Abstract

The genome of an organism encompasses the unique set of genetic instructions for every individual in a species. The genome, in totality, guides the course of evolution, development, genetic and epigenetic growth factors of an individual. Genomics, the study of genome, presents an interdisciplinary landscape, with a multistage data analytics pipeline. Understanding the genome involves determining the order of the four constituent nucleotides or bases or genetic alphabets, namely adenine (A), cytosine (C), guanine (G) and thymine (T), within the genome's DNA sequence, and the process is widely known as sequencing.

Next Generation Sequencing (NGS) involves massively parallel sequencing of genetic data with high throughput. NGS offers an unparalleled interrogation of the genome, throwing deeper insight into the functional and structural investigation of genetic data. The deductions from such a study leave a huge impact across fields, including medical diagnostics, therapeutics and drug discovery, and as well form the basis for genomic medicine. Data processing with NGS happens over an elaborated multi-stage data analytics pipeline. During the primary data analysis, the sequencing process produces billions of short fragments, called short reads, of the target genome. This amounts to petabytes of unprocessed genomic raw data. Short read mapping (SRM) is the process of mapping these short reads to their respective positions in the target genome.

Due to the sheer volume of data that needs to be handled, SRM serves as a major sequential bottleneck to the NGS data analytics pipeline in genomics, and presents profound technical and computing challenges. Classified as a complex big data engineering problem, SRM thus calls for innovative computational, scientific and statistical

approaches towards big data analysis. A strict validation of various algorithms and softwares in an NGS pipeline is essential, to ensure reliable and accurate results.

With growing volume of NGS big data, the SRM and subsequent analytic steps demand a High Performance Computing (HPC) environment for data storage and analyses. Existing solutions for accelerating SRM provide notable performance, while leveraging heuristics and incurring significant error rates. Given the impact of the results of SRM in subsequent diagnostics and therapeutics, such heuristics and error rates are not affordable. In this context, we need precise, affordable, reliable and actionable results from SRM, to support any application, with uncompromised accuracy and performance.In this work, we present a massively parallel and scalable archetype, for accurate alignment of short reads, at a fine-grained single nucleotide resolution. The significant contributions of this work are presented below:

1. *We present a robust and efficient indexing scheme for the reference genome, which is devoid of heuristics. The scheme reports all possible regions of mapping for a short read, inclusive of repeat regions. The lookup scheme efficiently handles the redundancy in reads. Though this leaves the rest of the pipeline with more data for SRM as compared to the heuristic solutions, it provides the end user with reliable and actionable results.*

2. *We present an efficient parallel implementation of an accurate sequence alignment algorithm based on the Dynamic Programming (DP) method. Our alignment kernels can seamlessly perform the traceback process in hardware simultaneously with the forward scan, thus eliminating the computational and memory bottlenecks associated with such algorithms. These kernels thus report alignment in a minimum deterministic time, which forms the first level of acceleration for SRM.*

3. *We present AccuRA, a hardware accelerator targeting reconfigurable hardware platforms. The model scales well at multiple levels of granularity, which precisely aligns short reads, at a fine-grained single nucleotide resolution, and offers full coverage of the genome.*

4. *We present GMAccS, a GPGPU based solution, for the SRM accelerator. This employs a platform independent model, capable of targeting a heterogeneous set of GPU hardware.*

5. *We present a performance and scalability analysis model for both the archetypes. The results from the prototypes substantiate the scalability of these architectures at multiple levels of granularity.*

6. *We present the generalization of our solution across applications which exhibit similar data patterns in terms of volume, variety, rate of production and analysis, randomness and uncertainty involved in data, and use Approximate String Matching (ASM) as the fundamental operation for data analytics.*

7. *We present the various problems within the biological domain, in terms of complexity and quantity of data, to which our solution can be customized and scaled, at various levels of granularity.*

We have presented the results from various prototype models of both AccuRA and GMAccS. The AccuRA prototype, hosting eight kernel units on a single reconfigurable device, aligns short reads with an alignment performance of 20.48 Giga Cell Updates Per Second (GCUPs). AccuRA can be ported onto devices as diverse as SoCs, ASICs or reconfigurable platform based hardware coprocessors or accelerators. The scalability analysis proved to substantiate the parallel AccuRA architecture, making it a promising target to accelerate the SRM process in the NGS pipeline.

The in-house supercomputing platform SahasraT, which is a Cray XC40 system, hosted the prototype for the GMAccS archetype. The GMAccS prototypes align with an optimal performance of 23.69 Million Maps Per Second (MMPS) to 528.69 MMPS, while scaling from a single GPU to 24 GPUs. The performance model for GMAccS, as well as the results from the prototypes, substantiates the scalability of the GMAccS archetype and the subsequent performance enhancement achieved by it.

Both AccuRA and GMAccS accommodate the big data of genomics, with uncompromised accuracy, precision and performance, while aligning the smaller archeal, bacterial

and fungal genomes, to the much larger mammalian human genomes. These models have successfully handled redundant reads and multiread alignments. The results from AccuRA and GMAccS are available in the Sequence Alignment/Map (SAM) format, making it compatible with the downstream applications in the NGS pipeline.

With a basic parameterized SRM model, and the results from its various prototypes for small and large genome benchmarks, we have gained the confidence that this solution can serve the requirements of accurate and scalable alignment of NGS big data. We believe that our model can serve as a reliable candidate in the future of genomics, called the "genomic highway", where data belonging to multiple applications can be streamed in, and can be aligned real time, with minimal memory and storage requirements, on a generalized alignment engine.