Accelerating Genomic Discoveries for Precision Medicine

Sissades Tongsima[1], Chumpol Ngamphiw[1], and Ankit Sethia[2]

[1]National Center for Genetic Engineering and Biotechnology (BIOTEC), [2]Parabricks, LLC

Abstract

DNA analysis is transforming the treatment of many diseases such as cancer, Alzheimer's, and monogenic diseases in infants by customizing patients' treatment based on their genetic characteristics. Due to the promise of precision medicine which is based on such analysis, the number of genomes to be sequenced is expected to double every year and newer technologies are needed to accelerate the analysis process to make these analyses the standard of care. In this work, we show how Parabricks' high performance Next Generation Sequencing (NGS) software, based on NVIDIA Graphics Processing Unit (GPU) accelerated computing, accelerates the process from two days to one hour at one-fourth the cost. Furthermore, we show the impact that this has on leading genomic research centers such as BIOTEC, which plans to analyze thousands of samples every year. By reducing the turnaround time from days to an hour, BIOTEC plans to gain deeper insights into data, which was not previously possible.

Introduction

Genomic analysis is transforming the understanding of disease mechanisms and the process of curing and preventing disease. It is driven by DNA sequencing whose cost and speed are improving faster than Moore's Law. With the advancements in NGS technologies, the number of human genomes sequenced has doubled every seven months and Illumina (the largest sequencer manufacturer) predicts it will continue to double every 12 months. This growth is further fueled by the ongoing transition of NGS into clinical applications where it is enabling genomic medicine and transforming the diagnosis and treatment of diseases, such as cancer, Alzheimer's, monogenic diseases in infants, HIV, diabetes, ADHD, and numerous other diseases and conditions. The use of DNA sequencing data for diagnosis and drug discovery promises to be a revolutionary change in modern medicine.

However, current state-of-the-art DNA analysis, which is driven by NGS, is often restricted to processing smaller panels of sequencing data which is derived from a targeted sequencing approach, such as Whole Exome Sequencing (WES). As NGS technologies continue to advance and processing time and cost are reduced, there is a major drive to move from targeted sequencing to Whole Genome Sequencing (WGS), which provides far greater information leading to better discovery and therapeutic actions.

On average, a whole human genome generates up to 1 terabyte of sequencing data which must be processed using nearly one thousand CPU-hours to align the genome to the reference sequence, identify the variants within the genome, and generate conclusions of biological significance for geneticists, bioinformaticians and physicians. This computationally-intensive task can become significant when processing large numbers of samples. Thus, a traditional approach of using a CPU-only compute solution requires scaling up or scaling out of an existing system, which may be impractical for many data centers. On the other hand, NVIDIA GPUs, with increasing numbers of CUDA cores (an order of 1000x), accelerate parallelizable sequencing

tasks in genome processing at different stages to greatly minimize total analytics time, meeting the genome sequencing processing time challenge.

Based on current trends of improvements in DNA sequencing, the computational analysis of the large volume of the digital DNA data is becoming the major bottleneck for using key WGS data to effectively offer treatment for affected patients. The number of patients that can benefit from analysis of their WGS data is expected to be one billion by 2026, demanding large computational costs and time. These analyses are the preliminary steps for precision medicine, and innovative technological disruption is required to bring the promise of precision medicine to fruition for the general public by accelerating the analysis of genomic data at reduced cost.

Common WGS for humans consists of three stages:

- Primary analysis converts biological DNA to raw digital DNA data using sequencing machines.
- Secondary analysis processes and scans the raw digital DNA data to identify potential biologically significant DNA variants which can be further analyzed by scientists and physicians.
- Tertiary analysis interprets these potential variants to confirm biological implications.

Secondary analysis is a required step to enable tertiary analysis which leads to biological insights. It processes large amounts of data and performs calculations that have been developed and refined over several decades. Highly accurate secondary analysis is a computationally expensive step that is commonly required; thus, it is imperative to improve processing times. Faster, more efficient secondary analysis processing results in earlier discoveries via higher analysis throughput and reduced processing costs and enables genomic analysis to become part of standard healthcare practices.

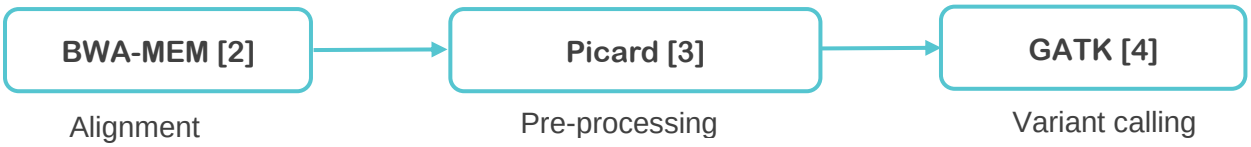
Nvidia GPUs and Accelerated Computing

GPU-accelerated computing is the use of a GPU together with a CPU to accelerate deep learning, genomic analytics, and engineering applications. GPUs use parallel processing techniques to break down complex computing problems into many smaller tasks that run simultaneously on CUDA cores. In genomics and related fields, where large-scale datasets are the norm, computing tasks can be handled in dramatically less time. In some cases, projects that once required supercomputers can now be run on individual machines like a NVIDIA® DGX-1 [1]. NVIDIA DGX-1 unlocks the full potential of the latest NVIDIA® Tesla® V100, including next-generation NVIDIA NVLink™, and the new Tensor Core architecture. DGX-1 delivers one full petaflop compute power, and supports container-docker software architecture, enabling both on-premise and on-cloud solution building.

High Performance Whole Genome Sequencing Analysis

Parabricks has developed a GPU-powered software solution that provides 15-40 times faster secondary analysis of raw sequencing data (FASTQ files) generated by sequencers to identify potential variants (VCF files) for tertiary analysis. The standard computational workflow shown below consists of three steps and are defined as the Genome Analysis Toolkit (GATK) from the

Broad Institute. Parabricks accelerates the existing GATK best practices pipeline to generate results that exactly match CPU-based results (100%), but are much faster.



Parabricks' technology (called DNA Bricks) has accelerated these standard computational tools by running them on GPUs. DNA Bricks can accelerate 42 hours of computation on a 32 vCPU machine running on a public cloud (AWS) to less than one hour on a local node/server with eight GPUs. Each stage of the pipeline has been accelerated to use all the resources on a machine, such as CPUs and GPUs, while implementing the exact same algorithm parallelized on a GPU server. This has resulted in each stage of the pipeline generating the exact same result using DNA Bricks. Users of DNA Bricks can choose which steps of the pipeline to run and can configure their own pipelines. The default settings of the Parabricks' GPU accelerated software run the exact GATK best practices pipeline in the prescribed order.

Pipeline Architecture

The GPU accelerated pipeline developed by Parabricks is shown below. It follows all the steps of the GATK best practices pipeline and can be used as an in-place replacement for the full pipeline. Because DNA Bricks uses GPUs for the major compute intensive portions of the pipeline, it can stream the output of one stage of the pipeline to the next stage. This allows multiple steps of the pipeline to run simultaneously. Overall, the Parabricks pipeline is broken into three stages and each stage is working on 2 to 3 steps of the GATK best practices pipeline simultaneously.

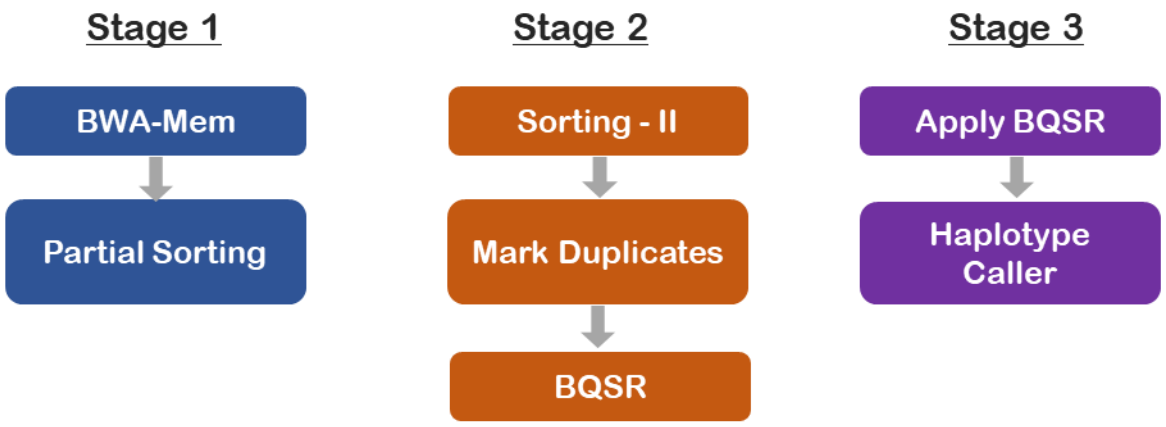


Figure 1: Architecture of the Parabricks Pipeline

Parabricks' Internal Performance

Following are results from Parabricks' internal performance testing. In the next section, we present and compare the results of BIOTECH's independent testing on a DGX-1 platform.

Pipeline Run Details:

Tools		Data	
BWA-Mem	v0.7.12	Type	WGS
Picard SortSam	v2.10.6	Sample	Human
Picard MarkDuplicates	v2.10.6	Read Length	151
GATK4	v4.0	Coverage	25x,40x

Baseline Run:

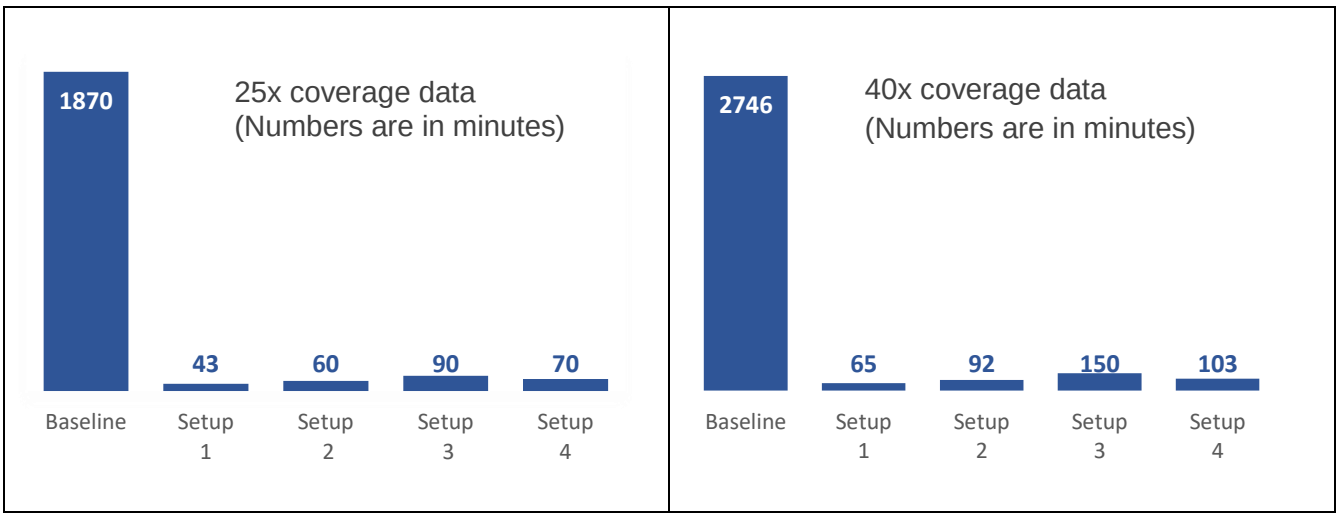
The baseline pipeline, comprising BWA-Mem version 0.7.12, SortSam, MarkDuplicates and GATK4, was run on a 32-vCPU machine on AWS (c3.8xlarge). The settings for the above tools were to generate the most accurate results while maximizing all computing resources (threads, cores). For BWA-Mem, 32 threads were used for alignment and for GATK Haplotype Caller, 16 threads were used per pairHMM.

Parabricks Run:

The Parabricks software was run in four settings with default parameters:

Setup #	Platform	GPU	Price/hr	Price/hr (Pre-emptive)
Setup 1	DGX-1	8 V100	N/A	N/A
Setup 2	AWS	16 K80	14.4	5
Setup 3	Google	4 P100	7.7	3.6
Setup 4	AWS	4 V100	12.2	5

The results below are for two samples of human WGS with read length 151 base-pairs and 25x and 40x coverages, respectively. The sequencer used for generating the data was an Illumina HiSeq X10.

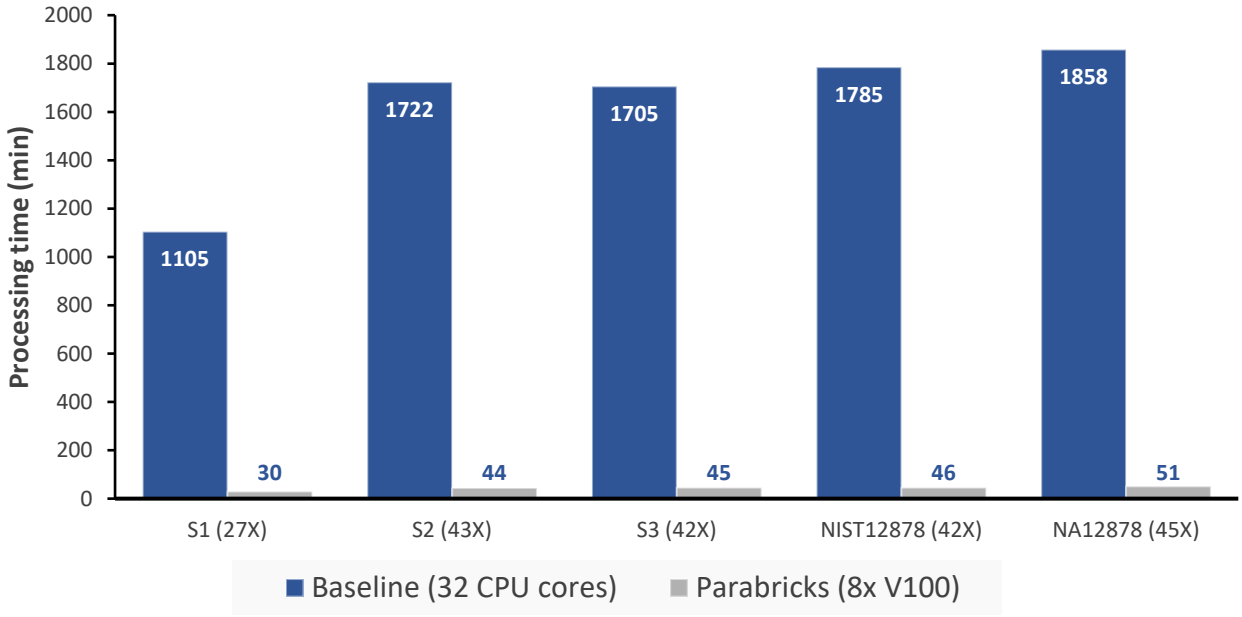


Accuracy

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 GATK is non-deterministic and can generate slightly different results based on certain parameters. For this step, all four Parabricks setups generated the same outptut which was within **0.99995** of the GATK execution. The differences are comparable to the variation in GATK execution due to random numbers, multiple threads, and similar factors.

BIOTEC Performance Test Result

BIOTEC performed its own internal testing of DNA Bricks on a DGX-1 platform to identify potential variants following the GATK Best Practices pipeline, including these four main tasks: 1) raw reads alignment with BWA-Mem 2) sorting and BAM conversion with Picard tools version 2.9 3) Picard mark duplication and 4) Base Quality Score Recalibration (BQSR) with GATK version 3.7. All experiments were performed on the new DGX-1 node with dual 20-core Intel® Xeon® E5-2698 v4, 20-core, 2.2GHz, 512 GB memory and eight Tesla V100s. Five human WGS datasets with different read coverages were used in this performance evaluation. The test mainly measured processing time subject to the varying sizes of the WGS data. BIOTECH's testing confirmed that DNA Bricks operating on DGX-1 significantly outperformed that of the CPU-only baseline case—with speed gains varying from 36x to 40x.



Cost Benefit

The on-premise server version of DNA Bricks used one DGX-1 server. The throughput achieved by this box would be nearly 8,500 whole genomes per year at full utilization. In comparison, the baseline CPU-only solution would require 35 servers, each with 32 vCPUs. This increases the

Total Cost of Ownership (TCO) significantly. Managing 35 servers requires a dedicated IT infrastructure and significantly higher power and cooling efforts. Furthermore, balancing the jobs and maintaining priority becomes a very complicated scenario. With one DGX-1 server, no dedicated IT infrastructure is required. Furthermore, DGX-1 boxes are the fastest servers for AI and Machine Learning, which no CPU-only solution can match. Therefore, the DGX-1 server with DNA Bricks is the most cost-effective and fastest solution for high-throughput secondary analysis on the market today. Parabricks reports that DNA Bricks scales nearly linearly from 4 GPU to 8 GPU platforms for higher volume sites.

Features and Benefits

Speedup is what to be expected when adopting a GPU solution. Parabricks also offers extra benefits which surpass bioinformatic traditional practices. The whole GATK suite is optimized in terms of both speed and required temporary disk space. After BWA alignment, the output BAM file is compressed using a much better compression ratio than the standard out-of-the-box BWA tool which requires extra space to store the SAM file. For example, 42x coverage of a NIST12878 dataset requires only 75GB of storage using DNA Bricks, compared with 624GB (360GB of SAM file, 58GB of sorted BAM, 59GB of mark-duplicated BAM and 147GB of BQSR BAM) produced by the baseline.

By following the best practice, the output BAM file is nicely sorted where duplicates are marked. DNA Bricks is very streamlined and can be run with minimal effort to generate the output. Splitting GATK best practices enables Parabricks to maximize the computational throughput via keeping each stage in the pipeline busy.

Summarizing the key features:

- **Equivalent results:** Each stage of the Parabricks' pipeline generates nearly identical results as baseline GATK best practices pipeline.
- **Supports all tool versions:** Parabricks' accelerated software supports multiple versions of BWA-Mem, Picard and GATK and will be updated in near real-time to support all future versions of these tools.
- **Leverage Machine Learning:** Deep learning on genomic data for tertiary analysis can be potentially applied on the same platforms. GPUs are the best platforms for such analysis.
- **Turnkey Solution:** DNA Bricks runs on standard GPU nodes available on the cloud and requires no additional setup steps by the user.
- **On-Premise and Cloud-Agnostic:** DNA Bricks can run on DGX-1 [1] on-premise or any public or private cloud with GPUs and is tested on DGX-1 [1] servers, Amazon Web Service (AWS), Google Cloud Services and Microsoft Azure.
- **Visualization:** Parabricks generates several key visualizations while performing secondary analysis that can improve the user's understanding of the data.

Related Publications

1. DGX-1V Datasheet https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-1/dgx-1-ai-supercomputer-datasheet-v4.pdf

2. Li H. and Durbin R. , “Fast and accurate long-read alignment with Burrows-Wheeler Transform”. Bioinformatics, Epub, 2010
3. Broad Institute. “Picard Tools.” Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/
4. Poplin *et al.* : Detailed description of HaplotypeCaller; best reference for germline joint calling, 2017

Acknowledgements

The authors would like to acknowledge NVIDIA for the use of a DGX-1 and providing feedback and support, which made this work possible.