TRANSLATIONAL RESEARCH
IN VETERINARY SCIENCE
Vol 3, No 1, 2020

DOI: http://dx.doi.org/10.12775/TRVS.2020.005

ISSN 2657-4489 (online)

Evaluation of analytical protocols of alignment mapping tools using high throughput next-generation genome sequencing data

Artur Bąk[1], Dawid Bodziony[1], Grzegorz Migdałek[2], Chandra S. Pareek[3], Kacper Żukowski[§,1,3]

[1]Department of Cattle Breeding and Genetics, National Research Institute of Animal Production, Balice, Poland; artur.bak@izoo.krakow.pl, dawid.bodziony@izoo.krakow.pl, kacper.zukowski@umk.pl, https://orcid.org/0000-0002-5690-3634

[2]Institute of Biology, Pedagogical University of Cracow, Poland; grzegorz.migdalek@up.krakow.pl, https://orcid.org/0000-0003-1458-2673

[3]Department of Fundamental and Preclinical Sciences, Faculty of Biology and Veterinary Sciences, Nicolaus Copernicus University in Toruń, Poland. pareekcs@umk.pl, https://orcid.org/0000-0002-0329-787X, kacper.zukowski@umk.pl, https://orcid.org/0000-0002-5690-3634

[§]**Corresponding author**: dr Kacper Żukowski, Adjunct, Department of Fundamental and Preclinical Sciences, Faculty of Biology and Veterinary Sciences, Nicolaus Copernicus University, ul. Gagarina 9, 87-100 Toruń, Poland. kacper.zukowski@umk.pl

Abstract

Background: Ever since the development of first next-generation genome sequencer (NGS) in 2005, there are rapid developments of high throughput next-generation genome

sequencing (HT-NGS) techniques and tools used in genetics and genomics has become much more comfortable and cheaper. The result is the generation of a massive amount of data sets, requiring detailed analysis, which becomes impossible without the use of appropriate bioinformatics tools. One of the crucial steps in the analysis of NGS data is to map readings to a reference sequence. Although the dominance of Illumina synthesis by sequencing (SBS) technology has been noticeable in recent years, the choice of the tools is hampered and the variety of input data and reference genomes. Moreover, the tools used are crucial for result files and further analysis.

Methods: The subject of this paper is the three most frequently used alignment mapping programs, which have functions to allow working with many platforms: BWA, Bowtie2 and SMALT. The task of the tested aligners is to match short sequences coming from NGS with reference sequences. The most popular: BWA and Bowtie2 use for this purpose the Burrows-Wheeler transformation and SMALT maps the sequences using hashing and dynamic programming. The presented paper aimed to compare the quality and efficiency of the alignment mapping programs under examination, due to three criteria: i) the quality of the compared sequences of different lengths and from different platforms; ii) coefficient of wrongly compared sequences; iii) the computational resources used.

Results: By comparing the results of the mapping analyses for all the programs used, the least popular SMALT is the best. Obtaining the highest percentage of mapped readings for each platform and maintaining the lowest computational memory usage, turns out to be the most optimal choice.

Conclusions: The results presented in this paper can be used to verify and rebuild data analysis pipelines from NGS based so far on other tools. We conclude that by using the tools under appropriate conditions, it is possible to improve the quality of the analyses, speed them up and reduce their cost.

Keywords: Next-generation sequencing, NGS, illumina, Aligners, Alignments, Mapping, Algorithm, Reads, Genome.

Introduction

The rapid development of competitive genome sequencing platforms such as illumina, Roche, and life technologies have made the high throughput next-generation genome sequencing (HT-NGS) much easier, faster and cheaper [**1**]. Ever-since the advent of these modern sequencing techniques and bioinformatics tools, identification of novel nucleic acid sequences has become a ubiquitous and essential approach across all areas of biological science. However, the biggest challenge is the interpret and analyze the NGS generated biological data using advanced bioinformatics tools. The

combined power of NGS data and bioinformatics is vital for identification of novel finding on the genome research. After the pre-processing of the NGS data, the very first step of bioinformatics analysis is the alignment of the NGS generated data on reference genome to perform the alignment mapping of the sequencing reads. Herein in this paper, we are comparing the quality and efficiency of the three alignment mapping programs *viz.,* Burrows-Wheeler Aligner (BWA) [**2–3**], Bowtie 2 [**4–5**] and SMALT (https://www.sanger.ac.uk/tool/smalt-0/).

Selection of mapping tools: Choosing the right mapping tool is not an easy task due to the large number of sequencing platforms and the variety of input data and reference genomes. In this paper, we compared and showed how different mapping algorithms cope with aligning reads from different sequencing platforms to the reference genome. The comparative analysis of three mapping tools (**Table 1**) were performed to examine the accuracy of individual programs and algorithms used by them depending on the input data provided by high throughput sequencing technologies.

Table 1. Important features of three selected alignment or mapping tools

Mapping tools	Algorithms used	Important features
BWA	Burrows–Wheeler transform and Smith-Waterman method with Prefix/Suffix Matching Algorithms	The BWA-backtrack is used for shorter reads, and BWA-SW and BWA-MEM are for longer reads. Generally it is used for mapping less divergent sequence. BWA-SW allows gaps.
Bowtie 2	Modified Ferragina and Manzini matching algorithm BWT--indexing with Prefix/Suffix Matching Algorithms, quality aware backtracking	It allows gapped alignment and compared with Bowtie 1, its sensitivity is high for reads >50bp
SMALT	Smith Waterman algorithm and short word hashing	It provide tuning parameters indexing word length and step size to improve sensitivity and accuracy.

The basic principle of NGS mapping is the parallel amplification and sequencing of a large number of matrix fragments in a relatively short time [**6**]. In general, the NGS technologies use different methods of DNA matrix amplification and different sequencing methods. The most popular NGS platforms are Illumina, 454, SOLiD and IonTorrent. Each of these platforms differs in reading length, accuracy, cost and sequencing errors. Regardless of the technology, the first step is to fragment the matrix and create a set of short matrix segments (libraries), then the ligation of adaptors, parallel clonal amplification, and parallel sequencing of fragments with simultaneous imaging [**7**]. This results in a large number of consequent short readings (single or paired) covering the entire analyzed sample.

Methodology

In this study, using three NGS platforms (Illumina, IonTorrent, 454 Roche), three mapping tools (BWA, Bowtie2 and SMALT) were compared using the following criteria **(Figure 1)**.

- Creating test readings from the Escherichia coli genome for Illumina, IonTorrent, 454 Roche platforms using DWGsim and GemSim.
- Building indexes based on the reference genome for each mapping program.
- Mapping of simulated sequences with a reference sequence using the tested programs.
- Sorting mapped files with SAMTools [**8**].
- Calculation of coverage depth and generation results using BAMStats.

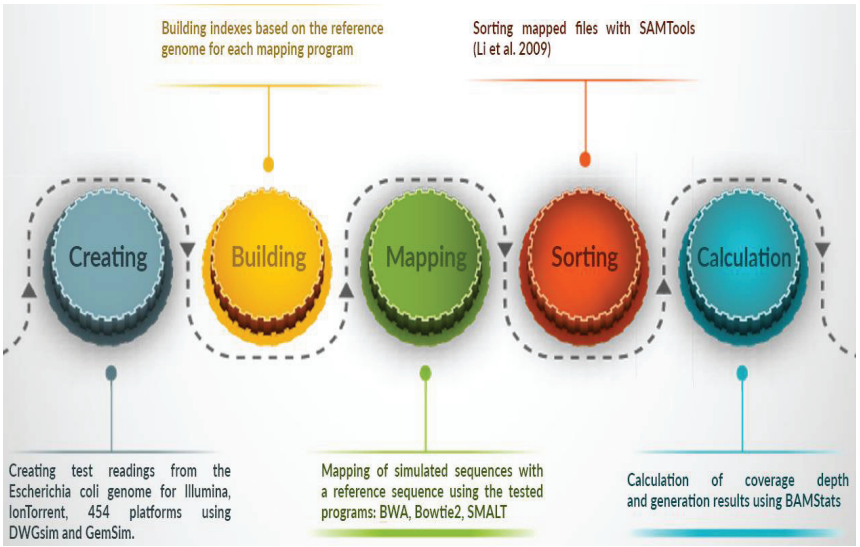

Figure 1. Mapping methodology adapted to compare the BWA, Bowtie2 and SMALT alignment tools using three NGS platforms

Results and discussions

By comparing the mapping analysis results of BWA, Bowtie2 and SMALT programs, SMALT tools was identified as the best, which obtained the highest percentage of mapped readings for each platform while maintaining the lowest computing memory usage. The another important aspect of the mapping analysis was the error rate. Based on this, comparative analysis of mapping results revealed that regardless of the mapping programs used, the error rate has always been the lowest for the readings produced by the 454 platform and the IonTorrent platform has always had the highest value of this parameter. Finally, study conclude with a noticeable trend with the fact that the percentage of mapped readings increased as the length of readings increased. The longer the reading length, the longer the string is, which increases its uniqueness. For example, a short reading such as (ACTGAC) will be much less unique than a long reading

such as (ACTGACAACGTGACCGGGTA) because the shorter the string the greater the probability that it will appear in the reference sequence more times making it difficult for the mapping program to adjust [**9**].

Conclusion

The results presented in this paper can be used to verify and rebuild data analysis pipelines from NGS based data. Study conclude that by using the mapping tools under appropriate conditions, it is possible to improve the quality of the analyses, speed them up to reduce their cost.

References

[**1**] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52:413–35.
[**2**] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
[**3**] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
[**4**] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
[**5**] Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 2019;35:421–432.
[**6**] Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. J Invest Dermatol. 2013;133:e11.
[**7**] Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem. 2011;83:4327–41.
[**8**] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
[**9**] Reinert K, Langmead B, Weese D, Evers DJ. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet. 2015;16:133–51.