Annai Preps NGS Data Management Platform for Formal Launch; Unveils Early Customer

Jun 30, 2016 | Uduak Grace Thomas

NEW YORK (GenomeWeb) – Annai Systems is preparing for the commercial release of a new product for the market called the Secondary and Tertiary Analysis Reporting Platform (STARPlatform), a data management solution for processing, analyzing, and managing raw next-generation sequence data as well as for calling, annotating, and reporting on genomic variants.

Annai CEO Michael Penley told GenomeWeb that the company has done a soft launch of the platform and signed agreements with two customers including the Glioma Longitudinal Analysis (GLASS) consortium, an international tumor research initiative that is led by researchers from the University of Texas MD Anderson Cancer Center, the University of California, San Francisco, and the Samsung Medical Center. It plans to do a full commercial launch of the platform in the fall of this year.

The STARPlatform is comprised of four solutions. The STARBox appliance offers tools for aligning raw sequence reads and for calling variants. It also features tools for lossless compression of raw sequence files to make them easier to store. It can, for example, reduce roughly a terabyte of data to about 60 to 70 percent of its initial size, making it easier to transfer and helping researchers save on storage costs, Penley said.

The platform also includes STARInsight, which lets researchers combine their variant call files with curated public reference data and annotations. Researchers can set up private workspaces within the system where they can upload and store private datasets as well as upload and run whatever algorithms they choose to on their data. They can also share their data with approved collaborators through the platform. STARInsight also includes tools for generating research and clinical reports.

In addition, the platform includes STARVault for storing compressed data and for capturing metadata around these files to make them easier to locate in storage resources. Also available is the STARClient, which is used for downloading and decompressing files from the STARPlatform as well as for accessing the platform through application programming interfaces.

According to Penley, the STARPlatform includes elements of two of Annai's previous products, namely its ShareSeq platform and the Annai GNOS system, but is a completely different product. For example, STARBox uses the same file transfer technology, which was also in ShareSeq. The so-called GeneTorrent technology makes it possible for researchers to reliably and securely transfer data. Also, the newly renamed STARVault solution was previously Annai's Genomic Network Operating System (GNOS), a tool designed for storing data files and capturing metadata associated with those files.

Developed in partnership with Hitachi Data Systems, Annai launched ShareSeq in 2014 to provide academic and commercial researchers with cloud-based access to genomic data, bioinformatics pipelines and workflows, and compute power and storage. Both ShareSeq and Annai's GNOS have supported projects such as the Cancer Genomics hub, a National Cancer Institute-funded petabyte-scale data repository developed by researchers at the University of California, Santa Cruz to provide access to genomic and clinical data from NCI-funded cancer genome research programs. In total, CGHub holds more than a petabyte of data from both the Cancer Genome Atlas and the Therapeutically Applicable Research to Generate Effective Treatments projects.

Annai was also named as one of the technology partners involved in the International Cancer Genome Consortium's Pan Cancer Project. Its GNOS software was used by the six centers selected to house the data collected for the effort. The ICGC also tapped ShareSeq to host data from more than 10,000 cancer genomes generated by its projects, but it now uses compute resources provided by Amazon for the Pan Cancer project and is housing its datasets in repositories such as the NCI's Cancer Genomics Cloud and the Cancer Genome Atlas. With these projects migrating to other resources, "there wasn't really a big need to continue to offer ShareSeq as a platform," Penley said, and Annai officially retired the product in February of this year.

The company is now focusing its attention on bringing the STARPlatform to market. Compared to existing solutions from companies such as DNAnexus, WuXi NextCode, and Seven Bridges Genomics, "I think our differentiating feature would be the flexibility of the analytics," Penley said. "We have a solution where researchers [and] clinicians can really generate any type of report they are looking to generate and do any type of downstream analytics quickly without having to weed through a massive set of pre-processed filters on data." The company is not disclosing details about its pricing structure at this time but Penley said that the exact costs will vary depending on how the customer uses the software. He also said that the company hopes to secure partnerships with large research organizations who are willing to use platform to provide genomic analysis services to their clients, he said.

So far, Annai has signed agreements with one unnamed customer as well as with the GLASS consortium, which is using the STARPlatform as part of efforts to understand resistance mechanisms in three glioma subtypes with an eye towards developing more effective therapies for the tumors. They are currently gathering a longitudinal genomics dataset that represents patients across three specified diffuse glioma genomic subtypes: IDH-wild-type, IDH-mutant, and IDH-mutant 1p/19q co-deletion.

With the Cancer Genome Atlas project now completed, "we now know what the molecular basis of cancer is, and, I think, have a much better understanding of what cancer looks like," Roel Verhaak, GLASS lead investigator and assistant professor of bioinformatics and computational biology at the MD Anderson Cancer Center, said during an interview with GenomeWeb. The next set of questions for researchers to explore are the mechanisms by which these tumors change over time, he said. This is crucial, particularly for high-grade tumors, which are "notoriously" resilient and able to resist various therapies, he explained.

Initially, Verhaak and colleagues focused on analyzing tissue samples that they had stored in the MD Anderson tissue bank. However, they soon realized that they did not have enough samples for the sort of longitudinal analysis that they wanted to perform. "We

always intended to do this on samples from hundreds of patients, [but] as we went about in this project, we quickly realized that we were not going to be able to identify more than a few dozen tissue samples in our MD Anderson tissue bank," he said.

The MD Anderson researchers then reached out to colleagues at other institutions who were trying to perform similar projects and found that they, too, did not have access to sufficient numbers of samples for their projects. "It was a logical next step to talk about a widespread collaboration in order to pursue this goal of longitudinal molecular characterization … [and] to make this an international effort," Verhaak said.

That led to the formation of the GLASS consortium about a year and a half ago. To date, some 20 institutions from 10 countries have signed on to contribute glioma samples and data. "We are very much in the early phases of this consortium, which means that we are trying to aggregate existing datasets [and] identifying tissue samples in all tissue banks of those institutions," Verhaak said. In total, the researchers expect to analyze data from about 1,500 diffuse glioma patients including existing exome sequence from about 250 glioma patients and new data from at least 450 tissue samples housed in participating institutions' biobanks.

The consortium's data infrastructure and processing working group has also developed an early iteration of a computational pipeline, that includes tools such as the Broad Institute's MuTecT and VarScan, which participating researchers will use to process raw sequence and call variants. "The partnership with Annai came about because we were looking for a [platform] where we could do computing of the datasets that we have and the datasets we would generate moving forward in a way that each institution could process their own raw data using [the] computational pipeline developed by the GLASS consortium without the need to exchange our datasets," Verhaak said.

For example, a collaborating institution in Germany could load their raw sequences to the selected platform, process the data and call variants, take down the raw sequence files, and then share the variant calls with the rest of the consortium. This way, the consortium does not run afoul of the regulatory and ethical frameworks that different countries have put in place to govern the way patient data is shared and used. "This is why Annai [is] a good solution for us," he said. "It allows us to process data using standardized pipelines without exchanging data."

In addition to gathering data and patient samples, the consortium is also seeking more funding for its efforts. Earlier this year, the consortium received a $250,000 grant from the National Brain Tumor Society to cover administration costs as well as to support regulatory-related activities such as setting up international data exchange agreements, Verhaak said. The consortium also benefits from existing research grants awarded to individual investigators for internal projects. Currently, the consortium has two pending grant applications in different countries including one submitted by a collaborating institution based in the Netherlands