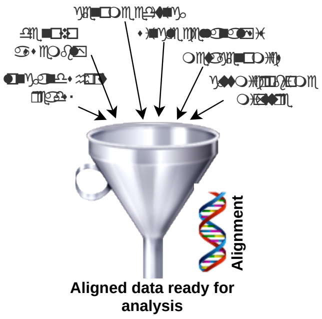
Development of a parallel algorithm for whole genome alignment for rapid delivery of personalized genomics

Sunita Chandrasekaran, Assistant Professor, University of Delaware, schandra@udel.edu

The Next Generation Sequencing (NGS) instruments are producing large volumes of data that is making the whole genome sequencing (WGS) a very important step for genomics research. Information gained from such volumes of data have been key to drug development and personalized medicine. Massive computing power offers tremendous capability to unwrap the complexity of biological systems and efficiently handle such massive genome (big) datasets. This challenge falls right into the intersection of computing systems and biology stimulating algorithmic innovation with sequence alignment on novel platforms.

With powerful sequencing data generated from instruments such as Illumina and Oxford Nanopore (long reads – a recent advancement) and a potentially transformative sequence alignment algorithm, we can make genomics a daily commodity. The state-of-the-art aligner is Burrows Wheeler Aligner (BWA) that is also tightly integrated into the gold-standard GATK best practices workflow, however BWA was not originally designed to take advantage of massively parallel processors. As a result, the algorithm is slow, memory inefficient, non-adaptable to hardware accelerators, non-portable across platforms and does now work well on long reads and only works well on short reads. An efficient aligner for long reads is very important as long reads are transforming our ability to assemble highly complex genomes, which the short reads cannot. BLASR is the state-of-the-art aligner for long reads but this algorithm has not been evaluated on massive computing systems, yet.

This demands the biology community along with the CS community to re-envision their goals and requirements in order to best exploit the tremendous opportunities the novel hardware platforms have to offer.



To address this challenge of bridging the gap between big data and HPC, our on-going research aims to create a novel parallel algorithm that can perform whole genome sequencing (WGS) faster while consuming less memory and not losing accuracy or sensitivity in comparison to classic sequence aligners. The most creative aspect of the proposed research is the seeding and the alignment algorithmic novelty that we carefully design after scanning a plethora of existing literature work and understanding the shortcomings of the same.

Our algorithm will be integrated into the widely popular and highly recommended industry standard Genome Analysis Toolkit (GATK) best practices workflow that is considered gold standard for variant calling (a process to identify variants from sequence data).

We plan to exploit the massive capability of hardware resources by creating a scalable algorithm that will not be suitable for just human genome but can also align sequences that are 3000 times larger than human genome, for example whiskfern. This way our algorithm will be propelling other areas of biology. Figure 1 indicates the benefits of such a faster sequence alignment.

The results of this fast WGS alignment with hybrid DNA read length will enable faster translation of basic scientific findings into personalized therapeutic interventions for patients thus increasing survival rates. Our algorithm will be used for sequence alignment of pediatric cancer dataset from children of age 0-12 years from Nemours/Alfred I. duPont Hospital for Children. The algorithm will also be used on project that uses millions of veterans' health data in order to understand the complex genetic underpinnings that affect medical disorders, drug interactions, drug specificity, and individuals' responses to pharmaceuticals. Our novel algorithm will also be used for sequencing 2500 human genome taken from 26 distinct groups of people from the world. In addition to human genome research where rapid alignment can cause significant improvement in the quality of care provided to the patients, our scalable algorithm can also propel other areas of biology such as sequencing whiskfern that is 3000 times larger than human genome. We aim to look into the field of genome editing (CRISPR) as well where we will soon see a huge leap in sequencing demand. By making our algorithm and the software open-source we are enabling wider adoption and participation from the community.