ADAM: Fast, Scalable Genomic Analysis

Frank Austin Nothaft[1], *, Matt Massie[1], *, Timothy Danford[4], Carl Yeksigian[4], Arun Ahuja[5], Neal Sidhwaney[5], Jey Kottalam[1], Christopher Hartl[2], Christos Kozanitis[1], André Schumacher[3], Jeff Hammerbacher[5], Michael D. Linderman[5], Anthony D. Joseph[1], and David Patterson[1]

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA
[2]The Broad Institute of MIT and Harvard, Cambridge, MA
[3]International Computer Science Institute (ICSI), University of California, Berkeley, CA
[4]GenomeBridge, Cambridge, MA
[5]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY
*These authors contributed equally.

http://www.bdgenomics.org
http://www.github.com/bigdatagenomics/adam

Abstract

ADAM is a high-performance distributed processing pipeline and API for DNA sequencing data. To allow computation to scale on clusters with more than a hundred nodes, ADAM uses Apache Spark as a computational engine and stores data using Apache Avro and the open-source Parquet columnar store. This scalability allows us to perform complex, computationally heavy tasks such as base quality score recalibration (BQSR), or duplicate marking on high coverage human genomes ($> 60\times$, 236GB) in under a half hour. In tests on the Amazon Elastic Compute platform, we achieve a $50\times$ speedup over current processing pipelines, and a lower processing cost.

To achieve scalability in a distributed setting, we rephrased conventional sequential DNA processing algorithms as data-parallel algorithms. In this talk, we'll discuss the general principles we used for making these algorithms scalable while achieving full concordance with the equivalent serial algorithms. Additionally, by adapting genomic analysis to a commodity distributed analytics platform like Apache Spark, it is easier to perform ad hoc analysis and machine learning on genomic data. We will discuss how this impacts the clinical use of DNA analysis pipelines, as well as population genomics.