Contribution ID : **75** Type : **not specified**

Towards Exabyte-Scale Genomics: Advanced Hardware-Software Solutions for Efficient Read Mapping

DNA analysis has become fundamental to various fields, including *disease treatment* [1], *outbreak surveillance* [2], *forensic investigations* [3], and *evolutionary studies* [4]. The rise of large-scale genomics and advanced sequencing technologies, such as Oxford Nanopore Technologies, has led to an exponential increase in data generation, far surpassing *Moore's Law* [5]. By 2024-2025, genomic data production is expected to range from exabytes (2^{50}) to zettabytes (2^{60}). This data explosion, coupled with progressively longer sequences constituting the new datasets, also called reads, poses significant challenges in terms of data storage, processing, and analysis. At the heart of these challenges lies the **read mapping** process, which is essential for aligning sequencing reads to a reference genome. The sheer volume of data now demands hours to days of computation, even on powerful servers with optimized tools. Input datasets often exceed hundreds of gigabytes, and peak memory requirements can reach tens of gigabytes, particularly for large genomes like the human genome. This massive computational demand highlights the need for more efficient solutions to handle the growing complexity and scale of genomic data.

Our goal is to tackle these challenges by proposing a hardware-software co-design approach that enhances the efficiency of read mapping in terms of both time and speed. Conventional General Purpose Graphics Processing Units (GPGPUs) and hardware accelerators like Field-Programmable Gate Arrays (FPGAs) are constrained by their limited memory capacity, making them insufficient for handling the entire read mapping workflow given the vast scale of the datasets generated. To overcome this limitation, we have developed a heuristic-based read mapping pipeline that significantly reduces the amount of data requiring computation. This reduction facilitates a dataflow-oriented model, optimizing the entire read mapping process for execution on massively parallel platforms such as GPUs and FPGAs.

[1] - M. M. Clark, A. Hildreth, S. Batalov, Y. Ding, S. Chowdhury, K. Watkins, K. Ellsworth, B. Camp, C. I. Kint, C. Yacoubian et al., "Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation," *Science translational medicine*

[2] - Ling-Hu, E. Rios-Guzman, R. Lorenzo-Redondo, E. A. Ozer, and J. F. Hultquist, "Challenges and opportunities for global genomic surveillance strategies in the covid-19 era," *Viruses*

[3] - Børsting and N. Morling, "Next generation sequencing and its applications in forensic genetics," *Forensic Science International: Genetics*

[4] - H. Ellegren and N. Galtier, "Determinants of genetic diversity," *Nature Reviews Genetics*

[5] - Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengx- iang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*

Department

DIETI - Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione - Università Degli Studi di Napoli Federico II

Primary author(s) : Mr. MERCOGLIANO, Stefano (Università degli Studi di Napoli Federico II)

Co-author(s) : Prof. CILARDO, Alessandro (Università degli Studi di Napoli Federico II)

Presenter(s) : Mr. MERCOGLIANO, Stefano (Università degli Studi di Napoli Federico II)