Cancer Research

The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research

Sheila M. Reynolds[1], Michael Miller[1], Phyliss Lee[1], Kalle Leinonen[1], Suzanne M. Paquette[1], Zack Rodebaugh[1], Abigail Hahn[1], David L. Gibbs[1], Joseph Slagel[1], William J. Longabaugh[1], Varsha Dhankani[1], Madelyn Reyes[2], Todd Pihl[2], Mark Backus[2], Matthew Bookman[3,4], Nicole Deflaux[3,4], Jonathan Bingham[3,4], David Pot[2], and Ilya Shmulevich[1]

Abstract

The ISB Cancer Genomics Cloud (ISB-CGC) is one of three pilot projects funded by the National Cancer Institute to explore new approaches to computing on large cancer datasets in a cloud environment. With a focus on Data as a Service, the ISB-CGC offers multiple avenues for accessing and analyzing The Cancer Genome Atlas, TARGET, and other important references such as GENCODE and COSMIC using the Google Cloud Platform. The open approach allows researchers to choose approaches best suited to the task at hand: from analyzing terabytes of data using complex workflows to developing new analysis methods in common languages such as Python, R, and SQL; to using an interactive web application to create synthetic patient cohorts and to explore the wealth of available genomic data. Links to resources and documentation can be found at www.isb-cgc.org. *Cancer Res; 77(21); e7–10. ©2017 AACR.*

Introduction

The National Cancer Institute's Cancer Genome Atlas (TCGA) program (1) has been a groundbreaking effort in many respects and has shown conclusively that cancer data analysis has outgrown the traditional model based on using only locally available resources. To explore cloud-based solutions while also seeking to democratize access to these valuable datasets, the National Cancer Institute (NCI) created three Cancer Genomics Cloud (CGC) Pilots. The ISB-CGC, a joint effort of the Institute for Systems Biology, with commercial partners Google and CSRA, is one of these.

The overarching goal of the ISB-CGC Cloud Pilot is to provide an environment in which different types of users can access and explore the full breadth and depth of TCGA and TARGET (Therapeutically Applicable Research to Generate Effective Treatments) data, and make use of a variety of cloud-based tools and technologies to run existing methods or develop new ones. Data-security is critical. Access to the low-level DNA and RNA sequence data must be carefully controlled and monitored. At the same time, we have created an open system that encourages cloud-based computational analyses of both the controlled-access data and the associated open-access data. Cloud technologies and bioinformatics standards have evolved rapidly in the past decade and will continue to do so; it is critical to be able to rapidly integrate advances in the field. The ISB-CGC has continuously evolved as we seek to maximize the use of existing infrastructure and tools, to collaborate with ongoing community-based efforts, and to offer users analytical approaches that can be easily tailored to their needs.

Use Cases

ISB-CGC aims to serve the full breadth of the cancer research community: algorithm developers who require hundreds or thousands of virtual machines (VM) to analyze terabytes of sequence data; computational research scientists who write custom scripts using languages such as R, Python, or SQL, and access data through APIs; and biologists or clinicians who prefer interactive web-based applications. These different access modalities are highlighted in Supplementary Video S1.

The original motivation for the cloud pilots was to support the needs of researchers to compute on the petabyte-scale DNA and RNA sequence datasets by bringing the computation to a single copy of the data in the cloud. Sequence data is typically provided in one of two standard file formats [BAM (2) and FASTQ (3)], and a broad range of algorithms and methods exist that operate on these types of files. Beyond the sequence data, however, exists a wealth of heterogeneous data, typically available in less-standardized form. As a pilot project, ISB-CGC is working with the cancer research community to explore use

[1]Institute for Systems Biology, Seattle, Washington. [2]CSRA Inc., Falls Church, Virginia. [3]Google, Mountain View, California. [4]Verily Life Sciences, South San Francisco, California.

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

Corresponding Authors: Sheila M. Reynolds, Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109. Phone: 206-732-1354; Fax: 206-732-1299; E-mail: sreynolds@systemsbiology.org; and Ilya Shmulevich, ilya.shmulevich@systemsbiology.org

doi: 10.1158/0008-5472.CAN-17-0617

cases and to provide examples illustrating approaches that take advantage of cloud-computing. We describe three broad categories of use cases below.

File-based analysis

The types of analyses that have been run on the raw sequence data hosted by the ISB-CGC have included variant-calling on normal DNA samples, targeted *de novo* assembly, antigen receptor analysis, DNA or RNA realignment to novel genomes or transcriptomes, and RNA quantification. The ISB-CGC is also being used to serve the ICGC-TCGA SMC-RNA DREAM challenge (4) data and to evaluate submissions. Between June 2016 and May 2017, over 14,000 terabytes of controlled-access TCGA sequence data have been accessed from the ISB-CGC object-store. Less than 1% of the data accessed was downloaded, evidence that ISB-CGC users are performing their analyses "near" the data, achieving NCI's objective for these Cloud Pilots.

A simple and cost-effective approach to running large-scale analyses on sequence data involves parallelizing tasks using the Google "Pipelines" service combined with low-cost preemptible VMs (5). Recently released methods enable more targeted usage of these large BAM files. IGV (6) supports direct, credentialed access to BAM files in Cloud Storage from anywhere, and SAMtools (2) allows "bam-slicing" of files in cloud-based object-stores.

Query-based analysis (SQL)

The "high-level" (derived) information such as somatic mutation calls, gene expression estimates, and copy-number segments is smaller in size by several orders of magnitude than the low-level sequence data, and is more heterogeneous, of greater interest to a larger number of researchers, and frequently open access. However, these data were originally distributed in hundreds of thousands of small files and were cumbersome to work with. To facilitate integrative and pan-cancer analyses of these types of data, the ISB-CGC team created a consistent set of tables that can be queried using SQL from a web-based interface, or from scripting languages, such as R and Python.

Alongside these data tables, ISB-CGC also hosts a variety of open-access reference sources [including GENCODE (7), Ensembl (8), miRBase (9), Kaviar (10), and others] to facilitate and encourage a wide range of use cases. Users can easily upload their own data (over which they control access) and perform integrative analyses of their data in the context of the ISB-CGC hosted data. As a model for making valuable data sources available in a commercial cloud, including those under restrictive licensing, the ISB-CGC has teamed up with the Welcome Trust Sanger Institute to provide the COSMIC (11) database. Users register with the Sanger Institute and are then granted access to this resource on ISB-CGC.

To encourage usage of this resource, the ISB-CGC documentation includes numerous example queries. From June 2016 through May 2017, the ISB-CGC open-access database resources (both data and reference tables) were accessed by over 250,000 unique queries.

Web-based interactive analysis

The ISB-CGC web application provides an interactive portal that allows users to create synthetic cohorts based on clinical and experimental metadata and then visualize and explore the associated molecular data. Within a visualization, users may also create cohort subsets by drag-selecting samples of interest. Built-in set operations allow users to create logical cohort combinations.

In a typical end-to-end scenario making use of many of the features of the ISB-CGC platform, a user might begin by using the web application to explore the available data and to create a custom cohort of TCGA cases diagnosed with gastric or esophageal cancer who harbor a *TP53* mutation. She may then switch to the application programming interface (API) to fine-tune the saved cohort and to obtain a list of relevant data files matching certain criteria, e.g., WXS DNA-Seq hg38 BAM files. Having already prepared a dockerized version of the analysis method to be applied to these files, she would then deploy this method to be run in parallel, using one VM per file, with inputs and outputs automatically fetched from and written to Cloud Storage. Finally, the analysis results are uploaded into a private BigQuery table where a colleague begins looking for significant associations between this new information and the existing clinical and molecular data for these TCGA cases.

System Architecture

The high-level architecture of the ISB-CGC is illustrated in Fig. 1. The cloud-based platform consists of these major components:

- Data as a Service (DaaS), a cloud-based data repository, which includes data from multiple large public cancer programs and is distributed between two Google Cloud Platform services, Cloud Storage and BigQuery.
- Authentication, authorization and accounting (AAA), a lightweight security layer that combines OAuth 2.0 with a SAML-based interface to the NIH federated login system, allowing users to authenticate using a Google identity, link it to their eRA Commons (or NIH) identity, and have their dbGaP authorization automatically verified.
- Interactive access. An interactive web-based application connects users to the authentication layer and allows them to explore and visualize the available data and create and share custom "cohorts" by filtering on clinical information, data-availability, and certain molecular features.
- Programmatic access. An API allows users to programmatically query and retrieve information stored in a Cloud SQL instance used by the security layer and the web application.
- Compute power. Users can provision hundreds or thousands of VMs, as needed, paying only for the cores, memory, and persistent disk space they need.
- Software tools, examples, and documentation. Users can learn how to use the system by running and extending open source code examples on GitHub and by following the online tutorials.

The ISB-CGC APIs are deployed with Google Cloud Endpoints and the web-application and authentication layer are deployed on App Engine. Both of these technologies provide automatic load balancing and scaling to ensure that users will experience consistent performance. In addition, the ISB-CGC resources can be used in conjunction with any existing Google-native tools, services, and interfaces, including BigQuery and Cloud Storage.

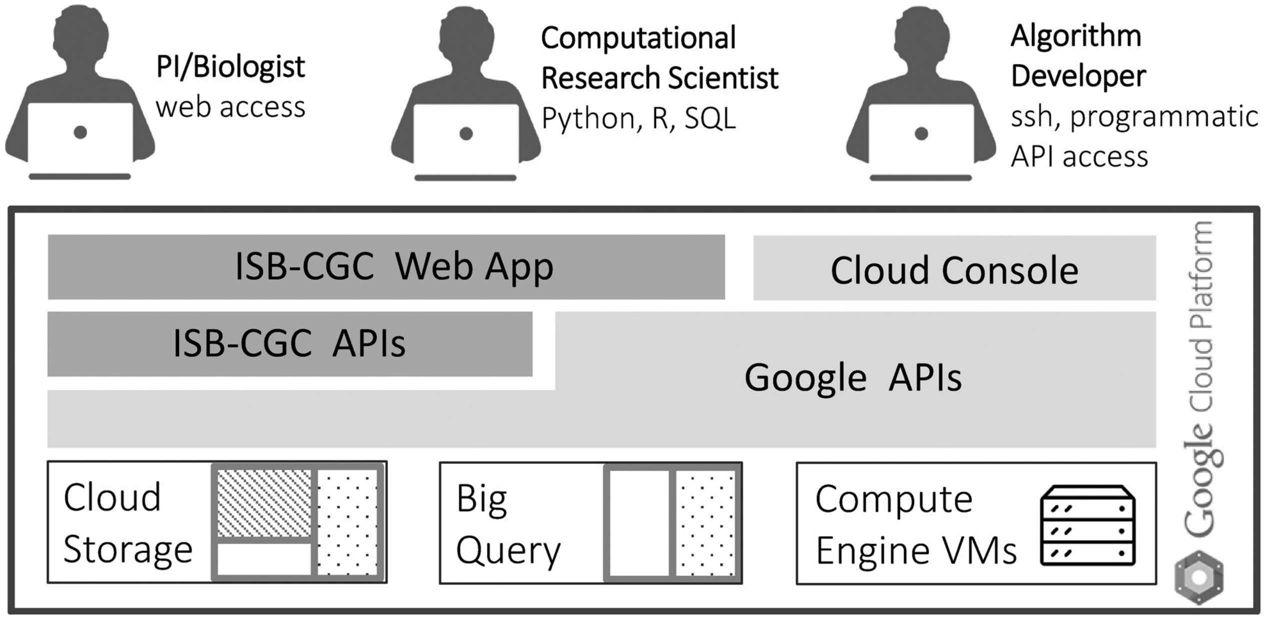

Figure 1.

The ISB-CGC interactive web application and programmatic interfaces provide access to over 2 petabytes of data hosted in Google Cloud Storage and BigQuery. Users can explore and analyze hosted datasets, upload their own data, and deploy workloads on Compute Engine VMs using a combination of ISB-CGC and Google interfaces.

Data resources

As of May 2017, the ISB-CGC data repository includes over 2 PB of data accessible to authenticated and authorized users, with a subset of the data available to all users. Depending on the data-type and the storage mechanism, these data are accessible as files or tables that can be queried using SQL and programmatically using APIs. In some cases, the same information is accessible using multiple approaches. Up-to-date information about the data hosted by ISB-CGC can be found in our online documentation.

Google Cloud Storage is a cost-effective large-scale object-store, and is used primarily to store the low-level sequence and image data. A typical workflow involves copying objects to a persistent disk attached to a VM, where they can be accessed as files using traditional methods. In addition, tools such as Cloud Storage Connector for Hadoop enable Spark and Hadoop jobs to access data directly in Cloud Storage, and Cloud Storage FUSE makes it possible to mount a Cloud Storage bucket as a file system (albeit with higher latency) on Linux and OS X systems.

BigQuery is a scalable analytic engine combining multilevel execution trees and a columnar datastore based on the open-access Dremel (12). It is well suited to storing, sharing, and analyzing the large-scale, heterogeneous data associated with programs such as TCGA. Information ranging from clinical and biospecimen information to molecular data such as gene-expression, copy-number, and mutation calls can be efficiently stored in "tidy data" format (13). BigQuery allows users to perform complex queries and joins using SQL supplemented with user-defined JavaScript functions and accessible from Python and R.

Discussion

Given the existence of other large-scale efforts, such as National Heart Lung and Blood Institute's "TOPMed" and NIH's Precision Medicine Initiative "All of Us," developing functional and interoperable cloud-based methods for large-scale data access and analysis is of critical importance. ISB-CGC has demonstrated the types of existing tools and approaches that can enable researchers with different backgrounds to access large-scale data and compute. By emphasizing Data as a Service, ISB-CGC enables a variety of approaches to application development. Pathology and radiology image data is a recent addition to the ISB-CGC resources, enabling cloud-scale deployment of image segmentation and feature extraction algorithms, combined with the ability to jointly analyze image-derived features with existing clinical and molecular data. Looking forward, interoperability between cloud systems will become an important factor. Currently, TCGA data are hosted on both AWS and Google, but paying to store the same data in multiple locations may not be the best long-term solution. As additional projects are moved to the cloud it is likely that other commercial providers will be included in the mix, whereas some projects may exist solely on private clouds or in private data centers. The ability to work seamlessly in this space will require standardized interfaces and methods for finding available data and tools. The ISB-CGC team is actively working with and supporting community-led efforts to define workflow and task-execution standards that will support reproducible analyses in this new era.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: S.M. Reynolds, M. Miller, P. Lee, S.M. Paquette, J. Bingham, D. Pot, I. Shmulevich
Development of methodology: S.M. Reynolds, M. Miller, P. Lee, K. Leinonen, S.M. Paquette, A. Hahn, D.L. Gibbs, J. Slagel, W.J. Longabaugh, M. Bookman, J. Bingham, D. Pot
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.M. Reynolds, D.L. Gibbs, V. Dhankani, T. Pihl, M. Bookman, D. Pot
Writing, review, and/or revision of the manuscript: S.M. Reynolds, T. Pihl, J. Bingham, D. Pot, I. Shmulevich
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Z. Rodebaugh, M. Reyes, M. Backus, N. Deflaux
Study supervision: S.M. Reynolds, D. Pot, I. Shmulevich
Other (writing of user facing documents and tutorials): D.L. Gibbs

Grant Support

All authors were funded by the National Cancer Institute, NIH, Department of Health and Human Services, under contract no. HHSN261201400007C.

Received February 28, 2017; revised June 23, 2017; accepted August 8, 2017; published online November 1, 2017.

References

1. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015;19:A68–77.
2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25: 2078–9.
3. Cock P, Fields C, Goto N, Heuer M, Rice P. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illlumina FASTQ variants. Nucleic Acids Res 2010;38:1767–71.
4. Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods. Ann N Y Acad Sci 2007;1115: 1–22.
5. Tatlow PJ, Piccolo SR. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci Rep 2016;6: 39259.
6. Thorvaldsdottir H, Robinson J, Mesirov J. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
7. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012;22;1760–74.
8. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, et al. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 2016;44:D574–80.
9. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014;42:D68–73.
10. Glusman G, Caballero J, Mauldin DE, Hood L, Roach J. KAVIAR: an accessible system for testing SNV novelty. Bioinformatics 2011;27:3216–7.
11. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45:D777–83.
12. Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, et al. Dremel: interactive analysis of web-scale datasets. Proc. of the 36th Int'l Conf on Very Large Data Bases, Singapore, Grand Copthorne Waterfront Hotel, September 13-17, 2010;330–9. http://www.vldb2010.org/.
13. Wickham H. Tidy Data. J Stat Softw 2014;59:1–23.