HHS Public Access

Author manuscript

Biotechniques. Author manuscript; available in PMC 2022 April 07.

Published in final edited form as:

Biotechniques. 2022 February ; 72(2): 36–38. doi:10.2144/btn-2021-0060.

Democratizing bioinformatics through easily accessible software platforms for non-experts in the field

Konstantinos Krampis[*,1]

[1]Department of Biological Sciences, Hunter College, City University of New York, NY, USA

Keywords

bioinformatics; cloud computing; data visualization; genomics; machine learning

Next-generation genome sequencing (NGS) technology is currently at a point where we can sequence genomes with unprecedented scale and accuracy, and it will be only a matter of time before we also have fully integrated, low-cost sequencing sample preparation, essentially being able to load DNA or RNA samples on the sequencer and generate data with the push of a button. Furthermore, with the sequencing cost significantly reduced in recent years, NGS technology has become a commodity technology in biomedical laboratories, and in the very near future it will be fully integrated into every aspect of medical and clinical practice. However, a significant bottleneck still exists, as the adoption of NGS technology for genomic research requires employing bioinformatics experts for deploying complex software and implementing computational infrastructure for bioinformatics data analysis. This is a key barrier to progress toward the democratization and broad adoption of the technology, especially by researchers in underrepresented institutions, lacking access to bioinformatics core facilities or the financial resources to hire experts for data analysis.

The development of standardized, scalable data analysis software by the bioinformatics community that is easily accessible by non-experts has alleviated some of this bottleneck. These computing platforms have accelerated biomedical research based on NGS data, by providing easy access to state-of-the-art data analysis algorithms applicable to a range of genomic applications. From the perspective of clinical practitioners and biomedical researchers without access to core facilities or software experts, these platforms require minimal effort during installation; provide an intuitive user interface for running bioinformatics algorithms and mining NGS data for answering basic research hypotheses or discovering clinical targets for genomic medicine; include a simple yet powerful software mechanism for algorithm version tracking, update and configuration; enable easy management of genomic data collections and supporting datasets for bioinformatics

[*]Author for correspondence: kk104@hunter.cuny.edu.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2021-0060

analysis, including genome indexes, assemblies and gene variants, among others; and provide a complete analysis and training infrastructure that will be accessible and portable across computing platforms ranging from institutional computer clusters to desktop computers in biomedical research laboratories, by leveraging cloud-computing and software virtualization technology.

Published platforms in the field that have these characteristics and leverage cloud computing for bioinformatics include Dugong [1], an implementation of Galaxy in Docker [2], BioPortainers [3], AlgoRun [4], GUIdock [5], DockerBIO [6] and Docker containers released through the Galaxy project [7]. However, while some of the aspects of a platform to help non-experts easily perform NGS data analysis can be found dispersed across those systems, namely the simplicity of using a desktop program, seamless deployment on various host computational platforms and a rich set of visualizations for mining the NGS data and the output of the bioinformatics pipelines. Specifically, BioPortainers and AlgoRun require users to perform command line installations [8,9], which, besides requiring significant technical expertise, are not available in all operating systems. In the case of GUIdock and especially DockerBIO, a significant effort has been made to provide easy installation and a user-friendly software interface. However, the former system is targeted only to a single application (cellular network analysis), while the latter depends on community developers of the Docker containers to providing the software and defining the user interface, while neither implements visualization capabilities. The Galaxy project has had great adoption by the community and requires a variable rate of a learning curve, depending on the user, while being an open-ended system for developing online bioinformatics portals or being used at sequencing centers where bioinformatics developers can customize it with pre-configured pipelines [10]. Finally, the Cloud BioLinux project [11] demonstrated early on that cloud-computing technology can increase the usability of bioinformatics software for non-experts, and Cloud BioLinux has been used for distributing pre-configured, accessible and ready-to-use bioinformatics software. While the funding for the project has expired and virtual machines are not actively maintained, the full open source code for Cloud BioLinux is still available through the project’s website, enabling users to utilize and improve the functionality in accordance with their specific application. Improving upon this technology, the BioDocklets [12] project demonstrated that further abstraction of complex bioinformatics operations can be achieved, helping non-experts to execute complex bioinformatics data analysis pipelines, with the simplicity of running a desktop software program.

In addition to the core bioinformatics data analysis pipelines, many of these platforms also provide intuitive visualizations of the data, something required for lowering the barrier for genomic discoveries from the NGS data. The visualizations can run, for example, on desktop computers or smartphone/tablet computing platforms, which are easy to use for clinicians and researchers to perform data-driven genomic discoveries. Some implementations also leverage the latest web development and smartphone application programming technologies, in order to provide a bioinformatics visualization system that is self-contained and runs without any required installation or external software dependencies. Supplementary Figure 1 shows the Visual Omics Explorer (VOE) [13], where users are able to simply load the application on a web browser or access it as a smartphone/tablet app without any remote

server connection. VOE visualizations are fully integrated with the bioinformatics pipelines and are easily shareable or exported as publication-ready graphics and integrative analysis of NGS data generated from their experiments, with data available in public databases such as The Cancer Genome Atlas (TCGA) or the ENCODE projects [14,15].

Users of such bioinformatics platforms will have variable knowledge of genomic technologies, and therefore will need to complement their backgrounds accordingly in order for them to efficiently perform bioinformatics data analysis. Therefore, a set of targeted training materials should be provided for all bioinformatics analysis pipelines, visualizations and related tasks to be performed using these platforms, covering a range of genomic technologies from metagenomics to cancer variant discovery. These could be structured as virtual short courses published using online teaching platforms such as Coursera [16-20] or MIT Courseware [21]. Following familiarization with a specific platform and corresponding training for their specific type of data, biomedical and clinical researchers should be able to easily perform integrative bioinformatics analysis without any prior bioinformatics expertise, achieving proficiency at performing analysis of large-scale genomic datasets with a small investment of time and resources.

In recent years, throughput from genomic sequencing projects has grown exponentially, following the constant drop in cost of the technology in addition to broad access to sequencing services. As a result, while genomic data released from large, publicly funded sequencing projects have reached the petabyte scale in size, data interpretation and extraction of scientific value from the research community still face significant bottlenecks. Sequencing instruments are typically bundled with only minimal computational and storage capacity, sufficient for data capture during runs of the instrument, and complex bioinformatics analyses are required in post-processing and the generation of scientific insights from the raw sequencing data. These analyses involve bioinformatics specialists and software engineers with specific technical skills and training; additionally, access to significant computational capacity is needed in order to process and store large-scale genomic datasets. Research laboratories in smaller, underrepresented institutions experience a significant bottleneck in finding the financial and time resources to put a bioinformatics infrastructure and teams of specialists in place, preventing them from participating fully in the genomics revolution and equal opportunities for scientific discoveries. A solution for alleviating this bottleneck has been the publication of genomic data analysis platforms by the bioinformatics community, providing easy access to pre-configured software that reduces the funds required to hire bioinformatics specialists and enables these institutions to budget the analysis as a fixed cost. A key factor in enabling this was the availability of cloud-computing services, which further helped democratize the bioinformatics field by providing access to the ample computational capacity required for genome data analysis to smaller, independent laboratories as well as large-scale bioinformatics core facilities.

In conclusion, the current and future genomic data analysis platforms will enable biomedical researchers and clinical practitioners at underfunded, minority universities and research institutions to integrate NGS and bioinformatics as a standard component of biomedical research and clinical applications. With further development of software platforms that fully abstract the complexity of bioinformatics operations by the research community, we

can remove the data analysis bottleneck and contribute to democratizing access to genomic-based, data-centric research for investigators at these underrepresented institutions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Financial & competing interests disclosure

This project was supported by TUFCCC/HC Regional Comprehensive Cancer Health Disparity Partnership, award number U54 CA221704 (5) from the National Cancer Institute of National Institutes of Health (NCI/NIH). Its contents are solely the responsibility of the author and do not necessarily represent the official views of the NCI/NIH. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1. Menegidio FB, Jabes DL, Costa de Oliveira R, Nunes LR. Dugong: a Docker image, based on Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses. Bioinformatics 34(3), 514–515 (2018). [PubMed: 28968637]
2. Digan W, Countouris H, Barritault M et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. GigaScience 6(11), gix099 (2017).
3. Menegidio FB, Aciole Barbosa D, Goncalves RD et al. Bioportainer Workbench: a versatile and user-friendly system that integrates implementation, management, and use of bioinformatics resources in Docker environments. GigaScience 8(4), giz041 (2019). [PubMed: 31222200]
4. Hosny A, Vera-Licona P, Laubenbacher R, Favre T. AlgoRun: a Docker-based packaging system for platform-agnostic implemented algorithms. Bioinformatics 32(15), 2396–2398 (2016). [PubMed: 27153722]
5. Hung LH, Kristiyanto D, Lee SB, Yeung KY. GUIdock: using Docker containers with a common graphics user interface to address the reproducibility of research. PLoS One 11(4), e0152686 (2016). [PubMed: 27045593]
6. Kwon C, Kim J, Ahn J. DockerBIO: web application for efficient use of bioinformatics Docker images. Peer J. 27(6), e5954 (2018).
7. Afgan E, Baker D, Van den Beek M et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44(W1), W3–10 (2016). [PubMed: 27137889]
8. BioPortainers installation instructions. https://bioportainer.github.io/BioPortainer/
9. Algorun installation instructions. http://algorun.org/documentation/
10. Galaxy Docker image. https://hub.docker.com/r/bgruening/galaxy-stable/
11. Krampis K, Booth T, Chapman B et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinform. 13(1), 1–8 (2012).
12. Kim B, Ali T, Lijeron C, Afgan E, Krampis K. Bio-Docklets: virtualization containers for single-step execution of NGS pipelines. GigaScience 6(8), gix048 (2017).
13. Kim B, Ali T, Hosmer S, Krampis K. Visual Omics Explorer (VOE): a cross-platform portal for interactive data visualization. Bioinformatics 32(13), 2050–2052 (2016). [PubMed: 27153572]
14. The Cancer Genome Atlas, TCGA data processing pipelines. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/
15. ENCODE project, data processing pipelines. https://www.encodeproject.org/pipelines/
16. Coursera. Bioinformatics specialization. https://www.coursera.org/specializations/bioinformatics/
17. Coursera. Data science specialization. https://www.coursera.org/specializations/jhu-data-science/

18. Coursera. Genomic data science specialization. https://www.coursera.org/specializations/genomic-data-science/
19. Coursera. Biology meets programming: bioinformatics for beginners. https://www.coursera.org/learn/bioinformatics/
20. Coursera. 'Bioinformatics' search results. https://www.coursera.org/search?query=bioinformatics/
21. edX. Bioinformatics courses. https://www.edx.org/learn/bioinformatics/