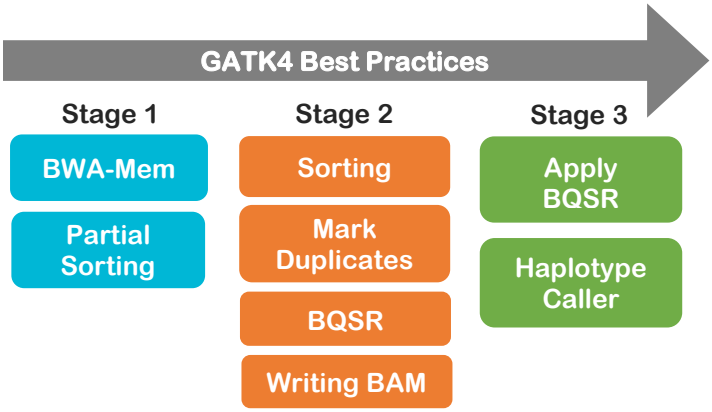
Parabricks

505 E. Liberty St.
Suite 400
Ann Arbor, MI - 48104

www.parabricks.com
info@parabricks.com

High Performance Secondary Analysis of Genomic Data

Parabricks provides 30-50 times faster secondary analysis of FASTQ files coming out of sequencer to VCF files that have the variants for tertiary analysis. The standard pipeline shown below consists of three steps and are defined as per the Genome Analysis Toolkit (GATK) best practices. Parabricks accelerates existing GATK best practices to generate equivalent results as the baseline CPU only solution.



Architecture of the Parabricks Pipeline

Parabricks has accelerated these software (BWA-Mem, GATK4) by running them on GPUs. Existing analysis which takes 30 hours of computation on a 32-vCPU machine can be completed in 45 mins, while implementing the exact same algorithm. The user of the Parabricks software can choose which steps of the pipeline to run and can configure her pipeline. By default, Parabricks runs the exact GATK best practices pipeline in the prescribed order.

Features

- **Turnkey Solution:** Parabricks software runs on standard CPU and GPU nodes available on the cloud and requires no additional setup steps by the user.
- **On-Premise and Cloud Agnostic:** The Parabricks software can run on local servers like DGX-1 servers or public cloud such as AWS, Google Cloud Services and Microsoft Azure.
- **100% Deterministic and Reproducible:** Any configuration of Parabricks software on any platform with any number of resources, generates the exact same results every execution.
- **Equivalent results:** Parabricks' pipeline generates equivalent results as baseline GATK 4 best practices pipeline as the same algorithm is used.
- **Support all tool versions:** Parabricks' accelerated software supports multiple versions of BWA-Mem, Picard and GATK and will support all future versions of these tools.
- **Visualization:** Parabricks generates several key visualizations real-time, while performing secondary analysis that can improve the user's understanding of the data.
- **Single Node Execution:** The entire pipeline is run using one computing node and does not incur any overhead of distributing data and work across multiple servers.

Computing Cost Reduction

The throughput achieved by one GPU Server (like DGX-1) with Parabricks software is nearly 12,000 whole genomes per year. In comparison, the baseline CPU only solution would require nearly 35 servers each with 32 vCPUs for similar throughput. Unlike a GPU Server, managing large number of CPU servers will require a dedicated IT infrastructure and significantly higher power and cooling efforts.

For users on the cloud, computing costs are proportional to execution time and, by reducing runtime by 30-50 times, Parabricks can reduce computing costs significantly compared to CPU-only solutions. Furthermore, cloud solutions provide significant discounts when their capacity is underutilized (like spot-bidding in AWS) with preemptive instances. By reducing the runtime to under one hour, the Parabricks solution reduces the probability of preemption and reduces the cost of computing significantly on the cloud.

Performance Comparison

Baseline Pipeline Setup:

Tools

BWA-Mem	0.7.12
GATK	4
CPU Threads	32
Node	c3.8xlarge

Data

Type	WGS
Sample	Human
Read Length	151
Coverage	25x -40x

Details

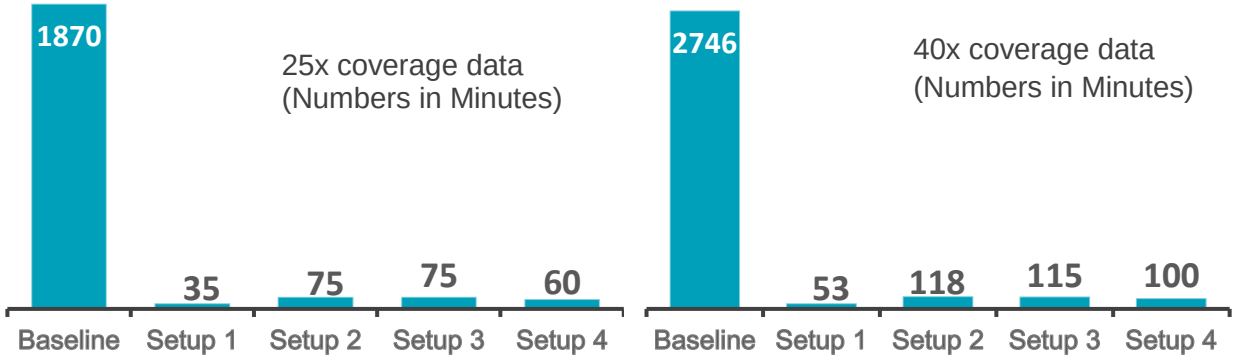
All steps shown above were run. BWA-Mem was run with 32 threads. GATK HaplotypeCaller with 16 threads

Parabricks Pipeline Setup:

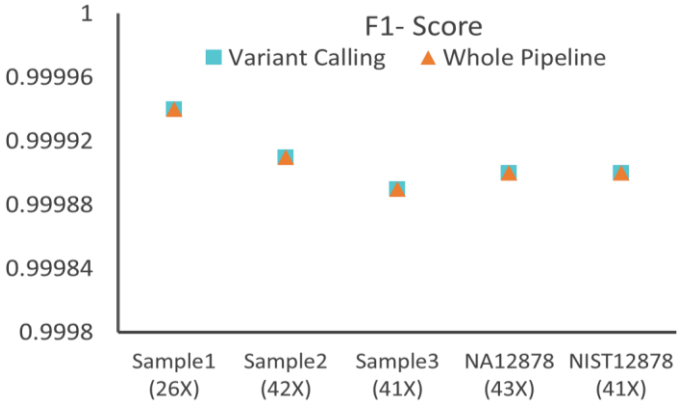
Setup	Platform	GPU
1	DGX-1	8 V100
2	Azure	4 P100
3	Google	4 P100
4	AWS	4 V100

**AWS has nodes with 8 V100 GPUs. The results are similar to Setup 1.*

The Parabricks solution can be deployed to private clouds or clusters as well. The only requirement for Parabricks solution to provide accelerated computing is the presence of GPUs. The platforms on which Parabricks has been tested are shown in the table on the left. The results below are for the GATK 4 best practices pipeline. However, the Parabricks pipeline is configurable and can run any stages. Small changes to the pipeline maybe incorporated on a need basis.



Accuracy



The results of all the runs for each sample were found to be identical for BWA-Mem, Sorting, Marking Duplicates and Apply BQSR Stage. The baseline variant caller in GATK 4 is non-deterministic and can generate slightly different results based on certain parameters. For this step, all four Parabricks setups generate the same outptut which is within 0.9999 of the GATK 4 execution. The differences are comparable to the variation in GATK execution due to multiple threads, nodes.

Parabricks software provides 30-50x times faster analysis at reduced computing costs while generating equivalent results. Limited free trials are available on AWS, Google, and Azure. Contact info@parabricks.com for a demo or more details.