strand

New Generation Healthcare

Strand was founded in 2000 by computer science and mathematics professors from India's prestigious Indian Institute of Science who recognized the need to automate and integrate life science data analysis through an algorithmic and computational approach. Strand's segue into the life sciences was through informatics products and services for research biologists, chemists, and toxicologists that combine advanced visualization, predictive systems modeling, data integration and scientific content management - over 2000 research laboratories worldwide (about 30% of global market share) are licensees of Strand's technology products, including leading pharmaceutical and biotechnology companies, research hospitals and academic institutions. With a recent investment by Biomark Capital, Strand has grown its established team to over 200 employees, many with multidisciplinary backgrounds that transcend computation and biology.

Since 2012, Strand has been expanding its focus to include clinical genomics, spanning sequencing, data interpretation, reporting and counseling. Strand operates a 10,000 square foot laboratory space with state-of-the-art clinical genomics capabilities and is also establishing Strand Centers for Genomics and Personalized Medicine in several hospitals around the world to serve as outreach points for genomic counseling. Based on the experience gained from sequencing, analyzing, interpreting and reporting on clinical samples over a wide variety of clinical indications, Strand has developed an end-to-end solution for clinical labs that handles all stagesfrom analysis to reporting. The interpretation and reporting software platform has been designed and developed specifically for the medical professional, ranging from the molecular pathologist to the physician. By enhancing sequence-based diagnostics and clinical genomic data interpretation using a strong foundation of computational, scientific, and medical expertise, Strand is bringing individualized medicine to the world.

For more information about Strand, please visit **www.strandls.com**,
or follow us on twitter **@StrandLife**.

INDIA
5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024

USA
548 Market Street, Suite 82804, San Francisco, CA 94104

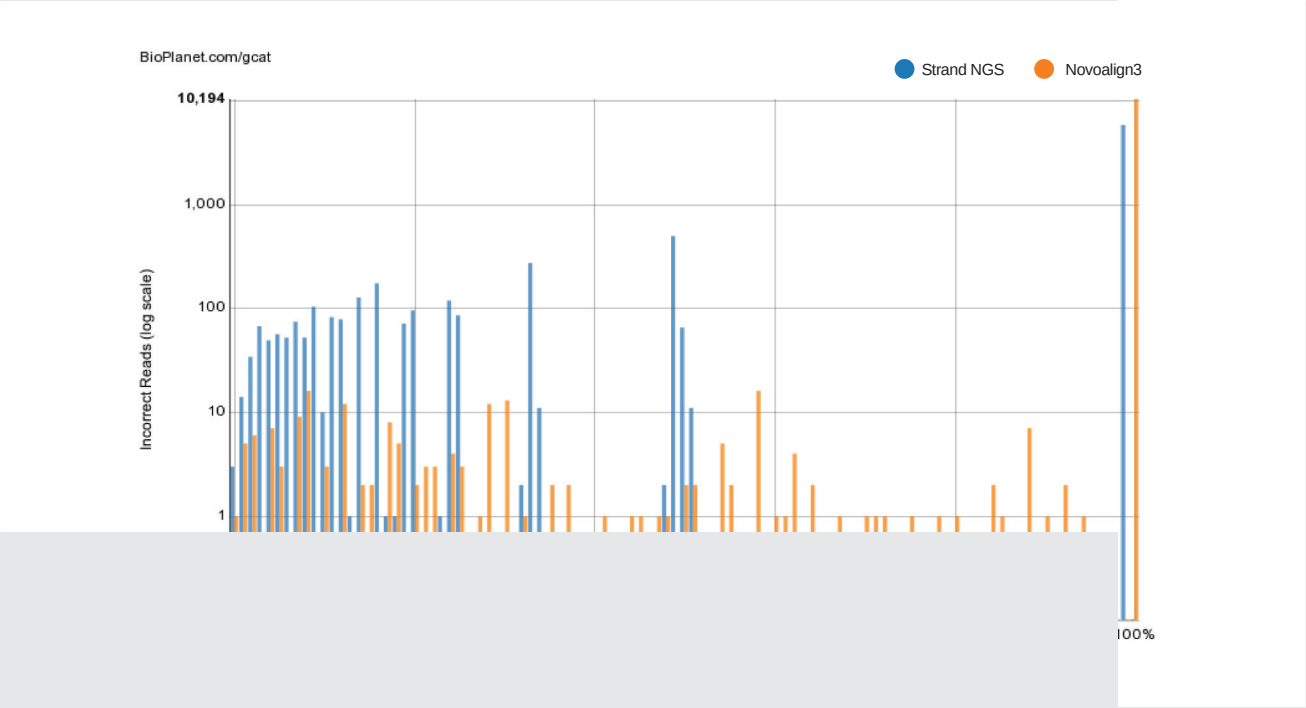

Figure 3(b):
Distribution of mapping qualities assigned to incorrectly mapped reads with InDels (results from only Strand NGS and Novoalign3 are shown).

4.4. Alignment Accuracy for Reads with Low Mapping Quality

The reads that originate from complex genomic locations, like repeat regions for example, are difficult to align. Typically these reads are classified as ambiguous by most algorithms as multiple locations in the genome are found to be equally likely for alignment of these reads. Depending on the algorithm or the choice of parameter settings, either all the genomic locations are reported or one random location is chosen and reported for these ambiguous reads. In either case, most of the algorithms reduce the mapping quality of these reads to reflect the ambiguity in their alignment. In this section, to assess the performance of algorithms in terms of accuracy on these ambiguous reads, we'll only compare Strand NGS and Novoalign3 because of the following reasons:

A) In terms of % correctly and % incorrectly mapped reads (which is used as the primary metric of accuracy), Strand NGS and Novoalign3 have similar accuracy, and both are found to be better than BWA and Bowtie2. Therefore, we believe that this additional analysis is valuable for comparing Strand NGS and Novoalign3 only.

B) Each algorithm has its own way of assigning mapping qualities. In addition, the mapping qualities assigned by different algorithms are in various ranges. This is the reason GCAT normalizes the mapping qualities between 0 and 100. Despite normalization, number of reads falling in different normalized mapping quality bins as per different algorithms, could be very different. Comparing algorithms on largely different number of reads may not be very useful. Strand NGS and Novoalign3 have similar number of reads that are assigned normalized mapping quality between 0 and 20, hence making the comparison more meaningful. On the other hand, Bowtie2 particularly has much higher number of reads with normalized mapping quality between 0 and 20. In this analysis, we have also included another algorithm, BWA-Mem which is a higher version of BWA and have assigned a normalized mapping quality between 0 – 20 to a comparable number of reads as Strand NGS and Novoalign.

A) **Strand NGS**: It is a BWT-based aligner, which in the first step, uses the Ferrazina and Manzini [10] algorithm to find the seeds of each read quickly in the reference. Then for every seed match found for the read, a Smith-Waterman type dynamic programming (DP) algorithm is applied to determine if the read matches around the area anchored by the seed. This match must satisfy the specified mismatches and gaps threshold.

B) **BWA**: It is a BWT-based aligner, which uses the Ferrazina and Manzini [10] matching algorithm to find seeds. Then a backtracking algorithm is used that searches for matches between a substring of the reference genome and the query within a certain defined distance.

C) **BWA-Mem**: It is conceptually based on BWA but reported to handle long reads and considered to be more accurate and faster.

D) **Bowtie2**: It is a BWT-based aligner, but just like BWA uses the Ferrazina and Manzini [10] matching algorithm to find seeds. Bowtie2 is designed to handle long reads and supports additional features that are not supported by Bowtie.

E) **Novoalign3**: It is a hash table-based tool, which first builds the index on the reference using hash tables and then uses the Neddleman-Wunsch algorithm with affine gap penalties to find the optimal alignment.

Other than Novoalign3, which uses the hash table-based index, others used BWT-based index. Despite the conceptual similarities among the BWT index-based algorithms, there are still differences in their seeding strategies and optimizations. For example, Bowtie2 search for the seeds of length 16 with a gap of 10 then prioritize the seed locations and perform DP around them. BWA on the other hand, does the inexact matching in the BWT index itself. In Strand NGS, we keep extending the seed till the seed is present in the reference. Then we jump by one-third of the seed and start searching for the next seed. In the end, we take the 4 longest seeds and find the seed locations in the reference. After removing the duplicate seed locations, DP is performed at each seed location.

3. Data Sets Used

The GCAT (Genome Comparison and Analytic Testing) [12] platform which is developed by Arpeggi Inc. and hosted on the bioplanet website, is heavily used in this study. GCAT provides a solid testing platform for comparing the performance of multiple genomic analysis tools across a standard set of metrics. For alignment tests, the GCAT website hosts different simulated data sets covering read lengths of 100, 150, 250, and 400; single-end and paired-end library protocols; and both short and long InDels. Assuming a typical use-case scenario where Illumina paired-end data is used, we selected four data sets from GCAT in this study for benchmarking purposes. The details of these data sets are shown in table 1. The base quality for each base in all the four simulated data sets is 17, producing an average read quality of 17.

For computational performance benchmarking, we used the data from whole-genome sequencing of the sample, NA12878 on Illumina HiSeq 2500 using paired-end library. This data is ~103GB and comprises of 1,165,216,818 (~1.16 billion) paired-end reads of length 150bp [13].

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

Table 1: Description of GCAT data sets

There are some limitations of using simulated data sets for accuracy benchmarking. First, it is difficult to mimic the error patterns in the generated reads that matches the true error model of the sequencer; and second it is not guaranteed that the true genomic location of the generated read with simulated sequencing errors is the one where the read is generated from. The read with simulated sequencing errors and/or SNPs/InDels can very well align at a different location with better accuracy. However barring few instances like these, still the ease of defining the truth makes this approach viable and useful for benchmarking studies. Therefore despite some shortcomings, most of the high-level comparisons that can be derived about different algorithms using simulated data sets still hold, thereby providing a useful resource to the users to understand the pros and cons of multiple alignment algorithms.

In order to address the above challenges, numerous approaches (MAQ [1], BWA [3], BWA-Mem, BWA-SW [11], Bowtie [9], Bowtie2 [2], SOAP2 [4], Novoalign [8], Novoalign3 [8], mrFAST [6], mrsFAST [7], SHRIMP [5], etc.) have been developed in the past. There have also been some benchmarking studies discussing and comparing results from different subsets of these approaches.

1.2. Evaluation Approach

Since the alignment task can very quickly become a bottleneck in the analysis pipeline due to the ever-increasing volume of the sequencing data, most of the above approaches adopt a trade-off between accuracy and speed. The choice of number of mismatches or gaps allowed, neglecting quality of data and known SNP information during alignment are some of the ways to trade-off accuracy in favor of speed or efficiency. Due to this trade-off, it is important to assess the performance of different algorithms on both metrics, i.e., accuracy of read alignment, and computational efficiency. However, since accuracy of the alignment process directly impacts the results of many downstream applications, accuracy is a more important metric than efficiency as long as the algorithm runs in a reasonable amount of time given the computational resources at hand.

To measure performance in terms of accuracy, we use simulated data sets for benchmarking and use the following four evaluation criteria:

A) Fraction (or %) of correctly, incorrectly and unmapped reads when alignment was done for

- All reads
- Only reads with SNPs
- Only reads with InDels
- Only reads with both SNPs and InDels

B) Trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). In this case, TPR and FPR correspond to correctly mapped and incorrectly mapped reads respectively.

C) Mapping quality distribution of incorrectly mapped reads.

D) Fraction (or %) of correctly, incorrectly and unmapped reads for special case of reads that are assigned low mapping qualities (representing ambiguous reads possibly originating from repeat regions in the genome).

To measure performance in terms of computational efficiency, we use a real data set (a whole-genome sequencing run for a widely used sample NA12878) and the total run-time of the chosen algorithms is used as an evaluation criterion. The total time taken for alignment includes the following times:

A) Time taken for Burrows Wheeler Transform (BWT) search to find the initial exact seeds.

B) Time taken for Dynamic Programming (DP) around the seeds found.

C) Time taken for post-processing to produce final alignment results.

Also, it is important to note that the total time reported should not include the time taken in the following tasks:

A) Time taken to build the BWT index on the reference genome (since this is a one-time task).
B) Time taken to import the sample and collect/calculate QC metrics.

2. Algorithms Compared for Benchmarking

To make the alignment process more efficient, most alignment algorithms start by building an index on the reference genome. This index is then used to find the genomic locations for each read. There are some algorithms that build index on the reads but most of the popular recent algorithms build index on the reference genome. This is because the same index once built on a reference genome can be used repeatedly for aligning different read sets. There are two main techniques used for building the index on reference genome: hash tables and Burrows-Wheeler transform (BWT). In this study, we only consider the following algorithms that build an index (either BWT- or hash table-based) on the reference genome:

Out of the total 7,963,500 reads in data set D4, Strand NGS, Novoalign3 and BWA-Mem assigns a normalized mapping quality of 0 - 20 to 164,071, 158,498 and 158,083 reads, respectively. Table 4 shows the alignment accuracy of the three algorithms on these reads.

	Strand NGS (164,071 reads)	**Novoalign3 (158,498 reads)**	**BWA-Mem (158,083 reads)**
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%)

Table 4: Alignment Accuracy of Strand NGS, Novoalign3 and BWA-Mem on reads with low mapping quality

It is clear from table 4 that Strand NGS has a higher percentage of correctly mapped reads (high TPR) compared to both Novoalign3 and BWA-Mem. On the other hand, percentage of incorrectly mapped reads (FPR rate) of Strand NGS is lower than BWA-Mem, thereby making it better than BWA-Mem on both metrics, but higher than Novoalign3 making the comparison between Strand NGS and Novoalign3 non-trivial. However, if we refer to the previous section and particularly figure 3(a) and figure 3(b) where distributions of mapping qualities assigned by these two algorithms to incorrectly mapped reads are examined, we also know that Strand NGS assigned relatively low mapping quality to incorrectly mapped reads. Therefore it is easy to filter out many incorrectly mapped reads by Strand NGS before the variant calling step, which is one of the most common downstream analysis steps. For example, a mapping quality threshold of 50% would filter out a large fraction of the incorrectly mapped reads, thereby reducing FPR to a large extent without sacrificing high TPR obtained using Strand NGS. Further, it is also important to note that incorrectly mapped reads by BWA-Mem cannot be filtered out using a threshold on normalized mapping quality because similar to Novoalign3, normalized mapping qualities assigned by BWA-Mem are more or less spread uniformly in the range 0 – 100 (data not shown here but accessible on the GCAT report, link to which is provided in table 2).

5. Computational Performance Benchmarking Results

For computational performance benchmarking, we performed the alignment of real whole-genome sample NA12878 (details in section 3), using the algorithms Strand NGS, BWA-Mem, and Bowtie2; on a machine with 64GB RAM and 15 cores. BWA-Mem is used instead of BWA because BWA-Mem is reported to be faster than BWA. Further, we did not include Novoalign3 in computational performance benchmarking because it is a proprietary algorithm and the academic version doesn't support multi-threading, making it harder to compare the run-time of Novoalign3 with the rest of the algorithms.

As per the evaluation criteria defined in section 1.2, table 5 shows the total time taken by different algorithms in aligning ~1.16 billion paired-end reads generated from the sample NA12878 against hg19 human genome reference.

Data Set: Whole-genome sequencing run with ~1.16 billion reads from sample NA12878	
Algorithm	**Total Time Taken**
Strand NGS	9 hours, 32 minutes
BWA-Mem	12 hours, 09 minutes
Bowtie2	11 hours, 0 minutes

Table 5: Computational performance of algorithms on whole-genome sequence data for sample NA12878

It is clear from table 5 that Strand NGS, in addition to offering higher alignment accuracy as demonstrated in the previous section, is computationally fast compared to other standard alignment algorithms.

Abstract

Background

Next generation sequencing technology has led to the generation of millions of short reads at an affordable cost. Aligning these short reads to a reference genome is a crucial task for many downstream applications. However due to the large size of such data, the process of alignment is computationally challenging and requires sophisticated algorithms which are both fast and accurate. In this work we will briefly discuss the Strand NGS (formerly Avadis NGS) alignment algorithm, but more importantly present the benchmarking results on several simulated data sets and a real whole-genome data to compare it with other standard state-of-the-art algorithms.

Results

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.

Conclusions

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.

1. Introduction

There has been an unprecedented growth in sequencing data due to the rapid advancements in the next-generation sequencing (NGS) technology. For a whole-genome sample, the number of reads can vary from few million long reads (≥400bp) generated by instruments like PacBio and 454, to 2-3 billion short reads (≥75bp) generated by instruments like Illumina/Solexa and SOLID. Further, there has been a significant cost reduction in DNA sequencing, attracting more and more researchers to use NGS technology in their own labs. However, unless we have proper bioinformatics tools to process and analyze this large amount of sequencing data, data by itself is not of much use.

1.1. Problem Statement and Challenges

One of the first steps in the analysis of NGS data, other than of course checking the quality and filtering out bad quality reads, is alignment or mapping of the generated sequencing reads to the respective reference genome. An accurate and efficient alignment of reads to a reference genome is crucial for many downstream applications, for example variant calling, structural variant detection including copy number changes, detecting protein-DNA binding sites using ChIP Sequencing, comparing expression of genes/transcripts across different biological conditions, understanding methylation patterns in DNA, determining species composition using metagenomics workflow, etc. However, alignment is a challenging problem due to the following reasons:

A) A reference genome is typically long (~billions) and has complex regions like repetitive elements, etc.

B) Reads are short in length (typically, 50 - 150bp), presenting issues with accuracy, and large in number, presenting issues with efficiency.

C) Reads have sequencing errors and must be mapped to unique positions in the reference genome.

D) The subject genome (for example, tumor sample) may inherently be different from the reference genome because of acquired alterations over time, making the alignment difficult.

6. Conclusions

In this work, we presented a benchmarking study to compare the performance, in terms of accuracy and computational run-time, of multiple alignment algorithms: Strand NGS, BWA, BWA-Mem, Bowtie2, and Novoalign3. In terms of accuracy measured as the fraction of correctly and incorrectly mapped reads, Strand NGS and Novoalign3 seem to be comparable to each other but superior to other algorithms considered in this study. Further for both Strand NGS and Novoalign3, FPR rate increases mildly with the increase in TPR rate, as opposed to steeper increase in case of BWA and Bowtie2. However when we focused closely on the assigned mapping qualities by Strand NGS and Novoalign3 to incorrectly mapped reads, we noticed that Strand NGS assigns lower mapping quality to them compared to Novoalign3, thereby providing a way to remove them before the downstream analysis.

In addition, we also focused on a subset of the reads, which are assigned very low mapping quality by both Strand NGS and Novoalign3. We found that Strand NGS has a higher true-positive rate, albeit at the cost of higher false-positive rate. However, we argue that since Strand NGS assigns low mapping quality to incorrectly mapped reads, false positive rate can be significantly reduced by choosing a lower cut-off on mapping quality without affecting the true-positive rate.

Overall, based on the results presented in this study, we strongly believe that Strand NGS offers a solid approach for DNA read alignment and either performs comparably well or outperforms most of the state-of-the-art algorithms, including Novoalign3 in one or more aspects. Further, since Strand NGS offers a GUI-based solution with a visually appealing and convenient way to change parameters, we believe it is one of the easiest tools to use by researchers with limited or even no knowledge of algorithms or bioinformatics.

7. References

1) Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Gen Res* 2008, 18(11):1851-1858. [MAQ]
2) Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Meth* 2012, 9:357-359. [Bowtie2]
3) Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinf* 2009, 25(14):1754-1760. [BWA]
4) Li R, Yu C, Li Y, *et al.*: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinf* 2009, 25(15):1966-1967. [SOAP2]
5) Rumble S, Lacroute P, Dalca A et al: **SHRiMP: accurate Mapping of Short Color-space Reads.** *PLoS Comput Biol* 2009, 5(5):e1000386 [SHRIMP]
6) Alkan C, Kidd J, Marques-Bonet T, *et al.*: **Personalized copy number and segmental duplication maps using next-generation sequencing.** *Nat Genet* 2009, 41(10):1061-1067. [mrFAST]
7) Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC: **mrsFast: a cache-oblivious algorithm for short-read mapping.** *Nat Method* 2010, 7(8):576-577. [mrsFAST]
8) **Novoalign3** [http://www.novocraft.com]
9) Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Gen Biol* 2009, 10(3):R25+ [Bowtie]
10) Ferragina P, Manzini G: **Opportunistic data structures with applications.** 41st Annual Symposium on Foundations of Computer Science, Washington, DC 2000, 390-398.
11) Li H, and Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler Transform.** *Bioinformatics* 2010. [BWA-SW]
12) **GCAT:** [http://www.bioplanet.com/gcat/]
13) Illumina Basespace data. [http://blog.basespace.illumina.com/2012/11/19/2x150bp-human-genome-in-record-time-with-the-hiseq-2500/]