A new ultra-fast and comprehensive NGS read aligner with high precision

Jos Lunenberg, Bas Tolhuis & Hans Karten
Genalice B.V. – Harderwijk, the Netherlands

GENALICE
TECHNOLOGY FOR PEOPLE & SCIENCE

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

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 longer, gapped alignments
3. Accept higher upper read lengths (be compatible with Roche/454 and Ion Torrent)
4. Be suitable for paired-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 new read aligner that uses a novel algorithm and reduces the storage footprint of the output file. To compare its speed and accuracy we performed a comparative analysis using the SEAL simulation and evaluation suite.

Methods

GENALICE alignment

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

Validation runs

We used the SEquence ALignment (SEAL) evaluation suite [3] to test the performance of GENALICE alignment. SEAL simulates short sequence reads extracted from locations in a randomly generated genome and introduces sequence variations, including single base variations, and short INsertion/DELetions (INDELs) in those reads. After alignment, SEAL parses the alignment tool output and evaluates runtimes and accuracy.

In our validation, we included three popular and fast short sequence read alignment tools, Bowtie2 [4], Bowtie [1] and BWA [2], to directly compare their performance to GENALICE alignment. When comparing the four alignment tools we used identical simulated genomes, sequence reads and commodity hardware.

Results & Discussion

High precision in complex genomes

We tested performance in simulated genomes with 50% repetitive DNA sequences to mimic natural genome complexity. We applied two sequence variations in short reads with respect to the reference genome, namely: single base variations, and short INsertion/DELetions (INDELS).

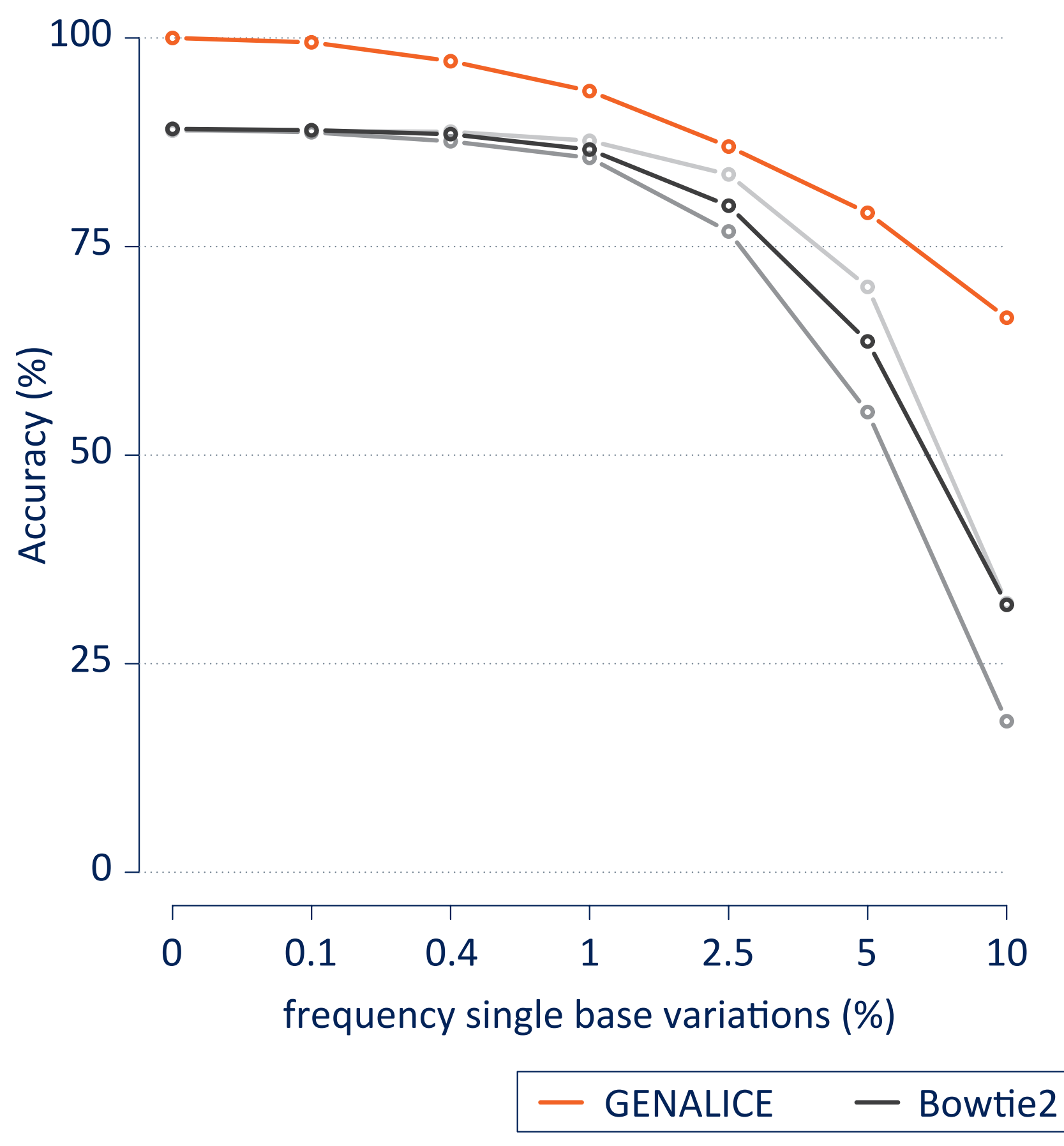

Figure 1. Accuracy of mapping reads in a complex reference genome with varying degrees of single base variations in sequence reads that need to be mapped by the alignment tools. GENALICE alignment (orange) always returns higher mapping precision than other tools (gray).

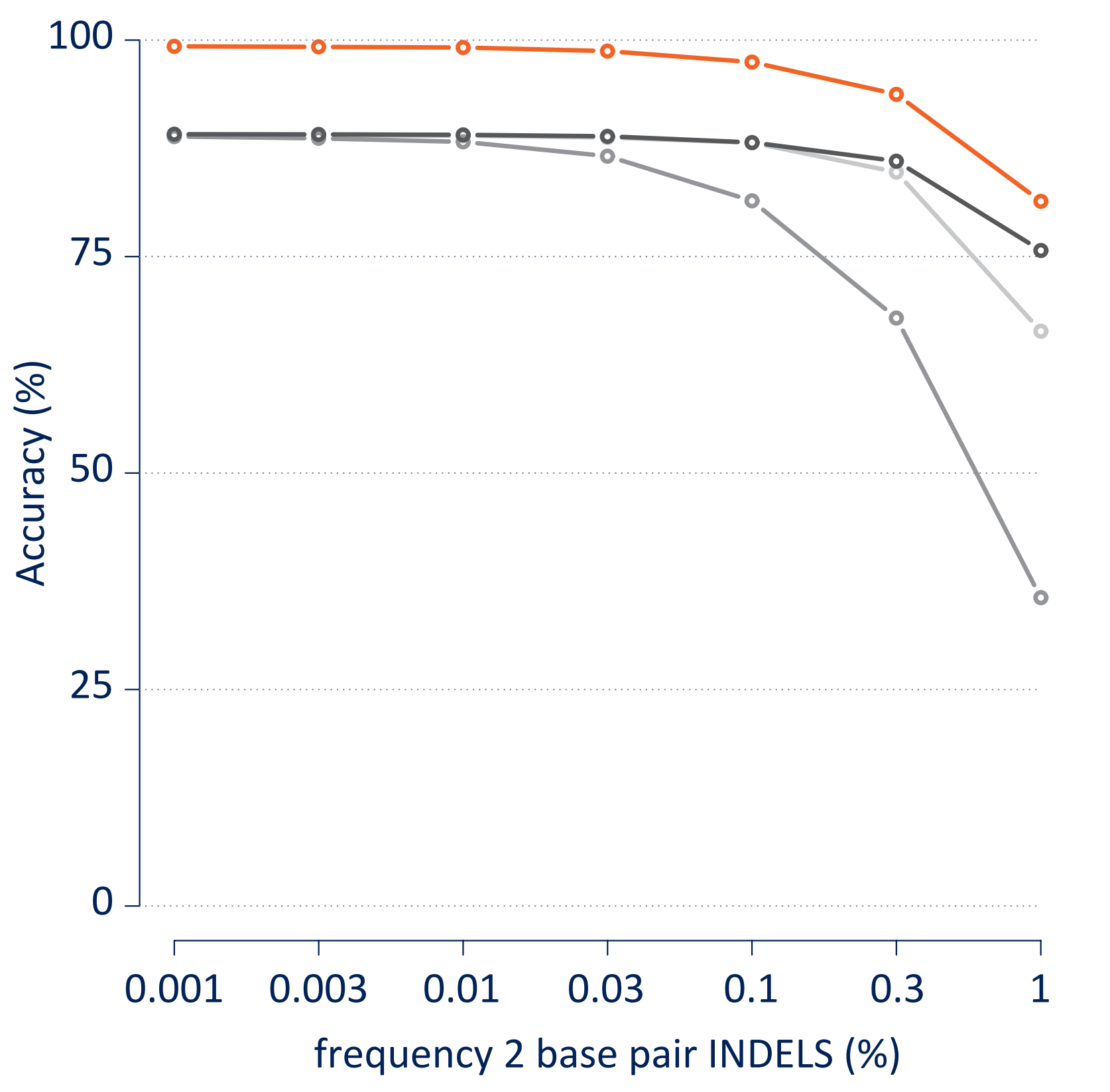

Figure 2. Accuracy of mapping reads that harbor INDELS with increasing frequencies in a complex reference genome. GENALICE alignment (orange) shows elevated mapping reliability compared to other tools (gray).

Results & Discussion *continued*

Ultra-fast alignment

Another important aspect of alignment is speed. For that purpose, we compared alignment runtimes between GENALICE alignment and other tools using commodity computer hardware.

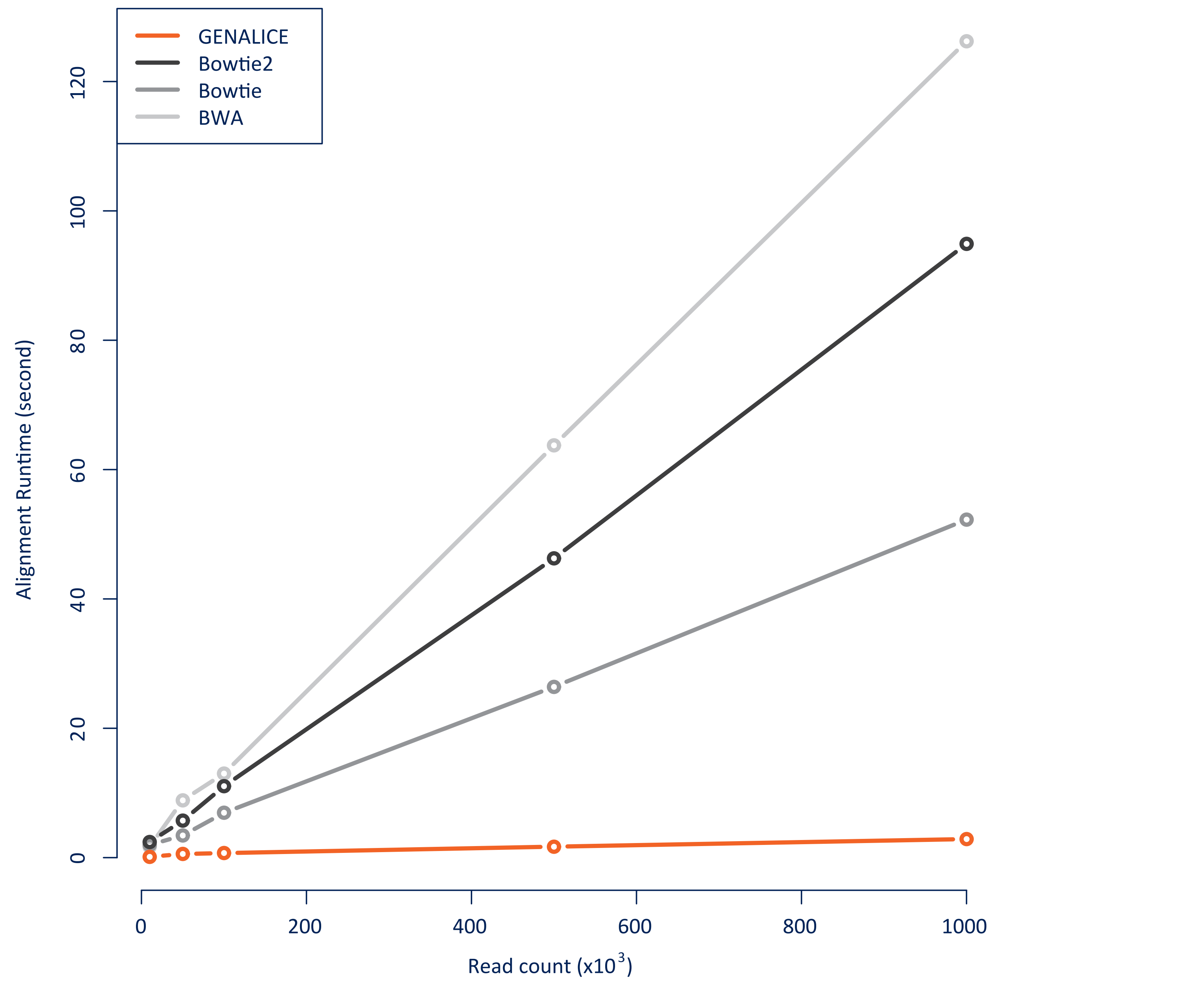

Figure 3. *Alignment speed of mapping increasing numbers of reads to a reference genome of 500 Mbp. GENALICE alignment (orange) performs better than other tools (gray).*

Alignment performance and accuracy are influenced by complexity of the reference genome and characteristics of sequence variations present in the short sequence reads. GENALICE alignment provides high precision and velocity mapping in complex genomes. This makes the new read aligner highly suitable for mapping sequence variations and mutations in complex genomes, such as the human genome.

In addition to its outstanding performance, GENALICE alignment has several other unique features. Its output file format allows a small storage footprint of only 4GB for all aligned reads (with 40x or 100x coverage) of a full human genome. In addition it comprehensively combines alignment and variant calling in one tool, requiring only a single pass to produce output for downstream analysis, and it is capable of mapping reads of any length.

We used a prototype of GENALICE alignment for the data presented here. Further improvements in speed and accuracy will be made, to sustain high quality alignment throughput of at least 600 Giga bases per day per CPU.

Conclusion

▲ High speed: much faster than existing tools on commodity hardware

▲ High precision: most accurate in mapping reads (SEAL provided objective measurements)

▲ Small storage footprint: 4GB for a realigned full human 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
3. Ruffalo M., LaFramboise T. and Koyutürk, Comparative analysis of algorithms for next-generation sequencing read alignment, (2011) *Bioinformatics* 27:2790-6
4. Langmead B. and Salzberg S., Fast gapped-read alignment with Bowtie-2, (2012) *Nature Methods*, 9:357-359