Developing an Arvados BWA-GATK pipeline

Pjotr Prins[1,2], Joep de Ligt[3], Isaac Nijman[1], Ward Vandewege[4], Bryan Cosca[4] and Peter Amstutz[4]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] University of Tennessee Health Science Center, Memphis, USA
[3] Hubrecht Institute, Utrecht, The Netherlands
[4] Curoverse, Boston, USA[†]

Contact: pjotr.public05@thebird.nl

Project website: `http://arvados.org/`
Source code: `https://arvados.org/projects/arvados/repository`
License: AGPLv3 for core, Apache 2.0 for the SDK

Arvados is a free and open source platform for bioinformatics and big data science. In this (lightning) talk, we present the porting of an existing HiSeq BWA-GATK based variant calling pipeline from Sun Grid Engine (SGE) to Arvados, resulting in a pipeline that is faster, scalable, simpler and more robust.

The main reason the pipeline runs faster is that the Arvados file system (named Keep) is decentralized. In most bioinformatics cluster environments, the central file storage is the bottleneck because of resource contention, i.e. many cluster nodes are hitting the storage for data requests. When uploading a file into Keep the file is chunked and distributed across nodes. When a node requires data, it can fetch it from different locations in the network.

Because of Keep, the new pipeline is scalable and runs in flat time, i.e. processing one genome takes the same amount of time as running ten or a hundred. Running pipelines in parallel is now only a function of adding new nodes. Keep is distributed and scales accordingly.

The original SGE pipeline consisted of hundreds of lines of Perl and bash code, which mostly dealt with the plumbing of submitting jobs, checking conditions and job completion. In contrast, the Arvados version is mostly single lines for command line invocation as Arvados takes care of the plumbing.

We also needed to speed up GATK using GATK/Queue. Queue chunks BAM files and fans work out to multiple nodes. For this, a special Queue adapter had to be written for Arvados similar to the one that exists for SGE. The Arvadas implementation is an improvement over the SGE implementation because it avoids using mounted NFS for sharing and collating Queue results, thereby making GATK/Queue more robust.

[†]Disclosure: Curoverse is a major contributor to the Arvados open source project and a sponsor of BOSC 2015.