Integrated Approach to Whole Genome Diagnostics

Zaid Al-Ars[1,2], Koen Bertels[1,2] and Edwin Cuppen[3]
[1]Computer Engineering Lab, Delft University of Technology, Delft, The Netherlands
[2]BlueBee Technologies, Delft, The Netherlands
[3]Dept. of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
E-mail: z.al-ars@tudelft.nl

1. Introduction

The application of genetic analyses in the clinic is being revolutionized as a result of fast developments in *next-generation DNA sequencing (NGS)* technologies. Patients with diseases with a suspected genetic basis can have their DNA sequenced in order to understand the basis of the disease and to assist treatment choice. Such personalized treatment, however, still faces three major challenges. First, data generation and storage is still laborious and costly. Second, data analysis currently requires highly specialized ICT infrastructure and expertise. And thirdly, data interpretation to guide clinical decisions remains a major challenge. This abstract presents an approach for an integrated solution to address these three challenges by establishing a generic out-of-the-box solution for NGS data storage, analysis and interpretation based on hardware acceleration. This should result in a strong reduction of genome analysis turn-around time as well as required ICT infrastructure and expertise, thereby facilitating broad implementation of NGS-based genome diagnostics.

2. Computational pipeline

Figure 1 shows a typical computational pipeline for DNA based diagnostics. Each stage of the pipeline is associated with one of the challenges mentioned above.

Generate
Analyze
Interpret

Figure 1. Different stages of DNA diagnostics pipeline

- Stage 1: Data storage---The objective of this stage is to read the genetic data stored in a DNA sample of a given patient. On average, each position in the genome is read between 30 and 50 times by hundreds of millions reads of 2 x 100 nucleotides. This results in huge amounts of data that needs to be stored, manipulated and archived.
- Stage 2: Data analysis---The objective of this stage is to reconstruct (mapping step) the patient DNA from the many millions short reads generated by Stage 1 and identify the variations of the patient DNA compared to a reference DNA (variant calling step). This stage is rather computationally intensive and requires top of the line ICT infrastructure.
- Stage 3: Data interpretation---The objective of this stage is to understand the analyzed genetic data and use it to make clinical decisions for the patient. This stage is challenging from a biological point of view, and requires having appropriate visualizations of the many giga bytes of genetic data and meaningful annotations of relevant information.

3. Approach and methods

We have been working on an integrated approach to address the challenges faced by the different pipeline stages. At the heart of this approach is the utilization of custom-made hardware units synthesized on FPGAs (field programmable gate arrays) to make it possible to implement complex algorithms and yet insure high-throughput at low total system cost.

- Solutions in Stage 1: To ensure efficient utilization of available storage, new specialized compression techniques are being investigated that go beyond standardized compression techniques such as gzip [1]. Compression rates of 15% can be expected.
- Solutions in Stage 2: On a standard computer cluster, Stage 2 can take a couple of days to complete, making it a bottleneck in the diagnostic pipeline. There are a couple of publications discussing hardware acceleration of genetic algorithms [2]. We are using similar techniques to achieve more than an order of magnitude acceleration for the analysis stage.
- Solutions in Stage 3: Visualization is still one of the less understood challenges in big data. The subjective nature of data interpretation makes solutions generated for this challenge as much art as they are science. A number of tools are available to enable insightful representation of genetic data while reducing the level of perceived complexity [3]. We are investigating visualization solutions in close consultation with the users to ensure their relevance to our pipeline.

References

1. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". Bioinformatics 25 (16): 2078–2079.
2. Hans Heideman, Kirby Collins, George Vacek, Jan Bot. “Eighteen-fold Performance Increase for Short-Read Sequence Mapping on Genome of the Netherlands Data using Hybrid-Core Architecture”, NBIC 2012.
3. Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski, “A survey of tools for variant analysis of next-generation genome sequencing data”, Brief Bioinform. (2013): bbs086v1-bbs086