PHYS.ORG

Genome Analysis Toolkit 4 (GATK4) released as open source resource to accelerate research

May 25 2017

Credit: Susanna M. Hamilton, Broad Communications

The Broad Institute of MIT and Harvard will release version 4 of the industry-leading Genome Analysis Toolkit under an open source software license. The software package, designated GATK4, contains new tools and rebuilt architecture. It is available currently as an alpha preview on the Broad Institute's GATK website, with a beta release expected in mid-June. Broad engineers announced the upgrade, as well as the decision to release the tool as an open source product, at Bio-IT World today.

The new version is built on a new architecture, allowing significant streamlining of individual tools and support for performance-enhancing technologies such as Apache SparkTM. This new framework brings improvements to parallelization, capitalizing on cloud deployment and making the process of analyzing vast amounts of genomic data easier, faster, and more efficient.

"We wanted to remove traditional barriers of scale while offering the same high level of data quality our users expect," said Eric Banks, Senior Director of Data Sciences and Data Engineering at Broad and a creator of the original GATK software package. "Thanks to the rapid adoption of cloud computing, researchers can finally do away with many of the infrastructure-related complications that have hampered progress, especially at smaller institutions and startups."

Today, more than 45,000 academic and commercial users worldwide rely on the GATK, running millions of analyses. The GATK is the industry standard for identifying SNPs and indels in germline DNA and RNAseq data. In addition to improving the performance of these established tools, GATK4 extends this scope of analysis to include copy number and structural variation, for both germline and somatic research applications.

Fully open source software

GATK4 will be released as a fully open source product, thanks in part to a collaboration between Broad Institute and Intel Corporation to advance high-performance analytics so researchers can study massive amounts of genomic data from diverse sources worldwide.

At the Intel-Broad Center for Genomic Data Engineering, software engineers and researchers have spent the last several months building, optimizing, and widely sharing new tools and infrastructure to help scientists integrate and process genomic data. GATK4 has benefited from this collaboration, which has helped engineers optimize best practices in hardware and software for genome analytics to make it possible to combine and use research data sets that reside on private, public, and hybrid clouds.

"Releasing GATK4 as open source was the obvious next step for our team," said Geraldine Van der Auwera, Associate Director of Outreach and Communications within the Data Science and Data Engineering group at the Broad Institute. "We believe it's the most effective way to support the community, and we hope it continues to grow, innovate, and help researchers make insights that are essential for future human health breakthroughs." "It is critical for progress in biomedicine that the software we use for analysing the genomes of millions of people is robust and well understood," said Ewan Birney, Director of EMBL-EBI and Chair of the Global Alliance for Genomics and Health (GA4GH). "Releasing GATK software with an open source license directly supports open innovation, data re-use and data re-analysis in the global biomedical community."

"The GATK tools are crucial for both germline and cancer analyses," said Robert L. Grossman of the University of Chicago Department of Medicine and an expert in biomedical informatics. "Releasing GATK4 as an open source software package will increase adoption, and benefit the community."

"Open sourcing the GATK is a big deal for open genomics, and for open science in general," said Jeremy Freeman, manager of computational biology at the Chan Zuckerberg Initiative (CZI). "Not only does it make this critical tool available to as broad as possible an audience for use, reuse, inspection, and contribution—it provides a powerful example to the community for how an existing project can embrace open source."

"Open source code is a foundation of efficient biomedical research," said Brad Chapman, a research scientist at the Harvard T.H. Chan School of Public Health. "It enables reproducibility, reuse and remixing by removing barriers for sharing and distributing analyses. The Broad Institute's GATK team leads in the development of scalable, sensitive and specific variant calling algorithms, and open sourcing GATK4 will allow frameworks like Blue Collar Bioinformatics to make these methods broadly available to the scientific research community."

"Cloudera has always been a supporter and believer in the power of open source code," said Tom White, data scientist at Cloudera and a member of the Apache Hadoop PMC. "We've been excited to contribute to the GATK codebase, to make it run smoothly on Apache Spark and Cloudera. This next phase of the GATK, powered by Spark and open source software, will expand access and improve collaboration among genomic data scientists."

"The open sourcing of GATK4 is a great step for genomics, allowing for scalability and performance gains to be openly available to the research, biotech and pharmaceutical communities," said Jason Waxman, corporate vice president and general manager of Data Center Solutions at Intel. "GATK4, when run on Intel's new reference architecture, can achieve a 5X speed-up compared to earlier versions of the software."

"We at Google are excited to see this new release," said Ilia Tulchinsky, Google Cloud Healthcare Engineering Lead. "We've been collaborating with the Broad Institute for the past three years to enhance genomic processing on Google Cloud Platform. As a strong supporter for open source technology, we believe that making GATK available this way will facilitate its use by genomic scientists everywhere. As fellow collaborators with Intel, we particularly look forward to enabling researchers to run GATK4 on Google Cloud using the upcoming Intel Xeon processor Scalable family."

"The GATK is one of the most widely utilized software packages in the life sciences, and our team has worked very productively with Broad to accelerate it for use on Azure," said Geralyn Miller, Director, AI & Research, Microsoft. "This new model will greatly facilitate this effort going forward, and we are excited to continue and expand our efforts around GATK on Azure."

"With the open source launch of GATK4, there is an opportunity to create a global community that can collaborate together and advance the state of art in bioinformatics," said Hong Tang, chief architect at Alibaba Cloud, the cloud computing arm of Alibaba Group. "We look forward to closely working with Broad Institute in bringing the cloud-based GATK service to genomics customers in China, as well as in ongoing GATK research and development."

In addition to offering GATK4 as an open source toolkit, Broad Institute will continue to offer user support, training, and outreach on its popular user support forum. GATK4, like many of the Broad Institute's genome analysis tools, will be available through the Broad Institute's cloud based analysis platform, FireCloud.

Provided by Broad Institute of MIT and Harvard research (2017, May 25) retrieved 8 April 2024 from https://phys.org/news/2017-05-genome-analysis-toolkit-gatk4-source.html