Benchmarking of DNA short read aligners on GCAT data sets

Rohit Gupta*, Ashwin Kalbhor, Pratibha Potla, Shanmukh Katragadda, Vamsi Veeramachaneni, Ramesh Hariharan

Strand Life Sciences, Bangalore India. *Contact: rohit@strandls.com

Abstract

Multiple aligners like Strand NGS, BWA, BWA-Mem, Bowtie2 and Novoalign3 are compared for accuracy and computational efficiency using 4 simulated data sets from the GCAT website and a real Illumina HiSeq 2500 whole-genome paired-end data of 1000 genomes CEU female sample, NA12878. Strand NGS and Novoalign3 showed comparable accuracy in terms of both, **%** correctly mapped reads and receiver operating curves (ROC). They also seem to outperform other algorithms especially on data sets with longer InDels. For reads potentially originating from complex genomic locations like repeat regions (and therefore assigned low mapping quality), Strand NGS aligner, with careful and intelligent filtering of false positives based on mapping qualities, produces a higher true positive rate compared to Novoalign3. As for the performance comparison based on computational efficiency, other than minor differences, practically all the included algorithms showed comparable performance.

Problem Statement and Challenges

- An accurate and efficient alignment of sequencing reads to a reference genome is crucial for many downstream applications. However, alignment is a challenging problem due to the following reasons:
  - A reference genome is typically long with repetitive elements
  - Reads are short in length (typically, 50 - 150bp)
  - Reads have sequencing errors
  - True alterations in the subject genome
- Numerous alignment approaches have been developed in the past to address these challenges. In this study, we compare the performance of Strand NGS aligner with several selected state-of-the-art algorithms.

Data Sets Used

- Description of simulated GCAT data sets used in this benchmarking study:

Data	Read length	InDel	Total # reads	With SNPs	With InDels	With both
D1	100bp	Short	11,945,250	1,202,587	313,753	31,731
D2	100bp	Long	11,945,250	1,195,002	308,029	31,135
D3	150bp	Short	7,963,500	1,158,693	304,788	44,449
D4	150bp	Long	7,963,500	1,164,160	310,750	45,782

- In addition, a real data from whole-genome sequencing on Illumina HiSeq of the sample NA12878 is used for benchmarking computational efficiency. This data is 103GB and comprises of 1,165,216,818 (1.16 billion) paired-end reads of length 150bp.

Algorithms Compared for Benchmarking

- Our Strand NGS aligner is benchmarked against other state-of-the-art algorithms such as:
  - Bowtie2
  - BWA
  - BWA-mem
  - Novoalign3
- Other than Novoalign3, which uses the hash table based index, others used BWT based index

Evaluation Approach

- Most alignment approaches consider trade-off between accuracy and efficiency. We assess the performance of different algorithms on both metrics, i.e., accuracy of read alignment, and computational efficiency. Accuracy is measured as:
  - Fraction (or **%**) of correctly, incorrectly and unmapped reads when alignment is done for all reads, reads with SNPs only, reads with InDels only, and reads with both SNPs and InDels.
  - Trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). TPR is **%** of correctly mapped reads and FPR is **%** of incorrectly mapped reads
  - Mapping quality distribution of incorrectly mapped reads
- To measure computational efficiency, total run-time of the algorithm is used as an evaluation criteria. This time includes the time taken for Burrows Wheeler Transform (BWT) search, Dynamic Programming (DP) around the seeds, and post-processing to produce final alignment results.

Results: Alignment Accuracy on Data Set D4

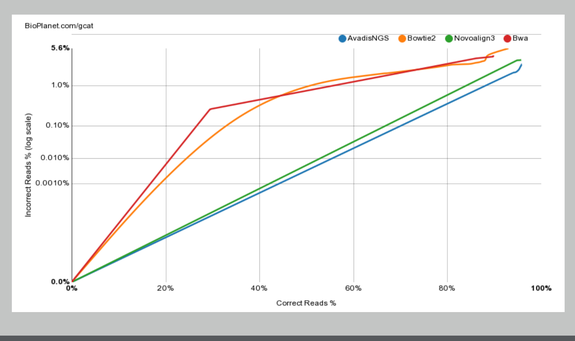
- Alignment accuracy of different aligners on GCAT data set D4

		All Reads	With SNPs	With InDels	With SNPs + InDels
Strand NGS	Correctly Mapped	99.22	99.11	96.58	95.88
	Incorrectly Mapped	0.5868	0.6452	2.550	2.728
	Unmapped	0.1938	0.2468	0.8692	1.396
BWA	Correctly Mapped	98.54	97.96	90.48	89.86
	Incorrectly Mapped	0.6914	0.6948	3.868	3.862
	Unmapped	0.7673	1.346	5.648	6.275
Bowtie2	Correctly Mapped	96.87	96.29	93.46	93.02
	Incorrectly Mapped	2.696	3.035	5.405	5.515
	Unmapped	0.4294	0.6718	1.139	1.468
Novoalign3	Correctly Mapped	98.99	98.94	95.80	95.72
	Incorrectly Mapped	0.1411	0.1441	3.34	3.355
	Unmapped	0.8672	0.9187	0.8608	0.9218

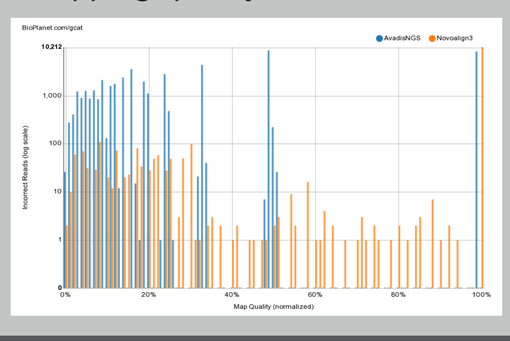
- On 45,782 reads that have both SNPs and InDels, Strand NGS and Novoalign3 produce both high TPR and low FPR compared to BWA and Bowtie2

Results: TPR versus FPR and Distribution of Mapping Qualities

- TPR and FPR



- Mapping quality distribution



Results: Accuracy on Reads with Low Mapping Quality

- Strand NGS, Novoalign3, and BWA-mem assigns a normalized mapping quality of 0 - 20 to 164,071, 158,498 and 158,083 reads respectively

	Strand NGS	Novoalign3	BWA-mem
Correctly Mapped	126,906 (77.35**%**)	88,834 (56.05**%**)	113,817 (72**%**)
Incorrectly Mapped	21,731 (13.24**%**)	603 (0.3804**%**)	44,266 (28**%**)
Unmapped	15,434 (9.407**%**)	69,061 (43.57**%**)	0 (0**%**)

- Strand NGS has a higher percentage of correctly mapped reads (high TPR) compared to both Novoalign3 and BWA-Mem

Results: Computational Efficiency

- Alignment of real whole-genome sample NA12878 is performed using the algorithms Strand NGS, BWA-Mem, and Bowtie2, on a machine with 64GB RAM and 15 cores
- Total time taken in aligning 1.16 billion paired-end reads against hg19 human genome reference is given below:

	Strand NGS	BWA-mem	Bowtie2
Total Time (in hrs)	9.5	12.18	11

Conclusion

- Alignment of millions of short reads to a large reference genome with many complex regions is still a hard problem and almost all current algorithms adopt some form of strategy to trade-off accuracy and computational efficiency. The benchmarking results presented in this study suggest that Strand NGS is a powerful approach for short read alignment and either compares well or even outperforms other state-of-the-art algorithms.