Deep Genomics: A Software/Hardware framework for integrating Deep learning into Genomics

Ramachandra Chakenalli Nanjegowda

October 15, 2018

1 Overview

Genome sequencing technologies such as Nanopore has not only democratized DNA sequencing but also caused dramatic increase in the number of genomes sequenced and sequencing capacity. However, with error rates as high as 30 percent, these technologies pose a challenge to obtain high accuracy using conventional methods and existing state-of-the art tools, and hence are not widely used for human genome analysis. To overcome this challenge some of the phases of the Nanopore genome analysis pipeline make use of deep learning tools. But these tools do not fit into the conventional genome sequencing workflow due to the heterogeneity of the data representation, computational resources needed and programming models used. Moreover, these tools are based on deep learning paradigm of train and predict, for which there is no well defined genome analysis workflow. In this project we propose a methodology and framework to integrate these deep learning tools into the conventional genome analysis pipeline. Through different contributions we aim not only to address the challenge of heterogeneity of the data representation, computational resources needed and programming models, but also provide a platform to reap the benefits of crossing the Genomic and deep learning domains together.

2 Intellectual merit

This project will address the challenges mentioned above, which are (1) heterogeneity of the data representation (2) variety of programming models used by different tools and (3) dynamism of the computational resources required. To address these challenges the project will build a software system to integrate some of the deep learning tools into the genome analysis pipeline by using a unified programming model of TensorFlow. The project will use an unified data representation to efficiently manage data formats, compression and encoding. On the hardware side, this project will attempt to build a prototype test bed, comprising of CPU, GPU and accelerators to cater to the variety of computational resource requirement.

3 Broader impact

Genome sequencing technologies such as Nanopore intend to enable analysis of any living thing, by any person, in any environment. By advancing the software pipeline of Nanopore technology, this project will influence precision medicine work flow. The project will impact several precision medicine workflows by improving the accuracy and efficiency of the analysis pipeline, and thus enable the use of Nanopore technology in clinical realm for human samples.
4 Introduction

Right now, genomic research is on the cusp of making breakthrough contribution to biomedical science through the benefits of total amount of sequence data[1]. The field is approximately where deep learning was a few years ago, which has seen dramatic advances driven in large part by availability of vast data sets. To be able to make advances in similar pace as deep learning, the genomic research should address looming 5V problem, which is quite similar to 5V problem of big data domain. Bioinformatics framework like Persona[3] has been architectured ground up to address these 5V problems.

• Volume: Due to the growth of genome sequencing technology, there has been a dramatic increase in both the number of genomes sequenced and the sequencing capacity. This has caused an explosion of genomic data and presents an unavoidable challenge of inconceivable amount of digital storage.

• Velocity: With the advent of genomics into clinical realm, the velocity or speed at which the analysis can be completed with minimum computational cost and high scalability is vital. But the heterogeneity of bioinformatics tools and programming models used to implement these tools poses a major hurdle to achieve velocity at cost in the time-critical clinical applications such as personalized cancer profiling and newborn screening.

• Variety: A crucial challenge in scaling computational genomics is that, genomic data is stored in multiple formats, none of which are suitable for parallel or distributed computation. Sequencing machines produce raw genomic data in one format (FASTQ) while aligned data uses different format (SAM/BAM) and downstream analysis produces more files with different format. This not only results in data duplication, but also presents challenge to effectively partitioning the data to achieve distributed computing.

• Veracity: The advent of third gen sequencing technologies like Nanopore has truly democratized genome sequencing. However high error rates, these technologies pose a challenge for accurate downstream analysis. Conventional error correction techniques are inadequate to achieve the required level of veracity when the error rates can go up to 30 percent using these technologies.

• Value: Comprehensive interpretation of human genome and deriving value out of it is a challenging problem. It typically requires weeks of analysis, with extensive hand-on expert involvement using wide range of tools, deployed on various setup ranging from single node server, gpus, cluster or cloud. This presents a steep learning curve for bioinformatician to extract actionable insights which could be identifying type of pathogen or mutation.

Persona is a cluster scale, high throughput, bioinformatics framework which is designed ground up to address these 5v problems. It implements a new Aggregate Genomic Data (AGD) format to unify different data formats (variety) and enables efficient columnar compression and storage (volume). Persona is built on Google TensorFlow, which is a state-of-the art distributed data flow framework and enables the computations to run on heterogeneous distributed platform at minimum computational cost and high scalability (velocity).
5 Proposed work

This project will have the following sub projects and contributions.

- Extend the existing implementation of Persona [3] to support the below tools and phases to achieve the project objective.

- The state-of-the-art sequencing tools Minimap 2[4] for third generation sequencing and Arioc[7] for second generation sequencing will be added to Persona to do sequence alignment at high precision and accuracy.

- Integrate Clairvoyant[5] or Deepvariant[2] which are the two state-of-the-art deep learning variant calling tools to detect genome variants.

- Integrate Chiron[6], a deep learning base calling tool into Persona to process raw signals in fast5 format and directly store data in AGD format.

- Use this framework to evaluate the full pipeline using Human genome data sequenced using the Nanopore MinION with an R9 flow cell.

- Using the proposed methodology we establish a baseline performance on an CPU, GPU and accelerator based test bed.

- Redesign the algorithms to take advantage GPU, TPU or accelerators and demonstrate the performance improvement over baseline.

6 Preliminary work

Some of the preliminary work and milestones accomplished so far in this project are:

- Completed the setup of different second and third generation genome analysis workflows. These workflows include (a) GATK and Sentieon best practices workflow for variant discovery (b) Nanopore workflow for De Novo assembly (c) Third generation sequencing and variant calling for discovery of structural variants.

- Completed the measurement of the baseline performance of the above mentioned second and third generation genome analysis pipelines on a canonical multi core platform setup on CHPC infrastructure.

- Established a basic methodology to measure the precision and accuracy of the genome analysis workflow using precision FDA dataset.

- Established a basic methodology to measure the efficiency of structural variant discovery workflow using dbvar dataset.

References


