Hututa Technologies Limited

Super-Fast Genome BWA-Bam-Sort on GLAD

Zhiqiang Ma, Wangjun Lv and Lin Gu

May 2016

Executive Summary

Aligning the sequenced reads in FASTQ files and converting the resulted SAM file into a sorted BAM file are a common practice and often the first stage of many genome analysis pipelines. Software such as BWA and Samtools has been developed by the bioinformatics community to achieve this on single computers. Although those tools usually can make use of multi-threading technologies to use multiple CPU cores in modern computers, the time used is still so long as hours, taking a large part of the whole analysis pipeline and limiting the overall speed of pipeline executions. As the application area of genome analysis is larger and larger, the analysis should finish in minutes instead of hours.

GLAD is a genome analysis system that makes use of the resources of many compute nodes in a cluster from private or public clouds to significantly accelerate the gnome analysis such as FASTQ to sorted BAM generating which we call BWA-BAM-Sort or BBS in this white paper. With a flexible programming interface, GLAD does not require modifications to software such as BWA and Samtools to run on GLAD.

This white paper describes performance results of BBS computation on GLAD in clusters of compute nodes in private or public clouds. The BBS step's time can be shortened to minutes from hours even with a cluster of a moderate size.

Background

Our products for fast large-scale genome storage and analysis consist of the GLAD genome analysis software and the Genes' Mind ready-to-use genome storage and analysis appliance.

GLAD: Super-fast software for parallel genome data analysis

The development of cloud computing makes it easy to own or use a private cloud or resources consisting of many compute nodes in public clouds at affordable prices. However, the fact that the traditional software such as BWA and Samtools was designed for single computers limits the ability of using resources such as CPU cores from a cluster of nodes. On the other hand, big data technologies, such as Data Thinker, have been developed in recent years to integrate many resources from many nodes to a single system for fast processing of large data.

Powered by D-thinker parallelization, GLAD speeds up processing by 10-100 times on normal PCs. When running GATK pipelines, GLAD reduces data processing time from days to hours.

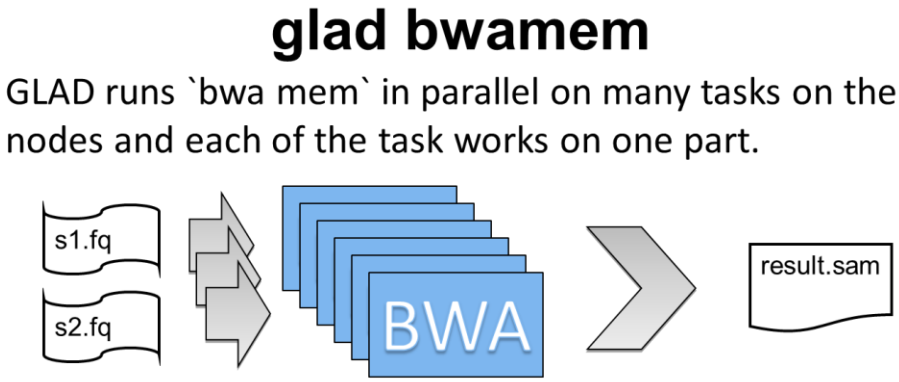

FIGURE 1: BWA MEM ON GLAD

GLAD helps biomedical scientists and practitioners store PBs of genome data safely and process them 100 times faster than before. Supporting standard software such as BWA and GATK, the GLAD technology is used in global projects including ICGC-PanCancer.

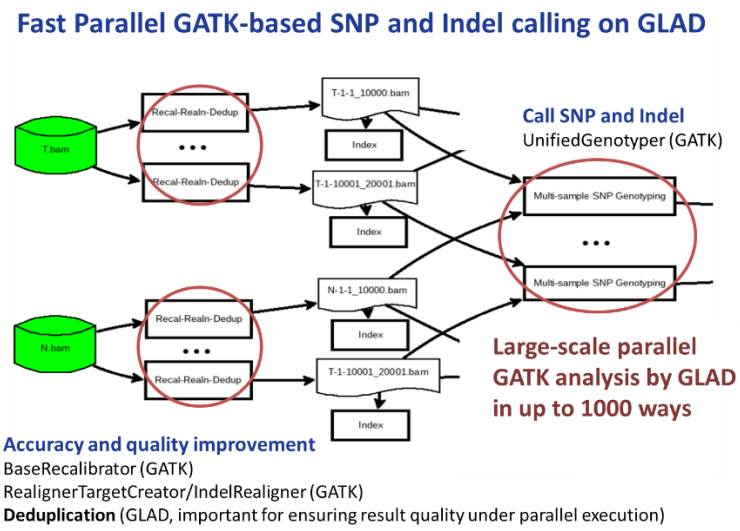

FIGURE 2: GATK-BASED VARIANTS CALLING ON GLAD

Genes' Mind: Inexpensive hardware for storing and processing genome data

Genes' Mind packages multiple compute servers, 10s-100s of processor cores, 100s of TBs of distributed storage and genome data processing software as a ready-to-use genome data storage and processing system, all at a very affordable price.

FIGURE 3: GENES' MIND GENOME STORAGE AND PROCESSING APPLIANCE

Genes' Mind has pre-installed big data and genome analysis systems such as Data Thinker and GLAD. It can also double-duty as an internal private cloud system for an organization, providing storage and virtual machines for R&D and office work.

BBS on GLAD in Private and Public Cloud

We use GLAD to speed up BBS computation on a dataset consisting of a pair of FASTQ files on 3 different clusters and compare the performance of BBS on GLAD and a single compute node.

Dataset

The input dataset, named S81, for the performance measurement is a human whole exome data.

The S81 dataset comprises two FAQSTQ files of 4.8GB each (9.6GB in total).

```
/thinker/dstore/world/gene/sapiens/study/exome/S81/S81_1.fastq
/thinker/dstore/world/gene/sapiens/study/exome/S81/S81_2.fastq
```

GLAD uses D-store, a distributed file system scaling to tens of petabytes, to store genome data and organize them as a file system hierarchy visible to authorized people around the world.

Clusters

We use 3 clusters for performance measurement of BBS on the S81 dataset. All these 3 clusters are installed with the GLAD/D-store/D-thinker systems.

tbg6: a small cluster with 8 inexpensive nodes inside of Hututa's private cloud. Each node has 4 Intel Core i3 CPU cores and 32GB memory.

gm15: a Genes' Mind appliance with 5 inexpensive nodes. Each compute node has 8 Intel Xeon E3 CPU cores and 32GB memory. It is a shared environment for development and testing from universities and cooperators.

ksc: a cluster in the Kingsoft Cloud, a public cloud for business and developers. The ksc cluster consists of 4 VMs allocated in the public cloud. Each VM has 8 Intel Xeon E5 CPU cores and 16GB memory.

Performance measurement

In each cluster, we run BBS on GLAD and on a single compute node and measure the time used to finish the analysis.

Multithreaded BBS computation: run native BWA and Samtools tools to do BBS computation on a single compute node. The computation uses multiple threads to utilize the multithreading capability of native BWA and Samtools tools to make full use of the CPU cores on a single node.

BBS computation on GLAD: run a BBS computation on GLAD. The computation is run in many containers on the compute nodes in the cluster used. GLAD runs unmodified BWA and Samtools in parallel on the compute nodes in the cluster.

Performance results

We measure and draw the time used for the BBS computation on the S81 datasets on these 3 clusters using GLAD and multi-threads on a compute node. Figure 4 shows the times used on all 3 clusters.

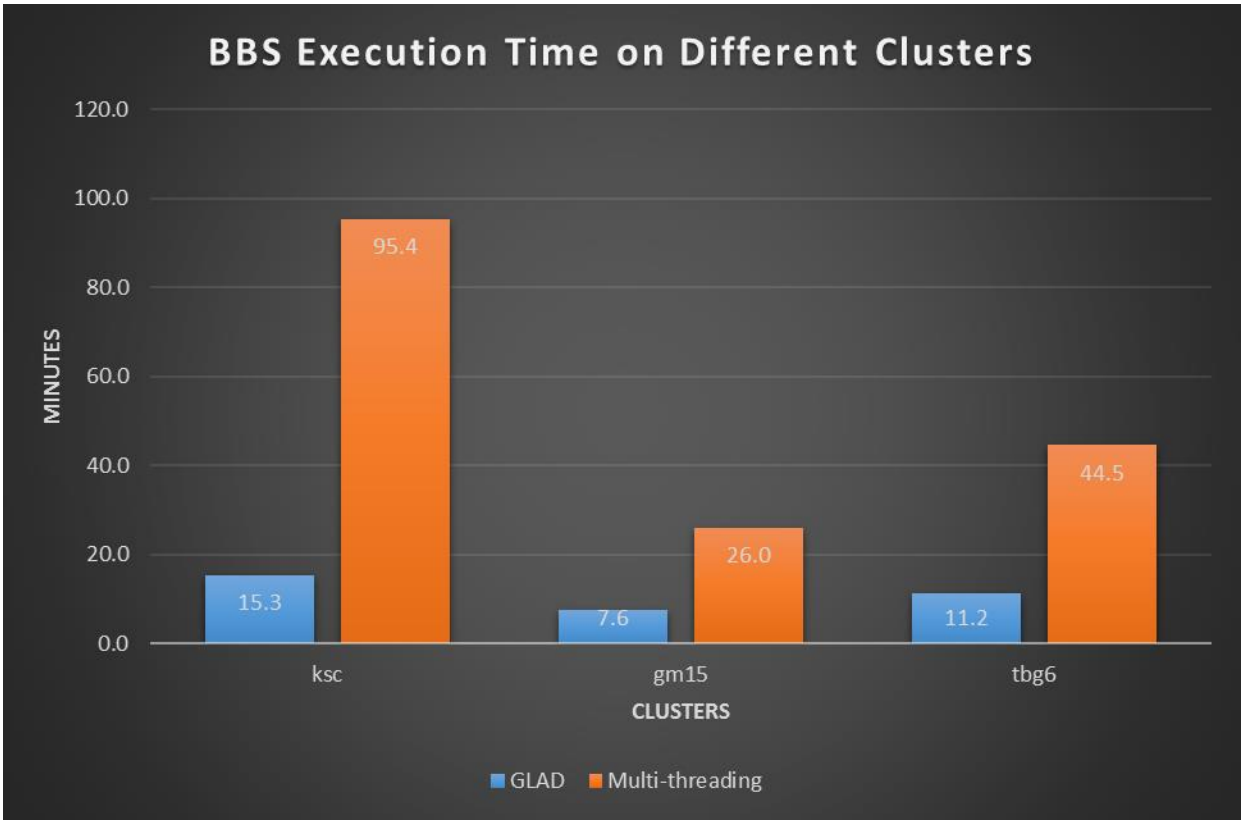

FIGURE 4: BBS EXECUTION TIME ON DIFFERENT CLUSTERS

On all 3 clusters, GLAD can finish the BBS computation of the S81 data in 8 to 16 minutes, while the multithreaded version takes 26 minutes to one and half hours. GLAD makes use of the CPU cores in the cluster to make BBS analysis super-fast, speeding up the BBS by 3 – 6 times in these clusters of 4-8 nodes. In the ksc cluster in the public cloud, GLAD even shows super-linear speedup.

Conclusions

GLAD, built on the super-fast Data Thinker technology, makes use of the compute power from many compute nodes in a cluster. In the performance comparison of multithreaded and GLAD-based BBS computation on all 3 different clusters with different configurations, GLAD delivers great performance and greatly accelerates the BBS computation. GLAD can also scale to larges clusters consisting of tens to hundreds of compute nodes for even higher performance.

Read more

Readers interested in more details can visit the following online info about GLAD and Data Thinker:

GLAD homepage: http://www.hututa.com/en/products/glad.html

Data Thinker website: http://d-thinker.org/

GLAD and Data Thinker forum: http://tab.d-thinker.org/forumdisplay.php?fid=515