Focus on Computer Resources

Cancer Research

The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research

Jessica W. Lau, Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick Groves-Kirkby, Aleksandar Mihajlovic, Jack DiGiovanna, Mladen Srdic, Dragan Bajcic, Jelena Radenkovic, Vladimir Mladenovic, Damir Krstanovic, Vladan Arsenijevic, Djordje Klisic, Milan Mitrovic, Igor Bogicevic, Deniz Kural, and Brandi Davis-Dusenbery; for The Seven Bridges CGC Team

Abstract

The Seven Bridges Cancer Genomics Cloud (CGC; www.cancergenomicscloud.org) enables researchers to rapidly access and collaborate on massive public cancer genomic datasets, including The Cancer Genome Atlas. It provides secure on-demand access to data, analysis tools, and computing resources. Researchers from diverse backgrounds can easily visualize, query, and explore cancer genomic datasets visually or programmatically. Data of interest can be immediately analyzed in the cloud using more than 200 preinstalled, curated bioinformatics tools and workflows. Researchers can also extend the functionality of the platform by adding their own data and tools via an intuitive software development kit. By colocalizing these resources in the cloud, the CGC enables scalable, reproducible analyses. Researchers worldwide can use the CGC to investigate key questions in cancer genomics. *Cancer Res; 77(21); e3–6. ©2017 AACR.*

Introduction

As the size and complexity of cancer genomic datasets continue to grow, the availability of scalable compute resources (i.e., the "cloud") facilitates rapid and cost-effective data analysis (1). The Seven Bridges Cancer Genomics Cloud (CGC; www.cancergenomicscloud.org) was funded as a pilot project by the NCI to explore novel approaches to democratize access to massive cancer genomic datasets alongside the tools and computational resources to analyze them. The CGC was publicly launched in February 2016 and is open to all cancer researchers worldwide, who can create a free profile online or log in via their eRA Commons or NIH Center for Information Technology account.

The CGC enables researchers to quickly access The Cancer Genome Atlas (TCGA; ref. 2), which contains genomic, transcriptomic, and clinical data from more than 11,000 cancer patients. TCGA has contributed greatly to understanding the molecular basis of cancer and identifying novel therapeutic targets (3, 4). However, downloading and storing TCGA requires significant time and resources. Furthermore, querying and using the data can be challenging for researchers without adequate computational resources or appropriate technical knowledge. The CGC addresses these challenges to make TCGA and other large cancer genomics datasets usable by a wide range of cancer researchers.

Methods: Scalable Cancer Genomics Analysis in the Cloud

The CGC comprises an intuitive interface for rapidly accessing and using large public genomic datasets, comprehensive security controls, preloaded bioinformatics tools and workflows, access to scalable cloud-based computation, a software development kit, an application programming interface (API) for automation, data visualization and querying tools, and extensive support for collaborative, reproducible research (Fig. 1A, Supplementary Video S1). Built on Amazon Web Services' enterprise level cloud services, the CGC provides researchers with secure access to public genomic datasets [including TCGA and the Cancer Cell Line Encyclopedia (CCLE; ref. 5)], alongside the high-performance computation needed to analyze them. Because the data are stored together with cloud computing resources, analysis is readily scalable on-demand, allowing thousands of samples to be quickly analyzed (Fig. 1B).

The hosted datasets are complemented by more than 200 biomedical data analysis tools and workflows, including pipelines for variant calling on whole genome and exome sequencing data, differential expression analysis on RNA sequencing data, and complex data visualization. These curated tools are continually revised to include updated versions, according to demand as determined by user interviews and surveys. Tools on the CGC are

Seven Bridges Genomics, Cambridge, Massachusetts.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corrected online July 31, 2018.

Corresponding Author: Brandi Davis-Dusenbery, Seven Bridges Genomics Inc., 1 Main St., Suite 500, Cambridge, MA 02142. Phone: 617-294-6582; E-mail: brandi@sevenbridges.com

doi: 10.1158/0008-5472.CAN-17-0387

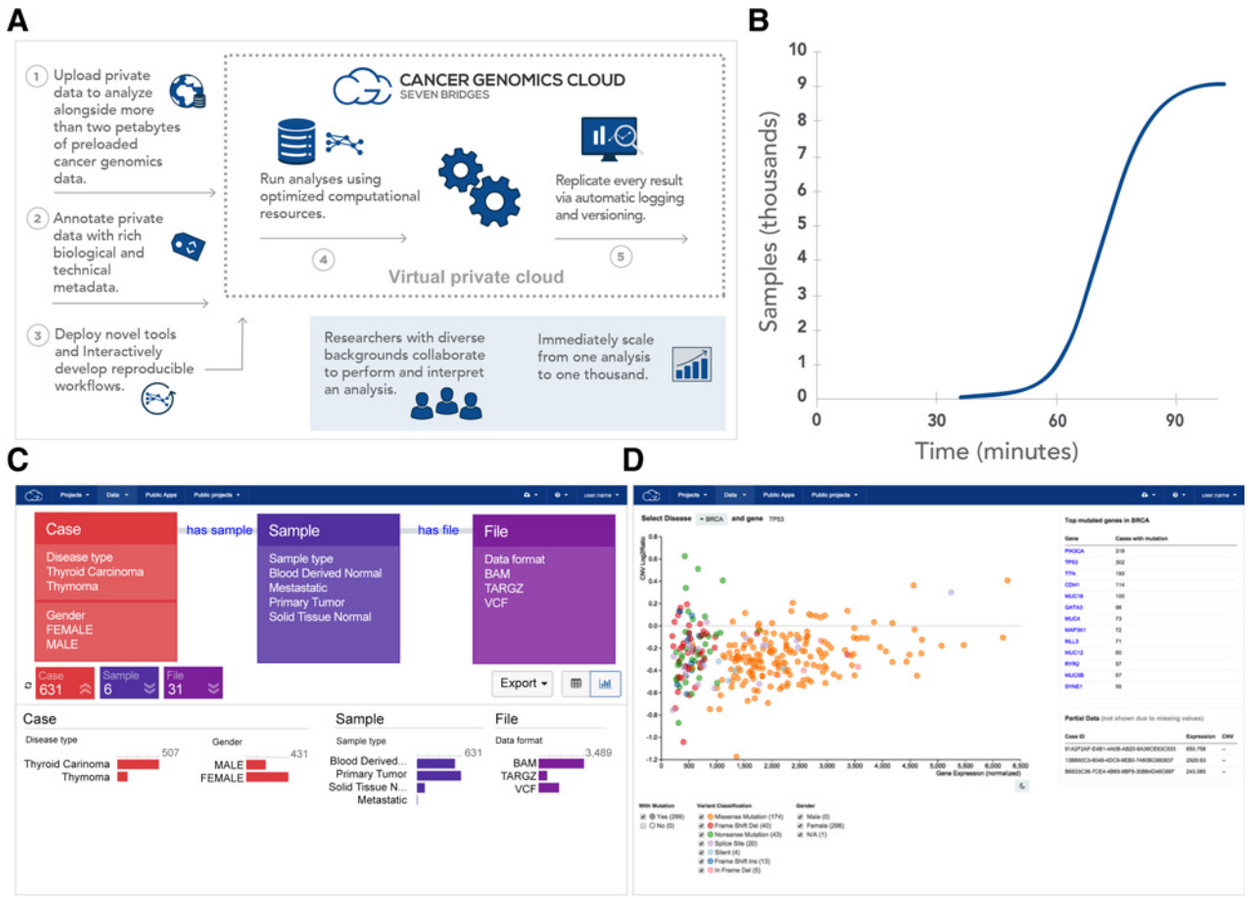

Figure 1.

The Cancer Genomics Cloud is designed to enable scalable cancer genomics research, with features that support usability. **A,** In addition to hosting public cancer genomics datasets and providing curated analysis tools, the CGC enables users to upload and annotate their own data, as well as integrate their own tools. Data analyses are run using optimized computing resources, and executions are recorded to ensure reproducibility. **B,** Time course of an RNA-Seq quantitation experiment in which more than 9,000 samples were analyzed in parallel. All samples were completed within 100 minutes. **C,** The visual Data Browser allows users to explore and select data by specifying properties of interest. Once selected, these files can be added to a project for analysis. **D,** The interactive Case Explorer allows users to visualize and select cases based on type of cancer, genes of interest, types of mutations, and more.

packaged within Docker containers, a lightweight software virtualization technology (www.docker.com). Execution instructions are described using Common Workflow Language (CWL; www.commonwl.org), an open-source, community-developed specification for describing analysis workflows and tools in a way that is portable and scalable across software and hardware environments. The CGC also offers researchers a robust software development kit, which enables them to easily describe their own tools and custom scripts in CWL for use on the platform. A visual workflow editor allows users to intuitively build reproducible workflows from individual tools.

The CGC offers a suite of technologies for visualizing, querying, and exploring to identify data of interest within complex datasets. By using a Semantic Web approach to link more than 140 clinical, biospecimen, and analysis metadata properties, the CGC enables researchers to build complex queries both visually and programmatically (Fig. 1C). This allows scientists to quickly access, for example, all RNA sequencing count files from normal and tumor samples taken from patients with thyroid cancer who were treated with local radiotherapy. Importantly, this approach is readily extendable to support multistudy integration. We have made data from the CCLE available with TCGA, and we will continue to add further important public datasets. Although the Data Browser allows users to explore datasets based on metadata, the Case Explorer focuses on the genetic properties of the data and allows global views of gene expression, copy number variation, and gene mutation status (Fig. 1D).

Methods: Collaborative, Reproducible, Extensible, and Scalable

The CGC has been designed to support best practice, reproducible scientific research at scale. On the basis of our engagements and collaborations with cancer researchers, and our experience in user-centric software development, we identified key principles to guide the design and implementation of the platform. We suggest that any modern biomedical "knowledge cloud" should follow similar principles.

First, public cancer genomic data need to be usable by all collaborators. Few institutions have the resources to download and manage TCGA, and specialized skills are required to manipulate the data. In contrast, users of the CGC can immediately begin

to explore and analyze more than a petabyte of cancer genomics data through a simple web interface. Users benefit from visual tools and guides, documentation, a dedicated user support team, and the continual development of the Seven Bridges core infrastructure upon which the CGC is built. We regularly review the usability of the CGC and incorporate feedback from researchers to create new solutions and features. In addition, collaboration between multiple researchers and institutions promotes scientific advances (6). By colocalizing data, computation, and analysis, the CGC enables distributed, multidisciplinary teams to collaborate on data analysis. Shared project spaces allow approved collaborators to access the same data and workflows, and to see the same results.

Second, reproducibility is required throughout the entire research workflow, from data management to analysis. Replication studies highlight the difficulty of reproducing a number of key studies in cancer biology (7, 8). The CGC ensures reproducibility of computational analyses by recording all aspects of data analysis, including files used, tool versions, and parameter settings. Moreover, because workflows are defined using CWL, they can be readily reproduced by collaborators, reviewers, and journal editors across different computational environments. The CGC currently supports CWL 1.0 via the API, and we are in the process of updating the CGC with Rabix (9), our open-source CWL executor capable of testing and running all CWL applications in any Unix-based environment. Rabix integration will further facilitate developer testing and interoperability with external platforms.

Third, the impact of large public cancer genomics datasets is extended by new tools and data. Besides providing curated biomedical data analysis tools and workflows, we created a software development kit for users to add their own tools and workflows, and a RESTful (10) API supported in multiple languages (Python, R). Using these tools, researchers on the CGC have developed custom pipelines for a variety of scientific applications, including characterizing tumor microbiomes, epitope identification, and segmenting patient populations by gene sequence. Researchers can extend the utility of TCGA data by adding their own private datasets and analyzing them together with the same workflows. Robust upload utilities (including a visual interface, command line, and API) use all available client bandwidth while ensuring secure data transfer. Once uploaded, data can be annotated with more than 40 standard metadata properties or any number of custom properties. Integration and semantic mapping of user-defined metadata to a common ontology is an area of active development.

Finally, data analyses must be scalable to fully make use of available datasets. The elasticity of the CGC infrastructure means that as data analysis scales up, additional computational resources are allocated to enable parallelization and processing of batch jobs. Cloud computing costs have decreased by more than 80% over the past 10 years (11), making it the most cost-efficient way to analyze large genomic datasets (1). As an example, one researcher was able to perform targeted variant calling across the 11,000 TCGA participants in about 3 hours for under $15.

Results: User Uptake and Impact

Within the 15 months since the launch of the CGC, over 1,900 researchers have registered on the platform, representing 150 institutions across 30 countries. In total, CGC users have deployed more than 5,000 tools or workflows and performed 80,000 executions, representing over 97 years of total computation. There is significant collaboration among users, with an average of seven members per project on the platform.

The CGC enables a diverse range of research (12, 13). For example, last year, Diermeier and colleagues (12) reported on potentially oncogenic mammary tumor–associated RNAs (MaTAR). After identifying and characterizing MaTARs using mouse models of breast cancer, they confirmed the relevance of human MaTAR orthologs in clinical breast cancer with the CGC. By analyzing RNA-Seq data available in TCGA, the researchers found that some of the MaTARs were upregulated in breast tumors.

Discussion

The CGC makes massive cancer genomic datasets available and usable for research, while safeguarding privacy and security. This approach is readily extensible. In addition to hosting TCGA, we have made CCLE data available and empowered users to integrate their private data with these public resources. New genomics datasets, including TARGET, CGCI, and Simons diversity data as well as new types of data (imaging and proteomic data), are being added to further extend the utility of the system. The CGC represents a successful model for democratizing access to and use of massive public datasets, allowing users to maximize their research productivity. Biomedical "knowledge clouds" like this can serve as the gateway to a wide ecosystem of interoperable cloud resources to support scientific discovery.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: E. Lehnert, G. Kaushik, Z. Onder, J. Radenkovic, D. Klisic, M. Mitrovic, D. Kural, B. Davis-Dusenbery

Development of methodology: E. Lehnert, R. Malhotra, G. Kaushik, Z. Onder, J. Radenkovic, D. Klisic, I. Bogicevic, D. Kural, B. Davis-Dusenbery

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A. Mihajlovic, D. Bajcic, V. Mladenovic, B. Davis-Dusenbery

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Sethi, R. Malhotra, A. Mihajlovic, D. Bajcic, J. Radenkovic, V. Arsenijevic, D. Klisic, B. Davis-Dusenbery

Writing, review, and/or revision of the manuscript: J.W. Lau, E. Lehnert, A. Sethi, R. Malhotra, G. Kaushik, Z. Onder, N. Groves-Kirkby, J. DiGiovanna, D. Bajcic, V. Mladenovic, D. Krstanovic, V. Arsenijevic, B. Davis-Dusenbery

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): E. Lehnert, A. Sethi, G. Kaushik, J. DiGiovanna, M. Srdic, D. Bajcic, V. Mladenovic, V. Arsenijevic, B. Davis-Dusenbery

Study supervision: D. Krstanovic, V. Arsenijevic, D. Kural, B. Davis-Dusenbery

Acknowledgments

We thank the entire Seven Bridges team, the Cancer Genomics Cloud Pilot teams from the NCI, the Broad Institute, and the Institute of Systems Biology, the Genomic Data Commons team, countless early users, and data donors. We also thank Laura Tramontozzi for assistance with the figure.

We also wish to further acknowledge the source of two of the datasets that are available to authorized users through the CGC and that were central to its development: The Cancer Genome Atlas (TCGA, phs000178). The resources described here were developed in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at https://cancergenome.nih.gov/. And Therapeutically Applicable

Research to Generate Effective Treatments (TARGET, phs000218). The resources described here were developed in part based on data generated by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative managed by the NCI. The data were obtained from the Genomic Data Commons (https://gdc.cancer.gov/) and corresponded to GDC Data Release 7.0 (https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-70). Information about TARGET can be found at https://ocg.cancer.gov/programs/target.

Grant Support

The Cancer Genomics Cloud is powered by Seven Bridges and has been funded in whole or in part with federal funds from the NCI, NIH, Department of Health and Human Services, under contract no. HHSN261201400008C.

Received February 6, 2017; revised April 5, 2017; accepted July 17, 2017; published online November 1, 2017.

References

1. Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO. Data analysis: create a cloud commons. Nature 2015;523:149–51.
2. The future of cancer genomics. Nat Med 2015;21:99.
3. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
4. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. Nature 2017;543:378–84.
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
6. Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. Science 2007;316:1036–9.
7. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. Elife 2014;3:e04333.
8. Nosek BA, Errington TM. Making sense of replications. Elife 2017;6;e23383.
9. Kaushik G, Ivkovic S, Simonovic J, Tijanic N, Davis-Dusenbery B, Kural D. Rabix: an open-source workflow executor supporting recomputability and interoperability of workflow descriptions. Pac Symp Biocomput 2016;22:154–65.
10. Fielding R. "CHAPTER 5: Representational State Transfer (REST)". Architectural Styles and the Design of Network-based Software Architectures. Irvine, CA: University of California; 2000. Available from: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm.
11. Barr J. AWS Storage Update – S3 & Glacier Price Reductions + Additional Retrieval Options for Glacier. Amazon Web Services; 2016. Available from: https://aws.amazon.com/blogs/aws/aws-storage-update-s3-glacier-price-reductions/.
12. Diermeier SD, Chang KC, Freier SM, Song J, El Demerdash O, Krasnitz A, et al. Mammary tumor-associated RNAs impact tumor cell proliferation, invasion, and migration. Cell Rep 2016;17:261–74.
13. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol 2017;35:319–21.