Cancer Genomics Cloud Pilot Developers Provide Update, Open Up Platforms for Early Testing

Nov 20, 2015 | Uduak Grace Thomas

Premium

NEW YORK (GenomeWeb) – The developers of the three proposals selected by the National Cancer Institute for the Cancer Genomics Cloud pilots — an effort to build sustainable computing infrastructure for analyzing omics data from the Cancer Genome Atlas and other NCI-funded projects — have begun various stages of testing ahead of the nine-month evaluation phase of the initiative, which is slated to begin in January 2016 when the first versions of all three platforms will be made available.

The proposals selected for the NCI initiative include one from Seven Bridges Genomics; one from the Institute for System Biology working in collaboration with Google and SRA International; and one from the Broad Institute in partnership with University of California, Berkeley and UC Santa Cruz.

Seven Bridges Genomics opened up its platform — based on Amazon Web Services — for broad early-access use by the cancer research community this week with the intent of further testing and refining it prior to evaluation. "Our overarching goal is to accelerate cancer research and discovery and treating patients," Brandi Davis Dusenbery, a senior scientist at the company and one of the lead researchers on the pilot, said explaining the rationale for the early release. "We think the best way to do that is to actually get our system into the hands of researchers as quickly as possible, get their feedback, and continue making it even better for them."

Dusenbery told GenomeWeb that the release features all of the functionality that the company detailed in its initial proposal to the NCI and also incorporates additional tools developed based on suggestions from a website, which the company set up to gather input from the research community as part of its development efforts. Furthermore, the company also uploaded additional datasets beyond those listed in its initial plans, she said.

Specifically, the new additions to the initial system, Dusenbery said, include a case explorer tool which lets users identify interesting research cases based on genomic, expression, and copy number variation data. The company also built a new triplestore-based browser which lets users

query more than 100 different clinical and biospecimen metadata properties. In addition, the company has built a number of functionalities to allow researchers who have been approved to use controlled-access data contained in dbGAP to collaborate with each other in a compliant way, she said.

Furthermore, Seven Bridges has uploaded both open and closed TCGA datasets to its platform, she said. That includes the required core datasets and bisulphite sequencing datasets which Seven Bridges selected as its orthogonal dataset but also mRNA sequencing data and clinical metadata. According to a blog post from the company announcing the opening of the EAP, the company worked with AWS to make the TCGA data available through Amazon's Public Data Set program, which covers the cost of storing TCGA data. That left more project funding available for use as computation and storage credits by the community, the post said. The CGC also offers standard bioinformatics pipelines for analysis tasks such as variant calling and researchers can also implement their own pipelines on the platform, she said. The system supports the Common Workflow Language so pipelines described using the language can be implemented easily on the CGC and they can also be transferred to other platforms that support language.

Seven Bridges hopes to sign on about 200 researchers for the EAP. As part of the EAP, the company will host weekly office hours to answer users' questions and gather feedback. The company currently has over $1 million in computation and storage credits to offer and it plans to distribute these to platform users under a fair-share model where the more a researcher uses the system, the more credits they receive, Dusenbery said.

Everyone who signs up for the CGC will receive at least $100 worth of computation and storage credits to use in the system, she said, which should be enough to analyze data from 40 to 100 RNA sequencing experiments. Moreover, researchers who enroll in the CGC Early Adoption program will receive $500 worth of AWS credits to use in the system. Research groups who bring their own tools or private data to the system will be able to apply for extra credits up to $7,500. Beyond that, the company will distribute credits on an incremental basis.

Meanwhile, the Google-based platform being developed by the Institute for System Biology in partnership with Google and SRA International has been opened up for alpha testing and the partners are on track to have the platform ready for evaluation in January, Ilya Shmulevich, an ISB professor and principal investigator on the NCI contract, told GenomeWeb in an email.

Specifically, Sheila Reynolds, ISB's project lead on its cancer genome pilot project, told GenomeWeb that the researchers are currently testing the platform with a select group of collaborators including the University of Texas MD Anderson Cancer Center, Oregon State University, and the British Columbia Cancer Agency's Genome Sciences Centre, she said.

In its current iteration, the ISB cloud contains all the level-three TCGA datasets and orthogonal datasets organized in Google BigQuery tables. The list of uploaded TCGA datasets includes clinical and biospecimen data; gene expression, mRNA, protein, DNA methylation, and copy number data; and there are also some cancer cell line datasets from the Cancer Genomics Hub in the platform, Reynolds said. The researchers have also uploaded controlled access TCGA data to Google storage but they are not making that available as part of the alpha run. In addition, the developers have released a series of iPython notebook tutorials that explain how to work with the

data, she said.

The data is currently accessible via an early version of an interactive web-based application and the developers have also implemented some basic functionality for accessing the data programmatically, Reynolds said. They'll continue to develop and improve these interfaces to the data ahead of the evaluation phase in January.

"From the beginning ... we wanted to host not just the very big files, [that is] the low-level BAM and FastQ files, and allow researchers to to run pipelines on that kind of data in the cloud but to also enable less computationally-savvy users to explore the open access data and drill down ... if they have access, to controlled data. That's still very much our vision," she said.

In terms of running computations on the cloud, the developers plan to absorb the costs of exploring the data using the web-based application, Reynolds said. However, users that want to run computationally heavy analyses will have to submit requests to set up so-called Google cloud projects within the platform. Once projects been approved, the developers will fund the analysis up until some "reasonable" amount of dollars have been used up. "We will be actively monitoring usage and if people have very large projects that they want to do, it'll be a back and forth between us and them about what they want to do and how much it would cost in terms of compute and storage," she said. "It's very much going to be an experiment to see what it is people want to do and how much it will cost so we will see how it goes."

GenomeWeb reached out to the Broad Institute to get an update on the third platform — dubbed FireCloud — which is being developed for the NCI-funded initiative, but was unable to get a response as of press time. Researchers at the Broad are collaborating with investigators at the UC Berkeley and UC Santa Cruz on the platform, which will provide a version of the Broad's Firehose analysis infrastructure on the Google cloud.

Filed Under Informatics Cancer Broad Institute ISB NCI UC Berkeley UCSC cloud computing software Google

Related Articles

Jul 27, 2015

NIH to Fund Research Testing Cancer Genomics Cloud Pilots

Jun 24, 2015

Broad Institute, Google Genomics to Offer Cloud-based GATK, Other Genomic Data Analysis Services

Dec 04, 2014

With Genomics Platform, Google Seeks to Bring Infrastructure, Other Capabilities to Bear on Biomedicine

Mar 30, 2015

ACMG: Google Views MSSNG Autism Database as Test Case for Enabling Large-Scale Genomic Analyses

Aug 26, 2015

FNLCR, FedCentric Partner to Explore Graph-Based Analytics for Cancer Research

Jul 31, 2015

NCI Names MD Anderson Genome Characterization Center Focused on Proteomic Analysis