIGMS Sandbox, which can be accessed publicly through https://github.com/NIGMS/NIGMS-Sandbox. The Sandbox is an open platform that, with continued improvements and researcher contributions, can support many more learning modules that NIGMS is committed to develop to serve the broader biomedical research community. This overarching GitHub repository includes an introduction of the overall design of the Sandbox, supporting documents, and how-to guides such as those on how to find a Google Cloud Project and how to spin up a Vertex AI notebook. A link for the GitHub repository for each of the 11 modules is also provided.

It is important to note that the computational requirements of each module differ. For example, some modules use the base Google Cloud Platform machine image with CPU-based machines with Python kernels, while others require R kernels; some modules launch virtual machines from custom Docker images, while others require GPU-based machines. As such, users are encouraged to review and evaluate the required compute environment for each module before using a module. Furthermore, when switching from one module to another, it is recommended to review differences in configuration in order to provision the appropriate environment. Each of the 11 protocols published in this supplemental issue of *Briefings in Bioinformatics* describes the development of one module and corresponding instructions for using it, including Allers *et al.* [4] on the analysis of bacterial gene expression data using bulk RNA sequencing, O'Connell *et al.* [5] on understanding proteome quantification, Wilkins *et al.* [6] on BASH coding for biologists, Qin *et al.* [7] on whole-genome bisulfite sequencing data analysis, Seaman *et al.* [8] on efficient de novo transcriptome assembly, Veerappa *et al.* [9] on the analysis of pooled-cell and single-cell ATAC-seq data, Nguyen *et al.* [10] on consensus pathway analysis, Ruprecht *et al.* [11] on transcriptomics and epigenetic data integration, Gnimpieba *et al.* [12] on biofilm marker discovery with dockerized metagenomics analysis of microbial communities, Hemme *et al.* [13] on biomarker discovery, and Woessner *et al.* [14] on identifying and training deep learning neural networks on biomedical-related datasets.

Table 1. The 11 NIGMS Sandbox learning modules included.

1	RNA-seq Differential Expression Analysis [4]
2	Proteome Quantification [5]
3	Fundamentals of Bioinformatics [6]
4	DNA Methylation Sequencing Analysis with WGBS [7]
5	Transcriptome Assembly Refinement and Applications [8]
6	ATAC-Seq and Single Cell ATAC-Seq Analysis [9]
7	Consensus Pathway Analysis in the Cloud [10]
8	Integrating Multi-Omics Datasets [11]
9	Metagenomics Analysis of Biofilm-Microbiome [12]
10	Analysis of Biomedical Data for Biomarker Discovery [13]
11	Biomedical Imaging Analysis using AI/ML approaches [14]

Deployment of the NIGMS Sandbox

A learning platform for biomedical research like the NIGMS Sandbox, while publicly available, is unlikely to be an effective gateway to cloud computing on its own, because the intended users typically do not have an institutionally sponsored cloud account to access the platform and pay for the cloud computing time required for learning. For the platform to reach students and investigators broadly, we need to bring the platform to them and develop account management methods to keep track of usage and pay user fees to cloud computing service providers. Taking advantage of the ready infrastructure supported by the NIGMS INBRE grants, we utilized the NIH Cloud Lab to provide students and investigators supported by these grants access to the Sandbox. NIH Cloud Lab is a program developed by NIH's CIT, and sponsored by NIH's ODSS as part of the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. The STRIDES initiative provides NIH and NIH-funded users secure access to cloud services, on-demand technical assistance, open-source GitHub documentation and tutorials, and tracking of the cloud service usage by users. Through NIH Cloud Lab, users have full access to the cloud console, including all offered services by Google Cloud, Amazon Web Services, and Microsoft Azure, with service

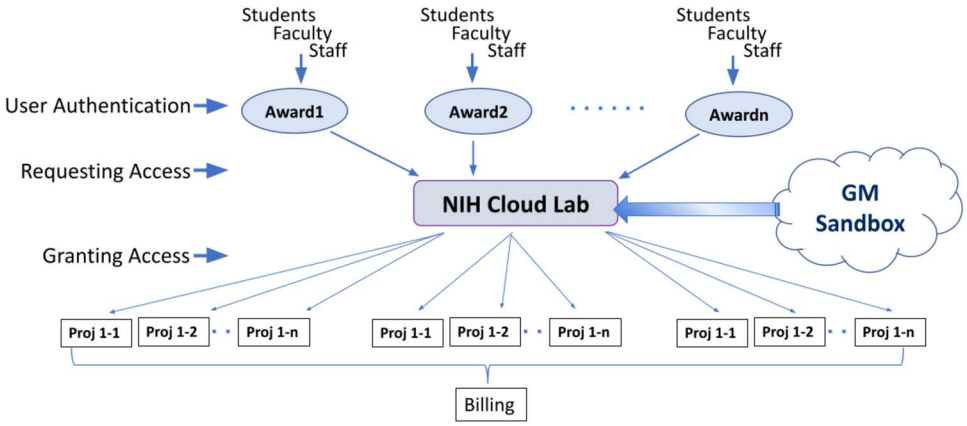

Figure 2. **Access and use of NIGMS Sandbox via the cloud lab; t**he following steps were followed to access and use the NIGMS Sandbox: first, institutions supported by NIGMS-funded programs identified users among their students, faculty, or staff (user Authentication); second, the institutions forwarded requests to the NIH cloud lab team (Requesting Access); third, the NIH cloud lab team granted each authenticated user an NIH cloud lab account (Granting Access), which enabled users to access modules in the Sandbox, identify module(s) of interest, and clone them into the cloud lab environment for learning.

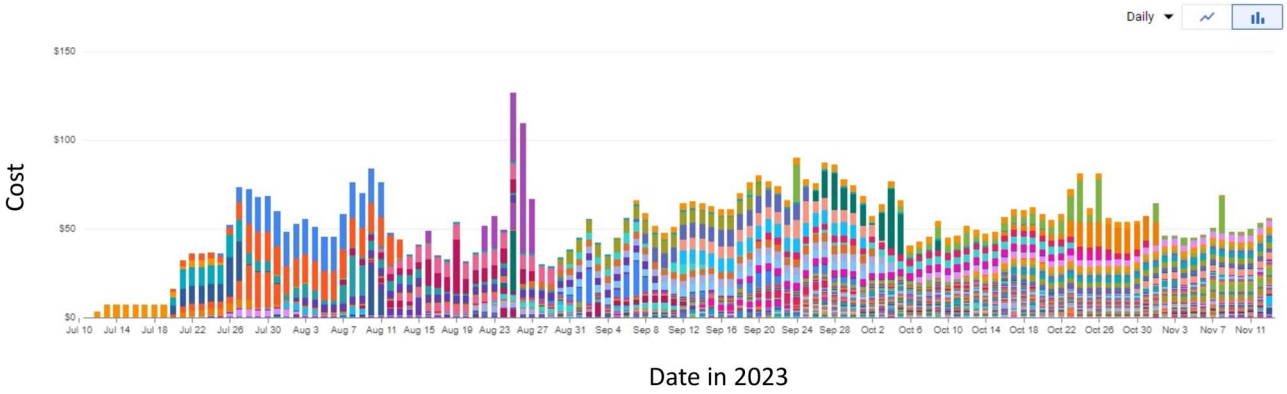

Figure 3. The daily billing graph provided by the Google Cloud console serves as a proxy for account usage; each colored band represents the spending of a unique user for a particular day; engagement per user has increased through time.

usage and incurred costs tracked, although the NIGMS Sandbox currently only uses Google Cloud.

How users access the Sandbox through NIH Cloud Lab is illustrated in Fig. 2. The cost incurred by each user in this pilot is paid by the NIH as an investment toward NIH's strategic goal of democratizing cloud computing for biomedical research. More than 300 students and faculty from NIGMS supported institutions were granted access to NIH Cloud Lab accounts to execute the Sandbox modules in three cohorts each lasted for 1 month. At the end of the 3 months, the NIGMS Sandbox modules collectively were cloned from GitHub over 1000 times for learning experiences. These cloud-based learning experiences were affordable since each learning modules only costs a few dollars to execute. Figure 3 shows the costs for cloud computing time incurred by each user during that period. The NIGMS Sandbox modules have also been used in bioinformatics workshops as well as classroom instructions by NIGMS-funded faculty investigators. Going forward, NIGMS plans to continue improving and expanding the Sandbox offering, including updating existing modules, developing new learning modules covering additional topics, and deploying the Sandbox through other CSPs. These efforts will explore innovative approaches to provide cost-affordable access of the Sandbox to students and faculty beyond NIGMS-supported institutions.

Conclusion

The ability to access and analyze big data is becoming essential for biomedical research. Cloud computing through commercial cloud services may provide this capacity broadly with manageable costs and significant economies of scale, thus enabling more students and investigators to participate in biomedical research and analyze big data. We hope our effort in developing a cloud-based learning platform will help catalyze the democratization of cloud computing for biomedical research.

Ming Lei[1,*,‡], *Lakshmi K Matukumalli*[1], *Krishan Arora*[1], *Nick Weber*[2], *Rachel Malashock*[2], *Fenglou Mao*[3], *Susan Gregurick*[3], *Jon Lorsch*[1]

[1]*National Institute of General Medical Sciences, National Institutes of Health, 9000 Rockville Pike, Bethesda, Marylnd 20892, USA*
[2]*Center for Information Technology, National Institutes of Health, 9000 Rockville Pike, Bethesda, Marylnd 20892, USA*
[3]*Office of Data Science Strategy, Office of Director; National Institutes of Health, 9000 Rockville Pike, Bethesda, Marylnd 20892, USA*

[*]*Corresponding author. Ming Lei, Office of Research and Graduate Education, West Virginia University Health Sciences, G102, 108 Biomedical Road, P O Box 9104 Morgantown, WV 26506-9104. Tel.: (304) 293-7206; E-mail: ming.lei@hsc.wvu.edu*
[‡]*Present address: West Virginia University Health Sciences, G102, 108 Biomedical Road, P O Box 9104 Morgantown, WV 26506–9104.*

Achnowledgements

We thank Dave Belardo, Anne Billak, Marcia Price, and Sam Russ of Google; Antej Nuhanovic, Thad Carlson, Madeline Cowan, Kyle O'Connell, Ross Campbell, Dina Mikdadi, Olga Robinson, Juergen Klenk, and Vivien Bonazzi of Deloitte; Joel Mills of Covalent Solutions; Dana Gaffney of NIH/CIT STRIDES; and Alison Lin and Raphael Isokpehi of NIH/ODSS for supporting the development and deployment of the NIGMS Sandbox. We also appreciate Kyle O'Connell for providing input to this manuscript.

Author contributions

All authors were involved in the conception and execution of the NIGMS Sandbox project. M.L., L.M., and K.A. wrote the manuscript, and all authors approved of the final version of the manuscript.

Conflict of interest: None declared.

Funding

Authors of this article are NIH staff members. Their efforts related to this work are parts of their regular duty supported by the NIH.

References

1. Dahlquist JM, Nelson SC, Fullerton SM. Cloud-based biomedical data storage and analysis for genomic research: landscape analysis of data governance in emerging NIH-supported platforms. *HGG Adv* 2023;**4**:1–13. https://doi.org/10.1016/j.xhgg.2023.100196.
2. Holko M, Weber N, Lunt C. *et al.* Biomedical research in the cloud: considerations for researchers and organizations moving to (or adding) cloud computing resources. *Pac Symp Biocomput* 2023, 536–40. http://psb.stanford.edu/psb-online/proceedings/psb23/wkshop_cloud.pdf.
3. Navale V, Kaeppler DV, McAuliffe M. An overview of biomedical platforms for managing research data. *J Data Inform Manag* 2021;**3**:21–7. https://doi.org/10.1007/s42488-020-00040-0.
4. Allers S, O'Connell KA, Carlson T. *et al.* Reusable tutorials for using cloud-based computing environments for the analysis of bacterial gene expression data from bulk RNA sequencing. *Brief Bioinform* 2024;**25**:bbae301. https://doi.org/10.1093/bib/bbae301.
5. O'Connell KA, Kopchick B, Carlson T. *et al.* Understanding proteome quantification in an interactive learning module on Google Cloud Platform. *Brief Bioinform* 2024;**25**:bbae235. https://doi.org/10.1093/bib/bbae235.
6. Wilkins OM, Campbell R, Yosufzai Z. *et al.* Cloud-based introduction to BASH coding for biologists. *Brief Bioinform* 2024;**25**:bbae244. https://doi.org/10.1093/bib/bbae244.
7. Qin Y, Maggio A, Hawkins D. *et al.* Whole genome bisulfite sequencing data analysis learning module on Google Cloud Platform. *Brief Bioinform* 2024;**25**:bbae236. https://doi.org/10.1093/bib/bbae236.
8. Seaman RP, Campbell R, Doe V. *et al.* A cloud-based training module for efficient *de novo* transcriptome assembly using Nextflow and Google cloud. *Brief Bioinform* 2024;**25**:bbae313. https://doi.org/10.1093/bib/bbae313.
9. Veerappa AM, Rowley MJ, Maggio A. *et al.* Cloud ATAC: a cloud-based interactive framework for ATAC-Seq data analysis. *Brief Bioinform* 2024;**25**:bbae090. https://doi.org/10.1093/bib/bbae090.
10. Nguyen H, Pham V-D, Nguyen H. *et al.* CCPA: cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and Reactome. *Brief Bioinform* 2024;**25**:bbae222. https://doi.org/10.1093/bib/bbae222.
11. Ruprecht NA, Kennedy JH, Bansal B. *et al.* Transcriptomics and epigenetic data integration learning module on Google Cloud. *Brief Bioinform* 2024;**25**:bbae352. https://doi.org/10.1093/bib/bbae352.
12. Gnimpieba EZ, Hartman TW, Do T. *et al.* Biofilm marker discovery with cloud-based dockerized metagenomics analysis of microbial communities. *Brief Bioinform* 2024;**25**:bbae429. https://doi.org/10.1093/bib/bbae429.
13. Hemme CL, Beaudry L, Yosufzai Z. *et al.* A cloud-based learning module for biomarker discovery. *Brief Bioinform* 2024;**25**:bbae126. https://doi.org/10.1093/bib/bbae126.
14. Woessner AE, Anjum U, Salman H. *et al.* Identifying and training deep learning neural networks on biomedical-related datasets. *Brief Bioinform* 2024;**25**:bbae232. https://doi.org/10.1093/bib/bbae232.