Original Article

Selection of Optimal Bioinformatic Tools and Proper Reference for Reducing the Alignment Error in Targeted Sequencing Data

Abstract

Background: Careful design in the primary steps of a next-generation sequencing study is critical for obtaining successful results in downstream analysis. **Methods:** In this study, a framework is proposed to evaluate and improve the sequence mapping in targeted regions of the reference genome. In this regard, simulated short reads were produced from the coding regions of the human genome and mapped to a Customized Target-Based Reference (CTBR) by the alignment tools that have been introduced recently. The short reads produced by different sequencing technologies aligned to the standard genome and also CTBR with and without well-defined mutation types where the amount of unmapped and misaligned reads and runtime was measured for comparison. **Results:** The results showed that the mapping accuracy of the reads generated from Illumina Hiseq2500 using Stampy as the alignment tool whenever the CTBR was used as reference was significantly better than other evaluated pipelines. Using CTBR for alignment significantly decreased the mapping error in comparison to other expanded or more limited references. While intentional mutations were imported in the reads, Stampy showed the minimum error of 1.67% using CTBR. However, the lowest error obtained by stampy too using whole genome and one chromosome as references was 3.78% and 20%, respectively. Maximum and minimum misalignment errors were observed on chromosome Y and 20, respectively. **Conclusion:** Therefore using the proposed framework in a clinical targeted sequencing study may lead to predict the error and improve the performance of variant calling regarding the genomic regions targeted in a clinical study.

Keywords: *Chromosomes, high-throughput nucleotide sequencing, sequence analysis*

Hannane Mohammadi Nodehi[1], Mohammad Amin Tabatabaiefar[2,3], Mohammadreza Sehhati[1]

[1]Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, [2]Department of Medical Genetics, School of Medicine, Isfahan University of Medical Sciences, [3]Department of Bioinformatics, Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

Submitted: 20-Jan-2020 **Revised:** 28-Jan-2020 **Accepted:** 12-Feb-2020 **Published:** 30-Jan-2021

Introduction

Knowing about next-generation sequencing (NGS) and its evolution during the past few years has led genomics to be more accurate in the diagnosis of hereditary disorders. Analyzing NGS data provides us with massive information, hence choosing the best processing steps to preserve valuable information is an important pace in detecting and understanding the genetic variants.[1] However, complexity and intensity of early steps in NGS analysis, lack of a standard and global pipeline, technical errors introduced during sequencing and analysis, and variety of abounding incomplete tools are the main limiting factors for a successful analysis.[2,3] The first step to have a reliable interpretation on the analysis results is to produce high-quality data using appropriate technology. Fortunately, there are computational simulation tools that can produce synthetic NGS data by emulating base error rate due to sequencing faults in different technologies, which could be used for comparison and evaluation of various NGS analytical pipelines.[4]

Today, the use of targeted panel sequencing that tags only a small and specific portion of the genome is very popular in clinical diagnostics.[5-7] Clinical targeted sequencing studies differ in the location of known hotspots and such data require an optimization procedure with a different standard from that of whole-genome (WG) sequencing. Furthermore, selecting an optimal tool and tuning its parameters for such variations in data is a very crucial process for individualized medicine.[5]

Alignment or mapping, which is the most important step of NGS analysis, especially for targeted sequencing, is the

Address for correspondence:
Dr. Mohammadreza Sehhati,
Department of Bioinformatics,
Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine,
Isfahan University of Medical Sciences, Azadi Square,
Isfahan, Iran.
E-mail: mr.sehhati@gmail.com

Access this article online
Website: www.jmssjournal.net
DOI: 10.4103/jmss.JMSS_7_20
Quick Response Code:

How to cite this article: Nodehi HM, Tabatabaiefar MA, Sehhati M. Selection of optimal bioinformatic tools and proper reference for reducing the alignment error in targeted sequencing data. J Med Sign Sens 2021;11:37-44.

process of finding the possible location of short reads along a reference genome that leads to finding possible variants. Alignment tools use built-in algorithms to locate matching reads or tags of a sample to different regions of the reference genome. In practice, <75% of short reads efficiently map to the reference genome because a single read may appear more than once in the reference genome.[8] On the other hand, an observed short read may not exactly pair any position in the reference genome due to mutations or a bad sequencing readout. As the number of acceptable mismatches increases, the number of positions for mapping tags goes up but so does the number of incorrectly aligned reads. Giving erroneous output of the alignment step, many false-positive variants could be called and a lot of variants in unmapped or misaligned regions would be missed. There are mathematically optimal solutions for this problem, but it requires a lot of hardware resources and a long processing time to reach the answer, which makes them infeasible.[8,9]

This study aimed to optimize a part of the bioinformatic pipeline for analyzing the targeted sequencing data. First, we have compared some recent NGS technologies using an appropriate NGS data simulator. Second, we have used the human coding genome constructed from standard databases as the reference to mapping. Third, we have compared recent alignment tools using simulated data with different depth of coverage and their performance on different chromosomes was reported individually for comparison. Our results have demonstrated error-prone targeted regions, which govern the stabilization of new mapping quality measures or statistical significance estimates for improving the performance in the next step of analysis (e.g., variant calling).

Subjects and Methods

Simulated data

Most of the studies hold WG data to follow the workflow and report the best aligner by simulating short-read sequences.[2,3] Whereas only a small portion (<1%) of the human genome would be usually focused on clinical tests, processing of the WG is a waste of time. On the other hand, a small part of the genome known as coding sequences (CDS) or exome includes 85% of the known mutations that cause Mendelian disorders.[10] Therefore, whole-exome sequencing as a massively parallel examination that strengthened by the high resolution of NGS technology provided a cost-effective method for analyzing of the exons to accelerate disease gene finding. Recently, some collaborative projects, such as GENCODE[11] and CCDS,[12] have been introduced to provide standard databases for organizing all gene features in the human and mouse protein-coding regions. The final goal of these projects is to care convergence to a standard set of gene interpretations for the benefit of biological research. However, the challenge is to afford extensive designs that supply more coverage of targets to increase confidence in variant calling. The SureSelect Human All Exon (SSHAE)[13] is a good design that targets updated records relevant to different clinical research (~60 megabases). Specifying a customized target-based reference (CTBR) genome, SSHAE version 7 (Agilent Technologies, Santa Clara, California, United States) was chosen to generate *in silico* short reads. It contains regions between the start and stop codons and splice junctions in the genome of Homo sapiens in the form of a BED file, which can be used for generating corresponding FASTA sequences using in-house written Linux scripts (https://github.com/mrsehhati/SNV-importer).

To simulate synthetic paired-end short reads, artificial read transcription (ART) is chosen because unlike other read simulators, it emulates errors caused by sequencing platforms and technologies and generates a sequence alignment mapping (SAM) file in addition to FASTQ files simultaneously.[14] To evaluate the performance of pipelines in both the presence and absence of mutations, we have used two versions of CTBR in FASTA format as inputs of ART. The first input is the FASTA file that is directly generated from hg38 according to SSHAEv7 using in-house written script. The second input is the modified version of the first input that includes intentional mutations, which are recently reported in ClinVar (released on 2019.6.29) database that imported in the CTBR using an appropriate script. Investigating system-specific errors, different Illumina systems were used for emulating base error rates due to sequencing faults. It mimics various error rates and distributions to arrange bases in short reads of different lengths. Hiseq 1000, 2000, and 2500 which adopt sequencing by synthesis strategy were selected because of their popularity. Coverage of ×10, ×25, ×50, and ×100 was chosen for data generation to check coverage effect on the number of correctly aligned reads on each chromosome. Fragment length and standard deviation were considered 175 and 51.85, namely. Using a fixed integer number for random seed parameter (–rndSeed) which stands for random seed, guaranteed identical short reads are generated for all different runs.[14] A 100 bp length for short reads was considered to simplify comparison. Using the SAM file obtained as an output of ART, simulated reads could be evaluated whether they were mapped correctly to the regions generated from.

Quality control

The first step in the processing of FASTQ files is quality control that can be performed using appropriate tools. A general preprocessing task in every NGS analysis pipeline is trimming the adapters and removing the low-quality bases (Q < 20) from two tails of short reads. A sample quality score diagram obtained by FaQCs tool[15] for one of the generated FASTQ files (Hiseq 1000, fold of coverage ×10) is illustrated in Figure 1. In contrast to traditional NGS analysis workflows and to be able to evaluate different NGS technologies, here, we did not perform any filtering/trimming on artificially generated tags.

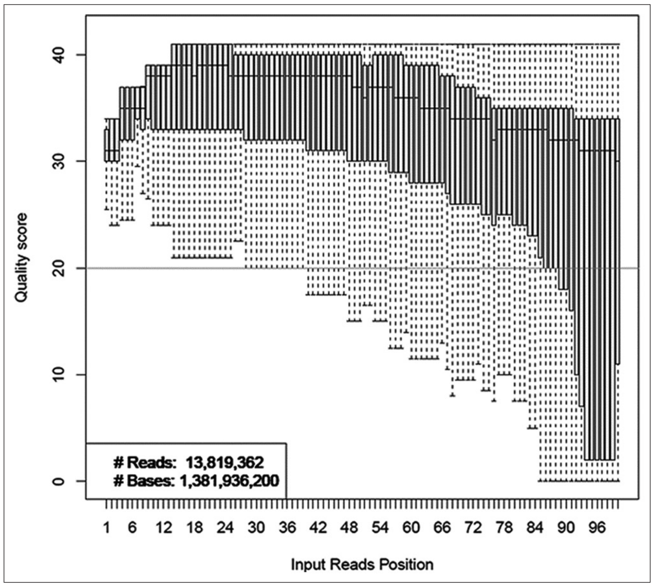

Figure 1: A sample quality score diagram. Quality boxplot obtained by FaQCs tool for the short reads simulated for Hiseq 1000 with fold of coverage ×10

Alignment

Algorithms applied for alignment fall into three major categories: hash table, array scanning, and Burrows–Wheeler Transform (BWT) backtracking. They create secondary data to facilitate mapping and to overcome the limitation in hardware resources. Hash table employs a seed-and-extend paradigm semi-similar to BLAST with spaced seeds which results in a table.[16] Two algorithms employed for seed alignment are Smith and Waterman[9] and Needleman and Wunsch[17] that are slow but accurate. The aforementioned combination of algorithms leads to an indel-sensitive alignment accepting a limited number of mismatches and gaps. In contrast to the hash table which localizes seeds in short reads and aligns them to subsequences of the reference genome, BWT backtracking[18] attempts mapping whole read to reference via compressing data structure of genome reference. As it seems to be a time-consuming process, suffixes of reference are built preliminary to speed the alignment up. Prefix/suffix tree,[19] suffix array,[20] and Ferragina-Manzini (FM) index[21] are different algorithms employed in BWT backtracking. No multiple identical reads shall be aligned to the same place and it is more advantageous than the hash table in some cases.[22]

Agrawal and Huang reported that using sequence-specific information in estimating pairwise statistical significance provides a more reliable measure to improve the alignment results rather than using database statistical significance estimates in calculating the alignment score.[23] Thus, aligners should be carefully chosen according to data being analyzed and using incompatible aligner may cause errors in secondary data and lead to mistaken alignment and misinterpretation. This study aims to align short reads without-SNV (with no single nucleotide variant) and SNV-imported ones to three kinds of references: WG reference (hg38), target regions of SSHAEv7 whose input short reads are generated from, and each chromosome separately. This helps to understand how different references affect mapping accuracy.

Two popular mappers known as Burrows-Wheeler Aligner (BWA)[20] and Bowtie2[24] utilizing the indexing method accept mismatches and gapped alignment. Bowtie2 utilizes BWT on FM-indexes of FASTA file, but for high-speed search through subsequences of reference, BWA[20] constructs suffix arrays. An academic version of Novoalign[3] mapping tool which adopts hash table performs optimized alignment for 30-300 bp reads. Stampy[25] is another alignment tool which uses a hybrid method to write a hash table and look it up to facilitate mapping. The divide and conquer strategy is employed in Kart[26] to create a faster alignment using BWT backtracking and hash table simultaneously. All the selected tools accessible for free and could perform paired-end alignment allowing mismatches, gaps, and indels. The five described mapping tools and their basic features are listed in Table 1. This study aims to assess the performance of each aligner in correct mapping on each chromosome using simulated short reads. All input parameters of different software were chosen based on the recommended default values in the utilized tools.

Performance evaluation

For every considered sequencing system and coverage folds, simulated short reads were mapped to the SSHAEv7 targets as the reference genome using BWA, Bowtie2, Kart, Novoalign, and Stampy. The aforementioned aligners were used to map 100 bp short reads compiled in FASTQ files of samples against the reference. Then, SAM files, which are the primary output of alignment tools, were converted to binary alignment mapping (BAM) format and sorted utilizing Sambamba tool.[27] Then, the sorted BAM was converted back to SAM format again to be comparable with the reference SAM. Then, the first, third, and fourth columns of the lastly sorted SAM files, which contained required information about the position of aligned reads that were extracted. To perform the evaluation, the extracted information (i.e., QNAME, RNAME and POS) was compared with the firstly produced (reference) SAM file by the ART from SSHAEv7 targets. The number of aligned short reads in each sorted SAM file was counted and compared to find exact matches of alignment. Finally, the number of unmapped tags and accuracy, which is defined as the ratio of properly aligned reads to the total number of short reads generated from each chromosome in the simulation step, was reported.

To discover the regions on the genome that are error prone in the alignment process, we defined a Normalized Matrix of Mapping Error (NMME) as Eq. 1. In this regard, we first calculated the Number of Misaligned Readss (NMR(i, j)) originated from chromosome i and erroneously aligned to

Table 1: List of selected mapping tools and their basic features

	BWA	Bowtie2	Kart	Stampy	Novoalign
Version	0.7.17	2.3.3.1	2.4.4	1.0.32	3.0802
Mapping algorithm	BWT	FM-Index and BWT	Hash table and BWT	Improved hash table and SIMD	Hash table
Multithreading	Yes	Yes	Yes	No	No
Optimized read length (bp)	4-200	4-5000k	150-7k	4-4k	30-300
Seed mismatches	Yes	Yes	Yes	(0.15 read length)	8
Indel	8	Yes	5	30	7
Gap	Yes	Yes	5	No	Yes
Alignment	Global	Global/local	Local	Global	Global
Mapping quality	0-60	0-42	0-60	0-99	0-70

FM – Ferragina-Manzini; BWT – Burrows-Wheeler Transform; SIMD – Single instruction, multiple data; BWA – Burrows-wheeler aligner

chromosome *j*. Afterward, *NMR (i, j)* is divided by the Total Length of Targeted Regions on Both Chromosomes i and j (T*LC (i, j))*. Then, the obtained values were divided into the maximum value of Misaligned Reads Ratio (MRR). MPR is the number of reads that originated from one chromosome or originated from multiple chromosomes and erroneously aligned to a common target chromosome and then divided by the length of targeted regions on that chromosome.

$$NMME\ (i, j) = (NMR\ (i, j)/TLC\ (i, j))/MPR \qquad (1)$$

Results

Mapping accuracy and average computational time are the two parameters to be satisfied, illustrating the basic performance of aligners. The first measure (Eq. 2) used to compute the efficiency of aligners in this study is based on how much coding regions of each whose based on CCDS, are correctly covered by the aligned short reads.

$$\text{Accuracy} = \frac{\text{Properly mapped reads}}{\text{Total number of short reads generated for each chromosome}} \qquad (2)$$

Figure 2 shows the accuracy of mapping for various aligners using data of Hiseq systems (10, 20, and 25) with coverage of × 10. It is obvious that by optimization of sequencing chemistry in developing newer technologies, the mapping accuracy was significantly improved. The results were similarly replicated for different depth of coverage (×10, ×25, ×50, and ×100). Based on the results shown in Figure 2, Stampy showed the best performance among five evaluated aligner tools regarding accuracy for all technologies and it could map almost all the simulated reads to the proper position in all chromosomes for Hiseq 25. BWA and Bowtie2 could correctly map about 60% of the short reads generated from Hiseq 20 and 25 where the coding genome of humans (CCDS-hg19) used as a reference. It is basically because of BWT-based algorithm applied in these aligners.[28] Furthermore, the academic version of Novoalign showed the lowest accuracy and could not correctly map more than 40% of short reads even for Hiseq 2500. It agrees with the Shang's study that states that Novoalign has low performance due to its overmapping at both ends of short reads.[29]

Based on the primary results demonstrated by Figure 2, which obtained for a limited mapping reference customized by CCDS on hg19, maximum and minimum mapping accuracies were observed on chromosome 19 and Y obtained by Stampy and Novoalign, respectively.

According to the average of mapping accuracy for all technologies and all examined fold of coverage, Stampy, Bowtie2, and BWA were selected as the best mappers which are able to cover >50% of short reads on autosomal chromosomes. It should be noted that there was no significant difference in accuracies between the four experiments using different folds of coverage (×10, ×25, ×50, and ×100) for each aligner.

Runtime is another important factor for comparison of the mapping tool, but speaking of other performance measures according to the results of different algorithms, one should compromise. Alignment time, which is the duration time between the time of execution of aligner tool command and the end time of generating the SAM output, usually depends on the reference genome size and the number of reads and their length.[3] These parameters were optimally chosen and remained constant in our study, and thus, the measured runtime would only depend on the aligner performance. Figure 3 shows the average runtime of each aligner on ×10 reads. According to Figure 3, Kart is the fastest tool, followed by BWA and Bowtie2. Stampy was the worst tool in terms of computational speed contrary to its superiority in mapping accuracy. The results approve the literature,[20,30] which expresses acceptable time consumption and memory occupancy (overall computational efficiency) and confirms that multithreading improves computational efficiency.

For further evaluation of the best aligners introduced in the previous experiment (Stampy, BWA, and Bowtie2), two versions of short reads containing without-SNV and SNV-imported simulated reads were mapped to three types of references: WG (hg38), CTBR (SSHAEv7-based reference), and each chromosome individually. Table 2 reports mapping error rate, which is the ratio of total number of misaligned short reads to the whole simulated short reads (1,315,770 reads), for three mappers evaluated on two

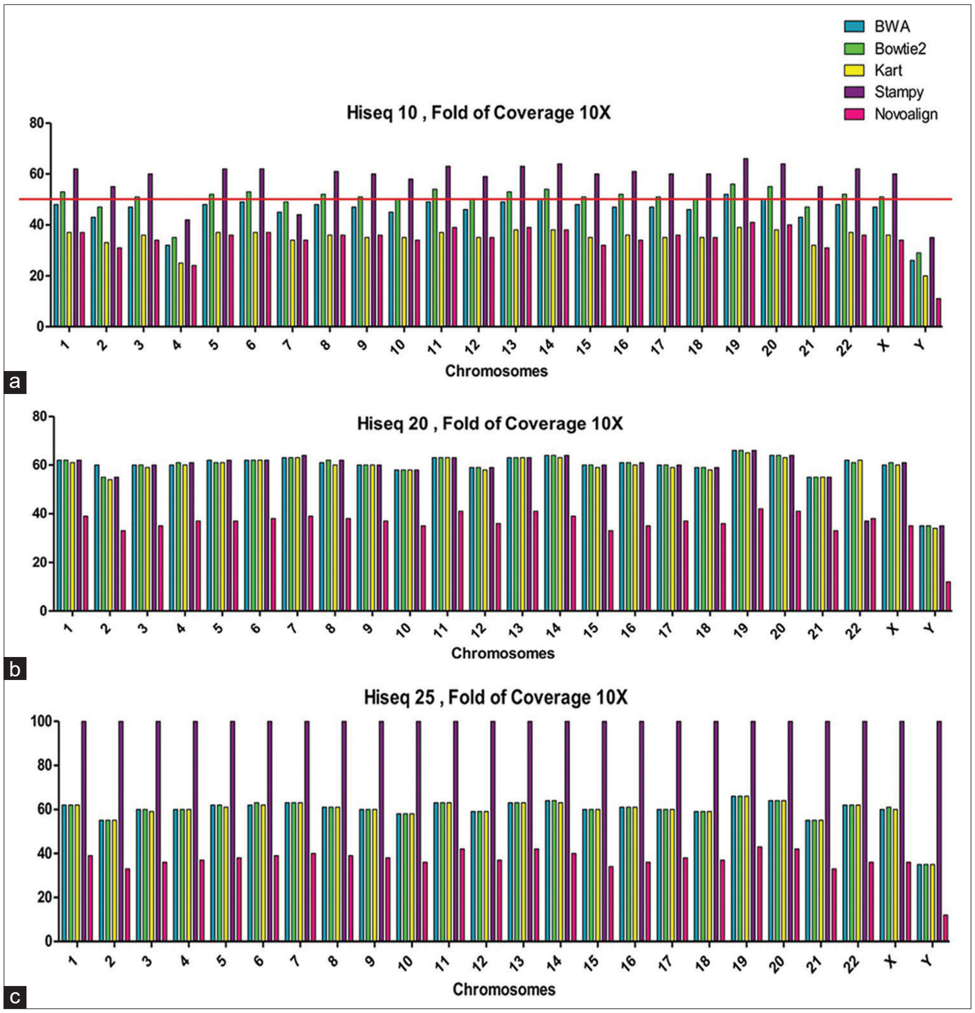

Figure 2: Mapping accuracy for various aligners using data of Hiseq systems ((a) 10,(b) 20, and (c) 25) with coverage of × 10

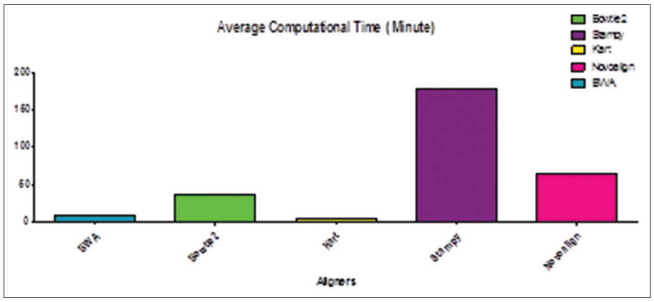

Figure 3: Average runtime in minutes for mapping ×10 reads using five selected aligners

Table 2: Mapping error rate (%) of different aligners in alignment of with and without mutation short reads using three types of references

	Without SNV		SNV imported		
	WG	CTBR	WG	CTBR	Chr[#]
BWA	3.60	1.31	4.2	1.84	> 30
Bowtie2	3.60	1.29	5.37	3.04	> 30
Stampy	3.35	1.30	3.78	1.67	> 20

[#]Reported value is the minimum error rate observed in the set of all 24 chromosomes individually (chr1-chr22, chrX and chrY). WG – Whole genome; CTBR – Customized target-based reference; SIMD – Single instruction, multiple data; BWA – Burrows-wheeler aligner

types of data (without SNV and SNV imported) using three references (WG, CTBR, and Chr#). It is obvious from the last column of Table 2, using single chromosomes as the reference led to a high error rate (minimum error of 20%) for all tools. Thus, limiting the reference to a target whenever short reads came from a wider region results in increasing the false-positive alignment error.

According to Table 2, all mapping tools reached the minimum error rate (1.3 ± 0.01%), where the simulated short reads did not contain intentional mutations and CTBR was used as the reference [third column of Table 2]. Whereas in the presence of mutations in FASTQ data (SNV-imported data), Stampy showed the best result (1.67%) using CTBR as reference. For SNV-imported data, Bowtie2 is the worst tool considering the error rate of 5.37% and 3.04% using WG and CTBR as reference, respectively. Regardless of selected mapping tools and the type of input data, choosing CTBR rather than WG as the reference for mapping the simulated short reads generated from coding regions of human genome led to more than 2% improvement in the mapping accuracy.

Figure 4 shows the scheme of normalized mapping error (Eq. 1) in different chromosomes for various aligners in which without-SNV data aligned to WG. However, a similar pattern was observed for all other conditions [as considered in Table 2] which is the same as the scheme obtained for different mapping tools [Figure 4]. Thus, for all tools in all conditions, maximum and minimum of normalized mapping error were observed for chromosome Y and 20, respectively. According to the observable points in the main diagonal of the displayed matrixes in Figure 4, the main part of misaligned reads is mispositioned on the right chromosome. Therefore, it can be concluded that the source of the most part of the measured error is the misalignment between neighbor genes/regions. It should be noted that the observed error is independent of the number of short reads generated from each chromosome.

To see the pattern of misaligned reads between different chromosomes, namely Cross-Chromosome Mapping Error (CCME) that is unobservable in Figure 4, the diagonal elements of NMME were removed and other elements were renormalized to make too small non-diagonal values observable. Due to the high misalignment error of chromosome Y to other chromosomes, we removed the last row and column of NMME to make other renormalized CCMEs observable as shown in Figure 5. According to Figure 5, using WG as a reference [first and third columns in Figure 5], there is a high tendency of mapping the reads originated from chromosome 16 to chromosome 1 using all mapping tools. Another common pattern among different aligners is the misalignment of reads from chromosomes 6 and 14 to chromosomes 7 and 22. Using CTBR as a reference, there is no evidence of mapping error between chromosomes 16 and 1, which was the main source of error using WG. However, CCME between chromosomes 6 and 7 is the most significant error pattern that remains in both cases of using WG and CTBR by all tools.

Discussion

Due to a variety of abounding incomplete bioinformatic tools that demonstrated inconsistent performance in different applications, it is necessary to have a framework for evaluating different tools to optimize an imperfect analytical pipeline. This work presents a straightforward approach for evaluation and selection of bioinformatic tools in a clinical application using simulated data. The results showed that limiting the mapping reference from WG to a customized one, considering the genomic region selected as the target of study, may lead to the improvement of mapping performance, while narrowing the reference to one chromosome, when the input data includes all exonic regions, caused high error rates in the alignment.

Providing a fair comparison among different mapping tools and demonstration of chromosome-specific pattern of alignment error, in this study, may provide an illustrative framework for designing a careful targeted sequencing study. The results of such a simulation can be used to construct a new quality measure, such as a mapping quality for different genomic regions according to this work; which guides us to focus on more reliable results for interpretation and ignore the low-quality ones. Based on the obtained results, independent sequencing of neighbor genes in different experiments may lead to a reduction in the alignment error, which improves the final performance of an NGS study. The main limitation of this study is relying on simulated data. In our future work, we aim to run the proposed framework for evaluating a list of specific genes on real datasets demonstrating misaligned reads between different genes.

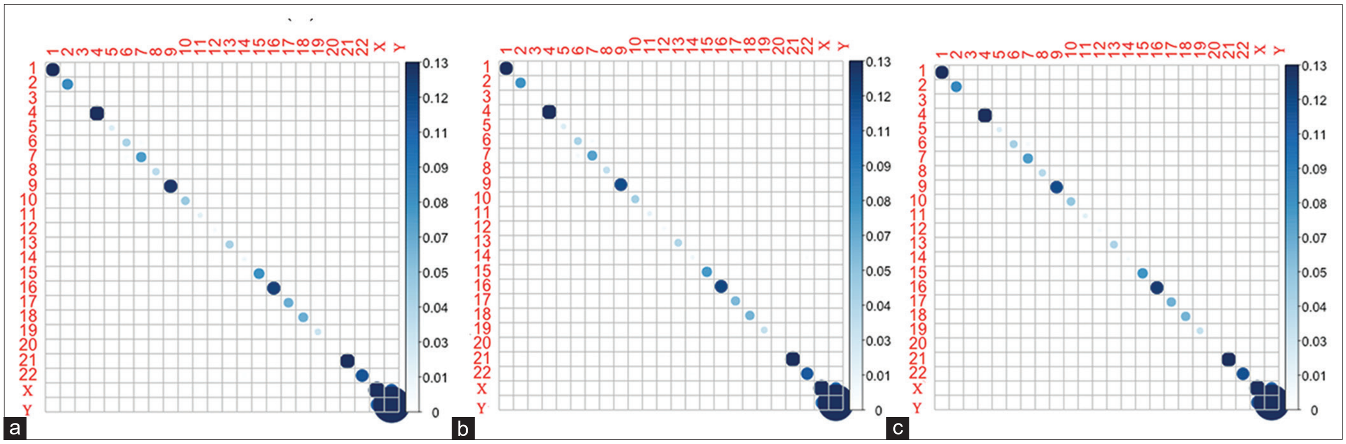

Figure 4: Scheme of normalized mapping error in different chromosomes using SNV-imported data and customized target-based reference as reference for (a) BWA, (b) Bowtie2, and (c) Stampy

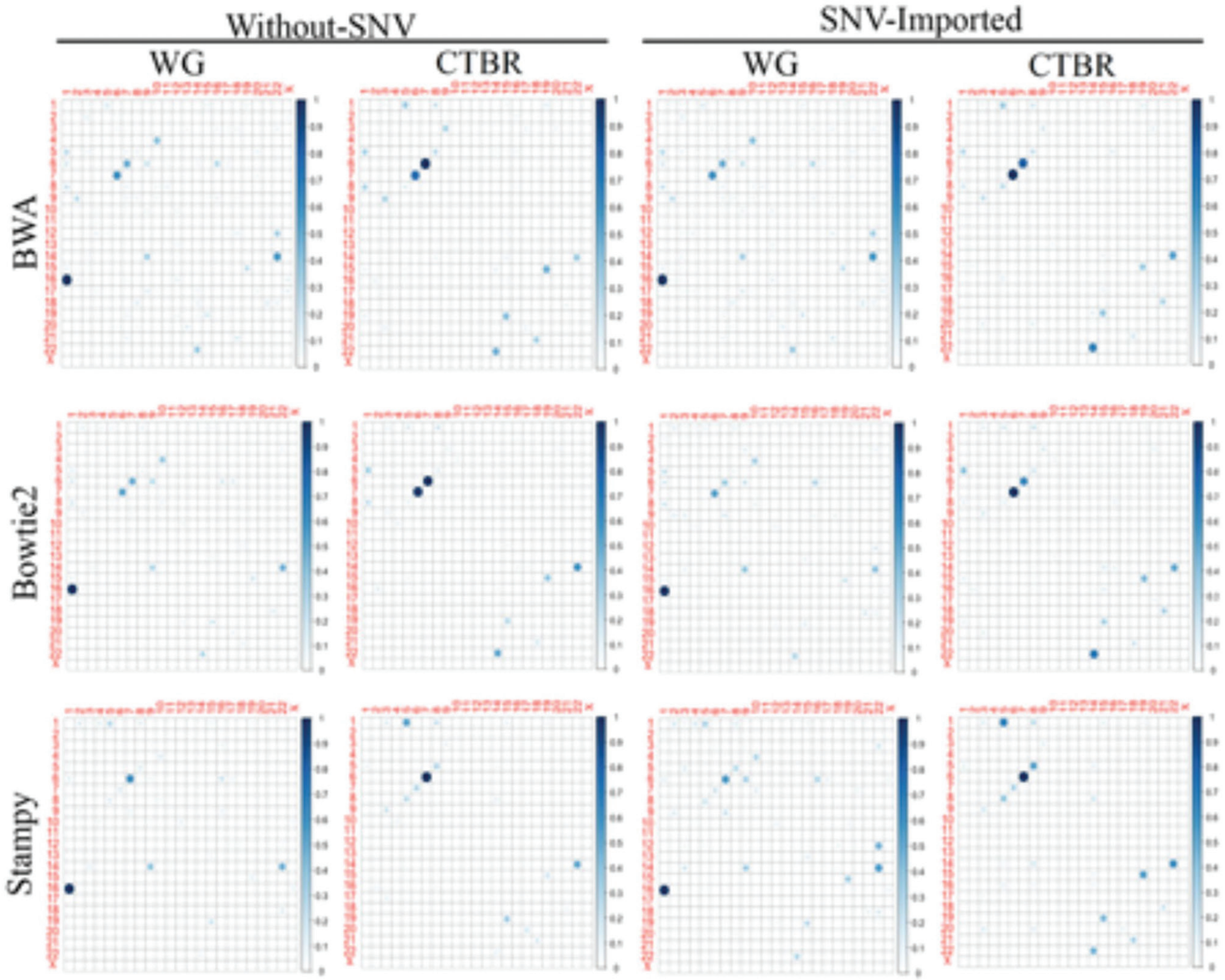

Figure 5: Scheme of normalized mapping error between different chromosomes, cross-chromosome mapping error, in all evaluated conditions using all tools

Acknowledgment

This research was financially supported by the Isfahan University of Medical Sciences, Isfahan, Iran (Grant no. 396282).

Conflicts of interest

There are no conflicts of interest.

References

1. Goh G, Choi M. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. Genomics Inform 2012;10:214-9.
2. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet 2017;18:473-84.
3. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics 2017;109:186-91.
4. Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet 2016;17:459-69.
5. Lee H, Lee KW, Lee T, Park D, Chung J, Lee C, *et al*. Performance evaluation method for read mapping tool in clinical panel sequencing. Genes Genomics 2018;40:189-97.
6. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, *et al*. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015;17:251-64.
7. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, *et al*. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 2015;372:2243-57.
8. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol 2009;27:455-7.
9. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195-7.
10. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 2012;20:490-7.
11. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, *et al*. GENCODE: The reference human

genome annotation for The ENCODE Project. Genome Res 2012;22:1760-74.

12. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, *et al*. Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res 2018;46:D221-8.
13. Chen R, Im H, Snyder M. Whole-exome enrichment with the agilent sure select human all exon platform. Cold Spring Harb Protoc 2015;2015:626-33.
14. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics 2012;28:593-4.
15. Lo CC, Chain PS. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics 2014;15:366.
16. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet 2016;57:71-9.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443-53.
18. Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 2011;27:2790-6.
19. Rheinländer A, Knobloch M, Hochmuth N, Leser U. Prefix Tree Indexing for Similarity Search and Similarity Joins on Genomic Data; 2010. p. 519-36.
20. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754-60.
21. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
22. Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 2014;9:e90581.
23. Agrawal A, Huang X. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM Trans Comput Biol Bioinform 2011;8:194-205.
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie2. Nat Methods 2012;9:357-9.
25. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 2011;21:936-9.
26. Lin HN, Hsu WL. Kart: A divide-and-conquer algorithm for NGS read alignment. Bioinformatics 2017;33:2281-7.
27. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics 2015;31:2032-4.
28. Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE. Comparing reference-based RNA-seq mapping methods for non-human primate data. BMC Genomics 2014;15:570.
29. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int 2014;2014.
30. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ. Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 2011;56:406-14.

BIOGRAPHIES

Hannane Mohammadi Nodehi is a PhD student in Bioelectrics and Biomedical Engineering at School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences.

Email: hannane.mohammadi@gmail.com

Mohammad Amin Tabatabaiefar received his Ph.D. degree in Medical Genetics in 2010 from Tehran University of Medical Sciences. He is currently an Associate professor at the Isfahan University of Medical Sciences. His research interests are focused on gene mapping of genetic diseases and analysis of genetics variants.

Email: tabatabaiefar@med.mui.ac.ir

Mohammadreza Sehhati received his Ph.D. degree in Biomedical Engineering in 2015 from Isfahan University of Medical Sciences. He/She is currently an assistant professor at the Isfahan University of Medical Sciences. His research interests are focused on Biological Data Mining and constructing predictive models with medical application using machine learning techniques.

Email: mr.sehhati@gmail.com