**The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**

McKenna, A.[1], Hanna, M.[1], Banks, E.[1], Sivachenko, A[1], Cibulskis, K.[1], Kernytsky, A.[1], Garimella, K.[1], Altshuler, D.[1,2], Gabriel, S. [1], Daly, M.[1,2], and DePristo, M.A.[1]
[1]Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Five Cambridge Center, Cambridge, Massachusetts 02142.
[2]Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts 02114, USA

Contact: Matt Hanna hanna@broadinstitute.org
Overview: http://www.broadinstitute.org/gsa/wiki/index.php/Main_Page
Source: https://svn.broadinstitute.org/Sting


Next-generation DNA sequencing (NGS) projects, such as the 1000 Genomes Project, are already revolutionizing our understanding of genetic variation among individuals. However, the massive data sets generated by NGS—the 1000 Genomes pilot alone includes nearly five terabases—make writing feature-rich, efficient and robust analysis tools difficult for even computationally sophisticated individuals. Indeed, many researchers are limited in the scope and the ease with which they can answer scientific questions by the complexity of accessing and manipulating the data produced by these machines.

Our solution to this issue is the Genome Analysis Toolkit (GATK), a structured programming framework designed to ease the development of analysis tools for next-generation DNA sequencers using the functional programming philosophy of MapReduce. The GATK provides a small but rich set of data access patterns that encompass the majority of analysis tool needs. Separating specific analysis calculations from common data management infrastructure enables us to optimize the GATK framework for correctness, stability, CPU and memory efficiency, and to enable distributed and shared memory parallelization. We highlight the capabilities of the GATK by describing the implementation and application of robust, scale-tolerant tools like coverage calculators and SNP calling. We present the techniques used to partition datasets, allowing the GATK to run effectively on both large-memory multiprocessor machines as well as the more restrictive but abundant nodes in a server farm. Finally, we outline the techniques we use to efficiently access and present genomic structure and variation data stored in a wide array of highly flexible file formats.

The GATK programming framework enables developers and analysts to quickly and easily write efficient, robust, and easy-to-use NGS tools. The GATK already underlies several critical tools in both the 1000 Genomes Project and The Cancer Genome Atlas, including: quality-score recalibration, multiple-sequence realignment, HLA typing, multiple sample SNP genotyping, and indel discovery. The GATK's robustness and efficiency has enabled these tools to be easily and rapidly deployed in recent projects to routinely process terabases of Solexa, SOLiD, and 454 sequencer data, as well as the hundreds of lanes processed each week in the production resequencing facilities at the Broad Institute.