American Research Journal of Computer Science and Information Technology

Volume 8, Issue 1, 1-5 Pages
Research Article | Open Access
ISSN (Online)- 2572-2921
DOI : 10.21694/2572-2921.25001

Evaluation of Software of Next-Generation Sequencing in Mapping

Farzon Nosiri

ABSTRACT

This study evaluates the performance of five next-generation sequencing (NGS) read mapping tools—BWA, Bowtie, Stampy, Mosaik, and Agile—focusing on their behavior under varying conditions, including read errors, insertions, deletions, and mutation rates. Reads from Illumina sequencing systems were simulated using dwgsim and further manipulated for controlled testing. The evaluation assessed mapping precision, sensitivity, and computational efficiency using a 2.7 GB reference genome to ensure compatibility across tools. Stampy demonstrated superior overall performance, particularly with complex variations, but exhibited limitations in handling perfect reads with base quality "!" scores. Bowtie and Mosaik performed well under specific conditions but struggled with paired reads and precision, respectively. Agile proved incompatible with short reads, while BWA emerged as the fastest tool, albeit with precision issues at higher mutation rates. Deletions posed the most significant challenge across all tools. These findings highlight critical trade-offs between speed, accuracy, and adaptability in NGS read mapping tools, providing insights for their application in population genomics and other domains.

INTRODUCTION

The objective is to study and evaluate programs that map reads from Illumina machines to a reference genome. We want to know how these programs behave with respect to speed, misplacement, and sensitivity under several different conditions. A special case to study is the behavior on very large reference genomes. Five software will be tested in this work. Details regarding these software is written on the Table 1.

Table 1. Software name, version and description are brought on this table.

Software	Version	Description
BWA [1]	0.5.9	BWA is a relatively fast light-weighted and also well-spread tool that aligns relatively short sequences to a reference sequences, such as the human reference genome. BWA implements two different algorithms, which are based on Burrows-Wheeler Transform.
bowtie [2]	0.12.7	Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small
Stampy [3]		Stampy is recently published sequence mapping software claiming to be a competitor to the BWA.
Mosaik [4]	1.1.0014	Mosaik is a large tool for reference guided assembling. One of its tool is aligning the reads, which will be mainly used in this evalutaion.
Agile [5]	0.4.0	AGILE is a new tool, which is a sequence mapping software specifically designed to map the longer reads to a reference genome.

DATA

For pragmatic reasons, the data to study are reads from Illumina systems. These reads are in the range of 50 -- 100 bases and both can be "single" and "paired". Paired reads are (in this project) assumed to have a distance from each other that is normally distributed with mean 300 and standard deviation 50.

We want to know the true source of reads, which means that the reads should be created by us from a reference genome. One way forward is to take a common and well-studied genome such as human or mouse, but to do something different from everyone else; some other data is used. Since we have a special interest in the spruce genome and which is also "stressing" software with its large size (21 GB), we should try using an early assembly of this genome as reference. For e.g. BWA, which cannot handle larger genomes than ca 4 GB, we can take the largest contigs that fit within 4 GB.

Conditions

The main use-case for read mappers is in population genomics when we want to identify sequence variation. In addition, read data is never perfect in itself, so there are small differences due to read errors as well. Studies show that different read mappers behave differently depending on how much variation there is and what type of variation (substitutions, insertions, deletions) there is. The task is to try different types and amount of variations in order to find the software limitations.

The suggestion is to the following test cases, with and without paired reads:

- Perfect reads
- Increasing amount of substitutions: 4, 8, 12, 16, 20 %
- Increasing length of one insert: from 1 to 20 bases
- Increasing length of one deletion: from 1 to 20 bases
- Varying the placement of the insert/deletion, from early to late in the reads

METHODS

To measure the precision or the number of positive predicted value (PPV) [6] the following formula will be used:

$$PPV = \frac{\text{Number of correctly mapped reads}}{\text{Number of correctly mapped reads} + \text{Number of incorrectly mapped reads}} \quad [7]$$

The incorrectly mapped read here is the read which have been mapped to a wrong position in the reference genome.

To generate the reads the tool dwgsim is used. This tool can perform a whole genome simulation. dwgsim is based off of wgsim found in SAMtools written by Heng Li [8].

To evaluate and check mapping quality the tool called dwgsim_eval is used. This utility evaluates the mappings from reads produced by dwgsim. It takes “SAM” or “BAM” file formats as input.

When varying the placement of indels, the 3 different placements have been chosen to test. The first placement was in the beginning of the reads starting from the 7th base pair. Next the second placement is done in the middle of the read starting from 45th basepair. The third placement is done at the end of the read starting at position 88th bp for short indels up to 8 bp, and 72nd bp for the longer indels longer than 8 bp up to 20bp.

RESULTS

One of the major difficulties was to define the size of the reference database. As it has been experienced in practice, Stampy cannot deal with the reference genome larger than 3Gb. In present work stampy was able to handle up to 2.7Gb, and since all the other software can easily deal with this size and larger, and to make sure conduct a fair test for all 5 tools, the size of reference dataset of 2.7 gb is selected. Reads also have been generated from it. Using **dwgsim** reads with different conditions and combinations according to the preset condition is generated. But to test under all conditions in this evaluation the test like varying placement of the insertions or deletions, an additional program have been written. This code allows controlling the position and length of insertions or deletions in the reads.

Unfortunately Agile wasn’t able to handle short reads. It was crushing down every single time dealing with our generated set of reads. None of the tweaks and adjustments for input parameters and input data helped to succeed. As it is written in description in the **software manual,** it is designed to map large reads with the length > 200 bp. This suggests that **Agile** does not suite for our evaluation conditions.

For the first step, we go through mapping perfect reads generated from refgenome without any errors or mutations/indels.

Table 2. The current table shows the results of mapping perfect reads to the reference genome. Here **mc** stands for the reads “mapped correctly”, **mi** – reads that “mapped incorrectly”, and the PPV here is calculated as ***mc/ (mc+mi)***.

soft	map type	mc	mi	mc+mi	total	PPV
bwa	single	5000	0	5000	5000	1
bowtie	single	5000	0	5000	5000	1
mosaik	single	3427	0	3427	5000	1
stampy	single	5000	0	5000	5000	1
bwa	paired	10000	0	10000	10000	1
bowtie	paired	0	2	2	10000	0
mosaik	paired	6883	0	6883	10000	1
stampy	paired	10000	0	10000	10000	1

All software mapped correctly all reads, and as one can see on the Table 2 Mosaik is less sensitive in regard of mapping, and bowtie did not align paired reads. So almost already we can grasp some picture about the performance. The next we evaluate the performance of software under different circumstances and conditions like substitutions or mutations with different rates. The evaluation under different mutation rates for single reads showed that **Mosaik** performs very accurately, though as we have seen from our first trial Mosaik is relatively less sensitive than other tools, but it shows very reliable results. This can be seen from the PPV value. Up to 16% mutation rate Mosaik showed 100% precision.

Comparing between Bowtie and BWA, bowtie shows better performance than BWA. It has aligned more reads than BWA and also the PPV value is much higher for Bowtie, because BWA has very extremely low precision the highest is 0.008 for 1% mutation. The suspicion was maybe there a little shift in base pair position numbering but after manual check it has been confirmed that BWA has mapped reads far from correct positions. Among all of them, the masterpiece results belong to **Stampy.** A slight deviation in PPV value and in alignment number can be observed after 16% mutation rate. Even so Stampy have showed much better performance for all mutation rates. For instance if we take the 8% mutation rate and look to the number of aligned reads we can see that BWA have mapped 2175 reads, Bowtie 3742, Mosaik 767 and Stampy 5000 reads. During the first test, one disturbing fact has been discovered for **Stampy.** When generating reads with *base* score of *fastq* set to "**!**", meaning perfect reads, Stampy haven't mapped any reads. However when this quality base score was changed to "**2**" Stampy has shown good results. The sequence of the reads remained the same only quality score differed.

The results for paired reads alignment with mutations is almost the same as for a single ones. Only few facts can be mentioned here that Bowtie overall haven't mapped any reads, except 4 reads for 1% mutation, but all four weren't mapped to the correct position.

The next two steps in testing performance of handling Insertions and Deletions have been combined. Since the dwgsim does not allow to control them separately. The length of them also controlled by introducing the probability of extension of indels, hence the precise amount of indels in the reads cannot be controlled manually. But in order to compensate these drawbacks of **dwgsim** the additional program has been written, which takes perfect reads generated by dwgsim and adds insertions or deletions according to the parameters given as an input. But this test will be described later. Now the results for combined insertions and deletions demonstrated. Reads generated with the indels is controlled by two parameters: mutation rate and indels extension probability. Mutation rate gives probability of occurrence of indels in the reads, and extension probability allows controlling the length of indels.

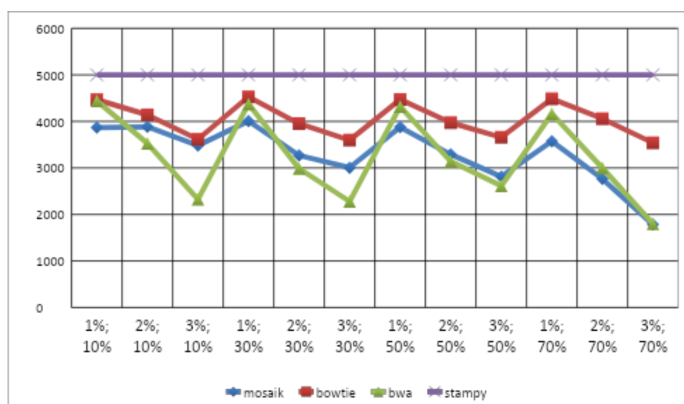

Figure 1. Number of aligned single reads on y axis for different mutation rate and extension probability on x axis.

The Figure 1 shows the number of aligned reads for different combination of mutation rate and indels extension probability. One can see that BWA is relatively unstable it goes rapidly down for each mutation rate change. The stability of Mosaik started to dwindle on higher extension probability of indels. Stampy shows extremely good results. Bowtie shows the same pattern for each extension probability change.

Up until now, the test like mapping a mutational reads and indels have been shown, but to test and evaluate these software under other conditions like adding deletion or insertion on different positions along the reads haven't been tested. For this purpose we use combination of dwgsim and additional written program. First we generate perfect reads by dwgsim and pass it to additional tool for further manipulation with them. One may find more specific detail how it is been done in the *methods* section.

The result for the single reads with insertion in the beginning, showed that BWA drops down from 100% alignment to almost 0 when insertion size reaches the size of 3. Bowtie outcomes with very low number of alignments starting from 1000 reads and decreases up to almost 0 when the insertion size reaches 5. Mosaik rapidly drops from 3000 aligned reads to the almost 0 when the size of insertion changes from 4 bp to 5 bp. Stampy aligns 100% reads for any insertion size. Similar results can be observed for paired reads. In respect of precision Bowtie showed the worst results comparing to other tools. Bowtie has PPV of 0.3 when the length of insertion is 1 bp, and this value goes down when the insertion length is increased, the values reaches 0.01 at insertion size of 5 bp. The PPV for BWA goes rapidly down from 1 to 0 at insertion size of 5 bp. As for Mosaik the picture is very similar to BWA the precision goes down to the 0 on 5th base pair mark which is expected since there are no aligned reads for Mosaik with insertion size larger than 5 bp. Stampy on other hand have showed again better results than the others. It has mapped reads with almost 100% precision when reads have up to 5 bp insertions but larger than 5bp long insertion Stampy's precision starts to fade. Increasing the length the precision goes down. The results show 14% precision when the insertion size of 20 bp is set. Here one should note that Stampy aligns almost 100% of the reads but the precision is only 14%. 86% of the reads just mapped to the wrong position by Stampy.

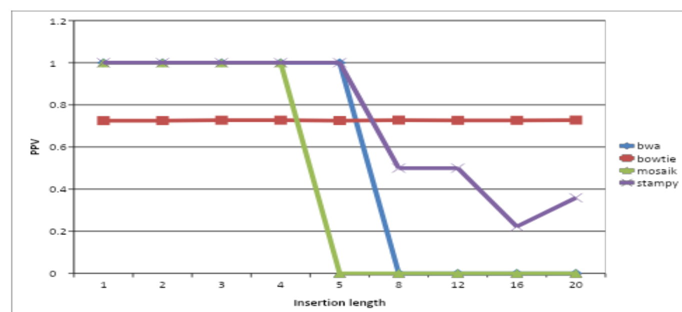

Figure 2. The figure is for single reads having insertions in the middle. It projects precision of mapped reads by the tools. Positive predicted value(PPV) is on the y-axis, and the length of the insertion in the reads shown on the x-axis.

In the contrast to the reads having insertions in the beginning, the reads with insertions in the middle have been aligned much better by all software. Here bowtie showed more aligned reads than he usually did before. BWA mapped all reads up to 5bp of insertion. Mosaik mapped about 3000 reads per each length change, but only mapping reads with up to 4 bp of insertions. Stampy and Bowtie were able to map reads with any size of insertions starting from 1bp up to 20bp. The Figure 2 shows the precision of the mapped reads by all 4 software. One can see bowtie has around 70% of precision for any length of insertion. Mosaik and BWA mapped 100% correctly all aligned reads. Stampy on other hand if it has almost 100% alignment for the reads of any insertion size but here one can see after 5bp long insertions the precision drops down rapidly, which means there are many false positive predictions for the reads with the insertion larger than 5bp.

Number of aligned paired reads is practically the same, except bowtie has 0 alignments. There is interesting fact can be worth mentioning about Stampy. On paired reads the precision of alignment of the reads with large insertion in the middle were slightly better than on single ones.

Aligning single reads with the insertion at the end of the reads have been much better comparing to the results of insertion at the beginning of the reads. BWA aligned reads with up to 5 bp long insertions. The pattern of Mosaik didn't change mapping up to 4 bp long insertions. Bowtie and Stampy mapped all tested reads. The precision for BWA and Mosaik again is 100%. The PPV for Bowtie resulted around 0.73 for all size of insertions. The PPV for Stampy dropped to 50% mark dealing with 8bp and larger insertion size. Similar results for paired reads. Bowtie again didn't map any read. Mosaik have mapped reads up to 4 bp long insertions, and one pair with 12 bp long insertion and mapped all of them correctly.

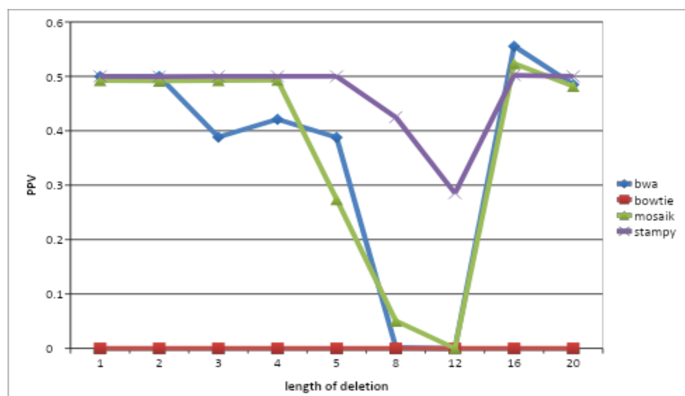

Figure 3. The figure projects the precision of aligned paired reads. The reads have a deletion in the beginning of the reads. Bowtie haven't mapped paired reads thus it has 0.

The evaluation of indels is divided to insertion and deletion, meaning the test is carried out separately for insertions and deletions. The results of insertion are already been described. Next step is to demonstrate the performance of the software for deletions. The approach is the same as for insertion but in this case instead of inserting the nucleotides, they are simply removed. In order to keep the length of 100 bp, reads with 120bp are generated by dwgsim then processed with addition written software to delete nucleotides in the beginning, middle or ending positions.

The Figure 3 shows precision of paired reads which have deletion in the beginning. In this situation the performance of all 4 software is not as good as it was for other conditions. Stampy aligned all reads for any deletion length, but more than a 50% of aligned are reads false positives. And there a steep concave from 5bp to 16bp mark for all software. Also the results showed the increase in number of aligned reads after 5bp long deletions, for instance Mosaik from 478 reads Mosaik jumped to 2118 reads and then aligned around 3000 reads for 12, 16 and 20 bp long deletions. BWA have much better results when the deletion is introduced in the middle than at the beginning. The picture is similar for single reads. Bowtie didn't map any reads for paired reads but has relatively low number of single aligned reads. 1681 reads for the 1bp deletion and 954 for 20bp deletion. In case when the deletion was introduced in the middle, the number of aligned reads has increased for all software. One should note that the precision is slightly better for paired reads alignment than single read alignment, although the number of reads is somewhat same, except bowtie have no alignment for paired reads.

On the Table 3 the running time of all 4 tools is shown. The fastest is BWA

Table 3. Software running times.

Type of reads	# of reads	Stampy	Mosaik	BWA	Bowtie
Manually generated	885 000	4:39:17	5:09:04	0:33:48	7:21:04
Dwgsim generated	420 000	0:59:44		0:07:07	

CONCLUSION

The evaluation showed that Stampy performed relatively better than other mapping tools. It has shown better performance almost in all kind of conditions. Bowtie showed also good results even in some cases better than other tools, but overall it lacked precision, thus resulting relatively larger number of false positives. Also Bowtie didn't map paired reads. BWA showed more accurate and precise results comparing to bowtie and BWA proved the name being fastest tool. Mosaik on other hand have shown fewer numbers of alignments but it has shown much better precision than the other tools. Additional tests were conducted for Stampy, to make sure that Stampy does not use title names of the reads, which contain full information about position and deviations and other parameters and information. For this purpose the headlines of the reads were hidden, so that it does not contain any clue or hint about the read positioning. The test

have shown and confirmed once more a good performance by Stampy.

In many cases tools have shown slightly better performance aligning paired reads, than single ones. Among all conditions, dealing with "deletion" was the hardest task for all 4 software.

REFERENCES

1. H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler Transform," *Bioinformatics,* vol. 25, pp. 1754-60, 2010.
2. B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology,* vol. 10, no. 3, 2009.
3. G. Lunter and M. Goodson, "Stampy: A Statistical algorithm for sensitive and fast mapping of illumina sequence reads," *Genonme research,* vol. 21, pp. 936-939, 2011.
4. "The MarthLab : Mosaik," 14 December 2010. [Online]. Available: http://bioinformatics.bc.edu/marthlab/Mosaik. [Accessed 13 April 2012].
5. S. Misra, A. Agrawal, W.-k. Liao and A. Choudhary, "Anatomy of a Hash-based Long Read Sequence Mapping Algorithm for Next Generation DNA Sequencing.," *Bioinformatics ,* vol. 10, p. 1093, 2010.
6. "Positive predictive value," 12 April 2012 . [Online]. Available: http://en.wikipedia.org/wiki/Positive_predictive_value. [Accessed 13 April 2012].
7. D. G. Altman and M. J. Bland, "Statistics Notes: Diagnostic tests 2: predictive values," *BMJ ,* vol. 309, no. 1, p. 102, 1994.
8. "Whole Genome Simulation," 8 April 2012. [Online]. Available: http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation. [Accessed 8 April 2012].

Citation: Farzon Nosiri, "Evaluation of Software of Next-Generation Sequencing in Mapping", American Research Journal of Computer Science and Information Technology, Vol 8, no. 1, 2025, pp. 1-5.

Copyright © 2025 Farzon Nosiri, This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.