**Abstract**

Next Generation Sequencing (NGS) has become an essential tool in the effort to better understand the relationships between genomic variants and phenotypes. Dynamic programming plays an important role in read alignment, as well as variant calling both requiring immense computational power due to sheer size of the genomic data in NGS sequencing. Employing Intel Skylake generation instances on Google Compute Engine allowed us to utilize Intel AVX-512 instruction set for vectorization of these algorithms and 64 cores. Compared to previously used AVX2 and 32 cores, we have achieved a 96% speedup of our alignment and a 95% speedup of variant calling. This approach resulted in significant reduction in duration for a Whole Genome Sequencing pipeline from 8h 12m to 4h 53m on a 50x sequencing run. These results show that despite inter and intra register dependencies, it is possible to increase performance by vectorization and utilization of larger registers.

**Genome Alignment: Smith Waterman Algorithm**

Smith Waterman(SW) algorithm is a local alignment algorithm that calculates all possible alignments in O(MN) time. An example SW alignment calculation is as follows: The score value is either +3 if the letters match, -3 if they mismatch and -2 from left or top sides. Final alignment can be found by backtracking highest score from the maximum score's location.

In order to use Smith Waterman on a graph data, alignment scores on the graph edges the their corresponding directions must be merged to maximize total score. The resulting score matrix is a 2D flattened structure while direction matrix consists of pointers to the respective graph edges.

This operation can be vectorized using AVX2 or AVX512 internals to achieve 5-10x performance improvement throughout WGS pipeline. However, insertion score calculations pose a great setback to performance because of their intra-register calculation dependencies.

**Variant Calling: Pair HMM Algorithm**

Determining Haplotypes

Variant calling algorithms work defining active regions based on the reads. Reads in the active region are assembled de novo methods and haplotypes are determined from these reassemblies.

Determining Likelihoods

For each active region, the program performs a pairwise alignment of each read against each haplotype using the Pair-HMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. This part was consuming approximately 60% of Reassembly Variant Caller’s running time.

Assigning Sample Genotypes

Sample genotypes are calculated by finding maximum value in likelihood matrix for each read.

**Conclusions**

We have showed that WGS applications can be leveraged to compete with by using Intel AVX instructions on Intel Skylake architecture. We have concluded that, Striped Smith Waterman algorithm can be scaled not only with number of CPUs, but also with register size of the platform. By using 8 bit unsigned values for alignment score and 1024 bit register size, we can achieve NGS sequence score calculation in one iteration in the future. Also Pair HMM algorithm can be ported into single precision while avoiding underflow in probability calculations.

**References**

- [https://software.broadinstitute.org/gatk/documentation/article.php?id=4148](https://software.broadinstitute.org/gatk/documentation/article.php?id=4148)
- “Multiple sequence alignment using partial order graphs” Christopher Lee, Catherine Grasso Mark F. Sharlow Bioinformatics, Volume 18, Issue 3, 1 March 2002, Pages 452–464, [https://doi.org/10.1093/bioinformatics/18.3.452](https://doi.org/10.1093/bioinformatics/18.3.452)