A SURVEY ON NGS - SHORT READ ALIGNMENT IN HIGH PERFORMANCE COMPUTING

G. Raja[1], U. Srinivasulu Reddy[2]

[1]Research Scholar, Department of Computer Applications, National Institute of Technology, Trichy, TamilNadu, India
[2]Asst. Professor, Department of Computer Applications, National Institute of Technology, Trichy, TamilNadu, India

Abstract

Next generation sequencing (NGS) is a new area for generating massive genome sequencing data at high speed with low cost. Due to this, there are huge problems for data storage, analysis, and management. At NGS, Short Read Alignment is a fundamental problem and it has a lot of applications, such as genetic variation, whole genome sequencing, and personalized medicine. Recently, many tools and algorithms have been developed to align the short-reads in NGS. This paper, describes a short survey on NSG-short read alignment in particular High Performance Computing environment. The authors discussed the performance of the most widely used tools like (BOWITE, BWA, SOAP, CLOUDBURST, etc.). Based on the results, CLOUDBURST gives the best performance compare to other methods.

Keywords: *NGS, Short read alignment, Hash, Tries, Cloud computing, and GPGPU*

--***--

1. INTRODUCTION

The modern NGS machines generate huge amount of short read sequences within few hours. A fundamental problem is, to aligning these short reads in proper formats such that they represent the actual genome. It has many applications such as genetic variation, whole genome sequencing, and personalized medicine [1]. NGS able to produce short reads, sizes ranging from 25basepairs (bp) to 400bp [2]. Roche 454, ABI Solid and Illumina are experimental technologies used in NGS. These methods consume a lot of time and having less accuracy rate for aligning short reads. Researchers have developed many computational tools like Bowtie [3], BWA [4], SOAP [5], Novoalign [5], Mr. Fast [6], MAQ [6], Cloudburst [7] and CUSHAW [8] in recent days for short read alignment.

The problems relating to genome sequencing are:
We don’t know the position information of the short reads, which part of the genome they came from and region of the corresponding short reads in the genome reference sequence.

- The genome reference sequence can be very long (3 billion bases for human) and making it an overwhelming task to find a matching region.
- Due to short reads, there may be many, similar places in the genome reference sequence which are matched to short reads are also read. This will occur mostly in repetitive regions.
- In the case of perfect matches to reference, we would not have any variation, so, we need to allow some structural variation and some mismatches in short reads.
- In case of errors in reads once should accept a lower level of sequencing errors.
- This process should be performed for lots of reads in sequencing data, which makes it tedious.

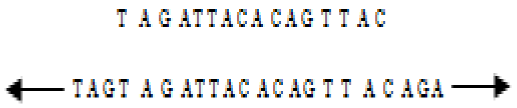

Fig -1: Short read alignment in NGS

Many quality problems arise due to Short Read Mapping Quality.

The low quality data represent a probabilistic wrong sequence, this wrong sequence will finally lead to wrong alignment. Sensitivity was used to determine the accuracy of an alignment algorithm. Mapping errors will happen because of low sensitivity, i.e., it misses the true matches in the short reads. Because of the repeated structure reads, repetitive regions generally, get very low mapping quality. [9].

2. A REVIEW ON SHORT READ ALIGNMENT TECHNIQUES

In this section, the authors described the indexing techniques like Hash table, Suffix tries and Burrows-Wheeler transform for short read alignment.

2.1 Short Read Alignment with Hash Table

Hash based methods can be performed by hashing the read and the reference genome sequences. The whole idea was to build a hash table for the short reads of a reference genome sequence. Here, key refers to short reads while value refers to the position where the short read appear in the reference genome.

BLAST uses a hash table and indexes every position in the query, with an index width of 9 nucleotides. It scans the genome and uses this index to find positions that needs to be examined closely for a match. Some hashes based algorithms index the genome, they are matched with, while others indexes the query.

Algorithms based on hash table must follow the seed-and-extend model. In case of BLAST, it keeps the position of k-mer short read of the query in a hash table along with the k-mer short read being the key, and scans the databases genome reference sequences for k-mer exact matching short reads [11]. In the larger short read sequences, increasing the size of hash table could reduce the execution time, but increases the memory size of hash table[12].

The seed –and –extend model used in various read mapping algorithms, it has fastest model for the short read mapping computation [7]. The key approach for this technique is that, so if a read maps to the genome at a particular location with a relatively small number of differences, then a significant fraction of the read must map, without any error. A conflict can arise due to exchange of character (mismatch), additional characters in the query (insertion), or leaving out characters from the reference (deletion).

2.2 Short Read Alignment with Suffix Tries

The structure of ordering tries into tree structures and practicing them to store strings for fast and exact matching was discussed in [11]. The applications requiring perfect genome matches can obtain fast results using tries. Imperfect matches would require backtracking the tires and can be costly if too many variances are taken into account. The main problem with using traditional tries for alignment is that the size of a tries for the full genome would end up being too large to be held in memory [12].

2.3 Short Read Alignment with Burrows Wheeler Transform

Burrow-Wheeler Transform is powerful data indexing technique and it maintains a very small memory when searching through a given data block. The FM - index technique was extended in BWT and support exact matching short read sequence to genome reference sequence. In the cases, genome sequencing transforms in to FM – index, where a single read matches multiple locations in the genome sequencing. BWT compared to hash table better quality in a short read alignment accuracy. An FM-index is just as effective as a tries, but is much more space efficient [14].

2.4 Short Read Alignment with GPGPU Computing

Currently, GPGPU developed can be performed by NGS short read alignment. GPGPU can reduce the execution time for short read alignment and sequencing error [15]. In the GPGPU sequence alignment based on BWT, to get faster mapping into sequence reads and genome reference sequence. GPGPU based short read alignment tools can be easily to align the lengthy reference sequence too short reads. The encoded reference sequence and short read sequence transfer from disk to GPU Memory. A GPGPU alignment kernel where the alignment task of each of the short read sequences is distributed to hundreds of processors within the GPU [17] [18].

2.5 Short Read Alignment with Map Reduce

In the mapping phase it generates key-value pairs of seeds which are patterns of the reads as well as the reference genome. In the shuffling stage the seeds are grouped based on the keys, which collects together similar keys between the short reads and the genome reference sequence. On the reduction stage an exact alignment with the allowed mismatches is found by extending the parallel keys, performing an end-to-end alignment, via these processes the seed and extend technique is implemented [19].

A review of some of the short read alignment tools including Bowtie, SOAP, Novoalign, MAQ, BWA and Mr. Fast is presented in Table 1.

Table-1: Short read alignment tools

Features	BWT indexing techniques		Hash indexing techniques			
	Bowtie	SOAP	Novoalign	MAQ	BWA	Mr. Fast
Fast	High	High	Low	Low	High	Low
Sensitivity	Low	Low	High	High	High	High
Accuracy	Low	Low	Low	Low	High	High
Memory	High	High	Low	Low	Low	Low

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this study, the short read alignment tools are analyzed by aligning short reads on the human genome to genome reference sequence. The system configuration details are listed in Table 2.

Table- 2: System Configuration details

Item	Equipment
CPU	Intel®core2Duo processor
Memory	@2. 20 GHZ
Hard Disk	3.00 GB
OS	500 GB
Compiler version	Ubuntu 13.04 desktop
	gcc version 4.7.0

The performance of short read alignment accuracy and CPU time has been evaluated by comparing it with Bowtie (version 0.12.7), BWA (version 0.6.2), MAQ (version 0.7.1) and Mr Fast (version 2.6.0.1), SOAP (version 1.0), NovoAlign (version 1.0), Cloudburst [7] and CUSHAW [8] using simulated and real short read sequence data sets

Table 3: Short read Alignment accuracy results (in percentage)

Datasets (Pair End)	Bowtie	BWA	MAQ	mrFast
SRR000129PE	98.3	98.7	97.8	98.4
SRR022868PE	98.7	99.6	96.7	96.3
SRR068211PE	97.2	98.4	98.1	98.5
SRR000133PE	96.8	96.4	95.1	95.3

Table -4: Short read Alignment accuracy results (in percentage)

Datasets (Pair End)	Novoalign	SOAP	CUSHAW	Cloud burst
SRR000129PE	97.5	98.3	98.1	98.6
SRR022868PE	96.1	99.3	99.5	99.1
SRR068211PE	97.9	98.3	98.7	98.9
SRR000133PE	95.3	96.1	98.6	99.1

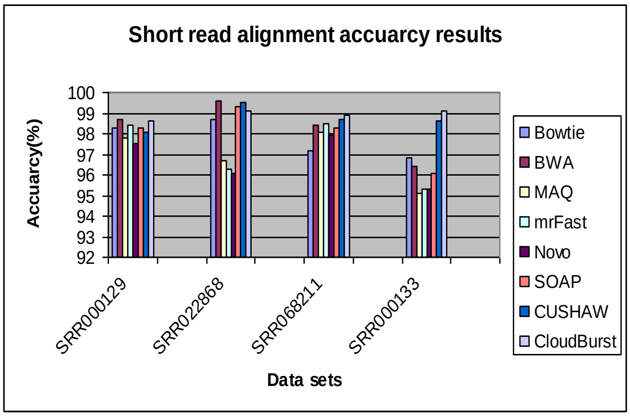

Chart -1: Short read alignment accuracy results

All aligns were assessed using four real data sets from 454, Ion Torrent, and Illumina sequencers. All these data sets are publicly existing and named after their accession numbers in the NCBI Sequence Read Archive (SRA). Measurements were carried out in terms of CPU time and clock run time. CUSHAW and CloudBurst shows consistently low Clock and CPU times, while Novo align demonstrates the highest clock cycles and CPU time (Chart 2, 3).

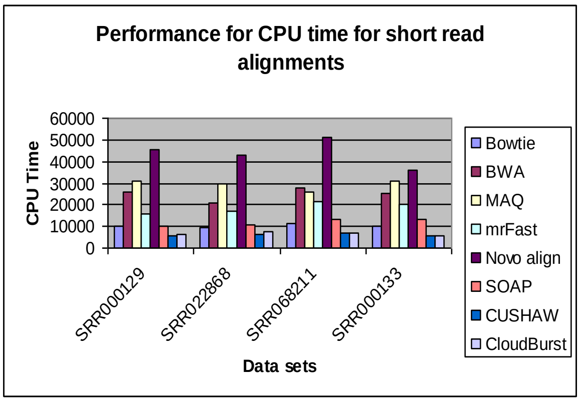

Chart -2: Performance for CPU time for short read alignments

Table -5: Performance for CPU time for short read alignments

Datasets (Pair End)	Bowtie	BWA	MAQ	mrFast
SRR000129 PE	10158	26786	31367	15989
SRR022868 PE	9874	21137	29876	16876
SRR068211 PE	11739	276.57	26780	21785
SRR000133 PE	10215	26321	32987	20438

Table -6: Performance for CPU time for short read alignments

Datasets (Pair End)	Novoalign	SOAP	CUSHAW	Cloudburst
SRR000129 PE	45276	10587	5835	6784
SRR022868 PE	42987	11342	6921	7598
SRR068211 PE	51456	13893	6739	7338
SRR000133 PE	35965	13659	6127	6073

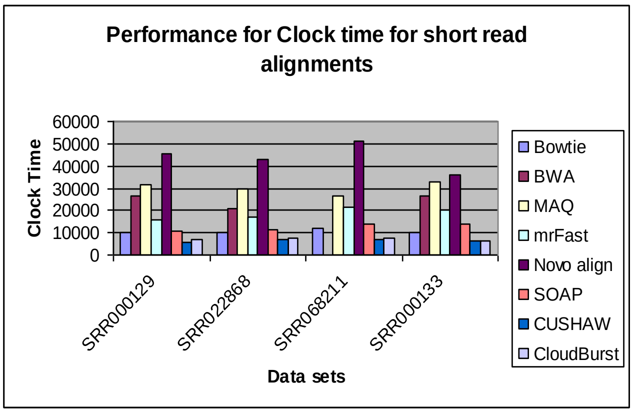

Chart -3: Performance for Clock time for short read alignments

Table -7: Performance for Clock time for short read alignments

Datasets (Pair End)	Bowtie	BWA	MAQ	mrFast
SRR000129 PE	10158	26786	31367	15989
SRR022868 PE	9874	21137	29876	16876
SRR068211 PE	11739	276.57	26780	21785
SRR000133 PE	10215	26321	32987	20438

Table -8: Performance for Clock time for short read alignments

Datasets (Pair End)	Novoalign	SOAP	CUSHAW	Cloudburst
SRR000129 PE	45276	10587	5835	6784
SRR022868 PE	42987	11342	6921	7598
SRR068211 PE	51456	13893	6739	7338
SRR000133 PE	35965	13659	6127	6073

All aligns were assessed using four real data sets from 454, Ion Torrent, and Illumina sequencers. All these data sets are publicly existing and named after their accession numbers in the NCBI Sequence Read Archive (SRA). Measurements were carried out in terms of CPU time and clock run time. CUSHAW and CloudBurst shows consistently low Clock and CPU times, while Novo align demonstrates the highest clock cycles and CPU time (Chart – 2,3).

It can be observed from chart 1 that on an average, Cloudburst performs well, and CUSHAW follows it with a very small lessening in performance with a few data sets, while lagging behind in others. Hence, it cannot be considered to provide consistent performance. Cloudburst and CUSHAW exhibit consistent performance, hence can be considered to be highly reliable when compared to other algorithms.

4. CONCLUSION

Next Generation Sequencing machines clue to computationally difficult alignment problems that can take many hours on a modern computer. Many studies have been passed out to analyze the performance of short read alignment. Short read alignment is supposed to be the computing bottleneck in analysis of new genome sequencing data. Fortunately, the dynamic development of short read alignment algorithms solved this problem along with better high throughput of sequencing machines. A survey of the short read alignment tools is presented here. Their efficiency and time taken to perform tasks were analyzed experimentally and presented in this paper. From our study, we found that most of the approaches are inherently parallel and can benefit from a distributed storage and processing architecture like Hadoop. We also consider these applications to CUDA C to make use of the massively parallel nature of GPGPU.

REFERENCES

[1] Ansorge WJ. Next-generation DNA sequencing techniques. *Nat. Biotechnol*. 2009. 25:195–203.

[2] Yongchao Liu et.al. CUSHAW2-GPU: Empowering Faster Gapped Short-Read Alignment Using GPU Computing, Design & Test, IEEE. Feb.2014. (Volume: 31, Issue: 1).

[3] Langmead B et.al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol*. 2009. Vol. 10, R25.

[4] Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*.2009. 25(14):17541760

[5] Li R, Li Y, Kristiansen K and Wang J. SOAP: short oligonucleotide alignment program, *Bioinformatics*. 2008. Vol. 24, pp. 713-714.

[6] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008. 18:1851–8.

[7] Schatz MC. Cloud Burst: Highly sensitive read mapping with MapReduce, *Bioinformatics*. 2009. 25:1363-1369

[8] Yongchao Liu et.al.,CUSHAW. A CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform, *Bioinformatics Advance Access published* May 9, 2012.

[9] Ayat Hate et.al. Benchmarking short sequence mapping tools, Hatem et al. *BMC Bioinformatics*. 2013. 14:184 .

[10] Paul Flicek & Ewan Birney. Sense from sequence reads: methods for alignment and assembly, *Nature Methods* 6. 2009. S6 - S12.

[11] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing, *briefings in Bioinformatics*. 2010. Vol 11. No 5.

[12] El-Metwally et.al. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. *PLoS Comput Biol* 9 (12): 2013. e1003345.

[13] Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical Report 124, *Digital Equipment Corporation*. CA: Palo Alto. 1994.

[14] Langmead and Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012. Vol. 9 (4) pp. 357-9.

[15] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. *In Proc. 41st Annual Symposium on Foundations of Computer Science*. 2000. Pages 390–398. IEEE.

[16] Petr Klus et.al. BarraCUDA - a fast short read sequence aligner using graphics processing units. SOURCE. *BMC Research Notes*. 2012, Vol. 5.

[17] Bloom, J. Et al. Exact and complete short read alignment to microbial genomes using GPU programming. *Bioinformatics*. 2011. 27(10), 1351-1358.

[18] Salavert Torres J et.al. Using GPUs for the exact alignment of short-read genetic sequences by means of the Burrows-Wheeler transform, *IEEE/ACM Trans Comput Biol Bioinform*. 2012 Jul-Aug; 9 (4): 1245-56. doi: 10.1109.

[19] Schatz M.C. BlastReduce: High Performance Short Read Mapping with MapReduce, *Bioinformatics* 2012.

[20] Hayan Lee, Michael C. Schatz., The reliability of short reads mappingillustrated by the genome mappability score. 2012. Vol. 28 no. 16, pages 2097–2105.

[21] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K and Wang J, SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*. 2009. Vol. 25, pp. 1966- 1967.

[22] Matthew Ruffalo et.al. Comparative analysis of algorithms for next-generation sequencing read alignment. 2011.Vol. 27 no. 20, pages 2790–2796 doi: 10.1093/*Bioinformatics*/btr477.

[23] Rasmussen KR et.al. Efficient qgram filters for finding all e-matches over a given length, *Lecture Notes in Computer Science, Springer*. 2005. Vol. 3500, pp. 189-203.

[24] Butler, J et.al., De novo assembly of whole-genome shotgun microreads. *Genome Res*18: 810–820.

[25] Homer N et.al., BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 2009, 4 (11)

[26] Li H, Durbin R, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*.2009. 25(14):17541760

[27] Schatz M et.al. High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics*. 2007. 8(1):474.

[28] Shaffer C. Next-generation sequencingoutpaces expectations. *Nat Biotechnol*, 2007.25 (2): p. 149.

[29] Wheeler, DA. Et al. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008.452 (7189): p. 872-6.

[30] Venture, JC. Et al., *The sequence of the human genome. Science*, 2001. 291 (5507): p. 1304-51.

[31] Hadoop. [Cited; Available from: http://hadoop.apache.org/.]

[32] Hayan Lee, Michael C. Schatz.,The reliability of short read mappingillustrated by the genome mappability score. 2012. Vol. 28 no. 16 pages 2097–2105.

[33] Jeffrey D and Sanjay G. MapReduce: simplified data processing on large clusters. Commun. *ACM*, 2008. 51 (1): p. 107-113.

[34] Matthew Ruffalo et.al., Accurate estimation of short read mapping quality for next-generation genome sequencing. 2012. Do: 10.1093*/Bioinformatics*/ Vol. 28 ECCB.

[35] Lunter and Goodson. Stampy., A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. 2011. *Genome Res* vol. 21 (6) pp. 936-9

[36] Liu, Y. Et al., GPU distributed error correction on massively parallel graphics processing units using CUDA and MPI. BMC *Bioinformatics*.2011. 12, 95.

[37] Metzker ML, Sequencing technologies – *the next generation. Nat Rev Genet* 2010. 11:31– 46.

[38] Ning Z, Cox AJ, Mullikin JC. SSAHA, A fast search method for large DNA databases. *Genome Res* 2001. 11:1725–9.

[39] Kent WJ. BLAT–the BLAST-like alignment tool. 2002. *Genome Res*.12: 656–64.

[40] Smith AD, Xuan Z, Zhang MQ., Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC *Bioinformatics* 2008; 9:128.

[41] Abouelhoda et.al., Replacing suffix trees with enhanced suffix arrays. *JDiscreteAlgorithms* 2004. 2: 53–86.

[42] Munro JI, Raman V, Rao SS. Space efficient suffix trees. *Algorithms* 2001.

[43] Hoffmann et al., Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoSComput Biol* 2009.

[44] Langmead B, Hansen KD, Leek JT, Cloud-scale RNA-sequencing differential expression analysis with Myrna, Genome *Biol* 2010, 11:R83.

[45] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL, Searching for SNPs with cloud computing, *Genome Biol*.2009. 10:R134.

[46] NVIDIACUDAZone. http://www.nvidia.com/object/cuda home new.html].

[47] CULA tools, EM Photonics. [http://www.culatools.com].

[48] Manavski SA, Valle G: CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinformatics*. 2008.

[49] Langmead, B. Et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. 2009. *Genome Biol*., 10 (3), R25.

[50] Lam, T.W. et al. Compressed indexing and local alignment of DNA. *Bioinformatics*. 2008. 4(6), 791-797.

[51] GenBank, ftp://ftp.ncbi.nih.gov/genbank/.

[52] EMBL, http://www.embl.org/ DDBJ,

[53] http://www.ddbj.nig.ac.jp/

[54] http://en.wikipedia.org

[55] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

BIOGRAPHIES

Dr. U. Srinivasulu Reddy received his Ph.D. degree from National Institute of Technology, Trichy, India in 2013. Since 2008, he has been working as Assistant Professor in Department of Computer Applications, National Institute of Technology, Trichy. His main research interests DNA computing and Big Data Analytics.

Mr. G.Raja received his M.Tech degree in Information Technology from School of Computer Science and Engineering, Bharathidasan University, Trichy, India in 2012. He is Research Scholar in Department of Computer Applications, National Institute of Technology, and Trichy. His current research interest includes Bioinformatics and Cloud Computing.