Ultra-Fast, Accurate and Cost-Effective NGS Read Alignment Validated for Complex Whole Plant Genomes

Jos Lunenberg, Bas Tolhuis & Hans Karten

Genalice B.V. – Harderwijk, the Netherlands

For more information:
www.genalice.com
bas.tolhuis@genalice.com

GENALICE
TECHNOLOGY FOR PEOPLE & SCIENCE

PAG XXII

Introduction

As the rate of next-generation sequencing increases, greater throughput is demanded from read aligners. The ideal read aligner needs to be capable of more than just simple alignment. It needs to:

1. Be fast, sensitive and accurate
2. Find long, gapped alignments
3. Accept high upper read lengths (be compatible with Roche/454 and Ion Torrent)
4. Be suitable for paired-end and split-end data
5. Run on commodity hardware

The full-text minute index is often used to make alignment as fast as possible and memory-efficient. The most widely used full-text minute index read aligners are Bowtie [1] and Burrows-Wheeler Alignment (BWA) [2].
Here we present the data of an innovative ultra-fast new read aligner that uses a novel algorithm and reduces the storage footprint of the output file in comparison to Bowtie and BWA.

Methods

GENALICE MAP

GENALICE developed an innovative next-generation sequence read alignment tool. The tool uses FASTQ format as input. It comprehensively combines read alignment and variant calling. Its output is flexible and can be our unique data footprint reducing format (GAR = GENALICE Aligned Reads), SAM/BAM output format or Variant Call Format (VCF).

Data description

Data is derived from whole genome sequencing using a cultivated tomato species. It contains 2x164 million paired-end reads with a minimum length of 50 and a maximum of 100 bases per read. This results in an average coverage of approximately 30 times the tomato genome (~950 megabases).

System configuration

We compared BWA and GENALICE MAP performance using the same hardware configuration. The system that was used ran 2x Intel Xeon E5 2620 CPUs with 6 cores per CPU and 2 threads per core (24 threads in total). Total RAM memory is 96 GB and the system ran on a Linux x86-64 operating system (Suse Linux Enterprise Server 11 SP2).

Reference index build

Both BWA and GENALICE MAP require that the reference genome is indexed. We used the SL2.40 build of the tomato genome as a reference. BWA builds its index in 17 minutes and 46 seconds using the hardware configuration described above. GENALICE MAP builds its index in 3 minutes and 6 seconds.

Alignment

FASTQ files containing paired-end reads were aligned using either BWA (version 0.6.2-r126) or GENALICE MAP. Both tools used the hardware described above. BWA was run using 24 threads with default parameters.

Results & Discussion

Fast alignment

GENALICE MAP aligns reads faster than BWA. The runtime of BWA on 30x coverage of the tomato genome is 340 minutes and 11 seconds. GENALICE MAP aligns that same data set in 6 minutes sharp. The runtime consists of three phases. First a short prepare stage (16 seconds) in which the reference index is loaded into memory. Second the alignment phase lasts 4 minutes and 52 seconds. Finally, the aligned reads are written in the GAR format (52 seconds).

The average alignment speed of GENALICE MAP is around 97.3 million bases per second, whereas BWA aligns with ~1.7 million bases per second (Figure 1). GENALICE MAP aligned the reads 57-fold faster than BWA (Figure 2).

Alignment result

The alignment results of BWA and GENALICE MAP are highly comparable. When we compare depth of coverage between BWA and GENALICE MAP aligned reads we notice a high degree of similarity (Figure 3). The vast majority (98.4%) of all reads mapped with high confidence by BWA (MAPQ 60) were aligned to identical genome positions by GENALICE MAP.

Storage footprint reduction

GENALICE MAP uses a novel format to report its aligned reads, namely GENALICE Aligned Reads (GAR). This format results in a significant storage footprint reduction compared to the commonly used BAM and the unaligned FASTQ format (Figure 4). The GAR format is fully realignable and as such can replace both BAM and FASTQ files as format to store.

Results & Discussion *continued*

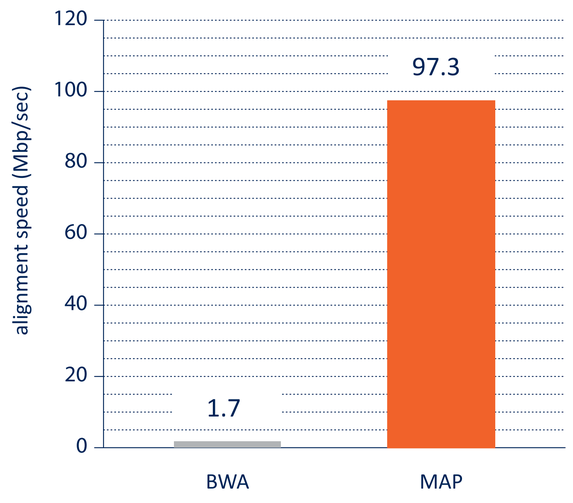

Figure 1. Average alignment speed
Performance of BWA (dark gray) and GENALICE MAP (orange) alignment speeds (megabases per second) are shown. For this data set BWA aligns at an average speed of 1.7 million bases per second. The average performance of GENALICE MAP is approximately 97.3 million bases per second.

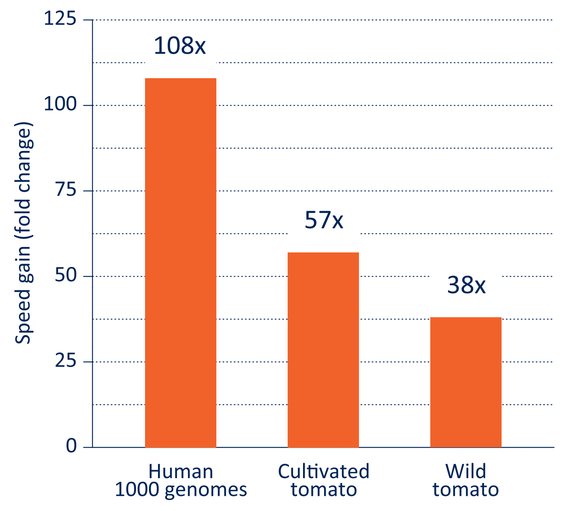

Figure 2. Alignment accelerations by GENALICE MAP
The speed gain (GENALICE MAP over BWA) is shown for three data sets. Visible are a human 1,000 genomes sample of 64 times coverage of chromosome 20 (NA12878), a 30 times full genome coverage of a cultivated tomato species and a 30 times full genome coverage of a wild tomato species. The latter sample is genetically very distant from the SL2.40 build of the tomato reference genome.

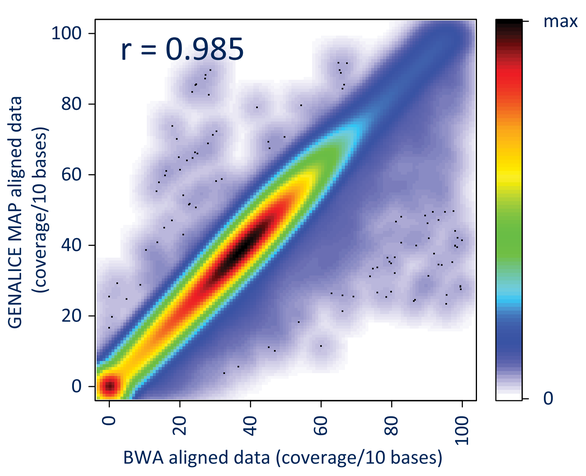

Figure 3. Depth of coverage analysis
Chromosome 2 was divided into consecutive bins that are 10 bases long. For each bin the average read coverage was calculated. The scatter plot compares depth of coverage of BWA aligned reads versus GENALICE MAP aligned reads. Heat map colors show density of the bins with white indicating no bins and black is the maximum density of bins. Pearson's correlation coefficient (r) is shown.

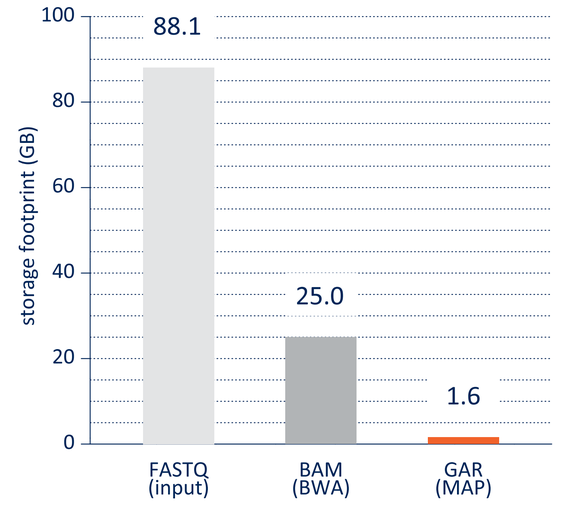

Figure 4. Read number comparison
NGS data formats for 2x164 million paired-end reads plotted as a function of the disk storage footprint in Gigabytes (GB). Three formats are compared. Input sequence reads in FASTQ format (light gray) require 88.1GB disk space. BWA aligned reads into BAM format (dark gray) use 25.0GB. GENALICE MAP's novel GAR format (orange) reduces disk space of all reads to 1.6GB.

Discussion

GENALICE MAP greatly reduces the computing power and processing time needed to align NGS short reads, while being highly accurate. Moreover, the GAR format minimizes data storage capacity and facilitates data sharing. We think that GENALICE MAP is a cost effective alternative for existing open source read alignment tools due to its high speed, using limited computing power and strongly reduced data storage footprint.

Conclusion

- **Speed:** much faster than existing tools on commodity hardware
- **High precision:** high accuracy in mapping reads
- **Small storage footprint:** 1,6GB for a realigned tomato genome
- **Comprehensive:** combines alignment and variant calling in one tool
- **Flexible:** mapping accuracy independent of read length

References

1. Langmead B., Trapnell C., Pop M., and Salzber S., Ultra-fast and memory-efficient alignment of short DNA sequences to the human genome, (2009) *Genome Biology* 10:R25
2. Li H. and Durbin R., Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, (2009) *Bioinformatics* 25:1754–1760